# Линейный дискриминантный анализ (LDA)

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 Матрицы разброса . . . 2
2.2 Дискриминантные функции . . . 2
2.3 Дискриминантные баллы . . . 3
2.4 Стандартизованные коэффициенты . . . 3
2.5 Каноническая корреляция . . . 3
2.6 Лямбда Уилкса . . . 4
2.7 Тест Бартлетта на значимость . . . 4
2.8 Кросс-валидация LOO (Leave-One-Out) . . . 4

**3 Разобранный пример: балансовые показатели** **4**
3.1 Исходные данные . . . 4
3.2 Шаг 1: Статистики групп . . . 5
3.3 Шаг 2: Дискриминантные функции . . . 5
3.4 Шаг 3: Стандартизованные дискриминантные коэффициенты . . . 5
3.5 Шаг 4: Центроиды групп в дискриминантном пространстве . . . 6
3.6 Шаг 5: Классификация и кросс-валидация . . . 6

**4 Работа с интерактивным инструментом LDA** **7**
4.1 Ввод данных . . . 7
4.2 Выбор параметров . . . 7
4.3 Вкладки результатов . . . 7
4.4 Типичный рабочий процесс . . . 8

**5 Сводка формул** **9**

**6 Рекомендуемая литература** **9**

## Введение и области применения

**Линейный дискриминантный анализ** (LDA) — это метод классификации *с учителем*: в отличие от кластерного анализа, принадлежность наблюдений к группам известна заранее. LDA находит **дискриминантные функции** — линейные комбинации переменных — которые наилучшим образом разделяют группы.

Двойная цель LDA:

- **Дескриптивная:** какие переменные способствуют разделению групп?
- **Классификационная:** как отнести новое наблюдение к группе?

**Типичные области применения:**

- Оценка кредитоспособности (хороший/плохой заёмщик)
- Медицинская диагностика (больной/здоровый)
- Сегментация рынка с известными целевыми группами
- Контроль качества (годная/бракованная деталь)
- Определение вида в биологии и археологии

## Математические основы

### Матрицы разброса

Пусть $\mathbf{X} \in \mathbb{R}^{n \times p}$ с $g$ известными группами, $n_i$ наблюдений в каждой группе и центроид группы $\bar{\mathbf{x}}_i$.

**Внутригрупповая матрица разброса $\mathbf{S}_W$:**

$$\mathbf{S}_W = \sum_{i=1}^{g} \sum_{j \in C_i} (\mathbf{x}_j - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_i)^\top \tag{1}$$

**Межгрупповая матрица разброса $\mathbf{S}_B$:**

$$\mathbf{S}_B = \sum_{i=1}^{g} n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^\top \tag{2}$$

### Дискриминантные функции

LDA максимизирует критерий отношения Релея:

$$\max_{\mathbf{w}} \frac{\mathbf{w}^\top \mathbf{S}_B \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_W \mathbf{w}} \tag{3}$$

Это сводится к обобщённой задаче на собственные значения:

$$\mathbf{S}_B\,\mathbf{w} = \lambda\,\mathbf{S}_W\,\mathbf{w} \quad\Longleftrightarrow\quad \mathbf{S}_W^{-1}\mathbf{S}_B\,\mathbf{w} = \lambda\,\mathbf{w} \tag{4}$$

Решения $\mathbf{w}_1, \mathbf{w}_2, \ldots$ — это **дискриминантные коэффициенты** функций DF1, DF2, …

Поскольку матрица $\mathbf{S}_W^{-1}\mathbf{S}_B$ несимметрична, инструмент решает задачу через разложение Холецкого $\mathbf{S}_W = \mathbf{L}\mathbf{L}^\top$: симметричная матрица $\mathbf{C} = \mathbf{L}^{-1}\mathbf{S}_B\mathbf{L}^{-\top}$ имеет те же собственные значения $\lambda_k$; из её собственных векторов $\mathbf{u}_k$ (метод Якоби) получаем $\mathbf{w}_k = \sqrt{n-g}\,\mathbf{L}^{-\top}\mathbf{u}_k$. При такой нормировке дискриминантные баллы имеют внутригрупповую (объединённую) дисперсию 1.

Максимальное число значимых дискриминантных функций: $\min(p,\ g-1)$.

## Дискриминантные баллы

$$d_{ik} = \mathbf{w}_k^\top \mathbf{x}_i = \sum_{j=1}^{p} w_{kj}\,x_{ij} \tag{5}$$

**Правило классификации:** наблюдение $\mathbf{x}_i$ относится к группе $g^*$, центроид $\bar{d}_g$ которой в дискриминантном пространстве ближайший (благодаря нормировке евклидово расстояние совпадает с расстоянием Махаланобиса), с поправкой на априорные вероятности $\pi_g$:

$$g^* = \arg\min_{g} \Big[ \sum_{k=1}^{K} (d_{ik} - \bar{d}_{gk})^2 - 2\ln \pi_g \Big] \tag{6}$$

При равных априорных вероятностях ($\pi_g = 1/g$) поправка исчезает; в противном случае $\pi_g = n_g/n$.

## Стандартизованные коэффициенты

$$w_{kj}^{\mathrm{std}} = w_{kj} \cdot s_j^{(W)}, \qquad s_j^{(W)} = \sqrt{\frac{(\mathbf{S}_W)_{jj}}{n-g}} \tag{7}$$

Умножение на объединённое внутригрупповое стандартное отклонение $s_j^{(W)}$ делает коэффициенты сопоставимыми независимо от шкалы измерения.

## Каноническая корреляция

$$r_k^* = \sqrt{\frac{\lambda_k}{1+\lambda_k}} \tag{8}$$

$r_k^*$ — корреляция между дискриминантными баллами DF$k$ и индикаторными переменными групп — мера силы разделения.

### Лямбда Уилкса

Лямбда Уилкса $\Lambda$ — центральный критерий качества LDA:

$$\Lambda = \prod_{k=1}^{K} \frac{1}{1+\lambda_k} = \frac{|\mathbf{S}_W|}{|\mathbf{S}_W + \mathbf{S}_B|} \tag{9}$$

$\Lambda = 0$: идеальное разделение; $\Lambda = 1$: разделения нет.

### Тест Бартлетта на значимость

$$\chi^2 \approx -\left(n - 1 - \frac{p+g}{2}\right) \ln \Lambda, \qquad \mathrm{df} = p(g-1) \tag{10}$$

### Кросс-валидация LOO (Leave-One-Out)

Для несмещённой оценки качества классификации:

1. Исключить наблюдение $i$ из обучающей выборки
2. Оценить LDA на оставшихся $n-1$ наблюдениях
3. Классифицировать наблюдение $i$
4. Повторить для всех $n$ наблюдений

## Разобранный пример: балансовые показатели

### Исходные данные

Используются те же 24 синтетических предприятия, что и в Руководстве 1 (полная таблица данных приведена там), с шестью показателями: $X_1$ доля собственного капитала, $X_2$ ликвидность 2-й степени, $X_3$ покрытие внеоборотных активов II, $X_4$ рентабельность продаж, $X_5$ рентабельность активов, $X_6$ рост выручки (все в %). В отличие от Руководств 1–3 здесь используется **известная принадлежность к группам**: $g = 3$ рейтинговых класса A, B, C по $n_g = 8$ предприятий (U01–U08: A, U09–U16: B, U17–U24: C). Расчёт с z-стандартизованными переменными и равными априорными вероятностями (по умолчанию); средние на шаге 1 приведены в исходных единицах, как в инструменте.

**Цель дискриминантного анализа:** определить, какие показатели лучше всего разделяют рейтинговые классы, и найти правило отнесения новых предприятий к классу.

### Шаг 1: Статистики групп

Таблица 1: Средние значения по группам (в %)

| Группа | $X_1$ Д. СК | $X_2$ Ликв. | $X_3$ Покр. | $X_4$ Р. пр. | $X_5$ Р. акт. | $X_6$ Рост |
|---|---|---|---|---|---|---|
| Рейтинг A ($n = 8$) | 45.14 | 127.63 | 125.13 | 5.26 | 7.26 | 5.20 |
| Рейтинг B ($n = 8$) | 34.13 | 99.38 | 96.75 | 8.55 | 11.16 | 11.71 |
| Рейтинг C ($n = 8$) | 28.00 | 84.38 | 88.88 | 1.88 | 4.60 | −0.36 |
| Итого ($n = 24$) | 35.75 | 103.79 | 103.58 | 5.23 | 7.68 | 5.52 |

Предприятия A отличаются высокой устойчивостью, предприятия B — высокой доходностью; у предприятий C все показатели самые низкие.

### Шаг 2: Дискриминантные функции

Число дискриминантных функций: $\min(p,\, g-1) = \min(6,\, 2) = 2$.

Таблица 2: Собственные решения и объяснённая дисперсия

| Функция | Собств. знач. $\lambda$ | $r_k^*$ | Дисп. | Накопл. |
|---|---|---|---|---|
| DF1 | 8.003 | 0.943 | 71.8 % | 71.8 % |
| DF2 | 3.146 | 0.871 | 28.2 % | 100.0 % |

Пример: $r_1^* = \sqrt{8.003/9.003} = 0.943$; доля дисперсии DF1 $= 8.003/(8.003+3.146) = 71.8\,\%$.

Лямбда Уилкса:

$$\Lambda = \frac{1}{1+8.003} \cdot \frac{1}{1+3.146} = 0.1111 \cdot 0.2412 = 0.0268 \quad \text{(очень хорошее разделение)},$$

тест Бартлетта: $\chi^2 = -\big(24 - 1 - \tfrac{6+3}{2}\big) \ln 0.0268 = 18.5 \cdot 3.620 = 66.96$, $\mathrm{df} = 6 \cdot 2 = 12$, $p < 0.001$.

### Шаг 3: Стандартизованные дискриминантные коэффициенты

Таблица 3: Стандартизованные коэффициенты

| Переменная | DF1 | DF2 |
|---|---|---|
| $X_1$ Доля СК | −0.038 | 0.193 |
| $X_2$ Ликвидность | 0.160 | 0.342 |
| $X_3$ Покрытие активов | 0.069 | **0.726** |
| $X_4$ Рент. продаж | **0.748** | −0.184 |
| $X_5$ Рент. активов | **0.488** | −0.281 |
| $X_6$ Рост выручки | 0.339 | 0.149 |

**Интерпретация:**

- **DF1** (71.8 % разделения) — **функция доходности**: наибольшие веса у рентабельности продаж, рентабельности активов и роста.
- **DF2** (28.2 %) — **функция устойчивости**: доминируют покрытие активов и ликвидность.
- Доля собственного капитала имеет в обеих функциях лишь малые веса, хотя заметно различается между группами. Причина: она тесно коррелирует с ликвидностью и покрытием активов; стандартизованные коэффициенты измеряют вклад *сверх* остальных переменных.

Таким образом, дискриминантный анализ находит те же два измерения, что и МГК и факторный анализ, — но взвешенные так, чтобы рейтинговые классы разделялись максимально. Знаки дискриминантных функций произвольны; зеркальное отражение функции не меняет результатов.

### Шаг 4: Центроиды групп в дискриминантном пространстве

Таблица 4: Центроиды групп

| Группа | Центроид DF1 | Центроид DF2 |
|---|---|---|
| Рейтинг A | 0.22 | 2.34 |
| Рейтинг B | 3.13 | −1.29 |
| Рейтинг C | −3.35 | −1.05 |

DF1 отделяет высокодоходные предприятия B от предприятий C (расстояние 6.5), DF2 выделяет предприятия A с их высокой устойчивостью среди двух других групп. На графике групп (DF1 vs. DF2) три класса образуют чётко разделённые облака.

### Шаг 5: Классификация и кросс-валидация

Таблица 5: Матрица ошибок классификации (обучающие данные)

| | Прогноз | | |
|---|---|---|---|
| Факт | A | B | C |
| Рейтинг A | **8** | 0 | 0 |
| Рейтинг B | 0 | **8** | 0 |
| Рейтинг C | 0 | 0 | **8** |

**Точность (ресубституция): 100 %** (24 из 24 правильно).

Эта оценка слишком оптимистична, поскольку каждое предприятие классифицируется функцией, в оценке которой оно само участвовало. **Кросс-валидация leave-one-out** (включается флажком в инструменте) даёт **91.7 %** (22 из 24): два предприятия A классифицируются неверно, если они не участвуют в оценке:

- **U03** → рейтинг C: устойчивость и рентабельность ниже среднего (ср. кластерный анализ, Руководство 3).

- **U07** → рейтинг B: очень высокие показатели доходности (рентабельность продаж 8.8 %, рентабельность активов 9.5 %, рост 10.1 %) — профиль доходности, как у предприятия B.

Разница в 8.3 процентного пункта между ресубституцией и LOO-CV типична при $n = 24$ и $p = 6$ и показывает, почему кросс-валидация рекомендуется для малых выборок.

## Работа с интерактивным инструментом LDA

### Ввод данных

Особенность инструмента LDA: **столбец GRP** слева в таблице ручного ввода содержит метки групп (текст или числа, например «Setosa», «Хорошая», «1»). Каждая строка должна содержать метку группы, чтобы попасть в анализ.

По умолчанию загружены балансовые показатели (3 группы, рейтинг A/B/C, $n = 24$, $p = 6$). Дополнительные примеры: Iris (3 группы, $n = 15$, $p = 4$), кредитоспособность (2 группы, $n = 21$, $p = 4$) и рыночные сегменты (3 группы, $n = 24$, $p = 4$). 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

**Стандартизация**
: Z-преобразование рекомендуется. При одинаковых единицах измерения можно выбрать «Без стандартизации» — нестандартизованные коэффициенты тогда интерпретируются напрямую.

**Число дискриминантных функций**
: *Автоматически:* $\min(p, g-1)$ функций. Вручную: 1–3 функции (не более $\min(p, g-1)$); классификация и графики тогда используют только эти функции. Лямбда Уилкса всегда вычисляется по всем функциям.

**Априорные вероятности групп**
: Флажок установлен: равные априорные вероятности (все группы равновероятны). Стандарт и рекомендация, если априорные вероятности неизвестны. Флажок снят: априорные вероятности пропорциональны размерам групп ($n_g/n$).

**LOO-кросс-валидация**
: Флажок: вычислить точность LOO (требует больше ресурсов). Рекомендуется при малых выборках ($n < 50$).

### Вкладки результатов

**Обзор**
: Показатели (n, группы, дискр. функции, точность) + таблица лямбды Уилкса с оценкой значимости + таблица собственных значений ДФ с каноническими корреляциями + таблица средних по группам (в исходных единицах).

**Коэффициенты**
Стандартизованные и нестандартизованные дискриминантные коэффициенты по функциям. Центроиды групп в дискриминантном пространстве.

**Классификация**
Матрица ошибок с цветовой кодировкой (диагональ — зелёная, ошибки — красные) + общая точность + таблица классификации каждого наблюдения.

**Графики**
*График групп (DF1 vs. DF2):* диаграмма рассеяния всех наблюдений и центроидов групп (ромбы) в дискриминантном пространстве. *Столбцы дисперсии:* объяснённая дисперсия по дискр. функциям. *График коэффициентов:* столбчатая диаграмма стандартизованных абсолютных коэффициентов для DF1.

**Интерпретация**
Автоматический текст: лямбда Уилкса, тест Бартлетта, важнейшие дискриминаторы, сравнение LOO-CV.

**Экспорт**
CSV: коэффициенты, классификация, дискриминантные баллы. PNG: график групп, столбцы коэффициентов.

### Типичный рабочий процесс

1. Загрузить данные со столбцом GRP (в каждой группе $\geq 2$ наблюдений)
2. Оставить стандартизацию Z-преобразование
3. Активировать флажок LOO-CV (при $n < 100$)
4. Нажать **«Вычислить LDA»**
5. Вкладка «Обзор»: проверить $\Lambda$ Уилкса и p-значение Бартлетта
6. Вкладка «Коэффициенты»: какие переменные разделяют сильнее всего?
7. Вкладка «Классификация»: анализ матрицы ошибок — какие группы путаются?
8. Вкладка «Графики»: график групп — разделимы ли группы в пространстве DF1-DF2?
9. LOO-CV vs. обучение: большая разница указывает на переобучение

**Важные замечания:**
- Каждая группа должна содержать минимум 2 наблюдения.
- Практическое правило: $n \geq 5 \cdot p$ на группу для стабильного LDA.
- Предпосылки: многомерная нормальность и однородность матриц ковариации (гомоскедастичность) по группам.
- При нарушенной гомоскедастичности: квадратичный дискриминантный анализ (QDA).
- Точность в инструменте — это *субституционная* оценка, которая завышает истинную точность; LOO-CV даёт несмещённую оценку.

- Вырожденная $\mathbf{S}_W$ при $n < p$ или идеальной мультиколлинеарности: использовать больше наблюдений или меньше переменных.

## Сводка формул

Таблица 6: Формулы дискриминантного анализа

| **Обозначение** | **Формула** |
|---|---|
| Внутригрупп. матрица | $\mathbf{S}_W = \sum_i \sum_{j \in C_i} (\mathbf{x}_j - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_i)^\top$ |
| Межгрупповая матрица | $\mathbf{S}_B = \sum_i n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^\top$ |
| Задача на собств. знач. | $\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$ |
| Станд. коэффициент | $w_{kj} \sqrt{(\mathbf{S}_W)_{jj}/(n-g)}$ |
| Дискриминантный балл | $d_{ik} = \mathbf{w}_k^\top \mathbf{x}_i$ |
| Каноническая корреляция | $r_k^* = \sqrt{\lambda_k/(1+\lambda_k)}$ |
| Лямбда Уилкса | $\Lambda = \prod_k 1/(1+\lambda_k)$ |
| Бартлетт $\chi^2$ | $-(n-1-(p+g)/2)\,\ln\Lambda$, $\mathrm{df} = p(g-1)$ |
| Классификация | $g^* = \arg\min_g \left[\sum_k (d_{ik} - \bar{d}_{gk})^2 - 2\ln \pi_g\right]$ |

## Рекомендуемая литература


- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188.
- Backhaus, K. et al. (2021). *Multivariate Analysemethoden*, 16. Aufl. Springer Gabler.
- McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley.