1. Переход к новым признакам

Как раньше, $\mathbb{X} \in \mathbb{R}^{n \times p}$ — матрица данных, столбцы — вектора признаков, строки — индивиды (наблюдения).

$Z_i = \mathbb{X}A_i \in \mathbb{R}^n$, $i = 1, \ldots, d$ — новые признаки (было в прошлый раз), A_i — коэффициенты линейной комбинации. В матричном виде

$$\mathbb{Z} = \mathbb{X}\mathbb{A} \in \mathbb{R}^{n \times d},$$

где столбцы матрицы $\mathbb{A}$ — вектора из коэффициентов линейных комбинаций.

Пусть новые признаки Z_i — ортогональные и их количество равно рангу d матрицы данных $\mathbb{X}$. Тогда эти вектора составляют ортогональный базис пространства столбцов (пространства признаков).

Нормируем их $Q_i = \frac{Z_i}{\|Z_i\|}$, получим ортонормированный базис.

В ортонормированном базисе удобно вычислять координаты вектора через скалярные произведения с базисными векторами.

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

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

***Задание**: Рассмотрим пример: пространство $\mathbb{R}^2$, соответствующий базис $(1,1)^{\mathrm{T}}/\sqrt{2}$, $(1,-1)^{\mathrm{T}}/\sqrt{2}$. Это ортонормированный базис, да? Как вычислить координаты вектора $(5,4)^{\mathrm{T}}$ в данном базисе? Нарисуйте картинку, где отметьте координаты.*

Таким образом, $\{Q_i\}_{i=1}^d$ — ортонормированный базис в пространстве $\mathrm{span}(X_1, \ldots, X_p)$.

Исходные признаки выражаются через новые (есть пространство, в нём базис, каждый вектор этого пространства раскладывается по базису):

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

Если ввести матрицу факторных весов (факторных нагрузок) $\mathbb{F} = \{f_{ij}\}_{i=1,j=1}^{p,d} = [F_1, : \ldots : F_d] \in \mathbb{R}^{p \times d}$, то можно записать разложение (1) в матричном виде:

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

(Здесь строки матрицы $\mathbb{F}$ — коэффициенты разложения исходных признаков по новым, ортонормированным, поэтому она транспонированная.)

Важно: разложение матрицы данных тесно связано с введением новых признаков.

После нормировки $P_j = \frac{F_j}{\|F_j\|}$, получим

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

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

Вывод: для нахождения новых признаков, которые самые лучшие в каком-то смысле, нам понадобятся матричные разложения. Далее мы займемся изучением самого лучшего матричного разложения, но сначала обсудим несколько фактов из линейной алгебры.

2. Пара фактов из линейной алгебры

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

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

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

Обычно, унитарная матрица строится из ортонормированного базиса, который составляет столбцы матрицы. (**Задание**. Проверьте, что в примере с поворотом на ϕ это так, т.е., матрица составлена из базисных векторов.)

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

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

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

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

Задание: какой вид тогда имеет матрица проектора?

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

Поэтому $\mathbf{\Pi} = \mathbb{P}\mathbb{P}^{\mathrm{T}}$.

Задание. Для трехмерного вектора проверить, что его проекция на плоскость первых двух координат вычисляется по такой формуле.

Глава 1

Сингулярное разложение (SVD — Singular Value Decomposition)

Итак, приступаем к изучению самого лучшего, самого красивого, самого оптимального, самого симметричного разложения матриц.

1.1. Как строится сингулярное разложение

Мы сейчас поменяем обозначения и будем раскладывать транспонированную матрицу данных $\mathbb{Y} = \mathbb{X}^{\mathrm{T}}$, представляйте $\mathbb{Y}$ как широкую матрицу с небольшим числом строк и большим числом столбцов.

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. По определению,

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

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

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

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

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

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

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.* $\{V_i\}_{i=1}^{d}$ — *ортонормированная 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.*

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

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

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

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

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.
- $\dfrac{\lambda_i}{\sum_{i=1}^{d} \lambda_i}$ — вклад i-ого признака.

Терминология: $\sqrt{\lambda_i}$ — *сингулярные числа* матрицы $\mathbb{Y}$, U_i — *левые сингулярные вектора*, V_i — *правые сингулярные вектора.*

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

Замечание 1. *Сингулярное разложение — единственное разложение с двумя ортонормированными базисами. Оно симметрично в след. смысле: можем* $\mathbb{Y}$ *транспонировать, проделать всё то же самое, а поменяются местами только* U_i *и* V_i.

Задание. Транспонируйте матрицу $\mathbb{Y}$ и покажите, что, действительно, U_i и V_i поменяются местами (то, что называлось U, станет называться V, и наоборот).

Вернемся к SVD $\mathbb{Y} = \sum\limits_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}} = \sum\limits_{i=1}^{d} \mathbb{Y}_i$.

Введем норму Фробениуса матрицы, квадрат которой равен сумме квадратов элементов матрицы. Так как U_i и V_i по норме равны 1, то можно показать, что $\|\mathbb{Y}_i\|_F^2 = \lambda_i$. А так как λ_i упорядочены по убыванию, то норма $\|\mathbb{Y}_1\|_F$ самая большая, у второй матрицы норма поменьше и т.д. А так как, напомню, разложение связано с введением новых признаков, то и первый новый признак самый важный, и т.д. по убыванию.

Также можно показать, что $\langle \mathbb{Y}_i, \mathbb{Y}_j \rangle_F = 0$ при $i \neq j$; поэтому получаем что-то вроде теоремы Пифагора: $\|\mathbb{Y}\|_F^2 = \sum\limits_{i=1}^{d} \|\mathbb{Y}_i\|_F^2 = \sum\limits_{i=1}^{d} \lambda_i$.

Отсюда становится очевидным, почему вклад i-й матрицы (и, соответственно, i-го признака) можно определить как $\dfrac{\lambda_i}{\sum_{i=1}^{d} \lambda_i}$.

Вопрос: Очевидно?

Задание: доказать, что столбцы матрицы $\mathbb{Y}_i$ состоят из проекций столбцов матрицы $\mathbb{Y}$ на подпространство $\operatorname{span} U_i$.

---

[3] Разложение в сумму элементарных матриц.

[4] Самый важный пункт утверждения.

[5] Они длинные :)

Отступление: SVD можно использовать для компактного хранения данных, если ранг матрицы маленький.

Задание: Посчитайте объем памяти для хранения всей матрицы или ее сингулярного разложения.

Ответ. 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.[6] Таким образом, если, к примеру, матрица близка к квадратной ($L = K$), то при $L > 2d$, выгоднее хранить сингулярное разложение.

1.2. 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,

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

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.

   Задание. Доказать это.

   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.[7]

Задание. Постройте сингулярное разложение матрицы, на диагонали которой стоит число 2, остальные нули.

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

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

[7] 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.

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.***

Замечание 2. *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* $\mathbb{Y}^{\mathrm{T}}\mathbb{Y}$,
- $Q_i = \frac{\mathbb{Y}^{\mathrm{T}} P_i}{\sqrt{\lambda_i}}$, $i = 1, \ldots, d$ *(*$d = \operatorname{rank} \mathbb{Y}\mathbb{Y}^{\mathrm{T}}$*).*

Задание. Является ли разложение матрицы

$$\mathbb{Y} = (1,1)(1,1,1)^{\mathrm{T}} + (-1,1)(1,-1,1)^{\mathrm{T}}$$

сингулярным? Заодно посчитайте, какая получается матрица $\mathbb{Y}$.

А это (матрица другая)

$$\mathbb{Y} = (1,1)(1,1,1)^{\mathrm{T}} + (-1,1)(2,-1,-1)^{\mathrm{T}}?$$

Если разложение сингулярное, выпишите сингулярные тройки, упорядочите их по λ_i.

1.3. Матричный вид сингулярного разложения

Можно записать двумя способами:

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

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

2. Возьмём $\mathbb{U} = [U_1 : \ldots : U_d : U_{d+1} : \ldots : U_L]$ — ортонормированный базис в $\mathbb{R}^L$.[8]

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

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \ldots & \ldots & 0 \\ 0 & \ddots & & 0 & 0 \\ 0 & & \lambda_d & & 0 \\ 0 & & & \ddots & 0 \\ 0 & 0 & \ldots & & 0 \end{pmatrix} \in \mathbb{R}^{L \times K}.$$

   Тогда[9]

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

Задание. Записать в матричной форме предыдущий пример.

1.4. Оптимальные свойства сингулярного разложения

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

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

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

   $\min\limits_{\widetilde{\mathbb{Y}} \in M_r} \|\mathbb{Y} - \widetilde{\mathbb{Y}}\|_F^2 = \sum\limits_{i=r+1}^{d} \lambda_i$ *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 $\leq r$. Тогда

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

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

Задание. Есть две точки на плоскости, размерность пространства $\mathcal{L}_r$ равна $r = 1$. Что минимизируется в этом случае, нарисуйте.

Задание. Есть две точки в трехмерном пространстве, размерность пространства $\mathcal{L}_r$ равна $r = 2$. Чему равен минимум?

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

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

1.5. 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: $\langle Y_i, P\rangle P$, а ее длина равна $|\langle Y_i, P\rangle|$. Проекция измеряет то, насколько это направление соответствует нашим данным. Поставим задачу найти направление, которое лучше всего описывает нашу совокупность точек. Чем больше проекция, тем лучше. Задача:

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

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

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

Предложение 5. *Верно следующее:*

1. $\max\limits_P \sum\limits_{i=1}^{K} \langle Y_i, P\rangle^2 = \lambda_1$ *и достигается на* $P = U_1$,
2. $\max\limits_{P:\ P\perp U_1} \sum\limits_{i=1}^{K} \langle Y_i, P\rangle^2 = \lambda_2$ *и достигается на* $P = U_2$,

...

r. $\max\limits_{P:\ P\perp U_j,\ j=1,\ldots,r-1} \sum\limits_{i=1}^{K} \langle Y_i, P\rangle^2 = \lambda_r$ *и достигается на* $P = U_r$.

U_i — *главные направления.*

Задание. Найти главные направление для примера с двумя точками на плоскости.

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:*

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

Задание. Найти вектор первых главных компонент для примера с двумя точками на плоскости.

Глава 2

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

2.1. Анализ главных компонент на выборочном языке

2.1.1. Изменение нормы в пространстве признаков

Итак, мы умеем строить сингулярное разложение в виде

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

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. Отсюда следует, что квадрат нормы $\|\cdot\|_1^2$ вектора-индивида из $\mathbb{R}^L$ равен сумме квадратов его компонент всегда, а норма в квадрате $\|\cdot\|_2^2$ вектора-признака из $\mathbb{R}^K$ в анализе главных компонент считается не как сумма, а как среднее арифметическое квадратов компонент вектора. Далее, 1 у нормы вектора означает, что там сумма квадратов, а 2 — что среднее арифметическое. Если индекса у нормы нет, то это означает, что рассматривается вариант 1, т.е., обычная евклидова норма.

Полученное разложение с измененной нормой в пространстве признаков обозначим:

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

2.1.2. 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.Знаем, что

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

Таким образом, получаем следующие соотношения:

- $U_i = \widetilde{U}_i$,

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

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

Также заметим, что

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

Задание: Переделать численный пример SVD из задания с изменением нормы.

Далее будем предполагать, что признаки центрированы, т.е. среднее по строкам $\mathbb{Y}$ (или, что то же самое, среднее по столбцам $\mathbb{X}$) равно 0.

Задание: Показать, что если признаки центрированы, то и V_i центрированы.

Задание: Что будет с примерами с двумя и тремя точками, если данные центрировать.

Подытожим отличия сингулярного разложения от анализа главных компонент.

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

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

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

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

2.2. Чему АГК соответствует на статистическом языке?

Перейдём к $\mathbb{X} \in \mathbb{R}^{n\times p}$ ($L \to p$ — количество признаков, $K \to n$ — число индивидов). Пусть столбцы $\mathbb{X}$ (признаки) центрированы. Надо бы обозначить эту матрицу $\mathbb{X}^{(c)}$, но обозначения будут перегружены. *Берём признак, какую норму нужно рассматривать?* Вероятностную норму. *Что означает характеристика $\|X_i\|_2^2$ на статистическом языке?*

$$\|X_i\|_2^2 = \frac{1}{n}\sum_{j=1}^{n}\left((X_i)_j - 0\right)^2 = s^2(X_i) \text{ — выборочная дисперсия } X_i.$$

Заметим, что выше $s^2(X_i)$ — выборочная дисперсия признака X_i как до, так и после центрирования, так как вычитание константы не влияет на дисперсию.

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

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

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

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

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

Разложение можем записать следующим образом:

$$\mathbb{X}^{\mathrm{T}} = \sum_{i=1}^{d}\sqrt{\lambda_i}U_iV_i^{\mathrm{T}} = \sum_{i=1}^{d}F_iV_i^{\mathrm{T}} = \sum_{i=1}^{d}U_iZ_i^{\mathrm{T}},$$

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

2.3. Вклад главных компонент

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

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

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

$\frac{\lambda_j}{\sum\limits_{i=1}^{d}\lambda_i}$ — *вклад j-ой главной компоненты в общую дисперсию.*[3]

$s^2(X_i)$ — 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 варианты:

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).

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

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

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

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

Вернемся к признакам. Пусть признаки стандартизованы. Аналогично можем получить, что

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

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

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

так как квадрат множественного коэффициента корреляции между вектором-признаком и набором векторов-признаков — это косинус угла между центрированным вектором-признаком и подпространством, натянутым на вектора-признаки из набора. Интерпретация множественного коэффициента корреляции — мера зависимости между одним признаком и набором признаков.

Замечание 3. $\mathbb{U} = [U_1 : \ldots : U_p],\ \mathbb{F} = [F_1 : \ldots : F_d]$

1. $\sum\limits_{i=1}^{p} u_{ij}^2 = \|U_j\|^2 = 1$
2. $\sum\limits_{j=1}^{p} u_{ij}^2 = 1$ *(так как $\mathbb{U}$ — тоже ортогональная матрица)*
3. $\sum\limits_{i=1}^{p} f_{ij}^2 = \|F_j\|^2 = \lambda_j$
4. $\sum\limits_{j=1}^{d} f_{ij}^2 = \sum\limits_{j=1}^{d} \langle X_i, V_j \rangle_2^2 = \|X_i\|_2^2 = \begin{cases} s^2(X_i), & \text{если АГК по ковариац. матр.} \\ 1, & \text{если АГК по корреляц. матр.} \end{cases}$

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.

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

Формула, на основе которой происходит интерпретация главных компонент — это $Z_i = \mathbb{X}U_i$. Т.е., i-й собственный вектор задает коэффициенты линейной комбинации исходных признаков. Таким образом, мы получаем новые признаки, на основе собственных векторов их интерпретируем и знаем, что первый признак имеет максимально возможную дисперсию (информативность), второй с ним не коррелирует и имеет среди некоррелирующих также максимально возможную дисперсию, и т.д.

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 коэффициентами одинакового знака. Это можно проинтерпретировать как некий общий уровень чего-либо (например, общий уровень ученика, если признаки — оценки в школе). От того, какой знак, положительный или отрицательный, зависит то, с какой стороны по оси первой главной компоненты находятся в среднем большие значения исходных признаков, а с какой - меньшие.

Остальные главные компоненты можно также интерпретировать исходя из коэффициентов перед старыми признаками.[4]

2.6. 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.

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

2. Правило Кайзера. Выбираются главные компоненты, информативность которых больше средней информативности:

$$i: \quad \lambda_i > \frac{\sum\limits_{i=1}^{p} s^2(X_i)}{p} = \frac{\sum\limits_{i=1}^{d} \lambda_i}{p}.$$

Задание. Что получается, если АГК по корреляц.матрице?

3. Правило сломанной трости. Пусть $\mu_k = \frac{\sum\limits_{i=1}^{k} \lambda_i}{\sum\limits_{i=1}^{d} \lambda_i}$. Числа μ_i делят отрезок $[0, 1]$ на неравные части. Получаем разбиение $0 < \mu_1 < \mu_2 < \ldots < \mu_{d-1} < 1$. Но это же разбиение можно построить, случайно бросая $d-1$ точку на единичный отрезок (как бы случайно ломаем трость на d кусочков). Выбираются те компоненты, длины которых 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 существенно отличаются друг от друга.

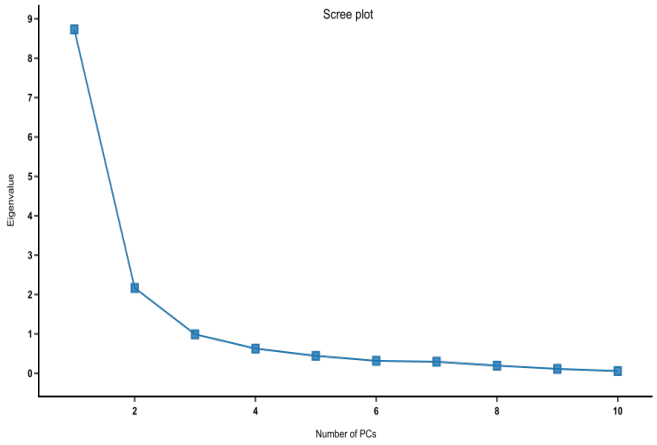

Рис. 2.1. Scree plot

5. Интерпретируем столько компонент, сколько можем.

2.7. Оптимизация в АГК в терминах ковариационных матриц

Напомним, что $\mathbb{Y} = \mathbb{X}^{\mathrm{T}}$, где $\mathbb{X}$ — матрица данных.

Предложение 6. *Задача* $\|\mathbb{Y} - \widetilde{\mathbb{Y}}\| \to \min\limits_{\widetilde{\mathbb{Y}}:\ \mathrm{rank}\,\widetilde{\mathbb{Y}} \le r}$ *эквивалентна задаче* $\|\mathbb{Y}\mathbb{Y}^{\mathrm{T}} - \widetilde{\mathbb{Y}}\widetilde{\mathbb{Y}}^{\mathrm{T}}\| \to \min\limits_{\widetilde{\mathbb{Y}}:\ \mathrm{rank}\,\widetilde{\mathbb{Y}} \le r}$. *Решение этой задачи* $\widetilde{\mathbb{Y}}^*$ *строится как сумма собственных троек, соответствующих первым* r *главным компонентам.*

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

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

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

Глава 3

Факторный анализ

(EFA — exploratory factor analysis)

3.1. Модель в факторном анализе

Главное отличие факторного анализа от анализа главных компонент — это наличие модели в факторном анализе.

Пусть $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)^{\mathrm{T}} \in \mathbb{R}^p$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_r)^{\mathrm{T}} \in \mathbb{R}^r$, $r < p$. Предполагаем, что $\mathbb{E}\xi_i = 0$, $\mathbb{E}\eta_i = 0$, $\mathbb{D}\eta_i = 1$, η_i и η_j некоррелированы. Пусть $\mathbb{F}_r = [F_1 : \ldots : F_r] \in \mathbb{R}^{p \times r}$. (Помните — у нас есть p старых признаков и мы строим новые. Новые r признаков будут в каком-то смысле самыми лучшими.)

Итак, *модель в факторном анализе*:

$$\boldsymbol{\xi} = \mathbb{F}_r \boldsymbol{\eta} + \boldsymbol{\varepsilon},$$

где ковариационная матрица $\operatorname{Cov} \boldsymbol{\varepsilon} = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ диагональная, $\boldsymbol{\eta}$ и $\boldsymbol{\varepsilon}$ некоррелированы. Случайному вектору $\boldsymbol{\xi}$ соответствует матрица данных $\mathbb{X}^{\mathrm{T}}$ (признаки, которые мы наблюдаем), а $\boldsymbol{\eta}$ соответствует $\mathbb{V}_r^{\mathrm{T}}$ (скрытые признаки, которые мы не наблюдаем). Получаем соотношение

$$\mathbb{X} = \mathbb{V}_r \mathbb{F}_r^{\mathrm{T}} + \mathbb{E}.$$

Перед нами стоит задача выявить эти скрытые признаки (задача факторного анализа). Причем, даже нет цели найти значения факторов. Цель — найти их смысл.

Заметим, что модель очень похожа на то, что можно получить с помощью АГК:

$$\mathbb{X} = \mathbb{V}_{1,r} \mathbb{F}_{1,r}^{\mathrm{T}} + \mathbb{V}_{r+1,d} \mathbb{F}_{r+1,d}^{\mathrm{T}},$$

но есть разница (увидим позже).

Модель можно переписать в следующем виде:

$$\Sigma = \operatorname{Cov}(\boldsymbol{\xi}) = \operatorname{Cov}(\mathbb{F}_r \boldsymbol{\eta}) + \operatorname{Cov}(\boldsymbol{\varepsilon}) = \mathbb{F}_r \mathbb{F}_r^{\mathbb{T}} + \Psi,$$

где $\Psi = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, $\operatorname{rank} \mathbb{F}_r \mathbb{F}_r^{\mathbb{T}} \leq r$. Тогда данные — это выборочная ковариационная матрица.

Вообще, факторный анализ всегда делается на основе стандартизованных данных. Т.е. всегда мы получаем модель корреляционной матрицы. Тогда $f_{ij} = \rho(\xi_i, \eta_j)$ и матрица факторных нагрузок выглядит следующим образом:

	η_1	$\dots$	η_r
ξ_1	f_{11}	$\dots$	f_{1r}
$\vdots$	$\vdots$		$\vdots$
ξ_p	f_{p1}	$\dots$	f_{pr}

Модель:

$$\xi_1 = f_{11}\eta_1 + \ldots + f_{1i}\eta_i + \ldots + f_{1r}\eta_r + (\varepsilon_1 + 0\varepsilon_2 + \ldots + 0\varepsilon_p)$$

$$\xi_2 = f_{21}\eta_1 + \ldots + f_{2i}\eta_i + \ldots + f_{2r}\eta_r + (0\varepsilon_1 + \varepsilon_2 + \ldots + 0\varepsilon_p)$$

$$\ldots$$

$$\xi_p = f_{p1}\eta_1 + \ldots + f_{pi}\eta_i + \ldots + f_{pr}\eta_r + (0\varepsilon_1 + 0\varepsilon_2 + \ldots + \varepsilon_p)$$

Мы предполагаем, что в каждом столбце (и строчке) перед ε_i только один из коэффициентов не равен нулю, а среди f_{ij}, $i = 1, \ldots, p$, хотя бы два не равны нулю.

3.2. Общности и уникальности

В этом пункте введем понятие общности и уникальности признаков.

Обозначим $\boldsymbol{\xi}' = \mathbb{F}_r\boldsymbol{\eta}$, тогда $\boldsymbol{\xi} = \boldsymbol{\xi}' + \boldsymbol{\varepsilon}$. То есть $\boldsymbol{\xi}'$ — это та часть данных, которая описывается факторами. Предполагаем, что $\mathrm{D}\xi_i = 1$, $\mathrm{D}\varepsilon_i = \sigma_i^2$, тогда $\mathrm{D}\xi_i' = \mathrm{D}\xi_i - \mathrm{D}\varepsilon_i = 1 - \sigma_i^2$.

Можно заметить, что $\mathrm{Cov}(\xi_i, \xi_j) = \mathrm{Cov}(\xi_i' + \varepsilon_i, \xi_j' + \varepsilon_j) = \mathrm{Cov}(\xi_i', \xi_j')$, так как остальные слагаемые обнуляются из-за некоррелированности. Получаем, что $\mathrm{Cov}(\xi_i', \xi_j') = \rho_{ij}$. Также известно, что ковариационная матрица $\boldsymbol{\xi}'$ равна $\mathrm{Cov}(\boldsymbol{\xi}') = \mathbb{F}_r\mathrm{Cov}(\boldsymbol{\eta})\mathbb{F}_r^{\mathrm{T}} = \mathbb{F}_r\mathbb{F}_r^{\mathrm{T}}$. **Почему**?

Тогда

$$\mathbb{F}_r\mathbb{F}_r^{\mathrm{T}} = \mathrm{Cov}(\boldsymbol{\xi}') = \begin{pmatrix} 1-\sigma_1^2 & \dots & \rho_{1p} \\ \vdots & \ddots & \vdots \\ \rho_{p1} & \dots & 1-\sigma_p^2 \end{pmatrix}.$$

Мы получили, что

$$1 = \mathrm{D}\xi_i = \mathrm{D}\xi_i' + \mathrm{D}\varepsilon_i = \underbrace{(1-\sigma_i^2)}_{\substack{\text{commonality,} \\ \text{общность признака}}} + \underbrace{\sigma_i^2}_{\substack{\text{uniquence,} \\ \text{уникальность}}}.$$

Замечание 4. *Общность является множественным коэффициентом корреляции:*

$$\underbrace{\mathrm{D}\xi_i'}_{\textit{общность}} = (\mathbb{F}_r\mathbb{F}_r^{\mathrm{T}})_{ii} = f_{i1}^2 + f_{i2}^2 + \ldots + f_{ir}^2 = \sum_{j=1}^{r} f_{ij}^2 = \sum_{j=1}^{r}(\rho(\xi_i, \eta_j))^2 = R^2(\xi_i|\eta_1, \ldots, \eta_r),$$

т.к. η_i и η_j некоррелированы.

Таким образом, общность — это то, что описывается факторами, а уникальность признака — то, что не описывается факторами.

3.3. Корректность поставленной задачи

Теперь посмотрим, корректно ли поставлена задача?

Во-первых, напомним, что в матрице $\mathbb{F}$ не может быть такого столбца, где только один элемент не равен нулю, то есть это факторы, соответствующие хотя бы двум признакам. Случайные величины из вектора $\boldsymbol{\varepsilon}$ в каком-то смысле тоже факторы, но каждый из них соответствует только одному признаку. Факторный анализ ищет общие факторы.

Определим $\widetilde{\mathbb{F}}_r = \mathbb{F}_r\mathbb{W}$, где $\mathbb{W} \in \mathbb{R}^{r\times r}$ — ортогональная матрица. Тогда

$$\widetilde{\mathbb{F}}_r\widetilde{\mathbb{F}}_r^{\mathrm{T}} = \mathbb{F}_r\mathbb{W}\mathbb{W}^{\mathrm{T}}\mathbb{F}_r^{\mathrm{T}} = \mathbb{F}_r\mathbb{F}_r^{\mathrm{T}},$$

то есть решение не единственно, и любое вращение — тоже решение. Необходимо добавить ограничения, чтобы как-то зафиксировать решение. Матрица $\mathbb{W}$ содержит $\frac{r(r-1)}{2}$ параметров (известный факт — ортогональную матрицу размерности $r \times r$ можно задать таким количеством параметров). Следовательно, для однозначности модели необходимо добавить столько же ограничений, чтобы убрать свободу.

Пример 2. *Пусть $\mathbb{S}$ — выборочная ковариационная матрица.*

Тогда ограничение: $\mathbb{F}_r^{\mathrm{T}}\mathbb{S}\mathbb{F}_r$ — диагональная матрица $r\times r$. Тут $\frac{r(r-1)}{2}$ ограничений (из-за симметричности число ограничений равно числу нулей под диагональю матрицы).

Теперь посчитаем число параметров в модели $\Sigma = \mathbb{F}_r\mathbb{F}_r^{\mathbb{T}} + \Psi$. Тут

$$\underbrace{pr}_{\text{для } \mathbb{F}_r} + \underbrace{p}_{\text{для } \Psi} - \underbrace{\frac{r(r-1)}{2}}_{\text{вычитаем число ограничений}}$$

параметров с учетом ограничений.

Модель задается $\frac{p(p+1)}{2}$ равенствами. Тогда нужно требовать, чтобы число параметров не превосходило число уравнений, то есть

$$pr + p - \frac{r(r-1)}{2} \le \frac{p(p+1)}{2}.$$

Если записать это неравенство в другом виде, то получим

$$\frac{(p-r)^2 - (p+r)}{2} \ge 0.$$

3.4. Задача оценивания параметров

Теперь можем перейти к задаче оценивания параметров. Можно оценивать несколькими методами.

3.4.1. Метод наименьших квадратов (OLS)

Пусть $\mathbb{S} = \{s_{ij}\}_{i,j=1}^{p}$ — выборочная ковариационная матрица.

Решаем задачу $\|\mathbb{S} - \widetilde{\mathbb{S}}\|_2^2 \to \min\limits_{\widetilde{\mathbb{S}}:\ \widetilde{\mathbb{S}}=\widetilde{\mathbb{F}}_r\widetilde{\mathbb{F}}_r^{\mathrm{T}}+\widetilde{\Psi}}$, что эквивалентно задаче

$$\sum_{i,j=1}^{p}(s_{ij} - \widetilde{s}_{ij})^2 \to \min_{\widetilde{\mathbb{S}}:\ \widetilde{\mathbb{S}}=\widetilde{\mathbb{F}}_r\widetilde{\mathbb{F}}_r^{\mathrm{T}}+\widetilde{\Psi}}$$

Алгоритм MINRES (OLS для ковариационной матрицы):

1. Рассматриваем не всю матрицу, а ее часть, и решаем следующую задачу:[1]

$$\sum_{i\neq j}\underbrace{(s_{ij} - \widetilde{s}_{ij})^2}_{\text{residuals}} \to \min_{\widetilde{\mathbb{S}}:\ \widetilde{\mathbb{S}}=\widetilde{\mathbb{F}}_r\widetilde{\mathbb{F}}_r^{\mathrm{T}}}.$$

2. Находим $\hat{\mathbb{F}}_r$.

3. $$\begin{pmatrix}\sigma_1^2\\ \vdots\\ \sigma_p^2\end{pmatrix} = \underbrace{\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}}_{\text{Диагональ }\mathbb{S}} - (\text{Диагональ } \widetilde{\mathbb{F}}_r\widetilde{\mathbb{F}}_r^{\mathrm{T}})$$

3.4.2. Взвешенный метод наименьших квадратов (WLS)

Решается задача

$$\sum_{i,j=1}^{p}\frac{(s_{ij} - \widetilde{s}_{ij})^2}{\widehat{\sigma}_i^2\widehat{\sigma}_j^2} \to \min_{\widetilde{\mathbb{S}}:\ \widetilde{\mathbb{S}}=\widetilde{\mathbb{F}}_r\widetilde{\mathbb{F}}_r^{\mathrm{T}}+\widetilde{\Psi}}$$

Получаем, что чем меньше уникальность, тем больше вес слагаемого.

3.4.3. Метод максимального правдоподобия (MLE)

Выписывается распределение ковариационной матрицы данных в модели факторного анализа и находятся оценки параметров ее максимизацией. Результат примерно такой же, как в WLS.

[1] Матрицу Ψ убрали, т.к. она диагональна, а мы суммируем по всем $i \neq j$.

3.5. Разница между АГК и факторным анализом

Первое отличие между факторным анализом и анализом главных компонент заключается в наличии модели в факторном анализе.

Второе отличие состоит в том, что у АГК и факторного анализа разные оптимизационные задачи. **Задача факторного анализа**:

$$\sum_{i\neq j}(s_{ij}-\widetilde{s}_{ij})^2 \to \min_{\widetilde{\mathbb{S}}:\ \widetilde{\mathbb{S}}=\widetilde{\mathbb{F}}_r\widetilde{\mathbb{F}}_r^{\mathrm{T}}},$$

где минимизируется сумма по всем элементам, кроме диагональных, $\widetilde{\mathbb{S}}$ — симметричная, неотрицательно определенная матрица ранга, не превосходящего r.

Задача в анализе главных компонент:

$$\sum_{i,j}(s_{ij}-\widetilde{s}_{ij})^2 \to \min_{\widetilde{\mathbb{S}}:\ \widetilde{\mathbb{S}}=\widetilde{\mathbb{F}}_r\widetilde{\mathbb{F}}_r^{\mathrm{T}}},$$

минимизируется сумма по всем элементам, $\widetilde{\mathbb{S}}$ — симметричная, неотрицательно определенная матрица ранга, не превосходящего r.

Таким образом, цель факторного анализа — как можно лучше воспроизвести ковариации. В анализе главных компонент стоит аппроксимизационная задача, и АГК не решает именно задачу факторного анализа, хотя все равно используется для поиска факторов на практике.

Есть еще третье различие, которое рассмотрим ниже — в факторном анализе порядок факторов не важен (в АГК они упорядочены по вкладу), а важен их смысл.

3.6. Проверка гипотез

3.6.1. Проверка значимости модели

Рассмотрим гипотезу $H_0:\ \Sigma = \mathbb{F}_r\mathbb{F}_r^{\mathrm{T}} + \Psi$ о том, что Σ действительно имеет такой вид. Статистика критерия

$$t = (n-1-\frac{2p+4r-5}{6})\log\frac{\overbrace{|\widehat{\mathbb{F}}_r\widehat{\mathbb{F}}_r^{\mathrm{T}}+\widehat{\Psi}|}^{\text{То, что ожидаем}}}{\underbrace{|\mathbb{S}|}_{\text{То, что есть}}} \sim \chi^2(\frac{(p-r)^2-(p+r)}{2})$$

(в нормальной модели).

Идеальное значение равно 0 (логарифм будет равен нулю, если данные идеально соответствуют гипотезе, критическая область справа).

3.6.2. Тест сферичности Бартлетта

Для начала необходимо проверить, есть ли вообще структура. Гипотеза о том, что общей структуры нет: $H_0 : \ \Sigma = \mathbb{I}_{p\times p}$ (единичная матрица).

Тест сферичности Бартлетта: статистика критерия

$$t = (n - 1 - \frac{2p-5}{6})(-\log|\mathbb{S}|) \sim \chi^2(\frac{p(p-1)}{2})$$

(в нормальной модели).

Для выбора порядка модели можно применять информационные критерии AIC, BIC (в нормальной модели) — подход со штрафом за число параметров. Можно промоделировать, посчитать вклады (несколько раз) и строить доверительный интервал.

3.7. Ортогональные вращения

Имеем $\mathbb{F}_r = \{f_{ij}\} = \{\rho(\xi_i, \eta_j)\}$. Мы интерпретируем признаки в факторном анализе, смотря на элементы f_{ij} (как исходные признаки выражаются через факторы). Нам хотелось бы, чтобы матрица $\mathbb{F}_r$ имела простую структуру (то есть чтобы в ней было много нулей).

Заметим, что если $\mathbb{F}_r$ — решение, то вращение $\widetilde{\mathbb{F}}_r = \mathbb{F}_r\mathbb{W}$ — тоже решение, только будут другие корреляции между исходными признаками и факторами, так как меняется $\mathbb{F}_r$. (Матрица вращения $\mathbb{W}$ ортогональна.)

На выборочном языке: $\mathbb{X}'$ — та часть исходных данных, которая описывается факторами (соответствует $\boldsymbol{\xi}' = \mathbb{F}_r\boldsymbol{\eta}$).

$$\mathbb{X}' = \mathbb{V}\mathbb{F}_r^{\mathrm{T}} = \underbrace{\mathbb{V}\mathbb{W}}_{\widetilde{\mathbb{V}}}\underbrace{\mathbb{W}^{\mathrm{T}}\mathbb{F}_r^{\mathrm{T}}}_{\widetilde{\mathbb{F}}_r},$$

$\widetilde{\mathbb{V}}$ — новые факторы, $\widetilde{\mathbb{F}}_r$ — новые факторные нагрузки. Хотелось бы найти такую $\mathbb{W}$, чтобы интерпретация была как можно проще.

Пример 3. *Пусть у нас какие-то признаки отвечают за знание математики, а другие за знание физики.*

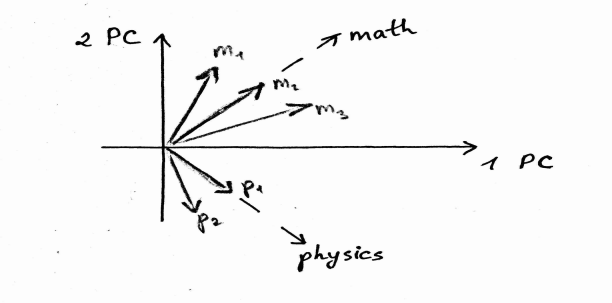

Рис. 3.1. Вращение в факторном анализе

Мы можем повернуть данные так, что у нас одна ось совпадет, к примеру, с признаком m_2, а другая с признаком p_1. Тогда интерпретировать станет проще.

Метод вращений Varimax: Хотим получить более понятную интерпретацию, для этого нам будет удобно, если в каждом столбце матрицы $\mathbb{F}_r$ было как можно больше нулей. Задача выглядит следующим образом:

$$\sum_{j=1}^{r}\Big[\frac{1}{p}\sum_{i=1}^{p}(\widetilde{f}_{ij}^2)^2 - (\frac{1}{p}\sum_{i=1}^{p}\widetilde{f}_{ij}^2)^2\Big] \to \max_{\mathbb{W}:\ \widetilde{\mathbb{F}}_r = \mathbb{F}_r\mathbb{W}},$$

где мы рассматриваем f_{ij}^2, так как f_{ij} могут быть отрицательны, и считаем разброс значений f_{ij}^2, $i = 1, \ldots, p$, как выборочную дисперсию, так как чем больше разброс, тем больше шансов, что значения будут разными, т.е. маленькие (почти нулевые) и большие.

На выборочном языке: $\mathbb{V} = [V_1 : \ldots : V_r],\ V_1, \ldots, V_r \in \mathbb{R}^n$ — факторные вектора. На выборочном языке $\boldsymbol{\xi}$ переходит в $\mathbb{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\xi}'$ переходит в $\mathbb{X}' \in \mathbb{R}^{n \times p}$, $\boldsymbol{\eta}$ (соответствует факторам) переходит в $\mathbb{V} \in \mathbb{R}^{n \times r}$. Тогда $\mathbb{F}\boldsymbol{\eta}$ соответствует $\mathbb{V}\mathbb{F}^{\mathrm{T}}$.

Поворачиваем факторы: $\widetilde{\boldsymbol{\eta}} = \mathbb{W}\boldsymbol{\eta}$ — факторы после вращения, $\widetilde{\mathbb{F}} = \mathbb{F}\mathbb{W}^{-1}$, где $\mathbb{W} \in \mathbb{R}^{r \times r}$ — матрица поворота. Что означает на языке векторов то, что мы поворачиваем факторы? Это значит, что $\widetilde{\mathbb{V}} = \mathbb{V}\mathbb{W} \in \mathbb{R}^{n \times r}$.

Идея: мы хотим, чтобы $\mathbb{F}\boldsymbol{\eta} = \widetilde{\mathbb{F}}\widetilde{\boldsymbol{\eta}}$ (то есть $\boldsymbol{\xi}'$ не должно меняться). И действительно,

$$\widetilde{\mathbb{F}}\widetilde{\boldsymbol{\eta}} = \mathbb{F}\mathbb{W}^{-1}\mathbb{W}\boldsymbol{\eta} = \mathbb{F}\boldsymbol{\eta},$$

мы поменяли факторные веса и получили то же самое (условно можно сказать, что в одну сторону поворачиваем факторы и в другую сторону поворачиваем $\mathbb{F}$).

В случае ортогональных вращений: $\mathbb{W}^{-1} = \mathbb{W}^{\mathrm{T}}$ и тоже является матрицей поворота.

3.8. Косоугольные вращения

Если мы допускаем косоугольные вращения, то, вообще говоря, $\mathbb{W}^{-1} \neq \mathbb{W}^{\mathrm{T}}$. Мы хотим, чтобы новые факторы после вращения были по-прежнему нормированы (то есть дисперсия была равна 1 на языке случайных величин). Мы предполагаем, что все центрировано. Тогда

$$\mathrm{Cov}(\widetilde{\boldsymbol{\eta}}) = \mathrm{E}\widetilde{\boldsymbol{\eta}}\widetilde{\boldsymbol{\eta}}^{\mathrm{T}} = \mathrm{E}(\mathbb{W}\boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}}\mathbb{W}^{\mathrm{T}}) = \mathbb{W}\underbrace{\mathrm{Cov}\,\boldsymbol{\eta}}_{\mathbb{I}_{r\times r}}\mathbb{W}^{\mathrm{T}} = \mathbb{W}\mathbb{W}^{\mathrm{T}},$$

будем требовать, чтобы на диагонали этой матрицы стояли единицы.

Замечание 5. *Необходимо проверять, чтобы в корреляционной матрице новых факторов были не очень большие по модулю корреляции, так как чем больше корреляции между новыми признаками, тем более подозрительный результат.*

Интерпретация после косоугольных вращений уже становится непонятна.

3.9. Факторная структура и факторный паттерн

Введем определения для двух матриц.

Определение 1. *Факторная структура (factor structure) — это матрица корреляций исходных признаков с новыми, то есть* $\Phi = \{\rho(\xi_i, \widetilde{\eta}_j)\}_{i=1,j=1}^{p,r}$.

Определение 2. *Факторный паттерн (factor pattern, factor loadings) — это матрица* $\widetilde{\mathbb{F}}$, *коэффициенты линейной комбинации в* $\boldsymbol{\xi}' = \widetilde{\mathbb{F}}\widetilde{\boldsymbol{\eta}}$ *(коэффициенты линейной комбинации, с которыми исходные признаки выражаются через новые).*

Предложение 7. *В случае ортогональных вращений факторная структура совпадает с факторным паттерном.*

Доказательство. Распишем, чему равно Φ:

$$\Phi = \mathrm{Cov}(\boldsymbol{\xi}, \widetilde{\boldsymbol{\eta}}) = \mathrm{E}(\boldsymbol{\xi}'\widetilde{\boldsymbol{\eta}}^{\mathrm{T}}) = \mathrm{E}(\widetilde{\mathbb{F}}\widetilde{\boldsymbol{\eta}}\widetilde{\boldsymbol{\eta}}^{\mathrm{T}}) = \widetilde{\mathbb{F}}\mathrm{Cov}(\widetilde{\boldsymbol{\eta}}) = \widetilde{\mathbb{F}}\mathbb{W}\mathbb{W}^{\mathrm{T}}.$$

В случае ортогональных вращений $\mathbb{W}^{-1} = \mathbb{W}^{\mathrm{T}}$, следовательно $\Phi = \widetilde{\mathbb{F}}\underbrace{\mathbb{W}\mathbb{W}^{\mathrm{T}}}_{\mathbb{I}_{r\times r}} = \widetilde{\mathbb{F}}$. □

3.10. Методы нахождения факторных значений

Перейдем на выборочный язык. $\mathbb{X} = \mathbb{V}\mathbb{F}^{\mathrm{T}} + \mathcal{E}$. Пусть мы уже оценили $\mathbb{F}$. Возникает вопрос — как найти $\mathbb{V}$ (factor scores)? Есть несколько методов.

1. (OLS) Распишем по индивидам $Y_i \in \mathbb{R}^p$, получим n уравнений

$$Y_i = \mathbb{F}\begin{pmatrix} v_{i1} \\ \vdots \\ v_{ir} \end{pmatrix} + \mathcal{E}_i, \quad i = 1, \ldots, n$$

   Факторные значения для i-ого индивида можно найти по методу наименьших квадратов, тогда решение выглядит следующим образом:

$$(v_{i1}, \ldots, v_{ir})^{\mathrm{T}} = (\mathbb{F}^{\mathrm{T}}\mathbb{F})^{-1}\mathbb{F}^{\mathrm{T}}Y_i.$$

2. (Метод Бартлетта (WLS)) Пусть $\Psi = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ — диагональная матрица, в которой на диагонали стоят уникальности. Тогда решение:

$$(v_{i1}, \ldots, v_{ir})^{\mathrm{T}} = (\mathbb{F}^{\mathrm{T}}\Psi^{-1}\mathbb{F})^{-1}\mathbb{F}^{\mathrm{T}}\Psi^{-1}Y_i.$$

3. (Regression method) В нормальной модели, основываясь на ОМП,

$$(v_{i1}, \ldots, v_{ir})^{\mathrm{T}} = \mathbb{F}^{\mathrm{T}}\mathbb{S}^{-1}Y_i.$$