Multivariate Outlier Detection

基于鲁棒统计（MCD）、结构降维（Robust PCA）及机器学习（Isolation Forest）的多种多元异常值判定方法，以提高异常判定结果的稳健性与可信度。

1. 不同方法关注异常的不同方面：

1.1 MCD：“这个观测点在统计意义上是否偏离总体中心？”；

1.2 PCA / Robust PCA：“这个观测点是否破坏了数据的主要结构模式？”

1.3 Isolation Forest：“这个观测点在样本空间中是否容易被孤立？”

2. 在实际数据分析与教学中：

2.1 只用 MCD：可能漏掉“结构异常但不远离中心”的点；

2.2 只用 PCA：阈值解释困难，统计意义较弱；

2.3 只用 Isolation Forest：缺乏可检验性，难以学术表述；

多方法一致判定的异常点，其可信度通常更高，这是当前学术界的普遍共识。

3. 在多元异常值判定的教学与应用中，推荐采用以下分层策略：

3.1 MCD 作为统计基准方法，提供权威且可解释的异常判定；

3.2 Robust PCA 用于揭示异常点的结构来源，并辅助可视化解释；

3.3 Isolation Forest 作为补充探索方法，用于发现潜在的非线性或复杂异常。

维度	MCD（Robust Mahalanobis）	PCA / Robust PCA	Isolation Forest
方法学本质	鲁棒协方差估计的统计方法	基于低维结构的几何方法	基于随机划分的机器学习方法
异常的定义	偏离总体中心的统计异常	偏离主成分结构的几何异常	在样本空间中易被孤立的异常
对数据分布的要求	近似多元正态（弱）	无严格分布假设	完全无分布假设
判定依据	鲁棒 MD^2 与 χ^2 分位数	Score Distance / Orthogonal Distance	Isolation score（相对值）
是否有明确阈值	是（χ^2 cutoff）	部分（经验或文献建议）	否（排序或比例）
是否参数法	是	半参数	否
对异常值的鲁棒性	高（设计初衷）	高（使用 Robust PCA 时）	高
对高维数据的适应性	一般（p 不能接近 n）	较好	很好
是否需要标准化	必须	必须	强烈建议
对变量相关性的依赖	强（协方差结构）	强（相关性越强越有效）	弱
可解释性	很强（统计意义清晰）	强（结构与方向可解释）	较弱（算法黑箱）
结果稳定性	高	中等（依赖 PC 数）	较低（随机性）
主要优势	权威、可检验、易解释	可视化好、揭示结构异常	捕捉复杂非线性异常
主要局限	不适合极高维或多峰分布	阈值不统一、解释需经验	无显著性、解释困难
典型误用风险	忽略标准化或样本量不足	将 PCA 当“异常值专用工具”	将 score 当统计显著性
教学使用建议	作为基准与主方法	作为结构补充方法	作为探索与验证方法