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

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 $p \times n$:

$$\begin{array}{|c|c|c|} \hline \\ \mathbf{x}_1 & \ldots & \mathbf{x}_n \\ \\ \hline \end{array}.$$

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.

- $\overline{x}_j = \frac{1}{n}\sum\limits_{i=1}^{n}(X_j)_i$ — *выборочное среднее*, $j = 1, \ldots, p$.

  $\overline{X}_j = (\overline{x}_j, \ldots, \overline{x}_j)^{\mathrm{T}} \in \mathbb{R}^n$.

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

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

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

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

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

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

- $\widehat{\mathrm{Corr}}\xi = \frac{1}{n}(\mathbb{X}^{(s)})^{\mathrm{T}}\mathbb{X}^{(s)}$ — *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:*[3]

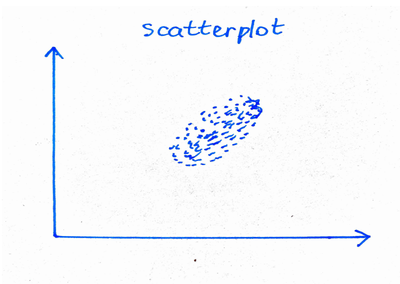

Рис. 1. Данные до стандартизации

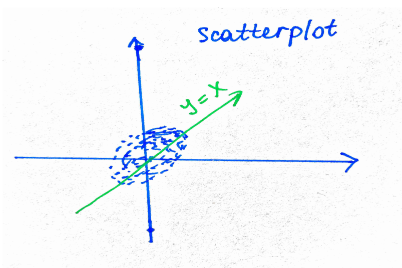

Рис. 2. Данные после стандартизации

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

[3] 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:*

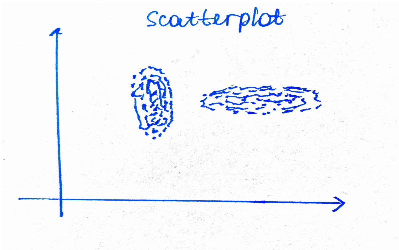

Рис. 3. Данные до стандартизации

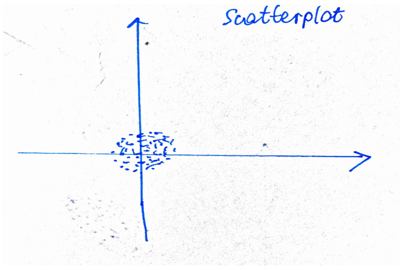

Рис. 4. Данные после стандартизации

4. Многомерное нормальное распределение

Рассмотрим $\xi \sim \mathrm{N_p}(\mu, \Sigma)$, где $\mu = \mathrm{E}\xi$ — среднее значение ξ, $\Sigma = \mathrm{Cov}\xi$ — ковариационная матрица. Предположим, что Σ — невырожденная матрица.[4] *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 ($p = 1$):

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

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

На выборочном языке:

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

Таким образом, если у нас есть нормальное распределение, то в одномерном случае мы можем измерять расстояние в сигмах, а в случае многомерного нормального распределения мы смотрим на линии уровня (точки, находящиеся на одном расстоянии Махаланобиса от центра).

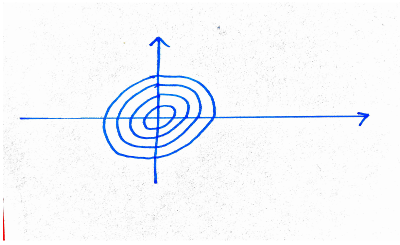

Рис. 5. Линии уровня

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

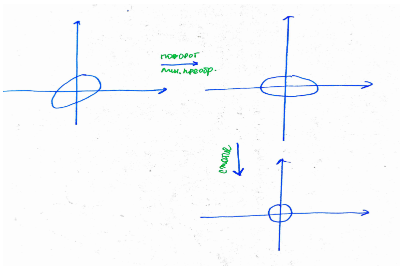

Рис. 6. Отбеливание

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. Можем считать, что $\mathrm{E}\zeta = \mathbb{O}$, так как

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

Посчитаем ковариацию ζ:[5]

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

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.

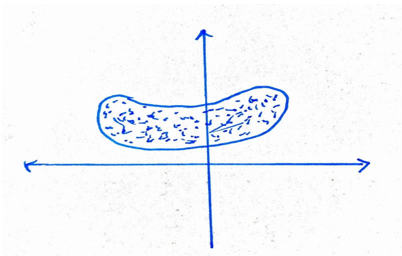

Рис. 7. Пример данных, когда зависимость есть, но корреляция нулевая

3. Условное математическое ожидание является линейной функцией.

4. Ковариационная матрица Σ может быть вырожденной:[6]

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

[6] Наблюдения лежат на одной прямой.

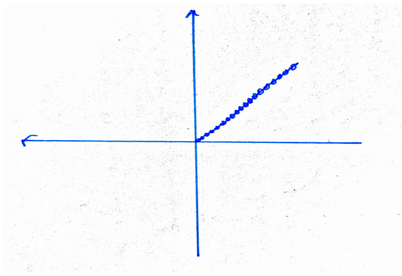

Рис. 8. Ковариационная матрица вырождена

5. Переход к новым признакам

Пусть ξ — p-мерная случайная величина.[7] Считаем, что среднее равно нулю. Хотим образовать новый признак, который является линейной комбинацией старых признаков.[8]

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

Теперь пусть $\eta = (\eta_1, \ldots, \eta_d)^{\mathrm{T}} \in \mathbb{R}^d$ — d-мерная случайная величина,[9] 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 $\mathbb{Z} = \mathbb{X}\mathbb{A} \in \mathbb{R}^{p\times d}$.[10]

Приведем примеры перехода к новым признакам.

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;

[7] p — количество признаков.
[8] Новый признак будет в некотором смысле лучше, чем старые.
[9] d новых признаков.
[10] $\mathbb{X} \in \mathbb{R}^{n\times p}$ — матрица данных.

- 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.

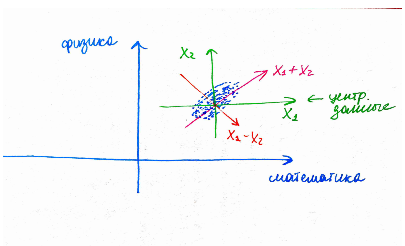

Рис. 9. Scatterplot: X_1 and X_2

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.[11][12]

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

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

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

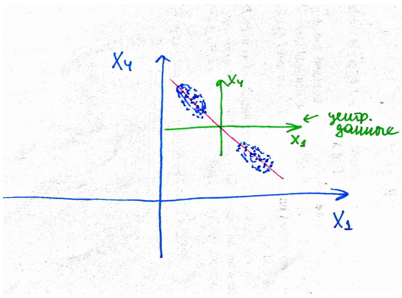

Рис. 10. Scatterplot: X_1 and X_4

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 предположим, что:

- на выборочном языке: $z_i \perp z_j$;[13]
- на генеральном языке: $\rho(\eta_i, \eta_j) = 0$.[14]

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

Пусть $\mathrm{rank}(\mathbb{X}) = d$, $(z_1, \ldots, z_d)$ — ортогональный базис в в подпространстве размерности d — $\mathrm{span}(X_1, \ldots, X_p) = \mathrm{colspan}(\mathbb{X})$. Превратим базис в *ортонормированный базис*: $Q_i = \frac{z_i}{||z_i||}$.[15] Таким образом, $\{Q_i\}_{i=1}^d$ — ортонормированный базис в пространстве $\mathrm{span}(X_1, \ldots, X_p)$.

Новые признаки на самом деле интерпретируются двумя способами:

[13] Признаки ортогональны.

[14] Признаки хотя бы некоррелированы.

[15] Рассмотрим пример: пространство $\mathbb{R}^2$, соответствующий базис $(1,1)^{\mathrm{T}}/\sqrt{2}$, $(1,-1)^{\mathrm{T}}/\sqrt{2}$. Это ортонормированный базис. Как вычислить координаты вектора $(5,4)^{\mathrm{T}}$ 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:[16]

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

- Исходные признаки выражаются через новые:[17]

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

Определение 1. $f_{ij} =< X_i, Q_j >$ — *факторные веса, 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 ортонормированному 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 $Q_j P_j^{\mathrm{T}} \in \mathbb{R}^{n\times p}$*?* $\mathrm{rank}(Q_j P_j^{\mathrm{T}}) = 1$.[18] Таким образом, у нас была исходная матрица ранга d, а мы превратили её в сумму d элементарных матриц ранга 1.

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

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:[19]

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

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* (вклад j-ой элементарной матрицы) — $\frac{||\mathbb{X}^{(j)}||^2}{||\mathbb{X}||^2}$.[20]

[16] Новые признаки есть линейная комбинация старых.

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

[18] Столбцы пропорциональны.

[19] Чтобы уметь измерять вклад признака, необходимо уметь измерять одним числом, то есть ввести норму матрицы. Норма по Фробениусу: $||\mathbb{A}||_F^2 = \sum\limits_{ij} a_{ij}^2$.

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

Замечание 1. $||\sum\limits_{j=1}^{d} \mathbb{X}^{(j)}||_F^2 = \sum\limits_{j=1}^{d} ||\mathbb{X}^{(j)}||_F^2$ *(скалярное произведение равно 0, потому что* $Q_i \perp Q_j$*).*

Так как Q_j нормированы, 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 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. □

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.[21]

   Умножение на матрицу $\mathbb{U}$ означает *поворот* или *отражение*. Пусть есть вектора Y, Z, после умножения на матрицу $\mathbb{U}$ получим $\widetilde{Y} = \mathbb{U}Y$, $\widetilde{Z} = \mathbb{U}Z$.[22]

   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 $\mathcal{L}_r = \mathrm{span}\{P_1, \ldots, P_r\}$ в $\mathbb{R}^L$, $\Pi : \mathbb{R}^L \to \mathcal{L}_r$ — проектор на $\mathcal{L}_r$. Матрица Π:

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

[21] 2 пункт эквивалентен 3. *Почему?* Если матрица ортогонональная, то и транспонированная к ней тоже ортогональная (следует из пункта 1).

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

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

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

Глава 1

Сингулярное разложение (SVD — Singular Value Decomposition)

1.1. Как строится сингулярное разложение

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

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

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

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

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

Введём вектор

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

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

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

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

4. $\mathbb{Y} = \sum\limits_{i=1}^d \sqrt{\lambda_i} U_i V_i^\mathrm{T}$ — SVD (Сингулярное разложение матрицы).[3][4]

---

[1] Положительные, т.к. 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 $\mathbb{Y} = \mathbb{X}^{\mathrm{T}}$*?* 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]
- $\frac{\lambda_i}{\sum_i \lambda_i}$ — вклад i-ого признака.

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

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 $\mathbb{U}_d = [U_1 : \ldots : U_d]$, $\mathbb{V}_d = [V_1 : \ldots : V_d]$, $\Lambda_d = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. 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 $\mathbb{R}^L$.[7]
$\mathbb{V}^{\mathrm{T}} = [V_1 : \ldots : V_d : V_{d+1} : \ldots : V_K]$ — 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[8]
$$\mathbb{Y} = \mathbb{U} \Lambda^{1/2} \mathbb{V}^{\mathrm{T}}.$$

---

[5] Они длинные :)

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

[7] $U_{d+1}, \ldots, U_L$ 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]

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

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 собственного вектора.[10]

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

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

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

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

Тогда $\mathbb{Y} = \sum\limits_{i=1}^{L} c_i P_i Q_i^{\mathrm{T}}$ — 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:*

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

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* $r \le d$.

1. *(Аппроксимация матрицей (Low-rank approximation))*
   $\min\limits_{\widetilde{\mathbb{Y}} \in M_r} \|\mathbb{Y} - \widetilde{\mathbb{Y}}\|_F^2 = \sum\limits_{i=r+1}^{d} \lambda_i$ *и достигается на* $\widetilde{\mathbb{Y}} = \sum\limits_{i=1}^{r} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$.
2. *(Аппроксимация подпространством)*
   Пусть $\mathcal{L}_r \subset \mathbb{R}^L$ — *подпространство размерности* $\le r$. *Тогда*
   $$\min_{\mathcal{L}_r} \sum_{i=1}^{K} dist^2(Y_i, \mathcal{L}_r) = \sum_{i=r+1}^{d} \lambda_i$$
   и достигается на $\mathcal{L}_r^{(0)} = span(U_1, \ldots, U_r)$.

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, $U_i \in \mathbb{R}^L$, $V_i \in \mathbb{R}^K$.

Глава 2

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

2.1. Главные направления

Пусть $Y_1, \ldots, Y_K \in \mathbb{R}^L$. Рассмотрим вектор $P : \|P\| = 1$. Этот вектор задает направление (прямую, подпространство размерности 1). Проекция на данное направление выглядит следующим образом: $\langle Y_i, P\rangle^2$. 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:

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

Вектор P_1, на котором достигается максимум, называется *первым главным направлением.*

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

Предложение 5.

$$1.\quad \max_P \sum_{i=1}^{K} \langle Y_i, P\rangle^2 = \lambda_1 \text{ и достигается на } P = U_1,$$

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

$$\ldots$$

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

U_i — главные направления.

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

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

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

2.2. Анализ главных компонент. Построение

2.2.1. Разложение Гильберта-Шмидта

Пусть есть множества $D_1 = \{1, \ldots, L\}$, $D_2 = \{1, \ldots, K\}$, $\mu_1,\ \mu_2$ — считающие меры (то есть $\mu_1(\{i\}) = 1\ \ \forall i = 1, \ldots, L$, $\mu_2(\{j\}) = 1\ \ \forall j = 1, \ldots, K$). Введем два гильбертовых пространства: $L_1 = L^2(D_1, \mu_1)$ и $L_2 = L^2(D_2, \mu_2)$ со скалярными произведениями $\langle \cdot, \cdot \rangle_1$ и $\langle \cdot, \cdot \rangle_2$ и нормами $\| \cdot \|_1$ и $\| \cdot \|_2$ соответственно. Заметим, что L_1 — обычное пространство векторов длины L, а L_2 — пространство векторов длины K со стандартным евклидовым скалярным произведением.

Можем ввести отображение $\mathcal{G} : L_2 \to L_1$. В наших обозначениях это отображение переводит вектор размерности K в вектор размерности L. В качестве $\mathcal{G}$ можем брать отображение, которое задается умножением на матрицу $\mathbb{G} \in \mathbb{R}^{L \times K}$, то есть $\forall Y \in \mathbb{R}^K\ \ \mathbb{G}Y = X \in \mathbb{R}^L$.

Зададим сопряженное отображение $\mathcal{G}^* : L_1 \to L_2$ такое, что $\langle X, \mathbb{G}Y \rangle_1 = \langle \mathcal{G}^* X, Y \rangle_2$ $\forall\ X \in L_1,\ Y \in L_2$. Тогда действие $\mathcal{G}^*$ — это умножение на матрицу $\mathbb{G}^{\mathrm{T}}$.

Оператор $\mathcal{G}$ по определению задается ядром $g(x, s)$:

$$(\mathcal{G}h)(x) = \int\limits_{D_2} g(x, s) h(s) \mu_2(ds).$$

При введенных нами D_1, D_2 $g(x, s)$ — это просто элементы матрицы $\mathbb{G}$.

Далее рассмотрим операторы $\mathcal{G}\mathcal{G}^* : L_1 \to L_1$ и $\mathcal{G}^*\mathcal{G} : L_2 \to L_2$.

Замечание 5. *Пусть меры не считающие, то есть $\mu_1(\{i\}) = w_i\ \ \forall i = 1, \ldots, L$, а $\mu_2(\{j\}) = q_i\ \ \forall j = 1, \ldots, K$. Обозначим диагональные матрицы $\mathbb{W} = diag(w_i)$ и $\mathbb{Q} = diag(q_i)$. Пусть оператор $\mathcal{G}$ задается умножением на матрицу $\mathbb{G}$. Тогда оператор $\mathcal{G}\mathcal{G}^*$ задается матрицей $\mathbb{G}\mathbb{Q}\mathbb{G}^{\mathrm{T}}\mathbb{W}$, а оператор $\mathcal{G}^*\mathcal{G}$ задается матрицей $\mathbb{G}^{\mathrm{T}}\mathbb{W}\mathbb{G}\mathbb{Q}$.*

Ядро оператора $\mathcal{G}\mathcal{G}^*$ выглядит следующим образом:

$$g_{1,1}(x, s) = \int\limits_{D_2} g(x, s) g(y, s) \mu_2(ds).$$

Обозначим $\{\phi_n\}$, $\phi_n \in L_1$ — ортонормированная система собственных функций оператора $\mathcal{G}\mathcal{G}^*$; $\{\psi_n\}$, $\psi_n \in L_2$ — ортонормированная система собственных функций оператора $\mathcal{G}^*\mathcal{G}$. Известно, что им соответствуют одни и те же собственные числа λ_n и что $\psi_n = \frac{\mathcal{G}^* \phi_n}{\sqrt{\lambda_n}}$ и $\phi_n = \frac{\mathcal{G}\psi_n}{\sqrt{\lambda_n}}$. Обратим внимание, что для операторов $\mathcal{G}$, заданных одним и тем же ядром, собственные числа и вектора могут различаться, если меры в пространствах заданы разные. Разложение Шмидта ядра оператора $\mathcal{G}$:

$$g(x,s) = \sum_n \sqrt{\lambda_n}\phi_n(x)\psi_n(s)$$

Замечание 6. *Если $g(i,j)$ — элементы матрицы $\mathbb{G} \in \mathbb{R}^{L\times K}$, а меры считающие, тогда разложение Шмидта ядра оператора $\mathcal{G}$ — это в точности сингулярное разложение матрицы $\mathbb{G}$.*

2.2.2. Анализ главных компонент на выборочном языке

Вернемся к сингулярному разложению. Имеем разложение:

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

Теперь перейдём на выборочный язык анализа главных компонент. Помним, что $\mathbb{Y} = \mathbb{X}^{\mathrm{T}} \in \mathbb{R}^{L\times K}$, где столбцы — индивиды, строки — признаки; индивидов K, а признаков L.[1] В случае анализа главных компонент $D_1 = \{1,\ldots,L\}, D_2 = \{1,\ldots,K\}$, $\mu_1(\{i\}) = 1$ — считающая мера, $\mu_2(\{i\}) = \frac{1}{K}$ — вероятностная мера, где i — номер индивида. Предполагаем, что $\mathbb{Y}$ — центрированная по строчкам, то есть среднее по признакам равно 0, и тогда получаем другое разложение:

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

2.3. Связь между SVD и АГК. Общее и различия

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 — о.н.с с.в. матрицы $\mathbb{Y}\mathbb{Y}^{\mathrm{T}} = \mathbb{X}^{\mathrm{T}}\mathbb{X}$,
- $\widetilde{U}_i$ — 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:

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

[2] В SVD веса 1, а в АГК 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.

- $U_i = \widetilde{U}_i$,
- $\lambda_i = K\widetilde{\lambda}_i$,
- $V_i = \frac{\widetilde{V}_i}{\sqrt{K}}$.[4]

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. 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 признаков.

Когда нормируем признаки? Когда признаки измерены в разных шкалах.[5]

- 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 сравняем эти две вещи.

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. *Что означает характеристика $||X_i||_2^2$ на статистическом языке?*

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

[4] Во втором пространстве мера другая, понятие ортонормированности разное.

[5] Если что-то измерено в шкале от 0 до 1, а что-то от 0 до 100, то результат АГК будет странным. Всегда первая главная компонента будет там, где сотни.

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

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

Посчитаем норму вектора главных компонент (считаем, что АГК: $\mathbb{Y} = \sum\limits_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$): $Z_i = \mathbb{X} U_i = \sqrt{\lambda_i} V_i$, где Z_i — проекция на i-ое направление и $i = 1 \ldots, n$.

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

$$||Z_i||^2_2 = s^2(Z_i) = \|\mathbb{X} U_i\|^2_2 = \|\sqrt{\lambda_i} V_i\|^2_2 = \lambda_i.$$

Разложение можем записать следующим образом:

$$\mathbb{X}^{\mathrm{T}} = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}} = \sum_{i=1}^{d} F_i V_i^{\mathrm{T}} = \sum_{i=1}^{d} U_i Z_i^{\mathrm{T}},$$

где $F_i = \sqrt{\lambda_i} U_i$ — вектор i-х факторных весов (нагрузок), $Z_i = \sqrt{\lambda_i} V_i$ — вектор главных компонент.

2.5. Вклад главных компонент

Вычислим норму $\mathbb{Y} = \mathbb{X}^{\mathrm{T}}$.

$$\|\mathbb{X}^{\mathrm{T}}\|^2_{1,2} = \sum_{i=1}^{p} s^2(X_i) = \sum_{i=1}^{d} \|(\sqrt{\lambda_i} U_i V_i^{\mathrm{T}})^{\mathrm{T}}\|^2_2 = \sum_{i=1}^{d} \lambda_i = \sum_{i=1}^{d} s^2(Z_i).$$

Получилось, что total variance не меняется при переходе к новым признакам (при повороте норма векторов не меняется).

$\frac{\lambda_j}{\sum_{i=1}^{d} \lambda_i}$ — *вклад j-ой главной компоненты в общую дисперсию.*[6]

$s^2(X_i)$ — информативность i-ого признака, $s^2(Z_i)$ — информативность i-ой главной компоненты. Таким образом, чем больше разброс, тем больше эта характеристика информативна. Первая главная компонента имеет наибольшую норму, поэтому эта компонента самая информативная.

Напомним, что $Z_i = \mathbb{X} U_i$, то есть Z_i — линейная комбинация X_j с коэффициентами, взятыми из U_i. Таким образом, *главные компоненты* — это ортогональные между собой линейные комбинации исходных признаков, обладающие свойством оптимальности.

Возможны случаи:

[6] В числителе — дисперсия нового признака, в знаменателе — общая дисперсия.

1. Матрица $\mathbb{X}$ — центрирована. Тогда U_i — это собственные векторы матрицы $\frac{1}{n}\mathbb{Y}\mathbb{Y}^{\mathrm{T}} = \frac{1}{n}\mathbb{X}^{\mathrm{T}}\mathbb{X}$ (выборочная ковариационная матрица).

2. Матрица $\mathbb{X}$ — центрирована и нормирована. Тогда U_i — собственные векторы матрицы $\frac{1}{n}\mathbb{Y}\mathbb{Y}^{\mathrm{T}} = \frac{1}{n}\mathbb{X}^{\mathrm{T}}\mathbb{X}$ (выборочная корреляционная матрица).

2.6. АГК с точки зрения построения базиса в пространстве индивидов и в пространстве признаков

Анализ главных компонент на выборочном языке: $\mathbb{X} \in \mathbb{R}^{n\times p},\ U_i \in \mathbb{R}^p,\ V_i \in \mathbb{R}^n$.

$$\mathbb{X}^{\mathrm{T}} = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}} = \sum_{i=1}^{d} F_i V_i^{\mathrm{T}} = \sum_{i=1}^{d} U_i Z_i^{\mathrm{T}},$$

$F_i = \sqrt{\lambda_i} U_i$ — вектор i-х факторных весов (нагрузок),

$Z_i = \sqrt{\lambda_i} V_i$ — вектор главных компонент.

1. $\mathbb{X}^{\mathrm{T}} = \sum\limits_{i=1}^{d} F_i V_i^{\mathrm{T}}$, где $\{V_i\}_{i=1}^{d}$ — ортонормированный базис в пространстве признаков.

   Пусть $\mathbb{F} = \{f_{ij}\}_{i=1,j=1}^{p,d} = [F_1 : \ldots, F_d]$. Тогда

$$f_{ij} = \underbrace{\langle X_i, V_j\rangle_2}_{\substack{j\text{-я коорд. } i\text{-ого признака} \\ \text{в базисе новых признаков}}} = \begin{cases} \mathrm{Cov}(X_i, V_j), & \text{если АГК по ковариац. матр.} \\ \rho(X_i, V_j) = \rho(X_i, Z_j), & \text{если АГК по корреляц. матр.} \end{cases}$$

2. $\mathbb{X}^{\mathrm{T}} = \sum\limits_{i=1}^{d} U_i Z_i^{\mathrm{T}}$, где $\{U_i\}_{i=1}^{d}$ — ортонормированный базис в пространстве индивидов.

   Пусть $\mathbb{Z} = \{z_{ij}\}_{i=1,j=1}^{n,d} = [Z_1 : \ldots, Z_d]$. Тогда

$$z_{ij} = \underbrace{\langle \mathbf{x}_i, U_j\rangle_1}_{\substack{j\text{-я коорд. } i\text{-ого индивида} \\ \text{в новом базисе}}}.$$

   Так как индивиды не центрированы и не нормированы, это равенство продолжить по аналогии с предыдущим пунктом не можем. Но можем выписать следующее. Пусть $\alpha(\mathbf{x}_i, U_j)$ — угол между $\mathbf{x}_i$ и U_j. Тогда

$$\cos(\alpha(\mathbf{x}_i, U_j)) = \frac{\langle \mathbf{x}_i, U_j\rangle_1}{\|\mathbf{x}_i\|\|U_j\|} = \frac{\langle \mathbf{x}_i, U_j\rangle_1}{\|\mathbf{x}_i\|}.$$

Эти два пункта помогают ответить на вопрос — *как выявить индивидов, которые плохо описываются плоскостью первых двух компонент?*

Чтобы посчитать, как хорошо индивид описывается плоскостью, надо посчитать косинус угла между плоскостью и индивидом. Очевидно, что индивиды, перпендикулярные плоскости, плохо описываются этой плоскостью. Если есть ортогональный базис, то верно следующее: $\cos^2$(угла между вектором и проекцией на плоскость)$=\cos^2$(угла между вектором и 1-ым элементом базиса)$+\cos^2$(угла между вектором и 2-ым элементом базиса).

Пусть $Y_i \in \mathbb{R}^p$ — индивид. Тогда косинус угла между ним и плоскостью первых двух главных компонент:

$$\cos^2(\alpha(Y_i, \mathrm{span}(U_1, U_2))) = \frac{\overbrace{\langle Y_i, U_1\rangle^2}^{z_{i1}^2}}{\|Y_i\|^2} + \frac{\overbrace{\langle Y_i, U_2\rangle^2}^{z_{i2}^2}}{\|Y_i\|^2} = \frac{z_{i1}^2}{\sum\limits_{j=1}^{d} z_{ij}^2} + \frac{z_{i2}^2}{\sum\limits_{j=1}^{d} z_{ij}^2}.$$

Мы получили, что, складывая квадраты нормированных строк $\mathbb{Z}$, мы можем получать квадраты косинусов углов между индивидами и плоскостью. Пусть признаки стандартизованы. Аналогично можем получить, что

$$\cos^2(\alpha(X_j, \mathrm{span}(V_1, V_2))) = f_{j1}^2 + f_{j2}^2.$$

Если все центрировано, то косинус можно назвать корреляцией (формулы для косинуса и коэффициента корреляции совпадут). Тогда получим, что множественный коэффициент корреляции равен сумме квадратов обычных корреляций, то есть

$$R^2(X_j; V_1, V_2) = \rho^2(X_j, V_1) + \rho(X_j, V_2).$$

Замечание 7. $\mathbb{U} = [U_1 : \ldots : U_d]$, $\mathbb{F} = [F_1 : \ldots : F_d]$

1. $\sum\limits_{i=1}^{p} u_{ij}^2 = \|U_j\|^2 = 1$
2. $\sum\limits_{j=1}^{d} u_{ij}^2 = 1$
3. $\sum\limits_{i=1}^{p} f_{ij}^2 = \|F_j\|^2 = \lambda_j$
4. $\sum\limits_{j=1}^{d} f_{ij}^2 = \sum\limits_{j=1}^{d} \langle X_i, V_j\rangle_2^2 = \|X_i\|_2^2 = \begin{cases} s^2(X_i), & \textit{по ковариац. матр.} \\ 1, & \textit{по корреляц. матр.} \end{cases}$

Скалярное произведение $\langle X_i, V_1\rangle_2^2$ характеризует то, насколько 1-ый новый признак описывает исходный. Если рассмотреть $\langle X_i, V_1\rangle_2^2 + \langle X_i, V_2\rangle_2^2$, то это то, насколько первых два новых признака описывают старый.

2.7. Интерпретация главных компонент. Смысл первой главной компоненты в случае положительных ковариаций

Теорема 1 (Перрона-Фробениуса)**.** *Пусть $\mathbb{A}$ — симметричная, неотрицательно определенная матрица, ее элементы положительны. Тогда все компоненты ее первого собственного вектора U_1 будут одного знака.*

Таким образом, *если все корреляции (ковариации) положительны, то все компоненты U_1 одного знака.* Это определяет смысл первой главной компоненты (в случае положительных ковариаций). Тогда первая главная компонента является линейной комбинацией старых признаков с положительными коэффициентами. Это можно проинтерпретировать как некий общий уровень чего-либо (например, общий уровень ученика, если признаки — оценки в школе). Остальные главные компоненты можно также интерпретировать исходя из коэффициентов перед старыми признаками.[7]

2.8. Выбор числа главных компонент

Приведем некоторые варианты выбора числа главных компонент.

1. Задается процент P и берется τ компонент:
$$\frac{\sum\limits_{i=1}^{\tau}\lambda_i}{\sum\limits_{i=1}^{d}\lambda_i} > P\%,$$
то есть чтобы компоненты несли в себе не менее $P\%$ информации.

2. Правило Кайзера. Выбираются главные компоненты, информативность которых больше средней информативности:
$$i: \quad \lambda_i > \frac{\sum\limits_{i=1}^{p} s^2(X_i)}{p} = \frac{\sum\limits_{i=1}^{d}\lambda_i}{p}.$$
Если АГК по корреляционной матрице, то $i: \quad \lambda_i > 1$.

3. Правило сломанной трости. Пусть $\mu_i = \frac{\lambda_i}{\sum\limits_{i=1}^{d}\lambda_i}$. Числа μ_i делят отрезок $[0,1]$ на неравные части. Получаем разбиение $0 < \mu_1 < \mu_2 < \ldots < \mu_{d-1} < 1$. Выбирается

[7] Напомним, что коэффициентами линейной комбинации являются элементы векторов U_i.

компонента, длина которой больше средней длины кусочка случайно сломанной трости.

4. Правило каменистой осыпи. Строим график упорядоченных по убыванию собственных чисел (scree plot). С какого-то момента собственные числа начинают медленно меняться. Берем компоненты до этого момента, то есть пока собственные числа отличаются друг от друга.

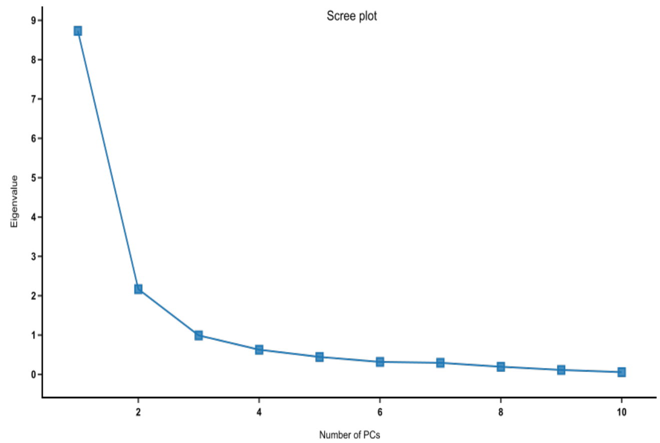

Рис. 2.1. Scree plot

5. Интерпретируем столько компонент, сколько можем.

2.9. Оптимизация в АГК в терминах ковариационных матриц

Предложение 6. $\mathbb{Y} = \mathbb{X}^{\mathrm{T}}$.

Задача $\|\mathbb{Y} - \widetilde{\mathbb{Y}}\| \to \min\limits_{\widetilde{\mathbb{Y}}:\ rank\widetilde{\mathbb{Y}} \leq r}$ [8] *эквивалентна задаче* $\|\mathbb{Y}\mathbb{Y}^{\mathrm{T}} - \widetilde{\mathbb{Y}}\widetilde{\mathbb{Y}}^{\mathrm{T}}\| \to \min\limits_{\widetilde{\mathbb{Y}}:\ rank\widetilde{\mathbb{Y}} \leq r}$.

Если матрица центрирована, то задача $\|\mathbb{Y}\mathbb{Y}^{\mathrm{T}} - \widetilde{\mathbb{Y}}\widetilde{\mathbb{Y}}^{\mathrm{T}}\| \to \min\limits_{\widetilde{\mathbb{Y}}:\ \mathrm{rank}\widetilde{\mathbb{Y}} \leq r}$ эквивалентна следующей задаче:

$$\|\mathbb{S} - \widetilde{\mathbb{S}}\| \to \min_{\widetilde{\mathbb{S}}:\ \mathrm{rank}\widetilde{\mathbb{S}} \leq r},$$

где $\mathbb{S}$ — ковариационная матрица для наших данных, а $\widetilde{\mathbb{S}}$ — какая-то ковариационная матрица (симметричная, неотрицательно определенная).

[8] решение этой задачи — сумма собственных троек, соответствующих первым r главным компонентам.