

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
5.2. Вклад новых признаков	11
6. Пара фактов из линейной алгебры	12
 Глава 1. Сингулярное разложение	
(SVD — Singular Value Decomposition)	14
1.1. Как строится сингулярное разложение	14
1.2. Матричный вид сингулярного разложения	15
1.3. Единственность сингулярного разложения	16
1.4. Оптимальные свойства сингулярного разложения	17
 Глава 2. Анализ главных компонент (АГК)	
(PCA — principal component analysis)	18
2.1. Главные направления	18
2.2. Анализ главных компонент. Построение	19
2.2.1. Анализ главных компонент на выборочном языке	19
2.3. Связь между SVD и АГК. Общее и различия	19

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:

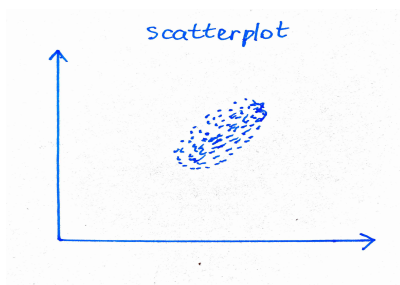


Рис. 1. Данные до стандартизации

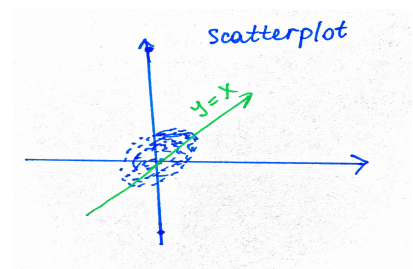


Рис. 2. Данные после стандартизации

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

2. Данные с нулевой корреляцией:

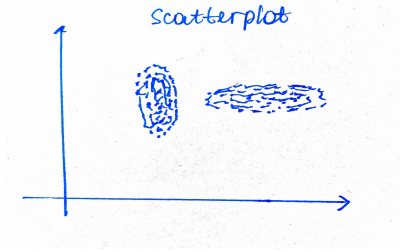


Рис. 3. Данные до стандартизации

После стандартизации — круг.

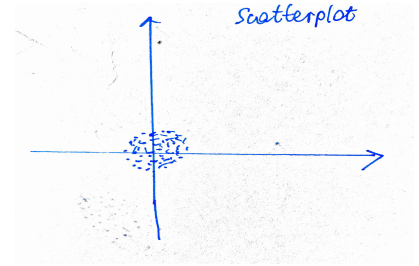


Рис. 4. Данные после стандартизации

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).

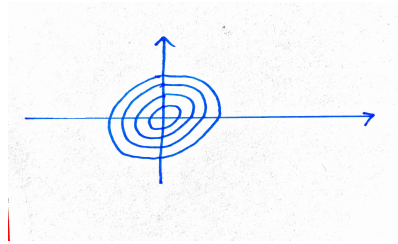


Рис. 5. Линии уровня

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

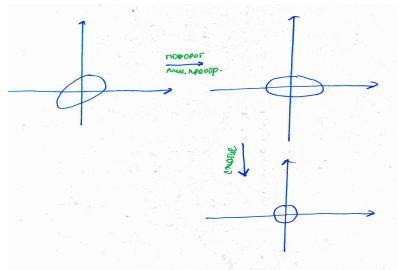


Рис. 6. Отбеливание

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.

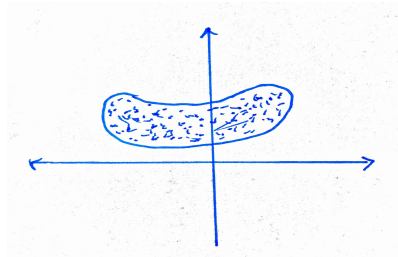


Рис. 7. Пример данных, когда зависимость есть, но корреляция нулевая

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

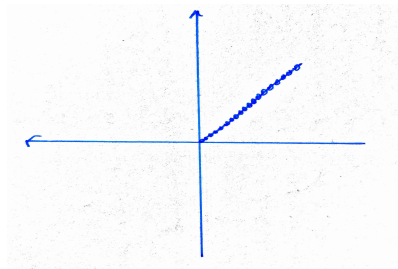


Рис. 8. Ковариационная матрица вырождена

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 .

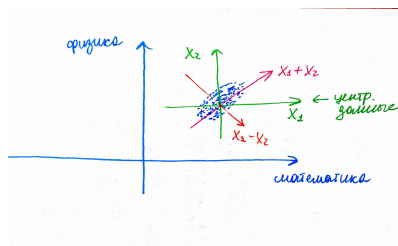
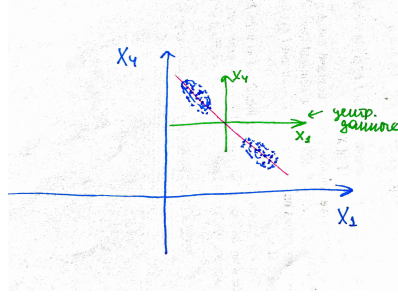


Рис. 9. Scatterplot: X_1 and X_2

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.

Составим матрицу $\mathbb{F} = \{f_{ij}\}_{i=1, j=1}^{p, d} = [F_1, \dots, F_d] \in \mathbb{R}^{p \times d}$ — коэффициенты разложения по ортонормированному базису, тогда можем записать разложение (1) в матричном виде:

$$\mathbb{X} = \sum_{j=1}^d Q_j F_j^T = \mathbb{Q}\mathbb{F}^T. \quad (2)$$

$\|Q_j\| = 1$, $\|F_j\| \neq 1$ в разложении (2). Давайте нормируем F . Пусть $\sigma_j = \|F_j\|$, $P_j = \frac{F_j}{\|F_j\|}$, тогда $\|Q_j\| = \|P_j\| = 1$ и

$$\mathbb{X} = \sum_{j=1}^d \sigma_j Q_j P_j^T = \mathbb{Q}\Sigma\mathbb{P}^T, \text{ где } \Sigma = \text{diag}\{\sigma_1, \dots, \sigma_d\}. \quad (3)$$

Что за матрица $Q_j P_j^T \in \mathbb{R}^{n \times p}$? Можно заметить, что $\text{rank}(Q_j P_j^T) = 1$; действительно, столбцы матрицы пропорциональны; то же самое для строк. Таким образом, у нас была исходная матрица ранга d , а мы превратили её в сумму d элементарных матриц ранга 1.

Пусть $\mathbb{X}^{(j)} = \sigma_j Q_j P_j^T$, тогда $\mathbb{X} = \sum_{j=1}^d \mathbb{X}^{(j)}$.

5.2. Вклад новых признаков

Ввели новый признак, теперь хотелось бы понять, какую важную часть информации он в себя включает. Введём скалярное произведение двух матриц, которой соответствует норма Фробениуса ($\|\mathbb{A}\|_F^2 = \sum_{ij} a_{ij}^2$):

$$(\mathbb{Y}, \mathbb{Z})_F = \sum_{i,j} y_{ij} z_{ij}.$$

Если верно $\|\mathbb{X}\|^2 = \left\| \sum_{j=1}^d \mathbb{X}^{(j)} \right\|_F^2 \stackrel{?}{=} \sum_{j=1}^d \|\mathbb{X}^{(j)}\|_F^2$, то вклад j -ого признака (вклад j -ой элементарной матрицы) равен $\frac{\|\mathbb{X}^{(j)}\|^2}{\|\mathbb{X}\|^2}$.⁶

Замечание 1. Для разложения (3), $\left\| \sum_{j=1}^d \mathbb{X}^{(j)} \right\|_F^2 = \sum_{j=1}^d \|\mathbb{X}^{(j)}\|_F^2$ (скалярное произведение равно 0, потому что $Q_i \perp Q_j$).

Так как Q_j и P_j нормированы, то можем продолжить равенство:

$$\|\mathbb{X}\|^2 = \left\| \sum_{j=1}^d \mathbb{X}^{(j)} \right\|_F^2 = \sum_{j=1}^d \|\mathbb{X}^{(j)}\|_F^2 = \sum_{j=1}^d \sigma_j^2.$$

Замечание 2. Пусть старые признаки X_i центрированные, тогда новые признаки Z_i и Q_i тоже центрированные.

Доказательство. Новые признаки — это линейная комбинация старых. Среднее линейной комбинации есть линейная комбинация средних по признакам. Средние по признакам равны 0, получаем центрированные новые признаки. $\square$

6. Пара фактов из линейной алгебры

1. Унитарная матрица $\mathbb{U}$ — ортогональная матрица в комплексном случае. Ее свойства:

- $\mathbb{U}$ — квадратная матрица: $\mathbb{U}^T = \mathbb{U}^{-1}$.
- Столбцы $\mathbb{U}$ ортонормированы.
- Строки $\mathbb{U}$ ортонормированы.⁷
- Умножение на матрицу $\mathbb{U}$ означает *поворот* или *отражение*.
- При умножении на матрицу $\mathbb{U}$ (и на $\mathbb{U}^T$) не меняются нормы векторов и углы между векторами. Пусть есть вектора Y, Z , после умножения на матрицу $\mathbb{U}$ получим $\tilde{Y} = \mathbb{U}Y, \tilde{Z} = \mathbb{U}Z$ и $\|Y\| = \|\tilde{Y}\|, \|Z\| = \|\tilde{Z}\|, \langle Y, Z \rangle = \langle \tilde{Y}, \tilde{Z} \rangle$ (почему это означает равенство углов между векторами?).

Пример 1. $\mathbb{U} = \begin{pmatrix} \cos \phi & \sin \phi \\ -\sin \phi & \cos \phi \end{pmatrix}$ — матрица поворота на угол ϕ .

⁶ Если отношение равно 1/2, то значит, что j -ый признак измеряет 50% всей информации.

⁷ 2 пункт эквивалентен 3. Почему? Если матрица ортогональная, то и транспонированная к ней тоже ортогональная (следует из пункта 1).

Обычно, унитарная матрица строится из ортонормированного базиса, который составляет столбцы матрицы. (Проверьте, что в примере с поворотом на ϕ это так.)

2. Пусть $\{P_i\}_{i=1}^r$ — система независимых векторов, рассмотрим линейную оболочку $\mathcal{L}_r = \text{span}\{P_1, \dots, P_r\}$ в $\mathbb{R}^L$, $\Pi : \mathbb{R}^L \rightarrow \mathcal{L}_r$ — ортогональный проектор на $\mathcal{L}_r$ (он сопоставляет вектору ближайшую точку из подпространства, что делается опусканием перпендикуляра). Матрица Π :

$$\Pi = \mathbb{P}(\mathbb{P}^T \mathbb{P})^{-1} \mathbb{P}^T.$$

Пусть $\{P_i\}_{i=1}^r$ — ортонормированный базис $\mathcal{L}_r$, т.е.

$$\mathbb{P}^T \mathbb{P} = \mathbb{I}_{r \times r} = \begin{pmatrix} 1 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{r \times r}.$$

Тогда $\Pi = \mathbb{P} \mathbb{P}^T$.

Сингулярное разложение (SVD — Singular Value Decomposition)

1.1. Как строится сингулярное разложение

Пусть L — число признаков, K — количество индивидов, $\mathbb{Y} = \mathbb{X}^T \in \mathbb{R}^{L \times K}$ — ненулевая матрица. Обозначим $\mathbb{S} = \mathbb{Y}\mathbb{Y}^T \in \mathbb{R}^{L \times L}$ — симметричная неотрицательно определённая матрица. По определению,

$$\mathbb{S}U_i = \lambda_i U_i, \text{ где}$$

$\{U_i\}_{i=1}^L$ — ортонормированный набор из собственных векторов матрицы $\mathbb{S}$,
 $\lambda_1 \geq \dots \geq \lambda_L \geq 0$ — собственные числа матрицы $\mathbb{S}$.¹

Пусть $d = \text{rank } \mathbb{Y} = \text{colrank } \mathbb{Y} = \text{rowrank } \mathbb{Y}$. Знаем, что $d \leq \min(L, K)$.

Предложение 1. 1. $d = \text{rank } \mathbb{Y}\mathbb{Y}^T$.

2. $\lambda_d > 0$; $\lambda_i = 0$ при $i > d$.²

3. $\{U_i\}_{i=1}^d$ образуют ортонормированный базис $\text{colspan } \mathbb{Y}$.

Введём вектор

$$V_i \stackrel{\text{def}}{=} \frac{\mathbb{Y}^T U_i}{\sqrt{\lambda_i}} \in \mathbb{R}^K, \quad i = 1, \dots, d.$$

Предложение 2. 1. $\{V_i\}_{i=1}^d$ — ортонормированная система векторов.

2. V_i — собственные вектора $\mathbb{Y}^T \mathbb{Y}$, соответствующие тем же собственным числам λ_i . Остальные собственные вектора $\mathbb{Y}^T \mathbb{Y}$ соответствуют нулевым собственным числам.

3. $U_i = \frac{\mathbb{Y}V_i}{\sqrt{\lambda_i}}.$

4. $\mathbb{Y} = \sum_{i=1}^d \sqrt{\lambda_i} U_i V_i^T$ — SVD (Сингулярное разложение матрицы).³⁴

¹ Неотрицательные, так как матрица $\mathbb{S}$ неотрицательно определена.

² Упорядочили собственные числа: первые d строго положительные, а остальные все нули.

³ Разложение в сумму элементарных матриц.

⁴ Самый важный пункт утверждения.

Что здесь считать новыми признаками, если $\mathbb{Y} = \mathbb{X}^T$? V_i ,⁵ так как $U_i \in \mathbb{R}^L$, $V_i \in \mathbb{R}^K$.

- U_i — ортонормированный базис в пространстве столбцов.
- V_i — ортонормированный базис в пространстве строк.⁶
- $\frac{\lambda_i}{\sum_{i=1}^d \lambda_i}$ — вклад i -ого признака.

Определение 2. $\sqrt{\lambda_i}$ — сингулярные числа матрицы $\mathbb{Y}$, U_i — левые сингулярные вектора, V_i — правые сингулярные вектора.

Тройка $(\sqrt{\lambda_i}, U_i, V_i)$ называется i -ой собственной тройкой сингулярного разложения.

Замечание 3. Сингулярное разложение — единственное разложение с двумя ортонормированными базисами.

1.2. Матричный вид сингулярного разложения

Можно записать двумя способами:

1. Введём $\mathbb{U}_d = [U_1 : \dots : U_d]$, $\mathbb{V}_d = [V_1 : \dots : V_d]$, $\Lambda_d = \text{diag}(\lambda_1, \dots, \lambda_d)$. Тогда

$$\mathbb{Y} = \mathbb{U}_d \Lambda_d^{1/2} \mathbb{V}_d^T.$$

2. Возьмём $\mathbb{U} = [U_1 : \dots : U_d : U_{d+1} : \dots : U_L]$ — ортонормированный базис в $\mathbb{R}^L$.⁷

$\mathbb{V}^T = [V_1 : \dots : V_d : V_{d+1} : \dots : V_K]$ — ортонормированный базис в $\mathbb{R}^K$.

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & \dots & 0 \\ 0 & \ddots & & & 0 \\ 0 & & \lambda_d & & 0 \\ 0 & & & \ddots & 0 \\ 0 & 0 & \dots & & 0 \end{pmatrix} \in \mathbb{R}^{L \times K}. \text{ Тогда}^8$$

$$\mathbb{Y} = \mathbb{U} \Lambda^{1/2} \mathbb{V}^T.$$

⁵ Они длинные :)

⁶ Можем $\mathbb{X}$ транспонировать, проделать всё то же самое, а поменяются местами только U_i и V_i .

⁷ $U_{d+1}, \dots, U_L$ соответствуют нулевому собственному числу матрицы.

⁸ $\mathbb{U}, \mathbb{V}$ — ортогональные матрицы.

1.3. Единственность сингулярного разложения

Насколько единственно разложение SVD (оно одно существует для матрицы или нет)? Можно подумать, что разложение не единственное, так как

1. Собственные вектора не единственные, то есть если U_i — собственный вектор, то $(-U_i)$ — собственный вектор,⁹

$$\mathbb{Y} = \sum_{i=1}^d \sqrt{\lambda_i} U_i V_i^T = \sum_{i=1}^d \sqrt{\lambda_i} (-U_i) (-V_i)^T.$$

2. Пусть есть два одинаковых собственных числа $\lambda = \lambda_1 = \lambda_2$, U_1 и U_2 — два ортонормированных вектора, соответствующих собственному числу λ . Тогда любая линейная комбинация U_1 и U_2 будет являться также собственным вектором и будет соответствовать тому же собственному числу, то есть $\forall \alpha, \beta: \alpha U_1 + \beta U_2$ — с.в. с с.ч. λ . Таким образом, если у нас есть два одинаковых собственных числа, то они порождают подпространство размерности 2, и любой ортонормированный базис в этом подпространстве подходит нам в качестве собственного вектора.¹⁰

Получаем, что единственности в буквальном смысле не получается. Поэтому сформулируем необходимое нам утверждение.

Предложение 3 (Единственность SVD). Пусть $L \leq K$. Пусть $\mathbb{Y} = \sum_{i=1}^L c_i P_i Q_i^T$ — некоторое разложение в сумму элементарных матриц (биортогональное разложение), такое что:

1. $c_1 \geq \dots \geq c_L \geq 0$;

2. $\{P_i\}_{i=1}^L$ — ортонормированные, $\{Q_i\}_{i=1}^L$ — ортонормированные.

Тогда $\mathbb{Y} = \sum_{i=1}^L c_i P_i Q_i^T$ — SVD, то есть любое биортогональное разложение с неотрицательными коэффициентами является сингулярным.

⁹ В вещественном случае собственный вектор определяется единственным образом с точностью до константы, модуль которой равен 1 (это всего 1 и -1). Но сингулярное разложение обобщается в комплексном случае, а здесь уже констант, по модулю равных 1, много.

¹⁰ Если у нас есть два одинаковых собственных числа, то мы можем брать любой базис, но сумма двух матриц постоянна, то есть она не меняется от выбора базиса.

Замечание 4. В частности:

- $c_d > 0, c_{d+1} = \dots = c_L = 0$,
- $c_i^2 = \lambda_i$ — собственные числа $\mathbb{Y}\mathbb{Y}^T$,
- P_i — собственные вектора $\mathbb{Y}\mathbb{Y}^T$,
- Q_i — собственные вектора $\mathbb{Y}^T\mathbb{Y}$,
- $Q_i = \frac{\mathbb{Y}^T P_i}{\sqrt{\lambda_i}}, i = 1, \dots, d$ ($d = \text{rank } \mathbb{Y}\mathbb{Y}^T$).

1.4. Оптимальные свойства сингулярного разложения

Обозначим $M_r \subset \mathbb{R}^{L \times K}$ — пространство матриц ранга, меньшего или равного r .

Предложение 4 (Оптимальные свойства сингулярного разложения). Пусть $r \leq d$.

1. (Аппроксимация матрицей (Low-rank approximation))

$$\min_{\tilde{\mathbb{Y}} \in M_r} \|\mathbb{Y} - \tilde{\mathbb{Y}}\|_F^2 = \sum_{i=r+1}^d \lambda_i \text{ и достигается на } \tilde{\mathbb{Y}} = \sum_{i=1}^r \sqrt{\lambda_i} U_i V_i^T.$$

2. (Аппроксимация подпространством)

Пусть $\mathcal{L}_r \subset \mathbb{R}^L$ — подпространство размерности $\leq r$. Тогда

$$\min_{\mathcal{L}_r} \sum_{i=1}^K \text{dist}^2(Y_i, \mathcal{L}_r) = \sum_{i=r+1}^d \lambda_i$$

и достигается на $\mathcal{L}_r^{(0)} = \text{span}(U_1, \dots, U_r)$.

Попробуем ответить на вопрос — что выгоднее хранить в памяти — всю матрицу или ее сингулярное разложение? Чтобы хранить матрицу данных размера $L \times K$, требуется хранить LK элементов. Чтобы хранить вектора сингулярного разложения, требуется $d(L + K)$ элементов.¹¹ Таким образом, если, к примеру, матрица близка к квадратной ($L = K$), то при $L > 2d$, выгоднее хранить сингулярное разложение.

¹¹ Всего d сингулярных троек, $U_i \in \mathbb{R}^L$, $V_i \in \mathbb{R}^K$, а сингулярные числа приписываем к одному из векторов.

Анализ главных компонент (АГК) (PCA — principal component analysis)

2.1. Главные направления

Пусть $Y_1, \dots, Y_K \in \mathbb{R}^L$. Рассмотрим вектор $P : \|P\| = 1$. Этот вектор задает направление (прямую, подпространство размерности 1). Проекция на данное направление выглядит следующим образом: $\langle Y_i, P \rangle P$, а ее длина равна $|\langle Y_i, P \rangle|$. Проекция измеряет то, насколько это направление соответствует нашим данным. Поставим задачу найти направление, которое лучше всего описывает нашу совокупность точек. Чем больше проекция, тем лучше. Задача:

$$\sum_{i=1}^K \langle Y_i, P \rangle^2 \rightarrow \max_P.$$

Вектор P_1 , на котором достигается максимум, задает *первое главное направление*.

Далее будем искать максимум по всевозможным векторам, ортогональным P_1 и так далее.

Предложение 5. *Верно следующее:*

1. $\max_P \sum_{i=1}^K \langle Y_i, P \rangle^2 = \lambda_1$ и достигается на $P = U_1$,
2. $\max_{P: P \perp U_1} \sum_{i=1}^K \langle Y_i, P \rangle^2 = \lambda_2$ и достигается на $P = U_2$,

...

- r . $\max_{P: P \perp U_j, j=1, \dots, r-1} \sum_{i=1}^K \langle Y_i, P \rangle^2 = \lambda_r$ и достигается на $P = U_r$.

U_i — *главные направления*.

Разложение Y_j по главным направлениям: $Y_j = \sum_{i=1}^r \langle Y_j, U_i \rangle U_i$.

$\langle Y_j, U_i \rangle$ — i -я компонента вектора Y_j (коэффициент разложения вектора по главным направлениям). Составим *вектор i -х главных компонент*:

$$Z_i = \begin{pmatrix} \langle Y_1, U_i \rangle \\ \dots \\ \langle Y_K, U_i \rangle \end{pmatrix} = \mathbb{Y}^T U_i = \sqrt{\lambda_i} V_i = \mathbb{X} U_i.$$

2.2. Анализ главных компонент. Построение

2.2.1. Анализ главных компонент на выборочном языке

Вернемся к сингулярному разложению. Имеем разложение:

$$\mathbb{Y} = \sum_{i=1}^d \sqrt{\lambda_i} U_i V_i^T.$$

Теперь перейдём на выборочный язык анализа главных компонент. Помним, что $\mathbb{Y} = \mathbb{X}^T \in \mathbb{R}^{L \times K}$, где столбцы — индивиды, строки — признаки; индивидов K , а признаков L .¹ В случае анализа главных компонент $D_1 = \{1, \dots, L\}$, $D_2 = \{1, \dots, K\}$, $\mu_1(\{i\}) = 1$ — считающая мера, $\mu_2(\{i\}) = \frac{1}{K}$ — вероятностная мера, где i — номер индивида. Отсюда следует, что квадрат нормы $\|\cdot\|_1^2$ вектора-индивида из $\mathbb{R}^L$ равен сумме квадратов его компонент всегда, а норма в квадрате $\|\cdot\|_2^2$ вектора-признака из $\mathbb{R}^K$ в анализе главных компонент считается не как сумма, а как среднее арифметическое квадратов компонент вектора. Далее, 1 у нормы вектора означает, что там сумма квадратов, а 2 — что среднее арифметическое. Если индекса у нормы нет, то это означает, что рассматривается вариант 1, т.е., обычная евклидова норма.

Предполагаем, что $\mathbb{Y}$ — центрированная по строчкам, то есть среднее по признакам равно 0, и тогда получаем другое разложение:

$$\mathbb{Y} = \sum_{i=1}^d \sqrt{\tilde{\lambda}_i} \tilde{U}_i \tilde{V}_i^T.$$

2.3. Связь между SVD и АГК. Общее и различия

Необходимо найти связь между $\tilde{\lambda}_i$, $\tilde{U}_i$, $\tilde{V}_i$ и λ_i , U_i , V_i соответственно. Знаем, что

- U_i — ортонормированная система собственных векторов матрицы $\mathbb{Y}\mathbb{Y}^T = \mathbb{X}^T\mathbb{X}$,
- $\tilde{U}_i$ — ортонормированная система собственных векторов матрицы $\frac{1}{K}\mathbb{Y}\mathbb{Y}^T = \frac{1}{K}\mathbb{X}^T\mathbb{X}$.

Таким образом, получаем следующие соотношения:

- $U_i = \tilde{U}_i$,
- $\lambda_i = K\tilde{\lambda}_i$,

¹ Визуально: $\mathbb{Y}$ — горизонтальная матрица, а $\mathbb{X}$ — вертикальная.

- $V_i = \frac{\tilde{V}_i}{\sqrt{K}}$.
- Так как коэффициент (веса) $1/K$ не влияет на равенство нулю скалярного произведения, то условие ортогональности векторов V_i то же самое.

Также заметим, что

$$\|\mathbb{Y}\|_{1,2}^2 = \frac{1}{K} \sum_{ij} y_{ij}^2 = \frac{\|\mathbb{Y}\|_F^2}{K}.$$

Рассмотрим теперь отличия сингулярного разложения от анализа главных компонент.

1. Столбцы в матрице $\mathbb{Y}$ неравноправны, то есть SVD полностью симметрично, а АГК нет. В частности, это из-за разных норм в пространстве признаков.
2. В АГК предполагается, что признаки центрированы, а индивиды нет.
3. В АГК рекомендована нормировка признаков.

Когда нормируем признаки? Когда признаки измерены в разных шкалах.²

- Нормируем признаки, если есть, например, данные в сантиметрах и метрах.
- Не нормируем признаки, если есть, например, баллы за задачи и хотим, чтобы главная компонента отражала уровень по результатам задач. Пусть есть сложные (от 0 до 10) и простые (от 0 до 5) задачи. Ясно, что получить 2.5 балла за простую задачу и 5 баллов за сложную — это разные вещи, но если мы нормируем данные, то мы сравниваем эти две вещи.

² Если что-то измерено в шкале от 0 до 1, а что-то от 0 до 100, то результат АГК будет неинтересным. Всегда первое главное направление будет перетягиваться тем признаком, которые принимает значения с существенно большим разбросом.