14 主成分分析

Principal Component Analysis

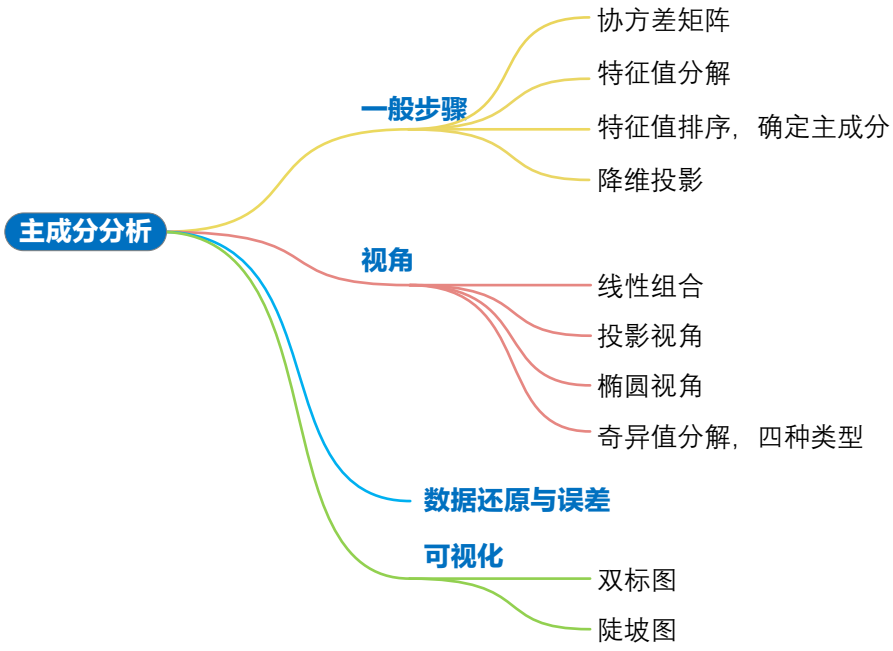
处理多维数据，通过降维发现数据隐藏规律

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

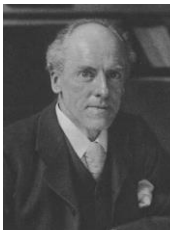
- `numpy.corrcoef()` 计算相关性系数矩阵
- `numpy.cov()` 计算协方差矩阵
- `numpy.linalg.eig()` 特征值分解
- `numpy.linalg.svd()` 奇异值分解
- `numpy.mean()` 计算均值
- `numpy.random.multivariate_normal()` 产生多元正态分布随机数
- `numpy.std()` 计算均方差
- `numpy.var()` 计算方差
- `numpy.zeros_like()` 产生形如输入矩阵的全 0 矩阵
- `seaborn.heatmap()` 绘制热图
- `seaborn.jointplot()` 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.lineplot()` 绘制线图
- `seaborn.pairplot()` 绘制成对分析图
- `sklearn.decomposition.PCA()` 主成分分析函数
- `yellowbrick.features.PCA()` 绘制 PCA 双标图



14.1 主成分分析

几何视角

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

读过《编程不难》的读者对图 1 应该很熟悉。如图所示，平面散点朝 16 个不同方向投影，并计算投影结果的方差值。

从图 1 中每个投影结果的分布宽度，用标准差量化，我们就可以得知 C、K 这两个方向就是我们要找的第一主成分方向。G、O 这两个方向也值得我们关注，因为这两个方向上投影结果的方差（标准差的平方）最小。

请大家格外注意，图 1 中样本数据质心位于原点，也就是说数据经过中心化，即去均值。比较 A、E 两个方向，我们可以发现标准差几乎相同；我们可以认为数据经过了标准化。

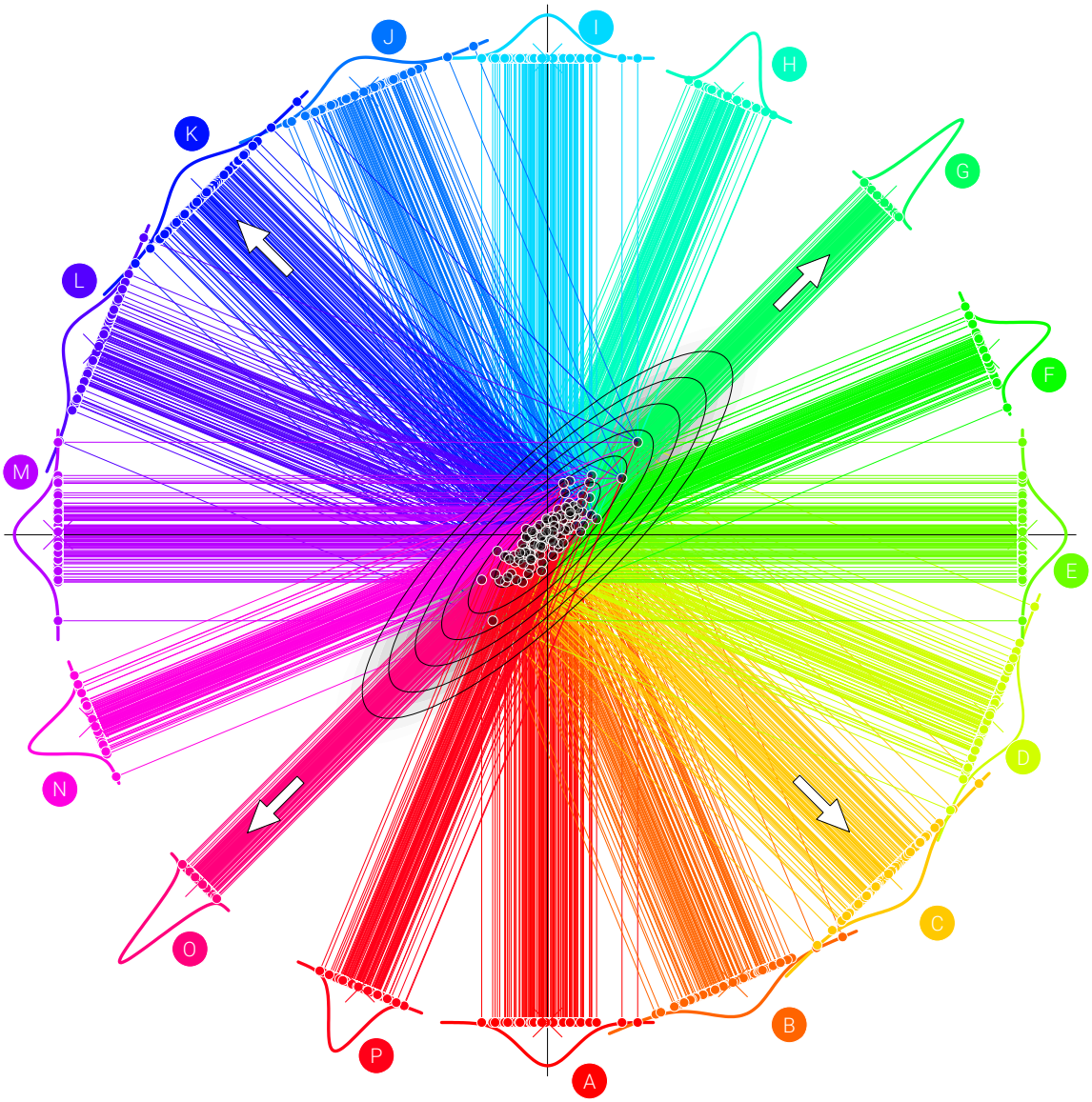


图 1．二维数据分别朝 16 个不同方向投影，图片来自《编程不难》

更通俗地讲，主成分分析实际上寻找数据在主元空间内投影。图 2 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 3 展示杯子六个平面上投影结果。

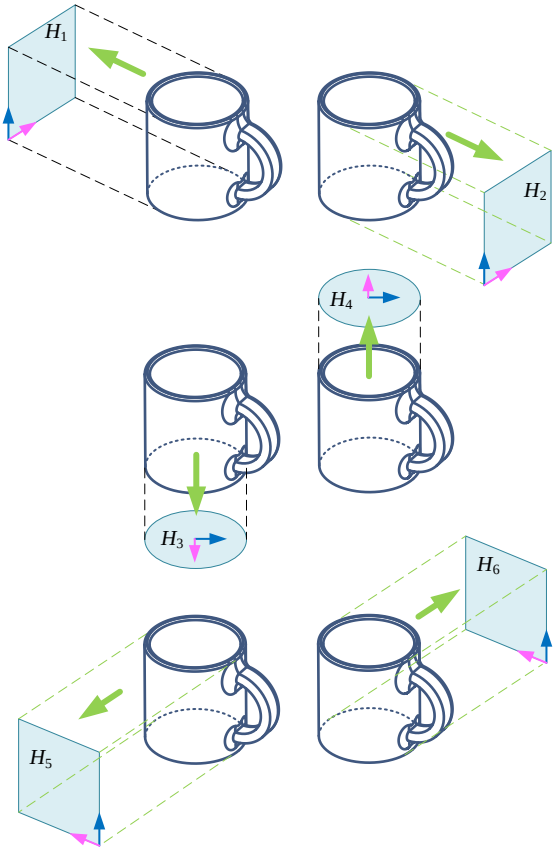


图 2. 咖啡杯六个投影方向

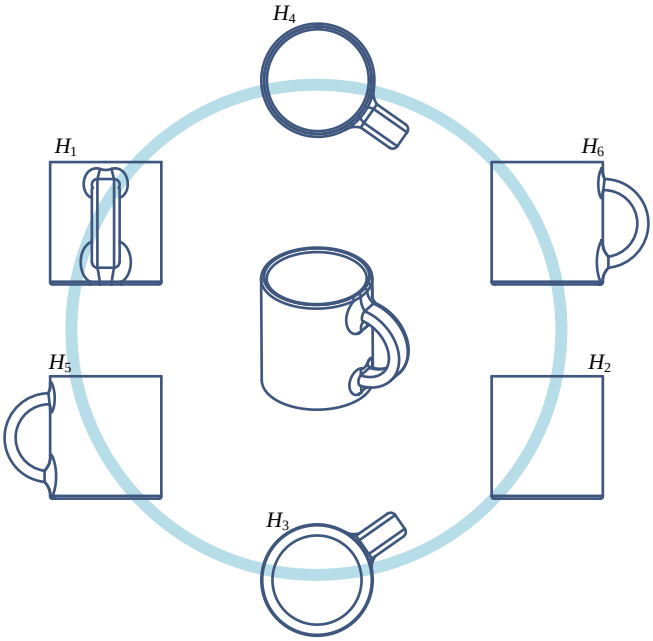


图 3. 咖啡杯在六个方向投影图像

14.2 原始数据

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 4 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。

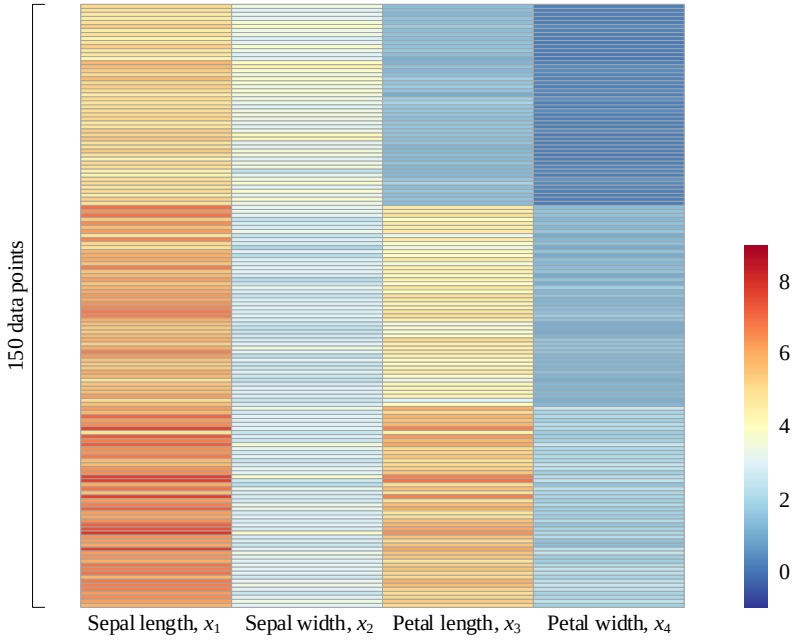


图 4. 鸢尾花数据，原始数据矩阵 $\boldsymbol{X}$

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：

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

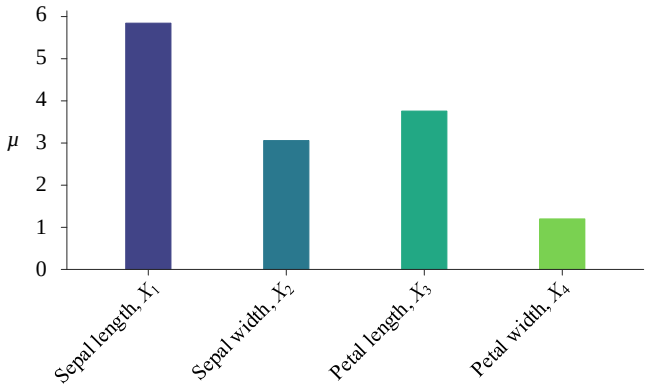


图 5. 鸢尾花数据四个特征上均值

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 6 所示为 KDE 估计得到的鸢尾花四个特征分布图。

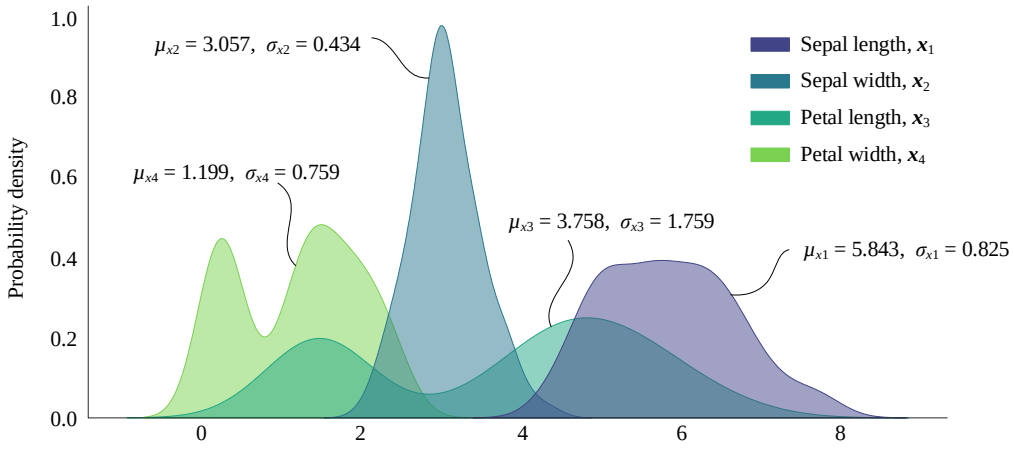


图 6. 鸢尾花数据四个特征上分布，KDE 估计

利用 seaborn.pairplot() 函数可以绘制如图 7 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 8 所示。图 8 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。

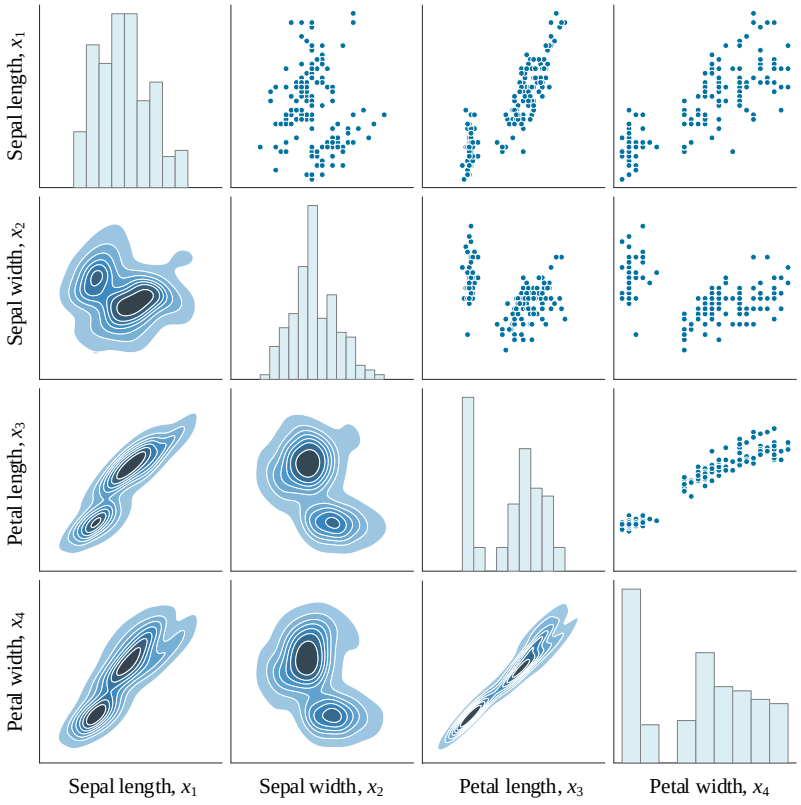


图 7. 鸢尾花数据成对特征分析图，不分类

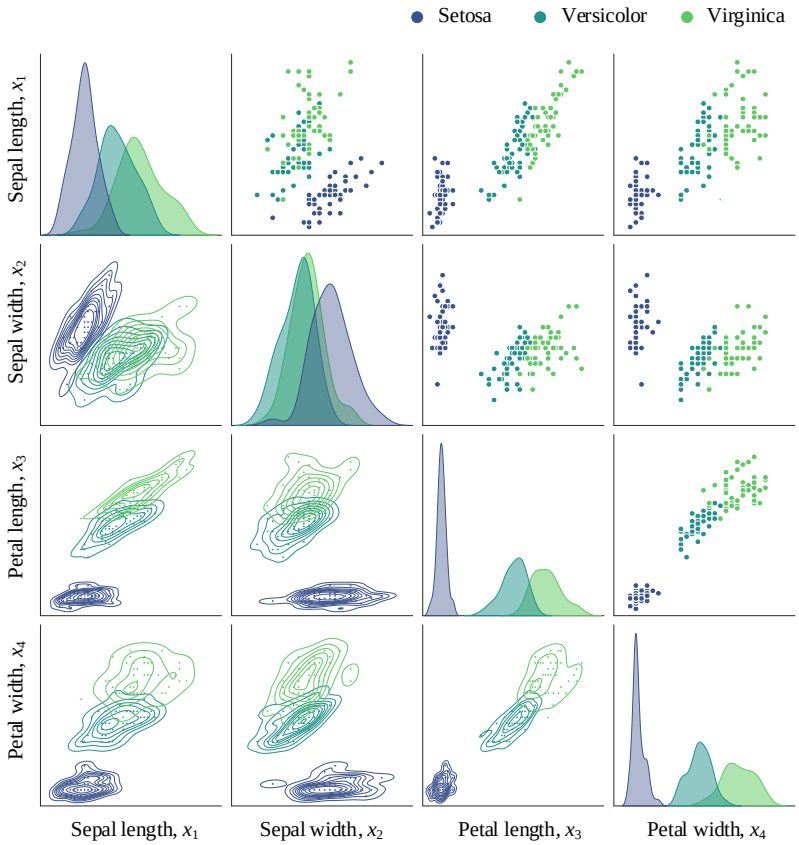


图 8. 鸢尾花数据成对特征分析图，分类

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

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

$$\begin{matrix} \text{Sepal length, } x_1 & \text{Sepal width, } x_2 & \text{Petal length, } x_3 & \text{Petal width, } x_4 \end{matrix}$$

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

在 PCA 中，有时候会对数据进行标准化是因为不同特征的单位和尺度不同，可能会对 PCA 的结果产生影响。如果不进行标准化处理，那么在协方差矩阵的计算过程中，某些特征的方差较大，将会对 PCA 的结果产生更大的影响，而这些特征不一定是最重要的。因此，为了消除这种影响，我们需要对数据进行标准化处理。

标准化的目的是将不同特征的值域缩放到相同的范围，使得所有特征的平均值为 0，标准差为 1，从而消除不同特征间的单位和尺度差异，使得所有特征具有相同的重要性。原始数据标准化的结果是 Z 分数。Z 分数的协方差矩阵实际上是原始数据的相关性系数矩阵。

总结来说，在进行 PCA 之前，如果数据中的特征具有不同的度量单位，或者特征值的范围变化很大，那么就应该考虑进行标准化。标准化可以使得 PCA 的结果更加准确和可靠，避免某些特征在主成分分析中被过度强调或者忽略。但是需要注意的是，有些情况下，标准化并不适用于所有数据集，例如当数据中的特征已经被精心设计或处理过时，标准化可能会使得信息损失或降低 PCA 的效果。

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

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

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。

14.3 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 $\boldsymbol{\Sigma}$ 为对称矩阵，因此上式为谱分解，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：

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

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 9 所示为不同主成分的系数线图。

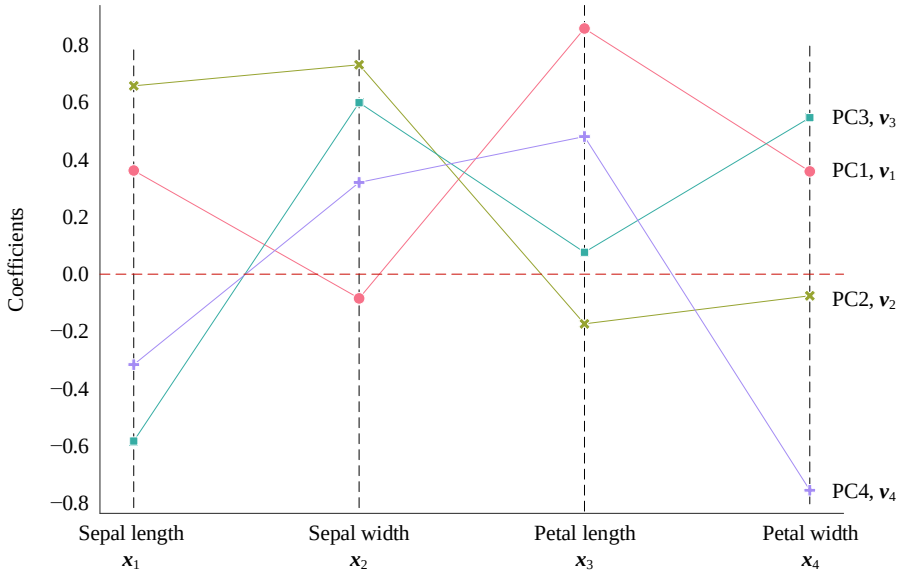


图 9. V 系数线图

如图 10 所示，V 和自己转置 V^T 乘积为单位阵 I，即：

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

展开上式得到：

$$\begin{bmatrix} v_1 & v_2 & v_3 & v_4 \end{bmatrix}^T \begin{bmatrix} v_1 & v_2 & v_3 & v_4 \end{bmatrix} = \begin{bmatrix} v_1^T \\ v_2^T \\ v_3^T \\ v_4^T \end{bmatrix} \begin{bmatrix} v_1 & v_2 & v_3 & v_4 \end{bmatrix}$$
$$= \begin{bmatrix} v_1^T v_1 & v_1^T v_2 & v_1^T v_3 & v_1^T v_4 \\ v_2^T v_1 & v_2^T v_2 & v_2^T v_3 & v_2^T v_4 \\ v_3^T v_1 & v_3^T v_2 & v_3^T v_3 & v_3^T v_4 \\ v_4^T v_1 & v_4^T v_2 & v_4^T v_3 & v_4^T v_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = I \tag{9}$$

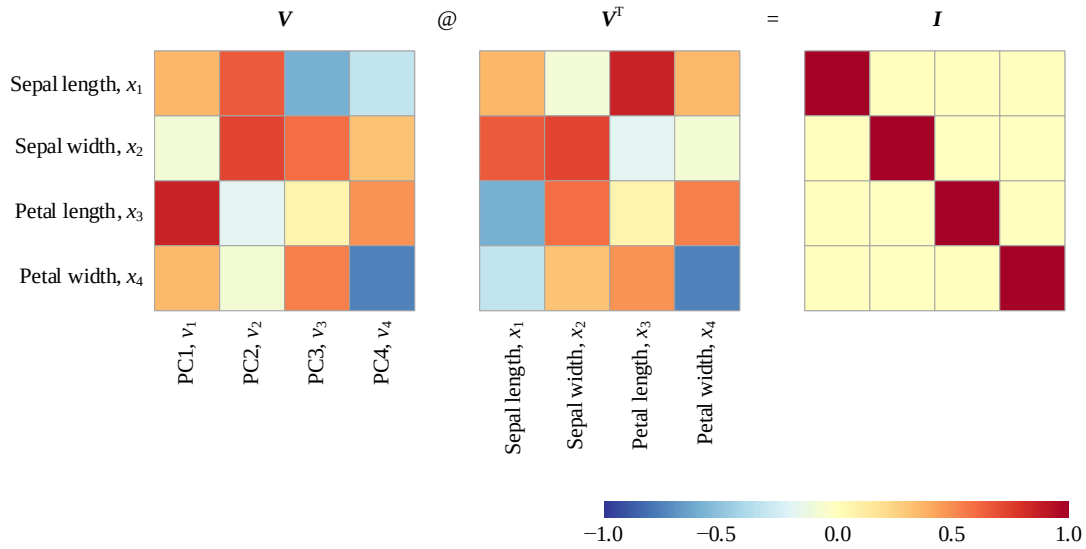


图 10. 特征矩阵 $\boldsymbol{V}$ 和自身转置的乘积为单位矩阵 $\boldsymbol{I}$

对相关性系数矩阵进行特征值分解得到的 $\boldsymbol{V}$ 为：

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

可以发现 (7) 和 (10) 明显不同，本书第 16 章将对比这两种技术路线。

14.4 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 $[\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_D]$ 成分，即列向量 $\boldsymbol{v}_i$ 为 $[\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_D]$ 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 11 所示为线性组合构造正交空间 $[\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_D]$。注意，$[\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_D]$ 类似于 $[\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_D]$，它们代表方向向量，而不是具体的数据。

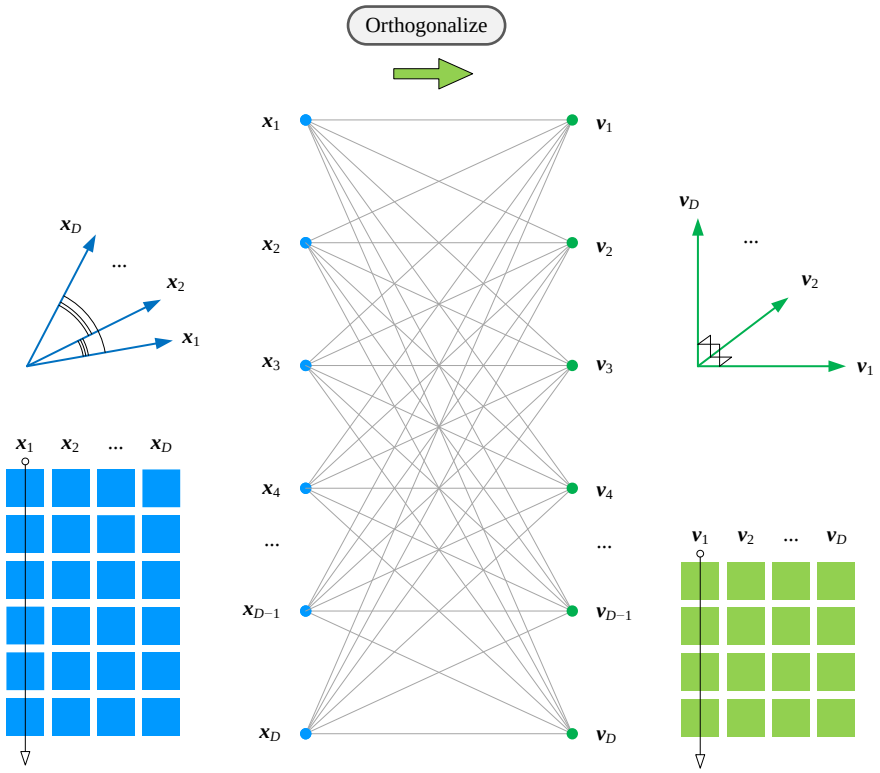


图 11. 线性组合构造正交空间 $[\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_D]$

如图 12 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，$[\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_D]$ 在向量 $\boldsymbol{v}_1$ 上标量投影值分别为 $v_{1,1}, v_{2,1}, ..., v_{D,1}$。图 13 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。

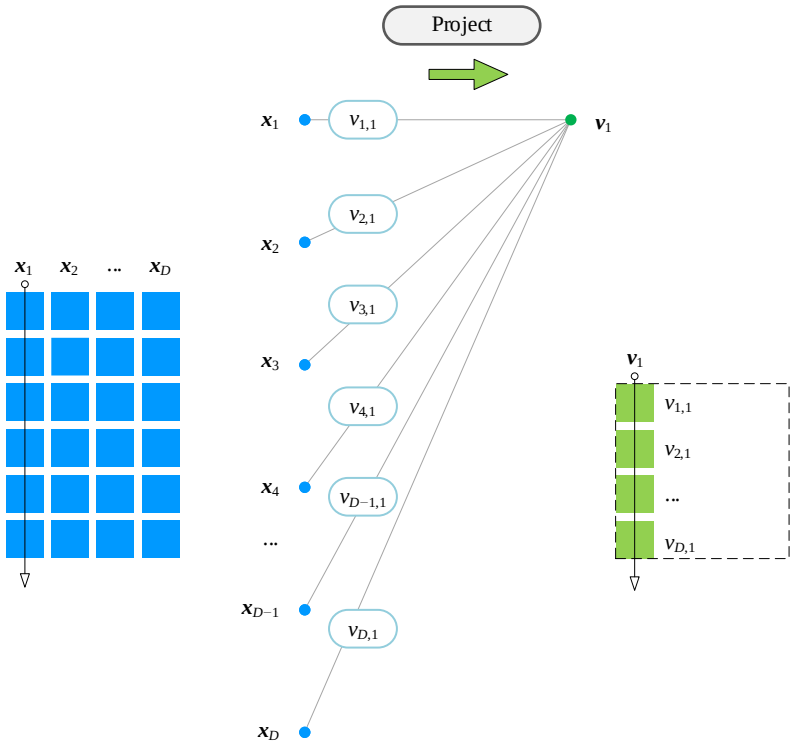


图 12. 构造第一主成分 $\boldsymbol{v}_1$

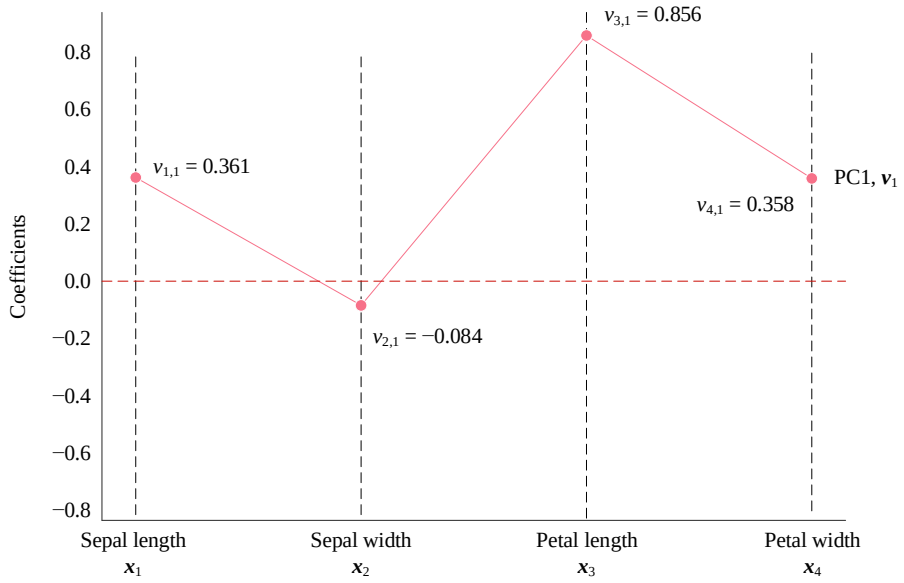


图 13. 构造第一主成分 $\boldsymbol{v}_1$，鸢尾花数据

如图 14 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 15 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。

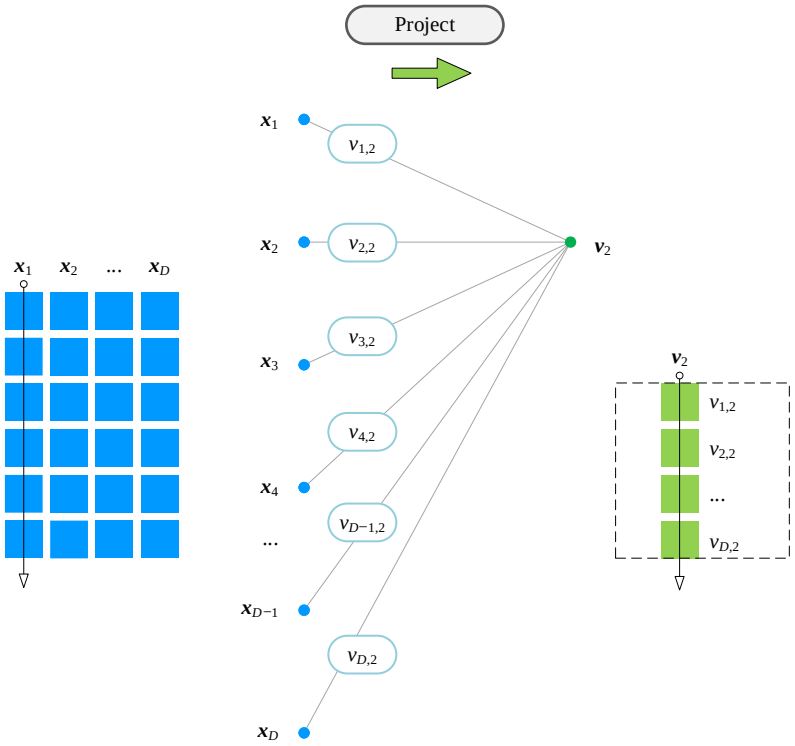


图 14. 构造第二主成分 $\boldsymbol{v}_2$

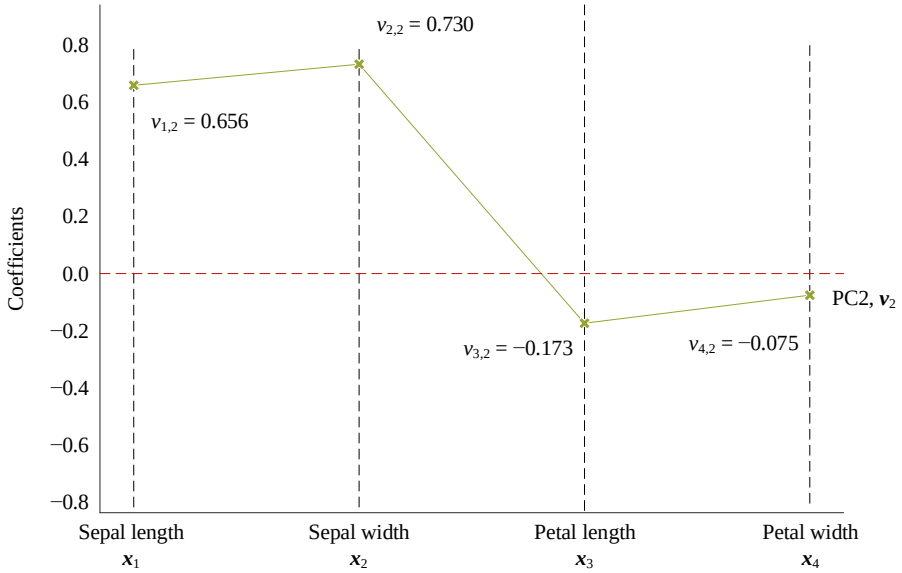


图 15. 构造第二主成分 $\boldsymbol{v}_2$，鸢尾花数据

如图 16 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 17 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。

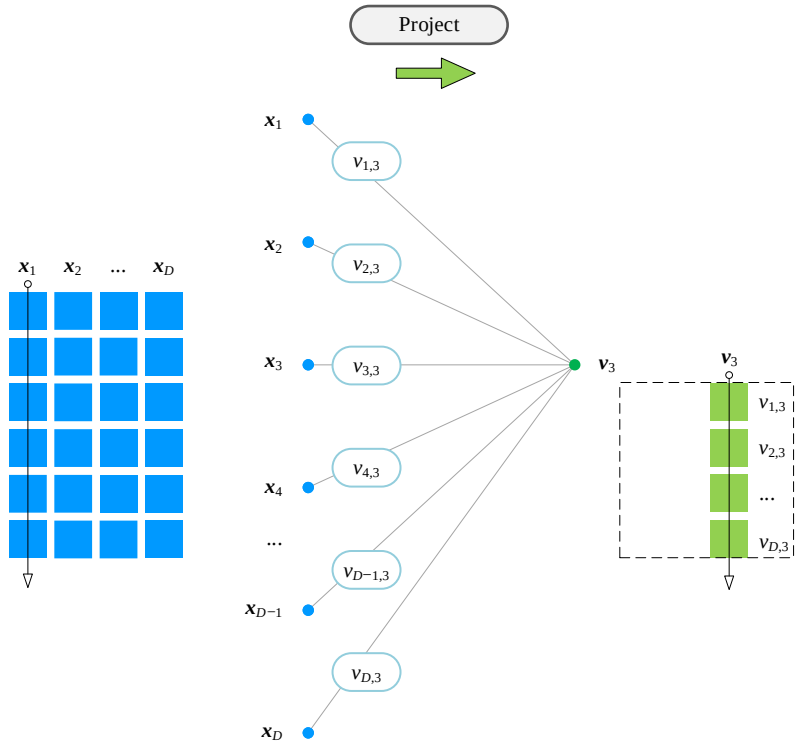


图 16. 构造第三主成分 $\boldsymbol{v}_3$

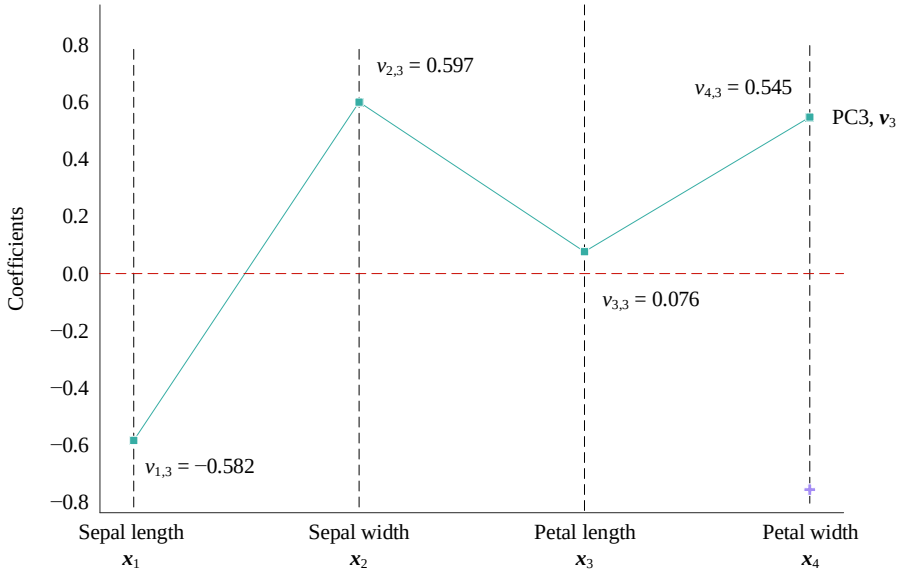


图 17. 构造第三主成分 $\boldsymbol{v}_3$，鸢尾花数据

14.5 投影结果

图 18 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。

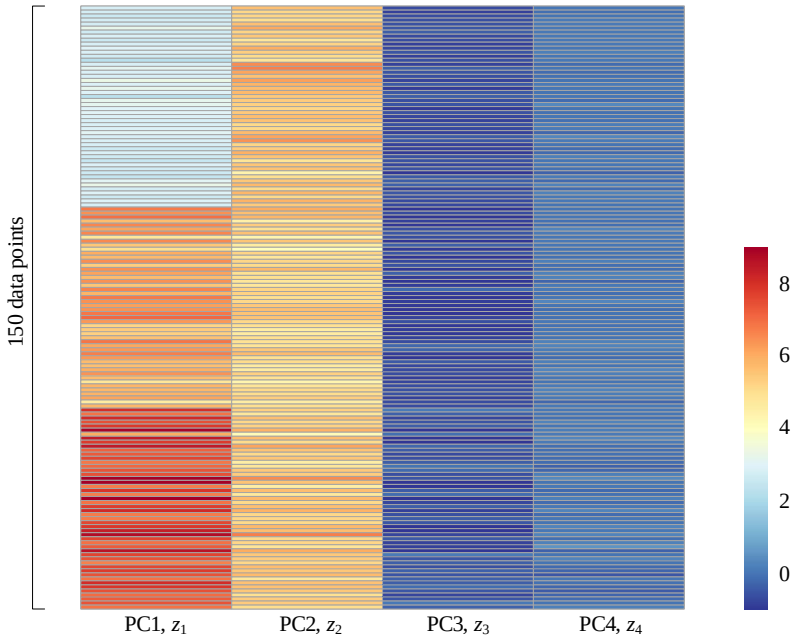


图 18. 新特征数据矩阵 $\boldsymbol{Z}$

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：

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

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

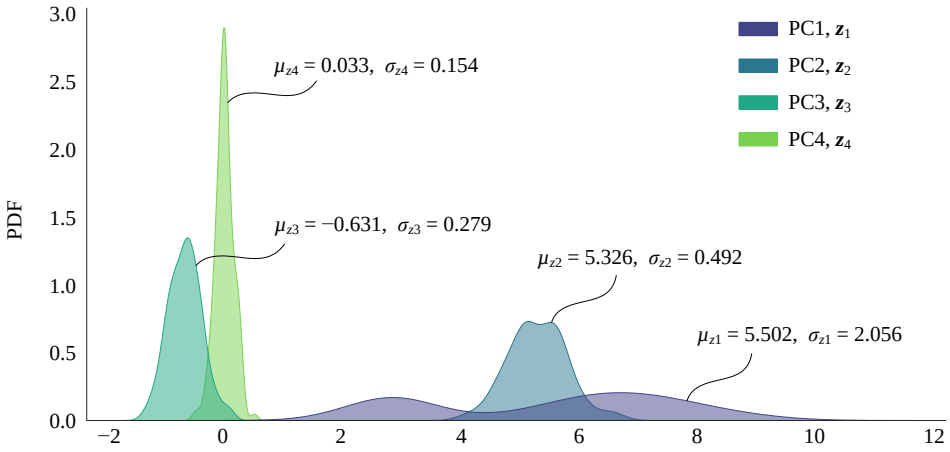


图 19. 转换数据 $\boldsymbol{Z}$ 四个特征上分布，KDE 估计

作为对比，图 20 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 19 和图 20，我们可以发现方差没有变化。唯一的区别是，图 20 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。

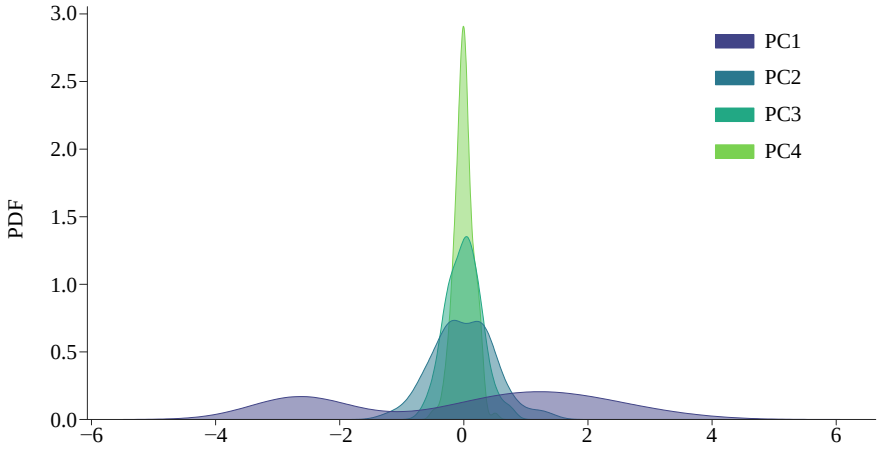


图 20. 转换数据 $\boldsymbol{Z}$ 四个特征上分布，KDE 估计；数据已经中心化

图 21 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 22 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 23 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 21 所示为转换数据 $\boldsymbol{Z}$ 协方差矩阵和相关性系数矩阵热图。

有关协方差运算，请大家回顾《统计至简》第 14 章。

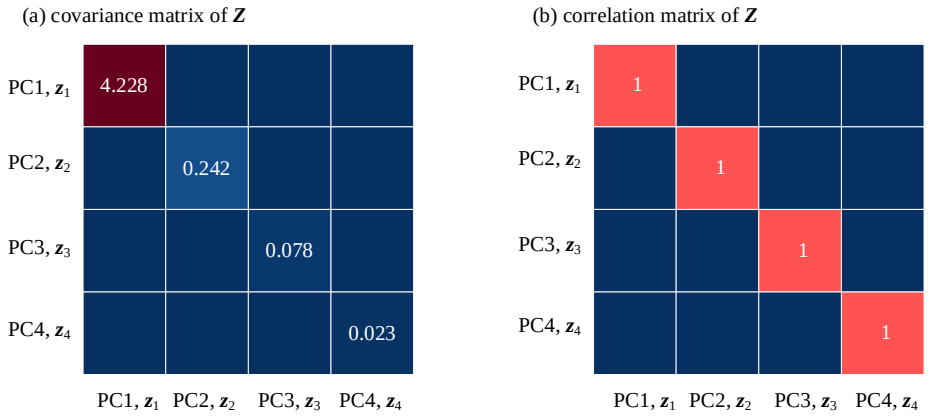


图 21. 转换数据 $\boldsymbol{Z}$ 协方差矩阵和相关性系数矩阵热图

图 22 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 23 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。

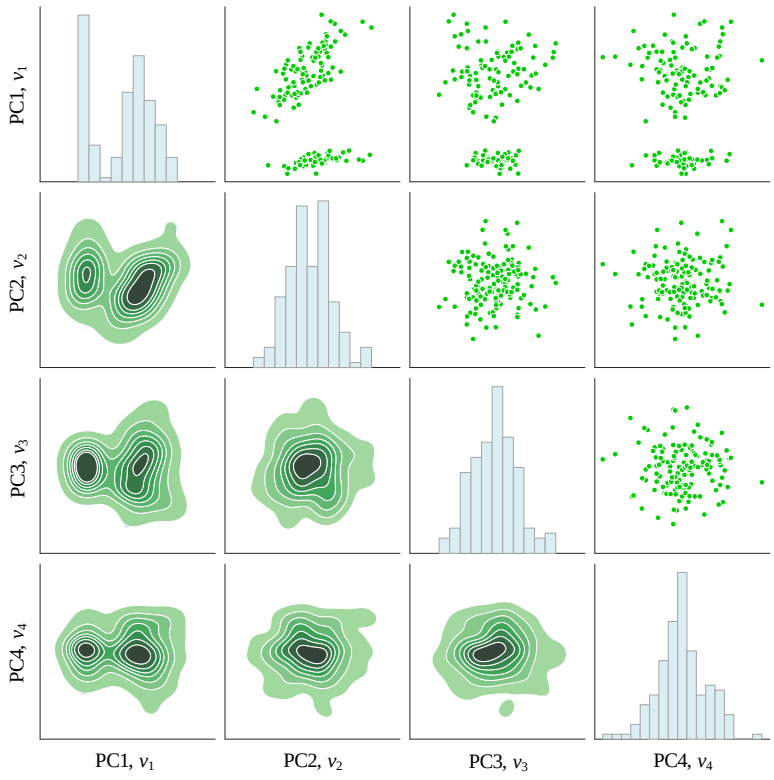


图 22. 转换数据 $\boldsymbol{Z}$ 成对特征分析图，不分类

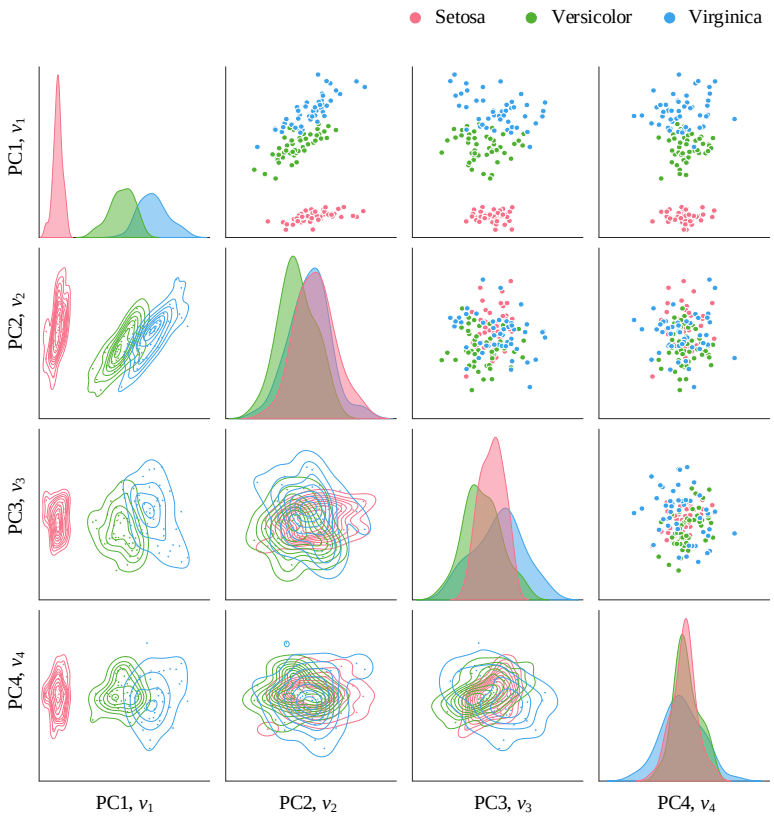


图 23. 转换数据 $\boldsymbol{Z}$ 成对特征分析图，分类

14.6 还原

主成分 $\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$ 上的投影结果可以用来还原部分原始数据。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：

$$\boldsymbol{E} = \boldsymbol{X} - \hat{\boldsymbol{X}} \tag{18}$$

图 24 所示为 $\boldsymbol{z}_1$ 还原 $\boldsymbol{X}$ 部分数据。图 25 所示为 $\boldsymbol{z}_2$ 还原 $\boldsymbol{X}$ 部分数据。图 26 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 26 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 27 所示。

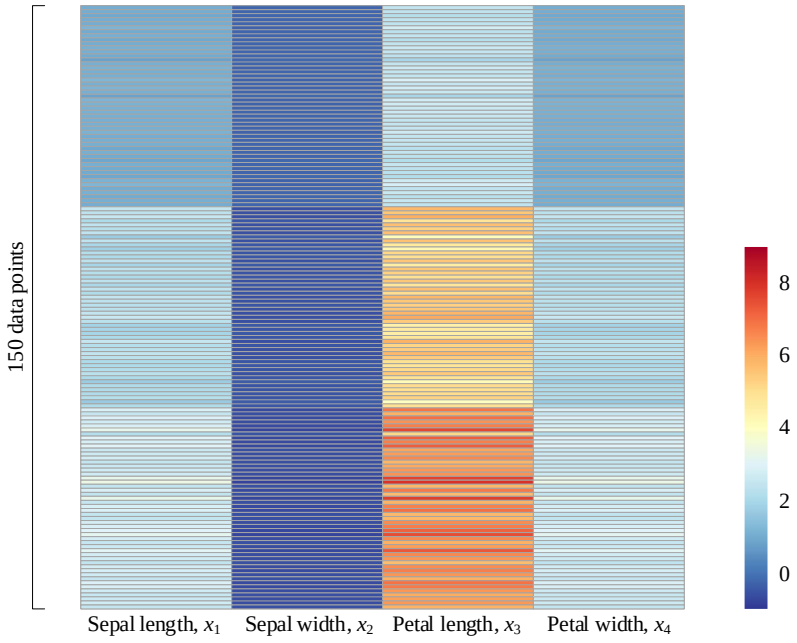


图 24. $\boldsymbol{z}_1$ 还原 $\boldsymbol{X}$ 部分数据

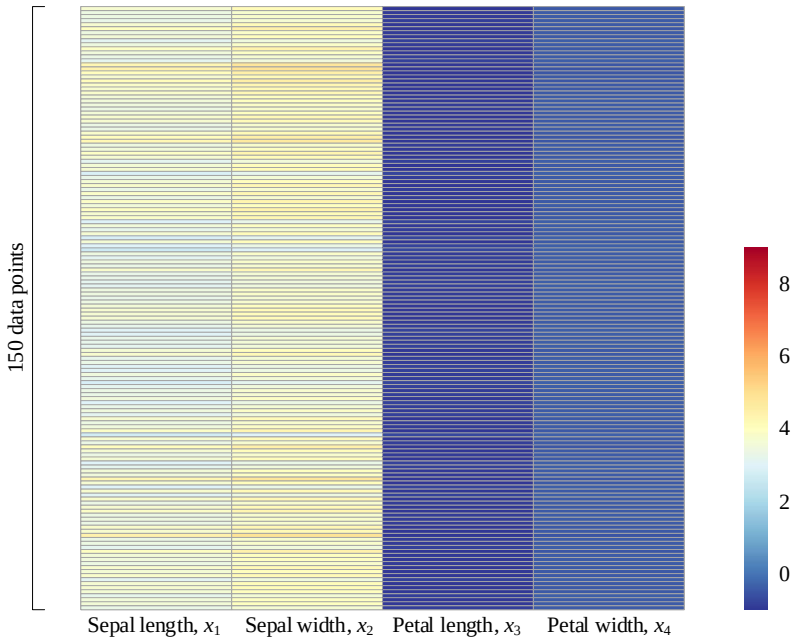


图 25. $\boldsymbol{z}_2$ 还原 $\boldsymbol{X}$ 部分数据

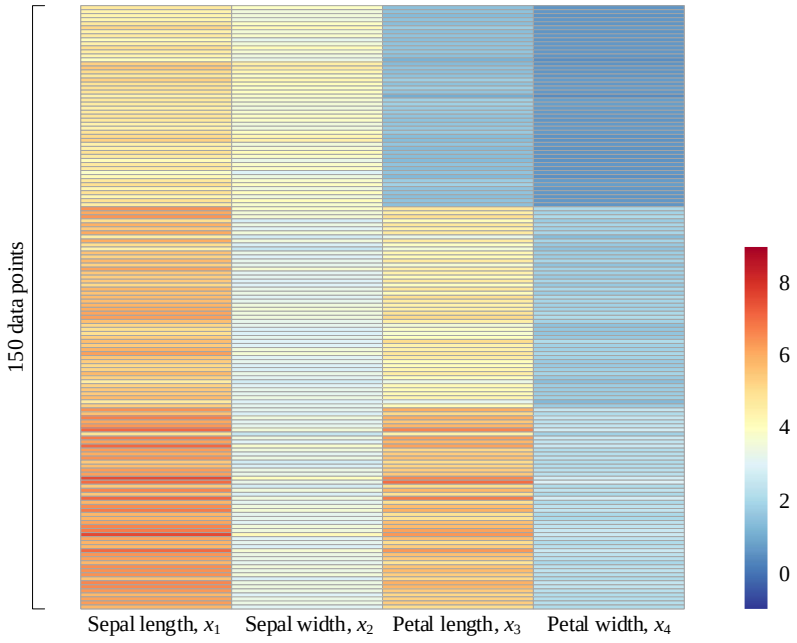


图 26.[$\boldsymbol{z}_1$, $\boldsymbol{z}_2$] 还原 $\boldsymbol{X}$ 部分数据

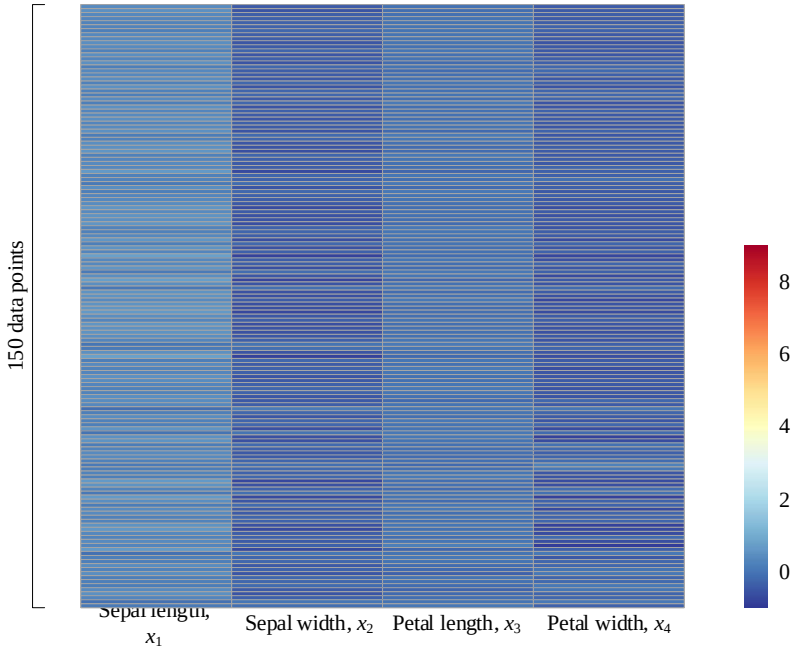


图 27. 误差 $\boldsymbol{E}$

14.7 双标图

双标图 (biplot) 是主成分分析中常用的可视化方案。它能够将高维数据投影到二维或三维空间中，并用散点图的形式展示出来，同时还能够显示原始数据和主成分的信息。一般情况，平面双标图的横坐标和纵坐标分别表示 PCA 的前两个主成分，每个点代表一个样本数据。

通过观察双标图，可以发现不同样本之间的相似性和差异性。如果两个点在双标图上非常接近，那么它们在原始数据中的特征值也可能非常接近，反之亦然。同时，双标图还能够帮助我们找出数据中的异常值和离群点，这些点在双标图上往往会距离其他点较远。

除了用于可视化，双标图还能够用来评估 PCA 的效果。如果双标图中的数据点分布较为均匀且没有聚集在一起，那么说明 PCA 的效果较好，主成分能够较好地解释数据的方差；如果双标图中的数据点呈现出聚集或者明显的分块现象，那么说明 PCA 的效果可能不太理想，主成分并不能完全解释数据的方差。

如图 28 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。

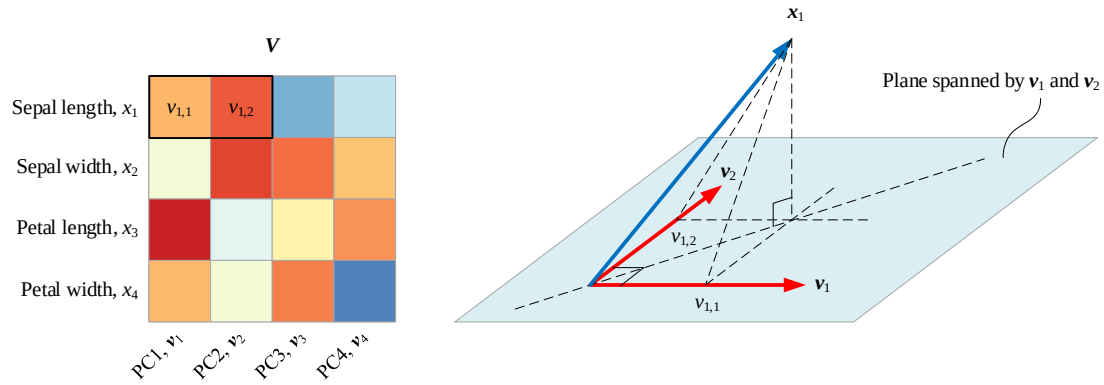


图 28. 双标图原理

图 29 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 29 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，花瓣长度 x_3、花瓣宽度 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。

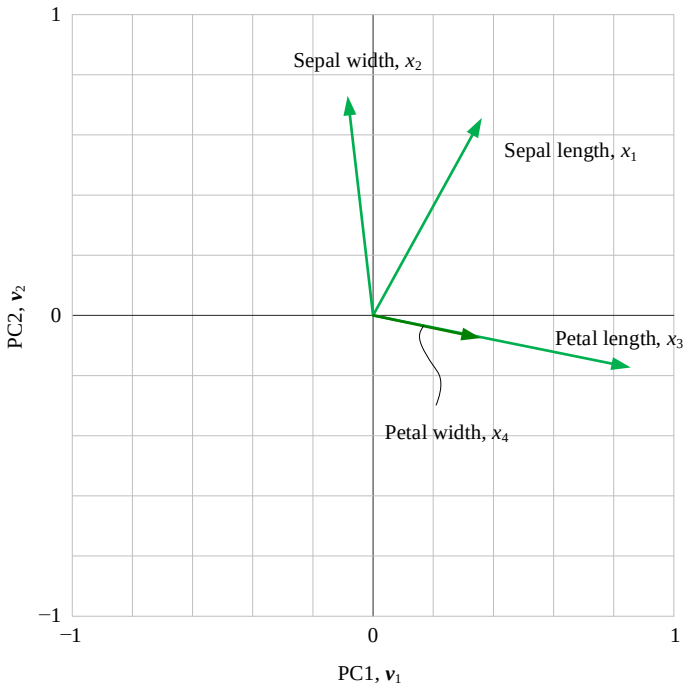


图 29. v_1-v_2 平面双标图，基于鸢尾花原始数据

图 30 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。

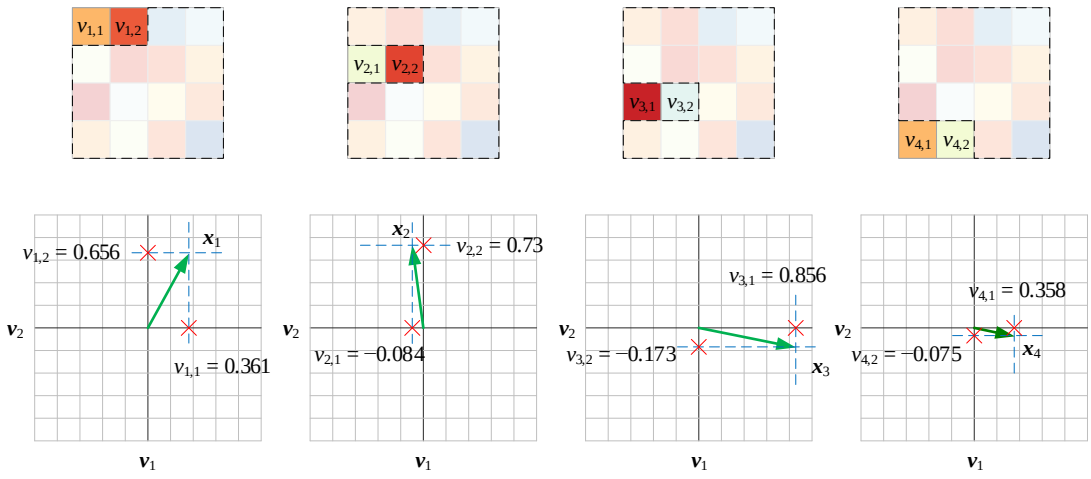


图 30. 向量 x_1、x_2、x_3 和 x_4 向 v_1-v_2 平面投影结果

图 31 所示为向量 x_1、x_2、x_3 和 x_4 向 v_3-v_4 平面投影结果。

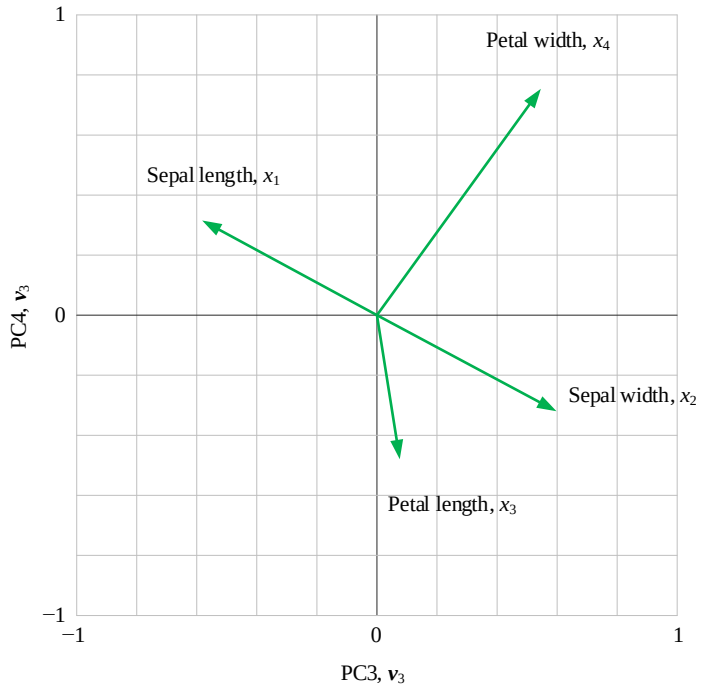


图 31. v_3-v_4 平面双标图，基于鸢尾花原始数据

双标图还可以基于标准化后数据；图 32 所示为基于鸢尾花标准化数据后的双标图，投影值对应 (10)。

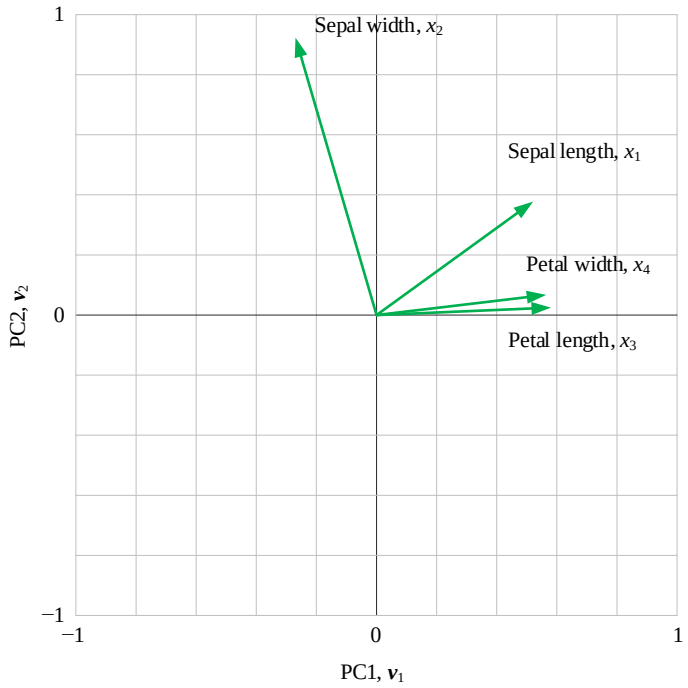


图 32. 平面双标图，基于鸢尾花标准化数据

此外，除了特征向量之外，双标图还会绘制数据点投影，如图 33 所示。图 33 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。

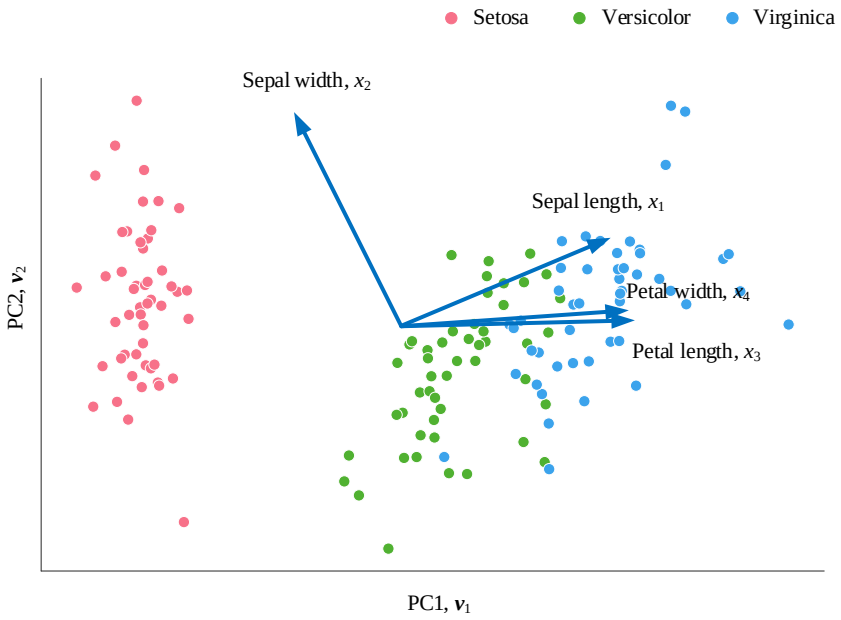


图 33. 平面双标图，标准化数据

图 34 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。

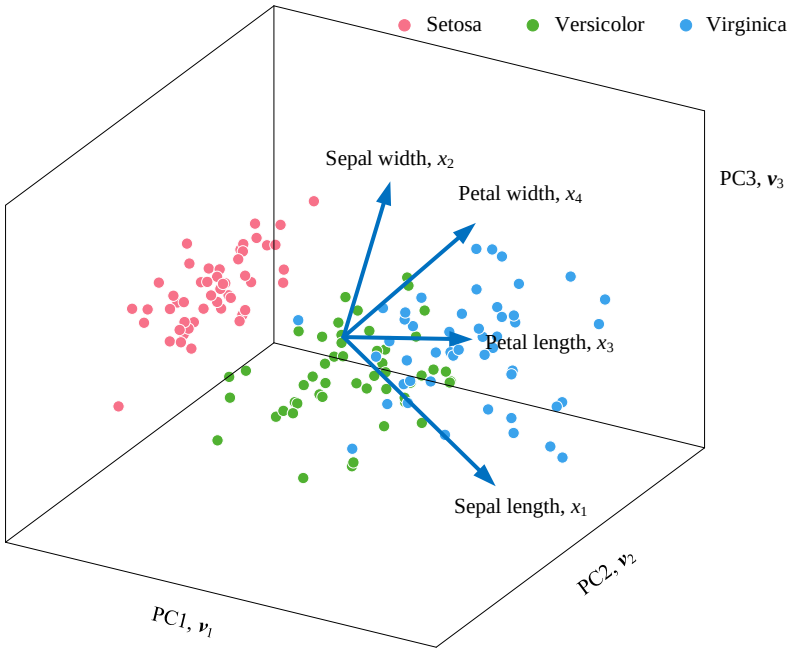


图 34. 三维双标图

14.8 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：

$$\frac{\lambda_j}{\sum_{i=1}^{D}\lambda_i}\times 100\% \tag{19}$$

上式分母是数据总方差。

协方差矩阵 $\boldsymbol{\Sigma}$ 的迹——方阵对角线元素之和——等于特征值之和，请大家回顾《统计至简》第 13 章。

(19) 这个比值可以用来衡量第 j 个主成分对数据的解释能力。如果已释方差较大，那么说明第 j 个主成分能够较好地解释数据的方差，即它包含了较多的信息。如果已释方差较小，那么说明第 k 个主成分对数据的解释能力较弱，不足以对数据进行有效的降维和特征提取。

前 p 个特征值累积解释总方差的百分比为：

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

这个比值代表前 p 个主成分所能解释的已释方差之和占所有主成分已释方差之和的比例。累计已释方差和百分比能够用来评估 PCA 的降维效果，它衡量了前 p 个主成分能够解释数据方差的比例。

通常来说，我们希望通过选择适当的主成分数 p，使累计已释方差和百分比达到预设的阈值 (比如 80%或 90%)，以保留尽可能多的原始数据信息。通过观察累计已释方差和百分比的变化趋势，我们可以得出选择适当主成分数的建议，以及对 PCA 的降维效果进行评估和比较。

图 35 给出图像可视化 (19) 和 (20) 。鸢尾花数据的主成分分析特征值如下：

$$\lambda_1=4.228,\quad \lambda_2=0.242,\quad \lambda_3=0.078,\quad \lambda_4=0.023 \tag{21}$$

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 35 所示，第一和第二主成分两者已释方差之和为 72.5%。

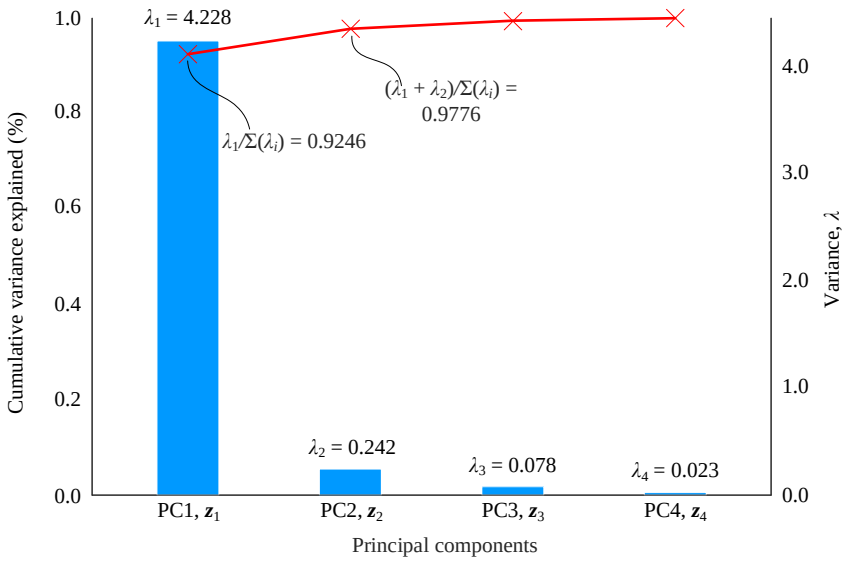


图 35. 陡坡图

Bk7_Ch14_01.ipynb 绘制本章前文大部分图片。

读过《编程不难》的读者对图 36 这个 App 都应该不陌生。这个 App 展示主元数量对数据还原的影响，图中的数据为不同期限的利率数据。这个 App 中用来完成主成分分析的函数来自 Statsmodels。

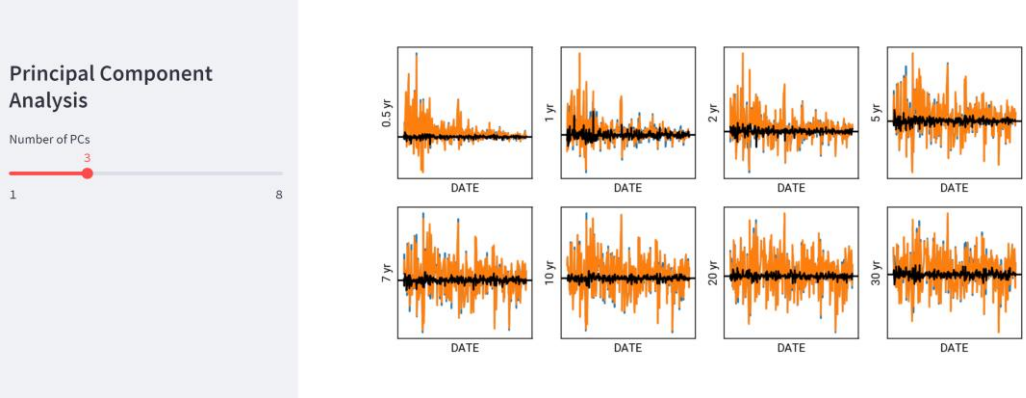


图 36. 展示主元数量对数据还原影响的 App | Streamlit_Bk7_Ch14_02.py

主成分分析是一种广泛使用的数据降维和特征提取技术，它可以将高维数据降至低维，同时保留数据的主要特征和结构。PCA 通过寻找一组最能解释数据变异性的线性组合，即主成分，来实现数据降维和特征提取。主成分是原始特征的线性组合，它们的排序代表了它们的重要性。通常，我们只需要保留前几个主成分，因为它们可以解释大部分数据的变异性。

一般的 PCA 步骤包括：中心化 (标准化) 数据、计算协方差矩阵、计算特征值和特征向量、排序特征值和对应的特征向量、选择前 p 个主成分、计算投影矩阵并对数据进行降维。在计算特征值和特征向量时，我们通常使用特征值分解，当然也可以使用奇异值分解，这是下一章要介绍的内容。

PCA 的投影可以帮助我们理解数据的结构和关系。投影到第一二主成分方向上的投影数据通常成椭圆形状，其中椭圆的长轴方向表示最大的方差方向，短轴方向表示最小的方差方向。通过线性组合，我们可以将主成分重新组合成原始数据，并通过双标图和陡坡图来分析 PCA 的效果。双标图可以帮助我们了解主成分之间的相关性，陡坡图可以帮助我们了解主成分的贡献程度。

在 PCA 中，理解数据和分析结果的视角非常重要。这涉及到如何选择主成分和如何解释它们，以及如何应用 PCA 的结果。选择主成分时，我们通常考虑主成分的贡献程度和解释能力，以及降维后的数据能否保留足够的信息。解释主成分时，我们需要考虑主成分的物理意义和应用背景。应用 PCA 的结果时，我们可以利用降维后的数据进行可视化、聚类、分类等分析。

总之，主成分分析是一种强大的数据降维和特征提取技术，它可以帮助我们更好地理解和分析数据。在应用 PCA 时，需要注意数据预处理、主成分选择和解释、以及降维后的数据应用等问题。本书第 15 章将介绍用截断奇异值分解完成 PCA，而第 16 章将比较六种不同的 PCA 技术路线。