

Министерство образования и науки Российской Федерации  
НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

---

В. М. НЕДЕЛЬКО

ОСНОВЫ МАТЕМАТИЧЕСКОЙ  
СТАТИСТИКИ.  
МЕТОДЫ АНАЛИЗА ДАННЫХ

Утверждено Редакционно-издательским советом  
Университета в качестве учебно-методического пособия

НОВОСИБИРСК  
2008

УДК 519.23

Рецензенты:

д-р техн. наук, профессор *Г.С. Лбов*;

д-р техн. наук, профессор *А.А. Попов*

Работа подготовлена кафедрой ПС и БД.

Н-... В. М. Неделько. **Основы математической статистики. Методы анализа данных** : учебно-метод. пособие / В. М. Неделько. – Новосибирск : Изд-во НГТУ, 2008. – 44 с.

ISBN \_\_\_\_\_

Пособие для студентов ФПМИ НГТУ. Содержит учебный материал и задания для практических работ по второй части курса «Основы математической статистики», посвященной задачам анализа данных.

Владение методами статистического анализа данных позволяет выявлять в массивах эмпирических наблюдений статистические закономерности и строить вероятностные модели исследуемых явлений, а также оценивать достоверность эмпирических обобщений.

УДК 519.23

ISBN \_\_\_\_\_

© Неделько В.М., 2008

© Новосибирский государственный  
технический университет, 2008

## Содержание

Введение.....	4
§ 1. Проверка гипотез о параметрах нормальных распределений.....	5
§ 2. Регрессионный и корреляционный анализ.....	7
2.1. Одномерная линейная регрессия.....	7
2.2. Многомерная линейная регрессия.....	11
2.3. Нахождение доверительных интервалов.....	12
2.4. Корреляционный анализ.....	14
§ 3. Дискриминантный анализ.....	17
3.1. Разделяющая функция при известных нормальных распределениях.....	17
3.2. Оценивание параметров распределений.....	19
§ 4. Диагонализация ковариационной матрицы.....	22
Задание 1.....	31
Задание 2.....	33
Задание 3.....	40
Приложение. Критические точки распределения Фишера.....	42
Литература.....	44

## Введение.

Теория вероятностей вводит понятие вероятности и изучает ее свойства. Формально вероятность определяется как функция события, удовлетворяющая аксиомам Колмогорова. Аксиоматическое определение вероятности позволяет устанавливать различные соотношения, которым она удовлетворяет. Однако математическая теория ничего не говорит о ее содержательном смысле.

Вероятность можно интерпретировать как степень «ожидаемости» события. Это понятие может использоваться во всех ситуациях, когда есть неопределенность (в смысле недетерминированности).

Знание вероятностей позволяет наилучшим образом подготовиться к возможным сценариям развития событий, однако достоверно узнать вероятности на практике редко когда представляется возможным, поэтому их приходится оценивать.

Способы задания (оценивания) вероятностей:

- аналитический (классическое определение вероятности),
- эвристический (субъективные вероятности),
- статистический (частотное определение вероятности).

Заметим, что слово «определение» здесь используется не в смысле определения понятия вероятность, а в смысле определения (нахождения или оценивания) ее значений.

Аналитический подход использует информацию о свойствах объектов, определяющих исходы, в частности, о симметрии. Эвристический подход использует интуитивный выбор значений вероятностей, применяется при экспертном оценивании.

Статистический подход основан на том, что на практике можно считать, что событие с достаточно малой вероятностью в эксперименте не произойдет. При этом используются факты о связи вероятностей различных событий, например, закон больших чисел.

Данное пособие посвящено методам статистического анализа данных. Рекомендуется использовать совместно с [1], где рассматриваются базовые статистические методы.

## §1. Проверка гипотез о параметрах нормальных распределений

### *Исходные данные*

Пусть имеется  $k$  выборок  $(X_{j1}, \dots, X_{jn_j})$ ,  $j = \overline{1, k}$ , каждая из распределения  $\Phi_{\mu_j, \sigma_j}$  соответственно. Задача состоит в проверке гипотез совпадения средних и (или) дисперсий.

Введем обозначения:

$$N = \sum_{j=1}^k n_j, \quad \tilde{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ji},$$
$$\tilde{\bar{X}} = \frac{1}{N} \sum_{j=1}^k \sum_{i=1}^{n_j} X_{ji}, \quad \tilde{S}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} (X_{ji} - \tilde{X}_j)^2.$$

### *Критерий равенства дисперсий двух выборок*

Пусть  $k = 2$ . Проверяем согласие эмпирических данных с гипотезой  $\sigma_1 = \sigma_2$ .

В случае справедливости данной гипотезы случайная величина

$$f = \frac{n_1(n_2 - 1)\tilde{S}_1}{n_2(n_1 - 1)\tilde{S}_2}$$

подчиняется распределению Фишера  $F_{n_1-1, n_2-1}$ .

Пусть  $\hat{f}$  — значение величины  $f$ , полученное для конкретных выборок в задаче. Критической вероятностью будет

$$P_q(\hat{f}) = 2 \min(F_{n_1-1, n_2-1}(f < \hat{f}), F_{n_1-1, n_2-1}(f > \hat{f})).$$

Если полученная вероятность меньше уровня значимости, то гипотеза о равенстве дисперсий отвергается.

### *Критерий равенства средних двух выборок*

Пусть  $k = 2$ . Проверяем согласие эмпирических данных с гипотезой  $\mu_1 = \mu_2$ .

Рассматриваемый ниже критерий может применяться, только если  $\sigma_1 = \sigma_2 = \sigma$ .

В случае справедливости данной гипотезы случайная величина

$$t = \frac{\tilde{X}_1 - \tilde{X}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cdot \sqrt{\frac{n_1 \tilde{S}_1 + n_2 \tilde{S}_2}{n_1 + n_2 - 2}}}.$$

подчиняется распределению Стьюдента  $T_{n_1+n_2-2}$ .

Пусть  $\hat{t}$  — значение величины  $t$ , полученное для конкретных выборок в задаче.

Критической вероятностью будет:

$$P_q(\hat{t}) = T_{n_1+n_2-2}(|t| > |\hat{t}|).$$

Если полученная вероятность меньше уровня значимости, то гипотеза о равенстве средних отвергается.

### ***Однофакторный дисперсионный анализ***

Пусть  $k > 1$  и  $\sigma_1 = \sigma_2 = \dots = \sigma_k = \sigma$ . Задача факторного анализа заключается в проверке согласия эмпирических данных с гипотезой  $\mu_1 = \mu_2 = \dots = \mu_k = \mu$ .

В случае справедливости данной гипотезы случайная величина

$$g = \frac{(N-k) \sum_{j=1}^k n_j (\tilde{X}_j - \tilde{X})^2}{(k-1) \sum_{j=1}^k \sum_{i=1}^{n_j} (\tilde{X}_{ji} - \tilde{X}_j)^2}.$$

подчиняется распределению Фишера  $F_{k-1, N-k}$ .

Пусть  $\hat{g}$  — значение величины  $g$ , полученное для конкретных выборок в задаче.

Критической вероятностью будет

$$P_q(\hat{g}) = F_{k-1, N-k}(g > \hat{g}).$$

Если полученная вероятность меньше уровня значимости, то гипотеза о равенстве всех средних отвергается.

***Пример***

Для выборок из вариантов 0 и 1 получаем  $\hat{f} = 0,067$ .

Находим  $F_{10-1,8-1}(f > \hat{f}) = 0,99973$ , откуда

$$P_q = 2 \cdot \min(1 - 0,99973, 0,99973) = 0,00054.$$

Гипотеза о равенстве дисперсий отвергается.

Для выборок из вариантов 0 и 10 вычисляем  $\hat{t} = -1,52$ .

Получаем

$$T_{10+10-2}(|t| > 1,52) = 0,15.$$

Гипотеза о равенстве средних не отвергается.

Для выборок из вариантов 0, 10 и 25 вычисляем  $\hat{g} = 4,13$ .

Получаем

$$F_{3-1,30-3}(g > \hat{g}) = 0,027.$$

Гипотеза о равенстве всех трех средних отвергается, поскольку  $0,027 < 0,1$ .

## **§2. Регрессионный и корреляционный анализ**

Исторически термином регрессия называется функция условного математического ожидания некоторой случайной величины в зависимости от значений другой величины. Таким образом, под регрессионным анализом можно понимать статистическое оценивание регрессионной функции. Наиболее распространена линейная регрессия.

### **2.1. Одномерная линейная регрессия**

***Постановка задачи***

Предположим, что имеется случайная величина  $Y$ , которая зависит от некоторой величины  $X$ . При этом мы формально не

считаем  $X$  случайной величиной, хотя на практике методы регрессионного анализа можно применять и в случае, когда значения  $X$  фактически случайны.

Будем рассматривать следующую модель

$$Y = \kappa X + \varepsilon.$$

Здесь  $\kappa$  — некоторая константа, называемая коэффициентом регрессии, а  $\varepsilon$  — случайная величина, имеющая нормальное распределение с параметрами 0 и  $\sigma$ . Таким образом, величина  $Y$  формально представляется параметрическим семейством случайных величин  $Y(\omega, X)$ .

Переменную  $X$  принято называть *регрессором*, а величину  $Y$  — *откликом*.

Предположим, что каким-то образом выбраны  $N$  значений  $X_1, \dots, X_i, \dots, X_N$ , и при каждом  $X_i$  получены соответствующие случайные значения  $Y_i$ .

Задача состоит в том, чтобы по набору  $(X_i, Y_i), i = \overline{1, N}$ , оценить неизвестный параметр регрессии  $\kappa$ .

Если бы не было слагаемого  $\varepsilon$ , которое играет роль случайной ошибки, то имело бы место  $Y_i = \kappa X_i$ , откуда бы элементарным образом определялось  $\kappa$ .

На практике же  $Y_i = \kappa X_i + \varepsilon_i$ , и нахождению  $\kappa$  препятствует наличие неизвестных значений ошибки  $\varepsilon_i$ .

### ***Метод наименьших квадратов***

Предположим, что мы смогли каким-то образом вычислить  $\tilde{\kappa}$  — оценку коэффициента  $\kappa$ . Тогда величина  $\tilde{Y}_i = \tilde{\kappa} X_i$  будет оценкой ожидаемого значения  $Y_i$ , поэтому естественно выбирать  $\tilde{\kappa}$  так, чтобы погрешность такой оценки была минимальной. В качестве меры погрешности возьмем сумму квадратов отклонений

$$S(\kappa) = \sum_{i=1}^N (Y_i - \kappa X_i)^2.$$

В качестве  $\tilde{\kappa}$  возьмем значение, обеспечивающее минимальную ошибку, то есть  $S(\tilde{\kappa}) = \min_{\kappa} S(\kappa)$ .

Один из методов нахождения экстремума функции построен на нахождении нулей ее производной. Имеем

$$S'(\kappa) = \frac{\partial S(\kappa)}{\partial \kappa} = \sum_{i=1}^N 2X_i(Y_i - \kappa X_i) = 2\sum_{i=1}^N X_i Y_i - 2\kappa \sum_{i=1}^N X_i^2.$$

Из условия  $S'(\tilde{\kappa}) = 0$  находим

$$\tilde{\kappa} = \frac{\sum_{i=1}^N X_i Y_i}{\sum_{i=1}^N X_i^2}.$$

### ***Геометрическая иллюстрация***

Введем для эмпирических данных векторные обозначения, то есть  $X_1, \dots, X_N$  обозначим как вектор  $\vec{X}$ , а  $Y_1, \dots, Y_N$  как  $\vec{Y}$ .

В этих обозначениях  $S(\kappa)$  есть длина вектора  $\vec{S} = \vec{Y} - \kappa \vec{X}$ .

Из геометрических соображений легко понять, что длина вектора  $\vec{S}$  будет минимальной, когда он перпендикулярен вектору  $\vec{X}$ . Это означает, что их скалярное произведение равно нулю.

Имеем  $\vec{X}\vec{S} = \vec{X}\vec{Y} - \kappa \vec{X}^2 = 0$ , откуда  $\tilde{\kappa} = \frac{\vec{X}\vec{Y}}{\vec{X}^2}$ , что совпадает с выражением, полученным ранее.

Введя обозначение  $\vec{\varepsilon} = \varepsilon_1, \dots, \varepsilon_N$ , регрессионную модель можем записать в виде

$$\vec{Y} = \kappa \vec{X} + \vec{\varepsilon}.$$

### ***Общая линейная зависимость***

Общий вид линейной зависимости есть  $y = \kappa x + b$ , поэтому будем рассматривать модель

$$Y = \kappa X + b + \varepsilon,$$

где помимо параметра  $\kappa$  присутствует неизвестный параметр  $b$ .

Оценки обоих параметров можно найти, минимизируя средний квадрат отклонения

$$S(\kappa, b) = \sum_{i=1}^N (Y_i - \kappa X_i - b)^2.$$

Находим производные

$$S'_\kappa(\kappa, b) = \frac{\partial S(\kappa, b)}{\partial \kappa} = \sum_{i=1}^N 2X_i(Y_i - \kappa X_i - b) = 2\bar{X}\bar{Y} - 2\kappa\bar{X}^2 - 2bN\bar{X},$$

$$S'_b(\kappa, b) = \frac{\partial S(\kappa, b)}{\partial b} = \sum_{i=1}^N 2(Y_i - \kappa X_i - b) \cdot (-1) = 2N(b + \kappa\bar{X} - \bar{Y}),$$

где  $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$ ,  $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$ .

Из условия  $S'_b(\tilde{\kappa}, \tilde{b}) = 0$  находим  $\tilde{b} = \bar{Y} - \tilde{\kappa}\bar{X}$ . Подставляя последнее выражение в уравнение  $S'_\kappa(\tilde{\kappa}, \tilde{b}) = 0$ , получаем  $\bar{X}\bar{Y} - \tilde{\kappa}\bar{X}^2 - N\bar{X}\tilde{Y} + \tilde{\kappa}N\bar{X}^2 = 0$ , откуда

$$\tilde{\kappa} = \frac{\bar{X}\bar{Y} - N\bar{X}\tilde{Y}}{\bar{X}^2 - N\bar{X}^2}.$$

При этом

$$\tilde{b} = \bar{Y} - \tilde{\kappa}\bar{X} = \frac{\tilde{Y}\bar{X}^2 - \bar{X}\bar{X}\tilde{Y}}{\bar{X}^2 - N\bar{X}^2}.$$

Видим, что если  $\bar{X} = 0$ , то  $\tilde{b} = \bar{Y}$ , а  $\tilde{\kappa} = \frac{\bar{X}\bar{Y}}{\bar{X}^2}$ .

Выполнение условия  $\bar{X} = 0$  легко обеспечить, если вместо переменной  $X$  использовать переменную  $X' = X - \bar{X}$ .

### **Пример**

Построим регрессию для варианта 0.

Обозначим переменные  $X \equiv X_1$ ,  $Y \equiv X_2$ .

Модель строим в виде:

$$Y = \tilde{\kappa}(X - \bar{X}) + \tilde{b}.$$

Находим  $\tilde{X} = 0,35$ ,  $\tilde{b} = \tilde{Y} = -0,15$ ,  $\tilde{\kappa} = -0,55$ .

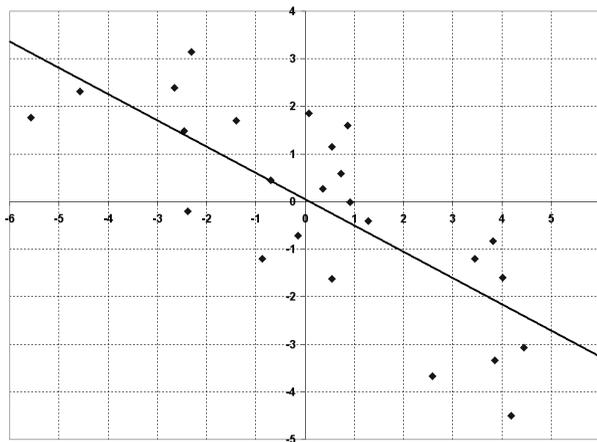


График оцененной зависимости приведен на рисунке.

## 2.2. Многомерная линейная регрессия

### Линейная регрессия

Предположим теперь, что имеется несколько регрессоров  $\bar{X}_1, \dots, \bar{X}_n$ , и рассматривается модель

$$\bar{Y} = \kappa_1 \bar{X}_1 + \dots + \kappa_n \bar{X}_n + \bar{\varepsilon}.$$

Будем искать линейную оценку вида  $\bar{Y}^* = \kappa_1 \bar{X}_1 + \dots + \kappa_n \bar{X}_n$ .

Требуется минимизировать длину вектора  $\bar{Y} - \bar{Y}^*$ . Из геометрических соображений очевидно, что минимум обеспечивается в случае, когда этот вектор перпендикулярен всем векторам  $\bar{X}_j$ , то есть  $(\bar{Y} - \bar{Y}^*) \bar{X}_j = 0$ .

Подставляя выражение для  $\bar{Y}^*$ , получаем на  $\kappa_j$  систему линейных уравнений

$$\sum_{j=1}^n a_{ij} \kappa_j = b_i, \quad i = \overline{1, n}, \quad \text{где } a_{ij} = \bar{X}_i \bar{X}_j, \quad b_i = \bar{X}_i \bar{Y}.$$

Решая указанную систему, можно получить искомые коэффициенты регрессии  $\kappa_j^* = \sum_{i=1}^n a_{ij}^{(-1)} \bar{X}_i \bar{Y}$ , где  $a_{ij}^{(-1)}$  — элементы матрицы  $A^{-1}$ , которая есть обратная к матрице  $A = (a_{ij})$ .

В случае, когда векторы  $\bar{X}_j$  ортогональны друг другу, полученная система упрощается, поскольку все недиагональные элементы матрицы  $A = (a_{ij})$  равны 0. При этом решением системы будет  $\kappa_j^* = \frac{\bar{X}_j \bar{Y}}{\bar{X}_j^2}$ , то есть все коэффициенты регрессии находятся независимо друг от друга.

### ***Ортогонализация выборки***

Заметим, что если исходная матрица  $A$  не является диагональной, всегда можно подобрать такую замену переменных  $X$ , что в новом пространстве ковариационная матрица будет диагональной. Более того, можно подобрать замену переменных с ортогональным оператором. Задача нахождения такого оператора не проще решения исходной системы, однако переход в каноническое пространство позволяет естественным образом построить доверительные интервалы для параметров модели.

### ***Одномерная полиномиальная регрессия***

Полиномиальная регрессия сводится к многомерной линейной регрессии путем выбора в качестве переменных  $X_1, \dots, X_n$  соответствующих степеней переменной  $X$ .

## **2.3. Нахождение доверительных интервалов**

### ***Оценки***

Несмещенной оценкой для  $\sigma^2$  будет  $\tilde{\sigma}^2 = \frac{1}{N-n} (\bar{Y} - \bar{Y}^*)^2$ .

Величина  $\frac{(\bar{Y} - \bar{Y}^*)^2}{\sigma^2}$  имеет распределение  $\chi_{N-n}^2$ , что позволяет построить доверительный интервал.

Если векторы  $\bar{X}_j$  ортогональны, то величина  $\frac{(\kappa_j^* - \kappa_j) |\bar{X}_j|}{\sigma}$  имеет нормальное распределение с параметрами  $(0, 1)$ , а величина  $t = \frac{(\kappa_j^* - \kappa_j) |\bar{X}_j|}{\tilde{\sigma}}$  имеет распределение  $T_{N-n}$ , что позволяет находить доверительные интервалы при известном и при неизвестном  $\sigma$ .

Приведенные оценки получены для многомерной регрессии без параметра  $b$ . Если данный параметр присутствует, модель можно свести к рассмотренной введением дополнительной переменной  $X_{n+1} \equiv 1$ .

Заметим, что если предварительная обработка данных включает центрирование выборки по переменной  $Y$ , это автоматически означает введение параметра  $b$ .

### **Пример**

Оценим доверительные интервалы для параметров по данным варианта 0.

Поскольку модель включает параметр  $b$ , вводим фиктивную переменную  $X_2 \equiv 1$ . Учитывая, что  $N = 25$ , а  $n = 2$ , получаем  $\tilde{\sigma}^2 = 1,75$ .

По таблице критических точек для распределения  $\chi_{25-2}^2$  находим критические значения величины  $\hat{w} = \frac{(N-k)\tilde{\sigma}^2}{\sigma^2}$  для вероятностей 0,05 и 0,95, разность которых есть доверительная вероятность  $\eta$ , которая задана равной 0,9.

Имеем  $\hat{w}_1 = 35,2$ ,  $\hat{w}_2 = 13,1$ , откуда получаем  $[1,07, 1,75]$  — доверительный для  $\sigma$ .

Из распределения  $T_{25-2}$  при доверительной вероятности 0,9 находим критическое значение  $\hat{t} = 1,71$ .

Вычислив  $|\bar{X}| = 13,8$ , получаем  $t = \frac{(-0,55 - \kappa) \cdot 13,8}{\sqrt{1,75}}$  и находим  $[-0,72, -0,39]$  – доверительный интервал для  $\kappa$ .

Из соотношения  $t_b = \frac{(-0,15 - b) \cdot \sqrt{5}}{\sqrt{1,75}}$  находим  $[-0,6, 0,31]$  – доверительный интервал для  $b$ .

## 2.4. Корреляционный анализ

### *Проецирование на линейное многообразие*

В данном разделе будем рассматривать случай зависимости от одной переменной. Ранее мы находили линейную аппроксимацию зависимости  $Y$  от  $X$ . Однако эту же зависимость можно рассматривать как зависимость  $X$  от  $Y$ , то есть  $X = \zeta Y + d + \varepsilon$ .

По выборке угловой коэффициент оценивается как  $\tilde{\zeta} = \frac{\bar{X}\bar{Y}}{\bar{Y}^2}$ .

Легко видеть, что, за исключением особых случаев,  $\tilde{\zeta} \neq \frac{1}{\tilde{\kappa}}$ .

Поэтому регрессионная модель  $X$  от  $Y$ , вообще говоря, не есть симметричное отражение модели  $Y$  от  $X$ .

Во многих случаях естественной выглядит задача построения линейной модели, в которой переменные равноправны:

$$aX + bY = \varepsilon, \quad a^2 + b^2 = 1.$$

Для простоты будем полагать, что выборка центрирована (среднее по каждой переменной равно нулю). Найдем прямую, сумма квадратов расстояний от выборочных точек до которой минимальна, т.е. Такую прямую будем называть главной компонентой.

$$\sum_{i=1}^N (aX_i + bY_i)^2 \rightarrow \min .$$

Делая замену  $a = \cos \alpha$ ,  $b = \sin \alpha$  и раскрывая квадрат, имеем

$$\bar{X}^2 \cos^2 \alpha + \bar{X}\bar{Y} \cos \alpha \sin \alpha + \bar{Y}^2 \sin^2 \alpha \rightarrow \min .$$

Приравняв производную по  $\alpha$  нулю, получаем

$$\operatorname{tg} 2\alpha = \frac{2\bar{X}\bar{Y}}{\bar{X}^2 - \bar{Y}^2} .$$

Углы  $\alpha$ , удовлетворяющие данному уравнению, задают направления главных осей эллипса рассеяния. В § 4 рассматривается более универсальный метод подобных построений, основанный на диагонализации ковариационной матрицы.

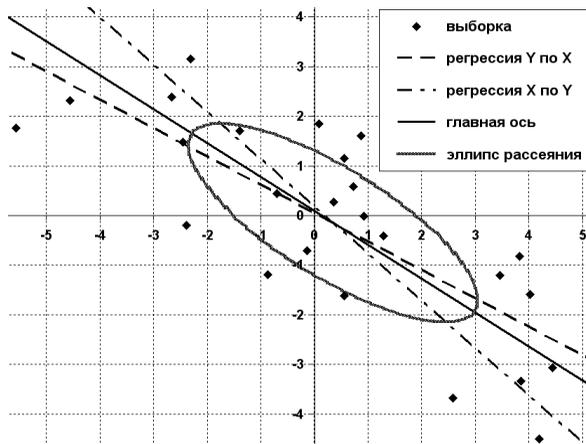
### **Пример**

Найдем главную компоненту для данных из варианта 0.

Проведя центрирование выборки, вычисляем

$$\operatorname{tg} 2\alpha = \frac{2\bar{X}\bar{Y}}{\bar{X}^2 - \bar{Y}^2} = \frac{2 \cdot (-4,19)}{7,3 - 4,01} = -2,55 .$$

Отсюда  $\alpha = -0,6$ , либо  $\alpha = 0,97$ . Полученные углы задают направления главных осей эллипса рассеяния (определение и метод построения см. в § 4).



Из двух полученных углов непосредственной проверкой выбираем тот, который доставляет минимум суммы квадратов от точек выборки до прямой. Находим искомую прямую

$$0,83X + 0,56Y = 0.$$

Чтобы от центрированных переменных вернуться к исходным, достаточно сделать соответствующий сдвиг по осям.

На рисунке изображены линии регрессии  $X$  по  $Y$  и  $Y$  по  $X$ , а также эллипс рассеяния и главная компонента.

### ***Проверка статистической значимости корреляции***

Коэффициент корреляции независимых случайных величин равен нулю, однако выборочные корреляции могут даже в этом случае значительно отличаться от нуля.

Чтобы оценить, действительно ли полученное ненулевое значение корреляции свидетельствует о зависимости переменных, можно использовать следующий статистический критерий.

Для нормальных случайных величин  $X$  и  $Y$  статистика

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}, \text{ где } r = \frac{\bar{X}\bar{Y}}{\sqrt{\bar{X}^2\bar{Y}^2}} \text{ — коэффициент корреляции,}$$

подчиняется распределению Стьюдента с  $N-2$  степенями свободы. Чем больше по модулю значение  $t$ , тем менее правдоподобна гипотеза независимости величин  $X$  и  $Y$ .

Критическая область задается соотношением

$$T(|t| > |t_\alpha|) = \alpha.$$

Для варианта 0 имеем  $r = -0,77$ ,  $t = -5,88$  и  $T(|t| > 5,88) = 5 \cdot 10^{-6}$ . Таким образом, гипотеза независимости переменных в данном примере отвергается при любых разумных уровнях значимости.

### §3. Дискриминантный анализ

Задача дискриминантного анализа заключается в том, что по выборке необходимо построить в пространстве переменных разделяющие (дискриминантные) поверхности, разбивающие пространство на области, соответствующие заданным классам. Число классов может быть два или более. Также для этой задачи используется термин классификация с учителем.

Мы ограничимся случаем двух классов при нормальных распределениях.

#### 3.1. Разделяющая функция при известных нормальных распределениях

##### *Постановка задачи*

Пусть имеется система случайных величин  $X_1, \dots, X_n, Y$ , где  $X_j$  — непрерывные случайные величины, а  $Y$  — дискретная величина, принимающая только значения 1 или 2, которые будем называть номерами классов. Совокупность  $X_1, \dots, X_n$  будем для краткости обозначать просто  $X$ , а вектор значений этих величин обозначим  $x$ .

Предположим, что условное распределение  $P(X/Y=1)$ , плотность вероятности для которого будем обозначать через  $\varphi_1(x)$ , является нормальным с параметрами  $\mu_1$  и  $\lambda_1$ , где  $\mu_1$  — вектор средних (математических ожиданий), а  $\lambda_1$  — ковариационная матрица. Аналогично, распределение  $P(X/Y=2)$  предположим нормальным с параметрами  $\mu_2$  и  $\lambda_2$ , соответствующую условную плотность вероятности обозначим  $\varphi_2(x)$ .

Задача заключается в том, чтобы найти такую функцию  $f(x)$ , чтобы вероятность события  $f(X) \neq Y$  была минимальной. Иными словами, требуется каждой точке пространства  $X$  приписать некоторый номер класса, то есть классифицировать точки пространства. При этом требуется, чтобы вероятность

$P(f(X) \neq Y)$ , называемая вероятностью ошибочной классификации, была наименьшей из возможных.

### **Оптимальное решение**

Чтобы минимизировать вероятность ошибочной классификации, нужно точке  $x$  приписывать тот класс, который в этой точке наиболее вероятен, то есть  $f(x) = y^*$ , где  $y^*$  выбирается из условия  $P(Y = y^*/x) = \max(P(Y = 1/x), P(Y = 2/x))$ .

Искомую функцию можно также записать в виде

$$f(x) = \begin{cases} 1, & P(Y = 1/x) \geq P(Y = 2/x) \\ 2, & P(Y = 1/x) < P(Y = 2/x) \end{cases}.$$

Более удобно представление

$$f(x) = \begin{cases} 1, & l(x) \geq 0 \\ 2, & l(x) < 0 \end{cases},$$

где введена функция  $l(x) = \ln P(Y = 1/x) - \ln P(Y = 2/x)$ , которая называется разделяющей. Решение уравнения  $l(x) = 0$  задает границу между классами и называется разделяющей поверхностью.

Воспользовавшись формулой Байеса, которую в рассматриваемом случае можно записать в виде  $P(Y = k/x) = \varphi_k(x) \frac{P(Y = k)}{\varphi(x)}$ , выражение для разделяющей функции можно привести к виду

$$l(x) = \ln \varphi_1(x) - \ln \varphi_2(x) + \ln P_1 - \ln P_2,$$

где введено обозначение  $P_k = P(Y = k)$ .

Теперь в последнюю формулу можно подставить аналитические выражения для нормальных плотностей

$$\varphi_k = \frac{1}{(2\pi)^{n/2} |\lambda_k|^{n/2}} e^{-\frac{1}{2} Q_k(x)}, \quad \text{где } Q_k(x) = (x - \mu_k)' (\lambda_k)^{-1} (x - \mu_k).$$

Штрих в записи квадратичной формы обозначает транспонирование вектора. В дальнейшем будем его опускать, считая левый вектор в произведении всегда транспонированным.

После преобразований имеем

$$2l(x) = Q_2(x) - Q_1(x) + \ln|\lambda_2| - \ln|\lambda_1| + 2 \ln P_1 - 2 \ln P_2 .$$

Далее преобразуем

$$\begin{aligned} Q_2(x) - Q_1(x) &= (x - \mu_2)(\lambda_2)^{-1}(x - \mu_2) - (x - \mu_1)(\lambda_1)^{-1}(x - \mu_1) = \\ &= x((\lambda_2)^{-1} - (\lambda_1)^{-1})x - (\mu_2(\lambda_2)^{-1} - \mu_1(\lambda_1)^{-1})x - \\ & x((\lambda_2)^{-1}\mu_2 - (\lambda_1)^{-1}\mu_1) + \mu_2(\lambda_2)^{-1}\mu_2 - \mu_1(\lambda_1)^{-1}\mu_1 . \end{aligned}$$

Наконец получаем

$$2l(x) = xAx + bx + c , \text{ где}$$

$$A = (\lambda_2)^{-1} - (\lambda_1)^{-1} , \quad b = 2\mu_1(\lambda_1)^{-1} - 2\mu_2(\lambda_2)^{-1} ,$$

$$c = \mu_2(\lambda_2)^{-1}\mu_2 - \mu_1(\lambda_1)^{-1}\mu_1 + \ln|\lambda_2| - \ln|\lambda_1| + 2 \ln P_1 - 2 \ln P_2 .$$

При получении выражения для  $b$  использована коммутативность скалярного произведения векторов и коммутативность произведения вектора на симметричную матрицу (напомним, что ковариационная матрица симметрична).

Уравнение  $l(x) = 0$  в общем случае задает поверхность второго порядка, которая при  $\lambda_1 = \lambda_2$ , то есть при равенстве ковариационных матриц, вырождается в гиперплоскость.

Заметим, что хотя ковариационные матрицы являются положительно определенными, ввиду чего линии (поверхности) уровня нормального распределения представляют собой эллипсоиды, разделяющая поверхность принадлежит широкому классу кривых второго порядка.

### 3.2. Оценивание параметров распределений

#### *Общий случай*

Пусть имеется выборка, представленная прямоугольной матрицей  $(x_{ij}, y_i)$ ,  $j = \overline{1, n}$ ,  $i = \overline{1, N}$ . Параметр  $N$  представляет

собой число независимых реализаций системы случайных величин и называется объемом выборки.

Сформируем множество индексов (номеров) объектов (строк матрицы), принадлежащих первому классу:  $I_1 = \{i \mid y_i = 1\}$ . Аналогично,  $I_2 = \{i \mid y_i = 2\}$ . Тогда  $N_1 = |I_1|$  есть число объектов первого, а  $N_2 = |I_2|$  — второго классов в выборке,  $N_1 + N_2 = N$ .

Проще всего оценить априорные вероятности классов, а именно:  $\tilde{P}_k = \frac{N_k}{N}$ , где  $\tilde{P}_k$  — оценка для  $P_k$ , а  $k$  — номер класса (1 или 2).

Обозначим через  $\tilde{\mu}^k = (\tilde{\mu}_1^k, \dots, \tilde{\mu}_n^k)$  оценку вектора  $\mu_k$ . В этом обозначении  $k$  рядом с  $\tilde{\mu}$  означает не степень, а верхний индекс. За счет переноса номера класса в верхний индекс мы освободили нижний индекс для обозначения компонент вектора.

Компоненты вектора средних для каждого класса оцениваются по формуле

$$\tilde{\mu}_j^k = \frac{1}{N_k} \sum_{i \in I_k} x_{ij}.$$

Оценку ковариационной матрицы  $\lambda_k$  обозначим через  $\tilde{\lambda}^k = (\tilde{\lambda}_{jl}^k)$ ,  $j = \overline{1, n}$ ,  $l = \overline{1, n}$ .

Компоненты этой матрицы вычисляются как

$$\tilde{\lambda}_{jl}^k = \frac{1}{N_k} \sum_{i \in I_k} (x_{ij} - \tilde{\mu}_j^k)(x_{il} - \tilde{\mu}_l^k).$$

Для построения выборочной разделяющей поверхности достаточно теперь в уравнение  $l(x) = 0$  подставить вместо параметров распределений найденные оценки.

### ***Случай равных матриц ковариации***

Если известно, что распределения обоих классов имеют одинаковые матрицы ковариации, то нужно всю выборку

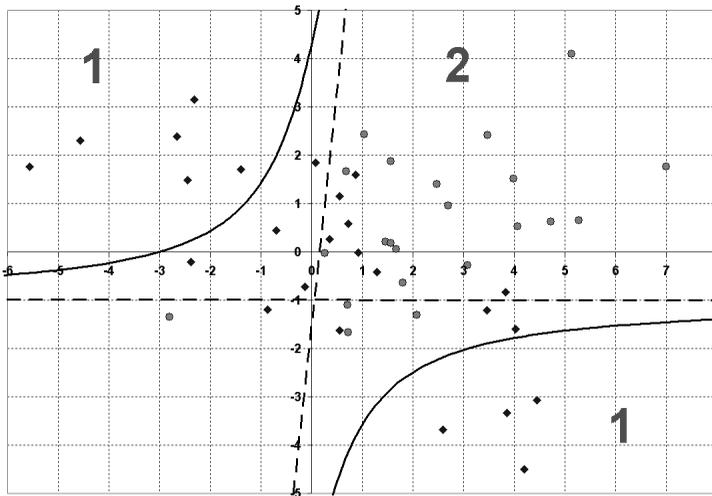
использовать для вычисления оценки  $\tilde{\lambda} = (\tilde{\lambda}_{jl})$  матрицы  $\lambda = \lambda_1 = \lambda_2$ . Оценка находится следующим образом

$$\tilde{\lambda}_{jl} = \frac{1}{N} \left( \sum_{i \in I_1} (x_{ij} - \tilde{\mu}_j^1)(x_{il} - \tilde{\mu}_l^1) + \sum_{i \in I_2} (x_{ij} - \tilde{\mu}_j^2)(x_{il} - \tilde{\mu}_l^2) \right).$$

Остальные параметры оцениваются так же, как в общем случае.

### Пример

На рисунке приведена разделяющая кривая между первым и вторым классами для задания из варианта 0. Темными маркерами отмечены точки первого, светлыми — второго класса. Сплошная линия — разделяющая кривая (гипербола), пунктирные — асимптоты.



Оценки векторов средних:

$$\tilde{\mu}^1 = (0,35, -0,15), \quad \tilde{\mu}^2 = (2,39, 0,64).$$

Оценки ковариационных матриц:

$$\tilde{\lambda}^1 = \begin{pmatrix} 7,6 & -4,19 \\ -4,19 & 4,18 \end{pmatrix}, \quad \tilde{\lambda}^2 = \begin{pmatrix} 4,5 & 1,51 \\ 1,51 & 2,06 \end{pmatrix}.$$

Параметры разделяющей кривой:

$$A = \begin{pmatrix} -0,00068 & -0,51 \\ -0,51 & 0,106 \end{pmatrix}, b = (-1,017, 0,26), c = -3,07.$$

Уравнение кривой:

$$-0,00068x_1^2 + 0,106x_2^2 - 1,02x_1x_2 - 1,017x_1 + 0,26x_2 - 3,07 = 0.$$

Первая асимптота полученной гиперболы проходит через точки  $(-0,36, -5)$  и  $(0,67, 5)$ , вторая – через  $(-6, -0,99)$  и  $(8, -1)$ .

Поскольку значение коэффициента при  $x_1^2$  близко к 0, его следует заменить нулем, упростив этим уравнение.

#### §4. Диагонализация ковариационной матрицы

Метод главных компонент основан на диагонализации матрицы ковариаций, т.е. выборе такой замены переменных, при которой ковариационная матрица становится диагональной.

Как известно, искомая замена переменных является ортогональным преобразованием, т.е. заключается в осуществлении подходящего поворота системы координат и сохраняет расстояния между точками.

##### *Постановка задачи*

Пусть имеется набор переменных  $X_1, \dots, X_n$ . Произвольное значение переменной  $X_j$  будем обозначать  $x_j$ . Совокупность значений переменных обозначим вектором  $x = (x_1, \dots, x_n)$ .

Предположим, что задана выборка значений  $v = (x_j^i)$ ,  $i = \overline{1, N}$ ,  $j = \overline{1, n}$ , где  $x_j^i$  —  $i$ -е выборочное значение переменной  $X_j$ . Чтобы отличать выборочные значения от других значений тех же переменных, используем верхний индекс.

По выборке  $v$  может быть вычислена матрица выборочных ковариаций

$$A = (a_{ij}), \quad a_{ij} = \frac{1}{N} \sum_{s=1}^N x_i^s x_j^s,$$

которая и является предметом анализа. Здесь  $i$  — номер строки матрицы,  $j$  — номер столбца.

Рассмотрим другой набор переменных  $Y_1, \dots, Y_n$ , такой что значения переменных  $X_1, \dots, X_n$  выражаются через них по формуле  $x_j = \sum_{k=1}^n t_{jk} y_k$ . Последнее выражение можно записать в матричной форме  $x = Ty$ , где  $y = (y_1, \dots, y_n)$ , а  $T = (t_{ij})$  — невырожденная квадратная матрица замены переменных.

Используя обратную матрицу  $T^{-1}$  можно выразить переменные  $Y_1, \dots, Y_n$  через переменных  $X_1, \dots, X_n$ .

Выясним, как при замене переменных меняется ковариационная матрица. Имеем

$$a_{ij} = \frac{1}{N} \sum_{s=1}^N x_i^s x_j^s = \frac{1}{N} \sum_{s=1}^N \left( \sum_{k=1}^n t_{ik} y_k^s \right) \left( \sum_{m=1}^n t_{jm} y_m^s \right).$$

Раскрывая скобки и меняя порядок суммирования, получаем

$$a_{ij} = \frac{1}{N} \sum_{s=1}^N \sum_{k=1}^n \sum_{m=1}^n t_{ik} y_k^s t_{jm} y_m^s = \sum_{k=1}^n \sum_{m=1}^n t_{ik} t_{jm} \frac{1}{N} \sum_{s=1}^N y_k^s y_m^s.$$

Введем обозначение  $b_{ij} = \frac{1}{N} \sum_{s=1}^N y_i^s y_j^s$ . Величины  $b_{ij}$  являются выборочными ковариациями в пространстве  $Y_1, \dots, Y_n$  и образуют матрицу  $B = (b_{ij})$ .

Используя данное обозначение, имеем

$$a_{ij} = \sum_{k=1}^n \sum_{m=1}^n t_{ik} t_{jm} b_{km}.$$

Обозначим  $r_{kj} = \sum_{m=1}^n b_{km} t_{jm}$ , тогда  $a_{ij} = \sum_{k=1}^n t_{ik} r_{kj}$ . Видим, что второе соотношение представляет собой произведение матриц, то

есть может быть записано в виде  $A = TR$ , где  $R = (r_{ij})$ . Первое соотношение отличается от произведения матриц  $B$  и  $T$  порядком индексов у  $t_{jm}$ , поэтому может быть записано в матричном виде с использованием транспонирования  $R = BT'$ , где  $T'$  — матрица, полученная из  $T$  транспонированием.

В результате получаем

$$A = TBT'.$$

Наша цель — найти такую систему переменных  $Y_1, \dots, Y_n$  или, что то же самое, такую матрицу перехода  $T$ , чтобы матрица  $B$  была диагональной (все элементы, кроме стоящих на главной диагонали, были нулевыми).

### *Диагонализация матрицы*

Из теории матриц известно, что для любой симметричной матрицы  $A$ , то есть такой матрицы, для которой  $A = A'$ , а ковариационная матрица как раз симметрична, существует матрица  $T$ , такая что  $A = TBT^{-1}$ , где  $B$  — диагональная матрица.

Выясним теперь, как находить матрицу  $T$ .

Заметим, что

$$A = TBT^{-1} \Leftrightarrow AT = TB.$$

Последнее равенство можно представить в виде

$$At_i = \lambda_i t^i,$$

где  $\lambda_i$  — элементы, стоящие на диагонали матрицы  $B$ , а через  $t^i$  обозначены столбцы матрицы  $T$ . Чтобы убедиться в справедливости данного представления, достаточно заметить, что умножение справа на диагональную матрицу приводит к умножению каждого столбца матрицы на соответствующий диагональный элемент.

Последнее уравнение является уравнением для собственных векторов, то есть векторы, ему удовлетворяющие, называются собственными векторами матрицы  $A$ , а  $\lambda_i$  — собственными

значениями. Введением единичной матрицы  $I$  уравнение можно переписать в виде

$$At^i = \lambda_i It^i \Leftrightarrow (A - \lambda_i I)t^i = 0.$$

Последнее соотношение является системой однородных линейных уравнений для собственных векторов  $t^i$ . Система однородных линейных уравнений имеет ненулевые решения, только когда ее определитель равен нулю, что дает уравнение для нахождения собственных значений  $\lambda_i$

$$|A - \lambda_i I| = 0.$$

Данное уравнение называется *характеристическим*. Известно, что все корни характеристического уравнения вещественны. Предположим, что все его  $n$  корней ненулевые и различны. Для каждого корня  $\lambda_i$  можем найти собственный вектор  $t^i$ , причем, поскольку решение однородной системы определено с точностью до константы, этот вектор возьмем единичной длины, т.е.

$$\|t^i\|^2 = t^i \cdot t^i = 1.$$

Легко показать, что в этом случае матрица  $T$  будет ортогональной, то есть что  $T^{-1} = T'$ .

Воспользуемся очевидным свойством операции транспонирования: для любых матриц  $C$  и  $D$  справедливо  $(CD)' = C'D'$ . Транспонируем обе части равенства  $AT = TB$ .

Получим  $(AT)' = (TB)' \Leftrightarrow T'A' = B'T' \Leftrightarrow T'A = BT'$  в силу симметричности матриц  $A$  и  $B$ . Умножим последнее равенство на  $T$  и заменим  $AT$  на равную ей  $TB$ . Получим  $T'A = BT' \Leftrightarrow T'AT = BT'T \Leftrightarrow T'TB = BT'T$ . Обозначим

$T'T = S = (s_{ij})$ , тогда последнее равенство запишется в виде  $s_{ij}\lambda_i = s_{ij}\lambda_j$ . Поскольку собственные числа мы предположили различными, имеем  $s_{ij} = 0$ , при  $i \neq j$ . Диагональные элементы

матрицы  $S$  равны 1 в силу нормированности векторов  $t^i$ , таким образом  $T'T = I$ , или  $T' = T^{-1}$ .

Если среди корней характеристического уравнения есть кратные, то также можно найти ортогональную матрицу  $T$ , такую что  $A = TBT^{-1}$ . Для этого в качестве решений однородных систем уравнений нужно выбирать ортогональные вектора.

Итак, для любой симметричной матрицы  $A$  можно найти матрицу  $T$ , такую что  $A = TBT^{-1}$ , где  $B$  — диагональная матрица, при этом  $T^{-1} = T'$ , а значит  $A = TBT'$  и  $T$  является матрицей замены переменных, приводящей ковариационную матрицу  $A$  к диагональному виду.

Это означает, что требуемая нам замена переменных всегда существует и, более того, является ортогональной, то есть представляет собой вращение системы координат, сохраняющее расстояния.

### **Пример**

Рассмотрим решение варианта 0 из задания 3.

Согласно условию из исходной таблицы следует выбрать столбцы для переменных  $Z_2$ ,  $Z_3$  и  $Z_8$ .

Для вычисления ковариаций нужно центрировать выборку, т.е. вычесть по каждой переменной среднее значение. Обозначим

$$X_1 = Z_2 - \bar{z}_2, \quad X_2 = Z_3 - \bar{z}_3, \quad X_3 = Z_8 - \bar{z}_8, \quad \text{где } \bar{z}_j = \frac{1}{N} \sum_{i=1}^N z_j^i, \quad z_j^i \text{ —}$$

значение из  $i$ -й строки и  $j$ -го столбца таблицы данных.

Получаем новую таблицу.

$i$	$X_1$	$X_2$	$X_3$	$i$	$X_1$	$X_2$	$X_3$	$i$	$X_1$	$X_2$	$X_3$
1	1,93	-0,85	0,45	10	-1,59	2,41	0,44	18	-1,11	2,67	0,58
2	-1,05	-2,48	0,37	11	-3,42	1,37	-1,97	19	0,59	-2,43	-4,69
3	0,70	2,33	1,08	12	-1,22	-1,46	1,13	20	2,32	0,23	2,75
4	-0,26	-2,84	-1,68	13	2,06	1,64	1,31	21	-3,58	1,17	-3,29

5	1,44	0,41	0,71	14	0,53	-0,83	0,45	22	1,24	-1,38	-1,98
6	1,40	-0,10	0,61	15	1,15	2,92	0,28	23	1,82	-3,92	3,54
7	-0,64	3,38	0,75	16	-1,15	1,46	-0,01	24	2,65	0,51	-0,74
8	2,14	-2,94	0,49	17	-0,23	-2,82	-1,79	25	-3,67	-0,12	1,55
9	-2,07	1,68	-0,35								

Вычисляем ковариационную матрицу

$$A = \begin{pmatrix} 3,47 & -0,99 & 0,97 \\ -0,99 & 4,28 & 0,36 \\ 0,97 & 0,36 & 3,07 \end{pmatrix}.$$

Решаем характеристическое уравнение

$$|A - \lambda I| = \begin{vmatrix} 3,47 - \lambda & -0,99 & 0,97 \\ -0,99 & 4,28 - \lambda & 0,36 \\ 0,97 & 0,36 & 3,07 - \lambda \end{vmatrix} = 0.$$

После приведения подобных (раскрывая определитель по формуле  $|A| = a_{11}a_{22}a_{33} + a_{13}a_{21}a_{32} + a_{12}a_{23}a_{31} -$

$a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}$ ) имеем

$$f(\lambda) = \lambda^3 - 10,83 \cdot \lambda^2 + 36,64 \cdot \lambda - 37,48 = 0.$$

Для решения уравнения можно воспользоваться формулами, выражающими корни в радикалах, однако данные формулы весьма громоздки и неудобны. Поэтому используем численный метод.

Выбираем наугад точки  $\lambda$ , пока не найдем любую пару, в которых функция имеет разные знаки. Имеем  $f(0) = -37,48$ ,  $f(1) = -10,67$ ,  $f(2) = 0,49$  — пара точек нашлась.

Далее ищем корень методом деления отрезка пополам. Повторяя процедуру последовательно, получаем

$$f(1,5) = -3,5, \quad f(1,9) = -0,09, \quad f(1,95) = 0,2, \quad f(1,92) = 0,03, \quad f(1,91) = -0,03.$$

Итак, первый корень есть  $\lambda_1 = 1,915$ .

Теперь многочлен можно разложить на множители

$$f(\lambda) = \lambda^3 - 10,83 \cdot \lambda^2 + 36,64 \cdot \lambda - 37,48 = (a\lambda^2 + b\lambda + c)(\lambda - 1,915) = a\lambda^3 + b\lambda^2 + c\lambda - 1,915a\lambda^2 - 1,915b\lambda - 1,915c.$$

Приравнявая коэффициенты при равных степенях, последовательно вычисляем:  $a = 1$ ,  $b = -10,83 + 1,915 = -8,915$ ,  $c = \frac{37,48}{1,915} = 19,58$ .

Решая уравнение  $\lambda^2 - 8,915 \cdot \lambda + 19,58 = 0$ , получаем оставшиеся корни:  $\lambda_2 = 5,0$ ,  $\lambda_3 = 3,92$ .

Для удобства и единообразия переупорядочим полученные корни по убыванию:  $\lambda_1 = 4,98$ ,  $\lambda_2 = 3,93$ ,  $\lambda_3 = 1,91$ . Здесь приведены результаты вычислений, производимых с большей точностью и последующим округлением до двух знаков.

Решим теперь систему уравнений  $At = \lambda t$ ,  $t = (t_1, t_2, t_3)$ , для найденных  $\lambda$ .

Подставляя  $\lambda_1 = 4,98$ , получаем систему

$$\begin{cases} -1,51t_1 - 0,99t_2 + 0,97t_3 = 0 \\ -0,99t_1 - 0,7t_2 + 0,36t_3 = 0 \\ 0,97t_1 + 0,36t_2 - 1,91t_3 = 0 \end{cases}$$

Поскольку уравнение линейно зависимо, любое из них можно исключить, например, третье. Так как решение определено с точностью до произвольного множителя, положим  $t_3 = 1$ . Если в процессе решения возникнет деление на близкие к нулю величины, то единице положим равной другую компоненту.

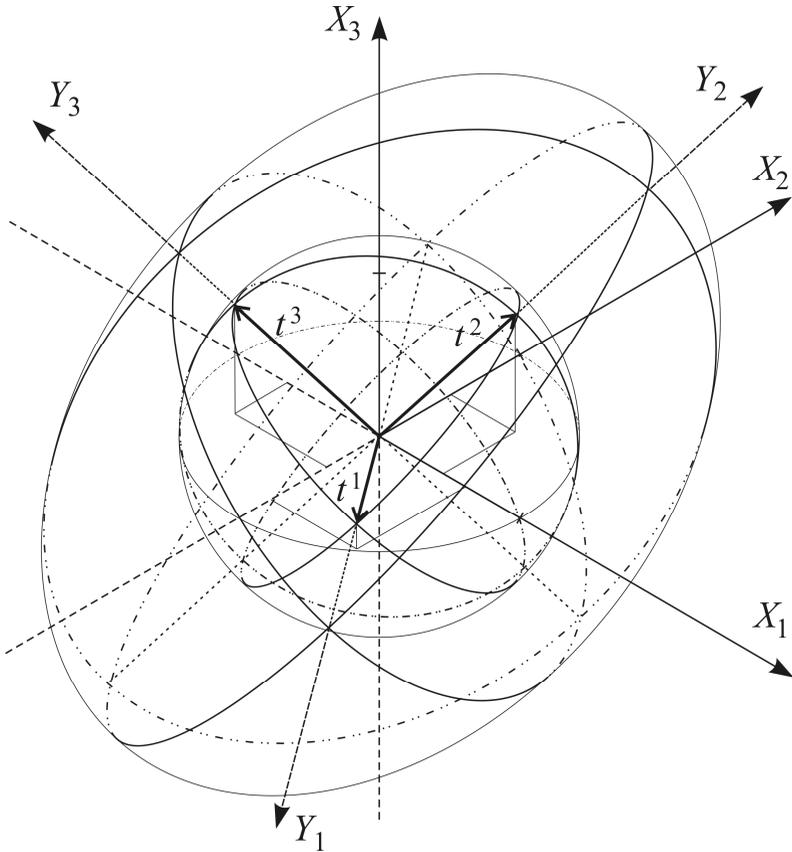
Решая систему двух уравнений, находим  $t_2 = -4,77$ ,  $t_1 = 3,75$ .

Вычислим норму вектора  $|t| = \sqrt{3,75^2 + 4,77^2 + 1} = 6,15$ . Разделив каждую компоненту на норму, получаем нормированный первый собственный вектор  $t^1 = (0,61; -0,78; 0,16)$ .

Аналогичным образом находим остальные собственные векторы  $t^2 = (0,45; 0,51; 0,73)$ ,  $t^3 = (-0,65; -0,37; 0,66)$ .

Используя полученные векторы в качестве столбцов, образуем матрицу

$$T = \begin{pmatrix} 0,61 & 0,45 & -0,65 \\ -0,78 & 0,51 & -0,37 \\ 0,16 & 0,73 & 0,66 \end{pmatrix}.$$



На рисунке изображен эллипсоид рассеяния и единичная сфера, на которой отмечены собственные вектора. Также показаны сечения сферы и эллипсоида координатными плоскостями системы переменных  $Y_1, Y_2, Y_3$ .

Эллипсоидом рассеяния мы называем поверхность, задаваемую уравнением  $x'A^{-1}x=1$ .

При замене переменных  $x = Ty$  имеем  $y'T'A^{-1}Ty = y'B^{-1}y = 1$ . В том, что  $B^{-1} = T'A^{-1}T$ , легко убедиться:  $BB^{-1} = T'ATT'A^{-1}T = I$ , учитывая ортогональность матрицы  $T$ . Получаем, что эллипсоид рассеяния инвариантен к ортогональной замене переменных и является функцией только взаимного расположения выборочных точек, поэтому его можно использовать для графической иллюстрации характерной формы рассеяния.

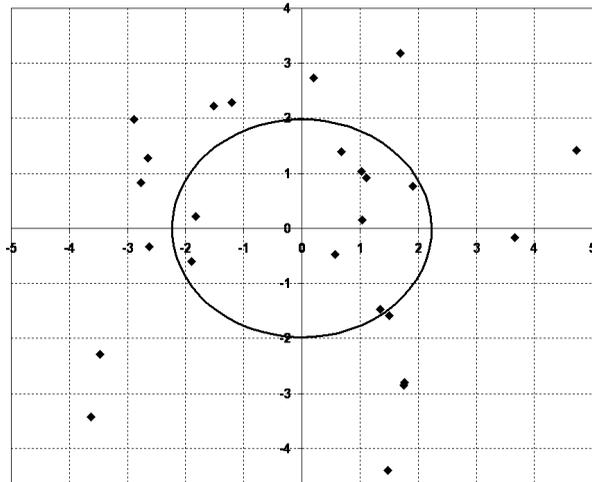
Поверхности, задаваемые уравнением  $x'A^{-1}x = C$ ,  $C \geq 0$  — константа, являются в случае нормального закона распределения поверхностями равной плотности, поэтому эллипсоид рассеяния может служить для наглядного представления «формы» распределения.

Вычислим координаты всех выборочных точек в пространстве  $Y_1, Y_2, Y_3$  по формуле  $y = T'x$ . Данное пространство называют также пространством *главных компонент*.

Результаты приведены в таблице.

$i$	$Y_1$	$Y_2$	$Y_3$	$i$	$Y_1$	$Y_2$	$Y_3$	$i$	$Y_1$	$Y_2$	$Y_3$
1	1,91	0,77	-0,63	10	-2,76	0,83	0,43	18	-2,65	1,28	0,11
2	1,34	-1,47	1,85	11	-3,47	-2,29	0,41	19	1,48	-4,40	-2,59
3	-1,20	2,29	-0,61	12	0,57	-0,47	2,09	20	1,69	3,18	0,22
4	1,77	-2,80	0,11	13	0,20	2,73	-1,08	21	-3,62	-3,43	-0,29
5	0,68	1,39	-0,62	14	1,04	0,15	0,26	22	1,51	-1,58	-1,60
6	1,03	1,03	-0,47	15	-1,51	2,22	-1,65	23	4,73	1,42	2,62
7	-2,89	1,98	-0,35	16	-1,83	0,22	0,20	24	1,10	0,92	-2,40
8	3,66	-0,17	0,03	17	1,76	-2,85	0,01	25	-1,90	-0,60	3,46
9	-2,62	-0,34	0,49								

Выборочные ковариации для полученных переменных равны нулю.



Выделение главных компонент используют как метод сокращения размерности данных, оставляя заданное число переменных, соответствующих наибольшим собственным значениям.

Также переход в пространство собственных векторов используется при визуализации данных. На рисунке изображена выборка в пространстве  $Y_1, Y_2$  и эллипс рассеяния.

### Задание 1

1. Проверить гипотезу равенства дисперсий текущей выборки и выборки под номером  $j_1$ .

2. Проверить гипотезу равенства математических ожиданий текущей выборки и выборки под номером, взятым из столбца 2.

3. Проверить гипотезу одновременного равенства математических ожиданий (факторный анализ) текущей выборки и выборок под номерами  $j_2$  и  $j_3$ .

Уровень значимости выбрать 0,1.

№	$n$	Выборка	$J_1$	$J_2$	$J_3$
0	10	0,7; 0,3; 0,5; 1,5; 0,2; 0,9; -0,5; 0,7; 1,0; 0,3;	1	10	25
1	8	0,2; -1,8; -0,8; 1,7; -2,0; -0,6; 4,2; -1,0;	30	6	13
2	12	0,1; -2,9; -0,4; 0,6; 3,3; -1,5; 0,2; -1,0; -4,3; -2,2; -0,9; -1,7;	24	1	6
3	8	10,6; 7,4; 9,4; 9,2; 11,0; 4,9; 0,9; -3,2;	15	24	26
4	12	3,8; 2,0; 1,8; 3,5; 1,5; 1,8; 1,5; 1,6; 1,0; 3,3; 2,8; 1,4;	8	10	23
5	15	10,2; 8,6; 11,0; 8,8; 13,0; 14,5; 12,0; 14,4; 14,2; 11,4; 12,2; 13,9; 9,2; 12,9; 12,4	29	16	2
6	14	0,2; -0,2; 0,4; 1,0; -0,8; 1,0; -2,0; -0,9; 1,9; -3,0; 2,3; 3,8; 0,3; 0,3;	13	19	5
7	8	2,5; 3,2; 2,6; 3,0; 2,4; 3,1; 3,8; 3,2;	17	18	28
8	12	-2,6; 2,4; -1,4; -1,3; 2,9; 0,7; 4,6; -3,2; -1,5; 1,1; 6,5; -0,2;	9	20	16
9	10	2,4; 1,8; -6,3; -0,2; 1,1; 0,1; -1,0; -3,6; -1,5; 0,5;	20	8	20
10	10	2,4; 1,9; 2,0; -0,7; -0,7; -0,3; 1,3; 2,5; 2,3; 1,5;	12	17	25
11	11	9,5; 9,1; 11,1; 10,6; 9,3; 11,7; 15,6; 11,3; 6,6; 5,3; 10,4;	3	3	24
12	12	2,8; 2,1; 3,2; 2,1; 2,8; 3,3; 3,6; 2,1; 2,7; 3,7; 3,5; 2,1;	28	7	28
13	9	-2,2; 0,4; -0,3; 0,0; 1,0; 0,1; -2,4; -2,4; -0,9;	5	6	19
14	10	-11,2; -10,1; -13,7; -6,8; -5,1; -12,3; -7,0; -14,4; -6,6; -8,0;	11	30	15
15	11	7,7; -13,7; -8,1; -0,1; -6,1; -5,9; -13,2; -12,2; -12,9; 0,0; -10,8;	30	27	14
16	12	4,8; 9,8; 1,3; 8,3; 4,1; 6,9; 10,3; 7,9; 11,2; 4,1; 2,5; 2,6;	17	21	8
17	15	-3,7; -5,7; -4,5; -6,2; -2,1; -4,9; -4,9; -6,9; -2,8; -4,4; -6,6; -4,6; -4,0; -4,8; -4,7	4	23	25
18	12	2,5; 3,3; 2,9; 2,8; 1,7; 3,7; 3,1; 2,3; 3,0; 3,0; 3,8; 3,6;	4	12	12

№	$n$	Выборка	$j_1$	$j_2$	$j_3$
19	11	4,2; 0,1; 1,5; 1,6; 0,3; -0,7; 1,8; -0,1; 0,3; 1,6; 2,2;	21	13	2
20	14	-6,3; -1,4; -0,4; -3,2; 4,5; -3,9; -0,3; -3,7; -1,1; -2,9; 0,7; -2,2; -1,0; 1,8;	29	9	16
21	12	10,8; 0,9; 4,4; 5,0; 6,9; 8,1; 2,8; 3,6; 5,7; -0,3; 3,1; 5,2;	8	9	20
22	10	-4,1; -2,3; -3,9; -6,0; -3,3; -3,4; -3,6; -4,3; -4,3; -2,4;	10	23	25
23	8	-4,9; -5,4; -4,1; -5,7; -6,0; -4,9; -4,3; -6,5;	25	10	22
24	12	12,7; 5,7; 7,2; 6,8; 18,1; 13,0; 13,2; 9,6; 18,3; 1,5; 10,5; 10,6;	11	26	11
25	10	1,7; 1,7; 1,9; 0,4; -0,3; 1,9; 2,8; 3,3; 3,2; 2,5;	5	4	10
26	12	3,1; 14,4; 12,8; 14,5; 10,6; 14,0; 10,5; 1,3; 10,1; 2,5; 6,8; 11,3;	27	3	11
27	15	-2,1; -3,3; -3,7; -4,7; -10,0; -6,1; -6,9; -13,7; -6,3; -4,4; -14,8; -9,5; -9,4; 1,7; 5,2	14	14	30
28	11	2,6; 3,2; 3,2; 3,2; 3,2; 3,1; 2,9; 3,3; 2,5; 2,8; 2,7;	7	12	18
29	10	-2,4; 5,0; -2,4; 8,2; 3,5; 0,3; 3,7; -3,6; -2,1; -6,0;	4	11	26
30	11	-1,3; -16,1; -11,6; -10,9; -8,4; 4,4; -5,5; -9,6; -4,6; -14,4; -19,5;	20	15	27

## Задание 2

1. Построить линейную регрессию  $X_2$  от  $X_1$ . Найти доверительные интервалы для параметров при  $\eta = 0,9$ . Построить эллипс рассеяния. Проверить гипотезу независимости  $X_2$  и  $X_1$  при уровне значимости 0,1.

2. Построить разделяющую кривую для двух классов в предположении нормальности распределений,  $j$  – номер выборки для второго класса.

$N_0$	0		1		2		3		4	
$j$	1		2		3		5		5	
$i$	$X_1$	$X_2$								
1	0,73	0,58	5,27	0,65	5,09	3,61	-1,90	-1,31	5,01	5,82
2	2,59	-3,68	1,80	-0,64	-4,29	-1,39	-0,67	-0,77	1,55	2,51
3	-2,66	2,39	0,26	-0,03	-0,53	-1,04	0,01	-1,03	3,08	0,67
4	-0,14	-0,72	3,07	-0,28	3,67	0,92	0,41	-1,67	6,37	4,81
5	3,82	-0,83	2,69	0,95	6,68	1,92	-0,11	2,77	2,84	3,48
6	-0,70	0,45	4,06	0,53	-1,46	1,48	-1,23	-5,79	2,98	1,92
7	-0,87	-1,20	1,47	0,21	-2,88	-5,75	2,03	2,14	6,27	4,74
8	0,55	-1,62	3,48	2,41	-0,15	1,55	0,56	0,07	-0,27	1,90
9	4,20	-4,50	4,73	0,63	4,64	3,31	0,39	-0,28	9,33	6,17
10	-2,38	-0,20	3,99	1,51	5,62	3,04	-0,58	1,23	0,53	2,43
11	-4,57	2,31	2,47	1,40	4,13	0,76	1,21	0,56	1,25	4,27
12	0,55	1,15	0,73	-1,66	4,34	1,10	1,21	-1,04	0,24	2,56
13	0,87	1,60	0,71	-1,10	0,89	-0,78	0,15	-1,88	7,02	4,43
14	4,45	-3,07	1,57	0,18	4,46	3,81	0,27	-2,01	1,53	3,40
15	3,46	-1,20	5,13	4,10	4,12	-1,50	-1,03	-1,76	0,97	4,00
16	-2,31	3,15	1,04	2,43	3,17	-1,75	-0,60	1,16	7,49	7,01
17	3,86	-3,34	0,69	1,67	-1,90	0,03	0,62	-3,64	2,91	3,48
18	-1,40	1,71	1,67	0,05	-1,95	1,70	1,37	-2,64	0,11	1,77
19	-2,45	1,48	1,57	1,87	9,27	2,96	-1,19	1,02	2,73	1,36
20	0,08	1,85	7,00	1,76	3,85	-0,23	1,20	-0,79	4,57	5,94
21	0,92	-0,02	-2,80	-1,34	3,76	2,86	0,09	-0,03	1,44	2,63
22	0,36	0,27	2,08	-1,31	1,27	-0,66	-0,93	0,62	5,72	3,11
23	1,29	-0,41			2,60	0,23	-0,19	0,21	0,62	2,63
24	4,02	-1,59			3,36	2,93	0,07	-3,85		
25	-5,57	1,77			1,47	0,56				

№	5		6		7		8		9	
$j$	6		4		8		9		6	
$i$	$X_1$	$X_2$								
1	-1,58	-1,66	-3,93	-2,52	2,77	-0,98	8,36	-2,03	-1,96	-0,93
2	2,39	-0,87	-2,08	0,77	3,88	-1,42	3,42	3,77	-0,36	-3,05
3	0,16	3,61	-0,17	1,48	5,43	2,75	1,42	2,79	-0,76	0,26
4	-0,14	-1,62	1,05	-0,39	5,01	0,93	5,24	-1,17	-2,90	-1,03
5	1,85	-1,19	-2,41	-0,74	0,12	-2,85	5,47	-1,84	-3,83	-0,21
6	0,96	0,75	-0,86	-4,93	11,54	5,20	5,95	-0,70	-1,72	-1,88
7	-1,40	1,89	-3,76	-1,16	3,77	-1,27	3,64	1,64	-2,93	2,14
8	0,31	1,01	-2,51	-2,12	4,88	2,71	6,15	1,04	-0,45	-1,87
9	1,44	0,48	4,51	0,36	5,18	1,94	6,59	-0,81	4,68	-3,60
10	0,29	1,18	-1,69	0,24	6,09	-0,56	3,79	1,53	-1,90	-3,56
11	-1,06	-1,18	-4,81	4,14	5,70	1,77	4,53	-1,56	-1,92	-0,56
12	-1,12	3,23	-1,11	-4,88	6,35	3,83	3,51	1,05	-1,94	-0,51
13	-0,67	-0,35	-0,86	-0,71	1,86	-0,76	2,51	0,64	-0,47	-1,13
14	-0,56	1,95	-0,40	-0,28	4,11	1,94	6,08	-0,71	0,33	-1,67
15	0,33	1,41	3,49	-6,27	8,45	3,09	3,37	2,39	-6,48	-0,20
16	0,43	1,54	-1,85	0,96	7,11	1,09	5,02	-1,70	-0,49	-1,26
17	-0,10	1,20	1,42	-1,46	6,42	-1,48	10,24	-6,73	-2,64	-0,27
18	-1,65	-0,02	-0,74	-0,34	4,18	-1,53	-0,74	5,58	-3,00	1,66
19	1,02	-0,48	-1,99	1,40	10,50	4,96	2,88	0,79	-0,34	-0,90
20	-1,16	1,86	3,73	-5,38	-1,13	-5,65	8,55	-2,71	-0,70	-3,16
21	-1,54	4,03			2,93	-3,74	0,34	2,90	-0,09	-1,54
22	-0,12	2,76			7,71	2,86	10,60	-5,81	-4,67	1,11
23	0,34	0,82			7,27	2,01			1,86	-3,05
24	-0,21	1,71			6,05	0,64				
25	0,41	-0,47								

$N_0$	10		11		12		13		14	
$j$		11		12		14		14		15
$i$	$X_1$	$X_2$								
1	-0,72	0,52	2,92	-3,25	1,64	5,12	-0,19	2,37	-0,90	-2,02
2	-7,08	2,99	-1,94	2,78	0,13	1,74	2,67	-2,78	-2,58	-3,72
3	1,85	0,57	-1,30	2,09	1,63	3,75	1,16	0,32	-6,44	-5,07
4	-3,86	4,46	-0,31	2,36	0,04	2,33	2,48	-3,35	-3,91	-3,62
5	2,24	-5,37	-3,74	6,27	0,45	2,63	-1,90	4,57	-0,49	-1,31
6	1,09	-1,28	-0,84	3,74	5,57	-0,15	-2,20	4,75	1,75	1,40
7	2,90	0,49	0,66	0,40	6,43	-1,05	-1,31	1,35	-4,48	-4,04
8	-4,09	2,36	0,16	2,95	-3,40	6,77	3,35	0,76	2,02	-1,14
9	-2,22	-0,51	-2,19	2,69	-3,20	6,87	0,93	-1,77	-3,58	-0,98
10	-5,63	-1,21	0,07	2,03	-0,49	2,04	5,13	-7,91	-1,18	-1,92
11	0,69	-0,26	-2,86	7,02	2,73	3,15	-1,65	2,24	-0,27	1,31
12	-2,36	-2,11	0,03	5,67	3,47	1,32	-1,64	1,72	-0,89	0,12
13	-1,40	1,04	2,38	0,60	0,91	2,96	-1,00	2,60	-0,84	-2,33
14	0,09	0,14	2,00	-0,10	3,14	3,59	0,03	1,49	3,17	-2,10
15	1,85	-4,66	-0,48	1,76	2,98	1,80	-1,58	2,70	-2,21	-0,33
16	2,73	-2,24	0,59	4,46	2,57	3,52	0,10	3,04	3,14	-0,29
17	-0,82	-2,47	-1,19	2,87	5,31	-1,53	-0,41	1,23	-2,20	-1,92
18	1,58	2,25	1,85	-2,40	0,46	1,64	0,00	2,17	3,15	1,96
19	-3,27	2,74	0,89	2,98	1,48	2,37	-2,03	3,88	-4,67	-4,30
20	-0,16	3,30	-4,41	5,59	6,12	-1,92	-0,24	0,96	1,78	-0,37
21	-0,24	2,35	6,39	-3,76	5,13	-0,97	-1,18	4,62	-3,35	-2,95
22	-0,24	1,40	-0,13	1,18	-0,56	5,16			-1,08	0,25
23	-2,01	-0,35	-1,32	4,42	2,41	0,80			-7,58	-4,15
24	5,19	0,31	3,04	-0,28					3,14	0,26
25	2,58	0,38							-2,39	-1,47

№	15		16		17		18		19	
$j$		16		17		20		19		20
$i$	$X_1$	$X_2$								
1	6,02	5,24	8,69	5,35	-9,65	9,18	1,70	5,87	-4,29	0,71
2	0,97	2,86	-6,07	-5,00	-10,1	4,90	4,67	-5,15	-0,81	-2,74
3	-3,60	-0,41	3,86	0,56	-4,54	5,52	2,90	-2,50	-2,96	-1,08
4	-3,42	0,34	-0,60	0,81	-4,80	5,47	3,93	-1,24	0,93	-1,67
5	-0,10	1,38	3,05	-2,76	-4,51	4,76	1,86	2,22	-4,18	0,97
6	0,04	1,61	-7,46	-2,02	-9,55	8,67	4,94	-0,98	-4,65	-1,92
7	-1,09	-0,71	1,08	2,76	-3,68	5,35	5,47	-5,22	0,21	-1,92
8	7,88	4,83	2,27	-0,82	1,91	2,37	3,86	-3,50	-6,72	-1,38
9	3,00	1,20	-6,37	-0,16	-5,57	5,42	3,40	-0,22	1,03	-3,33
10	-2,20	1,99	-3,45	1,56	-1,60	0,85	3,00	-0,58	-4,18	-0,29
11	-0,08	2,02	4,23	0,07	-6,33	4,95	2,27	-0,74	-3,98	3,14
12	-1,15	1,30	-2,59	0,86	2,46	1,44	4,13	-1,70	-1,45	-0,37
13	-1,55	0,51	-0,46	2,13	-10,9	8,11	3,72	1,97	-2,32	-4,67
14	-2,19	-0,74	-4,77	-2,82	-4,80	4,25	2,17	0,72	1,33	-2,92
15	-3,96	-0,80	-1,97	-1,14	-8,70	8,38	5,19	-4,20	-1,75	-1,92
16	2,13	2,93	-0,03	0,76	-5,53	8,33	2,07	-1,18	-4,63	3,20
17	1,94	3,18	4,57	4,36	-6,81	5,31	3,43	-2,78	-1,50	-2,15
18	3,28	5,18	-2,87	-1,83	-6,27	5,51	4,20	3,38	-7,12	2,65
19	-0,80	1,31	1,05	3,09	-5,90	5,04	2,29	-1,22	0,27	-1,68
20	0,26	0,63	5,28	1,26	-4,33	4,80	2,00	2,31	-3,60	2,16
21	4,74	3,30	1,58	-1,37	-4,23	4,30	3,58	-1,67	-9,32	5,19
22	-2,04	0,37	10,62	7,19	-6,34	4,83	0,69	5,45	-5,51	-1,80
23	5,92	3,51			2,63	1,59	5,11	-4,18	2,03	0,70
24					-8,31	7,41			-1,58	-0,12
25									-1,29	-2,73

№	20		21		22		23		24	
$j$		21		15		23		25		25
$i$	$X_1$	$X_2$								
1	-0,29	0,43	-1,01	-0,12	-10,4	-0,69	4,82	-0,94	-1,77	2,64
2	0,69	-3,20	0,09	-0,05	-5,90	2,58	13,09	4,19	1,34	-0,61
3	3,61	-2,10	-2,17	0,38	1,92	3,27	6,60	-0,07	-3,53	2,96
4	1,08	-0,14	3,48	-2,00	-8,47	1,63	11,63	1,23	2,86	-1,87
5	2,49	-3,23	-3,73	-2,01	-11,4	-1,22	14,59	0,94	0,60	-1,22
6	-0,58	0,47	6,43	1,15	-4,31	2,90	9,26	-2,16	0,34	-0,10
7	0,63	-0,92	1,58	-0,97	-2,46	3,47	3,91	-2,22	-0,91	0,60
8	1,57	-0,83	3,65	1,13	-4,63	-1,26	-1,80	-6,86	-1,83	0,83
9	-1,08	2,37	4,31	0,78	0,50	5,33	7,16	-0,31	-0,53	1,75
10	0,11	1,12	2,04	0,57	-5,92	-0,14	1,28	-6,98	-1,42	0,50
11	3,59	-1,55	3,07	0,17	-6,19	2,37	10,89	0,33	-1,58	0,61
12	2,44	-1,49	1,20	-0,16	-2,07	2,18	14,10	3,11	-3,24	0,59
13	1,43	-2,88	1,84	-0,69	-9,01	-0,18	9,27	1,34	0,48	1,64
14	-1,62	1,52	3,57	1,54	-2,34	2,38	10,54	0,22	-2,00	3,04
15	3,06	-1,90	2,09	-0,06	-4,06	1,19	5,35	-1,24	0,60	0,93
16	-1,83	3,18	5,42	-1,24	-10,1	-1,51	11,55	3,06	3,43	-4,35
17	-7,65	6,58	1,43	-1,56	-4,78	0,93	6,02	3,83	-1,46	2,21
18	-2,54	2,70	1,09	-1,86	-17,5	-2,37	16,38	5,77	-3,67	2,80
19	-2,75	1,95	-0,82	-1,11	0,01	4,64	6,52	-1,00	-4,35	2,75
20	1,30	-2,64	2,18	-0,55	-18,0	-3,09	5,91	-1,33	-3,13	4,17
21	-1,71	2,90			-2,08	4,36	-1,22	-4,95	1,44	-1,23
22	-0,77	1,08			-2,57	3,37	8,15	-0,59	-0,58	1,46
23	1,05	-1,55			-11,1	-0,57	10,97	4,55	-0,80	1,97
24	2,23	-2,44			-9,28	-1,33			-0,84	-0,63
25									3,05	-2,65

№	25		26		27		28		29		30	
$j$		26		27		28		30		30		27
$i$	$X_1$	$X_2$										
1	-2,4	1,96	-2,9	11,2	1,55	1,30	-6,4	1,72	-1,7	0,28	1,95	0,78
2	-1,9	2,73	0,72	1,31	-0,1	0,43	-3,8	0,36	-0,2	-0,1	4,89	3,28
3	0,17	0,09	0,26	5,34	-3,9	-2,4	-6,9	3,56	-5,2	-3,5	5,48	3,64
4	4,47	-2,7	-4,9	11,6	1,33	1,86	-6,0	1,70	4,38	1,76	2,60	3,28
5	0,62	0,34	-0,5	8,03	1,96	2,13	-5,4	3,50	1,05	0,11	4,72	2,14
6	2,92	-1,4	1,31	6,02	4,20	2,27	-5,4	-0,5	-3,1	-2,0	1,60	4,25
7	2,20	-0,2	-2,3	12,2	3,46	2,99	-6,1	0,39	1,93	0,70	4,33	2,59
8	6,13	-2,0	-9,9	19,8	5,58	3,34	-3,6	2,04	0,83	0,47	4,05	2,18
9	3,02	-0,7	-2,6	8,73	0,49	-0,3	-4,8	2,29	0,05	0,11	5,62	2,34
10	-0,8	0,68	0,35	3,46	-4,9	0,0	-3,7	-0,2	1,99	1,82	1,09	3,51
11	1,98	-1,1	-1,0	7,46	-2,1	-1,1	-8,0	3,41	-0,6	-0,4	4,14	2,04
12	2,85	-1,8	0,1	9,31	0,23	-0,8	-8,6	2,65	4,84	2,31	2,74	3,77
13	-0,5	1,23	-0,7	12,5	-0,2	1,59	-0,5	-3,1	-1,5	-0,7	3,95	1,12
14	1,18	-0,1	10,1	1,08	-2,0	-1,1	-2,7	1,63	-1,7	-0,5	4,22	3,88
15	2,54	0,13	-2,6	10,7	2,06	4,04	-9,8	3,56	-0,9	-0,5	2,43	3,43
16	1,00	1,29	0,0	8,26	-0,5	1,27	-4,8	-1,2	-4,5	-1,2	4,49	3,26
17	-0,1	0,72	0,99	6,32	-2,0	0,41	-4,9	1,17	-6,5	-2,2	6,65	2,89
18	-2,6	2,15	-2,2	11,7	2,90	2,72	-2,4	-2,2	2,94	2,78	3,75	3,08
19	0,71	0,22	0,30	9,36	3,55	1,21	-10	4,41	0,36	0,0	7,39	1,75
20	0,36	0,40	-1,5	10,6	-0,6	2,22	-2,5	-3,5	1,21	0,75	3,64	3,83
21	0,48	0,69	-3,7	13,1	-1,1	0,13	-4,5	1,09	-2,2	-2,0	-0,1	4,02
22	3,25	-0,1	-1,6	8,62	-2,4	-0,1	-3,8	1,14	-1,4	-0,9	6,00	2,21
23	-0,5	0,84	4,81	5,81			-6,7	0,96	9,83	4,19	1,19	2,09
24			4,17	5,87			-4,9	2,56			5,18	3,42
25							-5,0	2,83				

### Задание 3

1. Диагонализировать матрицу выборочных ковариаций в пространстве переменных  $Z_{j_1}, Z_{j_2}, Z_{j_3}$ .

2. Вычислить координаты всех точек выборки в базисе собственных векторов ковариационной матрицы.

3. Изобразить выборку и эллипс рассеяния в пространстве двух главных компонент.

Варианты заданий.

№	$j_1$	$j_2$	$j_3$	№	$j_1$	$j_2$	$j_3$
0	2	3	8	16	2	8	9
1	1	2	7	17	3	4	6
2	2	3	7	18	2	7	9
3	2	4	7	19	2	6	7
4	2	5	9	20	2	5	7
5	2	8	10	21	4	6	8
6	3	7	10	22	3	6	8
7	6	9	10	23	2	7	8
8	1	2	3	24	2	6	9
9	1	3	6	25	2	3	9
10	1	7	10	26	2	3	5
11	1	8	10	27	1	6	10
12	2	3	10	28	1	5	6
13	2	4	6	29	1	3	4
14	2	4	9	30	1	2	6
15	2	6	10				

Выборка.

$i$	$Z_1$	$Z_2$	$Z_3$	$Z_4$	$Z_5$	$Z_6$	$Z_7$	$Z_8$	$Z_9$	$Z_{10}$
1	-0,08	2,76	4,53	8,49	0,44	-1,66	1,87	3,57	3,22	-3,58
2	0,64	-0,21	2,91	10,76	1,66	-1,68	8,97	3,49	2,26	-3,31
3	1,87	1,53	7,72	8,87	-2,95	-3,95	8,84	4,20	-4,33	-5,13
4	4,80	0,58	2,55	7,13	1,79	-1,89	4,74	1,44	-2,79	0,41
5	-1,95	2,28	5,80	13,15	-2,09	-2,21	7,79	3,84	3,79	0,34
6	2,91	2,23	5,29	9,95	3,06	-5,53	5,83	3,73	0,00	-3,36
7	-0,80	0,19	8,76	10,73	1,80	-1,47	10,86	3,87	-0,97	-0,70
8	2,89	2,97	2,45	7,98	-3,52	0,88	2,80	3,61	0,44	-0,99
9	4,18	-1,24	7,06	9,84	-1,29	-3,19	7,32	2,77	-2,44	0,35
10	1,37	-0,76	7,79	8,72	5,66	-1,00	5,56	3,56	-2,62	-2,37
11	2,74	-2,58	6,75	5,55	2,49	0,98	6,47	1,15	-3,61	-1,85
12	2,30	-0,39	3,93	9,13	0,32	-5,31	7,78	4,25	2,09	-7,28
13	2,99	2,90	7,03	13,47	-1,83	-2,25	2,98	4,44	-1,95	-2,31
14	4,65	1,37	4,56	7,23	2,34	-4,32	5,18	3,57	3,21	-1,26
15	1,51	1,99	8,31	8,60	4,55	-3,30	8,01	3,40	3,09	-4,78
16	-1,75	-0,31	6,84	8,60	6,77	-2,27	11,40	3,12	-4,38	-0,10
17	3,17	0,61	2,57	10,14	-3,50	-2,45	5,04	1,33	2,40	0,84
18	0,59	-0,28	8,05	10,82	4,55	0,51	12,41	3,70	1,61	1,84
19	-1,91	1,43	2,95	8,71	-5,01	-0,98	3,76	-1,57	-1,04	-0,27
20	1,12	3,16	5,61	7,88	2,63	-2,72	3,26	5,87	3,87	-1,79
21	3,00	-2,74	6,56	11,53	-0,16	-2,25	12,72	-0,16	-1,04	-0,34
22	1,00	2,08	4,01	8,22	2,15	0,56	11,57	1,15	3,05	4,86
23	1,77	2,66	1,47	10,09	1,93	-6,83	5,04	6,67	-2,90	-5,12
24	5,09	3,48	5,89	12,13	0,14	-6,67	4,55	2,38	0,56	-3,68
25	4,43	-2,84	5,26	8,83	3,70	-0,26	9,37	4,67	-2,75	0,50

## Приложение. Критические точки распределения Фишера

Критические точки распределения Фишера. В таблице приведены значения  $x_\alpha$ , такие что  $F_{ij}(x > x_\alpha) = \alpha$ .

Имеет место тождество  $F_{ij}(x > x_\alpha) = F_{ji}(x < \frac{1}{x_\alpha})$ .

В левом верхнем углу таблицы указана пороговая вероятность  $\alpha$ , первое число степеней свободы  $i$  — определяется номером столбца, второе число степеней свободы  $j$  — номером строки.

0,025	1	2	3	5	7	10	15	20	30	50
1	648	799	864	922	948	969	985	993	1001	1008
2	38,5	39,0	39,2	39,3	39,4	39,4	39,4	39,4	39,5	39,5
3	17,4	16,0	15,4	14,9	14,6	14,4	14,3	14,2	14,1	14,0
4	12,2	10,6	9,98	9,36	9,07	8,84	8,66	8,56	8,46	8,38
5	10,0	8,43	7,76	7,15	6,85	6,62	6,43	6,33	6,23	6,14
7	8,07	6,54	5,89	5,29	4,99	4,76	4,57	4,47	4,36	4,28
8	7,57	6,06	5,42	4,82	4,53	4,30	4,10	4,00	3,89	3,81
10	6,94	5,46	4,83	4,24	3,95	3,72	3,52	3,42	3,31	3,22
12	6,55	5,10	4,47	3,89	3,61	3,37	3,18	3,07	2,96	2,87
14	6,30	4,86	4,24	3,66	3,38	3,15	2,95	2,84	2,73	2,64
16	6,12	4,69	4,08	3,50	3,22	2,99	2,79	2,68	2,57	2,47
18	5,98	4,56	3,95	3,38	3,10	2,87	2,67	2,56	2,44	2,35
20	5,87	4,46	3,86	3,29	3,01	2,77	2,57	2,46	2,35	2,25
25	5,69	4,29	3,69	3,13	2,85	2,61	2,41	2,30	2,18	2,08
30	5,57	4,18	3,59	3,03	2,75	2,51	2,31	2,20	2,07	1,97
40	5,42	4,05	3,46	2,90	2,62	2,39	2,18	2,07	1,94	1,83
50	5,34	3,97	3,39	2,83	2,55	2,32	2,11	1,99	1,87	1,75

<b>0,05</b>	1	2	3	5	7	10	15	20	30	50
1	161	199	216	230	237	242	246	248	250	252
2	18,5	19,0	19,2	19,3	19,4	19,4	19,4	19,4	19,5	19,5
3	10,1	9,55	9,28	9,01	8,89	8,79	8,70	8,66	8,62	8,58
4	7,71	6,94	6,59	6,26	6,09	5,96	5,86	5,80	5,75	5,70
5	6,61	5,79	5,41	5,05	4,88	4,74	4,62	4,56	4,50	4,44
7	5,59	4,74	4,35	3,97	3,79	3,64	3,51	3,44	3,38	3,32
8	5,32	4,46	4,07	3,69	3,50	3,35	3,22	3,15	3,08	3,02
10	4,96	4,10	3,71	3,33	3,14	2,98	2,85	2,77	2,70	2,64
12	4,75	3,89	3,49	3,11	2,91	2,75	2,62	2,54	2,47	2,40
14	4,60	3,74	3,34	2,96	2,76	2,60	2,46	2,39	2,31	2,24
16	4,49	3,63	3,24	2,85	2,66	2,49	2,35	2,28	2,19	2,12
20	4,35	3,49	3,10	2,71	2,51	2,35	2,20	2,12	2,04	1,97
25	4,24	3,39	2,99	2,60	2,40	2,24	2,09	2,01	1,92	1,84
30	4,17	3,32	2,92	2,53	2,33	2,16	2,01	1,93	1,84	1,76
40	4,08	3,23	2,84	2,45	2,25	2,08	1,92	1,84	1,74	1,66
50	4,03	3,18	2,79	2,40	2,20	2,03	1,87	1,78	1,69	1,60

<b>0,1</b>	1	2	3	5	7	10	15	20	30	50
1	39,9	49,5	53,6	57,2	58,9	60,2	61,2	61,7	62,3	62,7
2	8,53	9,00	9,16	9,29	9,35	9,39	9,42	9,44	9,46	9,47
3	5,54	5,46	5,39	5,31	5,27	5,23	5,20	5,18	5,17	5,15
4	4,54	4,32	4,19	4,05	3,98	3,92	3,87	3,84	3,82	3,80
5	4,06	3,78	3,62	3,45	3,37	3,30	3,24	3,21	3,17	3,15
7	3,59	3,26	3,07	2,88	2,78	2,70	2,63	2,59	2,56	2,52
8	3,46	3,11	2,92	2,73	2,62	2,54	2,46	2,42	2,38	2,35
10	3,29	2,92	2,73	2,52	2,41	2,32	2,24	2,20	2,16	2,12
12	3,18	2,81	2,61	2,39	2,28	2,19	2,10	2,06	2,01	1,97
14	3,10	2,73	2,52	2,31	2,19	2,10	2,01	1,96	1,91	1,87
16	3,05	2,67	2,46	2,24	2,13	2,03	1,94	1,89	1,84	1,79
20	2,97	2,59	2,38	2,16	2,04	1,94	1,84	1,79	1,74	1,69
25	2,92	2,53	2,32	2,09	1,97	1,87	1,77	1,72	1,66	1,61
30	2,88	2,49	2,28	2,05	1,93	1,82	1,72	1,67	1,61	1,55
40	2,84	2,44	2,23	2,00	1,87	1,76	1,66	1,61	1,54	1,48
50	2,81	2,41	2,20	1,97	1,84	1,73	1,63	1,57	1,50	1,44

## Литература

1. Основы математической статистики : учебно-метод. пособие / Г.Н. Миренкова, С.В. Неделько, В.М. Неделько, Т.В. Тренёва. – Новосибирск : Изд-во НГТУ, 2008. – 36 с.
2. Г.С. Лбов. Анализ данных и знаний : учебное пособие. – Новосибирск : Изд-во НГТУ, 2001. – 86 с.
3. В.И. Лотов. Теория вероятностей и математическая статистика : курс лекций. – Новосибирск : Изд-во НГУ, 2006. – 128 с.
4. В.М. Неделько, Т.А. Ступина. Основы теории вероятностей и математической статистики в примерах и задачах. Учебное пособие. – Новосибирск: Издательство НГУ, 2006. – 82с.
5. В.М. Бородихин, А.П. Ковалевский. Высшая математика. Том 4.2. Теория вероятностей и математическая статистика. Учебное пособие. – Новосибирск, 2005. – 256с.
6. Сборник задач по теории вероятностей, математической статистике и теории случайных функций / Под ред. А.А. Свешникова. – М.: Наука, 1965. – 632 с.
7. А.А. Боровков. Математическая статистика. Учебник. – М.: Наука, 1984. – 472 с.