**主成分计算**

**第一主成分**：最大化变量线性组合的方差

A black and white math symbols

Description automatically generated with medium confidence**优化方法**：通过求解样本协方差矩阵 V=X′X 的特征值和特征向量。

**几何解释**

**第一主成分方向**：是数据变化最大的方向。

**主成分得分**：将原始数据投影到主成分方向上得到的值。

**进一步的主成分**

**第二主成分**：与第一主成分正交，解释剩余方差。

**正交性**：所有主成分方向彼此正交。

**特征值关系**：主成分方差与样本协方差矩阵的特征值成正比。

**方差解释比例（PVE）**

定义：每个主成分解释的方差比例为

A mathematical equation with a number of letters

Description automatically generated with medium confidencePVE 之和为 1。

**累计解释比例**：通过累计 PVE 判断主成分数量。

**主成分数量选择**

**方法**：无明确答案，常用“碎石图”（Scree Plot），选择拐点处的主成分数。

**变量相关性和图示**

**相关性圆图（Correlation Circle）**：变量与主成分的相关性用图示表示。

**得分图（Score Plot）**：主成分得分的几何投影。

**两种主要的聚类方法**

**K-means 聚类特点**：需要预先指定簇的数量 KK。

**原理**：通过最小化组内差异（Within-Cluster Variation, WCV）来优化分组。

**算法步骤**：

随机分配样本到 KK 个簇中。

计算每个簇的中心（Centroid），并重新分配样本到离中心最近的簇。

迭代直到分配不再变化。

**注意点**：

组内差异通常用欧几里得距离衡量。

算法每步减少目标函数值，但可能陷入局部最优。

**层次聚类（Hierarchical）**

**特点**：不需要预先指定簇的数量 K。

**方法**：通常采用自底向上的聚合（Agglomerative Clustering）。

**目标**：构建树状结构（树状图，Dendrogram），展示观测值之间的层次关系。

**常用的相似性/距离度量**

欧几里得距离（Euclidean）。

编辑距离（Edit）或汉明距离（Hamming）等特定数据类型的度量。

相关性距离（Correlation-Based）：基于特征相关性衡量观测值的相似性。

**层次聚类的链接方式（Linkage）**

**单链接（Single）**：记录两组观测值间的最小距离。

**完全链接（Complete）**：记录两组观测值间的最大距离。

**平均链接（Average）**：记录两组观测值间的平均距离。

**质心链接（Centroid）**：基于两组中心点的距离。

完全链接和平均链接通常更受欢迎。

单链接容易生成不平衡的树状图。

**实践中的问题**

**特征标准化**：变量应标准化为均值为 0、标准差为 1 的形式。

**距离度量选择**：根据数据类型选择合适的相似性或距离度量。

**链接方式选择**：选择适合问题的链接方式。

**簇数量选择**：无统一方法，通常依赖数据可视化或领域知识。

**特征选择**：应谨慎选择驱动聚类的关键特征。

**降维的意义**

**降维目标**：

简化模型，提高解释性。

改善预测性能，尤其在高维数据中。

**降维方法**：

**变量选择**（subset selection）：枚举所有可能模型，选择最佳模型。

**系数收缩**（coefficient shrinkage）：如Lasso和Ridge回归，通过收缩系数实现降维。

**主成分回归（PCR）** 和 **偏最小二乘回归（PLS）**：通过线性变换减少变量数。

**主成分回归（PCR）**

**核心思想**：通过主成分分析（PCA）将原始变量转化为少数几个主成分，并使用这些主成分进行回归。

**步骤**：

使用PCA提取前 q 个主成分 Z1,…,Zq。

对主成分进行线性回归。

**优点**：

降低多重共线性问题。

提升预测性能。

**缺点**：

主成分的方向未受到响应变量 Y 的监督。

不进行特征选择。

主成分缺乏可解释性。

**适用场景**：

前几个主成分能捕捉大部分变量的变化并反映与响应变量的关系时表现较好。

**偏最小二乘回归（PLS）**

**核心思想**：与PCR相似，但PLS利用 Y 的信息来监督主成分的提取。

**步骤**：

标准化预测变量 X。

根据 Y 与每个 Xj 的相关性，计算主成分方向。

每次计算一个主成分，并对剩余变量正交化，直到提取 M 个主成分。

**优点**：

在降维的同时，保留与响应变量 Y 相关的特征。

**缺点**：

仍可能缺乏特征选择的能力。

**调参方法**

**主成分个数 q/M 的选择**：

使用总方差解释率：当新增主成分的解释率增量较小时停止。

交叉验证：评估模型预测性能以确定最优 q/M。

**比较与应用**

**PCR 与 PLS 的比较**：

PCR 是无监督降维，不考虑 Y 信息。

PLS 是监督降维，结合了 YY 的相关性。

**局限性**：

与惩罚回归方法（如Lasso、Ridge）相比，这些降维方法在实际应用中不如后者流行。

特别是在特征选择和模型解释方面，这些方法的优势有限。

**PCA 的重要性**：在无监督学习中仍然广泛应用。