

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



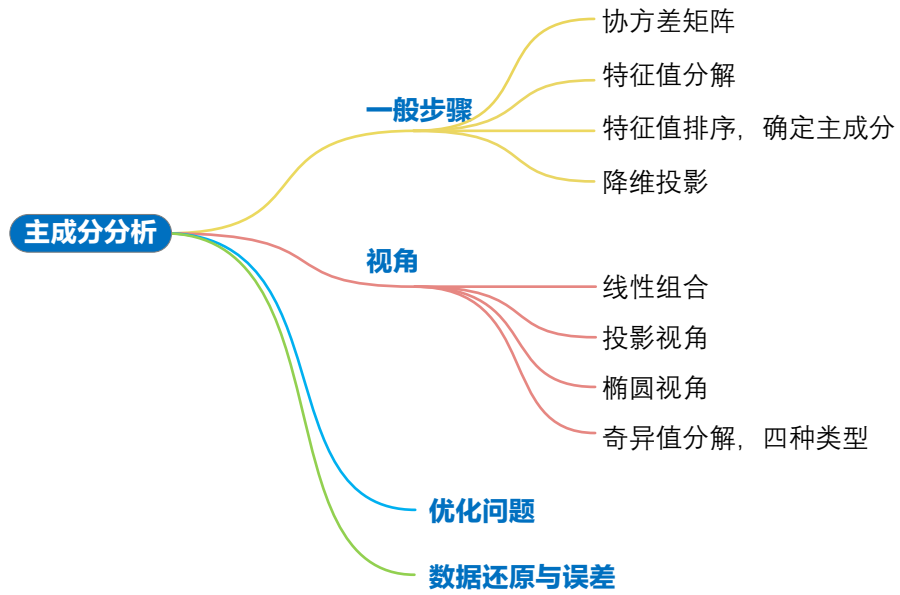
掌握我们的命运的不是星象，而是我们自己。

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



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 $X_{n \times D}$ 的协方差矩阵 $\Sigma_{D \times D}$ ；
- ▶ 对 Σ 特征值分解，获得特征值 λ_i 与特征向量矩阵 $V_{D \times D}$ ；
- ▶ 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 $Z_{n \times p}$ 。

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

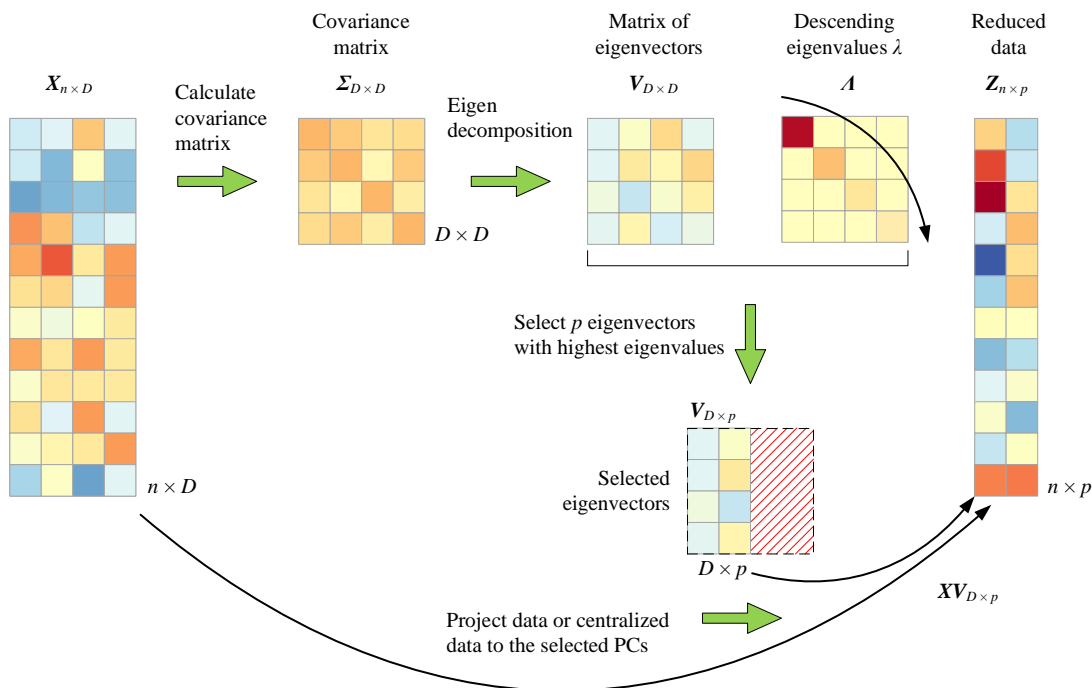


图 1. 主成分分析一般技术路线: 特征值分解协方差矩阵

➡ 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 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 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:

$$X = [x_1 \quad x_2 \quad x_3 \quad x_4] \quad (1)$$

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

$$\mathbf{x}^{(1)} = [x_{1,1} \quad x_{1,2} \quad x_{1,3} \quad x_{1,4}] \quad (2)$$

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

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

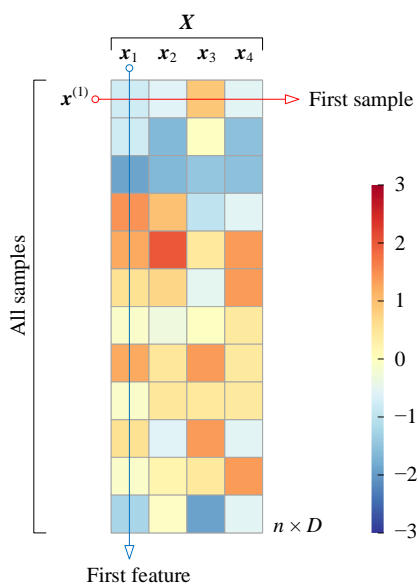
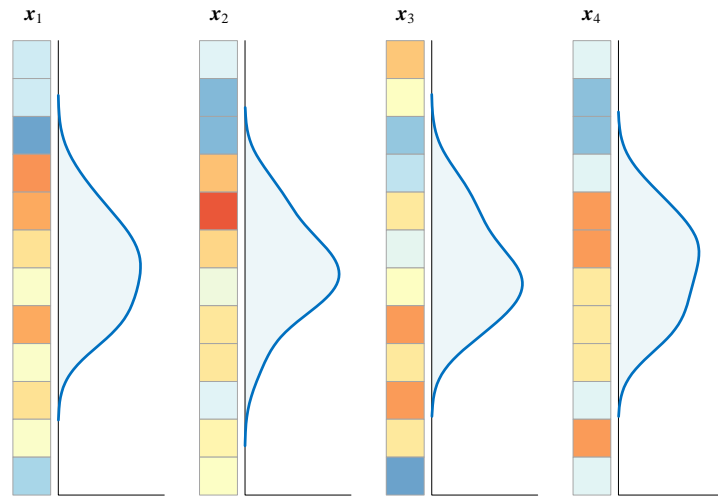


图 2. 原始数据 \mathbf{X} 热图， $D=4$ ， $n=12$ ， \mathbf{X} 已经去均值

分布特征

图 3 所示为矩阵 \mathbf{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. 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， X 的协方差矩阵 Σ 可以通过下式计算得到：

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

其中， $E(X)$ 也常被称作原始数据 X 的质心； $X - E(X)$ 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 $E(X) = \mathbf{0}^T$ ，即 $X = X_c$ 。

如图 5 所示， Σ 为实数对称矩阵，它的特征值分解 (谱分解) 可以写作：

$$\Sigma = V \Lambda V^T \quad (4)$$

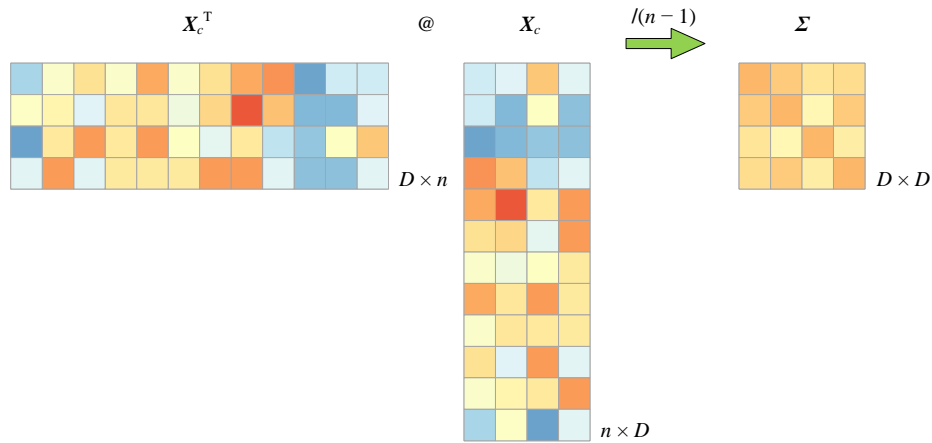
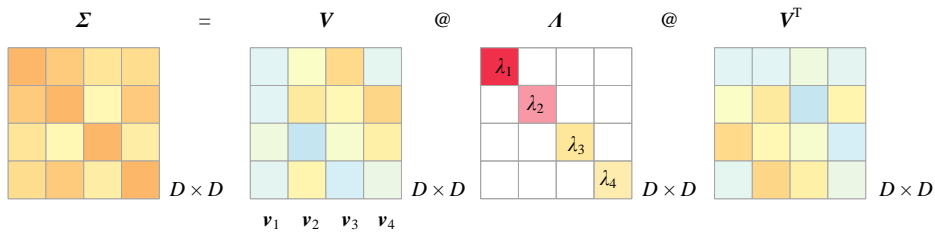
V 为正交矩阵。 V 和自己转置 V^T 乘积为单位阵 I ，即：

$$V^T V = I \quad (5)$$

特征值方阵 Λ 主对角线元素为特征值 λ ，特征值从大到小排列：

$$\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_d \end{bmatrix}, \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \quad (6)$$

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

V 为特征向量构造的 $D \times D$ 的方阵:

$$V = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_D \\ \text{PC1} & \text{PC2} & & \end{bmatrix} = \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,D} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ v_{D,1} & v_{D,2} & \cdots & v_{D,D} \end{bmatrix} \quad (7)$$

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

V 的列向量也叫载荷 (loadings)。注意, 有些文献中载荷定义为:

$$V\sqrt{A} = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_D] \begin{bmatrix} \sqrt{\lambda_1} & & & \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_D} \end{bmatrix} = [\sqrt{\lambda_1}\mathbf{v}_1 \quad \sqrt{\lambda_2}\mathbf{v}_2 \quad \cdots \quad \sqrt{\lambda_D}\mathbf{v}_D] \quad (8)$$

迹, 总方差

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

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

第 j 个特征值 λ_j 对**方差总和** (total variance) 的贡献百分比为：

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

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：

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

"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。

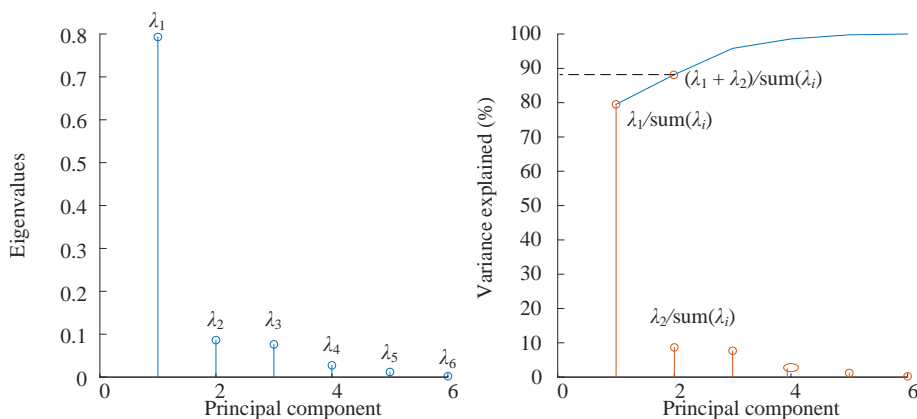


图 6. PCA 分析主元方差和陡坡图

25.4 投影

本节从投影角度介绍 PCA。数据矩阵 X 投影到矩阵 V 正交系 (v_1, v_2, \dots, v_D) 得到新特征数据矩阵 Z ，即：

$$\mathbf{Z} = \mathbf{X}\mathbf{V}$$

(12)

\mathbf{V} 常被称作**载荷** (loadings), \mathbf{Z} 常被称作**得分** (scores)。图 7 所示 $\mathbf{Z} = \mathbf{X}\mathbf{V}$ 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。

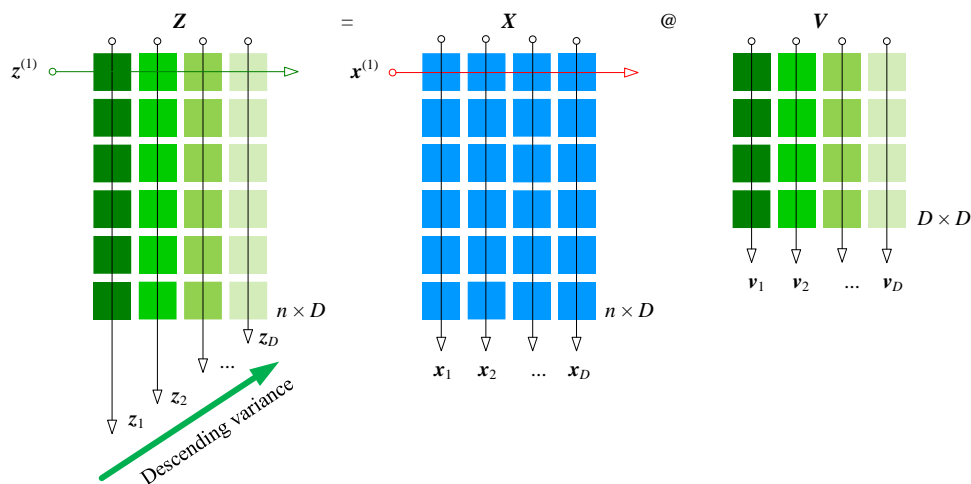


图 7. PCA 分解数据关系 $\mathbf{Z} = \mathbf{X}\mathbf{V}$

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

▲ 值得强调的一点是，把原始数据 \mathbf{X} 或中心化数据 \mathbf{X}_c 投影到 \mathbf{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。

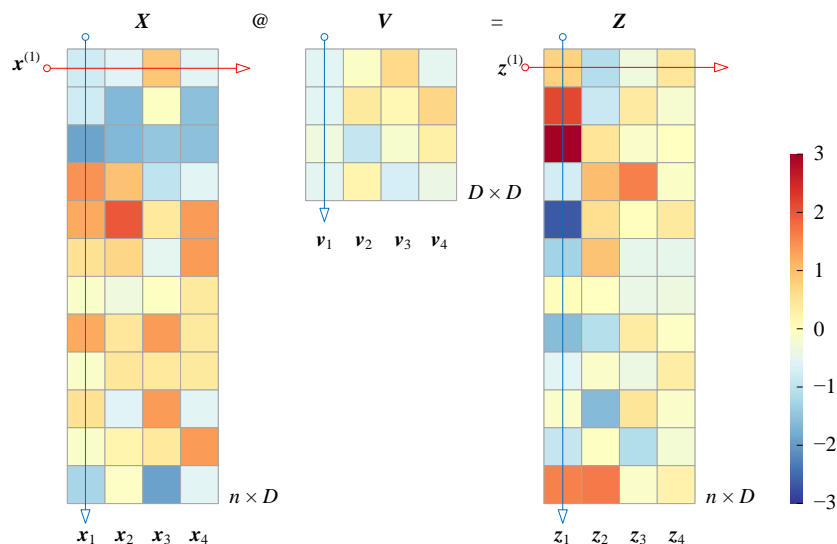


图 8. \mathbf{Z} 、 \mathbf{X} 和 \mathbf{V} 这三个矩阵关系和热图

\mathbf{Z} 的列向量

前文讨论过，矩阵 \mathbf{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 \mathbf{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。

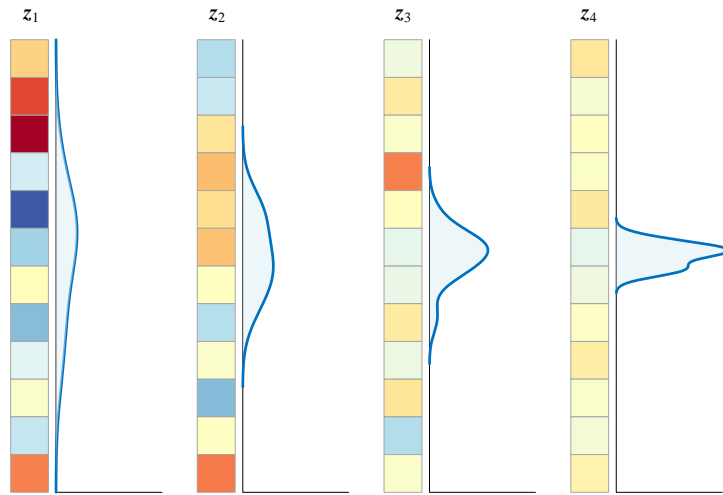


图 9. \mathbf{Z} 四个新特征数据分布

将 (12) 展开得到：

$$\begin{bmatrix} z_1 & z_2 & \cdots & z_D \end{bmatrix} = \mathbf{X} \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_D \\ \text{PC1} & \text{PC2} & & \end{bmatrix} \quad (13)$$

由此，得到图 10 所示主成分分析运算的数据关系：

$$\begin{cases} z_1 = \mathbf{X}\mathbf{v}_1 \\ z_2 = \mathbf{X}\mathbf{v}_2 \\ \vdots \\ z_D = \mathbf{X}\mathbf{v}_D \end{cases} \quad (14)$$

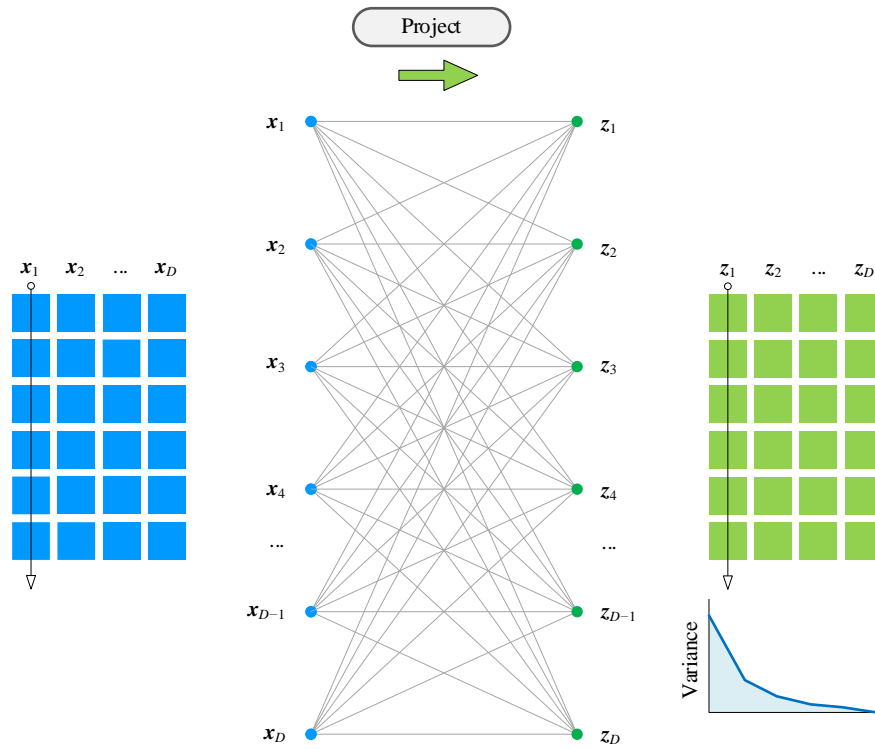


图 10. PCA 中数据关系

线性组合

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

$$\mathbf{z}_1 = \mathbf{X}\mathbf{v}_1 \quad (15)$$

(15) 展开得到：

$$\mathbf{z}_1 = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_D \end{bmatrix} \begin{bmatrix} v_{1,1} \\ v_{2,1} \\ \vdots \\ v_{D,1} \end{bmatrix} = v_{1,1}\mathbf{x}_1 + v_{2,1}\mathbf{x}_2 + \cdots + v_{D,1}\mathbf{x}_D \quad (16)$$

$\mathbf{v}_1, \text{PC1}$

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

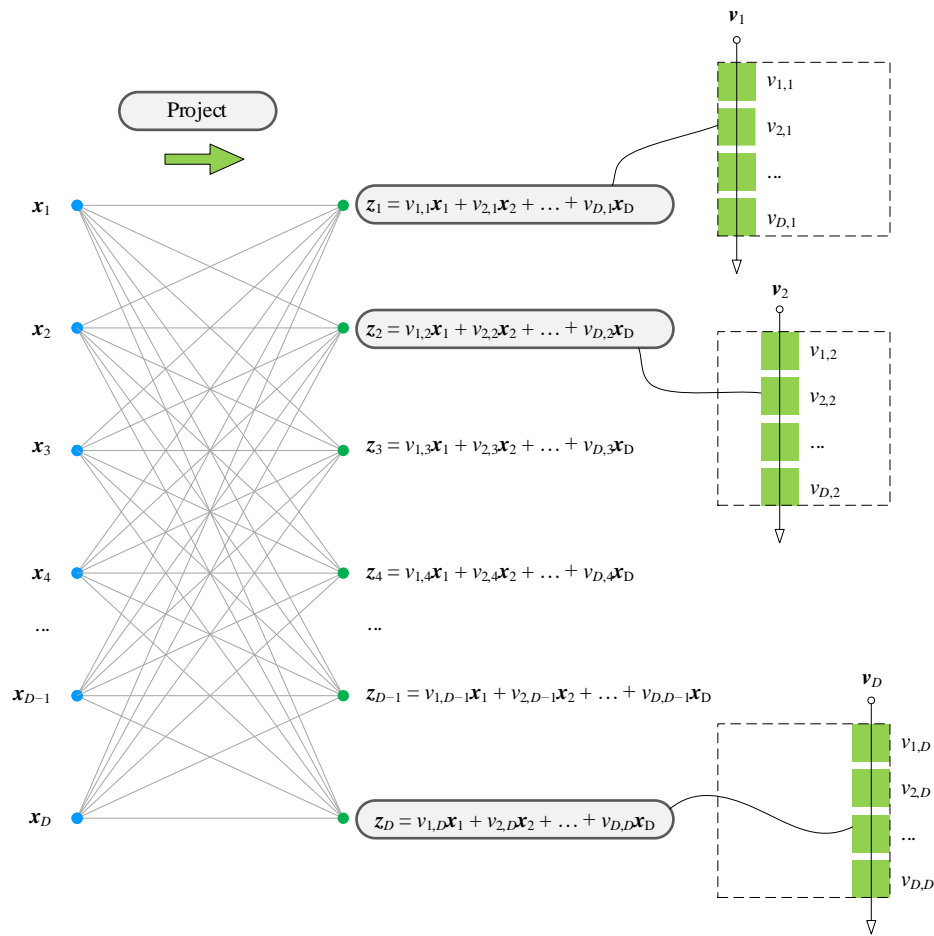


图 11. 线性组合角度看 PCA

朝向量投影

图 12 ~ 图 15 分别展示数据矩阵 X 向 v_1 、 v_2 、 v_3 和 v_4 向量投影。

图 12 所示 $z_1 = Xv_1$ 运算相当于数据 X 向 v_1 向量 (第一主成分) 投影获得 z_1 。图 13 展示 $z_2 = Xv_2$ 运算等价于数据 X 向 v_2 (第二主成分) 投影获得 z_2 。以此类推。

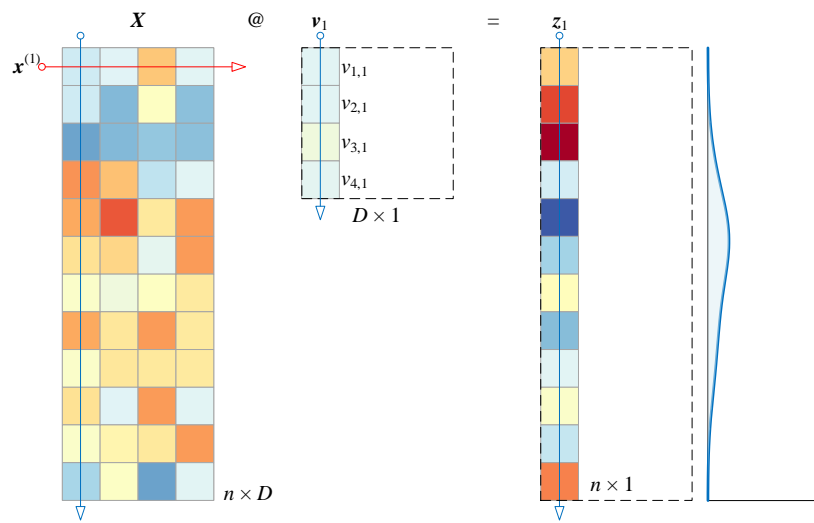


图 12. 数据 X 向 v_1 向量投影

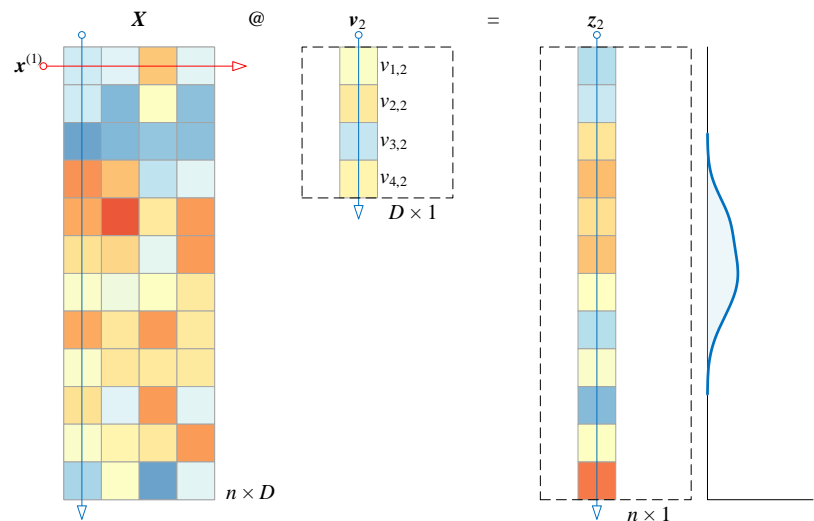


图 13. 数据 X 向 v_2 向量投影

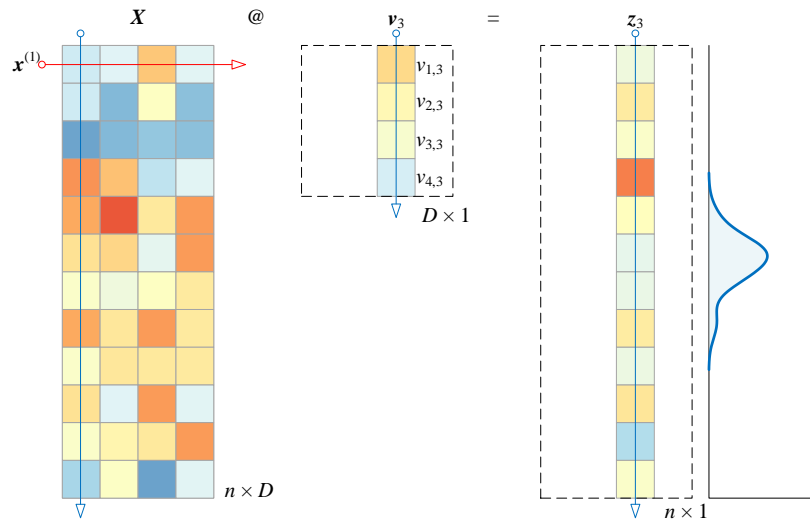


图 14. 数据 X 向 v_3 向量投影

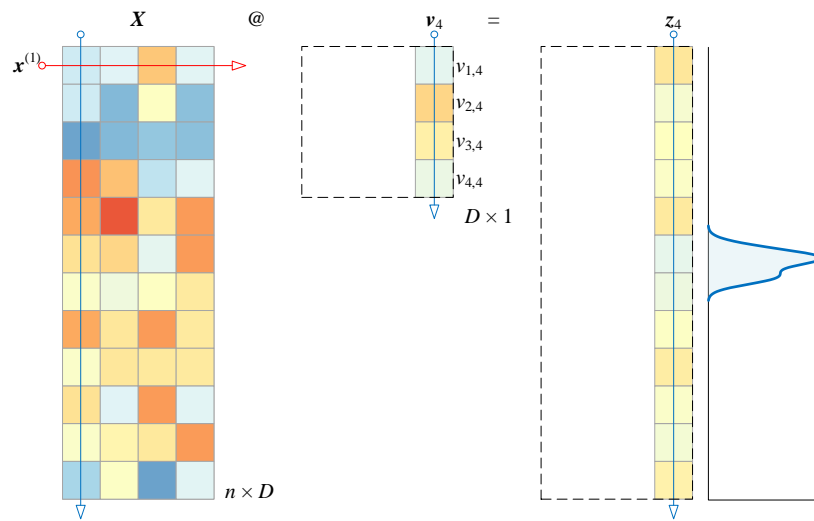


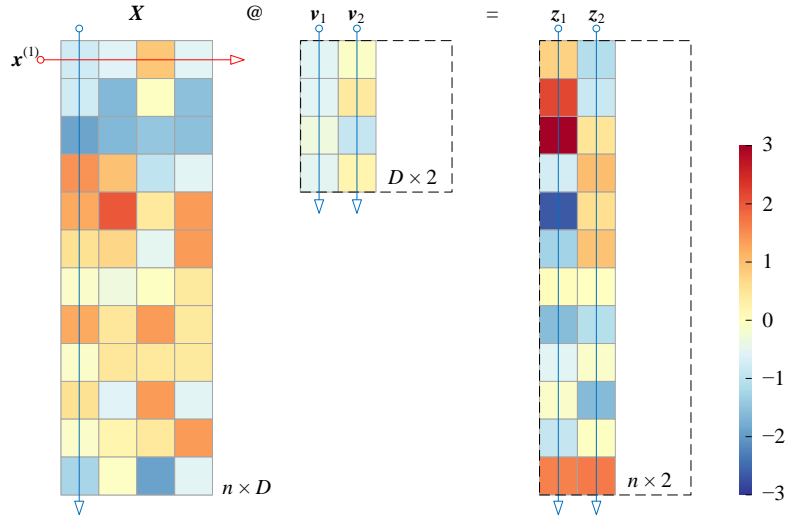
图 15. 数据 X 向 v_4 向量投影

朝平面投影

同样， $[z_1, z_2]$ 是 X 向 $[v_1, v_2]$ 投影结果，即四维数据 X 向二维空间投影。运算过程如下：

$$\begin{bmatrix} z_1 & z_2 \end{bmatrix} = X \begin{bmatrix} v_1 & v_2 \end{bmatrix} \quad (17)$$

图 16 所示为 (17) 运算过程及结果热图。

图 16. 数据 X 向 $[v_1, v_2]$ 投影

Z 的协方差矩阵

前文假设 X 已经中心化，因此 z_1 的期望值为 0。对 z_1 求方差，可以得到：

$$\text{var}(z_1) = \frac{(Xv_1)^T (Xv_1)}{n-1} = \frac{v_1^T X^T X v_1}{n-1} = v_1^T \frac{X^T X}{n-1} v_1 = v_1^T \Sigma v_1 \quad (18)$$

类似地，

$$\text{var}(z_2) = v_2^T \Sigma v_2, \quad \dots, \quad \text{var}(z_D) = v_D^T \Sigma v_D \quad (19)$$

这样， Z 的协方差矩阵可以通过下式计算得到：

$$\begin{aligned} \text{var}(Z) &= \frac{(XV)^T (XV)}{n-1} = \frac{V^T X^T X V}{n-1} \\ &= V^T \frac{X^T X}{n-1} V = V^T \Sigma V = \begin{bmatrix} v_1^T \Sigma v_1 & & \\ & v_2^T \Sigma v_2 & \\ & & \ddots \\ & & & v_D^T \Sigma v_D \end{bmatrix} = \Lambda = \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \\ & & & \lambda_D \end{bmatrix} \end{aligned} \quad (20)$$

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 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。

$Z_{n \times p}$ 的协方差矩阵为：

$$\text{var}(Z_{n \times p}) = \frac{(XV_{D \times p})^T (XV_{D \times p})}{n-1} = V_{D \times p}^T \frac{X^T X}{n-1} V_{D \times p} = V_{D \times p}^T \Sigma V_{D \times p} = \Lambda_{p \times p} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix} \quad (21)$$

➡ 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 μ ，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。

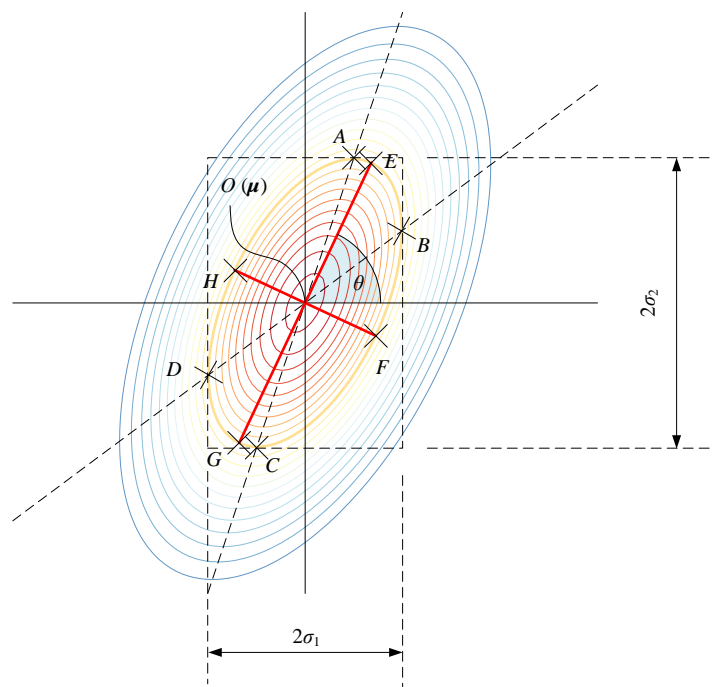


图 17. 主成分分析和椭圆的关系

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 Σ ，找到 Σ 对应椭圆的半长轴所在方向 v_1 。 v_1 对应的便是第一主成分 PC1。原始数据朝 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。

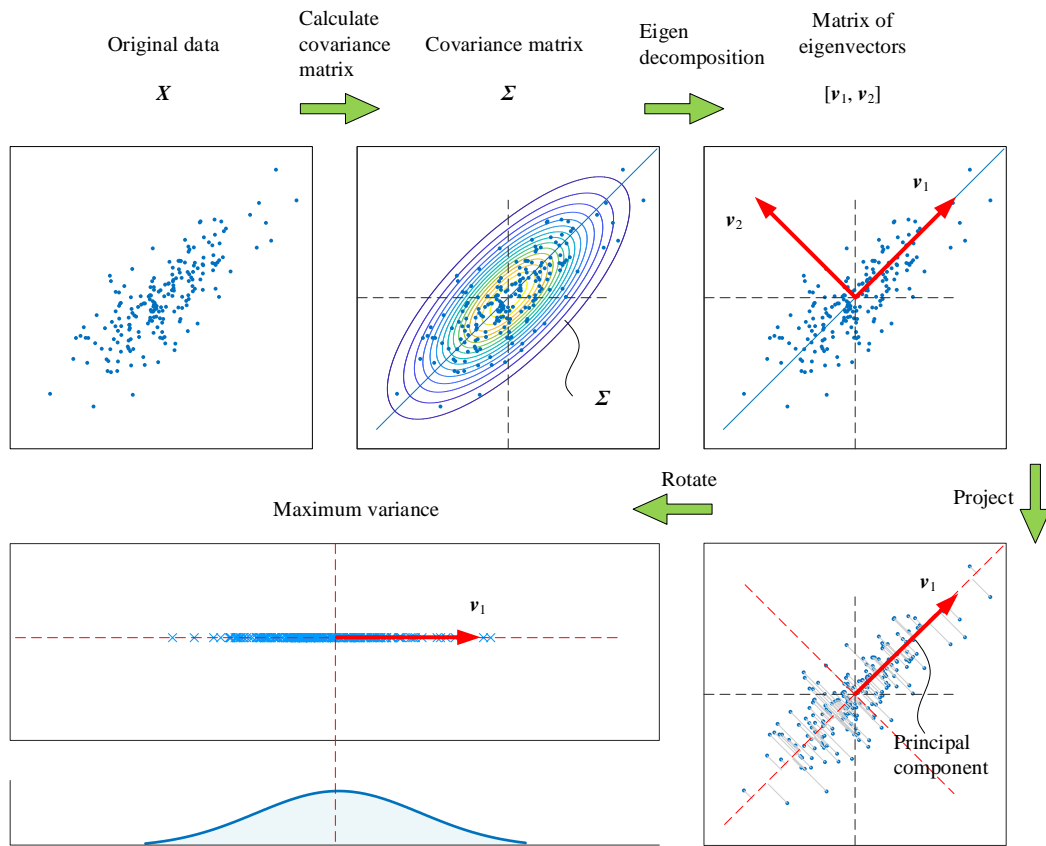


图 18. 几何视角下通过特征值分解协方差矩阵进行主成分分析

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。

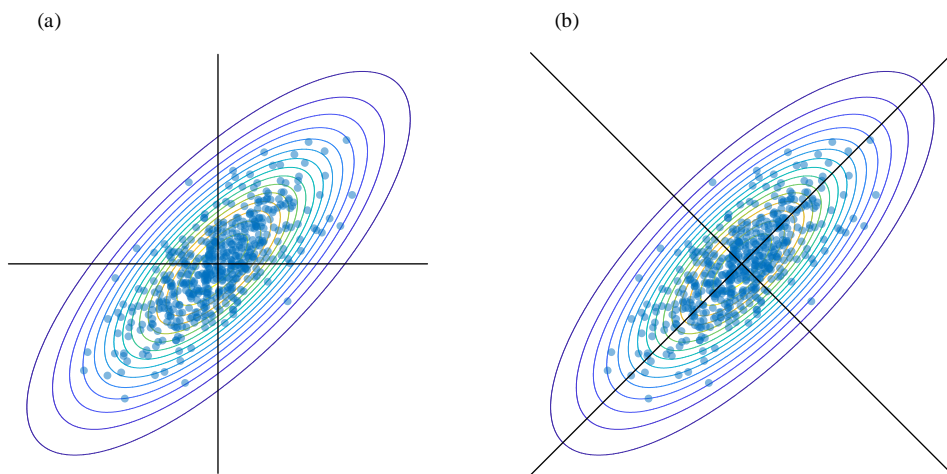


图 19. 两个角度看数据

举个例子

图 20 (a) 所示为原始二维数据 \mathbf{X} 的散点图，可以发现数据的质心位于 $[1, 2]^T$ 。分析数据 \mathbf{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。

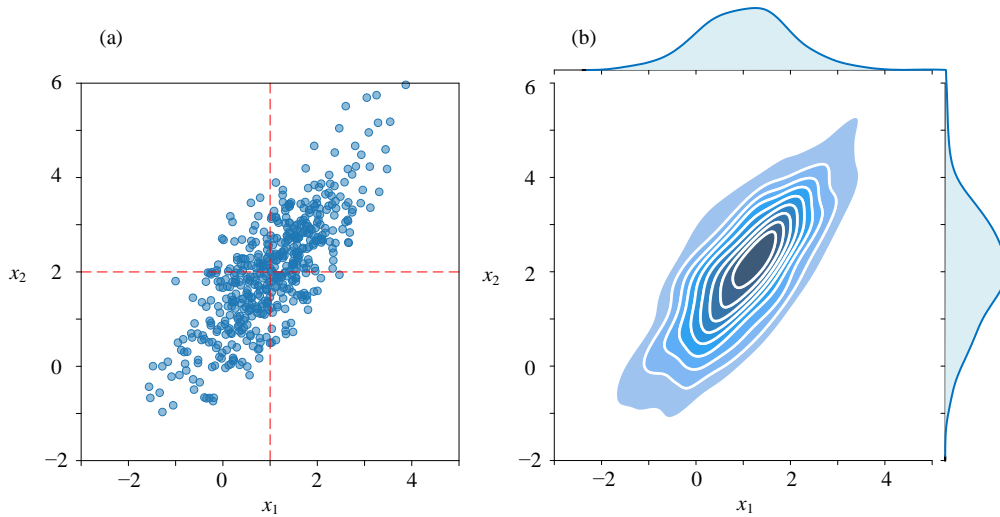


图 20. 原始二维数据 \mathbf{X}

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(X)` 可以获得投影后的数据 \mathbf{Y} 。图 21 对比 \mathbf{Y} 两列数据分布。图 22 所示为数据 \mathbf{Y} 在 $[\mathbf{v}_1, \mathbf{v}_2]$ 中散点图。

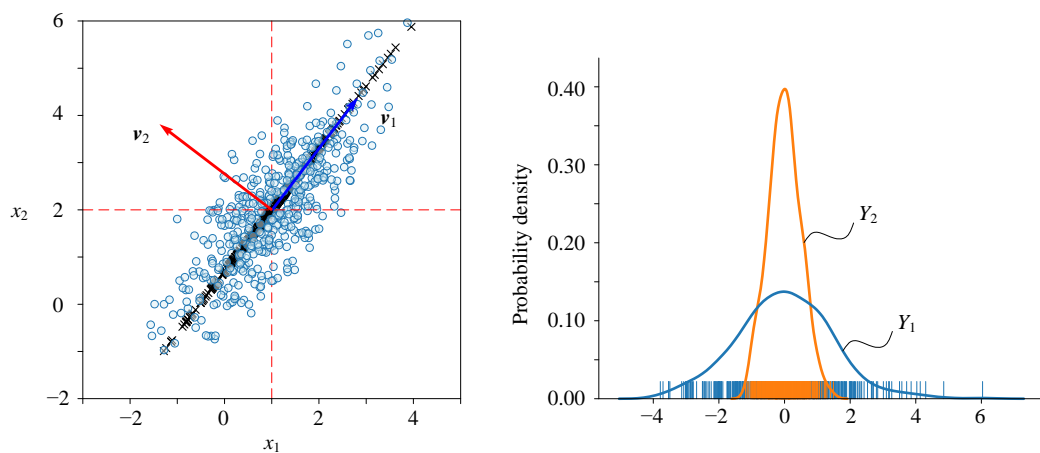
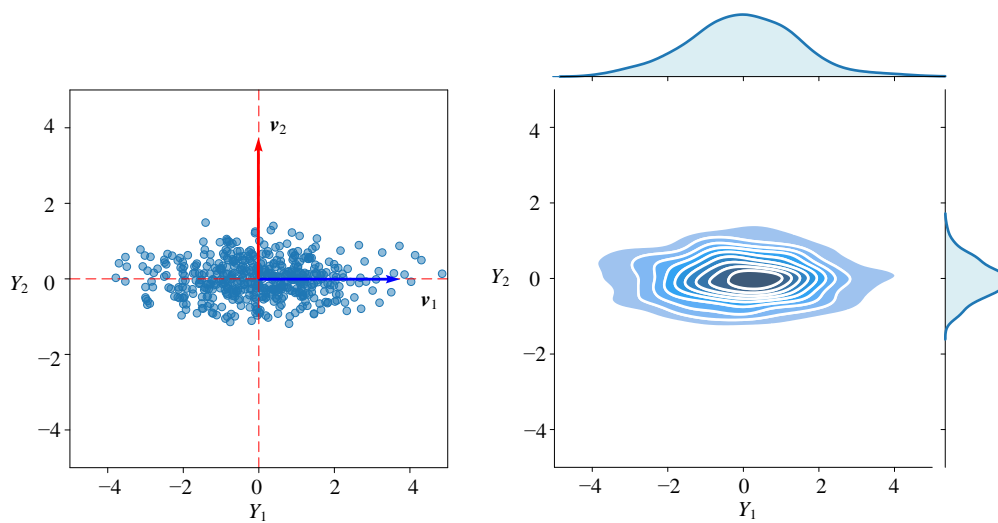


图 21. 主成分数据分布

图 22. 数据 Y 在 $[v_1, v_2]$ 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， U 为方阵， S 矩阵并非方阵。

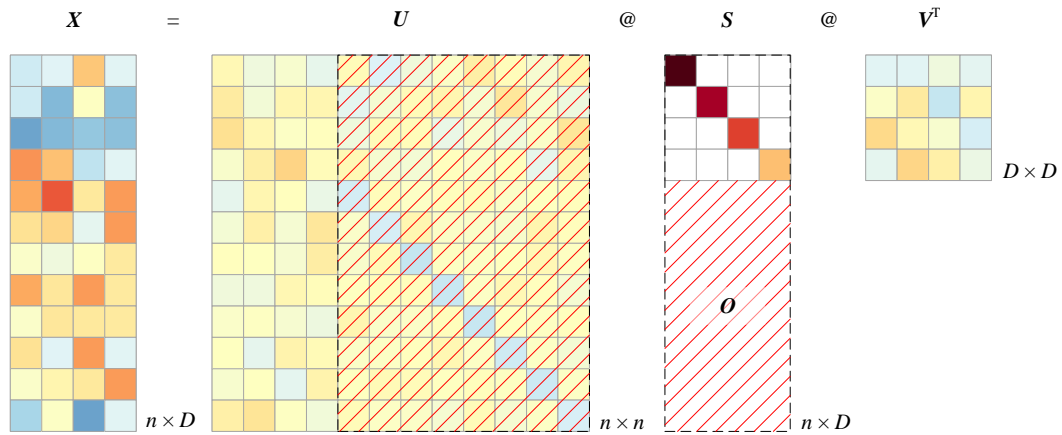


图 23. 完全 (full) 奇异值分解

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

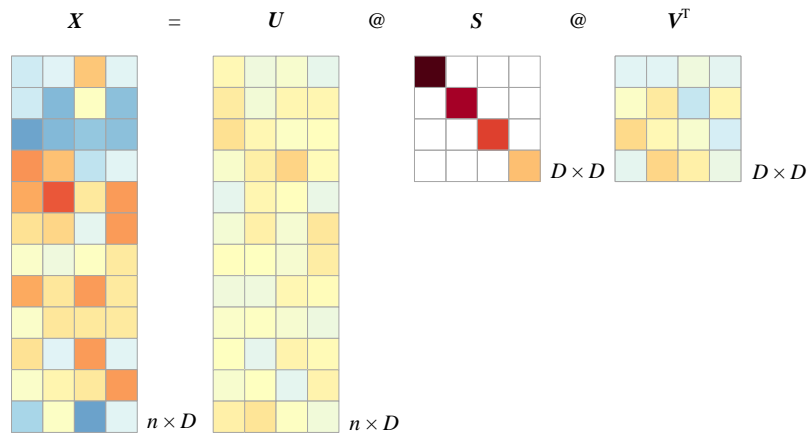


图 24. 经济型奇异值分解

当 X 非满秩时，即 $\text{rank}(X) = r < D$ ，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。

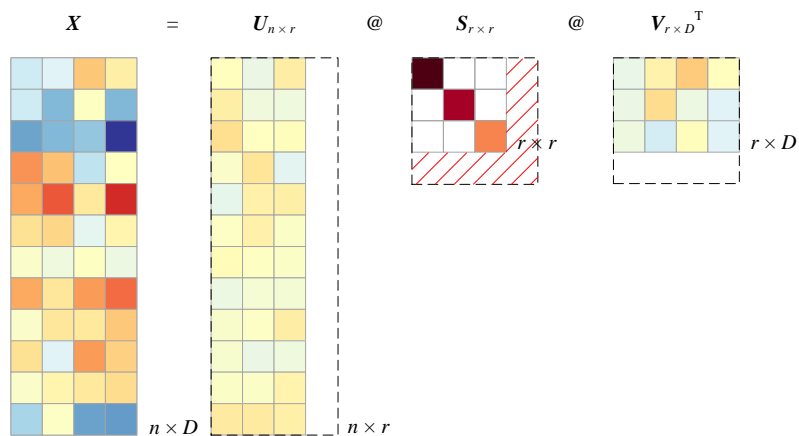


图 25. 紧凑型奇异值分解， X 非满秩

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， $S_{p \times p}$ 仅使用图 24 中 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“ \approx ”；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。

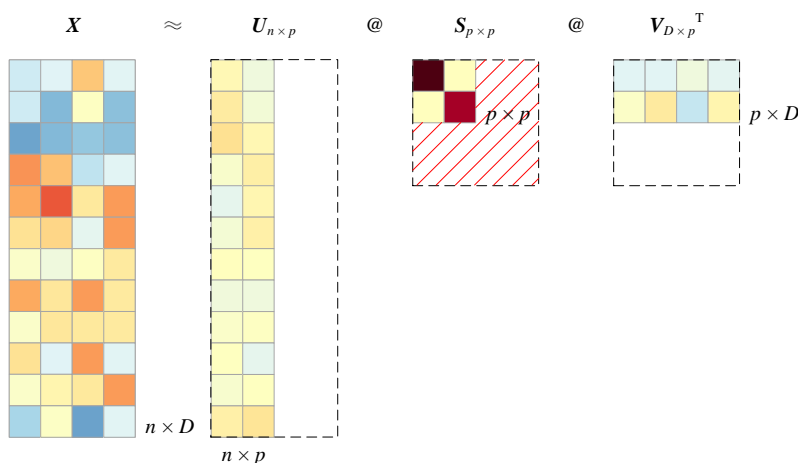


图 26. 截断型奇异值分解

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 $X_{n \times D}$ 进行完全型 SVD 分解，得到：

$$X = USV^T \quad (22)$$

V 和 U 均为正交矩阵，即满足：

$$\begin{aligned} UU^T &= U^T U = I \\ VV^T &= V^T V = I \end{aligned} \quad (23)$$

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：

$$\Sigma = \frac{X^T X}{n-1} \quad (24)$$

将(22)代入(24)得到：

$$\Sigma = \frac{(USV^T)^T USV^T}{n-1} = \frac{VS^T SV^T}{n-1} \quad (25)$$

对协方差矩阵进行特征值分解：

$$\Sigma = \Lambda V^T \quad (26)$$

联立 (25) 和 (26)，

$$\frac{VS^T SV^T}{n-1} = \Lambda V^T \quad (27)$$

对于经济型 SVD 分解， S 为对角方阵，(27) 整理得到：

$$\frac{S^2}{n-1} = \Lambda \quad (28)$$

即

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

注意， $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D$ 。

奇异值和特征值存在如下关系：

$$\frac{s_j^2}{n-1} = \lambda_j \quad (30)$$

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 U

Z 可以还原 X ：

$$X = ZV^{-1} = ZV^T \quad (31)$$

对比 (22) 和 $X = USV^T$ ，可以发现：

$$Z = US \quad (32)$$

也就是

$$\begin{bmatrix} z_1 & z_2 & \cdots & z_D \end{bmatrix} = \begin{bmatrix} u_1 & u_2 & \cdots & u_D \end{bmatrix} \begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_D \end{bmatrix} = \begin{bmatrix} s_1 u_1 & s_2 u_2 & \cdots & s_D u_D \end{bmatrix} \quad (33)$$

即：

$$s_1 \mathbf{u}_1 = \mathbf{z}_1, \quad s_2 \mathbf{u}_2 = \mathbf{z}_2, \quad \dots \quad (34)$$

对 \mathbf{z}_1 求方差：

$$\text{var}(\mathbf{z}_1) = \frac{\mathbf{z}_1^T \mathbf{z}_1}{n-1} = \frac{(s_1 \mathbf{u}_1)^T (s_1 \mathbf{u}_1)}{n-1} = \frac{s_1^2 \|\mathbf{u}_1\|^2}{n-1} = \frac{s_1^2}{n-1} = \lambda_1 \quad (35)$$

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

$$\mathbf{U} = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \dots \quad \mathbf{u}_D] = \begin{bmatrix} \frac{\mathbf{z}_1}{s_1} & \frac{\mathbf{z}_2}{s_2} & \dots & \frac{\mathbf{z}_D}{s_D} \end{bmatrix} \quad (36)$$

也就是：

$$\mathbf{U} = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \dots \quad \mathbf{u}_D] = \mathbf{Z} \mathbf{S}^{-1} \quad (37)$$

至此，我们理解了 SVD 分解中矩阵 \mathbf{U} 的内涵。

张量积

用张量积来展开 SVD 分解：

$$\begin{aligned} \mathbf{X} &= \mathbf{U} \mathbf{S} \mathbf{V}^T \\ &= [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \dots \quad \mathbf{u}_D] \begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_D \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_D^T \end{bmatrix} \\ &= s_1 \mathbf{u}_1 \mathbf{v}_1^T + s_2 \mathbf{u}_2 \mathbf{v}_2^T + \dots + s_D \mathbf{u}_D \mathbf{v}_D^T \\ &= s_1 \mathbf{u}_1 \otimes \mathbf{v}_1 + s_2 \mathbf{u}_2 \otimes \mathbf{v}_2 + \dots + s_D \mathbf{u}_D \otimes \mathbf{v}_D \end{aligned} \quad (38)$$

图 27 所示为 (38) 还原原始数据的过程。

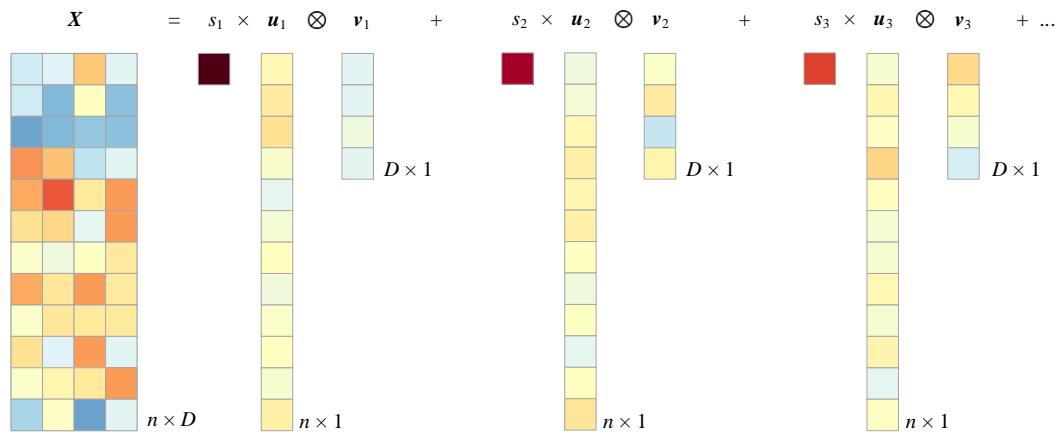


图 27. 张量积 $s_1 \mathbf{u}_1 \otimes \mathbf{v}_1$ 、 $s_2 \mathbf{u}_2 \otimes \mathbf{v}_2$ 等之和还原数据 \mathbf{X}

25.7 优化问题

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

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

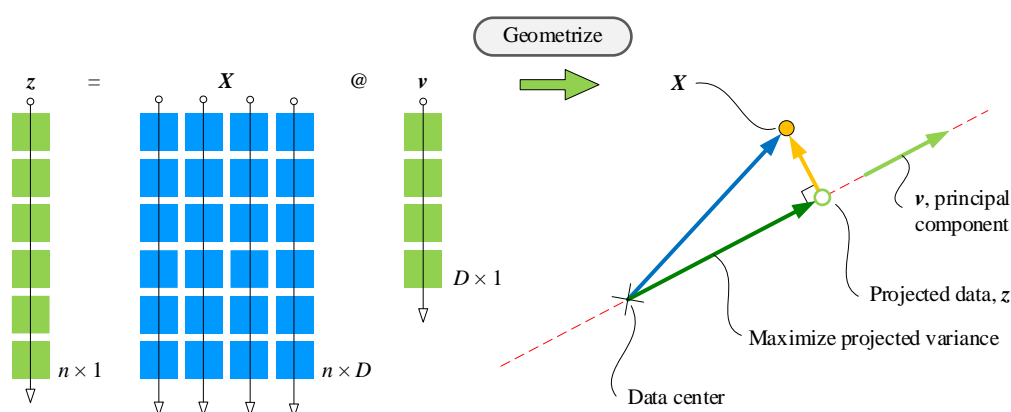


图 28. 主成分分析优化问题

由于 X 为中心化数据，因此 z 的均值也为 0；因此， z 方差为：

$$\text{var}(z) = \frac{\mathbf{z}^T \mathbf{z}}{n-1} = \mathbf{v}^T \frac{\text{Covariance matrix } \mathbf{X}^T \mathbf{X}}{n-1} \mathbf{v} \quad (39)$$

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

$$\text{var}(z) = \mathbf{v}^T \Sigma \mathbf{v} \quad (40)$$

\mathbf{v} 为单位列向量，即满足如下约束条件：

$$\mathbf{v}^T \mathbf{v} = 1 \quad (41)$$

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 \mathbf{v} 方向上数据投影方差最大化：

$$\begin{aligned} \arg \max_{\mathbf{v}} \quad & \mathbf{v}^T \Sigma \mathbf{v} \\ \text{subject to:} \quad & \mathbf{v}^T \mathbf{v} - 1 = 0 \end{aligned} \quad (42)$$

上式最大化优化问题等价于如下最小化优化问题：

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

构造拉格朗日函数 $L(\mathbf{v}, \lambda)$:

$$L(\mathbf{v}, \lambda) = -\mathbf{v}^T \Sigma \mathbf{v} + \lambda (\mathbf{v}^T \mathbf{v} - 1) \quad (44)$$

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

$$\nabla_{\mathbf{v}} L(\mathbf{v}, \lambda) = \frac{\partial L(\mathbf{v}, \lambda)}{\partial \mathbf{v}} = (-2\Sigma \mathbf{v} + 2\lambda \mathbf{v})^T = \mathbf{0} \quad (45)$$



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:

$$\Sigma \mathbf{v} = \lambda \mathbf{v} \quad (46)$$

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

$$\text{var}(z) = \mathbf{v}^T \Sigma \mathbf{v} = \mathbf{v}^T \lambda \mathbf{v} = \lambda \mathbf{v}^T \mathbf{v} = \lambda \quad (47)$$

也就是说， $\text{var}(z)$ 最大值对应 Σ 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， \mathbf{Z} 反向可以通过 $\mathbf{X} = \mathbf{Z}\mathbf{V}^T$ 还原 \mathbf{X} 。图 29 所示为还原得到 \mathbf{X} 过程。图 30 所示热图，矩阵 \mathbf{Z} 还原转化为原始数据矩阵 \mathbf{X} 。

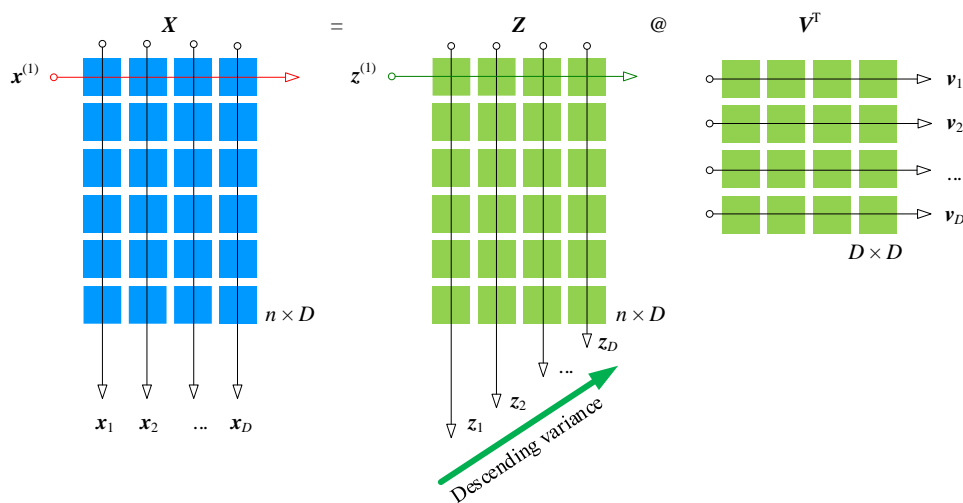


图 29. 反向还原数据 $\mathbf{X} = \mathbf{Z}\mathbf{V}^T$

再次强调，图 29 这种还原计算成立的条件是 X 的质心位于原点。

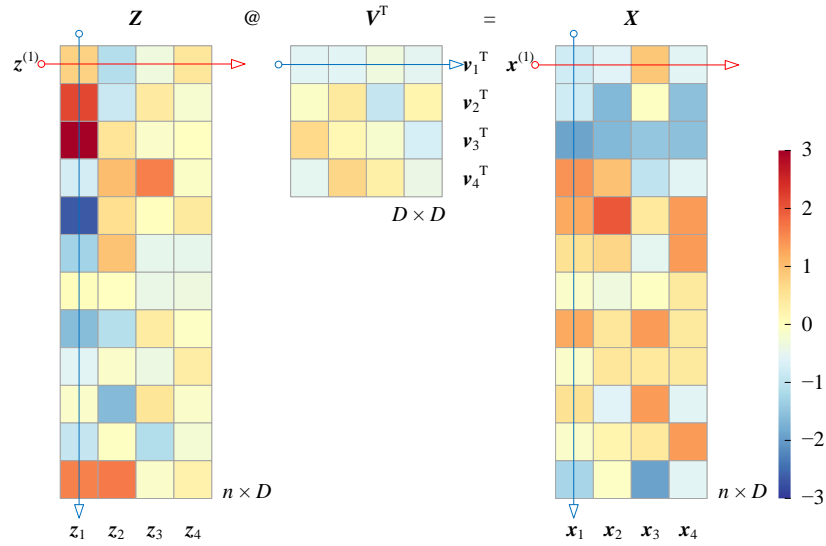


图 30. 新特征数据矩阵 Z 还原转化为原始数据矩阵 X

$X = ZV^T$ 展开得到下式：

$$X = \begin{bmatrix} z_1 & z_2 & z_3 & z_4 \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ v_3^T \\ v_4^T \end{bmatrix} = \begin{matrix} z_1 v_1^T & z_2 v_2^T & z_3 v_3^T & z_4 v_4^T \\ \hat{x}_1 & \hat{x}_2 & \hat{x}_3 & \hat{x}_4 \end{matrix} \quad (48)$$

(48) 所示运算过程如图 31 所示。

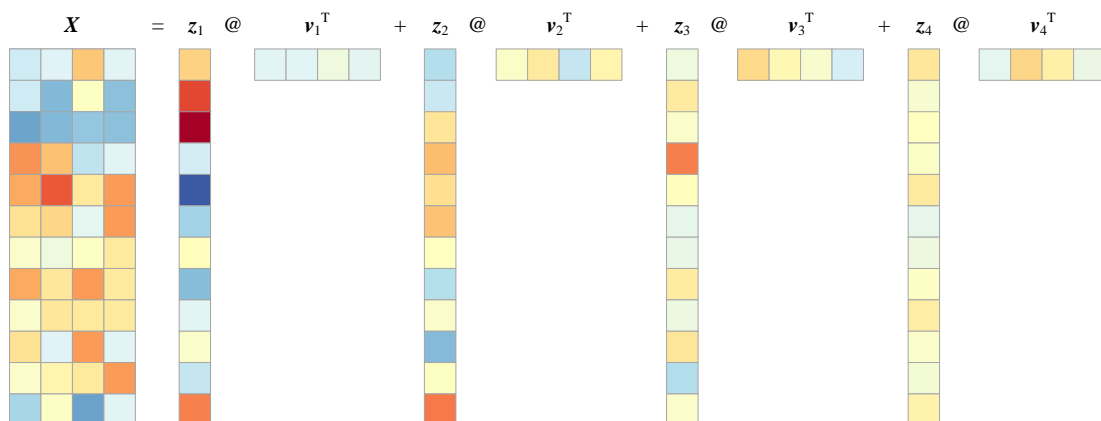


图 31. 还原原始数据运算

图 32 所示为 z_1 还原 X 部分数据，对应运算如下：

$$\mathbf{X}_1 = \mathbf{z}_1 \mathbf{v}_1^T \quad (49)$$

展开上式得到：

$$\begin{aligned} \mathbf{X}_1 &= \mathbf{z}_1 \mathbf{v}_1^T \\ &= \mathbf{z}_1 \begin{bmatrix} v_{1,1} & v_{2,1} & \cdots & v_{D,1} \end{bmatrix} \\ &= \begin{bmatrix} v_{1,1} \mathbf{z}_1 & v_{2,1} \mathbf{z}_1 & \cdots & v_{D,1} \mathbf{z}_1 \end{bmatrix} \end{aligned} \quad (50)$$

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， \mathbf{X}_1 的每一列均是标量乘以向量 \mathbf{z}_1 的结果。显然， \mathbf{X}_1 的秩为 1，即 $\text{rank}(\mathbf{X}_1) = 1$ 。

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

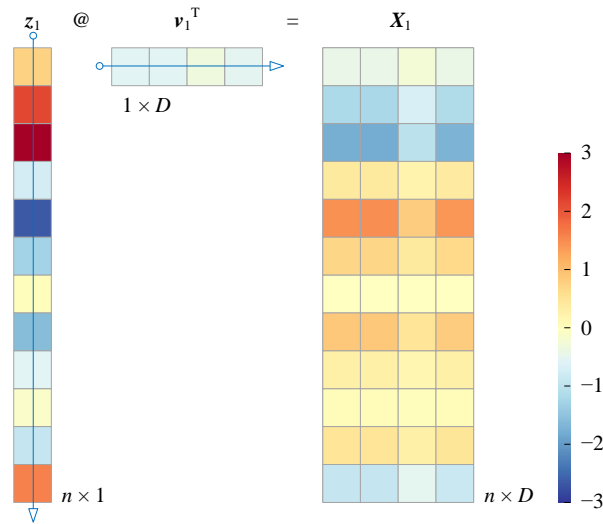


图 32. \mathbf{z}_1 还原 \mathbf{X} 部分数据

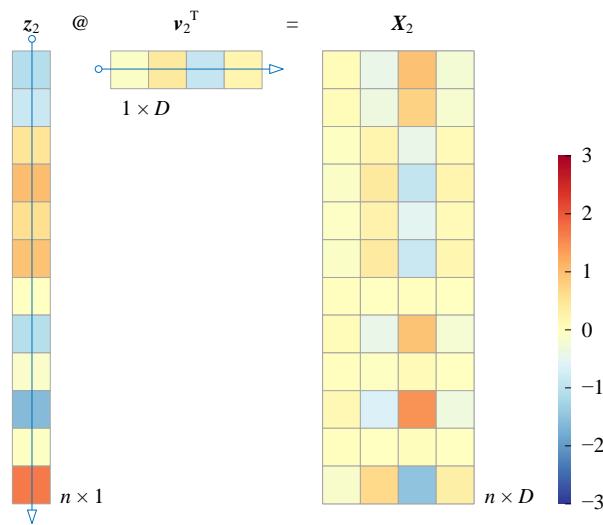


图 33. \mathbf{z}_2 还原 \mathbf{X} 部分数据

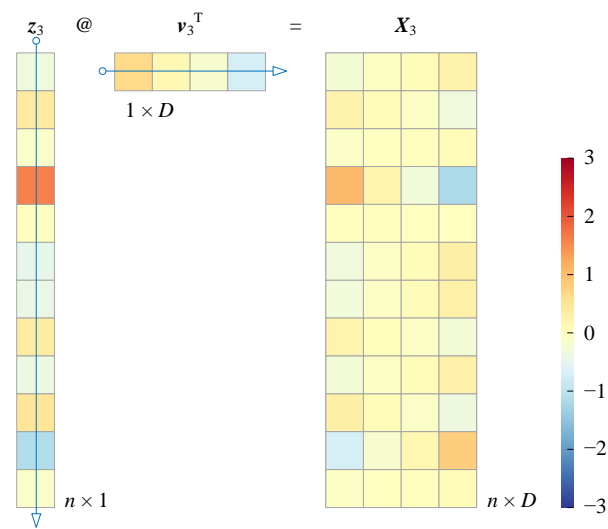


图 34. z_3 还原 X 部分数据

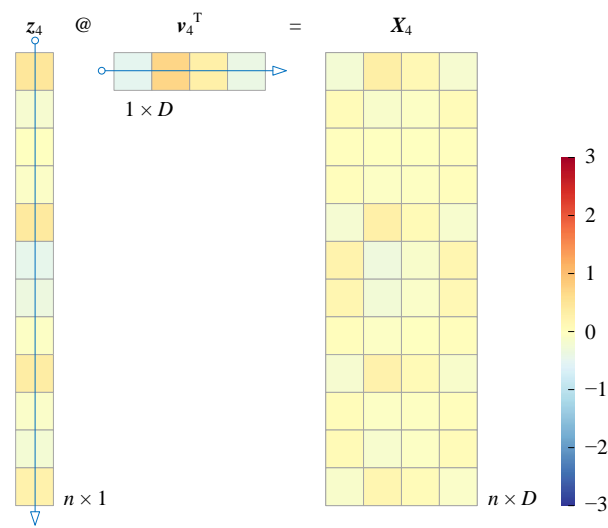


图 35. z_4 还原 X 部分数据

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。

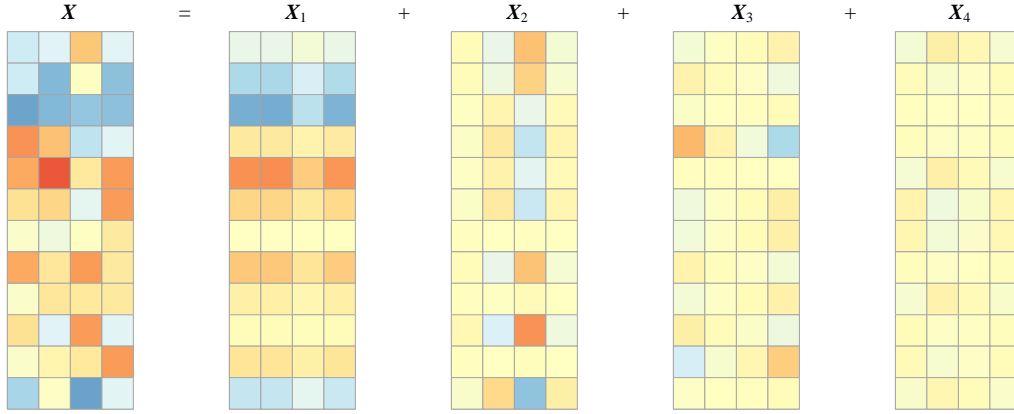


图 36. 原始数据矩阵 X 热图于四层热图叠加结果

张量积

另外，(48) 可以用张量积来表达：

$$X = \underbrace{z_1 \otimes v_1}_{\hat{x}_1} + \underbrace{z_2 \otimes v_2}_{\hat{x}_2} + \underbrace{z_3 \otimes v_3}_{\hat{x}_3} + \underbrace{z_4 \otimes v_4}_{\hat{x}_4} \quad (51)$$

利用 (14)，(48) 可以整理为：

$$X = Xv_1v_1^T + Xv_2v_2^T + \dots + Xv_Dv_D^T = \sum_{j=1}^D Xv_jv_j^T = X \left(\sum_{j=1}^D v_jv_j^T \right) \quad (52)$$

(52) 可以用张量积表达：

$$X = X(v_1 \otimes v_1) + X(v_2 \otimes v_2) + \dots + X(v_D \otimes v_D) = \sum_{j=1}^D Xv_j \otimes v_j = X \left(\sum_{j=1}^D v_j \otimes v_j \right) \quad (53)$$

图 37 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。

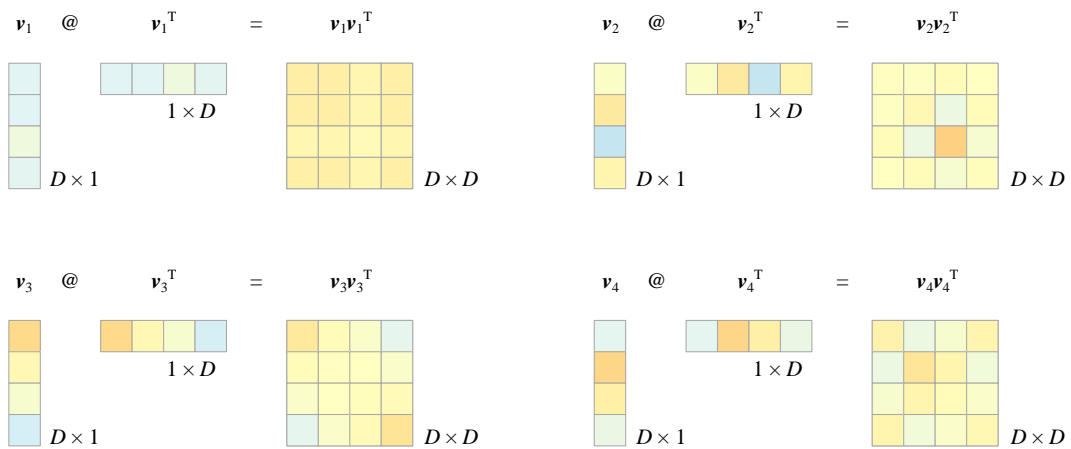


图 37. 列向量乘自身转置获得四个张量积

图 38 所示为张量积运算，和图 37 结果完全一致。

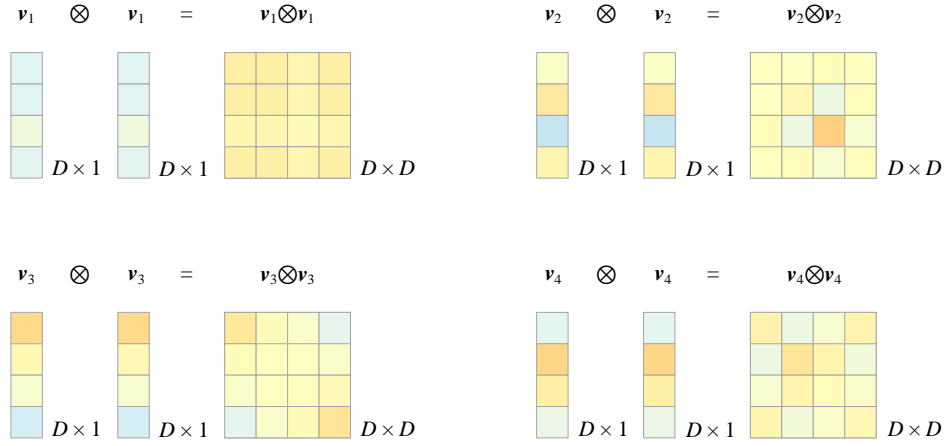


图 38. 内积计算获得四个张量积

容易推导得到，(53) 中张量积相加得到单位矩阵：

$$\mathbf{v}_1 \otimes \mathbf{v}_1 + \mathbf{v}_2 \otimes \mathbf{v}_2 + \dots + \mathbf{v}_D \otimes \mathbf{v}_D = \left(\sum_{j=1}^D \mathbf{v}_j \otimes \mathbf{v}_j \right) = \mathbf{I} \quad (54)$$

上式如图 39 热图所示。

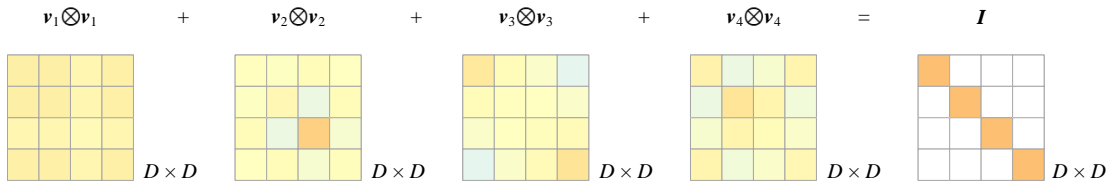


图 39. 张量积相加得到单位矩阵

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

$$\mathbf{X}_1 = \mathbf{z}_1 \mathbf{v}_1^T = \mathbf{X} \mathbf{v}_1 \mathbf{v}_1^T = \mathbf{X} \underbrace{(\mathbf{v}_1 \otimes \mathbf{v}_1)}_{\text{Tensor product}} \quad (55)$$

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

$$\mathbf{X}_2 = \mathbf{z}_2 \mathbf{v}_2^T = \mathbf{X} \mathbf{v}_2 \mathbf{v}_2^T = \mathbf{X} \underbrace{(\mathbf{v}_2 \otimes \mathbf{v}_2)}_{\text{Tensor product}} \quad (56)$$

图 40 所示为张量积 $\mathbf{v}_1 \otimes \mathbf{v}_1$ 和 $\mathbf{v}_2 \otimes \mathbf{v}_2$ 还原部分数据 \mathbf{X} ；图 41 所示为张量积 $\mathbf{v}_3 \otimes \mathbf{v}_3$ 和 $\mathbf{v}_4 \otimes \mathbf{v}_4$ 还原部分数据 \mathbf{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。

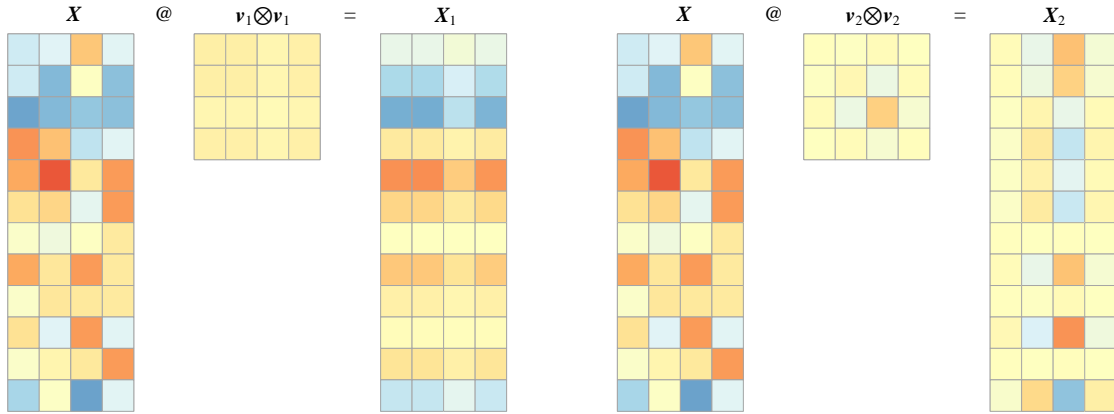


图 40. 张量积 $X(v_1 \otimes v_1)$ 和 $X(v_2 \otimes v_2)$ 还原部分数据 X

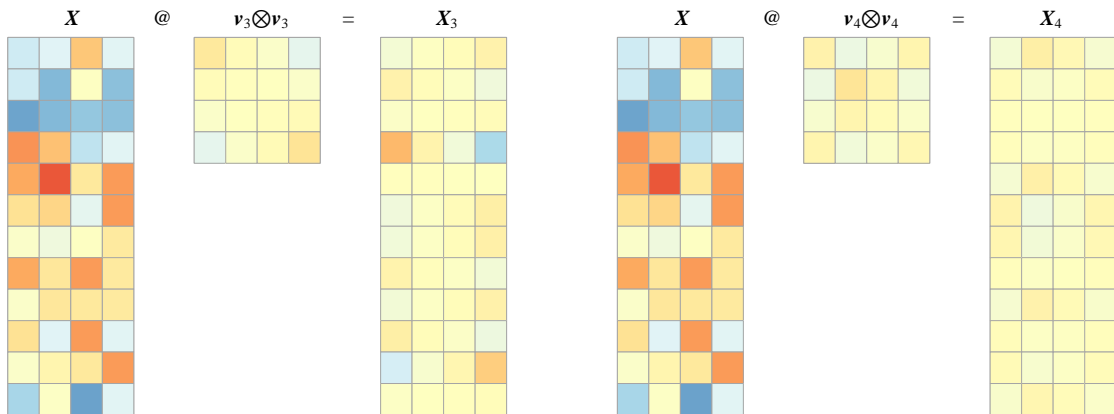


图 41. 张量积 $X(v_3 \otimes v_3)$ 和 $X(v_4 \otimes v_4)$ 还原部分数据 X

误差

图 42 所示为两个主成分 v_1 和 v_2 还原获得原始数据热图，具体计算如下：

$$\hat{X} = \begin{bmatrix} z_1 & z_2 \end{bmatrix} \begin{bmatrix} v_1 & v_2 \end{bmatrix}^T \quad (57)$$

相当于

$$\begin{aligned} \hat{X} &= X_1 + X_2 = z_1 v_1^T + z_2 v_2^T \\ &= X(v_1 v_1^T + v_2 v_2^T) = X(v_1 \otimes v_1 + v_2 \otimes v_2) \end{aligned} \quad (58)$$

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，

$$\mathbf{X} \approx \mathbf{X}(\mathbf{v}_1 \otimes \mathbf{v}_1 + \mathbf{v}_2 \otimes \mathbf{v}_2) = s_1 \mathbf{u}_1 \otimes \mathbf{v}_1 + s_2 \mathbf{u}_2 \otimes \mathbf{v}_2^T \quad (59)$$

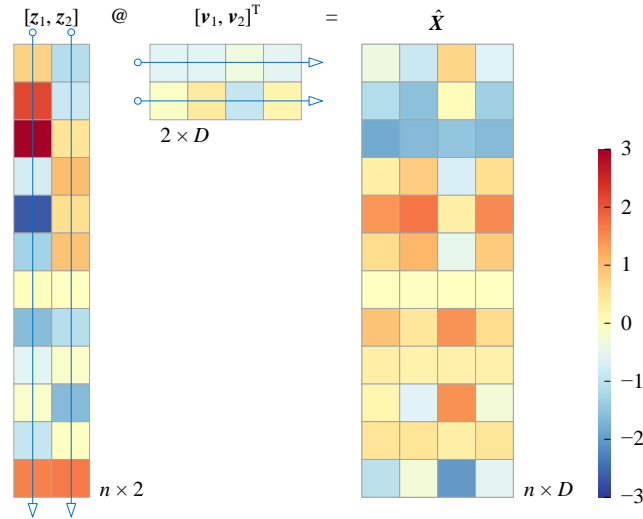
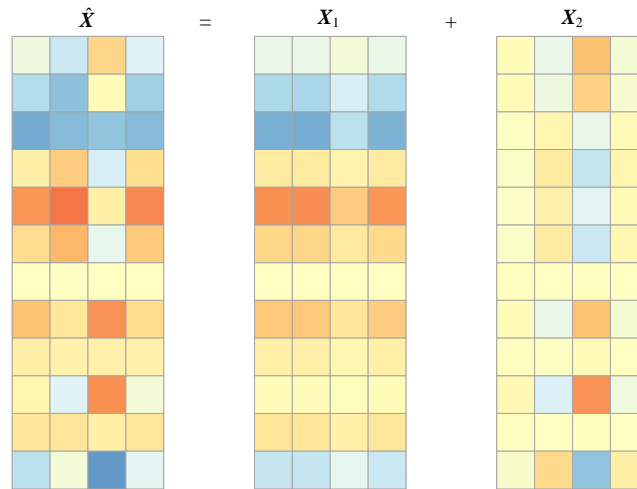
图 42. 前两个主成分 z_1 和 z_2 还原 \mathbf{X} 数据

图 43. 两个热图叠加还原原始数据

残差数据矩阵 \mathbf{E} ，即原始热图和还原热图色差，利用下式计算获得：

$$\mathbf{E} = \mathbf{X} - \hat{\mathbf{X}} \quad (60)$$

图 44 比较原始数据 \mathbf{X} 、拟合数据 $\hat{\mathbf{X}}$ 和残差数据矩阵 \mathbf{E} 热图，发现原始数据 \mathbf{X} 和拟合数据 $\hat{\mathbf{X}}$ 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。

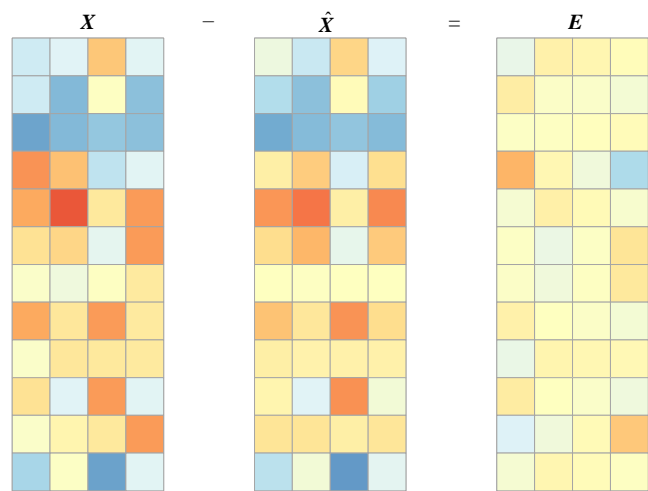


图 44. 原始数据、拟合数据和残差数据热图

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。

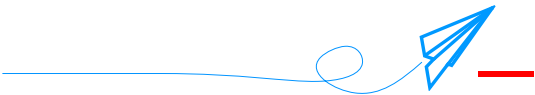
总结来说，通过 PCA 降维，我们可以减少数据的维度，从而简化模型和算法的复杂度，同时可以去除噪声和冗余信息，提高数据的可解释性和可视化效果，从而更好地理解数据和发现数据中的规律。PCA 广泛应用于数据挖掘、模式识别、图像处理、信号处理等领域。

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

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



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。希望大家日后在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。此外，让我们和鸡、兔、猪这三个伙伴说声感谢，感谢它们在学习数学路上的陪伴！再见！

下面，我们一起踏上《数据有道》、《机器学习》的“实践”之旅！