16 Principal Component Analysis
主成分分析

处理多维数据，通过降维发现数据隐藏规律

忽视数学会损害所有知识，因为不了解数学的人无法了解世界上的其他科学或事物。更糟糕的是，那些无知的人无法感知自己的无知，因此不寻求补救。

Neglect of mathematics work injury to all knowledge, since he who is ignorant of it cannot know the other sciences or things of this world. And what is worst, those who are thus ignorant are unable to perceive their own ignorance, and so do not seek a remedy.

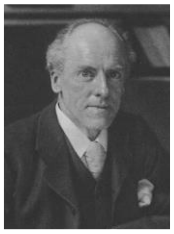
—— 罗吉尔·培根 (Roger Bacon) | 英国哲学家 | 1214 ~ 1294

- `numpy.cov()` 计算协方差矩阵
- `numpy.corrcoef()` 计算相关性系数矩阵
- `numpy.linalg.eig()` 特征值分解
- `numpy.linalg.svd()` 奇异值分解
- `numpy.mean()` 计算均值
- `numpy.var()` 计算方差
- `numpy.std()` 计算均方差
- `numpy.random.multivariate_normal()` 产生多元正态分布随机数
- `pca.components_` 获得 PCA 主成分向量，每一行代表一个向量
- `pca.inverse_transform()` 将数据 Z 还原成 X
- `pca.transform(X)` 将原始数据转化为数据 Z
- `seaborn.heatmap()` 绘制热图
- `seaborn.jointplot()` 绘制联合分布和边际分布
- `seaborn.kdeplot()` 绘制 KDE 核概率密度估计曲线
- `seaborn.pairplot()` 绘制成对分析图
- `sklearn.decomposition.PCA()` 主成分分析函数
- `seaborn.lineplot()` 绘制线图
- `numpy.zeros_like()` 产生形如输入矩阵的全 0 矩阵
- `yellowbrick.features.PCA()` 绘制 PCA 双标图

16.1 原始数据

主成分分析

主成分分析 (principal component analysis, PCA) 最初由**卡尔·皮尔逊** (Karl Pearson) 在 1901 提出。主成分分析是数据降维的重要方法之一。通过线性变换，主成分分析将原始多维数据投影到一个新的正交坐标系，将原始数据中的最大方差成分提取出来。



卡尔·皮尔逊 (Karl Pearson)
英国数学家 | 1857 ~ 1936
常被誉为现代统计科学的创立者；丛书关键词：● 相关性系数 ● 线性回归 ● 主成分分析

举个例子，主成分分析实际上寻找数据在主元空间内投影。图 1 所示杯子，它是一个 3D 物体，在一张图展示杯子，而且尽可能多地展示杯子细节，就需要从空间多个角度观察杯子并找到合适角度。这个过程实际上是将三维数据投影到二维平面过程。这也是一个降维过程，即从三维变成二维。图 2 展示杯子六个平面上投影结果。

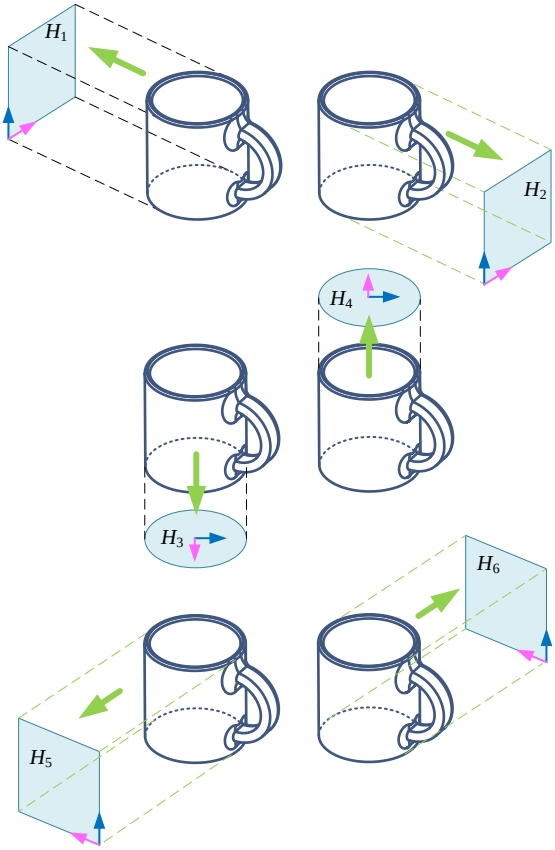


图 1. 咖啡杯六个投影方向

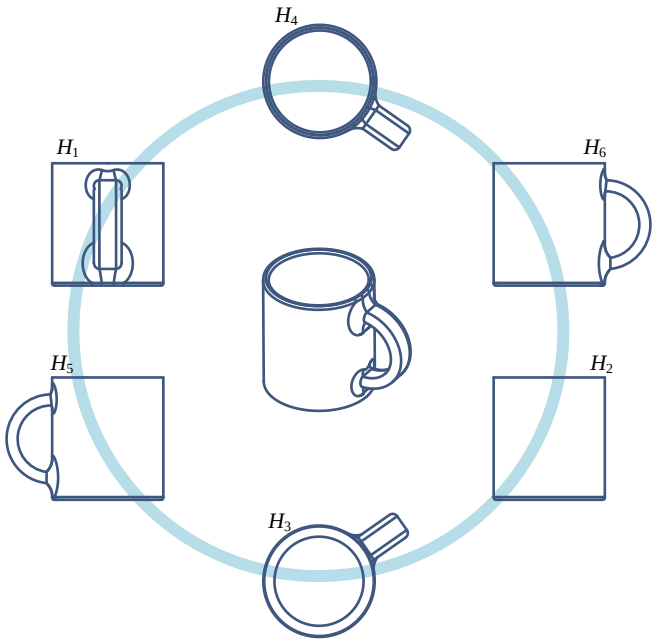


图 2. 咖啡杯在六个方向投影图像

以鸢尾花数据为例

本章以鸢尾花数据为例介绍如何利用主成分分析处理数据。图 3 所示为鸢尾花原始数据矩阵 $\boldsymbol{X}$ 构成的热图。数据矩阵 $\boldsymbol{X}$ 有 150 个数据点，即 150 行；矩阵 $\boldsymbol{X}$ 有 4 个特征，即 4 列。

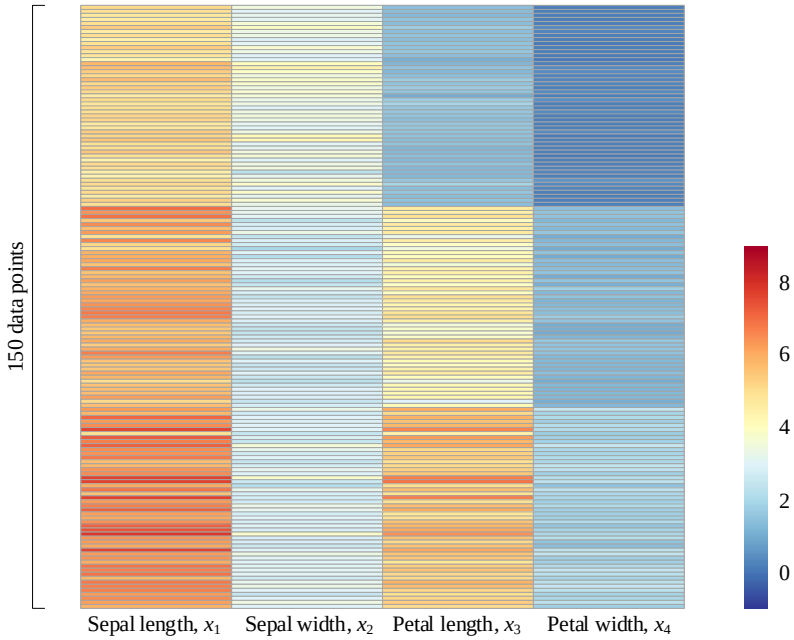


图 3. 鸢尾花数据，原始数据矩阵 $\boldsymbol{X}$

对原始数据进行统计分析。首先以行向量表达数据矩阵 $\boldsymbol{X}$ 质心：

$$\boldsymbol{\mu}_{\boldsymbol{X}} = \begin{bmatrix} \underset{\text{Sepal length, } x_1}{5.843} & \underset{\text{Sepal width, } x_2}{3.057} & \underset{\text{Petal length, } x_3}{3.758} & \underset{\text{Petal width, } x_4}{1.199} \end{bmatrix} \tag{1}$$

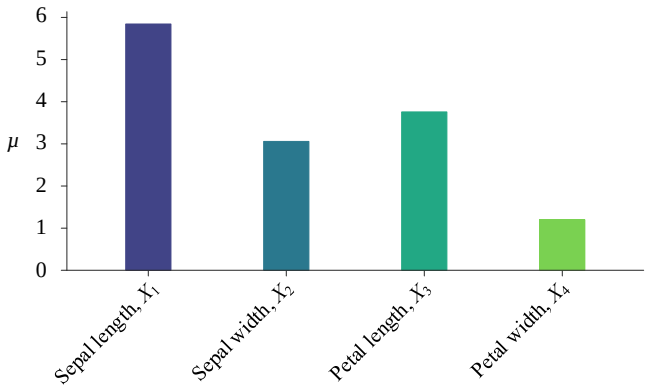


图 4. 鸢尾花数据四个特征上均值

然后，计算 $\boldsymbol{X}$ 每一列均方差，以行向量表达：

$$\boldsymbol{\sigma}_{\boldsymbol{X}} = \begin{bmatrix} \underset{\text{Sepal length, } x_1}{0.825} & \underset{\text{Sepal width, } x_2}{0.434} & \underset{\text{Petal length, } x_3}{1.759} & \underset{\text{Petal width, } x_4}{0.759} \end{bmatrix} \tag{2}$$

$\boldsymbol{X}$ 第三个特征，也就是花瓣长度 x_3 对应的均方差最大。图 5 所示为 KDE 估计得到的鸢尾花四个特征分布图。

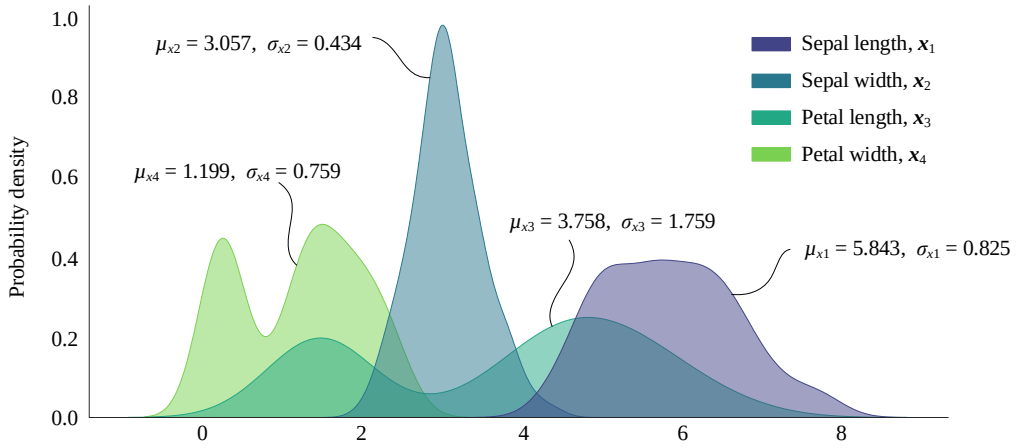


图 5. 鸢尾花数据四个特征上分布，KDE 估计

利用 seaborn.pairplot() 函数可以绘制如图 6 所示成对特征分析图；成对特征分析图方便展示每一对数据特征之间的关系，而对角线图像则展示每一个特征单独的统计规律。

由于鸢尾花数据存在三个分类，所以可以利用 seaborn.pairplot() 函数展示具有分类特征的成对分析图，具体如图 7 所示。图 7 这幅图让我们看到了每一类别数据特征之间和自身的分布规律。

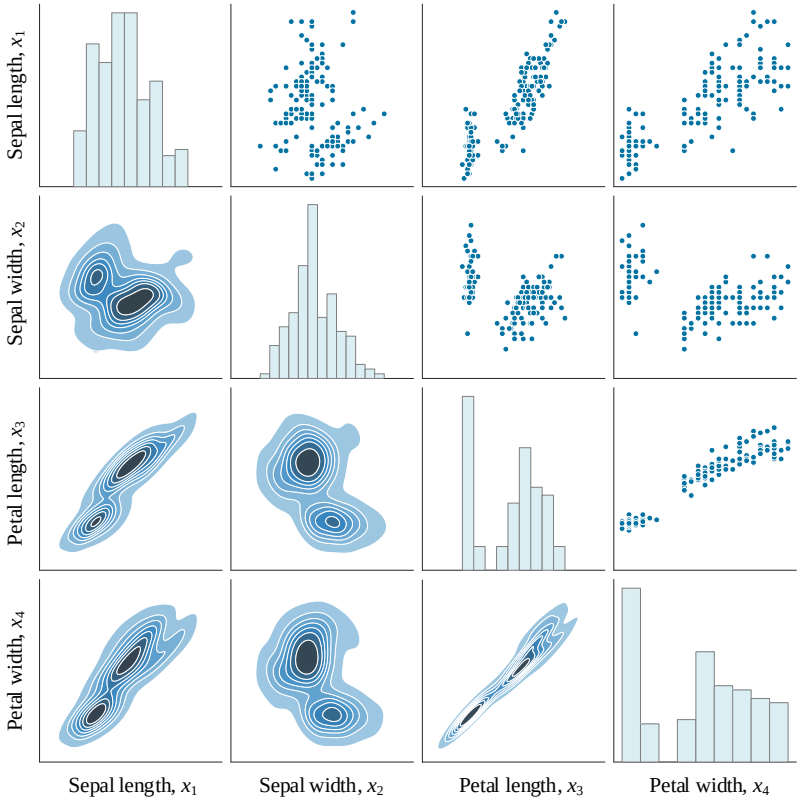


图 6. 鸢尾花数据成对特征分析图，不分类

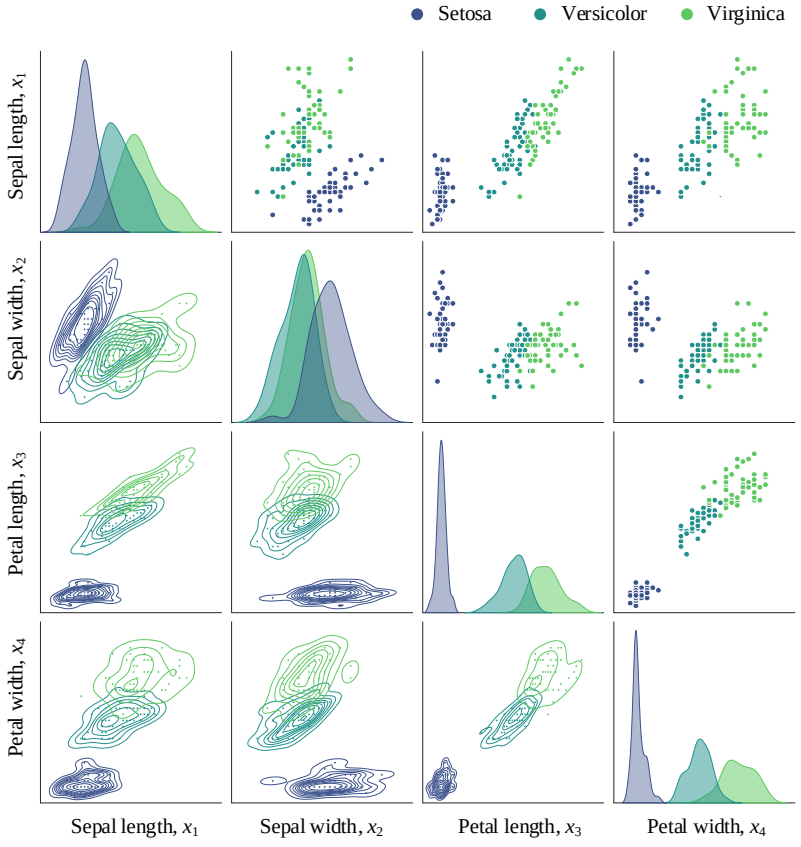


图 7. 鸢尾花数据成对特征分析图，分类

计算数据矩阵 $\boldsymbol{X}$ 协方差矩阵 $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.686 & -0.042 & 1.274 & 0.516 \\ -0.042 & 0.190 & -0.330 & -0.122 \\ 1.274 & -0.330 & 3.116 & 1.296 \\ 0.516 & \underbrace{-0.122}_{\text{Sepal width, } x_2} & 1.296 & 0.581 \\ \text{Sepal length, } x_1 & & \text{Petal length, } x_3 & \text{Petal width, } x_4 \end{bmatrix} \begin{matrix} \leftarrow \text{Sepal length, } x_1 \\ \leftarrow \text{Sepal width, } x_2 \\ \leftarrow \text{Petal length, } x_3 \\ \leftarrow \text{Petal width, } x_4 \end{matrix} \tag{3}$$

接下来，协方差矩阵 $\boldsymbol{\Sigma}$ 将用于特征值分解。

计算数据矩阵 $\boldsymbol{X}$ 相关性系数矩阵 $\boldsymbol{P}$:

$$\boldsymbol{P} = \begin{bmatrix} 1.000 & -0.118 & 0.872 & 0.818 \\ -0.118 & 1.000 & -0.428 & -0.366 \\ 0.872 & -0.428 & 1.000 & 0.963 \\ 0.818 & \underbrace{-0.366}_{\text{Sepal width, } x_2} & 0.963 & 1.000 \\ \text{Sepal length, } x_1 & & \text{Petal length, } x_3 & \text{Petal width, } x_4 \end{bmatrix} \begin{matrix} \leftarrow \text{Sepal length, } x_1 \\ \leftarrow \text{Sepal width, } x_2 \\ \leftarrow \text{Petal length, } x_3 \\ \leftarrow \text{Petal width, } x_4 \end{matrix} \tag{4}$$

观察相关性系数矩阵 $\boldsymbol{P}$，可以发现花萼长度 x_1 和花萼宽度 x_2 线性负相关，花瓣长度 x_3 和花萼宽度 x_2 线性负相关，花瓣宽度 x_4 和花萼宽度 x_2 线性负相关。

16.2 特征值分解

对 $\boldsymbol{\Sigma}$ 特征值分解得到：

$$\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1} \tag{5}$$

其中，$\boldsymbol{V}$ 是正交矩阵，满足 $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{I}$。实际上，上式为谱分解，即 $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}$。

特征值矩阵 $\boldsymbol{\Lambda}$ 为：

$$\boldsymbol{\Lambda} = \begin{bmatrix} 4.228 & & & \\ & 0.242 & & \\ & & 0.078 & \\ & & & 0.023 \end{bmatrix} \tag{6}$$

特征向量构成的矩阵 $\boldsymbol{V}$ 为：

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \boldsymbol{v}_3 & \boldsymbol{v}_4 \end{bmatrix} = \begin{bmatrix} v_{1,1} & v_{1,2} & v_{1,3} & v_{1,4} \\ v_{2,1} & v_{2,2} & v_{2,3} & v_{2,4} \\ v_{3,1} & v_{3,2} & v_{3,3} & v_{3,4} \\ v_{4,1} & v_{4,2} & v_{4,3} & v_{4,4} \end{bmatrix} \begin{matrix} \leftarrow \text{Sepal length, } x_1 \\ \leftarrow \text{Sepal width, } x_2 \\ \leftarrow \text{Petal length, } x_3 \\ \leftarrow \text{Petal width, } x_4 \end{matrix} = \begin{bmatrix} 0.361 & 0.656 & -0.582 & -0.315 \\ -0.084 & 0.730 & 0.597 & 0.319 \\ 0.856 & -0.173 & 0.076 & 0.479 \\ \underbrace{0.358}_{\text{PC1, } \boldsymbol{v}_1} & \underbrace{-0.075}_{\text{PC2, } \boldsymbol{v}_2} & \underbrace{0.545}_{\text{PC3, } \boldsymbol{v}_3} & \underbrace{-0.753}_{\text{PC4, } \boldsymbol{v}_4} \end{bmatrix} \tag{7}$$

矩阵 $\boldsymbol{V}$ 每一列代表一个主成分，该主成分中每一个元素相当于原始数据特征的系数。图 8 所示为不同主成分的系数线图。

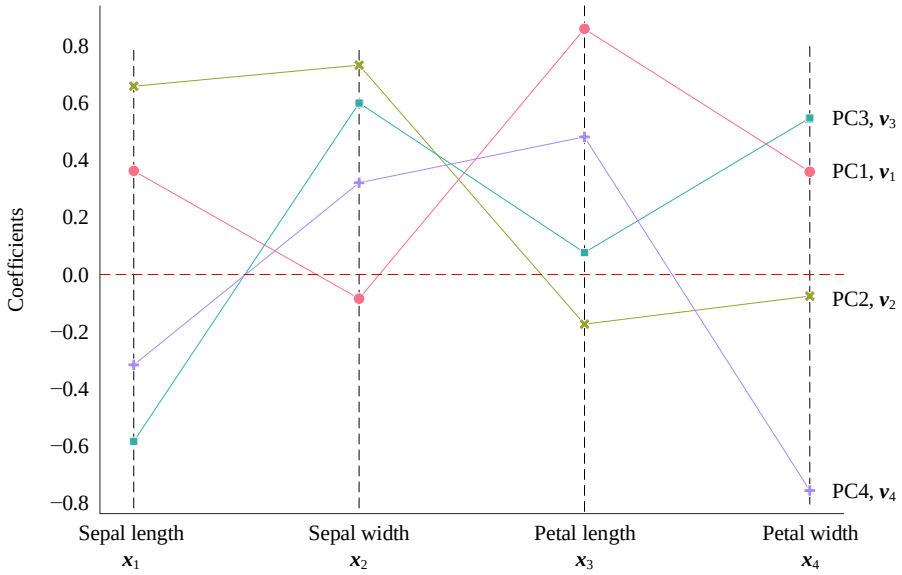


图 8. $\boldsymbol{V}$ 系数线图

如图 9 所示，$\boldsymbol{V}$ 和自己转置 $\boldsymbol{V}^T$ 乘积为单位阵 $\boldsymbol{I}$，即：

$$\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I} \tag{8}$$

展开上式得到：

$$\begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \boldsymbol{v}_3 & \boldsymbol{v}_4 \end{bmatrix}^T \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \boldsymbol{v}_3 & \boldsymbol{v}_4 \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1^T \\ \boldsymbol{v}_2^T \\ \boldsymbol{v}_3^T \\ \boldsymbol{v}_4^T \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \boldsymbol{v}_3 & \boldsymbol{v}_4 \end{bmatrix}$$
$$= \begin{bmatrix} \boldsymbol{v}_1^T\boldsymbol{v}_1 & \boldsymbol{v}_1^T\boldsymbol{v}_2 & \boldsymbol{v}_1^T\boldsymbol{v}_3 & \boldsymbol{v}_1^T\boldsymbol{v}_4 \\ \boldsymbol{v}_2^T\boldsymbol{v}_1 & \boldsymbol{v}_2^T\boldsymbol{v}_2 & \boldsymbol{v}_2^T\boldsymbol{v}_3 & \boldsymbol{v}_2^T\boldsymbol{v}_4 \\ \boldsymbol{v}_3^T\boldsymbol{v}_1 & \boldsymbol{v}_3^T\boldsymbol{v}_2 & \boldsymbol{v}_3^T\boldsymbol{v}_3 & \boldsymbol{v}_3^T\boldsymbol{v}_4 \\ \boldsymbol{v}_4^T\boldsymbol{v}_1 & \boldsymbol{v}_4^T\boldsymbol{v}_2 & \boldsymbol{v}_4^T\boldsymbol{v}_3 & \boldsymbol{v}_4^T\boldsymbol{v}_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \boldsymbol{I} \tag{9}$$

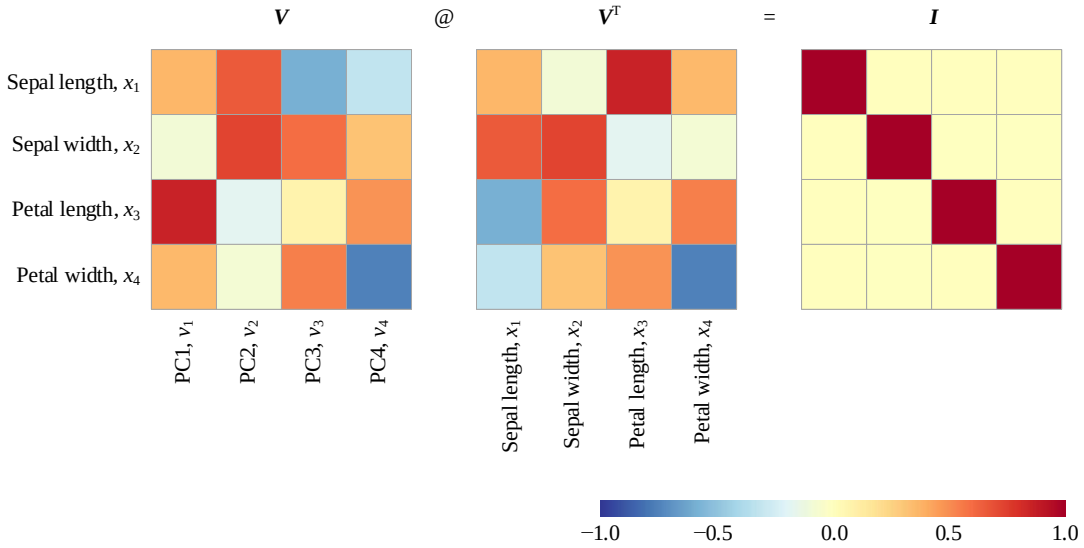


图 9. 特征矩阵 $\boldsymbol{V}$ 和自身转置的乘积为单位矩阵 $\boldsymbol{I}$

如果对鸢尾花数据先进行标准化处理，即使用每一列变成 z 分数；再计算得到的矩阵 $\boldsymbol{V}$ 则为：

$$\boldsymbol{V}=\begin{bmatrix} 0.521 & 0.377 & 0.720 & -0.261 \\ -0.269 & 0.923 & -0.244 & 0.124 \\ 0.580 & 0.024 & -0.142 & 0.801 \\ 0.565 & 0.067 & -0.634 & -0.524 \\ \text{PC1, } \boldsymbol{v}_1 & \text{PC2, } \boldsymbol{v}_2 & \text{PC3, } \boldsymbol{v}_3 & \text{PC4, } \boldsymbol{v}_4 \end{bmatrix} \begin{matrix} \\ \\ \\ \leftarrow \text{Petal width, } x_4 \end{matrix} \begin{matrix} \leftarrow \text{Sepal length, } x_1 \\ \leftarrow \text{Sepal width, } x_2 \\ \leftarrow \text{Petal length, } x_3 \\ \\ \end{matrix} \tag{10}$$

可以发现 (7) 和 (10) 明显不同，下一章将对比这两种技术路线。

16.3 正交空间

矩阵 $\boldsymbol{V}$ 有 D 个列向量，对应 D 个正交基，如下：

$$\boldsymbol{V}=\begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,D-1} & v_{1,D} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,D-1} & v_{2,D} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ v_{D-1,1} & v_{D-1,2} & \cdots & v_{D-1,D-1} & v_{D-1,D} \\ v_{D,1} & v_{D,2} & \cdots & v_{D,D-1} & v_{D,D} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \ldots & \boldsymbol{v}_{D-1} & \boldsymbol{v}_D \end{bmatrix} \tag{11}$$

任意列向量 $\boldsymbol{v}_i$ 每一个元素都包含 $\boldsymbol{X}$ 列向量 $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D]$ 成分，即列向量 $\boldsymbol{v}_i$ 为 $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D]$ 线性组合。

$$
\begin{aligned}
\boldsymbol{v}_1 &= v_{1,1}\boldsymbol{x}_1 + v_{2,1}\boldsymbol{x}_2 + \ldots + v_{D-1,1}\boldsymbol{x}_{D-1} + v_{D,1}\boldsymbol{x}_D \\
\boldsymbol{v}_2 &= v_{1,2}\boldsymbol{x}_1 + v_{2,2}\boldsymbol{x}_2 + \ldots + v_{D-1,2}\boldsymbol{x}_{D-1} + v_{D,2}\boldsymbol{x}_D \\
&\ldots \\
\boldsymbol{v}_D &= v_{1,D}\boldsymbol{x}_1 + v_{2,D}\boldsymbol{x}_2 + \ldots + v_{D-1,D}\boldsymbol{x}_{D-1} + v_{D,D}\boldsymbol{x}_D
\end{aligned} \tag{12}
$$

图 10 所示为线性组合构造正交空间 $[\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_D]$。注意，$[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D]$ 类似于 $[\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_D]$，它们代表方向向量，而不是具体的数据。

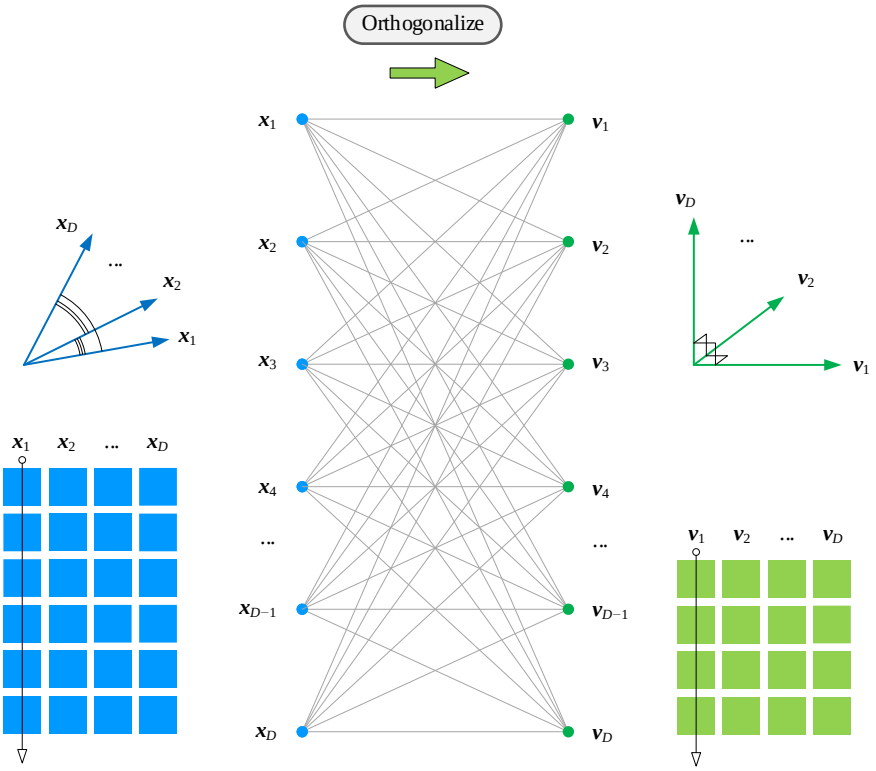


图 10. 线性组合构造正交空间 $[\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_D]$

如图 11 所示，以 $\boldsymbol{v}_1$ 为例，第一主成分方向上，$\boldsymbol{v}_1$ 等价于由 $v_{1,1}$ 比例 $\boldsymbol{x}_1$，$v_{2,1}$ 比例 $\boldsymbol{x}_2$，$v_{3,1}$ 比例 $\boldsymbol{x}_3$...以及 $v_{D,1}$ 比例 $\boldsymbol{x}_D$ 线性组合构造。从另外一个角度，$[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D]$ 在向量 $\boldsymbol{v}_1$ 上标量投影值分别为 $v_{1,1}, v_{2,1}, \ldots, v_{D,1}$。图 12 所示为鸢尾花数据主成分分析第一主成分 $\boldsymbol{v}_1$ 的构造情况。

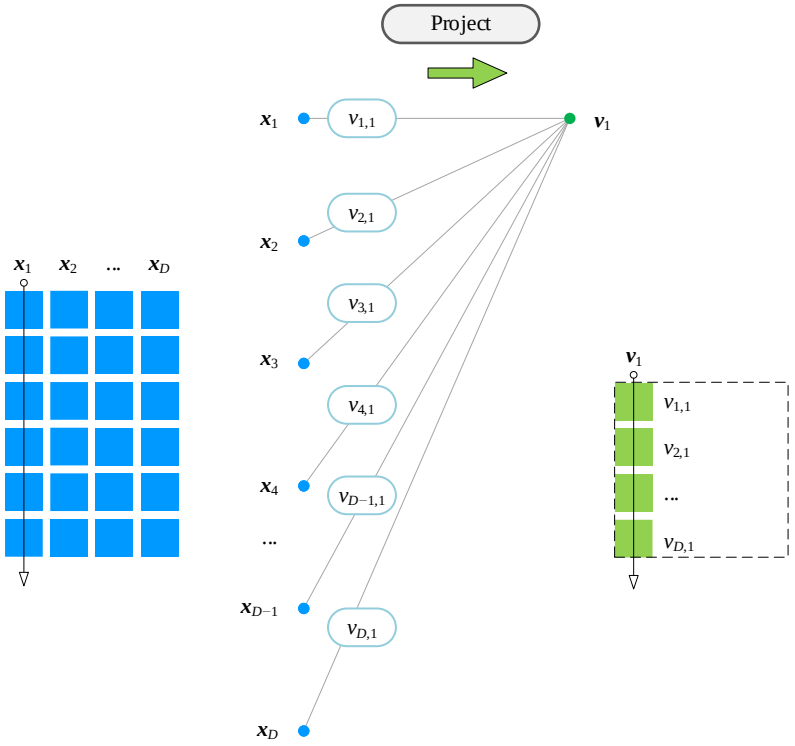


图 11. 构造第一主成分 $\boldsymbol{v}_1$

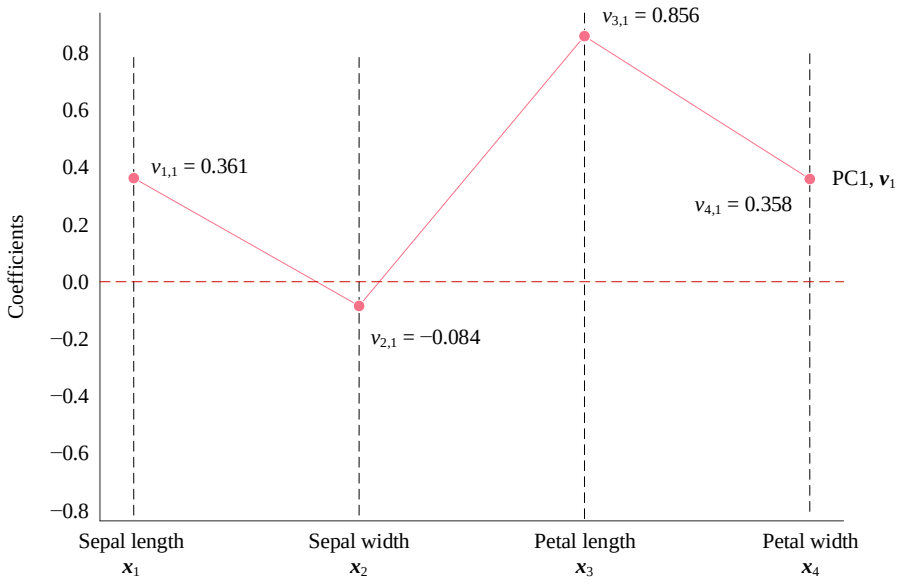


图 12. 构造第一主成分 $\boldsymbol{v}_1$，鸢尾花数据

如图 13 所示，第二主成分 $\boldsymbol{v}_2$ 方向上，$\boldsymbol{v}_2$ 等价于由 $v_{1,2}$ 比例 $\boldsymbol{x}_1$，$v_{2,2}$ 比例 $\boldsymbol{x}_2$，$v_{3,2}$ 比例 $\boldsymbol{x}_3$...以及 $v_{D,2}$ 比例 $\boldsymbol{x}_D$ 线性构造。图 14 所示为鸢尾花数据主成分分析第二主成分 $\boldsymbol{v}_2$ 的构造情况。

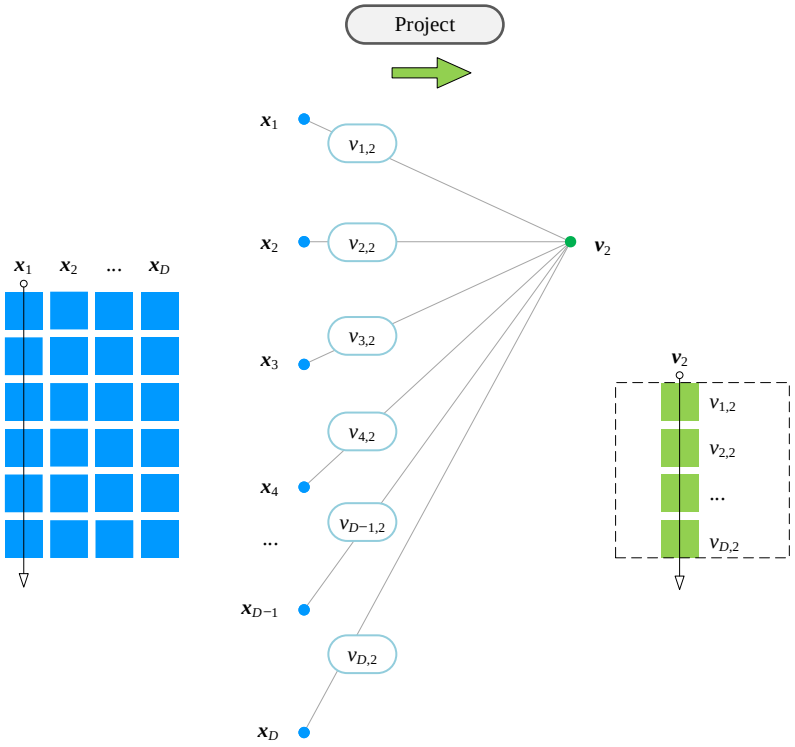


图 13. 构造第二主成分 $\boldsymbol{v}_2$

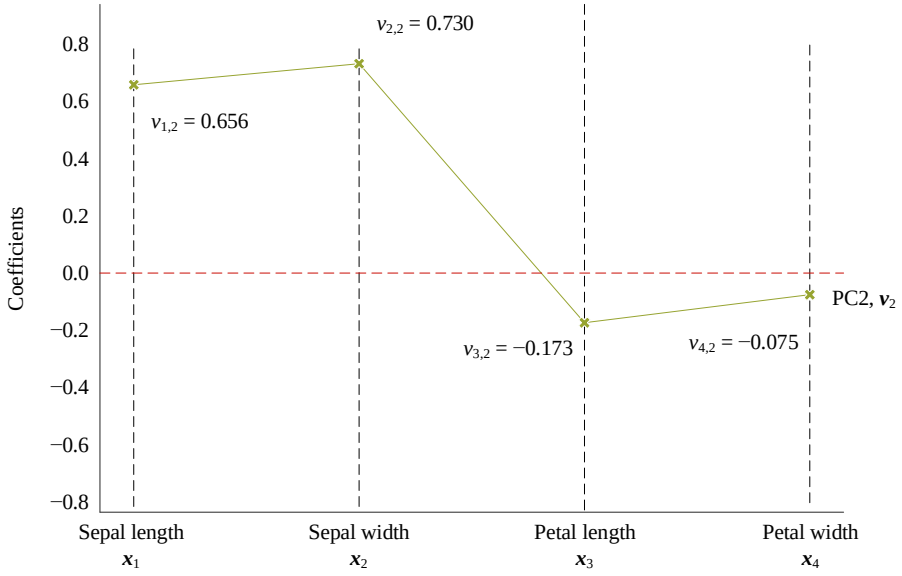


图 14. 构造第二主成分 $\boldsymbol{v}_2$，鸢尾花数据

如图 15 所示，第三主成分 $\boldsymbol{v}_3$ 方向上，$\boldsymbol{v}_3$ 等价于由 $v_{1,3}$ 比例 $\boldsymbol{x}_1$，$v_{2,3}$ 比例 $\boldsymbol{x}_2$，$v_{3,3}$ 比例 $\boldsymbol{x}_3$...以及 $v_{D,3}$ 比例 $\boldsymbol{x}_D$ 线性构造。图 16 所示为鸢尾花数据主成分分析第三主成分 $\boldsymbol{v}_3$ 的构造情况。

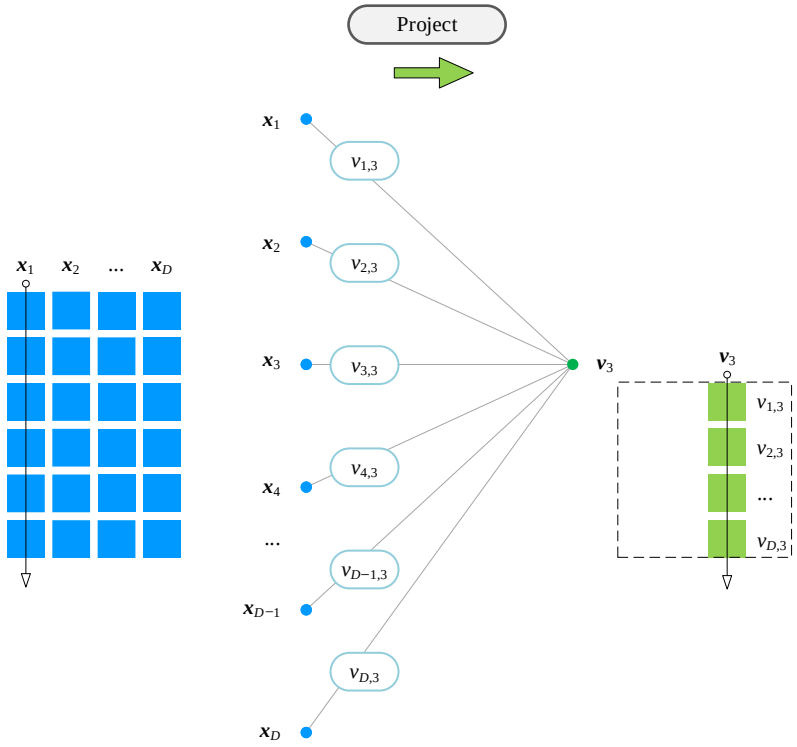


图 15. 构造第三主成分 $\boldsymbol{v}_3$

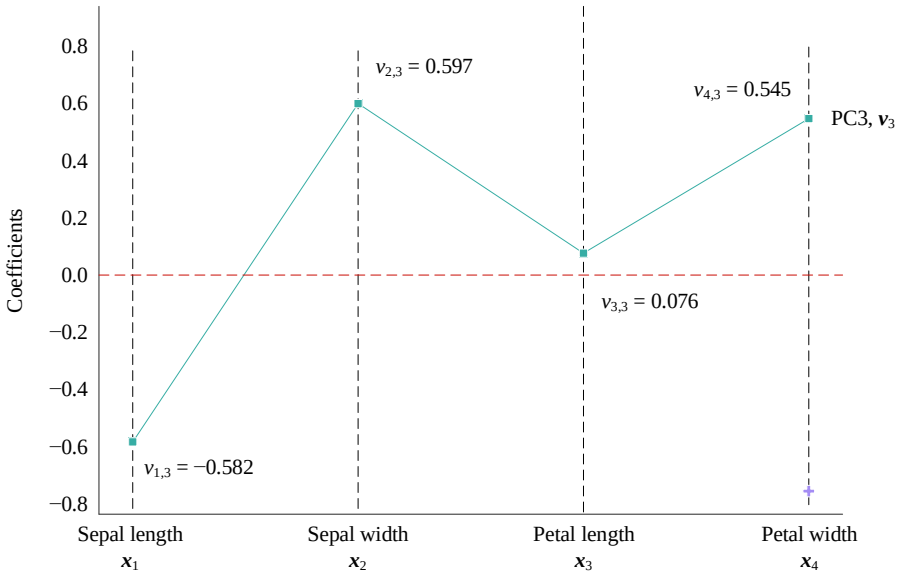


图 16. 构造第三主成分 $\boldsymbol{v}_3$，鸢尾花数据

16.4 投影结果

图 17 所示为投影后得到的新特征数据矩阵 $\boldsymbol{Z}$。这幅热图，蓝色色系数据接近 0，红色色系数据接近 8；可以发现矩阵 $\boldsymbol{Z}$ 四个新特征 (z_1, z_2, z_3 和 z_4) 从左到右颜色差异逐渐减小，即方差不断减小。

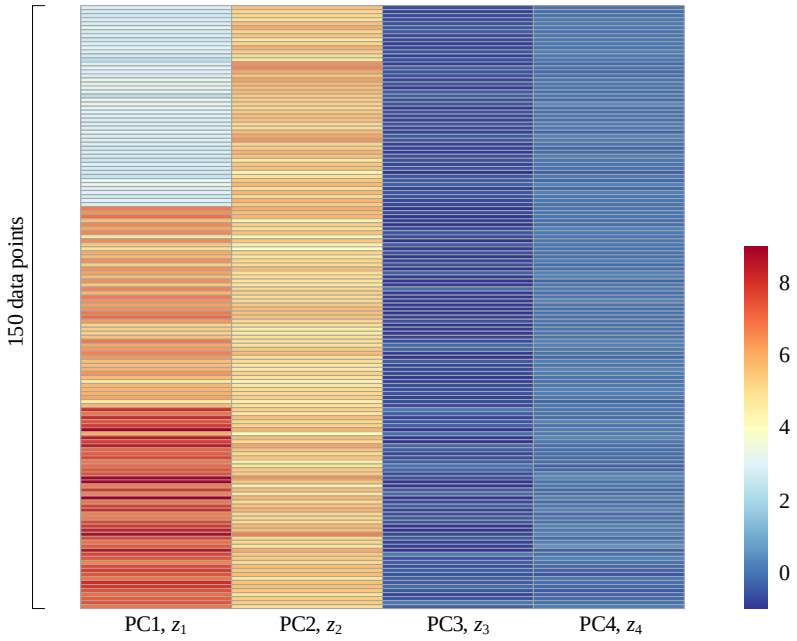


图 17. 新特征数据矩阵 $\boldsymbol{Z}$

对转换数据 $\boldsymbol{Z}$ 进行统计分析，以行向量表达数据矩阵 $\boldsymbol{Z}$ 质心：

$$\boldsymbol{\mu}_{\boldsymbol{Z}} = \begin{bmatrix} \underset{\text{PC1, } z_1}{5.502} & \underset{\text{PC2, } z_2}{5.326} & \underbrace{-0.631}_{\text{PC3, } z_3} & \underset{\text{PC4, } z_4}{0.033} \end{bmatrix} \tag{13}$$

数据矩阵 $\boldsymbol{Z}$ 质心和原始数据矩阵 $\boldsymbol{X}$ 质心之间的关系如下所示：

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{Z}} &= \boldsymbol{\mu}_{\boldsymbol{X}} \boldsymbol{V} \\ &= \begin{bmatrix} \underset{\text{Sepal length, } x_1}{5.843} & \underset{\text{Sepal width, } x_2}{3.057} & \underset{\text{Petal length, } x_3}{3.758} & \underset{\text{Petal width, } x_4}{1.199} \end{bmatrix} \begin{bmatrix} 0.521 & 0.377 & 0.720 & -0.261 \\ -0.269 & 0.923 & -0.244 & 0.124 \\ 0.580 & 0.024 & -0.142 & 0.801 \\ \underset{\text{PC1, } \boldsymbol{v}_1}{0.565} & \underset{\text{PC2, } \boldsymbol{v}_2}{0.067} & \underbrace{-0.634}_{\text{PC3, } \boldsymbol{v}_3} & \underbrace{-0.524}_{\text{PC4, } \boldsymbol{v}_4} \end{bmatrix} \\ &= \begin{bmatrix} \underset{\text{PC1, } z_1}{5.502} & \underset{\text{PC2, } z_2}{5.326} & \underbrace{-0.631}_{\text{PC3, } z_3} & \underset{\text{PC4, } z_4}{0.033} \end{bmatrix} \end{aligned} \tag{14}$$

注意，若使用 sklearn.decomposition.PCA() 函数进行主成分分析，则会发现数据矩阵 $\boldsymbol{Z}$ 质心均为 0；这是因为数据已经标准化。

$\boldsymbol{Z}$ 每一列均方差，以行向量表达：

$$\boldsymbol{\sigma}_{\boldsymbol{Z}} = \begin{bmatrix} \underset{\text{PC1, } z_1}{2.056} & \underset{\text{PC2, } z_2}{0.492} & \underset{\text{PC3, } z_3}{0.279} & \underset{\text{PC4, } z_4}{0.154} \end{bmatrix} \tag{15}$$

$\boldsymbol{Z}$ 每一列方差，以行向量表达：

$$\boldsymbol{\sigma}_{\boldsymbol{Z}}^2 = \begin{bmatrix} \underset{\text{PC1, } z_1}{4.228} & \underset{\text{PC2, } z_2}{0.242} & \underset{\text{PC3, } z_3}{0.078} & \underset{\text{PC4, } z_4}{0.023} \end{bmatrix} \tag{16}$$

图 18 所示为 KDE 估计得到的转换数据 $\boldsymbol{Z}$ 四个特征分布图。

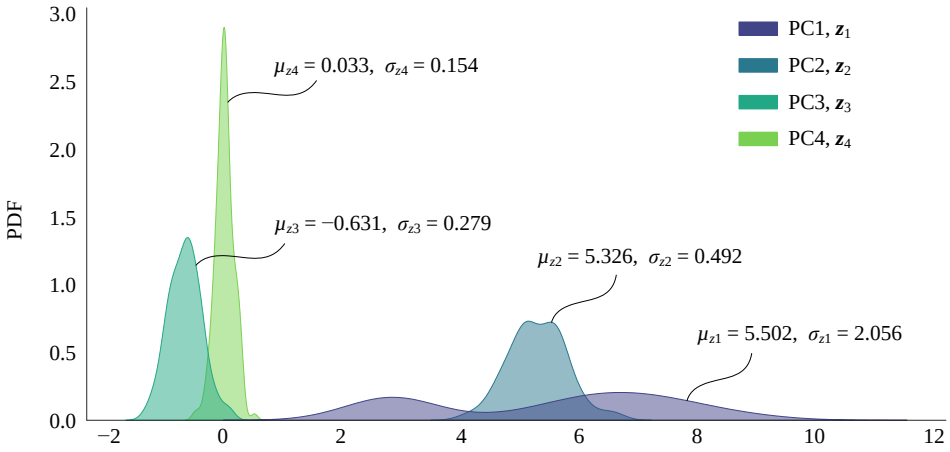


图 18. 转换数据 $\boldsymbol{Z}$ 四个特征上分布，KDE 估计

作为对比，图 19 所示为已经中心化的数据 $\boldsymbol{X}_c$ 朝 $\boldsymbol{V}$ 投影的结果。对比图 18 和图 19，我们可以发现方差没有变化。唯一的区别是，图 19 中所有特征的均值均为 0。注意，$\boldsymbol{V}$ 是通过对协方差矩阵特征值分解得到的。

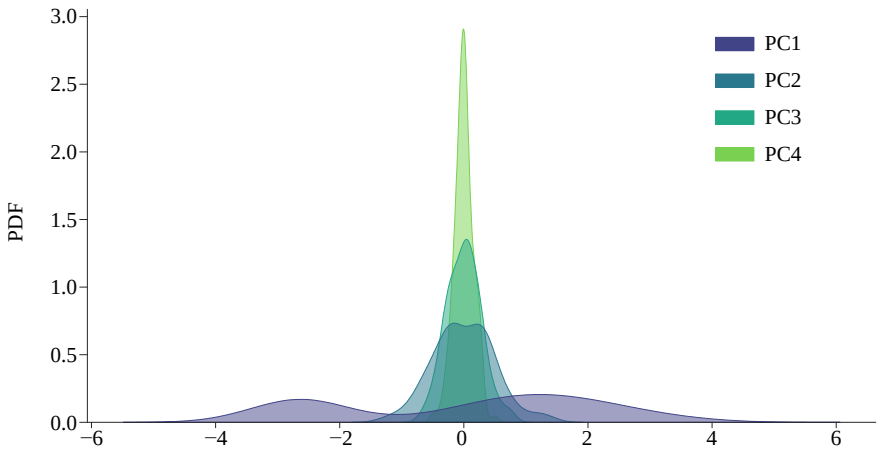


图 19. 转换数据 $\boldsymbol{Z}$ 四个特征上分布，KDE 估计；数据已经中心化

图 20 所示为转换数据 $\boldsymbol{Z}$ 协方差矩阵和相关性系数矩阵热图。

图 21 所示为不分类条件下，转换数据 $\boldsymbol{Z}$ 成对特征分析图；根据本节计算结果，可以知道转换数据 $\boldsymbol{Z}$ 任意两列数据之间的线性相关性系数为 0，也就是正交。图 22 所示为分类条件下，转换数据 $\boldsymbol{Z}$ 成对特征分析图。

$\boldsymbol{Z}$ 的协方差矩阵 $\boldsymbol{\Sigma_Z}$ 和 $\boldsymbol{X}$ 的协方差矩阵 $\boldsymbol{\Sigma_X}$ 之间关系如下：

$$\text{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}_x = \boldsymbol{V}^{\mathrm{T}} \boldsymbol{\Sigma}_x \boldsymbol{V} \tag{17}$$

图 20 所示为转换数据 $\boldsymbol{Z}$ 协方差矩阵和相关性系数矩阵热图。

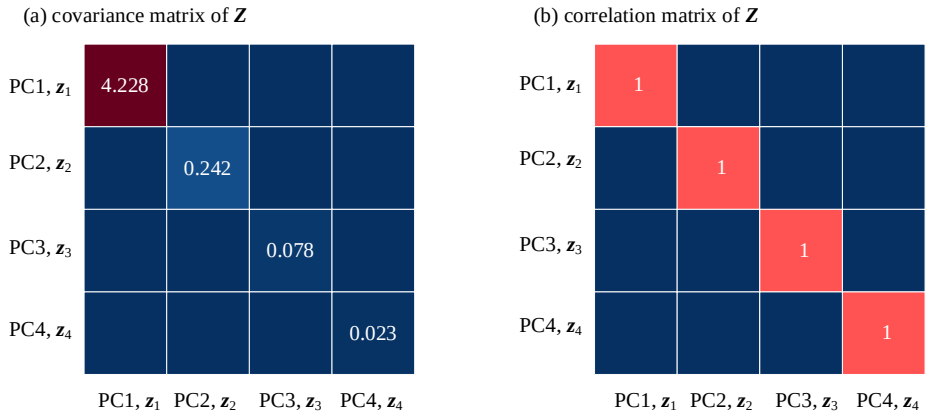


图 20. 转换数据 $\boldsymbol{Z}$ 协方差矩阵和相关性系数矩阵热图

图 21 所示为不分类条件下，转换数据 $\boldsymbol{Z}$ 成对特征分析图；根据本节计算结果，可以知道转换数据 $\boldsymbol{Z}$ 任意两列数据之间的线性相关性系数为 0，也就是正交。图 22 所示为分类条件下，转换数据 $\boldsymbol{Z}$ 成对特征分析图。

下一章还会用椭圆代表散点的分布情况。

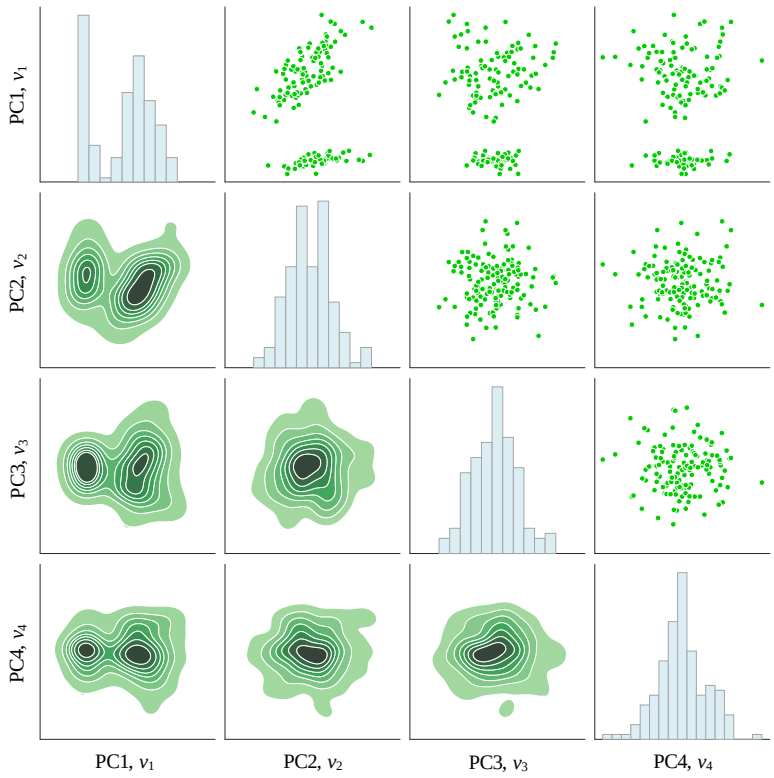


图 21. 转换数据 $\boldsymbol{Z}$ 成对特征分析图，不分类

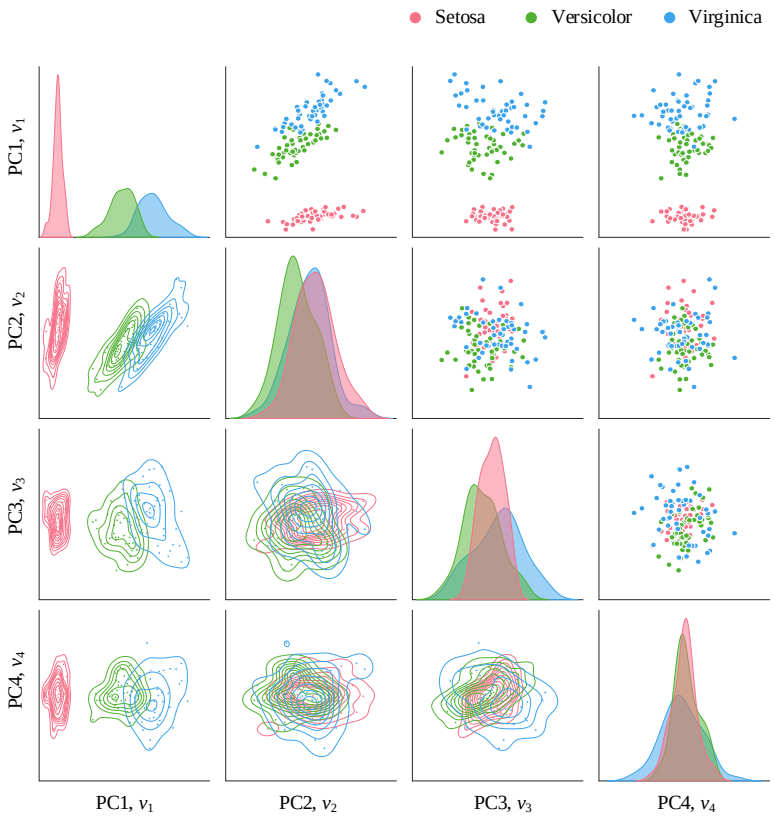


图 22. 转换数据 $\boldsymbol{Z}$ 成对特征分析图，分类

16.5 还原

根据本书前文介绍，主成分 $\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$ 还原部分原始数据：

$$\hat{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{z}_1 & \boldsymbol{z}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{bmatrix}^{\mathrm{T}} \tag{18}$$

残差数据矩阵 $\boldsymbol{E}$，即原始热图和还原热图色差，利用下式计算获得：

$$\boldsymbol{E} = \boldsymbol{X} - \hat{\boldsymbol{X}} \tag{19}$$

图 23 所示为 $\boldsymbol{z}_1$ 还原 $\boldsymbol{X}$ 部分数据。图 24 所示为 $\boldsymbol{z}_1$ 还原 $\boldsymbol{X}$ 部分数据。图 25 所示为 $[\boldsymbol{z}_1, \boldsymbol{z}_2]$ 还原 $\boldsymbol{X}$ 部分数据。比较原始数据和图 25 所示 $[\boldsymbol{z}_1, \boldsymbol{z}_2]$ 还原 $\boldsymbol{X}$ 部分数据，可以得到误差热图，如图 26 所示。

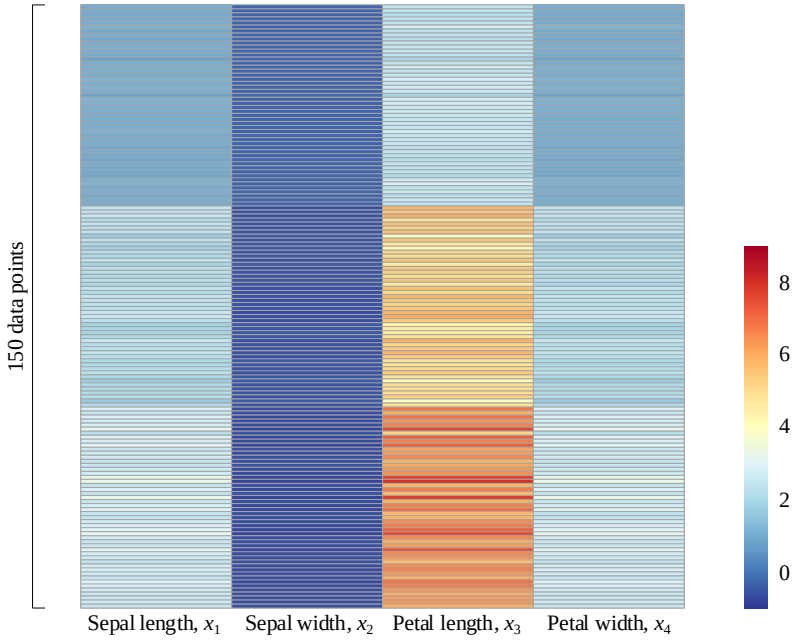


图 23. $\boldsymbol{z}_1$ 还原 $\boldsymbol{X}$ 部分数据

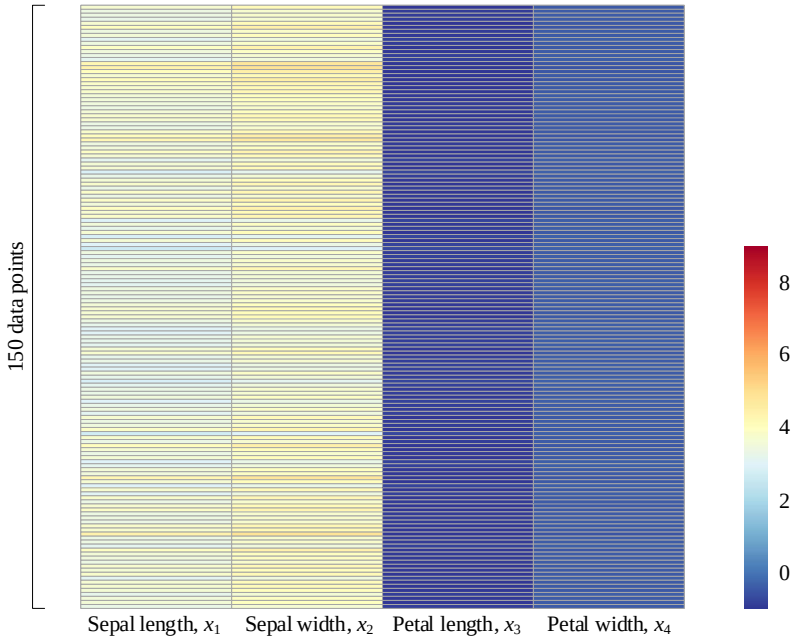


图 24. $\boldsymbol{z}_2$ 还原 $\boldsymbol{X}$ 部分数据

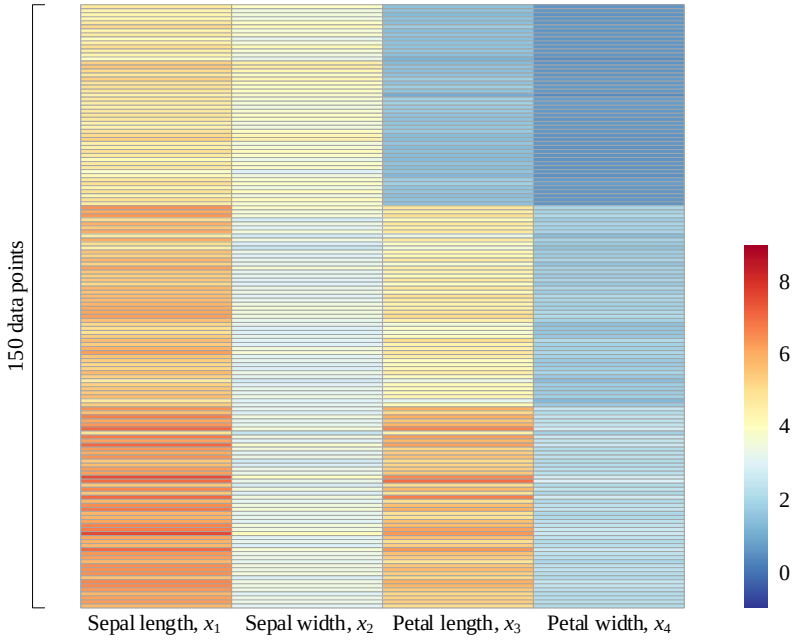


图 25.[$\boldsymbol{z}_1$, $\boldsymbol{z}_2$] 还原 $\boldsymbol{X}$ 部分数据

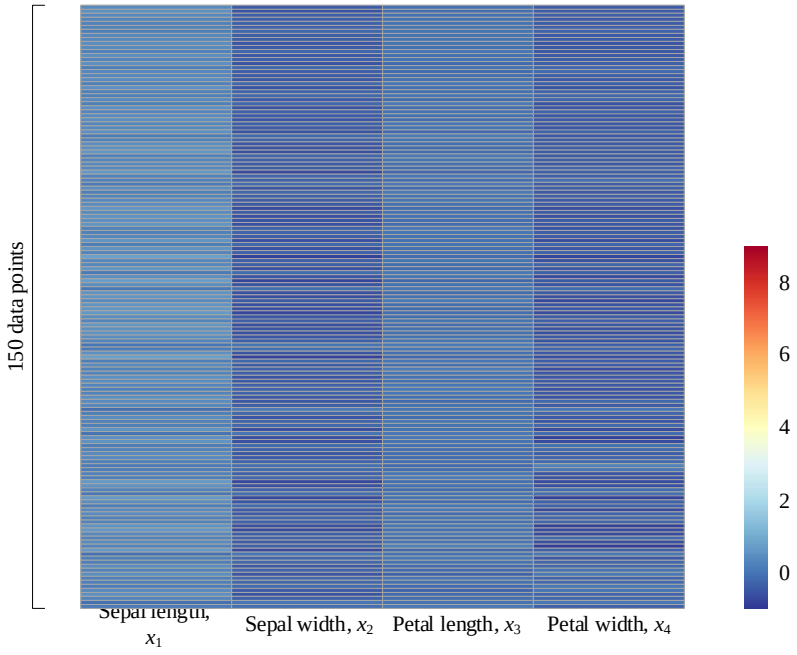


图 26. 误差 $\boldsymbol{E}$

16.6 双标图

双标图 (biplot) 是主成分分析中常用的可视化方案。如图 27 所示，双标图相当于原始数据特征向量向主成分构造的平面投影结果。比如，$\boldsymbol{x}_1$ 向量向 $\boldsymbol{v}_1$-$\boldsymbol{v}_2$ 平面投影，$\boldsymbol{x}_1$ 在 $\boldsymbol{v}_1$ 方向投影得到的标量值为 $v_{1,1}$，$\boldsymbol{x}_1$ 在 $\boldsymbol{v}_2$ 方向投影得到的标量值为 $v_{1,2}$。这两个值对应 $\boldsymbol{V}$ 矩阵第一行前两列数值。

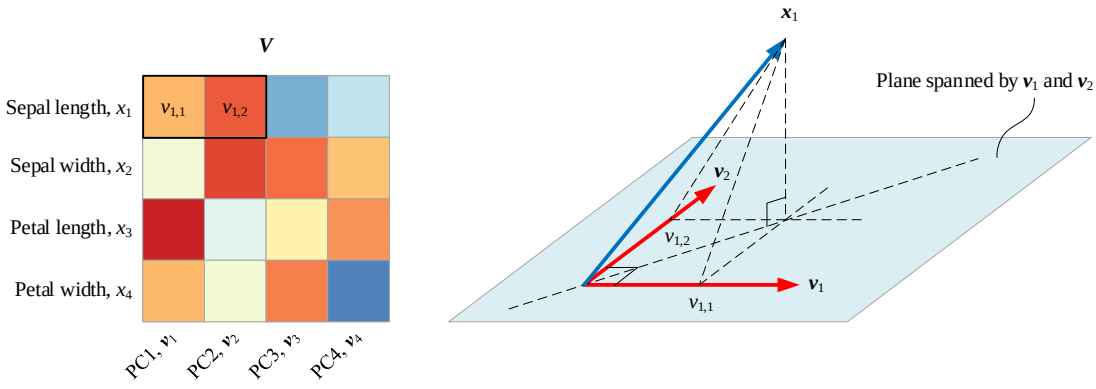


图 27. 双标图原理

图 28 所示为鸢尾花原始数据 PCA 分解后得到的双标图。该图横纵坐标分别是第一主成分 $\boldsymbol{v}_1$ 和第二主成分 $\boldsymbol{v}_2$。如图 28 所示，在双标图上，如果两个特征向量夹角越小，说明两个特征相似度越

高，也就是相关性系数越高。比如图中，花萼长度 x_3 和花萼宽度 x_4，在双标图上几乎重合，说明两者相关性极高，(4) 中给出的两者相关性高达 0.963，这也印证了这一点。

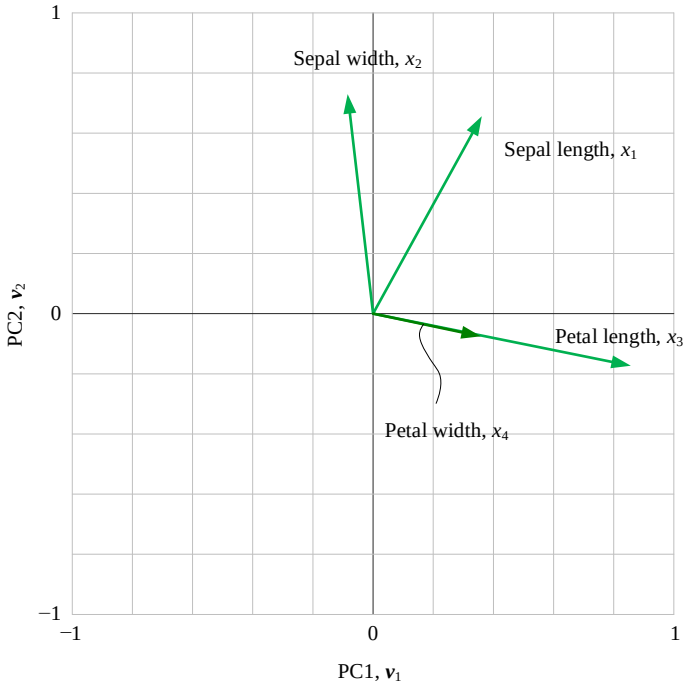


图 28. v_1-v_2 平面双标图，基于鸢尾花原始数据

图 29 所示为向量 x_1、x_2、x_3 和 x_4 向 v_1-v_2 平面投影结果和矩阵 V 之间的数值关系。

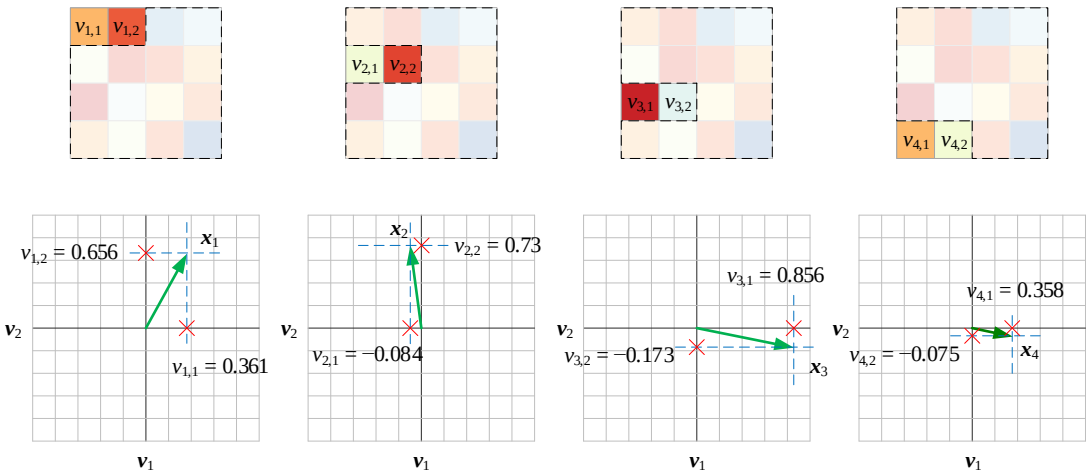


图 29. 向量 x_1、x_2、x_3 和 x_4 向 v_1-v_2 平面投影结果

图 30 所示为向量 $\boldsymbol{x}_1$、$\boldsymbol{x}_2$、$\boldsymbol{x}_3$和 $\boldsymbol{x}_4$向 $\boldsymbol{v}_3$-$\boldsymbol{v}_4$平面投影结果。

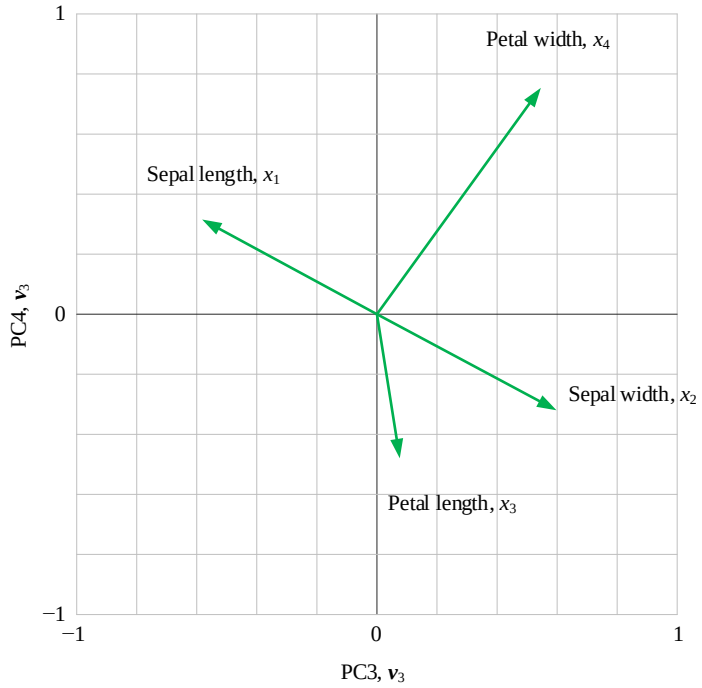


图 30. $\boldsymbol{v}_3$-$\boldsymbol{v}_4$平面双标图，基于鸢尾花原始数据

双标图还可以基于标准化后数据；图 31 所示为基于鸢尾花标准化数据后的双标图，投影值对应 (10)。

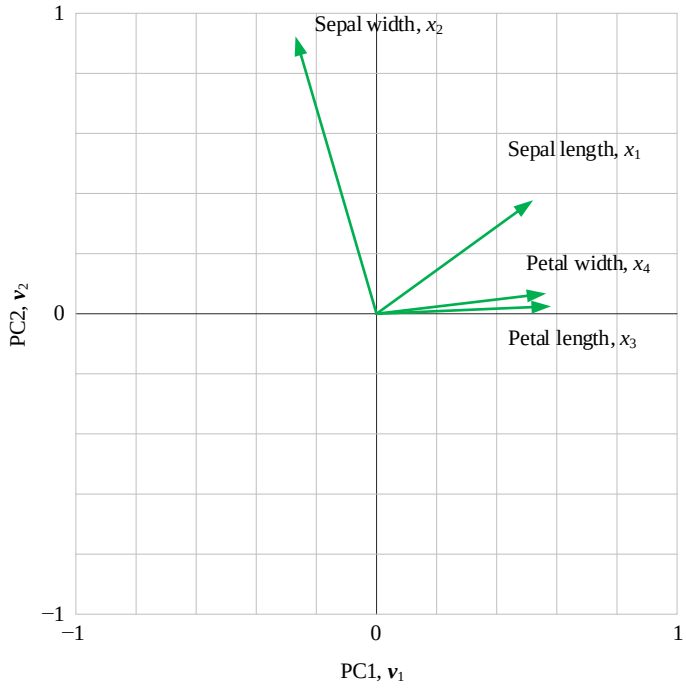


图 31. 平面双标图，基于鸢尾花标准化数据

此外，除了特征向量之外，双标图还会绘制数据点投影，如图 32 所示。图 32 采用 yellowbrick.features.PCA() 绘制。该函数绘制的双标图基于标准化鸢尾花数据。双标图中，点与点之间的距离，反映它们对应的样本之间的差异大小，两点相距较远，对应样本差异大；两点相距较近，对应样本差异小，存在相似性。

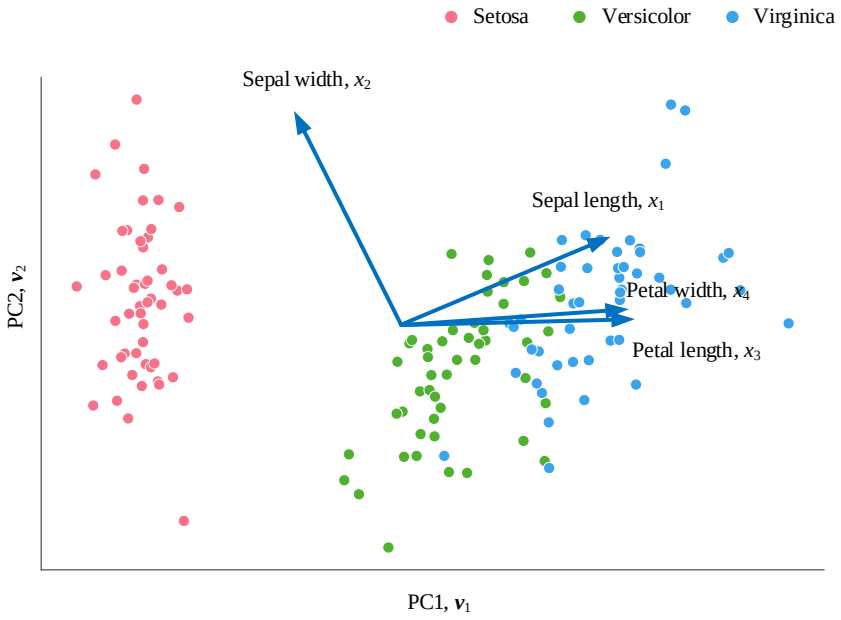


图 32. 平面双标图，标准化数据

图 33 给出的是由前三个主成分构造的空间，也就是将原始数据和它的四个特征向量投影到这个三维正交空间。该图也是采用 yellowbrick.features.PCA() 绘制。

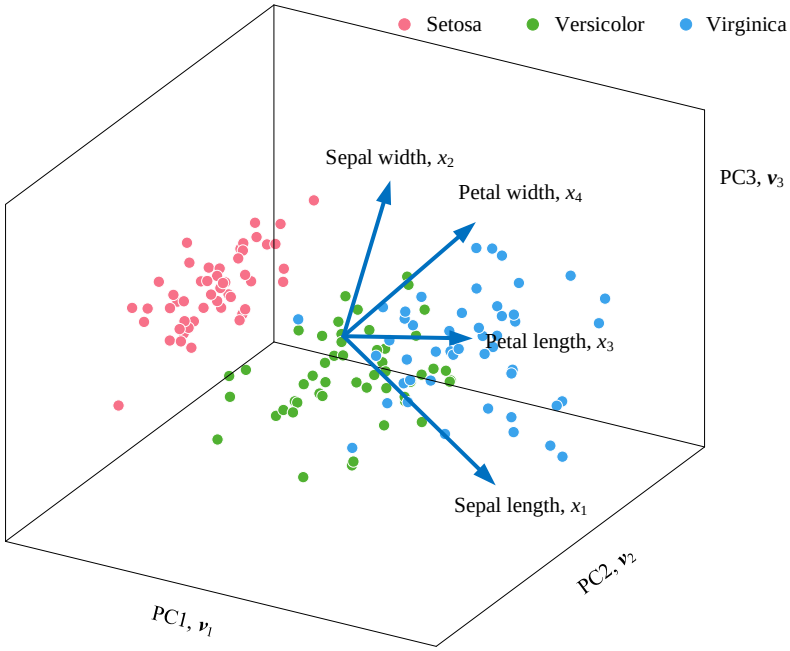


图 33. 三维双标图

16.7 陡坡图

《统计至简》第 25 章介绍过，第 j 个特征值 λ_j 对方差总和的贡献百分比为：

$$\frac{\lambda_j}{\sum_{i=1}^{D}\lambda_i}\times 100\% \tag{20}$$

前 p 个特征值累积解释总方差的百分比为：

$$\frac{\sum_{j=1}^{p}\lambda_j}{\sum_{i=1}^{D}\lambda_i}\times 100\% \tag{21}$$

(21) 分子一项是 p 个主成分已释方差 (explained variance)；(21) 分母是累计已释方差和百分比 (cumulative explained variance ratio)。

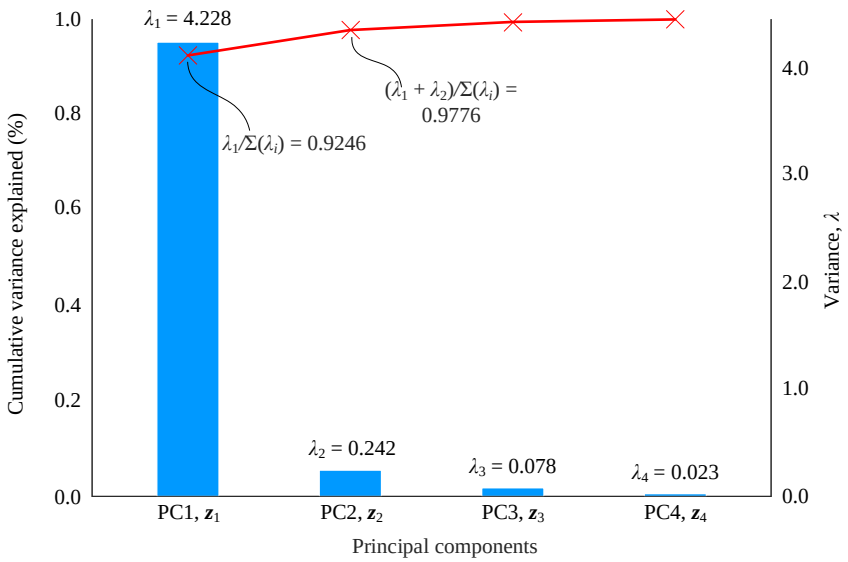


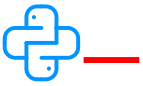
图 34. 陡坡图

图 34 给出图像可视化 (20) 和 (21) 。鸢尾花数据的主成分分析特征值如下：

$$\lambda_1=4.228,\quad \lambda_2=0.242,\quad \lambda_3=0.078,\quad \lambda_4=0.023 \tag{22}$$

PCA 主成分顺序根据各个主成分维度方向方差贡献大小排序。第一主成分方向上的方差最大，也就是这个方向最有力地解释了数据的分布。当第一主成分的方差贡献不足 (比如小于

50%)，我们就要依次引入其它主成分。如图 34 所示，第一和第二主成分两者已释方差之和为 72.5%。



Bk6_Ch16_01.py 绘制本章前文大部分图片。

16.8 分析鸢尾花照片

本节用 PCA 分析一章鸢尾花照片。图 35 所示为作者拍的一章鸢尾花照片，经过黑白化处理后的每个像素都是 [0, 1] 范围内的数字。所以整幅图片可以看成一个数据矩阵。

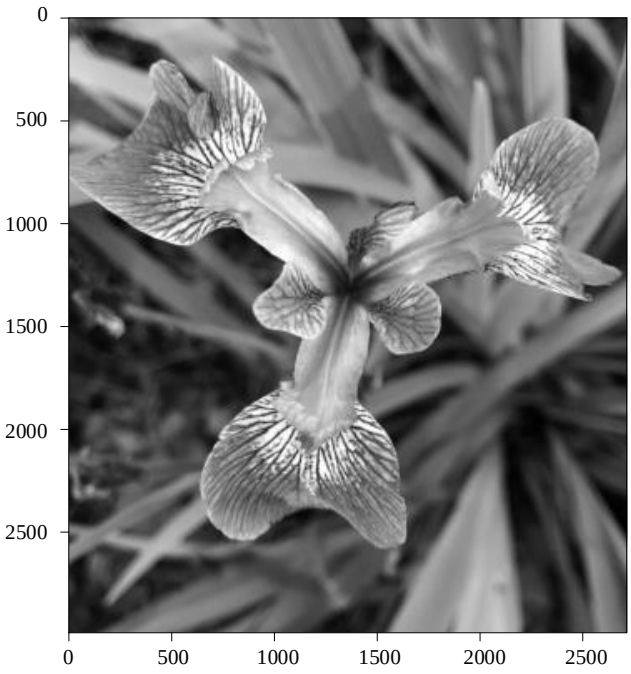


图 35. 鸢尾花图片，经过黑白处理

图 36 所示为利用 SVD 分解得到的奇异值随主成分变化。图 37 所示为特征值随主成分变化。图 38 所示为累积解释方差百分比随主成分变化。我们可以发现前 10 个主成分已经解释超过 90%的方差。

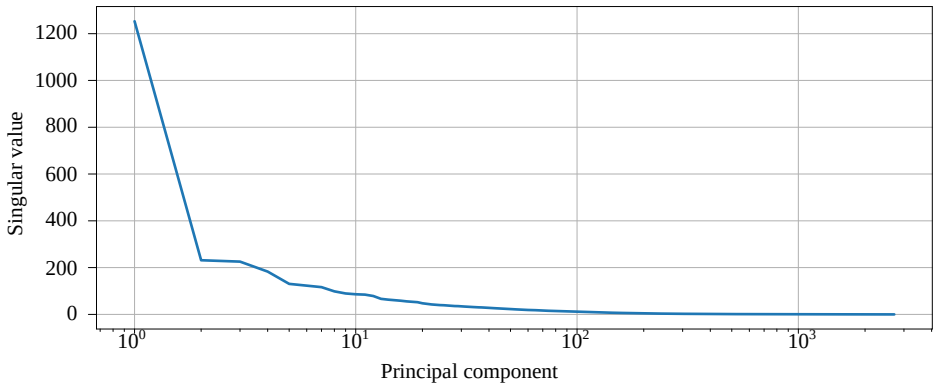


图 36. 奇异值随主成分变化

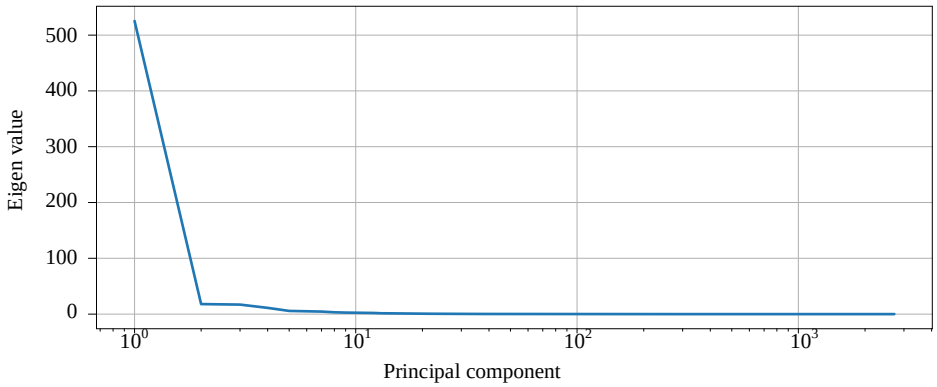


图 37. 特征值随主成分变化

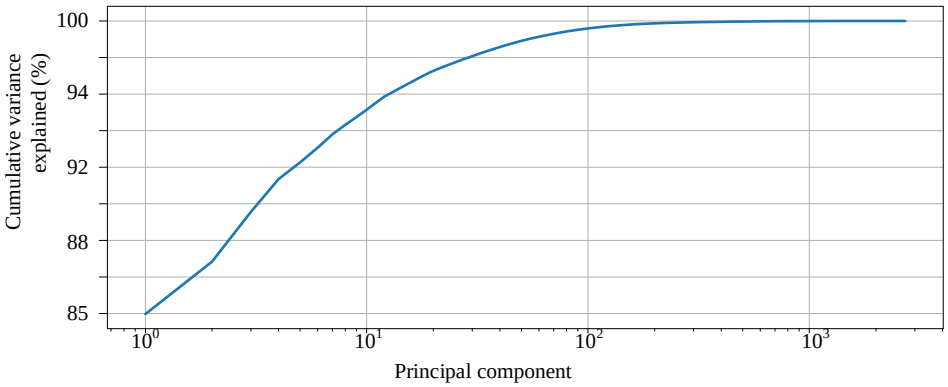


图 38. 累积解释方差百分比随主成分变化

图 39 所示为利用第 1 主元还原鸢尾花图片，左图为还原结果，右图为误差。左图中，鸢尾花还难觅踪影。图 40 所示为利用第 1、2 主元还原鸢尾花照片，图 41 所示为利用前 4 个主元还原鸢尾花照片，在两幅图的左图中我们仅仅能够看到“格子”。图 42 的左图利用前 16 个主元还原照片，我们已经能够看到鸢尾花的样子，注意这幅图的秩为 16。图 43 所示为利用前 64 个主元还原鸢尾花图片，图形已经很清晰。相比原图片，图 43 的数据发生大幅压缩。

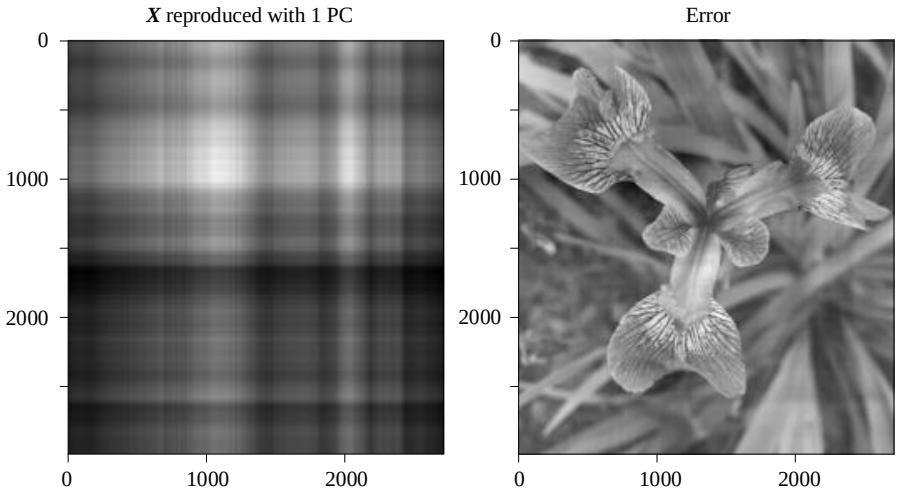


图 39. 利用第 1 主元还原鸢尾花照片

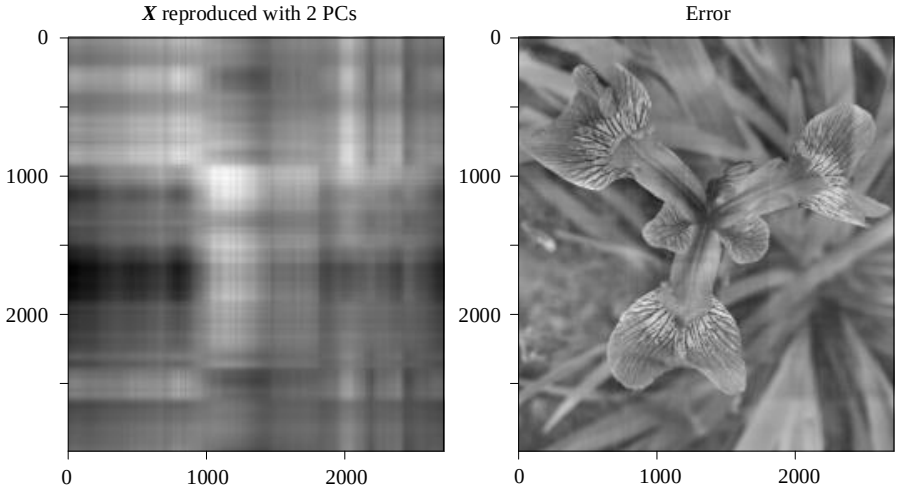


图 40. 利用第 1、2 主元还原鸢尾花照片

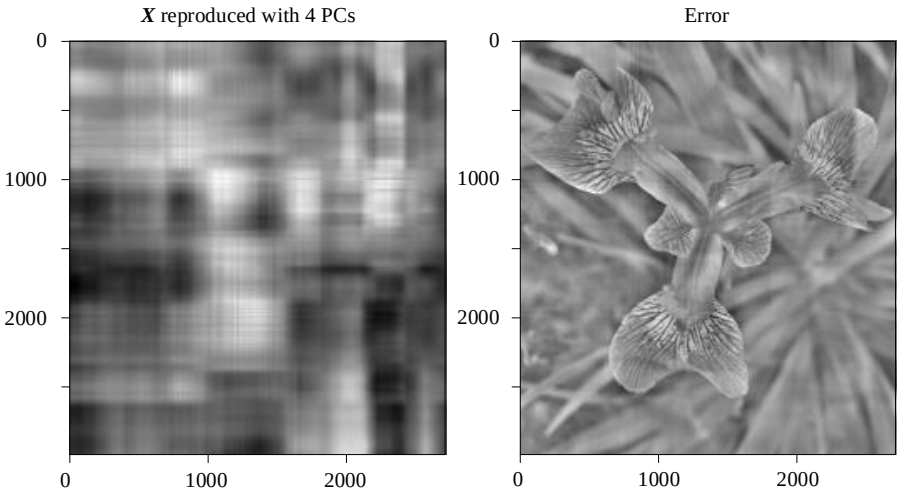


图 41. 利用第 1、2、3、4 主元还原鸢尾花照片

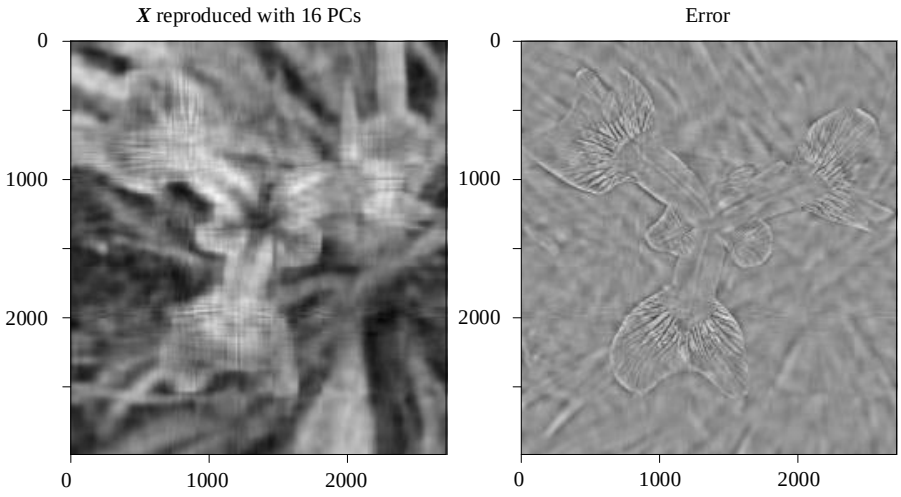


图 42. 利用前 16 个主元还原鸢尾花照片

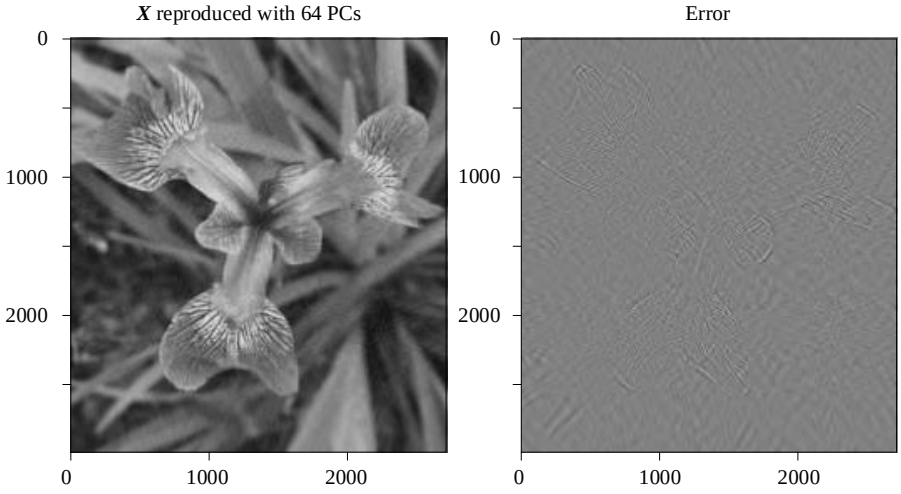


图 43. 利用前 64 个主元还原鸢尾花照片

Bk6_Ch16_02.py 绘制本节图片。鸢尾花照片也在文件夹中。

Statsmodels 也提供 PCA 分解的工具，请大家自行学习如下例子：

https://www.statsmodels.org/dev/examples/notebooks/generated/pca_fertility_factors.html