

Краткие ответы: коллоквиум по PCA

<https://github.com/Kapatsa/Multivariate-Stats>

4 декабря 2021 г.

Содержание

1	Общая информация	2
1.1	Введённые обозначения	2
1.2	Замечания про умножение матриц	2
1.3	Основные свойства ортогональных матриц	3
2	Краткие ответы	4
2.1	Многомерное нормальное распределение. Вектор мат. ож. и ковар. матрица при линейном преобразовании (умножении на матрицу)	4
2.2	Оценки вектора мат. ожиданий и ковар. матрицы. Несмещённая оценка ковар. матрицы	4
2.3	Распределение вектора выборочных средних	4
2.4	Переход к новым признакам с помощью ортогональной матрицы. Пример про способности по математике и физике (выписать матрицу вращения)	5
2.5	Сингулярное разложение, как строится	5
2.6	Сингулярное разложение, в каком смысле оно единственно	6
2.7	Выборочный анализ главных компонент и сингулярное разложение, общее и различия	6
2.8	Почему главные компоненты так называются, в каком смысле они главные	7
2.9	Оптимальность сингулярного разложения в смысле аппроксимации матрицей ранга r	7
2.10	Оптимальность сингулярного разложения в смысле аппроксимации подпространством размерности r	8
2.11	Оптимальность в АГК в статистической терминологии (через дисперсии)	8
2.12	Оптимизация в АГК в терминах ковариационных матриц	9
2.13	В двух статистических пакетах получились разные главные компоненты. Отчего так могло получиться?	9
2.14	Смысл первой ГК, если все ковариации (корреляции) исходных признаков положительны	10
2.15	АГК по корреляционной и по ковариационной матрице. Когда что использовать	10
2.16	Способы выбора числа главных компонент	10
2.17	Почему доля собственного числа по отношению к сумме собственных чисел называется объяснённой долей общей дисперсии?	11
2.18	На основе каких элементов сингулярного разложения интерпретируются главные компоненты как линейные комбинации исходных признаков? Привести формулу и пример	11

2.19	АГК с точки зрения построения базиса в пространстве индивидов и в пространстве признаков. Координаты в новых базисах	11
2.20	Как вычислить значения главных компонент для индивида, которого не было в исходной выборке. А как вычислить значения факторных значений? . .	12
2.21	В каком случае координаты в ортонормированном базисе можно назвать корреляциями?	12
2.22	Чему равны суммы квадратов по строкам и по столбцам в матрице, составленной из собственных векторов в АГК?	13
2.23	Чему равны суммы квадратов по строкам и по столбцам в матрице факторных нагрузок в АГК?	13
2.24	Как интерпретировать скалярное произведение строк в матрице факторных нагрузок в АГК?	13
2.25	Как нарисовать исходные орты в плоскости первых двух главных компонент? .	13
2.26	Зачем и когда первые две координаты факторных нагрузок рисуются в единичном круге?	13
2.27	Чему равна норма i -го вектора из главных компонент?	13

1 Общая информация

ВНИМАНИЕ! В конспекте есть опечатки. Если вы их заметите, можно помочь себе и остальным, исправив ошибку в гитхабе, сделав соответствующий коммит.

<https://github.com/Kapatsa/Multivariate-Stats>

1.1 Введённые обозначения

- a, b, c, \dots — скалярные величины;
- ξ, η, \dots — случайные величины;
- ξ, η, \dots — случайные векторы;
- $\mathbf{X}, \mathbf{Y}, \dots$ — матрицы;
- X_i — вектор-столбцы матрицы \mathbf{X} ;
- \mathbf{x}_i — вектор-строки матрицы \mathbf{X} ;

1.2 Замечания про умножение матриц

Умножение матриц бывает полезным понимать как некоторые операции над столбцами и строками перемножаемых матриц \mathbf{A} и \mathbf{B} . Рассмотрим две такие интерпретации.

Скалярное произведение Пожалуй, самая простая из интерпретаций — через скалярное произведение столбцов \mathbf{A} на строки \mathbf{B} . Рассмотрим наглядный пример.

$$\mathbf{AB} = \begin{pmatrix} \boxed{a_{11}} & \boxed{a_{12}} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} \boxed{b_{11}} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} \boxed{\langle \mathbf{a}_1, \mathbf{B}_1 \rangle} & \langle \mathbf{a}_1, \mathbf{B}_2 \rangle \\ \langle \mathbf{a}_2, \mathbf{B}_1 \rangle & \langle \mathbf{a}_2, \mathbf{B}_2 \rangle \end{pmatrix}$$

Очевидно, что если хотим рассмотреть скалярные произведения столбцов \mathbf{A} на столбцы \mathbf{B} , можно использовать транспонирование: $\mathbf{A}^T \mathbf{B}$. Аналогично для скалярного произведения

строк: \mathbf{AB}^T . Если вспомнить, что скалярные произведения имеют наглядную геометрическую интерпретацию, данный аппарат может быть весьма полезен.

Линейная комбинация Ещё одна очень полезная интерпретация заключается в том, что строки результата произведения матриц есть линейная комбинация столбцов исходной. Покажем это в двумерном случае:

$$\begin{aligned}\mathbf{AB} &= \begin{pmatrix} \boxed{a_{11}} & \boxed{a_{12}} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} \boxed{b_{11}} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} \boxed{a_{11}b_{11} + a_{12}b_{21}} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix} = \\ &= \begin{pmatrix} \boxed{b_{11}A_1 + b_{21}A_2} & b_{12}A_1 + b_{22}A_2 \end{pmatrix}\end{aligned}$$

Таким образом, если столбцы матрицы \mathbf{A} представляют из себя некоторый базис, то элементы столбцов матрицы \mathbf{B} есть координаты полученного вектор-столбца в этом базисе.

Новый признак как линейная комбинация старых

Пусть мы хотим получить новый признак η как линейную комбинацию старых признаков $\xi_1, \xi_2, \dots, \xi_n$. Так и запишем:

$$\eta = a_1\xi_1 + a_2\xi_2 + \dots + a_n\xi_n = \mathbf{a}^T \boldsymbol{\xi}.$$

При этом, если ковариационная матрица для $\boldsymbol{\xi}$ есть $\boldsymbol{\Sigma}$, то

$$\text{D}\eta = \mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a}.$$

Теперь с помощью линейных преобразований над элементами вектора $\boldsymbol{\xi}$ получим d -мерную случайную величину $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_d)^T$:

$$\boldsymbol{\eta} = \begin{pmatrix} \eta_1 = \mathbf{a}_1^T \boldsymbol{\xi} \\ \eta_2 = \mathbf{a}_2^T \boldsymbol{\xi} \\ \vdots \\ \eta_d = \mathbf{a}_d^T \boldsymbol{\xi} \end{pmatrix} = \mathbf{A}^T \boldsymbol{\xi}.$$

Если $\mathbf{A} = (\mathbf{a}_1 | \mathbf{a}_2 | \dots | \mathbf{a}_d) \in \mathbb{R}^{n \times d}$, то ковариационная матрица для $\boldsymbol{\eta}$ имеет вид:

$$\text{Cov}(\boldsymbol{\eta}) = \mathbf{A}^T \boldsymbol{\Sigma} \mathbf{A}.$$

1.3 Основные свойства ортогональных матриц

Ортогональной матрицей называется такая матрица \mathbf{A} , что $\mathbf{A}^T \mathbf{A} = \mathbf{A} \mathbf{A}^T = \mathbf{E}$, где \mathbf{E} — единичная матрица.

Из определения очевидно, что ортогональность матрицы \mathbf{A} равносильна тождеству

$$\mathbf{A}^T = \mathbf{A}^{-1}.$$

Подробнее рассмотрим произведение $\mathbf{A}^T \mathbf{A}$:

- С одной стороны, $\mathbf{A} \mathbf{A}^T$ — это матрица скалярных произведений строчек матрицы \mathbf{A} на строчки \mathbf{A} ;

- С другой стороны, так как \mathbf{A} — ортогональна, так что по определению $\mathbf{A}\mathbf{A}^T = \mathbf{E}$.

Отсюда сразу следует, что строки исходной матрицы \mathbf{A} ортогональны между собой. Более того, так как на диагонали полученной матрицы $\mathbf{A}\mathbf{A}^T$ единицы, то норма строчек $\|A_i\| = \sqrt{\langle A_i, A_i \rangle} = 1$, то есть они ещё и нормированы. Аналогично для столбцов.

Важное свойство ортогональных матриц: умножение на ортогональную матрицу означает поворот или отражение. При этом не меняются нормы векторов и углы между ними.

2 Краткие ответы

2.1 Многомерное нормальное распределение. Вектор мат. ож. и ковар. матрица при линейном преобразовании (умножении на матрицу)

Пусть с.в. $\xi \sim N(\mathbf{a}, \Sigma)$, где $\mathbf{a} \in \mathbb{R}^n$ — вектор средних, $\Sigma \in \mathbb{R}^{n \times n}$ — ковар. матрица. Плотность многомерного нормального распределения:

$$p(\mathbf{x}) = \frac{|\Sigma|^{-1/2}}{(2\pi)^{1/n}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{a})^T \Sigma^{-1}(\mathbf{x} - \mathbf{a})\right).$$

Пусть \mathbf{A} — некоторая матрица $m \times n$. Пусть есть с.в. $\xi \in \mathbb{R}^n$ с м.о. \mathbf{a} и ковариационной матрицей Σ . Как будут выглядеть м.о. и ковар. матрица с.в. $\mathbf{A}\xi$? С м.о. всё понятно:

$$E(\mathbf{A}\xi) = \mathbf{A}E(\xi) = \mathbf{A}\mathbf{a},$$

Теперь ковариация:

$$\begin{aligned} \text{Cov}(\mathbf{A}\xi) &= E(\mathbf{A}\xi - E\mathbf{A}\xi)(\mathbf{A}\xi - E\mathbf{A}\xi)^T = E(\mathbf{A}\xi - \mathbf{A}\mathbf{a})(\mathbf{A}\xi - \mathbf{A}\mathbf{a})^T = \\ &= E(\mathbf{A}(\xi - \mathbf{a}))(\mathbf{A}(\xi - \mathbf{a}))^T = E(\mathbf{A}(\xi - \mathbf{a}))((\xi - \mathbf{a})^T \mathbf{A}^T) = \mathbf{A}E((\xi - \mathbf{a})(\xi - \mathbf{a})^T) \mathbf{A}^T = \mathbf{A}\text{Cov}(\xi) \mathbf{A}^T \end{aligned}$$

2.2 Оценки вектора мат. ожиданий и ковар. матрицы. Несмещённая оценка ковар. матрицы

Здесь всё просто. Пусть векторы \mathbf{x}_i , $i \in 1 : n$ представляют из себя выборку некоторого многомерного случайного вектора ξ . С вектором средних всё понятно, обозначим его $\bar{\mathbf{x}}$. Оценка ковариационной матрицы строится по определению.

$$\widehat{\text{Cov}}(\xi) = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

С несмещённой оценкой вот так:

$$\widehat{\text{Cov}}(\xi)_{unbiased} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

2.3 Распределение вектора выборочных средних

Если случайный вектор имеет конечный второй момент, то верно следующее:

$$\bar{\mathbf{x}} \rightsquigarrow N\left(\mathbf{a}, \frac{\Sigma}{n}\right)$$

В частном случае нормального распределения значок асимптотики \sim можем убрать.

2.4 Переход к новым признакам с помощью ортогональной матрицы. Пример про способности по математике и физике (выписать матрицу вращения)

С помощью ортогональной матрицы можно перейти к новым признакам. Возьмём матрицу $\mathbf{X} = (X_1|X_2|\dots|X_p)$ и рассмотрим произведение \mathbf{AX} , где $\mathbf{A} \in \mathbb{R}^{n \times n}$ — ортогональная матрица. Хорошо известно, что применение ортогонального оператора к матрице задаёт некоторый поворот или отражение вектор-столбцов матрицы \mathbf{X} . При этом каждый столбец результирующей матрицы \mathbf{AX} будет линейной комбинацией столбцов матрицы \mathbf{A} .

Пример про способности по математике и физике

Пусть признаки такие:

- X_1 — оценки по математике;
- X_2 — оценки по физике;
- X_3 — оценки по русскому;
- X_4 — оценки по литературе.

Итак, $\mathbf{X} = (X_1|X_2|X_3|X_4)$. Домножением матрицы \mathbf{X} на \mathbf{A} получим матрицу с двумя новыми признаками Z_1 и Z_2 , которые отвечают за общую оценку по техническим и гуманитарным наукам.

$$\mathbf{XA} = \underbrace{(X_1|X_2|X_3|X_4)}_{\mathbf{X}} \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}}_{\mathbf{A}} = (X_1 + X_2 \quad X_3 + X_4).$$

2.5 Сингулярное разложение, как строится

Сингулярным разложением матрицы $\mathbf{Y} := \mathbf{X}^T$ называется её разложение в сумму следующего вида:

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

где λ_i — собственные числа матрицы \mathbf{YY}^T , U_i — её собственные векторы, V_i — собственные векторы матрицы $\mathbf{Y}^T\mathbf{Y}$, d — её ранг.

Пусть у нас L признаков и K индивидов. Пусть также $\mathbf{Y} := \mathbf{X}^T$. Возьмём матрицу

$$\mathbf{S} = \mathbf{YY}^T \in \mathbb{R}^{L \times L}.$$

Пусть λ_i — собственные числа матрицы, U_i — её собственные векторы, что по определению означает

$$\mathbf{S}U_i = \lambda_i U_i.$$

Собственных чисел у нас L штук; теперь, если упорядочить их по убыванию, то число *ненулевых* собственных чисел будет равно рангу матрицы \mathbf{S} . Число d — ранг матриц \mathbf{Y} и \mathbf{S} ,¹ он же ранг пространства столбцов и строк \mathbf{Y} , будет не больше, чем $\min(L, K)$.

¹ядра этих операторов совпадают, поэтому можно показать, что их ранги равны

Важно, что $\{U_i\}_{i=1}^d$ образуют *ортонормированный базис пространства столбцов* матрицы \mathbf{Y} .² Введём также векторы V_i , которые, как утверждается, сформируют ортонормированный базис пространства строк матрицы \mathbf{Y} и будут собственными векторами матрицы $\mathbf{Y}^T \mathbf{Y}$:

$$V_i = \frac{\mathbf{Y}^T U_i}{\sqrt{\lambda_i}}.$$

Можно проверить, что тогда

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

Формула сингулярного разложения имеет вид:

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

V_i — это новые признаки. В матричном виде:

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

2.6 Сингулярное разложение, в каком смысле оно единственно

Если у нас есть некоторое сингулярное разложение с векторами U_i и V_i , ничего не поменяется, если возьмём в качестве множителей векторы $-U_i$ и $-V_i$.

Также может возникнуть неединственность, когда у нас есть два или более равных собственных числа. Тогда любая нетривиальная линейная комбинация векторов, соответствующих этому собственному числу, будет давать собственный вектор, соответствующий тому же собственному числу. Другими словами, можно выбрать любой ортонормированный базис соответствующего подпространства в качестве новых собственных векторов.

Поэтому можно сформулировать следующее утверждение о единственности сингулярного разложения:

Предложение 1 (Единственность SVD) Пусть $L \leq K$. Если $\mathbf{Y} = \sum_{i=1}^L c_i P_i Q_i^T$ есть некоторое биортогональное разложение матрицы, то есть:

1. $c_1 \geq c_2 \geq \dots \geq c_L \geq 0$;
2. $\{P_i\}$ и $\{Q_i\}$ — ортонормированные.

Любое биортогональное разложение есть SVD.

2.7 Выборочный анализ главных компонент и сингулярное разложение, общее и различия

(?) Когда сравниваем разложение из SVD и PCA, у нас \mathbf{Y} в обоих случаях уже с центрированными столбцами? (иначе ведь они отличаются друг от друга; в конспекте написано не совсем понятно когда центрируем столбцы матрицы)

(?) Ещё про замечание про норму матрицы $L \times K$; $\|\cdot\|_{1,2}$ — это чисто сумма квадратов, делённая на K , верно?

Сначала про главные отличия, а потом уже про детали.

²если вернуться к матрице \mathbf{X} , то это означает *ортонормированный базис пространства строк-индивидов*

1. В SVD столбцы и строки равноправны, в PCA — нет. Подробности — в деталях;
2. В АГК имеется предположение о том, что признаки центрированы (АГК по ковариационной матрице), также возможна их нормировка (АГК по корреляционной матрице)

Если для сингулярного разложения матрицы \mathbf{Y} слова «ортонормированный вектор U_i » вполне ясны и не требуют разъяснений, ибо рассматривается стандартная евклидова норма $\|\mathbf{a}\|_1^2 = a_1^2 + \dots + a_n^2$, то в случае анализа главных компонент матрица данных $\mathbf{X} = \mathbf{Y}^T$ имеет определённую структуру, где строки есть индивиды, а столбцы — признаки. В статистике разумно назначать каждому индивиду вес $\frac{1}{K}$, где K — количество индивидов, а признакам — единицу. Поэтому норма для пространства строк-признаков матрицы \mathbf{Y} выбирается как корень из среднего арифметического квадратов, а норма для пространства столбцов-индивидов \mathbf{Y} — как корень из суммы квадратов.

Пусть у нас есть разложение $\mathbf{Y} = \sum_{i=1}^d \sqrt{\lambda_i} U_i V_i^T$. Так как $U_i \in \mathbb{R}^{L \times 1}$ есть ортонормированный базис пространства **столбцов** матрицы \mathbf{Y} , то норма в случае PCA не меняется, и следовательно в «новом» разложении $\mathbf{Y} = \sum_{i=1}^d \sqrt{\tilde{\lambda}_i} \tilde{U}_i \tilde{V}_i^T$ получается, что $\tilde{U}_i = U_i$. Для строк получается $\tilde{V}_i = \sqrt{K} V_i$, а для собственных значений $\tilde{\lambda}_i = \frac{\lambda_i}{K}$.

2.8 Почему главные компоненты так называются, в каком смысле они главные

Пусть у нас есть векторы $Y_1, \dots, Y_K \in \mathbb{R}^L$. Хотим найти вектор P с единичной нормой такой, который наилучшим образом описывает указанные элементы \mathbb{R}^L . Как известно, длина проекции некоторого вектора X на вектор P равна $|\langle X, P \rangle|$; чем больше длина проекции, тем «ближе» вектор P к вектору X .

Можем сформулировать оптимизационную задачу, которая найдёт такую точку в \mathbb{R}^L , которая будет самой «близкой» к Y_1, \dots, Y_K :

$$P_1 = \arg \max_{\tilde{P}} \sum_{i=1}^L \langle Y_i, \tilde{P} \rangle^2$$

Данную компоненту будем называть **главной**.

Для того, чтобы найти следующий близкий к Y_1, \dots, Y_K вектор, рассматриваем ту же задачу, только уже в ортогональном для P_1 подпространстве, и так далее, пока не исчерпаем размерность. Оказывается, что эти максимумы достигаются на векторах U_i , при этом максимумы суммы квадратов длин соответствующих проекций достигаются на λ_i . Так что компоненты главные, так как они наилучшим образом описывают совокупность старых признаков.

Заметим также, что длины соответствующих проекций можно найти в $Z_i = \mathbf{X} U_i = \sqrt{\lambda_i} V_i$.

2.9 Оптимальность сингулярного разложения в смысле аппроксимации матрицей ранга r

Сразу приведём предложение.

Предложение 2 (Аппроксимация матрицы матрицей меньшего ранга)

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

для произвольной матрицы $\mathbf{Y} \in \mathbb{R}^{L \times K}$ ранга d выполняется

$$\min_{\tilde{\mathbf{Y}} \in M_r} \|\mathbf{Y} - \tilde{\mathbf{Y}}\|_F^2 = \sum_{i=r+1}^d \lambda_i,$$

причём минимум достигается на матрице $\sum_{i=1}^d \sqrt{\lambda_i} U_i V_i^T$.

2.10 Оптимальность сингулярного разложения в смысле аппроксимации подпространством размерности r

Сразу приведём предложение.

Предложение 3 (Аппроксимация подпространством)

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

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

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

2.11 Оптимальность в АГК в статистической терминологии (через дисперсии)

(?) На странице 21: получается, что норма $\|\cdot\|_{1,2}$ одинакова у \mathbf{X} и \mathbf{X}^T . Меня немного смущает то, на что мы делим в этом случае: у нас же столбцы и строки неравноправны, то есть в случае \mathbf{X}^T мы должны делить сумму квадратов на K , а в случае \mathbf{X} на L . (исправлено)

(?) На стр. 22: знаем, что квадрат нормы Z_i есть его s^2 . Почему? Это потому, что Z_i является линейной комбинацией *центрированных столбцов* матрицы \mathbf{X} ?

(?) На стр. 22: не понял с нормой; $\|(\sqrt{\lambda_i} U_i V_i^T)^T\|_{1,2}^2$? (исправлено)

Можно перейти на выборочный язык, то есть $n = K, p = L$, но пока останемся в старом варианте.

Применяя АГК к матрице \mathbf{Y} , получаем новые признаки, которые ранжированы по убыванию своей значимости в контексте объяснённых различий между индивидами. Покажем, что это имеет прямое отношение к дисперсиям признаков.

По умолчанию, столбцы X_i центрированы, поэтому по определению

$$\|X_i\|_2^2 = \frac{1}{K} \sum_{j=1}^K ((X_i)_j)^2 = s^2(X_i).$$

В то же время, когда делаем АГК, ввиду ортонормированности V_i по норме $\|\cdot\|_2$ а также ввиду того, что Z_i есть линейная комбинация центрированных столбцов X_i

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

Поэтому, так как $\mathbf{X}^T \in \mathbb{R}^{L \times K}$, её ранг равен d , имеет место цепочка равенств

$$\begin{aligned} \|\mathbf{X}^T\|_{1,2}^2 &= \sum_{i=1}^L s^2(X_i) \\ &= \sum_{i=1}^d \|(\sqrt{\lambda_i} U_i V_i^T)\|_{1,2}^2 = \sum_{i=1}^d \lambda_i = \sum_{i=1}^d s^2(Z_i), \end{aligned}$$

то есть сумма дисперсий (*total variance*) не меняется при переходе к новым признакам.

Таким образом, можем первую главную компоненту интерпретировать как такой признак, который несёт в себе наибольшую часть информации о предыдущих признаках в совокупности. И так далее для второй, третьей, и т. д. Другими словами, главные компоненты являются ортогональными между собой линейными комбинациями исходных признаков, которые обладают свойством оптимальности.

2.12 Оптимизация в АГК в терминах ковариационных матриц

(?) $\|\cdot\|_{1,2}^2$? (исправлено)

Пусть $\mathbf{Y} = \mathbf{X}^T$ — транспонированная матрица данных.

Предложение 4 (Об эквивалентности двух задач оптимизации)

Рассмотрим две оптимизационные задачи:

$$\min_{\tilde{\mathbf{Y}} \in M_r} \|\mathbf{Y} - \tilde{\mathbf{Y}}\|^2, \quad (1)$$

$$\min_{\tilde{\mathbf{Y}} \in M_r} \|\mathbf{Y}\mathbf{Y}^T - \tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T\|^2. \quad (2)$$

Задачи 1 и 2 эквивалентны. Их решение строится как сумма первых собственных троек, которые соответствуют первым r главным компонентам.

Теперь, если матрица центрирована, то задача 2 будет эквивалентна задаче нахождения «ближайшей» по оговорённой норме ковариационной матрицы меньшего ранга r .

2.13 В двух статистических пакетах получились разные главные компоненты. Отчего так могло получиться?

Это может произойти по нескольким причинам:

1. Уже говорили про единственность разложения: компоненты U_i , V_i могут отличаться знаком, а также то, что некоторые λ_i могут совпадать;
2. Может быть такое, например, что вектор главных компонент пакетом отображается как V_i , а не $\sqrt{\lambda_i}V_i$;
3. Для нахождения собственных чисел и собственных векторов применяются разные численные методы, что влечёт различия в ошибке;
4. Процедура может стандартизировать данные перед нахождением компонент, а это изменит решение если в другом случае данные не приведены к такому же виду.

2.14 Смысл первой ГК, если все ковариации (корреляции) исходных признаков положительны

На этот вопрос отвечает

Теорема 1 (Перрона–Фробениуса) Пусть матрица A — симметричная, неотрицательно определённая с положительными элементами. Тогда все компоненты её первого собственного вектора будут одного знака.

В нашем конкретном случае теорема утверждает, что если выборочные корреляции/ковариации положительны, то все координаты U_1 будут положительными. Так как вектор U_1 состоит из коэффициентов линейной комбинации старых признаков (с одинаковыми знаками), то первую главную компоненту можем интерпретировать как общий уровень какой-то характеристики.

2.15 АГК по корреляционной и по ковариационной матрице. Когда что использовать

Если *только* центрируем наши признаки, то говорим, что АГК делаем **по ковариационной матрице**, так как

$$\frac{1}{n} \mathbf{X}_{(c)}^T \mathbf{X}_{(c)}$$

есть выборочная ковариационная матрица, если же *к тому же* стандартизируем, то получается выборочная корреляционная матрица

$$\frac{1}{n} \mathbf{X}_{(s)}^T \mathbf{X}_{(s)}$$

и мы говорим, что делаем АГК **по корреляционной матрице**.

Когда что использовать? Если достоверно известно, что разные признаки имеют разную шкалу, но при этом признаки рассматриваются как равноправные между собой, то их нужно стандартизировать. Если у признаков одинаковые шкалы, то это необязательно.

2.16 Способы выбора числа главных компонент

(?) Про правило сломанной кости, если можно, повторить. Там случайно не сумма должна стоять сверху? (там надо исправлять)

Есть несколько вариантов:

1. Берётся столько компонент, сколько в совокупности даёт информативность не менее определённого числа;
2. **Правило Кайзера**: берутся те компоненты, информативность которых больше средней;
3. **Правило сломанной трости**: (? , не принципиально)
4. **Правило каменистой осыпи**: строим график собственных чисел по убыванию; берём компоненты до момента, когда собственные числа начинают мало отличаться друг от друга.

2.17 Почему доля собственного числа по отношению к сумме собственных чисел называется объясненной долей общей дисперсии?

Для центрированных признаков дисперсия выглядит так:

$$\|X_i\|_2^2 = \frac{1}{K} \sum_{j=1}^K ((X_i)_j)^2 = s^2(X_i).$$

В то же время, когда делаем АГК, ввиду ортонормированности V_i по норме $\|\cdot\|_2$ а также ввиду того, что Z_i есть линейная комбинация центрированных столбцов X_i

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

Следующее равенство показывает, что суммы дисперсий и сумма собственных чисел это одно и то же:

$$\begin{aligned} \|\mathbf{X}^T\|_{1,2}^2 &= \sum_{i=1}^L s^2(X_i) \\ &= \sum_{i=1}^d \|(\sqrt{\lambda_i}U_iV_i^T)^T\|_{1,2}^2 = \sum_{i=1}^d \lambda_i = \sum_{i=1}^d s^2(Z_i), \end{aligned}$$

Поэтому для каждого нового признака Z_i выражение

$$\frac{\lambda_i}{\sum_{i=1}^d \lambda_i}$$

будет эквивалентно

$$\frac{s^2(Z_i)}{\sum_{i=1}^d s^2(Z_i)} = \frac{s^2(Z_i)}{\sum_{i=1}^d s^2(X_i)},$$

что и требовалось показать.

2.18 На основе каких элементов сингулярного разложения интерпретируются главные компоненты как линейные комбинации исходных признаков? Привести формулу и пример

Формула, которая это объясняет:

$$Z_i = \mathbf{X}U_i$$

Напомним, что U_i есть собственные векторы матрицы $\mathbf{X}^T\mathbf{X}$.

2.19 АГК с точки зрения построения базиса в пространстве индивидов и в пространстве признаков. Координаты в новых базисах

Итак, по построению

- $\{U_i\}_{i=1}^d \in \mathbb{R}^L$ — ортонормированный базис в пространстве индивидов/строчек матрицы \mathbf{X} ;

- $\{V_i\}_{i=1}^d \in \mathbb{R}^K$ — ортонормированный базис в пространстве признаков/столбцов матрицы \mathbf{X} .

Основная формула для АГК, которая нам должна помочь:

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

где $F_i = \sqrt{\lambda_i} U_i$ — вектор i -ых факторных компонент, $Z_i = \sqrt{\lambda_i} V_i$ — вектор главных компонент.

Базис в пространстве признаков Если составить матрицу \mathbf{F} из столбцов F_i , то элементы этой матрицы будут координатами *старых признаков* в новом базисе признаков.

$$f_{ij} = \langle X_i, V_j \rangle_2 = \begin{cases} \text{Cov}(X_i, V_j), & \text{если АГК по ковариационной матрице} \\ \rho(X_i, V_j) = \rho(X_i, Z_j), & \text{если по корреляционной матрице} \end{cases}$$

f_{ij} — j -ая координата i -го признака в новом базисе признаков. Таким образом, чтобы найти в матрице \mathbf{F} координаты первого старого признака, мы смотрим на первую строчку.

Базис в пространстве индивидов Если составить матрицу \mathbf{Z} из столбцов Z_i , то элементы этой матрицы будут координатами индивидов в новом базисе признаков.

$$z_{ij} = \langle \mathbf{x}_i, U_j \rangle_1$$

z_{ij} — j -ая координата i -го индивида в новом базисе признаков.

2.20 Как вычислить значения главных компонент для индивида, которого не было в исходной выборке. А как вычислить значения факторных значений?

(?) Что такое значения факторных значений?

Значения главных компонент для новых индивидов Здесь используем то, что написано в прошлом вопросе. Было:

$$z_{ij} = \langle \mathbf{x}_i, U_j \rangle_1.$$

z_{ij} — j -ая координата i -го индивида в новом базисе признаков. Поэтому точно так же можно взять нового индивида $\tilde{\mathbf{x}}$ и разложить его по базису U_i :

$$\sum_{j=1}^d \langle \tilde{\mathbf{x}}, U_j \rangle_1 U_j.$$

Значения факторных значений (?) мне ответили, но я не понял.

2.21 В каком случае координаты в ортонормированном базисе можно назвать корреляциями?

Координаты в ортонормированном базисе можно назвать корреляциями, если координаты есть некоторые скалярные произведения центрированных и стандартизированных векторов в пространстве с нормой $\|\cdot\|_2$. Можно привести пример с матрицей \mathbf{F} , которая состоит из скалярных произведений вида

$$\langle X_i, V_j \rangle_2.$$

2.22 Чему равны суммы квадратов по строкам и по столбцам в матрице, составленной из собственных векторов в АГК?

Спрашивается про собственные векторы U_i , которые объединили в матрицу U . По построению столбцы ортонормированы, поэтому сумма квадратов тут будет единица. Теперь по строчкам: сумма квадратов элементов по строке будет равна

$$\sum_{j=1}^d u_{ij}^2 = \sum_{j=1}^d f_{ij}^2 / \lambda_i = \frac{1}{\lambda_i} \sum_{j=1}^d f_{ij}^2 = \frac{\|X_i\|_2^2}{\lambda_i} = \frac{s^2(X_i)}{\lambda_i}.$$

(?)

2.23 Чему равны суммы квадратов по строкам и по столбцам в матрице факторных нагрузок в АГК?

По определению $F_i = \sqrt{\lambda_i} U_i$, а также

$$f_{ij} = \langle X_i, V_j \rangle_2 = \begin{cases} \text{Cov}(X_i, V_j), & \text{если АГК по ковариационной матрице} \\ \rho(X_i, V_j) = \rho(X_i, Z_j), & \text{если по корреляционной матрице} \end{cases}.$$

Отсюда

$$\sum_{i=1}^p f_{ij}^2 = \|F_j\|^2 = \lambda_j$$
$$\sum_{j=1}^d f_{ij}^2 = \langle X_i, V_j \rangle_2^2 = \|X_i\|_2^2 = \begin{cases} s^2(X_i), & \text{если АГК по ковариационной матрице} \\ 1, & \text{если по корреляционной матрице} \end{cases}.$$

2.24 Как интерпретировать скалярное произведение строк в матрице факторных нагрузок в АГК?

(?)

2.25 Как нарисовать исходные орты в плоскости первых двух главных компонент?

Берём матрицу факторных нагрузок F и её первые два столбца. Первый столбец — x -координаты старого признака, второй — y -координаты.

2.26 Зачем и когда первые две координаты факторных нагрузок рисуются в единичном круге?

Зачем Для того, чтобы понять, какую долю дисперсии старого признака объясняют первые два новых признака. Чем ближе к единичной окружности, тем лучше.

Когда Только когда АГК у нас по корреляционной матрице, иначе total variance не будет равен единице, а значит нет смысла смотреть что-то в единичном круге.

2.27 Чему равна норма i -го вектора из главных компонент?

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