Санкт-Петербургский государственный университет

Прикладная математика и информатика

Статистическое моделирование

Зенкова Наталья Валентиновна

Многомерный анализ данных

Лекции Голяндиной Нины Эдуардовны

Санкт-Петербург

2018

Оглавление

Введение

1. Генеральный язык

Пусть случайная величина $\xi = (\xi_1, \ldots, \xi_p)^{\mathrm{T}} \in \mathbb{R}^p$, где p — количество признаков. Обозначим

- *вектор средних* $\mu = (\mathrm{E}\xi_1, \ldots, \mathrm{E}\xi_p)^{\mathrm{T}} \in \mathbb{R}^p$;
- *ковариационную матрицу* $\Sigma = \mathrm{Cov}\xi = \mathrm{E}(\xi - \mathrm{E}\xi)(\xi - \mathrm{E}\xi)^{\mathrm{T}} \in \mathbb{R}^{p \times p}$.

Рассмотрим два случая:

- *Центрированные данные*: $\xi^{(c)} = \xi - \mathrm{E}\xi$. Тогда ковариационная матрица будет иметь вид: $\mathrm{Cov}\xi^{(c)} = \mathrm{E}\xi^{(c)}(\xi^{(c)})^{\mathrm{T}}$.
- *Стандартизованные данные*: $\xi_i^{(s)} = \frac{\xi_i - \mathrm{E}\xi_i}{\sqrt{\mathrm{D}\xi_i}}$.[1]
  Запишем теперь это в матричном виде: введём Δ диагональную матрицу, состоящую из элементов $\sqrt{\mathrm{D}\xi_i}$ для $i = 1, \ldots, p$. Тогда $\xi^{(s)} = \Delta^{-1}\xi^{(c)} \Rightarrow \mathrm{Corr}\xi = \mathrm{Cov}\xi^{(s)}$, то есть $\mathrm{Cov}\xi^{(s)}$ — это корреляционная матрица до стандартизации.

2. Выборочный язык

Перейдём теперь на **выборочный язык**.

Пусть $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ — 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 $p \times n$:

$\mathbf{x}_1$	$\ldots$	$\mathbf{x}_n$

.

Но в книгах встречается следующий вариант *матрицы данных*:

$$\mathbb{X} = \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{pmatrix} = [X_1 : \ldots : X_p] \in \mathbb{R}^{n \times p}.$$

[1] ξ_i — i-ая компонента вектора ξ.

Теперь знаем, как написать все характеристики, упомянутые выше, на выборочном языке.

- $\overline{X}_j = \frac{1}{n}\sum_{i=1}^{n}(X_j)_i$ — *выборочное среднее.*

  Центрированная матрица данных $\mathbb{X}^{(c)}$ [2]:

  $$\mathbb{X}^{(c)} = [X_1 - \overline{X}_1 : \ldots : X_p - \overline{X}_p].$$

- $s_j^2 = \frac{1}{n}||\mathbb{X}_j^{(c)}||^2$ — *выборочная дисперсия.*

  Матрица из стандартизованных признаков $\mathbb{X}^{(s)}$:

  $$\mathbb{X}_j^{(s)} = \frac{X_j - \overline{X}_j}{s_j} = \frac{\mathbb{X}_j^{(c)}}{s_j}, \qquad \frac{1}{n}||\mathbb{X}_j^{(s)}|| = 1 \quad \forall\, 1, \ldots, p.$$

- $\mathbb{S} = \frac{1}{n}(\mathbb{X}^{(c)})^{\mathrm{T}}\mathbb{X}^{(c)}$ — *выборочная ковариационная матрица.*[3]

- $\widehat{\mathrm{Corr}}\xi = \frac{1}{n}(\mathbb{X}^{(s)})^{\mathrm{T}}\mathbb{X}^{(s)}$ — *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:*[4]

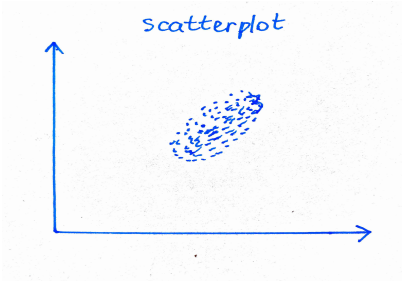

Рис. 1. Данные до стандартизации

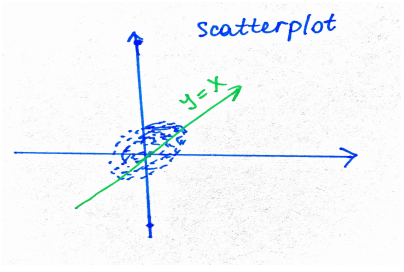

Рис. 2. Данные после стандартизации

[2] Посчитали среднее по каждому признаку и вычли. Таким образом, среднее по каждому признаку нулевое.

[3] Выписать несмещённую оценку ковариационной матрицы.

[4] Давайте подумаем, как будет выглядеть линия регрессии в первом и во втором случаях. Очевидно, что во втором случае она будет проходить через точку (0;0) и совпадать с прямой $y = x$.

2. *Данные с нулевой корреляцией:*

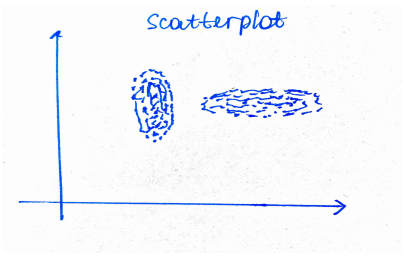

Рис. 3. Данные до стандартизации

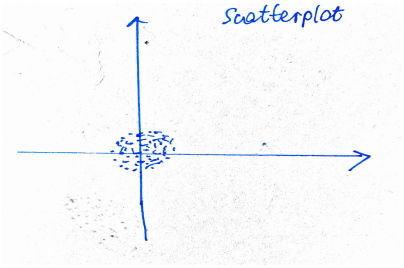

Рис. 4. Данные после стандартизации

4. Многомерное нормальное распределение

Рассмотрим $\xi \sim \mathrm{N}(\mu, \Sigma)$, где $\mu = \mathrm{E}\xi$ — среднее значение ξ, $\Sigma = \mathrm{Cov}\xi$ — ковариационная матрица. Предположим, что Σ — невырожденная матрица.[5] *Плотность многомерного нормального распределения* задаётся формулой:

$$\mathrm{p}(\mathbf{x}, \mu, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)\} \text{ для } \forall \mathbf{x} \in \mathbb{R}^p, \text{ где}$$

p — количество признаков. Очевидно, что в одномерном случае плотность будет иметь следующий вид:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}, \text{ где}$$

μ, σ — числа. Введём следующее понятие — *расстояние Махаланобиса*:

$$r_M^2(\mathbf{x}, \mu, \Sigma) = (\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu).$$

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 ($p = 1$):

$$\sqrt{\frac{(x-\mu)^2}{\sigma^2}} = \frac{|x-\mu|}{\sigma} \text{ — } \textit{расстояние, измеренное в } \sigma.$$

[5] $|\Sigma| \neq 0$.

На выборочном языке:

$$\frac{|x_i - \overline{x}|}{s_i}.$$

Таким образом, если у нас есть нормальное распределение, то в одномерном случае мы можем мерить расстояние в сигмах, а в случае многомерного нормального распределения мы смотрим на линии уровня.[6]

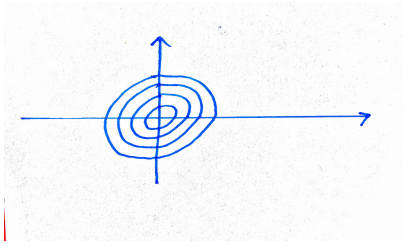

Рис. 5. Зависимые случайные величины

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

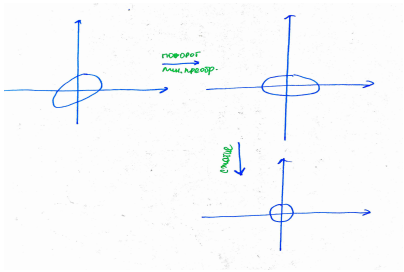

Рис. 6. Отбеливание

Пусть $\zeta = \Sigma^{-1/2}(\xi - \mu)$ — новая случайная величина, полученная из исходной линейными преобразованиями. Можем считать, что $\mathrm{E}\zeta = \mathbb{O}$, так как

$$\mathrm{E}\zeta = \Sigma^{-1/2}(\mathrm{E}\xi - \mu) = \mathbb{O}.$$

[6] Находимся на одном расстоянии от центра

Посчитаем ковариацию ζ:[7]

$$\mathrm{E}\zeta\zeta^{\mathrm{T}} = \Sigma^{-1/2} \underbrace{\Sigma^{1}}_{\mathrm{E}\xi\xi^{\mathrm{T}}} \Sigma^{-1/2} = \mathbb{I}.$$

Вспомним несколько свойств, которые характерны для нормально распределённых случайных величин:

1. Если $\xi \sim \mathrm{N}$, 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.

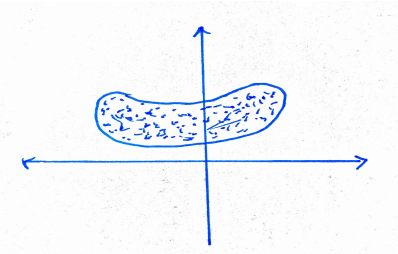

Рис. 7. Пример данных, когда зависимость есть, но корреляция нулевая

3. Условное математическое ожидание является линейной функцией.

4. Ковариационная матрица Σ может быть вырожденной:[8]

[7] В данном доказательстве воспользовались симметричностью и неотрицательной определённостью ковариационной матрицы Σ.

[8] Наблюдения лежат на одной прямой.

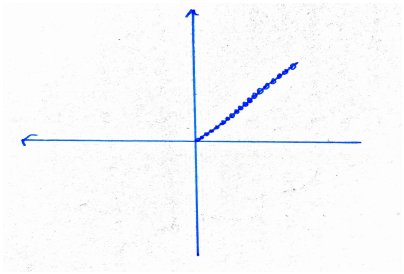

Рис. 8. Ковариационная матрица вырождена

5. Переход к новым признакам

Пусть ξ — p-мерная случайная величина.[9] Хотим образовать новый признак, который является линейной комбинацией старых признаков.[10]

$$\eta = \mathbf{A}^{\mathrm{T}}\xi = \xi^{\mathrm{T}}\mathbf{A} = a_1\xi_1 + \ldots + a_p\xi_p, \quad \mathbf{A} \in \mathbb{R}^p,$$
$$\mathrm{D}\eta = \mathbf{A}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{A}.$$

Теперь пусть $\eta = (\eta_1, \ldots, \eta_d)^{\mathrm{T}} \in \mathbb{R}^d$ — d-мерная случайная величина,[11] которая получается из исходной $\eta = \mathbb{A}^{\mathrm{T}}\xi$, где $\mathbb{A} \in \mathbb{R}^{p\times d}$. Соответствующая ковариационная матрица будет иметь вид $\mathrm{Cov}\eta = \mathbb{A}^{\mathrm{T}}\mathbf{\Sigma}\mathbb{A}$.

На выборочном языке: $Z = \sum\limits_{j=1}^{p} a_j X_j = \mathbb{X}A \in \mathbb{R}^n$ или $\mathbb{Z} = \mathbb{X}\mathbb{A} \in \mathbb{R}^{p\times d}$.[12]

5.1. Примеры

1. Возьмём в качестве индивида 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;

[9] p — количество признаков.

[10] Новый признак будет лучше, чем старые.

[11] d новых признаков.

[12] $\mathbb{X} \in \mathbb{R}^{n\times p}$ — матрица данных.

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 (рисунок 10).

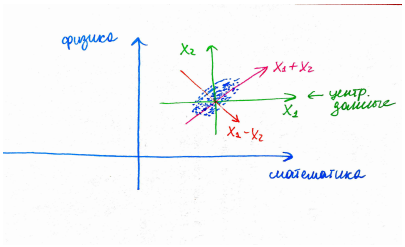

Рис. 9. Scatterplot: X_1 and X_2

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.[13][14]

3. Рассмотрим два признака: оценки по математике X_1 и литературе X_4. Максимальное отличие получаем по разности данных признаков.

[13] $X_1 - X_2$ — характеристика, по которой ученики минимально отличаются друг от друга.

[14] Необходимо выбрать признак, который бы наилучшим образом характеризовал данные. Такой характеристикой будет тот признак, по чему данные максимально различаются. Таким образом, приходим к анализу главных компонент.

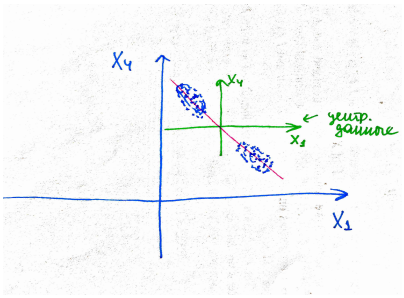

Рис. 10. Scatterplot: X_1 and X_4

5.2. Две интерпретации новых признаков. Факторные нагрузки

Обсудили, зачем необходимы новые признаки. Также до этого мы требовали, чтобы новые признаки были линейной комбинацией старых. Теперь предположим, что:

- на выборочном языке: $z_i \perp z_j$;[15]
- на генеральном языке: $\rho(\eta_i, \eta_j) = 0$.[16]

Зависимость признаков означает, что кусок одного признака входит в другой, то есть получается некоторая избыточность.

Пусть $\text{rank}(\mathbb{X}) = d$, $(z_1, \ldots, z_d)$ — ортогональный базис в в подпространстве размерности d — $\text{span}(X_1, \ldots, X_p) = \text{colspan}(\mathbb{X})$. Превратим базис в *ортонормированный*: $Q_i = \frac{z_i}{||z_i||}$.[17]

Новые признаки на самом деле интерпретируются двумя способами:

---

[15] Признаки ортогональны.

[16] Признаки хотя бы некоррелированы.

[17] Рассмотрим пример: пространство $\mathbb{R}^2$, соответствующий базис $(1,1)^{\text{T}}/\sqrt{2}$, $(1,-1)^{\text{T}}/\sqrt{2}$. Это ортонормированный базис по определению.[18] Как вычислить координаты вектора $(5,4)^{\text{T}}$ в данном базисе?

Если $U_1, \ldots, U_d$ — ортогональный базис в $\mathbb{R}^d$, тогда $\forall\ A \in \mathbb{R}^d$ раскладывается по ортонормированному базису:

$$A = \sum_{i=1}^{d} < A, U_i > U_i, \text{ где } < A, U_i > - i\text{-ая координата вектора } A \text{ в базисе } \{U_j\}_{j=1}^{d}.$$

- Новые признаки задаются равенством:[19]

$$\mathbb{X}\mathbb{A} = \mathbb{Z}.$$

- Исходные признаки выражаются через новые:[20]

$$\forall\, i = 1, \ldots, p \quad X_i = \sum_{j=1}^{d} < X_i, Q_j > Q_j. \tag{1}$$

Определение 1. $f_{ij} =< X_i, Q_j >$ — *факторные леса, факторные нагрузки.*

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

$$\mathbb{X} = \sum_{j=1}^{d} Q_j F_j^{\mathrm{T}} = \mathbb{Q}\mathbb{F}^{\mathrm{T}}. \tag{2}$$

$||Q_j|| = 1$, $||F_j|| \neq 1$ 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

$$\mathbb{X} = \sum_{j=1}^{d} \sigma_j Q_j P_j^{\mathrm{T}} = \mathbb{Q}\mathbf{\Sigma}\mathbb{P}^{\mathrm{T}}, \text{ где } \mathbf{\Sigma} = \mathrm{diag}\{\sigma_1, \ldots, \sigma_d\}. \tag{3}$$

Что за матрица $Q_j P_j^{\mathrm{T}}$*?*[21] $Q_j P_j^{\mathrm{T}} \in \mathbb{R}^{n\times p}$ и $\mathrm{rank}(Q_j P_j^{\mathrm{T}}) = 1$.[22] Таким образом, у нас была исходная матрица ранга d, а мы превратили её в сумму элементарных матриц ранга 1.

Предложение 1. $\mathbf{\Sigma}\mathbb{P}^{\mathrm{T}} = \mathbb{F}^{\mathrm{T}}$.

Доказательство. И это бы надо бы пояснить бы, гыыыыы. □

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

5.3. Важность новых признаков

Ввели новый признак, теперь хотелось бы понять, какую важную часть информации он в себя включает. Введём скалярное произведение двух матриц:[23]

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

[19] Новые признаки есть линейная комбинация старых.

[20] Есть пространство, в нём базис, каждый вектор этого пространства раскладывается по базису.

[21] Вектор умножаем на вектор.

[22] Столбцы пропорциональны.

[23] Чтобы уметь измерять вклад признака, необходимо уметь измерять одним числом, то есть ввести норму матрицы. Скалярное произведение по Фробениусу ($||\mathbb{A}||^2 = \sum_{ij} a_{ij}^2$).

Если верно $||\mathbb{X}||^2 = ||\sum\limits_{j=1}^{d} \mathbb{X}^{(j)}||_F^2 \stackrel{?}{=} \sum\limits_{j=1}^{d} ||\mathbb{X}^{(j)}||_F^2$, то *вклад j-ого признака* — $\frac{||\mathbb{X}^{(j)}||^2}{||\mathbb{X}||^2}$.[24]

Предложение 2. $||\sum\limits_{j=1}^{d} \mathbb{X}^{(j)}||_F^2 = \sum\limits_{j=1}^{d} ||\mathbb{X}^{(j)}||_F^2$, *если* $\mathbb{X}^{(j)}$ *ортогональны.*

Доказательство. Доказать как было в тесте. Домножить там на транспонированную матрицу и все дела. Мне лениво сейчас это делать. Скалярное произведение 0. Должны быть ортогональны либо Q, либо P. □

Так как Q_j нормированы (НАДО БЫ И ЭТО ПОЯСНИТЬ, ИБО ЩА Я ВАЩЕ НЕ СООБРАЖАЮ), то можем продолжить равенство:

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

Замечание 1. *Пусть старые признаки X_i центрированные, тогда новые признаки Z_i и Q_i тоже центрированные.*

Доказательство. АНАААААААЛОООООГИИИИИИИЧНООО, записать по-человечески. Новые — линейная комбинация старых. Среднее линейной комбинации есть линейная комбинация средних по признакам. Средние по признакам равны 0, получаем 0. □

6. Пара фактов из линейной алгебры

1. *Унитарная матрица* $\mathbb{U}$ — ортогональная матрица в комплексном случае.

   - $\mathbb{U}$ — квадратная матрица: $\mathbb{U}^{\mathrm{T}} = \mathbb{U}^{-1}$.
   - Столбцы $\mathbb{U}$ ортонормированы.
   - Строки $\mathbb{U}$ ортонормированы.[25]

   Умножение на матрицу $\mathbb{U}$ означает *поворот* или *отражение*. Пусть есть вектора Y, Z, после умножения на матрицу $\mathbb{U}$ получим $\widetilde{Y} = \mathbb{U}Y$ (ВСТАВИТЬ РИСУНОК).[26]

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

[25] 2 пункт эквивалентен 3. *Почему?* Если матрица ортогонональная, то и транспонированная к ней тоже ортогональная (следует из пункта 1).

[26] Поворачиваем вектора Y и Z на какой-то угол, а также при домножении на матрицу $\mathbb{U}$ не меняются нормы векторов.

Пример 1. $\mathbb{U} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix}$.

2. Пусть $\{P_i\}_{i=1}^r$ — система независимых векторов, рассмотрим линейную оболочку $\mathcal{L}_r = \operatorname{span}\{P_1, \ldots, P_r\}$ в $\mathbb{R}^L$, $\Pi : \mathbb{R}^L \to \mathcal{L}_r$ — проектор на $\mathcal{L}_r$.[27] Матрица Π:

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

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

$$\mathbb{P}^{\mathrm{T}}\mathbb{P} = \begin{pmatrix} 1 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{r\times r} \text{ и } \Pi = \mathbb{P}\mathbb{P}^{\mathrm{T}}.$$

[27] После проекции размерность вектора не меняется.

Глава 1

Сингулярное разложение (Singular Value Decomposition)

1.1. Как строится сингулярное разложение

Пусть L — число признаков, K — количество индивидов,[1] $\mathbb{Y} = \mathbb{X}^{\mathrm{T}} \in \mathbb{R}^{L \times K}$ — ненулевая матрица. Обозначим $\mathbb{S} = \mathbb{Y}\mathbb{Y}^{\mathrm{T}} \in \mathbb{R}^{L \times L}$ — симметричная неотрицательно определённая матрица.

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

$\{U_i\}_{i=1}^{L}$ — ортонормированный набор из собственных векторов матрицы $\mathbb{S}$,
$\lambda_1 \geq \ldots \geq \lambda_L \geq 0$ — собственные числа матрицы $\mathbb{S}$.[2]

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

Предложение 3. *1.* $d = \mathrm{rank}\mathbb{Y}\mathbb{Y}^{\mathrm{T}}$.

2. $\lambda_d > 0$, $\lambda_i = 0$, $i > d$.[3]

3. $\{U_i\}_{i=1}^{d}$ *образуют ортонормированный базис* $\mathrm{colspan}\mathbb{Y}$.[4]

Введём вектор[5]

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

Предложение 4. *1.* $\{V_i\}_{i=1}^{d}$ *— оронормированная система векторов.*

2. V_i *— собственные вектора* $\mathbb{Y}\mathbb{Y}^{\mathrm{T}}$*, соответствующие тем же самым сосбственным числам* λ_i*. Остальные собственные вектора* $\mathbb{Y}\mathbb{Y}^{\mathrm{T}}$ *соответствуют нулевому собственному числу.*

---

[1] Для понимания пусть будет так.

[2] Неотрицательная определённость матрицы $\mathbb{S}$ означает именно это.

[3] Упорядочили собственные числа: первые d строго положительные, а остальные все нули.

[4] Как понимаю, данное утверждение достаточно элементарное, поэтому все нужно будет как-то на экзамене пояснять, хоть оно и идёт без доказательства. И снова не сейчас мне что-то комментировать подробнее...

[5] «Длинный вектор». $i = 1, \ldots, d$ — на 0 делить не можем же.

3. $U_i = \frac{\mathbb{Y}V_i}{\sqrt{\lambda_i}}$ — *верно и в обратную сторону.*

4. $\mathbb{Y} = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$ — *SVD.*[6][7]

Что здесь считать новыми признаками, если $\mathbb{Y} = \mathbb{X}^{\mathrm{T}}$? V_i,[8] 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.[9]
- $\frac{\lambda_i}{\sum_i \lambda_i}$ — вклад нового признака.

Определение 2. $\sqrt{\lambda_i}$ — *сингулярные числа матрицы* $\mathbb{Y}$, U_i — *левый сингулярный вектор,* V_i — *правый сингулярный вектор.*

1.1.1. Матричный вид сингулярного разложения

Можно записать двумя способами:

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

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

2. Возьмём $\mathbb{U} = [U_1 : \ldots : U_d : U_{d+1} : \ldots : U_L]$ — ортонормированный базис в $\mathbb{R}^L$.[10]
$\mathbb{V} = [U_1 : \ldots : U_d : U_{d+1} : \ldots : U_K]$ — ортонормированный базис в $\mathbb{R}^K$.
$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \ldots & \ldots & 0 \\ 0 & \ddots & & 0 & 0 \\ 0 & & \lambda_d & & 0 \\ 0 & & & \ddots & 0 \\ 0 & 0 & \ldots & & 0 \end{pmatrix} \in \mathbb{R}^{L \times K}$. Тогда[11]

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

[6] Разложение в сумму элементарных матриц.

[7] Самый важный пункт утверждения. Ради него все и вводилось.

[8] Они длинные :)

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

[10] $U_{d+1}, \ldots, U_L$ соответствуют нулевому собственному числу матрицы.

[11] $\mathbb{U}, \mathbb{V}$ — ортогональные матрицы.

1.2. Единственность SVD

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.[12]

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

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 собственных векторов.[13]

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

Предложение 5 (Единственность SVD). *Пусть* $\mathbb{Y} = \sum_{i=1}^{L} c_i P_i Q_i^{\mathrm{T}}$ *— некоторое разложение в сумму элементарных матрицы (биортогональное разложение), такое что:*

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

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

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

Замечание 2. *В частности:*

- $c_d > 0$, $c_{d+1} = \ldots = c_L = 0$,

[12] 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.

[13] 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. Можно так-то даже в этом убедиться.

- $c_i^2 = \lambda_i$ — *собственные числа* $\mathbb{Y}\mathbb{Y}^{\mathrm{T}}$,
- P_i — *собственные вектора* $\mathbb{Y}\mathbb{Y}^{\mathrm{T}}$,
- Q_i — *собственные вектора* $\mathbb{Y}^{\mathrm{T}}\mathbb{Y}$,
- $Q_i = \frac{\mathbb{Y}^{\mathrm{T}} P_i}{\sqrt{\lambda_i}}$, $i = 1, \ldots, d$ *(*$d = \mathrm{rank}\mathbb{Y}$*).*

ДОПИСАТЬ SVD, пропущена часть

Глава 2

Анализ главных компонент (PCA — principal component analysis)

2.1. Анализ главных компонент на выборочном языке

Вспомним, как выглядит SVD на выборочном языке:

$$D_1 = \{1, \dots, L\},\ D_2 = \{1, \dots, K\}, \quad \mu_1,\ \mu_2 \text{ — считающие меры,}$$
$$\mathbb{Y} = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}.$$

Теперь перейдём на выборочный язык АГК. Помним, что $\mathbb{Y} = \mathbb{X}^{\mathrm{T}}$, где столбцы — индивиды, строки — признаки; индивидов K, а признаков L.[1]

$$D_1 = \{1, \dots, L\},\ D_2 = \{1, \dots, K\},$$
$$\mu_1 \text{ — считающая мера, } \mu_2(\{i\}) = \frac{1}{K} \text{ — вероятностная мера, где } i \text{ — номер индивида.}$$

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

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

2.2. Связь между SVD и АГК

Необходимо найти связь между $\widetilde{\lambda}_i$, $\widetilde{U}_i$, $\widetilde{V}_i$ и λ_i, U_i, V_i соответственно.[2] Знаем, что

$$U_i \text{ — о.н.с с.в. матрицы } \mathbb{Y}\mathbb{Y}^{\mathrm{T}} = \mathbb{X}^{\mathrm{T}}\mathbb{X},$$
$$\widetilde{U}_i \text{ — о.н.с с.в. матрицы } \frac{1}{K}\mathbb{Y}\mathbb{Y}^{\mathrm{T}} = \frac{1}{K}\mathbb{X}^{\mathrm{T}}\mathbb{X}.$$

То есть получили, что U_i и $\widetilde{U}_i$ совпадают с точностью до коэффициента.[3]

Таким образом, получаем следующие соотношения:[4]

---

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

[2] В SVD веса 1, а здесь 1 и $1/K$.

[3] Если в одном и том же пространстве есть два нормированных вектора: один нормирован с одними весами, другой с другими, то они не должны совпадать. Но здесь у нас понятие нормированности одинаковое за счёт того, что у U_i вес один и тот же — 1.

[4] Хм, считаю, что в конспекте была очепятка и так верно.

- $U_i = \widetilde{U}_i$,
- $\lambda_i = \frac{\widetilde{\lambda}_i}{K}$,
- $V_i = \sqrt{K}\widetilde{V}_i$.

Также заметим, что[5]

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

2.3. Чем отличается SVD от АГК?

1. Столбцы в матрице $\mathbb{Y}$ неравноправны, то есть SVD полностью симметрично, а АГК нет. Если формально, то получаем, что разные нормы в пространстве признаков.
2. В АГК предполагается, что признаки центрированы, а индивиды нет.
3. В АГК рекомендована нормировка.

Когда нормируем признаки? Когда признаки измерены в разных шкалах.[6]

2.3.1. Пример

- 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 1 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, поэтому если мы нормируем данные, то мы сравняем эти две вещи.

[5] Записать через интегралы.

[6] Если что-то измерено в шкале от 0 до 1 млн, а что-то от 0 до 100, то результат АГК будет странным :) Всегда первая главная компонента будет там, где миллионы.

2.4. Чему АГК соответствует на статистическом языке?

Перейдём к $\mathbb{X} \in \mathbb{R}^{n \times p}$ ($L \to p$ — количество индивидов, $K \to n$ — число признаков). Пусть $\mathbb{X}$ — центрированы. *Берём признак, какую норму нужно рассматривать?* Вероятностную норму.[7] *Что означает характеристика* $||X_i||_2^2$ *на статистическом языке?*

$$||X_i||_2^2 = \frac{1}{n}\sum_{j=1}^{n}\left((X_i)_j - \overline{X}_i\right)^2 = S^2(X_i) \text{ — выборочная дисперсия вектора } X_i.^{8}$$

Что означает матрица данных $||X_i||_{1,2}^2$ *на статистическом языке?* Используем верхнюю строчку и предполагаем, что $\mathbb{X}$ центрированы.

$$||X_i||_{1,2}^2 = \sum_{i=1}^{p}||X_i||_2^2 = \sum_{i=1}^{p}S^2(X_i) \text{ — total variance.}^{9}$$

Посчитаем норму вектора главных компонент (считаем, что АГК: $\mathbb{Y} = \sum\limits_{i=1}^{p}\sqrt{\lambda_i}U_iV_i^{\mathrm{T}}$):

$$Z_i = \mathbb{X}U_i = \sqrt{\lambda_i}V_i, \text{ где } Z_i \text{ — проекция на } i\text{-ое направление и } i = 1\ldots, n.$$

Знаем, что $||Z_i||_2^2 = S^2(Z_i)$. Учитывая, что V_i нормированы, то

$$||Z_i||_2^2 = S^2(Z_i) = \lambda_i.$$

2.5. АГК на выборочном языке (продолжение)

[7] У длинных векторов вторая норма.