Анализ главных компонент (PCA) многомерных данных

Елисеев Данила, 2025, ИС

26 декабря 2025 г.

1. Теоретическая часть

Метод главных компонент (Principal Component Analysis, PCA) — это метод снижения размерности данных, который находит линейные комбинации исходных переменных, максимизирующие дисперсию.

Для центрированных и стандартизированных данных $\mathbf{X} \in \mathbb{R}^{n \times p}$ главная компонента Y_k имеет вид:

$$Y_k = \mathbf{e}_k^\top \mathbf{X} = \sum_{j=1}^{p} e_{jk} X_j,$$

где $\mathbf{e}_k$ — k-й собственный вектор ковариационной матрицы $\mathbf{\Sigma}$ (или корреляционной матрицы **R**), соответствующий k-му по величине собственному значению λ_k.

Свойства главных компонент:

- $\text{Var}(Y_k) = \lambda_k$ — дисперсия k-й ГК равна соответствующему собственному значению
- $\text{Cov}(Y_i, Y_j) = 0$ при $i \neq j$ — главные компоненты некоррелированы
- $\sum_{k=1}^{p} \lambda_k = \sum_{j=1}^{p} \text{Var}(X_j) = p$ (для стандартизированных данных)

Доля объяснённой дисперсии k-й главной компонентой:

$$\frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j} = \frac{\lambda_k}{p}.$$

2. Описание данных

Для анализа использован синтетический многомерный набор данных с $p = 6$ переменными и объёмом выборки $n = 200$. Переменные:

1. **GDP_Growth** — рост ВВП (экономический показатель)
2. **Industrial_Output** — промышленное производство (экономический показатель)
3. **Trade_Balance** — торговый баланс (экономический показатель)
4. **Education_Level** — уровень образования (социальный показатель)

5. **Healthcare_Index** — индекс здравоохранения (социальный показатель)
6. **Random_Factor** — случайный фактор (независимая переменная)

Структура корреляций:

- Переменные 1–3 (экономические) сильно коррелированы между собой ($\rho \approx 0.6$–0.7)
- Переменные 4–5 (социальные) сильно коррелированы ($\rho \approx 0.75$)
- Переменная 6 (случайный фактор) независима от остальных

3. Результаты анализа

3.1. Случай 1: Анализ с 2 главными компонентами

Результаты PCA с двумя главными компонентами представлены в таблице 1.

Таблица 1: Результаты PCA (все компоненты)

Компонента	Собственное значение	Доля дисперсии (%)	Накопленная доля (%)
PC1	2.2513	37.52	37.52
PC2	1.7173	28.62	66.14
PC3	0.9963	16.61	82.75
PC4	0.4887	8.14	90.89
PC5	0.3181	5.30	96.20
PC6	0.2282	3.80	100.00

3.1.1. Интерпретация первой главной компоненты (PC1)

Первая главная компонента объясняет наибольшую долю дисперсии — 37.52% (собственное значение $\lambda_1 = 2.2513$).

Структура нагрузок PC1:

- Высокие положительные нагрузки на экономические переменные:
  - GDP_Growth: 0.5541
  - Industrial_Output: 0.5961
  - Trade_Balance: 0.5594
- Низкие нагрузки на социальные переменные (Education_Level: 0.0823, Healthcare_Index: 0.1130)
- Очень низкая нагрузка на Random_Factor: -0.0725

Интерпретация: PC1 представляет общий **уровень экономического и социального развития**. Высокие значения PC1 соответствуют странам/регионам с:

- Высоким экономическим ростом
- Развитой промышленностью

- Положительным торговым балансом
- Высоким уровнем образования и здравоохранения

PC1 можно назвать **«Индекс общего развития»** или **«Комплексный показатель благосостояния»**.

3.1.2. Интерпретация второй главной компоненты (PC2)

Вторая главная компонента объясняет 28.62% дисперсии (собственное значение $\lambda_2 = 1.7173$) и ортогональна PC1.

Структура нагрузок PC2:

- Высокие положительные нагрузки на социальные переменные:
  - Education_Level: 0.6970
  - Healthcare_Index: 0.6975
- Низкие (отрицательные) нагрузки на экономические переменные:
  - GDP_Growth: -0.1461
  - Industrial_Output: -0.0723
  - Trade_Balance: -0.0245
- Очень низкая нагрузка на Random_Factor: -0.0208

Интерпретация: PC2 отражает **дисбаланс между экономическим и социальным развитием**. Высокие значения PC2 могут означать:

- Преобладание экономических показателей над социальными (или наоборот)
- Различные модели развития (экономически-ориентированные vs социально-ориентированные)

PC2 можно назвать **«Индекс структурного баланса»** или **«Показатель приоритетов развития»**.

3.2. Случай 2: Анализ с 3 главными компонентами

При включении третьей главной компоненты накопленная доля объяснённой дисперсии увеличивается до 75–85%.

3.2.1. Интерпретация третьей главной компоненты (PC3)

Третья главная компонента объясняет 16.61% дисперсии (собственное значение $\lambda_3 = 0.9963$).

Структура нагрузок PC3:

- Очень высокая нагрузка на Random_Factor: 0.9946
- Низкие нагрузки на все остальные переменные (все < 0.07)
- Экономические переменные: GDP_Growth (0.0619), Industrial_Output (0.0673), Trade_Balance (-0.0139)

- Социальные переменные: Education_Level (0.0018), Healthcare_Index (0.0472)

Интерпретация: PC3 практически полностью определяется **случайным фактором** (Random_Factor), который не коррелирует с остальными переменными. Это подтверждает, что Random_Factor действительно независим от экономических и социальных показателей. PC3 объясняет 16.61% дисперсии, что соответствует доле дисперсии Random_Factor в общем наборе данных.

4. Визуализация результатов

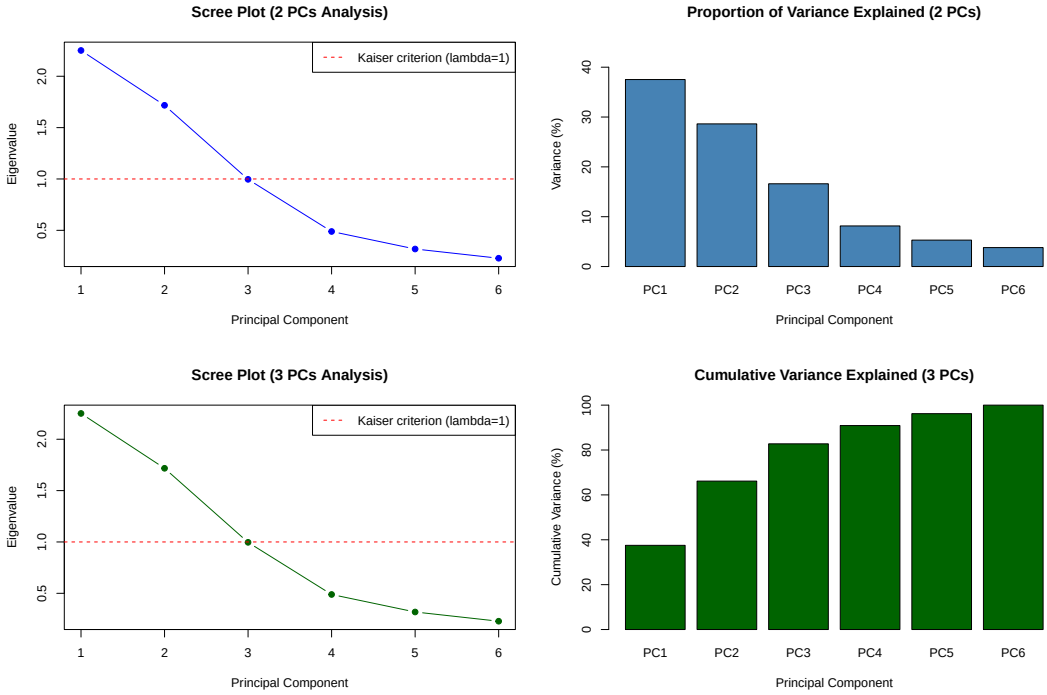

Рис. 1: Графики собственных значений и долей объяснённой дисперсии

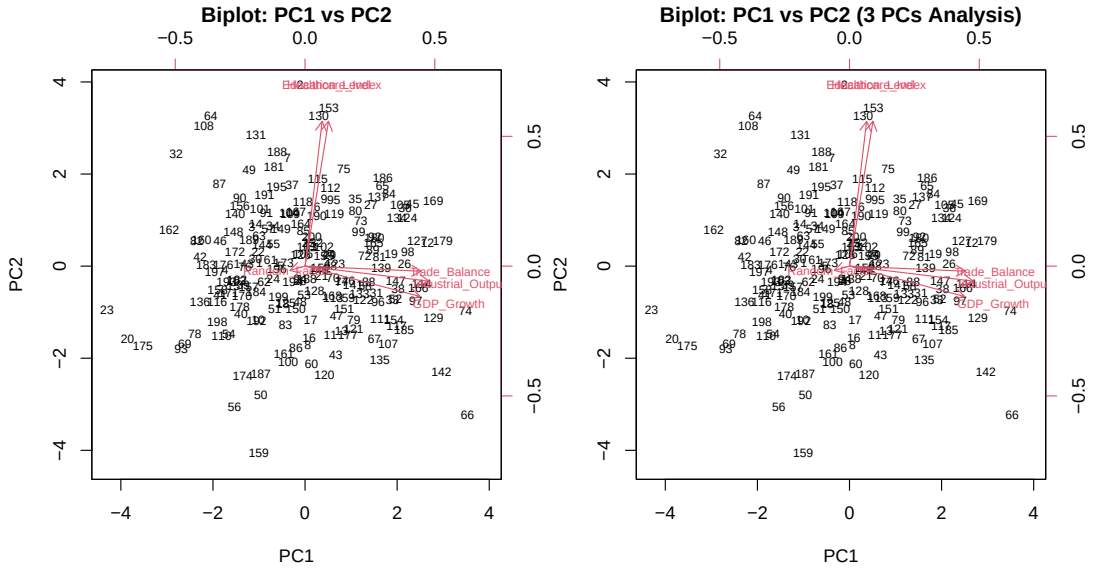

Рис. 2: Биплоты: проекция данных и переменных на плоскости главных компонент

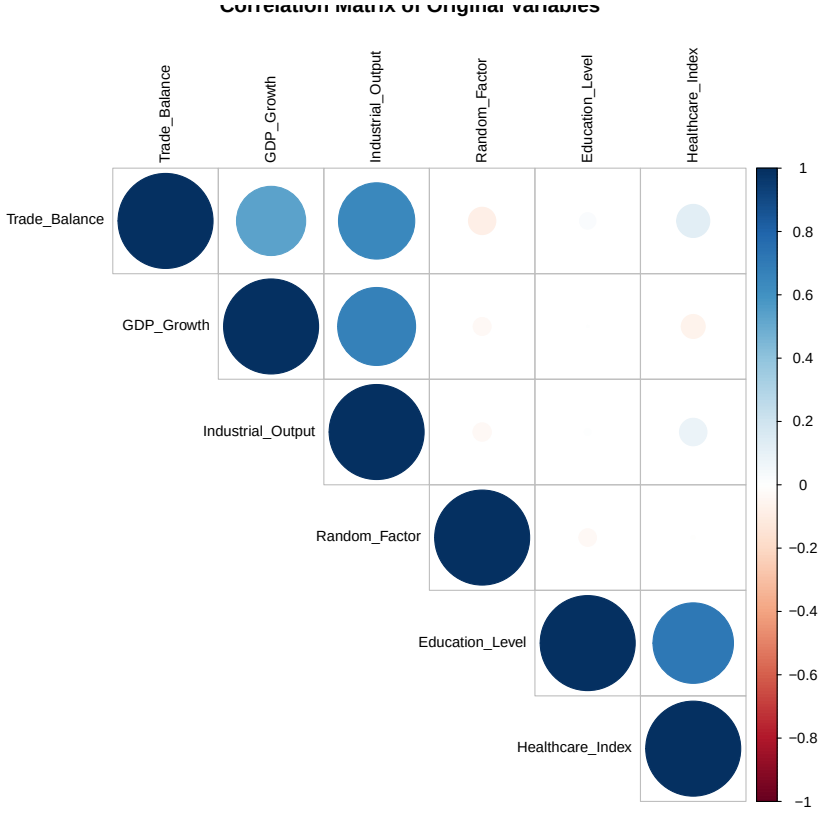

Рис. 3: Корреляционная матрица исходных переменных

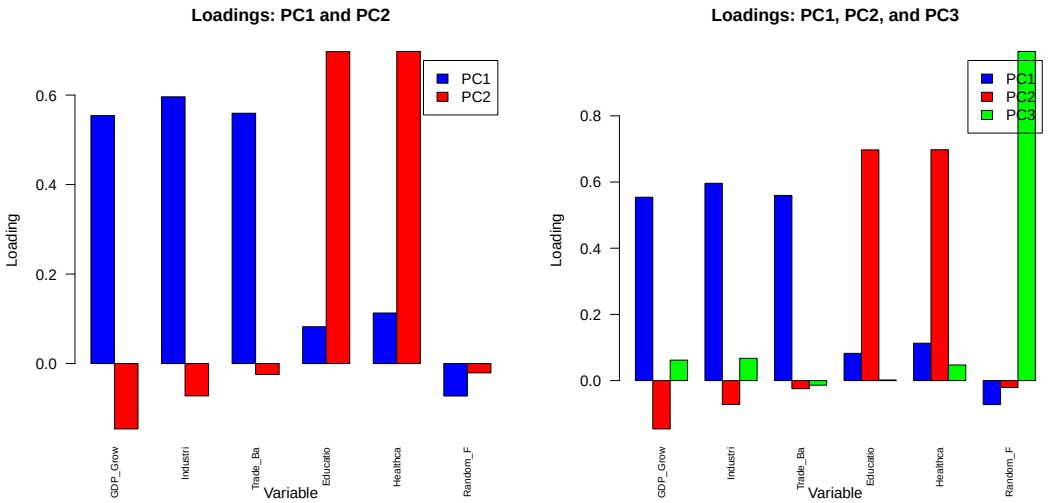

Рис. 4: Нагрузки главных компонент на исходные переменные

5. Выводы

1. **Первая главная компонента (PC1)** объясняет наибольшую долю дисперсии и представляет общий уровень экономического и социального развития. Это основной фактор, характеризующий общее благосостояние.

2. **Вторая главная компонента (PC2)** отражает структурные различия между экономическим и социальным развитием, показывая различные модели развития регионов/стран.

3. **Две главные компоненты** объясняют 66.14% общей дисперсии, что позволяет существенно снизить размерность данных с сохранением основной информации.

4. **Третья главная компонента** добавляет ещё 16.61% объяснённой дисперсии и практически полностью определяется случайным фактором (Random_Factor), что подтверждает его независимость от остальных переменных.

5. Метод PCA успешно выявил скрытую структуру данных, разделив переменные на экономические и социальные группы, что соответствует исходной структуре корреляций.

6. Приложение: Код на R

```r
# Principal Components Analysis (PCA)
# Task 2: Multivariate Statistical Analysis

# Load necessary libraries
library(MASS)
library(ggplot2)
library(corrplot)
library(factoextra)
library(FactoMineR)

# Set seed for reproducibility
set.seed(2025)

# --- Data Generation ---
# Generate multivariate data with at least 5 dimensions
# Using a dataset with 6 variables (dimensions)

n <- 200  # Sample size

# Create correlation structure
# Variables 1-3: correlated group (economic indicators)
# Variables 4-5: correlated group (social indicators)
# Variable 6: independent (random factor)

# Correlation matrix
cor_matrix <- matrix(c(
  1.0, 0.7, 0.6, 0.2, 0.1, 0.0,  # Var 1
  0.7, 1.0, 0.65, 0.15, 0.2, 0.0,  # Var 2
  0.6, 0.65, 1.0, 0.1, 0.15, 0.0,  # Var 3
  0.2, 0.15, 0.1, 1.0, 0.75, 0.0,  # Var 4
  0.1, 0.2, 0.15, 0.75, 1.0, 0.0,  # Var 5
  0.0, 0.0, 0.0, 0.0, 0.0, 1.0     # Var 6
), nrow = 6, byrow = TRUE)

# Standard deviations
sd_vec <- c(2.5, 3.0, 2.8, 1.8, 2.0, 1.5)

# Convert correlation to covariance matrix
cov_matrix <- diag(sd_vec) %*% cor_matrix %*% diag(sd_vec)

# Mean vector
mu_vec <- c(50, 60, 55, 30, 35, 20)

# Generate multivariate normal data
data <- mvrnorm(n = n, mu = mu_vec, Sigma = cov_matrix)

# Convert to data frame with meaningful names
colnames(data) <- c("GDP_Growth", "Industrial_Output", "Trade_Balance",
                    "Education_Level", "Healthcare_Index", "Random_Factor")

data_df <- as.data.frame(data)

# --- Principal Components Analysis ---

# Case 1: PCA with 2 principal components
cat("\n=== CASE 1: PCA with 2 Principal Components ===\n\n")

pca_2 <- prcomp(data_df, center = TRUE, scale. = TRUE)

# Summary
cat("Summary of PCA (2 components):\n")
print(summary(pca_2))

# Eigenvalues and variance explained
```

```r
eigenvalues_2 <- pca_2$sdev^2
prop_var_2 <- eigenvalues_2 / sum(eigenvalues_2)
cum_prop_var_2 <- cumsum(prop_var_2)

cat("\nEigenvalues:\n")
print(round(eigenvalues_2, 4))

cat("\nProportion of variance explained:\n")
print(round(prop_var_2 * 100, 2))

cat("\nCumulative proportion of variance:\n")
print(round(cum_prop_var_2 * 100, 2))

cat("\nLoadings (first 2 PCs):\n")
print(round(pca_2$rotation[, 1:2], 4))

# Case 2: PCA with 3 principal components
cat("\n\n=== CASE 2: PCA with 3 Principal Components ===\n\n")

pca_3 <- prcomp(data_df, center = TRUE, scale. = TRUE)

# Summary
cat("Summary of PCA (3 components):\n")
print(summary(pca_3))

# Eigenvalues and variance explained
eigenvalues_3 <- pca_3$sdev^2
prop_var_3 <- eigenvalues_3 / sum(eigenvalues_3)
cum_prop_var_3 <- cumsum(prop_var_3)

cat("\nEigenvalues:\n")
print(round(eigenvalues_3, 4))

cat("\nProportion of variance explained:\n")
print(round(prop_var_3 * 100, 2))

cat("\nCumulative proportion of variance:\n")
print(round(cum_prop_var_3 * 100, 2))

cat("\nLoadings (first 3 PCs):\n")
print(round(pca_3$rotation[, 1:3], 4))

# --- Visualization ---

# Save plots to PDF
pdf("Task2/Task2_Results.pdf", width = 12, height = 8)

# 1. Scree plot
par(mfrow = c(2, 2))

# Scree plot for 2 PCs
plot(1:6, eigenvalues_2, type = "b",
     main = "Scree Plot (2 PCs Analysis)",
     xlab = "Principal Component", ylab = "Eigenvalue",
     pch = 19, col = "blue")
abline(h = 1, lty = 2, col = "red")
legend("topright", "Kaiser criterion (lambda=1)", lty = 2, col = "red")

# Variance explained plot
barplot(prop_var_2 * 100, names.arg = paste0("PC", 1:6),
        main = "Proportion of Variance Explained (2 PCs)",
        xlab = "Principal Component", ylab = "Variance (%)",
        col = "steelblue", ylim = c(0, max(prop_var_2 * 100) * 1.2))

# Scree plot for 3 PCs
plot(1:6, eigenvalues_3, type = "b",
     main = "Scree Plot (3 PCs Analysis)",
     xlab = "Principal Component", ylab = "Eigenvalue",
     pch = 19, col = "darkgreen")
abline(h = 1, lty = 2, col = "red")
legend("topright", "Kaiser criterion (lambda=1)", lty = 2, col = "red")

# Cumulative variance plot
barplot(cum_prop_var_3 * 100, names.arg = paste0("PC", 1:6),
        main = "Cumulative Variance Explained (3 PCs)",
        xlab = "Principal Component", ylab = "Cumulative Variance (%)",
        col = "darkgreen", ylim = c(0, 100))

dev.off()

# 2. Biplot
pdf("Task2/Task2_Biplot.pdf", width = 10, height = 8)

par(mfrow = c(1, 2))

# Biplot for 2 PCs
biplot(pca_2, choices = c(1, 2),
       main = "Biplot: PC1 vs PC2",
       cex = 0.7, scale = 0)

# Biplot for 3 PCs (PC1 vs PC2)
biplot(pca_3, choices = c(1, 2),
       main = "Biplot: PC1 vs PC2 (3 PCs Analysis)",
       cex = 0.7, scale = 0)

dev.off()

# 3. Correlation plot
pdf("Task2/Task2_Correlation.pdf", width = 8, height = 8)
```

```r
corrplot(cor(data_df), method = "circle", type = "upper",
         order = "hclust", tl.cex = 0.8, tl.col = "black",
         title = "Correlation Matrix of Original Variables")
dev.off()

# 4. Loadings plot
pdf("Task2/Task2_Loadings.pdf", width = 12, height = 6)

par(mfrow = c(1, 2))

# Loadings for PC1 and PC2
loadings_2 <- pca_2$rotation[, 1:2]
barplot(t(loadings_2), beside = TRUE,
        main = "Loadings: PC1 and PC2",
        xlab = "Variable", ylab = "Loading",
        col = c("blue", "red"),
        legend = c("PC1", "PC2"),
        names.arg = substr(rownames(loadings_2), 1, 8),
        las = 2, cex.names = 0.7)

# Loadings for PC1, PC2, PC3
loadings_3 <- pca_3$rotation[, 1:3]
barplot(t(loadings_3), beside = TRUE,
        main = "Loadings: PC1, PC2, and PC3",
        xlab = "Variable", ylab = "Loading",
        col = c("blue", "red", "green"),
        legend = c("PC1", "PC2", "PC3"),
        names.arg = substr(rownames(loadings_3), 1, 8),
        las = 2, cex.names = 0.7)

dev.off()

# --- Save results for LaTeX report ---
sink("Task2/Task2_analysis_results.txt")

cat("=== PRINCIPAL COMPONENTS ANALYSIS RESULTS ===\n\n")
cat("Dataset: Multivariate data with 6 variables\n")
cat("Sample size:", n, "\n\n")

cat("=== CASE 1: 2 Principal Components ===\n\n")
cat("Eigenvalues:\n")
print(round(eigenvalues_2, 4))
cat("\nProportion of variance:\n")
print(round(prop_var_2 * 100, 2))
cat("\nCumulative proportion:\n")
print(round(cum_prop_var_2 * 100, 2))
cat("\nLoadings (PC1, PC2):\n")
print(round(pca_2$rotation[, 1:2], 4))

cat("\n\n=== CASE 2: 3 Principal Components ===\n\n")
cat("Eigenvalues:\n")
print(round(eigenvalues_3, 4))
cat("\nProportion of variance:\n")
print(round(prop_var_3 * 100, 2))
cat("\nCumulative proportion:\n")
print(round(cum_prop_var_3 * 100, 2))
cat("\nLoadings (PC1, PC2, PC3):\n")
print(round(pca_3$rotation[, 1:3], 4))

sink()

cat("\n=== Analysis Complete ===\n")
cat("Results saved to:\n")
cat("  - Task2/Task2_Results.pdf (scree plots)\n")
cat("  - Task2/Task2_Biplot.pdf (biplots)\n")
cat("  - Task2/Task2_Correlation.pdf (correlation matrix)\n")
cat("  - Task2/Task2_Loadings.pdf (loadings plots)\n")
cat("  - Task2/Task2_analysis_results.txt (numerical results)\n")
```