18 降维

Dimensionality Reduction

从书有关降维算法模型的综述

人类的历史，本质上是思想的历史。

Human history is, in essence, a history of ideas.

—— 赫伯特·乔治·威尔斯 (Herbert George Wells) | 英国小说家和历史学家 | 1866 ~ 1946

- `sklearn.decomposition.PCA()` 主成分分析函数
- `sklearn.decomposition.TruncatedSVD()` 截断奇异值分解
- `sklearn.decomposition.FastICA()` 独立成分分析
- `sklearn.decomposition.IncrementalPCA()` 增量主成分分析

18.1 一张“降维”版图

图 1 总结几种常见降维的算法。相信大家对下面这几种算法已经熟悉：主成分分析 (Principal Component Analysis, PCA)、典型相关分析 (Canonical Correlation Analysis, CCA)。

本书前文介绍的线性判别分析 (Linear Discriminant Analysis, LDA) 也可以视作一种降维方法。

本章还要简单介绍核主成分分析 (Kernel Principal Component Analysis, KPCA)、独立成分分析 (Independent Component Analysis)、流形学习 (Manifold Learning) 这几种方法。

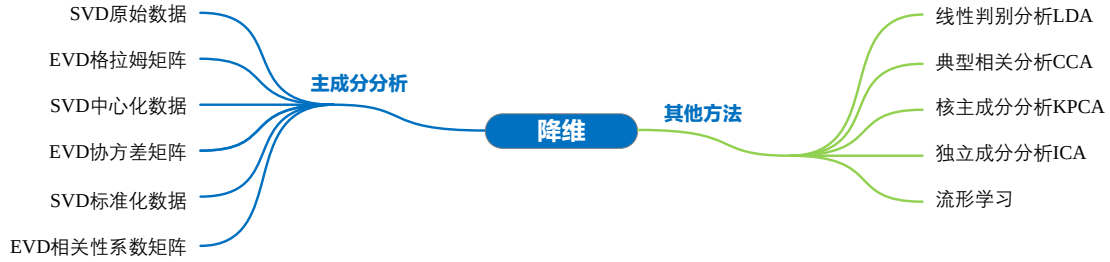


图 1. 降维方法分类

主成分分析

本系列丛书对主成分分析着墨颇多。和 OLS 线性回归类似，主成分分析也可以从几何 (图 2)、投影、数据、线性组合、特征值分解、SVD 分解、优化、概率统计等视角来理解。《数据有道》第 18、19 两章还介绍利用主成分分析进行回归的两种方法：正交回归、主元回归。

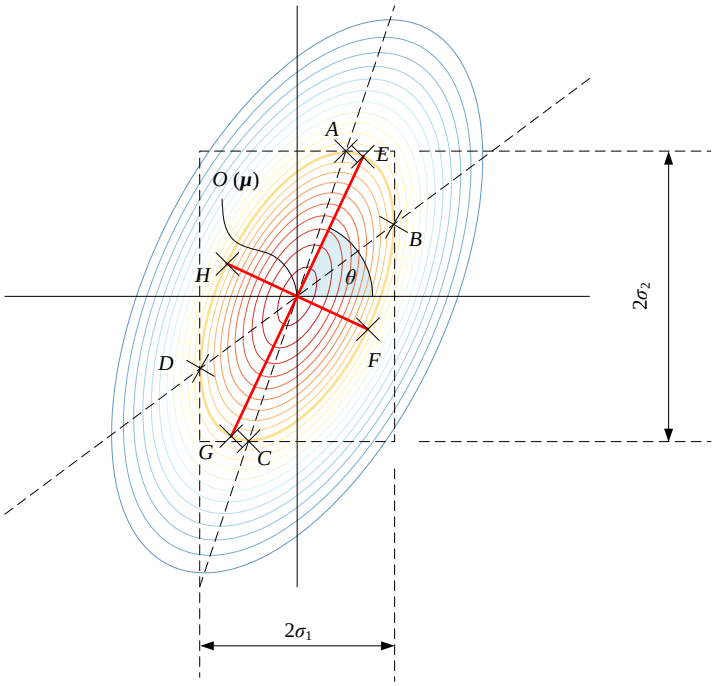


图 2. 主成分分析和椭圆的关系，图片来自《统计至简》第 25 章

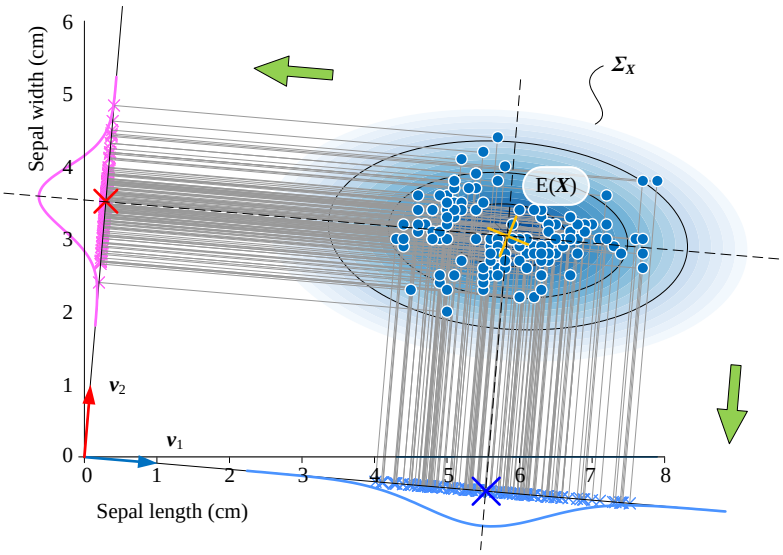


图 3. 投影视角看 PCA，图片来自《统计至简》第 14 章

此外，《数据有道》第 17 章还专门比较过主成分分析的六条技术路线，如表 1 所示。

表 1. 六条 PCA 技术路线，来自《矩阵分解》第 25 章

对象	方法	结果
原始数据矩阵 $\boldsymbol{X}$	奇异值分解	$\boldsymbol{X} = \boldsymbol{U_X S_X V_X}^{\mathrm{T}}$
格拉姆矩阵 $\boldsymbol{G} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 本章中用“修正”的格拉姆矩阵 $\boldsymbol{G} = \frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{n-1}$	特征值分解	$\boldsymbol{G} = \boldsymbol{V_X \Lambda_X V_X}^{\mathrm{T}}$
中心化数据矩阵 $\boldsymbol{X}_c = \boldsymbol{X} - \mathrm{E}(\boldsymbol{X})$	奇异值分解	$\boldsymbol{X}_c = \boldsymbol{U_c S_c V_c}^{\mathrm{T}}$
协方差矩阵 $\boldsymbol{\Sigma} = \frac{(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))^{\mathrm{T}}(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))}{n-1}$	特征值分解	$\boldsymbol{\Sigma} = \boldsymbol{V_c \Lambda_c V_c}^{\mathrm{T}}$
标准化数据 (z 分数) $\boldsymbol{Z}_{\boldsymbol{X}} = (\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))\boldsymbol{D}^{-1}$ $\boldsymbol{D} = \mathrm{diag}(\mathrm{diag}(\boldsymbol{\Sigma}))^{\frac{1}{2}}$	奇异值分解	$\boldsymbol{Z_X} = \boldsymbol{U_Z S_Z V_Z}^{\mathrm{T}}$
相关性系数矩阵 $\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{\Sigma}\boldsymbol{D}^{-1}$ $\boldsymbol{D} = \mathrm{diag}(\mathrm{diag}(\boldsymbol{\Sigma}))^{\frac{1}{2}}$	特征值分解	$\boldsymbol{P} = \boldsymbol{V_Z \Lambda_Z V_Z}^{\mathrm{T}}$

奇异值分解

表 1 中前两种 PCA 方法，又叫截断奇异值 (truncated SVD)。

sklearn.decomposition.TruncatedSVD() 这个函数支持这两种技术路线。

《矩阵力量》第 16 章介绍了四种奇异值分解，图 4 ~ 图 7 展示了它们之间的关系。此外，请大家回顾《矩阵力量》第 6 章有关分块矩阵乘法相关内容。

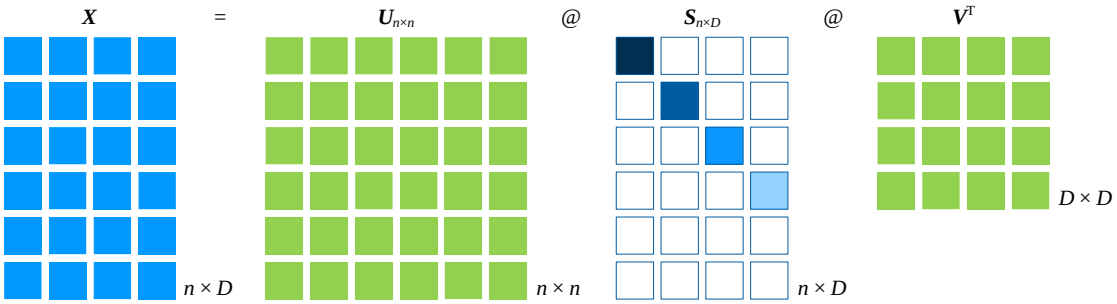


图 4. 完全型 SVD 分解

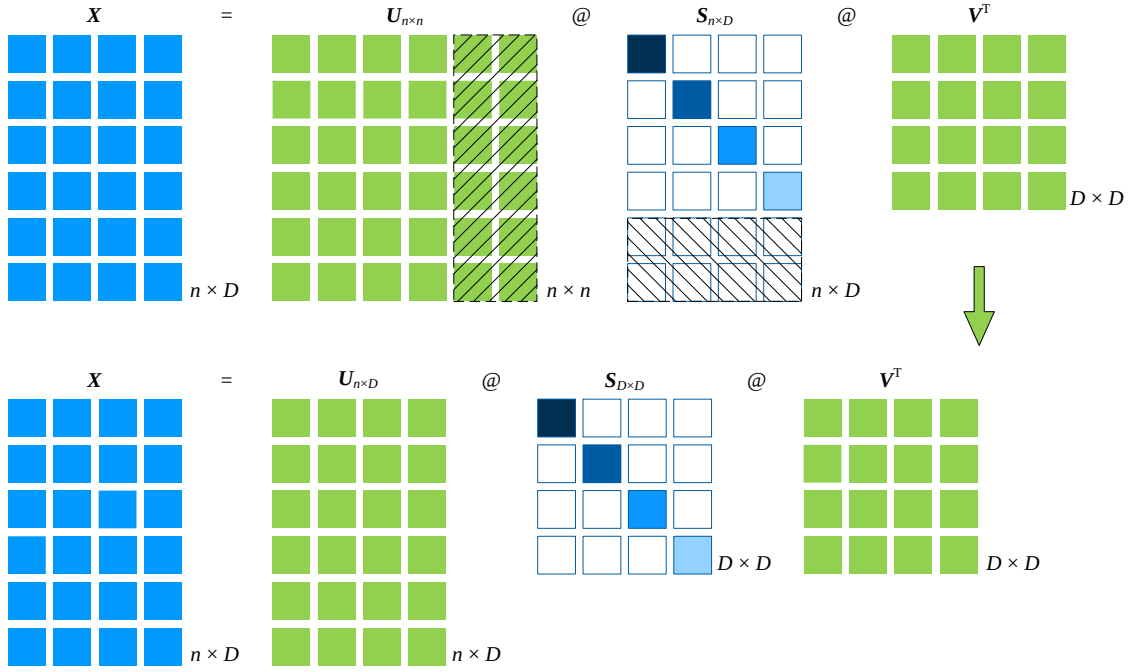


图 5. 从完全型到经济型

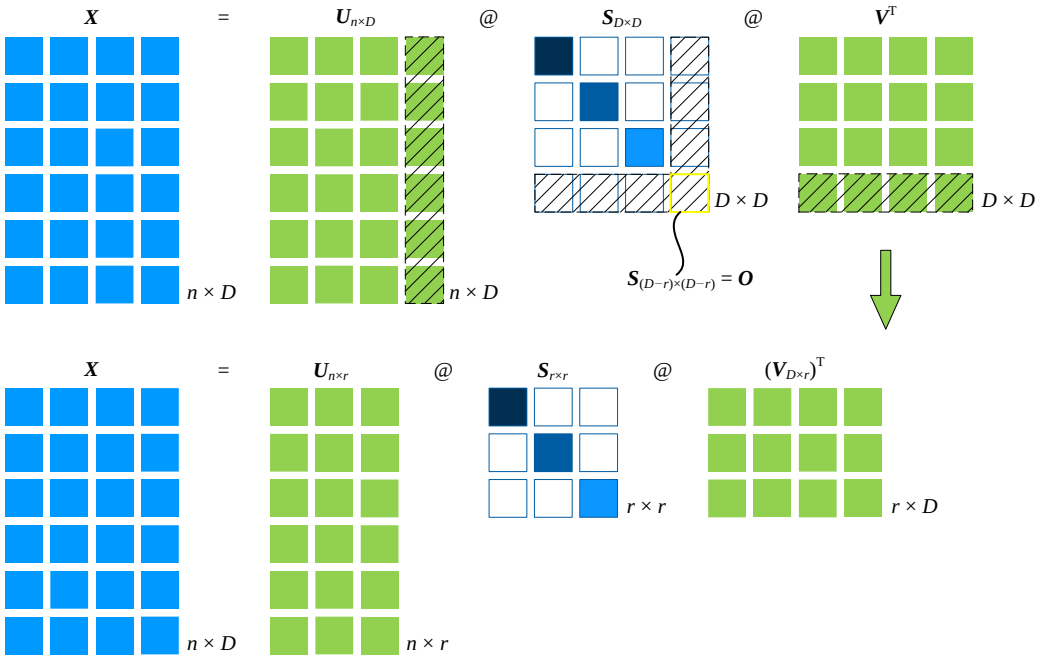


图 6. 从经济型到紧缩型

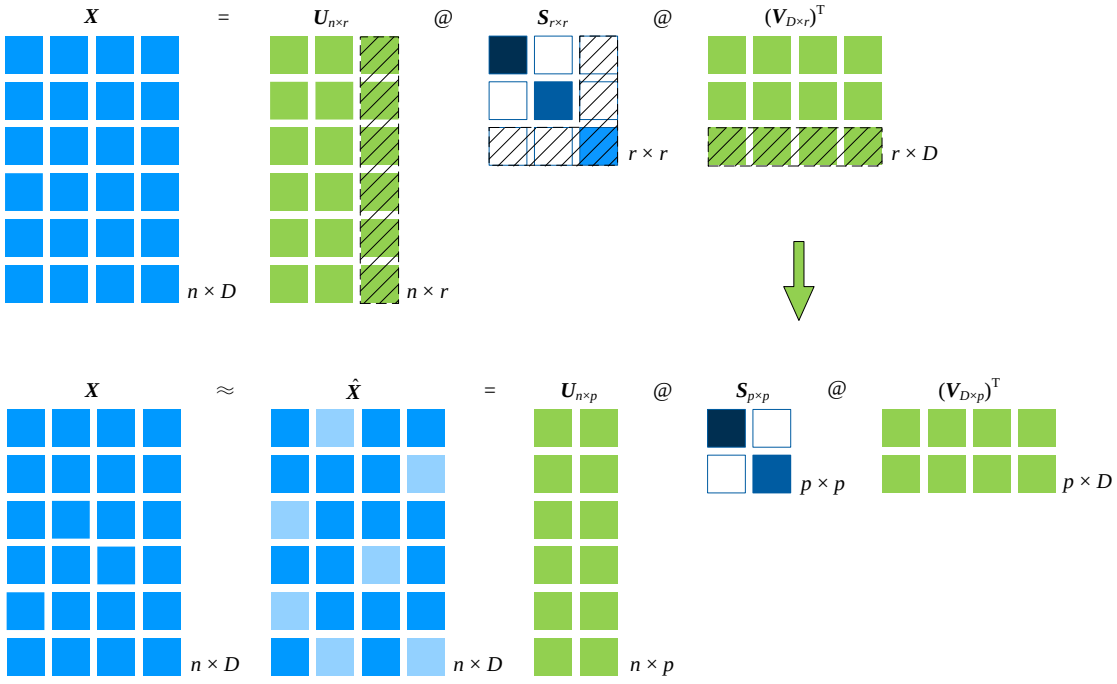


图 7. 从紧缩型到截断型

增量 PCA

当 PCA 需要处理的数据矩阵过大，以至于内存无法支持，可以使用增量主成分分析 (Incremental PCA, IPCA) 替代主成分分析。IPCA 分批处理输入数据，以便节省内存使用。Scikit-learn 中专门做增量 PCA 的函数为 sklearn.decomposition.IncrementalPCA()。

有关增量 PCA，大家可以参考下例：

https://scikit-learn.org/stable/auto_examples/decomposition/plot_incremental_pca.html

典型相关分析 CCA

典型相关分析也可以视作一种降维算法。CCA 也可以从几何、数据、优化、线性组合、统计

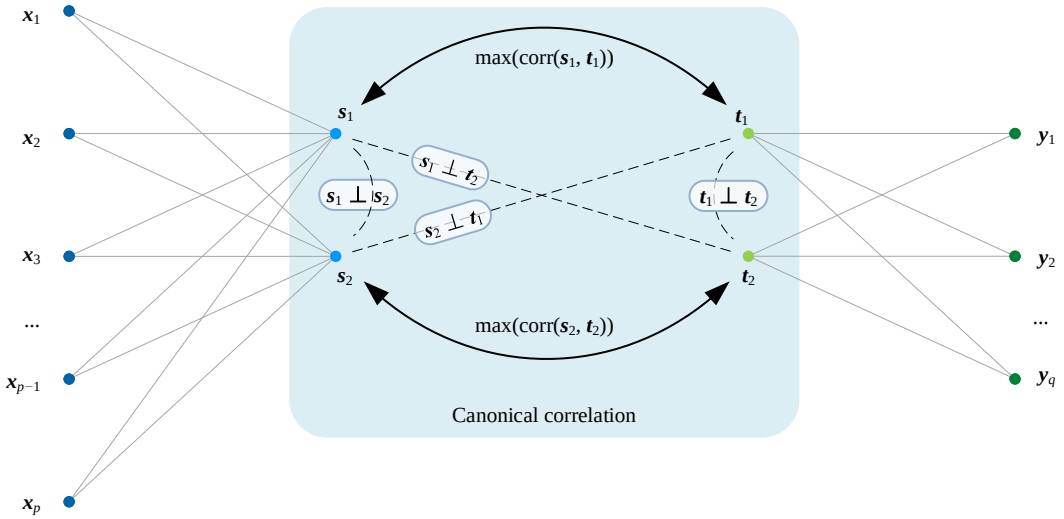


图 8. 线性组合角度看 CCA，图片来自《数据有道》第 20 章

下面这个例子比较偏最小二乘法 PLS、CCA，请大家参考：

https://scikit-learn.org/stable/auto_examples/cross_decomposition/plot_compare_cross_decomposition.html

18.2 核主成分分析

核主成分分析实现数据的非线性降维。图 9 (a) 所示数据线性不可分，我们先用非线性映射把数据映射到高维空间，使其线性可分。利用 KPCA 之后的结果如图 9 (b)。这一点和支持向量机中的核技巧颇为类似。

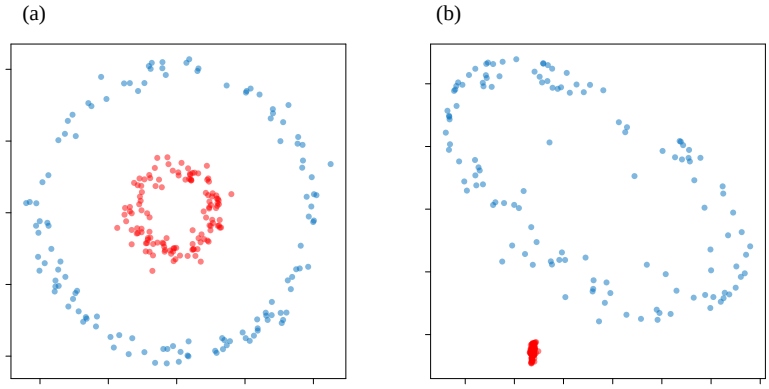


图 9. 核主成分分析

参考自如下示例，请大家自行学习：

https://scikit-learn.org/stable/auto_examples/decomposition/plot_kernel_pca.html

核主成分分析算法介绍，请参考：

https://people.eecs.berkeley.edu/~wainwrig/stat241b/scholkopf_kernel.pdf

18.3 独立成分分析

独立成分分析将一个多元信号分解成独立性最强的可加子成分。 因此，独立成分分析常用来分离叠加信号。图 10 比较 PCA 和 CCA 对同一组数据的分解结果。

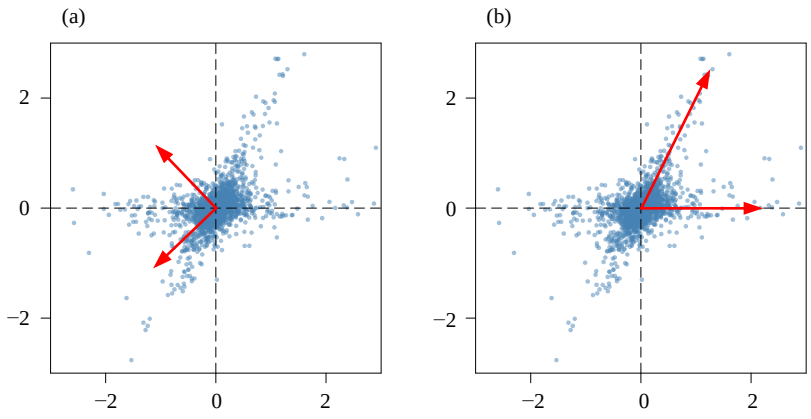


图 10. 比较 PCA 和 ICA

参考自如下示例，请大家自行学习：

https://scikit-learn.org/stable/auto_examples/decomposition/plot_ica_vs_pca.html

有关独立成分分析算法原理，请大家参考：

https://www.emerald.com/insight/content/doi/10.1016/j.aci.2018.08.006/full/html

18.4 流形学习

空间的数据可能是按照某种规则“卷曲”，度量点与点之间的“距离”要遵循这种卷曲的趋势。换一种思路，我们可以像展开“卷轴”一样，将数据展开并投影到一个平面上，得到的数据如图 12 所示。在图 12 所示平面上，A 和 B 两点的“欧氏距离”更好地描述了两点的距离度量，因为这个距离考虑了数据的“卷曲”。**流形学习** (manifold learning) 核心思想类似图 11 和图 12 所示展开“卷轴”的思想。流形学习也是非线性降维的一种方法。

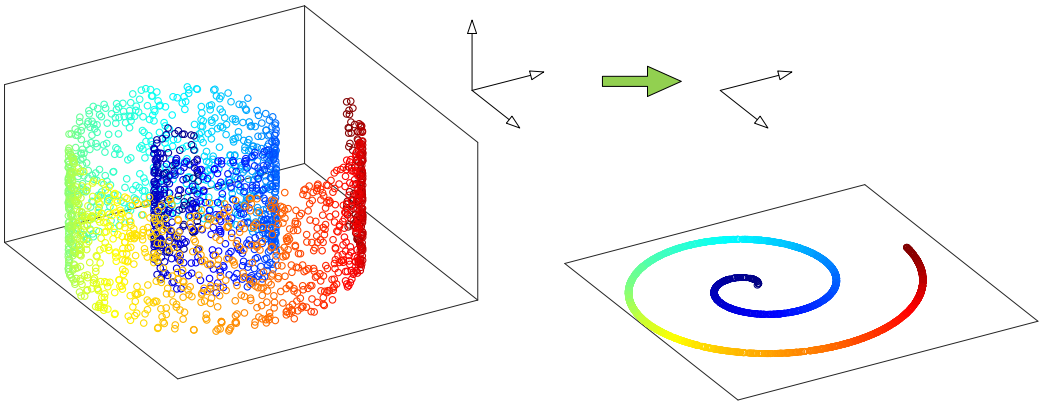


图 11. “卷曲”的数据

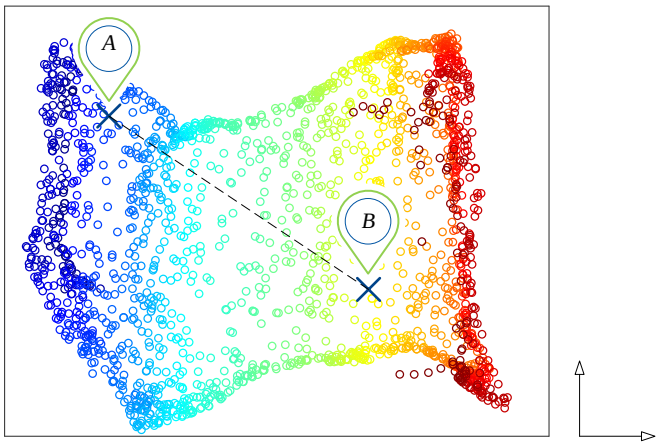


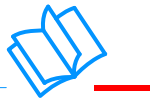
图 12. 展开“卷曲”的数据

想要深入了解 Scikit-learn 中的流形学习工具，请大家参考：

https://scikit-learn.org/stable/modules/manifold.html

如下这篇文献介绍了流形学习的数学基础，请大家参考：

https://arxiv.org/pdf/2011.01307.pdf



Scikit-learn 中更多有关降维工具，请大家参考：

https://scikit-learn.org/stable/modules/decomposition.html