САНКТ-ПЕТЕРБУРГСКИЙ
ГОСУДАРСТВЕННЫЙ
ЭКОНОМИЧЕСКИЙ
УНИВЕРСИТЕТ

Метод главных компонент (PCA)

Занятие 4

Глазунова Е.В.

Задачи, решаемые PCA

1. Сокращение числа переменных.
2. Измерение неизмеримого. Построение новых обобщенных показателей.
3. Наглядное представление многомерных наблюдений (проецирование данных).
4. Описание структуры взаимных связей между переменными, в частности выявление групп взаимозависимых переменных.
5. Преодоление мультиколинеарности переменных в регрессионном анализе
6. И так далее…

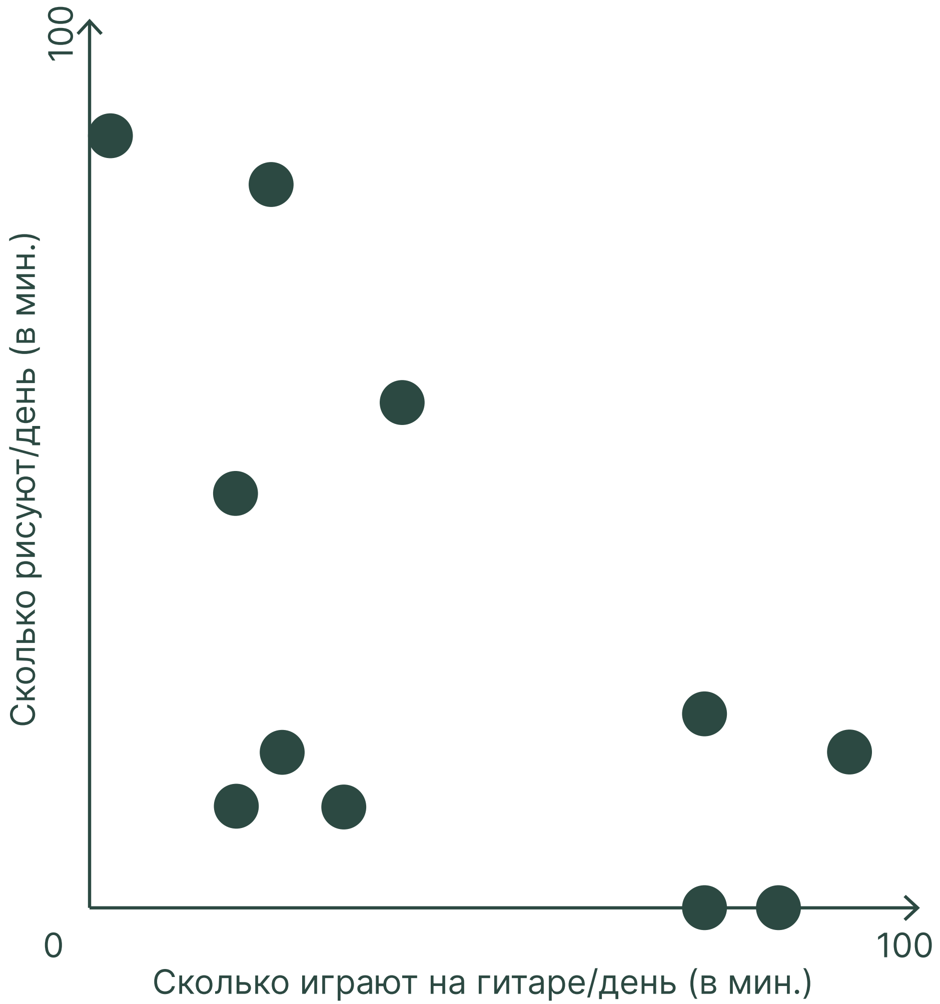
Задачи, решаемые PCA

Сокращение числа переменных, пример Б. Шоу

- Начало прошлого века
- Зависимость
- Носит цилиндр – шире грудная клетка
- Абонемент на место в церкви – дольше живет
- Чаще моется – любит оперы Вагнера

Задачи, решаемые PCA

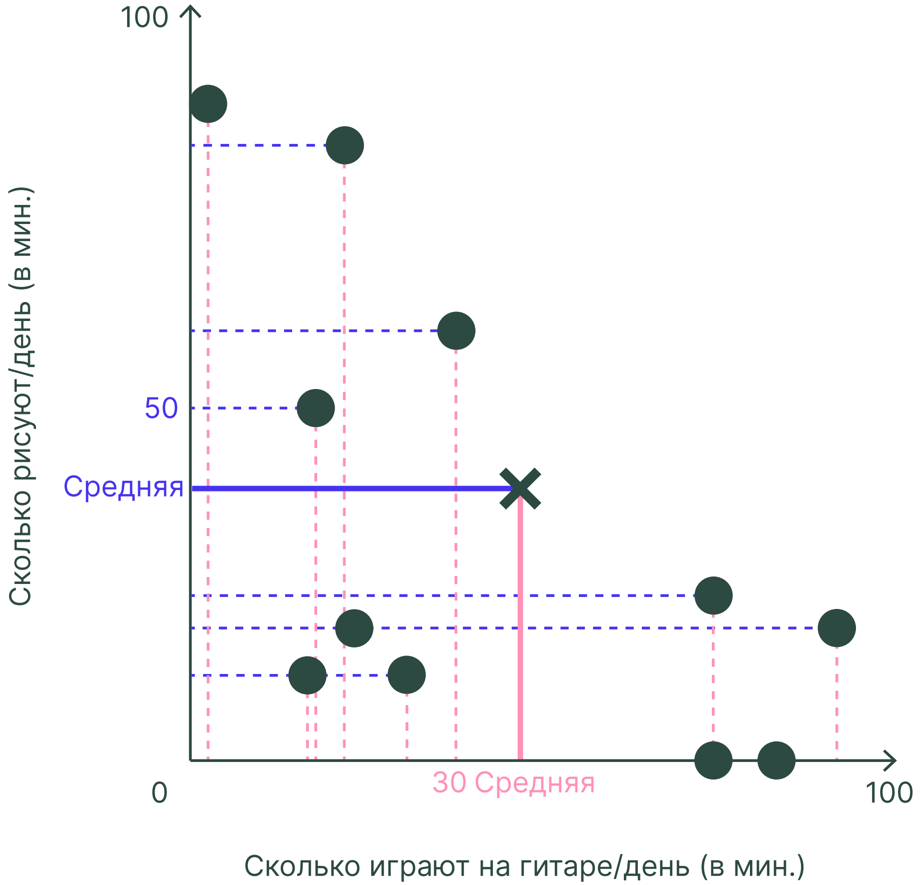
Пример



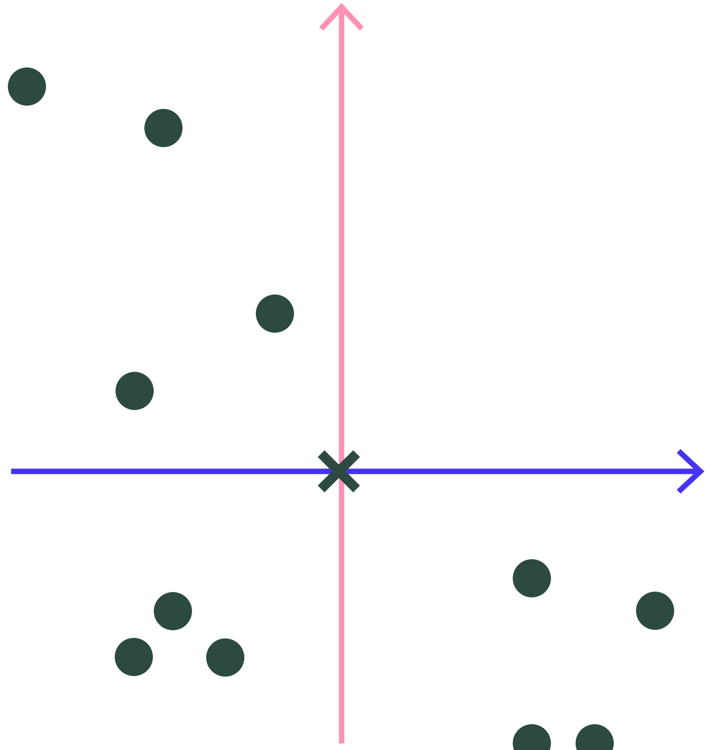
100
Сколько рисуют/день (в мин.)
0
100
Сколько играют на гитаре/день (в мин.)

Пример

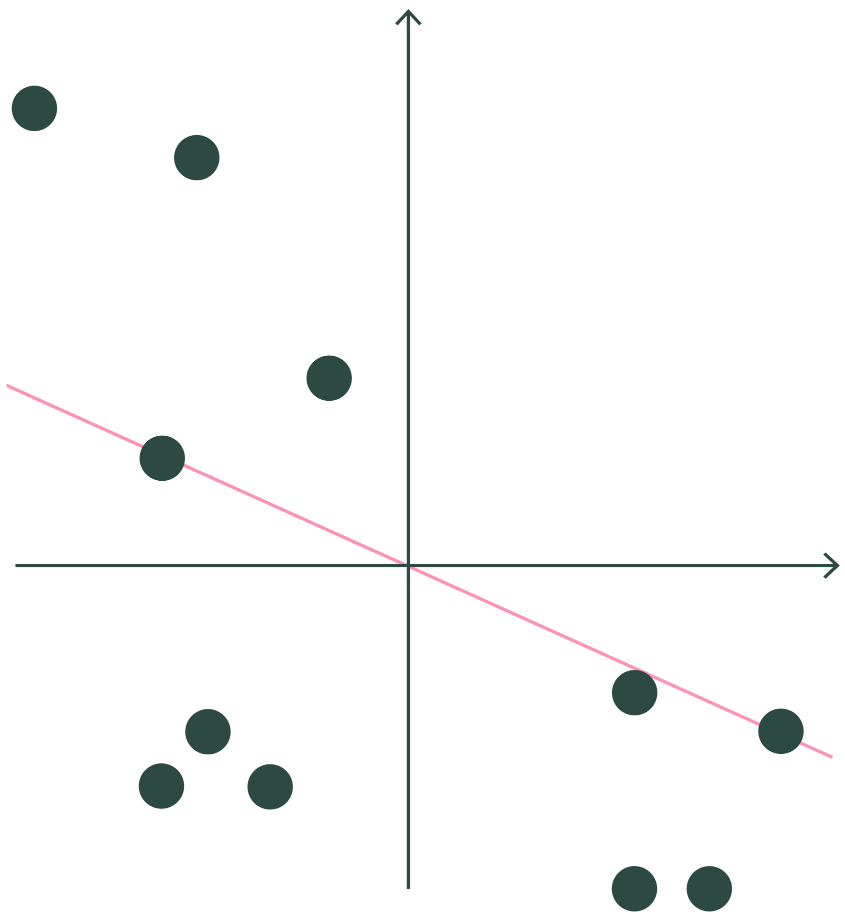
Нецентрированный график



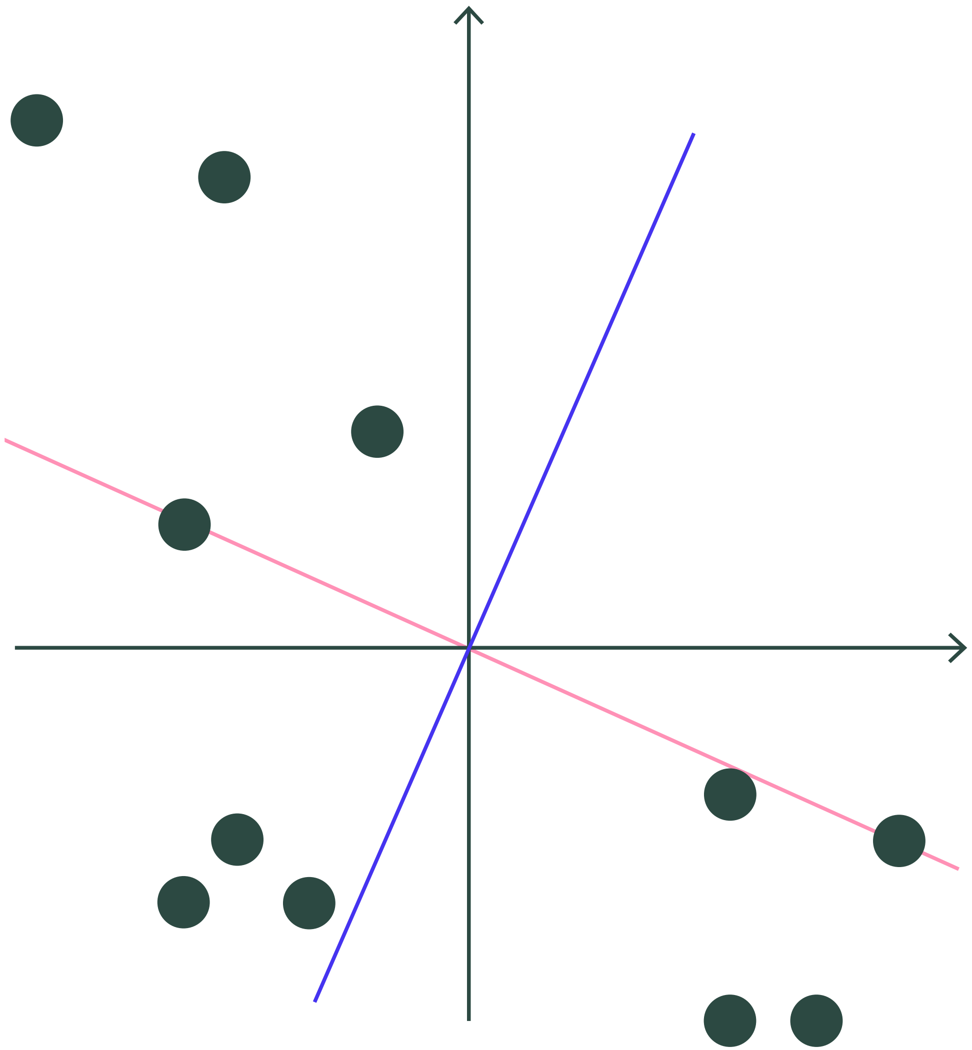
Центрированный график



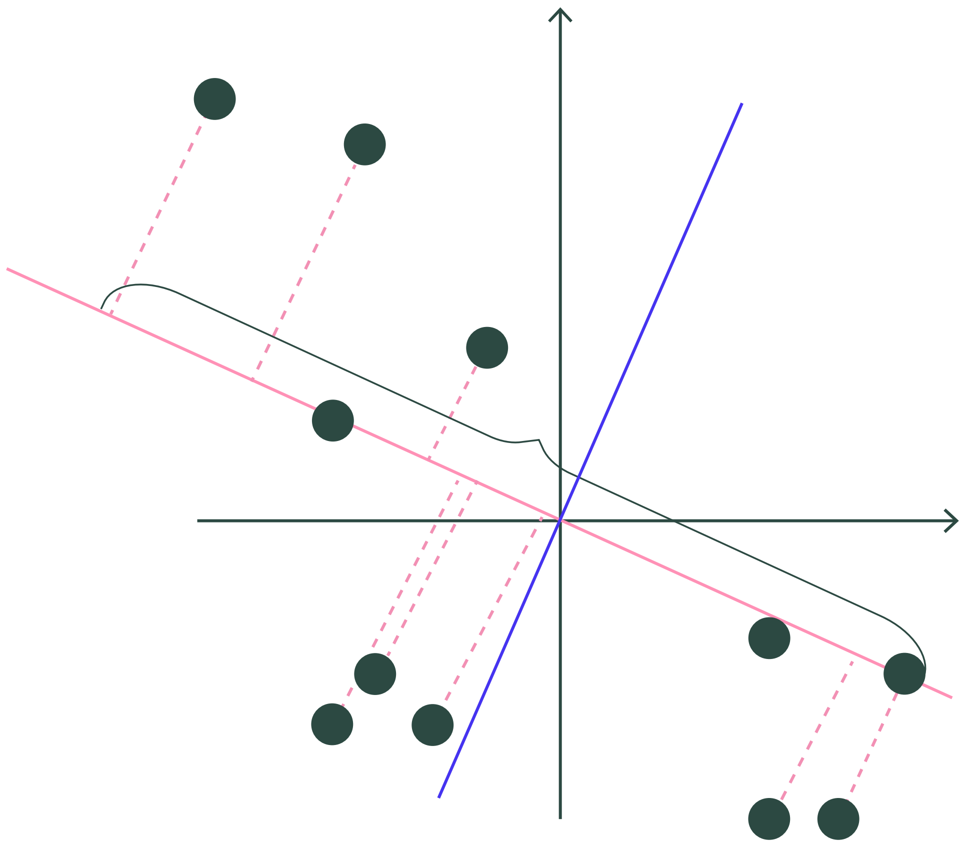
Пример



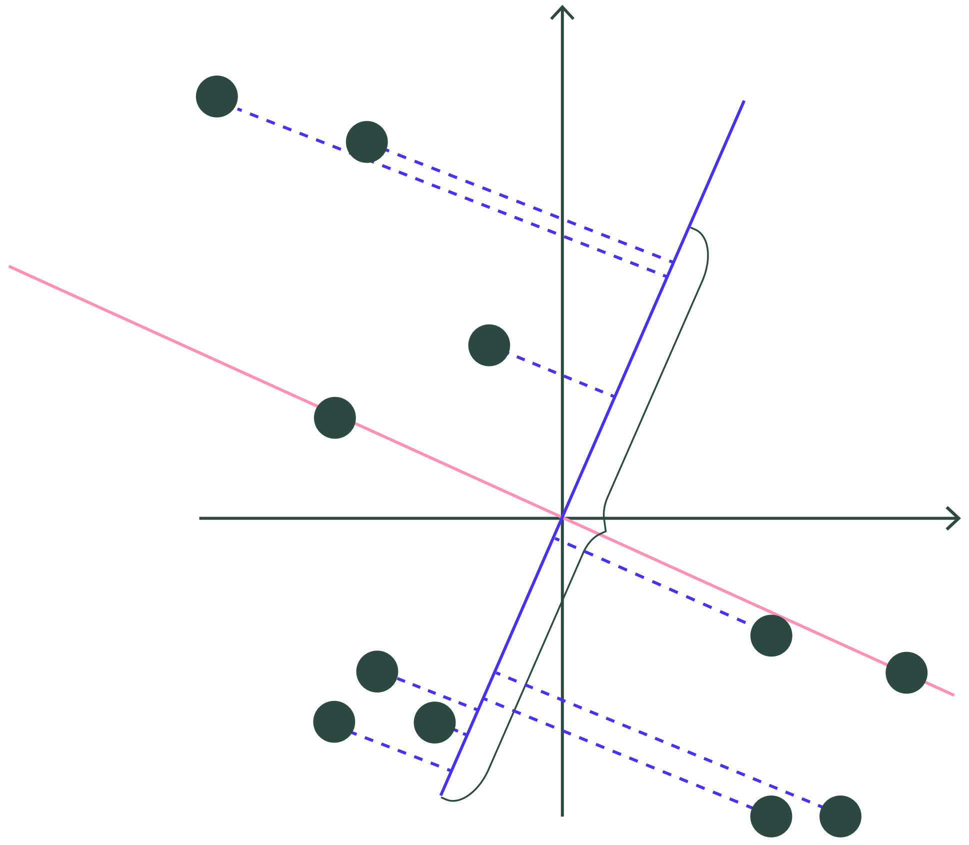
Пример



Пример

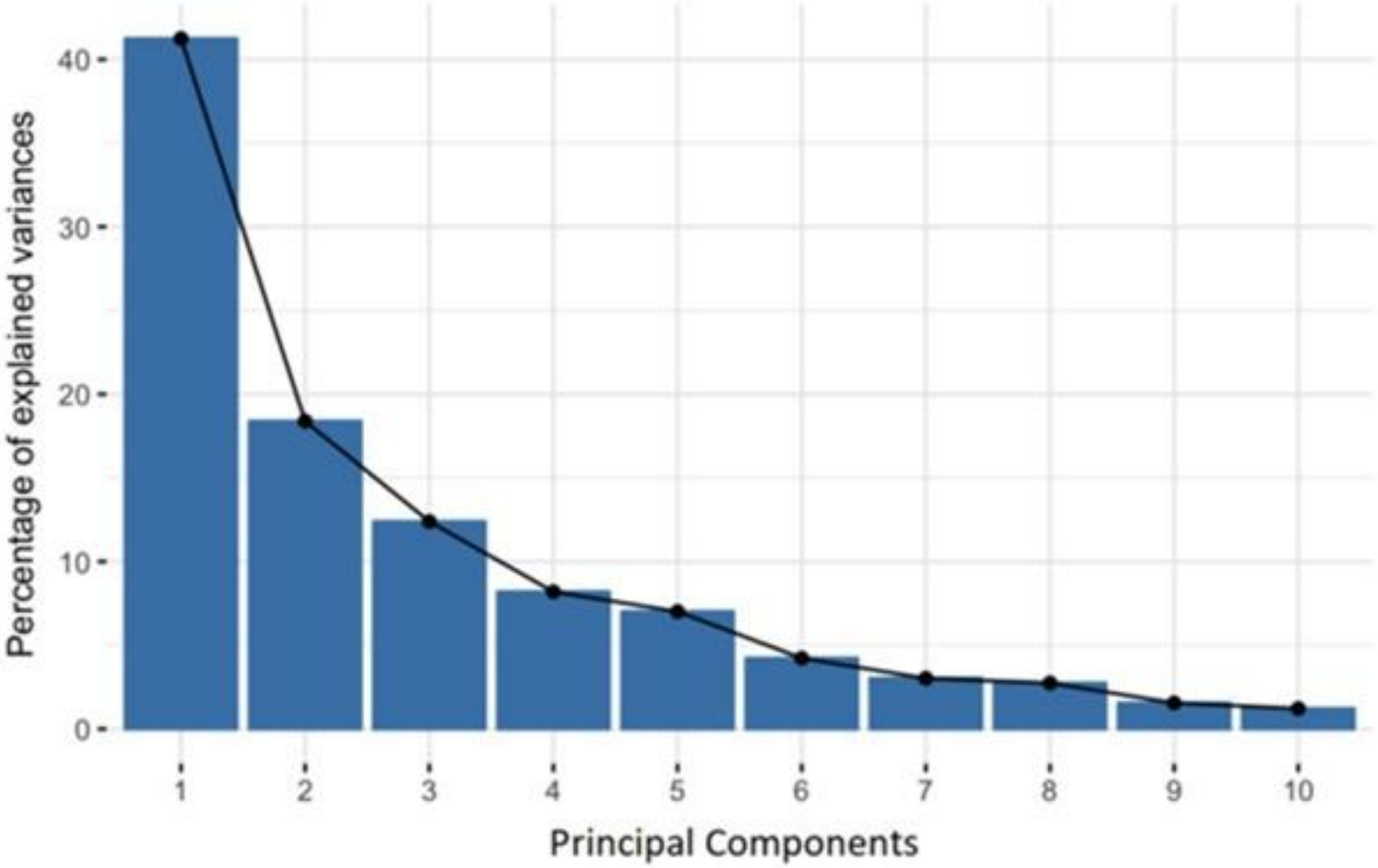


Диапазон дисперсии главного компонента 1



Диапазон дисперсии главного компонента 2

Пример



Percentage of explained variances
0
10
20
30
40
1
2
3
4
5
6
7
8
9
10
Principal Components

Постановка задачи и теорема

Рассмотрим матрицу F, строки которой соответствуют признаковым описаниям обучающих объектов:

$$F = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_l) & \dots & f_n(x_l) \end{pmatrix} = \begin{pmatrix} x_1 \\ \dots \\ x_l \end{pmatrix}$$

Пусть матрица G – признаковое описание тех же объектов в новом пространстве меньше размерности $m < n$

$$G = \begin{pmatrix} g_1(x_1) & \dots & g_m(x_1) \\ \dots & \dots & \dots \\ g_1(x_l) & \dots & g_m(x_l) \end{pmatrix} = \begin{pmatrix} z_1 \\ \dots \\ z_l \end{pmatrix}$$

Потребуем, чтобы исходные признаковые описания можно было восстановить по новым описаниям с помощью некоторого линейного преобразования, определяемого матрицей U.

Будем искать одновременно и матрицу новых признаковых описаний G, и матрицу линейного преобразования U, при которых суммарная невязка $\Delta_2(G,U)$ восстановленных описаний минимальна:

$$\Delta_2(G,U) = \|GU^T - F\|^2 \to \min_{G,U}$$

Теорема

Если $m \leq rankF$, то минимум $\Delta_2(G,U)$ достигается, когда столбцы матрицы U есть собственные векторы F^TF, соответствующие m максимальным собственным значениям. При этом $G = FU$, матрицы U и G ортогональны

PCA по шагам

1. Нормируем исходный набор данных, приводя их к единому масштабу (стандартизируем)

2. Вычисляем матрицу ковариации

3. Делаем разложение матрицы ковариации на собственные вектора и собственные числа

4. Ранжируем собственные числа по убыванию. Чем больше число - тем больше дисперсия.

5. Берем m первых собственных векторов, которые соответствуют первым m собственным числам. Это и есть искомые главные компоненты.

6. Проецируем данные на главные компоненты (находим FU)