Министерство образования Российской Федерации

Государственное образовательное учреждение
высшего профессионального образования
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ УПРАВЛЕНИЯ

Институт информационных систем управления

О д о б р е н о
Президиумом НМС ГУУ

В. Н. Калинина
кандидат технических наук, профессор

В. И. Соловьев
кандидат экономических наук, доцент

Введение в многомерный статистический анализ

Учебное пособие
для студентов всех специальностей

Москва — 2003

ББК 22.172я7
УДК 519.237 (075.8)
6Н1

К17 **Калинина В. Н., Соловьев В. И.** Введение в многомерный статистический анализ: Учебное пособие / ГУУ. – М., 2003. – 66 с.

Излагаются теоретические основы и алгоритмы методов многомерного статистического анализа. Рассмотрены два метода снижения размерности многомерного пространства (метод главных компонент и факторный анализ) и два метода классификации (кластерный и дискриминантный анализ). Изложение иллюстрируется решением практических задач, в том числе, с помощью современных пакетов прикладных программ. Приводятся задачи для самостоятельного решения.

Для студентов экономических специальностей. Может быть полезно аспирантам, преподавателям, научным сотрудникам, специалистам-практикам, интересующимся применением многомерных статистических методов в экономике.

Библиогр. 23 назв. Табл. 15. Ил. 11.

Ответственный редактор

заведующий кафедрой прикладной математики,
доктор экономических наук, профессор
В. А. КОЛЕМАЕВ

Рецензенты

д-р физ.-мат. наук, профессор В. В. ШЕВЕЛЕВ (МИТХТ)

канд. экон. наук Б. Г. МИХАЛЕВ (ЗАО «Баркли Строй»)

ISBN 5-215-01514-7

Введение

Социальные и экономические объекты, как правило, характеризуются достаточно большим числом параметров, образующих многомерные векторы, и особое значение в экономических и социальных исследованиях приобретают задачи изучения взаимосвязей между компонентами этих векторов, причем эти взаимосвязи необходимо выявлять на основании ограниченного числа многомерных наблюдений.

Многомерным статистическим анализом называется раздел математической статистики, изучающий методы сбора и обработки многомерных статистических данных, их систематизации и обработки с целью выявления характера и структуры взаимосвязей между компонентами исследуемого многомерного признака, получения практических выводов.

Предположим, что рассматривается некоторая совокупность, состоящая из n стран, для каждой из которых известны макроэкономические показатели: X_1 — валовой внутренний продукт, X_2 — площадь территории, X_3 — средняя продолжительность жизни населения и т. п. В результате получен набор из n наблюдений над k-мерным случайным вектором $\mathbf{X} = (X_1, X_2, \ldots, X_k)^{\mathrm{T}}$.

Похожая ситуация возникает, когда изучается совокупность предприятий, для каждого из которых рассчитаны показатели X_1 — валовая прибыль, X_2 — численность работников, X_3 — стоимость основных производственных фондов и т. п.

Типичные з а д а ч и, которые можно решить методами многомерного статистического анализа, таковы:

- по наблюдавшимся значениям случайного вектора $\mathbf{X} = (X_1, X_2, \ldots, X_k)^{\mathrm{T}}$ может понадобиться изучить связь между его компонентами $X_1, X_2, \ldots, X_k$;
- может понадобиться определить, какие из (большого числа) рассчитанных показателей $X_1, X_2, \ldots, X_k$ в наибольшей степени влияют на валовой внутренний продукт или на продолжительность жизни населения;
- может понадобиться классифицировать страны по какому-либо признаку.

Отметим, что способы сбора данных могут различаться. Так, если исследуется мировая экономика, то естественно взять в качестве объектов, на которых наблюдаются значения вектора **X**, страны, если же изучается национальная экономическая система, то естественно наблюдать значения вектора **X** на одной и той же (интересующей исследователя) стране в различные моменты времени.

Такие статистические методы, как множественный корреляционный и регрессионный анализ, традиционно изучаются в курсах теории вероятностей и математической статистики [3, 14], рассмотрению прикладных аспектов регрессионного анализа посвящена дисциплина «Эконометрика» [3, 18].

Другим методам исследования многомерных генеральных совокупностей на основании статистических данных посвящено данное пособие.

Методы снижения размерности многомерного пространства позволяют без существенной потери информации перейти от первоначальной системы большого числа наблюдаемых взаимосвязанных факторов к системе существенно меньшего числа скрытых (ненаблюдаемых) факторов, определяющих вариацию первоначальных признаков. В первой главе описываются методы компонентного и факторного анализа, с использованием которых можно выявлять объективно существующие, но непосредственно не наблюдаемые закономерности при помощи главных компонент или факторов.

Методы многомерной классификации предназначены для разделения совокупностей объектов (характеризующиеся большим числом признаков) на классы, в каждый из которых должны входить объекты, в определенном смысле однородные или близкие. Такую классификацию на основании статистических данных о значениях признаков на объектах можно провести методами кластерного и дискриминантного анализа, рассматриваемыми во второй главе.

Развитие вычислительной техники и программного обеспечения способствует широкому внедрению методов многомерного статистического анализа в практику. Пакеты прикладных программ с удобным пользовательским интерфейсом, такие как SPSS, Statistica, SAS и др., снимают трудности в применении указанных методов, заключающиеся в сложности математического аппарата, опирающегося на линейную алгебру, теорию вероятностей и математическую статистику, и громоздкости вычислений.

Однако применение программ без понимания математической сущности используемых алгоритмов способствует развитию у исследователя иллюзии простоты применения многомерных статистических методов, что может привести к неверным или необоснованным результатам. Значимые практические результаты могут быть получены только на основе профессиональных знаний в предметной области, подкрепленных владением математическими методами и пакетами прикладных программ, в которых эти методы реализованы.

Поэтому для каждого из рассматриваемых в данной книге методов приводятся основные теоретические сведения, в том числе алгоритмы; обсуждается реализация этих методов и алгоритмов в пакетах прикладных программ. Рассматриваемые методы иллюстрируются примерами их практического применения в экономике с использованием пакета SPSS.

Пособие написано на основе опыта чтения курса «Многомерные статистические методы» студентам Государственного университета управления. Для более подробного изучения методов прикладного многомерного статистического анализа рекомендуются книги [1 ~ 4, 8].

Предполагается, что читатель хорошо знаком с курсами линейной алгебры (например, в объеме учебника [10] и приложения к учебнику [18]), теории вероятностей и математической статистики (например, в объеме учебника [14]).

ГЛАВА 1. МЕТОДЫ СНИЖЕНИЯ РАЗМЕРНОСТИ МНОГОМЕРНОГО ПРОСТРАНСТВА

§ 1.1. СУЩНОСТЬ ЗАДАЧ СНИЖЕНИЯ РАЗМЕРНОСТИ

Во многих практических задачах исследователя интересуют главным образом признаки, которые обнаруживают наибольшую изменчивость (т. е. разброс, дисперсию) при переходе от одного объекта к другому, при этом такие признаки часто невозможно наблюдать непосредственно на объектах.

Приведем несколько примеров:

- при индивидуальном пошиве одежды портной замеряет на клиенте от восьми до одиннадцати различных параметров (рост, размах рук, длину предплечья, длину ног, окружности груди, бедер, талии и др.); при массовом производстве одежды ее размеры характеризуются всего двумя признаками: ростом и размером, являющимися производными от указанных параметров, и в большинстве случаев указание размера и роста при покупке одежды приводит к удовлетворительному выбору;
- склонность населения к миграции определяется по данным о достаточно большом числе социально-экономических, демографических, географических и др. показателей и результатам социологических опросов;
- только большая совокупность непосредственно измеряемых признаков позволяет сравнивать страны, регионы и города по уровню жизни, продукцию различных производителей — по качеству и т. п.

Приведенные примеры иллюстрируют **сущность задач снижения размерности** многомерного пространства, которая заключается в выражении большого числа исходных факторов, непосредственно измеренных на объектах, через меньшее (как правило, намного меньшее) число более емких, максимально информативных, но непосредственно не наблюдаемых внутренних характеристик объектов. При этом предполагается, что более емкие признаки будут отражать наиболее существенные свойства объектов.

Целью методов снижения размерности является исследование внутренней структуры изучаемой системы k случайных величин, «сжатие» этой системы без существенной потери содержащейся в ней информации путем выявления небольшого числа факторов, объясняющих изменчивость и взаимосвязи исходных случайных величин.

Метод главных компонент выявляет k компонент — факторов, объясняющих всю дисперсию и корреляции исходных k случайных величин; при этом компоненты строятся в порядке убывания объясняемой ими доли суммарной дисперсии исходных величин, что позволяет зачастую ограничиться несколькими первыми компонентами. Факторный анализ выявляет m ($m < k$) общих для всех исходных величин факторов, объясняя оставшуюся после этого дисперсию влиянием специфических факторов.

Среди **прикладных задач**, решаемых указанными методами, отметим следующие:

- поиск скрытых, но объективно существующих взаимосвязей между экономическими и социальными показателями, проверка гипотез о взаимосвязях этих показателей, выявление природы различий между объектами;
- описание изучаемой системы числом признаков, значительно меньшим числа исходных факторов, при этом выявленные факторы или главные компоненты содержат в среднем больше информации, чем непосредственно зафиксированные на объектах значения исходных факторов;
- построение обобщенных экономических и социальных показателей, таких как качество продукции, размер предприятия, интенсивность ведения хозяйства и т. п.;
- визуализация исходных многомерных наблюдений путем их проецирования на специально подобранную прямую, плоскость или трехмерное пространство;
- построение регрессионных моделей по главным компонентам; в социальных и экономических задачах исходные факторы часто обладают *мультиколлинеарностью* (т. е. являются коррелированными между собой), что затрудняет построение и интерпретацию регрессионных моделей, не позволяя часто получать сколь-нибудь точные прогнозы; главные компоненты, сохраняя всю информацию об изучаемых объектах, являются некоррелированными по построению;
- классификация по обобщенным экономическим показателям; практика показывает, что классификация объектов, проведенная по факторам или по главным компонентам, оказывается более объективной, чем классификация тех же объектов по исходным признакам; по одному — трем факторам или главным компонентам возможно проведение визуальной классификации, в случае большей размерности пространства обобщенных показателей, полученного в результате компонентного или факторного анализа, необходимо привлечение методов многомерной классификации, рассматриваемых во второй главе пособия;
- сжатие исходной информации, значительное уменьшение объемов информации, хранимой в базах данных, без существенных потерь в информативности.

§ 1.2. Компонентный анализ

1.2.1. Математическая модель главных компонент

Метод главных компонент состоит в разложении (с помощью ортогонального преобразования) k-мерного случайного вектора

$$\mathbf{X} = (X_1, X_2, \ldots, X_k)^{\mathrm{T}}$$

по системе линейно независимых векторов, в качестве которой выбирается ортонормированная система собственных векторов, отвечающих собственным значениям ковариационной матрицы вектора $\mathbf{X}$.

Линейная модель главных компонент для центрированного вектора-столбца

$$\overset{\circ}{\mathbf{X}} = \mathbf{X} - \mathbf{M}\mathbf{X}$$

записывается в виде

$$\overset{\circ}{\mathbf{X}} = \mathbf{A}\mathbf{F}, \qquad (1.2.1)$$

где

$$\mathbf{F} = (F_1, F_2, \ldots, F_k)^{\mathrm{T}} —$$

центрированный и нормированный случайный вектор-столбец некоррелированных *главных компонент* F_j $(j = 1, 2, \ldots, k)$,

$$\mathbf{A} = (a_{ij}) \in \mathbb{R}^{k \times k} —$$

неслучайная *матрица нагрузок* случайных величин X_i на компоненты F_j $(i = 1, 2, \ldots, k,\ j = 1, 2, \ldots, k)$.

Изложим **алгоритм** построения вектора **F** и расчета матрицы **A**.

Пусть

$$\mathbf{\Sigma} = \mathbf{M}(\overset{\circ}{\mathbf{X}}\overset{\circ}{\mathbf{X}}^{\mathrm{T}}) —$$

ковариационная матрица вектора **X**. Будучи симметричной и неотрицательно определенной, она имеет k вещественных неотрицательных собственных значений $\lambda_1, \lambda_2, \ldots, \lambda_k$. Предположим, что

$$\lambda_1 > \lambda_2 > \cdots > \lambda_k,$$

как и бывает обычно в большинстве приложений компонентного анализа.

Обозначим

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{pmatrix},$$

$$\mathbf{v}_j = (v_{1j}, v_{2j}, \ldots, v_{kj})^{\mathrm{T}} —$$

нормированные собственные векторы-столбцы матрицы $\mathbf{\Sigma}$, соответствующие собственным значениям λ_j $(j = 1, 2, \ldots, k)$. Тогда для всех $j = 1, 2, \ldots, k$ справедливы следующие равенства:

$$\det |\mathbf{\Sigma} - \lambda_j \mathbf{I}| = 0 \qquad (j = 1, 2, \ldots, k),$$

где **I** — единичная матрица порядка k;

$$\mathbf{\Sigma v}_j = \lambda_j \mathbf{v}_j \qquad (j = 1, 2, \ldots, k); \tag{1.2.2}$$

$$\mathbf{v}_p^{\mathrm{T}} \mathbf{v}_j = \sum_{i=1}^{k} v_{ip} v_{ij} = \delta_{pj} = \begin{cases} 1, & p = j, \\ 0, & p \neq j \end{cases} \tag{1.2.3}$$
$$(p = 1, 2, \ldots, k, \quad j = 1, 2, \ldots, k).$$

Введем матрицу

$$\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k).$$

Так как с учетом соотношений (1.2.2) и (1.2.3)

$$\mathbf{v}_j^{\mathrm{T}} \mathbf{\Sigma v}_p = \lambda_j \mathbf{v}_j^{\mathrm{T}} \mathbf{v}_p = \begin{cases} \lambda_j, & p = j, \\ 0, & p \neq j \end{cases}$$
$$(p = 1, 2, ..., k, \quad j = 1, 2, ..., k),$$

то

$$\mathbf{V}^{\mathrm{T}} \mathbf{\Sigma V} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{pmatrix} = \mathbf{\Lambda}. \tag{1.2.4}$$

Пусть

$$\overset{\circ}{\mathbf{F}} = \mathbf{V}^{\mathrm{T}} \overset{\circ}{\mathbf{X}}; \tag{1.2.5}$$

при этом, так как

$$\mathbf{M}\overset{\circ}{\mathbf{F}} = \mathbf{M}(\mathbf{V}^{\mathrm{T}} \overset{\circ}{\mathbf{X}}) = \mathbf{V}^{\mathrm{T}} \mathbf{M}\overset{\circ}{\mathbf{X}} = \mathbf{0}$$

(где $\mathbf{0} = (0, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^k$), то $\overset{\circ}{\mathbf{F}}$ — центрированный вектор, а поскольку

$$\mathbf{M}(\overset{\circ}{\mathbf{F}}\overset{\circ}{\mathbf{F}}^{\mathrm{T}}) = \mathbf{M}(\mathbf{V}^{\mathrm{T}} \overset{\circ}{\mathbf{X}}\overset{\circ}{\mathbf{X}}^{\mathrm{T}} \mathbf{V}) = \mathbf{V}^{\mathrm{T}} \mathbf{M}(\overset{\circ}{\mathbf{X}}\overset{\circ}{\mathbf{X}}^{\mathrm{T}}) \mathbf{V} = \mathbf{V}^{\mathrm{T}} \mathbf{\Sigma V},$$

то в силу (1.2.4) компоненты вектора $\overset{\circ}{\mathbf{F}}$ некоррелированы и

$$\mathbf{D}\overset{\circ}{F}_j = \lambda_j \qquad (j = 1, 2, \ldots, k).$$

Поэтому искомый центрированный и нормированный вектор $\mathbf{F}$ равен

$$\mathbf{F} = \mathbf{\Lambda}^{-\frac{1}{2}} \overset{\circ}{\mathbf{F}} = \mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{V}^{\mathrm{T}} \overset{\circ}{\mathbf{X}}. \tag{1.2.6}$$

Обратим внимание на следующие факты:

- так как след матрицы $\mathbf{\Sigma}$

$$\operatorname{tr} \mathbf{\Sigma} = \operatorname{tr} \mathbf{\Lambda},$$

то

$$\sum_{i=1}^{k} \mathbf{D}\overset{\circ}{X}_i = \sum_{i=1}^{k} \mathbf{D}X_i = \operatorname{tr}\boldsymbol{\Sigma} = \operatorname{tr}\boldsymbol{\Lambda} = \sum_{j=1}^{k} \lambda_j = \sum_{j=1}^{k} \mathbf{D}\overset{\circ}{F}_j , \qquad (1.2.7)$$

т. е. дисперсия исходных случайных величин $X_1, X_2, \ldots, X_k$ полностью исчерпывается дисперсией компонент $\overset{\circ}{F}_1, \overset{\circ}{F}_2, ..., \overset{\circ}{F}_k$; при этом, поскольку

$$\mathbf{D}\overset{\circ}{F}_1 > \mathbf{D}\overset{\circ}{F}_2 > \cdots > \mathbf{D}\overset{\circ}{F}_k ,$$

то дисперсией каждой следующей компоненты объясняется меньшая доля дисперсии исходных случайных величин, чем дисперсией предыдущей компоненты;

- так как

$$\mathbf{M}(\mathbf{F}^{\mathrm{T}}\mathbf{F}) = \mathbf{I},$$

то

$$\boldsymbol{\Sigma} = \mathbf{M}(\overset{\circ}{\mathbf{X}}\overset{\circ}{\mathbf{X}}{}^{\mathrm{T}}) = \mathbf{M}(\mathbf{A}\mathbf{F}^{\mathrm{T}}\mathbf{F}\mathbf{A}^{\mathrm{T}}) = \mathbf{A}\mathbf{M}(\mathbf{F}^{\mathrm{T}}\mathbf{F})\mathbf{A}^{\mathrm{T}} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$$

или

$$\operatorname{cov}(X_i, X_p) = \operatorname{cov}(\overset{\circ}{X}_i, \overset{\circ}{X}_p) = \sum_{j=1}^{k} a_{ij}a_{pj} \quad (i = 1, 2, ..., k, \;\; p = 1, 2, ..., k), \qquad (1.2.8)$$

в частности,

$$\mathbf{D}\overset{\circ}{X}_i = \mathbf{D}X_i = \sum_{j=1}^{k} a_{ij}^2 \qquad (i = 1, 2, ..., k),$$

т. е. ковариационная матрица вектора **X** полностью воспроизводится матрицей нагрузок **A**;

- так как

$$\mathbf{M}(\overset{\circ}{\mathbf{X}}\mathbf{F}^{\mathrm{T}}) = \mathbf{M}(\mathbf{A}\mathbf{F}\mathbf{F}^{\mathrm{T}}) = \mathbf{A}\mathbf{M}(\mathbf{F}\mathbf{F}^{\mathrm{T}}) = \mathbf{A},$$

то

$$\operatorname{cov}(X_i, F_j) = a_{ij} \quad (i = 1, 2, \ldots, k, \quad j = 1, 2, \ldots, k), \qquad (1.2.9)$$

т. е. ковариация случайной величины X_i и компоненты F_j равна нагрузке a_{ij}.

З а м е ч а н и е . Собственные значения и собственные векторы существенно зависят от выбора масштаба и единиц измерения случайных величин. Поэтому компонентный анализ эффективен, когда величины имеют одинаковую содержательную природу и измерены в одних и тех же единицах. К примерам таких величин можно отнести структуру бюджета времени индивидуумов или организаций (все X_i измеряются в единицах времени), структуру потребления семей, структуру затрат организаций (все X_i измеряются в денежных единицах) и т. п. При нарушении указанного ус-

ловия вектор **X** нормируют и центрируют, тогда **Σ** — это корреляционная матрица, и из соотношений (1.2.7) ~ (1.2.9) следует, что

$$k = \sum_{j=1}^{k} \lambda_j ,$$

т. е.

$$\frac{\lambda_j}{k} —$$

это доля суммарной дисперсии случайных величин $X_1, X_2, \ldots, X_k$, объясняемая компонентой F_j;

$$\rho(X_i, X_p) = \sum_{j=1}^{k} a_{ij} a_{pj} ;$$

$$\rho(X_i, F_j) = a_{ij} ;$$

$$\sum_{j=1}^{k} a_{ij}^2 = 1 .$$

Отметим, что использование в компонентном анализе корреляционной матрицы затрудняет проверку ряда гипотез.

Найдем матрицу нагрузок **A**. Из соотношения (1.2.5), используя ортогональность матрицы **V**, получим:

$$\mathbf{V}\overset{\circ}{\mathbf{F}} = \mathbf{V}\mathbf{V}^{\mathrm{T}}\overset{\circ}{\mathbf{X}} = \mathbf{V}\mathbf{V}^{-1}\overset{\circ}{\mathbf{X}} = \overset{\circ}{\mathbf{X}},$$

и с учетом соотношения (1.2.6):

$$\overset{\circ}{\mathbf{X}} = \mathbf{V}\overset{\circ}{\mathbf{F}} = \mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{F} .$$

Отсюда

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}},$$

$$F_j = \frac{\sum_{i=1}^{k} a_{ij}\overset{\circ}{X}_i}{\lambda_j} = \frac{\sum_{i=1}^{k} v_{ij}\overset{\circ}{X}_i}{\sqrt{\lambda_j}} \quad (j = 1, 2, ..., k). \qquad (1.2.10)$$

Как правило, для анализа используют k' первых главных компонент, которыми исчерпывается не менее 70% дисперсии исходных случайных величин ($k' < k$). Можно доказать (см., например, [1, 19]), что с помощью компонент $F_1, F_2, \ldots, F_{k'}$ достигается наилучший, в смысле метода наименьших квадратов, прогноз величин $X_1, X_2, \ldots, X_k$ среди всех прогнозов, которые можно построить с помощью k' линейных комбинаций набора из k произвольных величин (это свойство называется *свойством наилучшей самовоспроизводимости*), при этом относительная ошибка прогноза составляет

$$\delta = \frac{\sum_{i=k'+1}^{k} \lambda_i}{\sum_{i=1}^{k} \lambda_i}.$$

На практике обычно

$$\frac{k'}{k} \approx 10 \div 25\%.$$

Для наглядной интерпретации главных компонент наиболее удобны случаи $k' = 1$, 2 и 3.

1.2.2. Геометрическая интерпретация метода главных компонент

Метод главных компонент допускает следующую геометрическую интерпретацию:

- вначале (при переходе от исходного вектора **X** к центрированному вектору $\overset{\circ}{\mathbf{X}} = \mathbf{X} - \mathbf{MX}$) фактически производится перенос начала координат в точку **MX**, являющуюся центром эллипсоида рассеяния случайного вектора **X**;
- затем производится поворот осей координат таким образом, чтобы новые оси координат $Of^{(1)}$, $Of^{(2)}$, …были направлены вдоль осей эллипсоида рассеяния, причем разброс точек вдоль оси $Of^{(1)}$ должен быть не меньше, чем вдоль оси $Of^{(2)}$ и т. д.

При этом разброс наблюдений вдоль новой оси $Of^{(1)}$ для исследователя наиболее важен, менее важен разброс вдоль оси $Of^{(2)}$, а разбросом вдоль нескольких последних осей можно пренебречь. Графически это иллюстрирует рис. 1.2.1.

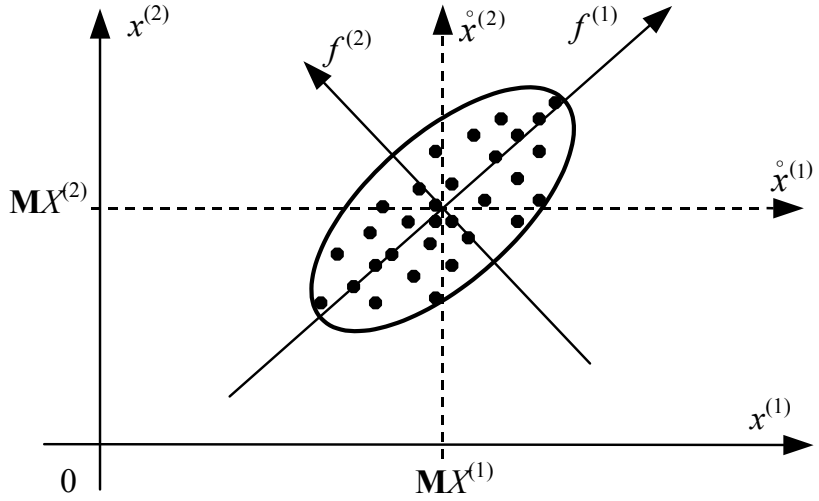


Рис. 1.2.1

1.2.3. Статистика модели главных компонент

В реальных задачах располагают лишь n наблюдениями

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik}) \qquad (i = 1, 2, \ldots, n)$$

k-мерного случайного вектора

$$\mathbf{X} = (X_1, X_2, \ldots, X_k)^{\mathrm{T}}$$

и оценками $\widehat{\mathbf{MX}}$ и $\hat{\boldsymbol{\Sigma}}$ вектора математических ожиданий $\mathbf{MX}$ и ковариационной матрицы $\boldsymbol{\Sigma}$. Будем предполагать, что наблюдения центрированы, т. е. произведен переход от x_{ij} к

$$\tilde{x}_{ij} = x_{ij} - \overline{x}_{\cdot j} \quad \left(\text{где } \overline{x}_{\cdot j} = \frac{\sum_{i=1}^{n} x_{ij}}{n}\right);$$

в дальнейшем значок «~» (тильда) над $\tilde{x}$ будем опускать. Если вектор $\mathbf{X}$ имеет нормальное распределение, наблюдения независимы, проведены в одинаковых вероятностных условиях, и все k собственных значений матрицы $\boldsymbol{\Sigma}$ различны, то справедливы следующие у т в е р ж д е н и я:

- оценки $\hat{\lambda}_j$ и $\hat{\mathbf{v}}_j = (\hat{v}_{1j}, \hat{v}_{2j}, \ldots, \hat{v}_{kj})^{\mathrm{T}}$ $(j = 1, 2, \ldots, k)$, найденные по матрице $\hat{\boldsymbol{\Sigma}}$, являются о ц е н к а м и м а к с и м а л ь н о г о п р а в д о п о д о б и я соответственно для λ_j и $\mathbf{v}_j$. Поэтому *выборочные главные компоненты*

$$\hat{\mathbf{F}}_j = (\hat{f}_{j1}, \hat{f}_{j2}, \ldots, \hat{f}_{jn}) \quad (j = 1, 2, \ldots, k),$$

где

$$\hat{f}_{ji} = \frac{1}{\sqrt{\hat{\lambda}_j}} \sum_{p=1}^{k} \hat{v}_{pj} \overset{\circ}{x}_{ip} \qquad (j = 1, 2, \ldots, k, \quad i = 1, 2, \ldots, n)$$

можно интерпретировать как о ц е н к и главных компонент F_j ($\hat{f}_{ji}$ — оценка j-й главной компоненты на i-м объекте);

- случайные величины

$$Y_j = \sqrt{n-1}(\hat{\lambda}_j - \lambda_j) \qquad (j = 1, 2, \ldots, k)$$

являются асимптотически нормальными с математическими ожиданиями $\mathbf{M}Y_j = 0$ и дисперсиями $\mathbf{D}Y_j = 2\lambda_j^2$. Поэтому при больших n д о в е р и т е л ь н ы й и н т е р в а л для собственного значения λ_j задается формулой

$$\mathbf{P}\left\{\frac{\hat{\lambda}_j}{1 + u_{\alpha/2}\sqrt{\frac{2}{(n-1)r}}} \leqslant \lambda_j \leqslant \frac{\hat{\lambda}_j}{1 - u_{\alpha/2}\sqrt{\frac{2}{(n-1)r}}}\right\} = 1 - \alpha, \qquad (1.2.11)$$

где $u_{\alpha/2}$ — квантиль уровня $\alpha/2$ стандартного нормального распределения (т. е. число, при котором $\mathbf{P}\{\mathcal{N}(0;\,1) < u_{\alpha/2}\} = \alpha/2$), а r — кратность собственного значения λ_j (в данном случае $r = 1$). Если $\hat{\lambda}_i$ попадает в доверительный интервал для λ_j при $i \neq j$, то возможно, что $\lambda_i = \lambda_j$.

Проверка гипотезы

$$H_0\colon \lambda_1 = \lambda_2 = \cdots = \lambda_k$$

[или равносильной ей гипотезы о диагональном виде ковариационной матрицы $\boldsymbol{\Sigma}$ (корреляционной матрицы $\mathbf{R}$)] основана на том, что при достаточно больших n статистика

$$\left[n - \frac{1}{6}\left(2k + 1 + \frac{2}{k}\right)\right]\left[-\ln\left(\det|\,\hat{\boldsymbol{\Sigma}}\,|\right) + k\ln\frac{\operatorname{tr}\hat{\boldsymbol{\Sigma}}}{k}\right]$$

(где $\det|\hat{\boldsymbol{\Sigma}}|$ — определитель матрицы $\hat{\boldsymbol{\Sigma}}$) имеет распределение $\chi^2\left(\frac{k(k+1)}{2} - 1\right)$

[статистика

$$-\left(n - \frac{2k+5}{6}\right)\ln\left(\det|\,\hat{\mathbf{R}}\,|\right)$$

имеет распределение $\chi^2\left(\frac{k(k-1)}{2}\right)$]. Принятие гипотезы H_0 означает, что переход к главным компонентам равносилен упорядочению исходных величин в порядке убывания их дисперсий.

Предположим, что k' первых главных компонент учтены; пусть $m = k - k'$. Проверка гипотезы

$$H_0\colon \lambda_{k'+1} = \lambda_{k'+2} = \cdots = \lambda_k$$

основана на статистике

$$a\left[-\ln\left(\det|\,\hat{\boldsymbol{\Sigma}}\,|\right) + \ln\left(\hat{\lambda}_1\hat{\lambda}_2\cdots\hat{\lambda}_{k'}\right) + m\ln b\right],$$

где

$$a = n - k' - \frac{1}{6}\left(2m + 1 + \frac{2}{m}\right),\ b = \frac{\operatorname{tr}\hat{\boldsymbol{\Sigma}} - \lambda_1 - \lambda_2 - \cdots - \lambda_{k'}}{m},$$

имеющей при больших n распределение $\chi^2\left(\frac{(m+2)(m-1)}{2}\right)$.

Если используется матрица $\hat{\mathbf{R}}$, и k' компонентами исчерпывается бóльшая доля суммарной дисперсии, то проверка гипотезы H_0 основана на аналогичной статистике с заменой $\hat{\boldsymbol{\Sigma}}$ на $\hat{\mathbf{R}}$ и a на n; однако в этом случае

аппроксимация распределения этой статистики распределением χ^2 менее точна, даже при бóльших n, чем при использовании матрицы $\hat{\boldsymbol{\Sigma}}$.

Возможно обобщение формулы (1.2.11) и на случай, когда собственное значение λ_j имеет кратность $r > 1$, т. е.

$$\lambda_j = \lambda_{j+1} = \cdots = \lambda_{j+(r-1)}:$$

в этом случае в формуле (1.2.11) следует заменить $\hat{\lambda}_j$ на

$$\overline{\lambda}_j = \frac{\hat{\lambda}_j + \hat{\lambda}_{j+1} + \cdots + \hat{\lambda}_{j+(r-1)}}{r}.$$

§ 1.3. ФАКТОРНЫЙ АНАЛИЗ

1.3.1. Модель факторного анализа

Как и прежде, будем полагать, что

$$\mathbf{X} = (X_1, X_2, \ldots, X_k)^{\mathrm{T}} —$$

исходный k-мерный случайный вектор.

Каноническая модель факторного анализа для центрированного вектора

$$\mathring{\mathbf{X}} = \mathbf{X} - \mathbf{M}\mathbf{X}$$

имеет следующий вид:

$$\mathring{\mathbf{X}} = \mathbf{A}\mathbf{F} + \mathring{\boldsymbol{\varepsilon}}, \tag{1.3.1}$$

где

$$\mathbf{F} = (F_1, F_2, \ldots, F_m)^{\mathrm{T}} —$$

центрированный и нормированный случайный вектор m ($m < k$) некоррелированных *общих факторов* для всех исходных случайных величин X_i ($i = 1, 2, \ldots, k$),

$$\mathbf{A} = (a_{ij}) \in \mathbb{R}^{k \times m} —$$

неслучайная *матрица нагрузок* случайных величин X_i на факторы F_j,

$$\mathring{\boldsymbol{\varepsilon}} = (\mathring{\varepsilon}_1, \mathring{\varepsilon}_2, \ldots, \mathring{\varepsilon}_k)^{\mathrm{T}} —$$

распределенный по k-мерному нормальному закону центрированный вектор *специфических факторов* $\mathring{\varepsilon}_1, \mathring{\varepsilon}_2, \ldots, \mathring{\varepsilon}_k$, некоррелированных как между собой, так и с общими факторами.

Пусть

$$\mathbf{\Sigma_X} = \mathbf{M}(\mathring{\mathbf{X}}\mathring{\mathbf{X}}^{\mathrm{T}}) \ —$$

ковариационная матрица вектора $\mathbf{X}$, а

$$\mathbf{\Sigma_\varepsilon} = \mathbf{M}(\mathring{\boldsymbol{\varepsilon}}\mathring{\boldsymbol{\varepsilon}}^{\mathrm{T}}) \ —$$

(диагональная) ковариационная матрица вектора $\mathring{\boldsymbol{\varepsilon}}$ с диагональными элементами, равными

$$\mathbf{M}\,\mathring{\varepsilon}_i^2 = \mathbf{D}\,\mathring{\varepsilon}_i = \upsilon_i .$$

Построим систему уравнений для определения матриц $\mathbf{A}$ и $\mathbf{\Sigma_\varepsilon}$. С учетом (1.3.1) и условий на векторах $\mathbf{F}$ и $\mathring{\boldsymbol{\varepsilon}}$ получим:

$$\mathbf{\Sigma_X} = \mathbf{M}[(\mathbf{AF} + \mathring{\boldsymbol{\varepsilon}})(\mathbf{AF} + \mathring{\boldsymbol{\varepsilon}})^{\mathrm{T}}] =$$

$$= \mathbf{M}(\mathbf{AFF}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}) + \mathbf{M}(\mathbf{AF}\,\mathring{\boldsymbol{\varepsilon}}^{\mathrm{T}}) + \mathbf{M}(\mathring{\boldsymbol{\varepsilon}}\,\mathbf{F}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}) + \mathbf{M}(\mathring{\boldsymbol{\varepsilon}}\mathring{\boldsymbol{\varepsilon}}^{\mathrm{T}}) =$$

$$= \mathbf{M}(\mathbf{AIA}^{\mathrm{T}}) + \mathbf{AM}(\mathbf{F}\,\mathring{\boldsymbol{\varepsilon}}^{\mathrm{T}}) + \mathbf{A}^{\mathrm{T}}\mathbf{M}(\mathring{\boldsymbol{\varepsilon}}\,\mathbf{F}^{\mathrm{T}}) + \mathbf{\Sigma_\varepsilon} = \mathbf{AA}^{\mathrm{T}} + \mathbf{\Sigma_\varepsilon}$$

(здесь $\mathbf{I}$ — единичная матрица порядка m).

Итак,

$$\mathbf{\Sigma_X} = \mathbf{AA}^{\mathrm{T}} + \mathbf{\Sigma_\varepsilon}$$

или

$$\begin{cases} \operatorname{cov}(X_i, X_p) = \sum\limits_{j=1}^{m} a_{ij}a_{pj} \quad (i = 1,2,...,k, \quad p = 1,2,...,k, \quad p \neq i), \\ \operatorname{cov}(X_i, X_i) = \mathbf{D}X_i = \sum\limits_{j=1}^{m} a_{ij}^2 + \upsilon_i \quad (i = 1,2,...,k). \end{cases} \tag{1.3.2}$$

Таким образом, здесь, в отличие от модели главных компонент (1.2.1), ковариации исходных случайных величин полностью воспроизводятся матрицей нагрузок, а для воспроизведения их дисперсий помимо нагрузок нужны дисперсии υ_i специфических факторов.

Далее, так как

$$\mathbf{M}(\mathring{\mathbf{X}}\mathbf{F}^{\mathrm{T}}) = \mathbf{M}[(\mathbf{AF} + \mathring{\boldsymbol{\varepsilon}})\mathbf{F}^{\mathrm{T}}] = \mathbf{M}(\mathbf{AFF}^{\mathrm{T}} + \mathring{\boldsymbol{\varepsilon}}\,\mathbf{F}^{\mathrm{T}}) =$$

$$= \mathbf{AM}(\mathbf{FF}^{\mathrm{T}}) + \mathbf{M}(\mathring{\boldsymbol{\varepsilon}}\,\mathbf{F}^{\mathrm{T}}) = \mathbf{A},$$

то здесь, как и в компонентном анализе, ковариации

$$\operatorname{cov}(X_i, F_j) = a_{ij}.$$

З а м е ч а н и е . Если исходный k-мерный вектор $\mathbf{X}$ не только центрирован, но и нормирован, то $\mathbf{\Sigma_X}$ — это корреляционная матрица $\mathbf{R_X}$, и система (1.3.2) примет вид

$$\mathbf{R_X} = \mathbf{AA}^{\mathrm{T}} + \mathbf{\Sigma_\varepsilon}$$

или

$$\begin{cases} \rho(X_i, X_p) = \sum\limits_{j=1}^{m} a'_{ij} a'_{pj} \quad (i = 1,2,...,k, \quad p = 1,2,...,k, \quad i \neq p), \\ \rho(X_i, X_i) = 1 = \sum\limits_{j=1}^{m} a'^{2}_{ij} + \upsilon'_i \quad (i = 1,2,...,k), \end{cases} \tag{1.3.3}$$

и в этом случае

$$a'_{ij} = \rho(X_i, F_j), \quad \upsilon'_i \in [0;1].$$

Числа

$$h_i = \sum_{j=1}^{m} a'^{2}_{ij} = 1 - \upsilon'_i \qquad (i = 1, 2, \ldots, k)$$

называют *общностями* случайных величин X_i, а матрицу

$$\mathbf{R'} = \mathbf{A'(A')}^{\mathrm{T}},$$

где $\mathbf{A'} = (a'_{ij}) \in \mathbb{R}^{k \times m}$, — *редуцированной матрицей* ($\mathbf{R'}$ отличается от $\mathbf{R}$ только тем, что ее диагональными элементами являются не единицы, а общности h_i).

В системе (1.3.2) k^2 уравнений, а число неизвестных (a_{ij} и υ_i) равно $mk + k < k(k + 1)$. Если допустить, что k, m и матрица $\mathbf{\Sigma_X}$ таковы, что решение этой системы существует (иначе построение модели (1.3.1) недопустимо), то это решение не единственно.

Действительно, пусть $\mathbf{V} \in \mathbb{R}^{m \times m}$ — некоторая ортогональная матрица. Проведем тождественные преобразования модели (1.3.1):

$$\mathring{\mathbf{X}} = \mathbf{AF} + \mathring{\boldsymbol{\varepsilon}} = \mathbf{AIF} + \mathring{\boldsymbol{\varepsilon}} = \mathbf{A}(\mathbf{VV}^{\mathrm{T}})\mathbf{F} + \mathring{\boldsymbol{\varepsilon}} = (\mathbf{AV})(\mathbf{V}^{\mathrm{T}}\mathbf{F}) + \mathring{\boldsymbol{\varepsilon}}. \tag{1.3.4}$$

В преобразованной модели (1.3.4) вектор общих факторов — это вектор

$$\tilde{\mathbf{F}} = \mathbf{V}^{\mathrm{T}}\mathbf{F},$$

а матрица нагрузок равна

$$\tilde{\mathbf{A}} = \mathbf{AV}.$$

Итак, если решение системы (1.3.2) существует, то оно не единственно: допустим целый класс матриц нагрузок, которые связаны между собой ортогональными преобразованиями.

З а м е ч а н и е . В методе главных компонент также допустимо ортогональное преобразование матрицы нагрузок. Однако вращение пространства главных компонент меняет вклады компонент в общую дисперсию исходных случайных величин: они становятся не равными собственным значениям ковариационной матрицы $\mathbf{\Sigma}$.

При каких дополнительных условиях на k, m и матрицу нагрузок $\mathbf{A}$ решение системы (1.3.2) единственно с точностью до ортогонального преобразования? Матрица $\mathbf{A}$ должна быть такой, чтобы при вычеркивании из нее любой строки оставшуюся матрицу можно было бы разбить на две подматрицы ранга m (откуда получаем, что $2m \leqslant k-1$). Сформулированное требование к матрице $\mathbf{A}$ является не только достаточным, но при $m=1$ и $m=2$ также и необходимым условием единственности решения системы (1.3.2).

Будем предполагать, что решение единственно с точностью до ортогонального преобразования. Тогда, вращая систему координат в m-мерном пространстве общих факторов, можно найти такую матрицу нагрузок, которая позволила бы дать содержательную интерпретацию общих факторов в терминах исходных случайных величин. Существует несколько вариантов дополнительных условий на класс матриц $\mathbf{A}$, обеспечивающих уже окончательную однозначность решения. От этих условий зависит и метод нахождения матриц $\mathbf{A}$, $\mathbf{\Sigma_\varepsilon}$, вектора $\mathbf{F}$, и, соответственно, способ их статистического оценивания.

Наиболее формализованы следующие варианты дополнительных **идентифицирующих требований** к виду матрицы $\mathbf{A}$.

Первое идентифицирующее требование. *Матрица* $\mathbf{J} = \mathbf{A}^{\mathrm{T}}\mathbf{\Sigma}_{\varepsilon}^{-1}\mathbf{A}$ *должна иметь диагональный вид с различными расположенными в порядке убывания диагональными элементами.*

Второе идентифицирующее требование. *Матрица* $\mathbf{A}^{\mathrm{T}}\mathbf{B}$ *(где матрица* $\mathbf{B} = (b_{ij}) \in \mathbb{R}^{k\times m}$ *задана заранее) должна иметь ранг* m *и быть верхней треугольной*.

В частности, второе идентифицирующее требование выполнено, если

$$\mathbf{A} = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m,2} & \cdots & a_{mm} \\ a_{m+1,1} & a_{m+1,2} & \cdots & a_{m+1,m} \\ a_{m+2,1} & a_{m+2,2} & \cdots & a_{m+2,m} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{km} \end{pmatrix} \begin{matrix} \\ \\ \\ \left.\begin{matrix}\\ \\ \\ \end{matrix}\right\} m \text{ строк} \\ \\ \\ \left.\begin{matrix}\\ \\ \\ \end{matrix}\right\} (k-m) \text{ строк} \end{matrix}, \quad \mathbf{B} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \begin{matrix} \\ \\ \\ \left.\begin{matrix}\\ \\ \\ \end{matrix}\right\} m \text{ строк} \\ \\ \\ \left.\begin{matrix}\\ \\ \\ \end{matrix}\right\} (k-m) \text{ строк} \end{matrix},$$

при этом исходный признак X_1 выражается только через общий фактор F_1, X_2 — через F_1 и F_2 и т. д.

Общая итерационная схема поиска параметров $(\mathbf{A}, \mathbf{\Sigma_\varepsilon})$ модели факторного анализа (при заданном числе общих факторов m) такова:

- задаются нулевым приближением $\upsilon_i^{(0)}$ дисперсий υ_i специфических факторов (т. е. нулевым приближением $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}^{(0)}$ диагональной матрицы $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$);
- получают нулевое приближение

$$\mathbf{\Psi}^{(0)} = \mathbf{\Sigma}_{\mathbf{X}} - \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}^{(0)}$$

матрицы $\mathbf{\Psi} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$;
- последовательно определяют нулевые приближения

$$\mathbf{a}_1^{(0)}, \mathbf{a}_2^{(0)}, \ldots, \mathbf{a}_m^{(0)}$$

столбцов $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ матрицы $\mathbf{A}$;

З а м е ч а н и е . Нетрудно убедиться в том, что

$$\mathbf{\Psi} = \mathbf{a}_1\mathbf{a}_1^{\mathrm{T}} + \mathbf{a}_2\mathbf{a}_2^{\mathrm{T}} + \cdots + \mathbf{a}_m\mathbf{a}_m^{\mathrm{T}}.$$

Исходя из этого равенства и учитывая специфику выбранного варианта идентифицирующих требований к матрице $\mathbf{A}$, сначала находят столбец $\mathbf{a}_1$. Затем переходят к матрице

$$\mathbf{\Psi}_1 = \mathbf{\Psi} - \mathbf{a}_1\mathbf{a}_1^{\mathrm{T}}$$

и определяют столбец $\mathbf{a}_2$ и т. д.
- определяют первые приближения

$$\upsilon_i^{(1)} = \mathbf{D}X_i - \sum_{j=1}^{m}\left(a_{ij}^{(0)}\right)^2$$

дисперсий υ_i (т. е. первое приближение $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}^{(1)}$ матрицы $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$) и переходят к следующей итерации;
- итерационный процесс заканчивают, когда очередное приближение матрицы $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ мало отличается от предыдущего (т. е. когда на очередном i-м шаге

$$\left\|\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}^{(i)} - \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}^{(i-1)}\right\| < \delta,$$

где δ — выбранное заранее малое число — *точность*).

В реальных задачах располагают лишь оценкой $\hat{\mathbf{\Sigma}}_{\mathbf{X}}$ ковариационной матрицы $\mathbf{\Sigma}_{\mathbf{X}}$. Заменив в рассмотренной общей схеме $\mathbf{\Sigma}_{\mathbf{X}}$ на $\hat{\mathbf{\Sigma}}_{\mathbf{X}}$, можно получить оценки $\hat{\mathbf{A}}$ и $\hat{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon}}$ соответственно матрицы нагрузок $\mathbf{A}$ и ковариационной матрицы $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ специфических факторов.

1.3.2. Метод максимального правдоподобия

Реализация описанной в предыдущем пункте общей итерационной схемы для первого идентифицирующего требования к виду матрицы $\mathbf{A}$ приводит к оценкам максимального правдоподобия параметров a_{ij} и υ_i — оценкам, получаемым *методом максимального правдоподобия* для модели (1.3.1) при постулировании нор-

мальности распределения наблюдений вектора **X** с учетом указанного требования. При достаточно общих ограничениях (регулярности) оценки максимального правдоподобия $\hat{\mathbf{A}}$ и $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}$ являются асимптотически нормальными, несмещенными и эффективными.

Можно доказать, что при максимизации логарифмической функции правдоподобия с учетом первого идентифицирующего требования собственными значениями матрицы

$$\hat{\mathbf{J}} = \hat{\mathbf{A}}^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}^{-1} \hat{\mathbf{A}}$$

(т. е. ее диагональными элементами) будут первые m наибольших собственных значений матрицы

$$\boldsymbol{\Phi} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}^{-1} (\hat{\boldsymbol{\Sigma}}_{\mathbf{X}} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}),$$

а соответствующие собственные векторы будут столбцами матрицы $\hat{\mathbf{A}}$. Поэтому **итерационная схема** нахождения оценок $\hat{\mathbf{A}}$ и $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}$ при заданном m примет следующий вид:

- задаются нулевыми приближениями $\hat{\upsilon}_i^{(0)}$ оценок дисперсий $\hat{\upsilon}_i$ специфических факторов (т. е. нулевым приближением $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}^{(0)}$ матрицы $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}$);
- получают нулевое приближение

$$\boldsymbol{\Phi}^{(0)} = \left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}^{(0)}\right)^{-1} \left(\hat{\boldsymbol{\Sigma}}_{\mathbf{X}} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}^{(0)}\right)$$

матрицы $\boldsymbol{\Phi}$;

- находят $\hat{\lambda}_1^{(0)}$ — наибольшее собственное значение матрицы $\boldsymbol{\Phi}^{(0)}$ и соответствующий ему собственный вектор $\hat{\mathbf{v}}_1^{(0)}$ — это столбец $\hat{\mathbf{a}}_1^{(0)}$ матрицы $\hat{\mathbf{A}}^{(0)}$; в $\boldsymbol{\Phi}^{(0)}$ матрицу $\hat{\boldsymbol{\Sigma}}_{\mathbf{X}}$ заменяют на

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{X}} - \hat{\mathbf{a}}_1^{(0)} (\hat{\mathbf{a}}_1^{(0)})^{\mathrm{T}},$$

и для новой матрицы $\boldsymbol{\Phi}^{(0)}$ находят наибольшее собственное значение $\hat{\lambda}_2^{(0)}$ и соответствующий ему собственный вектор — это будет столбец $\hat{\mathbf{a}}_2^{(0)}$ матрицы $\hat{\mathbf{A}}^{(0)}$; так продолжают до получения m столбцов матрицы $\hat{\mathbf{A}}^{(0)}$;

- получают первое приближение

$$\hat{\upsilon}_i^{(1)} = \widehat{\mathbf{D}X_i} - \sum_{j=1}^{m} (\hat{a}_{ij}^{(0)})^2$$

дисперсий $\hat{\upsilon}_i$ (т. е. первое приближение $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}^{(1)}$ матрицы $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}$), после чего переходят к следующей итерации.

Допустимо ли представление исходного вектора **X** с помощью модели (1.3.1) факторного анализа с числом общих факторов, равным m? Г и п о -

т е з а о том, что число общих факторов равно m, отвергается (на уровне значимости α), если

$$\chi^2 = n\left[\ln\left(\det|\hat{\boldsymbol{\Sigma}}_{\varepsilon}|\right) + \ln\left(\det|\mathbf{I} + \hat{\mathbf{J}}|\right)\right] > \chi^2_{\alpha}(q), \qquad (1.3.5)$$

где число степеней свободы

$$q = \frac{(k-m)^2 - (k+m)}{2}.$$

1.3.3. Ц е н т р о и д н ы й м е т о д

Одним из способов реализации общей итерационной схемы при в т о р о м и д е н т и ф и ц и р у ю щ е м т р е б о в а н и и является *центроидный метод*. Оценки, получаемые этим методом, близки к оценкам максимального правдоподобия, и являются более «устойчивыми» по отношению к отклонениям от нормальности распределения наблюдений вектора **X**; однако исследование их статистических свойств затруднено из-за использования в процедуре метода неформализуемых эвристических соображений. Метод менее трудоемок по сравнению с методом максимального правдоподобия и допускает наглядную г е о м е т р и ч е с к у ю и н т е р п р е т а ц и ю:

- исходные случайные величины $X_1, X_2, \ldots, X_k$ отождествляют с радиус-векторами $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ k-мерного пространства, построенными так, чтобы

$$\cos(\widehat{\mathbf{x}_i, \mathbf{x}_j}) = \rho(X_i, X_j) \qquad (i = 1, 2, \ldots, k, \quad j = 1, 2, \ldots, k),$$

а

$$\|\mathbf{x}_i\| = \sqrt{\mathbf{D}X_i} \qquad (i = 1, 2, \ldots, k);$$

- изменяя знаки отдельных векторов, добиваются, чтобы как можно больше корреляций между случайными величинами X_i были положительными (или, иначе, чтобы как можно больше векторов $\mathbf{x}_i$ образовывали о д н о н а п р а в л е н н ы й п у ч о к, т. е. чтобы как можно больше углов между векторами $\mathbf{x}_i$ были острыми);
- определяют вектор $\mathbf{f}_1$ (общий фактор F_1 — *первый центроид*) как нормированную сумму векторов пучка, при этом нагрузки (веса) равны

$$a_{i1} = \rho(X_i, F_1) = \cos(\widehat{\mathbf{x}_i, \mathbf{f}_1}) \qquad (i = 1, 2, \ldots, k);$$

- затем подсчитывают корреляционную матрицу

$$\mathbf{R}_{\mathbf{X}}^{(1)} = \mathbf{R}_{\mathbf{X}} - \mathbf{a}_1\mathbf{a}_1^{\mathrm{T}}$$

о с т а т о ч н ы х переменных

$$X_i^{(1)} = X_i - a_{i1}F_1 \qquad (i = 1, 2, \ldots, k),$$

где

$$\mathbf{a}_1 = (a_{11}, a_{21}, \ldots, a_{k1})^{\mathrm{T}},$$

и относительно $X_i^{(1)}$ и $\mathbf{R}_{\mathbf{X}}^{(1)}$ проделывают аналогичную процедуру, выделяя F_2 — *второй центроид* и т. д.

Для центроидного метода описанная выше **общая итерационная схема** факторного анализа конкретизируется (в терминах выборки) следующим образом:

- задаются нулевым приближением $\hat{\upsilon}_i^{(0)}$ дисперсий $\hat{\upsilon}_i$ (т. е. нулевым приближением $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}^{(0)}$ ковариационной матрицы $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}$ специфических факторов), обычно выбирают

$$\hat{\upsilon}_i^{(0)} = \hat{\sigma}_{X_i}^2 (1 - \max_{j \neq i} |\hat{\rho}_{ij}|);$$

- подсчитывают

$$\widehat{\boldsymbol{\Psi}}^{(0)} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}^{(0)};$$

- определяют нулевое приближение $\hat{\mathbf{a}}_1^{(0)}$ первого столбца $\hat{\mathbf{a}}_1$ матрицы $\hat{\mathbf{A}}$ по формуле

$$\hat{\mathbf{a}}_1^{(0)} = \frac{\widehat{\boldsymbol{\Psi}}^{(0)} \mathbf{b}_1^{(0)}}{\sqrt{\left(\mathbf{b}_1^{(0)}\right)^{\mathrm{T}} \widehat{\boldsymbol{\Psi}}^{(0)} \mathbf{b}_1^{(0)}}},$$

где

$$\mathbf{b}_1^{(0)} = (\underbrace{1;\ 1;\ \ldots;\ 1}_{k \text{ единиц}})^{\mathrm{T}} \in \mathbb{R}^k \ —$$

нулевое приближение первого столбца $\mathbf{b}_1$ вспомогательной матрицы $\mathbf{B}$; затем вычисляют

$$\widehat{\boldsymbol{\Psi}}_1^{(0)} = \widehat{\boldsymbol{\Psi}}^{(0)} - \hat{\mathbf{a}}_1^{(0)} \left(\hat{\mathbf{a}}_1^{(0)}\right)^{\mathrm{T}}$$

и определяют нулевое приближение $\hat{\mathbf{a}}_2^{(0)}$ второго столбца $\hat{\mathbf{a}}_2$ матрицы $\hat{\mathbf{A}}$ по формуле

$$\hat{\mathbf{a}}_2^{(0)} = \frac{\widehat{\boldsymbol{\Psi}}_1^{(0)} \mathbf{b}_2^{(0)}}{\sqrt{\left(\mathbf{b}_2^{(0)}\right)^{\mathrm{T}} \widehat{\boldsymbol{\Psi}}_1^{(0)} \mathbf{b}_2^{(0)}}},$$

где все координаты вектора $\mathbf{b}_2^{(0)} \in \mathbb{R}^k$ равны единице по модулю, а знаки координат подбираются так, чтобы знаменатель в последней дроби был максимальным; так продолжают до получения m столбцов матрицы $\hat{\mathbf{A}}^{(0)}$;

- получают диагональную матрицу $\hat{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon}}^{(1)}$, диагональные элементы которой вычисляются по формуле

$$\hat{\upsilon}_i^{(1)} = \widehat{\mathbf{D}X_i} - \sum_{j=1}^{m} \left(\hat{a}_{ij}^{(0)}\right)^2$$

и переходят к следующей итерации.

Замечание. Из изложенного алгоритма видно, что столбец $\mathbf{b}_1$ матрицы $\mathbf{B}$ задает веса, с которыми суммируются векторы одного пучка для получения общего вектора $\mathbf{F}_1$. Поскольку все веса по модулю равны единице, то определение очередного центроида состоит в простом суммировании векторов пучка; знаки же единиц определяют нужное направление каждого из векторов пучка. Вообще говоря, знаки устанавливаются на основе анализа знаков элементов остаточных матриц $\hat{\mathbf{\Psi}}_{i-1}$; в данном алгоритме предлагается при подборе знаков ориентироваться на максимизацию произведения

$$\left(\mathbf{b}_i^{(0)}\right)^T \hat{\mathbf{\Psi}}_{i-1} \mathbf{b}_i^{(0)},$$

что позволяет быстрее выделить m общих факторов, объясняющих возможно бóльшую часть общей дисперсии исходных величин.

Недостатком центроидного метода является зависимость получаемых с его помощью значений нагрузок от шкалы измерения исходных величин, поэтому перед применением центроидного метода исходные признаки обычно нормируют.

1.3.4. Метод Бартлетта оценки общих факторов

Согласно методу Бартлетта, в предположении, что известны оценки $\hat{\mathbf{A}}$ и $\hat{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon}}$, модель (1.3.1) для каждого фиксированного наблюдения i ($i = 1, 2, \ldots, n$) рассматривается как регрессия

$$x_{ip} = f_{1i}\hat{a}_{p1} + f_{2i}\hat{a}_{p2} + \cdots + f_{mi}\hat{a}_{pm} + \varepsilon_{pi} \qquad (i = 1,2,\ldots,n,\ p = 1,2,\ldots,k),$$

и величины $f_{1i}, f_{2i}, \ldots, f_{mi}$ интерпретируются как неизвестные коэффициенты регрессии. Тогда в соответствии с процедурой метода наименьших квадратов вектор

$$\hat{\mathbf{F}}_i = (\hat{f}_{1i}, \hat{f}_{2i}, \ldots, \hat{f}_{mi})^{\mathrm{T}}$$

определяется по формуле

$$\hat{\mathbf{F}}_i = \left(\hat{\mathbf{A}}^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon}}^{-1} \hat{\mathbf{A}}\right)^{-1} \hat{\mathbf{A}}^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon}}^{-1} \mathbf{x}_i \qquad (i = 1, 2, \ldots, n),$$

где

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^{\mathrm{T}} \ —$$

вектор значений исходных признаков $X_1, X_2, \ldots, X_k$ в i-м наблюдении ($i = 1, 2, \ldots, n$).

§ 1.4. Практическое использование методов снижения размерности

1.4.1. Различия методов снижения размерности

Отметим основные различия описанных методов снижения размерности:

- специфические факторы присутствуют только в модели факторного анализа, но не в модели главных компонент;
- общее число главных компонент равно числу исходных показателей; в отличие от этого, пространство общих факторов является пространством наименьшей возможной размерности, в котором можно представить исходные многомерные признаки;
- хотя в методе главных компонент для точного воспроизведения дисперсий и корреляций необходимо найти все компоненты, бóльшая доля дисперсии объясняется сравнительно небольшим числом главных компонент; кроме того, имеется возможность восстановления значений всех исходных признаков по значениям главных компонент и наоборот — восстановления значений главных компонент по значениям исходным признаков; для центроидного метода факторного анализа это принципиально невозможно; можно лишь минимизировать дисперсию остатков;
- метод главных компонент обладает свойствами линейности и аддитивности; центроидный метод, например, основан лишь на гипотезе аддитивности, и если эта гипотеза не принимается, то метод можно использовать лишь для получения приближенных оценок, уточняемых затем методом максимального правдоподобия;

На самом деле модель компонентного анализа и модель факторного анализа являются частными случаями более общей модели, различия состоят в выборе критерия оптимальности. Подробнее об этом можно прочитать в учебнике [3].

1.4.2. Особенности применения методов снижения размерности с использованием современных пакетов прикладных программ

В современных статистических пакетах прикладных программ наряду с методом главных компонент и рассмотренными методами факторного анализа реализованы методы, основанные на итерационных процедурах вычисления общностей. Суть этих методов при использовании корреляционной матрицы $\hat{\mathbf{R}}_{\mathbf{X}}$ состоит в нахождении заданного числа главных компонент редуцированной матрицы $\hat{\mathbf{R}}'_{\mathbf{X}}$, в которой стоящие на диагонали неизвестные общности заменяют некоторыми нулевыми приближениями (например, оценками квадратов множественных ко-

эффициентов корреляции); после нахождения нулевых приближений оценок нагрузок пересчитывают оценки общностей и переходят к следующей итерации; итерации продолжают до тех пор, пока не превышено максимальное число итераций или изменение в общностях не становится меньше заданной точности.

В **п. 1.3.1** мы отметили, что решение задачи факторного анализа при числе общих факторов, большем единицы, определяется неоднозначно, и можно получить целый набор полностью эквивалентных (с математической точки зрения) решений, получающихся одно из другого с помощью ортогональных преобразований. Поскольку начало координат у всех решений совпадает, фактически эти решения получаются друг из друга вращением координатных осей. В современных статистических пакетах реализованы следующие наиболее типичные стратегии вращения факторного пространства (current rotation), обеспечивающие максимально возможную концентрацию дисперсии исходных данных на координатных осях выделенных факторов и облегчающие предметную интерпретацию факторов:

- *варимакс* (максимизируют

$$V = \sum_{j=1}^{m} \frac{\sum_{i=1}^{k} \left(a_{ij}^2 - \frac{\sum_{h=1}^{k} a_{hj}^2}{k} \right)^2}{k} \quad —$$

различие столбцов матрицы нагрузок, тем самым обеспечивая разделение факторов за счет уменьшения числа исходных переменных, связанных с каждым фактором);

- *квартимакс* (максимизируют

$$Q = \sum_{i=1}^{k} \frac{\sum_{j=1}^{m} \left(a_{ij}^2 - \frac{\sum_{h=1}^{m} a_{jh}^2}{m} \right)^2}{m} \quad —$$

различие строк матрицы нагрузок, тем самым уменьшая число факторов, связанных с каждой переменной);

- *биквартимакс* (одновременно максимизируют различие и столбцов и строк матрицы нагрузок) и др.

Отметим также, что часто интересные и полезные результаты удается получить, поменяв ролями объекты и признаки и изучая, насколько близки друг к другу объекты.

1.4.3. Компонентный анализ мировой демографической статистики

Изучается зависимость ожидаемой продолжительности жизни при рождении в годах Y от пяти факторных признаков: среднего числа детей в семье X_1, доли молодежи X_2 — процента населения моложе 15 лет, валового внутреннего продукта X_3, приходящегося на душу населения в тыс. долл. США, плотности населения X_4 в тыс. чел. на кв. км и процента грамотных X_5 по числовым данным, собранным по 52 странам и приведенным в табл. 1.4.1.

Таблица 1.4.1

Страна	Y	x_1	x_2	x_3	x_4	x_5	Страна	Y	x_1	x_2	x_3	x_4	x_5
Австралия	79	1,7	20	23850	2,34	100	Кения	48	4,4	44	1010	51,95	69
Австрия	78	1,3	17	24600	98,05	99	Китай	71	1,8	23	3550	134,38	78
Беларусь	68	1,3	19	6880	48,83	99	Колумбия	71	2,6	32	5580	38,28	87
Бельгия	78	1,6	18	25710	340,63	99	Коста-Рика	77	2,6	32	7880	73,44	93
Боливия	62	4,2	40	2300	7,81	78	Мексика	75	2,8	34	8070	51,56	87
Ботсвана	41	3,9	41	6540	2,73	72	Нигерия	52	5,8	44	770	138,67	51
Бразилия	68	2,4	30	6840	20,31	81	Нидерланды	78	1,7	19	24410	397,66	99
Буркина-Фасо	47	6,8	48	960	45,31	18	Никарагуа	68	4,3	43	2060	40,63	57
Великобритания	77	1,7	19	22220	248,05	99	Новая Зеландия	77	2	23	17630	14,45	99
Венгрия	71	1,3	17	11050	108,59	99	Пакистан	60	5,6	42	1860	184,38	35
Вьетнам	66	2,3	33	1860	243,36	88	Парагвай	73	4,3	40	4380	14,06	90
Габон	52	4,3	40	5280	4,69	61	Перу	69	2,9	34	4480	20,70	85
Гаити	49	4,7	43	1470	253,91	53	Польша	73	1,4	20	8390	121,09	99
Гамбия	52	5,9	46	1550	126,17	27	Россия	66	1,2	18	6990	8,59	99
Гватемала	66	4,8	44	3630	120,70	55	Руанда	39	5,8	44	880	280,86	50
Германия	78	1,3	16	23510	233,20	99	Сенегал	52	5,7	44	1400	49,61	38
Гондурас	66	4,4	43	2270	60,55	73	Соединенные Штаты Америки	77	2,1	21	31910	30,08	97
Доминиканская Республика	69	3,1	35	5210	178,13	83	Таиланд	72	1,8	24	5950	123,05	93
Замбия	37	6,1	45	720	13,28	73	Уганда	42	6,9	51	1160	100,78	48
Индия	61	3,2	36	2230	317,97	52	Украина	68	1,1	18	3360	82,42	97
Индонезия	67	2,7	31	2660	109,38	77	Филиппины	67	3,5	37	3990	260,16	90
Испания	78	1,2	15	17850	79,69	95	Центрально-Африканская Республика	45	5,1	44	1150	5,86	27
Италия	79	1,3	14	22000	194,14	97	Швейцария	71	3,3	34	2820	46,09	88
Камбоджа	56	4	43	1350	73,05	35	Эквадор	40	5,9	46	4380	64,45	99
Камерун	55	5,2	43	1490	33,59	54	Эль-Сальвадор	70	3,5	36	4260	307,81	73
Канада	79	1,4	19	25440	3,13	97	Эфиопия	52	5,9	44	620	59,77	24

Матрица парных коэффициентов корреляции (первый столбец соответствует Y, оставшиеся столбцы — факторным признакам X_1, X_2, X_3, X_4, X_5) свидетельствует о том, что Y достаточно сильно связан со всеми факторными признаками, однако многие из регрессоров коллинеарны между собой:

$$\begin{pmatrix} 1,000 & -0,840 & -0,803 & 0,674 & 0,155 & 0,700 \\ -0,840 & 1,000 & 0,948 & -0,665 & -0,136 & -0,820 \\ -0,803 & 0,948 & 1,000 & -0,774 & -0,165 & -0,787 \\ 0,674 & -0,665 & -0,774 & 1,000 & 0,167 & 0,627 \\ 0,155 & -0,136 & -0,165 & 0,167 & 1,000 & 0,055 \\ 0,700 & -0,820 & -0,787 & 0,627 & 0,055 & 1,000 \end{pmatrix}$$

В результате проведения множественного линейного регрессионного анализа с последовательным исключением незначимых регрессоров было получено линейное уравнение регрессии

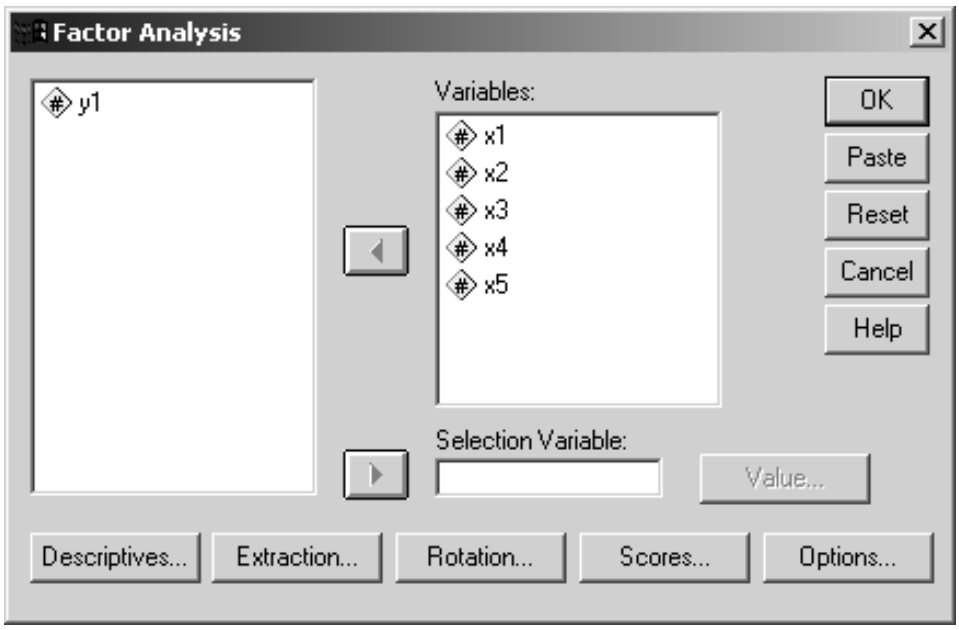
$$\hat{Y} = 78,75 - 5,02x_1 + 0,0003x_3,$$

оценка нормированного коэффициента детерминации оказалась при этом равной $\hat{R}^2_{\text{норм}} = 0,72$. Такое уравнение можно получить с помощью пакета SPSS путем выбора пункта «Regression | Linear…» меню «Statistics» и указания метода пошагового удаления регрессоров («Method: Stepwise»).

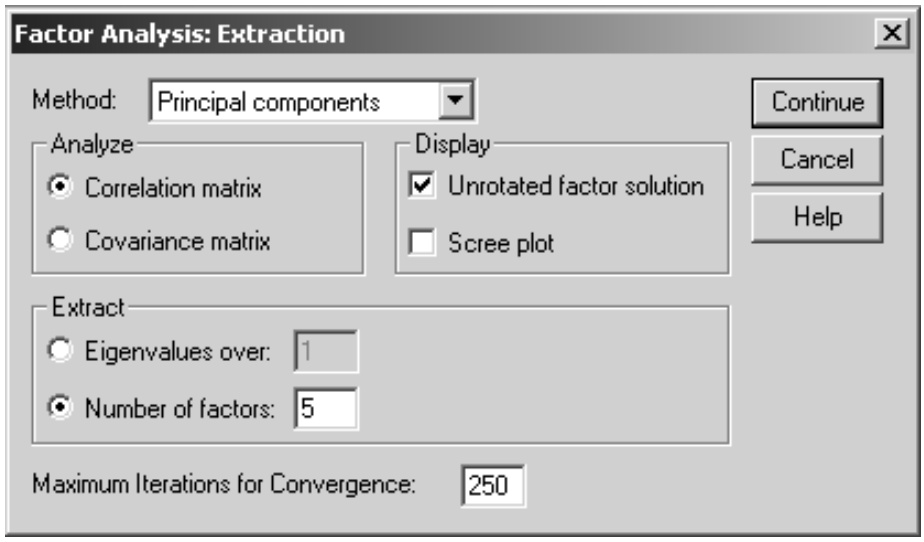
Выявим латентные факторы (компоненты), влиянием которых объясняется вариация факторных признаков X_1, X_2, X_3, X_4, X_5.

Для этого воспользуемся пакетом SPSS. Введем в окно ввода исходных данных пакета SPSS матрицу значений признаков, обратимся (с помощью выбора пункта «Data Reduction | Factor…» меню «Statistics») к программе «Factor Analysis» для компонентного анализа переменных X_1, X_2, X_3, X_4, X_5 (рис. 1.4.1, *а*) и реализуем метод главных компонент («Method: Principal components»), задав в окне, вызываемом нажатием кнопки «Extraction…», максимальное число факторов («Number of factors») равным пяти; поскольку исходные признаки разнородны по содержательному смыслу и имеют разные единицы измерения, компонентный анализ будем проводить с использованием корреляционной (а не ковариационной) матрицы («Analyze: Correlation matrix») (рис. 1.4.1, *б*). Установим флажок «Save as variables» в окне «Factor Analysis: Factor Scores», вызываемом нажатием кнопки «Scores…» — тогда значения пяти главных компонент на 52 объектах автоматически добавятся в виде переменных к исходным данным (рис. 1.4.1, *в*). Флажок «Display factor score coefficient matrix» позволяет получить в результате работы программы матрицу нагрузок компонент на признаки.

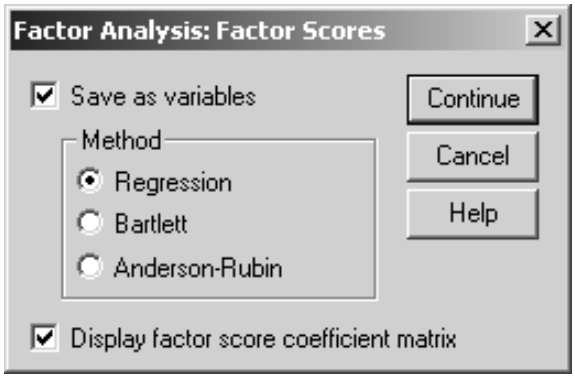
Обратимся к результатам работы программы (рис. 1.4.2). На основе анализа таблицы «Total Variance Explained» можно сделать вывод о том, что вклад первой главной компоненты в суммарную дисперсию составляет 66,99%, второй главной компоненты — 19,67%, третьей — 8,05%, и т. д., при этом общий вклад первых двух компонент в суммарную дисперсию равен 86,67%. Зависимость доли дисперсии исходных признаков, объясненной первыми k главными компонентами, от k, иллюстрируется рис. 1.4.3.


Factor Analysis
y1
Variables:
x1
x2
x3
x4
x5
OK
Paste
Reset
Cancel
Help
Selection Variable:
Value...
Descriptives...
Extraction...
Rotation...
Scores...
Options...

а)


Factor Analysis: Extraction
Method:
Principal components
Continue
Cancel
Help
Analyze
Correlation matrix
Covariance matrix
Display
Unrotated factor solution
Scree plot
Extract
Eigenvalues over:
1
Number of factors:
5
Maximum Iterations for Convergence:
250

б)


Factor Analysis: Factor Scores
Save as variables
Continue
Cancel
Help
Method
Regression
Bartlett
Anderson-Rubin
Display factor score coefficient matrix

в)

Рис. 1.4.1

Total Variance Explained

Component	Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %
1	3,35	66,99	66,99
2	0,98	19,67	86,67
3	0,40	8,05	94,72
4	0,23	4,58	99,30
5	0,04	0,70	100,00

Extraction Method: Principal Component Analysis.

Component Matrix[a]

	Component				
	1	2	3	4	5
X1	–0,947	0,065	0,203	0,208	–0,124
X2	–0,967	0,025	0,018	0,213	0,135
X3	0,836	0,035	0,539	0,091	0,034
X4	0,203	0,976	–0,075	0,037	0,001
X5	0,882	–0,160	–0,255	0,362	–0,017

Extraction Method: Principal Component Analysis.

[a] 5 components extracted.

Component Score Coefficient Matrix

	Component				
	1	2	3	4	5
X1	–0,283	0,066	0,504	0,908	–3,530
X2	–0,289	0,026	0,046	0,929	3,857
X3	0,250	0,036	1,338	0,397	0,971
X4	0,060	0,992	–0,187	0,161	0,021
X5	0,263	–0,163	–0,634	1,580	–0,484

Extraction Method: Principal Component Analysis.

Component Scores

Рис. 1.4.2

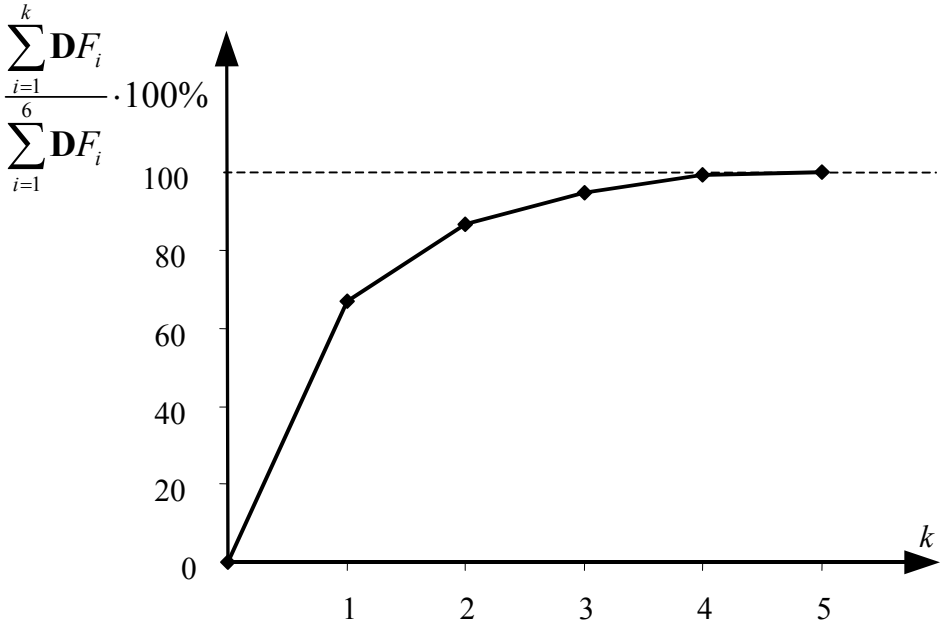


Рис. 1.4.3

В таблице «Component Matrix» приводится матрица

$$\mathbf{A} = \begin{pmatrix} \mathbf{-0,947} & 0,065 & 0,203 & 0,208 & -0,124 \\ \mathbf{-0,967} & 0,025 & 0,018 & 0,213 & 0,135 \\ \mathbf{0,836} & 0,035 & 0,539 & 0,091 & 0,034 \\ 0,203 & \mathbf{0,976} & -0,075 & 0,037 & 0,001 \\ \mathbf{0,882} & -0,160 & -0,255 & 0,362 & -0,017 \end{pmatrix}$$

нагрузок признаков X_1, X_2, X_3, X_4, X_5 (строк) на главные компоненты (столбцы) F_1, F_2, F_3, F_4, F_5 (жирным выделены элементы, бóльшие 0,6 по абсолютному значению, учитывающиеся при содержательной интерпретации главных компонент).

Запишем на основе матрицы нагрузок выражения исходных признаков через главные компоненты, например,

$$X_1 = -0,947F_1 - 0,065F_2 + 0,203F_3 + 0,208F_4 - 0,124F_5,$$

и выражения главных компонент через признаки: например,

$$F_1 = \frac{-0,947X_1 - 0,967X_2 + 0,836X_3 + 0,203X_4 + 0,882X_5}{3,35} =$$

$$= -0,283X_1 - 0,289X_2 + 0,250X_3 + 0,060X_4 + 0,263X_5.$$

Такой же результат получается при анализе матрицы нагрузок компонент на исходные признаки, которая выводится в таблице «Component Score Coefficient Matrix».

Поскольку две первые главные компоненты объясняют 86,67% общей дисперсии, снизим размерность исходных признаков, ограничившись первыми двумя главными компонентами. Установим содержательный смысл главных компонент. Первая главная компонента тесно связана со средним числом детей в семье X_1, долей молодежи среди населения X_2 (с отрицательным знаком), валовым внутренним продуктом на душу населения X_3 и процентом грамотных X_5 (с положительным знаком), поэтому ее можно интерпретировать как уровень развития страны, вторая главная компонента связана с плотностью населения X_4, она и интерпретируется как плотность населения.

Результатом регрессионного анализа (с исключением) признака Y на пять главных компонент (появившихся в результате работы программы в окне ввода исходных данных пакета SPSS под именами «fac1_1», «fac2_1», …, «fac5_1») стало уравнение

$$\hat{Y} = 64,08 + 10,43F_1,$$

т. е. первая главная компонента несет в себе достаточно информации для определения результативного признака (оценка нормированного коэффи-

циента детерминации $\hat{R}^2_{\text{норм}} = 0,69$, что не намного ниже оценки нормированного коэффициента детерминации уравнения регрессии по исходным признакам).

1.4.4. Компонентный и факторный анализ производственной деятельности предприятий

Проведем снижение размерности пространства шести показателей, применяемых в машиностроении для анализа использования производственной мощности и основных производственных фондов: коэффициента использования среднегодовой мощности по фактическому выпуску X_1, коэффициента загрузки металлорежущего оборудования по фактическому выпуску X_2, выпуска валовой продукции на один рубль основных промышленно-производственных фондов X_3, выпуска валовой продукции на один металлорежущий станок X_4, выпуска валовой продукции на один квадратный метр производственной площади основных цехов X_5, коэффициента сменности металлообрабатывающего оборудования X_6.

1. Компонентный анализ. Матрица выборочных коэффициентов парной корреляции этих показателей, вычисленная по данным 153 предприятий, такова:

$$\begin{pmatrix} 1,0000 & 0,4166 & 0,2206 & 0,1913 & 0,2573 & 0,3208 \\ 0,4166 & 1,0000 & 0,0206 & -0,0181 & 0,1994 & 0,3547 \\ 0,2206 & 0,0206 & 1,0000 & 0,2131 & 0,3076 & 0,0298 \\ 0,1913 & -0,0181 & 0,2131 & 1,0000 & 0,0724 & -0,0423 \\ 0,2573 & 0,1994 & 0,3076 & 0,0724 & 1,0000 & 0,1971 \\ 0,3208 & 0,3547 & 0,0298 & -0,0423 & 0,1971 & 1,0000 \end{pmatrix}.$$

Собственные значения этой матрицы

$$\hat{\lambda}_1 = 1,9986,\ \hat{\lambda}_2 = 1,2774,\ \hat{\lambda}_3 = 0,9084,\ \hat{\lambda}_4 = 0,6606,\ \hat{\lambda}_5 = 0,6390,\ \hat{\lambda}_6 = 0,5160.$$

Нормированный собственный вектор, соответствующий собственному значению $\hat{\lambda}_1$, имеет вид

$$\hat{\mathbf{v}}_1 = \begin{pmatrix} 0,5426 \\ 0,4657 \\ 0,3104 \\ 0,1745 \\ 0,4251 \\ 0,4255 \end{pmatrix}.$$

Оценки нагрузок показателей на компоненты и оценки долей вкладов главных компонент в суммарную дисперсию показателей приведены в табл. 1.4.2.

Таблица 1.4.2

Показатели	Оценки нагрузок на компоненты						$\sum_{j=1}^{6} \hat{a}_{ij}^2$
	F_1	F_2	F_3	F_4	F_5	F_6	
X_1	0,7671	–0,0197	–0,2754	–0,2233	–0,1387	0,5160	1
X_2	0,6584	–0,4401	–0,1610	–0,3391	0,0752	0,4267	1
X_3	0,4388	0,6459	0,3177	–0,0588	–0,4996	0,1903	1
X_4	0,2467	0,6399	–0,6331	0,2110	0,2476	0,1516	1
X_5	0,6010	0,2266	0,5524	0,0442	0,5249	–0,0690	1
X_6	0,6016	–0,4531	0,0033	0,6336	–0,1670	0,0689	1
Оценка доли вклада F_j в общую дисперсию, % $\left(\frac{\sum_{i=1}^{6} \hat{a}_{ij}^2}{6} \cdot 100\%\right)$	33,31	21,29	15,14	11,01	10,65	8,60	6

Зная оценки нагрузок, можно выразить каждый показатель через компоненты. Так, например,

$$X_1 = 0{,}7671F_1 - 0{,}0197F_2 - 0{,}2754F_3 - 0{,}2233F_4 - 0{,}1387F_5 + 0{,}5160F_6.$$

И наоборот, можно найти выражения компонент через показатели. Так,

$$F_1 = \frac{0{,}7671X_1 + 0{,}6584X_2 + 0{,}4388X_3 + 0{,}2467X_4 + 0{,}6010X_5 + 0{,}6016X_6}{1{,}9986}.$$

Так как нагрузки рассчитывались по корреляционной матрице, то в обоих выражениях все X_i и F_j центрированы и нормированы.

Из табл. 1.4.2 видно, что на первые две компоненты приходится 54,60% общей дисперсии показателей.

Дадим интерпретацию полученных результатов в рамках компонент F_1 и F_2, при этом будем иметь в виду, что нагрузка $\hat{a}_{ij}$ — это выборочный коэффициент корреляции между X_i и F_j:

- коэффициенты корреляции первой компоненты со всеми показателями положительны; судя по знакам нагрузок показателей на компоненту F_2, они делятся на две группы: с одной стороны показатели X_3, X_4, X_5, находящиеся в прямой зависимости с F_2, с другой — показатели X_6, X_2, X_1;
- компонента F_1 определяется показателем X_1 (коэффициент корреляции между ними максимальный); компонента F_2 — показателем X_3;
- наибольший вклад компоненты F_1 и F_2 вносят в дисперсию показателей X_2 (этот вклад равен $0{,}6584^2 + 0{,}4401^2 = 0{,}6271$). Таким образом, коэффициент загрузки металлорежущего оборудования X_2 является в рамках компонент F_1 и F_2 наиболее информативным среди всех шести анализируемых показателей. Второе и третье место по информативности занимают X_3 — выпуск продукции на один рубль основных фондов и X_1 — коэффициент использования производственной мощности.

В приведенном примере использовалась корреляционная матрица. Эту процедуру нельзя считать узаконенной, так как использование корреляционной матрицы создает трудности при проверке сформулированных в **п. 1.2.3** гипотез.

2. Факторный анализ. Проведем теперь факторный анализ тех же шести экономических показателей, приняв число общих факторов $m = 2$,

$$\hat{\upsilon}_i^{(0)} = 1 - \sum_{j=1}^{2} \left(\hat{a}_{ij}^{*} \right)^2 ,$$

где $\hat{a}_{i1}^{*}$ и $\hat{a}_{i2}^{*}$ — нагрузки показателя X_i на первые две компоненты. Результаты одной итерации приведены в табл. 1.4.3.

Таблица 1.4.3

$\hat{\upsilon}^{(0)}$	0,4112	0,3728	0,3903	0,5296	0,5875	0,4328
$\hat{\lambda}_1^{(0)}$	4,1433					
$\hat{\mathbf{a}}_1^{(0)}$	0,6952	0,6267	0,3835	0,1928	0,4604	0,5379
$\hat{\lambda}_2^{(0)}$	2,4759					
$\hat{\mathbf{a}}_2^{(0)}$	0,0170	–0,4266	0,5779	0,5065	0,1759	–0,3220
$\hat{\upsilon}^{(1)}$	0,5164	0,4252	0,5189	0,7063	0,7571	0,6070

Окончательные оценки нагрузок и дисперсий специфических факторов указаны в табл. 1.4.4.

Таблица 1.4.4

Показатель	Оценки нагрузок на факторы		$\sum_{j=1}^{2} \hat{a}_{ij}^2$	$\hat{\upsilon}_i = 1 - \sum_{j=1}^{2} \hat{a}_{ij}^2$
	F_1	F_2		
X_1	0,6627	–0,0519	0,4419	0,5581
X_2	0,5895	–0,3634	0,4796	0,5204
X_3	0,3967	0,5950	0,5114	0,4886
X_4	0,1749	0,2642	0,1004	0,8996
X_5	0,4463	0,1808	0,2319	0,7681
X_6	0,4618	–0,2575	0,2796	0,7204
Оценка доли вклада F_j в общую дисперсию, % $\left(\frac{\sum_{i=1}^{6} \hat{a}_{ij}^2}{6} \cdot 100\% \right)$	23,12	10,96		

Проверим справедливость гипотезы H_0 о наличии двух общих факторов. Имеем: $m = 2$, вместо ковариационной матрицы $\widehat{\mathbf{\Sigma}}_{\mathbf{X}}$ использовалась корреляционная матрица $\widehat{\mathbf{R}}_{\mathbf{X}}$. Учитывая это, в неравенстве (1.3.5)

$$n = 153, \quad k = 6, \quad m = 2,$$

$$\det|\hat{\boldsymbol{\Sigma}}_{\varepsilon}| = 0{,}0706, \quad \det|\mathbf{I} + \hat{\mathbf{J}}| = 7{,}388, \quad \det|\hat{\boldsymbol{\Sigma}}_{\mathbf{X}}| = 0{,}505,$$

$$\chi^2 = 4{,}95, \quad \chi^2_{0{,}05}(4) = 9{,}49.$$

Так как 4,95 < 9,49, то гипотезу о наличии двух общих факторов не отвергаем при $\alpha = 0{,}05$.

1.4.5. Компонентный анализ работы специалистов по окончании вуза

В работе [7] приводится пример компонентного анализа работы специалистов по окончании вуза.

Обсудим вначале место этой задачи в повышении эффективности системы управления производством. Стоимость одной аварии на сложном производстве оказывается намного выше стоимости системы управления производством и подготовки 100 специалистов для работы в этой системе. При этом 95% аварий происходят из-за ошибок персонала, а 5% — вследствие конструктивных недоработок.

Исходными данными в работе [7] послужили 17 признаков — оценок по дисциплинам, изучавшихся студентами в вузе, 18-й признак — оценка работы выпускника в течение года при текущем контроле на производстве, 19-й — оценка способностей к руководству подчиненными, 20-й признак — характеристика производственной дисциплины.

В результате проведения компонентного анализа оказалось, что первая главная компонента объясняет $33 \div 40\%$ суммарной дисперсии исходных признаков, что в четыре раза превосходит вклад второй главной компоненты.

Содержательный смысл этой первой главной компоненты сформулирован как «способность к обучению и работе на производстве». Оказалось, что оценки по некоторым предметам в вузе отрицательно коррелированы с первой главной компонентой. После анализа программ этих учебных дисциплин оказалось, что вследствие дублирования информации, получаемой студентами по различным дисциплинам, успех или неудача в изучении некоторых дисциплин нивелируется при изучении последующих курсов. Содержание других дисциплин оказалось не вполне соответствующим подготовке студентов к практической работе. Проведенный анализ позволил существенно изменить учебный процесс и повысить эффективность подготовки специалистов.

В результате проведенного анализа было также предложено вычислять индивидуальные значения первой главной компоненты для сотрудников, работающих в системе управления производством, и ранжировать этих сотрудников по первой главной компоненте, что позволит назначать на наиболее ответственные должности специалистов, обладающих наиболее высокой характеристикой по данному новому критерию.

В работе [7] приводятся и другие примеры применения компонентного и факторного анализа.

ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО РЕШЕНИЯ

1. Финансовая устойчивость предприятия характеризуется 8 показателями. Два наибольших собственных значения ковариационной матрицы равны $\lambda_1 = 6{,}0$, $\lambda_2 = 4{,}0$. Чему равен относительный вклад двух первых главных компонент?

2. Известны оценки $\hat{f}_{i1} = 0{,}661$, $\hat{f}_{i1} = -2{,}151$ главных компонент i-го наблюдения двух случайных величин X_1 и X_2 и оценки факторных нагрузок: $\hat{a}_{11} = -0{,}756$, $\hat{a}_{21} = 0{,}756$ (использовалась корреляционная матрица). Найдите значения x_{i1} и x_{i2} случайных величин X_1 и X_2, если выборочные оценки средних равны $\overline{x}_1 = 0{,}850$, $\overline{x}_2 = 0{,}877$, а выборочные оценки средних квадратичных отклонений равны $\hat{\sigma}_1 = 0{,}072$, $\hat{\sigma}_2 = 0{,}333$.

3. Используя асимптотическую нормальность распределения статистики $\sqrt{n-1}(\hat{\lambda}_j - \lambda_j)$, докажите соотношение (1.2.11).

4. В условиях компонентного анализа производственной деятельности предприятий (**п. 1.4.4**): выразите компоненту F_6 через исходные показатели и убедитесь в справедливости равенств (1.2.10); проранжируйте показатели в порядке убывания вклада компонент F_1 и F_2 в дисперсии показателей.

5. Запишите модель факторного анализа и систему уравнений (1.3.2) при $k = 5$ и $m = 2$; сколько уравнений и неизвестных в этой системе. Приведите пример ортогональной матрицы $\mathbf{V} \in \mathbb{R}^{2\times 2}$; убедитесь в неединственности решения системы.

6. Чему равны вклады общих факторов и специфического фактора в дисперсию случайной величины $X_i = 0{,}5F_1 + 0{,}9F_2 + \overset{\circ}{\varepsilon}_i$, если X_i центрирована и нормирована. Какова общность случайной величины X_i?

7. Найдите суммарную общность пяти случайных величин и долю этой общности, вносимую каждым из факторов, если матрица нагрузок

$$\mathbf{A} = \begin{pmatrix} 0{,}7 & 0{,}3 \\ 0{,}8 & 0{,}1 \\ 0{,}7 & 0{,}5 \\ 0{,}6 & 0{,}5 \\ 0{,}7 & 0{,}0 \end{pmatrix}$$

рассчитана по корреляционной матрице $\mathbf{R_X}$ (знаки нагрузок не указаны).

8. Продолжите итерационную процедуру метода максимального правдоподобия в условиях факторного анализа производственной деятельности предприятий (**п. 1.4.4**).

9. Проведите факторный анализ показателей мировой демографической статистики по данным табл. 1.4.1; при интерпретации общих факторов при необходимости используйте подходящую процедуру вращения факторного пространства.

10. Приведите алгоритм оценки общих факторов методом Бартлетта.

ГЛАВА 2. МЕТОДЫ КЛАССИФИКАЦИИ МНОГОМЕРНЫХ НАБЛЮДЕНИЙ

§ 2.1. СУЩНОСТЬ ЗАДАЧ КЛАССИФИКАЦИИ

Очень часто при исследовании больших совокупностей объектов необходимо выявить объекты, б л и з к и е между собой в определенном смысле.

Приведем несколько п р и м е р о в:

- производитель товара массового потребления не ориентируется, как правило, на каждого конкретного потенциального покупателя, а сегментирует рынок, разделяя покупателей на группы (или классы), внутри которых покупатели похожи;
- банкам удобно классифицировать заемщиков по уровню кредитоспособности, определяемом большим числом различным показателей, поскольку при решении о кредитовании необходимо для каждого заемщика установить лимит кредита, и удобнее задавать не индивидуальный лимит для каждого конкретного заемщика, а лимиты для групп сходных, близких между собой заемщиков; в свою очередь, регулирующие организации классифицируют банки по уровню надежности;
- при сравнении стран, регионов и городов по уровню жизни, продукции различных производителей — по качеству, семей — по структуре потребления и т. п. оказывается удобным отождествлять близкие друг к другу объекты.

Приведенные примеры иллюстрируют **сущность задач классификации** многомерных наблюдений, которая заключается в разбиении большого числа объектов на однородные группы.

Целью методов классификации является исследование внутренней структуры изучаемой системы n объектов, каждый из которых характеризуется m-мерным вектором признаков, «сжатие» этой системы без существенной потери содержащейся в ней информации путем выявления классов сходных между собой объектов и отождествления объектов внутри каждого класса.

К л а с т е р н ы й а н а л и з решает задачу классификации объектов при отсутствующей априорной информации о наблюдениях внутри классов, в д и с к р и м и н а н т н о м а н а л и з е предполагается наличие такой информации.

Среди **прикладных задач**, решаемых указанными методами, отметим следующие:

- проведение классификации объектов с учетом большого числа признаков, отражающих их природу;

- выявление структуры изучаемой совокупности объектов;
- снижение объема выборки путем отождествления каждого класса объектов с его типичным представителем;
- снижение размерности пространства путем отождествления близких признаков (в этом случае классифицируются не объекты, а признаки, в многомерном статистическом анализе этот прием перемены объектов и признаков местами, как уже указывалось ранее, часто приводит к интересным результатам).
- построение отдельных регрессионных моделей в каждом классе объектов; если дисперсия остатков регрессии сильно изменяется от класса к классу (такое явление называется *гетероскедастичностью*), то построение регрессии для каждого класса может помочь избавиться от гетероскедастичности.

Практика показывает, что классификация объектов исследования, проведенная по факторам или главным компонентам, полученным в результате снижения размерности исходного пространства признаков, оказывается более объективной, чем классификация тех же объектов по исходным признакам.

§ 2.2. Кластерный анализ

2.2.1. Меры однородности объектов

Методы *кластерного анализа* позволяют разбить изучаемую совокупность объектов на группы однородных в некотором смысле объектов, называемых *кластерами* или *классами*. Наибольшее распространение получили два подхода к задаче классификации: эвристический, реализующий некоторую схему разделения объектов на классы, исходя из интуитивных соображений, и экстремальный, реализующий схему разделения на основе заданного критерия оптимальности. Наиболее трудным в задаче классификации является определение меры однородности объектов.

Пусть каждый из исходных n объектов с номерами 1, 2, …, n задается как точка

$$\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{km}) \qquad (k = 1, 2, \ldots, n)$$

в m-мерном пространстве признаков $X_1, X_2, \ldots, X_m$. Совокупность этих точек можно трактовать как выборку объема n из многомерной генеральной совокупности $\mathbf{X} = (X_1, X_2, \ldots, X_m)$.

Замечание. Результаты классификации зависят от выбора масштаба и единиц изменения признаков. Чтобы исправить такое положение, прибегают к стандартной нормировке признаков:

$$x_{ki} \to \frac{(x_{ki} - \bar{x}_{\cdot i})}{\hat{\sigma}_{X_i}}.$$

Однако эта операция, уравнивая разделительные возможности всех признаков, может привести и к нежелательным последствиям.

В случае зависимых признаков и их различной значимости при классификации объектов за меру однородности объектов принимают *расстояние Махалонобиса*

$$l_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)\mathbf{\Lambda}\mathbf{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}} \quad (i = 1,2,...,n, \quad j = 1,2,...,n), \quad (2.2.1)$$

где

$$\mathbf{\Lambda} \in \mathbb{R}^{m \times m} —$$

симметричная (чаще всего диагональная) неотрицательно определенная матрица «весовых» коэффициентов признаков, $\mathbf{\Sigma}$ — ковариационная матрица генеральной совокупности, из которой извлекаются объекты. Частными случаями формулы (2.2.1) являются:

- *евклидово расстояние*

$$l_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}} = \sum_{k=1}^{m}(x_{ik} - x_{jk})^2 \quad (i = 1,2,...,n, \ \ j = 1,2,...,n), \quad (2.2.2)$$

использование которого оправдано, если генеральная совокупность распределена по многомерному нормальному закону с ковариационной матрицей

$$\mathbf{\Sigma} = \sigma^2\mathbf{I},$$

а признаки однородны по своему физическому смыслу и одинаково «весомы» с точки зрения решения вопроса об отнесении объекта к тому или иному кластеру;

- *взвешенное евклидово расстояние*

$$l_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)\mathbf{\Lambda}(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}} \quad (i = 1,2,...,n, \ \ j = 1,2,...,n); \quad (2.2.3)$$

- *хеммингово расстояние* (block distance)

$$l_{ij} = \sum_{k=1}^{m} | x_{ik} - x_{jk} | \quad (i = 1,2,...,n, \ \ j = 1,2,...,n), \quad (2.2.4)$$

в случае дихотомических признаков хеммингово расстояние l_{ij} — это число несовпадений значений соответствующих признаков на i-м и j-м объектах; это же расстояние можно использовать и для количественных признаков, при этом его называют *расстоянием городских кварталов* (или, в шутку, *дистанцией манхеттенского таксиста*), в этом случае l_{ij} — это не кратчайшее расстояние между двумя точками, а «путь, который должен преодолеть таксист, чтобы проехать от одной точки до другой по городским улицам, пересекающимся, как в Манхеттене, под прямым углом»).

Часто используются и другие формулы для вычисления расстояний между объектами:

- *расстояние Чебышёва*

$$l_{ij} = \max_{1 \leqslant k \leqslant m} | x_{ik} - x_{jk} | \quad (i = 1,2,...,n,\ \ j = 1,2,...,n); \qquad (2.2.5)$$

- *обобщенное расстояние Колмогорова — Минковского* (power distance)

$$l_{ij} = \left(\sum_{k=1}^{m} | x_{ik} - x_{jk} |^{p} \right)^{\frac{1}{r}} \quad (i = 1,2,...,n,\ \ j = 1,2,...,n), \qquad (2.2.6)$$

которое в качестве частных случаев включает в себя все рассмотренные выше виды расстояний.

Для оценки однородности объектов в пространстве $m = k + (m - k)$ разнотипных признаков (где k — число дихотомических признаков, к которым могут быть сведены и ранговые, и номинальные признаки, а $(m - k)$ — число количественных признаков), обычно используют не расстояние между объектами, а меру их сходства.

Сходство i-го и j-го объектов, например, можно рассчитать так:

$$r_{ij} = \frac{k}{m} r'_{ij} + \frac{(m-k)}{m} r''_{ij} \qquad (i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n),$$

где

$$r'_{ij} = \frac{a_{ij}}{k}$$

(a_{ij} — число дихотомических признаков, значение которых совпадают у объектов с номерами i и j),

$$r''_{ij} = \frac{\overline{l}}{\overline{l} + l_{ij}}$$

(l_{ij} — расстояние Махалонобиса в пространстве количественных признаков, $\overline{l}$ — среднее значение расстояния по всем парам объектов).

К мерам однородности (расстоянию l_{ij} и сходству r_{ij}) предъявляются следующие требования:

- *симметрии:*

$$l_{ij} = l_{ji}, \quad r_{ij} = r_{ji} \qquad (i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n);$$

- *максимального сходства объекта с самим собой:*

$$r_{ii} = \max_{1 \leqslant j \leqslant n} r_{ij} \qquad (i = 1, 2, \ldots, n);$$

- *монотонного убывания* r_{ij} *по* l_{ij}*:*

$$l_{kp} \geqslant l_{ij} \;\Leftrightarrow\; r_{kp} \leqslant r_{ij} \quad (i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n).$$

З а м е ч а н и е 1. В некоторых задачах классификации объектов мера однородности объектов содержательно интерпретируема. Например, при классификации n отраслей экономии используется матрица межотраслевого баланса $\mathbf{S} = (s_{ij}) \in \mathbb{R}^{n \times n}$, где s_{ij} — годовые поставки i-й отрасли в j-ю (в денежном выражении); здесь можно принять

$$r_{ij} = \frac{\dfrac{s_{ij}}{\sum\limits_{j=1}^{n} s_{ij}} + \dfrac{s_{ji}}{\sum\limits_{i=1}^{n} s_{ij}}}{2} \quad (i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n).$$

З а м е ч а н и е 2. Методы кластерного анализа могут использоваться и для к л а с с и ф и к а ц и и п р и з н а к о в. В качестве мер однородности признаков обычно используются характеристики степени их коррелированности (коэффициенты корреляции; коэффициенты, основанные на критерии χ^2 и т. д.). Задача классификации признаков может иметь как самостоятельный интерес (помимо кластерного анализа эта задача может решаться методами факторного и компонентного анализа), так предшествовать задаче классификации объектов в пространстве большого числа признаков: располагая кластерами — группами однородных признаков, можно в каждой группе выделить по одному представителю — признаку и классификацию объектов провести в пространстве представителей. При большом числе исходных признаков классификацию объектов можно провести также в пространстве главных компонент или в пространстве общих факторов.

З а м е ч а н и е 3. Если известен способ вычисления расстояний l_{ij}, то очень легко ввести естественную меру сходства:

$$r_{ij} = \frac{\overline{l}}{\overline{l} + l_{ij}} \quad (i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n).$$

Точно так же можно определить расстояние, зная сходство.

2.2.2. Р а с с т о я н и я м е ж д у к л а с т е р а м и

Пусть известна матрица

$$\mathbf{L} = (l_{ij}) \in \mathbb{R}^{m \times m}$$

расстояний между n объектами и некоторое их разбиение

$$\mathfrak{R}(p) = \{\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_p\}$$

на p кластеров.

Основным понятием *кластер-процедур* является р а с с т о я н и е

$$\rho_{st} = \rho(\mathcal{K}_s, \mathcal{K}_t)$$

м е ж д у к л а с т е р а м и $\mathcal{K}_s$ и $\mathcal{K}_t$ $(s = 1, 2, \ldots, p,\ t = 1, 2, \ldots, p)$.

Наиболее распространены следующие способы измерения расстояний между кластерами:

- *метод ближнего соседа:*

$$\rho_{st} = \min_{i \in \mathcal{K}_s, j \in \mathcal{K}_t} l_{ij} \quad (s = 1, 2, \ldots, p,\ t = 1, 2, \ldots, p) \tag{2.2.7}$$

(в зарубежной литературе этот метод чаще называют *методом одиночной связи*, single linkage);

- *метод дальнего соседа* (*метод полных связей*, complete linkage):

$$\rho_{st} = \max_{i \in \mathcal{K}_s, j \in \mathcal{K}_t} l_{ij} \quad (s = 1, 2, \ldots, p,\ t = 1, 2, \ldots, p); \tag{2.2.8}$$

Замечание. Объяснение смысла названий «одиночная связь», «полная связь» будет приведено на с. 52 ~ 53.

- *метод средней связи* (pair group average):

$$\rho_{st} = \frac{\sum_{i \in \mathcal{K}_s} \sum_{j \in \mathcal{K}_t} l_{ij}}{n_s n_t} \quad (s = 1, 2, \ldots, p,\ t = 1, 2, \ldots, p), \tag{2.2.9}$$

где n_s и n_t — количество объектов соответственно в кластерах $\mathcal{K}_s$ и $\mathcal{K}_t$;

- *центроидный метод* (измеряется расстояние между «центрами тяжести» кластеров):

$$\rho_{st} = l\left(\overline{x}_{\mathcal{K}_s}, \overline{x}_{\mathcal{K}_t}\right) \quad (s = 1, 2, \ldots, p,\ t = 1, 2, \ldots, p), \tag{2.2.10}$$

где

$$\overline{x}_{\mathcal{K}_s} = \frac{\sum_{i \in \mathcal{K}_s} x_i}{n_s} \ —$$

среднее арифметическое векторных наблюдений x_i при $i \in \mathcal{K}_s$.

2.2.3. Иерархические агломеративные методы

Рассмотренные методы вычисления расстояний между кластерами относятся к группе иерархических (т. е. деревообразующих) агломеративных (т. е. объединяющих) методов.

Иерархические агломеративные методы — это многошаговые методы классификации, работающие по следующему **алгоритму**.

- на нулевом шаге за разбиение принимается исходная совокупность n элементарных кластеров, матрица расстояний между которыми

$$(\rho_{ij}) = (l_{ij}) \in \mathbb{R}^{n \times n};$$

- на каждом следующем шаге происходит объединение (в соответствии с эвристическим или экстремальным подходом) двух кластеров $\mathcal{K}_s$ и $\mathcal{K}_t$, сформированных на предыдущем шаге, в один кластер

$$\mathcal{K}_{s\oplus t} = \mathcal{K}_s \cup \mathcal{K}_t,$$

при этом размерность матрицы расстояний уменьшается, по сравнению с размерностью этой матрицы на предыдущем шаге, на единицу.

При использовании рассмотренных в предыдущем пункте агломеративных методов рассчитать расстояние $\rho_{i,\,s\oplus t}$ между кластерами $\mathcal{K}_i$ и $\mathcal{K}_{s\oplus t}$ ($i \neq s$, $i \neq t$) можно, используя соответствующую методу формулу расстояния между кластерами, однако менее трудоемки расчеты по формуле:

$$\rho_{i,s\oplus t} = \alpha\rho_{is} + \beta\rho_{it} + \gamma\rho_{st} + \delta\,|\,\rho_{is} - \rho_{it}\,|, \qquad (2.2.11)$$

в которой значения коэффициентов α, β, γ, δ зависят от используемого метода определения расстояний между кластерами (табл. 2.2.1).

Таблица 2.2.1

Метод	α	β	γ	δ	$\rho_{i,\,s\oplus t}$
ближнего соседа	0,5	0,5	0	–0,5	$\min(\rho_{is},\rho_{it}) = \min\limits_{p\in\mathcal{K}_i,\,j\in\mathcal{K}_{s\oplus t}} l_{pj}$ (2.2.12)
дальнего соседа	0,5	0,5	0	0,5	$\max(\rho_{is},\rho_{it}) = \max\limits_{p\in\mathcal{K}_i,\,j\in\mathcal{K}_{s\oplus t}} l_{pj}$ (2.2.13)
средней связи	$\frac{n_s}{n_s+n_t}$	$\frac{n_s}{n_s+n_t}$	0	0	$\frac{n_s\rho_{is}+n_t\rho_{it}}{n_s+n_t} = \frac{1}{n_i n_{s\oplus t}}\sum\limits_{p\in\mathcal{K}_i}\sum\limits_{j\in\mathcal{K}_{s\oplus t}} l_{pj}$ (2.2.14)
центроидный	$\frac{n_s}{n_s+n_t}$	$\frac{n_s}{n_s+n_t}$	$-\alpha\beta$	0	$\frac{n_s\rho_{is}+n_t\rho_{it}}{n_s+n_t} - \frac{n_s n_t \rho_{st}}{(n_s+n_t)^2} = l\left(\overline{x}_{\mathcal{K}_i}, \overline{x}_{\mathcal{K}_{s\oplus t}}\right)$ (2.2.15)

В последнем столбце табл. 2.2.1 для каждого метода определения расстояний между кластерами слева приведена формула подсчета $\rho_{i,\,s\oplus t}$, следующая из (2.2.11), а справа — формула, следующая из определения соответствующего метода, рассмотренного в предыдущем пункте.

2.2.4. Параллельные кластер-процедуры. Методы, связанные с функционалами качества разбиения

Иерархические методы используются обычно в таких задачах классификации небольшого числа объектов (порядка нескольких десятков), где больший интерес представляет не число кластеров, а анализ структуры множества этих объектов и наглядная интерпретация проведенного анализа в виде дендрограммы. Если же число кластеров заранее задано или подлежит определению, то для классификации чаще всего используют *параллельные кластер-процедуры* — итерационные алгоритмы, на каждом шаге которых используются одновременно (параллельно) все наблюдения. Коль скоро эти алгоритмы на каждом шаге работают со всеми наблюдениями, то основной целью их конструирования является нахождение способов сокращения числа перебора вариантов (даже при числе наблюдений порядка нескольких десятков), что приводит зачастую лишь к приближенному, но не слишком трудоемкому решению задач. В параллельных кластер-процедурах реализуется обычно идея оптимизации разбиения в соответствии с некоторым функционалом качества.

З а м е ч а н и е . В рассмотренных ранее иерархических методах реализованы схемы объединения объектов на основе эвристических соображений. Однако, например, для методов ближнего и дальнего соседа можно указать функционалы качества $f(\mathfrak{R})$ разбиения $\mathfrak{R}$:

$$f(\mathfrak{R}) = \max_{p} \min_{i,j \in \mathcal{K}_p} l_{ij} \quad — \tag{2.2.16}$$

для метода ближнего соседа;

$$f(\mathfrak{R}) = \max_{p} \max_{i,j \in \mathcal{K}_p} l_{ij} \quad — \tag{2.2.17}$$

для метода дальнего соседа.

Докажем, что (2.2.16) соответствует расстоянию (2.2.7). Пусть $f(\mathfrak{R}_{s,t})$ — качество разбиения $\mathfrak{R}_{s,t}$, полученного из $\mathfrak{R}$ после объединения в нем кластеров $\mathcal{K}_s$ и $\mathcal{K}_t$; тогда

$$f_{st} = f(\mathfrak{R}_{s,t}) - f(\mathfrak{R}) = \max\left\{ \min_{i \in \mathcal{K}_s, j \in \mathcal{K}_t} l_{ij}, \max_{p \neq s,t} \min_{i,j \in \mathcal{K}_p} l_{ij} \right\} - f(\mathfrak{R}) =$$

$$= \begin{cases} 0, & \text{если } \min\limits_{i \in K_s, j \in K_t} l_{ij} \leqslant f(\mathfrak{R}), \\ \min\limits_{i \in \mathcal{K}_s, j \in \mathcal{K}_t} l_{ij} - f(\mathfrak{R}), & \text{если } \min\limits_{i \in \mathcal{K}_s, j \in \mathcal{K}_t} l_{ij} > f(\mathfrak{R}). \end{cases}$$

Поскольку для начального разбиения $\mathfrak{R}_0$ качество $f(\mathfrak{R}_0) = 0$, то величиной (2.2.7) определяется качество нового разбиения с точки зрения критерия (2.2.16), что и требовалось доказать.

Наиболее распространенными при заданном числе k кластеров являются следующие функционалы качества разбиения:

- *сумма внутриклассовых дисперсий*

$$f(\mathfrak{R}) = \sum_{p=1}^{k} \sum_{i \in \mathcal{K}_p} l^2(\mathbf{x}_i, \overline{\mathbf{x}}_{K_p}); \tag{2.2.18}$$

- *сумма попарных внутриклассовых расстояний*

$$f(\mathfrak{R}) = \sum_{p=1}^{k} \frac{1}{n_p} \sum_{i,j \in \mathcal{K}_p} l^2(\mathbf{x}_i, \mathbf{x}_j), \tag{2.2.19}$$

а при неизвестном числе кластеров —

- функционал

$$f(\mathfrak{R}) = \alpha f_1(\mathfrak{R}) + \beta f_2(\mathfrak{R});$$

- функционал

$$f(\mathfrak{R}) = \left[f_1(\mathfrak{R}) \right]^{\alpha} \left[f_2(\mathfrak{R}) \right]^{\beta},$$

где $f_1(\mathfrak{R})$ — некоторая монотонно невозрастающая функция числа классов, характеризующая средний внутриклассовый разброс наблюдений, $f_2(\mathfrak{R})$ — некоторая монотонно неубывающая функция числа классов, характеризующая взаимную удаленность классов или «меру концентрации» наблюдений.

Схема работы алгоритмов, связанных с функционалами качества, такова:

- для некоторого начального разбиения $\mathfrak{R}_0$ вычисляют значение $f(\mathfrak{R}_0)$;
- затем каждую из точек $\mathbf{x}_i$ поочередно перемещают во все кластеры и оставляют в том положении, которое соответствует наилучшему значению функционала качества;
- работу заканчивают, когда перемещение точек не дает улучшения качества.

Часто описанный алгоритм применяют несколько раз, начиная с разных начальных разбиений $\mathfrak{R}_0$, и выбирают наилучший вариант разбиения.

Остановимся на показателе качества разбиения и алгоритме «объединение», предложенных в работах [16, 20]. Это связано с тем, что к задаче нахождения разбиения, оптимизирующего этот показатель качества, можно свести задачу оптимизации организационной структуры производственной системы.

Пусть $\mathfrak{R}$ — некоторое разбиение совокупности n элементов с матрицей связи

$$(r_{ij}) \in \mathbb{R}^{n \times n},$$

а r_0 — некоторое пороговое значение связи. Связь $r_{s,t}$ кластеров $\mathcal{K}_s$ и $\mathcal{K}_t$ определим как

$$r_{s,t} = r'(\mathcal{K}_s, \mathcal{K}_t) = \sum_{i \in \mathcal{K}_s} \sum_{j \in \mathcal{K}_t} (r_{ij} - r_0) \qquad (2.2.20)$$

$$(s = 1, 2, \ldots, p,\ \ t = 1, 2, \ldots, p),$$

а в качестве критерия оптимальности разбиения выберем

$$f(\mathfrak{R}) = \sum_{p} \sum_{\substack{i,j \in \mathcal{K}_p^* \\ i \neq j}} (r_{ij} - r_0) \to \max. \qquad (2.2.21)$$

Нетрудно убедиться в том, что оптимальное разбиение $\mathfrak{R}^*$ обладает следующими свойствами «хорошей» классификации:

- оптимальное разбиение $\mathfrak{R}^*$, максимизируя сумму связей между элементами внутри образующих его кластеров, минимизирует сумму связей между элементами разных кластеров:

$$\sum_{s \neq t} \sum_{i \in \mathcal{K}_s^*} \sum_{j \in \mathcal{K}_t^*} (r_{ij} - r_0) = \min_{\{\mathfrak{R}\}} \sum_{m \neq l} \sum_{i \in \mathcal{K}_m} \sum_{j \in \mathcal{K}_l} (r_{ij} - r_0); \qquad (2.2.22)$$

- средняя связь между элементами, образующими кластер $\mathcal{K}_s^*$ разбиения $\mathfrak{R}^*$, не меньше средней связи элементов этого кластера с элементами любого другого кластера $\mathcal{K}_t^*$ $(t \neq s)$ разбиения $\mathfrak{R}^*$:

$$\frac{1}{n_s(n_s-1)}\sum_{\substack{i,j\in\mathcal{K}_s^*\\ i\neq j}}(r_{ij}-r_0)\geqslant\frac{1}{n_s n_t}\sum_{i\in\mathcal{K}_s^*}\sum_{j\in\mathcal{K}_t^*}(r_{ij}-r_0) \qquad (2.2.23)$$

$$(s=1,2,\ldots,p,\ t=1,2,\ldots,p,\ t\neq s);$$

- средняя связь между элементами, образующими кластер $\mathcal{K}_s^*$ разбиения $\mathfrak{R}^*$ не меньше средней связи этих элементов с элементами, не вошедшими в $\mathcal{K}_s^*$:

$$\frac{1}{n_s(n_s-1)}\sum_{\substack{i,j\in\mathcal{K}_s^*\\ i\neq j}}(r_{ij}-r_0)\geqslant\frac{1}{n_s(n-n_s)}\sum_{i\in\mathcal{K}_s^*}\sum_{j\notin\mathcal{K}_s^*}(r_{ij}-r_0) \qquad (2.2.24)$$

$$(s=1,2,\ldots,p).$$

Каков смысл порога r_0? Проведя тождественные преобразования (2.2.21), получим:

$$f(\mathfrak{R})=\sum_p\sum_{\substack{i,j\in\mathcal{K}_p\\ i\neq j}}r_{ij}-r_0\sum_p n_p^2+r_0 n\to\max,$$

т. е. максимизация $f(\mathfrak{R})$ достигается за счет максимизации суммы связей внутри кластеров (без учета порога r_0) и минимизации $\sum_p n_p^2$ (при условии $\sum_p n_p=n$). Но минимум $\sum_p n_p^2$ достигается при $n_1=n_2=n_3=\cdots$, т. е. при р а в н о м е р н о м распределении элементов по кластерам. Поэтому порог r_0 — это в е с т р е б о в а н и я р а в н о м е р н о с т и р а с п р е д е л е н и я элементов по кластерам в сравнении с требованием максимизации суммы связей внутри кластеров.

Алгоритм «объединение» является с у б о п т и м а л ь н ы м алгоритмом, использующим необходимое условие оптимальности разбиения по критерию (2.2.21): для оптимального разбиения $\mathfrak{R}^*$

$$\begin{cases} r'(\mathcal{K}_s^*,\mathcal{K}_t^*)=\sum_{i\in\mathcal{K}_s^*}\sum_{j\in\mathcal{K}_t^*}(r_{ij}-r_0)\leqslant 0, & s=1,2,\ldots,p,\ t=1,2,\ldots,p,\ t\neq s,\\ r'(\mathcal{K}_s^*,\mathcal{K}_s^*)=\sum_{\substack{i,j\in\mathcal{K}_s^*\\ i\neq j}}(r_{ij}-r_0)\geqslant 0, & s=1,2,\ldots,p. \end{cases} \qquad (2.2.25)$$

Алгоритм работает следующим образом:

- заменяют r_{ij} при $i\neq j$ на $r'_{ij}=r_{ij}-r_0$, а r_{ii} на $r'_{ii}=0$;
- за начальное принимают разбиение, состоящее из n одноэлементных кластеров;
- объединяют кластеры $\mathcal{K}_s$ и $\mathcal{K}_t$ $(s\neq t)$, связь $r'(\mathcal{K}_s,\mathcal{K}_t)$ между которыми максимальна (объединение сводится к суммированию строк (столбцов) с номерами s и t);

- процесс объединения кластеров прекращается, если в матрице связей между кластерами все связи $\rho'(\mathcal{K}_l, \mathcal{K}_m) \leqslant 0 \quad (l \neq m)$.

Рассмотрим *задачу оптимизации структуры некоторой производственной системы*. Пусть имеется n элементарных систем. Требуется найти такое разбиение

$$\mathfrak{R}(p) = \{\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_p\}$$

на p кластеров, которое минимизировано бы сумму затрат на управление внутри кластеров (они пропорциональны количеству $\sum_{s=1}^{p} n_s(n_s - 1)$ «внутренних связей») и затрат на управление «внешними связями» (они пропорциональны количеству «внешних связей», равному $\sum_{s=1}^{p} \sum_{i \in \mathcal{K}_q} \sum_{j \notin \mathcal{K}_q} r_{ij}$):

$$F(\mathfrak{R}) = \lambda \sum_{q=1}^{p} n_q(n_q - 1) + \lambda' \sum_{q=1}^{p} \sum_{i \in \mathcal{K}_q} \sum_{j \notin \mathcal{K}_q} r_{ij} \to \min. \qquad (2.2.26)$$

Но так как

$$n_q(n_q - 1) = \sum_{\substack{i,j \in \mathcal{K}_q \\ i \neq j}} 1,$$

а

$$\lambda' \sum_{q=1}^{p} \sum_{i \in \mathcal{K}_q} \sum_{j \notin \mathcal{K}_q} r_{ij} = \lambda' \sum_{\substack{i,j=1 \\ i \neq j}}^{n} r_{ij} - \lambda' \sum_{q=1}^{p} \sum_{\substack{i,j \in \mathcal{K}_q \\ i \neq j}} r_{ij},$$

где уменьшаемое не зависит от разбиения, то критерий (2.2.26) равносилен критерию

$$f(\mathfrak{R}) = \sum_{q=1}^{p} \sum_{\substack{i,j \in \mathcal{K}_q \\ i \neq j}} \left(r_{ij} - \frac{\lambda}{\lambda'} \right) \to \max,$$

идентичному (2.2.21) при

$$r_0 = \frac{\lambda}{\lambda'}.$$

2.2.5. Последовательные кластер-процедуры. Метод *K*-средних

Иерархические и параллельные кластер-процедуры практически реализуемы лишь в задачах классификации не более нескольких десятков наблюдений. К решению задач с бóльшим числом наблюдений применяют *последовательные кластер-процедуры* — итерационные алгоритмы, на каждом

шаге которых используется одно наблюдение (или небольшая часть исходных наблюдений) и результаты разбиения на предыдущем шаге.

И д е ю этих процедур поясним на примере *метода K-средних* (*K*-Means Clustering) с заранее заданным числом k классов:

- на нулевом шаге за центры искомых k кластеров принимают случайно выбранные k наблюдений — точки $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_k$; каждому кластеру присваивают единичный вес;
- на первом шаге находят расстояния от точки $\mathbf{x}_{k+1}$ до центров кластеров, построенных на предыдущем шаге, и точку $\mathbf{x}_{k+1}$ относят к кластеру, расстояние до которого минимально, после чего рассчитывают новый центр тяжести этого кластера (как взвешенное среднее по каждому показателю) и вес кластера увеличивают на единицу; все остальные кластеры остаются неизмененными (с прежними центрами и весами);
- на втором шаге аналогичную процедуру выполняют для точки $\mathbf{x}_{k+2}$ и т. д.

При достаточно большом числе n классифицируемых объектов или достаточно большом числе итераций пересчет центров тяжести практически не приводит к их изменению.

Если в какой-то точке не удается, рассмотрев все $n = k + (n - k)$ точек, достичь практически не изменяющихся центров тяжести, то либо, используя получившееся разбиение n точек на k кластеров в качестве начального, применяют изложенную процедуру к точкам $\mathbf{x}_1$, $\mathbf{x}_2$, …; либо в качестве начального разбиения принимают различные комбинации k точек из исходных n, а в качестве окончательного берут наиболее часто встречающееся финальное разбиение.

§ 2.3. ДИСКРИМИНАНТНЫЙ АНАЛИЗ И РАСЩЕПЛЕНИЕ СМЕСЕЙ ВЕРОЯТНОСТНЫХ РАСПРЕДЕЛЕНИЙ

2.3.1. Математическая модель дискриминантного анализа

Ограничимся рассмотрением *классической модели дискриминантного анализа*, не затрагивая вопросов статистического оценивания его результатов (эти вопросы обсуждаются, например, в [3]).

Пусть результатом наблюдения над объектом является реализация

$$\mathbf{x}^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, ..., x_m^{(0)} \right)$$

m-мерного случайного вектора

$$\mathbf{X} = (X_1, X_2, \ldots, X_m).$$

Известно, что этот объект относится к одной из l генеральных совокупностей, к одному из l классов K_j ($j = 1, 2, \ldots, l$), относительно которых предполагается, что:

- каждый класс $\mathcal{K}_j$ имеет m-мерное нормальное распределение:

$$\mathbf{X}(j) \sim \mathcal{N}(a_j, \mathbf{\Sigma}) \quad (j = 1, 2, \ldots, l),$$

где $\mathbf{X}^{(j)}$ — обозначение вектора $\mathbf{X}$ для класса $\mathcal{K}_j$, $\mathbf{a}_j = \mathbf{M}\mathbf{X}^{(j)}$, $\mathbf{\Sigma}$ — ковариационная матрица вектора $\mathbf{X}$, о д н а и т а ж е для всех l классов;

- каждый класс $\mathcal{K}_j$ представлен выборкой объема n_j (эти выборки называют *обучающими*).

Требуется построить *правило дискриминации* — правило распознавания класса, к которому относится не вошедший в выборки объект $\mathbf{x}^{(0)}$.

Поясним идею формирования такого правила для случая одномерной случайной величины X ($m = 1$) и двух классов ($l = 2$), предположив, что удельный вес объектов каждого класса в общей генеральной совокупности одинаков.

На рис. 2.3.1 изображены графики функций плотностей $f_1(x)$ и $f_2(x)$ нормальных случайных величин, различающихся только математическими ожиданиями a_1 и a_2.

Пусть d — некоторая точка на оси Ox, и правило дискриминации таково: классифицируемый объект относят к первому классу тогда и только тогда, когда

$$x \leqslant d$$

(где x — значение случайной величины X у данного объекта), и ко второму классу во всех остальных случаях.

Поступая таким образом, можно допустить о ш и б к и д в у х в и д о в:

- объект, принадлежащий к первому классу, будет отнесен ко второму; вероятность этой ошибки

$$p_1 = \mathbf{P}\{X > d \mid X \sim \mathcal{N}(a_1, \sigma)\} = \int_d^{\infty} f_1(x)dx = 1 - \int_{-\infty}^{d} f_1(x)dx;$$

- объект, принадлежащий ко второму классу, будет отнесен к первому; вероятность этой ошибки

$$p_2 = \mathbf{P}\{X \leqslant d \mid X \sim \mathcal{N}(a_2, \sigma)\} = \int_{-\infty}^{d} f_2(x)dx.$$

Точку d найдем как решение следующей э к с т р е м а л ь н о й з а д а ч и: при условии равенства вероятностей

$$p_1 = p_2,$$

что равносильно условию

$$\int_{-\infty}^{d} [f_1(x) + f_2(x)]dx = 1, \qquad (2.3.1)$$

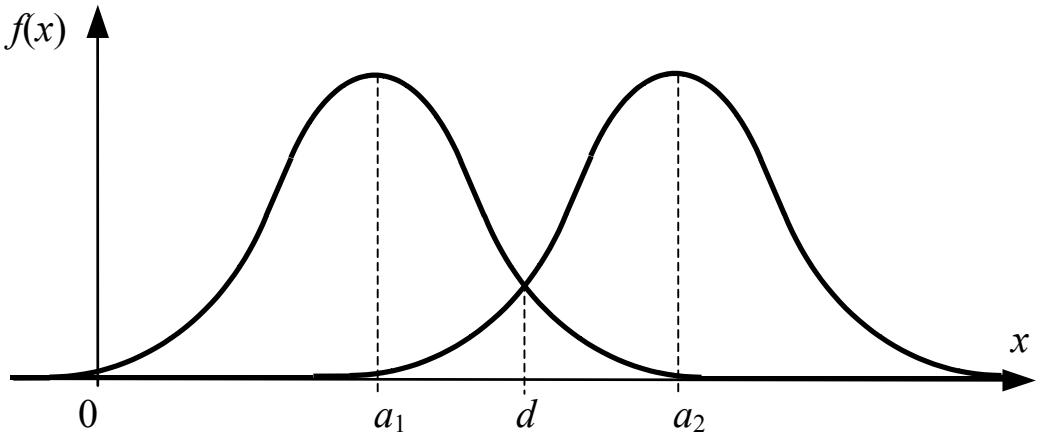


Рис. 2.3.1

требуется минимизировать вероятность p_2:

$$\int_{-\infty}^{d} f_2(x)dx \to \min . \qquad (2.3.2)$$

Используя для решения задачи (2.3.1) ~ (2.3.2) метод множителей Лагранжа, введем функцию Лагранжа

$$\mathcal{L}(d;\lambda) = \int_{-\infty}^{d} f_2(x)dx - \lambda\left(\int_{-\infty}^{d}[f_1(x)+f_2(x)]dx - 1\right)$$

и запишем необходимые условия экстремума:

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial d} = f_2(d) - \lambda[f_1(d)+f_2(d)] = 0, \\ \dfrac{\partial \mathcal{L}}{\partial \lambda} = -\displaystyle\int_{-\infty}^{d}[f_1(x)+f_2(x)]dx + 1 = 0, \end{cases}$$

откуда

$$\begin{cases} \dfrac{f_1(d)}{f_2(d)} = \dfrac{1-\lambda}{\lambda}, \\ \displaystyle\int_{-\infty}^{d}[f_1(x)+f_2(x)]dx = 1. \end{cases}$$

Таким образом, задача (2.3.1) ~ (2.3.2) равносильна системе, включающей уравнение (2.3.1) — требование равенства вероятностей ошибок — и уравнение

$$\frac{f_1(d)}{f_2(d)} = \text{const}$$

(где константа не зависит от d), которое с учетом нормальности распределений и равенства дисперсий равносильно уравнению

$$\left[x-\frac{1}{2}(a_1+a_2)\right](a_1-a_2)\frac{1}{\sigma^2}=C\,, \qquad (2.3.3)$$

где C — некоторая постоянная величина, не зависящая от d.

В рассматриваемом тривиальном случае из (2.3.1) следует, что

$$d=\frac{a_1+a_2}{2}\,.$$

Подставив $x = d$ в (2.3.3), получим $C = 0$. Таким образом, сформулированное выше классификационное правило эквивалентно следующему: объект относят к первому классу тогда и только тогда, когда

$$\left[x-\frac{1}{2}(a_1+a_2)\right](a_1-a_2)\frac{1}{\sigma^2}\geqslant 0\,, \qquad (2.3.4)$$

а во всех остальных случаях — ко второму.

Для m-мерного случайного вектора $\mathbf{X}$ классификационное правило в терминах выборки звучит так: объект с координатами

$$x^{(0)}=\left(x_1^{(0)},...,x_m^{(0)}\right)$$

относят к первому классу тогда и только тогда, когда

$$\left[\mathbf{x}^{(0)}-\frac{1}{2}(\hat{\mathbf{a}}_1+\hat{\mathbf{a}}_2)\right]\hat{\mathbf{\Sigma}}^{-1}(\hat{\mathbf{a}}_1-\hat{\mathbf{a}}_2)^{\mathrm{T}}\geqslant 0\,, \qquad (2.3.5)$$

и ко второму — во всех остальных случаях.

В соотношении (2.3.5)

$$\hat{\mathbf{a}}_1=\left(\hat{a}_1^{(1)},\hat{a}_2^{(1)},...,\hat{a}_m^{(1)}\right)\ —$$

вектор средних значений случайных величин $X_1, X_2, \ldots, X_m$ в выборке объема n_1 из первого класса,

$$\hat{\mathbf{a}}_2=\left(\hat{a}_1^{(2)},\hat{a}_2^{(2)},...,\hat{a}_m^{(2)}\right)\ —$$

вектор средних значений этих величин в выборке объема n_2 из второго класса, $\hat{\mathbf{\Sigma}}$ — рассчитанная по обучающим выборкам оценка ковариационной матрицы вектора $\mathbf{X}$, общая для двух классов.

Будем считать, что в первую выборку попали объекты с номерами $1, 2, \ldots, n_1$, во вторую выборку — объекты с номерами $1^*, 2^*, \ldots, n_2^*$, а x_{ij} — это значение случайной величины X_j для i-го объекта. Тогда

$$\hat{a}_j^{(1)}=\frac{1}{n_1}\sum_{i=1}^{n_1}x_{ij}\,,\quad \hat{a}_j^{(2)}=\frac{1}{n_2}\sum_{i=1^*}^{n_2^*}x_{ij} \qquad (2.3.6)$$

$$(j = 1, 2, \ldots, m),$$

$$\widehat{\operatorname{cov}}\left(X_j, X_k\right) = \frac{\widehat{\operatorname{cov}}^{(1)}(X_j, X_k)n_1 + \widehat{\operatorname{cov}}^{(2)}(X_j, X_k)n_2}{n_1 + n_2 - 2} \qquad (2.3.7)$$

$$(j = 1, 2, ..., m, \quad k = 1, 2, ..., m),$$

где

$$\widehat{\operatorname{cov}}^{(1)}(X_j, X_k) = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(x_{ij} - \hat{a}_j^{(1)}\right)\left(x_{ik} - \hat{a}_k^{(1)}\right), \qquad (2.3.8)$$

$$\widehat{\operatorname{cov}}^{(2)}(X_j, X_k) = \frac{1}{n_2^*}\sum_{i=1^*}^{n_2^*}\left(x_{ij} - \hat{a}_j^{(2)}\right)\left(x_{ik} - \hat{a}_k^{(2)}\right). \qquad (2.3.9)$$

Соотношением (2.3.5) задается вид *дискриминантной функции* для двух совокупностей, распределенных по нормальному закону.

Обобщение классификационного правила [дискриминантной функции (2.3.5)] на случай $l \geqslant 2$ классов, доли объектов которых в общей генеральной совокупности равны соответственно π_1, π_2, …, π_l, звучит так: объект с координатами $\mathbf{x}^{(0)} = \left(x_1^{(0)}, ..., x_m^{(0)}\right)$ относят к классу с номером j_0 тогда и только тогда, когда

$$\left[\mathbf{x}^{(0)} - \frac{1}{2}\left(\hat{\mathbf{a}}_{j_0} + \hat{\mathbf{a}}_j\right)\right]\hat{\mathbf{\Sigma}}^{-1}\left(\hat{\mathbf{a}}_{j_0} - \hat{\mathbf{a}}_j\right)^{\mathrm{T}} \geqslant \ln\frac{\pi_j}{\pi_{j_0}}$$

для всех $j = 1, 2, \ldots, l$.

При использовании этого правила потери от неправильной классификации постоянны и минимальны (см., например, [1, 3]).

2.3.2. Расщепление смесей вероятностных распределений

К задаче дискриминантного анализа может быть сведена и *задача расщепления смеси вероятностных распределений*

$$f(x) = \sum_{i=1}^{k}\pi_i f(x, \boldsymbol{\theta}_i): \qquad (2.3.10)$$

если известен лишь общий вид функций $f(x, \boldsymbol{\theta}_i)$, и в распоряжении исследователя есть обучающие выборки, то задача расщепления смеси, состоящая в том, что требуется построить оценки для числа k компонент смеси , их удельных весов π_1, π_2, …, π_k, и для каждой из компонент $f(x, \boldsymbol{\theta}_i)$ анализируемой смеси (2.3.10), то решение такой задачи можно провести согласно следующему **алгоритму**:

- вначале по i-й обучающей выборке оценить параметр $\boldsymbol{\theta}_i$;
- затем провести классификацию точно так же, как если бы функции $f(x, \boldsymbol{\theta}_i)$ были точно известны.

Подробнее о задачах расщепления смесей и методах их решения можно узнать в книге [3].

§ 2.4. Практическое использование методов классификации

2.4.1. Особенности применения методов классификации с использованием современных пакетов прикладных программ

В современных пакетах прикладных программ реализовано достаточно большое число методов кластерного анализа, существенно большее, чем рассмотрено в нашем пособии.

Важно отметить, что программы, как правило, выдают не только протокол объединения кластеров и дендрограмму, но и результаты дисперсионного анализа полученного разбиения по каждому признаку: в том числе, межклассовую и внутриклассовую вариацию, число степеней свободы, F-статистику и рассчитанный уровень значимости при проверке гипотезы о равенстве межклассовой и внутриклассовой дисперсий.

При использовании методов дискриминантного анализа в пакетах прикладных программ, как правило, можно не задавать априорное разбиение объектов по классам, программа сама построит такое разбиение. Если же разбиение по классам задано, программа укажет объекты из обучающих выборок, которые, возможно, классифицированы ошибочно.

2.4.2. Кластерный анализ торговых предприятий

Проиллюстрируем методы кластерного анализа на примере классификации пяти точек:

$$(4; 3), \quad (7; 4), \quad (2; 0), \quad (2; 1), \quad (0; 4),$$

изображенных на рис. 2.4.1.

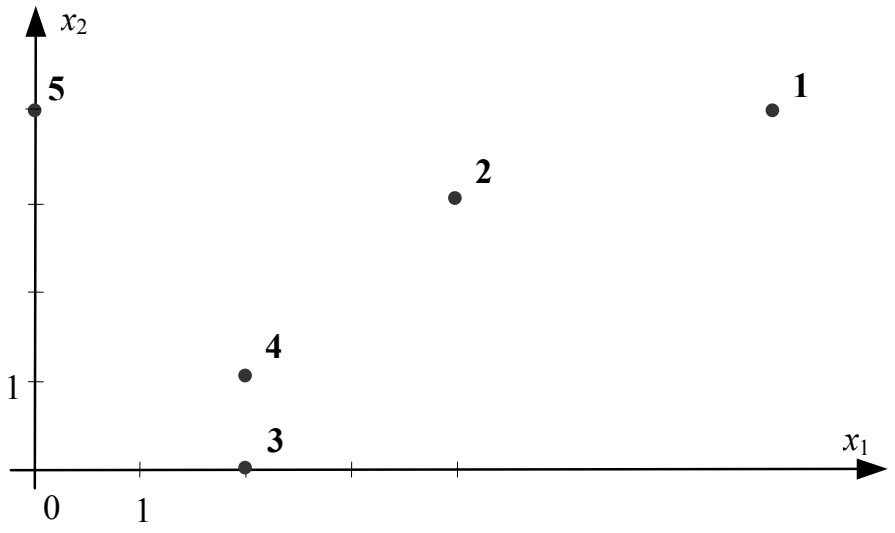


Рис. 2.4.1

Эти точки соответствуют объемам продаж в пяти магазинах небольшой розничной сети, торгующей холодильниками и стиральными машинами; первая координата соответствует продажам холодильников в магазине за неделю, а вторая координата — продажам стиральных машин.

За метрику расстояний примем квадрат евклидова расстояния. Матрица расстояний приведена в табл. 2.4.1.

Таблица 2.4.1

	1	2	3	4	5
1	0	10	13	8	17
2	10	0	41	34	49
3	13	41	0	**1**	20
4	8	34	**1**	0	13
5	17	49	20	13	0

1. Метод ближнего соседа. Начальное разбиение таково: {1, 2, 3, 4, 5}. Минимальное расстояние между кластерами $\rho_{3,4} = 1$, поэтому переходим к следующему разбиению: {1, 2, 3⊕4, 5}.

Последовательность разбиений

$$\{1, 2, 3, 4, 5\} \to \{1, 2, 3\oplus4, 5\} \to \{1\oplus(3\oplus4), 2, 5\} \to$$
$$\to \{1\oplus(3\oplus4)\oplus2, 5\} \to \{1\oplus2\oplus3\oplus4\oplus5\}$$

представлена в табл. 2.4.2, а на рис. 2.4.2, *а* изображена соответствующая дендрограмма (слева проставлены расстояния, при которых происходит переход к новому разбиению).

Таблица 2.4.2

	1	2	3⊕4	5
1	0	10	**8**	17
2	10	0	34	49
3⊕4	**8**	34	0	13
5	17	49	13	0

→

	1⊕(3⊕4)	2	5
1⊕(3⊕4)	0	**10**	13
2	**10**	0	49
5	13	49	0

→

	1⊕(3⊕4)⊕2	5
1⊕(3⊕4)⊕2	0	**13**
5	**13**	0

Поскольку данный метод объединяет кластеры, в которых расстояние между ближайшими элементами минимально по сравнению с другими кластерами, то два объекта попадают в один кластер, если существует соединяющая их цепочка ближайших друг к другу объектов (так называемый цепочечный эффект). Поэтому метод ближнего соседа называют методом одиночной связи. Для устранения цепочечного эффекта можно, например, ввести ограничение на максимальное расстояние между объектами кластера: в первый кластер включить два наиболее близких объекта, затем в этот кластер включить объект, который имеет минимальное расстояние с одним их объектов кластера, а его расстояние до другого объекта кластера не больше некоторого числа l_0 и т. д.; формирование первого кластера продолжают до тех пор, когда нельзя будет найти объект,

расстояние от которого до любого объекта кластера не превзойдет l_0; формирование второго и последующих кластеров осуществляется из оставшихся объектов аналогичным образом.

2. Метод дальнего соседа. Начальное разбиение таково: {1, 2, 3, 4, 5}. Минимальное расстояние между кластерами $\rho_{3,4} = 1$, поэтому, как и в методе ближнего соседа, переходим к следующему разбиению: $\{1, 2, 3 \oplus 4, 5\}$.

Последовательность разбиений

$$\{1, 2, 3, 4, 5\} \to \{1, 2, 3\oplus4, 5\} \to \{1\oplus2, 3\oplus4, 5\} \to$$
$$\to \{1\oplus2, 3\oplus4\oplus5\} \to \{1\oplus2\oplus3\oplus4\oplus5\}$$

представлена в табл. 2.4.3, а на рис. 2.4.2, *б* изображена соответствующая дендрограмма.

Таблица 2.4.3

	1	2	3⊕4	4
1	0	**10**	13	17
2	**10**	0	41	49
3⊕4	13	41	0	20
5	17	49	20	0

→

	1⊕2	3⊕4	5
1⊕2	0	41	49
3⊕4	41	0	**20**
5	49	**20**	0

→

	1⊕2	(3⊕4)⊕5
1⊕2	0	**49**
(3⊕4)⊕5	**49**	0

В этом методе объединяются кластеры, в которых минимально расстояние между самыми далекими друг от друга объектами. Это означает, что все остальные объекты в полученном после объединения кластере связаны друг с другом еще теснее, чем «соседи». Поэтому метод дальнего соседа называют методом полной связи.

3. Метод средней связи. Начальное разбиение таково: {1, 2, 3, 4, 5}. Минимальное расстояние между кластерами $\rho_{3,4} = 1$, поэтому, как и в методе ближнего соседа, переходим к следующему разбиению: $\{1, 2, 3 \oplus 4, 5\}$.

Последовательность разбиений

$$\{1, 2, 3, 4, 5\} \to \{1, 2, 3\oplus4, 5\} \to \{1\oplus2, 3\oplus4, 5\} \to$$
$$\to \{1\oplus2, 3\oplus4\oplus5\} \to \{1\oplus2\oplus3\oplus4\oplus5\}$$

представлена в табл. 2.4.4, а на рис. 2.4.2, *в* изображена соответствующая дендрограмма.

Таблица 2.4.4

	1	2	3⊕4	5
1	0	**10**	10,5	17
2	**10**	0	37,5	49
3⊕4	10,5	37,5	0	16,5
5	17	49	16,5	0

→

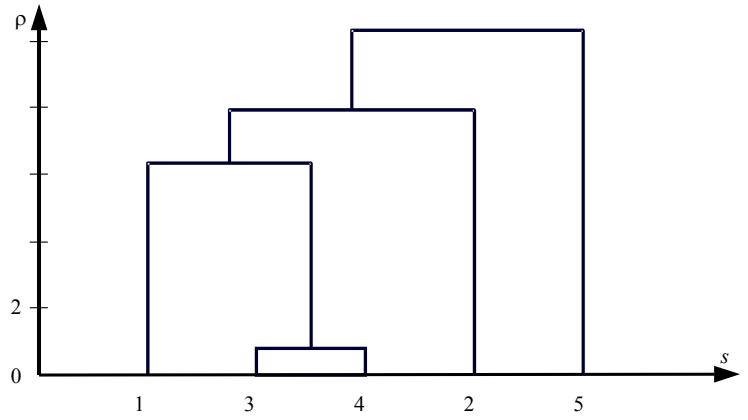
	1⊕2	3⊕4	5
1⊕2	0	24	33
3⊕4	24	0	**16,5**
5	33	**16,5**	0

→

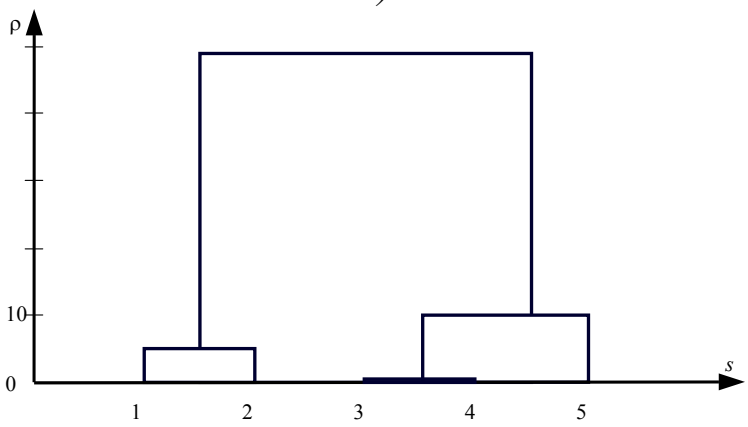
	1⊕2	(3⊕4)⊕5
1⊕2	0	**27**
(3⊕4)⊕5	**27**	0

В последней таблице

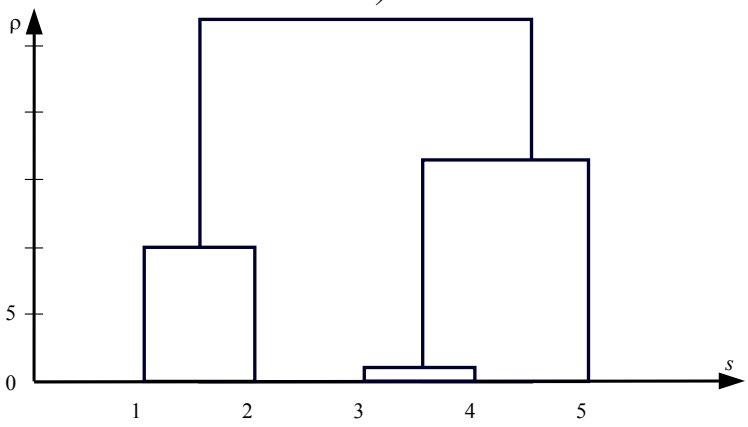
$$\rho_{1\oplus2,(3\oplus4)\oplus5} = \frac{2}{3}\rho_{1\oplus2,3\oplus4} + \frac{1}{3}\rho_{1\oplus2,5} = \frac{2}{3}\cdot 24 + \frac{1}{3}\cdot 33 = 27.$$



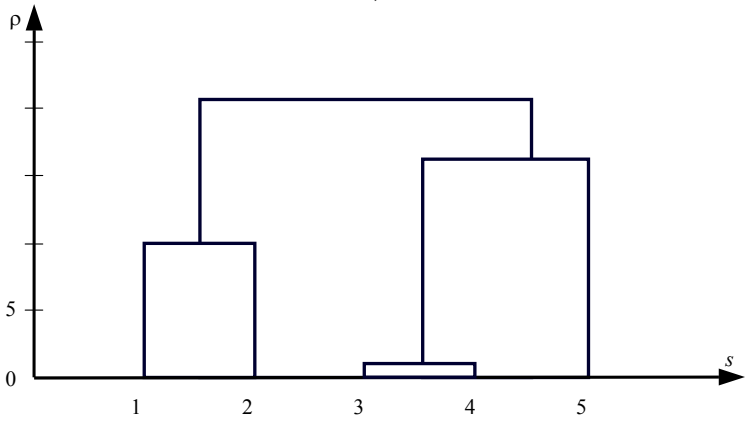
а)



б)



в)



г)

Рис. 2.4.2

4. Центроидный метод. Начальное разбиение таково: {1, 2, 3, 4, 5}. Минимальное расстояние между кластерами $\rho_{3,4} = 1$, поэтому, как и в методе ближнего соседа, переходим к следующему разбиению: $\{1, 2, 3 \oplus 4, 5\}$.

Последовательность разбиений

$$\{1, 2, 3, 4, 5\} \to \{1, 2, 3\oplus4, 5\} \to \{1\oplus2, 3\oplus4, 5\} \to$$
$$\to \{1\oplus2, 3\oplus4\oplus5\} \to \{1\oplus2\oplus3\oplus4\oplus5\}$$

представлена в табл. 2.4.5, а на рис. 2.4.2, *г* изображена соответствующая дендрограмма.

Таблица 2.4.5

	1	2	3⊕4	5
1	0	**10**	10,25	17
2	**10**	0	37,25	49
3⊕4	10,25	37,25	0	16,25
5	17	49	16,25	0

→

	1⊕2	3⊕4	5
1⊕2	0	21,25	30,5
3⊕4	21,25	0	**16,25**
5	30,5	**16,25**	0

→

	1⊕2	(3⊕4)⊕5
1⊕2	0	**373/18**
(3⊕4)⊕5	**373/18**	0

Дадим пояснение к расчетам некоторых расстояний. Расчеты проведем по тождественным формулам (2.2.15):

$$\rho_{1,3\oplus4} = \begin{cases} \dfrac{\rho_{1,3} + \rho_{1,4}}{2} - \dfrac{\rho_{3,4}}{4} = \dfrac{13+8}{2} - \dfrac{1}{4} = 10.25, \\ l(\overline{x}_1, \overline{x}_{3\oplus4}) = l((1,2),(-1,-0.5)) = (1+1)^2 + (2+0.5)^2 = 10.25; \end{cases}$$

$$\rho_{1\oplus2,(3\oplus4)\oplus5} = \begin{cases} \dfrac{2\rho_{1\oplus2,3\oplus4} + \rho_{1\oplus2,5}}{3} - \dfrac{2\rho_{3\oplus4,5}}{9} = \dfrac{2\cdot21,25+30,5}{3} - \dfrac{2\cdot16,25}{9} = \dfrac{373}{18}, \\ l(\overline{x}_{1\oplus2}, \overline{x}_{(3\oplus4)\oplus5}) = l\left[\left(\dfrac{5}{2},\dfrac{5}{2}\right),\left(-\dfrac{5}{3},\dfrac{2}{3}\right)\right] = \left(\dfrac{5}{2}+\dfrac{5}{3}\right)^2 + \left(\dfrac{5}{2}-\dfrac{2}{3}\right)^2 = \dfrac{373}{18}. \end{cases}$$

Если число кластеров заранее известно, то классификацию заканчивают, как только будет сформировано разбиение с нужным числом кластеров. При неизвестном числе кластеров правило остановки связывают с понятием п о р о г а (уже введенным ранее), т. е. некоторого расстояния l_0, определяемое условиями конкретной задачи.

Например, если

$$l_0 = \sum_{j>i} \frac{l_{ij}}{n(n-1)}$$

(в условиях примера $l_0 = 20,6$), то метод ближнего соседа объединит все пять объектов в один кластер, а остальные три метода дадут два кластера: 3⊕4⊕5 и 1⊕2.

5. Алгоритм «объединение». Применим алгоритм к матрице расстояний, приведенной в табл. 2.4.1, заменив ее на матрицу связей, например, с помощью преобразования

$$r_{ij} = \frac{\overline{l}}{\overline{l} + l_{ij}} = \frac{20,6}{20,6 + l_{ij}} \ —$$

такая матрица связей представлена в табл.2.4.6; примем $r_0 = 0,5$.

Таблица 2.4.6

	1	2	3	4	5
1	1	0,67	0,61	0,72	0,55
2	0,67	1	0,33	0,38	0,30
3	0,61	0,33	1	**0,95**	0,51
4	0,72	0,38	**0,95**	1	0,61
5	0,55	0,30	0,51	0,61	1

Последовательность разбиений

$$\{1, 2, 3, 4, 5\} \to \{1, 2, 3\oplus4, 5\} \to \{1\oplus3\oplus4, 2, 5\} \to$$
$$\to \{1\oplus3\oplus4\oplus5, 2\} \to \{1\oplus2\oplus3\oplus4\oplus5\}$$

представлена в табл. 2.4.7.

Таблица 2.4.7

	1	2	3	4	5
1	0	0,17	0,11	0,22	0,05
2	0,17	0	–0,17	–0,12	–0,20
3	0,11	–0,17	0	**0,45**	0,01
4	0,22	–0,12	**0,45**	0	0,11
5	0,05	–0,20	0,01	0,11	0

→

	1	2	3⊕4	5
1	0	0,17	**0,33**	0,05
2	0,17	0	–0,29	–0,20
3⊕4	**0,33**	–0,29	0,9	0,12
5	0,05	–0,20	0,12	0

→

	1⊕(3⊕4)	2	5
1⊕(3⊕4)	1,56	–0,12	**0,17**
2	–0,12	0	–0,20
5	**0,17**	–0,20	0

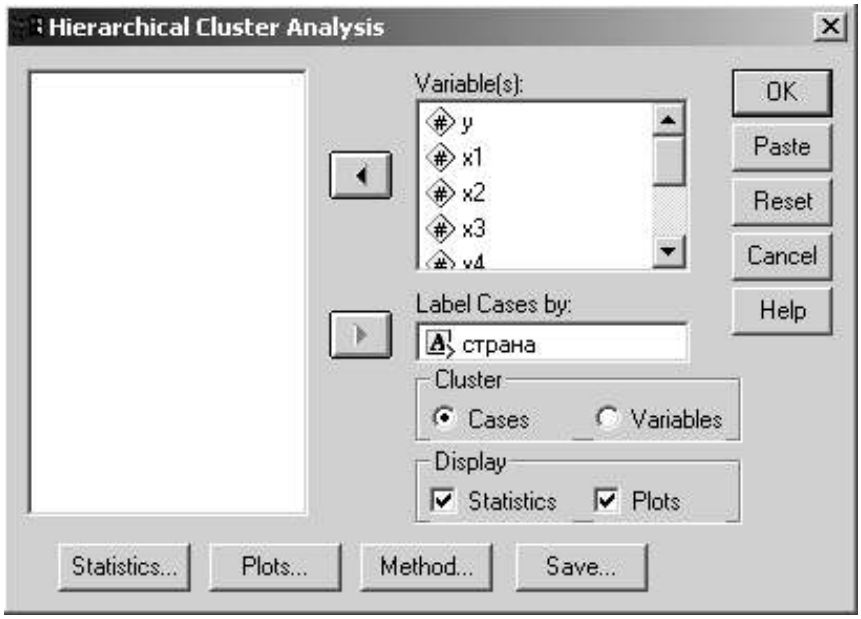
→

	(1⊕(3⊕4))⊕5	2
(1⊕(3⊕4))⊕5	1,9	–0,32
2	–0,32	0

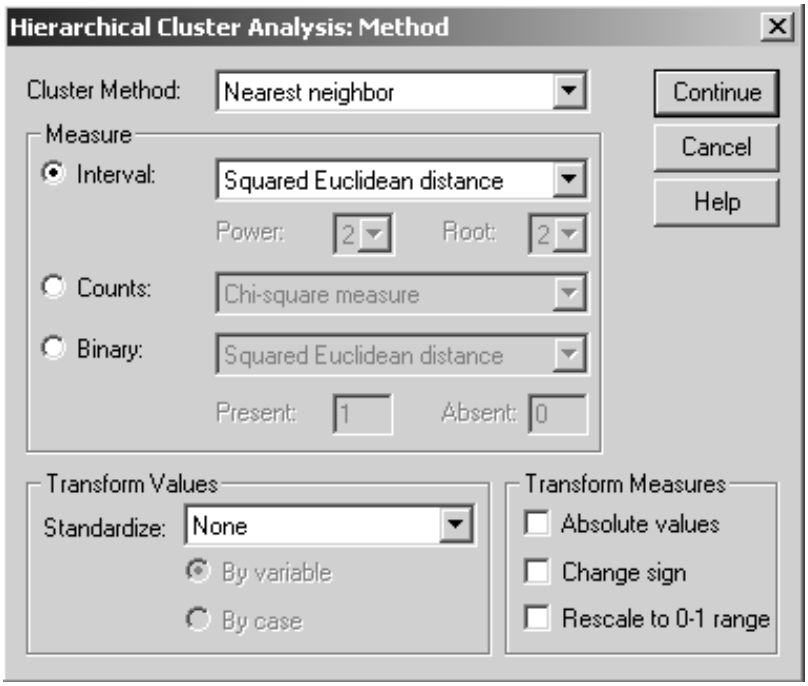
2.4.3. Кластерный анализ мировой демографической статистики

Проведем по исходным данным табл. 1.4.1 классификацию 20 стран: Австралии, Австрии, Беларуси, Бразилии, Великобритании, Венгрии, Германии, Замбии, Индии, Италии, Канады, Китая, Мексики, Нидерландов, Польши, России, Соединенных Штатов Америки, Украины, Филиппин и Эфиопии.

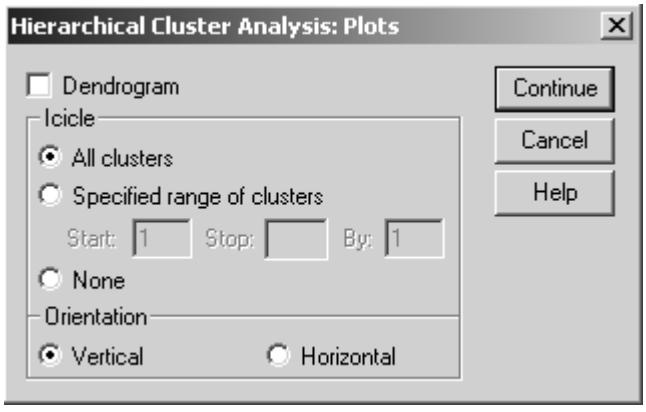
Введем в окно ввода исходных данных пакета SPSS матрицу значений признаков, оставив в ней только строки, соответствующие классифицируемым странам. Обратимся (с помощью выбора пункта «Classify | Hierarchical Cluster…» меню «Statistics») к программе «Hierarchical Cluster Analysis» (рис. 2.4.3, *а*) для иерархического кластерного анализа переменных Y, X_1, X_2, X_3, X_4, X_5.


Hierarchical Cluster Analysis
Variable(s):
y
x1
x2
x3
x4
OK
Paste
Reset
Cancel
Help
Label Cases by:
страна
Cluster
Cases
Variables
Display
Statistics
Plots
Statistics...
Plots...
Method...
Save...

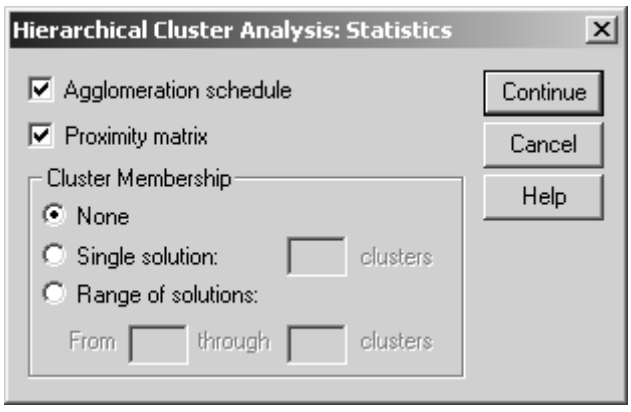
а)


Hierarchical Cluster Analysis: Method
Cluster Method:
Nearest neighbor
Continue
Cancel
Help
Measure
Interval:
Squared Euclidean distance
Power:
2
Root:
2
Counts:
Chi-square measure
Binary:
Squared Euclidean distance
Present:
1
Absent:
0
Transform Values
Standardize:
None
By variable
By case
Transform Measures
Absolute values
Change sign
Rescale to 0-1 range

б)


Hierarchical Cluster Analysis: Plots
Dendrogram
Continue
Cancel
Help
Icicle
All clusters
Specified range of clusters
Start:
1
Stop:
By:
1
None
Orientation
Vertical
Horizontal

в)


Hierarchical Cluster Analysis: Statistics
Agglomeration schedule
Proximity matrix
Continue
Cancel
Help
Cluster Membership
None
Single solution:
clusters
Range of solutions:
From
through
clusters

г)

Рис. 2.4.3

Укажем, что объекты (страны) расположены по строкам («Cluster: Cases»), а в результатах работы программы нужно обозначать их с помощью названий («Label cases by: страна»). В окне, вызываемом нажатием кнопки «Method…», укажем, что необходимо использовать метод ближнего соседа («Cluster Method: Nearest neighbor») с использованием евклидовой метрики расстояний («Measure | Interval: Euclidean distance») — данное окно представлено на рис. 2.4.3, *б*. В окне, вызываемом нажатием кнопки «Plots…» (рис. 2.4.3, *в*), установим флажок «Dendrogram», указав, что в результатах необходимо получить дендрограмму. В окне, вызываемом нажатием кнопки «Statistics…» (рис. 2.4.3, *г*), укажем, что необходимо сохранить в результатах работы программы протокол объединения кластеров («Agglomeration schedule») и матрицу расстояний между кластерами («Proximity matrix»).

Перейдем к рассмотрению результатов работы программы (рис. 2.4.4). В матрице расстояний (таблица «Proximity Matrix») указаны евклидовы расстояния между объектами, например, евклидово расстояние между первым объектом (Австралией) и 20-м (Эфиопией) равно

$$l_{1,20} = \sum_{i=0}^{5} (x_{1,i} - x_{20,i})^2 = 23230,2,$$

где $x_{1,0} = Y_1$, $x_{20,0} = Y_{20}$.

Судя по дендрограмме, исходную совокупность стран имеет смысл разбить на три кластера:

- в первый кластер вошли страны бывшего СССР, Восточной Европы, Африки и Латинской Америки: Беларусь, Бразилия, Россия, Мексика, Польша, Венгрия, Замбия, Эфиопия, Китай, Украина, Филиппины, Индия;
- во второй кластер вошли промышленно развитые страны Запада (кроме США): Великобритания, Италия, Австрия, Нидерланды, Австралия, Германия, Канада;
- в третий кластер вошла одна страна: США.

Средние значения и дисперсии каждого признака для каждого из кластеров приведены в табл. 2.4.8.

Таблица 2.4.8

Кластер	Характеристика	Y	X_1	X_2	X_3	X_4	X_5
1	среднее значение	78,29	1,49	17,71	23718,57	168,08	98,57
	дисперсия	0,57	0,04	4,57	1579714,29	20600,73	1,29
2	среднее значение	64,75	2,67	28,42	5224,17	102,25	81,50
	дисперсия	112,39	3,09	110,81	10798590,15	9447,38	529,91
3	среднее значение	77,00	2,10	21,00	31910,00	30,08	97,00
	дисперсия	0,00	0,00	0,00	0,00	0,00	0,00

Proximity Matrix

Case	Euclidean Distance					
	1:Австралия	2:Австрия	3:Беларусь	4:Бразилия	…	20:Эфиопия
1:Австралия		756,1	16970,1	17010,0	…	23230,2
2:Австрия	756,1		17720,1	17760,2	…	23980,2
3:Беларусь	16970,1	17720,1		53,5	…	6260,5
4:Бразилия	17010,0	17760,2	53,5		…	6220,4
…	…	…	…	…	…	…
20:Эфиопия	23230,2	23980,2	6260,5	6220,4	…	

This is a dissimilarity matrix

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	3	4	53,475	0	0	2
2	3	16	117,151	1	0	12
3	8	20	121,607	0	0	15
4	12	18	197,978	0	0	9
5	5	10	226,582	0	0	14
…	…	…	…	…	…	…

1. Dendrogram

* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *

Dendrogram using Single Linkage

Rescaled Distance Cluster Combine

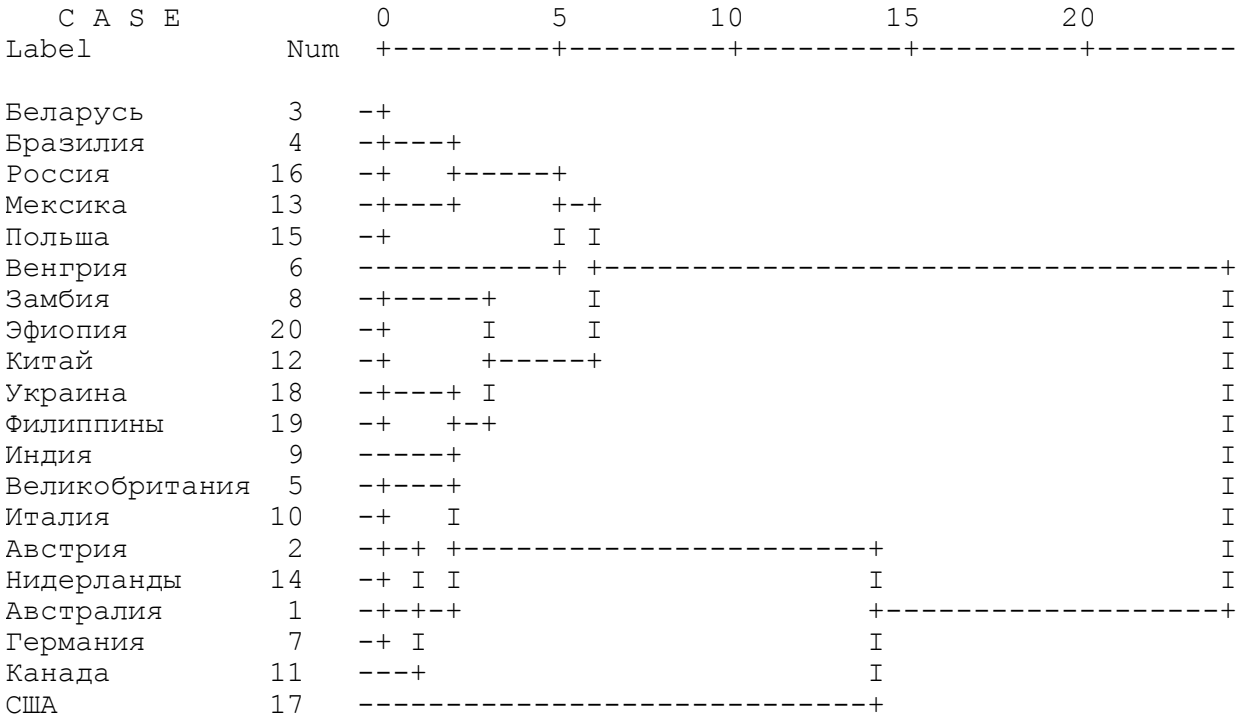


Рис. 2.4.4

2.4.4. Дискриминантный анализ надежности банков

Пусть $\mathbf{X} = (X_1, X_2, X_3)$, где X_1, X_2, X_3 — некоторые показатели, измеренные для каждого банка. Первая обучающая выборка образована из банков с самым высоким рейтингом надежности, а вторая обучающая выборка состоит из наименее надежных банков. Значения случайных величин X_1, X_2, X_3 для обучающих выборок приведены в табл. 2.4.9.

Таблица 2.4.9

Объект	Выборка из первого класса			Выборка из второго класса		
	X_1	X_2	X_3	X_1	X_2	X_3
1	9,9	0,34	1,68	5,5	0,05	1,02
2	9,1	0,09	1,89	5,6	0,48	0,88
3	9,4	0,21	2,3	4,3	0,41	0,62
4	9,4	0,28	2,03	7,4	0,62	1,09
5	—	—	—	6,6	0,5	1,32
Средние	$\hat{a}_1^{(1)} = 9,45$	$\hat{a}_2^{(1)} = 0,23$	$\hat{a}_3^{(1)} = 1,975$	$\hat{a}_1^{(2)} = 5,88$	$\hat{a}_2^{(2)} = 0,412$	$\hat{a}_3^{(2)} = 0,986$

Известно также, что число надежных банков приблизительно равно числу ненадежных.

Используя (2.3.7), (2.3.8) и (2.3.9), найдем оценки ковариационных матриц $\boldsymbol{\Sigma}^{(1)}$ и $\boldsymbol{\Sigma}^{(2)}$:

$$\hat{\boldsymbol{\Sigma}}^{(1)} = \begin{pmatrix} 0,0825 & 0,0242 & -0,0305 \\ 0,0242 & 0,009 & -0,006 \\ -0,0305 & -0,006 & 0,051 \end{pmatrix}, \quad \hat{\boldsymbol{\Sigma}}^{(2)} = \begin{pmatrix} 1,1096 & 0,1 & 0,199 \\ 0,1 & 0,037 & 0,006 \\ 0,199 & 0,006 & 0,054 \end{pmatrix},$$

а затем — оценку

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0,839 & 0,085 & 0,125 \\ 0,085 & 0,032 & 0,001 \\ 0,125 & 0,001 & 0,067 \end{pmatrix}$$

ковариационной матрицы вектора $\mathbf{X}$, общей для обоих классов.

Учитывая равенство долей объектов каждого класса в общей совокупности ($\pi_1 = \pi_2 = 0,5$), найдем составляющие дискриминантной функции (2.3.5):

$$\hat{\mathbf{a}}_1 = (9,45 \quad 0,23 \quad 1,975), \quad \hat{\mathbf{a}}_2 = (5,88 \quad 0,412 \quad 0,986),$$

$$\hat{\boldsymbol{\Sigma}}^{-1} = \begin{pmatrix} 2,566 & -6,669 & -4,688 \\ -6,669 & 48,600 & 11,718 \\ -4,688 & 11,718 & 23,496 \end{pmatrix}.$$

Правило отнесения к первому классу примет тогда следующий вид:

$$\left(x_1^{(0)} - 7,665; x_2^{(0)} - 0,321; x_3^{(0)} - 1,4805\right)\hat{\Sigma}^{-1}\begin{pmatrix} 3,57 \\ -0,182 \\ 0,989 \end{pmatrix} \geqslant 0$$

или

$$f(x_1^{(0)}, x_2^{(0)}, x_3^{(0)}) = 5,739x_1^{(0)} + 0,356x_2^{(0)} + 4,369x_3^{(0)} - 43,686 \geqslant 0.$$

Классифицируем шесть банков со значениями $x_1^{(0)}, x_2^{(0)}, x_3^{(0)}$ признаков X_1, X_2, X_3; представленными в табл. 2.4.10.

Таблица 2.4.10

Объект	$x_1^{(0)}$	$x_2^{(0)}$	$x_3^{(0)}$	f	Класс
1	9,4	0,15	1,91	18,66 > 0	1
2	5,5	1,20	0,68	–13,08 < 0	2
3	5,7	0,66	1,43	–4,49 < 0	2
4	5,2	0,74	1,82	–5,63 < 0	2
5	10,0	0,32	2,62	25,26 > 0	1
6	6,7	0,39	1,24	0,32 > 0	1

В табл. 2.4.10 для каждого объекта указаны значения функции

$$f = 5,739x_1^{(0)} + 0,356x_2^{(0)} + 4,369x_3^{(0)} - 43,686$$

и класс, к которому следует отнести объект.

2.4.5. Дискриминантный анализ мировой демографической статистики

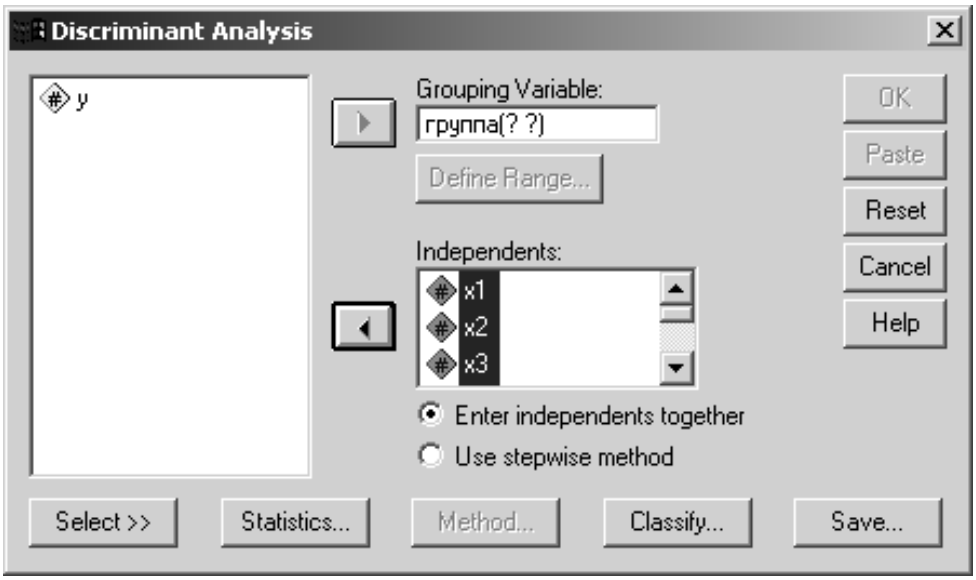
Продолжим анализ мировой демографической статистики, начатый в **пп. 1.4.2, 2.2.7**. Предположим теперь, что известно, что Бразилия, Мексика, Польша, Венгрия, Китай, Филиппины и Индия входят в группу **А**, Великобритания, Италия, Австрия, Нидерланды, Австралия, Германия, Канада и США входят в группу **Б**.

Требуется определить, в какую из групп (**А** или **Б**) входит каждая из стран: Россия, Беларусь, Украина, Замбия и Эфиопия (считая, что каждая из этих пяти стран входит в одну и только в одну из двух групп **А** и **Б**). Предположим, что каждая группа подчиняется пятимерному нормальному распределению с одинаковой для обеих групп ковариационной матрицей.

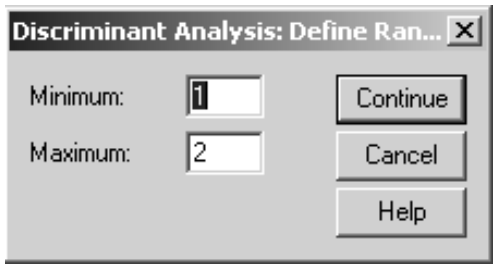
Введем в окно ввода исходных данных пакета SPSS матрицу значений признаков, оставив в ней только строки, соответствующие 20 названным странам. Добавим переменную «группа», значения которой установим равными единице для стран группы **А**, двойке для стран группы **Б**, а для остальных стран оставим пустыми.

Обратимся (с помощью выбора пункта «Classify | Discriminant…» меню «Statistics») к программе «Discriminant Analysis» (рис. 2.4.5, *а*). Укажем независимые переменные («Independents»), группирующую переменную

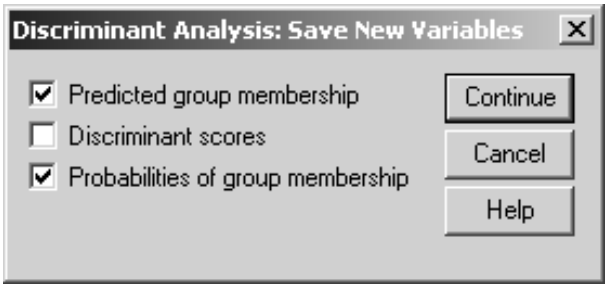
(«Grouping Variable») и диапазон ее значений от единицы до двух (в окне, появляющемся после нажатия кнопки «Define Range…» — рис. 2.4.5, *б*). В окне, появляющемся после нажатия кнопки «Save…» (рис. 2.4.5, *в*), укажем, что в результате работы программы необходимо сохранить предсказанные значения групп (установив флажок «Predicted group membership») и вероятности попадания в каждую из групп (установив флажок «Probabilities of group membership»).



а)



б)



в)

Рис. 2.4.5

В результате работы программы (рис. 2.4.6) получена дискриминантная функция

$$\widehat{D} = -2,733X_1 + 2,441X_2 + 1,360X_3 + 0,976X_4 + 0,042X_5,$$

ее среднее значение в группе **А** составило 1,547, в группе **Б** — (–1,354).

Появившиеся в окне ввода исходных данных новые переменные «dis_1», «dis_1_1» и «dis_2_1» содержат сведения о классификации, предложенной методом дискриминантного анализа, и вероятности попадания в первую и вторую группы соответственно.

В результате оказалось, что изначально все 15 стран были правильно разбиты по группам, и пять стран, которые необходимо было классифицировать, отнесены к группе **A**.

Standardized Canonical Discriminant Function Coefficients

	Function
	1
X1	–2,733
X2	2,441
X3	1,360
X4	0,976
X5	0,042

Functions at Group Centroids

	Function
	1
1	1,547
2	–1,354

а)

Group	dis_1	dis1_1	dis2_1
Австралия	2	0,38293	0,61707
Австрия	2	0,00908	0,99092
Беларусь	1	0,98194	0,01806
Бразилия	1	0,99781	0,00219
Великобритания	2	0,01948	0,98052
Венгрия	1	0,93529	0,06471
Германия	2	0,00260	0,99740
Замбия	1	1,00000	0,00000
Индия	1	0,99625	0,00375
Италия	2	0,13740	0,86260
Канада	2	0,01687	0,98313
Китай	1	0,99731	0,00269
Мексика	1	0,99410	0,00590
Нидерланды	2	0,00020	0,99980
Польша	1	0,83657	0,16343
Россия	1	0,99123	0,00877
Соединенные Штаты Америки	2	0,16997	0,83003
Украина	1	0,97856	0,02144
Филиппины	1	0,99973	0,00027
Эфиопия	1	1,00000	0,00000

б)

Рис. 2.4.6

Задачи для самостоятельного решения

1. Докажите равенства (2.2.12) ~ (2.2.15). Приняв за метрику расстояний квадрат евклидова расстояния, проведите классификацию пяти точек

$$(2; 4),\ (8; 6),\ (-2; -2),\ (-2; 0),\ (-6; 6)$$

иерархическими агломеративными методами; постройте дендрограммы.

2. Докажите, что для метода дальнего соседа функционал качества разбиения имеет вид (2.2.17).

3. Изобразите дендрограмму, соответствующую табл. 2.4.6.

4. В чем отличие параллельных и последовательных кластер-процедур?

5. Проверьте гипотезы о равенстве средних значений признаков для каждой пары кластеров, полученных в **п. 2.4.3**.

6. Исследуйте, различается ли линейная регрессионная зависимость результативного признака Y от факторных признаков X_1, X_2, X_3, X_4, X_5 в кластерах, полученных в **п. 2.4.3**.

7. Проведите классификацию стран, рассмотренных в **п. 2.4.3**, по стандартизованным значениям исходных признаков X_1, X_2, X_3, X_4, X_5. Сравните результаты с полученными в **п. 2.4.3**.

8. Проведите классификацию стран, рассмотренных в **п. 2.4.3**, по первой главной компоненте, полученной в **п. 1.4.3**. Сравните результаты с полученными в **п. 2.4.3**.

9. Сравните результаты классификации стран, рассмотренных в **п. 2.4.3**, полученные с использованием различных методов кластерного анализа, реализованных в пакете SPSS.

10. Определите, к какому из двух классов относятся каждый из шести объектов с номерами 10 ~ 16, при обучающих выборках: четыре объекта (1 ~ 4) из первого класса и пять объектов (5 ~ 9) из второго класса:

Объекты	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X_1	9,4	9,9	9,4	9,4	4,3	7,4	6,6	5,5	5,7	9,1	5,5	5,6	5,2	10,0	6,7
X_2	0,15	0,34	0,21	0,28	0,41	0,62	0,50	1,20	0,66	0,09	0,05	0,48	0,74	0,32	0,39
X_3	1,91	1,68	2,30	2,03	0,62	1,09	1,32	0,68	1,43	1,89	1,02	0,88	1,82	2,62	1,24

Литература

1. *Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика: Классификация и снижение размерности. – М.: Финансы и статистика, 1989.
2. *Айвазян С. А., Мхитарян В.С.* Практикум по прикладной статистике и эконометрике: Учеб. пособие. – М.: МЭСИ, 2002.
3. *Айвазян С. А., Мхитарян В.С.* Прикладная статистика. Основы эконометрики: Учебник: В 2-х тт. – М.: ЮНИТИ-ДАНА, 2001.
4. *Айвазян С. А., Мхитарян В.С.* Прикладная статистика в задачах и упражнениях: Учебник. – М.: ЮНИТИ-ДАНА, 2001.
5. *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. – М.: Наука, 1983.
6. *Бююль А., Цефель П.* SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей: Пер. с нем. – СПб.: ДиаСофтЮП, 2001.
7. *Дубров А. М.* Компонентный анализ и эффективность в экономике: Учеб. пособие. – М.: Финансы и статистика, 2002.
8. *Дубров А. М., Мхитарян В. С., Трошин Л. И.* Многомерные статистические методы: Учебник. – М.: Финансы и статистика, 2000.
9. *Жуковская В. М., Мучник И. Б.* Факторный анализ в социально-экономических исследованиях. – М.: Статистика, 1976.
10. *Ильин В. А., Ким Г. Д.* Линейная алгебра и аналитическая геометрия: Учебник. – М.: Издательство Московского университета, 2002.
11. *Иберла К.* Факторный анализ: Пер. с нем. – М.: Статистика, 1980.
12. *Калинина В. Н.* Многомерный статистический анализ в управлении: Учеб. пособие / МИУ. – М., 1987.
13. *Калинина В. Н., Соловьев В. И.* Методические указания к курсовому проектированию по дисциплинам «Методология исследования в социально-поведенческих науках» и «Теория вероятностей и математическая статистика». Раздел «Теория вероятностей и математическая статистика» / ГУУ. – М., 2001.
14. *Колемаев В. А., Калинина В. Н.* Теория вероятностей и математическая статистика: Учебник. – М.: ЮНИТИ-ДАНА, 2003.
15. *Колемаев В. А., Староверов О. В., Турундаевский В. Б.* Теория вероятностей и математическая статистика: Учеб. пособие. – М.: Высшая школа, 1991.
16. *Куперштох В. Л., Миркин Б. Г., Трофимов В. А.* Сумма внутренних связей как показатель качества классификации // Автоматика и телемеханика. – 1976. – № 3.
17. *Лоули Д., Максвелл А.* Факторный анализ как статистический метод: Пер. с англ. – М.: Мир, 1967.
18. *Магнус Я. Р., Катышев П. К., Пересецкий А. А.* Эконометрика: Начальный курс: Учебник. – М.: Дело, 2003.
19. *Магнус Я. Р., Нейдеккер Я. Р.* Матричное дифференциальное исчисление с приложениями к статистике и эконометрике: Пер. с англ. – М.: Физматлит, 2002.
20. *Миркин Б. Г.* Анализ качественных признаков и структур. – М.: Статистика, 1980.
21. *Окунь Я.* Факторный анализ: Пер. с польск. – М.: Статистика, 1974.
22. *Сошникова Л. А., Тамашевич В. Н., Уебе Г., Шефер М.* Многомерный статистический анализ в экономике: Учеб. пособие. – М.: ЮНИТИ-ДАНА, 1999.
23. *Харман Г.* Современный факторный анализ: Пер. с англ. – М.: Статистика, 1972.

Оглавление

Издается в авторской редакции, с авторского оригинал-макета.
Ответственность за сведения, представленные в издании, несут авторы.

Вера Николаевна КАЛИНИНА
Владимир Игоревич СОЛОВЬЕВ

ВВЕДЕНИЕ В МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ

УЧЕБНОЕ ПОСОБИЕ

Технический редактор *В. И. Соловьев*
Тематический план изданий ГУУ 2003 г.

Подп. в печ. 24.12.2003. Формат 60×90/16. Объем 4,25 печ. л.
Бумага офисная. Печать трафаретно-цифровая. Гарнитура Times.
Уч.-изд. л. 3,50. Изд. № 289/2003. Тираж 100 экз. Заказ № 24

ГОУВПО Государственный университет управления
Издательский центр ГОУВПО ГУУ
109542, Москва, Рязанский проспект, 99, Учебный корпус, ауд. 106
тел./факс (095) 371-95-10, e-mail: ic@guu.ru
http://www.guu.ru/

ДЛЯ ЗАМЕТОК