# 25 Principal Component Analysis

# 主成分分析

## 以概率统计、几何、矩阵分解、优化为视角

> 想象力无边界的人，才创造不可能的事。
>
> ***Those who can imagine anything, can create the impossible.***
>
> —— 艾伦·图灵 (Alan Turing) | 英国计算机科学家、数学家，人工智能之父 | 1912 ~ 1954

- ◄ 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
- ◄ 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.random.multivariate_normal() 产生多元正态分布随机数
- ◄ 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/散点图和边际分布
- ◄ 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

# 25.1 再聊主成分分析

主成分分析 (Principal Component Analysis, PCA) 是重要的降维工具。PCA 可以显著减少数据的维数，同时保留数据中对方差贡献最大的成分。另外对于多维数据，PCA 可以作为一种数据可视化的工具。PCA 还可以用来构造回归模型，这是《数据有道》一册要介绍的内容。

本章将以概率统计、几何、矩阵分解、优化为视角给大家全景展示主成分分析。此外，这一章大家可以把它看成丛书“数学”板块的一个总结。

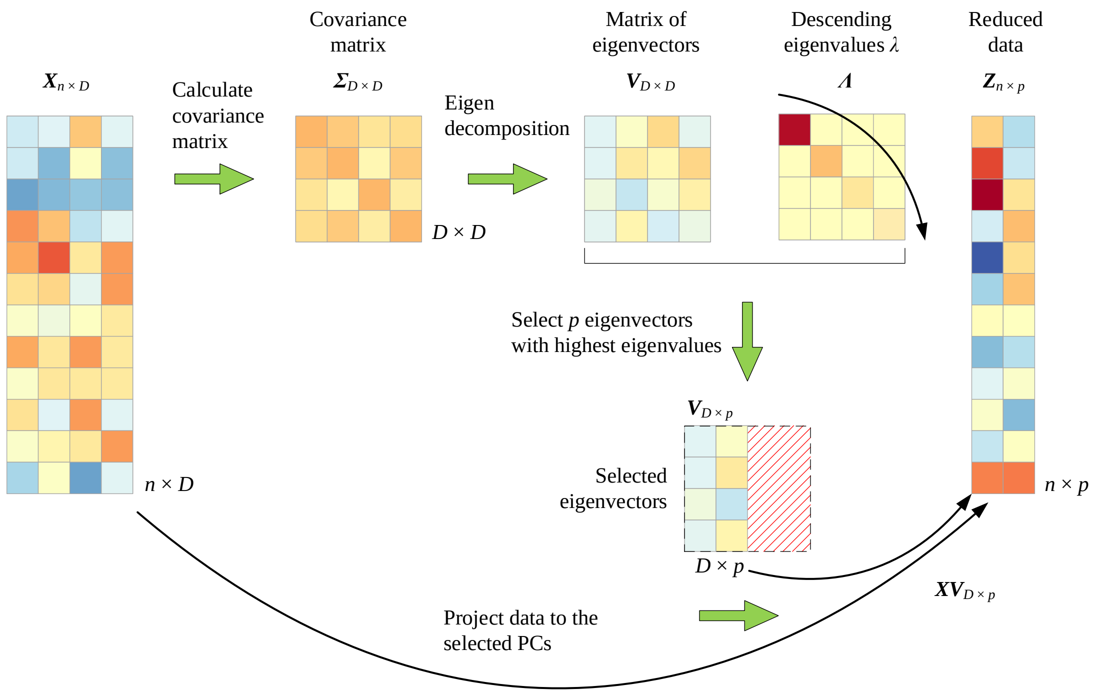


图 1. 主成分分析过程；技术路线：特征值分解协方差矩阵

PCA 的一般步骤如下：

- 对原始数据 $\boldsymbol{X}_{n\times D}$ 的协方差矩阵 $\boldsymbol{\Sigma}_{D\times D}$；
- 计算 $\boldsymbol{\Sigma}$ 特征值 $\lambda_i$ 与特征向量矩阵 $\boldsymbol{V}_{D\times D}$，即因子载荷；
- 对特征值 $\lambda_i$ 从大到小排序，选择其中特征值最大的 $p$ 个特征向量；
- 将原始数据投影到这 $p$ 个正交向量构建的新空间中，得到因子得分 $\boldsymbol{Z}_{n\times p}$。

很多时候，在第一步中，我们直接对原始数据进行标准化 (standardization) 处理，即计算 $\boldsymbol{X}$ 的 z 分数。标准化防止不同特征上方差差异过大。我们在《矩阵力量》第 25 章看到的就是利用标准化数据进行 PCA 分析的技术路线。标准化数据的协方差矩阵实际上就是原数据的相关性系数矩阵。

而有些情况，对原始数据 $X_{n \times D}$ 进行中心化 (去均值) 就足够了，即将数据质心移到原点。

图 1 所示为通过分解协方差矩阵进行主成分分析过程；当然，也可以通过奇异值分解中心化数据 $X_c$ 进行主成分分析。本章将主要介绍通过特征值分解协方差矩阵进行主成分分析。

## 25.2 原始数据

《矩阵力量》介绍过，样本数据矩阵 $X$ 可以分别通过行和列来解释。矩阵 $X$ 每一列代表一个特征向量：

$$X = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \tag{1}$$

$X$ 矩阵每一行代表一个样本。比如，$X$ 矩阵第一行对应是第一个数据点，它写成一个行向量 $x^{(1)}$：

$$x^{(1)} = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} \end{bmatrix} \tag{2}$$

图 2 展示原始数据矩阵 $X$ 热图，红色色系代表正数，蓝色色系代表负数，黄色接近 0。$X$ 矩阵有 12 行，即 12 个样本；$X$ 矩阵有 4 列，即 4 个特征。

注意，本例中假设 $X$ 已经中心化 $\mathrm{E}(X) = 0$，即质心位于原点。

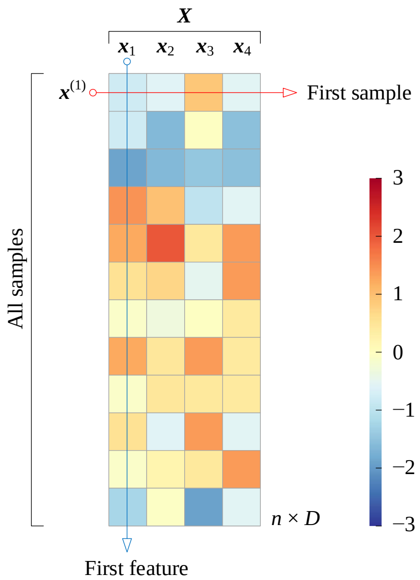


图 2. 原始数据 $X$ 热图，$D = 4$，$n = 12$，$X$ 已经去均值

### 分布特征

图 3 所示为矩阵 $\boldsymbol{X}$ 每一列特征数据的分布情况；可以发现它们之间的均方差区别不大。但是经过主成分分解之后，大家可以明显发现每一列新特征数据均方差大小完全不同。

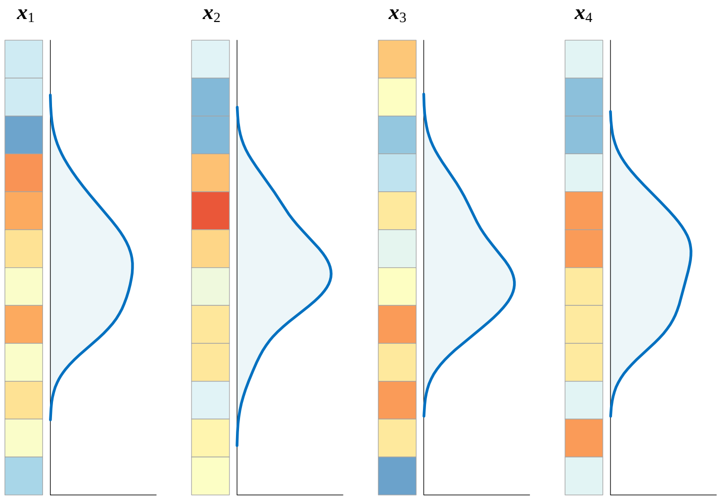


图 3. $\boldsymbol{X}$ 四个特征向量数据分布

## 25.3 特征值分解协方差矩阵

本书第 13 章介绍过，$\boldsymbol{X}$ 的协方差矩阵 $\boldsymbol{\Sigma}$ 可以通过下式计算得到：

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

其中，$\mathrm{E}(\boldsymbol{X})$ 也常被称作原始数据 $\boldsymbol{X}$ 的质心；$\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})$ 相当于数据中心化。当 $n$ 足够大，(3) 的分母可以用 $n$ 替换。本例设定 $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{0}^{\mathrm{T}}$，即 $\boldsymbol{X} = \boldsymbol{X}_c$。

如图 5 所示，$\boldsymbol{\Sigma}$ 为实数对称矩阵，它的特征值分解 (谱分解) 可以写作：

$$\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}} \tag{4}$$

$\boldsymbol{V}$ 为正交矩阵。$\boldsymbol{V}$ 和自己转置 $\boldsymbol{V}^{\mathrm{T}}$ 乘积为单位阵 $\boldsymbol{I}$，即：

$$\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{I} \tag{5}$$

特征值方阵 $\boldsymbol{\Lambda}$ 主对角线元素为特征值 $\lambda$，特征值从大到小排列：

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_D \end{bmatrix}, \quad \lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_D \tag{6}$$

本书前文介绍过，从统计学角度来讲，$\lambda_j$ 是第 $j$ 个主成分所贡献的方差。

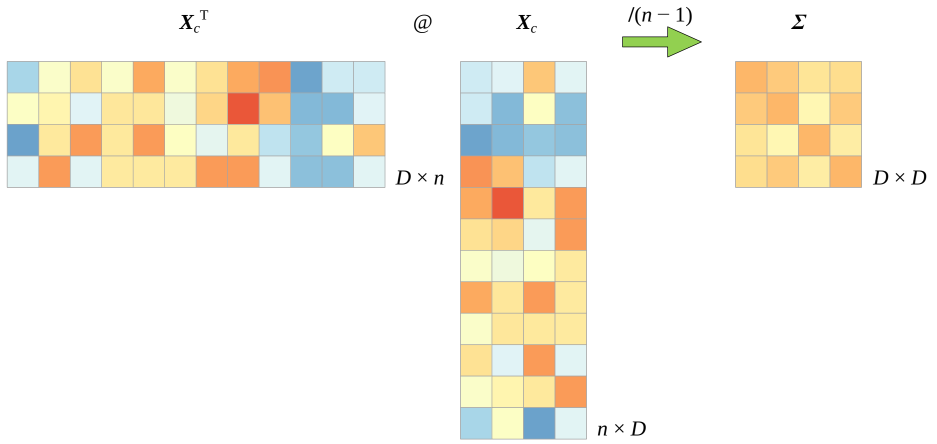


图 4. 计算原始数据协方差矩阵，$D = 4$，$n = 12$

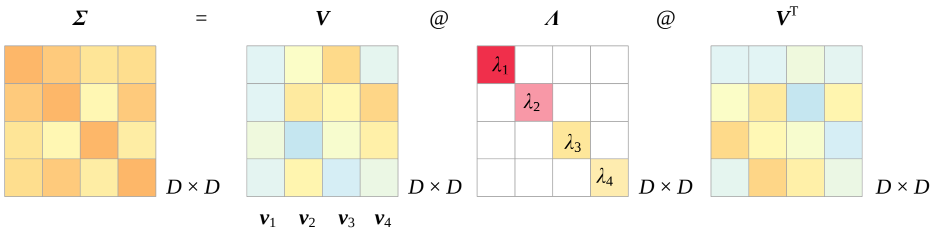


图 5. 协方差矩阵特征值分解，$D = 4$

## 主成分、因子载荷

$V$ 为特征向量构造的 $D \times D$ 的方阵：

$$V = \begin{bmatrix} \underset{\text{PC1}}{v_1} & \underset{\text{PC2}}{v_2} & \cdots & v_D \end{bmatrix} = \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,D} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ v_{D,1} & v_{D,2} & \cdots & v_{D,D} \end{bmatrix} \tag{7}$$

$v_1$ 被称作第一主成分 (first principal component)，本书常记做 PC1；$v_2$ 被称作第二主成分 (second principal component)，记做 PC2；以此类推。

$V$ 的列向量也叫因子载荷 (loadings)。注意，有些文献中因子载荷为：

$$V\sqrt{\Lambda} = \begin{bmatrix} v_1 & v_2 & \cdots & v_D \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} & & & \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_D} \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda_1} v_1 & \sqrt{\lambda_2} v_2 & \cdots & \sqrt{\lambda_D} v_D \end{bmatrix} \tag{8}$$

## 迹，总方差

本书前文介绍过，协方差矩阵 $\boldsymbol{\Sigma}$ 的迹 trace($\boldsymbol{\Sigma}$) 等于的特征值方阵 $\boldsymbol{\Lambda}$ 迹 trace($\boldsymbol{\Lambda}$)：

$$\operatorname{trace}(\boldsymbol{\Sigma})=\sigma_1^2+\sigma_2^2+\cdots+\sigma_D^2=\sum_{j=1}^{D}\sigma_j^2=\operatorname{trace}(\boldsymbol{\Lambda})=\lambda_1+\lambda_2+\cdots\lambda_D=\sum_{j=1}^{D}\lambda_j \tag{9}$$

第 $j$ 个特征值 $\lambda_j$ 对方差总和 (total variance) 的贡献百分比为：

$$\frac{\lambda_j}{\sum_{i=1}^{D}\lambda_i}\times 100\% \tag{10}$$

前 $p$ 个特征值，即 $p$ 个主成分解释总方差 (total variance explained) 的百分比为：

$$\frac{\sum_{j=1}^{p}\lambda_j}{\sum_{i=1}^{D}\lambda_i}\times 100\% \tag{11}$$

主成分分析中，我们常用陡坡图 (scree plot) 可视化这个百分比，《数据有道》一册中大家会看到很多实例。

## 25.4 投影

本节从投影角度介绍 PCA。数据矩阵 $\boldsymbol{X}_c$ 投影到矩阵 $\boldsymbol{V}$ 正交系 ($\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_D$) 得到新特征数据矩阵 $\boldsymbol{Z}$，即：

$$\boldsymbol{Z}=\boldsymbol{X}\boldsymbol{V} \tag{12}$$

$\boldsymbol{V}$ 常被称作**因子载荷** (factor loadings)，$\boldsymbol{Z}$ 常被称作**因子得分** (factor score)。图 6 所示 $\boldsymbol{Z}=\boldsymbol{X}\boldsymbol{V}$ 矩阵运算原理图。《矩阵力量》第 10 章特别介绍过这种数据投影，建议大家回顾。

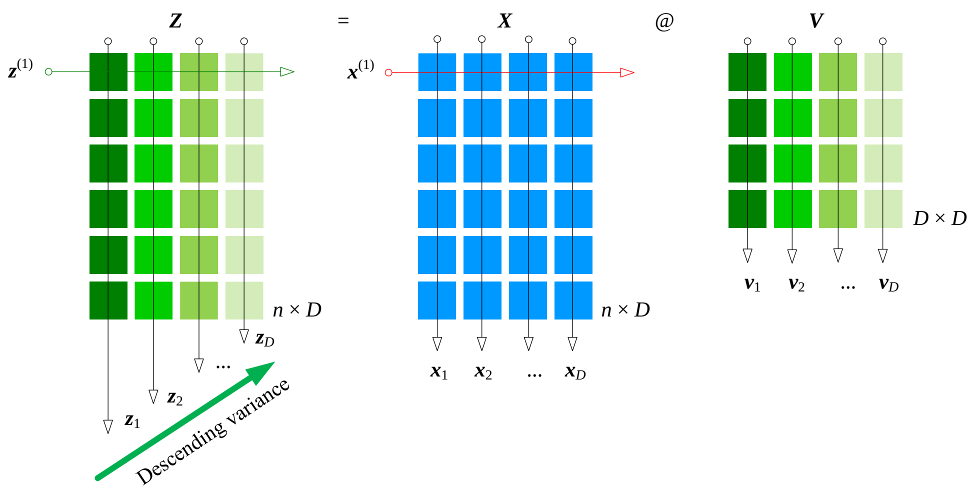


图 6. PCA 分解数据关系 $\boldsymbol{Z}=\boldsymbol{X}\boldsymbol{V}$

图 7 所示为将图 2 给出数据矩阵 $\boldsymbol{X}$ 投影到矩阵 $\boldsymbol{V}$，得到新特征数据矩阵 $\boldsymbol{Z}$。

值得强调的一点是，如果 $\boldsymbol{X}$ 没有中心化，把原始数据 $\boldsymbol{X}$ 或中心化数据 $\boldsymbol{X}_c$ 投影到 $\boldsymbol{V}$ 中结果不一样。从统计角度来看，差异主要体现在质心位置，而投影得到的数据协方差矩阵相同。

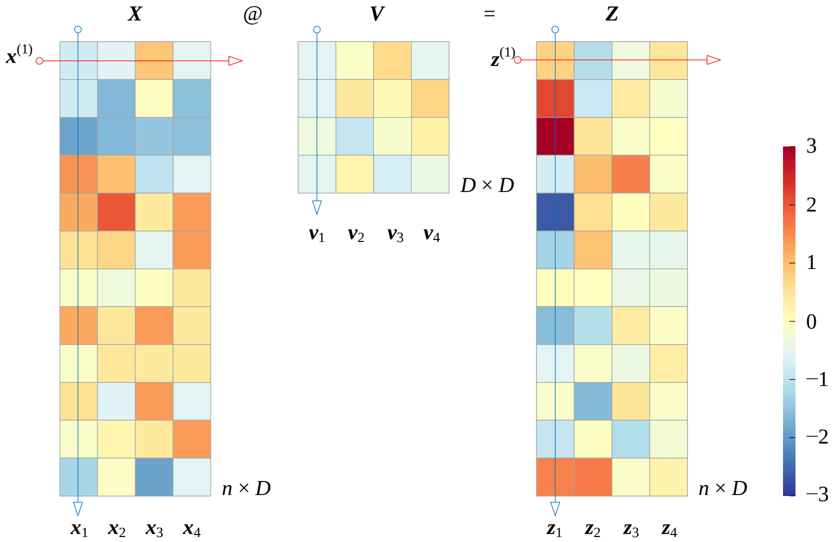


图 7. $\boldsymbol{Z}$、$\boldsymbol{X}$ 和 $\boldsymbol{V}$ 这三个矩阵关系和热图

## $\boldsymbol{Z}$的列向量

前文讨论过，矩阵 $\boldsymbol{X}$ 每一列特征数据方差区别不大 (见图 3)；而图 8 告诉我们，经过 PCA 分解得到的矩阵 $\boldsymbol{Z}$ 四个新特征数据分布差异巨大。

如图 8 所示，第一列 $\boldsymbol{z}_1$ 数据分布最为分散，也就是第一主成分 (first principal component) 解释了数据中最多方差。第一列 $\boldsymbol{z}_1$ 到第四列 $\boldsymbol{z}_4$ 数据分散情况逐渐降低，热图对应的色差从明显到模糊。

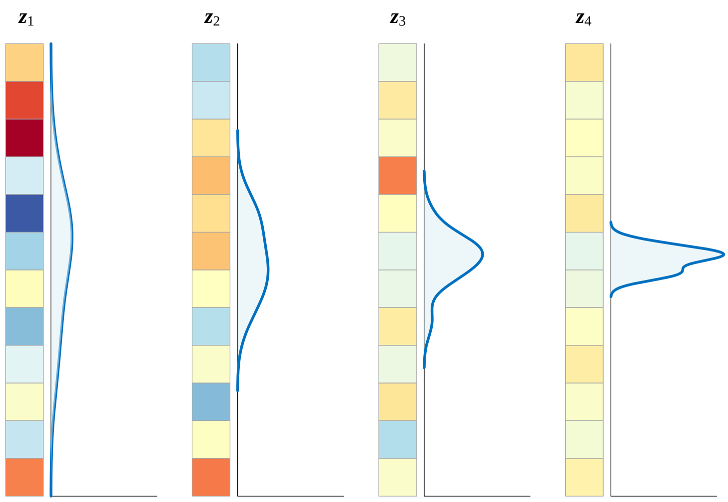

图 8. $Z$ 四个新特征数据分布

将 (12) 展开得到：

$$\begin{bmatrix} z_1 & z_2 & \cdots & z_D \end{bmatrix} = X \begin{bmatrix} \underset{\text{PC1}}{v_1} & \underset{\text{PC2}}{v_2} & \cdots & v_D \end{bmatrix} \tag{13}$$

由此，得到图 9 所示主成分分析运算的数据关系：

$$\begin{cases} z_1 = Xv_1 \\ z_2 = Xv_2 \\ \quad \vdots \\ z_D = Xv_D \end{cases} \tag{14}$$

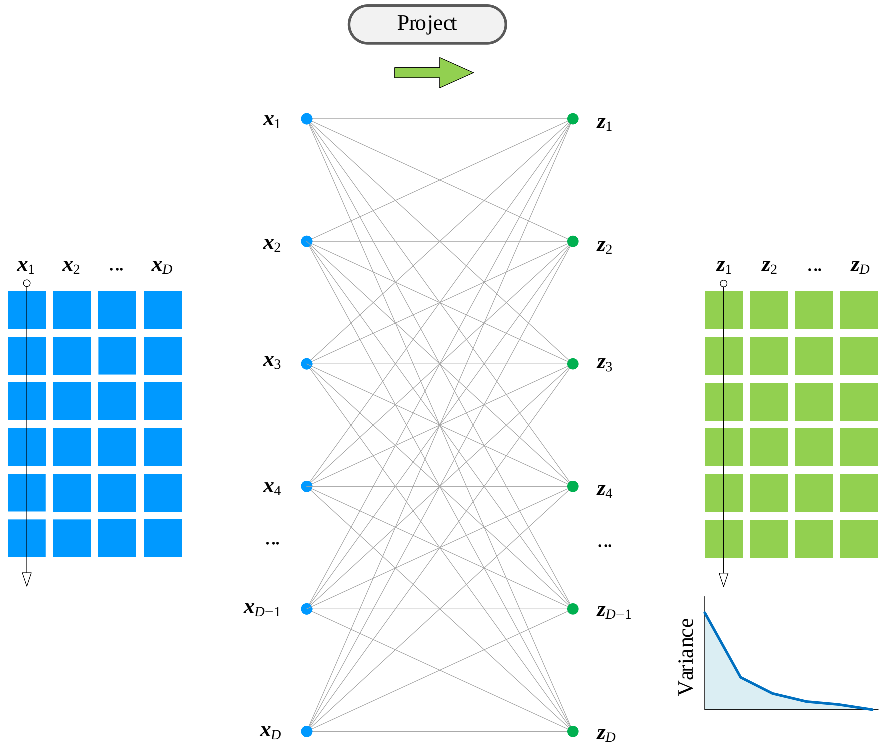


图 9. PCA 运算数据关系

注意，$z_i$ 和 $z_j$ $(i \neq j)$ 相互正交：

$$\begin{aligned} z_i^{\mathrm{T}} z_j &= \left(Xv_i\right)^{\mathrm{T}} Xv_j = v_i^{\mathrm{T}} X^{\mathrm{T}} Xv_j = \left(n-1\right) v_i^{\mathrm{T}} \Sigma v_j \\ &= \left(n-1\right) v_i^{\mathrm{T}} \Sigma v_j = \left(n-1\right) v_i^{\mathrm{T}} V \Lambda V^{\mathrm{T}} v_j = 0 \end{aligned} \tag{15}$$

## 线性组合

如图 10 所示，列向量 $\boldsymbol{v}_1$ 每一个元素相当于原始数据 $\boldsymbol{X}$ 每一列 $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D]$ 对应的系数。以第一主成分数据 $\boldsymbol{z}_1$ 为例，将 $\boldsymbol{X}$ 向 $\boldsymbol{v}_1$ 投影，展开得到：

$$\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{v}_1 \tag{16}$$

(16) 展开得到：

$$\boldsymbol{z}_1 = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_D \end{bmatrix} \underbrace{\begin{bmatrix} v_{1,1} \\ v_{2,1} \\ \vdots \\ v_{D,1} \end{bmatrix}}_{v_1,\ \mathrm{PC1}} = v_{1,1}\boldsymbol{x}_1 + v_{2,1}\boldsymbol{x}_2 + \cdots + v_{D,1}\boldsymbol{x}_D \tag{17}$$

白话讲，$\boldsymbol{z}_1$ 相当于 $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D]$ 每一列的混合成分，即线性组合。

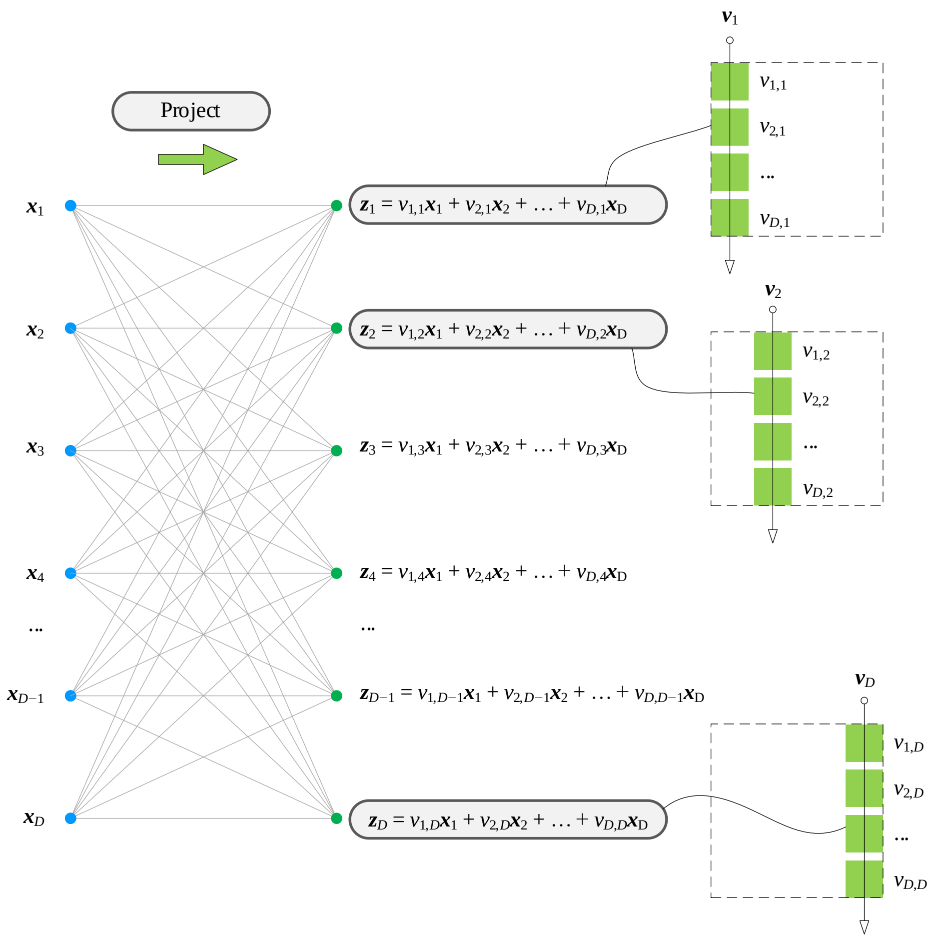


图 10. $\boldsymbol{V}$ 每一列元素相当于原始数据 $\boldsymbol{X}$ 每一列系数

## 朝向量投影

图 11 所示 $\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{v}_1$ 运算相当于数据 $\boldsymbol{X}$ 向 $\boldsymbol{v}_1$ 向量 (第一主成分) 投影获得 $\boldsymbol{z}_1$，即一个四特征数据 $\boldsymbol{X}$ 投影到 $\boldsymbol{v}_1$ 得到一维新特征数据。图 12 展示 $\boldsymbol{z}_2 = \boldsymbol{X}\boldsymbol{v}_2$ 运算等价于数据 $\boldsymbol{X}$ 向 $\boldsymbol{v}_2$ (第二主成分) 投影获得 $\boldsymbol{z}_2$。图 11 ~ 图 14 分别展示数据矩阵 $\boldsymbol{X}$ 向 $\boldsymbol{v}_1$、$\boldsymbol{v}_2$、$\boldsymbol{v}_3$ 和 $\boldsymbol{v}_4$ 向量投影。

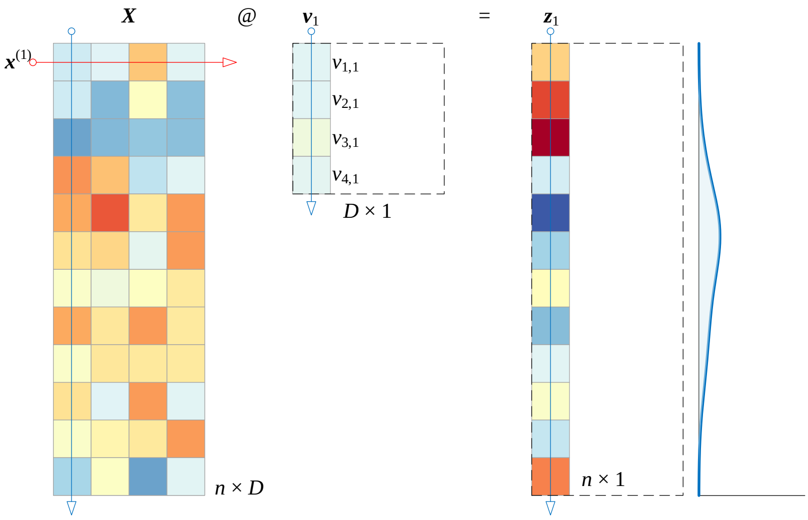


图 11. 数据 $\boldsymbol{X}$ 向 $\boldsymbol{v}_1$ 向量投影

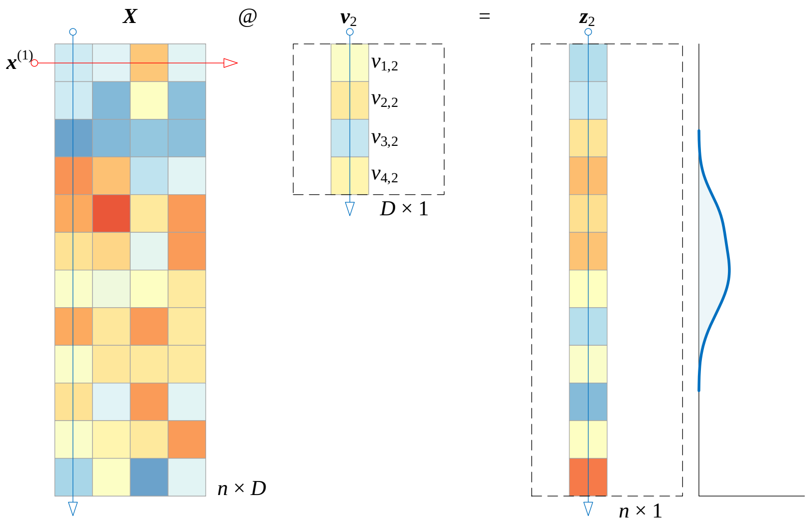


图 12. 数据 $\boldsymbol{X}$ 向 $\boldsymbol{v}_2$ 向量投影

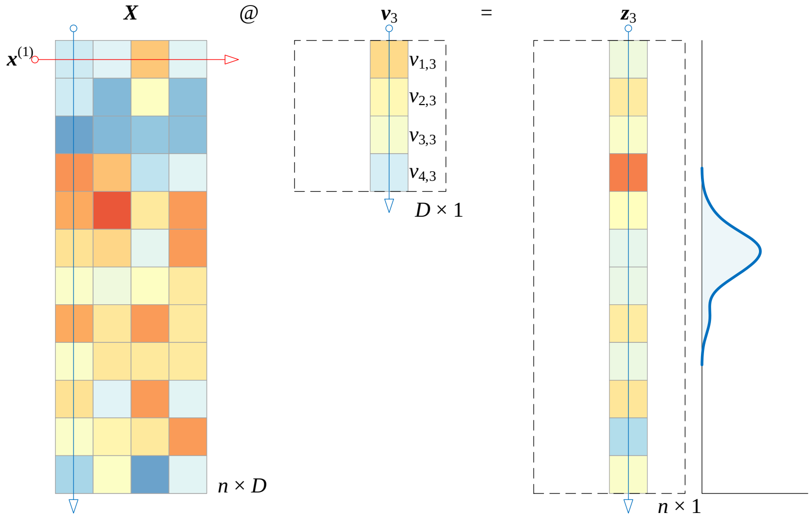


图 13. 数据 $\boldsymbol{X}$ 向 $\boldsymbol{v}_3$ 向量投影

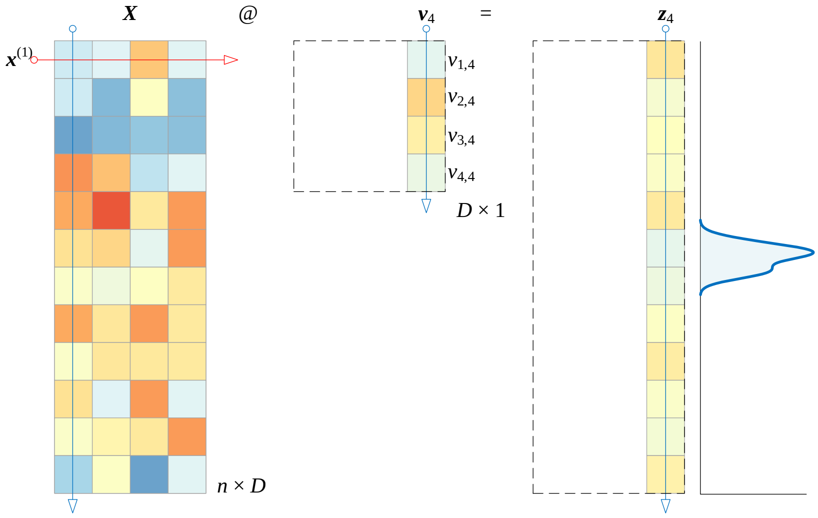


图 14. 数据 $\boldsymbol{X}$ 向 $\boldsymbol{v}_4$ 向量投影

## 朝平面投影

同样，$[\boldsymbol{z}_1, \boldsymbol{z}_2]$ 是数据 $\boldsymbol{X}$ 向 $[\boldsymbol{v}_1, \boldsymbol{v}_2]$ 投影结果，即相当于四维数据 $\boldsymbol{X}$ 向正交平面投影。运算过程如下：

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

图 15 所示为 (18) 运算过程及结果热图。

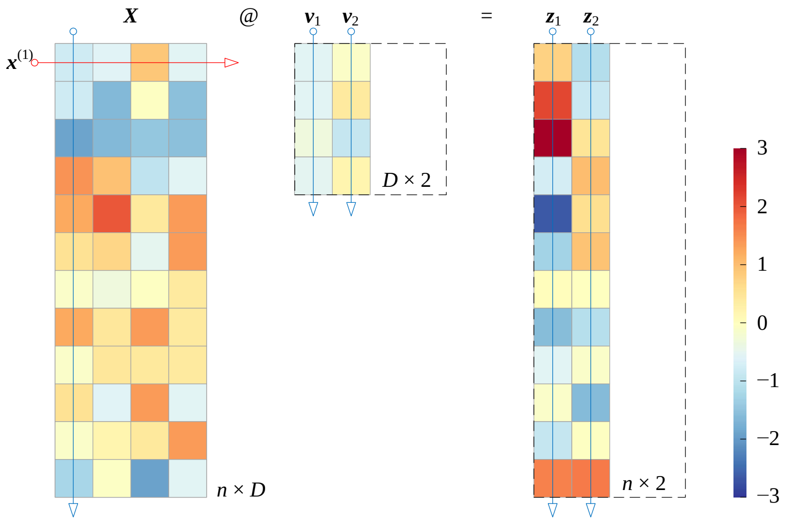


图 15. 数据 $\boldsymbol{X}$ 向 $[\boldsymbol{v}_1, \boldsymbol{v}_2]$ 投影

## $\boldsymbol{Z}$ 的协方差矩阵

前文假设 $\boldsymbol{X}$ 已经中心化，因此 $\boldsymbol{z}_1$ 的期望值为 0。对 $\boldsymbol{z}_1$ 求方差，可以得到：

$$\operatorname{var}(\boldsymbol{z}_1)=\frac{(\boldsymbol{X}\boldsymbol{v}_1)^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{v}_1)}{n-1}=\frac{\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{v}_1}{n-1}=\boldsymbol{v}_1^{\mathrm{T}}\underbrace{\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{n-1}}_{\boldsymbol{\Sigma}}\boldsymbol{v}_1=\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_1 \tag{19}$$

类似地，

$$\operatorname{var}(\boldsymbol{z}_2)=\boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_2,\quad \ldots,\quad \operatorname{var}(\boldsymbol{z}_D)=\boldsymbol{v}_D^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_D \tag{20}$$

这样，整个新特征数据 $\boldsymbol{Z}$ 的协方差矩阵，可以通过下式计算得到：

$$\begin{aligned}\operatorname{var}(\boldsymbol{Z})&=\frac{(\boldsymbol{X}\boldsymbol{V})^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{V})}{n-1}=\frac{\boldsymbol{V}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{V}}{n-1}\\&=\boldsymbol{V}^{\mathrm{T}}\underbrace{\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{n-1}}_{\boldsymbol{\Sigma}}\boldsymbol{V}=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{V}=\begin{bmatrix}\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_1 & & & \\ & \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{v}_D^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_D\end{bmatrix}=\boldsymbol{\Lambda}=\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_D\end{bmatrix}\end{aligned} \tag{21}$$

观察 (21) 所示协方差矩阵，可以发现主对角线以外元素均为 0，也就是 $\boldsymbol{Z}$ 的列向量两两正交，线性相关系数为 0。

$\boldsymbol{Z}_{n\times p}$ 的协方差矩阵为：

$$\operatorname{var}(\boldsymbol{Z}_{n\times p})=\frac{(\boldsymbol{X}\boldsymbol{V}_{D\times p})^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{V}_{D\times p})}{n-1}=\boldsymbol{V}_{D\times p}^{\mathrm{T}}\underbrace{\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{n-1}}_{\boldsymbol{\Sigma}}\boldsymbol{V}_{D\times p}=\boldsymbol{V}_{D\times p}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{V}_{D\times p}=\boldsymbol{\Lambda}_{p\times p}=\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_p\end{bmatrix} \tag{22}$$

对于投影数据的方差计算，我们已经在本书第 14 章详细介绍过，请感兴趣的读者自行回顾复习。

## 25.5 几何视角看 PCA

如图 16 所示，椭圆中心对应质心 $\boldsymbol{\mu}$，椭圆和±$\sigma$ 标准差构成的正方形相切，四个切点分别为 $A$、$B$、$C$ 和 $D$，对角切点两两相连得到两条直线 $AC$、$BD$。

$AC$ 相当于在给定 $X_2$ 条件下 $X_1$ 的条件概率期望值；$BD$ 相当于在给定 $X_1$ 条件下 $X_2$ 的条件概率期望值。

图 16 中，$EF$ 为椭圆长轴；$FH$ 为椭圆短轴。而 $EF$ 就相当于本章介绍的第一主成分，$FH$ 为第二主成分。

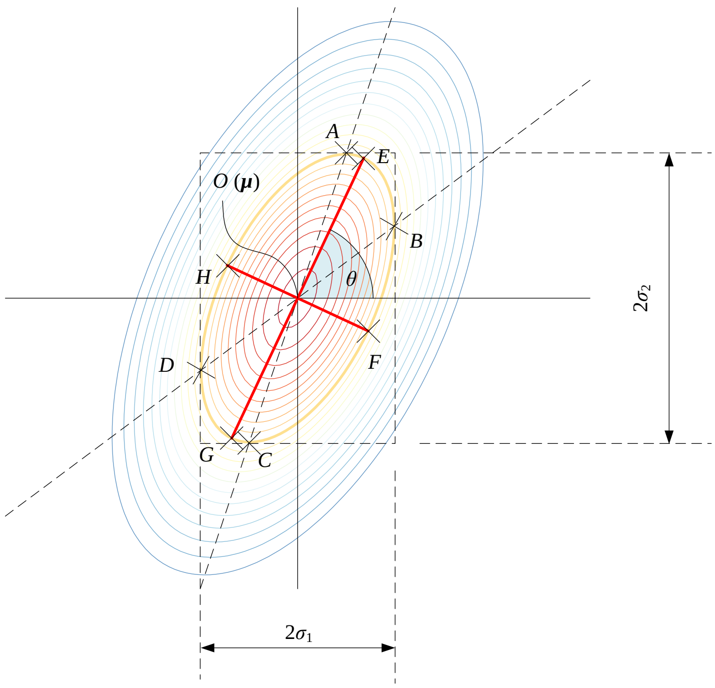


图 16. 椭圆和±$\sigma$ 标准差长方形的关系

$\boldsymbol{\Sigma}$ 特征值分解进行主成分分析的具体步骤如图 17 所示。假设图 17 原始数据已经标准化，计算得到协方差矩阵 $\boldsymbol{\Sigma}$，找到 $\boldsymbol{\Sigma}$ 对应椭圆的半长轴所在方向 $\boldsymbol{v}_1$。$\boldsymbol{v}_1$ 对应的便是第一主成分 (first principal component)。原始数据朝 $\boldsymbol{v}_1$ 投影得到的数据对应最大方差。

整个过程实际上用到了我们在丛书《矩阵力量》一本中介绍的平移、缩放、正交化、投影、旋转等线性变换操作。

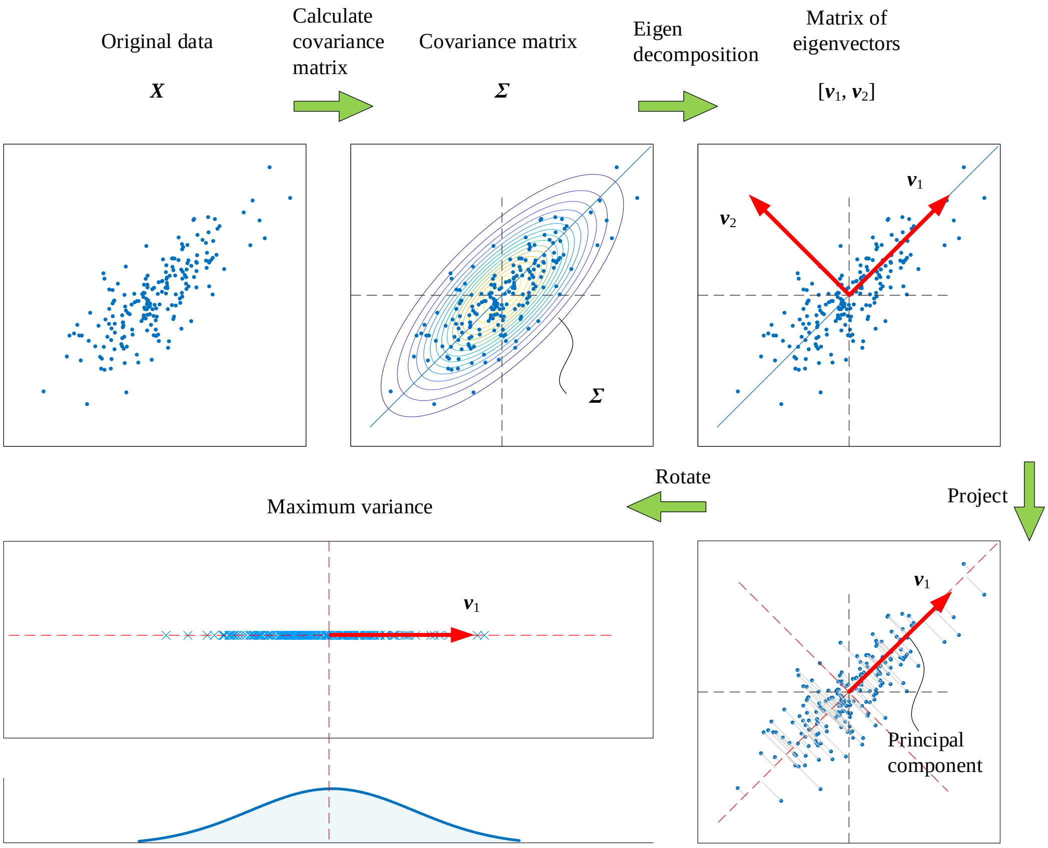


图 17. 几何视角下通过特征值分解协方差矩阵进行主成分分析

从线性变换角度来看，主成分分析无非就是，在不同的坐标系中看同一组数据。如图 18 所示，数据朝不同方向投影会得到不同的投影结果，对应不同的分布；朝椭圆长轴方向投影，得到的数据均方差最大；朝椭圆短轴方向投影得到的数据均方差最小。

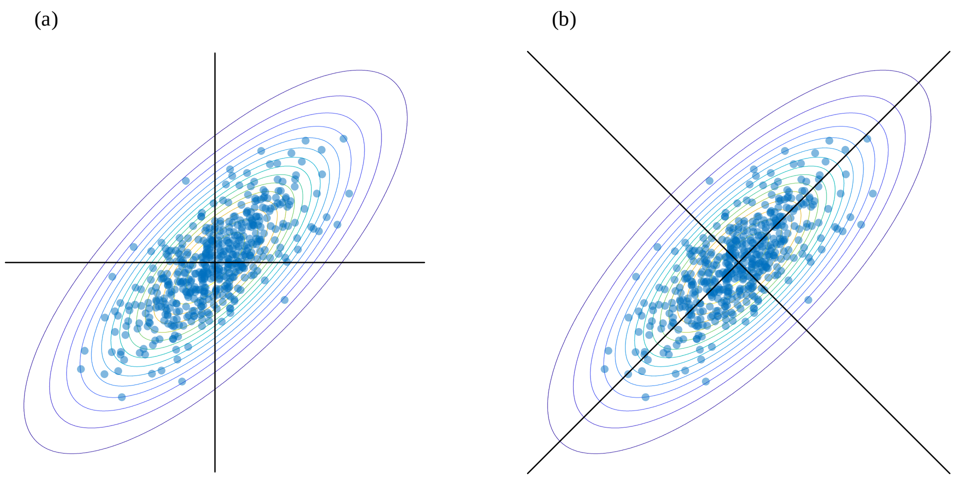


图 18. 两个角度看数据

## 举个例子

图 19 所示为原始二维数据 $\boldsymbol{X}$，可以发现数据的质心位于 $[1, 2]^T$。分析数据 $\boldsymbol{X}$，可以发现数据的两个特征上分布分散情况相似，也就是方差大小几乎相同。

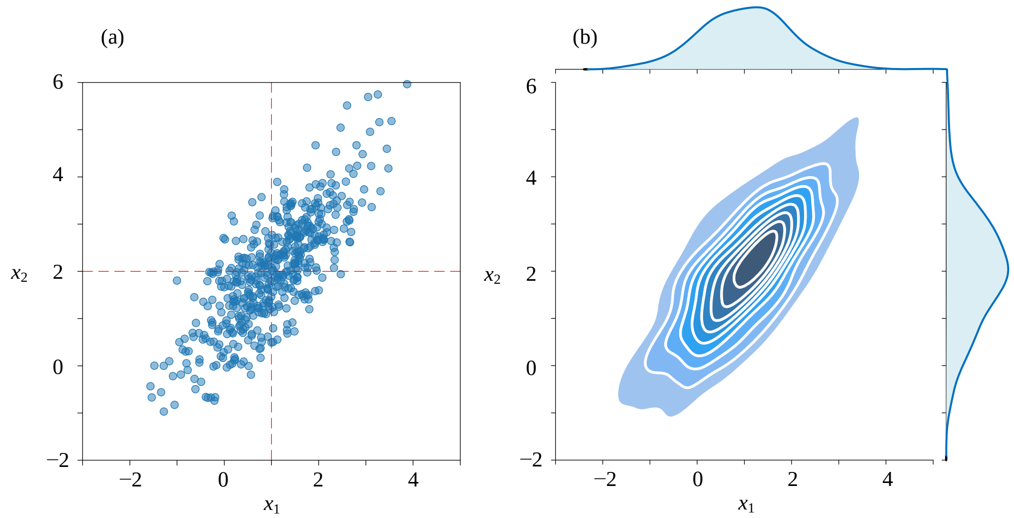


图 19. 原始二维数据 $\boldsymbol{X}$

利用 sklearn.decomposition.PCA() 函数，我们可以通过 pca.components_获得主成分向量。利用 pca.transform($\boldsymbol{X}$) 可以获得投影后的数据 $\boldsymbol{Y}$。图 20 对比 $\boldsymbol{Y}$ 两列数据分布。图 21 所示为数据 $\boldsymbol{Y}$ 在 $[\boldsymbol{v}_1, \boldsymbol{v}_2]$ 中散点图。

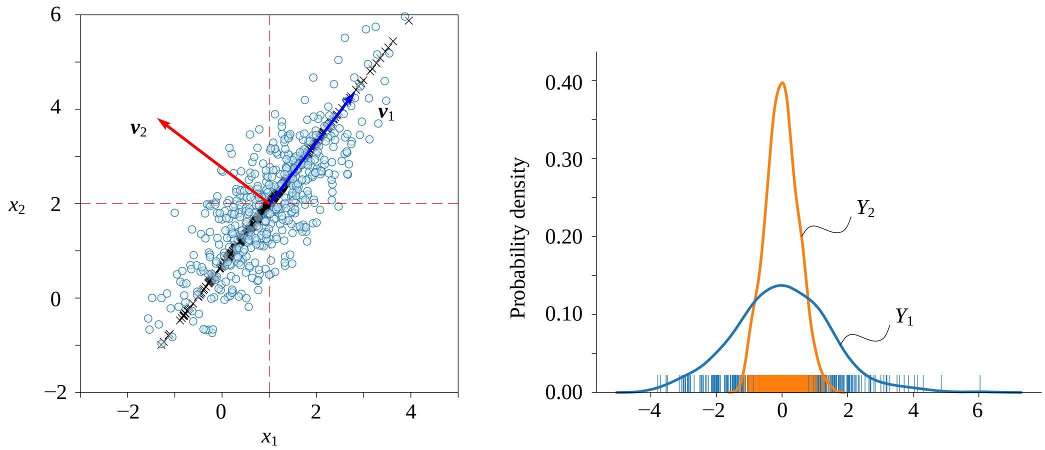


图 20. 主成分数据分布

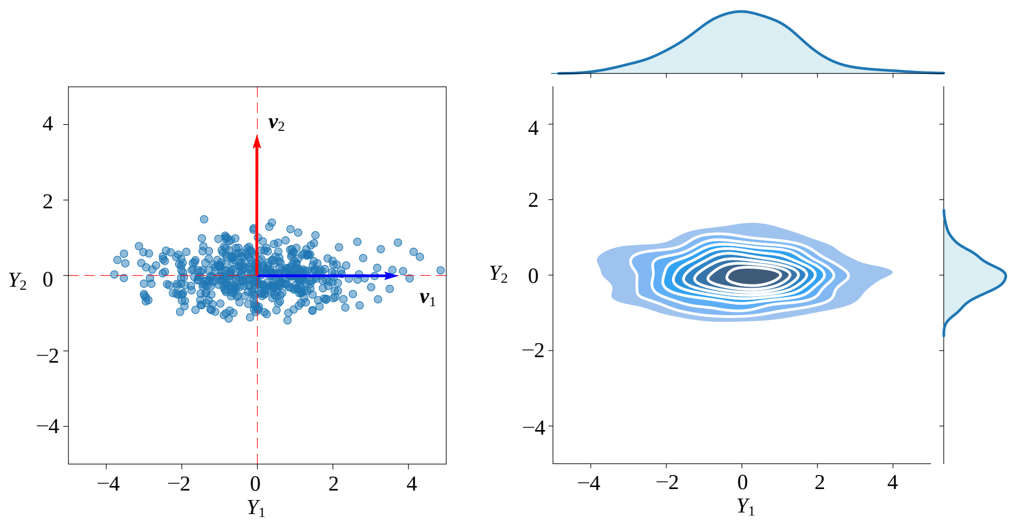


图 21. 数据 $\boldsymbol{Y}$ 在 $[\boldsymbol{v}_1, \boldsymbol{v}_2]$ 中散点图

| |
|---|
| Bk5_Ch25_01.py 绘制图 19 ~ 图 21。 |

# 25.6 奇异值分解

## 四种奇异值分解

从书在《矩阵力量》一本系统讲解过，奇异值分解有四种形式：

- **完全型** (full)
- **经济型** (economy-size, thin)
- **紧凑型** (compact)
- **截断型** (truncated)

如图 22 所示，完全型奇异值分解 $\boldsymbol{S}$ 矩阵并非方阵，相当于主成分特征值方阵 $\boldsymbol{S}$ 下面叠加一块全 0 矩阵 $\boldsymbol{O}$。

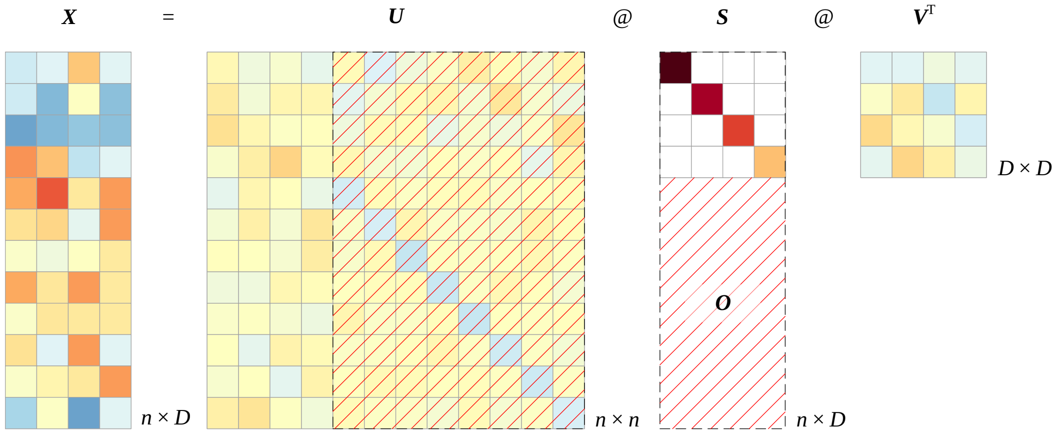


图 22. 完全 (full) 奇异值分解

去掉图 22 中这个全 0 矩阵 $\boldsymbol{O}$，便得到经济型奇异值分解。图 23 给出的是经济型奇异值分解，$\boldsymbol{S}$ 矩阵为方阵，形状为 $D \times D$。

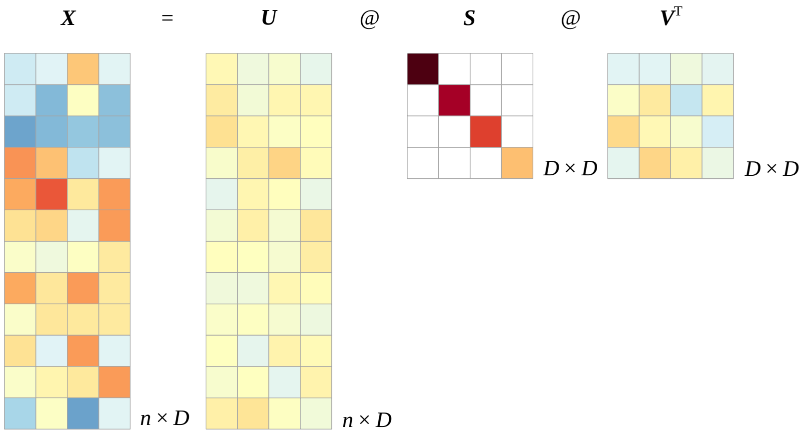


图 23. 经济型奇异值分解

当 $\boldsymbol{X}$ 不是满秩时，即 $\text{rank}(\boldsymbol{X}) = r < D$，图 23 经济型奇异值分解可以进一步简化为如图 24 所示的紧凑型 SVD 分解。

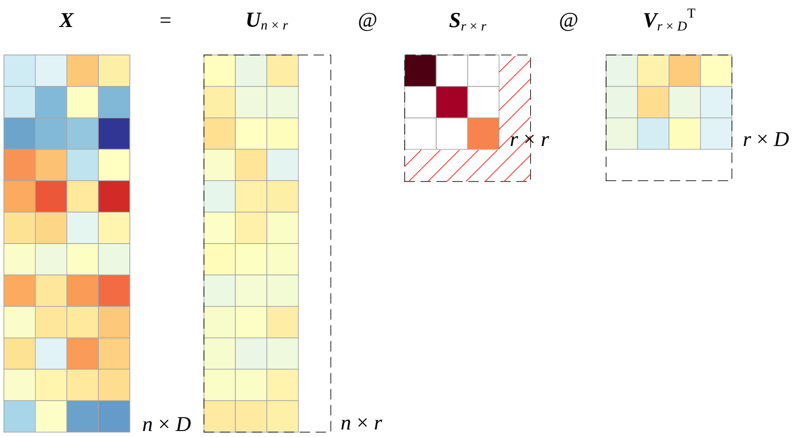


图 24. 紧凑型奇异值分解，$\boldsymbol{X}$ 不是满秩

图 25 给出的是截断型奇异值分解，$S_{p\times p}$ 仅使用图 23 中 $S$ 矩阵 $p$ 个主成分特征值，形状为 $p\times p$。注意，图 25 中使用的是约等号"≈"；这是因为，约等号右侧矩阵运算仅仅还原 $X$ 矩阵部分数据，并非还原全部信息。本章后续将会展开讲解数据还原和误差。

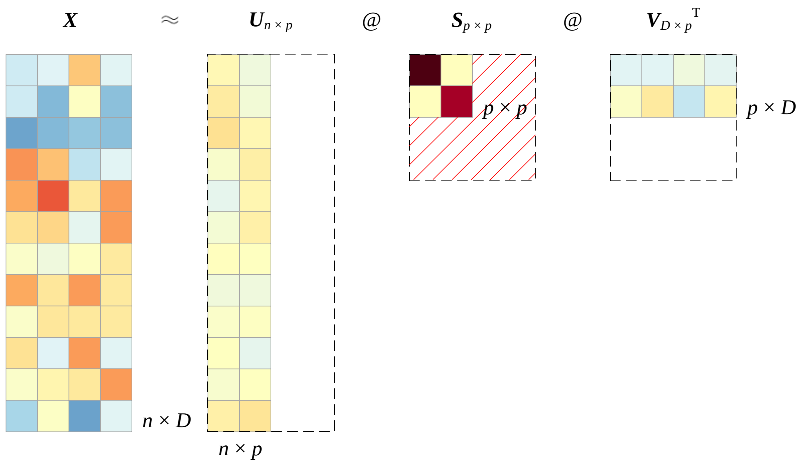


图 25. 截断型奇异值分解

## SVD 完成主成分分析

**奇异值分解** (singular value decomposition, SVD) 也可以用来做主成分分析。首先中心化 (去均值) 数据矩阵。对已经去均值的矩阵 $X_{n\times D}$ 进行完全型 SVD 分解，得到：

$$X = USV^{\mathrm{T}} \tag{23}$$

$V$ 和 $U$ 均为正交矩阵，即满足：

$$\begin{aligned} V^{\mathrm{T}}V &= I \\ U^{\mathrm{T}}U &= I \end{aligned} \tag{24}$$

Python 中常用奇异值分解函数为 numpy.linalg.svd()。

由于 $X$ 已经中心化，其协方差矩阵可以通过下式计算获得：

$$\Sigma = \frac{X^{\mathrm{T}}X}{n-1} \tag{25}$$

将(23) 代入 (25) 得到：

$$\Sigma = \frac{\left(USV^{\mathrm{T}}\right)^{\mathrm{T}}USV^{\mathrm{T}}}{n-1} = \frac{VS^{\mathrm{T}}SV^{\mathrm{T}}}{n-1} \tag{26}$$

对协方差矩阵进行特征值分解：

$$\Sigma = V\Lambda V^{\mathrm{T}} \tag{27}$$

联立 (26) 和 (27)，

$$\frac{\boldsymbol{V}\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}}{n-1}=\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}} \tag{28}$$

对于经济型 SVD 分解，$\boldsymbol{S}$ 为对角方阵，(28) 整理得到：

$$\frac{\boldsymbol{S}^{2}}{n-1}=\boldsymbol{\Lambda} \tag{29}$$

即

$$\frac{1}{n-1}\begin{bmatrix} s_1^2 & & & \\ & s_2^2 & & \\ & & \ddots & \\ & & & s_D^2 \end{bmatrix}=\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_D \end{bmatrix} \tag{30}$$

注意，$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_D$。

奇异值和特征值存在如下关系：

$$\frac{s_j^2}{n-1}=\lambda_j \tag{31}$$

$s_j$ 为第 $j$ 个主成分的**奇异值** (singular value)，$\lambda_j$ 为协方差矩阵的第 $j$ 个特征值。

## 理解 $\boldsymbol{U}$

$\boldsymbol{Z}$ 反向可以还原 $\boldsymbol{X}$：

$$\boldsymbol{X}=\boldsymbol{Z}\boldsymbol{V}^{-1}=\boldsymbol{Z}\boldsymbol{V}^{\mathrm{T}} \tag{32}$$

对比 (23) 和 $\boldsymbol{X}=\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}$，可以发现：

$$\boldsymbol{Z}=\boldsymbol{U}\boldsymbol{S} \tag{33}$$

也就是

$$\begin{bmatrix} \boldsymbol{z}_1 & \boldsymbol{z}_2 & \cdots & \boldsymbol{z}_D \end{bmatrix}=\begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_D \end{bmatrix}\begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_D \end{bmatrix}=\begin{bmatrix} s_1\boldsymbol{u}_1 & s_2\boldsymbol{u}_2 & \cdots & s_D\boldsymbol{u}_D \end{bmatrix} \tag{34}$$

即：

$$s_1\boldsymbol{u}_1=\boldsymbol{z}_1,\quad s_2\boldsymbol{u}_2=\boldsymbol{z}_2,\ \ldots \tag{35}$$

对 $\boldsymbol{z}_1$ 求方差：

$$\operatorname{var}\left(\boldsymbol{z}_1\right)=\frac{\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_1}{n-1}=\frac{\left(s_1\boldsymbol{u}_1\right)^{\mathrm{T}}\left(s_1\boldsymbol{u}_1\right)}{n-1}=\frac{s_1^2\left\|\boldsymbol{u}_1\right\|^2}{n-1}=\frac{s_1^2}{n-1}=\lambda_1 \tag{36}$$

可以发现矩阵 $\boldsymbol{U}$ 每一列数据相当于 $\boldsymbol{Z}$ 的标准化：

$$\boldsymbol{U}=\left[\boldsymbol{u}_1 \quad \boldsymbol{u}_2 \quad \cdots \quad \boldsymbol{u}_D\right]=\left[\frac{\boldsymbol{z}_1}{s_1} \quad \frac{\boldsymbol{z}_2}{s_2} \quad \cdots \quad \frac{\boldsymbol{z}_D}{s_D}\right] \tag{37}$$

也就是：

$$\boldsymbol{U}=\left[\boldsymbol{u}_1 \quad \boldsymbol{u}_2 \quad \cdots \quad \boldsymbol{u}_D\right]=\boldsymbol{Z}\boldsymbol{S}^{-1} \tag{38}$$

至此，我们理解了 SVD 分解中矩阵 $\boldsymbol{U}$ 的内涵。

## 张量积

用张量积来展开 SVD 分解：

$$\begin{aligned}\boldsymbol{X}&=\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}\\&=\left[\boldsymbol{u}_1 \quad \boldsymbol{u}_2 \quad \cdots \quad \boldsymbol{u}_D\right]\begin{bmatrix}s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_D\end{bmatrix}\begin{bmatrix}\boldsymbol{v}_1^{\mathrm{T}}\\ \boldsymbol{v}_2^{\mathrm{T}}\\ \vdots \\ \boldsymbol{v}_D^{\mathrm{T}}\end{bmatrix}\\&=s_1\boldsymbol{u}_1\boldsymbol{v}_1^{\mathrm{T}}+s_2\boldsymbol{u}_2\boldsymbol{v}_2^{\mathrm{T}}+\cdots+s_D\boldsymbol{u}_D\boldsymbol{v}_D^{\mathrm{T}}\\&=s_1\boldsymbol{u}_1\otimes\boldsymbol{v}_1+s_2\boldsymbol{u}_2\otimes\boldsymbol{v}_2+\cdots+s_D\boldsymbol{u}_D\otimes\boldsymbol{v}_D\end{aligned} \tag{39}$$

图 26 所示为 (39) 还原原始数据的过程。

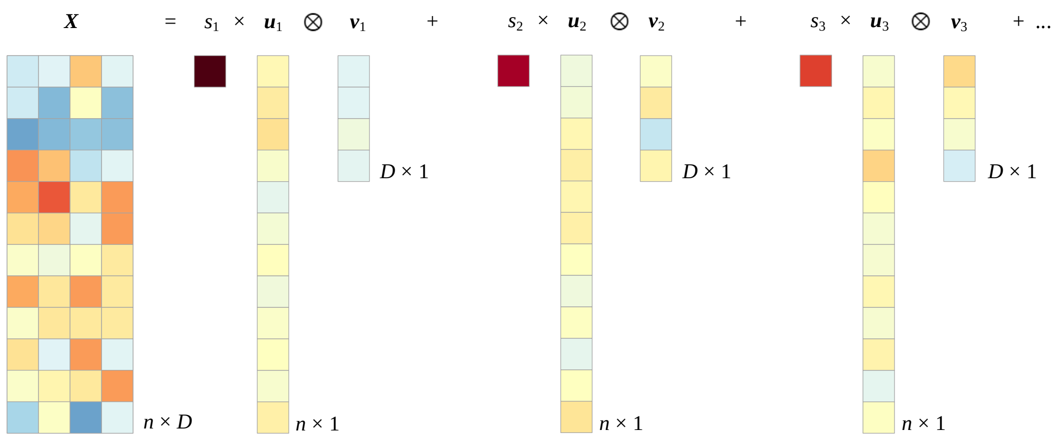


图 26. 张量积 $s_1\boldsymbol{u}_1\otimes\boldsymbol{v}_1$、$s_2\boldsymbol{u}_2\otimes\boldsymbol{v}_2$ 等之和还原数据 $\boldsymbol{X}$

# 25.7 优化问题

本节最后要从两个角度构造主成分分析优化问题角度，并剖析主成分分析理论内核。如图 27 所示，$\boldsymbol{X}$ 为中心化数据，数据中心为零向量；数据 $\boldsymbol{X}$ 协方差矩阵如下：

$$\boldsymbol{\Sigma} = \frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{n-1} \tag{40}$$

图 27 中，$\boldsymbol{v}$ 为某个主成分向量。数据 $\boldsymbol{X}$ 在 $\boldsymbol{v}$ 上投影结果为 $\boldsymbol{z}$：

$$\boldsymbol{z} = \boldsymbol{X}\boldsymbol{v} \tag{41}$$

主成分分析中，选取主成分 $\boldsymbol{v}$ 向量核心是在方向上数据投影值 $\boldsymbol{z}$ 方差最大化；这便是构造优化问题第一个角度。

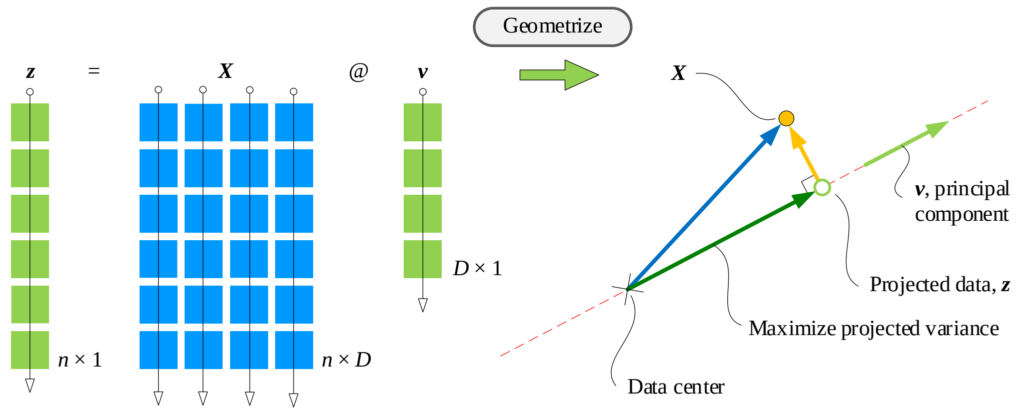


图 27. 主成分分析优化问题

由于 $\boldsymbol{X}$ 为中心化数据，因此 $\boldsymbol{z}$ 的均值也为 0；因此，$\boldsymbol{z}$ 方差为：

$$\operatorname{var}(\boldsymbol{z}) = \frac{\boldsymbol{z}^{\mathrm{T}}\boldsymbol{z}}{n-1} = \boldsymbol{v}^{\mathrm{T}} \overbrace{\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{n-1}}^{\text{Covariance matrix}} \boldsymbol{v} \tag{42}$$

发现上式隐藏着数据 $\boldsymbol{X}$ 协方差矩阵，因此 $\operatorname{var}(\boldsymbol{z})$ 为：

$$\operatorname{var}(\boldsymbol{z}) = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v} \tag{43}$$

$\boldsymbol{v}$ 为单位列向量，即下式成立：

$$\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} = 1 \tag{44}$$

有以上分析，构造主成分分析优化问题，优化目标为数据在 $\boldsymbol{v}$ 方向上数据投影值方差最大化：

$$\begin{aligned} &\underset{\boldsymbol{v}}{\arg\max}\ \ \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v} \\ &\text{subject to: } \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} - 1 = 0 \end{aligned} \tag{45}$$

上式最大化优化问题等价于如下最小化优化问题：

$$\begin{aligned} &\underset{\boldsymbol{v}}{\arg\min} \quad -\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v} \\ &\text{subject to: } \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}-1=0 \end{aligned} \tag{46}$$

构造拉格朗日函数 $L(\boldsymbol{v}, \lambda)$：

$$L(\boldsymbol{v},\lambda) = -\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v} + \lambda\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}-1\right) \tag{47}$$

$\lambda$ 为拉格朗日乘子。$L(\boldsymbol{x}, \lambda)$ 对 $\boldsymbol{v}$ 求偏导，最优解必要条件如下：

$$\nabla_{\boldsymbol{v}} L(\boldsymbol{v},\lambda) = \frac{\partial L(\boldsymbol{v},\lambda)}{\partial \boldsymbol{v}} = \left(-2\boldsymbol{\Sigma}\boldsymbol{v} + 2\lambda\boldsymbol{v}\right)^{\mathrm{T}} = \boldsymbol{0} \tag{48}$$

有关拉格朗日乘子法，请大家回顾《矩阵力量》第 18 章。

整理 (48) 得到：

$$\boldsymbol{\Sigma}\boldsymbol{v} = \lambda\boldsymbol{v} \tag{49}$$

由此，$\boldsymbol{v}$ 为数据 $\boldsymbol{X}$ 方差-协方差矩阵 $\boldsymbol{\Sigma}$ 特征向量。var($\boldsymbol{z}$) 整理为：

$$\operatorname{var}(\boldsymbol{z}) = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v} = \boldsymbol{v}^{\mathrm{T}}\lambda\boldsymbol{v} = \lambda\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} = \lambda \tag{50}$$

即说，var($\boldsymbol{z}$) 最大值对应 $\boldsymbol{\Sigma}$ 最大特征值。$\boldsymbol{\Sigma}$ 特征值分解：

$$\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}} \tag{51}$$

其中，

$$\begin{aligned} &\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_D \end{bmatrix} \\ &\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_D \end{bmatrix}, \quad \lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_D \end{aligned} \tag{52}$$

这一节从优化角度解释了为什么特征值分解能够完成主成分分析。

# 25.8 数据还原和误差

## 还原

前文介绍过，$\boldsymbol{Z}$ 反向可以通过 $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{V}^{\mathrm{T}}$ 还原 $\boldsymbol{X}$。

图 28 所示为还原得到 $\boldsymbol{X}$ 过程。

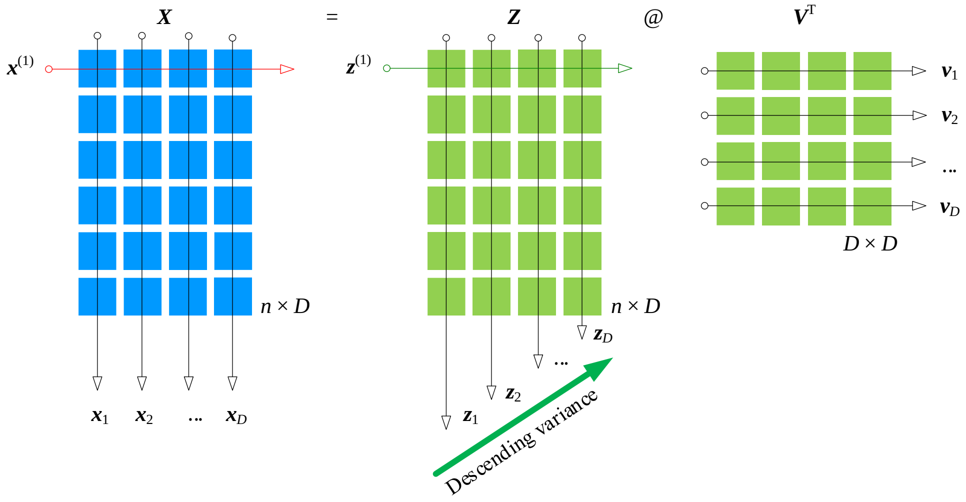


图 28. 反向还原数据 $X = ZV^T$

图 29 所示热图，为新特征数据矩阵 $Z$ 还原转化为原始数据矩阵 $X$。

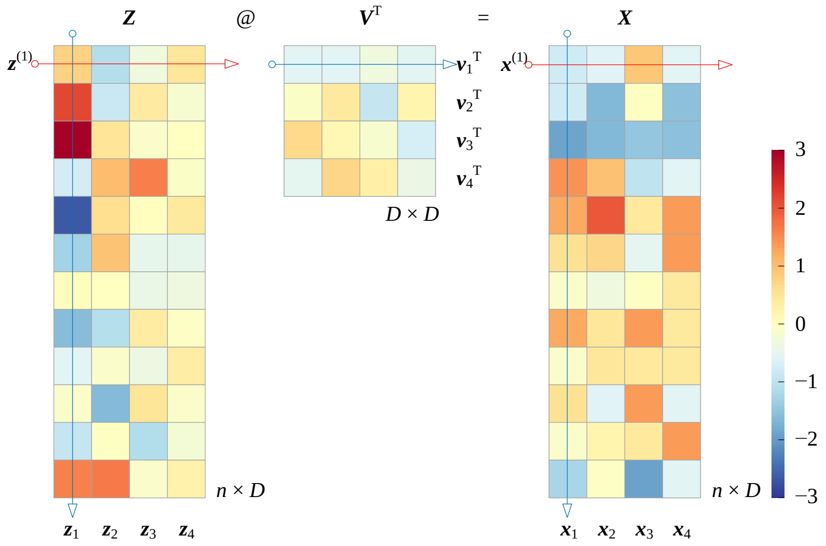


图 29. 新特征数据矩阵 $Z$ 还原转化为原始数据矩阵 $X$

$X = ZV^T$ 展开得到下式：

$$X = \begin{bmatrix} z_1 & z_2 & z_3 & z_4 \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ v_3^T \\ v_4^T \end{bmatrix} = \underbrace{z_1 v_1^T}_{\hat{x}_1} + \underbrace{z_2 v_2^T}_{\hat{x}_2} + \underbrace{z_3 v_3^T}_{\hat{x}_3} + \underbrace{z_4 v_4^T}_{\hat{x}_4} \tag{53}$$

(53) 所示运算过程如图 30 所示。

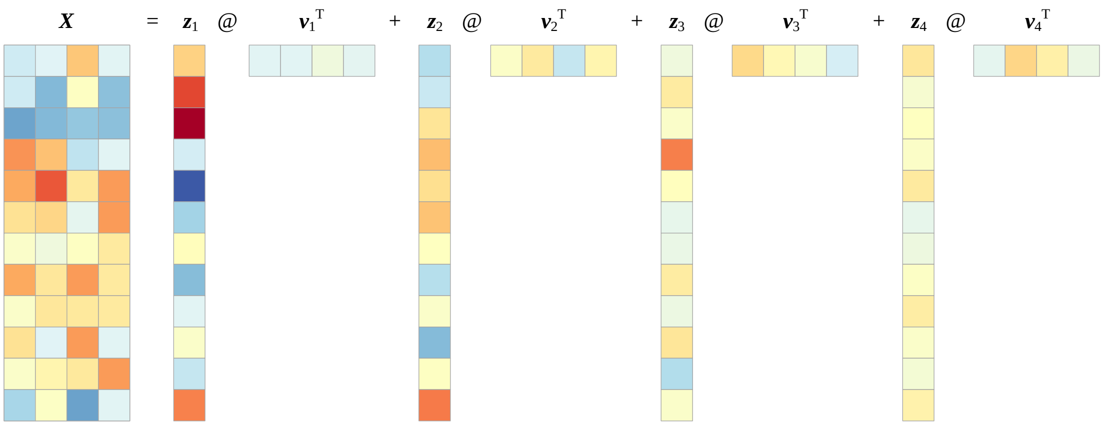


图 30. 还原原始数据运算

图 31 所示为 $z_1$ 还原 $X$ 部分数据，对应运算如下：

$$X_1 = z_1 v_1^{\mathrm{T}} \tag{54}$$

展开上式得到：

$$\begin{aligned} X_1 &= z_1 v_1^{\mathrm{T}} \\ &= z_1 \begin{bmatrix} v_{1,1} & v_{2,1} & \cdots & v_{D,1} \end{bmatrix} \\ &= \begin{bmatrix} v_{1,1} z_1 & v_{2,1} z_1 & \cdots & v_{D,1} z_1 \end{bmatrix} \end{aligned} \tag{55}$$

观察图 31 热图可以发现一些有意思的特点。还原得到的数据每一列热图模式高度相似；(55) 解释了这一点，$X_1$ 的每一列均是标量乘以向量 $z_1$ 的结果。显然，$X_1$ 的秩为 1，即 $\mathrm{rank}(X_1) = 1$。

图 32、图 33 和图 34 分别展示 $z_2$、$z_3$ 和 $z_4$ 还原 $X$ 部分数据。

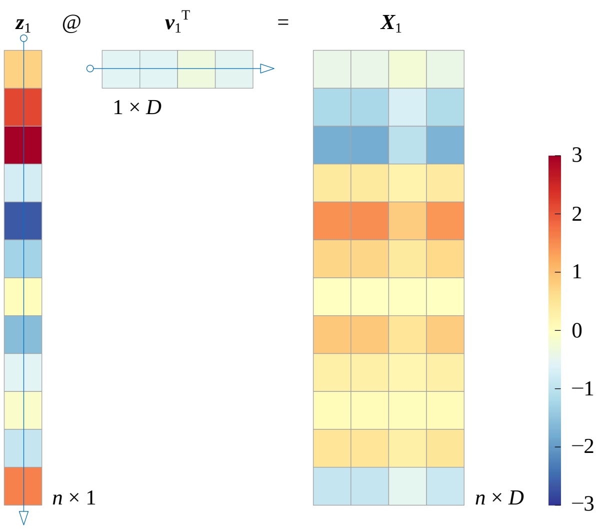


图 31. $z_1$ 还原 $X$ 部分数据

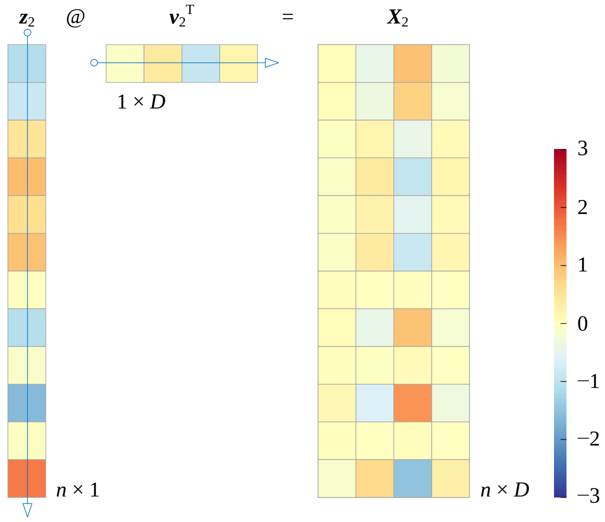


图 32. $z_2$ 还原 $X$ 部分数据

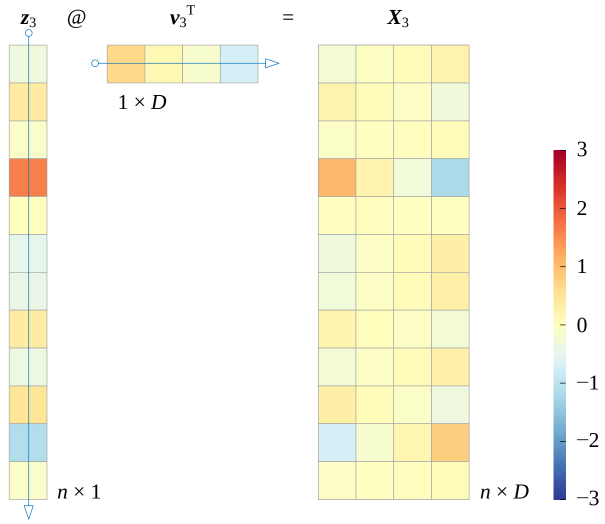


图 33. $z_3$ 还原 $X$ 部分数据

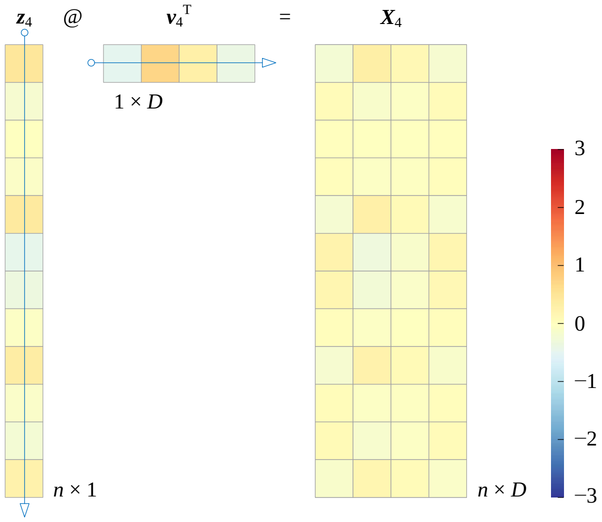


图 34. $z_4$还原 $X$ 部分数据

图 35 所示为原始数据矩阵 $X$ 热图于四层热图叠加结果。观察图 35，发现随着主成分次数降低，每个主成分各自对数据 $X$ 还原力度不断降低，看到还原热图颜色越来越浅；但是，把这些主成分各自还原生成热图不断叠加，获得热图就不断逼近原始热图。

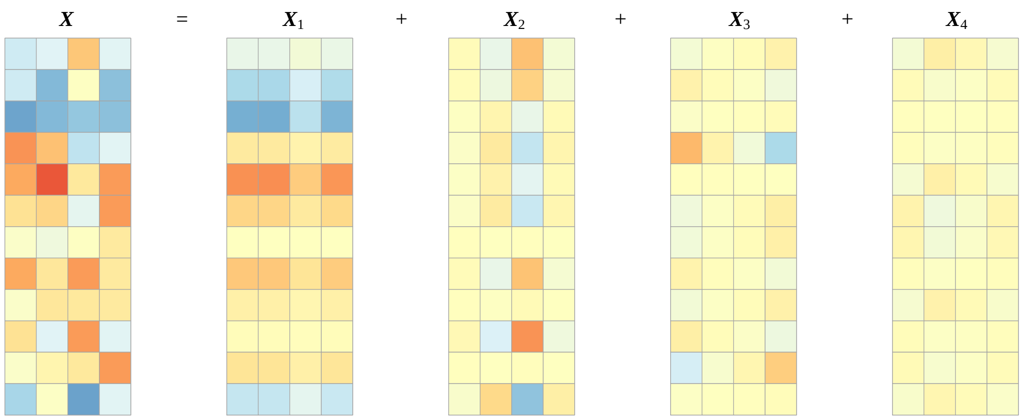


图 35. 原始数据矩阵 $X$ 热图于四层热图叠加结果

## 张量积

另外，(53) 可以用张量积来表达：

$$X = \underbrace{z_1 \otimes v_1}_{\hat{X}_1} + \underbrace{z_2 \otimes v_2}_{\hat{X}_2} + \underbrace{z_3 \otimes v_3}_{\hat{X}_3} + \underbrace{z_4 \otimes v_4}_{\hat{X}_4} \tag{56}$$

图 36 所示为通过主成分 $v_1, v_2, v_3, v_4$ 和其自身转置乘积计算张量积。

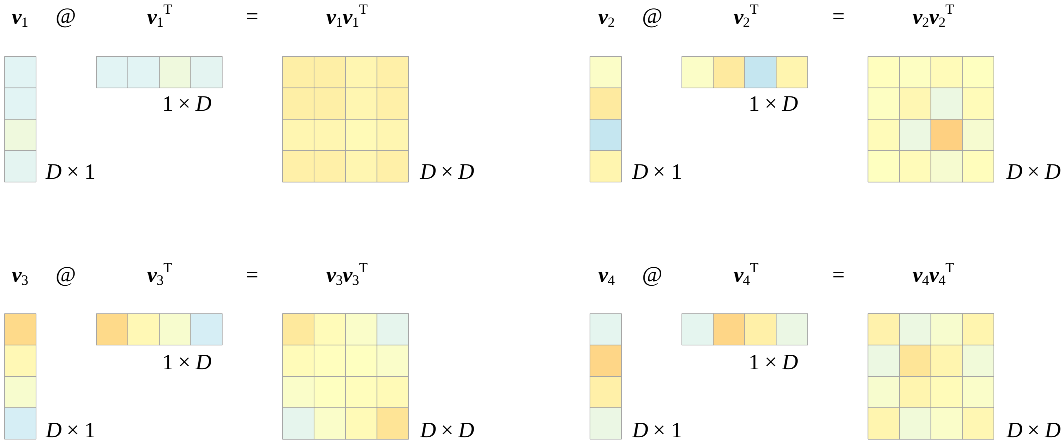


图 36. 列向量乘自身转置获得四个张量积

图 37 所示为张量积运算，和图 36 结果完全一致。

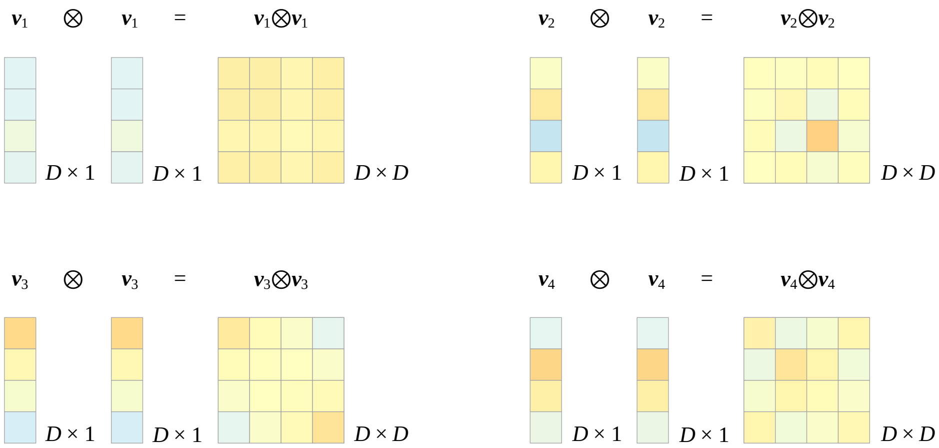


图 37. 内积计算获得四个张量积

利用 (14)，(53) 可以整理为：

$$\boldsymbol{X} = \boldsymbol{X}\boldsymbol{v}_1\boldsymbol{v}_1^{\mathrm{T}} + \boldsymbol{X}\boldsymbol{v}_2\boldsymbol{v}_2^{\mathrm{T}} + \ldots + \boldsymbol{X}\boldsymbol{v}_D\boldsymbol{v}_D^{\mathrm{T}} = \sum_{j=1}^{D} \boldsymbol{X}\boldsymbol{v}_j\boldsymbol{v}_j^{\mathrm{T}} = \boldsymbol{X}\left(\sum_{j=1}^{D} \boldsymbol{v}_j\boldsymbol{v}_j^{\mathrm{T}}\right) \tag{57}$$

(57) 可以用张量积表达：

$$\boldsymbol{X} = \boldsymbol{X}\left(\boldsymbol{v}_1 \otimes \boldsymbol{v}_1\right) + \boldsymbol{X}\left(\boldsymbol{v}_2 \otimes \boldsymbol{v}_2\right) + \ldots + \boldsymbol{X}\left(\boldsymbol{v}_D \otimes \boldsymbol{v}_D\right) = \sum_{j=1}^{D} \boldsymbol{X}\boldsymbol{v}_j \otimes \boldsymbol{v}_j = \boldsymbol{X}\left(\sum_{j=1}^{D} \boldsymbol{v}_j \otimes \boldsymbol{v}_j\right) \tag{58}$$

容易推导得到，(58) 中张量积相加得到单位矩阵。《矩阵力量》第 10 章给这种投影一个特别的名字——二次投影，建议大家回顾。

$$\boldsymbol{v}_1 \otimes \boldsymbol{v}_1 + \boldsymbol{v}_2 \otimes \boldsymbol{v}_2 + \ldots + \boldsymbol{v}_D \otimes \boldsymbol{v}_D = \left( \sum_{j=1}^{D} \boldsymbol{v}_j \otimes \boldsymbol{v}_j \right) = \boldsymbol{I} \tag{59}$$

上式如图 38 热图所示。

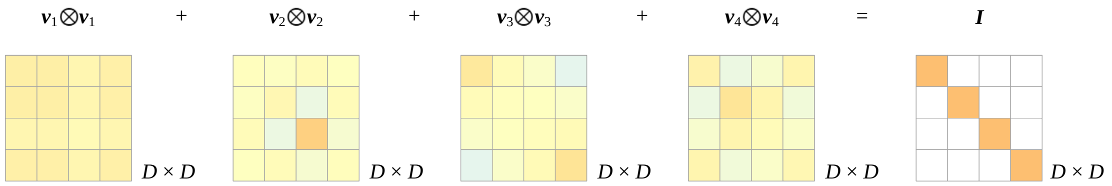


图 38. 张量积相加得到单位矩阵

联立 (16) 和 (54)，利用张量积 $\boldsymbol{v}_1 \otimes \boldsymbol{v}_1$ 还原部分原始数据：

$$\boldsymbol{X}_1 = \boldsymbol{z}_1 \boldsymbol{v}_1^{\mathrm{T}} = \boldsymbol{X} \boldsymbol{v}_1 \boldsymbol{v}_1^{\mathrm{T}} = \boldsymbol{X} \underbrace{\left( \boldsymbol{v}_1 \otimes \boldsymbol{v}_1 \right)}_{\text{Tensor product}} \tag{60}$$

类似，张量积 $\boldsymbol{v}_2 \otimes \boldsymbol{v}_2$ 也可以还原部分原始数据：

$$\boldsymbol{X}_2 = \boldsymbol{z}_2 \boldsymbol{v}_2^{\mathrm{T}} = \boldsymbol{X} \boldsymbol{v}_2 \boldsymbol{v}_2^{\mathrm{T}} = \boldsymbol{X} \underbrace{\left( \boldsymbol{v}_2 \otimes \boldsymbol{v}_2 \right)}_{\text{Tensor product}} \tag{61}$$

图 39 所示为张量积 $\boldsymbol{v}_1 \otimes \boldsymbol{v}_1$ 和 $\boldsymbol{v}_2 \otimes \boldsymbol{v}_2$ 还原部分数据 $\boldsymbol{X}$；图 40 所示为张量积 $\boldsymbol{v}_3 \otimes \boldsymbol{v}_3$ 和 $\boldsymbol{v}_4 \otimes \boldsymbol{v}_4$ 还原部分数据 $\boldsymbol{X}$。

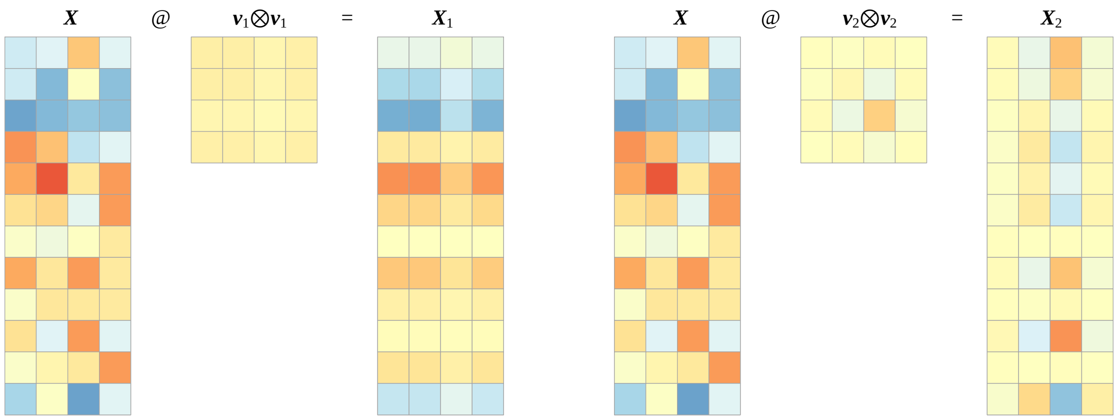


图 39. 张量积 $\boldsymbol{X}(\boldsymbol{v}_1 \otimes \boldsymbol{v}_1)$和 $\boldsymbol{X}(\boldsymbol{v}_2 \otimes \boldsymbol{v}_2)$还原部分数据 $\boldsymbol{X}$

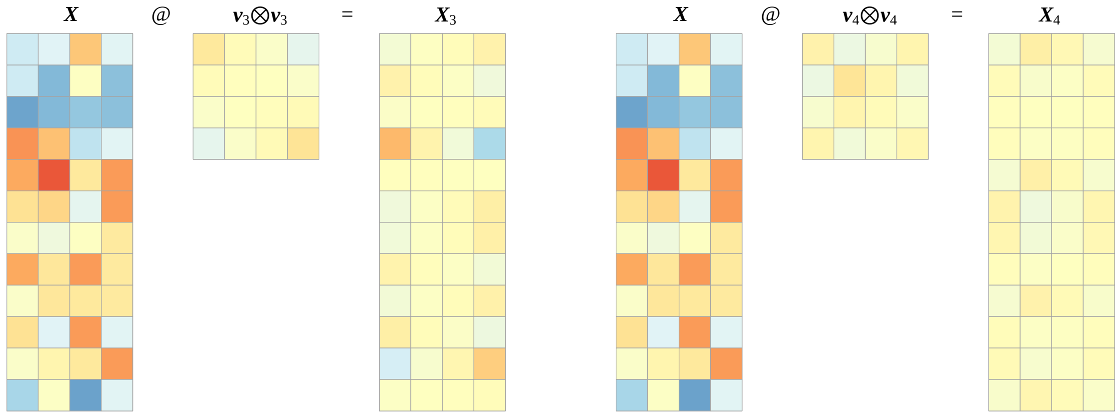


图 40. 张量积 $\boldsymbol{X}(\boldsymbol{v}_3 \otimes \boldsymbol{v}_3)$ 和 $\boldsymbol{X}(\boldsymbol{v}_4 \otimes \boldsymbol{v}_4)$ 还原部分数据 $\boldsymbol{X}$

## 误差

图 41 所示为两个主成分 $\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$ 还原获得原始数据热图，具体计算如下：

$$\hat{\boldsymbol{X}} = [\boldsymbol{z}_1 \quad \boldsymbol{z}_2][\boldsymbol{v}_1 \quad \boldsymbol{v}_2]^{\mathrm{T}} \tag{62}$$

相当于

$$\begin{aligned} \hat{\boldsymbol{X}} &= \boldsymbol{X}_1 + \boldsymbol{X}_2 = \boldsymbol{z}_1 \boldsymbol{v}_1^{\mathrm{T}} + \boldsymbol{z}_2 \boldsymbol{v}_2^{\mathrm{T}} \\ &= \boldsymbol{X}\left(\boldsymbol{v}_1 \boldsymbol{v}_1^{\mathrm{T}} + \boldsymbol{v}_2 \boldsymbol{v}_2^{\mathrm{T}}\right) = \boldsymbol{X}\left(\boldsymbol{v}_1 \otimes \boldsymbol{v}_1 + \boldsymbol{v}_2 \otimes \boldsymbol{v}_2\right) \end{aligned} \tag{63}$$

图 42 所示为通过叠加图 31 和图 32 两个热图还原原始数据矩阵。

从张量积角度来看图 42，

$$\boldsymbol{X} \approx s_1 \boldsymbol{u}_1 \otimes \boldsymbol{v}_1 + s_2 \boldsymbol{u}_2 \otimes \boldsymbol{v}_2^{\mathrm{T}} \tag{64}$$

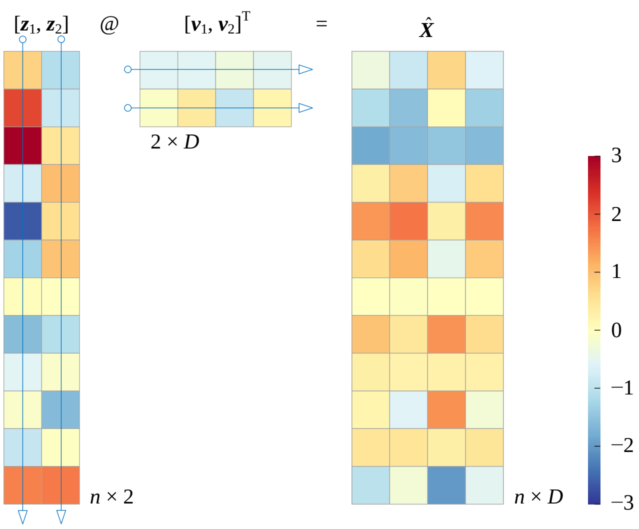

图 41. 前两个主成分 $\boldsymbol{z}_1$ 和 $\boldsymbol{z}_1$ 还原 $\boldsymbol{X}$ 数据

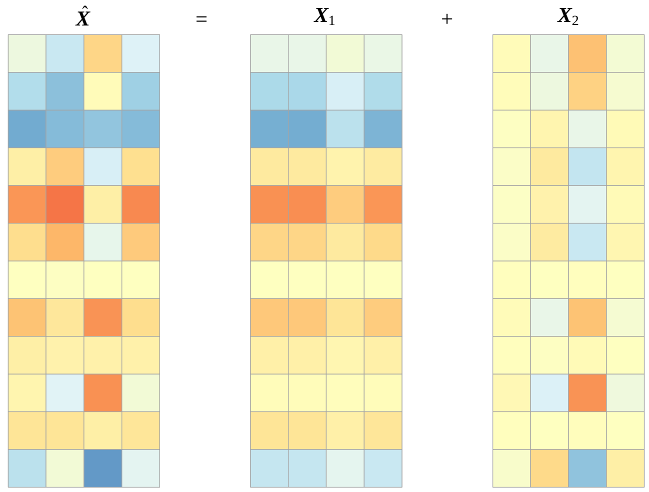


图 42. 两个热图叠加还原原始数据

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：

$$\boldsymbol{E} = \boldsymbol{X} - \hat{\boldsymbol{X}} \tag{65}$$

图 43 比较原始数据 $\boldsymbol{X}$、拟合数据 $\hat{\boldsymbol{X}}$ 和残差数据矩阵 $\boldsymbol{E}$ 热图，发现原始数据 $\boldsymbol{X}$ 和拟合数据 $\hat{\boldsymbol{X}}$ 已经相差无几。从图片还原角度，如图 43 所示，PCA 降维用更少维度、更少数据获得几乎一样画质图片。

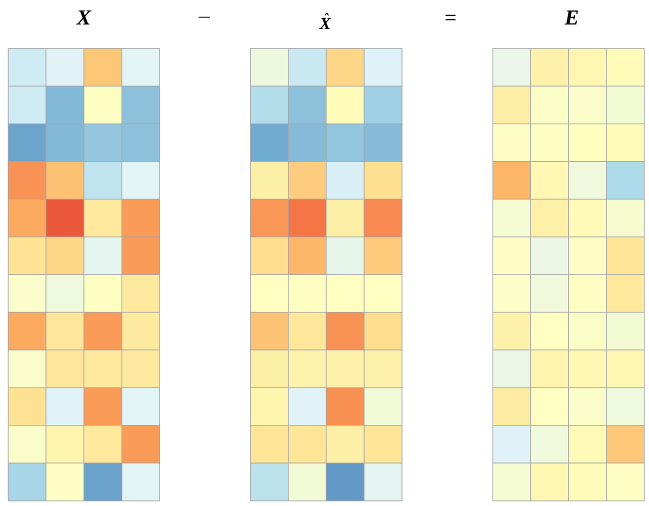


图 43. 原始数据、拟合数据和残差数据热图

## 六条技术路径

相信大家对表 1 并不陌生，大家都在《矩阵力量》第 25 章中见过这六条 PCA 技术路线。本章介绍的实际上是：a) 特征值分解协方差矩阵；b) 奇异值分解中心化数据矩阵。《数据有道》一册将比较表 1 这六种方法的异同。

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 $\boldsymbol{G} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$<br>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 $\boldsymbol{\Sigma} = \dfrac{\left(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})\right)^{\mathrm{T}}\left(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})\right)}{n-1}$ | 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) $\boldsymbol{Z}_X = \left(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})\right)\boldsymbol{D}^{-1}$<br>$\boldsymbol{D} = \mathrm{diag}\left(\mathrm{diag}(\boldsymbol{\Sigma})\right)^{\frac{1}{2}}$ | 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 $\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{\Sigma}\boldsymbol{D}^{-1}$<br>$\boldsymbol{D} = \mathrm{diag}\left(\mathrm{diag}(\boldsymbol{\Sigma})\right)^{\frac{1}{2}}$ | 特征值分解 | $\boldsymbol{P} = \boldsymbol{V}_Z \boldsymbol{\Lambda}_Z \boldsymbol{V}_Z^{\mathrm{T}}$ |

人类的思维方式天然具备概率统计属性。概率统计的背后的思想更贴近“生活常识”。人们在谈论可能性的时候，大脑就不自觉进入“概率统计”模式。看着天上云层很厚，可能两小时就会下雨。昨晚淋了雨，估计今天要感冒。估计这次考试通过率 80%以上。咱们剪刀石头布，三局两胜决胜负。

可惜的是，当数学家将这些生活常识“翻译成”数学语言之后，它们就变成了冷冰冰“火星文”。更遗憾的是，很多概率统计图书读起来更像是数学公式手册，满篇的公式、定理、推导。

概率统计与其说是工具，不如说是方法论、世界观。大家常说的“一命，二运，三风水，四读书”，体现的也是概率统计的思维。

命运不可问，命中没有莫强求。“小概率事件”能发生，得之我幸，不得我命。

风水轮流转，玄而又玄。

正所谓知识改变命运，只有读书成才对应“大概率事件”。大家捧起这本书的时候，就依靠统计思维做出了“最优化”选择。

《统计至简》是“数学”板块的三本中的最后一本。大家读到这里，也就完成了整个“数学”板块的修炼。希望大家日后再看到任何公式的时候，闭上眼睛，能够在脑海中“看见”矩阵和各种几何图形。

下面，我们将踏上《数据有道》、《机器学习》的“实践”之旅！期待和大家共同学习、成长！