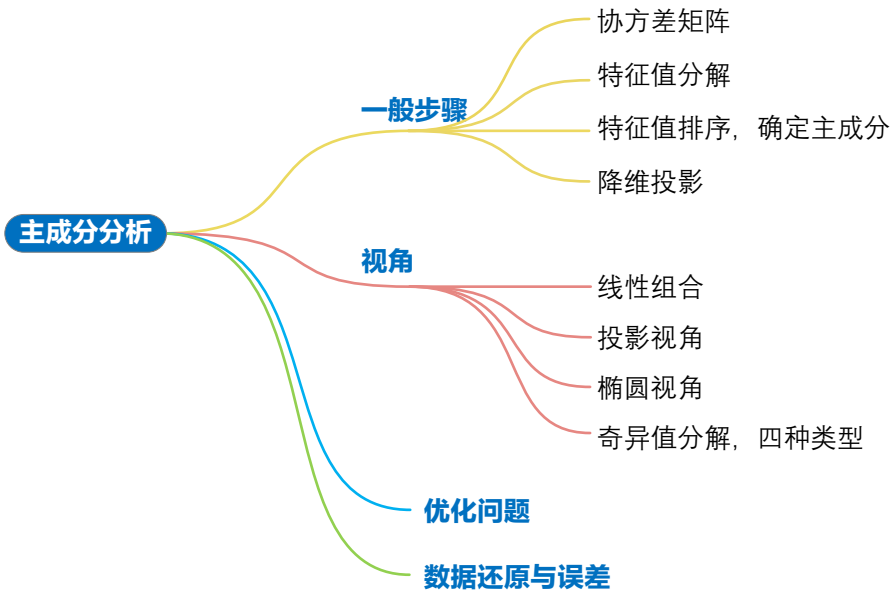
25 Principal Component Analysis
主成分分析

以概率统计、几何、矩阵分解、优化为视角

> 掌握我们的命运的不是星象，而是我们自己。
>
> ***It is not in the stars to hold our destiny but in ourselves.***
>
> —— 威廉·莎士比亚 (William Shakespeare) | 英国剧作家 | 1564 ~ 1616

- `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
- `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


主成分分析
一般步骤
协方差矩阵
特征值分解
特征值排序，确定主成分
降维投影
视角
线性组合
投影视角
椭圆视角
奇异值分解，四种类型
优化问题
数据还原与误差

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。简单来说，PCA 的核心思想是通过线性变换将高维数据映射到低维空间中，使得映射后的数据能够尽可能地保留原始数据的信息，同时去除噪声和冗余信息，从而更好地描述数据的本质特征。

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，大家可以把这一章看成丛书“数学”板块的一个总结。

无监督学习

主成分分析是重要的**无监督学习** (unsupervised learning) 算法。无监督学习是一种机器学习方法，它处理没有标签或输出值的数据。在无监督学习中，模型只能通过分析输入数据的内部结构、模式和相似性来发现数据的特征，从而自动学习数据的潜在结构和规律。

无监督学习通常用于**聚类** (clustering)、**降维** (dimensionality reduction)、**异常检测** (outlier detection) 和**关联规则挖掘** (association rule learning) 等问题。

在聚类问题中，目标是将相似的数据点分组到不同的簇中，从而将数据分割为具有内在结构的不同子集。

在降维问题中，目标是从高维数据中提取出具有代表性的低维特征，从而减少计算复杂度、提高数据可视化效果和去除噪声。主成分分析就是常用的降维算法。

在异常检测问题中，目标是检测数据集中的异常数据点，这些数据点与其它数据点存在显著的差异。本书第 23 章介绍的马氏距离就常用来发现数据中的离群值。

在关联规则挖掘问题中，目标是在大规模数据集中寻找频繁出现的关联项集，从而发现数据中的相关性和关联性。

《数据有道》一册将介绍异常检测、降维、关联规则挖掘等话题，而《机器学习》将关注常见的聚类算法。

一般步骤

如图 1 所示，PCA 的一般步骤如下：

- 计算原始数据 $\boldsymbol{X}_{n\times D}$ 的协方差矩阵 $\boldsymbol{\Sigma}_{D\times D}$；
- 对 $\boldsymbol{\Sigma}$ 特征值分解，获得特征值 λ_i 与特征向量矩阵 $\boldsymbol{V}_{D\times D}$；
- 对特征值 λ_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 (中心化数据) 投影到这 p 个正交向量构建的低维空间中，获得得分 $\boldsymbol{Z}_{n\times p}$。

很多时候，在第一步中，我们先**标准化** (standardization) 原始数据，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。而有些情况，对原始数据 $\boldsymbol{X}_{n\times D}$ 进行中心化 (去均值) 就足够了，即将数据质心移到原点。

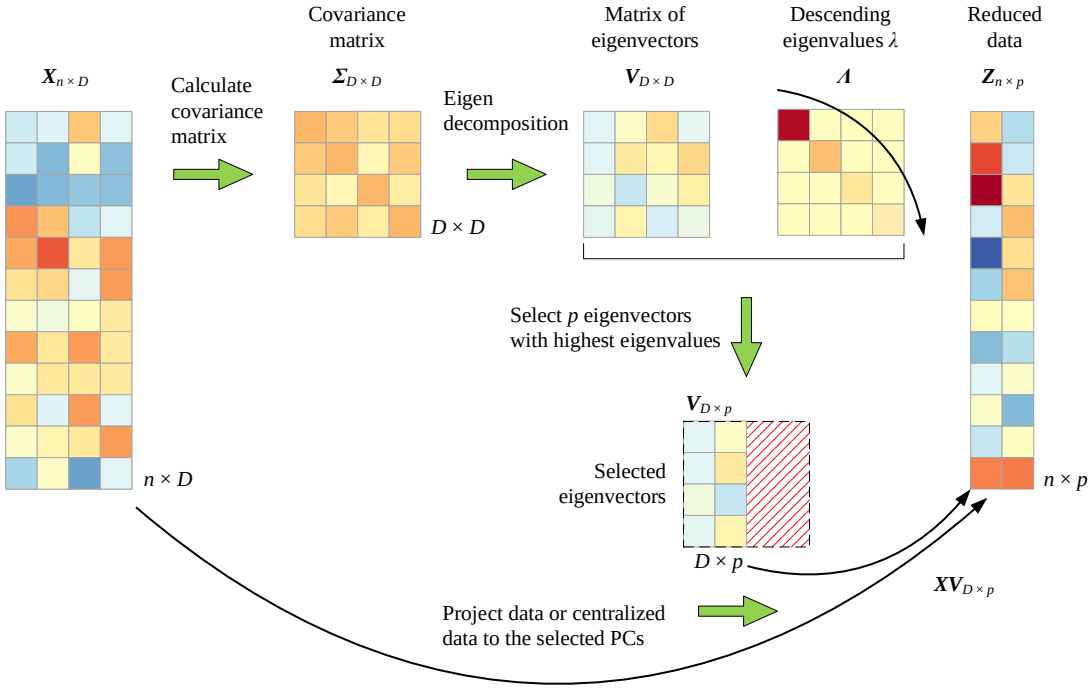


图 1. 主成分分析一般技术路线：特征值分解协方差矩阵

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。

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 $\boldsymbol{X}_c$ 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 $\boldsymbol{X}$ 可以分别通过行和列来解释。矩阵 $\boldsymbol{X}$ 每一列代表一个特征向量：

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3 & \boldsymbol{x}_4 \end{bmatrix} \tag{1}$$

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

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

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

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

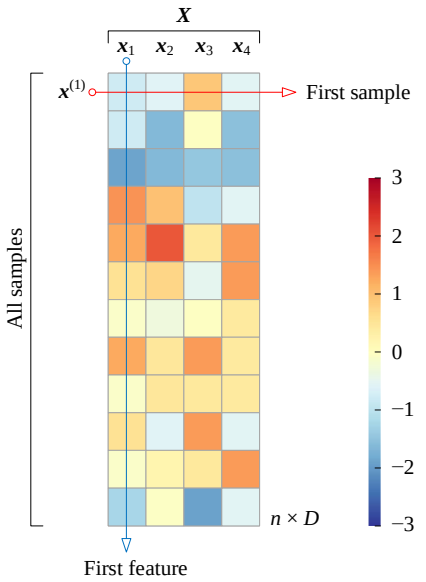


图 2. 原始数据 $\boldsymbol{X}$ 热图，$D = 4$，$n = 12$，$\boldsymbol{X}$ 已经去均值

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标准差区别不大。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标准差大小差异明显。

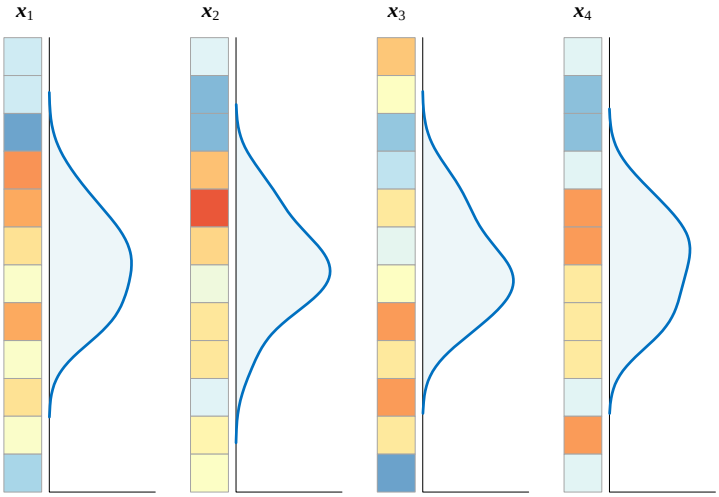


图 3. $\boldsymbol{X}$ 四个特征向量数据分布

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：

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

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 λ，特征值从大到小排列：

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

本书前文介绍过，从统计学角度来讲，λ_j 是第 j 个主成分所贡献的方差。

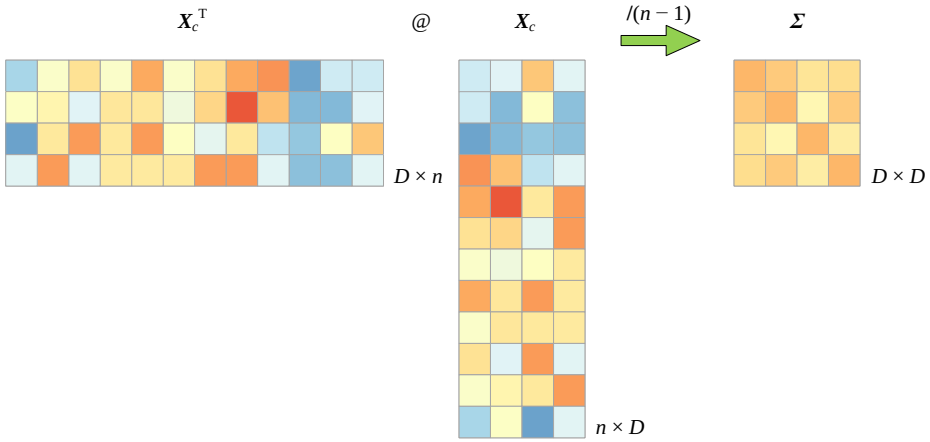


图 4. 计算原始数据协方差矩阵，$D = 4$，$n = 12$

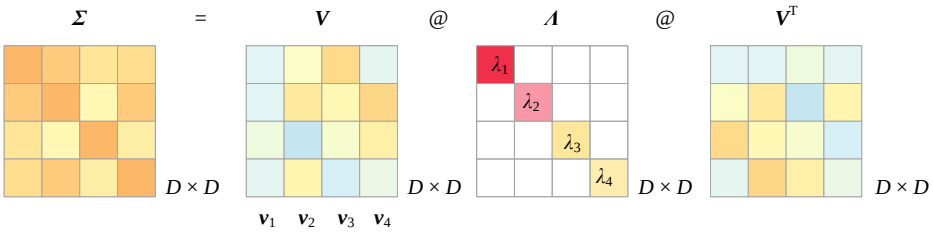


图 5. 协方差矩阵特征值分解，$D = 4$

主成分、载荷

$\boldsymbol{V}$ 为特征向量构造的 $D \times D$ 的方阵：

$$\boldsymbol{V} = \begin{bmatrix} \underset{\text{PC1}}{\boldsymbol{v}_1} & \underset{\text{PC2}}{\boldsymbol{v}_2} & \cdots & \boldsymbol{v}_D \end{bmatrix} = \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,D} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ v_{D,1} & v_{D,2} & \cdots & v_{D,D} \end{bmatrix} \tag{7}$$

$\boldsymbol{v}_1$ 被称作**第一主成分** (first principal component)，本书常记做 PC1；$\boldsymbol{v}_2$ 被称作**第二主成分** (second principal component)，记做 PC2；以此类推。

$\boldsymbol{V}$ 的列向量也叫载荷 (loadings)。注意，有些文献中载荷定义为：

$$\boldsymbol{V}\sqrt{\boldsymbol{\Lambda}} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_D \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} & & & \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_D} \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{v}_1 & \sqrt{\lambda_2}\boldsymbol{v}_2 & \cdots & \sqrt{\lambda_D}\boldsymbol{v}_D \end{bmatrix} \tag{8}$$

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($\boldsymbol{\Lambda}$)：

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

第 j 个特征值 λ_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**总方差解释** (total variance explained) 的百分比为：

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

"total variance" 指的是原始数据中所有变量的总方差，"explained" 意味着这个方差被 PCA 模型中所选的主成分所解释。因此，"total variance explained" 表示通过 PCA 转换后的主成分所解释的原始数据中总方差的比例。这个值通常以百分比的形式给出，可以帮助我们了解每个主成分对数据的解释程度，以及所有主成分的总体效果。

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陡坡图实例。

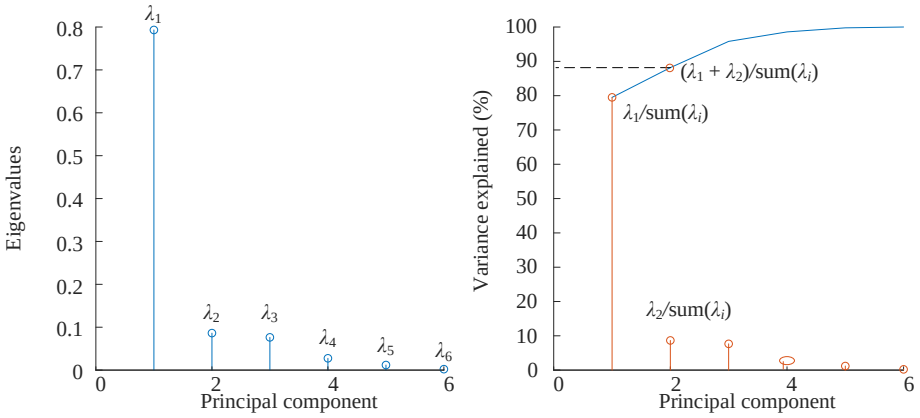


图 6. PCA 分析主元方差和陡坡图

25.4 投影

本节从投影角度介绍 PCA。数据矩阵 $\boldsymbol{X}$ 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**载荷** (loadings)，$\boldsymbol{Z}$ 常被称作**得分** (scores)。图 7 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。

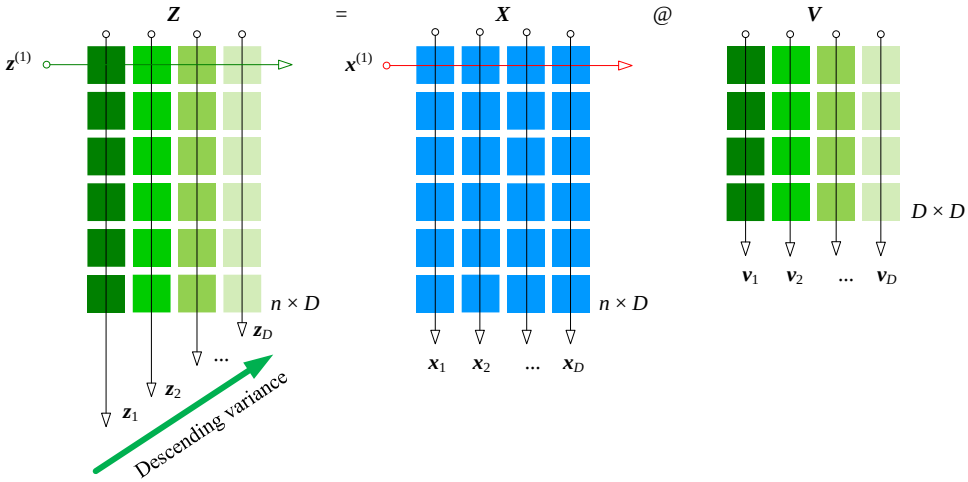


图 7. PCA 分解数据关系 $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{V}$

图 8 所示为将图 2 给出数据矩阵 $\boldsymbol{X}$ 投影到矩阵 $\boldsymbol{V}$，得到的得分 $\boldsymbol{Z}$。

值得强调的一点是，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。

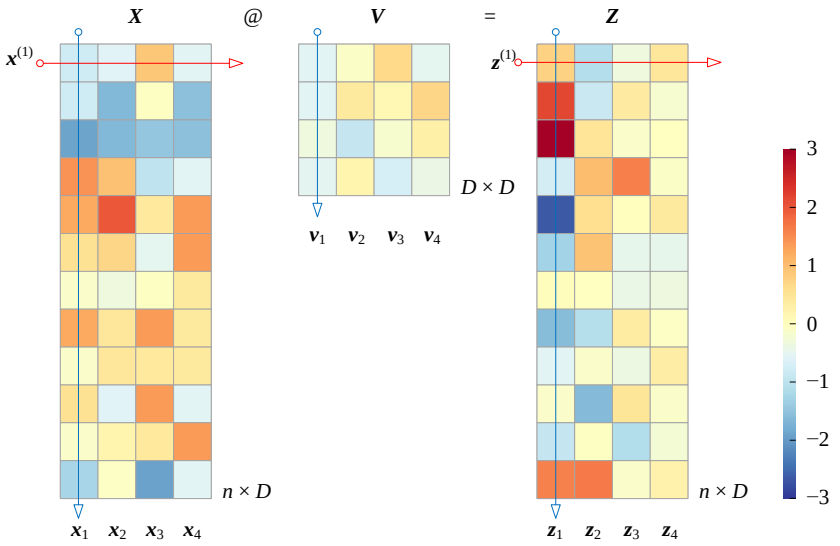


图 8. $\boldsymbol{Z}$、$\boldsymbol{X}$ 和 $\boldsymbol{V}$ 这三个矩阵关系和热图

Z的列向量

前文讨论过，矩阵 X 每一列特征数据方差区别不大 (见图 3)；而图 9 告诉我们，经过 PCA 分解得到的矩阵 Z 四个新特征数据分布差异显著。

如图 9 所示，第一列 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 z_1 到第四列 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。

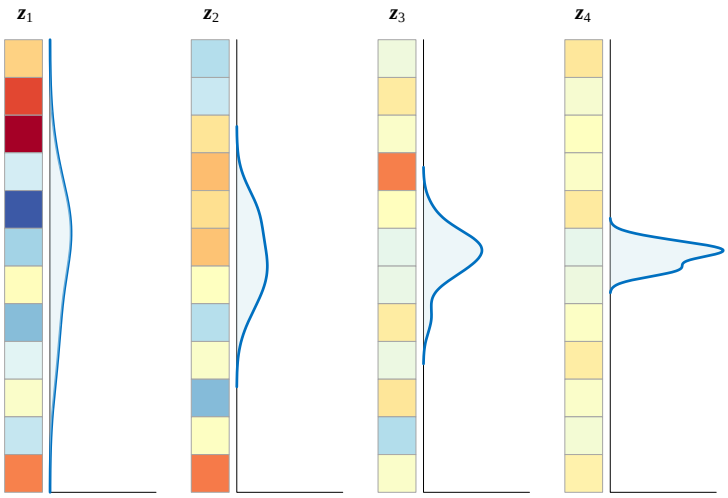


图 9. Z 四个新特征数据分布

将 (12) 展开得到：

$$\begin{bmatrix} \boldsymbol{z}_1 & \boldsymbol{z}_2 & \cdots & \boldsymbol{z}_D \end{bmatrix} = \boldsymbol{X} \begin{bmatrix} \underset{\text{PC1}}{\boldsymbol{v}_1} & \underset{\text{PC2}}{\boldsymbol{v}_2} & \cdots & \boldsymbol{v}_D \end{bmatrix} \tag{13}$$

由此，得到图 10 所示主成分分析运算的数据关系：

$$\begin{cases} \boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{v}_1 \\ \boldsymbol{z}_2 = \boldsymbol{X}\boldsymbol{v}_2 \\ \quad \vdots \\ \boldsymbol{z}_D = \boldsymbol{X}\boldsymbol{v}_D \end{cases} \tag{14}$$

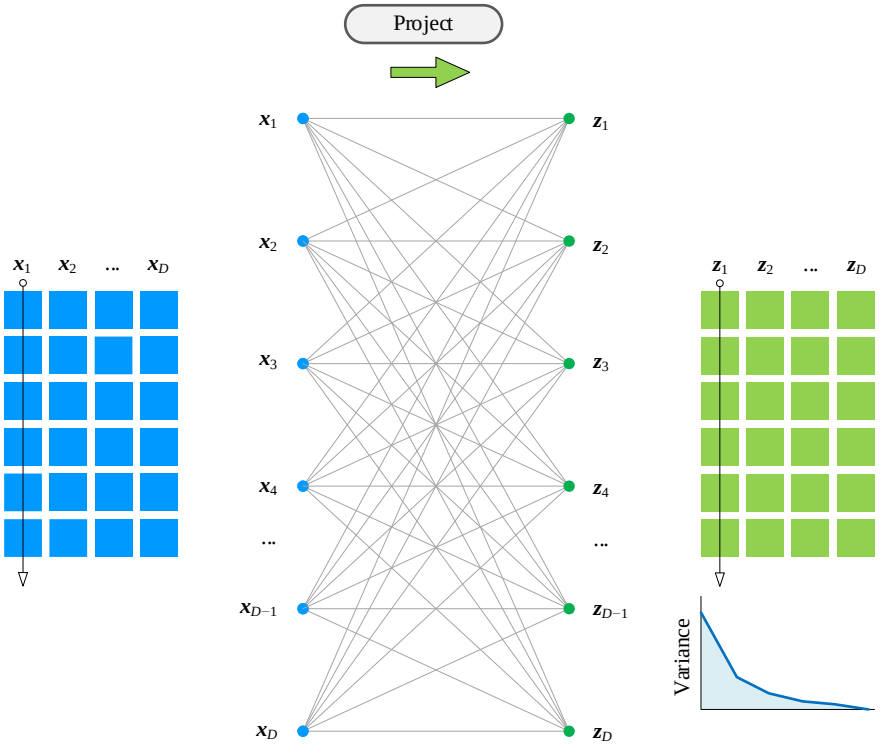


图 10. PCA 中数据关系

线性组合

如图 11 所示，以列向量 $\boldsymbol{v}_1$ 为例，它的每个元素相当于 $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D]$ 线性组合对应系数。将 $\boldsymbol{X}$ 向 $\boldsymbol{v}_1$ 投影：

$$\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{v}_1 \tag{15}$$

(15) 展开得到：

$$\boldsymbol{z}_1 = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_D \end{bmatrix} \underbrace{\begin{bmatrix} v_{1,1} \\ v_{2,1} \\ \vdots \\ v_{D,1} \end{bmatrix}}_{\boldsymbol{v}_1,\ \mathbf{PC1}} = v_{1,1}\boldsymbol{x}_1 + v_{2,1}\boldsymbol{x}_2 + \cdots + v_{D,1}\boldsymbol{x}_D \tag{16}$$

简单来讲，$\boldsymbol{z}_1$ 相当于 $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D]$ 的某种特殊线性组合。

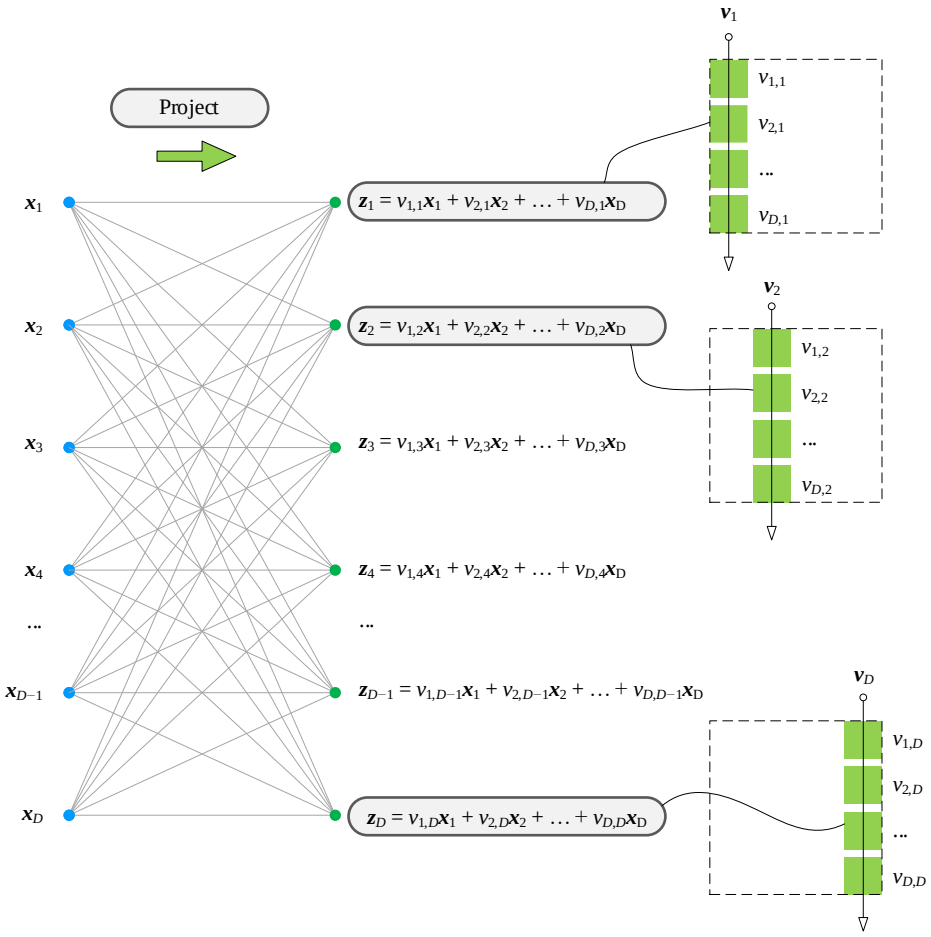


图 11. 线性组合角度看 PCA

朝向量投影

图 12 ~ 图 15 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 12 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 $\boldsymbol{z}_1$。图 13 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 $\boldsymbol{z}_2$。以此类推。

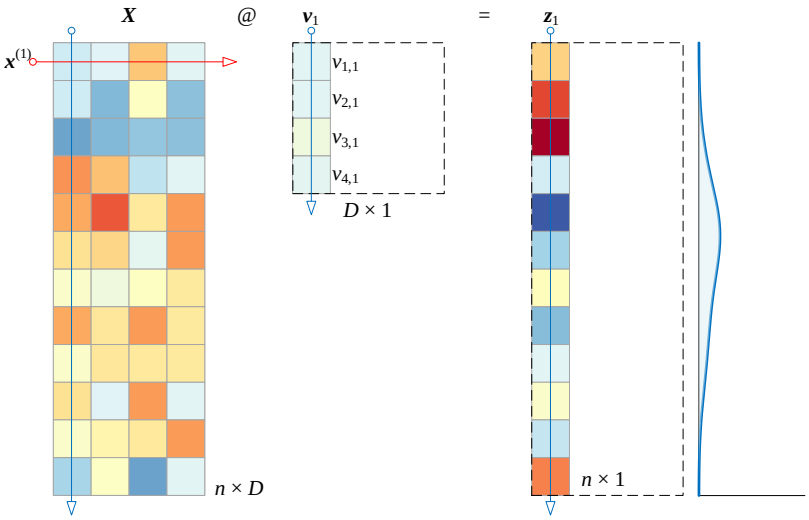


图 12. 数据 $\boldsymbol{X}$ 向 $\boldsymbol{v}_1$ 向量投影

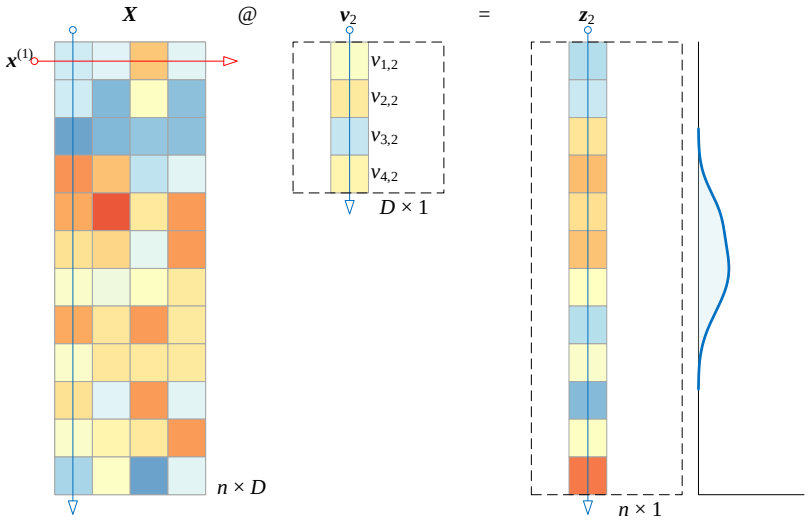


图 13. 数据 $\boldsymbol{X}$ 向 $\boldsymbol{v}_2$ 向量投影

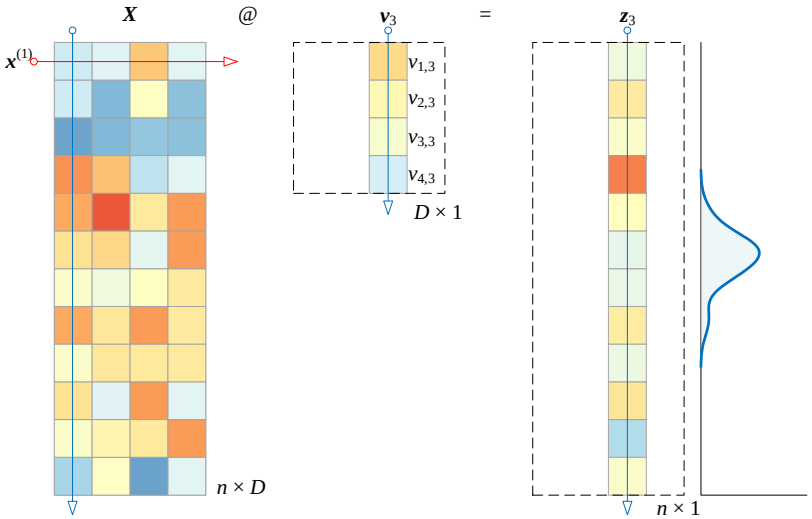


图 14. 数据 $\boldsymbol{X}$ 向 $\boldsymbol{v}_3$ 向量投影

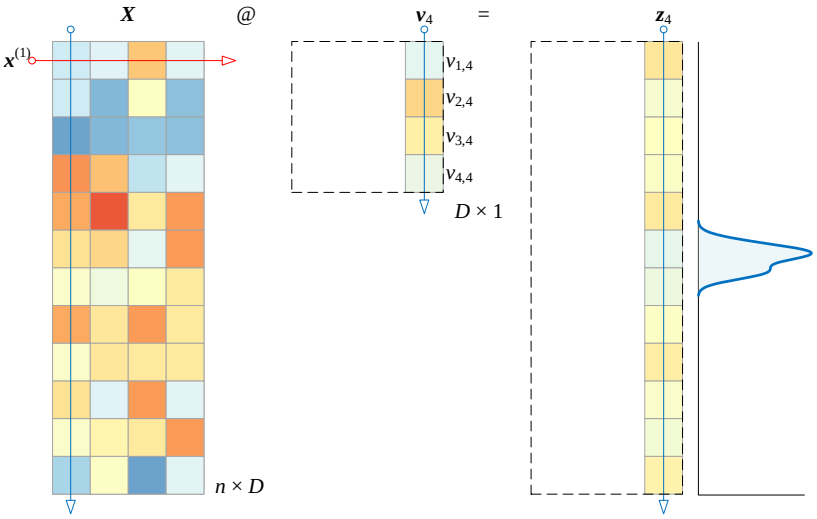


图 15. 数据 $\boldsymbol{X}$ 向 $\boldsymbol{v}_4$ 向量投影

朝平面投影

同样，$[\boldsymbol{z}_1, \boldsymbol{z}_2]$ 是 $\boldsymbol{X}$ 向 $[\boldsymbol{v}_1, \boldsymbol{v}_2]$ 投影结果，即四维数据 $\boldsymbol{X}$ 向二维空间投影。运算过程如下：

$$\begin{bmatrix} \boldsymbol{z}_1 & \boldsymbol{z}_2 \end{bmatrix} = \boldsymbol{X} \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{bmatrix} \tag{17}$$

图 16 所示为 (17) 运算过程及结果热图。

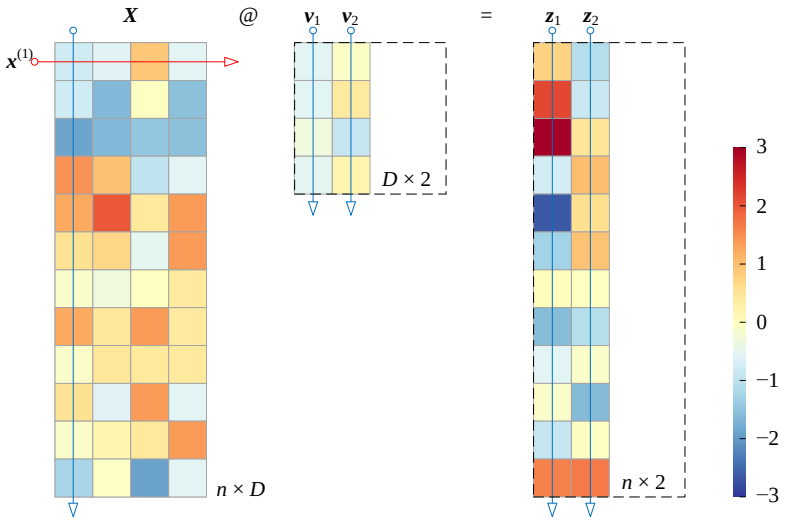


图 16. 数据 $\boldsymbol{X}$ 向 $[\boldsymbol{v}_1, \boldsymbol{v}_2]$ 投影

$\boldsymbol{Z}$ 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：

$$\operatorname{var}(\boldsymbol{z}_1) = \frac{(\boldsymbol{X}\boldsymbol{v}_1)^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{v}_1)}{n-1} = \frac{\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{v}_1}{n-1} = \boldsymbol{v}_1^{\mathrm{T}}\underbrace{\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{n-1}}_{\boldsymbol{\Sigma}}\boldsymbol{v}_1 = \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_1 \tag{18}$$

类似地，

$$\operatorname{var}(\boldsymbol{z}_2) = \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_2, \quad \ldots, \quad \operatorname{var}(\boldsymbol{z}_D) = \boldsymbol{v}_D^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_D \tag{19}$$

这样，$\boldsymbol{Z}$ 的协方差矩阵可以通过下式计算得到：

$$\begin{aligned}\operatorname{var}(\boldsymbol{Z}) &= \frac{(\boldsymbol{X}\boldsymbol{V})^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{V})}{n-1} = \frac{\boldsymbol{V}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{V}}{n-1} \\ &= \boldsymbol{V}^{\mathrm{T}}\underbrace{\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{n-1}}_{\boldsymbol{\Sigma}}\boldsymbol{V} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_1 & & & \\ & \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{v}_D^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}_D \end{bmatrix} = \boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_D \end{bmatrix}\end{aligned} \tag{20}$$

观察 (20) 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 (前提是其质心位于原点)，线性相关系数为 0。

$\boldsymbol{Z}_{n\times p}$ 的协方差矩阵为：

$$\operatorname{var}(\boldsymbol{Z}_{n\times p}) = \frac{(\boldsymbol{X}\boldsymbol{V}_{D\times p})^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{V}_{D\times p})}{n-1} = \boldsymbol{V}_{D\times p}{}^{\mathrm{T}}\underbrace{\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{n-1}}_{\boldsymbol{\Sigma}}\boldsymbol{V}_{D\times p} = \boldsymbol{V}_{D\times p}{}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{V}_{D\times p} = \boldsymbol{\Lambda}_{p\times p} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix} \tag{21}$$

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 17 所示，椭圆中心对应质心 $\boldsymbol{\mu}$，椭圆和$\pm\sigma$ 标准差构成的矩形相切，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。

本书前文介绍过，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 17 中，EF 为椭圆长轴；FH 为椭圆短轴。而 EF 就相当于 PCA 的第一主成分，FH 为第二主成分。

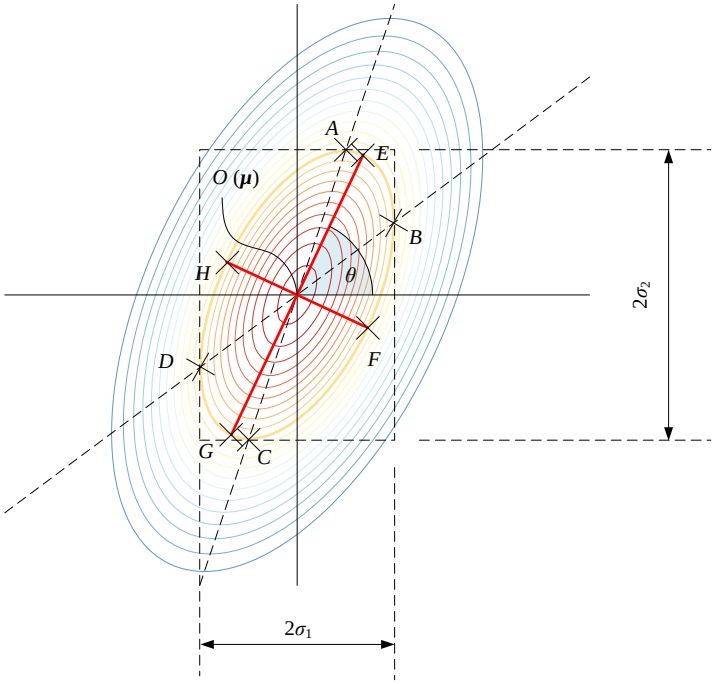


图 17. 主成分分析和椭圆的关系

图 18 则从椭圆视角解释主成分分析。假设图 18 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 PC1。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“鸢尾花书”《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。

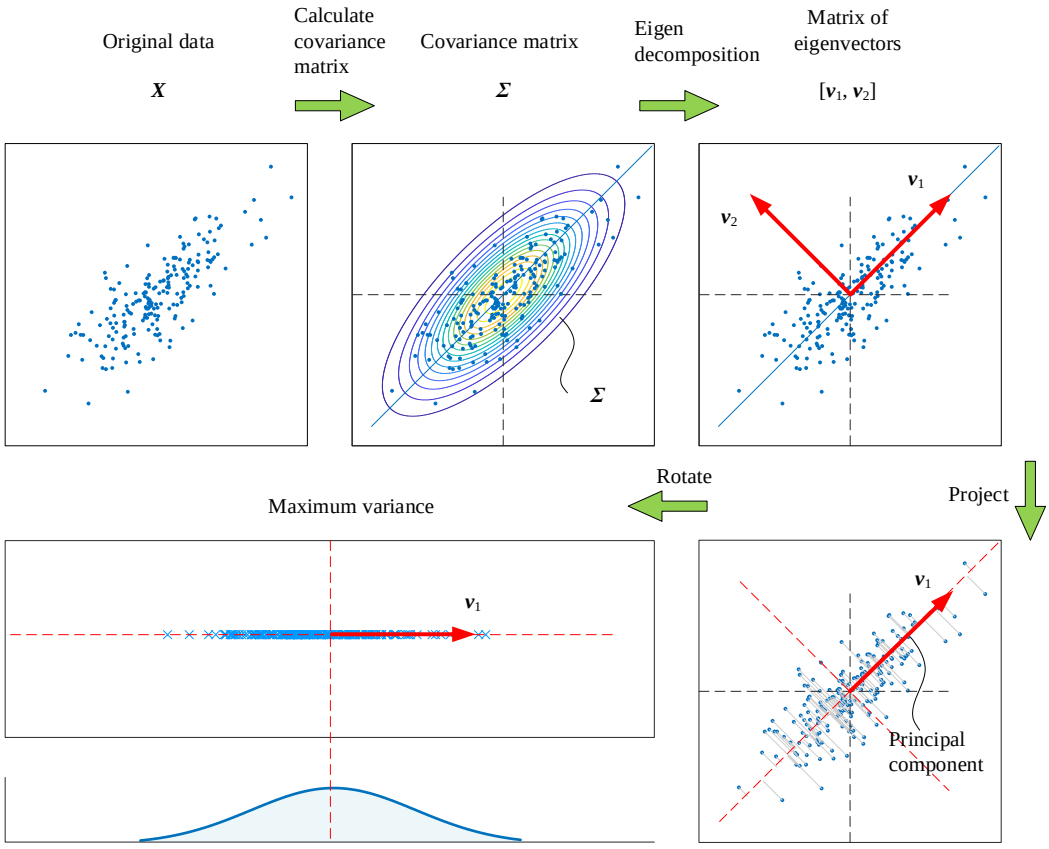


图 18. 几何视角下通过特征值分解协方差矩阵进行主成分分析

如图 19 所示，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。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，得到的数据标准差最大；朝椭圆短轴方向投影得到的数据标准差最小。

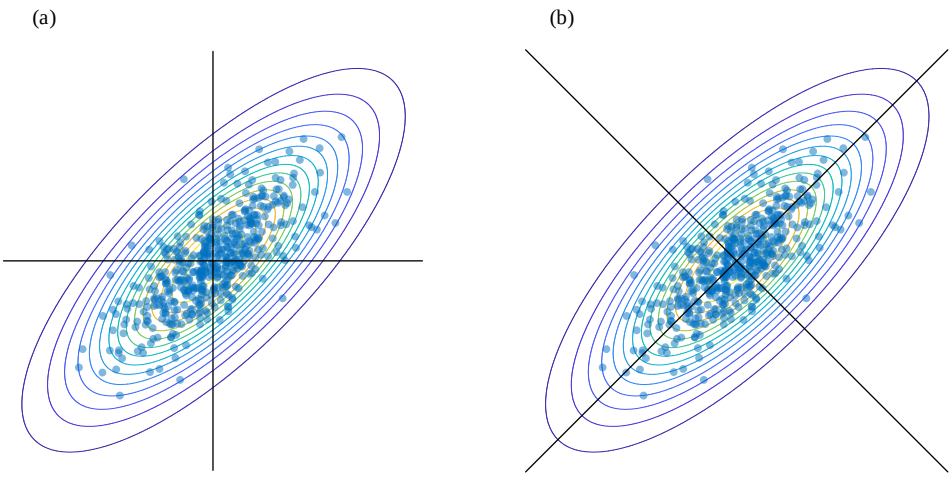


图 19. 两个角度看数据

举个例子

图 20 (a) 所示为原始二维数据 $\boldsymbol{X}$ 的散点图，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。

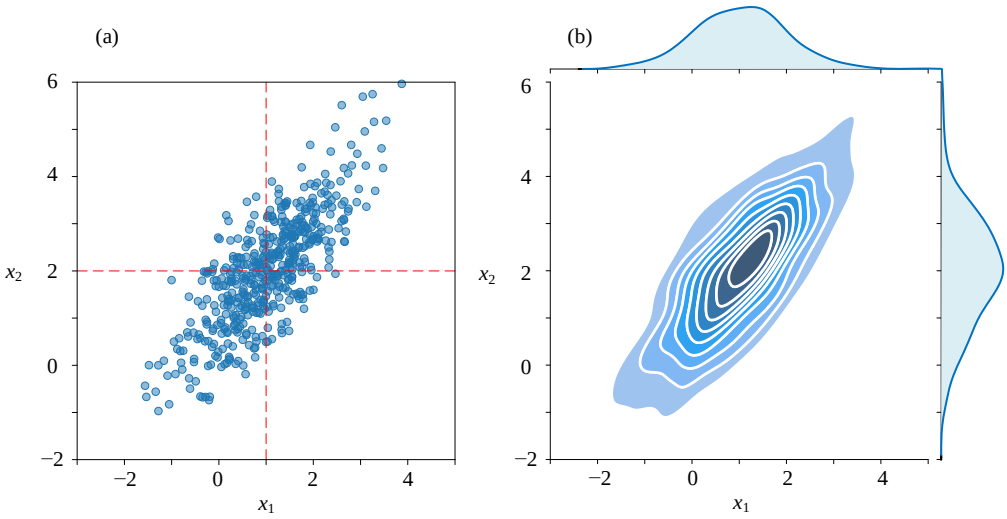


图 20. 原始二维数据 $\boldsymbol{X}$

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 21 对比 $\boldsymbol{Y}$ 两列数据分布。图 22 所示为数据 $\boldsymbol{Y}$ 在 $[\boldsymbol{v}_1, \boldsymbol{v}_2]$ 中散点图。

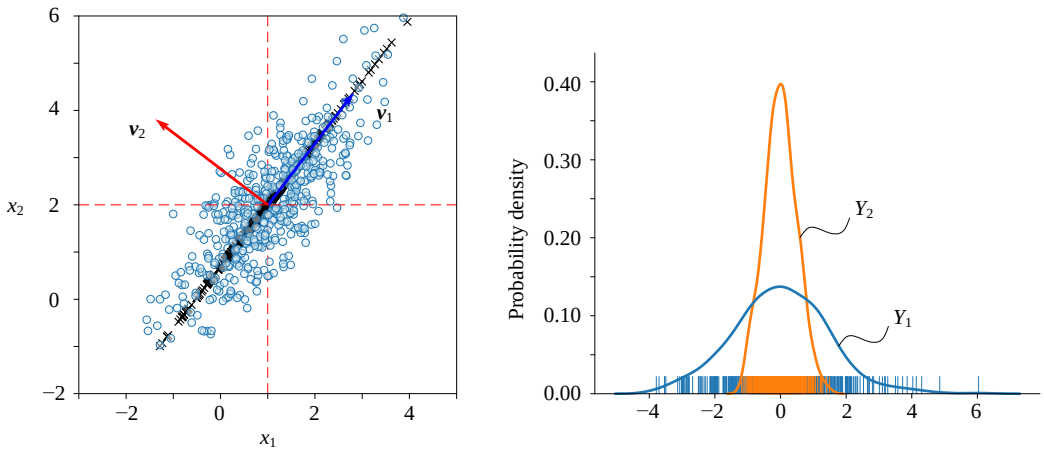


图 21. 主成分数据分布

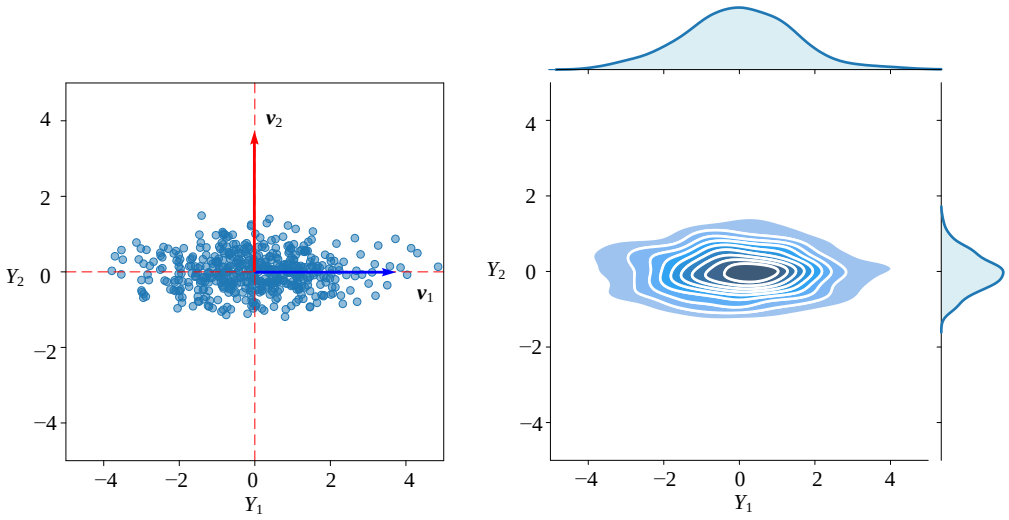


图 22. 数据 $\boldsymbol{Y}$ 在 [$\boldsymbol{v}_1$, $\boldsymbol{v}_2$] 中散点图

Bk5_Ch25_01.py 绘制图 20 ~ 图 22。

25.6 奇异值分解

四种奇异值分解

奇异值分解 (singular value decomposition, SVD) 也可以用来做主成分分析。丛书在《矩阵力量》一本系统讲解过奇异值分解的四种类型：

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

如图 23 所示，完全型奇异值分解中，$\boldsymbol{U}$ 为方阵，$\boldsymbol{S}$ 矩阵并非方阵。

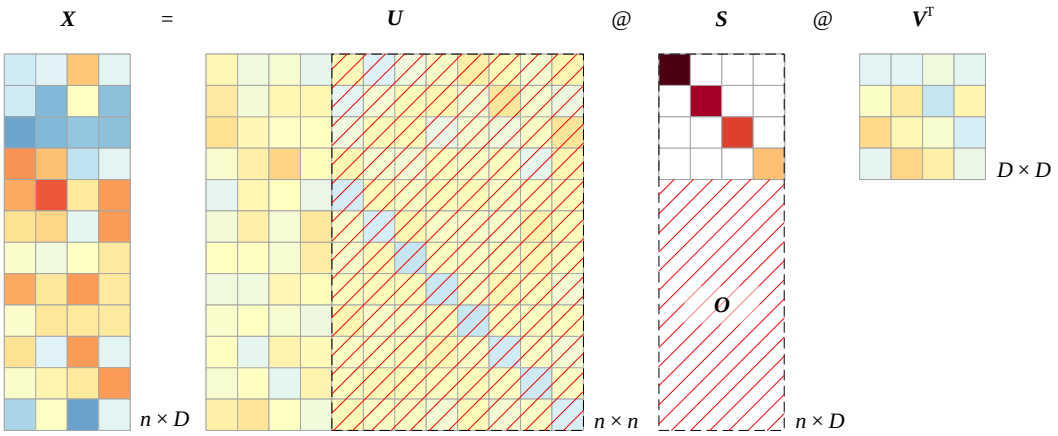


图 23. 完全 (full) 奇异值分解

去掉图 23 中这个全 0 矩阵 $\boldsymbol{O}$，便得到经济型奇异值分解，具体如图 24 所示。经济型 SVD 中，$\boldsymbol{U}$ 的形状和 $\boldsymbol{X}$ 相同，$\boldsymbol{S}$ 矩阵为对角方阵，形状为 $D \times D$。

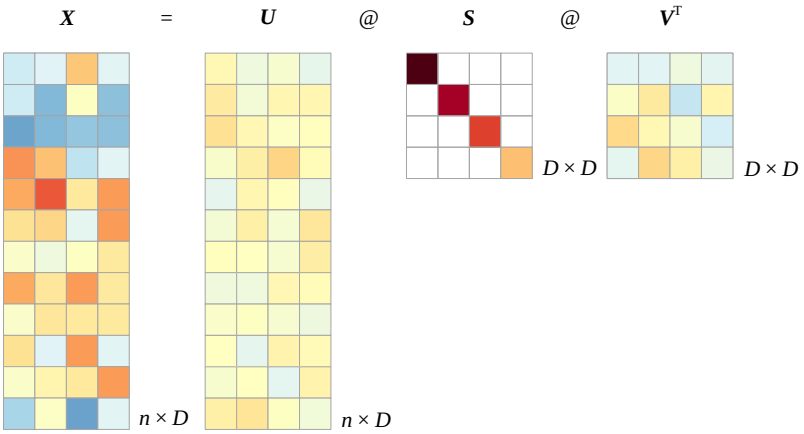


图 24. 经济型奇异值分解

当 $\boldsymbol{X}$ 非满秩时，即 $\text{rank}(\boldsymbol{X}) = r < D$，图 24 经济型奇异值分解可以进一步简化为如图 25 所示的紧凑型 SVD 分解。

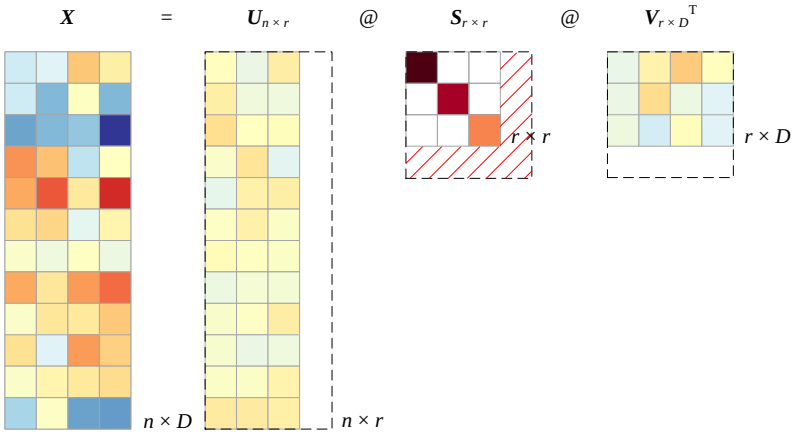


图 25. 紧凑型奇异值分解，$\boldsymbol{X}$ 非满秩

在线性代数中，矩阵的秩指的是其列向量或行向量的线性无关的数量。如果矩阵的秩等于它的行数或列数中的较小值，则称该矩阵为满秩矩阵。如果矩阵的秩小于它的行数或列数中的较小值，则称该矩阵为非满秩矩阵。

在机器学习中，非满秩的矩阵通常表示存在冗余或线性相关的特征或样本。这些冗余或线性相关的特征或样本可能会导致算法的过拟合，降低模型的准确性和稳定性。因此，在许多机器学习算法中，对于非满秩矩阵，通常需要进行一些特殊的处理，例如降维或正则化，以减少冗余或相关性，并提高模型的效果。

图 26 给出的是截断型奇异值分解，$\boldsymbol{S}_{p\times p}$ 仅使用图 24 中 $\boldsymbol{S}$ 矩阵 p 个主成分特征值，形状为 $p \times p$。注意，图 26 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 $\boldsymbol{X}$ 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。

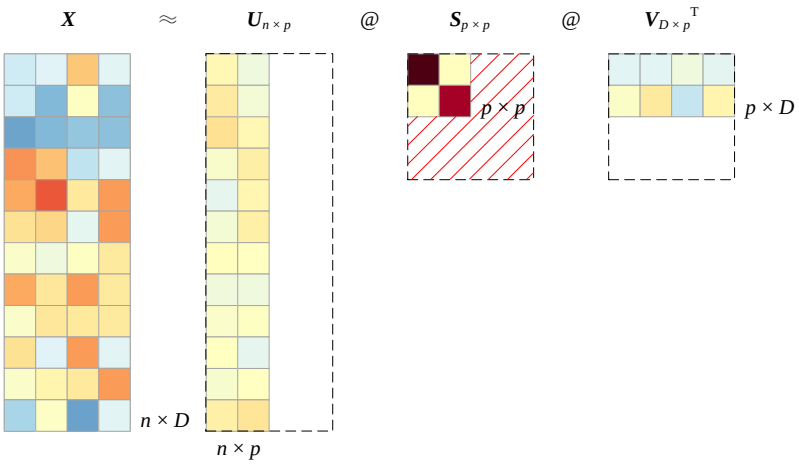


图 26. 截断型奇异值分解

SVD 完成主成分分析

首先中心化 (去均值) 数据矩阵。对已经去均值的矩阵 $\boldsymbol{X}_{n\times D}$ 进行完全型 SVD 分解，得到：

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}} \tag{22}$$

$\boldsymbol{V}$ 和 $\boldsymbol{U}$ 均为正交矩阵，即满足：

$$\begin{aligned}\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} &= \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I} \\ \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} &= \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{I}\end{aligned} \tag{23}$$

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

由于 $\boldsymbol{X}$ 已经中心化，其协方差矩阵可以通过下式计算获得：

$$\boldsymbol{\Sigma} = \frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{n-1} \tag{24}$$

将(22) 代入 (24) 得到：

$$\boldsymbol{\Sigma} = \frac{\left(\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}\right)^{\mathrm{T}}\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}}{n-1} = \frac{\boldsymbol{V}\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}}{n-1} \tag{25}$$

对协方差矩阵进行特征值分解：

$$\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}} \tag{26}$$

联立 (25) 和 (26)，

$$\frac{\boldsymbol{V}\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}}{n-1} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}} \tag{27}$$

对于经济型 SVD 分解，$\boldsymbol{S}$ 为对角方阵，(27) 整理得到：

$$\frac{\boldsymbol{S}^2}{n-1} = \boldsymbol{\Lambda} \tag{28}$$

即

$$\frac{1}{n-1}\begin{bmatrix} s_1^2 & & & \\ & s_2^2 & & \\ & & \ddots & \\ & & & s_D^2 \end{bmatrix} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_D \end{bmatrix} \tag{29}$$

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：

$$\frac{s_j^2}{n-1} = \lambda_j \tag{30}$$

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

理解 $\boldsymbol{U}$

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

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{V}^{-1} = \boldsymbol{Z}\boldsymbol{V}^{\mathrm{T}} \tag{31}$$

对比 (22) 和 $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}$，可以发现：

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{S} \tag{32}$$

也就是

$$\begin{bmatrix} \boldsymbol{z}_1 & \boldsymbol{z}_2 & \cdots & \boldsymbol{z}_D \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_D \end{bmatrix}\begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_D \end{bmatrix} = \begin{bmatrix} s_1\boldsymbol{u}_1 & s_2\boldsymbol{u}_2 & \cdots & s_D\boldsymbol{u}_D \end{bmatrix} \tag{33}$$

即：

$$s_1\boldsymbol{u}_1=\boldsymbol{z}_1,\quad s_2\boldsymbol{u}_2=\boldsymbol{z}_2,\ \ldots \tag{34}$$

对 $\boldsymbol{z}_1$ 求方差：

$$\operatorname{var}(\boldsymbol{z}_1)=\frac{\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_1}{n-1}=\frac{(s_1\boldsymbol{u}_1)^{\mathrm{T}}(s_1\boldsymbol{u}_1)}{n-1}=\frac{s_1^2\|\boldsymbol{u}_1\|^2}{n-1}=\frac{s_1^2}{n-1}=\lambda_1 \tag{35}$$

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

$$\boldsymbol{U}=\begin{bmatrix}\boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_D\end{bmatrix}=\begin{bmatrix}\frac{\boldsymbol{z}_1}{s_1} & \frac{\boldsymbol{z}_2}{s_2} & \cdots & \frac{\boldsymbol{z}_D}{s_D}\end{bmatrix} \tag{36}$$

也就是：

$$\boldsymbol{U}=\begin{bmatrix}\boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_D\end{bmatrix}=\boldsymbol{Z}\boldsymbol{S}^{-1} \tag{37}$$

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：

$$\begin{aligned}\boldsymbol{X}&=\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}\\&=\begin{bmatrix}\boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_D\end{bmatrix}\begin{bmatrix}s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_D\end{bmatrix}\begin{bmatrix}\boldsymbol{v}_1^{\mathrm{T}}\\ \boldsymbol{v}_2^{\mathrm{T}}\\ \vdots\\ \boldsymbol{v}_D^{\mathrm{T}}\end{bmatrix}\\&=s_1\boldsymbol{u}_1\boldsymbol{v}_1^{\mathrm{T}}+s_2\boldsymbol{u}_2\boldsymbol{v}_2^{\mathrm{T}}+\cdots+s_D\boldsymbol{u}_D\boldsymbol{v}_D^{\mathrm{T}}\\&=s_1\boldsymbol{u}_1\otimes\boldsymbol{v}_1+s_2\boldsymbol{u}_2\otimes\boldsymbol{v}_2+\cdots+s_D\boldsymbol{u}_D\otimes\boldsymbol{v}_D\end{aligned} \tag{38}$$

图 27 所示为 (38) 还原原始数据的过程。

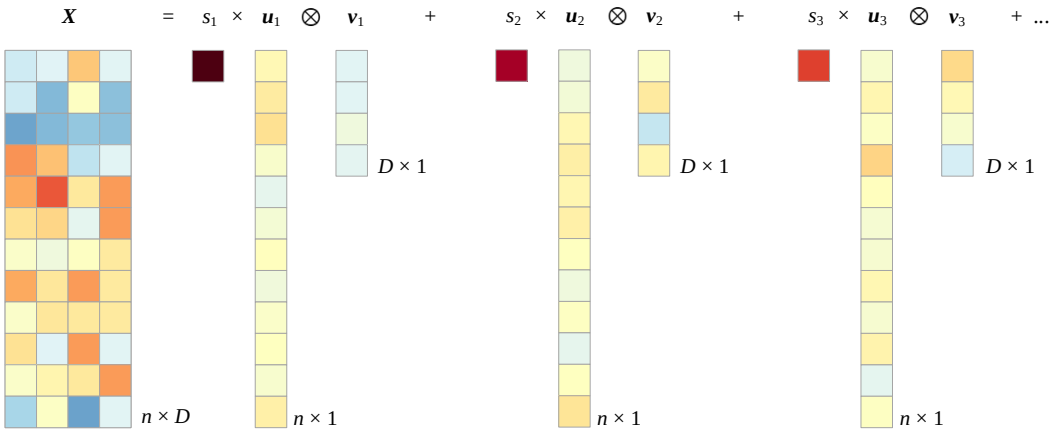


图 27. 张量积 $s_1\boldsymbol{u}_1\otimes\boldsymbol{v}_1$、$s_2\boldsymbol{u}_2\otimes\boldsymbol{v}_2$ 等之和还原数据 $\boldsymbol{X}$

25.7 优化问题

下面我们从优化角度理解 PCA。如图 28 所示，$\boldsymbol{X}$ 为中心化数据，即 $\boldsymbol{X}$ 质心零向量。$\boldsymbol{v}$ 为单位向量。数据 $\boldsymbol{X}$ 在 $\boldsymbol{v}$ 上投影结果为 $\boldsymbol{z}$，即 $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{v}$ 。

主成分分析中，选取 $\boldsymbol{v}$ 的标准是——$\boldsymbol{z}$ 方差最大化。这便是构造 PCA 优化问题的第一个角度。

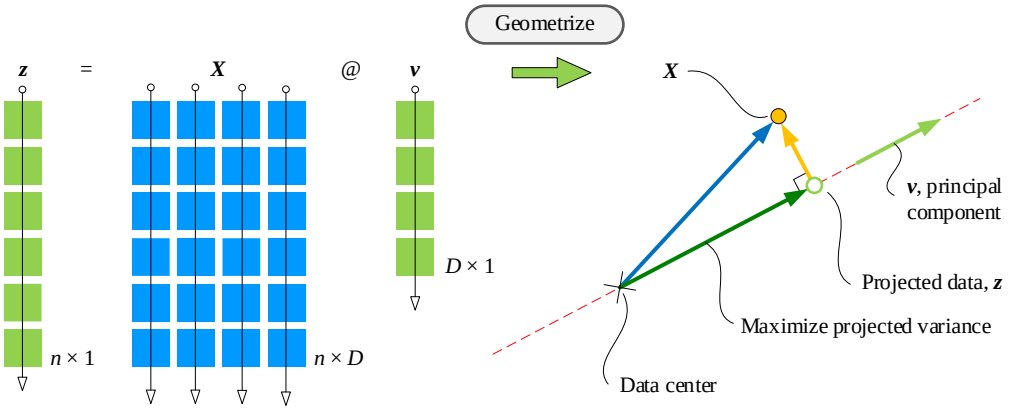


图 28. 主成分分析优化问题

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：

$$\operatorname{var}(\boldsymbol{z}) = \frac{\boldsymbol{z}^{\mathrm{T}}\boldsymbol{z}}{n-1} = \boldsymbol{v}^{\mathrm{T}} \overbrace{\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{n-1}}^{\text{Covariance matrix}} \boldsymbol{v} \tag{39}$$

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

$$\operatorname{var}(\boldsymbol{z}) = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v} \tag{40}$$

$\boldsymbol{v}$ 为单位列向量，即满足如下约束条件：

$$\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} = 1 \tag{41}$$

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

$$\begin{aligned} &\arg\max_{\boldsymbol{v}} \ \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v} \\ &\text{subject to: } \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} - 1 = 0 \end{aligned} \tag{42}$$

上式最大化优化问题等价于如下最小化优化问题：

$$\begin{aligned} &\arg\min_{\boldsymbol{v}} \quad -\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v} \\ &\text{subject to:} \quad \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}-1=0 \end{aligned} \tag{43}$$

构造拉格朗日函数 $L(\boldsymbol{v},\lambda)$：

$$L(\boldsymbol{v},\lambda)=-\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}+\lambda\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}-1\right) \tag{44}$$

λ 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：

$$\nabla_{\boldsymbol{v}}L(\boldsymbol{v},\lambda)=\frac{\partial L(\boldsymbol{v},\lambda)}{\partial \boldsymbol{v}}=\left(-2\boldsymbol{\Sigma}\boldsymbol{v}+2\lambda\boldsymbol{v}\right)^{\mathrm{T}}=\boldsymbol{0} \tag{45}$$

有关拉格朗日乘子法，请大家回顾《矩阵力量》第 18 章。

整理 (45) 得到：

$$\boldsymbol{\Sigma}\boldsymbol{v}=\lambda\boldsymbol{v} \tag{46}$$

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

$$\mathrm{var}(\boldsymbol{z})=\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{v}=\boldsymbol{v}^{\mathrm{T}}\lambda\boldsymbol{v}=\lambda\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}=\lambda \tag{47}$$

也就是说，var($\boldsymbol{z}$) 最大值对应 $\boldsymbol{\Sigma}$ 最大特征值。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 29 所示为还原得到 $\boldsymbol{X}$ 过程。图 30 所示热图，矩阵 $\boldsymbol{Z}$ 还原转化为原始数据矩阵 $\boldsymbol{X}$。

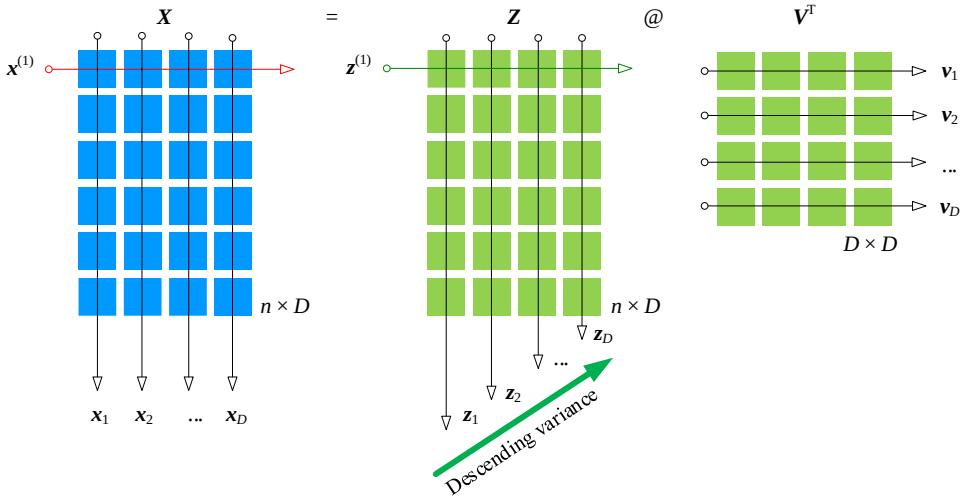


图 29. 反向还原数据 $\boldsymbol{X}=\boldsymbol{Z}\boldsymbol{V}^{\mathrm{T}}$

再次强调，图 29 这种还原计算成立的条件是 $\boldsymbol{X}$ 的质心位于原点。

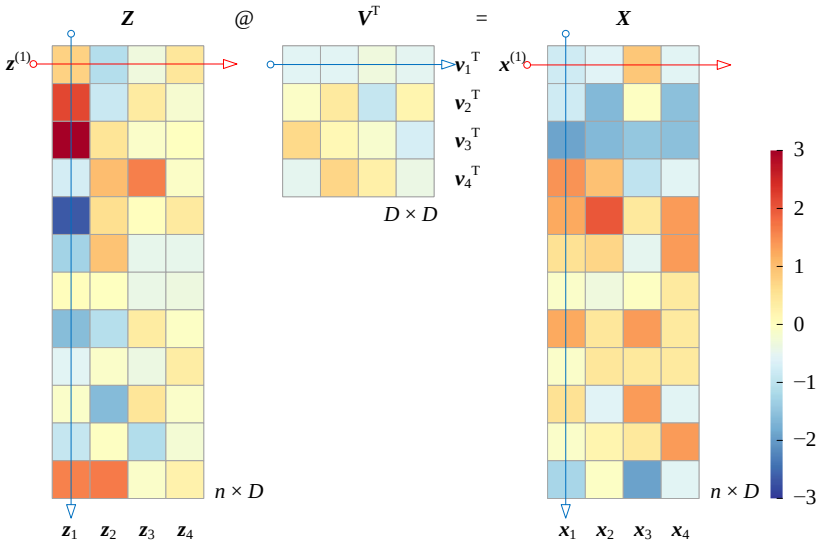


图 30. 新特征数据矩阵 $\boldsymbol{Z}$ 还原转化为原始数据矩阵 $\boldsymbol{X}$

$\boldsymbol{X}=\boldsymbol{Z}\boldsymbol{V}^{\mathrm{T}}$ 展开得到下式：

$$\boldsymbol{X}=\begin{bmatrix}\boldsymbol{z}_1 & \boldsymbol{z}_2 & \boldsymbol{z}_3 & \boldsymbol{z}_4\end{bmatrix}\begin{bmatrix}\boldsymbol{v}_1^{\mathrm{T}}\\ \boldsymbol{v}_2^{\mathrm{T}}\\ \boldsymbol{v}_3^{\mathrm{T}}\\ \boldsymbol{v}_4^{\mathrm{T}}\end{bmatrix}=\underbrace{\boldsymbol{z}_1\boldsymbol{v}_1^{\mathrm{T}}}_{\hat{\boldsymbol{X}}_1}+\underbrace{\boldsymbol{z}_2\boldsymbol{v}_2^{\mathrm{T}}}_{\hat{\boldsymbol{X}}_2}+\underbrace{\boldsymbol{z}_3\boldsymbol{v}_3^{\mathrm{T}}}_{\hat{\boldsymbol{X}}_3}+\underbrace{\boldsymbol{z}_4\boldsymbol{v}_4^{\mathrm{T}}}_{\hat{\boldsymbol{X}}_4} \tag{48}$$

(48) 所示运算过程如图 31 所示。

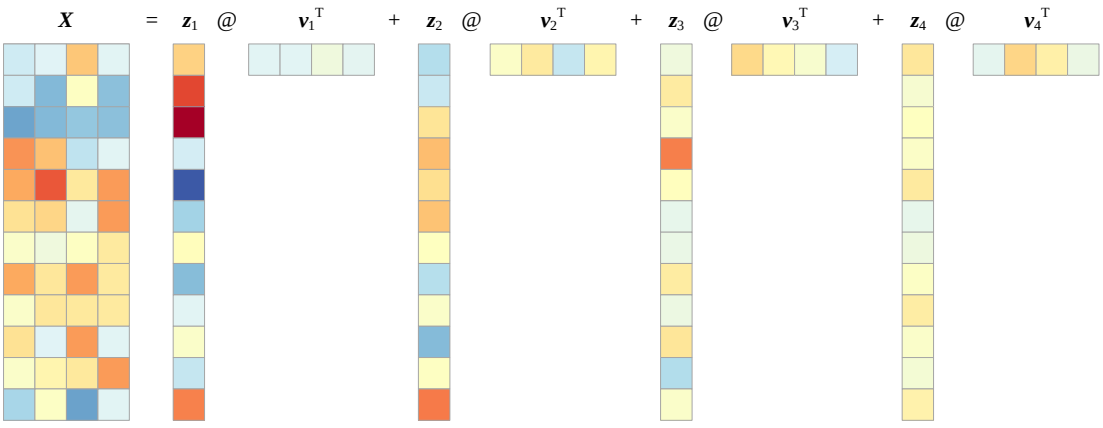


图 31. 还原原始数据运算

图 32 所示为 $\boldsymbol{z}_1$ 还原 $\boldsymbol{X}$ 部分数据，对应运算如下：

$$\boldsymbol{X}_1 = \boldsymbol{z}_1 \boldsymbol{v}_1^{\mathrm{T}} \tag{49}$$

展开上式得到：

$$\begin{aligned} \boldsymbol{X}_1 &= \boldsymbol{z}_1 \boldsymbol{v}_1^{\mathrm{T}} \\ &= \boldsymbol{z}_1 \begin{bmatrix} v_{1,1} & v_{2,1} & \cdots & v_{D,1} \end{bmatrix} \\ &= \begin{bmatrix} v_{1,1}\boldsymbol{z}_1 & v_{2,1}\boldsymbol{z}_1 & \cdots & v_{D,1}\boldsymbol{z}_1 \end{bmatrix} \end{aligned} \tag{50}$$

观察图 32 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；(50) 解释了这一点，$\boldsymbol{X}_1$ 的每一列均是标量乘以向量 $\boldsymbol{z}_1$ 的结果。显然，$\boldsymbol{X}_1$ 的秩为 1，即 $\mathrm{rank}(\boldsymbol{X}_1) = 1$。

图 33、图 34 和图 35 分别展示 $\boldsymbol{z}_2$、$\boldsymbol{z}_3$ 和 $\boldsymbol{z}_4$ 还原 $\boldsymbol{X}$ 部分数据。

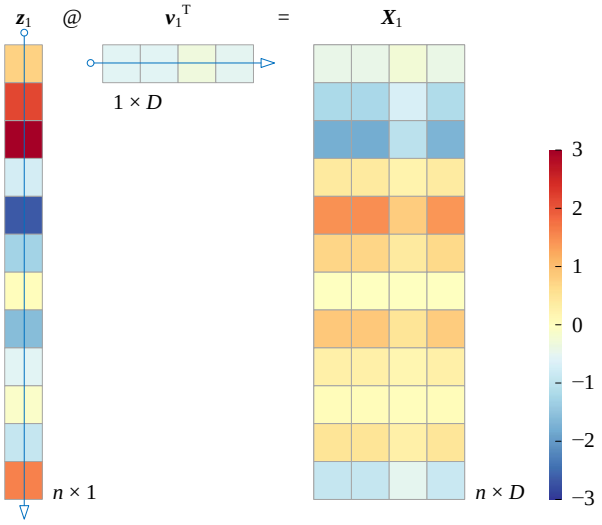


图 32. $\boldsymbol{z}_1$ 还原 $\boldsymbol{X}$ 部分数据

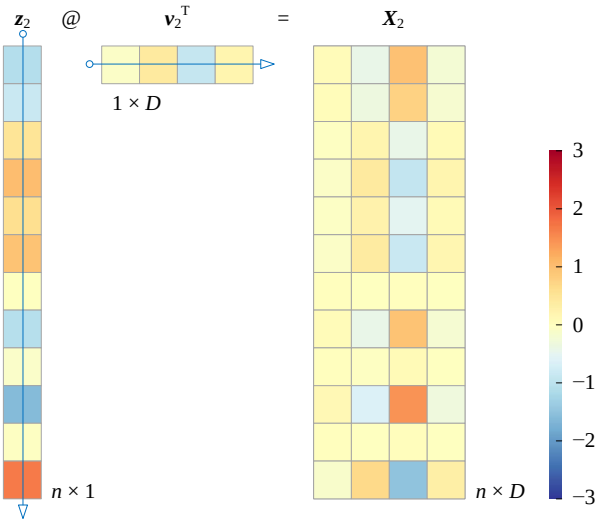


图 33. $\boldsymbol{z}_2$ 还原 $\boldsymbol{X}$ 部分数据

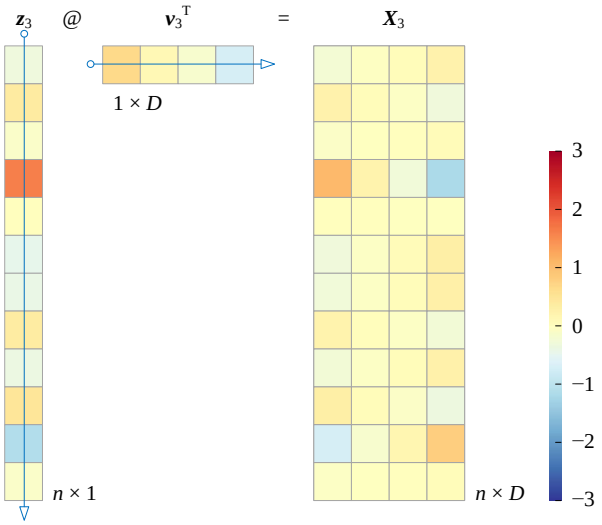


图 34. z_3 还原 X 部分数据

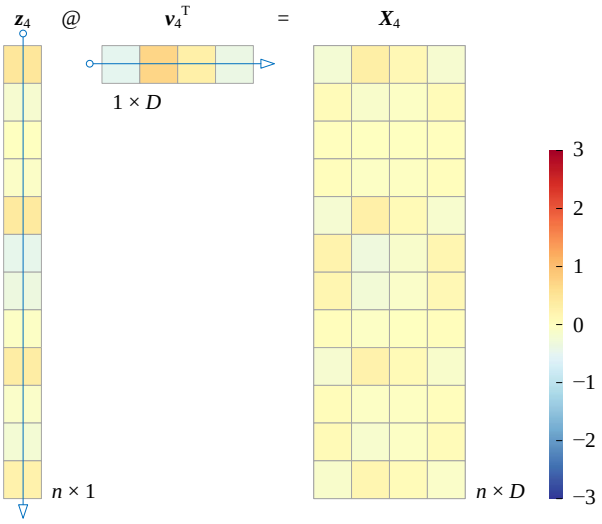


图 35. z_4 还原 X 部分数据

图 36 所示为原始数据矩阵 X 热图相当于四层热图叠加结果。观察图 36，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。

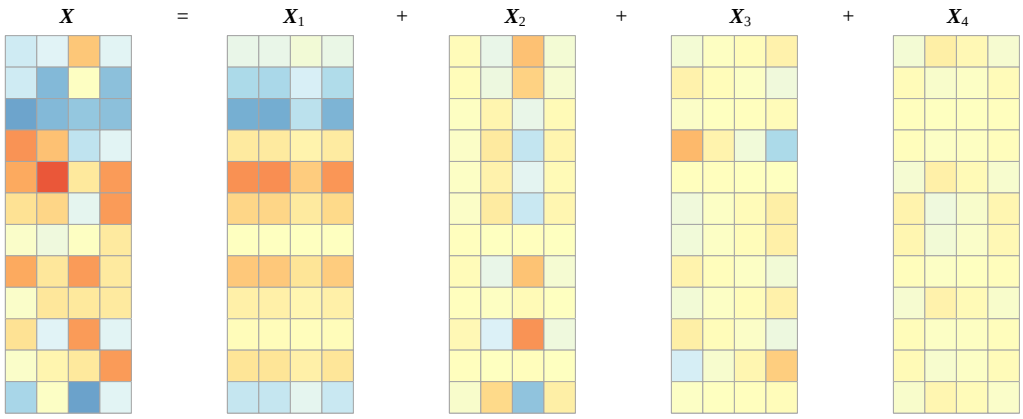


图 36. 原始数据矩阵 $\boldsymbol{X}$ 热图于四层热图叠加结果

张量积

另外，(48) 可以用张量积来表达：

$$\boldsymbol{X} = \underbrace{\boldsymbol{z}_1 \otimes \boldsymbol{v}_1}_{\hat{\boldsymbol{X}}_1} + \underbrace{\boldsymbol{z}_2 \otimes \boldsymbol{v}_2}_{\hat{\boldsymbol{X}}_2} + \underbrace{\boldsymbol{z}_3 \otimes \boldsymbol{v}_3}_{\hat{\boldsymbol{X}}_3} + \underbrace{\boldsymbol{z}_4 \otimes \boldsymbol{v}_4}_{\hat{\boldsymbol{X}}_4} \tag{51}$$

利用 (14)，(48) 可以整理为：

$$\boldsymbol{X} = \boldsymbol{X}\boldsymbol{v}_1\boldsymbol{v}_1^{\mathrm{T}} + \boldsymbol{X}\boldsymbol{v}_2\boldsymbol{v}_2^{\mathrm{T}} + \ldots + \boldsymbol{X}\boldsymbol{v}_D\boldsymbol{v}_D^{\mathrm{T}} = \sum_{j=1}^{D} \boldsymbol{X}\boldsymbol{v}_j\boldsymbol{v}_j^{\mathrm{T}} = \boldsymbol{X}\left(\sum_{j=1}^{D} \boldsymbol{v}_j\boldsymbol{v}_j^{\mathrm{T}}\right) \tag{52}$$

(52) 可以用张量积表达：

$$\boldsymbol{X} = \boldsymbol{X}\left(\boldsymbol{v}_1 \otimes \boldsymbol{v}_1\right) + \boldsymbol{X}\left(\boldsymbol{v}_2 \otimes \boldsymbol{v}_2\right) + \ldots + \boldsymbol{X}\left(\boldsymbol{v}_D \otimes \boldsymbol{v}_D\right) = \sum_{j=1}^{D} \boldsymbol{X}\boldsymbol{v}_j \otimes \boldsymbol{v}_j = \boldsymbol{X}\left(\sum_{j=1}^{D} \boldsymbol{v}_j \otimes \boldsymbol{v}_j\right) \tag{53}$$

图 37 所示为通过主成分 $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, $\boldsymbol{v}_3$, $\boldsymbol{v}_4$ 和其自身转置乘积计算张量积。

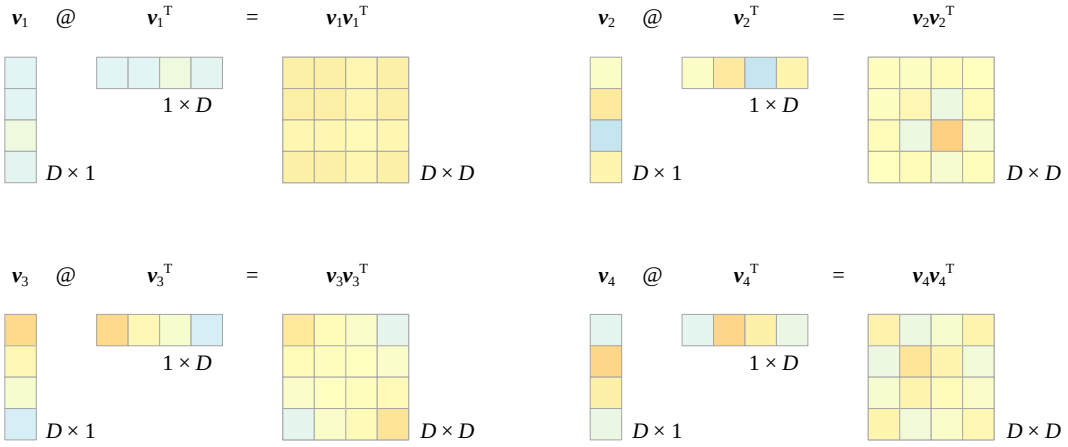


图 37. 列向量乘自身转置获得四个张量积

图 38 所示为张量积运算，和图 37 结果完全一致。

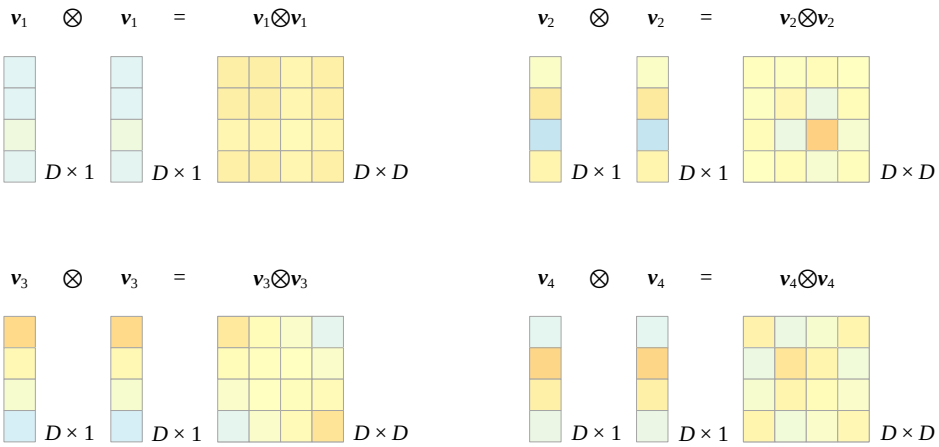


图 38. 内积计算获得四个张量积

容易推导得到，(53) 中张量积相加得到单位矩阵：

$$\boldsymbol{v}_1 \otimes \boldsymbol{v}_1 + \boldsymbol{v}_2 \otimes \boldsymbol{v}_2 + \ldots + \boldsymbol{v}_D \otimes \boldsymbol{v}_D = \left(\sum_{j=1}^{D} \boldsymbol{v}_j \otimes \boldsymbol{v}_j \right) = \boldsymbol{I} \tag{54}$$

上式如图 39 热图所示。

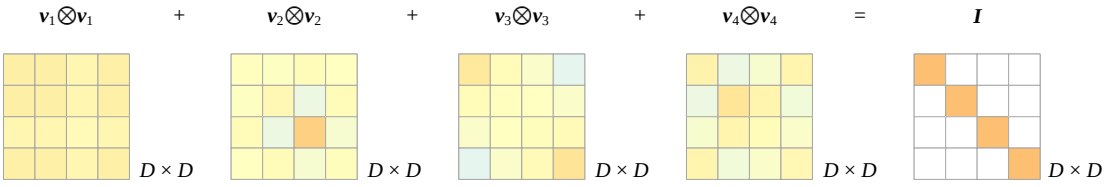


图 39. 张量积相加得到单位矩阵

联立 (15) 和 (49)，利用张量积 $\boldsymbol{v}_1 \otimes \boldsymbol{v}_1$ 还原部分原始数据：

$$\boldsymbol{X}_1 = \boldsymbol{z}_1 \boldsymbol{v}_1^{\mathrm{T}} = \boldsymbol{X} \boldsymbol{v}_1 \boldsymbol{v}_1^{\mathrm{T}} = \boldsymbol{X} \underbrace{\left(\boldsymbol{v}_1 \otimes \boldsymbol{v}_1 \right)}_{\text{Tensor product}} \tag{55}$$

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

$$\boldsymbol{X}_2 = \boldsymbol{z}_2 \boldsymbol{v}_2^{\mathrm{T}} = \boldsymbol{X} \boldsymbol{v}_2 \boldsymbol{v}_2^{\mathrm{T}} = \boldsymbol{X} \underbrace{\left(\boldsymbol{v}_2 \otimes \boldsymbol{v}_2 \right)}_{\text{Tensor product}} \tag{56}$$

图 40 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 41 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 $\boldsymbol{X}$。

《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。

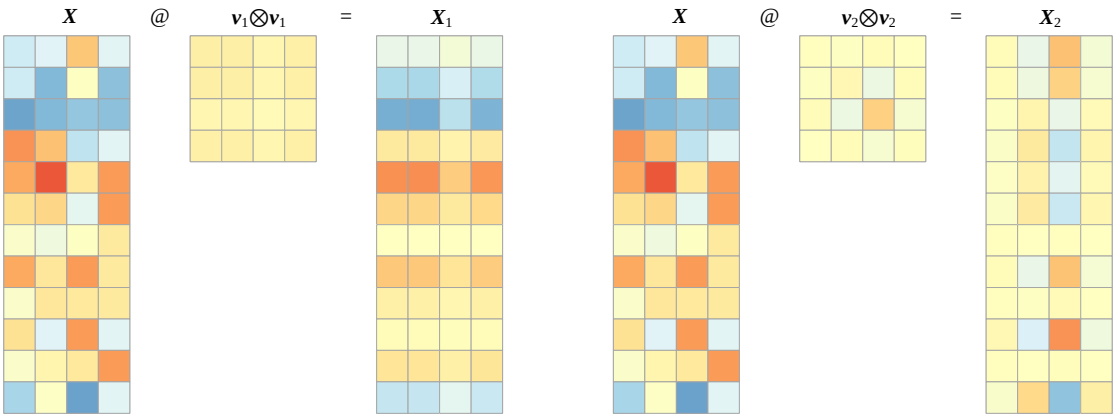


图 40. 张量积 $\boldsymbol{X}(\boldsymbol{v}_1\otimes\boldsymbol{v}_1)$ 和 $\boldsymbol{X}(\boldsymbol{v}_2\otimes\boldsymbol{v}_2)$ 还原部分数据 $\boldsymbol{X}$

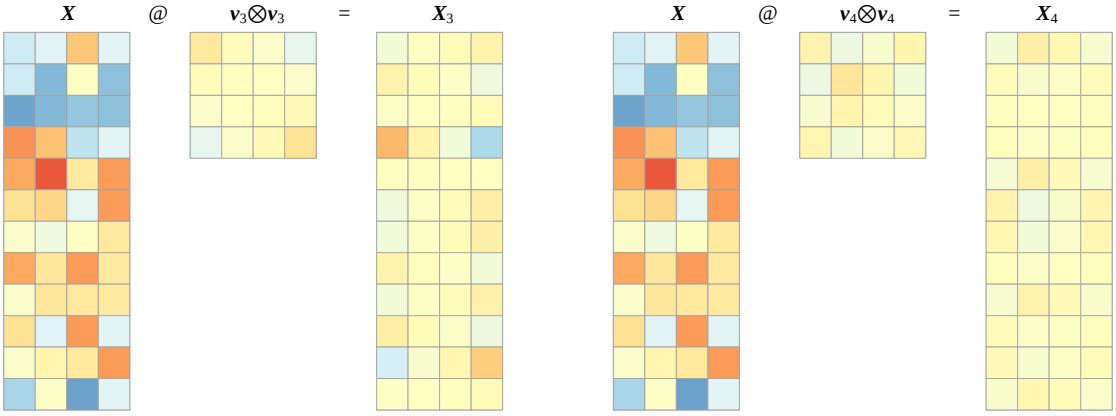


图 41. 张量积 $\boldsymbol{X}(\boldsymbol{v}_3\otimes\boldsymbol{v}_3)$ 和 $\boldsymbol{X}(\boldsymbol{v}_4\otimes\boldsymbol{v}_4)$ 还原部分数据 $\boldsymbol{X}$

误差

图 42 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：

$$\hat{\boldsymbol{X}} = \begin{bmatrix}\boldsymbol{z}_1 & \boldsymbol{z}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{v}_1 & \boldsymbol{v}_2\end{bmatrix}^{\mathrm{T}} \tag{57}$$

相当于

$$\begin{aligned}\hat{\boldsymbol{X}} &= \boldsymbol{X}_1 + \boldsymbol{X}_2 = \boldsymbol{z}_1\boldsymbol{v}_1^{\mathrm{T}} + \boldsymbol{z}_2\boldsymbol{v}_2^{\mathrm{T}} \\ &= \boldsymbol{X}\left(\boldsymbol{v}_1\boldsymbol{v}_1^{\mathrm{T}} + \boldsymbol{v}_2\boldsymbol{v}_2^{\mathrm{T}}\right) = \boldsymbol{X}\left(\boldsymbol{v}_1\otimes\boldsymbol{v}_1 + \boldsymbol{v}_2\otimes\boldsymbol{v}_2\right)\end{aligned} \tag{58}$$

图 43 所示为通过叠加图 32 和图 33 两个热图还原原始数据矩阵。

从张量积角度来看图 43，

$$\boldsymbol{X} \approx \boldsymbol{X}\left(\boldsymbol{v}_1 \otimes \boldsymbol{v}_1 + \boldsymbol{v}_2 \otimes \boldsymbol{v}_2\right) = s_1\boldsymbol{u}_1 \otimes \boldsymbol{v}_1 + s_2\boldsymbol{u}_2 \otimes \boldsymbol{v}_2^{\mathrm{T}} \tag{59}$$

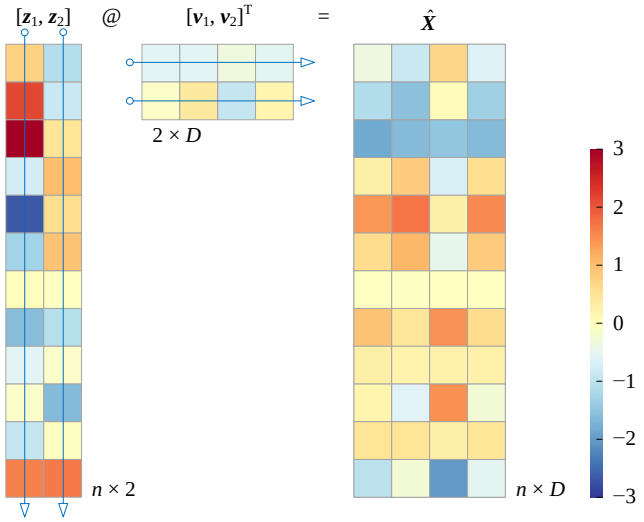


图 42. 前两个主成分 $\boldsymbol{z}_1$ 和 $\boldsymbol{z}_1$ 还原 $\boldsymbol{X}$ 数据

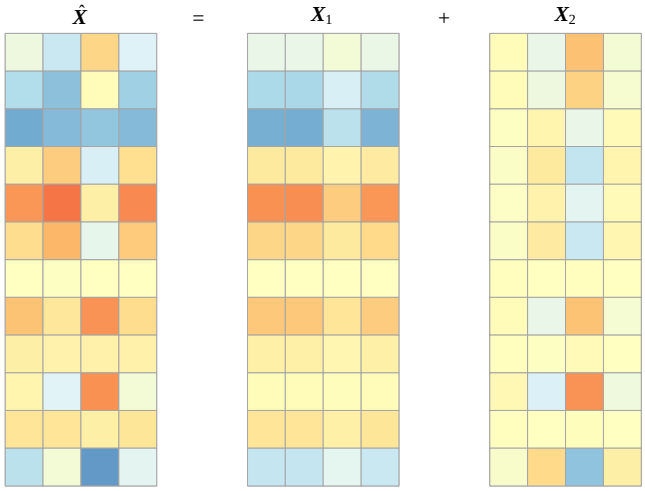


图 43. 两个热图叠加还原原始数据

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：

$$\boldsymbol{E} = \boldsymbol{X} - \hat{\boldsymbol{X}} \tag{60}$$

图 44 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。从图片还原角度来看，如图 44 所示，PCA 降维用更少维度、更少数据获得几乎一样画质图片。

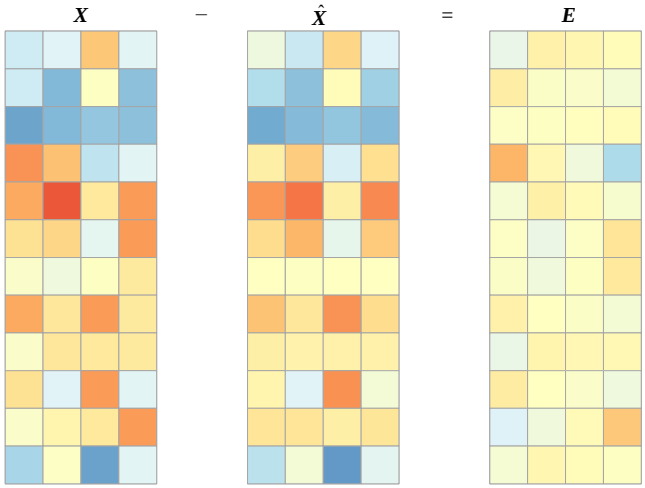


图 44. 原始数据、拟合数据和残差数据热图

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。

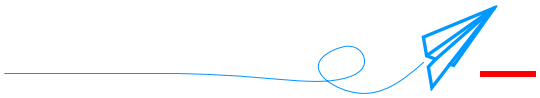
总结来说，通过 PCA 降维，我们可以减少数据的维度，从而简化模型和算法的复杂度，同时可以去除噪声和冗余信息，提高数据的可解释性和可视化效果，从而更好地理解数据和发现数据中的规律。PCA 广泛应用于数据挖掘、模式识别、图像处理、信号处理等领域。

《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	$\boldsymbol{X} = \boldsymbol{U_X S_X V_X}^{\mathrm{T}}$
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 $\boldsymbol{G} = \frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{n-1}$	特征值分解	$\boldsymbol{G} = \boldsymbol{V_X \Lambda_X V_X}^{\mathrm{T}}$
中心化数据矩阵 $\boldsymbol{X}_c = \boldsymbol{X} - \mathrm{E}(\boldsymbol{X})$	奇异值分解	$\boldsymbol{X}_c = \boldsymbol{U_c S_c V_c}^{\mathrm{T}}$
协方差矩阵 $\boldsymbol{\Sigma} = \frac{(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))^{\mathrm{T}}(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))}{n-1}$	特征值分解	$\boldsymbol{\Sigma} = \boldsymbol{V_c \Lambda_c V_c}^{\mathrm{T}}$
标准化数据 (z 分数) $\boldsymbol{Z}_X = (\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))\boldsymbol{D}^{-1}$ $\boldsymbol{D} = \mathrm{diag}(\mathrm{diag}(\boldsymbol{\Sigma}))^{\frac{1}{2}}$	奇异值分解	$\boldsymbol{Z_X} = \boldsymbol{U_Z S_Z V_Z}^{\mathrm{T}}$

相关性系数矩阵 $\boldsymbol{P}=\boldsymbol{D}^{-1}\boldsymbol{\Sigma}\boldsymbol{D}^{-1}$ $\boldsymbol{D}=\mathrm{diag}\left(\mathrm{diag}\left(\boldsymbol{\Sigma}\right)\right)^{\frac{1}{2}}$	特征值分解	$\boldsymbol{P}=\boldsymbol{V}_Z\boldsymbol{\Lambda}_Z\boldsymbol{V}_Z^{\mathrm{T}}$



人类思维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”。大脑涉及可能性判断时，就不自觉进入“贝叶斯推断”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。根据以往经验，估计这次考试通过率 80%以上。这种“先验 + 数据 → 后验”的思维模式比比皆是。

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”。

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。

天意从来高难问，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。

目不转睛地盯着社会财富分布曲线的“右尾”，对巨贾兜售的“成功学”布道言听计从，从统计角度来看都是痴人说梦。

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“鸢尾花书”的时候，就依靠统计思维做出了“优化”选择。

《统计至简》是“鸢尾花书”数学板块的三本中的最后一本，其中大家看到了代数、几何、线性代数、概率统计、优化等数学板块的合流。

读到这，大家便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，都能在脑中“看见”各种几何图形。

下面，我们一起踏上《数据有道》、《机器学习》的“实践”之旅！