

16

Dive into Principal Component Analysis

主成分分析进阶

区分六条基本 PCA 技术路线



我发现了！

Eureka!

—— 阿基米德 (Archimedes) | 数学家、发明家、物理学家 | 287 ~ 212 BC



- ▶ `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
- ▶ `seaborn.heatmap()` 绘制热图
- ▶ `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

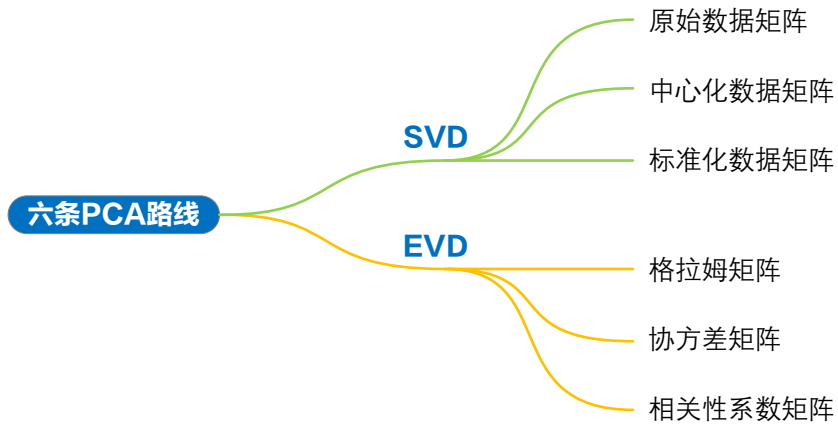
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com



本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

16.1 从“六条技术路线”说起

来自《矩阵力量》的表格

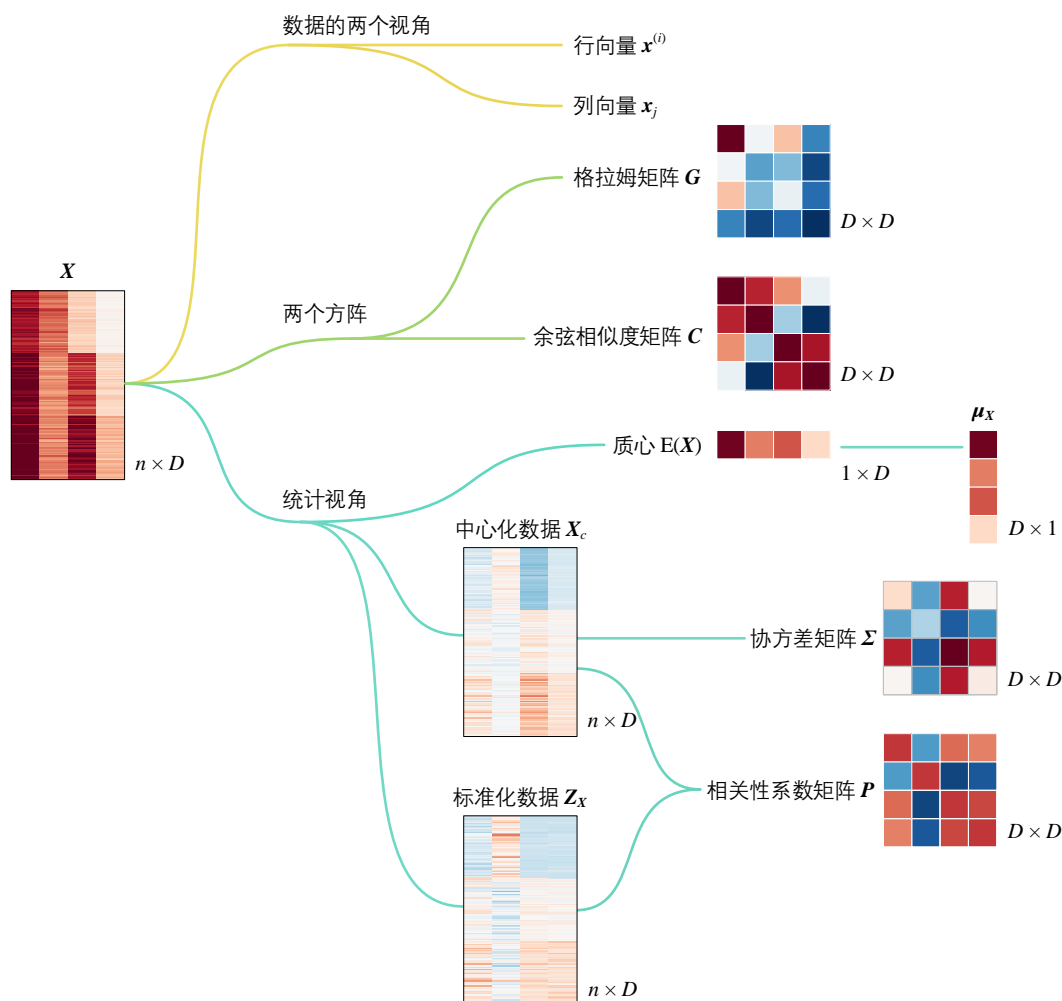
表 1 来自《矩阵力量》第 25 章，本章将讲解表 1 中六条 PCA 技术路线的细节，并比较它们的差异。

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 X	奇异值分解	$X = U_X S_X V_X^T$
格拉姆矩阵 $G = X^T X$ 本章中用“修正”的格拉姆矩阵 $G = \frac{X^T X}{n-1}$	特征值分解	$G = V_X A_X V_X^T$
中心化数据矩阵 $X_c = X - E(X)$	奇异值分解	$X_c = U_c S_c V_c^T$
协方差矩阵 $\Sigma = \frac{(X - E(X))^T (X - E(X))}{n-1}$	特征值分解	$\Sigma = V_c A_c V_c^T$
标准化数据 (z 分数) $Z_X = (X - E(X)) D^{-1}$ $D = \text{diag}(\text{diag}(\Sigma))^{\frac{1}{2}}$	奇异值分解	$Z_X = U_Z S_Z V_Z^T$
相关性系数矩阵 $P = D^{-1} \Sigma D^{-1}$ $D = \text{diag}(\text{diag}(\Sigma))^{\frac{1}{2}}$	特征值分解	$P = V_Z A_Z V_Z^T$

比较六个输入矩阵

表 1 中有六个输入矩阵，它们都衍生自原始数据矩阵 X 。如图 1 所示，原始数据矩阵 X 的形状为 $n \times D$ 。

图 1. \mathbf{X} 衍生得到的几个矩阵，来自《矩阵力量》

\mathbf{X} 的格拉姆矩阵 \mathbf{G} 为：

$$\mathbf{G} = \mathbf{X}^T \mathbf{X} \quad (1)$$

格拉姆矩阵 \mathbf{G} 形状为 $D \times D$ 。 \mathbf{G} 的主对角线元素是 \mathbf{X} 的每一列向量 L^2 模的平方。

中心化 (去均值) 矩阵 \mathbf{X}_c 为：

$$\mathbf{X}_c = \mathbf{X} - \mathbf{E}(\mathbf{X}) \quad (2)$$

即 \mathbf{X} 的每一列分别减去各自的均值得到 \mathbf{X}_c 。几何角度， \mathbf{X} 的质心位于 $\mathbf{E}(\mathbf{X})$ ， \mathbf{X}_c 的质心则位于原点 $\mathbf{0}$ 。

样本数据矩阵 \mathbf{X} 的协方差矩阵 $\mathbf{\Sigma}$ 为：

$$\mathbf{\Sigma} = \frac{\mathbf{X}_c^T \mathbf{X}_c}{n-1} = \frac{(\mathbf{X} - \mathbf{E}(\mathbf{X}))^T (\mathbf{X} - \mathbf{E}(\mathbf{X}))}{n-1} \quad (3)$$

容易发现，协方差相当于特殊的格拉姆矩阵。

请大家特别注意，为了方便和协方差比较，本章中 G 特别定义为：

$$G = \frac{X^T X}{n-1} \quad (4)$$

标准化 (standardization 或 z-score normalization) 数据矩阵 Z_X 为：

$$Z_X = (X - E(X))D^{-1} \quad (5)$$

其中 D 为：

$$D = \text{diag}(\text{diag}(\Sigma))^{\frac{1}{2}} = \begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & \ddots \\ & & & \sigma_D \end{bmatrix} \quad (6)$$

(5) 中的每一列都是每个特征的 Z 分数。 Z_X 的质心也位于原点，不同的是 Z_X 每个特征的标准差都是 1。

线性相关性系数矩阵 P 为：

$$P = D^{-1} \Sigma D^{-1} \quad (7)$$

P 实际上是 Z_X 的协方差，即：

$$P = \frac{Z_X^T Z_X}{n-1} \quad (8)$$

比较 SVD 和 EVD

主成分分析的核心数学工具为**奇异值分解** (Singular Value Decomposition, SVD) 和**特征值分解** (Eigen Decomposition, EVD)。



《矩阵力量》强调过 SVD 和 EVD 在主成分分析中具有等价性，这也就是为什么表 1 看上去是六种技术路线，实际上可以归纳为三大类技术路线。下面简单说明一下。

对原始矩阵 X 进行经济型 SVD 分解：

$$X = U_X S_X V_X^T \quad (9)$$

其中， S_X 为对角方阵。

将 (9) 代入 (1)：

$$G = V_X S_X^{-2} V_X^T \quad (10)$$

上式便是格拉姆 G 的特征值分解。

对中心化数据矩阵 \mathbf{X}_c 经济型 SVD 分解：

$$\mathbf{X}_c = \mathbf{U}_c \mathbf{S}_c \mathbf{V}_c^T \quad (11)$$

而协方差矩阵 $\mathbf{\Sigma}$ 则可以写成：

$$\mathbf{\Sigma} = \mathbf{V}_c \frac{\mathbf{S}_c^2}{n-1} \mathbf{V}_c^T \quad (12)$$

相信大家在上式中能够看到协方差矩阵 $\mathbf{\Sigma}$ 的特征值分解。请大家注意 (11) 中奇异值和 (12) 中特征值关系：

$$\lambda_{c-j} = \frac{s_{c-j}^2}{n-1} \quad (13)$$

同样，对标准化数据矩阵 \mathbf{Z}_x 进行经济型 SVD 分解：

$$\mathbf{Z}_x = \mathbf{U}_z \mathbf{S}_z \mathbf{V}_z^T \quad (14)$$

相关性系数矩阵 \mathbf{P} 则可以写成：

$$\mathbf{P} = \mathbf{V}_z \frac{\mathbf{S}_z^2}{n-1} \mathbf{V}_z^T \quad (15)$$

上式相当于对 \mathbf{P} 特征值分解。

本章下面将分别讲解特征值分解 1) 协方差矩阵、2) 格拉姆矩阵、3) 相关性系数矩阵，来完成主成分分析。并利用诸如热图、饼图、直方图、陡坡图、双标图、椭圆等可视化工具分析三种路线。本章以下三节将采用完全相似的结构，方便大家比较三大类不同 PCA 技术路线的异同。

16.2 协方差矩阵

本节讲解利用特征值分解协方差矩阵 $\mathbf{\Sigma}$ 完成主成分分析。

特征值分解

图 2 所示为特征值分解协方差矩阵 $\mathbf{\Sigma}$ 。 $\mathbf{\Sigma}$ 的对角线元素为方差，其他元素为协方差。 $\mathbf{\Sigma}$ 的迹代表方差之和：

$$\text{trace}(\mathbf{\Sigma}) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_D^2 = \sum_{j=1}^D \sigma_j^2 \quad (16)$$

图 2 中 $\mathbf{\Sigma}$ 为对称矩阵，因此对 $\mathbf{\Sigma}$ 的特征值分解实际上是谱分解。

Λ_c 为对角矩阵，对角线元素为特征值，特征值从大到小排列。 X_c 投影到规范正交基 V_c 中得到 Y_c ，即 $Y_c = X_c V_c$ 。 Λ_c 主对角线上的特征值实际上是 Y_c 的方差，也就是说 Λ_c 是 Y_c 的协方差矩阵。因此，在主成分分析中，特征值也叫主成分方差。

Λ_c 的方差，即特征值，之和为：

$$\text{trace}(\Lambda_c) = \lambda_1 + \lambda_2 + \cdots + \lambda_D = \sum_{j=1}^D \lambda_j$$

(17)

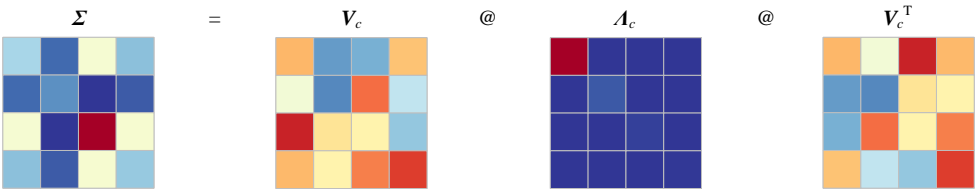


图 2. 特征值分解协方差矩阵 Σ

图 16 对比格拉姆矩阵 G 和 Λ_x 。

下面，我们进一步分析这两个矩阵。

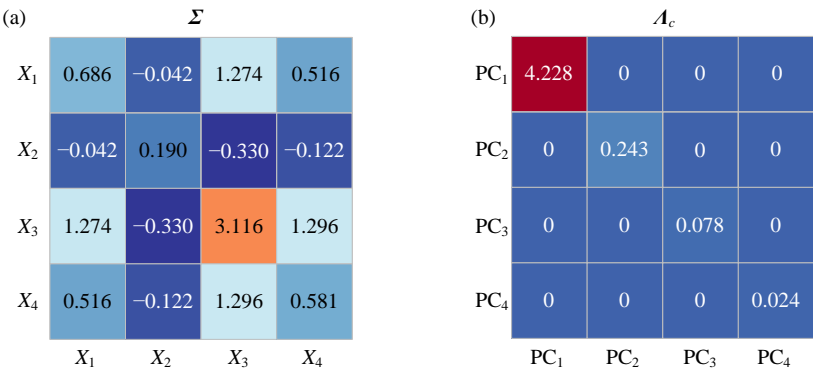
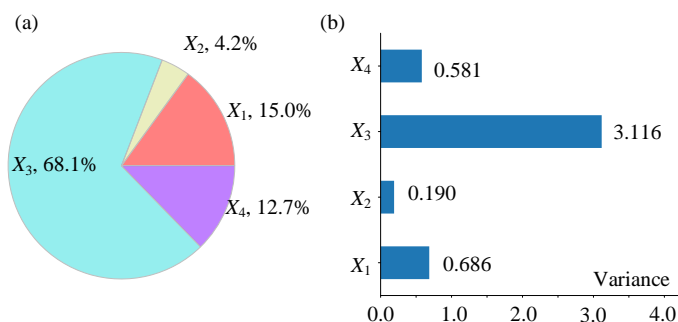


图 3. 对比协方差矩阵 Σ 和 Λ_c 热图

分解前后

大家在本书第 12 章已经见过图 4 和图 5。

如图 4 所示，数据矩阵 X 中第三列，即 X_3 ，的方差最大， X_3 对方差和 $\text{trace}(\Sigma)$ 贡献超过 68%。

图 4. 协方差矩阵 Σ 的主对角线成分，即方差

我们在《矩阵力量》第 13 章提过，特征值分解前后矩阵的迹不变，也就是说协方差矩阵 Σ 的迹 $\text{trace}(\Sigma)$ 等于的特征值方阵 Λ_c 迹 $\text{trace}(\Lambda_c)$ ：

$$\text{trace}(\Sigma) = \text{trace}(\Lambda_c) \quad (18)$$

即：

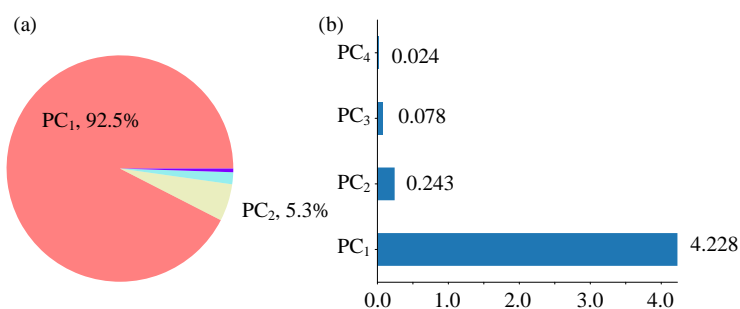
$$\sum_{j=1}^D \sigma_j^2 = \sum_{j=1}^D \lambda_j \quad (19)$$

也就是说，PCA 不改变数据各个特征方差总和。

而第 j 个特征值 λ_j 对 $\text{trace}(\Lambda_c)$ 的贡献百分比为：

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

如图 5 所示，第一主成分的贡献超过 92%，解释了数据中大部分“方差”。数据分析中，如果原始数据特征很多，彼此之间又具有复杂的相关性，那么我们就可以考虑利用主成分分析对数据进行“降维”，减少特征的数量。而这个过程又保留了原始数据主要的信息。

图 5. Λ_c 的主对角线成分，协方差矩阵 Σ 的特征值

陡坡图

上一章介绍过我们经常用陡坡图可视化前 p 个主成分解释总方差的百分比，即累积贡献率：

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

图 6 所示为特征值分解协方差矩阵 Σ 获得的陡坡图。观察陡坡图，可以帮助我们确定选取多少个主成分。

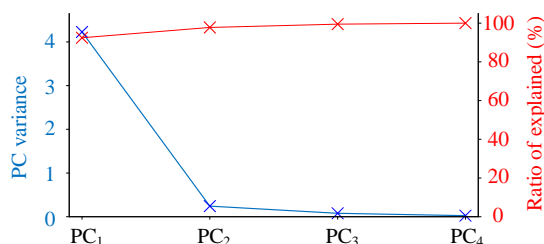


图 6. 陡坡图，特征值分解协方差矩阵 Σ

特征向量矩阵

图 7 所示为特征向量矩阵 V_c 热图。 V_c 的每一列便代表一个主成分的方向，即 $V_c = [v_{c_1}, v_{c_2}, v_{c_3}, v_{c_4}]$ 从左到右分别是第一、二、三、四主成分。这些主成分方向两两正交。

在主成分分析中， V_c 叫主成分系数，也称为载荷 (loading)。注意，有一些参考文献中，载荷还要乘上特征值的平方根，即 $v_j \sqrt{\lambda_j}$ 。

V_c 也可以通过经济型 SVD 分解中心化矩阵 X_c 得到。

V_c			
v_{c_1}	v_{c_2}	v_{c_3}	v_{c_4}
0.36	-0.66	-0.58	0.32
-0.085	-0.73	0.6	-0.32
0.86	0.17	0.076	-0.48
0.36	0.075	0.55	0.75
PC ₁	PC ₂	PC ₃	PC ₄

图 7. 特征向量矩阵 V_c 热图

投影

由于 V_c 为正交矩阵，满足 $V_c^T V_c = V_c V_c^T = I$ ，因此 V_c 本身也是规范正交基。如图 8 所示，将中心化矩阵 X_c 投影到 V_c 这个规范正交基中得到数据矩阵 Y_c ，即 $Y_c = X_c V_c$ 。通过图 8 中的 Y_c 每一列的色差，我们就可以看出来不同的次序主成分对数据总体方差的解释力度。



《矩阵力量》第 18 章介绍过 SVD 分解的优化视角。

利用 L^2 范数， V_c 的第一列列向量实际上是如下优化问题的解：

$$\begin{aligned} \mathbf{v}_{c_1} = \arg \max_{\mathbf{v}} \quad & \|\mathbf{X}_c \mathbf{v}\| \\ \text{subject to: } & \|\mathbf{v}\| = 1 \end{aligned} \quad (22)$$

前文提过， \mathbf{A}_X 本身是 \mathbf{Y}_c 的协方差矩阵。 \mathbf{A}_X 为对角方阵，因此 \mathbf{Y}_c 的任意两列之间线性相关系数为 0。也就是说， V_c 完成了 X_c 的正交化，注意不是原始数据矩阵 X 的正交化。

请大家思考 \mathbf{Y}_c 的每一列的均值是多少？ \mathbf{Y}_c 的质心位置是什么？为什么？

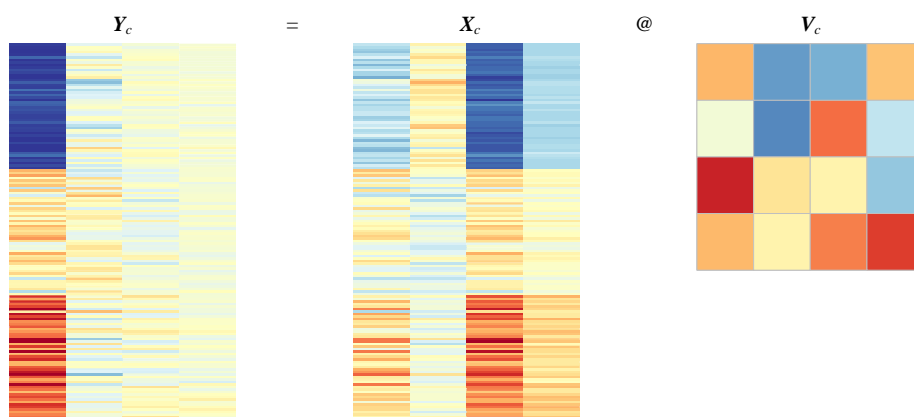


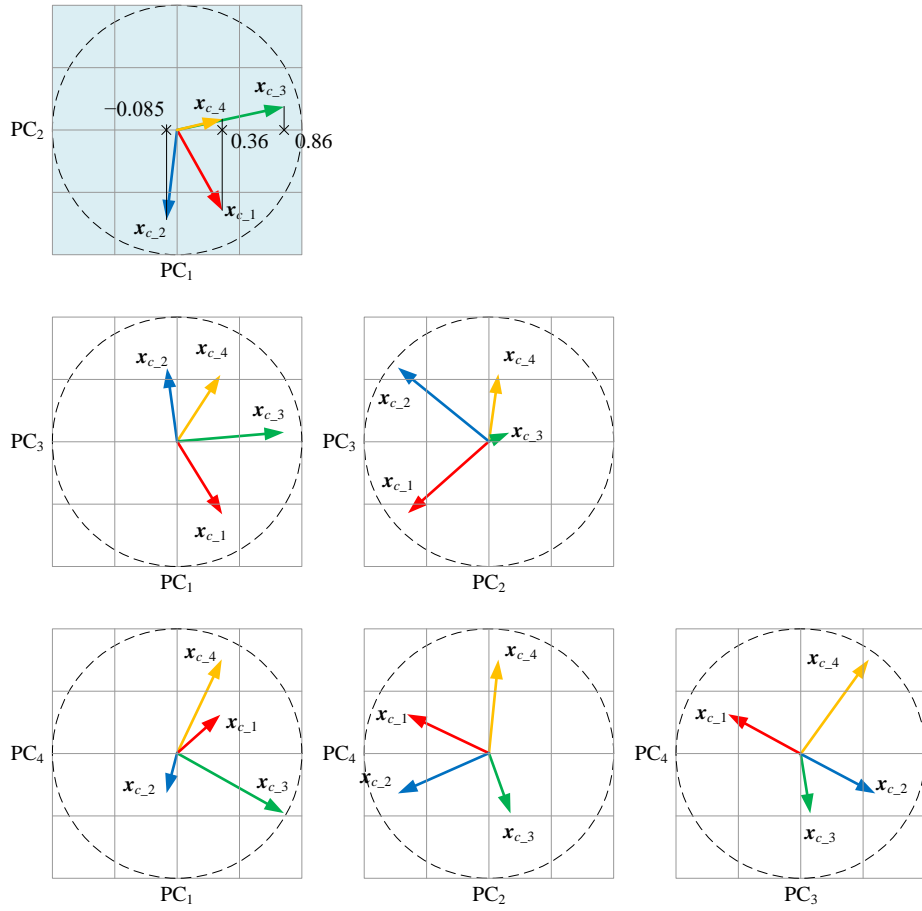
图 8. 将中心化数据 X_c 投影到 V_c

双标图

如图 9 所示，双标图是可视化特征向量矩阵 V_c 的重要方法。

以图 9 中蓝色背景的双标图为例，中心化数据 X_c 投影到第一、二主成分平面内的结果如四个箭头所示。比如， X_1 、 X_2 、 X_3 、 X_4 在 PC1 上贡献的分量分别为 0.36、-0.085、0.86、0.36，这正是如图 7 所示的 V_c 第一列 \mathbf{v}_{c_1} 。

我们还可以把投影数据的散点图也画在双标图上，大家已经在上一章看到很多例子，本章不再重复。

图 9. V_c 双标图，特征值分解协方差矩阵 Σ

数据还原、误差

将 (11) 展开写成：

$$\begin{aligned}
 \mathbf{X}_c &= \begin{bmatrix} \mathbf{u}_{c-1} & \mathbf{u}_{c-2} & \cdots & \mathbf{u}_{c-D} \end{bmatrix} \begin{bmatrix} s_{c-1} & & & \\ & s_{c-2} & & \\ & & \ddots & \\ & & & s_{c-D} \end{bmatrix} \begin{bmatrix} \mathbf{v}_{c-1}^T \\ \mathbf{v}_{c-2}^T \\ \vdots \\ \mathbf{v}_{c-D}^T \end{bmatrix} \\
 &= s_{c-1} \mathbf{u}_{c-1} \mathbf{v}_{c-1}^T + s_{c-2} \mathbf{u}_{c-2} \mathbf{v}_{c-2}^T + \cdots + s_{c-D} \mathbf{u}_{c-D} \mathbf{v}_{c-D}^T = \sum_{j=1}^D s_{c-j} \mathbf{u}_{c-j} \mathbf{v}_{c-j}^T
 \end{aligned} \tag{23}$$

图 10 所示为用第一主成分逼近估计 \mathbf{X}_c ，即：

$$\hat{\mathbf{X}}_c = s_{c-1} \underbrace{\mathbf{u}_{c-1} \mathbf{v}_{c-1}^T}_{\text{First principal}} \tag{24}$$

图中可以看到， $\hat{\mathbf{X}}_c$ 和 \mathbf{X}_c 非常相似；虽然 $\hat{\mathbf{X}}_c$ 是个 150×4 矩阵， $\hat{\mathbf{X}}_c$ 的秩还是 1。请大家回顾如何用张量积计算 $\hat{\mathbf{X}}_c$ 。图 10 中的 \mathbf{E} 为误差，即 $\mathbf{E} = \mathbf{X}_c - \hat{\mathbf{X}}_c$ 。

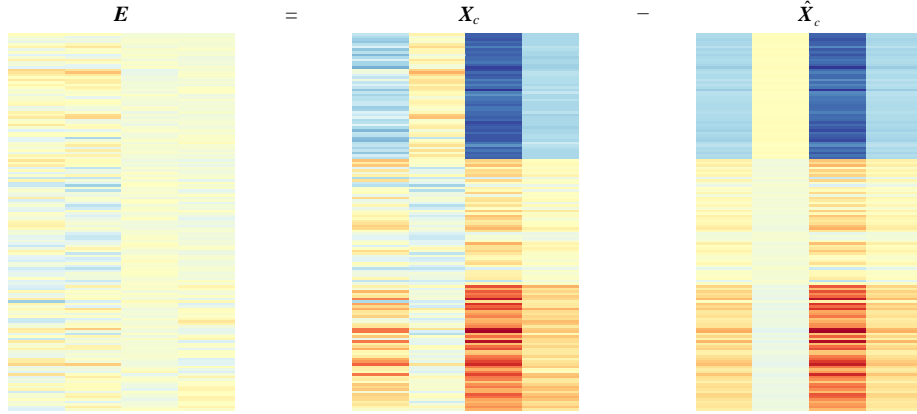


图 10. 第一主成分估计 \mathbf{X}_c

要想还原原始数据 \mathbf{X} ，我们还需要考虑 (2) 这个等式关系，即：

$$\mathbf{X} = \mathbf{X}_c + \mathbf{E}(\mathbf{X}) = \sum_{j=1}^D s_{c-j} \mathbf{u}_{c-j} \mathbf{v}_{c-j}^T + \mathbf{E}(\mathbf{X}) \quad (25)$$

如果利用第一主成分估计原始数据矩阵 \mathbf{X} 的话，可以利用：

$$\mathbf{X} \approx s_{c-1} \mathbf{u}_{c-1} \mathbf{v}_{c-1}^T + \mathbf{E}(\mathbf{X}) \quad (26)$$

上式中， $\mathbf{E}(\mathbf{X})$ 为行向量，计算用到了广播原则。

大家可能会问，图 2 中特征值分解仅仅获得了 \mathbf{V}_c ，没有 \mathbf{U}_c 。难道我们还需要再对 \mathbf{X}_c 做 SVD 分解？答案是不需要。

《矩阵力量》第 10 章介绍过“二次投影”，也就是说 \mathbf{X}_c 可以写成：

$$\mathbf{X}_c = \mathbf{X}_c \mathbf{I} = \mathbf{X}_c \mathbf{V}_c \mathbf{V}_c^T \quad (27)$$

将 \mathbf{V}_c 展开，上式可以写成：

$$\begin{aligned} \mathbf{X}_c &= \mathbf{X}_c \underbrace{\begin{bmatrix} \mathbf{v}_{c-1} & \mathbf{v}_{c-2} & \cdots & \mathbf{v}_{c-D} \end{bmatrix}}_{\mathbf{V}_c} \underbrace{\begin{bmatrix} \mathbf{v}_{c-1}^T \\ \mathbf{v}_{c-2}^T \\ \vdots \\ \mathbf{v}_{c-D}^T \end{bmatrix}}_{\mathbf{V}_c^T} \\ &= \mathbf{X}_c \mathbf{v}_{c-1} \mathbf{v}_{c-1}^T + \mathbf{X}_c \mathbf{v}_{c-2} \mathbf{v}_{c-2}^T + \cdots + \mathbf{X}_c \mathbf{v}_{c-D} \mathbf{v}_{c-D}^T = \mathbf{X}_c \sum_{j=1}^D \mathbf{v}_{c-j} \mathbf{v}_{c-j}^T \end{aligned} \quad (28)$$

所以，(24) 可以写成：

$$\hat{\mathbf{X}}_c = \mathbf{X}_c \mathbf{v}_{c-1} \mathbf{v}_{c-1}^T = \mathbf{X}_c \mathbf{v}_{c-1} \otimes \mathbf{v}_{c-1} \quad (29)$$

(26) 则可以写成：

$$X \approx X_c v_{c-1} \otimes v_{c-1} + E(X) \quad (30)$$

如果用第一、二主成分还原 X ，上式需要再加一项：

$$X \approx \underbrace{X_c v_{c-1} \otimes v_{c-1}}_{\text{First principal}} + \underbrace{X_c v_{c-2} \otimes v_{c-2}}_{\text{Second principal}} + \underbrace{E(X)}_{\text{Centroid}} \quad (31)$$

鸢尾花书在不同位置反复强调数据单位，也就是量纲。如果原始数据的每列数据的量纲不一致，比如高度、质量、时间、温度、密度、百分比、股价、收益率、GDP 等等。利用特征值分解协方差矩阵完成 PCA 就会有麻烦，因为大家通过图 9 可以看到每一个主成分是若干特征的“线性融合”。哪怕每一列数据的量纲一致，比如鸢尾花前四列的单位都是厘米 cm，这种 PCA 技术路线还会受到不同特征方差大小影响。解决这些问题的方法是特征值分解线性相关系数矩阵，这是本章后文要讨论的话题。

椭圆：投影之前

如图 11 所示，协方差矩阵 Σ 椭球 (马氏距离为 1) 在六个平面上的投影。

通过旋转椭圆的形状、位置、旋转角度，我们可以读出标准差、相关性系数等重要信息。

图 12 比较数据 X 的分类和合并协方差矩阵对应的椭圆。



对椭圆、合并方差这些概念感到陌生的话，请回顾《统计至简》第 13 章。

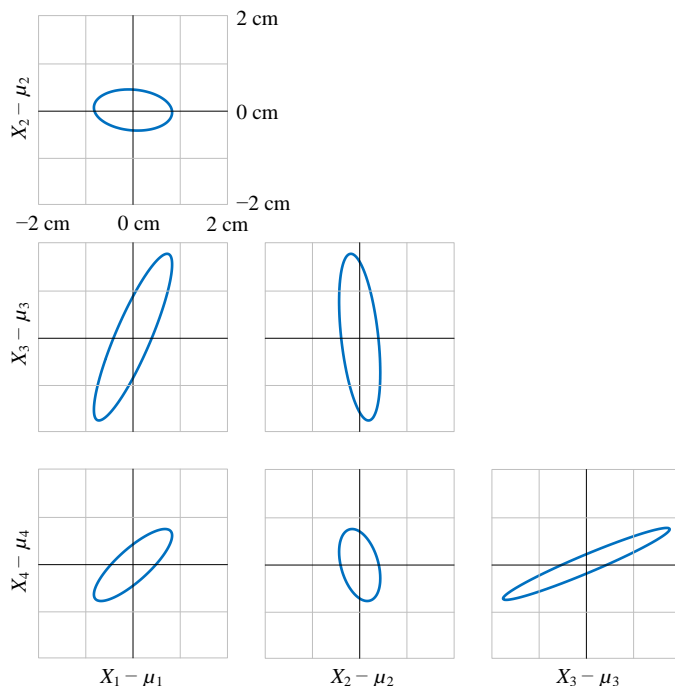
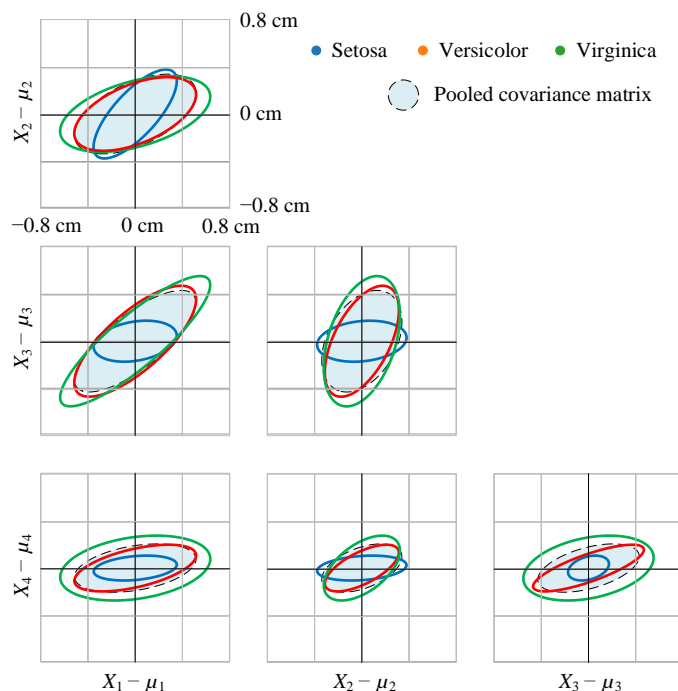


图 11. 马氏距离 1 椭圆，协方差矩阵 Σ

图 12. 马氏距离 1 椭圆，数据 \mathbf{X} 的分类、合并协方差矩阵 Σ

椭圆：投影之后

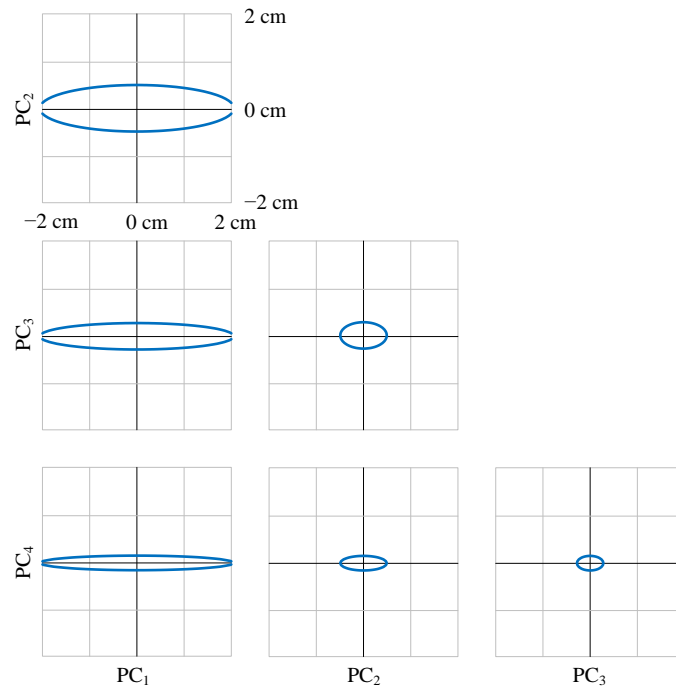
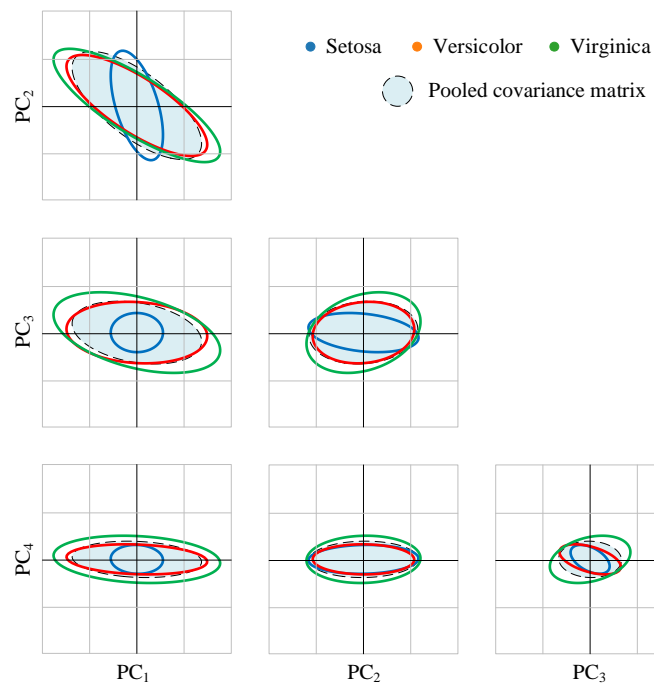
将中心化数据 \mathbf{X}_c 投影到 \mathbf{V}_c 得到的结果为 \mathbf{Y}_c ：

$$\mathbf{Y}_c = \mathbf{X}_c \mathbf{V}_c \quad (32)$$

\mathbf{Y}_c 的协方差矩阵就是 \mathbf{X} 的协方差矩阵的特征值矩阵。

图 13 所示为 \mathbf{Y}_c 的协方差矩阵在六个平面上的投影，这些椭圆都是正椭圆。 \mathbf{Y}_c 的协方差矩阵实际上就是 Σ 的特征值矩阵。

图 14 比较数据 \mathbf{Y}_c 的分类和合并协方差矩阵对应的椭圆。

图 13. 马氏距离 1 椭圆, \mathbf{Y}_c 的协方差矩阵图 14. 马氏距离 1 椭圆, 数据 \mathbf{Y}_c 的分类、合并协方差矩阵 Σ

16.3 格拉姆矩阵

特征值分解

图 15 所示为特征值分解格拉姆矩阵 G 。

注意，前文提过为了便于和协方差矩阵比较，本章中用的格拉姆矩阵 G 实际上是 $X^T X / (n - 1)$ 。

图 15 中的格拉姆矩阵 G 为对称矩阵，因此这个特征值分解同样是谱分解。

V_X 为正交矩阵，满足 $V_X^T V_X = V_X V_X^T = I$ 。 A_X 为对角矩阵，对角线元素为特征值，特征值从大到小排列。图 16 对比格拉姆矩阵 G 和 A_X 。下面，我们进一步分析这两个矩阵。

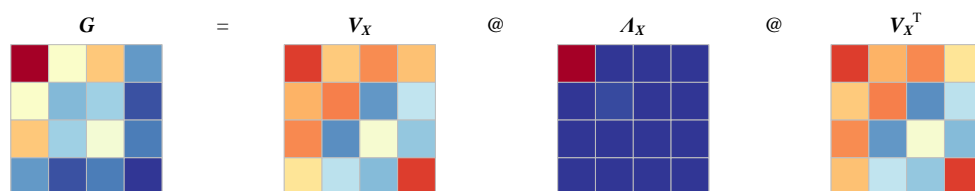


图 15. 特征值分解格拉姆矩阵 G

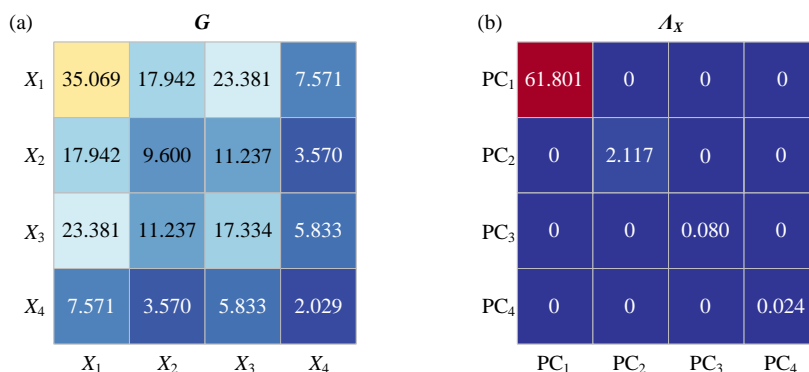
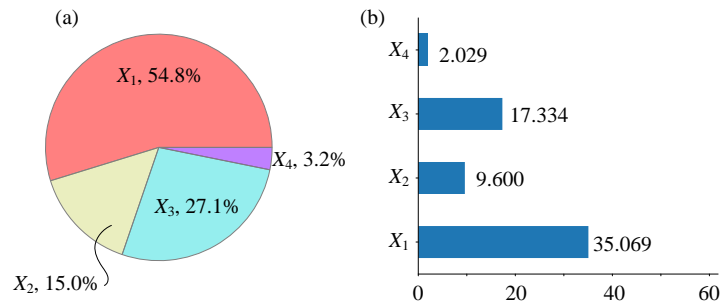
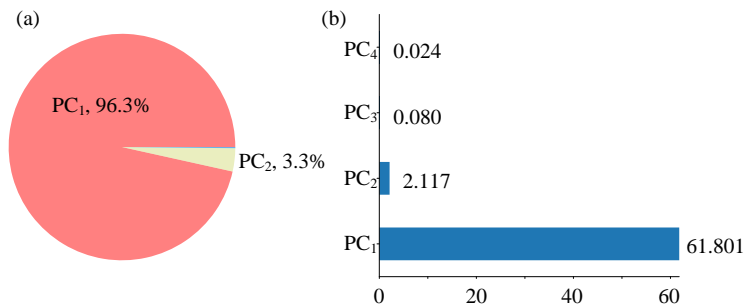


图 16. 对比 G 和 A_X 热图

分解前后

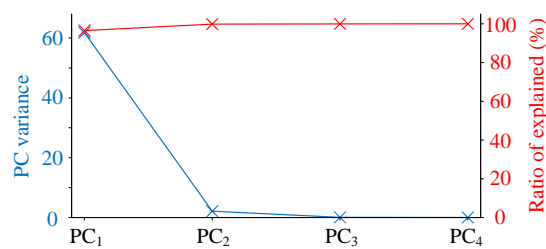
G 和 A_X 的主对角线之和相同，即 $\text{trace}(G) = \text{trace}(A_X)$ 。如图 17 所示，矩阵 G 的主对角线元素为矩阵 X 的每一列向量的模除以 $n - 1$ ，代表某个特征相对于原点的分散情况，即“不去均值”的方差。

而 $\text{trace}(\mathbf{G})$ 相当于数据整体相对于原点的分散度量。如图 17 所示，矩阵 \mathbf{X} 的第一列和第二列贡献最大。经过特征值分解之后，如图 18 所示，第一主成分解释了大部分数据分散情况，占比高达 96.3%。

图 17. \mathbf{G} 的主对角线成分图 18. \mathbf{X} 的主对角线成分，格拉姆矩阵 \mathbf{G} 的特征值

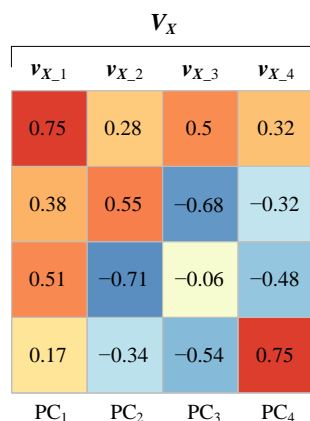
陡坡图

图 19 所示为在特征值分解格拉姆矩阵 \mathbf{G} 主成分分析的陡坡图。

图 19. 陡坡图，特征值分解格拉姆矩阵 \mathbf{G}

特征向量矩阵

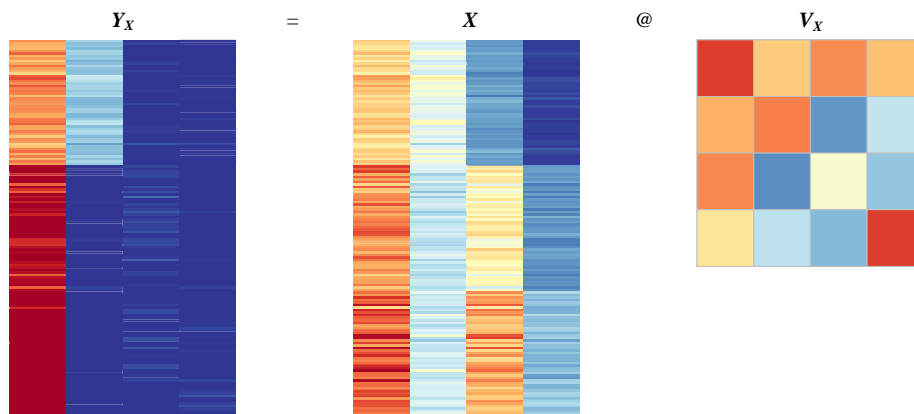
图 20 所示为特征向量矩阵 \mathbf{V}_X 热图。显然，图 20 不同于图 7。

图 20. 特征向量矩阵 V_X 热图

投影

图 21 是将原始数据 X 投影到 V_X ，即 $Y_X = XV_X$ 。 Y_X 的特点是其格拉姆矩阵为对角方阵，也就是说 Y_X 的列向量两两正交。

注意，两两正交不代表线性无关。

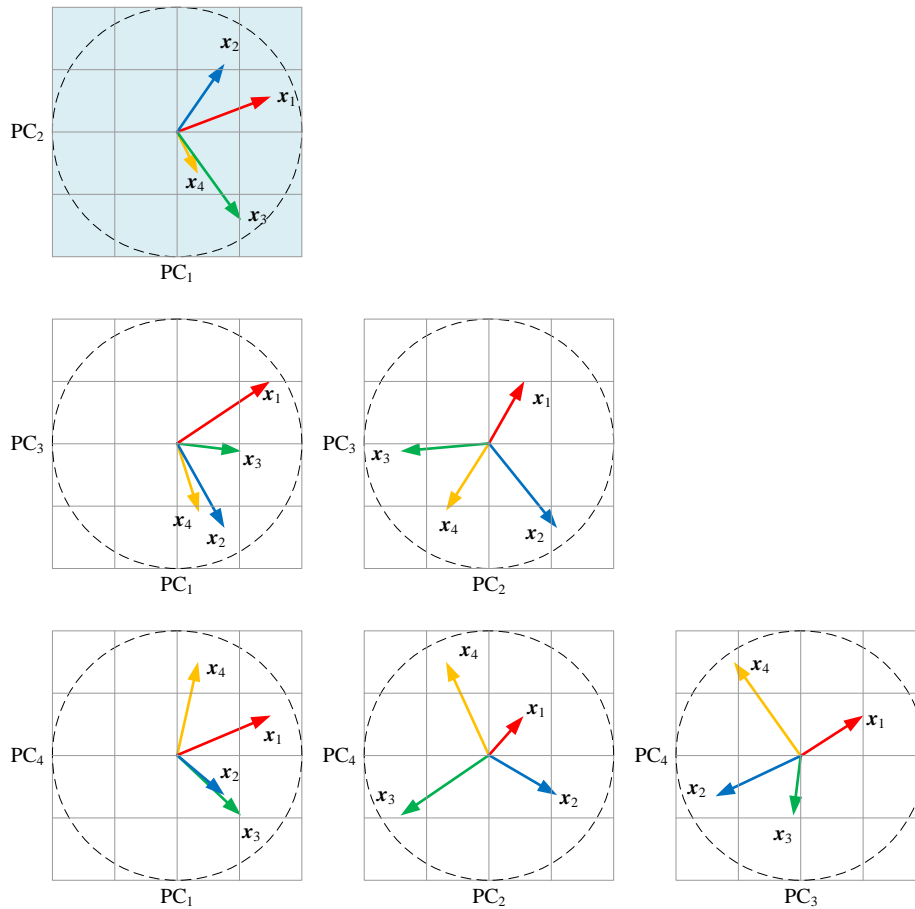
图 21. 将原始数据 X 投影到 V_X

正交矩阵 V_X 也是一个规范正交基， V_X 是因原始数据 X 而生。前文提到， V_c 同样是一个规范正交基，但是 V_c 是因中心化数据矩阵 X_c 而生。

我们当然可以将 X 投影到 V_c 这个规范正交基中，大家可以自行验证 XV_c 的协方差和 X_cV_c 相同，都是对角方阵。也就是说， XV_c 的列向量也是线性无关。但是， XV_c 的质心不再是原点。

双标图

图 22 所示为 V_X 的双标图。请大家自行比较图 9 和图 22。

图 22. V_X 双标图，特征值分解格拉姆矩阵 G

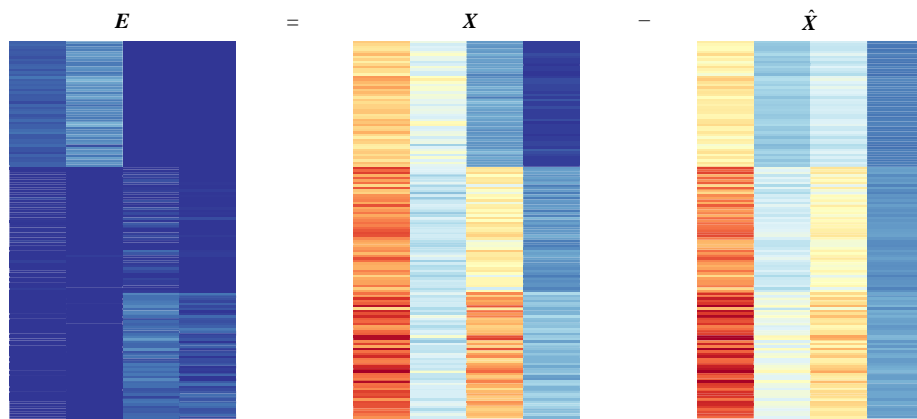
数据还原、误差

由于本节中 PCA 分析直接采用特征值分解格拉姆矩阵 G ，根据 (1)，利用第一主成分还原原始数据 X 时我们不需要加入质心成分：

$$X \approx X v_{X-1} \otimes v_{X-1} \quad (33)$$

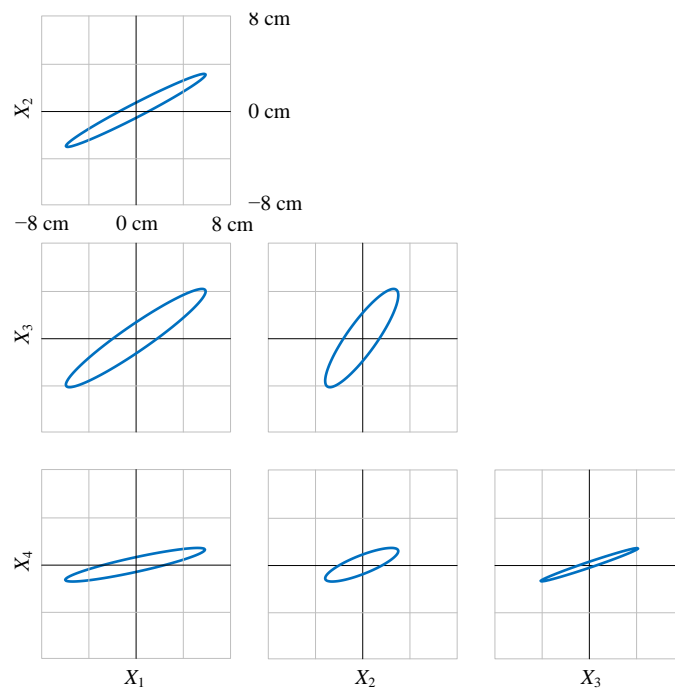
如果用第一、二主成分还原 X ，上式也需要再加一项：

$$X \approx \underbrace{X v_{X-1} \otimes v_{X-1}}_{\text{First principal}} + \underbrace{X v_{X-2} \otimes v_{X-2}}_{\text{Second principal}} \quad (34)$$

图 23. 第一主成分估计 \hat{X}

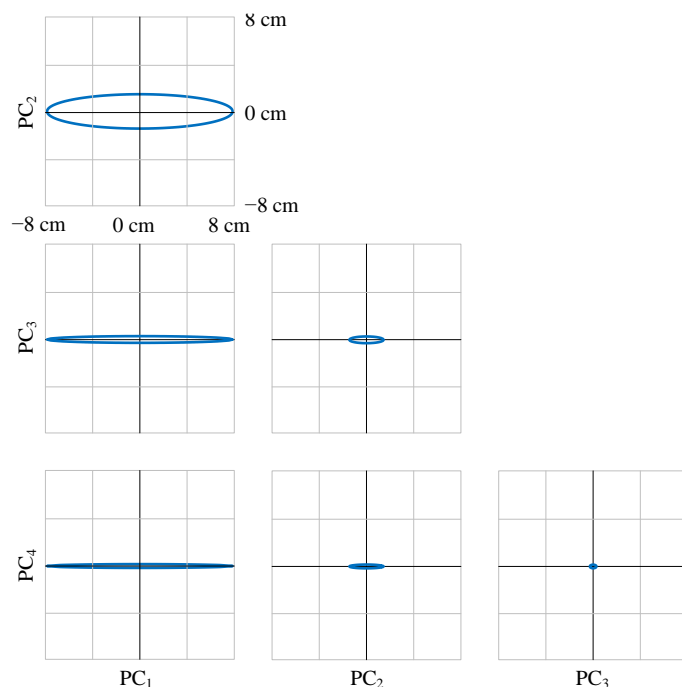
椭圆：投影之前

图 24 所示为格拉姆矩阵 G 对应的旋转椭圆。 G 相当于“不去均值”的协方差矩阵。观察图 24，我们发现椭圆的朝向都是一三象限，而且椭圆都细长。比较图 11 和图 24，大家应该理解为什么需要去均值。

图 24. 马氏距离 1 椭圆，“不去均值”的协方差矩阵 Σ

椭圆：投影之后

经过 $Y_X = X V_X$ 投影之后，图 25 所示 Y_X 协方差矩阵对应的椭圆。

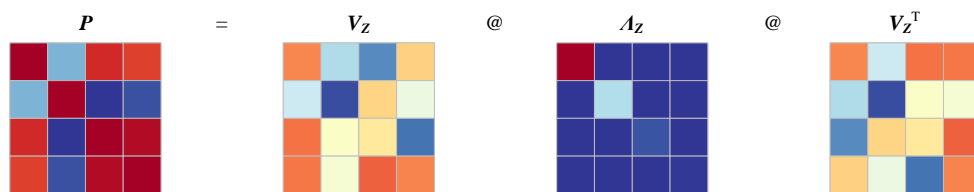
图 25. 马氏距离 1 椭圆, \mathbf{Y}_X 的协方差矩阵

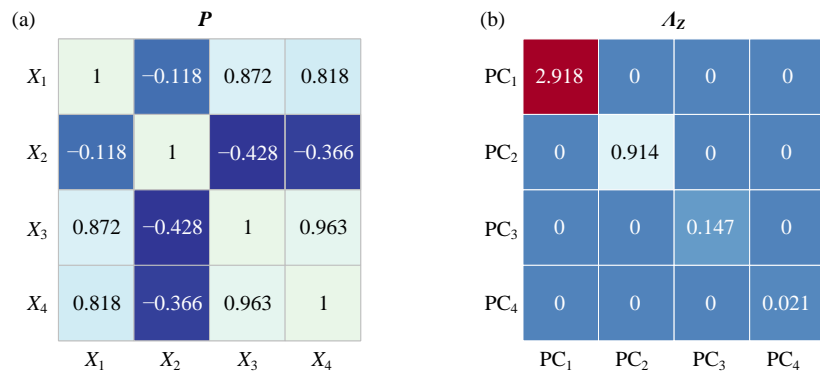
16.4 相关性系数矩阵

标准化数据 \mathbf{Z}_X 相当于 Z 分数, 因此消除了特征量纲影响。因此, 特征值分解相关性系数矩阵不再受量纲影响。此外, 标准化数据每一列特征数据均值均为 0, 方差为 1。这也消除了较大方差特征的影响。

特征值分解

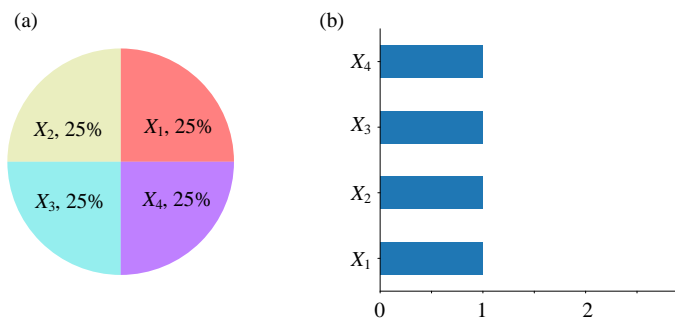
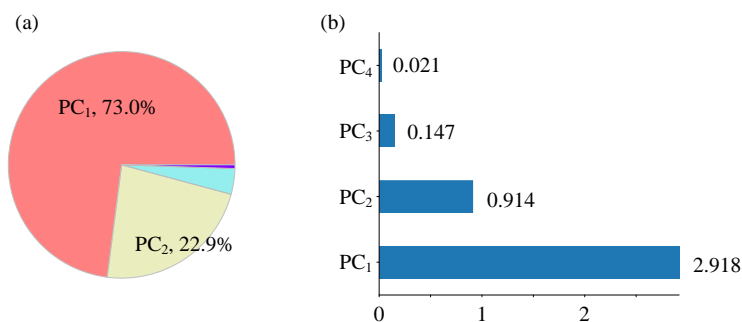
图 26 所示为特征值分解相关性系数矩阵 \mathbf{P} , \mathbf{P} 的主对角线都是 1, \mathbf{P} 对角线之外的元素都是线性相关系数。图 27 对比相关性系数矩阵 \mathbf{P} 和 \mathbf{A}_Z 热图。同样地, \mathbf{P} 和 \mathbf{A}_Z 主对角线之和相同, 即 $\text{trace}(\mathbf{P}) = \text{trace}(\mathbf{A}_Z)$ 。

图 26. 特征值分解相关性系数矩阵 \mathbf{P}

图 27. 对比相关性系数矩阵 P 和 A_z 热图

分解前后

图 4 中， X_3 对方差和 $\text{trace}(\Sigma)$ 贡献超过 68%，而 X_3 的贡献小于 5%。而图 28 中每个特征经过标准化之后，贡献率完全相同。方差小特征也可能含有重要的信息，利用特征值分解相关性系数完成 PCA，可以消除这种顾虑。

图 28. 相关性系数矩阵 P 主对角线成分图 29. A_z 的主对角线成分，相关性系数矩阵 P 特征值

陡坡图

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 30 所示为特征值分解相关性系数矩阵 P 主成分分析结果陡坡图。第一主成分贡献小于 80%。

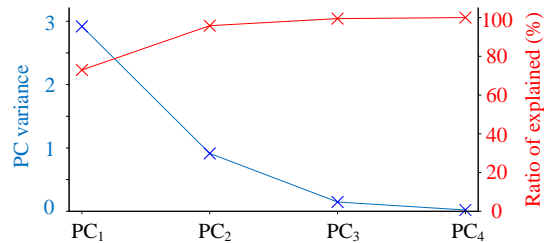


图 30. 陡坡图，特征值分解相关性系数矩阵 P

特征向量矩阵

图 31 所示为特征向量矩阵 V_Z 热图。这幅图和图 7、图 20 均不同。

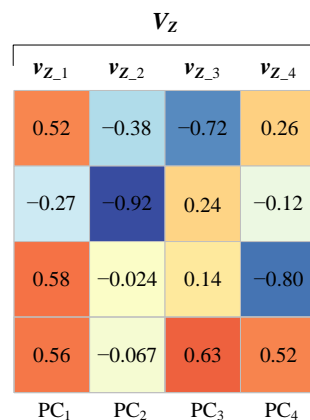
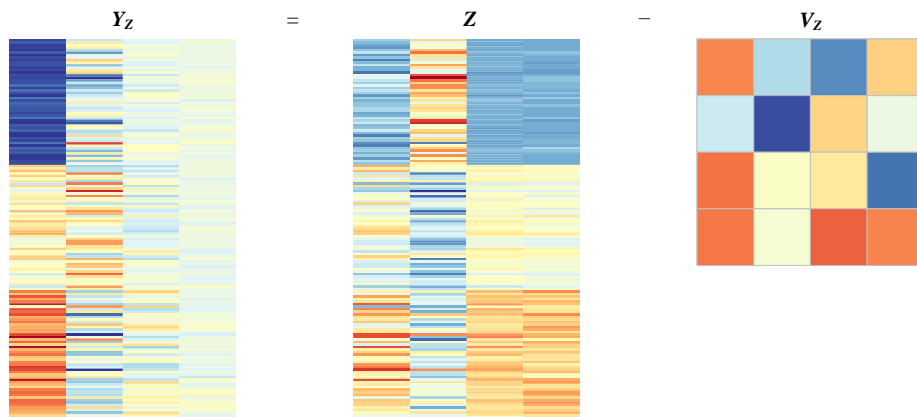


图 31. 特征向量矩阵 V_Z 热图

投影

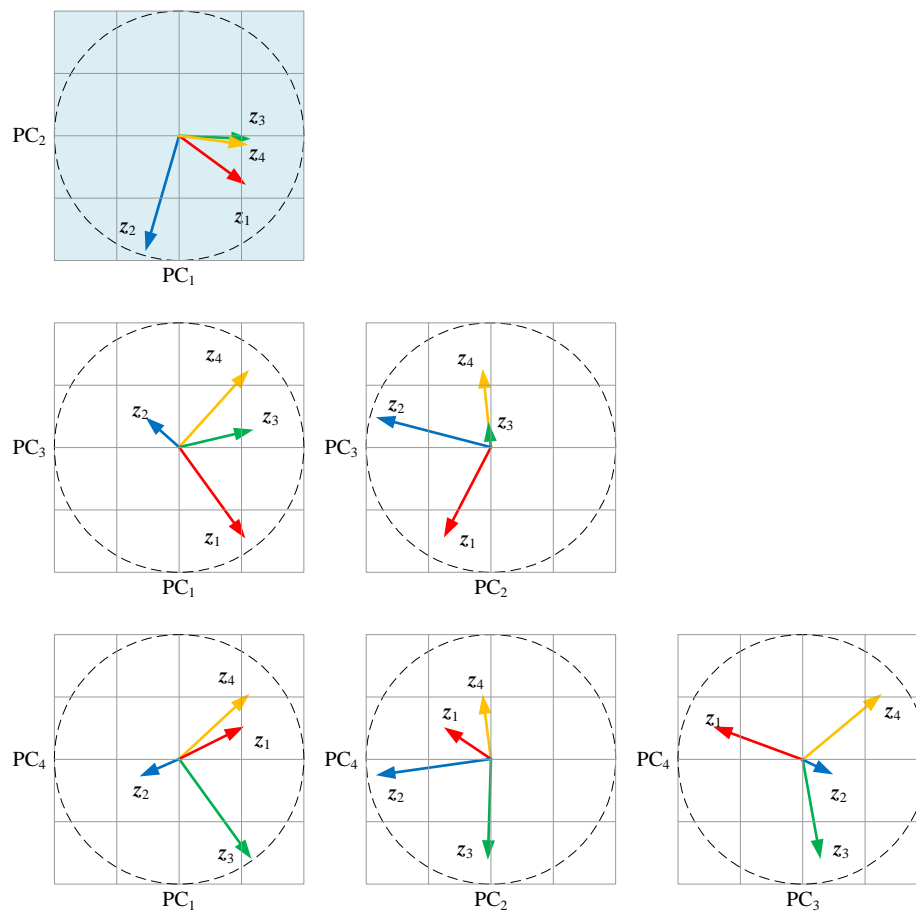
图 32 所示为标准化数据 Z 投影到 V_Z 得到数据矩阵 Y_Z 。同样地，正交矩阵 V_Z 也是一个规范正交基，而 V_Z 是因中心化数据 Z_X 而生。

请大家将原数据 X 、中心化 X_C 也投影到 V_Z 中，并检验结果的协方差矩阵和质心。

图 32. 中心化数据 Z 投影到 V_Z

双标图

图 33 所示为 V_Z 双标图，请大家比较本章三幅双标图。

图 33. V_Z 双标图，特征值分解格相关性系数矩阵 P

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

数据还原、误差

图 34 所示为第一主成分估计 Z_X :

$$Z_X \approx Z_X v_{X-1} \otimes v_{X-1} \quad (35)$$

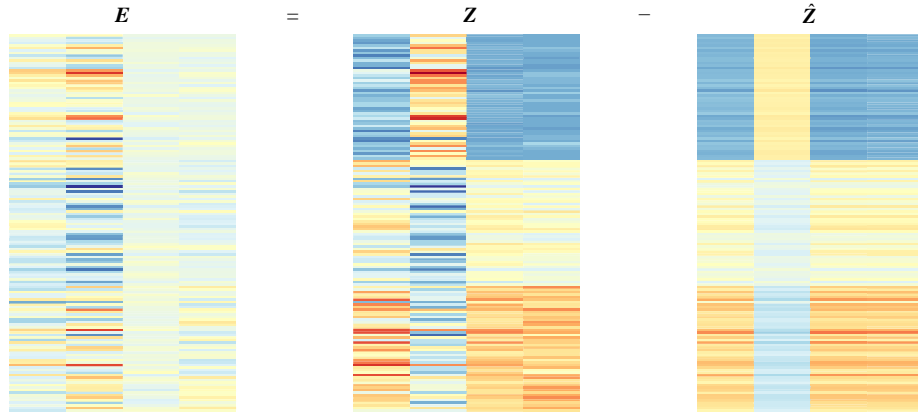


图 34. 第一主成分还原 Z_X

Z_X 可以写成:

$$Z_X = (X - E(X))D^{-1} = \sum_{j=1}^D Z_X v_{X-j} \otimes v_{X-j} \quad (36)$$

用 V_Z 还原 X :

$$X = \left(\sum_{j=1}^D Z_X v_{X-j} \otimes v_{X-j} \right) D + E(X) \quad (37)$$

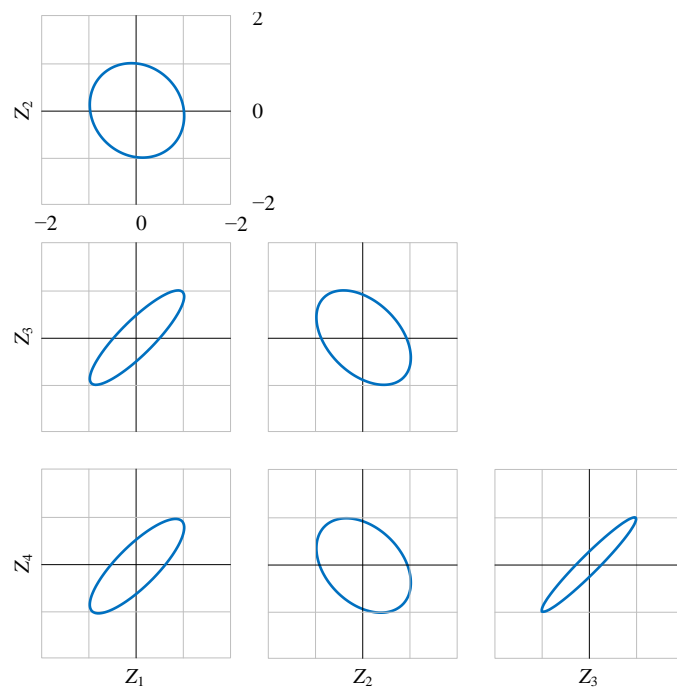
用 V_Z 第一主成分估计 X :

$$X \approx \underbrace{(Z_X v_{X-1} \otimes v_{X-1})}_{\text{First principal}} D + E(X) \quad (38)$$

其中, D 起到缩放的作用, $E(X)$ 是平移的作用。

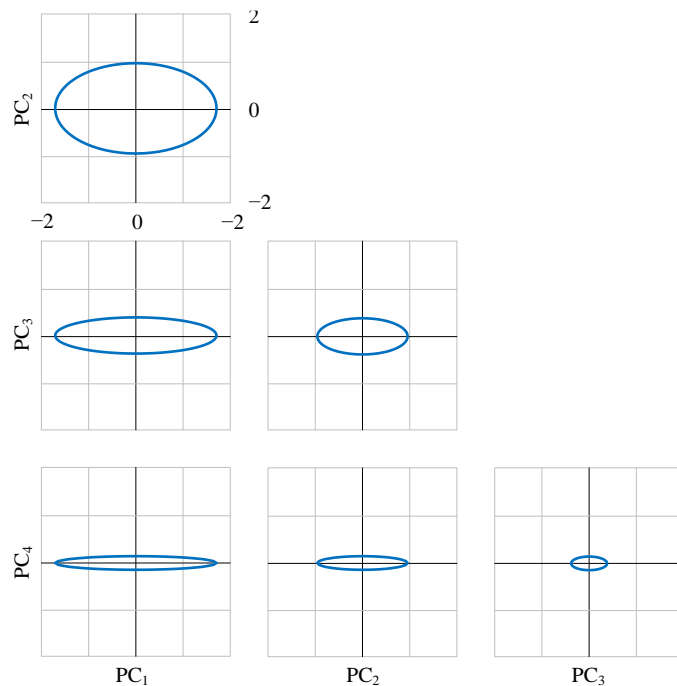
椭圆：投影之前

图 35 所示为投影之前相关性系数矩阵 P 对应的椭圆。请大家特别和前文协方差矩阵对应椭圆进行比较。

图 35. 马氏距离 1 椭圆，相关性系数矩阵 P

椭圆：投影之后

图 36 所示为投影之后正椭圆的位置和形状。

图 36. 马氏距离 1 椭圆， Yz 的协方差矩阵

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com



Bk6_Ch16_01.py 绘制本章大部分图片。



主成分分析是鸢尾花书的“常客”，我们用椭圆、数据、格拉姆矩阵、协方差矩阵、特征值分解、奇异值分解、线性组合、优化、随机变量的线性函数等等视角探讨过主成分分析。换句话说，机器学习常用的数学工具在主成分分析处达到了一种融合，大家也看到了数学板块实际上不是一个孤立的个体，它们有其内在联系和网络。

鸢尾花书有关主成分分析专题内容到此为止，下两章我们将主要介绍和主成分分析相关的回归算法。此外，本书还会在最后一章比较奇异值分解和因子分析的异同。《机器学习》一册还要综述常见降维算法，其中还包括核主成分分析 KPCA，KPCA 相当于 PCA 的升级版。



在用椭圆理解数据、解释主成分分析方面，以下论文给本章很多启发，欢迎大家阅读：

<https://arxiv.org/pdf/1302.4881.pdf>