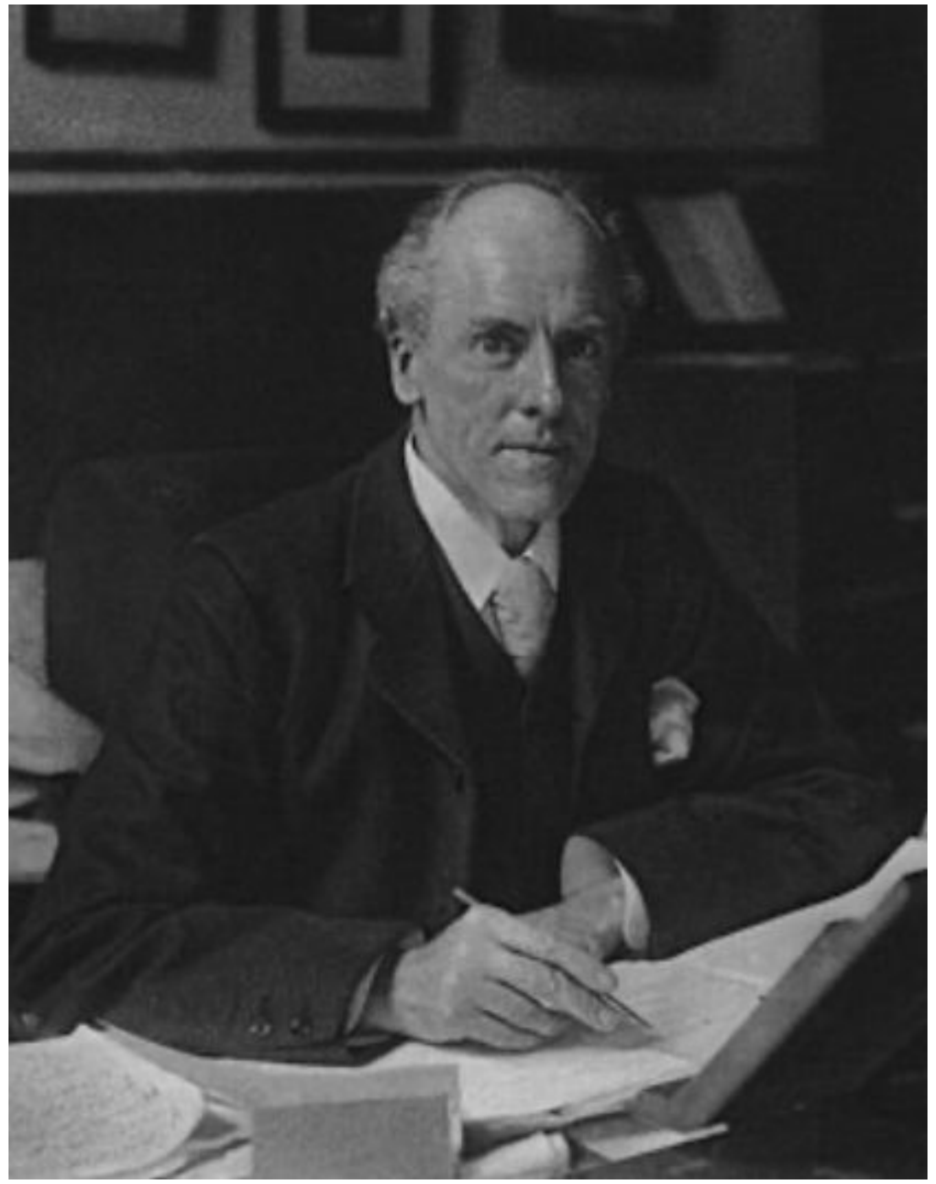
УНИВЕРСИТЕТ ИННОПОЛИС

Снижение размерности. tSNE

Воробьёва Мария

- maria.vorobyova.ser@gmail.com
- @SparrowMaria

Методы понижения размерности

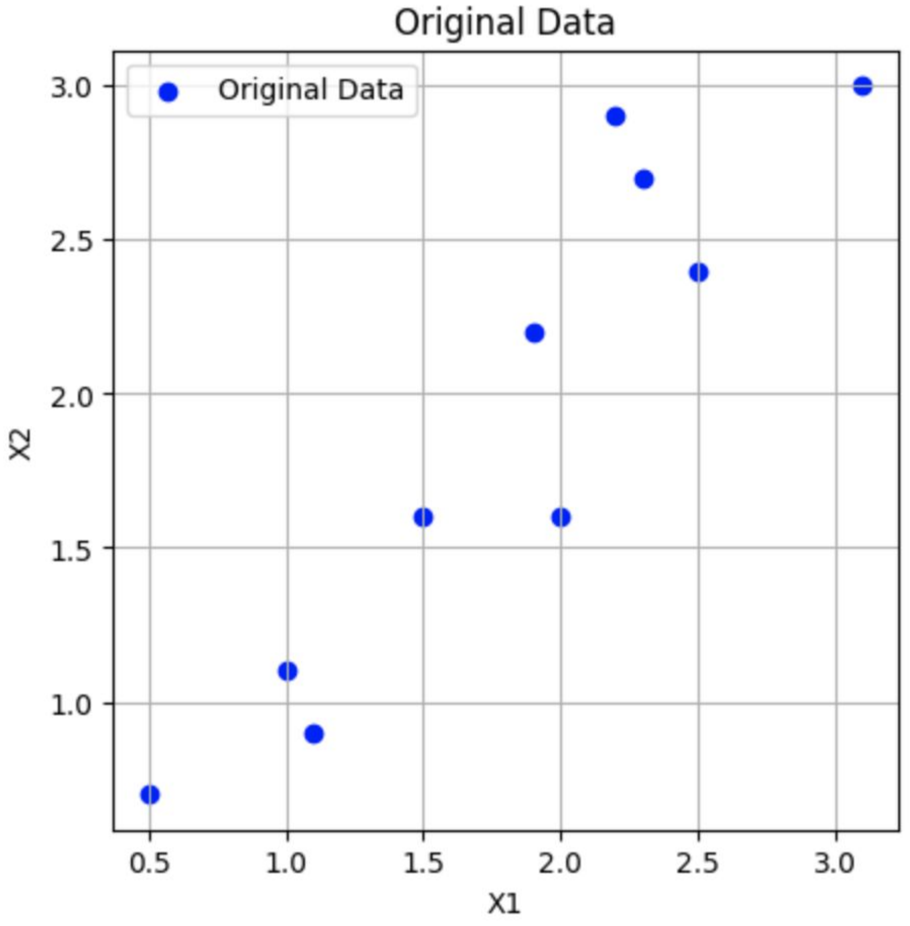


PCA - один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации.

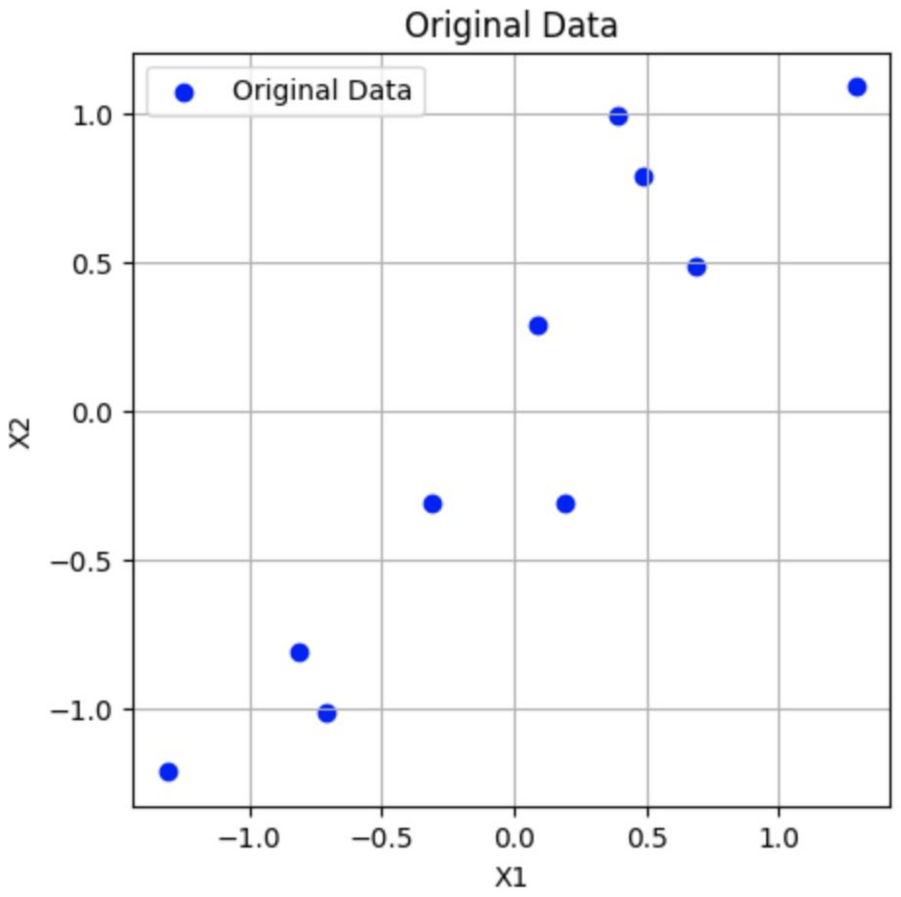
Изобретён Карлом Пирсоном в 1901 году. Применяется во многих областях, в том числе в эконометрике, биоинформатике, обработке изображений, для сжатия данных, в общественных науках

Методы понижения размерности. PCA

```
array([[2.5, 2.4],
       [0.5, 0.7],
       [2.2, 2.9],
       [1.9, 2.2],
       [3.1, 3. ],
       [2.3, 2.7],
       [2. , 1.6],
       [1. , 1.1],
       [1.5, 1.6],
       [1.1, 0.9]])
```




```
array([[ 0.69,  0.49],
       [-1.31, -1.21],
       [ 0.39,  0.99],
       [ 0.09,  0.29],
       [ 1.29,  1.09],
       [ 0.49,  0.79],
       [ 0.19, -0.31],
       [-0.81, -0.81],
       [-0.31, -0.31],
       [-0.71, -1.01]])
```



- **Центрирование данных: вычитание среднего значения каждой переменной**
- Выборка ковариационной матрицы: расчет ковариаций между всеми парами переменных.
- Нахождение собственных векторов и собственных значений ковариационной матрицы: собственные векторы определяют направление главных компонентов, а собственные значения — их значимость.

Методы понижения размерности. PCA

```python
# Ковариационная матрица
cov_matrix = np.cov(X_centered, rowvar=False)
cov_matrix
```

```
array([[0.61655556, 0.61544444],
       [0.61544444, 0.71655556]])
```

- Центрирование данных: вычитание среднего значения каждой переменной
- **Выборка ковариационной матрицы: расчет ковариаций между всеми парами переменных**
- Нахождение собственных векторов и собственных значений ковариационной матрицы: собственные векторы определяют направление главных компонентов, а собственные значения — их значимость.

Методы понижения размерности. PCA

```
eigenvalues
```

```
array([0.0490834 , 1.28402771])
```

```
eigenvectors
```

```
array([[-0.73517866,  0.6778734 ],
       [ 0.6778734 ,  0.73517866]])
```

- Центрирование данных: вычитание среднего значения каждой переменной
- Выборка ковариационной матрицы: расчет ковариаций между всеми парами переменных
- **Нахождение собственных векторов и собственных значений ковариационной матрицы: собственные векторы определяют направление главных компонентов, а собственные значения — их значимость.**

Методы понижения размерности. PCA

```
sorted_eigenvalues, sorted_eigenvectors
```

```
(array([1.28402771, 0.0490834 ]),
 array([[ 0.6778734 , -0.73517866],
        [ 0.73517866,  0.6778734 ]]))
```

```
# Проекция данных на главные компоненты
Z = np.dot(X_centered, sorted_eigenvectors)
Z
```

```
array([[ 0.82797019, -0.17511531],
       [-1.77758033,  0.14285723],
       [ 0.99219749,  0.38437499],
       [ 0.27421042,  0.13041721],
       [ 1.67580142, -0.20949846],
       [ 0.9129491 ,  0.17528244],
       [-0.09910944, -0.3498247 ],
       [-1.14457216,  0.04641726],
       [-0.43804614,  0.01776463],
       [-1.22382056, -0.16267529]])
```

Сортировка собственных значений и векторов:

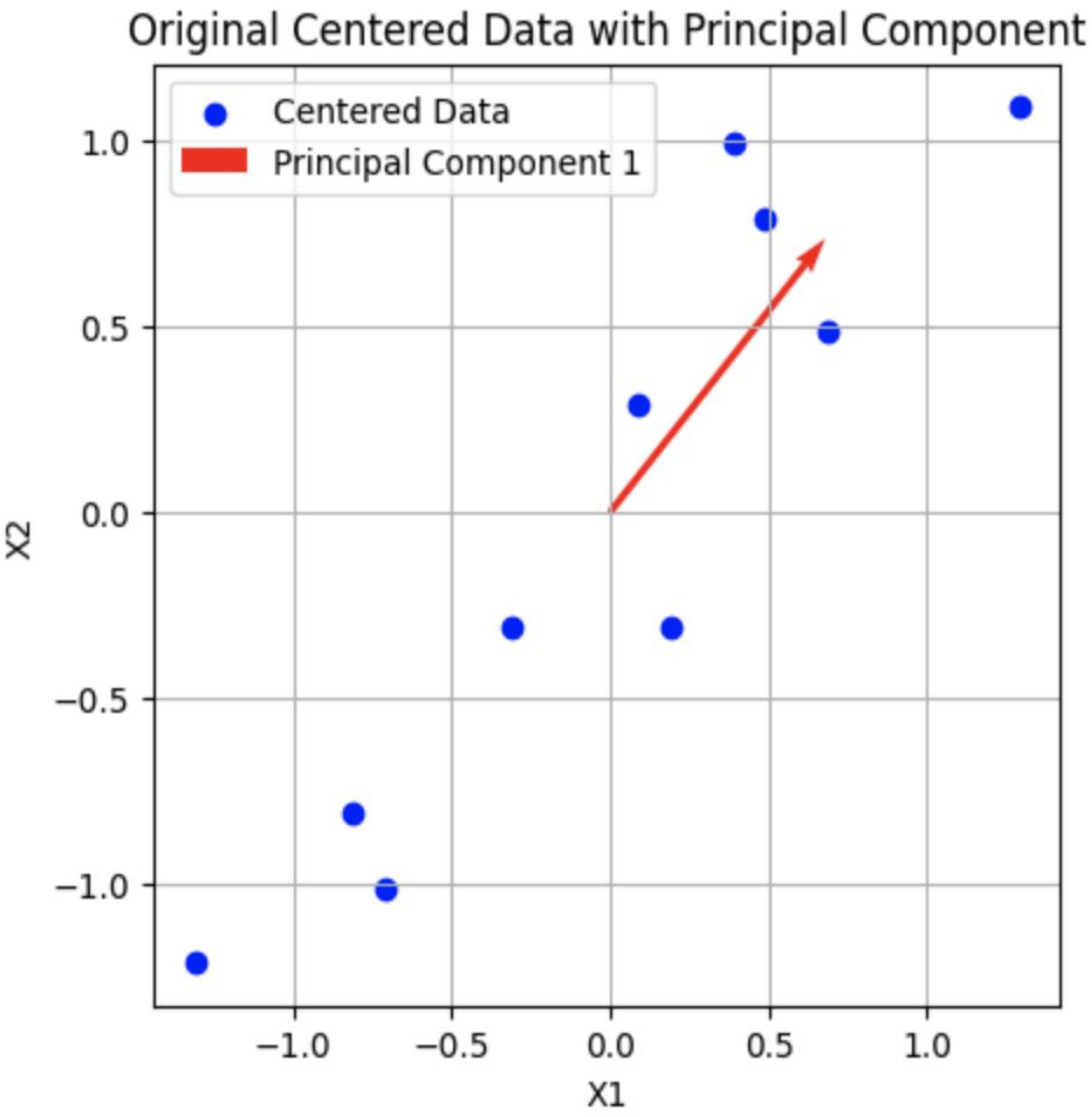
- Главные компоненты будут векторами, соответствующими наибольшим собственным значениям.

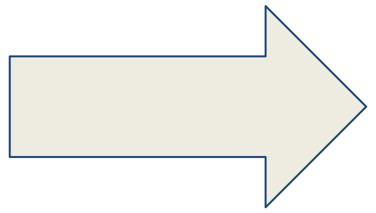
Проекция данных на главные компоненты:

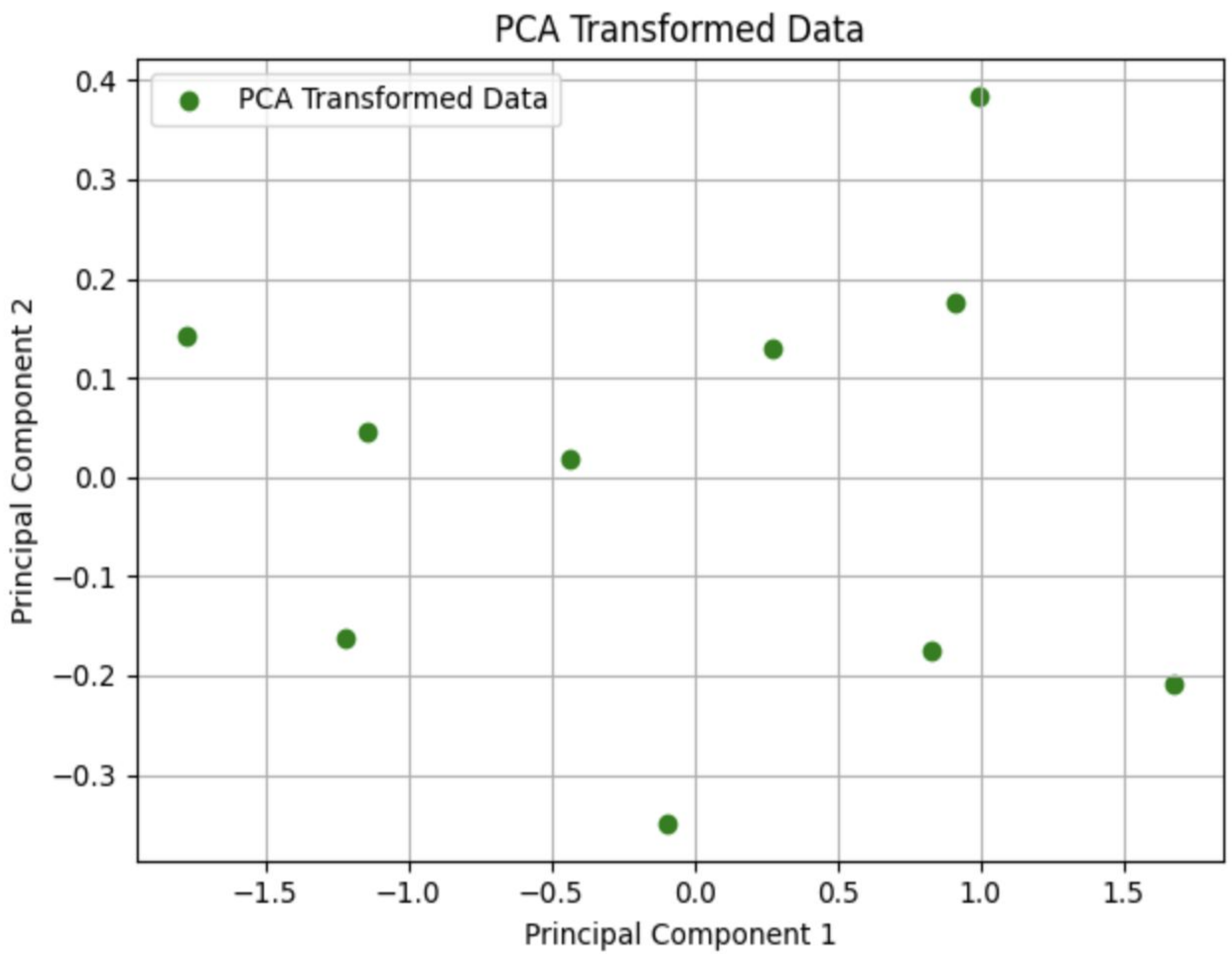
- Для получения новых переменных (главных компонент) спроецируйте центрированные данные X на собственные векторы: Z=XV

Z — новые координаты данных в пространстве главных

Методы понижения размерности







Методы понижения размерности. SVD (Singular Value Decomposition)

Для матрицы A размером $m \times n$, сингулярное разложение выглядит следующим образом:

$$A = U\Sigma V^T$$

где:

- A – исходная матрица.
- U – ортогональная матрица размером $m \times m$, содержащая левые сингулярные векторы.
- Σ – диагональная матрица размером $m \times n$, содержащая сингулярные значения.
- V^T – транспонированная ортогональная матрица V размером $n \times n$, содержащая правые сингулярные векторы.

Методы понижения размерности. SVD (Singular Value Decomposition)

Собственные значения λ_i ковариационной матрицы C равны квадратам сингулярных значений σ_i матрицы данных X, деленных на число наблюдений минус один:

$$\lambda_i = \frac{\sigma_i^2}{n-1}$$

Главные компоненты PCA соответствуют правым сингулярным вектором V из SVD матрицы данных X.

Методы понижения размерности.
T-SNE (t-distributed Stochastic Neighbor Embedding)

- T-SNE - метод снижения размерности
- T-SNE стремится расположить точки, которые были близки друг к другу в исходном высокоразмерном пространстве, также близко друг к другу и в низкоразмерном пространстве
- T-SNE использует нелинейные преобразования для отображения многомерных данных в пространство низкой размерности

Методы понижения размерности.
T-SNE (t-distributed Stochastic Neighbor Embedding)

- Перплексия — это гиперпараметр, который можно рассматривать как приближение к числу ближайших соседей для каждого точки. Он сильно влияет на результаты T-SNE. Более низкая перплексия подчеркивает локальную структуру, в то время как более высокая перплексия может выявить более глобальные структуры

- Чувствительность к масштабированию данных: Перед применением T-SNE данные часто необходимо масштабировать.

Методы понижения размерности
https://scikit-learn.org/stable/modules/manifold.html#manifold

Isomap: Сохраняет геодезические расстояния в низких измерениях.

Locally Linear Embedding (LLE): Сохраняет локальные расстояния.

Modified LLE: Решает проблемы регуляризации стандартного LLE.

Hessian LLE: Использует квадратичные формы на основе гессиана.

Spectral Embedding: Использует граф Лапласиан для спектрального разложения.

Local Tangent Space Alignment (LTSA): Выравнивает локальные касательные пространства.

Multi-dimensional Scaling (MDS): Сохраняет расстояния.

t-SNE: Подчеркивает локальную структуру для кластеризации.

Методы понижения размерности
UMAP (Uniform Manifold Approximation and Projection)

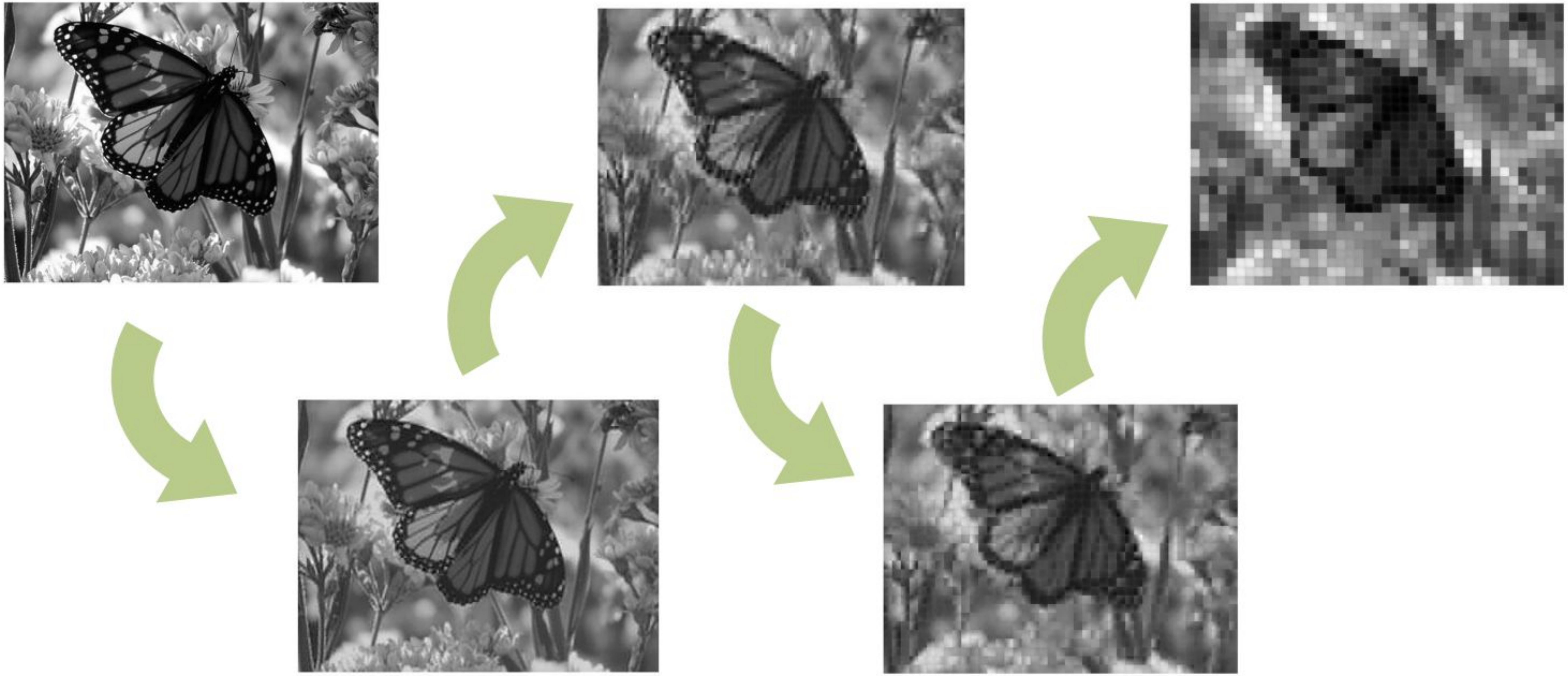
UMAP (Uniform Manifold Approximation and Projection) – это метод для уменьшения размерности, который работает за счет создания графа соседей в высокомерном пространстве и его проекции в низкоразмерное пространство. Основные особенности UMAP:

Эффективность: Быстрее и масштабируемее t-SNE.

Гибкость: Может использоваться как для визуализации, так и для предварительной обработки данных.

Сохранение структуры: Сохраняет глобальную и локальную структуры данных лучше, чем многие другие методы.

Методы понижения размерности



УНИВЕРСИТЕТ ИННОПОЛИС

ВОПРОСЫ И ОТВЕТЫ