16 主成分分析

用少数正交方向表示数据的主要变化

本讲学习目标

- 解释 PCA 的降维问题及中心化的作用
- 说明最大投影方差如何确定主方向
- 区分主方向、主成分得分与重构，解释保留与丢失的变化
- 说明保留方差与平方重构误差的关系及成立条件
- 根据给定计算示范，解释训练、变换、重构与新样本复用
- 理解 SVD 求解接口，以及选维、特征尺度、典型用途和适用边界

1 降维与中心化

用少数坐标保留主要变化，先中心化再选择方向

PCA 解决的降维问题

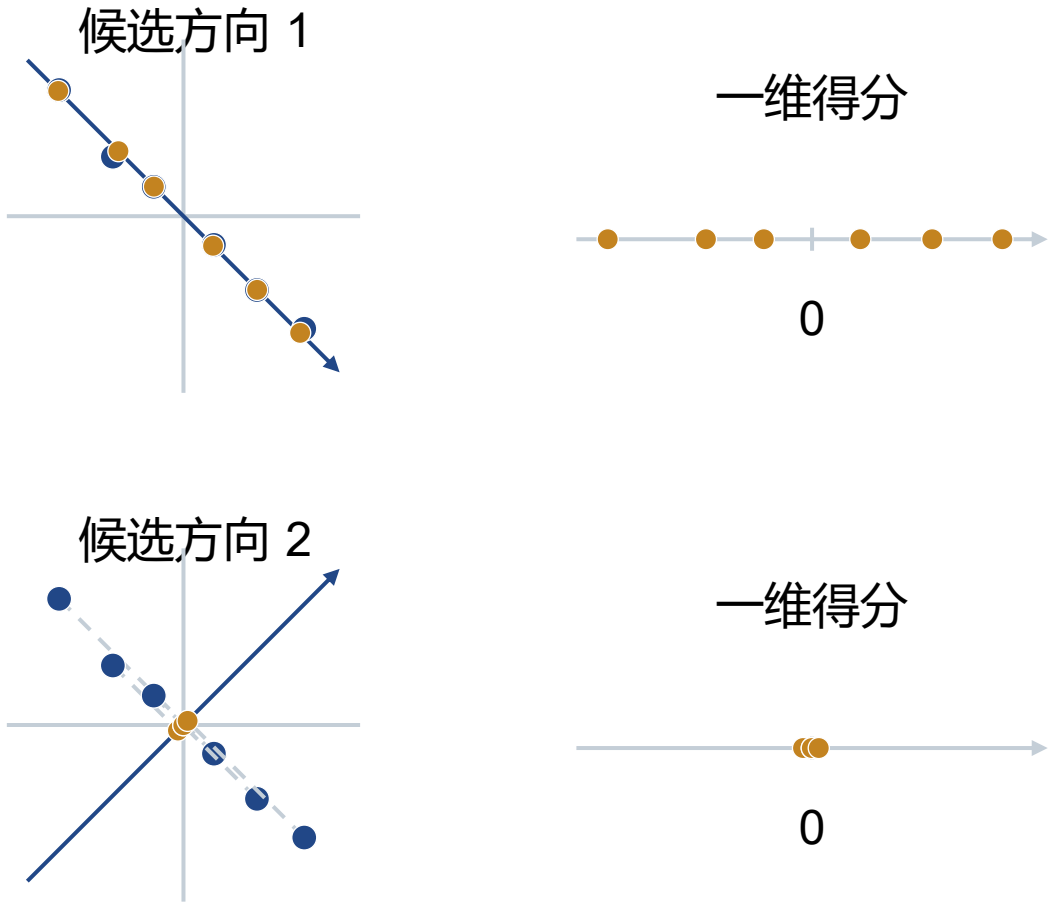
每个样本由多个特征描述。降维用较少坐标表示样本，同时尽量保留样本之间的变化。

- **低维观察与紧凑表示**提出减少坐标的需求
  - 二维或三维坐标便于观察样本关系
  - 允许一定近似时，可用较少数值存储或传递样本；高维本身不意味着必须降维
- 多个特征可能共同变化，少数线性方向可能概括大部分变化
  - 保留这些方向上的坐标，可以形成较低维的表示

减少坐标可能丢失差异：依据什么选择保留的方向，怎样衡量保留与丢失的变化？

点云中的候选方向

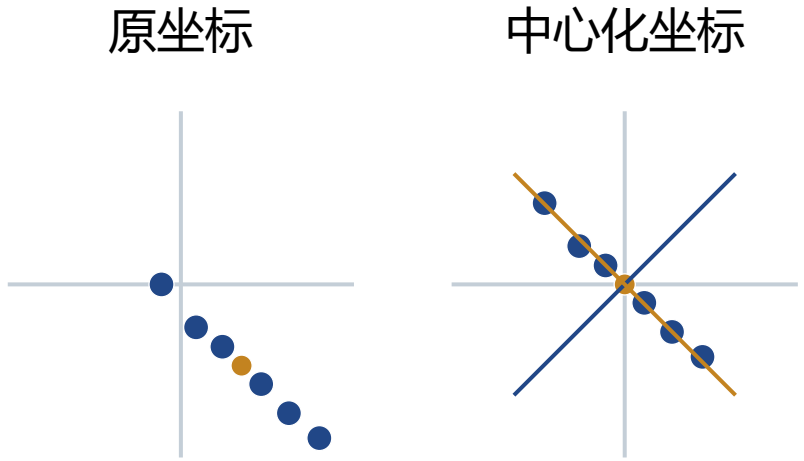
- 两个特征共同变化时，点云可能沿倾斜方向延展
  - 特征的线性组合定义新的坐标；候选坐标轴不限于原来的某个特征轴
- **投影方向决定保留的差异**
  - 沿延展方向投影，主要差异仍表现为坐标变化
  - 沿较窄方向投影，许多差异被压缩到较小范围
- 由此形成降维思路：换一组方向，保留变化较多的方向，用少数坐标概括点云
  - 目测只提供候选方向；比较依据还需区分均值位置与样本变化，并用投影方差量化



两条得分轴使用相同尺度

中心化：区分均值位置与样本变化

- 训练数据 $X \in \mathbb{R}^{N\times d}$ 按行记录样本、按列记录特征
  - N 为样本数，d 为特征数；第 i 行是样本列向量 x_i 的转置
- **中心化**使每个样本减去同一训练均值
  - $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i , X_c = X - \mathbf{1}\mu^T$
  - $\mathbf{1}$ 是 N 维全 1 列向量；X_c 的每列均值为 0
- 中心化改变坐标原点，保持样本相对位置与任意方向的投影方差
  - 整体平移数据后再中心化，得到相同的 X_c
- 原始投影的平均平方值等于投影方差加上投影均值的平方
  - 直接用该平均平方值选择方向，均值位置可能影响结果
  - **中心化后投影均值为 0，平均平方值等于投影方差**



橙点：样本均值

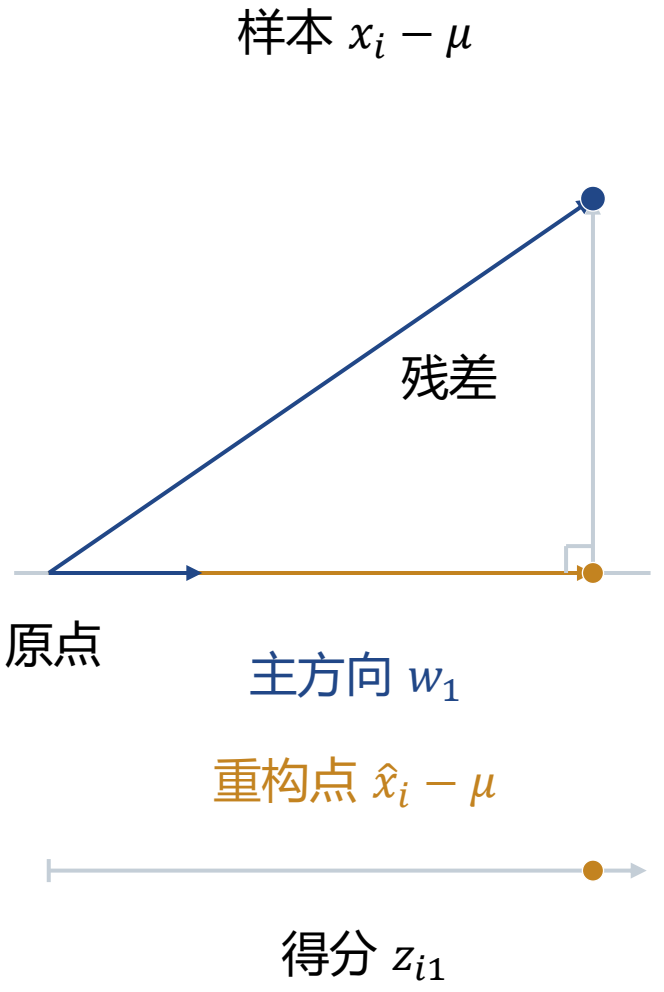
右图蓝轴、橙轴为候选方向

主方向、主成分得分与重构

主成分分析（Principal Component Analysis，PCA）是一种无监督线性降维方法：对中心化数据选择彼此正交的单位方向，优先保留投影方差较大的方向。正交单位方向两两垂直、长度为 1。

- **主方向** $w_j \in \mathbb{R}^d$ 定义特征空间中的新坐标轴
  - 前 k 个正交单位方向组成 $W_k = [w_1, \dots, w_k] \in \mathbb{R}^{d \times k}$
- **主成分得分** z_{ij} 是预处理后样本在主方向 w_j 上的标量坐标
  - 样本按训练参数预处理；只中心化而未标准化时，$z_{ij} = (x_i - \mu)^T w_j$
- **重构** $\hat{x}_i \in \mathbb{R}^d$ 用保留得分形成原始特征空间中的近似样本

当前讨论只中心化的 PCA；标准化后的表示与重构见使用条件。



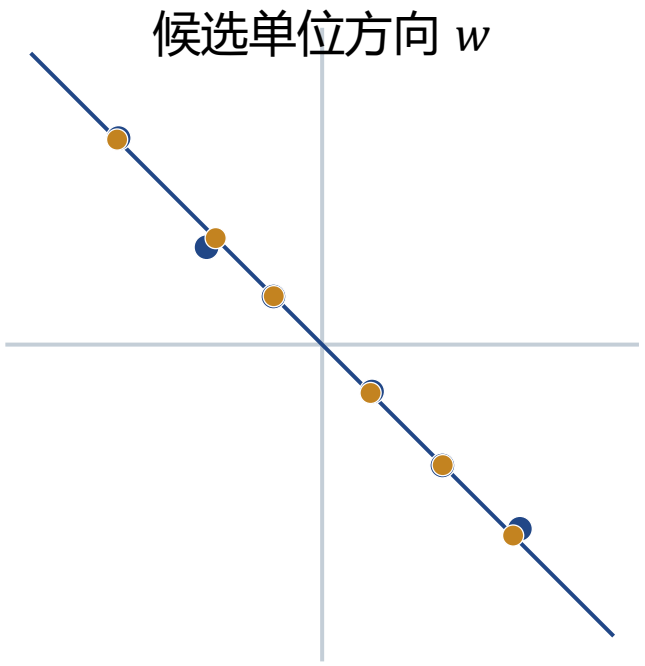
2 方差准则、主方向与重构误差

以投影方差选择方向，用正交分解联系保留与丢失

候选方向上的投影得分

对中心化样本选择单位候选方向，正交投影形成该方向上的标量坐标。

- 图中橙点的位置对应**投影得分**
  - 得分集中时，原有差异被压缩；得分分散时，沿该方向保留的变化较多
- **投影方差**是得分围绕均值的平均平方偏离
  - PCA 以最大投影方差选择方向，将投影中的分散程度转为可比较的量
- 方差准则衡量数据变化，未使用标签或下游任务
  - 较大方差本身不保证更高任务价值



蓝点：样本　橙点：投影位置

灰线：正交残差

投影方差的协方差矩阵表达

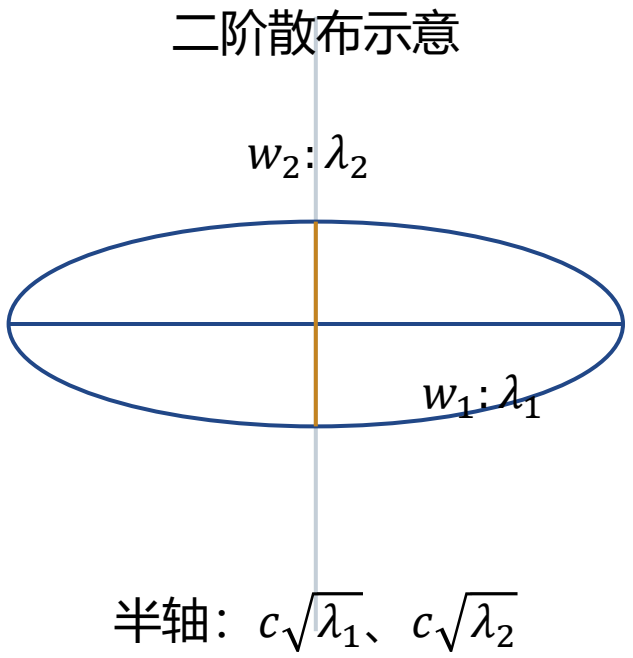
- 对中心化数据 X_c，任意单位方向 w 的投影得分向量为 $z = X_c w$
  - 第 i 项对应第 i 个样本
  - 因为 X_c 的各列均值为 0，投影得分 z 的均值也为 0
- 采用 $1/N$ 约定，定义**协方差矩阵** $S = X_c^T X_c / N$
  - 对角元素是各特征的方差，非对角元素是两两特征的协方差
  - 它把各特征的变化及线性共同变化汇总在同一个矩阵中
- 投影得分的方差可由同一个 S 计算
  - $Var(z) = \frac{1}{N}\|X_c w\|^2 = w^T S w$
  - 改变单位方向 w，即可比较各方向的投影方差
- 单位长度约束使比较只针对方向
  - 若任意放大 w，得分及其方差也会随之放大，无法据此选择方向

协方差矩阵的主轴与解释方差

协方差矩阵 S 的单位特征向量满足 $Sw_i = \lambda_i w_i$：矩阵作用后，向量仍在同一轴上。

- 彼此正交的单位特征向量作为 PCA 的主方向，也是**二阶散布的主轴**
- 特征值 λ_i 等于方向 w_i 上的投影方差，称为该方向的**解释方差**
- 示意椭圆的半轴长为 $c\sqrt{\lambda_i}$
  - 共同系数 $c > 0$ 决定轮廓尺度；椭圆描述二阶散布，概率或置信含义需要额外条件

主轴给出候选方向；最大方差主轴的最优性，还需用方差二次型比较所有单位方向。



不是点云外轮廓或置信区域

第一主方向达到投影方差上界

$S = X_c^T X_c / N$ 是实对称半正定矩阵，可选择正交单位特征向量基 w_j，对应特征值非负。

- 任意单位方向都可在这组基中展开
  - $w = \sum_j a_j w_j, \sum_j a_j^2 = 1$
- 投影方差是特征值的加权平均
  - $w^T S w = \sum_j \lambda_j a_j^2 \leq \lambda_1$
  - 按 $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ 排序，$\lambda_j \leq \lambda_1$；权重 $a_j^2 \geq 0$，总和为 1
- 取 $w = w_1$，由 $S w_1 = \lambda_1 w_1$ **达到方差上界**
  - 因而最大特征值对应的单位特征向量可取作第一主方向

后续主方向的正交约束

PCA 建立一组正交坐标轴。后续方向需与已选方向正交，避免重复选择同一轴。

- 第二主方向在与 w_1 正交的单位方向中使投影方差最大
  - $w_2 = argmax_{\|w\|=1, w^T w_1 = 0} w^T S w$
  - 正交约束使特征基展开中 w_1 的系数为 0；投影方差由剩余特征值加权构成
  - 上界为 λ_2，取对应单位特征向量 w_2 达到
- 第 j 个方向在**已选方向的正交补空间**中取得最大方差
  - 约束 $w^T w_\ell = 0 (\ell < j)$ 使前 $j-1$ 项系数为 0
  - 剩余方差上界为 λ_j，由对应单位特征向量达到
- 依次得到按方差降序排列的正交主方向
  - $S w_j = \lambda_j w_j$，$\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$
  - 前 k 个方向组成 $W_k = [w_1, \ldots, w_k]$

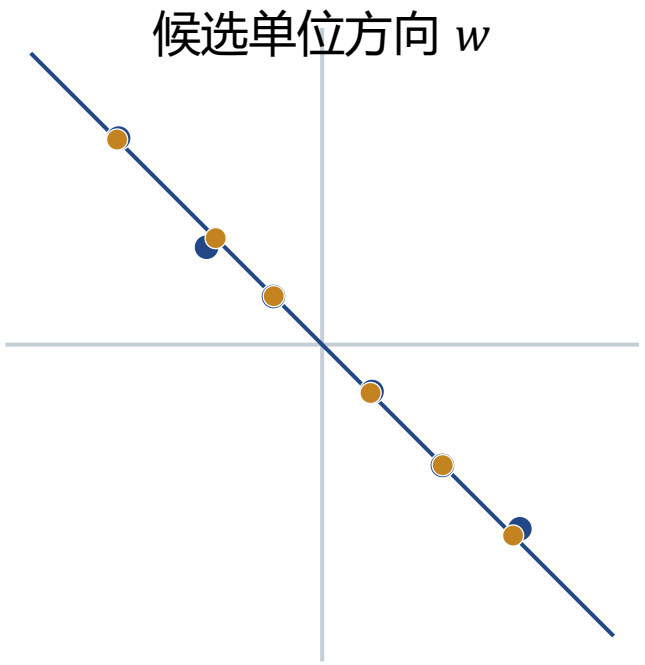
主成分得分的方差与不相关性

- 第 i 列主成分得分 $X_c w_i$ 的方差等于对应特征值
  - $Var(X_c w_i) = w_i^T S w_i = \lambda_i$
- 不同列的主成分得分**不相关**
  - 对 $i \neq j$，由主方向正交性得到
  - $Cov(X_c w_i, X_c w_j) = w_i^T S w_j = \lambda_j w_i^T w_j = 0$
  - 不相关指协方差为 0，一般不能推出统计独立；独立要求一个得分的信息不改变另一个得分的分布
- 协方差的分母约定改变方差尺度
  - 本讲用 $1/N$，对应训练样本平均平方得分；第 15 讲用 $1/(N-1)$
  - 当 $N > 1$ 时，后者的特征值乘以 $N/(N-1)$；主方向、得分及非零总方差下的解释方差比保持不变

正交投影的保留分量与重构残差

W_k 的 k 个正交单位候选方向张成保留子空间。正交投影分离保留分量与垂直残差。

- **投影形成低维坐标**
  - $Z = X_c W_k \in \mathbb{R}^{N \times k}$
  - 当 W_k 取前 k 个主方向时，Z 是保留的主成分得分矩阵
- 沿保留方向重构特征空间中的分量
  - $\hat{X}_c = Z W_k^T = X_c W_k W_k^T$
  - 加回训练均值，单个样本为 $\hat{x}_i = \mu + W_k W_k^T (x_i - \mu)$
- **正交残差**是被舍弃的分量
  - 橙点为保留分量；虚线残差的平方长度决定平方重构误差

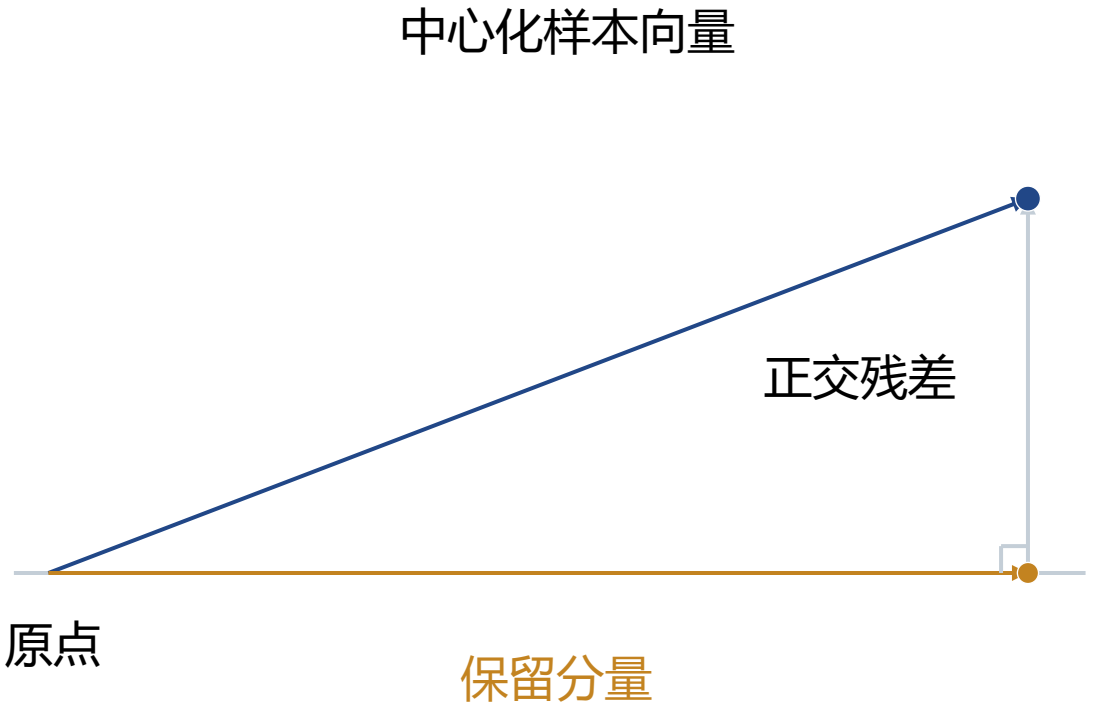


蓝点：样本　橙点：投影位置

灰线：正交残差

正交分解中的保留分量与残差

- 中心化样本分为子空间内的**保留分量**与垂直于子空间的**残差**
  - 图中从原点到样本的向量，等于保留向量与残差向量之和
- 两个分量正交，勾股关系将样本的平方长度分成保留量与残差量
  - 对所有样本求和，得到总平方量的同样分解
- 给定同一中心化数据，总平方量固定；保留量增加时，残差量等量减少
  - 两个优化目标的联系还需固定保留维度，并采用正交投影与平方误差



单个样本的平方长度相加

对全部样本求和得到总量分解

最大保留方差与最小平方重构误差的等价条件

对同一中心化数据，固定保留维度 k，候选方向满足 $W_k^T W_k = I$。记重构为 $\hat{X}_c = X_c W_k W_k^T$。

- **总量固定，保留量与残差此消彼长**
  - $\|X_c\|_F^2 = \|\hat{X}_c\|_F^2 + \|X_c - \hat{X}_c\|_F^2$
  - Frobenius 范数平方是全部样本、全部特征的平方和
- **保留平方量对应投影方差之和**
  - 记 $W_k = [q_1, \dots, q_k]$，$S = X_c^T X_c / N$
  - $\|\hat{X}_c\|_F^2 / N = \|X_c W_k\|_F^2 / N = \sum_{\ell=1}^{k} q_\ell^T S q_\ell$
- 因此最大保留方差等价于最小平方重构误差
  - $min_{W_k^T W_k = I} \|X_c - X_c W_k W_k^T\|_F^2$
  - 等价依赖同一中心化数据、固定 k、正交投影与平方误差

前 k 个主方向的联合最优性

比较 k 维候选子空间在各主方向上保留的平方长度。

- 令 $c_j = \|W_k^T w_j\|^2$，表示单位主方向 w_j 投影后保留的平方长度
  - 正交投影不增加长度，故 $0 \le c_j \le 1$
  - 候选子空间有 k 个正交单位轴，在完整主方向基中的平方系数总和为 k，即 $\sum_j c_j = k$
- 总保留方差不超过前 k 个特征值之和
  - $\frac{1}{N}\|X_c W_k\|_F^2 = \sum_j \lambda_j c_j \le \sum_{j=1}^{k} \lambda_j$
  - 各权重最多为 1、总和为 k；权重集中于最大的 k 个特征值时达到上界
- 取前 k 个主方向组成 W_k，保留方差等于 $\sum_{j=1}^{k} \lambda_j$
  - 因而这些方向张成的子空间**达到联合最优值**

舍弃方差与平方重构误差

对同一中心化数据、固定 k 与正交投影，前 k 个主方向给出最大保留方差。

- 总方差是各特征方差之和
  - $\|X_c\|_F^2/N = \sum_{j=1}^{d} \lambda_j$
  - 保留全部 d 个主方向时，完整正交基恢复全部中心化数据
- **舍弃方差等于每样本平均平方重构误差**
  - 取前 k 个主方向，总方差减去保留方差得到
  - $\frac{1}{N}\|X_c - \hat{X}_c\|_F^2 = \sum_{j=k+1}^{d} \lambda_j$

3 PCA 的计算示例与新样本变换

沿同一四点例解释主方向、得分、重构及新样本使用

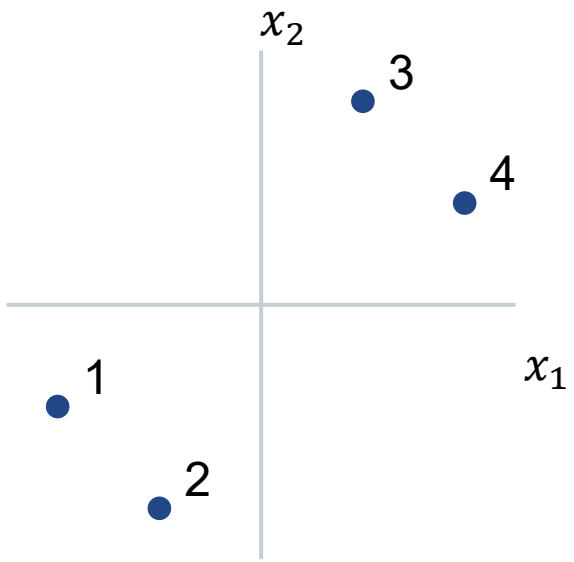
二维中心化样本与降维目标

四个二维训练样本按行排列，各列均值为 0：

$$X_c = \begin{bmatrix} -2 & -1 \\ -1 & -2 \\ 1 & 2 \\ 2 & 1 \end{bmatrix}$$

降维目标是用一个主成分得分表示每个二维样本。

两个特征大体同向变化，候选第一主方向接近直线 $x_2 = x_1$。



编号对应数据矩阵的行

二维算例的协方差矩阵

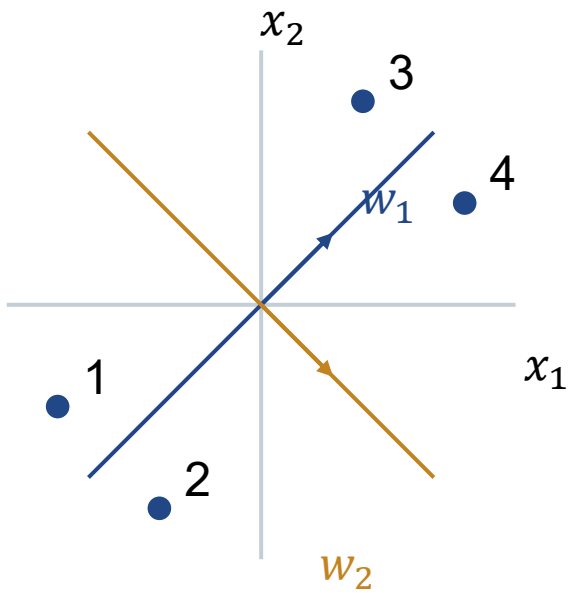
$$S = \frac{1}{4} X_c^T X_c = \begin{bmatrix} 2.5 & 2 \\ 2 & 2.5 \end{bmatrix}$$

- 两个对角元素都是 2.5，表示两个特征各自的方差
- 非对角元素为 2，表示两个特征的协方差为正，倾向同向变化
- 两个特征方差相同且协方差为正，共同支持“第一主方向沿正对角线”的预判

二维算例的主方向与解释方差

由 $Sw = \lambda w$ 求得特征值与单位主方向。

- 协方差特征值给出各方向的**解释方差**
  - $\lambda_1 = 4.5, \lambda_2 = 0.5$
- 对应的单位特征向量给出**主方向**
  - $w_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}, w_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}$
- 第一主方向对应的解释方差比，即该方向的解释方差占总方差的比例
  - $\frac{4.5}{4.5+0.5} = 90\%$

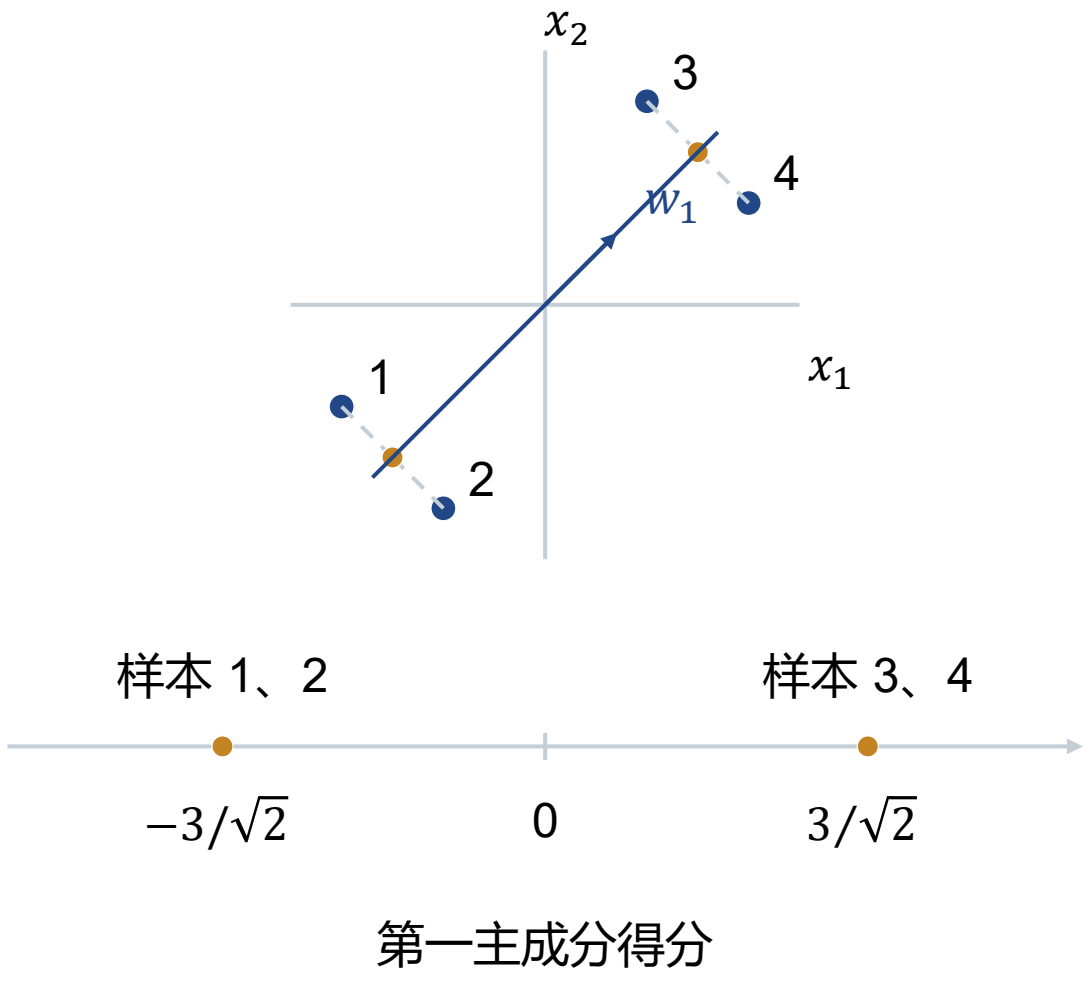


$\lambda_1 = 4.5, \lambda_2 = 0.5$

一维主成分得分

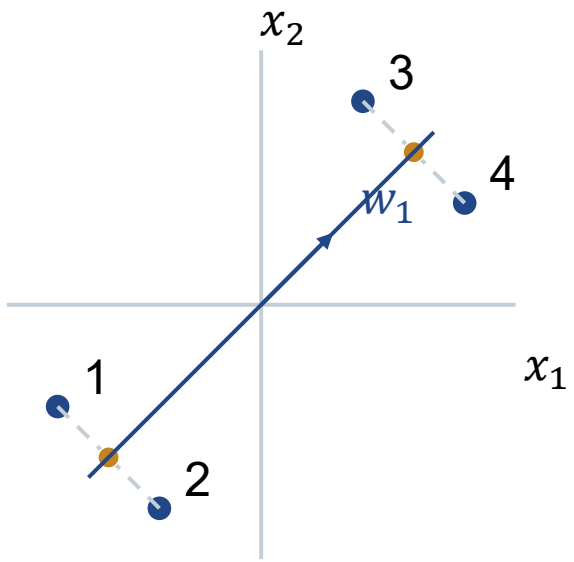
$$z = X_c w_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} -3 \\ -3 \\ 3 \\ 3 \end{bmatrix}$$

- 按当前选定的 w_1 正向，前两个样本的得分为负，后两个样本的得分为正
- 一维主成分得分保留了两个特征共同增减的主要变化
- 舍弃第二主方向后，两个不同样本可能得到相同的一维得分



主成分得分相同的样本与被舍弃的差异

- 各对样本在第一主方向上的得分相同，在第二主方向上的得分不同
  - 样本 1、2 的第一主成分得分都为 $-3/\sqrt{2}$
  - 样本 3、4 的第一主成分得分都为 $3/\sqrt{2}$
- 只用第一主成分得分重构时，每对样本重构到同一点
  - 原先沿第二主方向的差异被舍弃



蓝点：样本　橙点：重构位置

灰线：正交残差

保留一维时的样本重构与平方误差

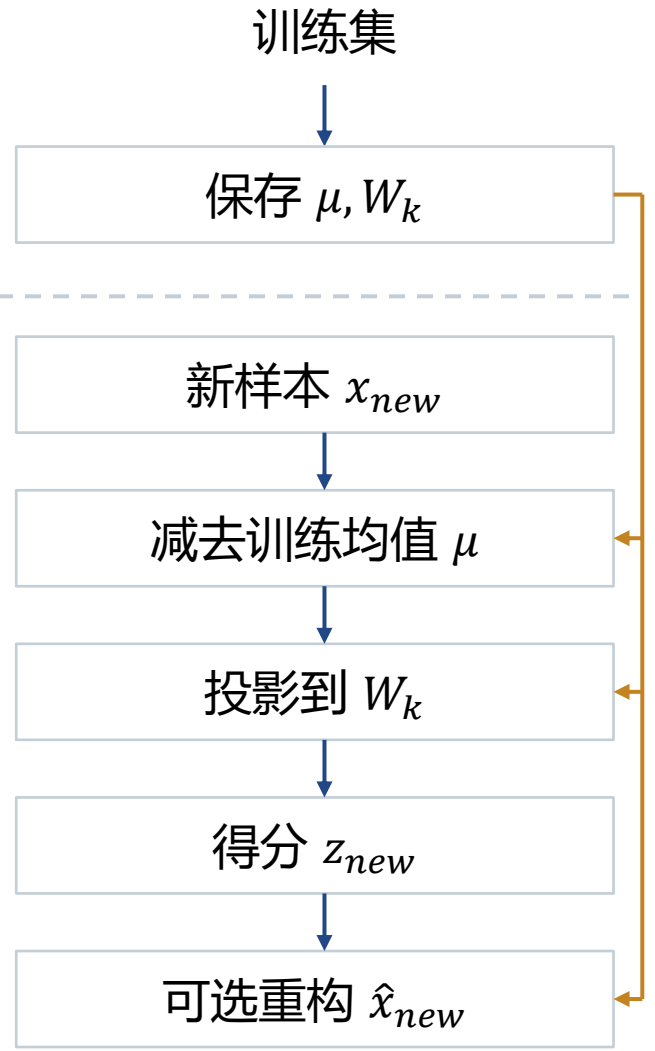
$$\hat{X}_c = zw_1^T = \begin{bmatrix} -1.5 & -1.5 \\ -1.5 & -1.5 \\ 1.5 & 1.5 \\ 1.5 & 1.5 \end{bmatrix}$$

- 重构点都落在直线 $x_2 = x_1$ 上，与上一页看到的两个橙色重构点一致
- 原样本在第二主方向 w_2 上的差异被舍弃
- 保留第一主方向后，每样本平均平方重构误差为
  - $\frac{1}{N}\|X_c - \hat{X}_c\|_F^2 = \lambda_2 = 0.5$
  - 这一误差先计算各样本二维残差的平方长度，再对四个样本求平均

新样本变换与训练参数复用

本例训练均值 $\mu = (0,0)^T$，保留 $k = 1$ 维，方向矩阵 $W_1 = w_1$。一般 k 维变换沿用相同的参数复用关系。

1. 训练阶段保存均值 μ 与前 k 个主方向 W_k
   - 参数只从训练集估计；测试集参与估计会将测试信息带入训练，造成数据泄漏
2. 新样本复用训练参数，形成低维坐标
   - $z_{new} = (x_{new} - \mu)^T W_k$
   - $x_{new} \in \mathbb{R}^d$ 是样本列向量，$z_{new} \in \mathbb{R}^{1\times k}$ 是得分行向量
   - 均值确定原点，主方向确定坐标轴；复用两者使新样本与训练样本采用**同一套坐标**
3. 只中心化时，可按保留得分重构新样本
   - $\hat{x}_{new} = \mu + W_k z_{new}^T$



4 PCA 的 SVD 求解、特征尺度与保留维度

用SVD计算PCA，区分尺度影响与保留维度

中心化数据的 SVD 求解接口

$X_c w$ 将特征方向 w 映为 N 个样本的投影得分。中心化数据的 SVD 因子分别对应特征空间与样本空间。

- 对 $X_c \in \mathbb{R}^{N\times d}$，完整奇异值分解（Singular Value Decomposition，SVD）为
  - $X_c = U\Sigma V^T$
  - U, V 均为正交矩阵
- $V \in \mathbb{R}^{d\times d}$ 的列是**特征空间中的右奇异向量**
  - 每列有 d 个坐标，对应输入特征；$V^T V = I$，单位矩阵尺寸为 $d\times d$
- $U \in \mathbb{R}^{N\times N}$ 的列是**样本空间中的左奇异向量**
  - 每列有 N 个坐标，对应训练样本；$U^T U = I$，单位矩阵尺寸为 $N\times N$
- $\Sigma \in \mathbb{R}^{N\times d}$ 可为矩形，对角项为奇异值
  - 按 $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ 降序排列

SVD 给出的主方向、解释方差与主成分得分

- 利用 $U^TU = I$，SVD 给出协方差矩阵的正交特征分解
  - $S = \frac{1}{N}X_c^TX_c = V\frac{\Sigma^T\Sigma}{N}V^T$
  - $\Sigma^T\Sigma \in \mathbb{R}^{d\times d}$ 为对角矩阵，对角项是对应奇异值的平方
- V 的列向量给出**主方向**，奇异值平方给出**解释方差**
  - 按奇异值降序，$\lambda_i = \sigma_i^2/N$；矩形 Σ 未覆盖的特征维度对应零解释方差
- 利用 $V^TV = I$，得到全部训练样本的主成分得分
  - $X_cV = U\Sigma$
  - 非零奇异值对应列为 $X_cv_i = \sigma_iu_i$：单位左奇异向量乘以奇异值，得到实际得分；零解释方差方向的得分为零
- 二维例中 V 的两列可取为 w_1, w_2
  - $\sigma_1 = \sqrt{4\lambda_1} = \sqrt{18}$，$\sigma_2 = \sqrt{4\lambda_2} = \sqrt{2}$

中心化与标准化的区别

- 中心化减去训练均值，保留特征围绕均值的方差
  - 长度单位从米改为厘米，数值乘以 100，方差乘以 100^2；中心化仍保留这种单位影响
  - 逐特征改变尺度会改变协方差矩阵，主方向与解释方差谱也可能改变
- **标准化**用各特征自身的标准差衡量变化
  - 中心化后，逐特征除以训练标准差，消除原始量纲的影响
  - 按本讲方差定义计算标准差后，非恒定特征的训练方差为 1；标准差为 0 的特征须单独处理，不能直接作除数
- 是否标准化取决于**特征单位与分析目标**
  - 原始单位和方差是否有直接意义，是否希望原始方差较大的特征对准则贡献更多
  - 改变预处理尺度后，须重新拟合并读取新的解释方差谱

标准化后的 PCA 变换与重构

采用标准化时，解释方差与平方重构误差都针对标准化后的输入坐标。

1. 从训练数据估计均值与标准差，与主方向一起保存
2. 新样本复用训练参数：中心化、标准化，再投影到主方向
3. 需要回到原始单位时，在标准化坐标中重构，再恢复原始尺度与均值
   - 逐特征乘回训练标准差，再加回训练均值
   - 两个坐标系的尺度不同，标准化坐标的平方误差不能直接作为原始单位下的平方误差

解释方差比与平方重构误差

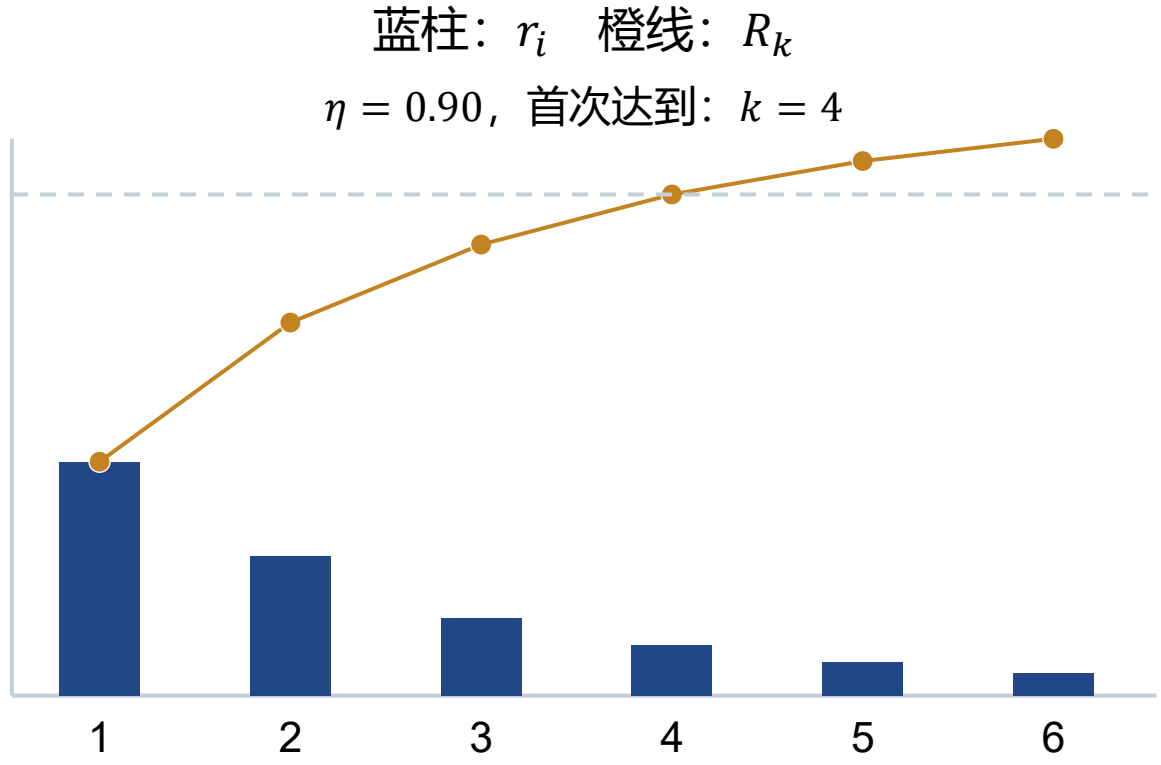
保留维度 k 是保留的主方向数量。

- 总方差非零时，解释方差比与累计解释方差比分别为
  - $r_i = \frac{\lambda_i}{\sum_j \lambda_j}, R_k = \sum_{i=1}^{k} r_i$
  - r_i 对应方向 i，R_k 对应前 k 个方向
- 对同一输入尺度，保留前 k 个主方向时，方差—重构关系给出
  - R_k 是保留方差占总方差的比例
  - $1 - R_k$ 是每样本平均平方重构误差占总方差的比例
  - 总方差为零时，各特征在训练样本中没有变化，比例无定义
- 二维例保留一维时，$R_1 = 90\%$，舍弃比例为 10%
  - 每样本平均平方误差为 0.5，总方差为 5，两者比值与舍弃比例一致

解释方差谱与保留维度

图中为输入尺度已确定的独立六维构造谱，与前述二维算例不同。

- 蓝柱表示各方向的解释方差比，橙线表示累计比例
  - 示意阈值 0.90 首次在 $k = 4$ 达到
- 碎石图按方向顺序显示特征值或比例
  - 曲线由陡转缓的“肘部”表示后续方向新增的方差较少
- **方差比例说明保留了多少变化**
  - 阈值与肘部不构成通用选维推荐；实际选用还需任务证据



构造示意谱，非前述二维算例

5 PCA 的用途、适用边界与总结

按任务与条件理解PCA用途，回收本讲核心关系

PCA 的典型用途与条件

用途	PCA 操作与输出	使用条件
可视化	投影到前两个或三个主方向，得到低维主成分得分	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 个主成分得分表示样本，可选重构	检查保留维度与平方重构误差
降噪	舍弃解释方差较小的主方向，再由保留的主成分得分重构	被舍弃的主方向主要承载扰动时才有降噪依据
模型预处理	用训练所得主方向变换输入特征	用验证集判断是否改善下游任务

PCA 的方法与任务边界

- PCA 用固定主方向张成的**线性子空间**表示数据
  - 沿曲线或曲面变化的结构可能无法由低维线性子空间充分表示
- 协方差可能受到远离均值的样本的显著影响，并改变主方向
  - 协方差由中心化特征的乘积求平均，对角项包含平方偏离，远离均值的样本可能产生较大贡献
  - 标准化调整特征尺度，也不自动消除异常值的影响
- 保留较大方差不保证保留**判别信息**
  - 方向选择只使用输入数据的投影方差，没有利用标签
  - 即使某个小方差方向能区分类别，它仍可能在截断时被舍弃
- 主成分得分的含义与下游效果需要**任务解释与验证**
  - 主成分得分是预处理后特征的线性组合，各特征的权重来自方差准则
  - 解释方差说明保留了多少变化

方向与子空间的非唯一性

- w 与 $-w$ 表示同一条轴
  - 改变方向符号会改变得分符号，投影位置和重构保持不变
- 重复特征值对应的特征子空间中，可以选择不同的正交基
  - 同一子空间的不同基具有相同投影
- 若保留维度只截取重复特征值块的一部分，最优保留子空间也可能不唯一
  - 不同最优子空间可产生不同投影与重构，但达到相同的最优目标值

方向或基的表示不同，本身不足以判定计算错误。

本讲总结：PCA 的核心关系与使用条件

- PCA 以中心化数据的主要正交方向形成低维表示，服务于低维观察或允许近似的紧凑表示
- **中心化与方差准则**分别明确所分析的变化及方向选择依据
  - 中心化分析围绕训练均值的变化；最大投影方差准则给出主方向
- **方向、得分与重构**对应坐标轴、样本坐标与保留分量
  - 正交残差是被舍弃的部分；对同一中心化数据、固定 k、正交投影与平方误差，保留最多方差对应最小平方重构误差
- 中心化数据的 SVD 提供另一条求解路线；新样本复用训练均值、主方向与可选训练尺度
- **解释方差与任务效果分别评价**
  - 解释方差说明保留了多少变化；特征尺度、用途与任务证据限定低维表示的使用