

16 主成分分析

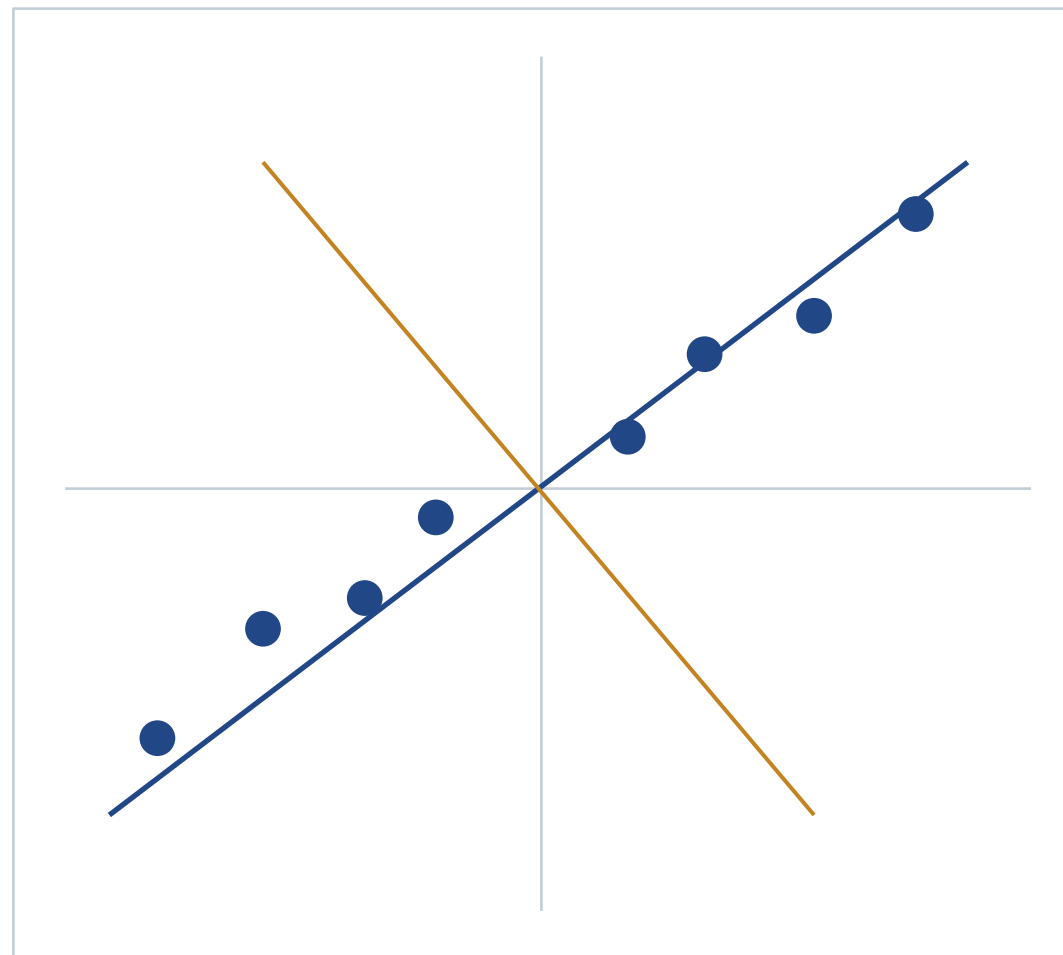
用少数正交方向表示数据的主要变化

1 降维问题

相关特征常能用较少的线性坐标近似表示

- **高维数据可能存在冗余**
 - 多个特征共同变化，记录了相近的信息
- **低维表示的用途**
 - 可视化、压缩、去除部分扰动或作为后续模型输入
- **PCA 的选择**
 - 寻找一组正交方向，优先保留样本方差较大的方向
- **需要检验的假设**
 - 大方差方向对当前分析是否确实重要？

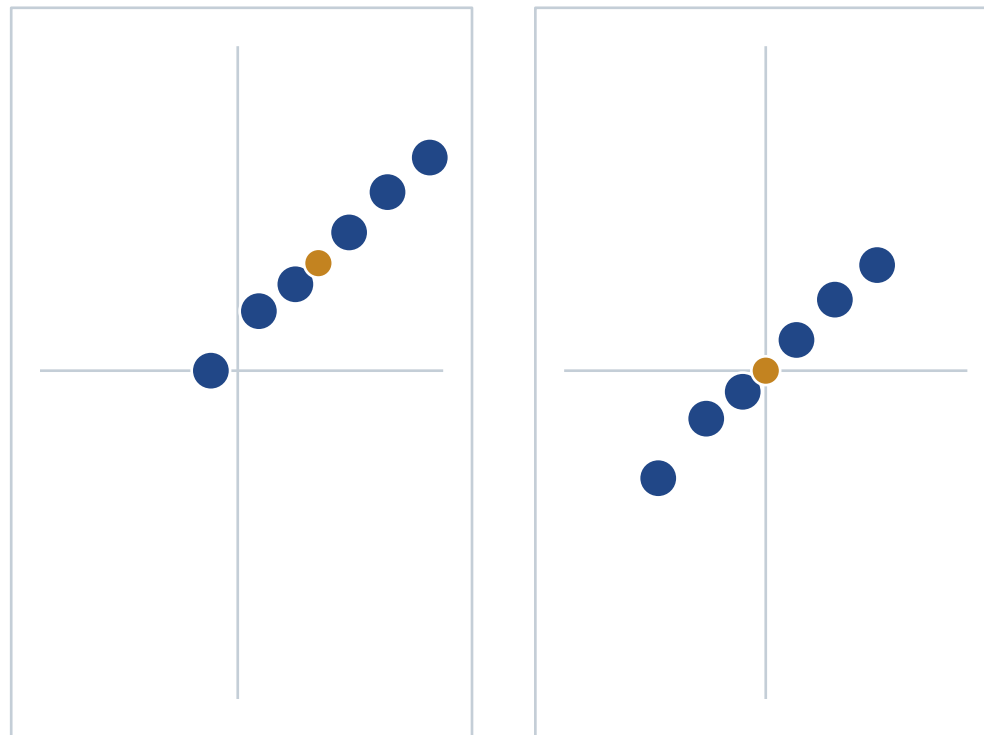
- 两个特征共同变化，点云呈现倾斜延展
- 比较两个候选方向上的投影跨度：哪一个更大？
- 几何长轴只是猜想；PCA 还需要中心化和方差目标



方向 1：投影跨度大 方向 2：投影跨度小

- 训练数据 $X \in \mathbb{R}^{N \times d}$: 样本为行、特征为列
- 训练均值: $\mu = \frac{1}{N} \sum_{i=1}^N x_i$
- 减去均值: $X_c = X - \mathbf{1}\mu^T$
- 整体平移改变均值; 中心化后的散布和主方向不变

橙点: 样本均值



原坐标: 均值偏离原点

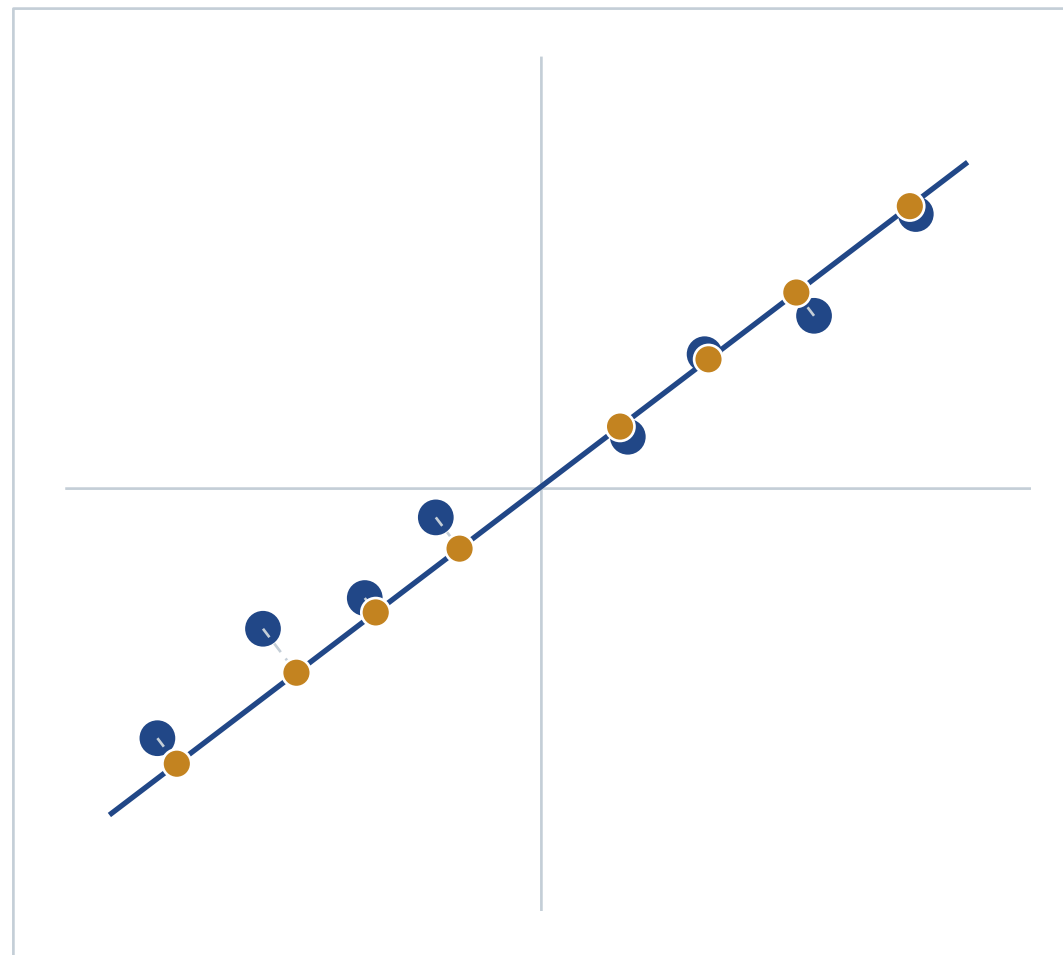
中心化: 均值在 origin

- **主方向** $w_j \in \mathbb{R}^d$
 - 特征空间中的单位向量，定义新的坐标轴
- **主成分得分** z_{ij}
 - 第 i 个样本在方向 w_j 上的坐标
 - $z_{ij} = (x_i - \mu)^T w_j$
- **用前 k 个方向重构**
 - $\hat{x}_i = \mu + W_k W_k^T (x_i - \mu)$

2 PCA 的目标

最大投影方差与最小平方重构误差给出同一子空间

- 选一个单位候选方向 w
- 把每个中心化样本正交投到该方向
- 橙点在方向上的坐标是投影得分
- 投影越分散，得分的方差越大

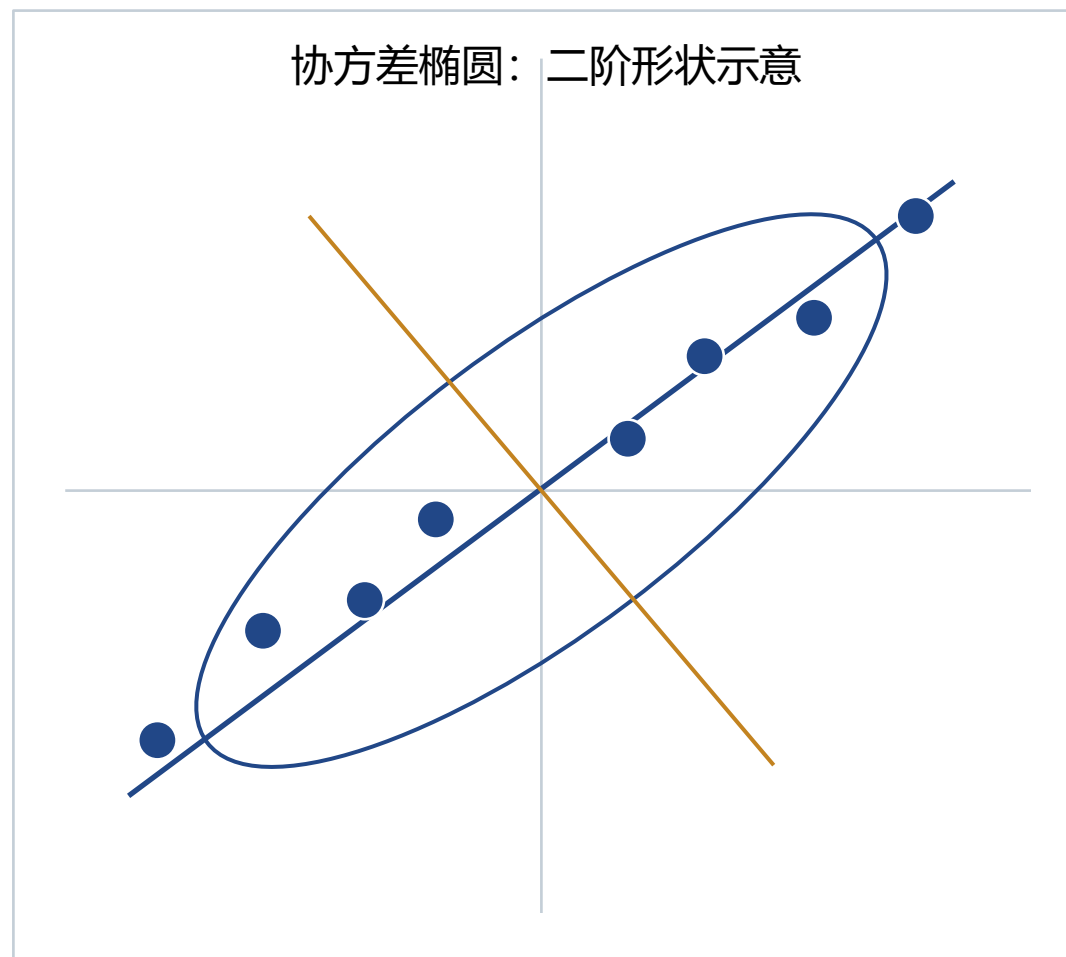


橙点：正交投影

- 对中心化数据 X_c , 任意单位方向 w 的得分为 $z = X_c w$
- 采用 $1/N$ 约定, 定义协方差 $S = X_c^T X_c / N$
- 得分的方差为
 - $Var(z) = \frac{1}{N} \|X_c w\|^2 = w^T S w$
- 同一个 S 可比较所有单位方向, 不只图中的两条轴

约定说明: 本节采用 $1/N$ 。若采用 $1/(N-1)$, 主方向和解释方差比不变; 特征值整体乘以 $N/(N-1)$ 。

- 特征方向满足 $Sw_i = \lambda_i w_i$, 是二阶主轴
- 特征值 λ_i 是方向 w_i 上的投影方差
- 协方差椭圆在该方向上的半轴长度为 $c\sqrt{\lambda_i}$
- 比例系数 c 由所选等方差轮廓或置信水平决定
- 椭圆只描述二阶散布, 不决定完整分布形状



w_1 : 方差 λ_1 , 半轴 $c\sqrt{\lambda_1}$ w_2 : 方差 λ_2 , 半轴 $c\sqrt{\lambda_2}$

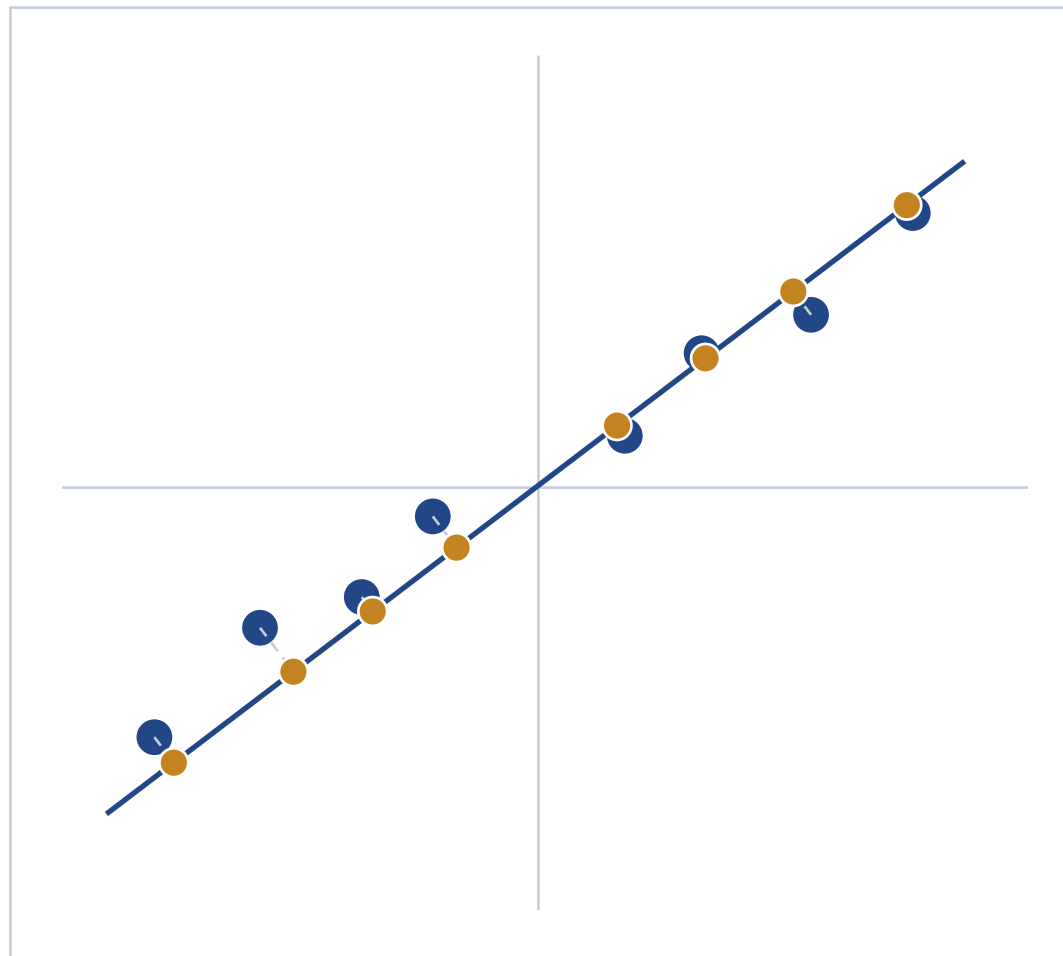
只表示二阶散布

第一主方向的选择

- S 的特征方向 w_j 构成正交基
- 任意单位方向都可展开:
 - $w = \sum_j a_j w_j, \sum_j a_j^2 = 1$
- 因为 $\lambda_j \leq \lambda_1$, 其投影方差满足
 - $w^T S w = \sum_j \lambda_j a_j^2 \leq \lambda_1$
- 选 $w = w_1$ 达到上界: $S w_1 = \lambda_1 w_1$

- 第二主方向与 w_1 正交, 并使剩余投影方差最大
- 依次得到
 - $Sw_j = \lambda_j w_j$
 - $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$
- 主方向彼此正交
- 前 k 个主方向组成 $W_k = [w_1, \dots, w_k]$

- 橙点是中心化样本在所选子空间上的投影
- 虚线是样本到子空间的正交残差
- 投影后的矩阵为 $\hat{X}_c = X_c W_k W_k^T$
- 舍弃的方向决定平方重构损失



橙点：正交投影

虚线：正交残差

- W_k 的列是 k 个正交单位方向, 满足 $W_k^T W_k = I$
- 对给定 W_k , 重构残差为 $X_c - X_c W_k W_k^T$
- PCA 选取平方残差最小的 k 维子空间:
 - $\min_{W_k^T W_k = I} \|X_c - X_c W_k W_k^T\|_F^2$
- Frobenius 范数平方是所有样本、所有特征残差的平方和

- 对正交投影，中心化数据分成保留分量与残差
- 勾股关系给出
 - $\|X_c\|_F^2 = \|X_c W_k W_k^T\|_F^2 + \|X_c - X_c W_k W_k^T\|_F^2$
- 左侧总变化固定
- 因此最大化保留方差与最小化平方重构误差选择同一子空间

3 二维到一维的贯穿例子

同一组数据贯穿中心化、训练、投影、重构和新样本变换

- $X_c = \begin{bmatrix} -2 & -1 \\ -1 & -2 \\ 1 & 2 \\ 2 & 1 \end{bmatrix}$
- 两个特征的均值都为 0
- 两个特征大体同向变化
- 目标：把每个二维样本变成一个一维得分
- 几何预判：第一主方向接近直线 $x_2 = x_1$

- $S = \frac{1}{4} X_c^T X_c = \begin{bmatrix} 2.5 & 2 \\ 2 & 2.5 \end{bmatrix}$
- 对角线 2.5：两个特征各自的方差
- 非对角线 2：两个特征倾向同向变化
- 较大的正协方差支持 “主要方向接近对角线” 的预判

- 特征值

- $\lambda_1 = 4.5, \lambda_2 = 0.5$

- 对应主方向

- $w_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, w_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

- 第一主成分解释方差比

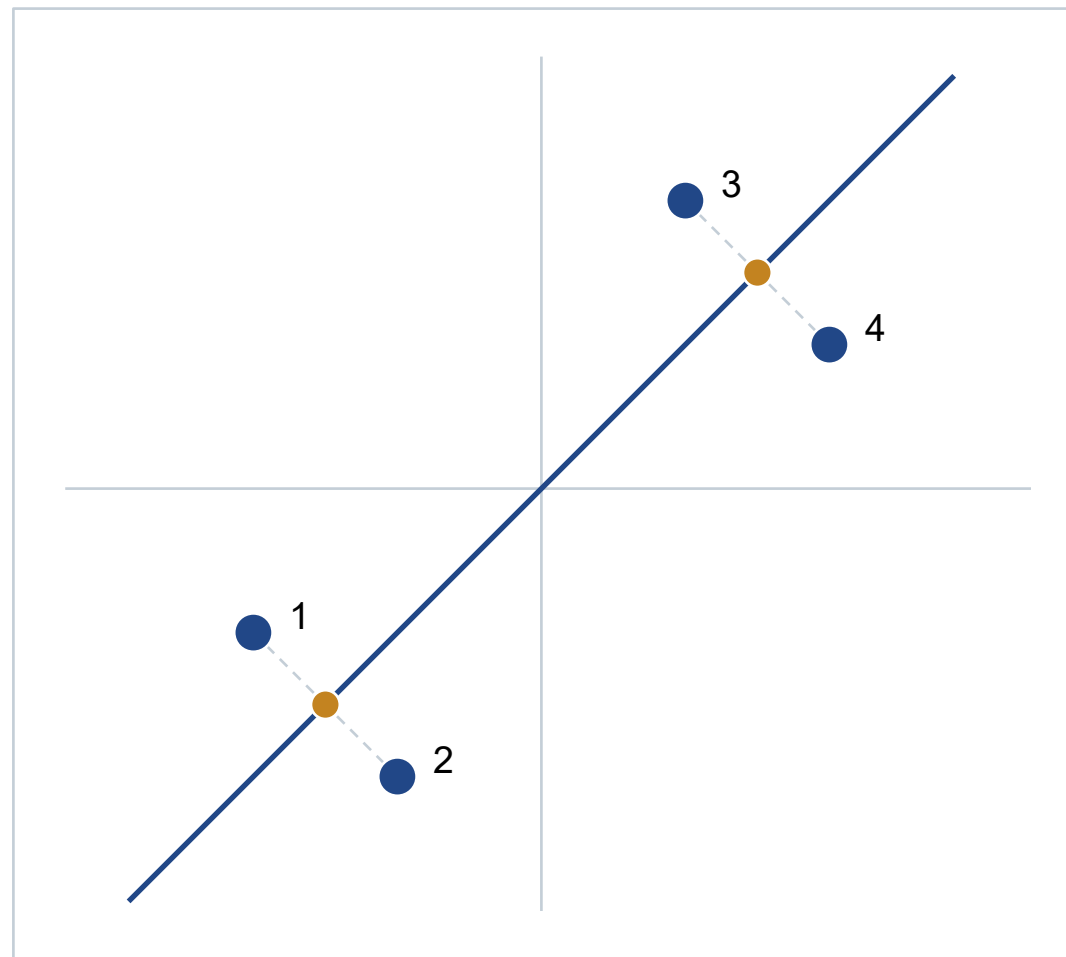
- $\frac{4.5}{4.5+0.5} = 90\%$

- $z = X_c w_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} -3 \\ -3 \\ 3 \\ 3 \end{bmatrix}$
- 前两个样本位于主方向的负侧
- 后两个样本位于主方向的正侧
- 一维得分保留了两个特征共同增减的主要变化
- 两个不同样本可能得到相同得分，因为次要方向已被舍弃

- $\hat{X}_c = zw_1^T = \begin{bmatrix} -1.5 & -1.5 \\ -1.5 & -1.5 \\ 1.5 & 1.5 \\ 1.5 & 1.5 \end{bmatrix}$
- 重构点都落在直线 $x_2 = x_1$ 上
- 原样本在 w_2 方向上的差异被舍弃
- 保留一维时
 - $\frac{1}{N} \|X_c - \hat{X}_c\|_F^2 = \lambda_2 = 0.5$
- 这里对样本求平均；每个样本的损失是其二维重构残差的平方长度

例子：投影为何合并样本

- 样本 1、2 的一维得分都为 $-3/\sqrt{2}$
- 样本 3、4 的一维得分都为 $3/\sqrt{2}$
- 各对样本在主方向上相同，在次方向上不同
- 重构保留主方向坐标，丢失这两对差异



蓝点：原样本

橙点：重构点

- **训练阶段保存**

- 训练均值 μ 与前 k 个主方向 W_k

- **变换新样本**

- $z_{new} = (x_{new} - \mu)^T W_k$

- **需要时重构**

- $\hat{x}_{new} = \mu + W_k z_{new}^T$

- 均值与主方向只从训练集估计；测试集参与估计会造成数据泄漏

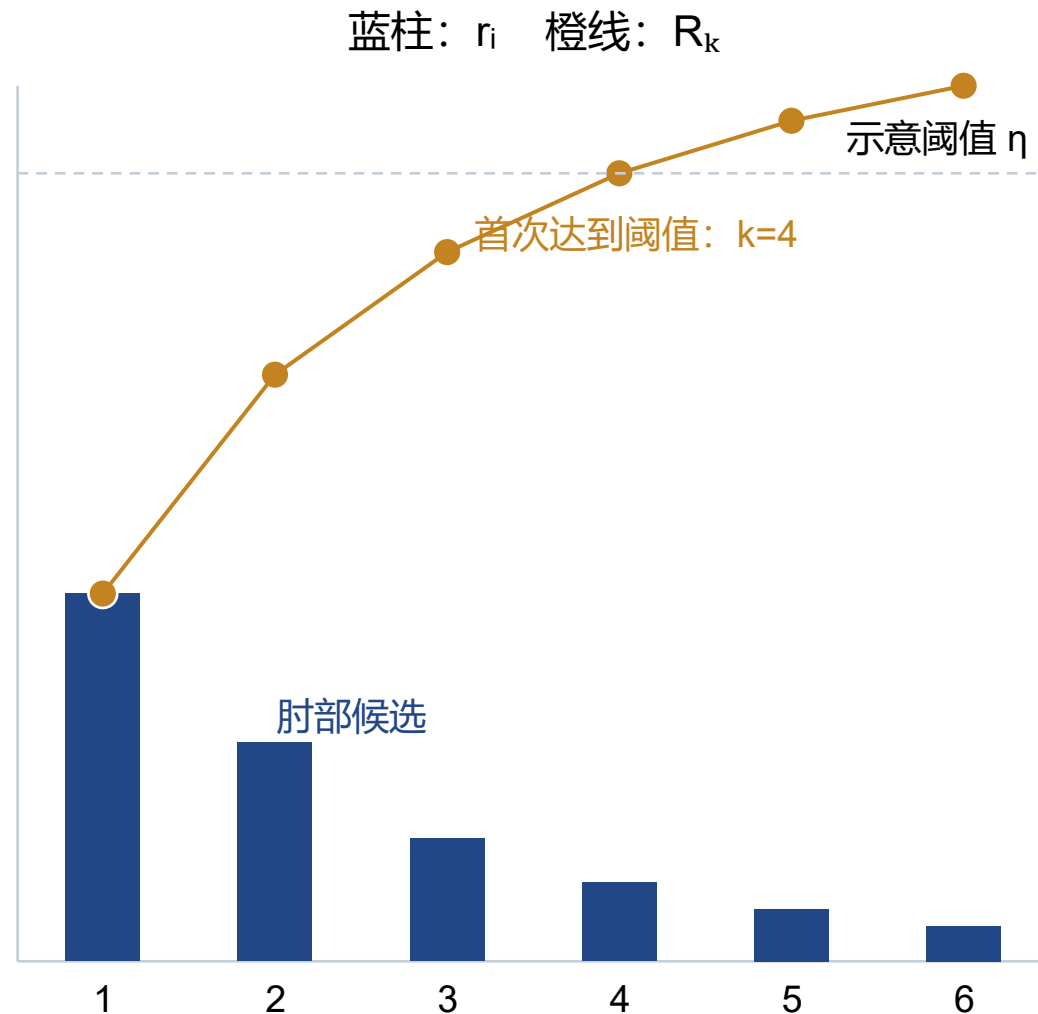
4 SVD 与使用选择

中心化数据的 SVD 直接给出主方向、得分和解释方差

- 中心化数据 $X_c \in \mathbb{R}^{N \times d}$
- 完整 SVD: $X_c = U \Sigma V^T$
- $U \in \mathbb{R}^{N \times N}$ 与 $V \in \mathbb{R}^{d \times d}$ 为正交矩阵
- $\Sigma \in \mathbb{R}^{N \times d}$ 可为矩形, 对角项是奇异值 $\sigma_i \geq 0$
- V 的列在特征空间, U 的列在样本空间

- 中心化协方差满足
 - $S = \frac{1}{N} X_c^T X_c = V \frac{\Sigma^T \Sigma}{N} V^T$
- $\Sigma^T \Sigma \in \mathbb{R}^{d \times d}$, 其对角项是 σ_i^2
- V 的列是 PCA 的特征空间主方向
- $X_c V = U \Sigma$ 是样本在全部方向上的得分
- 采用 $1/N$ 约定时, 解释方差 $\lambda_i = \sigma_i^2 / N$

- 总方差 $\sum_j \lambda_j > 0$ 时, 第 i 个方向的解释方差比
 - $r_i = \frac{\lambda_i}{\sum_j \lambda_j}$
- 前 k 个方向的累计解释方差比
 - $R_k = \sum_{i=1}^k r_i$
- 可以结合累计解释方差、碎石图、重构质量和下游验证选择 k
- 不存在适用于所有任务的固定阈值



示意谱, 不对应前述二维算例

■ 中心化

- 每个特征减去训练均值
- 让 PCA 分析围绕均值的变化

■ 标准化

- 常进一步除以训练标准差
- 让不同量纲的特征具有可比较尺度

■ 选择依据

- 原始单位和方差是否具有直接意义
- 是否允许大尺度特征自然占更大权重

5 经典应用

同一个线性子空间可以服务于观察、表示、压缩与预处理

用途	PCA 操作与输出	关键判断
可视化	投影到前两个或三个方向，得到低维坐标	观察总体结构；投影重叠可能掩盖原始差异
压缩与特征提取	用前 k 个得分表示样本，可选重构	检查表示维度与重构损失
降噪	舍弃弱方向，再由保留得分重构	弱方向主要承载扰动时才有依据
模型预处理	用训练所得方向变换输入特征	用验证集判断是否改善下游任务

- PCA 只寻找线性子空间，可能错过非线性结构
- 大方差方向不一定包含标签判别信息或业务价值
- 协方差及其主方向对特征尺度和异常值敏感
- 主方向的符号不唯一；特征值重复时，主子空间确定，但其中的单个基方向不唯一
- 主成分是原特征的线性组合，解释性和下游效果都需要结合任务验证

- **准备**：用训练均值中心化数据
- **训练**：求协方差矩阵的特征向量，或对中心化数据做 SVD
- **变换**：把样本投影到前 k 个主方向，得到低维得分
- **解释**：在本节 $1/N$ 约定下，特征值给出各方向的平均平方变化；舍弃特征值之和给出平均平方重构损失
- **使用**：结合尺度、任务目标和验证结果决定是否采用 PCA