Метод главных компонент (МГК)

Многомерный анализ

Математические основы, разобранный пример, работа с инструментом

Русский

Вольфганг Бёме

wolfgang.boehme@economic-modeling.de

www.economic-modeling.de

8 ноября 2026 г.

Документация на этом сайте разработана при поддержке ИИ Claude (Anthropic).

Содержание

1 Введение и области применения **2**

2 Математические основы **2**
2.1 Матрица данных и стандартизация 2
2.2 Разложение матрицы корреляций по собственным значениям 2
2.3 Факторные нагрузки 3
2.4 Объяснённая дисперсия и критерий Кайзера 3
2.5 Баллы компонент (scores) 3
2.6 Коммунальности 3
2.7 Вращение Varimax 3

3 Разобранный пример: балансовые показатели **4**
3.1 Исходные данные ($n = 24$, $p = 6$) 4
3.2 Шаг 1: Описательные статистики 5
3.3 Шаг 2: Матрица корреляций $\mathbf{R}$ 6
3.4 Шаг 3: Собственные значения 6
3.5 Шаг 4: Собственные векторы и нагрузки 6
3.6 Шаг 5: Баллы компонент (выборка) 7

4 Работа с интерактивным инструментом МГК **8**
4.1 Ввод данных 8
4.2 Выбор параметров (левая панель) 8
4.3 Вкладки результатов 8
4.4 Типичный рабочий процесс 9

5 Сводка формул **9**

6 Рекомендуемая литература **9**

Введение и области применения

Метод главных компонент (англ. Principal Component Analysis, PCA) — это метод многомерной статистики для **снижения размерности**. Исходя из p коррелированных переменных, МГК порождает $k \leq p$ новых, некоррелированных переменных — **главных компонент** — которые объясняют максимальную долю общей дисперсии.

Типичные области применения:

- Построение индексов: объединение нескольких индикаторов в один сводный индекс
- Предобработка данных: устранение мультиколлинеарности перед регрессионным анализом
- Визуализация: отображение многомерных данных в двумерном биплоте
- Сжатие данных: уменьшение объёма хранения сигнальных данных
- Разведочный анализ: выявление скрытых структур в наблюдаемых данных

Математические основы

Матрица данных и стандартизация

Пусть задана матрица данных $\mathbf{X} \in \mathbb{R}^{n \times p}$ с n наблюдениями и p переменными. При различных шкалах измерения рекомендуется Z-преобразование:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \qquad \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \qquad s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} \tag{1}$$

После Z-преобразования матрица ковариации $\mathbf{S}$ совпадает с матрицей корреляций $\mathbf{R}$.

Разложение матрицы корреляций по собственным значениям

Симметричная, положительно полуопределённая матрица корреляций $\mathbf{R} \in \mathbb{R}^{p \times p}$ разлагается следующим образом:

$$\mathbf{R} = \mathbf{V}\,\mathbf{\Lambda}\,\mathbf{V}^{\top} \tag{2}$$

где $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, и ортонормированные собственные векторы $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$.

Численный расчёт выполняется в инструменте с помощью **метода Якоби** для симметричных матриц (циклический вариант). В каждом проходе (sweep) все внедиагональные элементы поочерёдно обнуляются вращениями Якоби, пока сумма квадратов внедиагональных элементов не станет меньше 10^{-12}.

Факторные нагрузки

$$l_{jk} = v_{jk} \cdot \sqrt{\lambda_k} \tag{3}$$

l_{jk} — это корреляция между j-й исходной переменной и k-й главной компонентой.

Объяснённая дисперсия и критерий Кайзера

$$_k = \frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j} \cdot 100\% \tag{4}$$

Критерий Кайзера сохраняет все компоненты с $\lambda_k \geq 1$: при стандартизованных данных компонента с $\lambda_k < 1$ объясняет меньше дисперсии, чем одна исходная переменная.

Баллы компонент (scores)

$$\mathbf{T} = \mathbf{Z}\,\mathbf{V}_k, \qquad \mathbf{V}_k \in \mathbb{R}^{p \times k} \tag{5}$$

Баллы — это проекции стандартизованных наблюдений на пространство главных компонент. Они попарно некоррелированы.

Коммунальности

$$h_j^2 = \sum_{k=1}^{K} l_{jk}^2 \tag{6}$$

$h_j^2 \in [0, 1]$ — доля дисперсии j-й переменной, объяснённой K компонентами.

Вращение Varimax

Вращение Varimax максимизирует дисперсию квадратов нагрузок по факторам:

$$\max \sum_{k=1}^{K} \left[\frac{1}{p} \sum_{j=1}^{p} l_{jk}^4 - \left(\frac{1}{p} \sum_{j=1}^{p} l_{jk}^2 \right)^2 \right] \tag{7}$$

Вращение ортогонально: коммунальности и общая объяснённая дисперсия остаются неизменными, однако доли дисперсии перераспределяются между компонентами. Структура нагрузок становится более интерпретируемой (простая структура).

Разобранный пример: балансовые показатели

Исходные данные ($n = 24$, $p = 6$)

Этот пример загружен по умолчанию во всех четырёх калькуляторах набора и разбирается во всех четырёх руководствах. Речь идёт о **синтетических** балансовых показателях 24 предприятий (U01–U24), все в процентах:

- **Устойчивость:** X_1 доля собственного капитала (СК/валюта баланса), X_2 ликвидность 2-й степени ((денежные средства + краткосрочная дебиторская задолженность)/краткосрочные обязательства), X_3 покрытие внеоборотных активов II ((СК + долгосрочный заёмный капитал)/внеоборотные активы)
- **Доходность:** X_4 рентабельность продаж, X_5 рентабельность активов, X_6 рост выручки к предыдущему году

Данные сгенерированы так, что образуются три типа предприятий по восемь в каждом: устойчивые предприятия с крепкой капитальной базой (рейтинг A), высокодоходные, но более закредитованные растущие предприятия (рейтинг B) и предприятия в зоне риска (рейтинг C). Рейтинг используется как групповая переменная только в дискриминантном анализе (Руководство 4); МГК, факторный и кластерный анализ работают только с шестью показателями.

Цель МГК: свести частично сильно коррелированные показатели к нескольким некоррелированным главным компонентам.

Таблица 1: Балансовые показатели 24 предприятий (в %, синтетические)

№	Рейтинг	X_1 Д. СК	X_2 Ликв.	X_3 Покр. II	X_4 Р. прод.	X_5 Р. акт.	X_6 Рост
U01	A	46.6	127	124	4.3	8.1	9.0
U02	A	49.4	141	127	3.4	7.0	5.5
U03	A	32.0	109	100	4.8	5.0	4.5
U04	A	33.7	134	129	5.7	4.8	2.7
U05	A	51.1	120	122	6.4	5.8	2.7
U06	A	41.2	143	144	3.1	9.3	5.0
U07	A	50.5	117	125	8.8	9.5	10.1
U08	A	56.6	130	130	5.6	8.6	2.1
U09	B	38.3	119	86	7.7	12.2	4.3
U10	B	40.8	114	98	8.5	10.7	15.0
U11	B	35.5	102	99	7.0	10.9	8.7
U12	B	39.0	128	108	11.8	11.9	13.5
U13	B	20.2	65	93	8.1	11.9	18.0
U14	B	33.2	107	97	8.3	11.9	11.1
U15	B	31.5	82	100	7.7	12.9	11.0
U16	B	34.5	78	93	9.3	6.9	12.1
U17	C	16.5	86	81	1.1	8.0	5.6
U18	C	26.6	74	103	1.8	4.7	-3.0
U19	C	33.1	82	98	1.7	2.5	-3.4
U20	C	23.5	80	72	1.6	7.9	-2.1
U21	C	32.5	58	82	2.3	2.3	2.7
U22	C	23.2	96	85	2.8	3.2	-3.6
U23	C	34.3	101	95	1.6	5.6	-1.9
U24	C	34.3	98	95	2.1	2.6	2.8

Шаг 1: Описательные статистики

Таблица 2: Средние значения и стандартные отклонения

	X_1	X_2	X_3	X_4	X_5	X_6
$\bar{x}$	35.754	103.792	103.583	5.229	7.675	5.517
s	10.074	24.270	18.547	3.084	3.379	6.083

Стандартные отклонения различаются почти на порядок (ликвидность и рентабельность продаж). Без Z-преобразования ликвидность и покрытие активов доминировали бы в анализе, поэтому стандартизация здесь обязательна. Пример U01: $z_{11} = (46.6 - 35.754)/10.074 = 1.077$.

Шаг 2: Матрица корреляций R

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.710 & 0.743 & 0.311 & 0.122 & 0.113 \\ 0.710 & 1.000 & 0.771 & 0.233 & 0.225 & 0.079 \\ 0.743 & 0.771 & 1.000 & 0.172 & 0.087 & 0.108 \\ 0.311 & 0.233 & 0.172 & 1.000 & 0.699 & 0.793 \\ 0.122 & 0.225 & 0.087 & 0.699 & 1.000 & 0.727 \\ 0.113 & 0.079 & 0.108 & 0.793 & 0.727 & 1.000 \end{pmatrix} \tag{8}$$

Матрица распадается на два блока: внутри показателей устойчивости (X_1–X_3) и внутри показателей доходности (X_4–X_6) корреляции составляют 0.70–0.79, между блоками — не более 0.31. Поэтому следует ожидать **две** главные компоненты.

Шаг 3: Собственные значения

Таблица 3: Собственные значения и объяснённая дисперсия

Компонента	Собств. знач. λ	Доля дисп.	Накопл.	Кайзер
PC1	2.971	49.52 %	49.52 %	≥ 1
PC2	2.010	33.50 %	83.01 %	≥ 1
PC3	0.397	6.61 %	89.62 %	< 1
PC4	0.289	4.81 %	94.43 %	< 1
PC5	0.200	3.34 %	97.77 %	< 1
PC6	0.134	2.23 %	100.00 %	< 1

Критерий Кайзера: **2 компоненты**, вместе объясняющие 83.01 % дисперсии. График осыпи также показывает чёткий излом после PC2 (2.010 → 0.397). Сумма собственных значений равна $p = 6$; например, $2.971/6 = 49.52\,\%$.

Шаг 4: Собственные векторы и нагрузки

Первые два собственных вектора (нормированные):

$$\mathbf{v}_1 = (\ 0.416,\ \ 0.418,\ \ 0.397,\ 0.442,\ 0.389,\ 0.384)^\top$$
$$\mathbf{v}_2 = (-0.384,\ -0.391,\ -0.436,\ 0.362,\ 0.410,\ 0.459)^\top$$

Нагрузки по формуле $l_{jk} = v_{jk}\sqrt{\lambda_k}$, например $l_{11} = 0.416 \cdot \sqrt{2.971} = 0.717$ и $l_{12} = -0.384 \cdot \sqrt{2.010} = -0.544$.

Таблица 4: Матрица нагрузок: без вращения и после вращения Varimax

	без вращения		Varimax			
Переменная	PC1	PC2	RC1	RC2	h_j^2	$1-h_j^2$
X_1 Доля СК	0.717	−0.544	**0.893**	0.115	0.810	0.190
X_2 Ликвидность	0.720	−0.555	**0.902**	0.109	0.826	0.174
X_3 Покрытие активов	0.684	−0.617	**0.921**	0.040	0.850	0.150
X_4 Рент. продаж	0.762	0.513	0.184	**0.900**	0.844	0.156
X_5 Рент. активов	0.671	0.581	0.071	**0.885**	0.789	0.211
X_6 Рост выручки	0.662	0.651	0.015	**0.928**	0.862	0.138
Объясн. дисп.	49.5 %	33.5 %	41.6 %	41.4 %	Σ 83.0 %	

Пример коммунальности: $h_1^2 = 0.717^2 + (-0.544)^2 = 0.810$ — столько же и для решения после вращения ($0.893^2 + 0.115^2 = 0.810$). Решение с вращением получают в инструменте, выбрав *Вращение* = **Varimax**.

Интерпретация:

- **Без вращения:** PC1 положительно нагружает все шесть показателей (0.66–0.76) — общий фактор «совокупной финансовой силы». PC2 биполярна: отрицательна для показателей устойчивости и положительна для показателей доходности. Она описывает «профиль» предприятия между капитальной базой и рентабельностью.
- **Varimax:** после вращения возникает простая структура. RC1 = **устойчивость** (X_1–X_3, нагрузки 0.89–0.92), RC2 = **доходность** (X_4–X_6, нагрузки 0.89–0.93). Объяснённая дисперсия распределяется почти поровну (41.6 % и 41.4 %), сумма 83.0 % не меняется.

Шаг 5: Баллы компонент (выборка)

$$t_{i1} = 0.416\,z_{i1} + 0.418\,z_{i2} + 0.397\,z_{i3} + 0.442\,z_{i4} + 0.389\,z_{i5} + 0.384\,z_{i6}$$

Таблица 5: Баллы выбранных предприятий (решение без вращения)

Предприятие	Рейтинг	Балл PC1	Балл PC2
U08	A	1.822	−1.939
U07	A	2.307	−0.291
U03	A	−0.576	−0.309
U12	B	2.578	1.268
U13	B	0.150	3.258
U21	C	−2.602	0.161
U19	C	−2.270	−1.131

На биплоте (PC1 vs. PC2) три рейтинговые группы образуют хорошо различимые облака: у предприятий C все баллы PC1 отрицательны (низкая совокупная сила), предприятия A находятся преимущественно в области отрицательных, а предприятия B — положительных баллов PC2. Выделяются два предприятия A: **U03** с

$t_1 = -0.576$ ближе к предприятиям C, а **U07** с высокой рентабельностью имеет почти профиль B. Оба снова появляются как пограничные случаи в кластерном и дискриминантном анализе (Руководства 3 и 4).

Работа с интерактивным инструментом МГК

Ввод данных

Вкладка »Вручную«
Установить количество строк и столбцов. Ввести имена столбцов в строку заголовков; числовые значения — в поля ввода. При изменении размеров введённые данные сохраняются.

Вкладка »CSV«
Загрузить CSV-файл кликом или перетаскиванием. Первая строка должна содержать имена столбцов. Разделители: запятая, точка с запятой и табуляция распознаются автоматически.

Вкладка »Пример«
Доступны четыре подготовленных датасета: балансовые показатели (24 предприятия, 6 пер., загружен по умолчанию), Iris (15 набл., 4 пер.), экономические индикаторы (12 стран, 5 пер.), удовлетворённость трудом (20 чел., 6 пунктов). Для каждого примера над таблицей ввода появляется пояснительный блок (данные, цель анализа, ожидаемый результат); он исчезает при нажатии «Очистить» или загрузке CSV-файла.

Выбор параметров (левая панель)

Стандартизация
Z-преобразование (рекомендуется): обязательно при разных единицах измерения.
Без стандартизации: использовать матрицу ковариации (только при одинаковых единицах).

Количество компонент
Автоматически: критерий Кайзера ($\lambda \geq 1$). Или фиксированное количество 2–5.

Вращение
Без вращения: невращённое решение. *Varimax:* ортогональное вращение при $k \geq 2$ компонентах.

Порог нагрузок
Нагрузки $\geq |0.3|$ выделяются (по умолчанию). Выделение можно отключить флажком.

Вкладки результатов

Обзор
Карточки показателей (n, p, число компонент, объяснённая дисп.) + полная таблица собственных значений с маркером Кайзера.

Нагрузки
Матрица факторных нагрузок с цветовым выделением, коммунальности и объяснённая дисперсия по компонентам.

Графики
График осыпи: собственные значения с пунктирной линией Кайзера. *Накоплен. дисп.:* столбчатый и линейный графики. *Биплот:* баллы (PC1 vs. PC2) со стрелками нагрузок.

Баллы
Таблица всех баллов компонент по наблюдениям.

Интерпретация
Автоматически сформированный текст с оценкой качества.

Экспорт
CSV: нагрузки, баллы, собственные значения. PNG: график осыпи, биплот.

Типичный рабочий процесс

1. Загрузить данные или выбрать пример
2. Проверить стандартизацию (Z-преобразование при разных единицах)
3. Оставить количество факторов *Автоматически*
4. Нажать »**Вычислить МГК**«
5. Проверить график осыпи: найти «локоть»
6. Интерпретировать таблицу нагрузок: какие переменные определяют каждую компоненту?
7. При $k \geq 2$: активировать вращение Varimax и сравнить структуру нагрузок
8. При необходимости экспортировать баллы в CSV

Важные замечания:
- Требуется минимум $n = 3$ полных наблюдения и $p = 2$ переменных.
- Строки с пропущенными значениями исключаются автоматически.
- МГК неустойчив к выбросам — предварительно проверяйте данные.
- Практическое правило: $n \geq 5 \cdot p$ для стабильных результатов.
- Биплот требует минимум $k = 2$ извлечённых компонент.

Сводка формул

Рекомендуемая литература

- Jolliffe, I. T. (2002). *Principal Component Analysis*, 2nd ed. Springer.
- Backhaus, K. et al. (2021). *Multivariate Analysemethoden*, 16. Aufl. Springer Gabler.

Таблица 6: Формулы МГК

Обозначение	**Формула**
Z-преобразование	$z_{ij} = (x_{ij} - \bar{x}_j)/s_j$
Коэфф. корреляции	$r_{jk} = \frac{1}{n-1}\sum_i z_{ij} z_{ik}$
Задача на собств. знач.	$\mathbf{R}\mathbf{v} = \lambda \mathbf{v}$
Факторная нагрузка	$l_{jk} = v_{jk}\sqrt{\lambda_k}$
Объяснённая дисперсия	$\lambda_k / \sum_j \lambda_j \cdot 100\,\%$
Коммунальность	$h_j^2 = \sum_{k=1}^{K} l_{jk}^2$
Балл компоненты	$t_{ik} = \sum_{j=1}^{p} z_{ij} v_{jk}$
Критерий Кайзера	$\lambda_k \geq 1$

- Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler*. Springer.