ТЕМА 2

ОТБОР ПРИЗНАКОВ

Понижение размерности

- С ростом размерности данных растет вычислительная сложность, а также и сложность визуализации.
- При изучении объектов, каждый из который характеризуется большим количеством признаков, часто возникает необходимость описать эти объекты значительно меньшим числом признаков, сохранив при этом как можно больше важной информации об объектах.
- Отбор наиболее значимых признаков.
- Понижение размерности пространства признаков.
- Уменьшение объема исходных данных, сохраняя полезную информацию.

Цели

- Удаление лишних (нерелевантных) признаков
- Повышение качества решения поставленной задачи
- Уменьшение стоимости данных
- Увеличение скорости последующего анализа
- Повышение интерпретируемости моделей

Сокращение числа переменных

- при массовом пошиве одежды используются размер, полнота, рост.
- Ввести **искусственные переменные** (факторы).
- Например, по одной из формул
- полнота = (длина окружности груди - длиной окружности талии)/2.

Сокращение числа переменных

Две переменные: рост ста людей в дюймах и сантиметрах.

- Дублирование информации.
- Одну переменную отбрасываем.
- Сокращение данных.

Значения одной переменной вычисляются по значениям другой с помощью линейного преобразования.

- Линейная зависимость между переменными ⇔ коэффициент корреляции между ними равен единице.

- **Факторный анализ**

- отыскание скрытых, но объективно существующих закономерностей, которые определяются воздействием внутренних и внешних причин на изучаемый процесс;

- сжатие информации путем описания процесса при помощи общих факторов или главных компонент, число которых значительно меньше количества первоначально взятых признаков;

- выявление и изучение статистической связи признаков с факторами или главными компонентами.

- **Кластерный анализ**

- разбить изучаемую совокупность объектов на группы схожих, близких в некотором смысле объектов, называемых кластерами.

- **Многомерное шкалирование**

- использование мер различия между объектами

Факторный анализ

По целям исследования

- **EFA**, **Эксплораторный (**разведочный) факторный анализ пытается выявить внутреннюю структуру довольно широкого набора переменных.
  - *Априорное* допущение исследователя состоит в том, что любые признаки (переменные) могут ассоциироваться с любым фактором. Наиболее распространенная форма факторного анализа. В ней отсутствует предварительная теория, и используются факторные нагрузки, чтобы интуитивно понять факторную структуру данных.
- **CFA**, **Конфирматорный** (подтверждающий) факторный анализ пытается определить, соответствует ли количество факторов и нагрузки измеряемых переменных (признаков, индикаторов) тому, что ожидается, на основе предварительной теории.
  - Интегрирован в моделирование структурными уравнениями SEM.

Факторный анализ

- Главными целями факторного анализа:

(1) *сокращение* числа переменных (редукция данных), EFA;

(2) *определение структуры* взаимосвязей между переменными, т.е. *классификация переменных*, CFA.

- Поэтому факторный анализ используется или как метод сокращения данных или как метод классификации.

Факторный анализ как метод редукции данных

- Позволяет свести большое количество исходных *переменных* к значительно меньшему числу *факторов*, каждый из которых объединяет исходные переменные, имеющие сходный смысл.
- Каждый фактор интерпретируется как некоторая общая причина взаимосвязи группы переменных.

Сокращение переменных

- исходные переменные (не все) заменяют на меньшее число новых искусственных переменных
- новые переменные называют главными **факторами /** главными **компонентами.**
- далее работают с факторами, а не с исходными переменными (признаками).
- Основные понятия факторного анализа: *фактор* — скрытая переменная и *факторная нагрузка* — корреляция между исходной переменной и фактором.

Измерение неизмеримого или факторный анализ как поиск латентных (скрытых) переменных, не поддающихся непосредственному измерению (CFA)

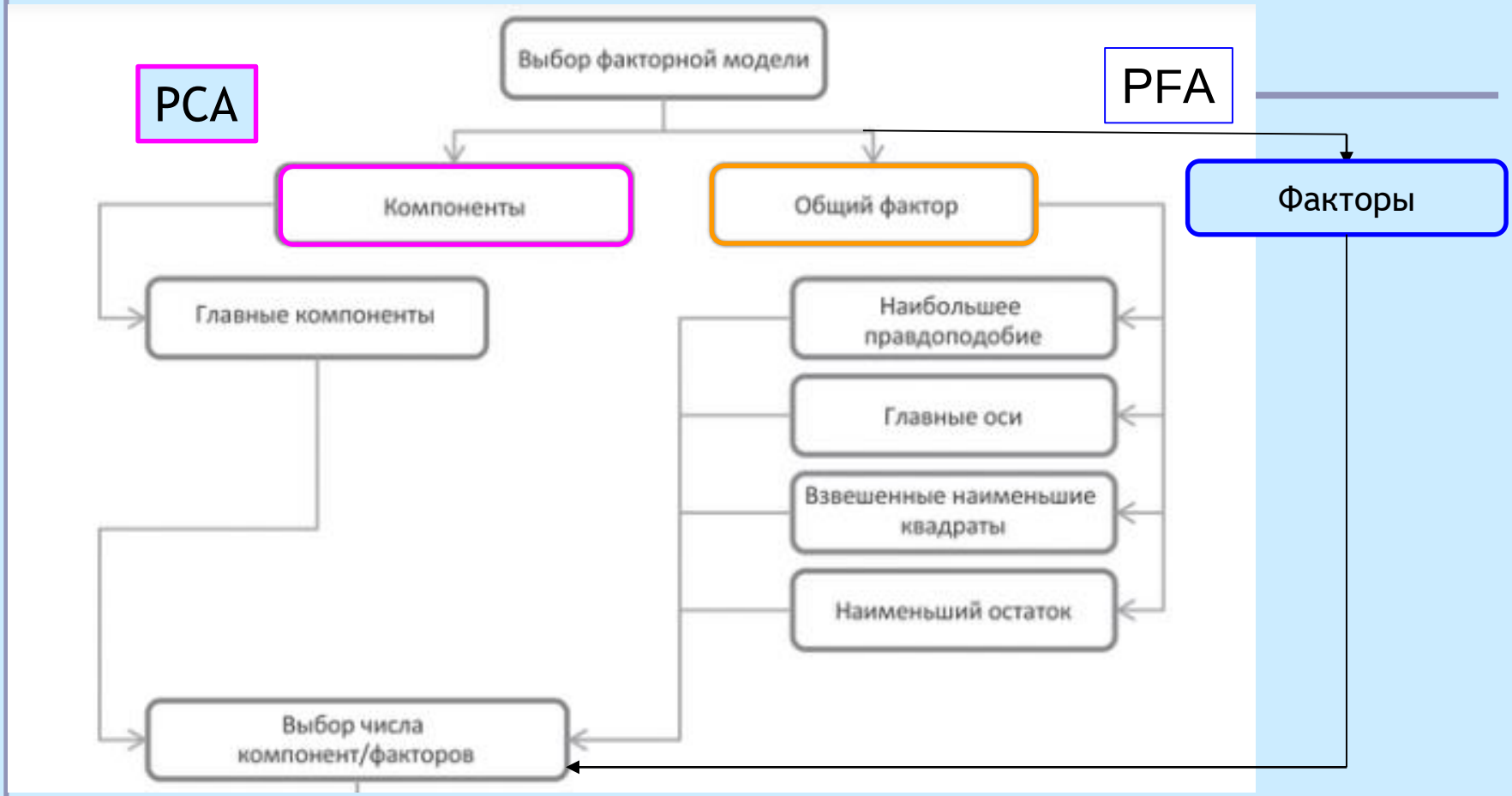
- Отношение пациента к своему доктору?
- Удовлетворенность сортом кофе?
- Как определить степень депрессии человека?
- Степень приверженности курению?
- Лояльность торговой марке?
- Вероятность разорения фирмы в течение следующего года?

ПРОЦЕДУРА ФАКТОРНОГО АНАЛИЗА

- Подготовка исходной матрицы
  - Очистка данных, Обработка выбросов
  - **Стандартизация**
- ВЫБОР факторной модели
- **Факторизация**, ИЗВЛЕЧЕНИЕ важных признаков:

  первичных компонент/факторов
- ВРАЩЕНИЕ факторов
- Оценка факторный значений и ИНТЕРПРЕТАЦИЯ факторов

ШАГ 1. **Факторизация** - метод выделения факторов



PCA – метод главных компонент

PFA – общий факторный анализ

- Источник. Кабаков. R в действии. Глава 14.

Извлечение факторов. Факторизация

- Метод главных компонент (PCA)
- Общий факторный анализ (PFA)
  - Методы вращения
  - Максимального правдоподобия
- Нелинейные методы
  - Нейронные сети
  - SVM

Факторизация

PCA, Метод главных компонент

- Ищет набор факторов, способных объяснить всю (общую и уникальную) дисперсию в наборе переменных.
- Предпочитается для целей сокращения данных.
- Используется в EFA – **эксплораторный (**разведочный) факторный анализ

PFA, общий факторный анализ

- Ищет наименьшее количество факторов, способных объяснить общую дисперсию (корреляцию) набора переменных.
- Используется в **CFA – конфирматорный** факторный анализ.

PCA - главные компоненты	**PFA - главные факторы**
Ковариационная матрица или Корреляционная матрица – по диагонали 1.	**Корреляционная матрица, По диагонали общности – h2 *квадраты множественной корреляции***
Используется в EFA	Используется в CFA

Если исходные данные **стандартизированы**
Ковариационная матрица = Корреляционная матрица

Количество “компонент”=количеству переменных	Ищется наименьшее количество “факторов”, объясняющих вариацию.
Количество главных компонент (факторов) отбирается по критериям.	Отбор факторов зависит от поставленной задачи.
Легко реализуется	Пользоваться специализированными библиотеками

Имеется набор признаков

$$\boldsymbol{X}^T=(X_1, X_2, \ldots, X_p)$$

- PCA

Найти $Y_k = a_{k1} X_1 + a_{k2} X_2 + \ldots + a_{kk} X_k$

Количество главных компонент (факторов) равно количеству исходных признаков (переменных). $k = p$

- PFA

$\boldsymbol{F}^T=(F_1, F_2, \ldots F_k)$ - факторы (общие факторы, латентные факторы)

$k < p$

$$X_i = a_{i1} F_1 + a_{i2} F_2 + a_{ik} F_k + U_i \qquad i = 1, \ldots, p$$

Количество факторов меньше количества исходных признаков.

Анализ главных компонент

Определим понятие главной компоненты. Пусть имеется k признаков $X_1, \ldots, X_k$. Первой главной компонентой Y_1 называется сохраняющая расстояние между точками линейная комбинация исходных признаков

$$Y_1 = \alpha_{11}X_1 + \ldots + \alpha_{k1}X_k,$$

где коэффициенты $\alpha_{11}, \ldots, \alpha_{k1}$ выбираются таким образом, чтобы дисперсия $D(Y_1) = \lambda_1$ была максимальной. Это соответствует тому, что по первой главной компоненте индивиды должны отличаться наибольшим образом.

Вторая главная компонента также является линейной комбинацией исходных признаков:

$$Y_2 = \alpha_{12}X_1 + \ldots + \alpha_{k2}X_k,$$

где коэффициенты $\alpha_{12}, \ldots, \alpha_{k2}$ выбираются таким образом, что компоненты Y_1 и Y_2 некоррелированы, а дисперсия $D(Y_2) = \lambda_2$ является максимальной из всех линейных комбинаций, некоррелированных с Y_1, то есть вторая компонента должна нести наибольшую новую информацию, не имеющую отношения к первой главной компоненте. Аналогично строятся остальные главные компоненты

$$Y_j = \sum_{i=1}^{k} \alpha_{ij}X_i, \ j = 1, \ldots, k.$$

Суммарная дисперсия остается неизменной:

$$V = D(X_1) + \ldots D(X_k) = \lambda_1 + \ldots + \lambda_k.$$

Анализ главных компонент

Рассмотрим случайный вектор $X_1, X_2, \ldots, X_k$

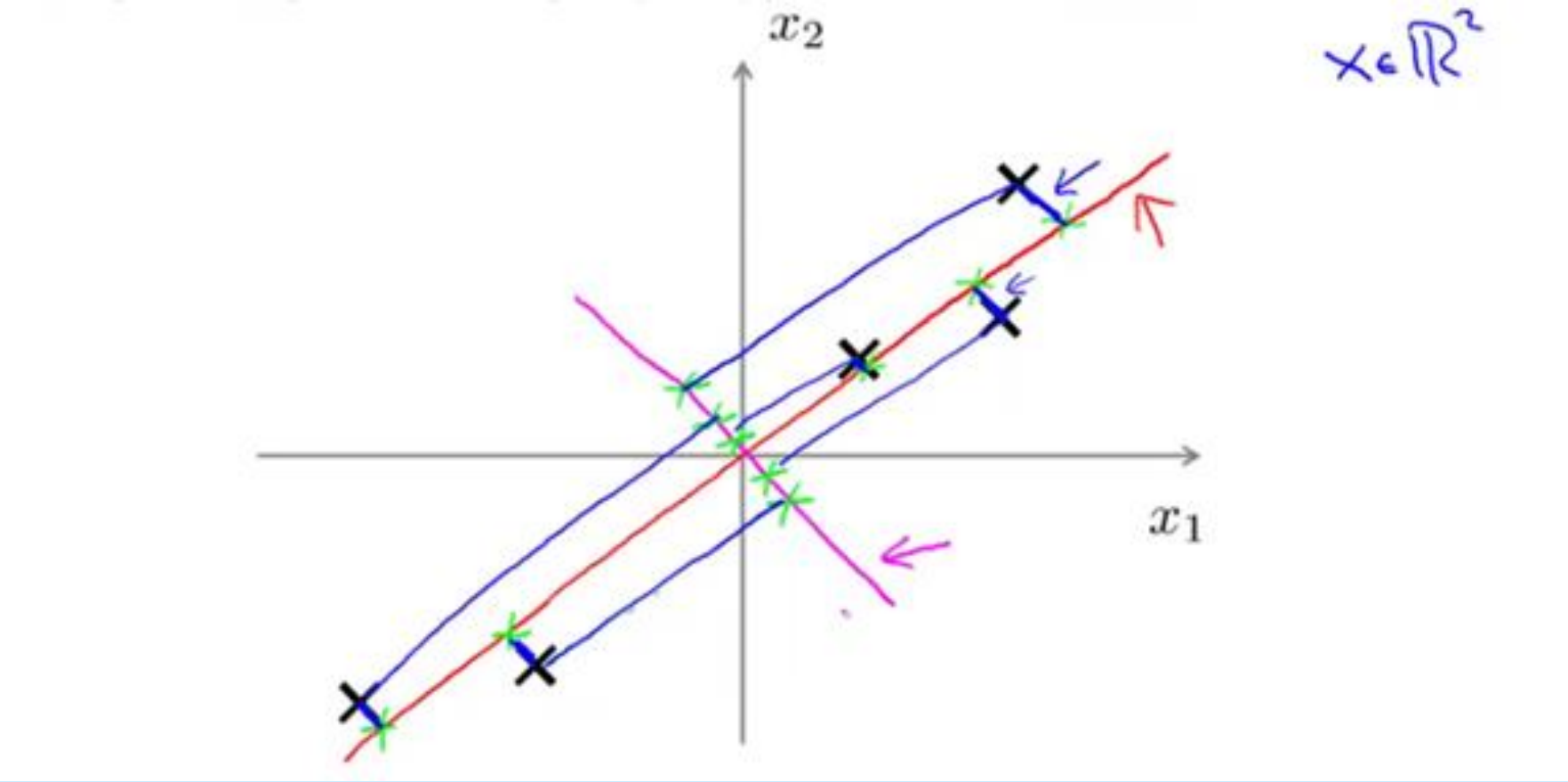
Задача 1. Найти $Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1k}X_k$ такую, что $D(Y_1)$ максимальна.
Дополнительное условие 1: $\vec{a}_1 \vec{a}_1^T = 1$, где $\vec{a}_1 = (a_{11}, a_{12}, \ldots, a_{1k})$

Задача 2. Найти $Y_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2k}X_k$ такую, что $D(Y_2)$ максимальна.
Дополнительное условие 2.1: $\vec{a}_2 \vec{a}_2^T = 1$, где $\vec{a}_2 = (a_{21}, a_{22}, \ldots, a_{2k})$
Дополнительное условие 2.2: $corr(Y_2, Y_1) = 0$

Задача k. Найти $Y_k = a_{k1}X_1 + a_{k2}X_2 + \ldots + a_{kk}X_k$ такую, что $D(Y_k)$ максимальна.
Дополнительное условие k.1: $\vec{a}_k \vec{a}_k^T = 1$, где $\vec{a}_k = (a_{k1}, a_{k2}, \ldots, a_{kk})$
Дополнительное условие k.2: $corr(Y_k, Y_1) = 0$
Дополнительное условие k.3: $corr(Y_k, Y_2) = 0$
…
Дополнительное условие k.k: $corr(Y_k, Y_{k-1}) = 0$

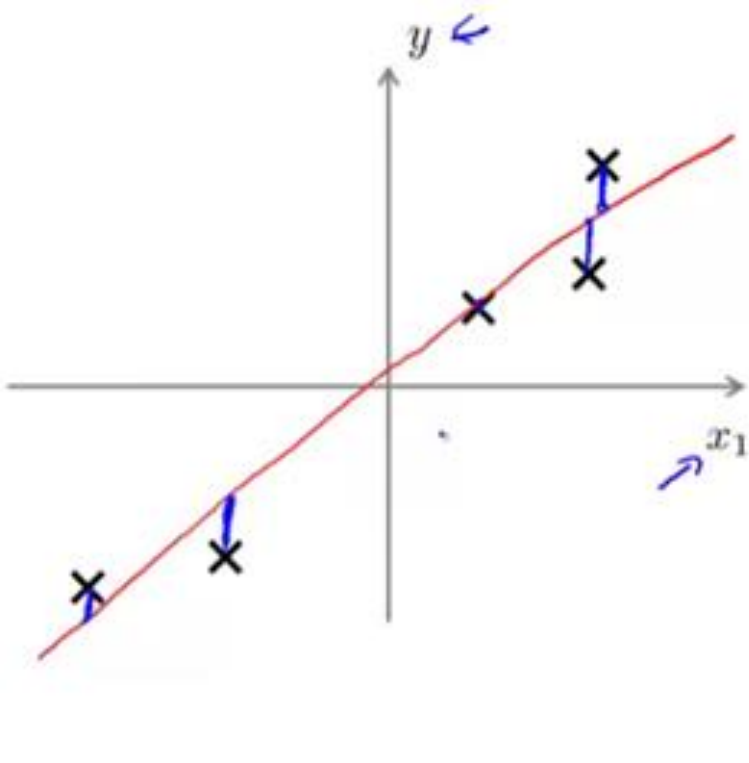
Алгоритм 1. Решение оптимизационной задача

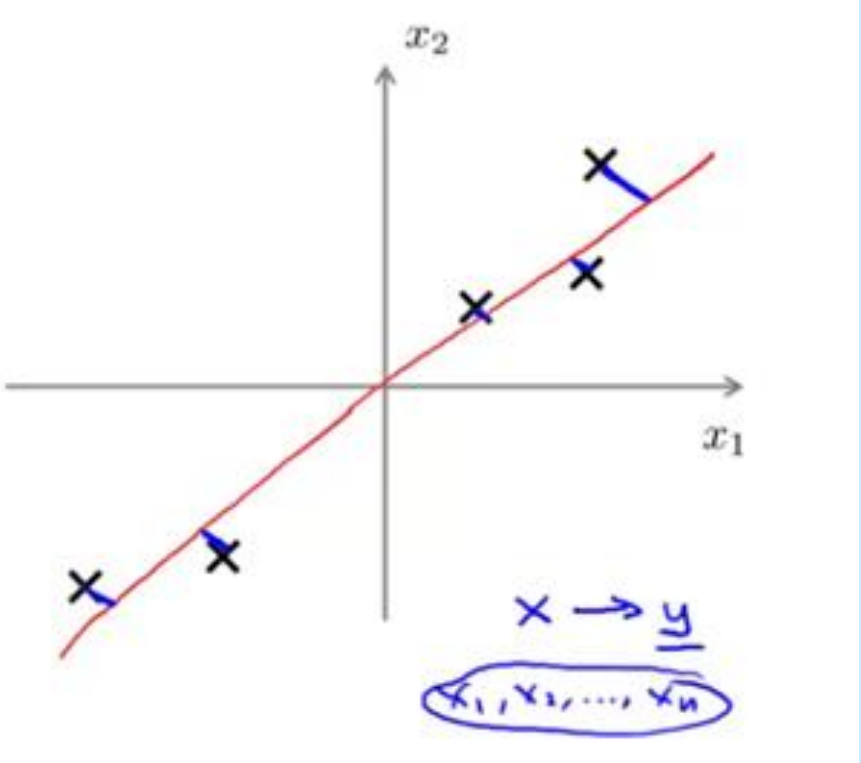
В основе метода главных компонент – **максимизация дисперсии**, то есть поиск осей максимальных изменений входных данных.



- Максимизация дисперсии – минимизация расстояний до прямой.
- То есть решается «регрессионная» задача.

PCA не является линейной регрессией (МНК)





PCA

Алгоритм 2. Использование ковариационной матрицы.

В методе главных компонент задача снижения размерности сводится к нахождению **собственных чисел** и **собственных векторов** **ковариационной** или корреляционной матрицы исходных признаков.

- Если переменная одна, то мерой разброса ее значений является **дисперсия** – средний квадрат отклонений от среднего значения этой переменной. Мера изменчивости.
- Мерой линейной зависимости двух переменных служит ковариация. Мера совместной изменчивости двух случайных величин.
- Для многомерных данных используется **ковариационная матрица**. Обобщение дисперсии на случай многомерных случайных величин.

Ковариация

- Ковариация является мерой **совместной** изменчивости двух случайных величин (мерой линейной зависимости).

$$Cov(X_i, X_j) = E\Big[\big(X_i - E(X_i)\big) \cdot \big(X_j - E(X_j)\big)\Big] = E(X_i X_j) - E(X_i) \cdot E(X_j)$$

- если Xi=Xj - дисперсия

$$Cov(X_i, X_i) = Var(X_i)$$

матрицей ковариации векторов $\mathbf{X}, \mathbf{Y}$ называется

$$\Sigma = \operatorname{cov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{Y} - \mathbb{E}\mathbf{Y})^{\top}\right],$$

то есть

$$\Sigma = (\sigma_{ij}),$$

где

$$\sigma_{ij} = \operatorname{cov}(X_i, Y_j) \equiv \mathbb{E}\left[(X_i - \mathbb{E}X_i)(Y_j - \mathbb{E}Y_j)\right],\ i = 1, \ldots, n,\ j = 1, \ldots, m.$$

$\mathbb{E}$ — математическое ожидание.

Коэффициент корреляции — параметр, который характеризует степень линейной взаимосвязи между двумя выборками, рассчитывается по формуле:

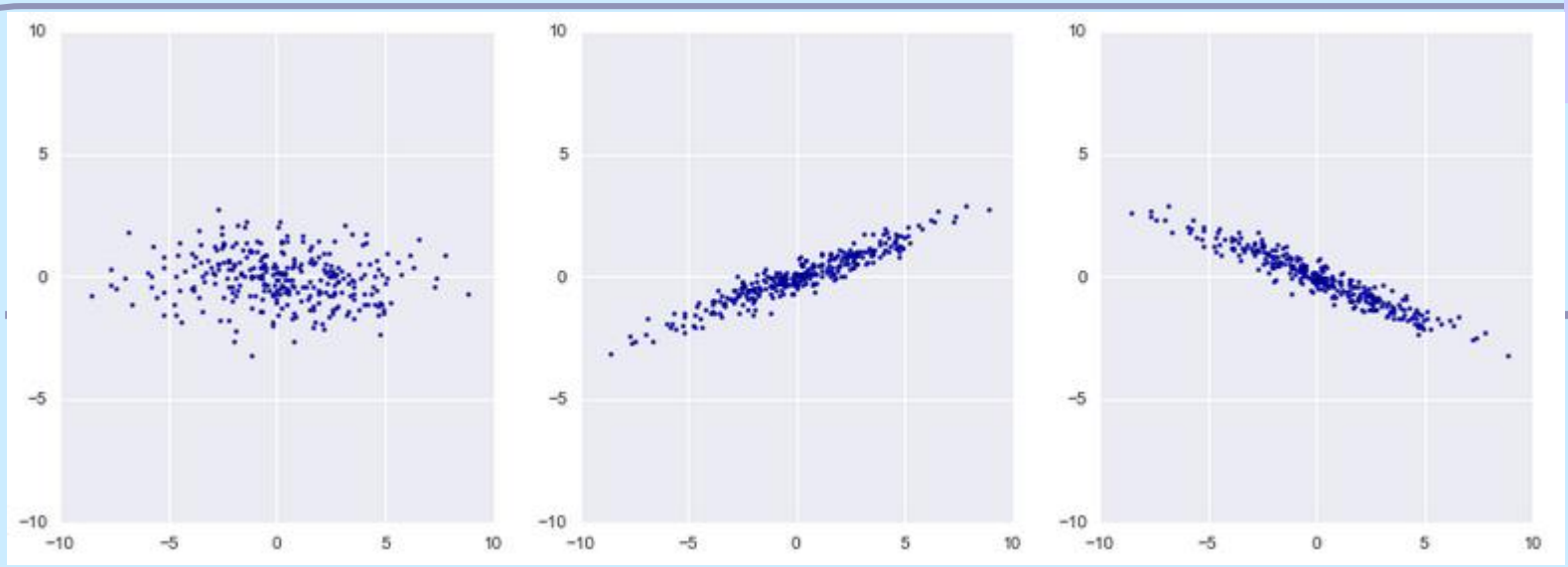
$$r_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

Пусть центрированные случайные величины $x_1,\ldots,x_r$ - исходные признаки, A - их ковариационная матрица. A - симметрична, следовательно, все ее собственные числа вещественны. Обозначим их $\lambda_1,\ldots,\lambda_r$ в порядке убывания. Предположим, что все собственные числа различны и положительны. Для большинства практических задач это предположение обычно верно.

Математическое ожидание центрированной случайной величины равно нулю.

Ковариация центрированных случайных величин

$$Cov(X_i, X_j) = E(X_i X_j)$$



Для описания формы случайного вектора необходима ковариационная матрица.

для описания формы распределений недостаточно только ее дисперсий по осям, у всех трех случайных величин одинаковые мат.ожидания и дисперсии, а их проекции на оси в целом окажутся одинаковы!

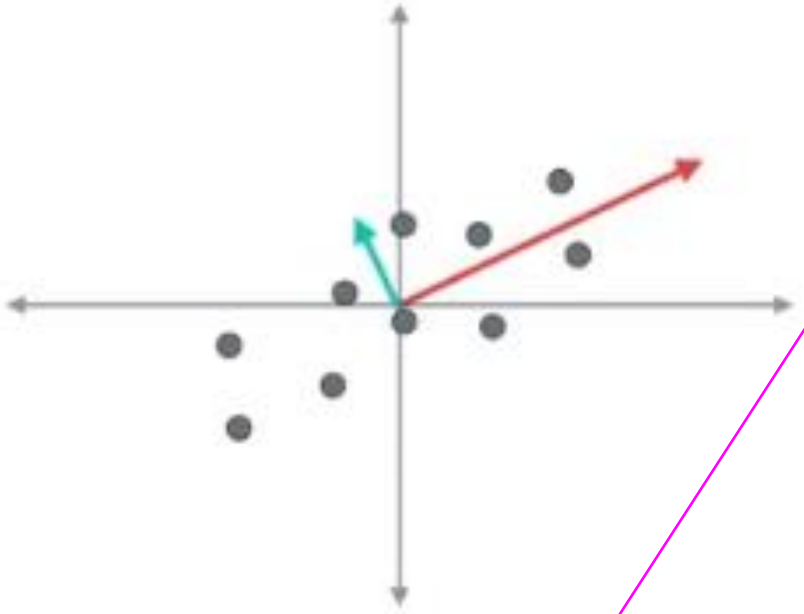
Ковариационная матрица является обобщением дисперсии на случай многомерных случайных величин – она так же описывает форму (разброс) случайной величины, как и дисперсия.

Ковариационная матрица

- По диагонали – дисперсия
- Попарные ковариации

$$\begin{bmatrix} V_a & C_{ab} & C_{ac} & C_{ad} \\ C_{ab} & V_b & C_{bc} & C_{bd} \\ C_{ac} & C_{bc} & V_c & C_{cd} \\ C_{ad} & C_{bd} & C_{cb} & V_d \end{bmatrix}$$

Principal Component Analysis (PCA)



$$\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

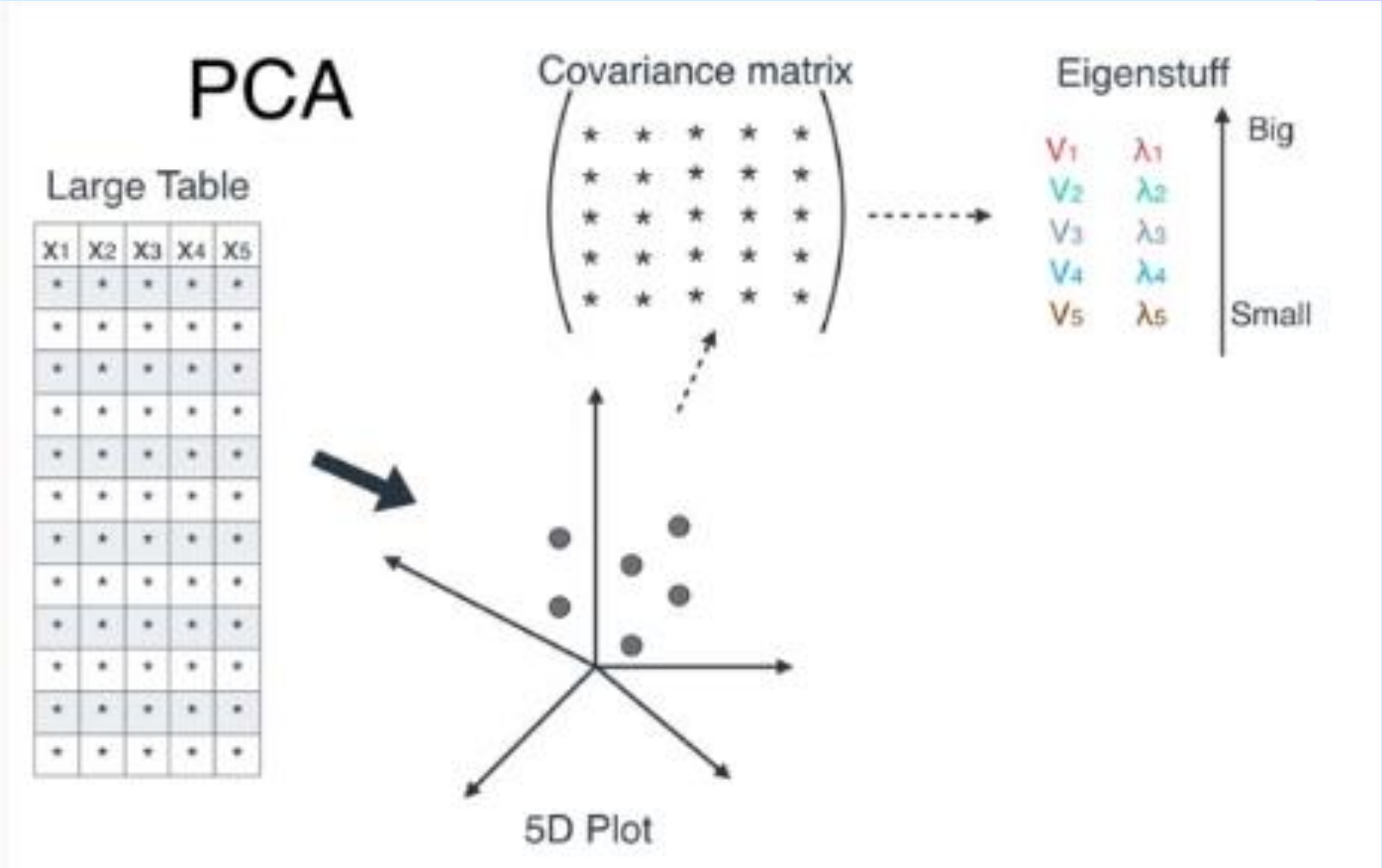
$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$

Eigenvectors (direction)

11 1

Eigenvalues (magnitude)

- Ковариационная матрица.


PCA
Large Table
X1 X2 X3 X4 X5
Covariance matrix
Eigenstuff
V1 λ1
V2 λ2
V3 λ3
V4 λ4
V5 λ5
Big
Small
5D Plot

Можно показать, что

- Собственные векторы ковариационной матрицы соответствуют направлению максимальной изменчивости.
- Геометрически – набор собственных векторов является матрицей поворота перехода от одного базиса к другому.
- Собственные числа соответствуют масштабированию исходных признаков по каждой из осей.

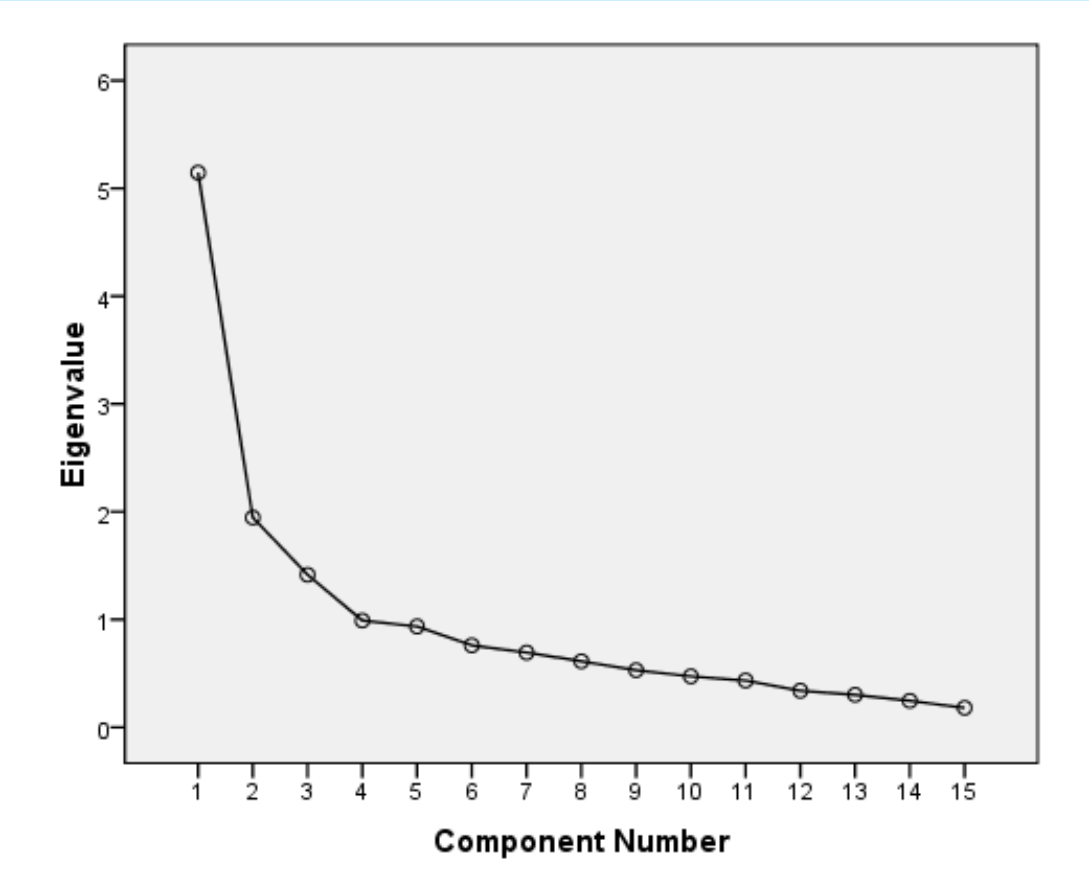
PCA. С чего начать?

- Подготовить данные. Без выбросов.
- Стандартизировать переменные.
- Проверить применимость метода главных компонент.
- Определить количество главных компонент (факторов).

Методы определения числа факторов

- Сколько собственных чисел больше 1?
- Сколько собственных чисел больше 0.8?
- График каменистая осыпь

Графический метод определения количества факторов



Определение числа факторов

- Анализ главных компонент, анализируется ковариационная (или корреляционная) матрица
- Собственные числа == дисперсии главных компонент (Eugenvalues)
- Полная дисперсия (= числу переменных)
- Объясненная дисперсия (70%, 80%, 90%)

Анализ главных компонент. Определение числа факторов:

- Собственные значения сортируются в порядке убывания, для чего обычно отбирается столько факторов, сколько имеется собственных значений, превосходящих по величине единицу.
- Собственные векторы, соответствующие этим собственным значениям, образуют *факторы*; элементы собственных векторов получили название *факторной нагрузки*.

Их можно понимать как коэффициенты корреляции между соответствующими переменными и факторами.

ПРИМЕР 1.

	L	M	P	A	V
1970	68.90	1060	7.80	5.50	25.30
1975	68.10	1101	9.50	15.30	28.00
1980	67.60	1147	10.10	30.20	30.00
1985	69.20	1204	10.00	44.50	23.50
1990	69.20	1602	9.80	58.60	18.00
1995	64.60	1893	5.50	93.30	38.40
1998	67.00	2777	6.20	122.00	29.60

Данные о средней продолжительности жизни и сопутствующих факторах.

Признаки: L средняя продолжительность жизни; M количество чиновников; A количество автомобилей; P доходы бедных; V объемы продажи водки. Вклад первого фактора равен 72%.

Стандартизируем переменные.
В этом случае ковариационная матрица совпадает с корреляционной.

Еф

	L	M	P	A	V
L	1.00	-0.50	0.77	-0.60	-0.93
M	-0.50	1.00	-0.70	0.95	0.30
P	0.77	-0.70	1.00	-0.67	-0.68
A	-0.60	0.95	-0.67	1.00	0.37
V	-0.93	0.30	-0.68	0.37	1.00

Корреляционная матрица.

	1	2	3	4	5
1	3.60	1.09	0.24	0.05	0.02
2	72.00	93.80	98.70	99.63	100.00

Собственные числа и суммарный вклад компонент в общую дисперсию.

При восстановлении переменных по m главным компонентам, меньшему количества исходных признаков k, значения признаков могут восстанавливаться с ошибками. Чем больше вклад используемых в восстановлении главных компонент, тем меньше ошибки восстановления.

Вычисление коэффициентов главных компонент

Пусть признаки $X = (X_1, \ldots, X_k)^T$ центрированы $EX_i = 0$ и имеют ковариационную матрицу $\Sigma = EXX^T$. Обозначим через A_i собственные векторы матрицы Σ, соответствующие собственным числам λ_i

определим главную компоненту как

$$Y_j = A_j^T X = \sum_{i=1}^{k} a_{ij} X_i.$$

Для всех главных компонент справедливо выражение

$$Y = \begin{bmatrix} Y_1 \\ \cdot \\ Y_k \end{bmatrix} = \begin{bmatrix} A_1^T X \\ \cdot \\ A_k^T X \end{bmatrix} = \mathcal{A}^T X, \quad \text{откуда} \quad X = \mathcal{A} Y.$$

Упорядочим собственные числа по убыванию: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$. Соберем все собственные вектора в одну ортогональную матрицу

$$\mathcal{A} = [A_1, \ldots, A_k] = \begin{bmatrix} a_{11}, & \ldots, & a_{1k} \\ \cdot & \cdot & \cdot \\ a_{k1}, & \ldots, & a_{kk} \end{bmatrix}, \quad \mathcal{A}^T \mathcal{A} = I = \begin{bmatrix} 1, & \ldots, & 0 \\ \cdot & \cdot & \cdot \\ 0, & \ldots, & 1 \end{bmatrix}$$

	1	2	3	4	5
1	0.47	-0.38	0.30	0.50	-0.54
2	-0.43	-0.54	-0.03	0.54	0.48
3	0.48	-0.04	-0.85	0.17	0.13
4	-0.45	-0.47	-0.38	-0.33	-0.57
5	-0.41	0.59	-0.20	0.57	-0.35

Собственные векторы.

	1	2	3	4	5
1970	0.60	0.47	2.01	0.17	-0.06
1975	0.54	0.71	-0.16	0.66	0.89
1980	0.37	0.79	-1.28	0.67	0.36
1985	0.72	-0.34	-0.46	-0.11	-2.02
1990	0.66	-1.30	-0.23	-1.41	0.98
1995	-1.59	1.02	-0.04	-1.23	-0.21
1998	-1.31	-1.34	0.16	1.26	0.07

Значения главных компонент

- Матрица собственных векторов ковариационной (или корреляционной) является **матрицей перехода** от исходных переменных к главным компонентам.

Интерпретация

Корреляция $\beta_{ij} = \mathrm{cor}(X_i, Y_j)$ между признаком X_i и главной компонентой Y_j называется факторной нагрузкой.

	1	2	3	4	5
L	0.90	-0.40	0.15	0.11	-0.07
M	-0.82	-0.56	-0.02	0.12	0.07
P	0.90	-0.04	-0.42	0.04	0.02
A	-0.85	-0.49	-0.19	-0.07	-0.08
V	-0.77	0.61	-0.10	0.12	-0.05

Матрица факторных нагрузок

факторные нагрузки - коэффициенты корреляции между признаками и факторами

Признаки: L средняя продолжительность жизни; M количество чиновников; A количество автомобилей; P доходы бедных; V объемы продажи водки. Вклад первого фактора равен 72%.

Значения в i-й строке и j-м столбце соответствуют коэффициенту корреляции между i-м признаком и j-й главной компонентой. Чем больше первый фактор, тем больше продолжительность жизни и доходы бедных, меньше чиновников и автомобилей и не много водки фактор какого-то благополучия.

Факторные нагрузки
Главные компоненты

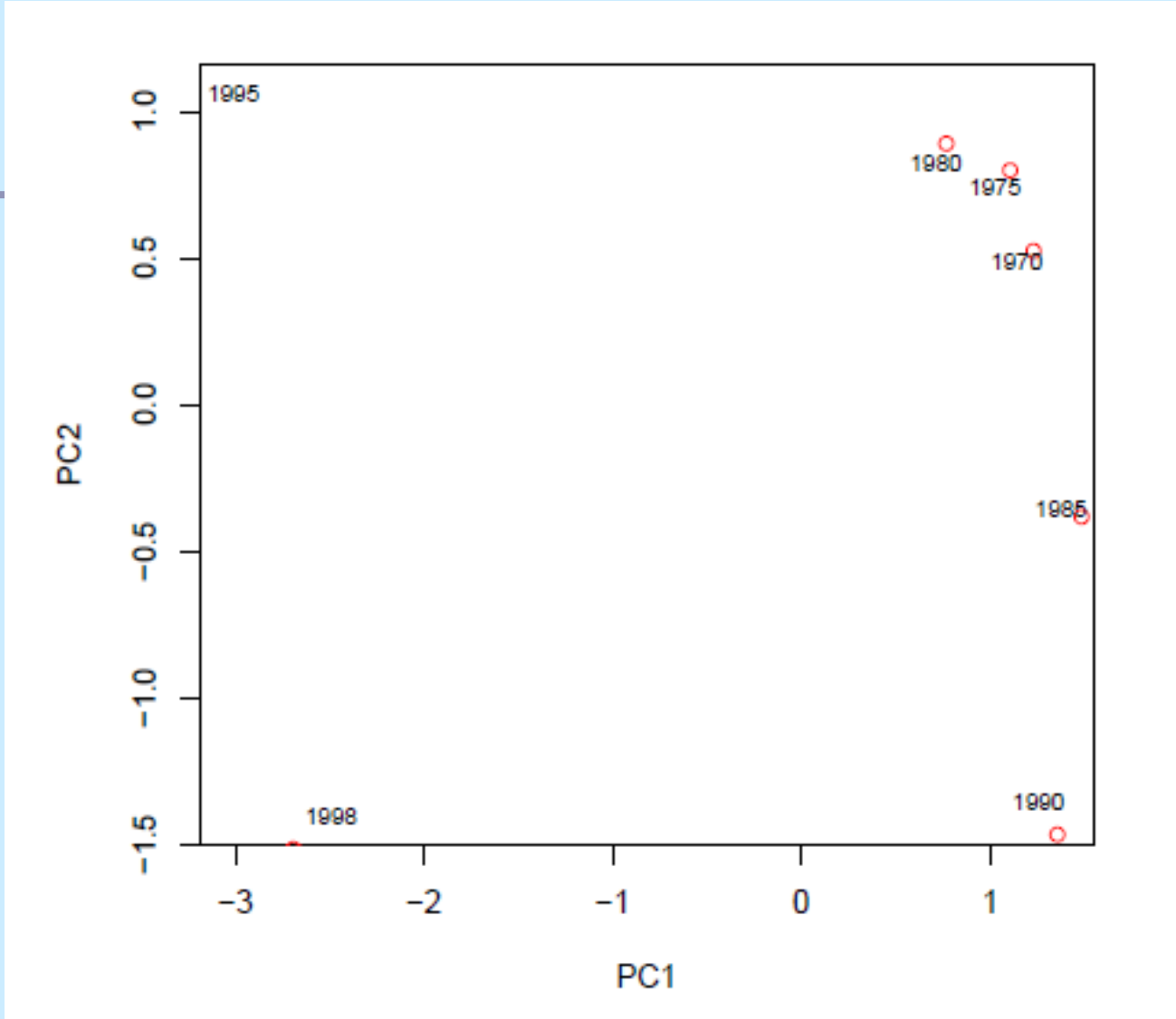
Корреляция $\beta_{ij} = \mathrm{cor}(X_i, Y_j)$ между признаком X_i и главной компонентой Y_j называется факторной нагрузкой.

факторные нагрузки

	Factor 1	Factor 2
L	0.896	−0.398
M	−0.815	−0.564
P	0.905	−0.045
A	−0.847	−0.486
V	−0.772	0.613
Дисп.гл.комп. λ	3.60	1.06
вклад в дисп.	72%	21.8%

Значения факторов

годы	f_1	f_2
1970	0.600	0.465
1975	0.540	0.710
1980	0.375	0.790
1985	0.724	−0.339
1990	0.662	−1.301
1995	−1.587	1.019
1998	−1.314	−1.345


1995
1980
1975
1970
1985
1990
1998
PC2
PC1
1.0
0.5
0.0
−0.5
−1.0
−1.5
−3
−2
−1
0
1