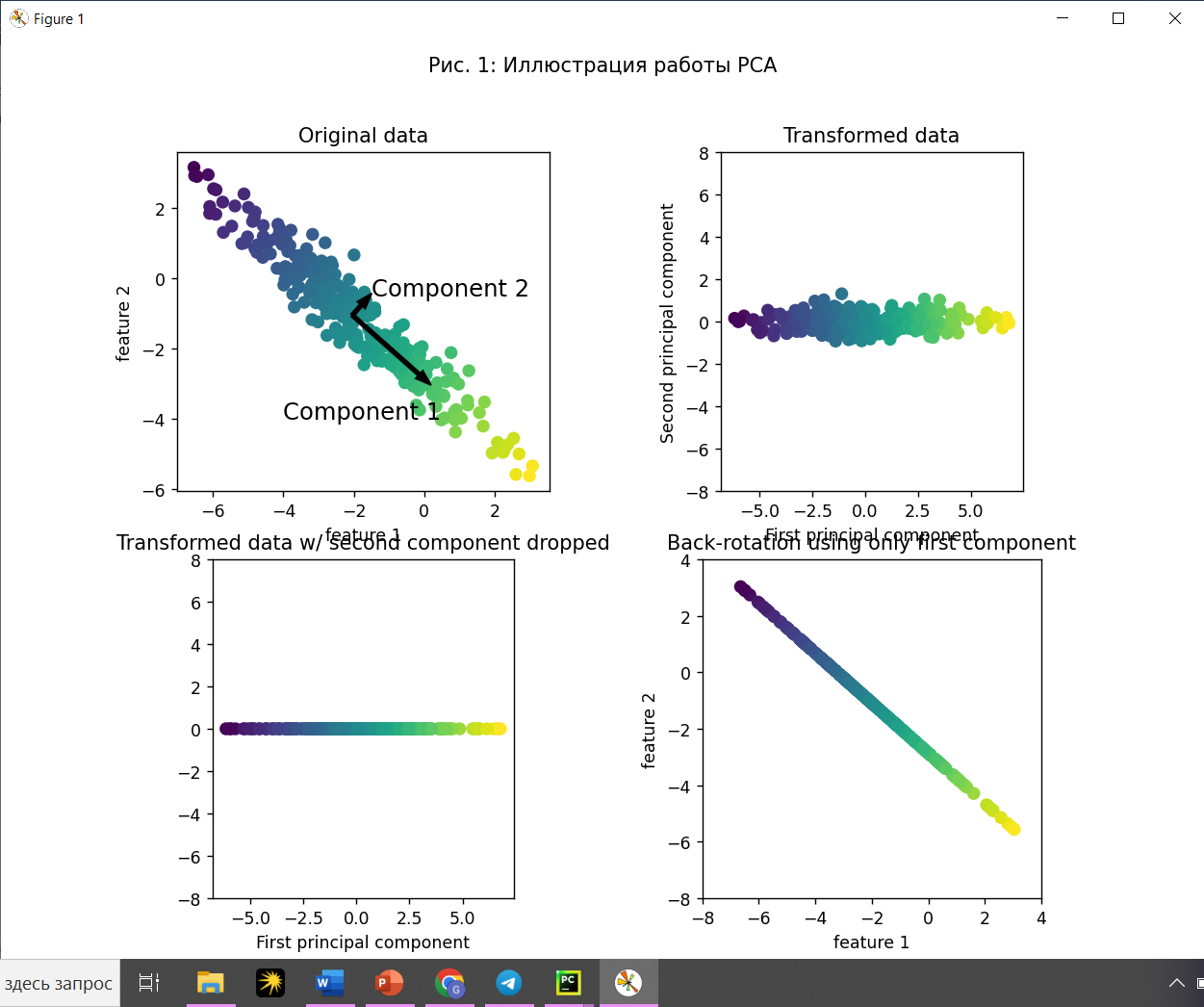
Анализ главных компонент (Principal Component Analysis, PCA) — это метод преобразования данных. Его основная цель — "повернуть" данные таким образом, чтобы новые, преобразованные признаки (называемые главными компонентами) не были скоррелированы друг с другом. Это означает, что они становятся независимыми в линейном смысле.

Часто такой "поворот" сопровождается выбором только нескольких наиболее важных новых признаков. Это позволяет уменьшить количество признаков (размерность данных), сохранив при этом максимальное количество информации, что упрощает дальнейший анализ и интерпретацию.



Первый график показывает исходные точки данных.

Алгоритм сначала находит направление в данных, вдоль которого наблюдается наибольшая дисперсия (разброс). Это направление, где данные "растянуты" сильнее всего. Оно содержит большую часть информации о вариативности данных. Это направление и называется "первой главной компонентой".

Затем алгоритм ищет следующее направление, которое также содержит максимальное количество оставшейся информации (дисперсии), но при этом оно должно быть ортогонально (перпендикулярно) первому найденному направлению.

Стрелки, изображающие компоненты, показывают направление.

Найденные таким образом направления называются **главными компонентами** (principal components), так как они отражают основные направления изменчивости (дисперсии) в данных.

Максимально возможное количество главных компонент равно количеству исходных признаков. Если у вас было 5 признаков, вы можете найти до 5 главных компонент.

2 график

Те же данные, но "повернутые" так, что:

* Первая главная компонента теперь совпадает с осью X (названной "Первая главная компонента").
* Вторая главная компонента совпадает с осью Y (названной "Вторая главная компонента").

3 график. Здесь показан результат уменьшения размерности. Мы "выбросили" вторую главную компоненту и оставили только первую. Данные теперь одномерные — все точки лежат на одной линии

4 график. Здесь мы пытаемся восстановить данные в исходном пространстве признаков, используя только информацию, сохраненную в первой главной компоненте.

Когда у нас много признаков (например, в наборе данных Breast Cancer их 30), построить обычные диаграммы рассеяния становится сложно. Матрица диаграмм рассеяния (pairplot), где показаны все возможные пары признаков, стала бы огромной (для 30 признаков, если брать пары, это 30\*29/2 = 435 графиков, а если учитывать что каждая пара может быть (x,y) и (y,x), то в mglearn это 30\*30, но многие будут дублироваться или быть гистограммами; в тексте указано 30\*14=420, что, видимо, соответствует реализации в mglearn). Анализировать такое количество графиков невозможно.

* Можно построить гистограммы для каждого признака, показывающие распределение его значений отдельно для каждого класса (например, для доброкачественных и злокачественных опухолей).
* **Синяя гистограмма:** Распределение значений признака для доброкачественных опухолей.
* **Зеленая гистограмма:** Распределение значений признака для злокачественных опухолей.

3 график

* Гистограммы показывают распределение каждого признака по отдельности, но не дают информации о взаимодействии между признаками или о том, как их комбинации связаны с классами.
* PCA может помочь учесть главные взаимодействия и получить более полную картину. Идея: найти первые две главные компоненты (которые являются комбинациями всех 30 исходных признаков) и визуализировать данные в этом новом 2D-пространстве.
* Каждая точка на графике представляет одну опухоль.
* Положение точки определяется значениями её первых двух главных компонент.
* Цвет точки указывает на класс опухоли: синий (malignant - злокачественная), зеленый (benign - доброкачественная).

На графике видно, что два класса (синие и зеленые точки) довольно хорошо разделены в этом новом 2D-пространстве. Это говорит о том, что первые две главные компоненты смогли уловить информацию, полезную для различения классов.

4 график

Главные компоненты — это линейные комбинации *всех* исходных признаков.

* + **Первая главная компонента (верхняя строка):**  
    Видно, что все коэффициенты для этой компоненты имеют примерно одинаковый знак (на карте они все одного "теплого" цвета, что соответствует положительным значениям по шкале справа; но знак сам по себе не так важен, как их относительное единообразие). Это означает, что первая главная компонента представляет собой некое "средневзвешенное" всех признаков или направление, вдоль которого все признаки коррелируют положительно друг с другом. Если значение одного признака велико, то и значения других признаков, скорее всего, тоже будут велики.
  + **Вторая главная компонента (нижняя строка):**  
    Здесь коэффициенты имеют разные знаки (некоторые "теплые", некоторые "холодные"). Это означает, что вторая главная компонента отражает некий **контраст** между группами признаков. Например, высокие значения одних признаков (с положительными коэффициентами) и низкие значения других (с отрицательными коэффициентами) будут давать большое значение этой компоненте.
  + **Сложность:** Обе главные компоненты включают в себя вклады от *всех 30 исходных признаков*. Это "смешивание" и делает интерпретацию осей сложной. Мы не можем сказать, что "Первая главная компонента = это просто mean radius", это всегда их комбинация.

5 график

* Еще одно применение PCA — это не просто уменьшение размерности, а поиск нового представления данных, которое лучше подходит для анализа.
* Изображения состоят из пикселей. Каждый пиксель может иметь интенсивность (для черно-белых) или значения для красного, зеленого и синего каналов (RGB). Объекты на изображениях — это совокупности тысяч пикселей.
* Мы можем применить PCA к набору изображений лиц.
  + Каждое изображение "разворачивается" в длинный вектор пикселей (например, если изображение 87x65 пикселей, то это вектор длиной 87\*65 = 5655). Этот вектор становится одной "точкой" данных.
  + Набор таких векторов (от разных изображений лиц) образует наш датасет.
  + PCA, примененный к этому датасету, найдет главные компоненты. Эти главные компоненты, если их "свернуть" обратно в формат изображения (87x65), будут выглядеть как некие "усредненные" или "характерные" лица. Их и называют **собственными лицами (eigenfaces)**.
* После нахождения главных компонент и поворота данных, выбеливание дополнительно масштабирует каждую компоненту так, чтобы она имела единичную дисперсию.
* В результате, если данные образовывали эллипс в пространстве главных компонент, после выбеливания они образуют "круг" (в 2D) или "сферу" (в 3D и выше). Все главные компоненты становятся одинаково "важными" с точки зрения их дисперсии.

В консоли увидим что есть задача

* Определить, принадлежит ли новое (незнакомое) фото одному из известных людей в базе данных.
* Сначала точность 20% после выбеливания 30%

6 график

Поскольку исходные данные — это изображения, то и главные компоненты, которые являются направлениями в этом пространстве данных, также можно визуализировать как изображения.

Компоненты:

1. component: Обычно это самое "усредненное" лицо или основной паттерн вариации.

Последующие компоненты: Фиксируют более тонкие вариации (освещение с разных сторон, контуры, и т.д.).

7 график

Можно восстановить (реконструировать) исходные данные, используя только некоторое количество главных компонент. Это делается с помощью метода pca.inverse\_transform().

8 график

* Визуализировать *все* лица из обучающего набора (X\_train\_pca) на диаграмме рассеяния, используя их значения для первых двух главных компонент. Каждая точка — это лицо, цвет/форма точки может соответствовать конкретному человеку (классу).
* Из рис. видно, когда мы используем лишь первые две главные компоненты, все данные представляют собой просто одно большое скопление данных без видимого разделения классов. Данный факт неудивителен, учитывая, что даже при использовании 10 компонент, как уже было показано ранее на рис. 9, PCA фиксирует самые общие характеристики лиц.