



# Конспект > 15 урок > Понижение размерности признакового пространства

## >Оглавление

### >Оглавление

#### >Понижение размерности

Общая постановка задачи:

Мотивация:

Отбор признаков:

Извлечение признаков:

#### >Преимущество извлечения признаков

Мотивация:

#### >Метод главных компонент (PCA)

Основная идея:

#### >Геометрическая интерпретация PCA

#### >PCA. Визуализация

#### >T-SNE

Основная идея:

Возникает вопрос:

Что же нужно оптимизировать?

#### >T-SNE vs PCA

PCA

T-SNE

## >Понижение размерности

### Общая постановка задачи:

Допустим, имеется матрица объект-признак  $n * M$ , но по какой-то из причин нам нужно перевести объекты в пространство  $n * m$ , где  $m < M$ .

### Мотивация:

- Борьба с переобучением (например, через устранение мультиколлинеарности)
- Интерпретируемость и скорость обучения модели
- Извлечение новых, более сильных признаков на основе предыдущих
- Визуализация классов, когда признаков много

Приемы понижения размерности признакового пространства можно разделить на два направления: Отбор признаков и Извлечение признаков.

### Отбор признаков:

Под отбором признаков мы подразумеваем те приемы, которые были изучены в предыдущей лекции (методы фильтрации, методы обертки и т.д). Они позволяют отобрать  $m$  важных признаков среди того множества, которое у нас уже есть. Не создаются новые фишки, просто убираются самые ненужные из существующих.

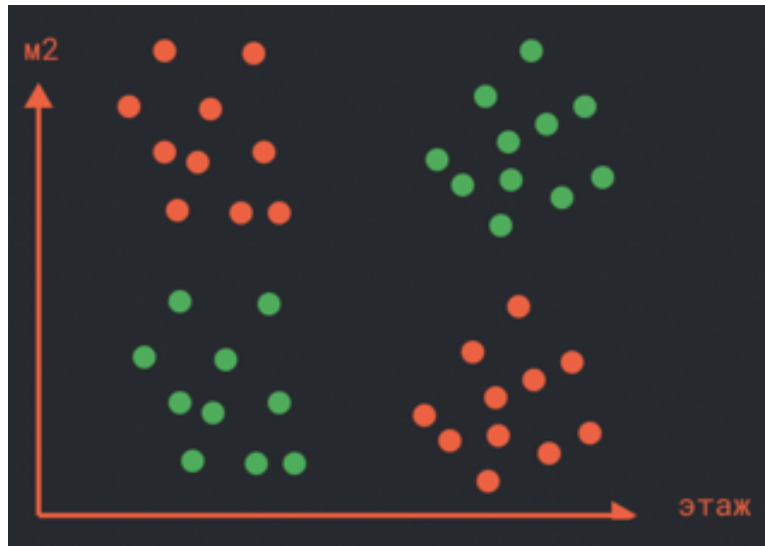
### Извлечение признаков:

Приемы извлечения признаков не просто пытаются сократить количество признаков путем удаления "плохих", а путем комбинирования предыдущих.

## >Преимущество извлечения признаков

Приведем пример, когда извлечение признаков может оказаться выигрышнее, чем просто их отбор.

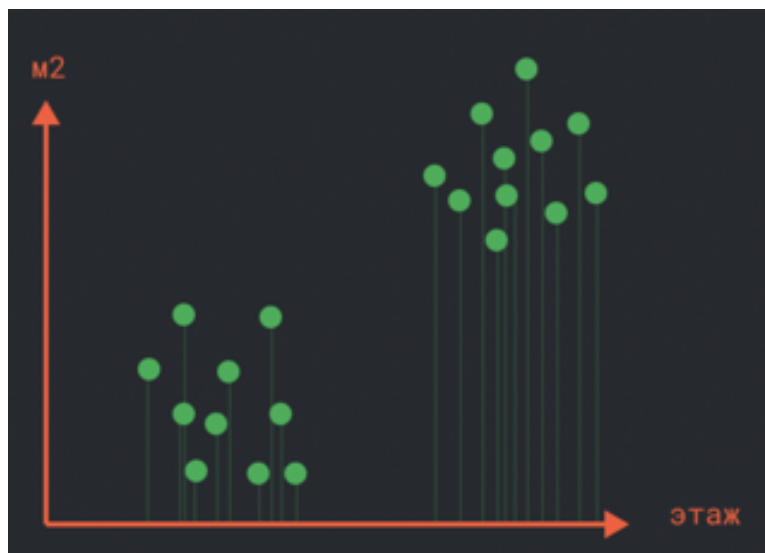
Допустим, мы решаем задачу классификации: есть всего два признака и два класса. Также мы пытаемся использовать уже известные нам методы визуализации для того, чтобы понять важны ли в отдельности признаки "этаж" и "квадратный метр".



Какой график можно построить, если решается задача классификации?

Мы можем, например, построить гистограммы распределения каждого признака для зеленого и красного классов в отдельности и потом сравнить полученные распределения.

Возьмем сначала зеленый класс и признак этаж:



Давайте произведем проекцию этих точек на ось с признаком "этаж". Получим реализацию признака "этаж" для класса зеленых объектов.

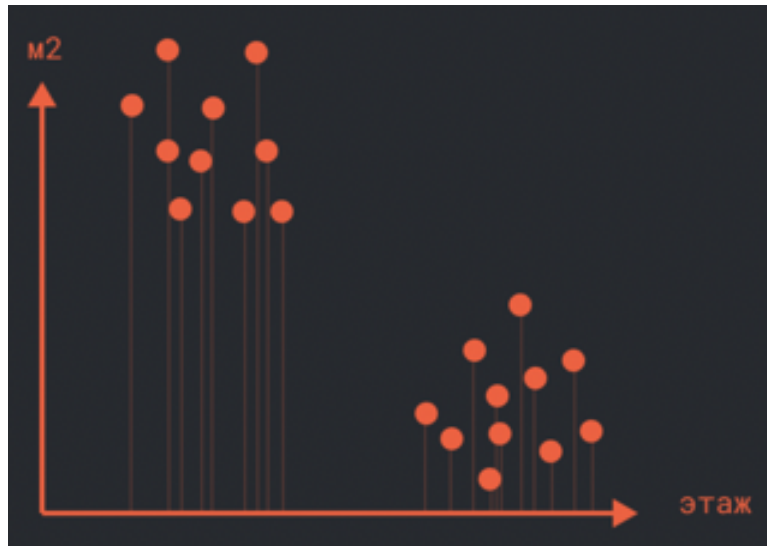


На основании точек, их количества, частот можем построить гистограмму.

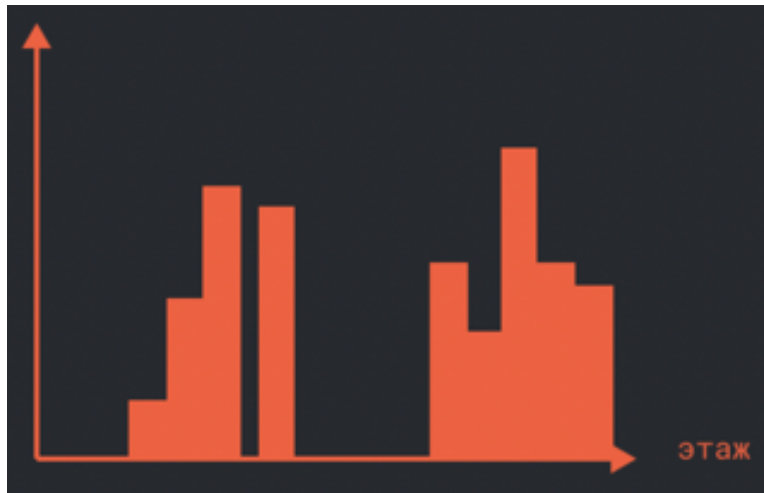


Теперь давайте сделаем то же самое для красных объектов:

Посмотрим, с какой частотой признак "этаж" и в каких числах реализуется для красного класса.



Построим гистограмму:



Давайте сравним гистограммы распределения признака "этаж" для класса зеленых и для класса красных. Оказывается, гистограммы очень похожи между собой. Может сложиться впечатление, что признак неважный и его можно исключить, но в дальнейшем мы узнаем, что их можно скомбинировать в один более сильный признак

### Мотивация:

- Методы извлечения признаков позволяют сохранить потенциально полезную информацию изначального датасета
- Это может дать не только техническое преимущество, но и улучшить модель

## >Метод главных компонент (PCA)

### Основная идея:

Во-первых, нужно решить, какое количество признаков мы хотим получить. Новые признаки, которые будут образовывать новое признаковое пространство будем получать как линейные комбинации изначальных фичей. Т. е в итоге у нас будет  $m$  новых признаков, полученных следующим образом:

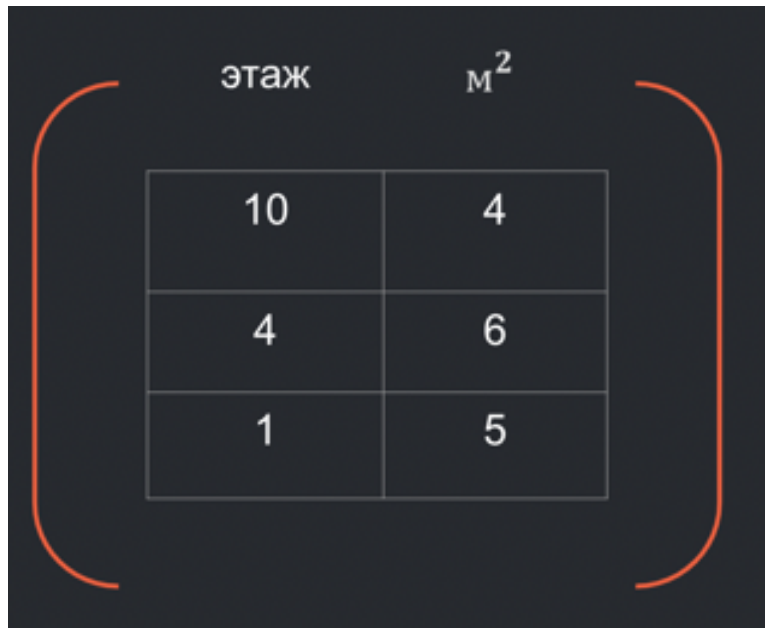
$$z_j = \sum_i^M w_{ij} * d_i$$

Как подобрать веса  $w_{ij}$ ?

Как минимум мы можем сказать, что нужно подобрать такие веса, чтобы в новых признаках осталось как можно больше полезной информации из изначального

датасета.

Например, в выборке было три объекта с двумя признаками. Изначально мы находимся в двумерном пространстве, но нам нужно, чтобы остался всего один признак. При этом мы понимаем, что отбор признаков может работать плохо. Поэтому мы хотим получить новый более сильный признак как комбинацию из двух старых.



этаж	м²
10	4
4	6
1	5

Пусть новый признак будет  $z_1$

Т.е мы должны взять старый признак "этаж", старый признак "квадратный метр", умножить их на какие-то веса и просуммировать полученные результаты. Но на каком основании выбирать веса?

Повторим старую идею:

Мы хотим получить как можно более полное отображение в новое признаковое пространство, т.е сохранить как можно больше информации о старом датасете.

В качестве критерия информативность можно использовать дисперсию в новом пространстве: чем она меньше, тем меньше информации дает новый признак.

Как можно найти данный коэффициент?

$$\sigma_i^2 = \sum_i z_j^2 = \sum_i (w_{\text{этаж}} * d_{\text{этаж}}^i + w_{\text{м}^2} * d_{\text{м}^2}^i)^2$$

В целом, мы знаем эту формулу, но есть небольшая модификация. Можно заметить, что от  $z_j^2$  не отнимается среднее. Если мы будем еще отнимать среднее, то формула получится достаточно сложной, в ней придется достаточно сложно считать производную, поэтому мы среднее опускаем. А чтобы опустить среднее, нужно применить один прием к изначальному датасету.

Нужно изначальные признаки центрировать, т.е. отнять от каждого значения признака среднее значение этой фичи.

Для примера выше получим следующую таблицу:



этаж	м²
5	-1
-1	1
-4	0

Давайте теперь для нашей выборки выпишем коэффициент дисперсии:



$$\begin{aligned}
 & (w_{\text{этаж}} \cdot 5 + w_{M^2} \cdot (-1))^2 + \\
 & (w_{\text{этаж}} \cdot (-1) + w_{M^2} \cdot 1)^2 + \\
 & (w_{\text{этаж}} \cdot (-4) + w_{M^2} \cdot 1)^2 \rightarrow \max \\
 & \text{s.t} \\
 & w_{\text{этаж}}^2 + w_{M^2}^2 = 1
 \end{aligned}$$

Мы получили какую-то функцию, которая должна быть как можно больше. Давайте ее промаксимизируем и на этой основе найдем нужные коэффициенты. Но только данную функцию мы будем оптимизировать с некоторым ограничением.

В общем случае можно проделать все те же самые действия. Мы будем максимизировать не дисперсию одной новой фичи, а их сумму.

$$\begin{aligned}
 z_1 &= \sum_i^M w_{i1} \cdot d_i \\
 &\dots \\
 z_m &= \sum_i^M w_{im} \cdot d_i \\
 \\ 
 VaR_{sum} &= \sum_j^m \sigma_j^2 = \sum_j^m \sum_i z_j^2 = \sum_j^m \sum_i (\sum_l^M w_{ij} \cdot d_l)^2 \rightarrow \max \\
 &\text{s.t} \\
 &\sum_k^M w_k^2 = 1
 \end{aligned}$$

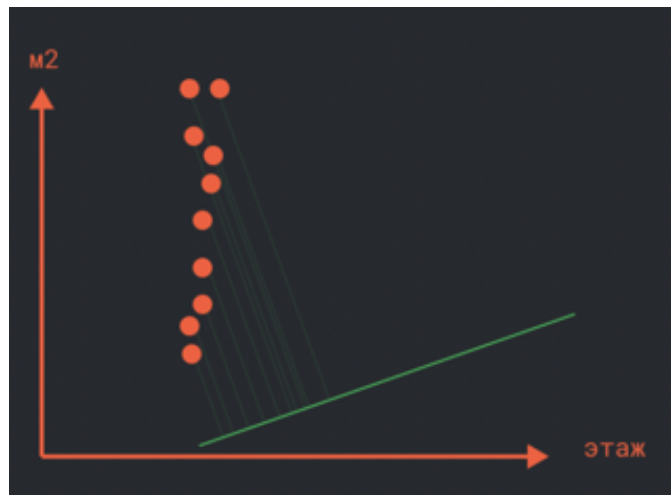
## >Геометрическая интерпретация PCA

Давайте попробуем дать методу PCA (метод главных компонент) геометрическую интерпретацию.

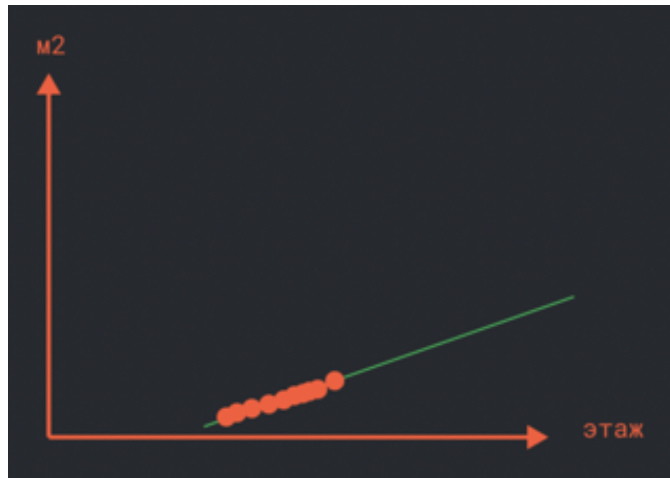
Пусть есть объекты в двумерном пространстве, которые распределены следующим образом:



Когда мы подбираем коэффициенты  $w$  для перевода старых признаков в новые, мы находим в изначальном признаковом пространстве гиперплоскость для проецирования наших точек. Например, мы могли получить следующий результат:

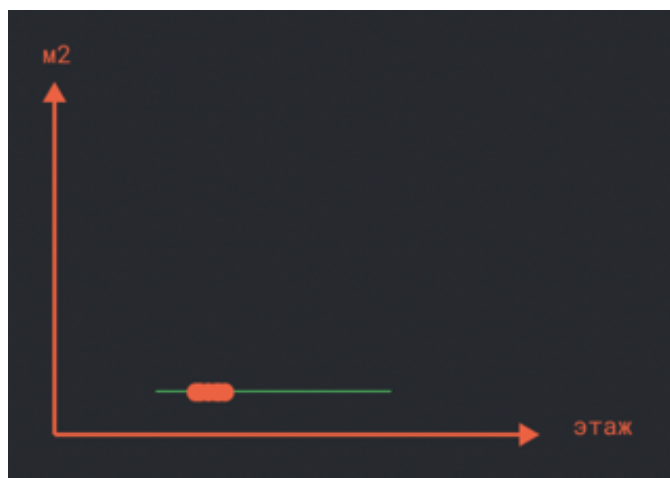
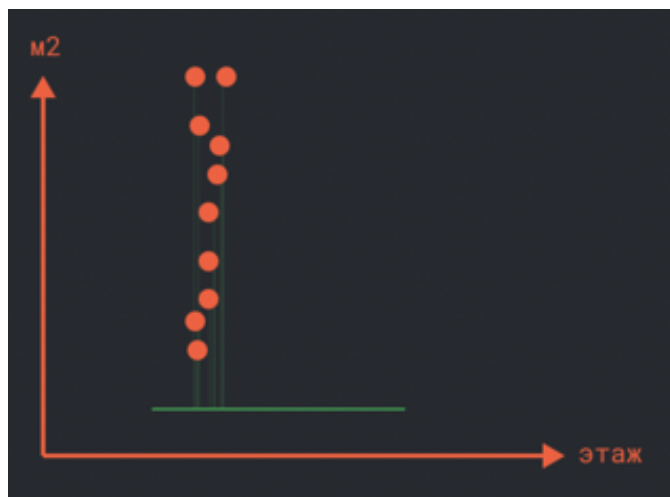


Давайте теперь изначальные точки спроецируем перпендикулярно на эту гиперплоскость. После того, как спроецировали, можем теперь эти точки по проекции разместить. И то расположение этих точек, которое получится графически, это и будет реализацией нашего нового признака.

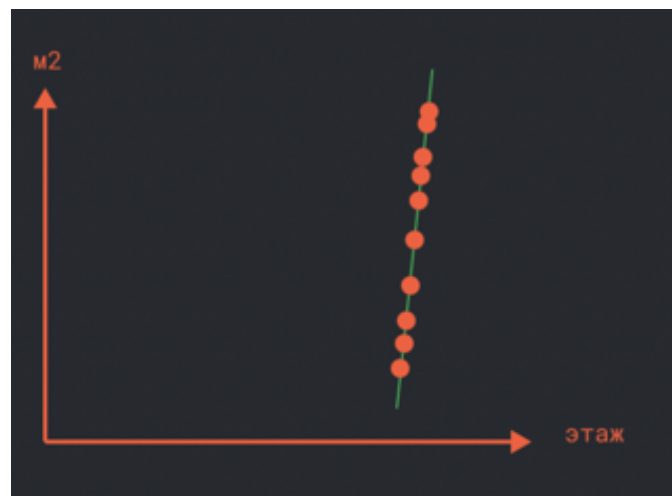
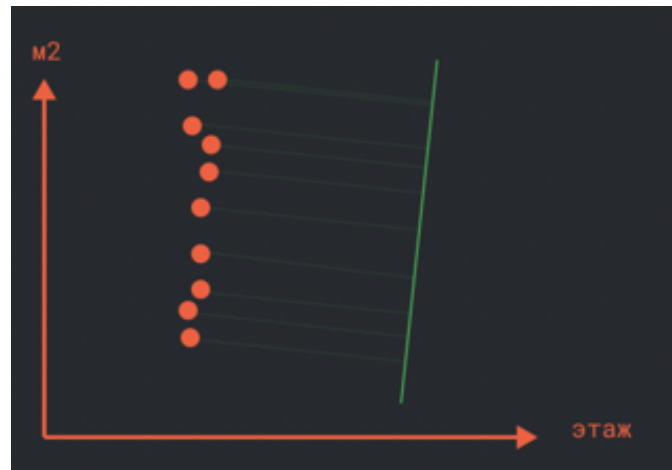


Рассмотрим еще несколько примеров.

Пример 2:



Пример 3:



Давайте их сравним:

Вспомним, что мы максимизировали дисперсию. Очевидно, что максимальной дисперсия оказалась в Примере 3, а самый плохой результат показал Пример 2.

## >РСА. Визуализация

Как применить данный метод для визуализации?

Представим, что есть выборка, содержащая 20 вещественных и бинарных признаков. Мы решили задачу классификации, получили какие-то результаты. Но как их визуализировать?

Давайте применим метод главных компонент, и из 20 изначальных фичей выделим всего лишь две, а затем изобразим их на двумерной плоскости.



Но у данного метода есть ограничения:

Иногда очевидных направлений у данных нет, тогда линейная проекция может быть неудачной и для визуализации данный метод не очень хорошо подходит.

## >T-SNE

### Основная идея:

Пусть есть изначальное пространство признаков, тогда давайте для каждой пары объектов  $(x_i, x_j)$  научимся считать расстояние между ними. Введем какую-то меру, которая позволит оценивать насколько объекты близки друг к другу и насколько они похожи.

Мы хотим отобразить  $(x_i, x_j) \rightarrow (z_i, z_j)$ , чтобы расстояние  $\rho(x_i, x_j)$  было сильно похоже на  $\rho(z_i, z_j)$ . Если мы хотим сохранить в новом пространстве информацию о данных, тогда нам нужно, чтобы расстояния тоже были похожи.

Невозможно требовать абсолютной идентичности расстояний, но можно попробовать сохранить пропорции.

$$\frac{\rho(x_1, x_2)}{\rho(x_1, x_3)} \approx \frac{\rho(z_1, z_2)}{\rho(z_1, z_3)}$$

А как вообще при помощи этой функции  $p$  мерить, насколько похожи объекты друг на друга.

Измерять схожесть мы будем по следующей формуле:

$$p(i | j) = \frac{e^{-\frac{|x_i - x_j|^2}{2\sigma_j^2}}}{\sum_{k \neq j} e^{-\frac{|x_k - x_j|^2}{2\sigma_j^2}}}$$

В отдельности расстояние между каждой парой объектов будем мерить как экспонента в степени минус длина вектора между двумя точками в старом пространстве делить на некоторый коэффициент  $2 * \sigma_j^2$ .

Как мы будем понимать, что один объект похож на другой?

Мы посчитаем меру, которая является экспонентой в числителе, а затем поделим на схожесть этого объекта со всеми остальными, кроме того, с которым мы сравниваем.

В итоге мы получим, какое-то числовое значение, которое будет показывать сходятся ли наши объекты или нет.

### **Возникает вопрос:**

Почему мы не можем просто измерить расстояние? Идея следующая: формула выше имеет некоторое преимущество. Такая формула позволяет учитывать те объекты, которые изначально ближе к нашему объекту находятся, с большей степенью важности, чем те, которые находятся далеко.

Коэффициент  $2 * \sigma_j^2$  индивидуально подбирается для каждого объекта. Зачем он нужен?

Идея в том, что мы не хотим, чтобы распределение всех расстояний до объекта не было равномерным. Кроме того, распределение должно быть невырожденное.

Кроме того есть еще один дополнительный технический шаг, который необходимо сделать для всех попарных схожестей объектов. Формула выше несимметрична: если  $i$  и  $j$  поменять местами, то получится другое число. Поэтому мы хотим некоторым образом симметризовать схожести.

$$\rho(i | j) = \rho(j | i)$$

Модифицируем немного формулу:

$$\rho_{ij} = \frac{\rho(i|j) + \rho(j|i)}{2 * l}$$

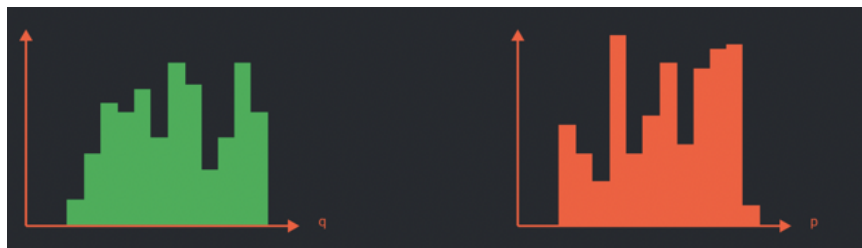
В новом пространстве мерить схожесть мы будем по другой формуле:

$$q(i | j) = \frac{(1 + |z_i - z_j|)^{-1}}{\sum_{k \neq m} (1 + |z_k - z_m|)^{-1}}$$

## Что же нужно оптимизировать?

Мы хотим, чтобы распределение точек было друга на друга поожим, чтобы схожесть двух объектов была примерно такой же в новом пространстве.

Представим, что есть набор объектов  $\{(q_{ij}, p_{ij})\}$ . В отдельности можно посмотреть, какие значения принимают эти переменные и нарисовать гистограммы. Например,



Наша задача: подобрать такие координаты, чтобы гистограммы были более похожи друг на друга.

Посмотреть, насколько одно распределение похоже на другое, можно при помощи расстояния Кульбака-Лейблера:

$$KL(p||q) = \sum_{i \neq j} p_{ij} \cdot \log\left(\frac{p_{ij}}{q_{ij}}\right) \rightarrow \min$$

Нашей задачей встает минимизация этого расстояния по координатам  $z$

## >T-SNE vs PCA

### PCA

- Дает некоторую формулу получения новых признаков через старые
- Ограничена линейной природой
- Фокус - на оптимизацию обучения
- Визуализация вряд ли хорошо сохранит расстояние между объектами

### T-SNE

- Просто переводит координаты в другое пространство
- Не может быть использовано для генерации новых признаков
- Обучение не оптимизирует
- Часто помогает хорошо визуализировать даже сложно устроенные данные