



Министерство науки и высшего образования РФ

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ДОНСКОЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
(ДГТУ)**

**Методические указания
и контрольная работа
по дисциплине Многомерный статистический анализ**

для магистрантов направления подготовки 38.04.05

Ростов-на-Дону
2026

УДК 517(07)

Многомерный статистический анализ. Методические указания — Ростов-на-Дону: Донской государственный технический университет, 2026 – 27 с.

Методические указания предназначены для студентов очной и заочной формы подготовки магистров. Содержит основные понятия многомерных статистических методов, формулировки теорем, соответствующие формулы, примеры с подробными решениями, задания контрольной работы.

Кафедра математики и информатики
Составители: к.ф.-м.н. Мисюра В.В.

©Донской государственный
технический университет, 2026

ВВЕДЕНИЕ

Многомерные статистические методы (МСМ), представленные в данном пособии, активно используются в аналитической практике. Они являются своего рода интеллектуальным инструментарием современного исследователя, позволяющим находить решение широкого спектра экономических задач. В первую очередь это задачи статистического исследования динамики структурного изменения, выявления латентных факторов, определяющих течение того или иного социально–экономического процесса, построения интегральных индикаторов качества и эффективности функционирования социально–экономической системы, типологизации социально–экономических объектов.

Теоретические основы МСМ базируются на обобщении классической одномерной статистики. Тем не менее, их спецификой являются трудоемкие алгоритмы реализации вычислительных процедур, что приводит к активному использованию компьютерной техники. В то же время, аналитические результаты МСМ обладают сложной интерпретируемостью. Все это требует от исследователя или студента достаточно высокой математической подготовки.

МНОГОМЕРНЫЕ НОРМАЛЬНЫЕ СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

Нормальные случайные величины играют основополагающую роль в теории МСМ.

Определение: многомерной случайной величиной (МСВ) называется функция $\xi = (\xi_1, \xi_2, \dots, \xi_n): \Omega \rightarrow \mathfrak{R}^n$, отображающая вероятностное пространство в n -мерное евклидово пространство.

Определение: матрицей ковариаций A для МСВ ξ называется матрица с элементами

$$a_{ij} = E\{(\xi_i - E(\xi_i))(\xi_j - E(\xi_j))\}.$$

Матрица ковариаций характеризует степень случайного разброса компонент ξ_i МСВ ξ , поэтому ее определитель служит для нахождения обобщенной дисперсии.

Определение: матрицей корреляций R для МСВ ξ называется нормированная ковариационная матрица, элементы которой задаются соотношением

$$r_{ij} = \frac{a_{ij}}{\sqrt{a_{ii}} \cdot \sqrt{a_{jj}}}.$$

Замечание: матрицу R часто называют еще матрицей парных корреляций между одномерными случайными величинами ξ_i и ξ_j .

Для оценки A и R на некоторой совокупности эмпирических данных часто используют выборочные матрицы ковариаций и корреляций.

Определение: выборочной матрицей ковариаций $\hat{A}$ назовем матрицу вида:

$$\hat{A} = \frac{1}{k}(X - \mu)^T(X - \mu),$$

где k – число векторов в генеральной совокупности (размер выборки), X – выборочная матрица, μ – вектор выборочного математического ожидания, каждая компонента которого вычислена по соответствующему столбцу матрицы X : $\mu_j = \frac{1}{k} \sum_{i=1}^k X_{ij}$, $j = \overline{1, n}$. При

этом предполагается, что исходные данные выборок записаны в столбцах X .

Определение: выборочной матрицей корреляций $\bar{R}$ называется матрица, состоящая из элементов вида

$$\bar{r}_{ij} = \bar{a}_{ij}(\sqrt{\bar{a}_{ii}} \cdot \sqrt{\bar{a}_{jj}})^{-1}.$$

Определение: если $\xi_i \sim N(a, \sigma^2)$, $i = \overline{1, n}$ – независимые нормально распределенные случайные величины с плотностями $f(x_i) = (1/\sigma\sqrt{2\pi}) \cdot \exp(-(x_i - a)^2/2\sigma^2)$, то многомерной нормально распределенной случайной величиной $\xi = (\xi_1, \xi_2, \dots, \xi_n) \in \mathfrak{R}^n$ называется вектор $\xi \sim N(M, A)$, где M – вектор математических ожиданий, A – матрица ковариаций, имеющий плотность распределения

$$f_{\xi}(x_1, \dots, x_n) = \left(1/2\pi\sigma^2\right)^{n/2} \cdot \exp\left(-\frac{1}{2} \cdot \sum_{i=1}^n \frac{(x_i - a)^2}{\sigma^2}\right).$$

Можно показать, что при выполнении условий данного определения функция $f_{\xi}(x)$ может быть представлена в виде

$$f_{\xi}(x) = (1/2\pi)^{n/2} \cdot \exp\left(-\frac{1}{2} \cdot \left\{ (X - M)^T \cdot A^{-1} \cdot (X - M) \right\}\right).$$

Для нормально распределенных МСВ справедливы следующие утверждения:

- 1) МСВ η , полученная линейным преобразованием $\eta = C \cdot \xi + d$ из нормально распределенной МСВ $\xi \sim N(M, A)$, так же является распределенной нормально, причем $\eta \sim N(C \cdot M + d, C^T A C)$.
- 2) Если $\xi \sim N(M, A)$, то любой подвектор вектора ξ будет нормально распределенной случайной величиной.
- 3) Пусть $\xi, \eta \sim N(M, A)$. Тогда для независимости ξ и η необходимо и достаточно, чтобы корреляция между ними была равна нулю.

Замечание: независимые случайные величины ξ и η , имеющие произвольное распределение, не обязательно нормальное, всегда некоррелированы. Однако обратное утверждение, вообще говоря, неверно.

Нормальное распределение МСВ применяется в теории многомерного статистического анализа повсеместно. Наиболее естественной и простой областью его использования является множественный регрессионный анализ.

РЕГРЕССИОННЫЙ АНАЛИЗ

Регрессионный анализ – это статистический метод исследования функциональной связи случайной величины y от переменных x_i , $i = \overline{1, n}$, рассматриваемых как неслучайные (известные) МСВ с произвольной функцией распределения. При этом предполагается, что $y \in \mathbb{R}^m$ имеет нормальный закон распределения с условным математическим ожиданием $y = E(x_1, \dots, x_n)$ и с постоянной, не зависящей от x_i , $i = \overline{1, n}$, дисперсией σ^2 .

Функциональную связь принято называть уравнением регрессии, переменные x_i – «входными», y называют откликом или выходной переменной.

В большинстве случаев функциональную зависимость полагают линейной, т.е. решают следующее уравнение регрессии:

$$\bar{y} = \Theta_0 + \Theta_1 x_1 + \Theta_2 x_2 + \dots + \Theta_n x_n + \xi, \quad (1)$$

где Θ_i – неизвестные параметры, $\bar{y}$ – оценка для МСВ y , $\xi = (\xi_1, \dots, \xi_n)$ – вектор ошибок, причем $\xi_i \sim N(0, \sigma^2)$.

Вектор ξ характеризует неучтенные в (1) переменные (скрытые факторы), а также случайные ошибки измерений.

Наряду с линейными моделями вида (1) используют и нелинейные [1], например, полиномиальную, логарифмическую, инверсионную, тригонометрическую, степенную и другие модели. Их применяют, если в результате анализа данных получают статистически ненадежную регрессионную модель (1).

Для нахождения вектора параметров $\Theta = (\Theta_0, \Theta_1, \dots, \Theta_n)$ используют метод наименьших квадратов, для чего минимизируют вектор ошибок:

$$D(\xi) = \frac{1}{n} \cdot \sum_{i=1}^n \xi_i^2 = \frac{1}{n} \cdot \sum_{i=1}^n (y_i - \bar{y}_i)^2 \rightarrow \min \quad (2)$$

Так как из (1) $\xi = y - \Theta X$, то решение задачи (2) можно получить из соотношения

$$\Theta^T = (X^T X)^{-1} X^T \bar{y}.$$

Определение: пусть R – матрица корреляций переменных $\{y = x_0, x_1, x_2, \dots, x_n\}$. Пусть R_{00} – алгебраическое дополнение элемента r_{00} этой матрицы. Множественным коэффициентом корреляции называется число

$$R_y^2 = 1 - [\det R] R_{00}^{-1}.$$

Определение: коэффициентом детерминации называется число

$$K_d = R_y^2 = 1 - D(\xi) D^{-1}(y).$$

Коэффициент детерминации определяет наличие функциональной связи вида (1) или более сложной. Если

$K_d = 0$, то $D(\xi) = D(y)$, т.е. вектор ошибки сравним с откликом и неучтенные в (1) переменные будут определяющими. Следовательно, линейная связь между $\{x_1, x_2, \dots, x_n\}$ и y отсутствует. Если $K_d = 1$, то $D(\xi) = 0$. Значит, переменные $\{x_1, x_2, \dots, x_n\}$ однозначно определяют вектор y .

После нахождения параметров регрессионной модели для определения типа связи в модели (1) требуется вычислить коэффициенты K_d или R_y^2 . Если $0,01 \leq K_d \leq 0,09$, то связь между $\{x_1, x_2, \dots, x_n\}$ и y слабая, недостаточно подтвержденная. Если $0,10 \leq K_d \leq 0,49$, то говорят о наличии средней связи. При $K_d > 0,5$ применение (1) теоретически обосновано и связь сильная. В этом случае необходимо провести дополнительное статистическое исследование. Оно состоит из статистического оценивания регрессионной модели, статистического оценивания надежности коэффициентов регрессии Θ_i и множественного коэффициента корреляции R_y^2 .

Статистическое оценивание регрессионной модели проводится по F-критерию Фишера [2]. Вычисляют статистику $F_H = \frac{m-n-1}{n+1} \cdot (X\Theta)^T (X\Theta) \left[(\bar{y} - (X\Theta))^T (\bar{y} - (X\Theta)) \right]^{-1}$, где m – число компонент вектора y .

По таблицам определяют F_T при заданном уровне значимости α и числе степеней свободы $v_1 = n+1$ и $v_2 = m-n-1$. Надежность регрессионной модели подтверждается, если $F_H > F_T$.

Статистическое оценивание надежности коэффициентов регрессии Θ_i производится с помощью t – критерия Стьюдента [2]. Вычисляют статистику:

$$t_H(i) = \frac{\theta_i}{s_i},$$

где $s_i = \sqrt{\frac{(\bar{y} - X\Theta)^T \cdot (\bar{y} - X\Theta) \cdot c_{ii}}{m-n-1}}$ – средняя ошибка для θ_i , c_{ii} – диагональные элементы матрицы $(X^T X)^{-1}$.

Наблюдаемое значение $t_H(i)$ сравнивают с табличным t_T при заданном уровне значимости α и числе степеней свободы $v = m-n-1$. Значимость θ_i подтверждается, если $|t_H(i)| > t_T$.

Статистическое оценивание множественного коэффициента корреляции или коэффициента детерминации производится с помощью F-критерия Фишера. Вычисляют статистику Снедекора:

$$F_H = \frac{m-n}{n-1} \cdot (R_y^2) \cdot [1 - R_y^2]^{-1}.$$

Наблюдаемое значение F_H сравнивают с табличным F_T при заданном уровне значимости α и числе степеней свободы $\nu_1 = n - 1$ и $\nu_2 = m - n$. Значимость R_y^2 или K_d подтверждается, если $F_H > F_T$.

МЕТОД КАНОНИЧЕСКИХ КОРРЕЛЯЦИЙ

Метод канонических корреляций является обобщением регрессионного анализа на случай нескольких откликов. Он предназначен для статистического анализа связей между массовыми явлениями и процессами. Цель применения метода заключается в нахождении максимальных корреляционных связей между группами исходных переменных: факторами $x_1, \dots, x_n$ и качественными показателями $y_1, \dots, y_m$, $m < n$. В случае линейной зависимости между какими-либо элементами двух групп корреляция достигает максимального значения, равного единице. Поэтому канонический анализ позволяет оценить степень тесноты различных внутригрупповых функциональных связей, а так же определить количество малозначительных факторов и откликов, имеющих между собой наименьшую корреляцию. В связи с малой информативностью последние можно исключить из дальнейшего анализа и тем самым сократить объем данных.

Пусть имеются нецентрированные исходные переменные X и Y .

Пусть выполнены соотношения вида:

$$u = \Theta_1 x_1 + \Theta_2 x_2 + \dots + \Theta_n x_n, \quad (3)$$

$$v = b_1 y_1 + b_2 y_2 + \dots + b_m y_m, \quad (4)$$

причем $u = (u_1, \dots, u_k)$, $v = (v_1, \dots, v_k)$, Θ_i и b_j – неизвестные параметры, $i = 1, n$, $j = 1, m$.

Задача метода заключается в нахождении таких пар векторов $\{u_i, v_j\}$, что корреляция между ними будет наибольшей и u_i будет наилучшим образом предсказываться значениями v_j . Вектора u, v принято называть каноническими переменными. Для нахождения канонических переменных составляется блочная выборочная матрица ковариаций вида:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{pmatrix},$$

где A_{11} – матрица ковариаций МСВ $x_1, \dots, x_n$, A_{22} – матрица ковариаций МСВ $y_1, \dots, y_m$, A_{12} – матрица ковариаций МСВ $x_1, \dots, x_n, y_1, \dots, y_m$.

На практике достаточно объединить данные в одно множество $\{x_1, \dots, x_n, y_1, \dots, y_m\}$ и построить матрицу A . Затем, учитывая размерности A_{11} и A_{22} , вырезать готовые блоки.

Пусть $E(u) = E(v) = 0$, $D(u) = D(v) = 1$. Используя эти соотношения, можно показать, что

$$\text{corr}(u, v) = \Theta^T A_{12} b = \text{cov}(u, v). \quad (5)$$

Для того чтобы (5) достигало максимума, необходимо определить Θ, b из условия экстремума соответствующей функции Лагранжа

$$L = \Theta^T A_{12} b - \frac{\lambda}{2} (\Theta^T A_{11} \Theta - 1) - \frac{\mu}{2} (b^T A_{22} b - 1).$$

Находя частные производные по Θ, b и приравнявая их нулю, получаем систему

$$\begin{cases} A_{12} b - \lambda A_{11} \Theta = 0 \\ A_{12}^T \Theta - \mu A_{22} b = 0, \end{cases}$$

решение которой будет нетривиальным в случае равенства нулю ее определителя. Имеют место соотношения:

$$(A_{11}^{-1}A_{12}A_{22}^{-1}A_{12}^T - \lambda^2 E)\Theta = 0, (A_{22}^{-1}A_{12}^T A_{11}^{-1}A_{12} - \lambda^2 E)b = 0.$$

Таким образом, задача определения максимальной корреляции между каноническими переменными сведена к задаче определения собственных значений матриц $A_{11}^{-1}A_{12}A_{22}^{-1}A_{12}^T$ и $A_{22}^{-1}A_{12}^T A_{11}^{-1}A_{12}$ и их собственных векторов. Учитывая, что размерность матриц равна $n \times n$ и $m \times m$, получаем n собственных чисел $\lambda_1^2 \geq \lambda_2^2 \geq \dots \geq \lambda_n^2$, m собственных векторов b и n собственных векторов Θ .

Теорема 1: числа $\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_n$ равны корреляциям между соответствующими каноническими переменными.

Согласно теореме, корреляция $\text{corr}(u, v)$ достигает максимального значения при наибольшем λ . Поэтому для определения пар $\{u_i, v_j\}$ достаточно найти все λ , упорядочить их по убыванию, затем вычислить соответствующие им b , Θ и, используя соотношения (3), (4), найти требуемые пары канонических переменных.

Замечание: согласно методу канонических корреляций $\text{corr}(u_i, u_j) = \text{corr}(v_i, v_j) = \delta_{ij}$.

Для статистической проверки значимости найденных канонических переменных используется χ^2 – критерий. При этом если определены первые n канонические корреляции, то последовательно для каждого s , $s = \overline{2, n}$, вычисляется статистика

$$\chi_H^2 = - \left\{ k - s - \frac{1}{2}(m + n + 1) + \sum_{p=1}^{s-1} K_{d,p} \right\} \cdot \ln \left(\prod_{p=s}^n (1 - K_{d,p}) \right),$$

где k – размерность векторов $x_1, \dots, x_n, y_1, \dots, y_m$, $K_{d,p}$ – множественный коэффициент корреляции, найденный по переменным $x_1, \dots, x_n$ с алгебраическим дополнением R_{0p} .

Заметим, что вычисление статистики χ_H^2 следует проводить до тех пор, пока подтверждается значимость пар. Как только значимость какой-либо пары, например p -ой, не подтверждается, вычисления прекращаются, так как в этом случае все корреляции, начиная с p -ой, будут равны нулю. Это происходит в силу того, что $0 = \text{corr} = \lambda_p \geq \lambda_{p+1} \geq \dots \geq \lambda_m$.

Проведенный анализ позволяет отсеять /3/ слабокоррелированные факторы и показатели. Полученная таким образом компактная, максимально информативная система данных может служить основой для дальнейших исследований, например, при помощи методов факторного анализа.

МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Обычно в начале исследования некоторого явления с интересующего нас объекта снимается большое число параметров и измеряется большое число характеристик $x_1, \dots, x_n$. Среди них не все являются линейно независимыми друг относительно друга. Если существует достаточно много зависимых между собой признаков, то их нужно исключить, уменьшая избыточную информацию и переходя к признакам $y_1, \dots, y_m$, $m < n$.

Сформулируем задачу метода главных компонент более точно:

- 1) Коррелированность $y_1, \dots, y_m$ уменьшает информацию об объекте, так как корреляция говорит о связи между признаками;

2) чем в более широких пределах меняются признаки, тем более они информативны и тем меньше их число требуется. Поэтому от выбранной совокупности $y_1, \dots, y_m$ потребуем максимальной дисперсии.

Пусть $\xi = (x_1, x_2, \dots, x_n)^T \in \mathcal{R}^n$ – многомерная случайная величина, $\xi \sim N(M, A)$. Пусть $u = (y_1, y_2, \dots, y_m)^T$ или $u = (u_1, u_2, \dots, u_m)^T$ – матрица признаков меньшей размерности или вектор главных компонент, $u \sim N(0, A)$. Рассмотрим систему вида:

$$u_i = (\xi - M)\beta^{(i)}, \quad (6)$$

где $\beta^{(i)}$ – неизвестные коэффициенты, $i = \overline{1, m}$.

Так как $D(u_i) = \beta^{(i)T} A \beta^{(i)}$, то задача нахождения максимума дисперсии $D(u_i)$ выглядит следующим образом:

$$\begin{cases} \beta^{(i)T} A \beta^{(i)} \rightarrow \max \\ \beta^{(i)T} \beta^{(i)} = 1. \end{cases} \quad (7)$$

Последнее равенство является условием ортонормированности коэффициентов системы (6), что гарантирует существование и единственность ее решения.

Задача (7) на нахождение условного экстремума решается методом множителей Лагранжа. Функция Лагранжа при этом будет иметь следующий вид:

$$L = \beta^{(i)T} A \beta^{(i)} - \lambda \beta^{(i)T} \beta^{(i)}.$$

Можно показать, что условный экстремум достигается на решениях матричного уравнения $(A - \lambda E)\beta^{(i)} = 0$. Условие ортонормированности обеспечивает выполнение равенств $D(u_i) = \lambda$, $i = \overline{1, m}$. Значит, для достижения максимума дисперсии необходимо найти все собственные числа λ матрицы ковариаций A и выбрать максимальное. В силу симметричности A будет n различных λ .

Пусть $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq \lambda_{m+1} \geq \dots \geq \lambda_n \geq 0$. Тогда первая компонента u вычисляется в соответствии с (6):

$$u_1 = (\xi - M)\beta^{(1)},$$

где $\beta^{(1)}$ – собственный вектор, соответствующий λ_1 .

Для получения следующей компоненты u_2 наряду с (7) потребуем некоррелированности u_2 с u_1 . Так как $\text{corr}(u_1, u_2) = \frac{\text{cov}(u_1, u_2)}{\sqrt{D(u_1)}\sqrt{D(u_2)}} = 0$ тогда и только тогда, когда

$\text{cov}(u_1, u_2) = \beta^{(1)T} A \beta^{(2)} = 0$, то задача нахождения максимума дисперсии $D(u_2)$ будет следующей:

$$\begin{cases} \beta^{(2)T} A \beta^{(2)} \rightarrow \max \\ \beta^{(2)T} \beta^{(2)} = 1 \\ \beta^{(1)T} A \beta^{(2)} = 0. \end{cases} \quad (8)$$

Решение (8) производится по аналогии с решением (7). Можно показать, что условие некоррелированности не дает новой информации о структуре решения. Как и ранее, для достижения максимума дисперсии необходимо взять максимальное собственное число матрицы ковариаций A . Так как λ_1 задействовано на первом шаге, выбирается λ_2 . Тогда вторая компонента имеет вид:

$$u_2 = (\xi - M)\beta^{(2)},$$

где $\beta^{(2)}$ – собственный вектор, соответствующий λ_2 .

Повторяя процедуру $(n - 2)$ раз, находим $u_1, \dots, u_n$.

После определения всех компонент вектора u возникает вопрос о сокращении их числа, требуемого по условию задачи, т.е. о выделении главных компонент. Для решения этого вопроса рассмотрим так называемый след матрицы ковариаций $sp A = \sum_{i=1}^n a_{ii} = \sum_{i=1}^n \lambda_i$. Известно [2], что $\sum_{i=1}^n D(u_i) = sp A = \sum_{i=1}^n D(x_i) = \sum_{i=1}^n a_{ii}$. Зная $sp A$, выбирают только те главные компоненты $u_1, \dots, u_m$, для которых соответствующие им собственные числа обеспечивают выполнение неравенства:

$$\sum_{i=1}^m \lambda_i \geq \frac{4}{5} sp A,$$

т.е. главные компоненты должны объяснять не менее 80% всей дисперсии.

Наряду со следом матрицы ковариаций можно использовать и ее определитель, так как

$$\det(A_U) = \det(A_X),$$

где A_U – матрица ковариаций признаков, A_X – матрица ковариаций исходных данных.

В этом случае необходимо выполнение неравенства $\prod_{i=1}^m \lambda_i \geq \frac{4}{5} \det A_X$.

ФАКТОРНЫЙ АНАЛИЗ

Факторный анализ является естественным обобщением и развитием метода главных компонент. Если объект описывается с помощью n признаков, то в результате действия метода получается математическая модель, зависящая от меньшего числа переменных. При этом предполагается, что на исходные измеряемые данные $x_1, \dots, x_n$ оказывает влияние небольшое число латентных признаков. Цель факторного анализа заключается в выявлении этих скрытых характеристик (факторов) и оценивании их числа.

Запишем факторную модель в общем виде:

$$X_i = \alpha^{(i)T} U + \varepsilon_i, \quad (9)$$

где $X_i \sim N(0, \sigma^2)$, $U = (u_1, \dots, u_m)$ – факторы, $\alpha^{(i)}$ – факторные нагрузки, ε_i – латентные факторы, $i = \overline{1, n}$.

Техника факторного анализа направлена на определение факторных нагрузок, дисперсий характерных факторов и значений факторов для каждого наблюдаемого объекта.

Однофакторная модель

Пусть вектор U содержит одну компоненту u_1 . Обозначим ее через $f = (f_1, \dots, f_m)$. Тогда (9) можно переписать в виде:

$$X_j^{(i)} = \alpha_j f_i + \varepsilon_{ij},$$

где i – индекс наблюдения, j – индекс компоненты централизованного вектора исходных данных X , $j = \overline{1, n}$, $i = \overline{1, m}$.

Нахождение факторной нагрузки α и фактора f осуществляется методом наименьших квадратов из условия минимизации ε_{ij} :

$$D(\varepsilon) = \sum_{j=1}^n \sum_{i=1}^m \varepsilon_{ij}^2 = \sum_{j=1}^n \sum_{i=1}^m (X_j^{(i)} - \alpha_j f_i)^2 \rightarrow \min \quad (10)$$

Накладывая на $f = (f_1, \dots, f_m)$ условие нормировки $D(f) = \frac{1}{m} \cdot \sum_{i=1}^m f_i^2 = 1$ и дифференцируя (10) по α_j, f_i , получаем следующие соотношения:

$$f_i = \frac{1}{\lambda} \sum_{j=1}^n X_j^{(i)} \alpha_j, \quad \alpha_j = \frac{1}{m} \sum_{i=1}^m X_j^{(i)} f_i, \quad (11)$$

где $\lambda = \sum_{j=1}^n \alpha_j^2$.

Подставляя первое равенство из (11) во второе и учитывая, что $\bar{a}_{sj} = \frac{1}{m} \cdot \sum_{i=1}^m X_s^{(i)} X_j^{(i)}$ – выборочная матрица ковариаций $\bar{A}$ для нормально распределенных случайных величин $X_i \sim N(0, \sigma^2)$, имеем:

$$(\bar{A} - \lambda E) \alpha = 0.$$

Таким образом, получили задачу на собственные числа и собственные вектора выборочной матрицы ковариаций.

По аналогии, подставляя второе равенство из (11) в первое, имеем:

$$\frac{1}{m} \cdot \sum_{s=1}^m \sum_{j=1}^n f_s X_j^{(s)} X_j^{(i)} = \lambda f_i. \quad (12)$$

Если теперь ввести в рассмотрение матрицу P с нецентрированными элементами матрицы S исходных данных $p_{si} = \frac{1}{m} \cdot \sum_{j=1}^n S_j^{(s)} S_j^{(i)}$, то (12) может быть записано в виде:

$$(P - \lambda E) f = 0.$$

Значит, фактор f и факторная нагрузка α могут быть найдены как собственные вектора матриц P и $\bar{A}$ соответственно. Следует отметить, что при этом необходимо использовать такое λ , чтобы дисперсия латентных факторов (10) принимала наименьшее значение. Можно показать, что минимум дисперсии достигается при наибольшем собственном числе матриц $\bar{A}$ и P .

Двухфакторная модель

Запишем двухфакторную модель в виде:

$$X_j^{(i)} = \alpha_j^{(1)} f_i^{(1)} + \alpha_j^{(2)} f_i^{(2)} + \varepsilon_{ij}, \quad j = \overline{1, n}, \quad i = \overline{1, m}, \text{ причем на } f_i^{(1)}, f_i^{(2)} \text{ накладывают условие}$$

взаимной некоррелированности $\sum_{i=1}^m f_i^{(1)} f_i^{(2)} = 0$ и условие нормировки

$$D(f^{(1)}) = D(f^{(2)}) = 1.$$

Нахождение факторных нагрузок и факторов осуществляется с помощью метода неопределенных множителей Лагранжа из условия минимизации ε_{ij} . Так как

$$D(\varepsilon) = \sum_{j=1}^n \sum_{i=1}^m (X_j^{(i)} - \alpha_j^{(1)} f_i^{(1)} - \alpha_j^{(2)} f_i^{(2)})^2, \text{ то функция Лагранжа определяется следующим}$$

равенством:

$$L = \sum_{i,j} (X_j^{(i)} - \alpha_j^{(1)} f_i^{(1)} - \alpha_j^{(2)} f_i^{(2)})^2 + \tilde{\lambda} \sum_{i=1}^m (f_i^{(2)})^2 + 2\tilde{\mu} \sum_{i=1}^m (f_i^{(1)} f_i^{(2)}), \text{ где } \tilde{\lambda}, \tilde{\mu} \text{ — неизвестные}$$

множители.

Для нахождения условного экстремума дифференцируем функцию Лагранжа по $\alpha_j^{(2)}, f_i^{(2)}$ и приравняем найденные производные нулю:

$$\lambda f_i^{(2)} + \mu f_i^{(1)} = \sum_{j=1}^n X_j^{(i)} \alpha_j, \quad \alpha_j^{(2)} = \frac{1}{m} \sum_{i=1}^m X_j^{(i)} f_i^{(2)}, \quad (13)$$

где $\lambda = \tilde{\lambda} + \sum_{j=1}^n (\alpha_j^{(2)})^2$, $\mu = \tilde{\mu} + \sum_{j=1}^n \alpha_j^{(1)} \alpha_j^{(2)}$.

Можно показать, что $\mu = 0$. Тогда (13) совпадает с (11) и имеет место случай однофакторной модели, причем за фактор взят вектор $f^{(2)}$. Решение этой задачи известно: $\alpha^{(2)}$ – собственные вектора матрицы ковариаций $\sigma_{sj} = \frac{1}{m} \cdot \sum_{i=1}^m X_s^{(i)} X_j^{(i)}$, $\lambda = \sum_{j=1}^n (\alpha_j^{(2)})^2$ – ее собственные числа. При этом дисперсия $D(\varepsilon)$ достигает минимума, если $\lambda = \lambda_1 = \sum_{j=1}^n (\alpha_j^{(1)})^2$ и $\lambda = \lambda_2 = \sum_{j=1}^n (\alpha_j^{(2)})^2$ будут первым и вторым наибольшими по величине собственными числами матрицы $\bar{A}$.

Вводя в рассмотрение матрицу P с центрированными элементами исходных данных по формуле $p_{si} = \frac{1}{m} \sum_{j=1}^n X_j^{(s)} X_j^{(i)}$ и учитывая выполнение соотношения $(P - \lambda E)f = 0$, полученного ранее в однофакторной модели, определим факторы $f^{(1)}$, $f^{(2)}$ как собственные вектора матрицы P , соответствующие ее первым максимальным по абсолютной величине собственным числам λ_1 и λ_2 .

Рассмотрение общего случая p -факторной модели производится в полном соответствии с двух- или однофакторной моделями. Однако следует помнить, что при вычислении факторов и факторных нагрузок необходимо задействовать уже p наибольших по величине собственных чисел и p соответствующих им собственных векторов.

Найденные $f^{(1)}$, $f^{(2)}$ образуют в пространстве признаков новый базис, а $\alpha_j^{(1)}$, $\alpha_j^{(2)}$ играют роль координат $X_1, \dots, X_n$ в этом базисе.

После определения факторов исследователю зачастую требуется оценить уровень информативности или вклад фактора в суммарную дисперсию всех признаков.

Определение: пусть имеется n -факторная модель. Пусть $f^{(p)}$ – некоторый фактор, $p \leq n$. Вкладом фактора $f^{(p)}$ в суммарную дисперсию всех признаков называется число $V^{(p)} = \sum_{j=1}^n [\alpha_j^{(p)}]^2 = \lambda_p$, где λ_p – собственное число выборочной матрицы ковариаций $\bar{A}$.

Очевидно, что для n -факторной модели общая дисперсия есть $V_0 = \sum_{p=1}^n V^{(p)}$. В этом случае V_0 называют еще суммарной общностью факторов $f^{(1)}, \dots, f^{(n)}$.

Определение: долей фактора $f^{(p)}$ в суммарной общности называется отношение $V^{(p)}/V_0$. Оно характеризует долю, которую вносит фактор $f^{(p)}$ в факторную модель.

Определение: пусть $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ – собственные числа выборочной матрицы ковариаций $\bar{A}$. Пусть собственному числу λ_p соответствует фактор $f^{(p)}$ и имеются факторные нагрузки $\alpha_{s_1}^{(p)}, \alpha_{s_2}^{(p)}, \dots, \alpha_{s_k}^{(p)}$, которые, в свою очередь, являются наибольшими по абсолютной величине координатами вектора $\alpha^{(p)}$. Тогда число

$K_u(\alpha^{(p)}) = \frac{\sum_{j=1}^k (\alpha_{s_j}^{(p)})^2}{\|\alpha^{(p)}\|^2}$ называется коэффициентом информативности признаков $X^{(s_1)}, \dots, X^{(s_k)}$.

Данное число определяет, какие вектора из множества $X_1, \dots, X_n$ вносят наибольший вклад в название $f^{(p)}$. Принято считать набор объясняющих признаков $X_{s_1}, \dots, X_{s_k}$ удовлетворительным, если $0,7 \leq K_u(\alpha^{(p)}) \leq 1$.

КЛАСТЕРНЫЙ АНАЛИЗ

Кластерный анализ – это анализ, позволяющий получить разбиение большого объема данных на классы или группы (от англ. *cluster*) согласно некоторому критерию или их совокупности. При этом считается, что отсутствует дополнительная информация о характере исходных данных. Кластерный анализ позволяет

- 1) провести классификацию объектов с учетом признаков, отражающих их сущность;
- 2) проверить предположения о наличии некоторой структуры в совокупности этих объектов;
- 3) построить новые классификации для слабоизученных явлений, когда необходимо установить наличие связей внутри совокупности и структурировать их.

Методы кластерного анализа подразделяются на агломеративные (объединяющие), дивизимные (разделяющие) и итеративные. Особенностью последних является формирование кластеров исходя из условий разбиения (т.н. параметров), которые могут быть изменены в процессе работы алгоритма для улучшения качества разбиения.

Для проведения классификации данных $X_1, \dots, X_n$ используют понятие метрики или расстояния.

Определение: метрикой называется функция $\rho: M \rightarrow \mathbb{R}$, отображающая некоторое метрическое пространство в пространство действительных чисел и обладающая следующими свойствами (аксиомами метрики):

- 1) $\rho(X, Y) \geq 0$, 2) $\rho(X, Y) = \rho(Y, X)$, 3) $\rho(X, Y) = 0 \Leftrightarrow X = Y$, 4) $\rho(X, Y) \leq \rho(X, Z) + \rho(Z, Y)$.

В теории кластерного анализа используются следующие метрики:

- 1) Евклидово расстояние

$$\rho^2(X^{(i)}, X^{(j)}) = \sum_{s=1}^m (X_s^{(i)} - X_s^{(j)})^2;$$

- 2) Хеммингово (city-block) расстояние

$$\rho(X^{(i)}, X^{(j)}) = \sum_{s=1}^m |X_s^{(i)} - X_s^{(j)}|$$

- 3) расстояние (или угол) Махаланобиса

$$\rho^2(X^{(i)}, X^{(j)}) = (X^{(i)} - X^{(j)})^T \Lambda^T A^{-1} \Lambda (X^{(i)} - X^{(j)}),$$

где Λ – симметричная положительно–определенная матрица весовых коэффициентов, A – матрица ковариаций $X_1, \dots, X_n$

- 4) расстояние Минковского

$$\rho^p(X^{(i)}, X^{(j)}) = \sum_{s=1}^m |X_s^{(i)} - X_s^{(j)}|^p.$$

Расстояния 1) или 4) используют в случае нормального распределения независимых случайных величин $X_1, \dots, X_n$ или их однородности по физическому смыслу, когда каждый вектор одинаково важен для классификации.

Задача выбора метрики для проведения кластерного анализа не имеет единственного решения, так как процесс выбора неформализуем и зависит от многих факторов, в частности, от ожидаемого результата, от опыта исследователя, уровня его математической подготовки и т.д.

В ряде алгоритмов наряду с расстояниями между векторами используются расстояниями между кластерами и объединениями кластеров. Более подробно об этом можно узнать в [1].

Агломеративные методы

Агломеративные методы являются наиболее простыми и распространенными среди алгоритмов кластерного анализа. На первом шаге каждый вектор или объект $X_1, \dots, X_n$ исходных данных рассматривается как отдельный кластер или класс. По вычисленной матрице расстояний $R = (-\rho(X^{(i)}, X^{(j)}))$, ρ – некоторая метрика, выбираются наиболее близкие друг к другу и объединяются. Очевидно, что процесс завершится через $(n - 1)$ шаг, когда в результате все объекты будут объединены в один кластер.

К агломеративным относят методы одиночной, средней, полной связи и метод Уорда. С математической точки зрения последний из них наиболее интересен. Остановимся на нем более подробно.

Пусть $X_1, \dots, X_n$ – данные, причем каждый вектор образует один кластер. Найдем матрицу расстояний, используя какую-нибудь метрику, определяем по ней наиболее близкие друг к другу кластеры. Вычисляем сумму квадратов отклонений векторов внутри кластера S_k по формуле:

$$V_k = \sum_{i=1}^{n_k} \sum_{j=1}^p (X_{ij} - \bar{X}_{jk})^2,$$

где k – номер кластера, i – номер вектора в кластере, j – номер координаты $X_i \in \mathbb{R}^p$, n_k – число векторов в кластере, $\bar{X}_{jk}$ – выборочное среднее X_j в S_k .

В дальнейшем в кластер S_k добавляются вектора или кластеры, приводящие к наименьшему изменению V_k и, как следствие, расположенные на минимальном расстоянии от S_k .

Итеративные методы

Сущность итеративных методов заключается в том, что кластеризация начинается с задания некоторых начальных условий. Требуется задать число кластеров, которые необходимо получить, расстояние, определяющее конец процесса образования кластеров и т.д. Начальные условия выбираются согласно результату, который нужен исследователю. Однако обычно они задаются по решению, найденному одним из агломеративных методов.

1) Метод k – средних.

Пусть имеются вектора $X_1, \dots, X_n \in \mathbb{R}^p$ и их необходимо разбить на k кластеров. На нулевом шаге из n векторов случайным образом выбираем k элементов. Пусть каждый образует один кластер. Получаем множество кластеров-эталонов $e_1^{(0)}, \dots, e_k^{(0)}$ с весами $\omega_1^{(0)}, \dots, \omega_k^{(0)}$, определяющими число элементов в них. На следующем шаге из оставшегося набора данных выбираем некоторый вектор, например, X_i и вычисляем матрицу расстояний между X_i и $e_1^{(0)}, \dots, e_k^{(0)}$ по евклидовой метрике. Затем X_i помещается в тот эталон, расстояние до которого минимально. Допустим для определенности, что это $e_m^{(0)}$. Он заменяется новым, пересчитанным с учетом присоединенной точки, по формуле:

$$e_m^{(1)} = \begin{cases} \frac{\omega_m^{(0)} e_m^{(0)} + X_i}{\omega_m^{(0)} + 1}, & X_i \text{ включен} \\ e_m^{(0)} & , X_i \text{ не включен.} \end{cases}$$

Кроме того, пересчитывается и вес:

$$\omega_m^{(1)} = \begin{cases} \omega_m^{(0)} + 1, & X_i \text{ включен} \\ \omega_m^{(0)} & , X_i \text{ не включен.} \end{cases}$$

Если в матрице встречается два или более минимальных расстояния, то X_i включают в кластер с наименьшим порядковым номером.

На следующем шаге выбирают следующий вектор из оставшихся и процедура повторяется. Таким образом, через $(n-k)$ шагов каждому эталону $e_m^{(n-k)}$ будет соответствовать вес $\omega_m^{(n-k)}$ и процедура кластеризации завершится.

Можно показать, что при большом n и малом k алгоритм быстро сходится к устойчивому решению. Тем не менее, алгоритм всегда пересчитывают несколько раз, используя полученное разбиение в качестве векторов – эталонов (как начальное приближение).

2) Метод поиска сгущений.

Этот алгоритм не требует априорного задания числа кластеров. На первом шаге вычисляется матрица расстояний между $X_1, \dots, X_n \in \mathbb{R}^p$. Затем случайным образом выбирают один вектор, который будет играть роль центра первого кластера. Положим, что этот вектор лежит в центре p -мерной сферы радиуса R , задаваемого исследователем. После этого определяют вектора, попавшие внутрь этой сферы, и по ним вычисляется выборочное математическое ожидание $\bar{X}$. Центр сферы переносится в $\bar{X}$ и процедура повторяется. Первый кластер считается образованным окончательно, если вектор средних $\bar{X}$ при вычислениях на предыдущем и последующем шаге остается неизменным.

Попавшие внутрь сферы элементы $X_{s_1}, \dots, X_{s_k}$ заключаем в кластер и исключаем из дальнейшего исследования. Для оставшихся точек алгоритм повторяется. Можно показать [2], что алгоритм сходится при любом выборе начального приближения и любом объеме исходных данных. Однако для получения устойчивого разбиения рекомендуется повторить алгоритм несколько раз при различных значениях радиуса сферы R . Признаком устойчивого разбиения будет образование одного и того же числа кластеров с одним и тем же составом.

Функционалы качества разбиения

Заметим, что задача кластеризации может иметь континуальное число решений, если число исходных данных счетно. Как следствие, перебрать все возможные разбиения данных на классы не представляется возможным. Для того чтобы оценить качество различных способов кластеризации вводят понятие функционала качества разбиения, который принимает минимальное значение на наилучшем (с точки зрения исследователя) разбиении.

Пусть $S = \{S_1, \dots, S_k\}$ – некоторая совокупность кластеров, k известно. Тогда основные функционалы качества разбиения при известном числе кластеров имеют вид:

1) Взвешенная сумма внутриклассовых дисперсий $Q_1(S) = \sum_{l=1}^k \sum_{X_m \in S_l} \rho^2(X_m, a(l))$, где $a(l)$ –

выборочное математическое ожидание класса S_l .

Функционал $Q_1(S)$ позволяет оценить меру однородности всех кластеров в целом.

2) Сумма попарных внутриклассовых расстояний между элементами

$$Q_2(S) = \sum_{l=1}^k \sum_{X_i, X_j \in S_l} \rho^2(X_i, X_j).$$

3) Обобщенная внутриклассовая дисперсия $Q_3(S) = \det\left(\sum_{l=1}^k n_l A_l\right)$, где n_l – число элементов в S_l , A_l – выборочная ковариационная матрица для S_l .

Функционал $Q_3(S)$ является средней арифметической характеристикой обобщенных внутриклассовых дисперсий, посчитанных для каждого кластера. Как известно, обобщенная дисперсия позволяет оценить степень рассеивания многомерных наблюдений. Поэтому $Q_3(S)$ определяет средний разброс векторов наблюдений в классах $S_1, \dots, S_k$. Он применяется в случае, когда необходимо решить задачу о сжатии данных или о сосредоточении наблюдений в пространстве с размерностью меньше исходной.

Замечание: оценить качество разбиения на классы можно и эмпирически. Например, можно сравнивать выборочные средние, найденные для каждого класса, со средним всей совокупности наблюдений. Если они разнятся в два раза и более, то разбиение хорошее.

Если число кластеров в $S = \{S_1, \dots, S_k\}$ заранее неизвестно, то используют следующие функционалы качества разбиения:

1) $Q_1(S) = \alpha I_m(S) + \frac{\beta}{Z_m(S)}$, $\alpha, \beta > 0$, где

$$I_m(S) = \left\{ \frac{1}{n} \sum_{i=1}^k \frac{1}{n_i} \sum_{X_j \in S_i} \sum_{X_l \in S_i} \rho^m(X_j, X_l) \right\}^{\frac{1}{m}} - \text{средняя мера внутриклассового рассеяния,}$$

$$Z_m(S) = \left\{ \frac{1}{n} \sum_{i=1}^n \left(\frac{V(X_i)}{n} \right)^m \right\}^{\frac{1}{m}} - \text{мера концентрации точек множества } S, \quad V(X_i) - \text{число}$$

элементов в кластере, содержащем точку X_i .

2) $Q_2(S) = (I_m(S))^\alpha (Z_m(S))^\beta$, $\alpha, \beta > 0$.

Заметим, что в случае неизвестного числа кластеров функционалы качества разбиения $Q(S)$ можно выбирать в виде алгебраической комбинации (суммы, разности, произведения, отношения) двух функционалов $I_m(S)$, $Z_m(S)$, так как первый является убывающей, а другой – возрастающей функцией числа классов k . Такое поведение $I_m(S)$, $Z_m(S)$ гарантирует существование экстремума $Q(S)$.

ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Задача дискриминантного анализа заключается в разработке методов различения объектов наблюдения по известным признакам, называемым эталонами. Например, совокупность клиентов банка можно разбить на два подмножества людей: «кредитоспособны» и «некредитоспособны». Более точно задача различения заключается в следующем: пусть X_1 – наблюдение над некоторым объектом. Требуется установить правило, которое по значению X_1 переводит его в одну из возможных совокупностей объектов $\pi_i, i = \overline{1, n}$. Для построения этого правила переходят от векторов признаков, характеризующих объект, к линейной функции от них. Эту функцию называют дискриминантной функцией или решающим правилом или процедурой классификации.

Пусть имеются классы объектов $\pi_i, i = \overline{1, n}$, причем каждый класс описывается k -мерной функцией плотности нормального распределения

$f_i(X) = (2\pi)^{-k/2} \cdot \det(A)^{-1/2} \exp\left(-\frac{1}{2}(X - M^{(i)})^T A^{-1}(X - M^{(i)})\right)$, где $M^{(i)}$ – теоретическое математическое ожидание размерности k для π_i , A – теоретическая матрица ковариаций векторов из π_i , $M^{(i)}$, A известны. Будем относить точку $X \in \mathfrak{R}^k$ классу π_i , если $f_i(X)$ принимает наибольшее значение среди $f_1(X), \dots, f_n(X)$. Можно показать, что в этом случае $X \in \mathfrak{R}^k$ будет доставлять наибольшее значение дискриминантной функции:

$$L_i(X) = X^T A^{-1} \bar{M}^{(i)} - \frac{1}{2} (\bar{M}^{(i)})^T A^{-1} \bar{M}^{(i)}, \quad (14)$$

где $\bar{M}^{(i)}$ – выборочное среднее.

Таким образом, получили правило, по которому каждая точка $X \in \mathfrak{R}^k$ будет к заданному классу π_i .

Так как тема дискриминантного анализа не входит в лабораторный практикум, приведем поясняющий пример.

Пример: исследование налоговой службы показало, что склонность фирм к утаиванию части доходов определяется двумя показателями: X_1 – соотношением «быстрых активов» и текущих пассивов и X_2 – соотношением прибыли и просроченных платежей. Известно, что двумерный признак $X = (X_1, X_2)$ распределен нормально внутри совокупностей π_1 = «фирма уклоняется от уплаты налогов» и π_2 = «фирма платит налоги». Пусть ковариационные матрицы для π_1 и π_2 совпадают и равны $A = \begin{pmatrix} 8467 & 2041 \\ 2041 & 4273 \end{pmatrix}$. Пусть $M^{(1)} = (576, 596)$, $M^{(2)} = (598, 5; 710, 8)$ – теоретические математические ожидания π_1 и π_2 соответственно. Для фирмы, не проходившей проверку налоговой службы, с вектором средних $M = (740, 590)$ определить, в какую совокупность она попадет.

Решение: воспользуемся линейной дискриминантной функцией (14). В нашем случае $X = (740, 590)$, $X_1 = 740$, $X_2 = 590$, обратная матрица ковариации $A^{-1} = \begin{pmatrix} 1,3 \cdot 10^{-4} & -6,3 \cdot 10^{-5} \\ -6,3 \cdot 10^{-5} & 2,6 \cdot 10^{-4} \end{pmatrix}$. Поэтому $L_1(X) = 52,88$; $L_2(X) = 50,387$. Так как $L_1(X) > L_2(X)$ на $X = (740, 590)$, то фирму относим к классу π_1 .

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ВЫПОЛНЕНИЮ ЛАБОРАТОРНОЙ РАБОТЫ №1

Рассмотрим приложение теории МСМ для расчетов лабораторного задания №1 (см. приложение 1). Технические вычисления будем производить каким-нибудь математическим пакетом, например MathCad, функциями рабочего листа MS Excel, Gretl и т.д.

Пусть исходные данные расположены в столбцах Y3, X8, X9, X10, X11, X17. Они образуют шестимерную МСВ. Для нахождения вектора выборочного математического ожидания для каждого столбца находим его среднее. Тогда $M=(13.7, 1.072, 0.486, 1.53, 14707.8, 19.57)$. После этого центрируем данные, вычитая из каждого столбца соответствующее среднее. Получаем центрированную матрицу X, которую можно использовать при нахождении выборочной ковариационной матрицы:

$$A = \frac{1}{n} \cdot X^T \cdot X, \text{ т.е.}$$

$$A = \begin{pmatrix} 38.902 & 3.005 & -0.04 & 0.92 & 1.899 \times 10^3 & -9.578 \\ 3.005 & 0.448 & 0.024 & -0.03 & 2.073 \times 10^3 & -0.902 \\ -0.04 & 0.024 & 0.116 & -0.039 & -76.133 & -0.6 \\ 0.92 & -0.03 & -0.039 & 0.163 & -406.947 & 0.07 \\ 1.899 \times 10^3 & 2.073 \times 10^3 & -76.133 & -406.947 & 9.63 \times 10^7 & 51.3 \\ -9.578 & -0.902 & -0.6 & 0.07 & 51.3 & 21.696 \end{pmatrix}$$

Для нахождения частных коэффициентов корреляции необходимо воспользоваться формулой

$$r_{ij} = \frac{-R_{ij}}{\sqrt{R_{ii} R_{jj}}},$$

где R_{ij} – алгебраическое дополнение элемента r_{ij} матрицы корреляций вида:

$$R = \begin{pmatrix} 1 & 0.719 & -0.019 & 0.365 & 0.031 & -0.33 \\ 0.719 & 1 & 0.104 & -0.111 & 0.316 & -0.289 \\ -0.019 & 0.104 & 1 & -0.282 & -0.023 & -0.378 \\ 0.365 & -0.111 & -0.282 & 1 & -0.103 & 0.037 \\ 0.031 & 0.316 & -0.023 & -0.103 & 1 & 1.122 \times 10^{-3} \\ -0.33 & -0.289 & -0.378 & 0.037 & 1.122 \times 10^{-3} & 1 \end{pmatrix}.$$

Отметим, что для нахождения алгебраических дополнений R_{ij} необходимо использовать известную теорему Лапласа о вычислении определителя. Так, для нахождения R_{ij} достаточно скопировать найденную выше выборочную матрицу корреляций, обнулить в ней все элементы j – столбца, кроме r_{ij} , найти определитель и умножить его на $(-1)^{i+j}$. Поступая таким образом, имеем:

$$r_{11} = -1, r_{12} = -0.592, r_{13} = 10^{-3}, r_{14} = -0.234, r_{15} = -9.8 \cdot 10^{-3}, r_{16} = -7.6 \cdot 10^{-2}.$$

Аналогично вычисляем и множественный коэффициент корреляции $R_y^2 = 0.762$. Так как $R_y^2 > 0.5$, то связь между векторами исходных данных Y3 и X8, X9, X10, X11, X17 сильная и использование регрессионной модели для анализа теоретически обоснованно.

Для построения регрессионной модели воспользуемся равенством (1). Заметим, что матрица X наряду с нецентрированными векторами $X_8, X_9, X_{10}, X_{11}, X_{17}$ должна содержать дополнительный столбец, составленный из одних единиц. Это необходимо для определения параметра θ_0 в (1).

Вставить единичный столбец в X можно с помощью команды *augment*.

Находим θ :

$$\Theta^T = (X^T X)^{-1} X^T Y = (0.98; 7.3; -0.59; 6.7; -1.1 \cdot 10^{-4}; -0.17).$$

Проведем статистическое оценивание регрессионной модели. Вычисляем статистику

$$F_H = \frac{m-n-1}{n+1} \cdot (X\Theta)^T (X\Theta) [(Y - (X\Theta))^T (Y - (X\Theta))]^{-1},$$

где $m=53, n=5$.

Так как $F_H = 288.234$, а $F_T(0.05; 6; 47) = 2.298956$, то надежность регрессионной модели подтверждается.

Статистическое оценивание надежности коэффициентов регрессии θ_i произведем с помощью t – критерия Стьюдента. Вычисляем статистику:

$$t_H(i) = \frac{\theta_i}{s_i},$$

где $s_i = \sqrt{\frac{(Y - X\Theta)^T \cdot (Y - X\Theta) \cdot c_{ii}}{m-n-1}}$ – средняя ошибка для θ_i , c_{ii} – диагональные элементы матрицы $(X^T X)^{-1}$.

Получаем, что $t_H(0) = 0.273$, $t_H(1) = 9.978$, $t_H(2) = -0.4$, $t_H(3) = 5.743$, $t_H(4) = -2.303$, $t_H(5) = -1.615$.

Наблюдаемое значение $t_H(i)$ сравниваем с табличным $t_T(0.05; 47) = 1.677927$. Очевидно, что значимость коэффициентов $\theta_1, \theta_3, \theta_4$ подтверждается. Коэффициенты $\theta_0, \theta_2, \theta_5$ незначимы.

Наконец, произведем статистическое оценивание вычисленного ранее множественного коэффициента корреляции $R_y^2 = 0.762$. Определяем величину статистики Снедекора:

$$F_H = \frac{m-n}{n-1} \cdot (R_y^2) \cdot [1 - R_y^2]^{-1}.$$

Наблюдаемое значение $F_H = 38.42$ сравниваем с табличным $F_T(0.05; 4; 48) = 2.565241$. Очевидно, что значимость R_y^2 подтверждается, так как $F_H > F_T$.

В заключение лабораторной работы №1 выберем результативный показатель.

Наибольший по абсолютной величине коэффициент уравнения регрессии (1) равен $\theta_1 = 7.3$. Ему соответствует столбец X_8 исходных данных (вектор X_1 в (1)). Значит, это и есть искомый результативный показатель.

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ВЫПОЛНЕНИЮ ЛАБОРАТОРНОЙ РАБОТЫ №2

Рассмотрим приложение теории МСМ для расчетов лабораторного задания №2 (см. приложение 2). Технические вычисления будем производить каким-нибудь математическим пакетом, например MathCad, функциями рабочего листа MS Excel, Gretl и тд..

Пусть исходные данные расположены в столбцах

Y2, Y3, X1 – X8 матрицы X , т.е. заданы 2-х и 8-мерные МСВ. Для нахождения объединенного вектора выборочного математического ожидания для каждого столбца найдем его среднее. Тогда $M=(5152.2; 40.2; 2402.6; 2.78; 285.8; 334.8; 29.3; 224.7; 432.4; 40.6)$. После этого центрируем данные, вычитая из каждого столбца соответствующее среднее. Получаем центрированную матрицу, которую можно использовать при нахождении выборочной ковариационной матрицы.

$$A = \frac{1}{30} X^T \cdot X.$$

В методе канонических корреляций она будет иметь блочный вид. Для «вырезания» блоков используется команда *submatrix*. Тогда, например, ковариационная матрица для Y следующая:

$$A_{22} = \begin{pmatrix} 6.165 \times 10^6 & 2.902 \times 10^4 \\ 2.902 \times 10^4 & 192.195 \end{pmatrix}.$$

Вид остальных матриц достаточно объемен, поэтому они приводиться не будут.

Строим матрицы $A_{11}^{-1}A_{12}A_{22}^{-1}A_{12}^T$, $A_{22}^{-1}A_{12}^TA_{11}^{-1}A_{12}$ и находим их собственные числа (команда *eigenvals*). У $A_{11}^{-1}A_{12}A_{22}^{-1}A_{12}^T$ это $\lambda_1 = 0.492$, $\lambda_2 = 0.723$, $\lambda_3 = 0$, у $A_{22}^{-1}A_{12}^TA_{11}^{-1}A_{12}$ – $\lambda_1 = 0.492$, $\lambda_2 = 0.723$.

Находим соответствующие им собственные вектора (команда *eigenvec*):

$$\Theta^{(1)} = (0; 1; 0; 0; -0.02; 0; 0; 0),$$

$$\Theta^{(2)} = (0; -0.98; 0.02; -0.01; 0.08; -0.04; 0; -0.19),$$

$$\Theta^{(3)} = (0; -0.85; 0; 0.08; -0.06; 0.12; 0.05; 0.5), \quad b^{(1)} = (0; 1), \quad b^{(2)} = (0.08; 1).$$

Согласно (3), (4), для первых двух собственных векторов строим канонические переменные $U^{(1)}, U^{(2)}, V^{(1)}, V^{(2)} \in \mathbb{R}^{30}$. Из-за их объемности точный координатный вид приводиться не будет.

Из теории метода канонических корреляций следует, что корреляция между $U^{(1)}$ и $V^{(1)}$ равна $\lambda_1 = 0.723$ – наибольшему собственному числу. Корреляция между $U^{(2)}$ и $V^{(2)}$ равна $\lambda_2 = 0.492$. Остальные корреляции равны нулю.

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ВЫПОЛНЕНИЮ ЛАБОРАТОРНОЙ РАБОТЫ №3

Рассмотрим приложение теории МСМ для расчетов лабораторного задания №3 (см. приложение 3). Технические вычисления будем производить каким-нибудь математическим пакетом, например MathCad, функциями рабочего листа MS Excel, Gretl и тд..

Пусть исходные данные расположены в столбцах X3, X8, X9, X10, X11, X17. Согласно этим данным строим матрицу ковариаций A . Она будет совпадать с аналогичной матрицей из лабораторной работы №1 (это же касается вектора математического ожидания и корреляционной матрицы), поэтому достаточно ее скопировать из отчета.

Так как X разнородны, переходим к стандартизированным данным:

$$Z_{it} = (X_{it} - \mu_i) \sigma_i^{-1}, \quad 1 \leq i \leq n, \quad 0 \leq t \leq N,$$

где μ_i – математическое ожидание признака X_i , $\sigma_i = [E((X_{it} - \mu_i)^2)]^{1/2}$ – его волатильность, $n=6$, $N=53$.

Проведем обработку Z_{it} методом главных компонент. Для этого вычисляем собственные числа матрицы ковариаций для Z (которая будет совпадать с ее матрицей

корреляций) и упорядочиваем их по убыванию: $\lambda_6 = 0.12, \lambda_5 = 0.52, \lambda_4 = 0.69, \lambda_3 = 1.16, \lambda_2 = 1.48, \lambda_1 = 2.02$. Определяем след матрицы A . Он будет равен $sp A = 6$. Затем выбираем λ_i таким образом, чтобы выполнялось неравенство $\sum_{i=1}^m \lambda_i \geq \frac{4}{5} sp A$, где $m < 6$. Однако при осуществлении такого выбора не стоит забывать об основной цели метода главных компонент – сжатии информации. Поэтому всегда следует добиваться минимального количества таких λ_i -х.

Для выполнения неравенства следует выбрать четыре собственных числа, так как $\sum_{i=1}^4 \lambda_i = 0.89$, а $\sum_{i=1}^3 \lambda_i = 0.78$. Поэтому для нахождения главных компонент Y стандартизированных данных Z выберем собственные вектора–столбцы, соответствующие $\lambda_1, \dots, \lambda_4$, и составим из них матрицу L_1 . Тогда $Y = Z \cdot L_1$. Для перехода к искомым главным компонентам U воспользуемся соотношением вида:

$$u_{it} = \mu_i + \sigma_{it} \sum_{s=1}^n y_{is} l_{st}^T, \quad 1 \leq i \leq n, \quad 0 \leq t \leq N,$$

где l_{it}^T – элементы матрицы L_1^T .

В связи с объемом получаемых результатов вид главных компонент приведен не будет.

Заметим, что уже на этапе метода главных компонент можно сделать выводы о числе факторов, которые нужно использовать в факторном анализе, о факторных нагрузках, о векторах, вносящих наибольший вклад в название факторов. Если говорить о данном примере, то в дальнейшем следует остановить свой выбор на построении четырехфакторной модели. Далее, для вычисления матрицы факторных нагрузок S нужно найти матрицу $\Lambda = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_6})$, объединить все собственные вектора–столбцы, соответствующие $\lambda_i, i = \overline{1,6}$, в матрицу L и применить формулу $S = L\Lambda$, причем, для проверки, $S^T S = \Lambda^2 = \text{diag}(\lambda_1, \dots, \lambda_6)$. Так, для рассматриваемого нами примера нагрузки первых двух главных компонент будут иметь вид:

$$\alpha_1 = (0.85; 0.88; 0.27; 0.08; 0.28; -0.62), \quad \alpha_2 = (-0.42; -0.02; 0.74; -0.79; 0.09; -0.34).$$

Доли факторов в суммарной общности равны 0,34; 0,25; 0,19 и 0,115 соответственно.

Определим название, например, первого фактора. Для этого посчитаем для него коэффициент информативности признаков. В нашем случае $\alpha_1 = (0.85; 0.88; 0.27; 0.08; 0.28; -0.62)$, $\|\alpha_1\|^2 = 2.023$. Информационное множество выберем в виде (0,85; 0,88). Поэтому

$$K_u(\alpha_1) = \frac{0,85^2 + 0,88^2}{2,023} \approx 0,74 > 0,7,$$

т.е. набор признаков можно считать удовлетворительным. Кроме того, первый и второй вектор исходных данных вносят наибольший вклад в название первого фактора (а это столбцы X_3 и X_8), поэтому их названия можно перенести на первый фактор.

ЛИТЕРАТУРА

1. Сошникова Л.А., Тамашевич В.Н., Уебе Г., Шефер М. Многомерный статистический анализ в экономике. М.: ЮНИТИ-ДАНА, 2015. – 598 с.
2. Айвазян С.А., Мхитарян В.С. Теория вероятностей и прикладная статистика. М.: ЮНИТИ-ДАНА, т.1, 2001. – 656 с.
3. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. М.: Финансы и статистика, 2018. – 352 с.
4. Гантмахер Ф.Р. Теория матриц. М.: Наука, 1967 г.

Контрольная работа**Задание №1****РЕГРЕССИОННЫЙ АНАЛИЗ,
ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ
МНОГОМЕРНЫХ СЛУЧАЙНЫХ ВЕЛИЧИН**

1. В соответствии с предложенными данными:
 - а) определить вектор выборочного математического ожидания;
 - б) определить выборочную матрицу ковариаций, корреляций;
 - в) найти шесть частных коэффициентов корреляции,
 - г) найти множественный коэффициент корреляции.
2. По матрице исходных данных построить уравнение регрессии. Выбрать результативный показатель (вектор данных, которому соответствует наибольший коэффициент уравнения регрессии).
3. Произвести статистическое оценивание регрессионной модели, статистическое оценивание надежности коэффициентов регрессии, статистическое оценивание множественного коэффициента корреляции.
4. Сравнить результаты с результатами, полученными с помощью пакета программ Gretl.

Таблица 1

ВАРИАНТЫ РАСЧЕТА ЗАДАЧ

№ варианта	Y	Номера векторов X
1	Y_1	6,8, 11, 12, 17
2	Y_1	6,8, 11, 13, 17
3	Y_1	8,11,12,13,17
4	Y_1	6,8, 13, 14, 17
5	Y_1	8,11,13,14, 17
6	Y_1	6,8, 12, 13, 17
7	Y_1	7,11,12,13,17
8	Y_1	7,9, 12, 13, 17
9	Y_1	8,11,12,13,17
10	Y_1	8,9,13, 14, 17
11	Y_3	8,10,15, 16, 17
12	Y_3	5,6,10,15, 17
13	Y_3	5,6,7, 11, 12
14	Y_3	8,9,10,11,17
15	Y_3	8, 9,10, 12, 17
16	Y_2	4,5,6,8,9
17	Y_2	4,5,6,7,9
18	Y_2	4,5,6,8,9
19	Y_2	4,5,8,9,17
20	Y_2	4,5,7,9,17

ТАБЛИЦА ИСХОДНЫХ ДАННЫХ

№	Y ₁	Y ₂	Y ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀
1	9.26	204.2	13.26	0.23	0.78	0.40	1.37	1.23	0.23	1.45
2	9.38	209.6	10.16	0.24	0.75	0.26	.49	1.04	0.39	1.30
3	12.11	222.6	13.72	0.19	0.68	0.40	.44	1.80	0.43	1.37
4	10.81	236.7	12.85	0.17	0.70	0.50	.42	0.43	0.18	1.65
5	9.35	62.0	10.63	0.23	0.62	0.40	.35	0.88	0.15	1.91
6	9.87	53.1	9.12	0.43	0.76	0.19	.39	0.57	0.34	1.68
7	8.17	172.1	25.83	0.31	0.73	0.25	.16	1.72	0.38	1.94
8	9.12	56.5	23.39	0.26	0.71	0.44	.27	1.70	0.09	1.89
9	5.88	52.6	14.68	0.49	0.69	0.17	.16	0.84	0.14	1.94
10	6.30	46.6	10.05	0.36	0.73	0.39	.25	0.60	0.21	2.06
11	6.22	53.2	13.99	0.37	0.68	0.33	.13	0.82	0.42	1.96
12	5.49	30.1	9.68	0.43	0.74	0.25	.10	0.84	0.05	1.02
13	6.50	146.4	10.03	0.35	0.66	0.32	.15	0.67	0.29	1.85
14	6.61	18.1	9.13	0.38	0.72	0.02	.23	1.04	0.48	0.88
15	4.32	13.6	5.37	0.42	0.68	0.06	.39	0.66	0.41	0.62
16	7.37	89.8	9.86	0.30	0.77	0.15	.38	0.86	0.62	1.09
17	7.02	62.5	12.62	0.32	0.78	0.08	.35	0.79	0.56	1.60
18	8.25	46.3	5.02	0.25	0.78	0.20	.42	0.34	1.76	1.53
19	8.15	103.5	21.18	0.31	0.81	0.20	.37	1.60	1.31	1.40
20	8.72	73.3	25.17	0.26	0.79	0.30	.41	1.46	0.45	2.22
21	6.64	76.6	19.40	0.37	0.77	0.24	.35	1.27	0.50	1.32

продолжение таблицы 2

22	8.10	73.01	21.0	0.29	0.78	0.10	.48	1.58	0.77	1.48
23	5.52	32.3	6.57	0.34	0.72	0.11	.24	0.68	1.20	0.68
24	9.37	199.6	14.19	0.23	0.79	0.47	.40	0.86	0.21	2.30
25	13.17	598.1	15.81	0.17	0.77	0.53	.45	1.98	0.25	1.37
26	6.67	71.2	5.23	0.29	0.80	0.34	.40	0.33	0.15	1.51
27	5.68	90.8	7.99	0.41	0.71	0.20	.28	0.45	0.66	1.43
28	5.22	82.1	17.50	0.41	0.79	0.24	.33	0.74	0.74	1.82
29	10.02	76.2	17.16	0.22	0.76	0.54	1.22	0.03	0.32	2.62
30	8.16	119.5	14.54	0.29	0.78	0.40	.28	0.99	0.89	1.75
31	3.78	21.9	6.24	0.51	0.62	0.20	.47	0.24	0.23	1.54
32	6.48	48.4	12.08	0.36	0.75	0.64	.27	0.57	0.32	2.25
33	10.44	173.5	9.49	0.23	0.71	0.42	.51	1.22	0.54	1.07
34	7.65	74.1	9.28	0.26	0.74	0.27	.46	0.68	0.75	1.44
35	8.77	68.6	11.42	0.27	0.65	0.37	.27	1.00	0.16	1.40
36	7.00	60.8	10.31	0.29	0.66	0.38	.43	0.81	0.24	1.31
37	11.06	355.6	8.65	0.01	0.84	0.35	.50	1.27	0.59	1.12
38	9.02	264.8	10.94	0.02	0.74	0.42	.35	1.14	0.56	1.16
39	13.28	526.6	9.87	0.18	0.75	0.32	.41	1.89	0.63	0.88
40	9.27	118.6	6.14	0.25	0.75	0.33	.47	0.67	1.10	1.07
41	6.70	37.1	12.93	0.31	0.79	0.29	.35	0.96	0.39	1.24
42	6.69	57.7	9.78	0.38	0.72	0.30	.40	0.67	0.73	1.49
43	9.42	51.6	13.22	0.24	0.70	0.56	.20	0.98	0.28	2.03
44	7.24	64.7	17.29	0.31	0.66	0.42	.15	1.16	0.10	1.84
45	5.39	48.3	7.11	0.42	0.69	0.26	.09	0.54	0.68	1.22
46	5.61	15.0	22.49	0.51	0.71	0.16	.26	1.23	0.87	1.72
47	5.59	87.5	12.14	0.31	0.73	0.45	.36	0.78	0.49	1.75
48	6.57	108.4	15.25	0.37	0.65	0.31	.15	1.16	0.16	1.46
49	6.54	267.3	31.34	0.16	0.82	0.08	.87	4.44	0.85	1.60
50	4.23	34.2	11.56	0.18	0.80	0.68	.17	1.06	0.13	1.47
51	5.22	26.8	30.14	0.43	0.83	0.03	.61	2.13	0.49	1.38
52	18.00	43.6	19.71	0.40	0.70	0.02	.34	1.21	0.09	1.41
53	11.03	72.0	23.56	0.31	0.74	0.22	.22	2.20	0.79	1.39

Таблица 3

ТАБЛИЦА ИСХОДНЫХ ДАННЫХ

№	X ₁₁	X ₁₂	X ₁₃	X ₁₄	X ₁₅	X ₁₆	X ₁₇
1	26006	167.69	47750	6.40	166.32	10.08	17.72
2	23935	186.10	50391	7.80	92.88	14.76	18.39
3	22589	220.45	43149	9.76	158.04	6.48	26.46
4	21220	169.30	41089	7.90	93.96	21.96	22.37
5	7394	39.53	14257	5.35	173.88	11.88	28.13
6	11586	40.41	22661	9.90	162.30	12.60	17.55
7	26609	102.96	52509	4.50	88.56	11.52	21.92
8	7801	37.02	14903	4.88	101.16	8.28	19.52
9	11587	45.74	25587	3.46	166.32	11.52	23.99
10	9475	40.07	16821	3.60	140.76	32.40	21.76
11	10811	45.44	19459	3.56	128.52	11.52	25.68
12	6371	41.08	12973	5.65	177.84	17.28	18.13
13	26761	136.14	50907	4.28	114.48	16.20	25.74
14	4210	42.39	6920	8.85	93.24	13.32	21.21
15	3557	37.39	5736	8.52	126.72	17.28	22.97
16	14148	101.78	26705	7.19	91.80	9.72	16.38
17	9872	47.55	20068	4.82	69.12	16.20	13.21
18	5975	32.61	11487	5.46	66.24	24.84	14.48
19	16662	103.25	32029	6.20	67.68	14.76	13.38
20	9166	38.95	18946	4.25	50.40	7.56	13.69
21	15118	81.32	28025	5.38	70.56	8.64	16.66
22	11429	67.26	20968	5.88	72.00	8.64	15.06
23	6462	59.92	11049	9.27	97.20	9.00	20.09
24	24628	107.34	45893	4.36	80.28	14.76	15.98
25	49727	512.60	99400	10.31	51.48	10.08	18.27
26	11470	53.81	20719	4.69	105.12	14.76	14.42
27	19448	80.83	36813	4.16	128.52	10.44	22.76
28	18963	59.42	33956	3.13	94.68	14.76	15.41
29	9185	36.96	17016	4.02	85.32	20.52	19.35
30	17478	91.43	34873	5.23	76.32	14.40	16.83
31	6265	17.16	11237	2.74	153.00	24.84	30.53
32	8810	27.29	17306	3.10	107.64	11.16	17.98
33	17659	184.33	39250	10.44	90.72	6.48	22.09
34	10342	58.42	19074	5.65	82.44	9.72	18.29
35	8901	59.40	18452	6.67	79.92	3.24	26.05
36	8402	49.63	17500	5.91	120.96	6.48	26.20
37	32625	391.27	7888	11.99	84.60	5.40	17.26
38	31160	258.62	58947	8.30	85.32	6.12	18.83
39	46461	75.66	94697	1.63	101.52	8.64	19.70
40	13833	123.68	29626	8.94	107.64	11.88	16.87
41	6391	37.21	11688	5.82	85.32	7.92	14.63
42	11115	53.37	21955	4.80	131.76	10.08	22.17
43	6555	32.87	12243	5.01	116.64	18.72	22.62
44	11085	45.63	20193	4.12	138.24	13.68	26.44
45	9484	48.41	20122	5.10	156.96	16.56	22.26
46	3967	13.58	7612	3.49	137.52	14.76	19.13
47	15283	63.99	27404	4.19	135.72	7.92	18.28
48	20874	104.55	39648	5.01	155.52	18.36	28.23
49	19418	222.11	43799	11.44	48.60	8.28	12.39
50	3351	25.76	6235	7.67	42.84	14.04	11.64
51	6338	29.52	11524	4.66	142.20	16.92	8.62
52	9756	41.99	17309	4.30	145.80	11.16	20.10
53	11795	78.11	22225	6.62	120.52	14.76	19.41

Задание №2

МЕТОД КАНОНИЧЕСКИХ КОРРЕЛЯЦИЙ

1. В соответствии с предложенными данными:
 - а) построить блочную матрицу ковариаций переменных $X_2, X_3, \dots, X_n$ и $Y_1, Y_2, Y_3, \dots, Y_m$. Определить по ней матрицы $A_{11}, A_{12}, A_{22}, A_{21}$;
 - б) найти все собственные числа матриц $A_{11}^{-1}A_{12}A_{22}^{-1}A_{21}$ и $A_{22}^{-1}A_{21}A_{11}^{-1}A_{12}$;
 - в) определив параметры регрессионной модели θ, b , для каждого собственного числа найти соответствующие канонические переменные;
2. Найти корреляции между каноническими переменными.
3. Сравнить результаты с результатами, полученными с помощью пакета программ Gretl.

Таблица 1

ТАБЛИЦА ИСХОДНЫХ ДАННЫХ

№ банка	X1	X2	X3	X4	X5	X6	X7	X8	Y1	Y2	Y3
1	879	2.80	311	250	146	198	114	34.7	3090	23	19.0
2	1533	3.24	250	420	57	212	212	30.9	3388	23.9	20.0
3	2526	3.66	220	400	18.7	260	103	41.4	7525	56.9	46.5
4	1646	1.55	307	200	62	243	582	33.8	3795	27.4	23.4
5	1997	3.26	360	400	123	115	2978	3.60	0.4	22.9	25.3
6	6005	2.60	275	420	21.0	166	940	20.2	0.00	18.7	50.9
7	1824	2.96	343	120	19.68	160	858	10.4	0.00	44.3	36.5
8	2693	2.53	280	420	9.92	128	206	3.8	5360	40.2	33.1
9	3416	3	190	800	14.0	168	509	34.2	6514	47.5	40.2
10	1386	2.44	258	439	27	93	194	24.7	4995	37.8	30.7
11	2547	3.00	330	335	39.5	222	164	67.0	4889	35.9	30.1
12	4217	2.75	250	520	16.0	190	439	55.8	7361	54.7	45.4
13	3506	3.17	247	320	15	195	409	40.3	4878	36.1	30
14	2531	3.63	280	380	29.5	179	222	37.1	5344	39.4	32.8
15	2229	3.09	260	430	10	28	177	45.8	5938	43.4	36.2
16	2161	2.99	240	90	27	251	58	68.7	5268	39.0	32.4
17	3169	2.97	126	400	4.50	313	325	43.8	8299	62.3	51.1
18	3552	2.10	88.0	475	29	178	93	99.8	7041	52.0	43.1
19	1332	2.72	310	92	12.0	208	117	57.2	5155	40.6	31.8
20	3419	3.18	145	250	5.30	445	241	56.6	9011	66.8	55.5
21	1194	3.03	290	165	58.0	184	150	31.9	4697	35.8	28.7
22	2828	2.86	330	260	29.0	257	219	48.0	7008	51.5	42.7
23	2748	2.17	502	180	17.0	329	613	36.4	5467	36.3	31.5
24	1196	3.21	263	485	7.00	193	127	39.6	3535	25.7	21.7
25	2265	3.69	100	128	1.00	177	11	105.1	10644	76.7	65.7
26	1241	1.80	720	320	43.0	294	432	23.6	2768	19.2	16.7
27	3217	1.77	420	400	7.00	465	920	36.3	5225	37.0	32.1
28	1997	1.82	360	220	8.00	311	643	31.0	5167	38.1	31.5
29	1983	1.66	230	474	1	414	742	31.9	8393	45.8	39.1
30	842	3.47	289	250	21.0	165	175	23.6	3810	28.4	23.2

ВАРИАНТЫ РАСЧЕТА ЗАДАЧ

№	Y	X	№	Y	X
1	Y_1, Y_2	1-5	11	Y_2, Y_3	2-7
2	Y_1, Y_3	1-5	12	Y_1, Y_2	4-8
3	Y_2, Y_3	1-5	13	Y_1, Y_3	4-8
4	Y_1, Y_2, Y_3	1-6	14	Y_2, Y_3	1-8
5	Y_1, Y_2, Y_3	1-7	15	Y_1, Y_2	1-8
6	Y_1, Y_2	3-8	16	Y_1, Y_3	1-8
7	Y_1, Y_3	3-8	17	Y_2, Y_3	1-8
8	Y_2, Y_3	3-8	18	Y_1, Y_2	1-3 и 5-8
9	Y_1, Y_2	2-7	19	Y_1, Y_3	1-3 и 5-8
10	Y_1, Y_3	2-7	20	Y_2, Y_3	1-3 и 5-8

Задание №3 МЕТОД ГЛАВНЫХ КОМПОНЕНТ ФАКТОРНЫЙ АНАЛИЗ

1. В соответствии с предложенными данными:
 - перейти к стандартизированным данным, найти выборочную ковариационную матрицу новых векторов – признаков Z ;
 - найти собственные числа матрицы ковариаций и упорядочить их по убыванию;
 - найти все ее собственные вектора;
 - с помощью метода главных компонент выбрать главные компоненты, используя или $\text{sp } A$, или $\det(A)$;
 - найти факторные нагрузки $S = L\Lambda$;
2. С помощью факторного анализа построить многофакторную модель (число факторов детерминировать самостоятельно). Для этого:
 - найти собственные числа матрицы $P = (-p_{i,j}-)$, $p_{i,j} = \frac{1}{53} \sum_{k=1}^6 x_k^{(i)} x_k^{(j)}$, $i, j = \overline{1,53}$ и упорядочить их по убыванию;
 - выбрать максимальные собственные числа матриц ковариации и $P = (-p_{i,j}-)$, $i, j = \overline{1,53}$;
 - найти соответствующие им собственные вектора матрицы ковариаций $\bar{A} = (-\bar{a}_{i,j}-)$, $i, j = \overline{1,6}$ (факторные нагрузки) и матрицы $P = (-p_{i,j}-)$, $i, j = \overline{1,53}$ (факторы);
 - составить из факторных нагрузок матрицу S и проверить правильность их вычисления, найдя $S^T S$.
 - вычислить нормы для векторов факторных нагрузок и их факторов;
 - посчитать доли факторов в суммарной общности;
 - определить название факторов, вычислив коэффициенты информативности признаков.
3. Перейти к исходным признакам X .
4. Сравнить результаты для стандартизированных данных с результатами, полученными с помощью пакета программ STATISTICA 6.0.

Таблица 1

ВАРИАНТЫ РАСЧЕТА ЗАДАЧ

№ варианта	Номер векторов X
1	1,6,8, 11, 12, 17
2	1,6,8, 11, 13, 17
3	1,8,11,12,13,17
4	1,6,8, 13, 14, 17
5	1,8,11,13,14, 17
6	1,6,8, 12, 13, 17
7	1,7,11,12,13,17
8	1,7,9, 12, 13, 17
9	1,8,11,12,13,17
10	1,8,9,13, 14, 17
11	3,8,10,15, 16, 17
12	3,5,6,10,15, 17

№ варианта	Номер векторов X
13	3,5,6,7, 11, 12
14	3,8,9,10,11,17
15	3,8, 9,10, 12, 17
16	2,4,5,6,8,9
17	2,4,5,6,7,9
18	2,4,5,6,8,9
19	2,4,5,8,9,17
20	2,4,5,7,9,17

Исходные данные находятся в таблицах 2,3 приложения 1. Следует учесть, что вектора X_1 , X_2 , X_3 текущего задания упомянуты там как вектора Y_1 , Y_2 , Y_3 .