

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Статистическое моделирование

Третьякова Александра, Зенкова Наталья

МНОГОМЕРНЫЙ АНАЛИЗ ДАННЫХ

Лекции Голяндиной Нины Эдуардовны

Санкт-Петербург

2019

Оглавление

Введение	3
1. Генеральный язык	3
2. Выборочный язык	3
3. Примеры данных до стандартизации и после	4
4. Многомерное нормальное распределение	5
5. Переход к новым признакам	8
5.1. Две интерпретации новых признаков. Факторные нагрузки	10

Введение

1. Генеральный язык

Пусть случайная величина $\xi = (\xi_1, \dots, \xi_p)^T \in \mathbb{R}^p$, где p — количество признаков. Обозначим

- *вектор средних* $\mu = (E\xi_1, \dots, E\xi_p)^T \in \mathbb{R}^p$;
- *ковариационную матрицу* $\Sigma = \text{Cov}\xi = E(\xi - E\xi)(\xi - E\xi)^T \in \mathbb{R}^{p \times p}$.

Рассмотрим два случая:

- *Центрированные данные*: $\xi^{(c)} = \xi - E\xi$. Тогда ковариационная матрица будет иметь вид: $\text{Cov}\xi^{(c)} = E\xi^{(c)}(\xi^{(c)})^T$.
- *Стандартизованные данные*: $\xi_i^{(s)} = \frac{\xi_i - E\xi_i}{\sqrt{D\xi_i}}$ для каждой ξ_i , i -ой компоненты вектора ξ . Запишем теперь это в матричном виде: введём Δ — диагональную матрицу, состоящую из элементов $\sqrt{D\xi_i}$ для $i = 1, \dots, p$. Тогда $\xi^{(s)} = \Delta^{-1}\xi^{(c)} \Rightarrow \text{Corr}(\xi) = \text{Cov}(\xi^{(s)})$, то есть $\text{Cov}(\xi^{(s)})$ — это корреляционная матрица до стандартизации.

2. Выборочный язык

Перейдём теперь на выборочный язык. Пусть $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ — выборка объёма n . Можем задать эмпирическое распределение как $\{\mathbf{x}_i$ с вероятностью $1/n\}$. Возникает вопрос: как задать *матрицу данных*? Удобно было бы ее задать как матрицу размера $p \times n$:

$$\begin{array}{|c|c|c|} \hline & & \\ \hline \mathbf{x}_1 & \dots & \mathbf{x}_n \\ \hline \end{array}.$$

Но в книгах встречается следующий вариант *матрицы данных*:

$$\mathbb{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = [X_1 : \dots : X_p] \in \mathbb{R}^{n \times p}.$$

Теперь знаем, как написать все характеристики, упомянутые выше, на выборочном языке.

- $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n (X_j)_i$ — выборочное среднее, $j = 1, \dots, p$.

$$\bar{X}_j = (\bar{x}_j, \dots, \bar{x}_j)^T \in \mathbb{R}^n.$$

Центрированная матрица данных $\mathbb{X}^{(c)}$ ¹:

$$\mathbb{X}^{(c)} = [X_1 - \bar{X}_1 : \dots : X_p - \bar{X}_p] = [X_1^{(c)} : \dots : X_p^{(c)}].$$

- $s_j^2 = \frac{1}{n} \|X_j^{(c)}\|^2$ — выборочная дисперсия.

Матрица из стандартизованных признаков $\mathbb{X}^{(s)}$:

$$X_j^{(s)} = \frac{X_j - \bar{X}_j}{s_j} = \frac{X_j^{(c)}}{s_j}, \quad \frac{1}{n} \|X_j^{(s)}\|^2 = 1 \quad \forall j = 1, \dots, p.$$

- $\mathbb{S} = \frac{1}{n} (\mathbb{X}^{(c)})^T \mathbb{X}^{(c)}$ — выборочная ковариационная матрица.
- $\widehat{\text{Corr}}\xi = \frac{1}{n} (\mathbb{X}^{(s)})^T \mathbb{X}^{(s)}$ — выборочная корреляционная матрица.

3. Примеры данных до стандартизации и после

Зададимся вопросом, как выглядят данные до стандартизации и после неё?

1. Данные с ненулевой корреляцией:

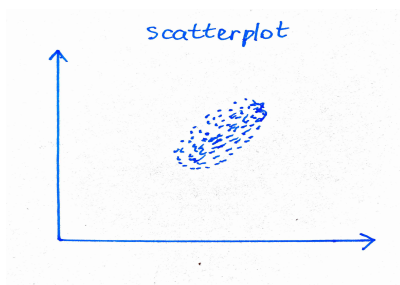


Рис. 1. Данные до стандартизации

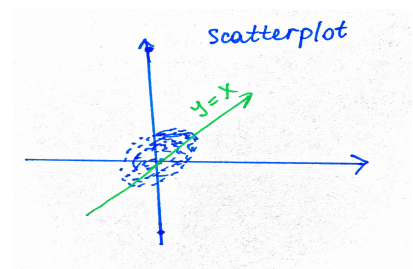


Рис. 2. Данные после стандартизации

¹ Посчитали среднее по каждому признаку и вычли. Таким образом, среднее по каждому признаку нулевое.

2. Данные с нулевой корреляцией:

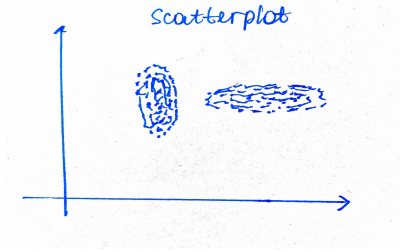


Рис. 3. Данные до стандартизации

После стандартизации — круг.

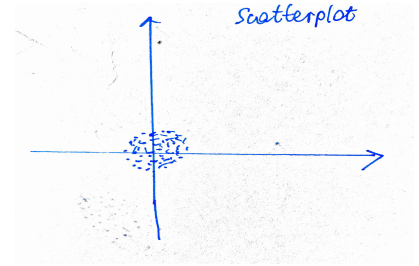


Рис. 4. Данные после стандартизации

4. Многомерное нормальное распределение

Рассмотрим $\xi \sim N_p(\mu, \Sigma)$, где $\mu = E\xi$ — среднее значение ξ , $\Sigma = \text{Cov}\xi$ — ковариационная матрица. Предположим, что Σ — невырожденная матрица.² Плотность многомерного нормального распределения задаётся формулой:

$$p(\mathbf{x}, \mu, \Sigma) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right\} \text{ для } \forall \mathbf{x} \in \mathbb{R}^p, \text{ где}$$

p — количество признаков. Очевидно, что в одномерном случае плотность будет иметь следующий вид:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}, \text{ где}$$

μ, σ — числа. Введём следующее понятие — *расстояние Махаланобиса*:

$$r_M^2(\mathbf{x}, \mu, \Sigma) = (\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu).$$

Отметим, что Σ в данном случае является неотрицательно определённой симметричной матрицей. Но так как ковариационная матрица удовлетворяет данному условию, то получаем, что если расстояния Махаланобиса совпадают, то и плотности многомерного нормального распределения тоже. Посмотрим, что будет являться расстоянием Махаланобиса в одномерном случае ($p = 1$):

$$\sqrt{\frac{(x - \mu)^2}{\sigma^2}} = \frac{|x - \mu|}{\sigma} = r_M(x, \mu, \sigma^2) \text{ — расстояние до мат.ож., измеренное в } \sigma.$$

² $|\Sigma| \neq 0$.

На выборочном языке:

$$\frac{|x_i - \bar{x}|}{s_i}.$$

Таким образом, если у нас есть нормальное распределение, то в одномерном случае мы можем измерять расстояние в сигмах, а в случае многомерного нормального распределения мы смотрим на линии уровня (точки, находящиеся на одном расстоянии Махаланобиса от центра).

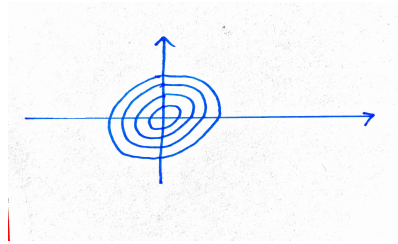


Рис. 5. Линии уровня

Ответим на следующий вопрос: *какое преобразование необходимо сделать со случайной величиной, имеющей нормальное распределение, чтобы из зависимости получить независимость?*

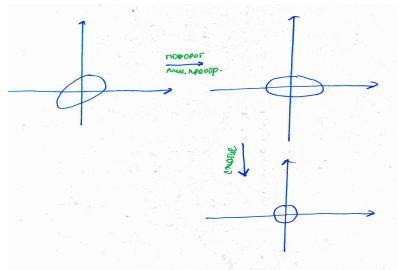


Рис. 6. Отбеливание

Сначала делаем поворот, а затем стандартизуем. Такое преобразование еще называется “отбеливанием”.

Пусть $\zeta = \Sigma^{-1/2}(\xi - \mu)$ — новая случайная величина, полученная из исходной линейным преобразованием. несложно сосчитать, что $E\zeta = \mathbb{O}$, так как

$$E\zeta = \Sigma^{-1/2}(E\xi - \mu) = \mathbb{O}.$$

Посчитаем ковариацию ζ :

$$\text{Cov}(\zeta) = \mathbb{E}\zeta\zeta^T = \Sigma^{-1/2}\text{Cov}(\xi)\Sigma^{-1/2} = \Sigma^{-1/2} \underbrace{\Sigma^1}_{\mathbb{E}\xi\xi^T} \Sigma^{-1/2} = \mathbb{I}.$$

Вспомним несколько свойств, которые характерны для нормально распределённых случайных величин:

1. Если $\xi \sim N$, то и всякое линейное преобразование ξ тоже имеет нормальное распределение.
2. Знаем, что из независимости следует некоррелированность. В случае нормального распределения верно и обратное. В общем случае это неверно. Приведём пример, когда зависимость есть, но корреляция равна 0.

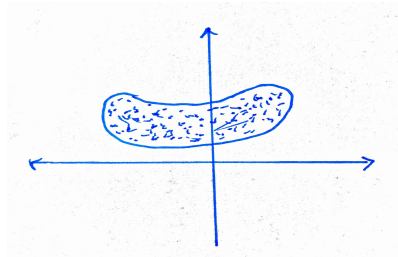


Рис. 7. Пример данных, когда зависимость есть, но корреляция нулевая

3. Условное математическое ожидание является линейной функцией.
4. Ковариационная матрица Σ может быть вырожденной.³

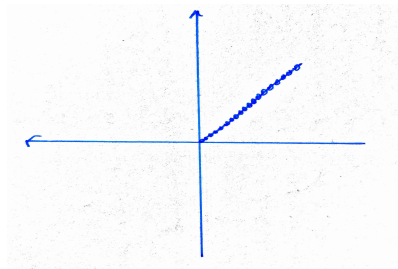


Рис. 8. Ковариационная матрица вырождена

³ Например, когда двумерное распределение сосредоточено на одной прямой.

5. Переход к новым признакам

Пусть ξ — p -мерная случайная величина, p — количество признаков. Считаем, что среднее равно нулю. Хотим образовать новый признак, который является линейной комбинацией старых признаков (смысл в том, чтобы сделать так, чтобы новый признак был в некотором смысле лучше, чем старые):

$$\eta = \mathbf{A}^T \xi = \xi^T \mathbf{A} = a_1 \xi_1 + \dots + a_p \xi_p, \quad \mathbf{A} \in \mathbb{R}^p,$$

$$D\eta = \mathbf{A}^T \Sigma \mathbf{A}.$$

Теперь пусть $\eta = (\eta_1, \dots, \eta_d)^T \in \mathbb{R}^d$ — d -мерная случайная величина, т.е. d — число новых признаков, которая получается из исходной как $\eta = \mathbb{A}^T \xi$, где $\mathbb{A} \in \mathbb{R}^{p \times d}$. Соответствующая ковариационная матрица будет иметь вид $\text{Cov}(\eta) = \mathbb{A}^T \Sigma \mathbb{A}$.

На выборочном языке: $Z = \sum_{j=1}^p a_j X_j = \mathbb{X} \mathbf{A} \in \mathbb{R}^n$ или $\mathbb{Z} = \mathbb{X} \mathbb{A} \in \mathbb{R}^{p \times d}$, где $\mathbb{X} \in \mathbb{R}^{n \times p}$ — матрица данных.

1. Возьмём в качестве индивидов школьников, а в качестве признаков — оценки по четырём школьным предметам:

- X_1 — оценка по математике;
- X_2 — оценка по физике;
- X_3 — оценка по русскому;
- X_4 — оценка по литературе.

Сами по себе оценки мало несут информации про общее положение в школе или классе, поэтому давайте создадим новые признаки, которые будут более информативными. Пусть это будут

- $Z_1 = X_1 + X_2 + X_3 + X_4$ — сумма оценок по всем предметам;
- $Z_2 = X_1 + X_2 - (X_3 + X_4)$ — разность между оценками по естественным и гуманитарным наукам.

Выясним, как будет выглядеть матрица $\mathbb{A}$ в этом случае:

$$\mathbb{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \end{pmatrix} \in \mathbb{R}^{4 \times 2}.$$

2. Рассмотрим два признака: оценки по математике X_1 и физике X_2 .

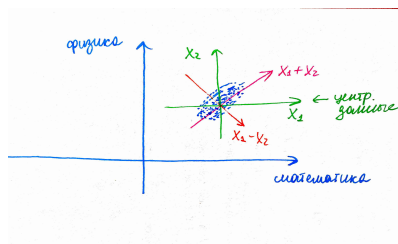
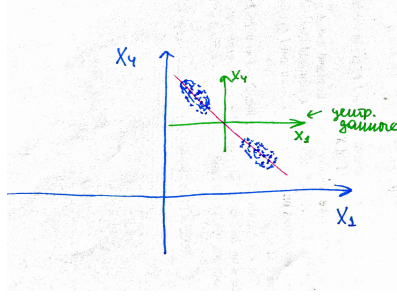


Рис. 9. Scatterplot: X_1 and X_2

В данной ситуации получается, что новый признак $X_1 + X_2$ лучше всего характеризует данные, то есть это та характеристика, по которой ученики максимально различаются; второй признак — $X_1 - X_2$, по нему ученики отличаются меньше. В анализе главных компонент (АГК) необходимо выбрать признак, который бы наилучшим образом характеризовал данные. Такой характеристикой будет тот признак, по которому данные максимально различаются.

3. Рассмотрим два признака: оценки по математике X_1 и литературе X_4 . Максимальное отличие получаем по разности данных признаков.⁴
4. Возможно, что данные неоднородные. Тогда новый признак строится так, чтобы по нему группы максимально различались. Например, если взять математиков и филологов, то данные образуют две группы.

⁴ Нарисуйте scatterplot в этом случае

Рис. 10. Scatterplot: X_1 and X_4

5.1. Две интерпретации новых признаков. Факторные нагрузки

Обсудили, зачем необходимы новые признаки. Также до этого мы требовали, чтобы новые признаки были линейной комбинацией старых. Теперь добавим еще условие, что:

- на выборочном языке: $Z_i \perp Z_j$ (вектора-признаки ортогональны).
- на генеральном языке: $\rho(\eta_i, \eta_j) = 0$ (случайные величины-признаки некоррелированы).

Зависимость признаков по смыслу означает, что часть одного признака входит в другой, то есть получается некоторая избыточность. Поэтому и требуется независимость (в данном случае, более слабое условие — некоррелированность).

Пусть $\text{rank}(\mathbb{X}) = d$, $(Z_1, \dots, Z_d)$ — ортогональный базис в подпространстве размерности d — $\text{span}(X_1, \dots, X_p) = \text{colspan}(\mathbb{X})$. Превратим базис в *ортонормированный базис*: $Q_i = \frac{Z_i}{\|Z_i\|}$. В ортонормированном базисе удобно вычислять координаты вектора через скалярные произведения с базисными векторами. Если $U_1, \dots, U_d$ — ортономированный базис в $\mathbb{R}^d$, то $\forall A \in \mathbb{R}^d$ раскладывается по ортонормированному базису:

$$A = \sum_{i=1}^d \langle A, U_i \rangle U_i, \text{ где } \langle A, U_i \rangle - i\text{-ая координата вектора } A \text{ в базисе } \{U_j\}_{j=1}^d.$$

⁵ Таким образом, $\{Q_i\}_{i=1}^d$ — ортонормированный базис в пространстве $\text{span}(X_1, \dots, X_p)$.

Новые признаки на самом деле интерпретируются двумя способами:

⁵ Рассмотрим пример: пространство $\mathbb{R}^2$, соответствующий базис $(1, 1)^T/\sqrt{2}$, $(1, -1)^T/\sqrt{2}$. Это ортонормированный базис, да? Как вычислить координаты вектора $(5, 4)^T$ в данном базисе?

- Новые признаки задаются равенством (новые признаки есть линейная комбинация старых):

$$\mathbb{X}\mathbb{A} = \mathbb{Z}.$$

- Исходные признаки выражаются через новые (есть пространство, в нём базис, каждый вектор этого пространства раскладывается по базису):

$$\forall i = 1, \dots, p \quad X_i = \sum_{j=1}^d \langle X_i, Q_j \rangle Q_j. \quad (1)$$

Определение 1. $f_{ij} = \langle X_i, Q_j \rangle$ — факторные веса, факторные нагрузки.