25 Principal Component Analysis

主成分分析

以随机变量线性变换为视角

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
- sklearn.decomposition.PCA() 主成分分析函数
- 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

25.1 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 $\boldsymbol{X}$	奇异值分解	$\boldsymbol{X} = \boldsymbol{U}_X\boldsymbol{S}_X\boldsymbol{V}_X^{\mathrm{T}}$
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 $\boldsymbol{G} = \dfrac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{n-1}$	特征值分解	$\boldsymbol{G} = \boldsymbol{V}_X\boldsymbol{\Lambda}_X\boldsymbol{V}_X^{\mathrm{T}}$
中心化数据矩阵 $\boldsymbol{X}_c = \boldsymbol{X} - \mathrm{E}(\boldsymbol{X})$	奇异值分解	$\boldsymbol{X}_c = \boldsymbol{U}_c\boldsymbol{S}_c\boldsymbol{V}_c^{\mathrm{T}}$
协方差矩阵 $\boldsymbol{\Sigma} = \dfrac{(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))^{\mathrm{T}}(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))}{n-1}$	特征值分解	$\boldsymbol{\Sigma} = \boldsymbol{V}_c\boldsymbol{\Lambda}_c\boldsymbol{V}_c^{\mathrm{T}}$
标准化数据 (z 分数) $\begin{gathered}\boldsymbol{Z}_X = (\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))\boldsymbol{D}^{-1}\\ \boldsymbol{D} = \mathrm{diag}(\mathrm{diag}(\boldsymbol{\Sigma}))^{\frac{1}{2}}\end{gathered}$	奇异值分解	$\boldsymbol{Z}_X = \boldsymbol{U}_Z\boldsymbol{S}_Z\boldsymbol{V}_Z^{\mathrm{T}}$
相关性系数矩阵 $\begin{gathered}\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{\Sigma}\boldsymbol{D}^{-1}\\ \boldsymbol{D} = \mathrm{diag}(\mathrm{diag}(\boldsymbol{\Sigma}))^{\frac{1}{2}}\end{gathered}$	特征值分解	$\boldsymbol{P} = \boldsymbol{V}_Z\boldsymbol{\Lambda}_Z\boldsymbol{V}_Z^{\mathrm{T}}$

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 $\boldsymbol{X}$。如图 1 所示，原始数据矩阵 $\boldsymbol{X}$ 的形状为 $n \times D$。

$\boldsymbol{X}$ 的格拉姆矩阵 $\boldsymbol{G}$ 为：

$$\boldsymbol{G} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \tag{1}$$

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

中心化 (去均值) 矩阵 $\boldsymbol{X}_c$ 为：

$$\boldsymbol{X}_c = \boldsymbol{X} - \mathrm{E}(\boldsymbol{X}) \tag{2}$$

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

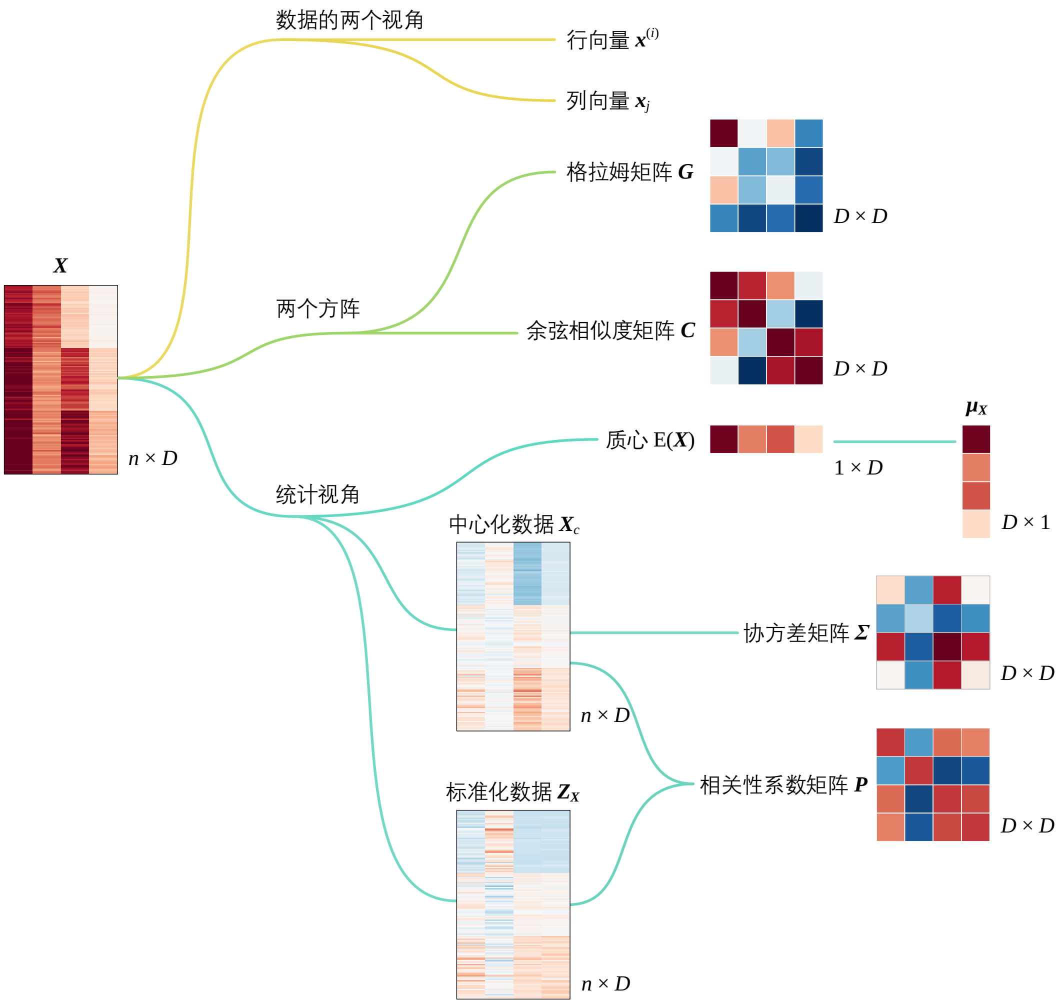

图 1. $\boldsymbol{X}$ 衍生得到的几个矩阵，来自《矩阵力量》

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

$$\boldsymbol{\Sigma} = \frac{\boldsymbol{X}_c^{\mathrm{T}} \boldsymbol{X}_c}{n-1} = \frac{\left(\boldsymbol{X} - \mathrm{E}\left(\boldsymbol{X}\right)\right)^{\mathrm{T}} \left(\boldsymbol{X} - \mathrm{E}\left(\boldsymbol{X}\right)\right)}{n-1} \tag{3}$$

请大家特别注意，为了方便和协方差比较，本章中 $\boldsymbol{G}$ 特别定义为：

$$\boldsymbol{G} = \frac{\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}}{n-1} \tag{4}$$

标准化 (standardization 或 z-score normalization) 数据矩阵 $\boldsymbol{Z}_X$ 为：

$$\boldsymbol{Z}_X = \left(\boldsymbol{X} - \mathrm{E}\left(\boldsymbol{X}\right)\right) \boldsymbol{D}^{-1} \tag{5}$$

其中 $\boldsymbol{D}$ 为：

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

(5) 中的每一列都是 z 每个特征的 z 分数。$\boldsymbol{Z}_X$ 的质心也位于原点，不同的是 $\boldsymbol{Z}_X$ 每个特征的方差都是 1。

线性相关性系数矩阵 $\boldsymbol{P}$ 为：

$$\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{\Sigma}\boldsymbol{D}^{-1} \tag{7}$$

$\boldsymbol{P}$ 实际上是 $\boldsymbol{Z}_X$ 的协方差，即：

$$\boldsymbol{P} = \frac{\boldsymbol{Z}_X^{\mathrm{T}}\boldsymbol{Z}_X}{n-1} \tag{8}$$

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 $\boldsymbol{X}$ 进行经济型 SVD 分解：

$$\boldsymbol{X} = \boldsymbol{U}_X\boldsymbol{S}_X\boldsymbol{V}_X^{\mathrm{T}} \tag{9}$$

其中，$\boldsymbol{S}_X$ 为对角方阵。

将 (9) 代入 (1)：

$$\boldsymbol{G} = \boldsymbol{V}_X\boldsymbol{S}_X^2\boldsymbol{V}_X^{\mathrm{T}} \tag{10}$$

上式便是格拉姆 $\boldsymbol{G}$ 的特征值分解。

对中心化数据矩阵 $\boldsymbol{X}_c$ 经济型 SVD 分解：

$$\boldsymbol{X}_c = \boldsymbol{U}_c\boldsymbol{S}_c\boldsymbol{V}_c^{\mathrm{T}} \tag{11}$$

而协方差矩阵 $\boldsymbol{\Sigma}$ 则可以写成：

$$\boldsymbol{\Sigma} = \boldsymbol{V}_c\frac{\boldsymbol{S}_c^2}{n-1}\boldsymbol{V}_c^{\mathrm{T}} \tag{12}$$

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

$$\lambda_{c\_j} = \frac{s_{c\_j}^2}{n-1} \tag{13}$$

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

$$\boldsymbol{Z}_X = \boldsymbol{U}_Z \boldsymbol{S}_Z \boldsymbol{V}_Z^{\mathrm{T}} \tag{14}$$

相关性系数矩阵 $\boldsymbol{P}$ 则可以写成：

$$\boldsymbol{P} = \boldsymbol{V}_Z \frac{\boldsymbol{S}_Z^2}{n-1} \boldsymbol{V}_Z^{\mathrm{T}} \tag{15}$$

上式相当于对 $\boldsymbol{P}$ 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。

25.2 协方差矩阵

本节讲解利用特征值分解协方差矩阵 $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\Sigma}$。$\boldsymbol{\Sigma}$ 的对角线元素为方差，其他元素为协方差。$\boldsymbol{\Sigma}$ 的迹代表方差之和：

$$\operatorname{trace}(\boldsymbol{\Sigma}) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_D^2 = \sum_{j=1}^{D} \sigma_j^2 \tag{16}$$

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

$\boldsymbol{\Lambda}_c$ 为对角矩阵，对角线元素为特征值，特征值从大到小排列。$\boldsymbol{X}_c$ 投影到规范正交基 $\boldsymbol{V}_c$ 中得到 $\boldsymbol{Y}_c$，即 $\boldsymbol{Y}_c = \boldsymbol{X}_c\boldsymbol{V}_c$。$\boldsymbol{\Lambda}_c$ 的特征值实际上是 $\boldsymbol{Y}_c$ 的方差。因此，在主成分分析中，特征值也叫主成分方差。

$\boldsymbol{\Lambda}_c$ 的方差，即特征值，之和为：

$$\operatorname{trace}(\boldsymbol{\Lambda}_c) = \lambda_1 + \lambda_2 + \cdots \lambda_D = \sum_{j=1}^{D} \lambda_j \tag{17}$$

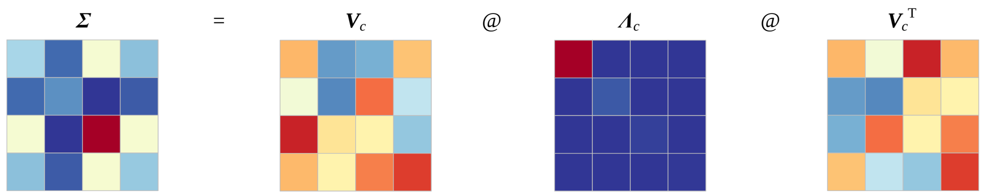

图 2. 特征值分解协方差矩阵 $\boldsymbol{\Sigma}$

图 12 对比格拉姆矩阵 $\boldsymbol{G}$ 和 $\boldsymbol{\Lambda}_X$。

下面，我们进一步分析这两个矩阵。

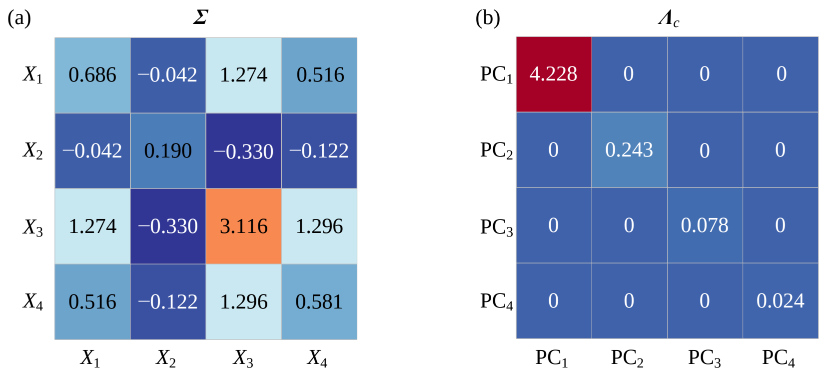

图 3. 对比协方差矩阵 $\boldsymbol{\Sigma}$ 和 $\boldsymbol{\Lambda}_c$ 热图

分解前后

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

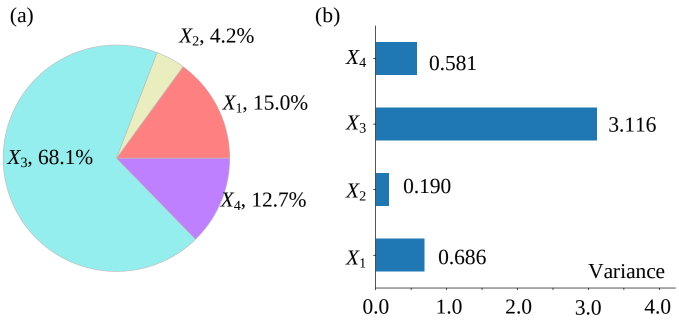

图 4. 协方差矩阵 $\boldsymbol{\Sigma}$ 的主对角线成分，即方差

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 $\boldsymbol{\Sigma}$ 的迹 trace($\boldsymbol{\Sigma}$) 等于的特征值方阵 $\boldsymbol{\Lambda}_c$ 迹 trace($\boldsymbol{\Lambda}_c$)：

$$\operatorname{trace}(\boldsymbol{\Sigma}) = \operatorname{trace}(\boldsymbol{\Lambda}_c) \tag{18}$$

即：

$$\sum_{j=1}^{D} \sigma_j^2 = \sum_{j=1}^{D} \lambda_j \tag{19}$$

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 trace($\boldsymbol{\Lambda}_c$) 的贡献百分比为：

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

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。

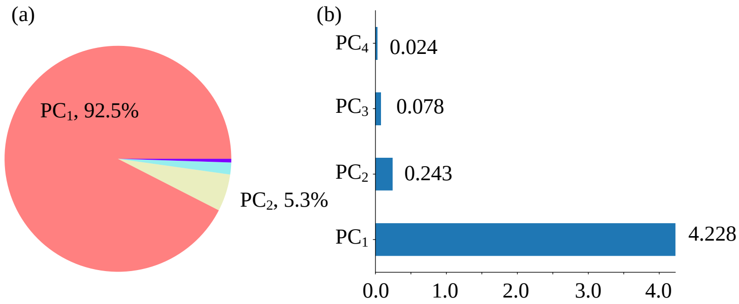

图 5. $\boldsymbol{\Lambda}_c$ 的主对角线成分，协方差矩阵 $\boldsymbol{\Sigma}$ 的特征值

陡坡图

主成分分析中，我们经常用陡坡图 (Scree plot) 可视化前 p 个主成分解释总方差的百分比，即累积贡献率：

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

图 6 所示为特征值分解协方差矩阵 $\boldsymbol{\Sigma}$ 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。

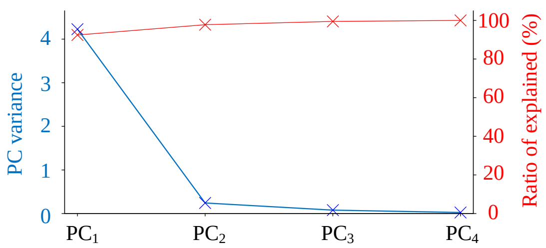

图 6. 陡坡图，特征值分解协方差矩阵 $\boldsymbol{\Sigma}$

特征向量矩阵

图 7 所示为特征向量矩阵 $\boldsymbol{V}_c$ 热图。$\boldsymbol{V}_c$ 的每一列便代表一个主成分的方向，即 $\boldsymbol{V}_c = [\boldsymbol{v}_{c\_1}, \boldsymbol{v}_{c\_2}, \boldsymbol{v}_{c\_3}, \boldsymbol{v}_{c\_4}]$ 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，$\boldsymbol{V}_c$ 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 $\boldsymbol{v}_j\sqrt{\lambda_j}$ 。

$\boldsymbol{V}_c$也可以通过经济型 SVD 分解中心化矩阵 $\boldsymbol{X}_c$得到。

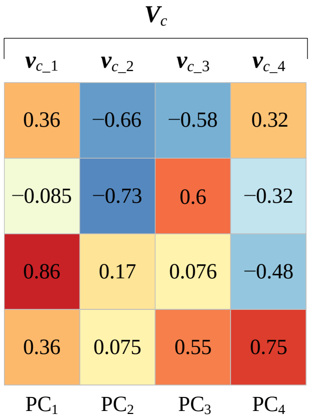

图 7. 特征向量矩阵 $\boldsymbol{V}_c$热图

投影

由于 $\boldsymbol{V}_c$为正交矩阵，满足 $\boldsymbol{V}_c^{\mathrm{T}}\boldsymbol{V}_c = \boldsymbol{V}_c\boldsymbol{V}_c^{\mathrm{T}} = \boldsymbol{I}$，因此 $\boldsymbol{V}_c$本身也是规范正交基。如图 8 所示，将中心化矩阵 $\boldsymbol{X}_c$投影到 $\boldsymbol{V}_c$这个规范正交基中得到数据矩阵 $\boldsymbol{Y}_c$，即 $\boldsymbol{Y}_c = \boldsymbol{X}_c\boldsymbol{V}_c$。通过图 8 中的 $\boldsymbol{Y}_c$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，$\boldsymbol{V}_c$的第一列列向量实际上是如下优化问题的解：

$$\begin{aligned} \boldsymbol{v}_{c\_1} &= \arg\max_{\boldsymbol{v}} \left\|\boldsymbol{X}_c\boldsymbol{v}\right\| \\ \text{subject to: } & \left\|\boldsymbol{v}\right\| = 1 \end{aligned} \tag{22}$$

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

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

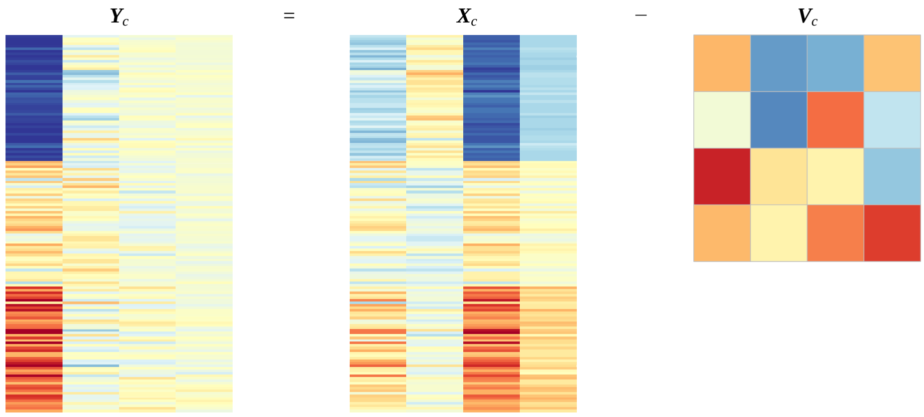

图 8. 将中心化数据 $\boldsymbol{X}_c$投影到 $\boldsymbol{V}_c$

双标图

如图 9 所示，双标图 (biplot) 是可视化特征向量矩阵 $\boldsymbol{V}_c$ 的重要方法。以图 9 中蓝色背景的双标图为例，中心化数据 $\boldsymbol{X}_c$ 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 $\boldsymbol{V}_c$ 第一列 $\boldsymbol{v}_{c\_1}$。我们还可以把投影数据的散点图也画在双标图上，在《数据有道》一册我们还会深入探讨这些可视化方案。

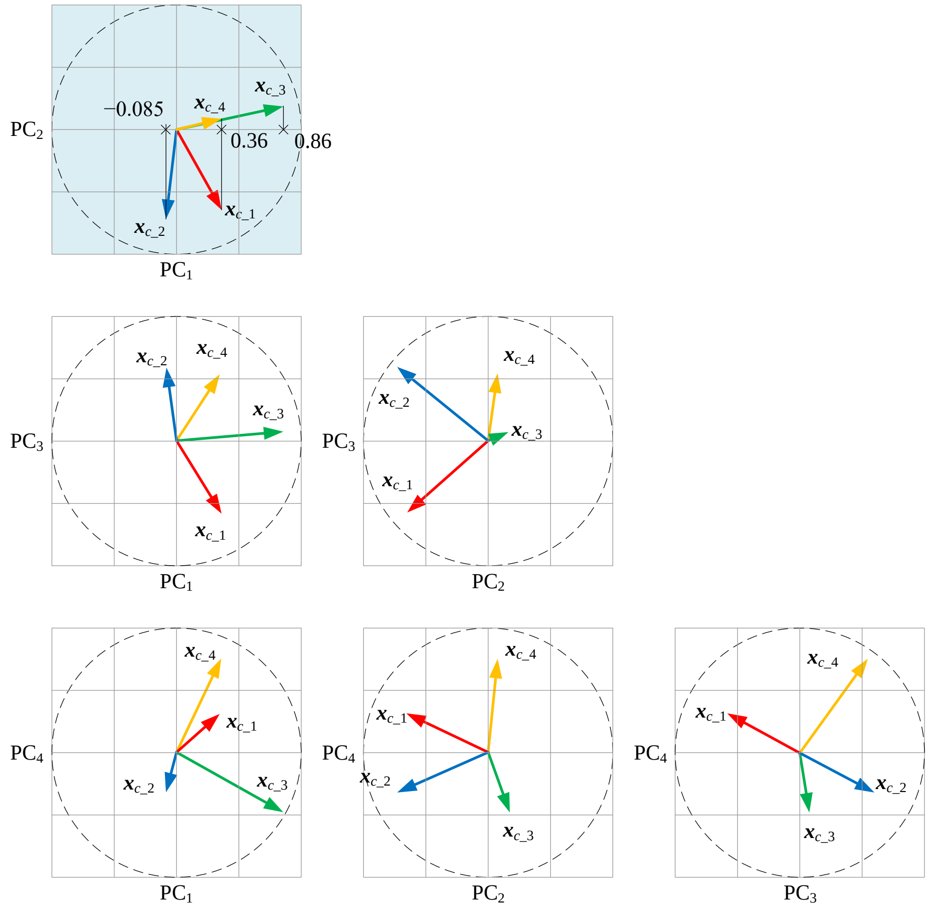

图 9. $\boldsymbol{V}_c$ 双标图，特征值分解协方差矩阵 $\boldsymbol{\Sigma}$

数据还原、误差

将 (11) 展开写成：

$$
\begin{aligned}
\boldsymbol{X}_c &= \underbrace{\begin{bmatrix} \boldsymbol{u}_{c\_1} & \boldsymbol{u}_{c\_2} & \cdots & \boldsymbol{u}_{c\_D} \end{bmatrix}}_{\boldsymbol{U}_c} \underbrace{\begin{bmatrix} s_{c\_1} & & & \\ & s_{c\_2} & & \\ & & \ddots & \\ & & & s_{c\_D} \end{bmatrix}}_{\boldsymbol{S}_c} \underbrace{\begin{bmatrix} \boldsymbol{v}_{c\_1}{}^{\mathrm{T}} \\ \boldsymbol{v}_{c\_2}{}^{\mathrm{T}} \\ \vdots \\ \boldsymbol{v}_{c\_D}{}^{\mathrm{T}} \end{bmatrix}}_{\boldsymbol{V}_c^{\mathrm{T}}} \\
&= s_{c\_1}\boldsymbol{u}_{c\_1}\boldsymbol{v}_{c\_1}{}^{\mathrm{T}} + s_{c\_2}\boldsymbol{u}_{c\_2}\boldsymbol{v}_{c\_2}{}^{\mathrm{T}} + \cdots + s_{c\_D}\boldsymbol{u}_{c\_D}\boldsymbol{v}_{c\_D}{}^{\mathrm{T}} = \sum_{j=1}^{D} s_{c\_j}\boldsymbol{u}_{c\_j}\boldsymbol{v}_{c\_j}{}^{\mathrm{T}}
\end{aligned}
\tag{23}
$$

图 10 所示为用第一主成分逼近估计 $\boldsymbol{X}_c$，即：

$$\hat{\boldsymbol{X}}_c = \underbrace{s_{c\_1}\boldsymbol{u}_{c\_1}\boldsymbol{v}_{c\_1}^{\mathrm{T}}}_{\text{First principal}} \tag{24}$$

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

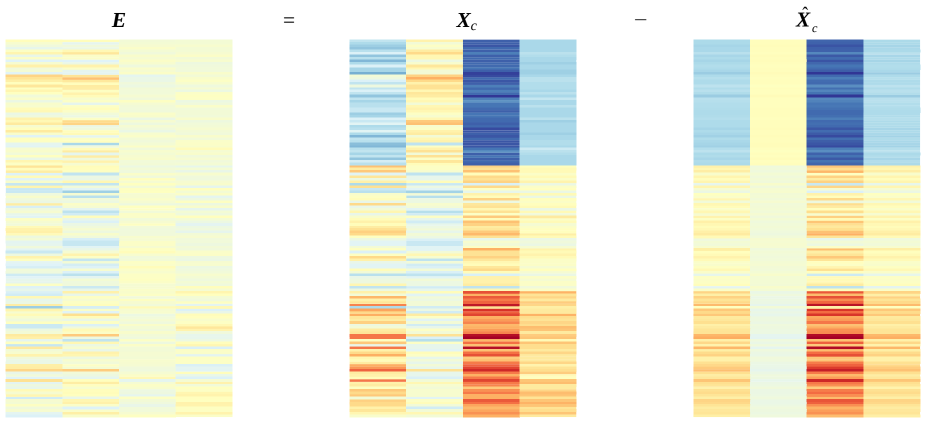

图 10. 第一主成分估计 $\boldsymbol{X}_c$

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

$$\boldsymbol{X} = \boldsymbol{X}_c + \mathrm{E}(\boldsymbol{X}) = \sum_{j=1}^{D} s_{c\_j}\boldsymbol{u}_{c\_j}\boldsymbol{v}_{c\_j}^{\mathrm{T}} + \mathrm{E}(\boldsymbol{X}) \tag{25}$$

如果利用第一主成分估计原始数据矩阵 $\boldsymbol{X}$ 的话，可以利用：

$$\boldsymbol{X} \approx s_{c\_1}\boldsymbol{u}_{c\_1}\boldsymbol{v}_{c\_1}^{\mathrm{T}} + \mathrm{E}(\boldsymbol{X}) \tag{26}$$

上式中，$\mathrm{E}(\boldsymbol{X})$ 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 $\boldsymbol{V}_c$，没有 $\boldsymbol{U}_c$。难道我们还需要再对 $\boldsymbol{X}_c$ 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 $\boldsymbol{X}_c$ 可以写成：

$$\boldsymbol{X}_c = \boldsymbol{X}_c\boldsymbol{I} = \boldsymbol{X}_c\boldsymbol{V}_c\boldsymbol{V}_c^{\mathrm{T}} \tag{27}$$

将 $\boldsymbol{V}_c$ 展开，上式可以写成：

$$\begin{aligned}\boldsymbol{X}_c &= \boldsymbol{X}_c \underbrace{\begin{bmatrix} \boldsymbol{v}_{c\_1} & \boldsymbol{v}_{c\_2} & \cdots & \boldsymbol{v}_{c\_D} \end{bmatrix}}_{\boldsymbol{V}_c} \underbrace{\begin{bmatrix} \boldsymbol{v}_{c\_1}^{\mathrm{T}} \\ \boldsymbol{v}_{c\_2}^{\mathrm{T}} \\ \vdots \\ \boldsymbol{v}_{c\_D}^{\mathrm{T}} \end{bmatrix}}_{\boldsymbol{V}_c^{\mathrm{T}}} \\ &= \boldsymbol{X}_c\boldsymbol{v}_{c\_1}\boldsymbol{v}_{c\_1}^{\mathrm{T}} + \boldsymbol{X}_c\boldsymbol{v}_{c\_2}\boldsymbol{v}_{c\_2}^{\mathrm{T}} + \cdots + \boldsymbol{X}_c\boldsymbol{v}_{c\_D}\boldsymbol{v}_{c\_D}^{\mathrm{T}} = \boldsymbol{X}_c\sum_{j=1}^{D}\boldsymbol{v}_{c\_j}\boldsymbol{v}_{c\_j}^{\mathrm{T}}\end{aligned} \tag{28}$$

所以，(24) 可以写成：

$$\hat{X}_c = X_c v_{c\_1} v_{c\_1}^{\mathrm{T}} = X_c v_{c\_1} \otimes v_{c\_1} \tag{29}$$

(26) 则可以写成：

$$X \approx X_c v_{c\_1} \otimes v_{c\_1} + \mathrm{E}(X) \tag{30}$$

如果用第一、二主成分还原 X，上式需要再加一项：

$$X \approx \underbrace{X_c v_{c\_1} \otimes v_{c\_1}}_{\text{First principal}} + \underbrace{X_c v_{c\_2} \otimes v_{c\_2}}_{\text{Second principal}} + \underset{\text{Centroid}}{\mathrm{E}(X)} \tag{31}$$

注意，本书一直强调方差的单位，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。

25.3 格拉姆矩阵

特征值分解

图 11 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 $X^{\mathrm{T}}X/(n-1)$。图 11 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 $V_X^{\mathrm{T}} V_X = V_X V_X^{\mathrm{T}} = I$。$\Lambda_X$ 为对角矩阵，对角线元素为特征值，特征值从大到小排列。图 12 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。

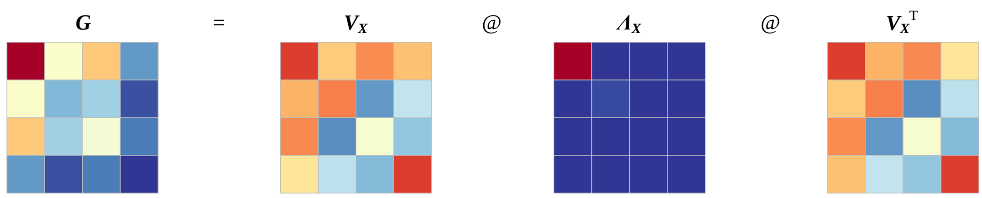

图 11. 特征值分解格拉姆矩阵 G

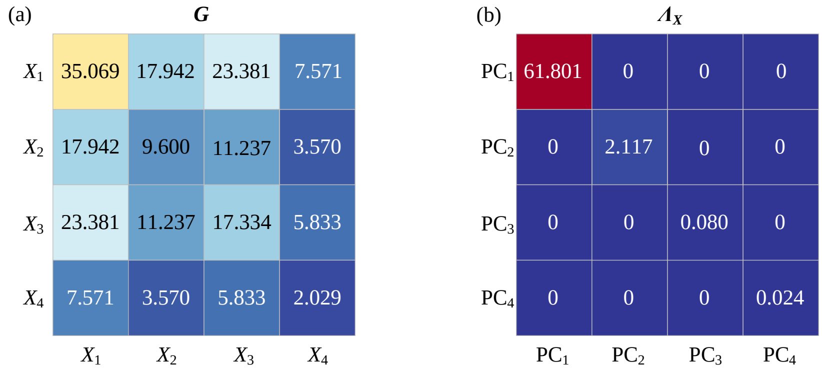

图 12. 对比 $\boldsymbol{G}$ 和 $\boldsymbol{\Lambda}_X$ 热图

分解前后

$\boldsymbol{G}$ 和 $\boldsymbol{\Lambda}_X$ 的主对角线之和相同，即 $\operatorname{trace}(\boldsymbol{G}) = \operatorname{trace}(\boldsymbol{\Lambda}_X)$。如图 13 所示，矩阵 $\boldsymbol{G}$ 的主对角成分为矩阵 $\boldsymbol{X}$ 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 $\operatorname{trace}(\boldsymbol{G})$ 相当于数据整体相对于原点的分散度量。如图 13 所示，矩阵 $\boldsymbol{X}$ 的第一列和第二列贡献最大。经过特征值分解之后，如图 14 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%。

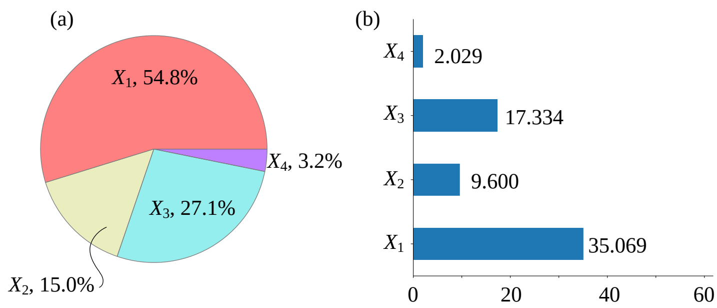

图 13. $\boldsymbol{G}$ 的主对角线成分

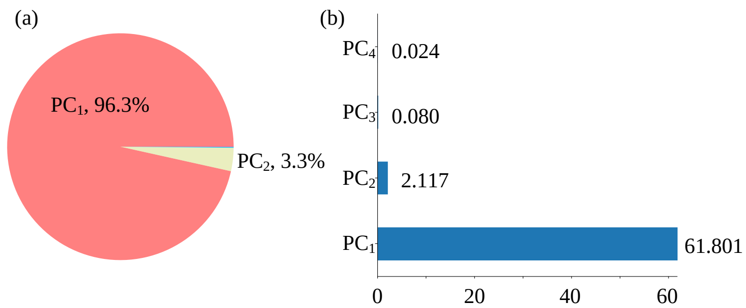

图 14. $\boldsymbol{\Lambda}_X$ 的主对角线成分，格拉姆矩阵 $\boldsymbol{G}$ 的特征值

陡坡图

图 15 所示为在特征值分解格拉姆矩阵 $\boldsymbol{G}$ 主成分分析的陡坡图。

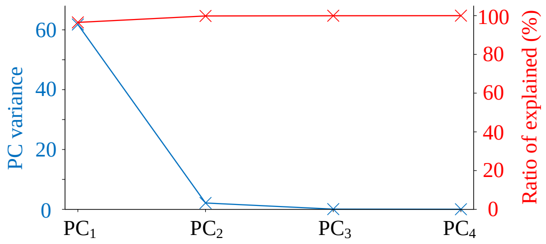

图 15. 陡坡图，特征值分解格拉姆矩阵 $\boldsymbol{G}$

特征向量矩阵

图 16 所示为特征向量矩阵 $\boldsymbol{V}_X$ 热图。显然，图 16 不同于图 7。

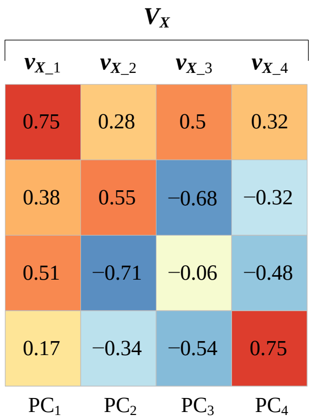

图 16. 特征向量矩阵 $\boldsymbol{V}_X$ 热图

投影

图 17 是将原始数据 $\boldsymbol{X}$ 投影到 $\boldsymbol{V}_X$，即 $\boldsymbol{Y}_X = \boldsymbol{X}\boldsymbol{V}_X$。$\boldsymbol{Y}_X$ 的特点是其格拉姆矩阵为对角方阵，也就是说 $\boldsymbol{Y}_X$ 的列向量两两正交。注意，两两正交不代表线性无关。

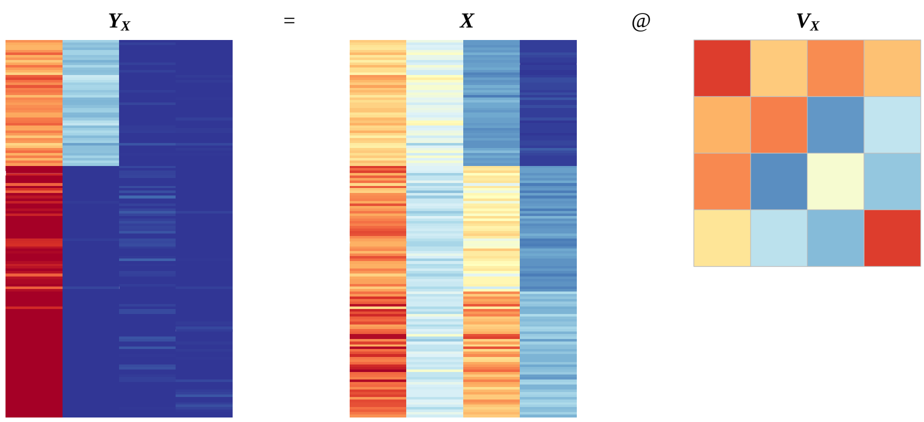

图 17. 将原始数据 $\boldsymbol{X}$ 投影到 $\boldsymbol{V}_X$

正交矩阵 $\boldsymbol{V}_X$ 也是一个规范正交基，$\boldsymbol{V}_X$ 是因原始数据 $\boldsymbol{X}$ 而生。前文提到，$\boldsymbol{V}_c$ 同样是一个规范正交基，但是 $\boldsymbol{V}_c$ 是因中心化数据矩阵 $\boldsymbol{X}_c$ 而生。

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

双标图

图 18 所示为 $\boldsymbol{V}_X$ 的双标图。请大家自行比较图 9 和图 18。

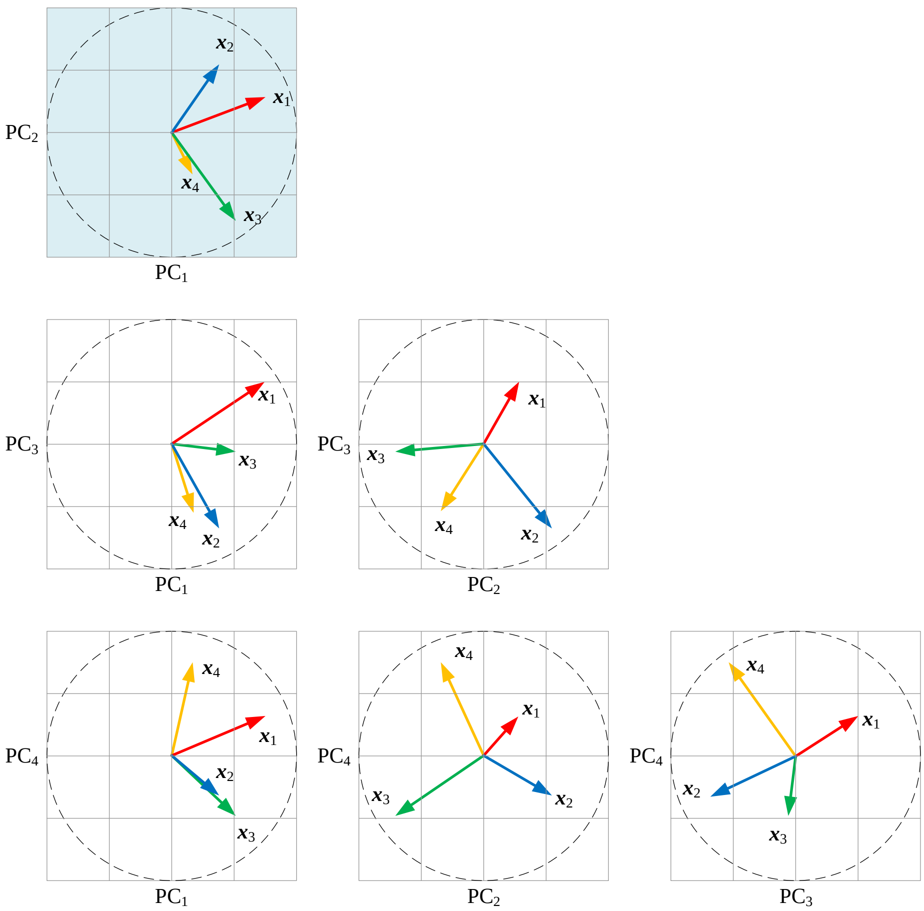

图 18. $\boldsymbol{V}_X$ 双标图，特征值分解格拉姆矩阵 $\boldsymbol{G}$

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 $\boldsymbol{G}$，根据 (1)，利用第一主成分还原原始数据 $\boldsymbol{X}$ 时我们不需要加入质心成分：

$$\boldsymbol{X} \approx \boldsymbol{X}\boldsymbol{v}_{X\_1} \otimes \boldsymbol{v}_{X\_1} \tag{32}$$

如果用第一、二主成分还原 $\boldsymbol{X}$，上式也需要再加一项：

$$\boldsymbol{X} \approx \underbrace{\boldsymbol{X}\boldsymbol{v}_{X\_1} \otimes \boldsymbol{v}_{X\_1}}_{\text{First principal}} + \underbrace{\boldsymbol{X}\boldsymbol{v}_{X\_2} \otimes \boldsymbol{v}_{X\_2}}_{\text{Second principal}} \tag{33}$$

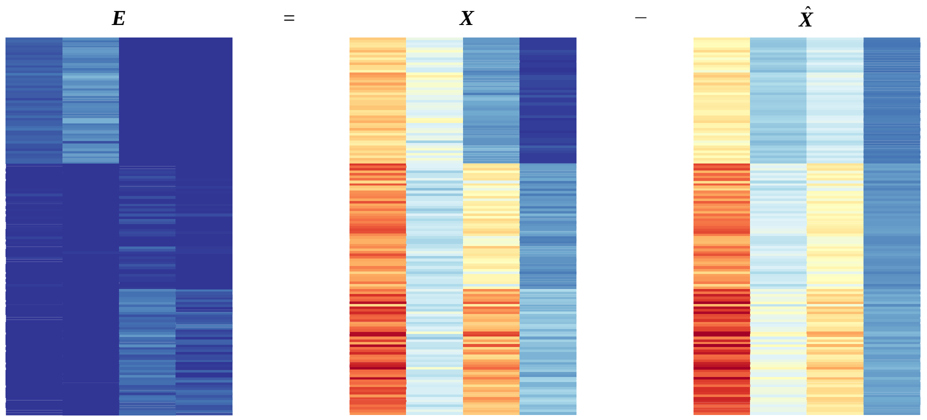

图 19. 第一主成分估计 $\boldsymbol{X}$

25.4 相关性系数矩阵

标准化数据 $\boldsymbol{Z}_X$ 相当于是 z 分数，因此消除了特征量纲影响。因此，特征值分解相关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 20 所示为特征值分解相关性系数矩阵 $\boldsymbol{P}$，$\boldsymbol{P}$ 的主对角线都是 1，$\boldsymbol{P}$ 对角线之外的元素都是线性相关系数。图 21 对比相关性系数矩阵 $\boldsymbol{P}$ 和 $\boldsymbol{\Lambda}_Z$ 热图。同样地，$\boldsymbol{P}$ 和 $\boldsymbol{\Lambda}_Z$ 主对角线之和相同，即 $\text{trace}(\boldsymbol{P}) = \text{trace}(\boldsymbol{\Lambda}_Z)$。

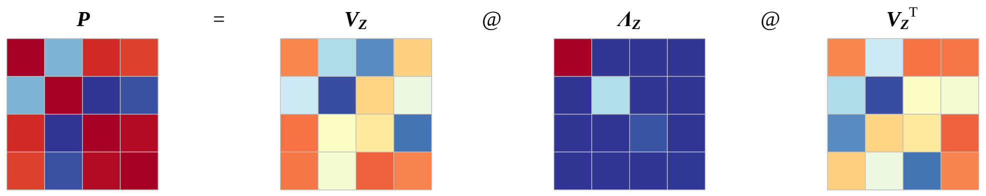

图 20. 特征值分解相关性系数矩阵 $\boldsymbol{P}$

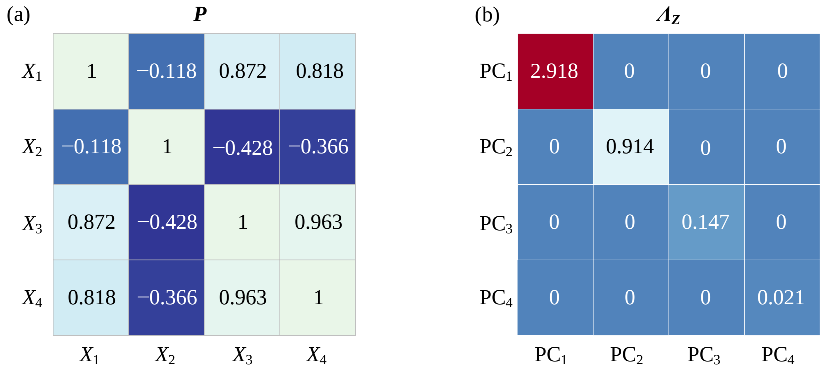

图 21. 对比相关性系数矩阵 $\boldsymbol{P}$ 和 $\boldsymbol{\Lambda}_Z$ 热图

分解前后

图 4 中，X_3 对方差和 trace($\boldsymbol{\Sigma}$) 贡献超过 68%，而 X_3 的贡献小于 5%。而图 22 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。

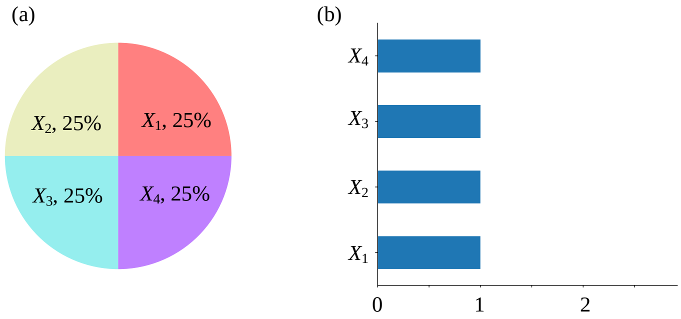

图 22. 相关性系数矩阵 $\boldsymbol{P}$ 主对角线成分

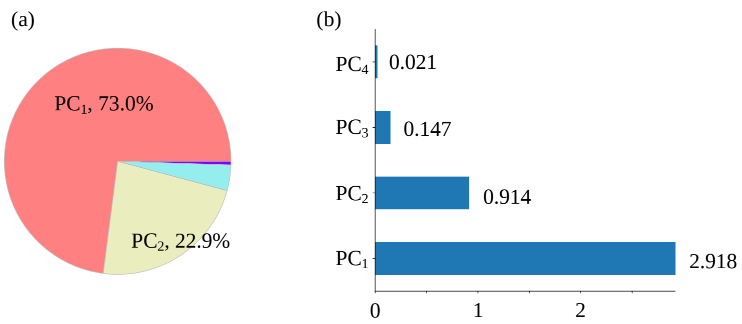

图 23. $\boldsymbol{\Lambda}_Z$ 的主对角线成分，相关性系数矩阵 $\boldsymbol{P}$ 特征值

陡坡图

图 24 所示为特征值分解相关性系数矩阵 $\boldsymbol{P}$ 主成分分析结果陡坡图。第一主成分贡献小于 80%。

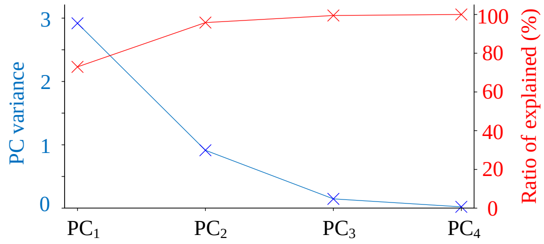

图 24. 陡坡图，特征值分解相关性系数矩阵 $\boldsymbol{P}$

特征向量矩阵

图 25 所示为特征向量矩阵 $\boldsymbol{V}_Z$ 热图。这幅图和图 7、图 16 均不同。

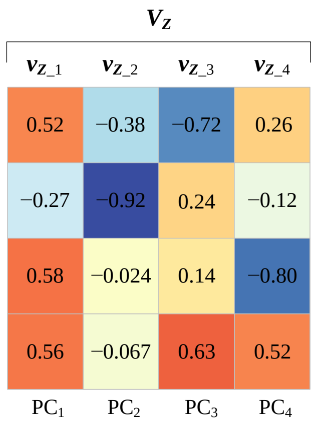

图 25. 特征向量矩阵 $\boldsymbol{V}_Z$ 热图

投影

图 26 所示为标准化数据 $\boldsymbol{Z}$ 投影到 $\boldsymbol{V}_Z$ 得到数据矩阵 $\boldsymbol{Y}_Z$。同样地，正交矩阵 $\boldsymbol{V}_Z$ 也是一个规范正交基，而 $\boldsymbol{V}_Z$ 是因中心化数据 $\boldsymbol{Z}_X$ 而生。

请大家将原数据 $\boldsymbol{X}$、中心化 $\boldsymbol{X}_C$ 也投影到 $\boldsymbol{V}_Z$ 中，并检验结果的协方差矩阵和质心。

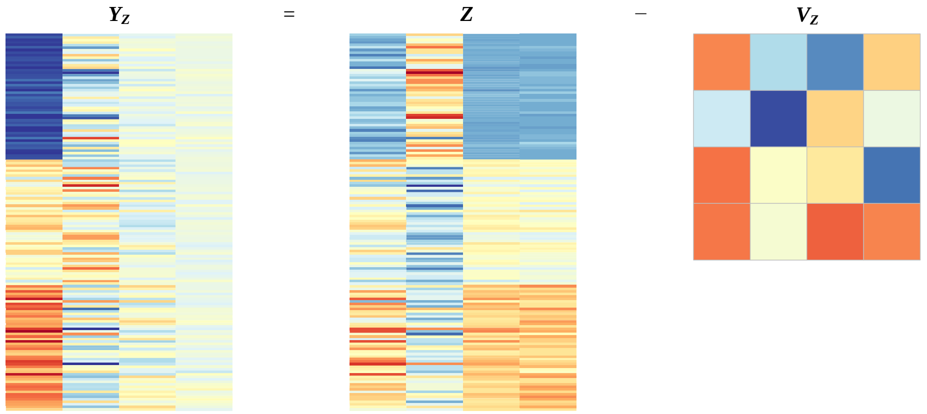

图 26. 中心化数据 Z 投影到 V_Z

双标图

图 27 所示为 V_Z 双标图，请大家比较本章三幅双标图。

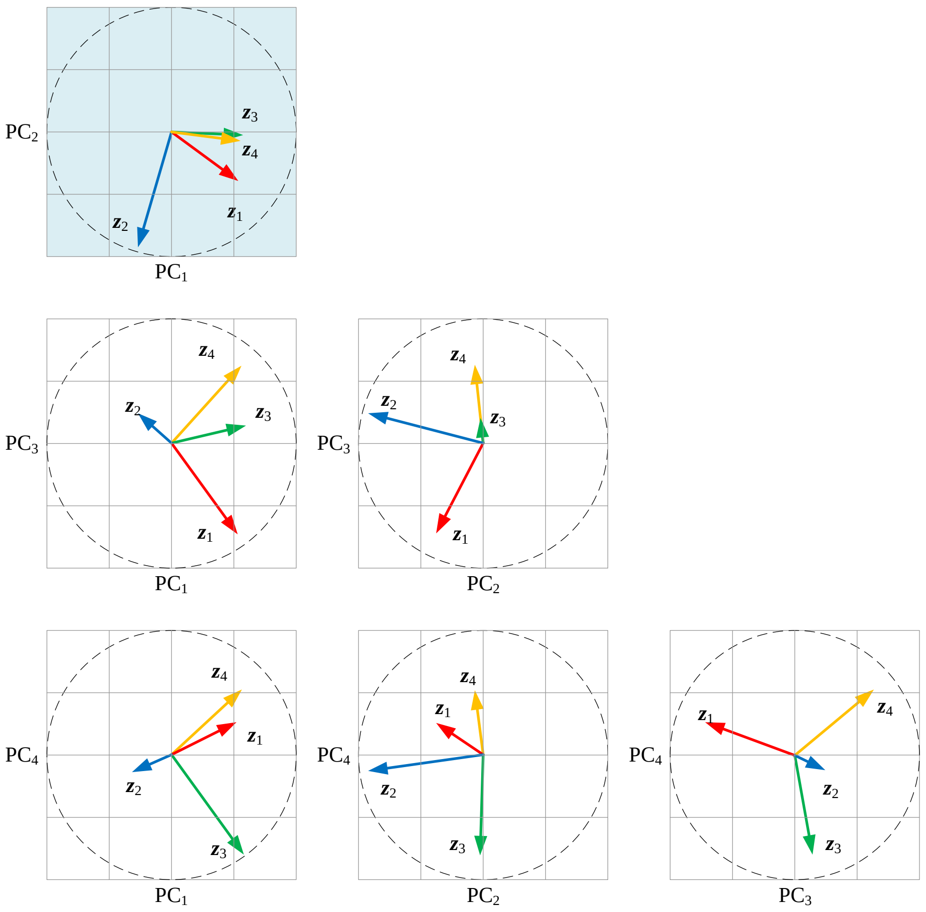

图 27. V_Z 双标图，特征值分解格相关性系数矩阵 P

数据还原、误差

图 28 所示为第一主成分估计 $\boldsymbol{Z}_X$:

$$\boldsymbol{Z}_X \approx \boldsymbol{Z}_X \boldsymbol{v}_{X\_1} \otimes \boldsymbol{v}_{X\_1} \tag{34}$$

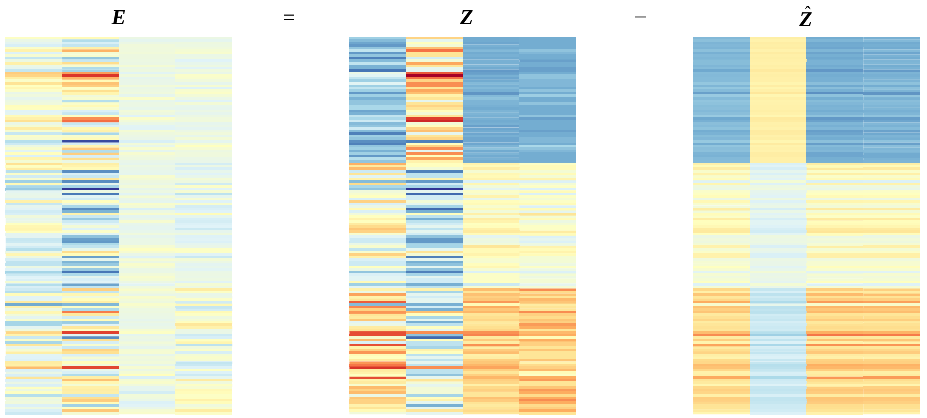

图 28. 第一主成分还原 $\boldsymbol{Z}_X$

$\boldsymbol{Z}_X$ 可以写成:

$$\boldsymbol{Z}_X = \left(\boldsymbol{X} - \mathrm{E}\left(\boldsymbol{X}\right)\right)\boldsymbol{D}^{-1} = \sum_{j=1}^{D} \boldsymbol{Z}_X \boldsymbol{v}_{X\_j} \otimes \boldsymbol{v}_{X\_j} \tag{35}$$

用 $\boldsymbol{V}_Z$ 还原 $\boldsymbol{X}$:

$$\boldsymbol{X} = \left(\sum_{j=1}^{D} \boldsymbol{Z}_X \boldsymbol{v}_{X\_j} \otimes \boldsymbol{v}_{X\_j}\right)\boldsymbol{D} + \mathrm{E}\left(\boldsymbol{X}\right) \tag{36}$$

用 $\boldsymbol{V}_Z$ 第一主成分估计 $\boldsymbol{X}$:

$$\boldsymbol{X} \approx \underbrace{\left(\boldsymbol{Z}_X \boldsymbol{v}_{X\_1} \otimes \boldsymbol{v}_{X\_1}\right)}_{\text{First principal}}\boldsymbol{D} + \mathrm{E}\left(\boldsymbol{X}\right) \tag{37}$$

其中，$\boldsymbol{D}$ 起到缩放的作用，$\mathrm{E}(\boldsymbol{X})$ 是平移的作用。