主成分分析：原理、公式、应用与实战

2025 年 9 月 17 日

1 引言

主成分分析（Principal Component Analysis, PCA）通过寻找方差最大的正交方向，将高维数据投影到低维子空间，从而实现降维、可视化与噪声抑制。只需保留贡献最大的主成分即可兼顾数据结构与压缩率。

2 原理与公式

2.1 协方差矩阵与特征分解

对中心化后的数据矩阵 $\mathbf{X} \in \mathbb{R}^{n\times d}$，经验协方差为

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^\top\mathbf{X}. \tag{1}$$

PCA 通过求解特征值问题 $\mathbf{S}\mathbf{u}_k = \lambda_k\mathbf{u}_k$ 获得按 $\lambda_1 \geq \lambda_2 \geq \dots$ 排序的主方向。由前 k 个特征向量组成的矩阵 $\mathbf{U}_k$ 构成主子空间。

2.2 投影与重构

主成分得分（scores）为

$$\mathbf{Z} = \mathbf{X}\mathbf{U}_k, \tag{2}$$

秩 k 的重构为 $\hat{\mathbf{X}} = \mathbf{Z}\mathbf{U}_k^\top$。前 k 个主成分的方差贡献率为

$$\mathrm{ExplainedVariance}(k) = \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{j=1}^{d}\lambda_j}. \tag{3}$$

2.3 奇异值分解视角

若对 $\mathbf{X}$ 做奇异值分解 $\mathbf{X} = \mathbf{U}\ \mathbf{V}^\top$，则 $\mathbf{V}$ 的列向量对应协方差矩阵的特征向量，奇异值满足 $\sigma_i^2 = (n-1)\lambda_i$。当样本数远小于特征数时，SVD 是更高效的实现方式。

3 应用与技巧

- **可视化**：将高维数据映射到前两三个主成分，便于观察聚类或趋势。
- **预处理**：在聚类、回归前先降维以减轻多重共线性和噪声。
- **压缩存储**：仅保存主成分得分和载荷矩阵，用于推荐系统或图像压缩。
- **实用建议**：特征需居中，必要时做标准化；关注方差贡献率曲线，并在解释轴向时注意主成分符号可能翻转。

4 Python 实战

脚本 `gen_pca_figures.py` 构造相关特征数据，拟合 PCA，并输出主成分投影图与方差贡献率曲线。

Listing 1: 脚本 gen$_{p}ca_{f}igures.py$

```python
from sklearn.decomposition import PCA

pca = PCA(n_components=3, whiten=False, random_state=7)
pca.fit(points)
projected = pca.transform(points)

explained = np.cumsum(pca.explained_variance_ratio_)
```

5 实验结果

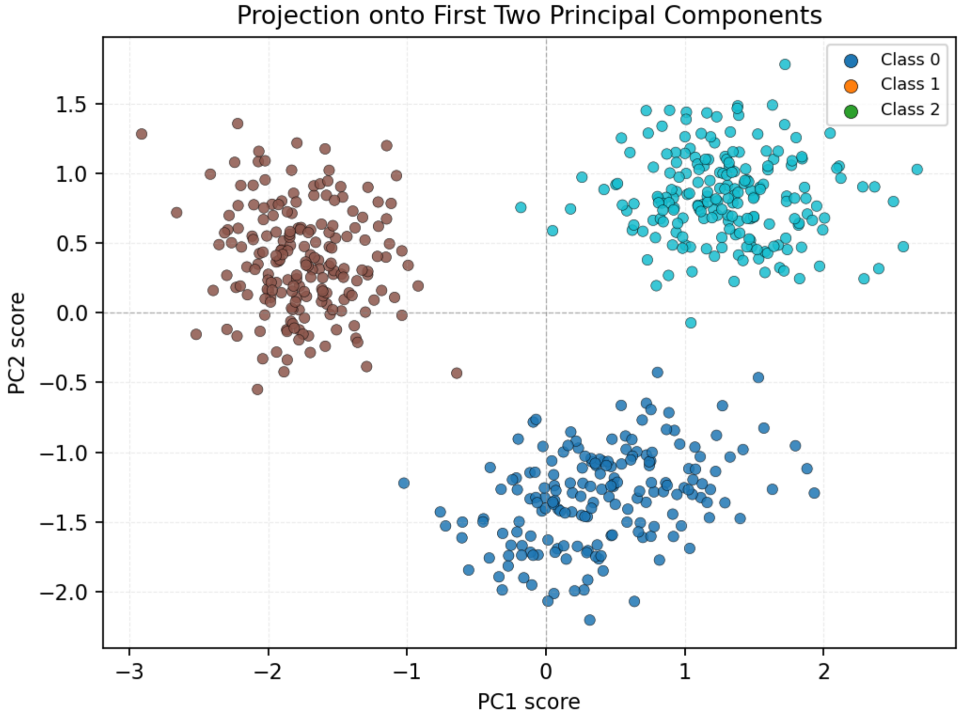

图 1: 前两个主成分的散点图（按类别着色）

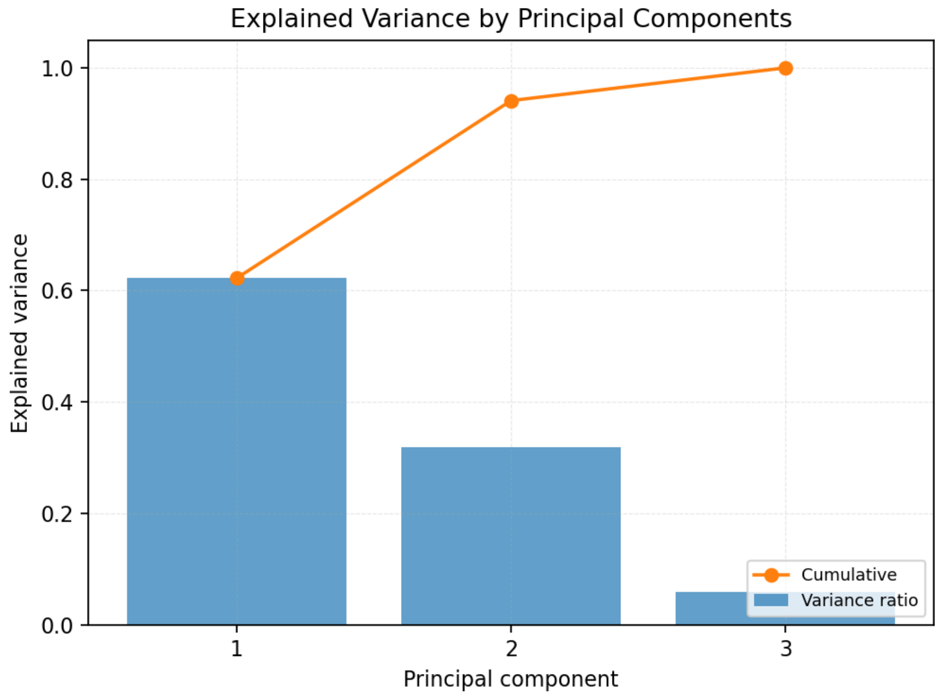

图 2: 各主成分方差贡献率与累计曲线

6 总结

PCA 通过特征分解或 SVD 提取最大方差方向，实现信息保留与降维之间的权衡。示例展示了主成分散点与方差曲线如何辅助选择保留成分数量。