Генерация и визуализация трёхмерных нормальных подвыборок с различной структурой ковариационной матрицы

Елисеев Данила, 2025, ИС

26 декабря 2025 г.

1. Теоретическая часть

Случайный вектор

$$\mathbf{X} = (X_1, X_2, X_3)^\top \sim \mathcal{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

характеризуется вектором средних $\boldsymbol{\mu} \in \mathbb{R}^3$ и симметричной положительно определённой ковариационной матрицей $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times 3}$. Структура $\boldsymbol{\Sigma}$ полностью определяет форму распределения: размеры, ориентацию и степень вытянутости эллипсоидов постоянной плотности.

Если $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, его плотность:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}. \tag{1}$$

Множества постоянной плотности — эллипсоиды, задаваемые уровневыми уравнениями

$$(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2. \tag{2}$$

Оси этого эллипсоида направлены вдоль собственных векторов $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ матрицы $\boldsymbol{\Sigma}$, а длины полуосей равны $c\sqrt{\lambda_1}, c\sqrt{\lambda_2}, c\sqrt{\lambda_3}$, где λ_i — соответствующие собственные значения.

Для нормального распределения 95%-эллипсоид (т.е. содержащий 95% вероятностной массы) соответствует уровню $c^2 = \chi^2_{3;0.95} \approx 7.815$, откуда $c = \sqrt{7.815} \approx 2.795$.

i-я главная компонента (ГК) популяции имеет вид:

$$Y_i = \mathbf{e}_i^\top \mathbf{X}, \quad \mathrm{Var}(Y_i) = \lambda_i, \quad \mathrm{Cov}(Y_i, Y_j) = 0 \ (i \neq j).$$

Таким образом, ГК — это проекции на оси эллипсоида рассеяния. Форма и ориентация облака напрямую отражают спектр $\{\lambda_i\}$ и собственные векторы.

2. Описание подвыборок

Генерируются три подвыборки объёма $n = 100$ каждая из $\mathcal{N}_3(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$:

2.1. Подвыборка 1: Сферическое распределение

$\boldsymbol{\mu}_1 = (0, 0, 0)^\top$,

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{pmatrix}.$$

Диагональная матрица, т.е. X_1, X_2, X_3 независимы. Все дисперсии равны единице, что даёт сферическое распределение.

2.2. Подвыборка 2: Вытянутое распределение

$\boldsymbol{\mu}_2 = (5, 5, 5)^\top$,

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 5.0 \end{pmatrix}.$$

Диагональная матрица с различными дисперсиями. Переменные независимы, но наибольший разброс по оси Z (дисперсия равна 5), что приводит к вытянутому эллипсоиду вдоль оси Z.

2.3. Подвыборка 3: Коррелированное распределение

$\boldsymbol{\mu}_3 = (10, 0, 0)^\top$,

$$\boldsymbol{\Sigma}_3 = \begin{pmatrix} 1.0 & 0.8 & 0.8 \\ 0.8 & 1.0 & 0.8 \\ 0.8 & 0.8 & 1.0 \end{pmatrix}.$$

Полная взаимная корреляция всех компонент. Корреляция между всеми парами переменных равна 0.8, что приводит к повороту эллипсоида в пространстве. Обратим внимание, что такая матрица положительно определена (все собственные значения положительны).

3. Алгоритм генерации

1. Устанавливается `set.seed(123)` для воспроизводимости.
2. С помощью функции `mvrnorm()` из пакета `MASS` генерируются подвыборки.
3. Вычисляется спектральное разложение для каждой ковариационной матрицы:

   $$\boldsymbol{\Sigma}_k = \mathbf{U}_k \, \mathrm{diag}(\lambda_1^{(k)}, \lambda_2^{(k)}, \lambda_3^{(k)}) \mathbf{U}_k^\top,$$

   где $\mathbf{U}_k$ — матрица собственных векторов, $\lambda_i^{(k)}$ — собственные значения.
4. Строится 3D-диаграмма рассеяния с наложенными 95%-эллипсоидами рассеяния.

4. Результаты спектрального разложения

Для каждой подвыборки вычислены собственные значения и собственные векторы ковариационной матрицы. Результаты представлены в таблице 1.

Для матрицы $\mathbf{\Sigma}_3$ с корреляцией $\rho = 0.8$ между всеми парами переменных собственные значения можно вычислить аналитически. Характеристическое уравнение:

$$\det(\mathbf{\Sigma}_3 - \lambda \mathbf{I}) = 0.$$

Для матрицы вида

$$\begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$$

собственные значения: $\lambda_1 = 1 + 2\rho = 2.6$, $\lambda_2 = \lambda_3 = 1 - \rho = 0.2$.

Таблица 1: Собственные значения ковариационных матриц

Подвыборка	λ_1	λ_2	λ_3
1 (Сферическая)	1.0000	1.0000	1.0000
2 (Вытянутая)	5.0000	1.0000	1.0000
3 (Коррелированная)	2.6000	0.2000	0.2000

Доли объяснённой дисперсии главными компонентами:

- Подвыборка 1: $\frac{1.0}{3.0} = 33.3\%$ (все компоненты равнозначны)
- Подвыборка 2: $\frac{5.0}{7.0} \approx 71.4\%$ (первая ГК), $\frac{1.0}{7.0} \approx 14.3\%$ (вторая и третья ГК)
- Подвыборка 3: $\frac{2.6}{3.0} \approx 86.7\%$ (первая ГК), $\frac{0.2}{3.0} \approx 6.7\%$ (вторая и третья ГК)

Для подвыборки 3 первая главная компонента объясняет более 86% дисперсии, что указывает на сильную линейную зависимость между переменными.

5. Визуализация

На рис. 1 показаны три подвыборки:

- красные точки и полупрозрачный красный эллипсоид — подвыборка 1;
- синие точки и синий эллипсоид — подвыборка 2;
- тёмно-зелёные точки и зелёный эллипсоид — подвыборка 3.

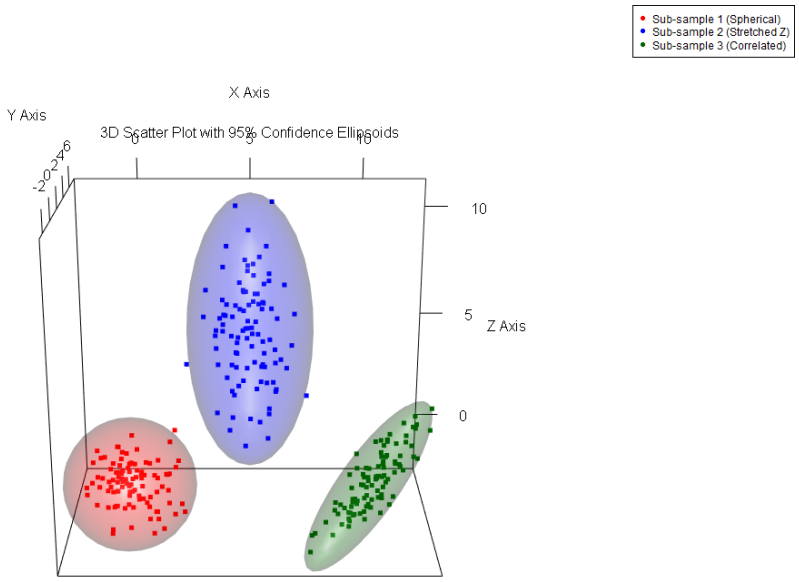

Рис. 1: 3D-облака и 95%-эллипсоиды рассеяния

5.1. Подвыборка 1 (диагональная Σ_1)

- Поскольку Σ_1 диагональна и имеет равные элементы, главные направления совпадают с координатными осями.
- Эллипсоид имеет сферическую форму: все полуоси равны $c\sqrt{1.0} \approx 2.80$.
- Отсутствие наклона — прямое следствие нулевых ковариаций $s_{12} = s_{13} = s_{23} = 0$.
- Для нормального распределения некоррелированность $\Leftrightarrow$ независимость.

5.2. Подвыборка 2 (вытянутая по оси Z)

- Диагональная матрица с различными дисперсиями.
- Эллипсоид вытянут вдоль оси Z: $\lambda_1 = 5.0$, полуось длиной $c\sqrt{5.0} \approx 6.25$.
- По осям X и Y полуоси равны $c\sqrt{1.0} \approx 2.80$.
- Переменные независимы, но масштабы различны.

5.3. Подвыборка 3 (общая корреляция)

- Собственные векторы уже не совпадают ни с одной координатной осью.
- Ориентация эллипсоида — произвольная в $\mathbb{R}^3$; его оси образуют базис ГК.
- Длины полуосей: $c\sqrt{2.6} \approx 4.52$, $c\sqrt{0.2} \approx 1.25$, $c\sqrt{0.2} \approx 1.25$, где $c \approx 2.795$.

- Большое различие между $\lambda_1 = 2.6$ и $\lambda_2 = \lambda_3 = 0.2$ говорит о сильной вытянутости вдоль первого главного направления.
- Первый собственный вектор для такой матрицы имеет вид $\mathbf{e}_1 = \frac{1}{\sqrt{3}}(1,1,1)^\top$ — направление максимальной дисперсии совпадает с главной диагональю пространства.

Этот случай наиболее сложен для интуитивного восприятия — именно он демонстрирует необходимость PCA: проекция на главные оси позволяет получить некоррелированные координаты Y_1, Y_2, Y_3, в которых распределение «распрямляется».

6. Выводы

Ковариационная матрица $\mathbf{\Sigma}$ полностью управляет геометрией нормального облака точек:

- диагональные элементы $\Rightarrow$ масштаб по осям;
- недиагональные элементы $\Rightarrow$ поворот и сдвиг в пространстве.

Наличие корреляций ведёт к повороту эллипсоида рассеяния в подпространствах, соответствующих парам связанных переменных. Геометрическая интерпретация главных компонент подтверждается экспериментально: главные направления $\Leftrightarrow$ оси эллипсоидов.

Визуализация в 3D (особенно с наложенными эллипсоидами) является эффективным инструментом для диагностики структуры данных и проверки адекватности многомерных моделей.

7. Приложение: Код на R

```
# Основной код находится в файле task1.1.r
# Код включает:
# - Генерацию данных с помощью mvrnorm()
# - Спектральное разложение ковариационных матриц
# - Построение 3D графиков с эллипсоидами
# - Анализ главных компонент
```