arXiv:2305.05371v1 [stat.ME] 9 May 2023

用于局部异常值检测的空间平滑稳健协方差估计

Patricia Puchhammer维也纳工业大学
patricia.puchhammer@tuwien.ac.at

Peter Filzmoser维也纳工业大学
peter.filzmoser@tuwien.ac.at

2023年5月10日

摘要

大多数多变量异常值检测方法都忽略了观测值的空间依赖性，而这种依赖性存在于来自各类应用领域的许多真实数据集中。本文提出了一种新的异常值检测方法，该方法考虑了随观测值的空间邻域而（连续）变化的协方差结构。因此，所采用的估计量在统一的全局协方差估计与针对各个邻域估计的局部协方差之间取得了折衷。本文介绍了该估计量的理论性质，尤其是与稳健性性质相关的性质，并提出了一种高效的计算算法。本文基于模拟数据和奥地利气象站记录的数据集，对该方法的性能进行了评估和比较。

关键词 空间异常值检测 · MRCD · 成对马哈拉诺比斯距离

1 引言

识别多变量离群值可能是多变量数据分析中最重要的任务之一。为了使后续分析更加可靠而需要寻找离群值，或是对离群值本身的直接关注，推动了众多多变量异常值检测方法的发展。识别出的离群值应在一定程度上偏离数据主体的主要趋势或结构，因此也称为“全局异常值”（Filzmoser 等人，2013 年）。相比之下，“局部异常值”指的是一种可获取某类邻域相关附加信息的情形，例如观测值的空间坐标。此时，局部异常值是指与其空间邻居的多变量测量值明显不同的观测值，这些差异表明存在值得关注的局部异常，因此有必要进行进一步分析。尽管如此，这些值本身仍可能处于数据集的常规范围内，因此从全局角度来看，该观测值并不属于离群值。

现有的多变量局部异常值检测统计方法通常基于距离度量和邻域结构。邻域 a 定义为观测索引集合的一个子集，即 $\{1,\ldots,n\}$。对于 $i \in \{1,\ldots,n\}$，p变量观测值 $\boldsymbol{x}_i$ 当且仅当满足 $i \in a$ 时，被定义为处于邻域 a 中。判断某个观测值 $\boldsymbol{x}$ 是否处于某个邻域中，通常基于其空间坐标 s（$\boldsymbol{x}$）。构建空间邻域的一种方法，是为每个点 $\boldsymbol{x}$取其空间k近邻，其中 $k \in \mathbb{N}$。对于固定的 $\boldsymbol{x}$，点 $\boldsymbol{y}$ 到另一点的空间距离定义为欧几里得距离

$$d_{\boldsymbol{x}}(\boldsymbol{y}) = \|s(\boldsymbol{x}) - s(\boldsymbol{y})\| = \left[(s(\boldsymbol{x}) - s(\boldsymbol{y}))'\,(s(\boldsymbol{x}) - s(\boldsymbol{y}))\right]^{1/2}.$$

空间　　邻域 a_k（$\boldsymbol{x}$）随后可以定义为包含 k 个空间最近观测值的集合，

$$a_k(\boldsymbol{x}) = \{\boldsymbol{x}_j :\ d_{\boldsymbol{x}}(\boldsymbol{x}_j) \le d_{\boldsymbol{x}(k)}\}, \tag{1}$$

其中，$d_{\boldsymbol{x}(1)} \le d_{\boldsymbol{x}(2)} \le \ldots \le d_{\boldsymbol{x}(k)} \le d_{\boldsymbol{x}(n)}$是到所有观测值的排序后距离 $\boldsymbol{x}_i$，$i = 1\ldots\ldots n$。在局部异常值检测中，常用于评估异常程度的距离度量是成对马氏距离（MD）。对于具有协方差 $\boldsymbol{\Sigma}_a$ 的邻域 a 以及邻域 a中的观测值 $\boldsymbol{x}_i$，成对马氏距离定义为

$$\mathrm{MD}_{\boldsymbol{\Sigma}_a}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left[(\boldsymbol{x}_i - \boldsymbol{x}_j)'\boldsymbol{\Sigma}_a^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)\right]^{1/2} \quad \text{for all } j \in a.$$

一般而言，MD描述了两个观测值之间的距离，其中，特征空间中的欧几里得距离会根据局部分布特性进行调整。

用于确定 $\mathbf{\Sigma}_a$ 所有邻域的估计方法，是获得良好且可靠结果的关键。必须确保离群观测本身不会影响估计，因为这可能会掩盖离群点，使其无法被检测到。因此，有必要在邻域层面进行稳健协方差估计。最广泛使用的协方差稳健估计量之一是最小协方差行列式（MCD）估计量（Rousseeuw，1984，1985），其需要找出 h 个观测值的子样本（其中 h 固定为观测值总数的一半等），使其样本协方差的行列式最小。随后，将 h 个观测值子集的样本协方差乘以一致性因子（Croux andHaesbroeck，1999），即可得到MCD协方差估计量。为计算该估计量，Rousseeuw and Driessen（1999）提出的fastMCD算法引入了一种迭代集中步骤，即所谓的C步；该步骤可确保目标函数不断下降，直至收敛到一个（局部）最小值，从而使MCD估计量的计算更快，也更受欢迎。随后，通过迭代使用若干个随机初始值并选取最小的局部最小值，来近似求得全局最小值。通过为fastMCD选取少量优质的确定性初始值，Hubert 等人（2012）提出的detMCD算法进一步缩短了运行时间。尽管MCD估计量具有仿射等变性和高度稳健性等出色特性，但其缺点之一是，当 h 个观测值子集的样本协方差矩阵奇异时，该方法便不适用，而这种情况很容易发生。与许多方法一样，计算马氏距离也需要估计得到的协方差矩阵具有正则性。尤其是在只能使用由可能数量较少的观测值构成的局部邻域时，可能无法满足正则性要求，因此无法对局部协方差矩阵求逆。一种解决方案是基于正则化稳健协方差估计量进行局部估计，例如Boudt 等人（2020）最近提出的最小正则化协方差行列式（MRCD）估计量（或者也可采用Fritsch估计量，Fritsch 等人（2012））。MRCD的一项诸多优点之一是，同样可以使用基于C步、经过轻微调整的fastMCD算法。

现有的局部异常值检测方法采用不同的方式定义协方差矩阵，以确保正则性。Filzmoser 等（2013）的方法通过使用基于整个数据集 $\mathbf{\Sigma}_a = \mathbf{\Sigma}$ 计算的 MCD估计量来处理正则性问题，即施加全局协方差结构。他们证明，对于均值为 $\boldsymbol{\mu}$、协方差矩阵为 $\mathbf{\Sigma}$ 的独立同分布高斯随机向量 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$，在给定 $\boldsymbol{x}_i$ 的条件下，$j = 1, \ldots, n$ 对应的成对平方马氏距离 $\mathrm{MD}_{\mathbf{\Sigma}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 的条件分布为自由度为 p、非中心参数为马氏距离 $\mathrm{MD}^2_{\mathbf{\Sigma}}(\boldsymbol{x}_i, \boldsymbol{\mu})$ 的非中心卡方分布。他们不采用固定的成对马氏距离截断值，而是使用不同的复杂可视化方法。通过绘制基于成对马氏距离的孤立度以及非中心卡方分布的分位数，可以发现可疑且高度孤立的观测值，并进行更详细的分析。与 Filzmoser 等（2013）截然相反，Ernst 和 Haesbroeck（2016）的方法通过分别考虑每个点的 k近邻（kNN）邻域，进行非常局部的协方差估计。为解决正则性问题，他们针对每个 kNN 邻域分别采用正则化协方差估计（最初使用 Fritsch 等人（2012）提出的估计量；关于 MRCD，也可参见 Bellino 等人（2019）所作的改进）。此外，他们引入了“次近距离”的概念，该距离同样基于马氏距离，并将所有次近距离的 Hubert 和 Vandervieren（2008）调整箱线图的上围栏用作检测离群点的非参数截断值。

由于寻找局部异常值并不需要使用马氏距离（MD）概念，因此不妨简要探讨一下机器学习技术。机器学习中用于检测多变量局部离群值的最著名方法之一，是由Breunig 等人（2000）提出的局部离群因子（LOF）。最初，局部性指的是特征空间中的多变量取值和欧几里得距离，但这种方法也可以直接应用于空间局部离群检测。在 Schubert 等人（2012）中，作者讨论了这种应用以及其他基于 LOF 的方法。有趣的是，LOF 及其变体同样基于距离和邻域的概念。
现有方法在多个方面存在尚未得到妥善解决的不足。Filzmoser 等人（2013 年）的方法具有较强的全局性，能够稳健可靠地估计协方差。然而，估计所得的协方差是否适用、能否代表局部层面的协方差结构，仍有些值得商榷。为解决估计中缺乏局部性的问题，Ernst 和 Haesbroeck（2016）采用了非常局部的方法，针对每个观测值分别重新计算协方差矩阵。尽管局部性有所增强，但该方法没有考虑到协方差矩阵不太可能在相邻邻域之间发生突变。即使局部协方差结构确实发生突变，估计参数的数量也极其庞大，而且其估计依据的观测值相对较少。对局部协方差矩阵进行更具全局性的估计，或许会更加稳定可靠，也可能避免重复计算。迄今为止，协方差估计的局部性似乎只有两个极端可供选择。

我们通过提供一种基于 MRCD 的协方差估计器，弥合了完全全局方法与完全局部方法之间的差距。该估计器一方面解决了局部性缺失的问题，另一方面解决了空间平滑度不足的问题。通过

设置空间平滑程度和邻域大小，我们得到了两种检测方法的推广，旨在基于新的局部协方差估计实现良好的异常值检测性能。此外，当数据集具有额外的子结构时，该协方差估计也可以视为 MRCD 的推广。

本文的组织结构如下：第 2 节介绍一种新的协方差估计量，推导其性质以及原始 MRCD 的性质，并建立检测局部异常值的方法论。第 3 节讨论一种算法以及广义 C 步的推导。第 4 节给出有关运行时间、收敛性和异常值检测的模拟结果；第 5 节则使用新开发的局部离群检测方法分析一个真实数据集。最后，在结论部分呈现并总结主要研究结果。

2 方法论

2.1 空间平滑 MRCD 估计量

假设 $i = 1, \ldots, n$ 的 p 维观测值 $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ 按行排列在 $n \times p$ 矩阵 $\boldsymbol{X}$ 中，其中 $n > 2p$。此外，每个观测值都有可用的空间信息，例如空间坐标，并被分配到 N 个邻域之一，这些邻域由大小为 $n_1, \ldots, n_N$ 的索引集 $a_1, \ldots, a_N$ 定义。

所提方法的目标是为每个邻域获取局部协方差估计，以便计算成对马氏距离，并在一定程度上使相邻邻域之间平滑。由于 MCD估计量至少需要 $n > 2p$ 才能提供正则协方差矩阵，这对较小的邻域尤其构成了严重限制，因此我们着重研究其正则化扩展，即 MRCD估计量。与 MCD 方法中最小化样本协方差矩阵的行列式不同，这里的优化目标是样本协方差与一个对称正定矩阵（即所谓的目标矩阵 $\boldsymbol{T}$）的凸组合的行列式。Boudt 等人(2020) 提出一种基于协方差矩阵条件数的数据驱动方法，用于设定凸组合中的正则化程度 ρ。对于目标矩阵 $\boldsymbol{T}$，选择稳健的正则协方差矩阵是合理的，例如基于单变量稳健尺度估计的对角矩阵。

接下来，我们将 MRCD 估计量的思路应用于局部平滑协方差估计的情形。令 $\mathcal{H} = (H_1, \ldots\ldots, H_N)$ 为定义邻域的索引集 $a_1, \ldots\ldots, a_N$ 的子集。每个子集H_i的大小为 $h_i = |H_i| = \lceil \alpha n_i \rceil$，其中 $i = 1, \ldots\ldots, N$，α 选自区间 $[0.5, 1]$，而 $\lceil \cdot \rceil$ 为向上取整函数，将数值向上舍入到下一个整数。α 的值越小，对异常值的稳健性越强；也可以针对每个邻域单独调整该值。子集 H_i 的观测值写作矩阵 $\boldsymbol{X}_{H_i}$，其维度为 $h_i \times p$。对于 $i = 1, \ldots\ldots, N$，邻域特定的基于 MRCD 的协方差矩阵 $\boldsymbol{K}_i(\mathcal{H})$ 定义为

$$\boldsymbol{K}_i(\mathcal{H}) = \rho_i \boldsymbol{T} + (1 - \rho_i) c_\alpha Cov(\boldsymbol{X}_{H_i}), \tag{2}$$

其中 $Cov(\boldsymbol{Y})$是 $\boldsymbol{Y}$ 的样本协方差矩阵，而 c_α 是比例 α 对应的一致性因子（参见 Crouxand Haesbroeck, 1999）。正则化参数 ρ_i 针对每个邻域单独设置；如果估计协方差矩阵已经可逆，也可以将其设为零。最后，由于我们希望对协方差矩阵进行平滑，选择各邻域特有的目标矩阵似乎有悖直觉，而且还会需要更多参数估计。因此，我们假设存在一个全局目标矩阵 $\boldsymbol{T}$，该矩阵是基于完整数据集 $\boldsymbol{X}$估计得到的稳健正则协方差矩阵。由于我们假设 $n > 2p$，因此提议使用 $\boldsymbol{X}$ 的 MCD估计量作为目标矩阵。

我们希望找出 $\mathcal{H}$ 中的子集组合，使目标函数达到最小值

$$f(\mathcal{H}) = \sum_{i=1}^{N} \det \left((1 - \lambda) \boldsymbol{K}_i(\mathcal{H}) + \lambda \sum_{j=1, j \neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}) \right). \tag{3}$$

调节参数 $\lambda \in [0, 1]$ 用于平衡单个局部邻域与其余邻域对协方差估计的影响。当 $\lambda = 0$时，完全不存在空间影响，这相当于在使用全局目标矩阵的情况下，分别对每个邻域进行 MRCD 估计。另一个极端是$\lambda = 1$，此时特定邻域中的协方差矩阵是周围邻域的协方差估计值的平均值，不会加入该邻域自身的局部信息。此外，行列式中的第二部分也可以解释为惩罚项。由于行列式需要最小化，来自 a_i 且与具有正权重的邻域中的观测值主要趋势 ω_{ij} 相符的观测值，在 λ 增大时更有可能被纳入最优 H 集。权重 ω_{ij} 应为非负数，且我们令 $\omega_{ii} = 0$。权重向量 $\boldsymbol{\omega}_i = (\omega_{i1}, \ldots\ldots, \omega_{iN})'$ 中的元素表示其他邻域的协方差估计值对该

第 i 个邻域的协方差估计的相对影响。此外，每个权重向量的元素之和必须为 1，即对所有 $i = 1, \ldots\ldots, N$，均有 $\sum_{j=1}^{N} \omega_{ij} = 1$。所有这些权重都需要预先指定，例如根据地理距离的倒数确定，并作为各行汇集在加权矩阵 $\boldsymbol{W} \in \mathbb{R}^{N \times N}$中。

请注意，对于目标函数 (3)，存在全局最小值 $\mathcal{H}^* = (H_i^*)_{i=1,\ldots,N}$，因为其定义域由有限个子集组合构成。对于该全局最小值，每个邻域的估计协方差矩阵a_i 是

$$\hat{\boldsymbol{\Sigma}}_{SSM,i} = (1 - \lambda)\boldsymbol{K}_i(\mathcal{H}^*) + \lambda \sum_{j=1, j \neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}^*), \tag{4}$$

且位置估计 $\hat{}\boldsymbol{\mu}_{SSM,\,i}$ 是选定观测值 $\boldsymbol{X}_{H_i^*}$ 的样本均值。我们将这些估计量称为空间平滑MRCD（ssMRCD）位置与协方差估计量。

尽管邻域结构和 λ 的取值通常取决于当前的数据，但对于 $\boldsymbol{W}$，仍有一些合理且自然的选择。如果邻域结构没有其他特殊含义，仅用于将空间划分为若干子集，那么基于距离倒数的权重矩阵（第4节和第5节中也采用了这种矩阵）可能是一个不错的选择。其他选择还包括二元矩阵：如果两个邻域共享一条边界，则对应元素为1，否则为0，并对各行进行适当的缩放。此外，正则性参数可以采用默认设置。对于邻域 a_i，我们建议将正则化参数 ρ_i 设为MRCD算法中提出的数据驱动值，具体方法是将该邻域视为其自身的数据集，如Boudt 等人(2020所述。

2.2 理论性质

下文将证明，与原始MRCD估计量不同，本文提出的空间平滑 MRCD 估计量具有仿射等变性，并推导其崩溃点。对于数据集 $\boldsymbol{X} \in \mathbb{R}^{n \times p}$，若对于所有邻域，位置和协方差估计量满足：对任意非奇异矩阵 $\boldsymbol{A} \in \mathbb{R}^{p \times p}$、任意向量 $\boldsymbol{b} \in \mathbb{R}^p$ 以及所有 $i = 1$，2、……、N，均有

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{SSM,i}(\boldsymbol{X}\boldsymbol{A}' + \mathbf{1}_n \boldsymbol{b}') &= \hat{\boldsymbol{\mu}}_{SSM,i}(\boldsymbol{X})\boldsymbol{A}' + \boldsymbol{b}', \\ \hat{\boldsymbol{\Sigma}}_{SSM,i}(\boldsymbol{X}\boldsymbol{A}' + \mathbf{1}_n \boldsymbol{b}') &= \boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{SSM,i}(\boldsymbol{X})\boldsymbol{A}'. \end{aligned} \tag{5}$$

定理1（仿射等变性）。设 $\boldsymbol{T}$ 是数据集 $\boldsymbol{X}$ 的任意稳健、正则且仿射等变的协方差估计量，记为 $\boldsymbol{T}$（$\boldsymbol{X}$）。则以目标矩阵 $\boldsymbol{T}$（$\boldsymbol{X}$）为基础的位置和协方差空间平滑 *MRCD* 估计量具有仿射等变性。

证明。证明见附录。 □

该定理假设估计量 $\boldsymbol{T}(\boldsymbol{X})$ 正则、稳健且具有仿射等变性。对于完整数据集 $\boldsymbol{X}$，当 $n > 2p$ 时，最小协方差行列式可以满足这些性质。然而，在观测值不足或需要额外正则化的情况下，这并非易事。由于 MRCD 正是为这种情况提供补救方法，原始 MRCD 的目标矩阵很可能不满足定理中陈述的假设。因此，MRCD 通常不具有仿射等变性。

稳健估计量的另一项重要性质是有限样本崩溃点，它定义为使估计量失效所需替换的数据点的最小比例。在考察空间平滑MRCD之前，我们必须推导原始MRCD（即未经预先缩放观测值的MRCD）的崩溃点，以下称为原始*MRCD*。位置估计量$\hat{}\boldsymbol{\mu}_n$ 的崩溃点正式定义为

$$\epsilon_n^*(\hat{\boldsymbol{\mu}}_n; \boldsymbol{X}_n) = \frac{1}{n} \min\{m : \sup ||\hat{\boldsymbol{\mu}}_n(\boldsymbol{X}_{n,m}) - \hat{\boldsymbol{\mu}}_n(\boldsymbol{X}_n)|| = +\infty\},$$

其中$\boldsymbol{X}_{n,\,m}$ 是数据 矩阵 $\boldsymbol{X}_n$，其中 m个观测值被替换为任意值（Maronna等人，2006 ）.

$\hat{}$ 对于协方差估计 $\boldsymbol{\Sigma}_n$，有限样本崩溃点定义为

$$\epsilon_n^*(\hat{\boldsymbol{\Sigma}}_n; \boldsymbol{X}_n) = \frac{1}{n} \min\{m : \sup \max_j |\ln(\lambda_j(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_{n,m}))) - \ln(\lambda_j(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_n)))| = +\infty\},$$

其中，$\lambda_1(\boldsymbol{\Sigma}) \ldots, \lambda_p(\boldsymbol{\Sigma})$ 表示矩阵 $\boldsymbol{\Sigma}$ 的特征值，按降序排列。由于特征值已排序，我们只需考虑最大的特征值 $\lambda_1(\boldsymbol{\Sigma}_n(\boldsymbol{X}_{n,m}))$，它在将观测值替换为任意值时可能会爆炸（爆炸崩溃点）；以及最小的特征值 $\lambda_p(\boldsymbol{\Sigma}_n(\boldsymbol{X}_{n,m}))$，它可能变为零（内爆崩溃点），从而导致奇异性（Maronna等人, 2006）。

定理 2. 考虑具有固定 $\rho > 0$、正则且固定的 $\boldsymbol{T} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$ 和数据矩阵$\boldsymbol{X}_n$的原始*MRCD*估计量。则以下结论成立：

a. MRCD 位置估计量$\hat{}\boldsymbol{\mu}_n$ 的有限样本崩溃点为$\min(h, n-h+1)/n$。

b. MRCD 协方差估计量$\hat{}\ \boldsymbol{\Sigma}_n$ 的有限样本爆炸崩溃点为$(n-h+1)/n$。

$\hat{}$*c. MRCD* 协方差估计量 $\boldsymbol{\Sigma}_n$ 的有限样本内爆崩溃点为1。

证明。 证明见附录。 □

关于包含多个估计量的位置与协方差估计量的有限样本崩溃点，我们将有限样本崩溃点 ϵ_n^* 定义为：为使至少一个估计量失效，需要替换同一任意邻域中的点所占的最小比例。对于位置估计量 $\hat{}\boldsymbol{\mu}_{SSM,n,i}$,
$i = 1 \ldots N$,,, 其形式化定义为

$$\epsilon_n^*(\left(\hat{\boldsymbol{\mu}}_{SSM,n,i}\right)_{i=1}^N; \boldsymbol{X}_n) = \min_{i=1,\ldots,N} \frac{1}{n_i} \min\{m : \sup ||\hat{\boldsymbol{\mu}}_{SSM,n,i}(\boldsymbol{X}_{n,m}^i) - \hat{\boldsymbol{\mu}}_{SSM,n,i}(\boldsymbol{X}_n)|| = +\infty\},$$

其中，$\boldsymbol{X}_{n,\ m}^i$ 是矩阵 $\boldsymbol{X}_n$，其中邻域 a_i 的 m个观测值被任意值替换。

定理3. 空间平滑*MRCD*的位置估计量 $\left(\hat{}\boldsymbol{\mu}_{SSM,\ n,\ i}\right)_{i=1}^N$ 的有限样本崩溃点为

$$\epsilon_n^*(\left(\hat{\boldsymbol{\mu}}_{SSM,n,i}\right)_{i=1}^N; \boldsymbol{X}_n) = \min_{i=1,\ldots,N} \min(n_i - h_i + 1, h_i)/n_i.$$

证明. 证明见附录。 □

对于协方差 $\hat{}$估计$\boldsymbol{\Sigma}_{SSM,\ n,\ i}$, $i = 1, \ldots, N$，有限样本崩溃点相应定义为 y,

$$\epsilon_n^*(\left(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}\right)_{i=1}^N; \boldsymbol{X}_n) = \min_{i=1,\ldots,N} \frac{1}{n_i} \min\{m : \sup \max_j |\ln(\lambda_j(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}(\boldsymbol{X}_{n,m}^i))) - \ln(\lambda_j(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}(\boldsymbol{X}_n)))| = +\infty\}.$$

同样，我们可以区分爆炸崩溃点和内爆崩溃点。

定理 4. 给定一个固定且正则的目标矩阵 $\boldsymbol{T}$，空间平滑 *MRCD* 协方差估计量 $\left(\hat{}\boldsymbol{\Sigma}_{SSM,n,i}\right)_{i=1}^N$ 的有限样本内爆崩溃点等于

$$\epsilon_n^*(\left(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}\right)_{i=1}^N; \boldsymbol{X}_n) = 1.$$

有限样本爆炸崩溃点为

$$\epsilon_n^*(\left(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}\right)_{i=1}^N; \boldsymbol{X}_n) = \min_{i=1,\ldots,N}(n_i - h_i + 1)/n_i.$$

证明。 证明见附录。 □

请注意，对于原始 MRCD 和空间平滑 MRCD 的所有崩溃性质，均假设目标矩阵 $\boldsymbol{T}$ 是正则且固定的。在实际应用中，目标矩阵会是某个协方差估计量 $\boldsymbol{T}(\boldsymbol{X})$。因此，应当注意， $\boldsymbol{T}(.)$的内爆和爆炸崩溃点也可能与此相关。

2.3 局部离群点检测

检测离群点的最后一步，是确定如何将邻域 $a_i, i = 1, \ldots, N$对应的空间平滑协方差与成对马氏距离关联起来。

该方法基于 Ernst 和 Haesbroeck (2016)。为了将每个观测值 $\boldsymbol{x}$ 与其局部邻居进行比较，我们需要第二种邻域结构，以提供空间上距离较近的邻居，与用于平滑协方差估计的结构邻域a_i 形成对照。因此，我们选取一些 $k \in \mathbb{N}$ 并计算空间 k 最近邻

邻居， $b_k(\boldsymbol{x})$，见公式 (1)，其中一个典型值可能是 $k = 10$。然而，在应用该方法时，还应考虑数据集的空间结构。

对于每个观测值 $\boldsymbol{x} \in a_i$，次近距离 定义为

$$d(\boldsymbol{x}) = min_{\boldsymbol{y} \in b_k(\boldsymbol{x})} \left[(\boldsymbol{x} - \boldsymbol{y})' \hat{\boldsymbol{\Sigma}}^{-1}_{SSM,i} (\boldsymbol{x} - \boldsymbol{y}) \right]^{1/2},$$

ˆ，即基于协方差矩阵 $\boldsymbol{\Sigma}_{SSM,i}$，计算其与空间$k$ 最近邻域 $b_k(\boldsymbol{x})$内所有观测值的成对马氏距离中的最小值。邻域 $b_k(\boldsymbol{x})$不一定是 a_i的子集。然而，由于协方差矩阵的空间平滑以及常规观测值的空间相关性，与 $\boldsymbol{x}$ 在空间上接近的观测值 $\boldsymbol{y}$ 应当足够相似，即使使用另一个相近邻域 a_i 的协方差矩阵，也不会被归类为离群点。如果 $\boldsymbol{x}$ 和 $\boldsymbol{y}$之间差异很大，那么观测值 $\boldsymbol{y}$ 仍会被归类为异常值。

次近距离 $d(\boldsymbol{x})$ 用作衡量异常程度的指标。如果次近距离较大，则空间 k 最近邻域内没有任何观测值与观测值 $\boldsymbol{x}$相似，这意味着 $\boldsymbol{x}$ 将被标记为局部异常值。为了明确次近距离达到多大才算高，可以使用基于调整箱线图上围栏的非参数截断值（Hubert和Vandervieren，2008），该值根据所有邻域中的所有次近距离计算得出 a_i， $i = 1, \ldots\ldots, N$。高于截断值的观测值被视为局部异常值。
例如，Ernst 和 Haesbroeck（2016中提出的将范围限制为同质邻域等可能的进一步扩展尚未纳入，但值得在未来研究中加以探讨。

3 算法与C步

为了计算空间平滑MRCD的位置与协方差估计量，我们需要最小化目标函数（3）。然而，由于可能的子集组合数量与MCD或MRCD的情况相当，通常无法逐一计算所有这些组合对应的目标函数值，再选出最优组合。为此，我们将把Rousseeuw 和 Driessen (1999) 为 MCD估计量提出的 C 步（集中步骤）策略 H_1 调整 h到这一问题情境中 $\boldsymbol{X}$。给定数据矩阵中对应于 H_2 个 h 观测值的索引集，C 步会选取后续子集：基于H_1中的观测值所计算的算术平均数和样本协方差矩阵，选取马氏距离最小的观测值。Rousseeuw 和 Driessen (1999) 已证明，该过程会收敛到局部最小值。C 步的思想也已扩展至 MRCD（Boudt 等人，2020），现在我们将把广义 C 步调整到我们的情境中。

定理 *5*。对于每个 $j = 1, \ldots, N$，令 $\rho_j \in (0, 1)$，并令 H_j^0为 a_j 中任意一个大小为 $h_j \in (n_j/2, n_j)$ 的初始子集。令 α_j 为所用观测值所占的相应比例， $\alpha_j = h_j/n_j$。令 $\mathcal{H}^0 = (H_1^0, \ldots, H_N^0)$ 为这些子集的组合，并固定 $\lambda \in [0, 1)$。假设目标矩阵 $\boldsymbol{T}(\boldsymbol{X})$ 为正定、对称且固定，并且 $\boldsymbol{K}_j(\mathcal{H}^0)$ 按方程 (2) 定义，其中 $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$。基于各自的子集，计算每个邻域 $a_j, j = 1, \ldots, N$ 的样本均值， $\boldsymbol{m}_j(\mathcal{H}^0) = \frac{1}{h_j} \sum_{k \in H_j^0} \boldsymbol{x}_k$。

固定邻域 a_i。对于 $\boldsymbol{x} \in a_i$，令基于子集 $\mathcal{H}^0$ 定义的基于马氏距离的度量为

$$d(\boldsymbol{x}; \mathcal{H}^0) = (\boldsymbol{x} - \boldsymbol{m}_i(\mathcal{H}^0))' \left[(1 - \lambda) \boldsymbol{K}_i(\mathcal{H}^0) + \lambda \sum_{j=1, j \neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}^0) \right]^{-1} (\boldsymbol{x} - \boldsymbol{m}_i(\mathcal{H}^0)).$$

对于一个新的大小为 h_i 的子集 H_i^1（属于 a_i）和

$$\sum_{k \in H_i^1} d(\boldsymbol{x}_k; \mathcal{H}^0) \leq \sum_{k \in H_i^0} d(\boldsymbol{x}_k; \mathcal{H}^0),$$

表示 $\tilde{\mathcal{H}} = (H_1^0、\ldots\ldots、H_i^1、\ldots\ldots、H_N^0)$（注意， $\boldsymbol{K}_j(\tilde{\mathcal{H}}) = \boldsymbol{K}_j(\mathcal{H}^0)$ 对于 $j \neq i$）。然后，

$$\det \left((1 - \lambda) \boldsymbol{K}_i(\tilde{\mathcal{H}}) + \lambda \sum_{j=1, j \neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\tilde{\mathcal{H}}) \right) \leq \det \left((1 - \lambda) \boldsymbol{K}_i(\mathcal{H}^0) + \lambda \sum_{j=1, j \neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}^0) \right)$$

当且仅当 $\boldsymbol{K}_i(\tilde{\mathcal{H}}) = \boldsymbol{K}_i(\mathcal{H}^0)$ 且 $\boldsymbol{m}_i(\tilde{\mathcal{H}}) = \boldsymbol{m}_i(\mathcal{H}^0)$时，等号成立。

证明。 证明见附录。 □

C步定理指出，只要其他协方差估计量保持不变，目标函数就会在每一步递减。下面介绍的实现算法通常并非如此。不过，该定理及其证明应足以说明以下算法的动机，并为其提供依据。

该算法利用C步，根据 Boudt 等人 (2020)提出的方法求解最小化问题。假设我们可以通过仿射等变MCD估计量估计目标矩阵 $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ ，那么我们也可以获得空间平滑MRCD的仿射等变性。因此，我们可以忽略 Boudt 等人 (2020) 中的缩放步骤，从而减少参数估计的数量。利用 $\boldsymbol{T} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$ 的特征分解，其中 $\boldsymbol{Q}$ 的列包含特征向量，并且 $\boldsymbol{\Lambda}$是由正特征值组成的对角矩阵，我们对观测值进行变换，

$$\boldsymbol{z}_i = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q}'\boldsymbol{x}_i, \tag{6}$$

对于 $i = 1, \ldots, n$。因此，在接下来的步骤中，我们可以使用 $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)'$ 作为数据矩阵，并使用 $\boldsymbol{T} = \boldsymbol{I}_p$ 作为目标矩阵，这在数值上简化了 ρ_j 的数据驱动的选择过程。
为了利用C步启动迭代过程，我们需要初始值，即为每个邻域提供稳健且正则的协方差估计。与原始MRCD估计量的建议做法一致，我们也将对每个邻域分别使用Hubert 等人(2012)提出的确定性MCD算法。该方法为每个邻域生成六组位置与散布估计值，并表现出尤为良好的收敛性质。此外，对于每个邻域 a_j，ρ_j 的值是根据Boudt 等人（2020）提出的基于最大条件数的数据驱动选择过程，按照第3.2至3.4步计算得出的。每个邻域的六组确定性初始值仅保留条件数足够小的那些值（更多详情请参见Boudt 等人，2020）中的第3.5步。

另一个新出现的问题是初始子集的可能组合数量：对于 N 个邻域，最终可能得到多达 6^N 种子集组合，作为可能的初始值。由于大量初始值对于获得良好的近似结果可能并非必不可少，因此，在下文中，我们将考虑初始值数量的上限为 $6N$，并从经过 ρ选择步骤后剩余的可能组合中随机选取这些初始值。

假设现在我们从初始子集 $\mathcal{H}^0 = (H_1^0, \ldots, H_N^0)$ 开始该过程，然后对每个邻域 a_i 应用 C步，并得到一个新子集 H_1^i。这些子集 $\mathcal{H}^1 = (H_1^1, \ldots, H_N^1)$ 的组合用作下一次迭代的起始点，依此类推。当子集不再发生变化时，迭代过程便停止（见图1）。对所有初始值应用 C 步迭代后，我们选择目标函数值最小的子集组合作为最终子集组合 $\mathcal{H}^* = (H_1^*, \ldots, H_N^*)$。
尽管定理 5 并未证明目标函数在每一步都会下降，因为每次迭代后，额外的协方差矩阵会针对各个邻域分别进行调整，但模拟结果表明，在大多数情况下，该算法能产生稳定的结果，并具有良好的单调性（见第 4 节）。

收到 fin 后 对每个邻域中的子集进行组合 $\mathcal{H}^*$ 后，矩阵 $\boldsymbol{K}_i^Z$将被反变换为 t o

$$\boldsymbol{K}_i^* = \boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}\left[\boldsymbol{K}_i^Z(\mathcal{H}^*)\right]\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}'. \tag{7}$$

邻域 a_i 的协方差估计为

$$\hat{\boldsymbol{\Sigma}}_{SSM,i} = (1-\lambda)\boldsymbol{K}_i^* + \lambda\sum_{j=1, j\neq i}^{N}\omega_{ij}\boldsymbol{K}_j^*, \tag{8}$$

位置估计则为最优子集 $\boldsymbol{X}_{\mathcal{H}_i^*}$的算术平均数。详细的数值计算过程以伪代码形式概述于算法1中。

对于调节参数 λ，目前没有类似于正则化参数 ρ_i计算过程那样的标准化自动选取方法。不过，第5节展示了多种可用于实际应用中选择 λ 的方法。

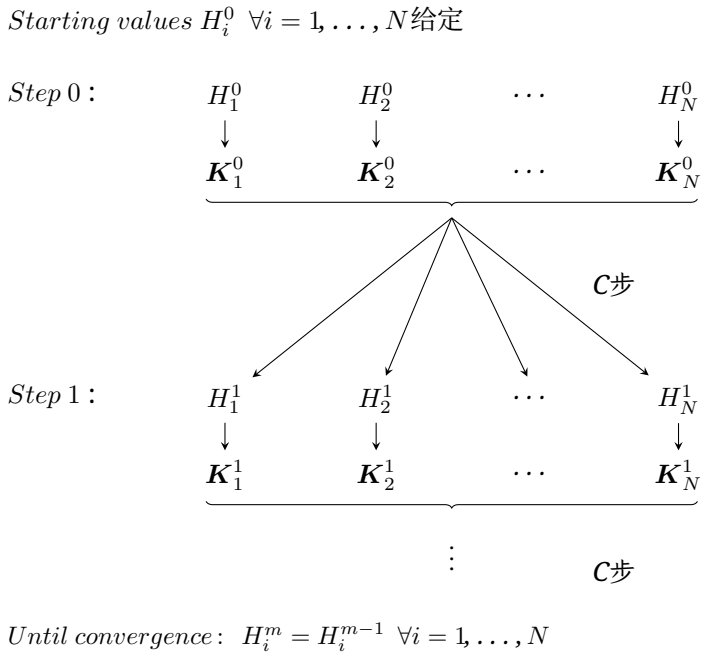

图1：每次迭代步骤后 C步中使用的矩阵。 H_i^j 是步骤 i中邻域 a_j 的选定子集， $\boldsymbol{K}_i^j$ 是相应的正则协方差矩阵。

算法1：空间平滑MRCD估计量的算法。1 步骤1.1：使用MCD估计量对 $\boldsymbol{X}$计算目标矩阵 $\boldsymbol{T}$；2 步骤1.2：对 $\boldsymbol{T} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$ 进行特征分解，并根据公式（6）变换观测值；3 步骤2：针对每个邻域，按照Boudt等人（2020）的步骤3.1至3.5进行初始化（不包括C步）；4 对于 $i = 1, .. . , N$，执行 5 固定邻域 a_i；6 根据Hubert等人（2012），从 a_i 中获取6个由 h_i 个观测值组成的初始确定性集合；7 计算6个初始协方差矩阵和均值估计；8 通过数据驱动方法选择特定于邻域的 ρ_i；9 根据条件数筛选初始起始估计子集（Boudt等人（2020）的步骤3.5）；10 结束 11 随机选择一组初始H 集合组合，作为初始值；12 步骤3.1：C步：针对每个初始子集组合进行迭代，直至收敛；13 步骤3.2：根据目标函数值选择最佳子集组合；14 步骤4：计算 $\boldsymbol{K}_i\left(\mathcal{H}^*\right) \ \forall i$，并使用公式（7）和（8）得到最终估计；

4 数值模拟

为了测试这种新方法，我们设置了两个模拟情景，其中也纳入了协方差估计和异常值检测所需的邻域结构。对于这两种设置，我们模拟依赖于参数 δ 的协方差矩阵，记为 $\mathbf{\Sigma}(\delta)$，其元素定义为 $(\mathbf{\Sigma}(\delta))_{jk} = \delta^{(j-k)}$，其中 $j,\ k \in \{1,\ldots,p\}$，从而保证正定性和对称性。

设置 *1*： 第一种设置的灵感来自协方差矩阵随空间平滑变化这一最初构想。我们首先设置观测值 $\boldsymbol{x}_i$ 的二维坐标 $(s_i^1,\ s_i^2)$，每个坐标轴上有 $n_{sim} = 41$ 个观测值，均匀分布在 0 到 20 之间，因此总计得到 $n = n_{sim}^2 = 1681$ 个数据点。

对于 N_{sim} 个包含相似观测值的区域 $a_{lm}, l, m \in \{1,\ldots,\sqrt{N_{sim}}\}$，我们构建第二个空间网格，平均每个网格单元包含 n_{sim}/N_{sim} 个观测值。第一坐标轴上各区域的边界定义为 $b_l^1 = l\frac{20}{\sqrt{N_{sim}}}$，适用于 $l = 0$、……、$\sqrt{N_{sim}}$。类似地，第二坐标轴上的边界定义为$b_m^2 = m\frac{20}{\sqrt{N_{sim}}}$，适用于 $m = 0$、……、$\sqrt{N_{sim}}$，从而形成均匀间隔的网格。因此，区域 a_{lm} 包含观测值$\{\boldsymbol{x}_i:\ s_i^1 \in (b_{l-1}^1,\ b_l^1] \cap s_i^2 \in (b_{m-1}^2,\ b_m^2]\}$，适用于 $l,\ m \neq 0$。对于 l 或 m 等于 1 的情况，区间的左端点（即零）包含在内。区域 a_{lm} 的坐标中心定义为 $c_{lm}^1 = \frac{b_{l-1}^1 + b_l^1}{2}$ 和 $c_{lm}^2 = \frac{b_{m-1}^2 + b_m^2}{2}$。

随后从$\mathcal{N}(\boldsymbol{\mu}_{lm}, \mathbf{\Sigma}(\delta_{lm}))$这一 p维正态分布中随机抽取区域 a_{lm} 的观测值，其中 $\boldsymbol{\mu}_{lm} := ((c_{lm}^1 + c_{lm}^2)/2, \ldots, (c_{lm}^1 + c_{lm}^2)/2) \in \mathbb{R}^{p\times 1}$，因此其元素值介于 0 和 20 之间。对于区域 a_{lm} 的协方差矩阵 $\mathbf{\Sigma}(\delta_{lm})$，我们采用上文所述的结构，并将参数 δ_{lm} 定义为

$$\delta_{lm} = \left(0.1 + l\frac{0.9-0.1}{\sqrt{N_{sim}}}\right)\left(0.1 + m\frac{0.9-0.1}{\sqrt{N_{sim}}}\right) \in [0.01, 0.81],$$

从左下角到右上角平滑递增。图 2（左）展示了一个含有 $p = 2$ 的模拟数据集，其中网格结构和均值的变化清晰可见。结果中也显示了各个区域的 δ_{lm}。

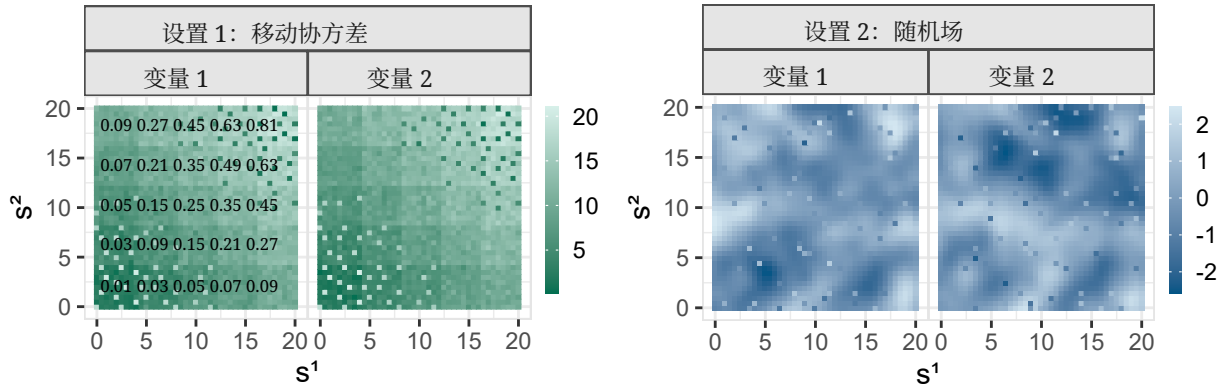

图 2：带有模拟情景的 $p = 2$ 和 5% 污染率。左侧展示了模拟设置1，通过 Ernst 和 Haesbroeck Ernst 和 Haesbroeck (2016) 中所述的交换过程实现污染，$N_{sim} = 25$ 以及 $n_{sim} = 41$。最左侧面板上标出的数值对应于参数 δ_{lm}。右侧展示了带有 $\nu = 3$的设置 2，并采用完全随机交换。

设置 2：为了获得更灵活的模拟设置，我们采用了一个随机场，具体来说是简约多元 Matérn 模型（参见 Gneiting 等人，2010；Ernst 和 Haesbroeck（2016）），这与 Ernst 和 Haesbroeck（2016）以及 Harris 等人（2013）的建议一致。我们没有使用 Ernst 和 Haesbroeck（2016）及 Harris 等人（2013）中构造的矩阵，而是再次选择了如上所述的矩阵结构 $\mathbf{\Sigma}(\delta)$，其中包含 $\delta = 0.7$，因为它可以扩展到更高维度。将 δ 设为 0.7 后，我们得到从高到低的一系列相关性，反映变量之间的不同关系。假设所有变量的空间平滑度相同，并由一个平滑度参数 ν 控制，其取值为 $\{0.5, 1.5, 3\}$。较高的 ν 会带来更高的空间平滑度，并且通常会在污染后产生更明显的离群点。Matérn 模型的空间尺度参数 a 设为 1。观测值的坐标采用网格结构，取值范围为 0 到 20，网格间距为 0.5，因此总共得到 $41^2 = 1681$ 个观测值，与设置 1 中的标准设置类似。

最后，通过相互交换观测值的坐标来实现离群值污染。 Filzmoser 等人(2013 年) 交换完全随机选取的观测值，而 Ernst 和 Haesbroeck (2016) 则提出交换全局稳健主成分中第一得分最极端的观测值。为避免因空间相关性较高而导致整片区域的观测值相互交换的问题，观测值交换后，会将其 15 个最近邻从交换过程中移除。这使离群观测得以清晰区分，同时避免其他离群值靠近它们（另见图 2）。因此，按照 Ernst 和 Haesbroeck(2016) 进行交换，通常应能使所有考察的方法取得更好的表现。将在两种设置下分析这两种交换方法，并改变污染水平 β，范围为 1%至 10%。

对于 ssMRCD协方差估计，我们采用基于网格的邻域结构。与设置 1 的描述类似，我们使用等间距边界，并将观测值分配到邻域 n_i，对于 $i = 1$、.....、 N。因此，设置 1 中的情形 $N = N_{sim}$ 体现了为 ssMRCD 选定的邻域与真实潜在协方差结构完美匹配。对于 $N_{sim} > N$，ssMRCD 使用较少的邻域，从而使协方差估计更加平滑。加权矩阵 $\boldsymbol{W}$ 基于各中心之间的欧几里得距离的倒数，这些中心的定义方式与设置 1 相同。

4.1 协方差估计

首先，我们希望更详细地分析第 3 节中描述并阐释的算法。由于定理 5 仅适用于一个协方差矩阵变化的情形，而不适用于多个协方差矩阵同时变化的情形，因此有必要进一步考察其收敛性质。图 3 展示了在两种模拟设置下，采用不同参数设置时，所有初始 H 集合组合沿迭代过程的目标函数值。我们选择 $p = 5$，并通过完全随机交换实现 5% 的污染率。加权矩阵基于距离倒数构建，如第 2 节所述。每个面板都反映了一个模拟数据集的目标函数收敛行为。

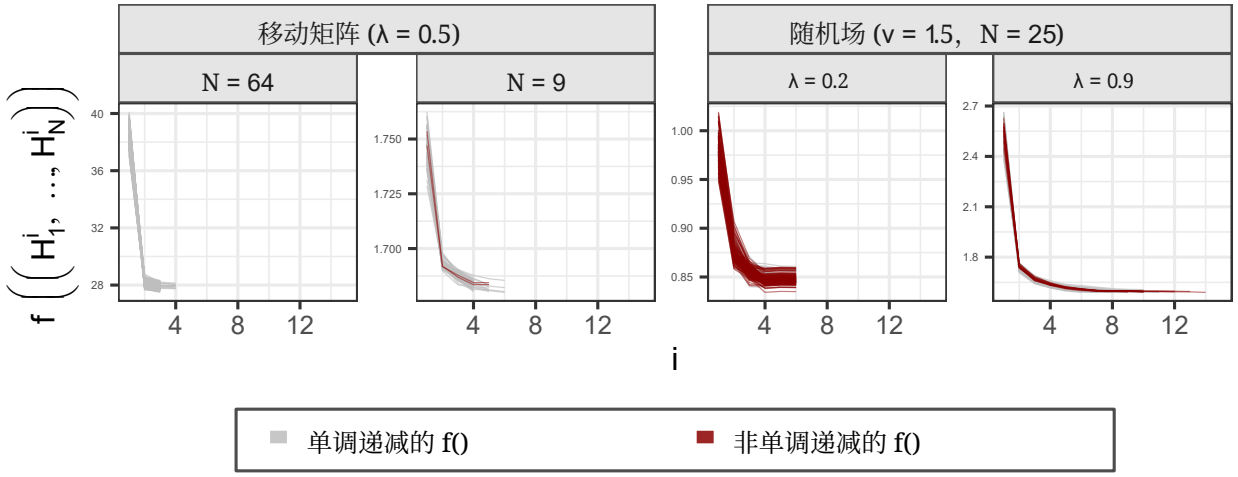

图 3：目标函数的不同收敛行为。维度为 $p = 5$，通过完全随机交换实现 5% 的污染率。每条线表示算法第 3 节中沿 C 步迭代，从一个初始 H 集合出发的路径。

尽管整体表现有所不同，但显而易见，该算法总体上具有非常良好的收敛性质。通常可以实现很高比例的目标函数值单调递减，而非单调递减的路径也只会略微上升。模拟中的收敛结果之所以可靠，可能是因为空间相关值使算法过程中的协方差矩阵变化相当小。因此，对于我们的模拟数据集，固定协方差矩阵这一假设似乎得到了充分满足。由于现实中的局部异常值检测大多只有在数据具有空间相关性时才有意义，定理 5 的理论结果因而更显重要。

此外，收敛速度很快，这可能是由 detMCD 算法所选用的良好初始估计值导致的（Hubert 等人，2012）。另一个原因可能是，邻域分配将可用观测值的数量限制在比完整数据集的协方差估计更小的范围内。另一个非常有希望的发现是，在最初阶段改进迅速，而且这一点不受模拟数据集的影响。

另一个需要讨论的问题是计算效率。迭代过程以及随着 N 增大而不断增加的初始值数量，都会对运行时间产生相当大的影响。尽管如此，只要邻域数量 N 不太大，基于 ssMRCD 的异常值检测方法就能与其他局部方法相媲美（Ernst 和 Haesbroeck，2016），尤其是在 p 较大的情况下（另见图 10）。对于模拟设置 1，采用$N_{sim} = 25$，通过完全随机切换引入 5% 的污染率，并进行 100 次重复，参数 p、 λ、 N和 n 分别进行单变量变化。未进行变化的参数的默认值为 $p = 5$、$\lambda = 0.5$、 $N = 25$ 和 $n = 1681$。

如图4所示，邻域数量 N和观测值数量 n 对运行时间的影响最大，且增幅超过线性增长。 N近似呈二次增长，部分原因是初始值数量随 N线性增加；如有必要，也可以减少初始值数量以提高效率。有趣的是，维度 p对运行时间的影响近似呈线性，整体影响适中。平滑参数 λ不会显著改变运行时间。

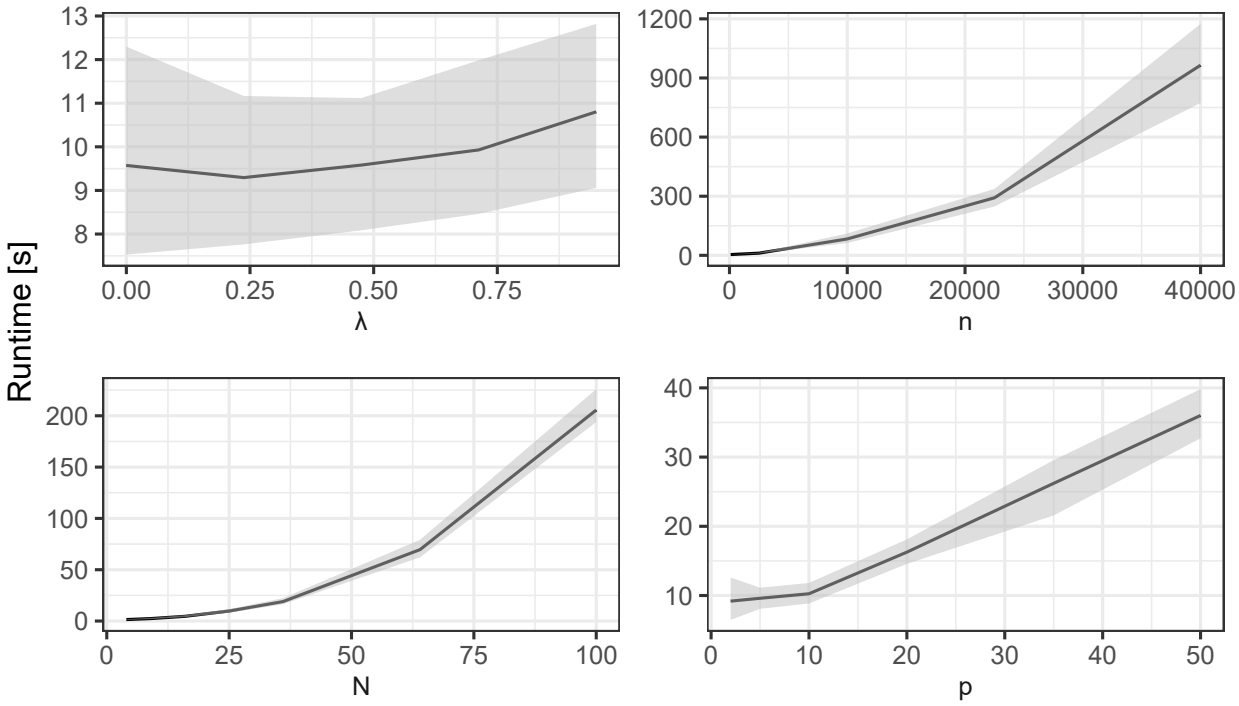

图4：污染率为 5% 时，设置 1 下 ssMRCD 运行时间分析； $N_{sim} = 25$ 以及参数取不同值时的结果，未变化的参数则采用默认值 $\lambda = 0.5$、 $p = 5$、 $N = 25$ 和 $n = 1681$ 。实线表示 100 次重复的均值，均值周围灰色带的边界表示第 5 和第 95 分位数。

4.2 异常值检测

在将局部异常值检测的性能与其他方法进行比较之前，我们将更详细地分析 ssMRCD 的参数敏感性。这项模拟研究旨在简化 λ和 N的选择，尤其是对于真实世界数据，并重点关注与参数设置欠佳相关的潜在问题。

为此，采用设置 2（随机场）作为模拟设置，参数为 $\nu = 3$，并且 $\beta = 5\%$观测值完全随机交换。重点关注 λ 和 N 的选择，同时也分析维度 p 的影响。其他参数则根据可能的默认设置进行选择。加权矩阵基于邻域中心之间的欧氏距离的倒数 a_{ij}。由于所有考虑的方法都建议将 $k = 10$ 作为默认值，因此目前我们采用这一设置。每种参数组合均进行了100 次不同的实现模拟。尽管 Ernst 和 Haesbroeck (2016) 建议使用 Cohen's Kappa 作为混淆矩阵的汇总统计量，但鉴于 F1 分数具有良好的可解释性，且也适用于类别不平衡分类数据，我们将采用 F1 分数。

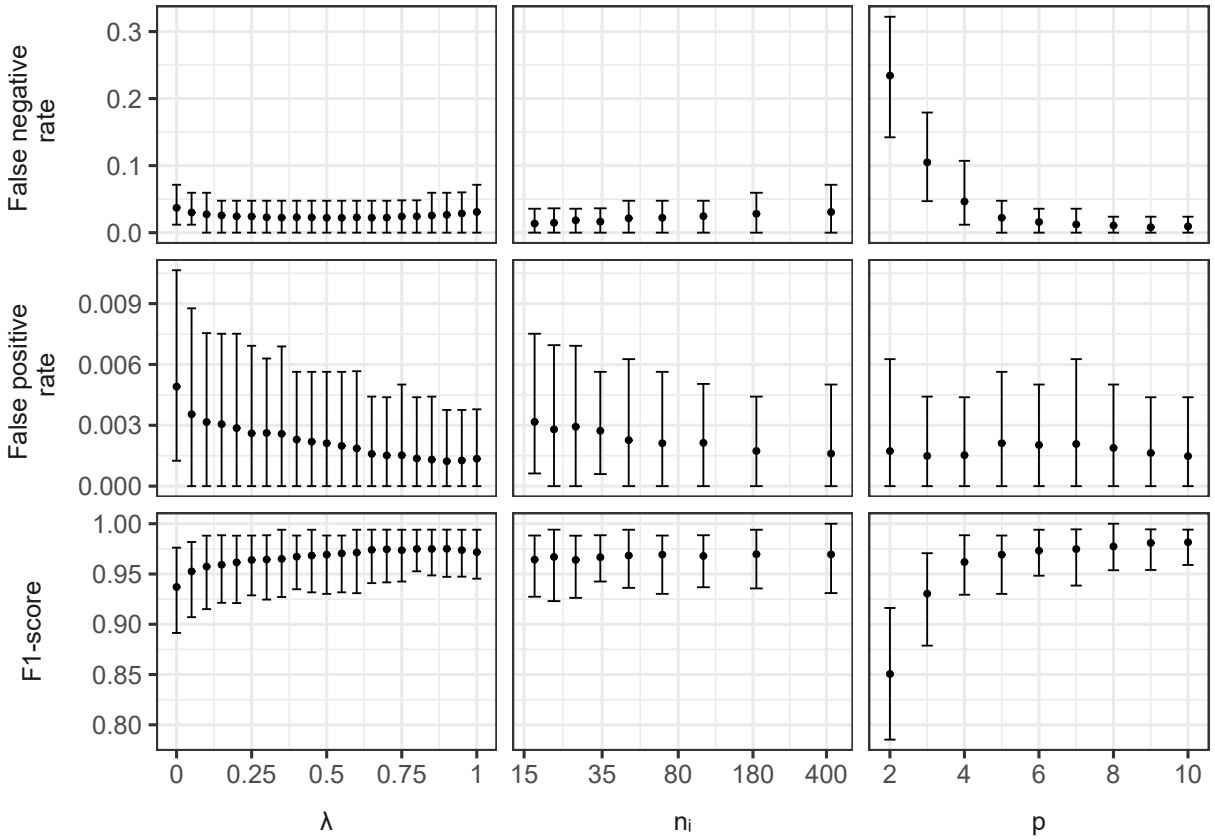

图 5：在不同参数设置下，基于 ssMRCD 异常值检测方法的假阴性率、假阳性率和 F1 分数评估异常值检测性能。每个点表示 100 次模拟的算术平均数，相应的误差线表示第 5 和第 95 分位数。对于不变的参数，默认设置为$p = 5$、 $N = 25$（与 $n_i \approx 67$相当）和 $\lambda = 0.5$。

图 5 展示了假阴性率、假阳性率以及相应的 F1 分数随 λ、 N 和 p 变化的曲线，其中 $p = 5$、 $N = 25$ （意味着平均而言为 $n_i \approx 67$ ）和 $\lambda = 0.5$ 取默认值。这些数值应能反映一种相当通用且不具特异性的参数设置。为了便于说明，绘制平均邻域大小 $n_i \approx 1681/N$ 比绘制邻域数量更有信息量。

模拟结果表明，较高的 λ 会降低假阳性率；在平滑程度过高、掩盖真实离群点之前，它对假阴性率的降低作用很小。总体性能在 λ中有所提升，但当 λ 高于 0.5 时，提升幅度很小。因此，我们建议将 λ 的默认值设为 0.5，以在降低假阳性率的同时，避免数值过高造成的掩盖效应。与λ的影响相比，维度 p 的影响更为显著。维度很小时，似乎能更有效地覆盖离群点，这可能是因为可用信息较少。有趣的是，邻域的大小似乎无关紧要，至少在本次模拟设置中是如此。仅出现轻微的掩盖效应，与 λ的影响类似。邻域过大会导致过度平滑。因此，这可能意味着应选择中等大小的邻域，以提高计算效率并减少不必要的正则化。这些参数选择建议可能偏向于这一特定模拟设置，在其他设置下未必最优；不过，将参数固定为至少合理的值，可以简化整体流程。

关于 ssMRCD 协方差对于局部离群检测的适用性，我们将其异常值检测性能与 Filzmoser 等人(2013 年) (F)、Ernst 和 Haesbroeck (2016)(EH) 的局部离群检测方法，以及 Breunig 等人(2000) 的局部离群因子方法论进行比较，并按照 Schubert 等人(2012) 所述，以规范方式将其调整为空间邻域 (LOF)。两种模拟设置在参数方面各不相同 p、 ν 和 N_{sim}，并分别使用两种交换过程；每种组合均模拟 100 次。所有纳入比较的方法都会将每个观测值与其 k 个邻近观测值中的许多个进行比较，我们假定所有方法的该数量均为 10 。对于 ssMRCD，我们

将假设沿用先前设置中的默认值（$\lambda = 0.5$、 $N = 25$ 和 $\boldsymbol{W}$，基于逆距离）进行参数敏感性分析。

Filzmoser 等人提出的离群点分类方法 Filzmoser 等人 (2013 年) 有一个参数 β_F，表示局部离群点所允许与之相似的邻近观测值所占的百分比。这里我们采用 $\beta_F = 0.1$，正如 Ernst 和 Haesbroeck (2016) 所提出的，这意味着 $0.1k = 1$ 观测值可以与其 10 个最近邻中的观测值相似。对于内点，孤立度的期望值为 β_F。如果实际孤立度高于期望值，则表明存在局部离群性。我们将观测值归类为局部离群点的截断值设为 β_F的两倍，即 20%。这个截断值比 Ernst 和 Haesbroeck (2016) 模拟设置中采用的标准宽松，他们使用期望值的三倍作为截断值，即 30%。请注意，Filzmoser 等人 (2013 年) 的方法论主要侧重于可视化异常值检测工具，因此此处选取的截断值可能并非最优。

对于Ernst 和 Haesbroeck (2016提出的正则化空间检测方法，参数 β_{EH}表示纳入异常值检测过程的最同质邻域所占比例，该参数设为 1。在模拟中，我们关注所有离群值，并且模拟数据中的异质性应对所有观测值而言都具有可比性。因此，只考虑其中一部分会导致结果不可比较。此外，Ernst 和 Haesbroeck(2016的模拟结果表明，在所有考察的设置中，$\beta_{EH} = 1$ 同样是最优的。作为正则协方差矩阵估计量，我们使用采用默认目标矩阵（等相关目标矩阵）的 MRCD，并设 $\alpha = 75\%$。
最后但同样重要的是，非参数方法局部离群因子（LOF）根据每个观测值与其 k 近邻的所谓“局部可达密度”进行比较，计算局部离群因子；该方法需要一个截断值。由于没有固定规则来选择截断值，因此局部离群因子高于1。5 时，即判定为离群点。原始论文Breunig 等人(2000)也采用了这一数值。

两种模拟设置以及完全随机切换的结果分别见图 6 和图 7。移动矩阵方法的结果 Ernst 和 Haesbroeck (2016) 见图 8 和图 9。
首先看假阳性率（FPR），我们发现 Ernst 和 Haesbroeck(2016) 的方法在几乎所有设置下都存在问题，会将过多正常观测值判定为离群点。这可能是因为其协方差估计非常局部化，通常可能过于严格，从而导致严重的淹没效应，尤其是在观测值的空间相关性较弱的设置中。有趣的是，Filzmoser 等人 (2013) 方法的假阳性率（FPR）表现取决于数据模拟设置。在移动矩阵设置下，假阳性率较高；在随机场下则很低。局部离群因子（LOF）和基于 ssMRCD 的异常值检测方法在所有设置下的假阳性率都稳定地保持在较低水平。
假阴性率（FNR）的结果则截然不同。对于 Ernst 和 Haesbroeck (2016)，其假阴性率（FNR）在许多设置中都低于所有其他方法，但这可能只是因为其假阳性率（FPR）较高。 Filzmoser 等人 (2013 年) 的方法在大多数情景中的假阴性率都非常高，即使在假阳性率较高的情景中也是如此。对于局部离群因子（LOF），模拟设置对其性能有很大影响。在移动矩阵设置中，与其他方法相比，该方法表现相当不错；而在随机场中，除了 Filzmoser 等人(2013 年) 的方法外，其假阴性率几乎无法与其他方法相提并论。ssMRCD 方法的表现介于两者之间。尽管对于完全随机交换的移动矩阵设置，其对应的假阴性率并不突出，但对于 Ernst 和 Haesbroeck (2016) 的切换方法而言，仍处于合理范围内。此外，ssMRCD 离群检测技术在其他情景中的假阴性率都低得令人信服。

比较不同方法的 F1 分数后可以发现，通常最佳方法取决于具体场景。尽管所选的三种方法似乎在至少一种模拟设置中存在不足，但基于 ssMRCD 的方法始终表现可靠，且大多位列表现最好的两种方法之中。此外，在假阴性率（FNR）和假阳性率（FPR）方面，其表现也较少出现极端情况。因此，基于 ssMRCD 的局部异常值检测方法可作为标准异常值检测任务的首选方法。

此外，当维度和观测值数量较多时，计算效率就变得十分重要。图 10 展示了在参数设置依据前述分析的情况下，针对移动矩阵场景 $N_{sim} = 100$，对四种异常值检测技术运行时间进行的一项简短模拟研究结果。随着维度 p 增大，所有方法都需要更多计算时间，尤其是基于协方差估计和求逆的方法（EH、ssMRCD、F）。此外，观测值数量似乎也会对运行时间产生很大影响，尤其是Filzmoser 等人 (2013 年) 提出的方法，这可能是由于寻找 k 近邻的实现效率不高所致。尽管局部离群因子（LOF）方法的运行速度稳定且很快，ssMRCD 的效率似乎也不逊色，尽管它涉及复杂的协方差估计过程。

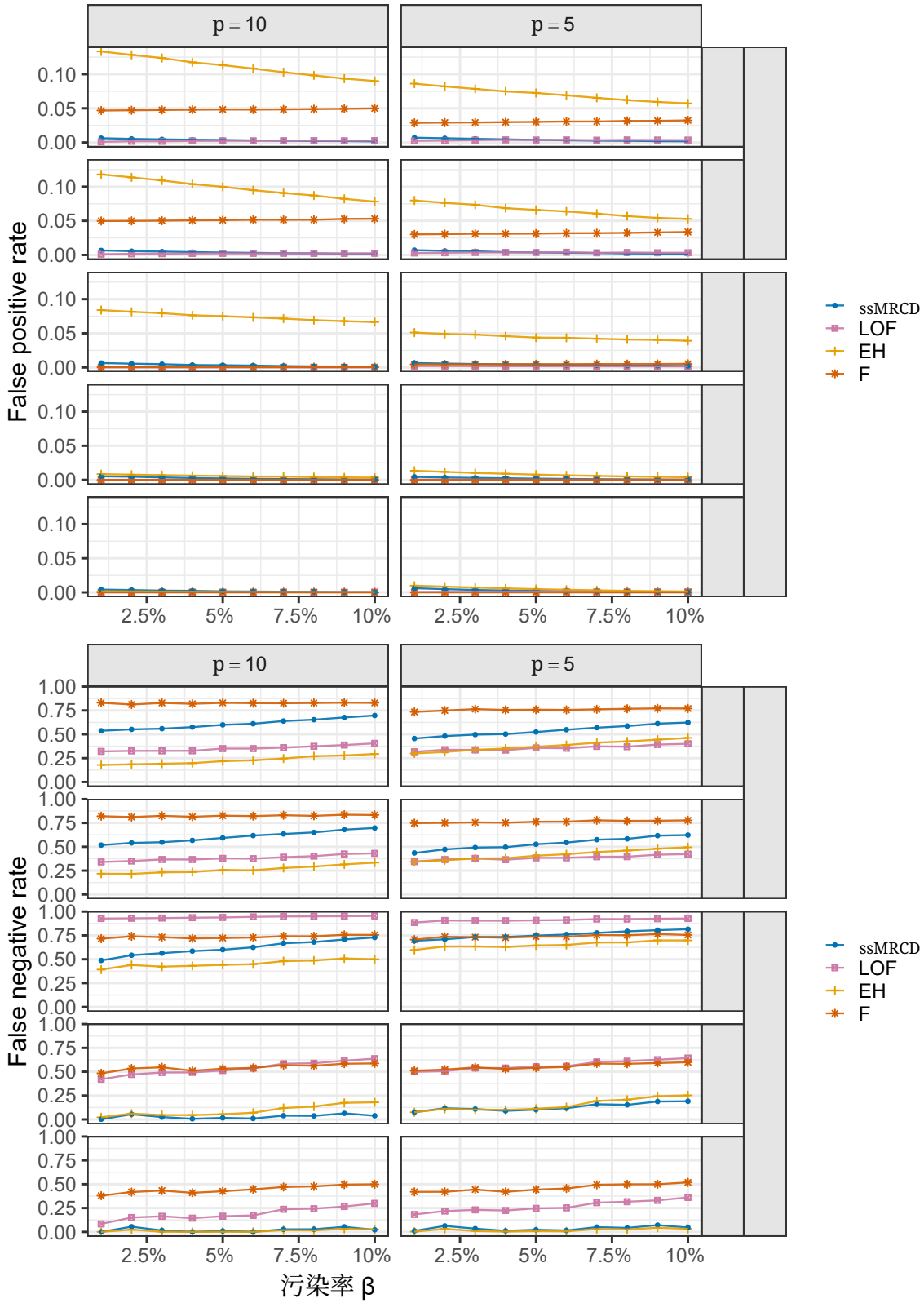

图6：在不同情景下，通过完全随机切换实现污染时，四种离群点检测方法的假阳性率和假阴性率。每个点代表100次重复的均值。

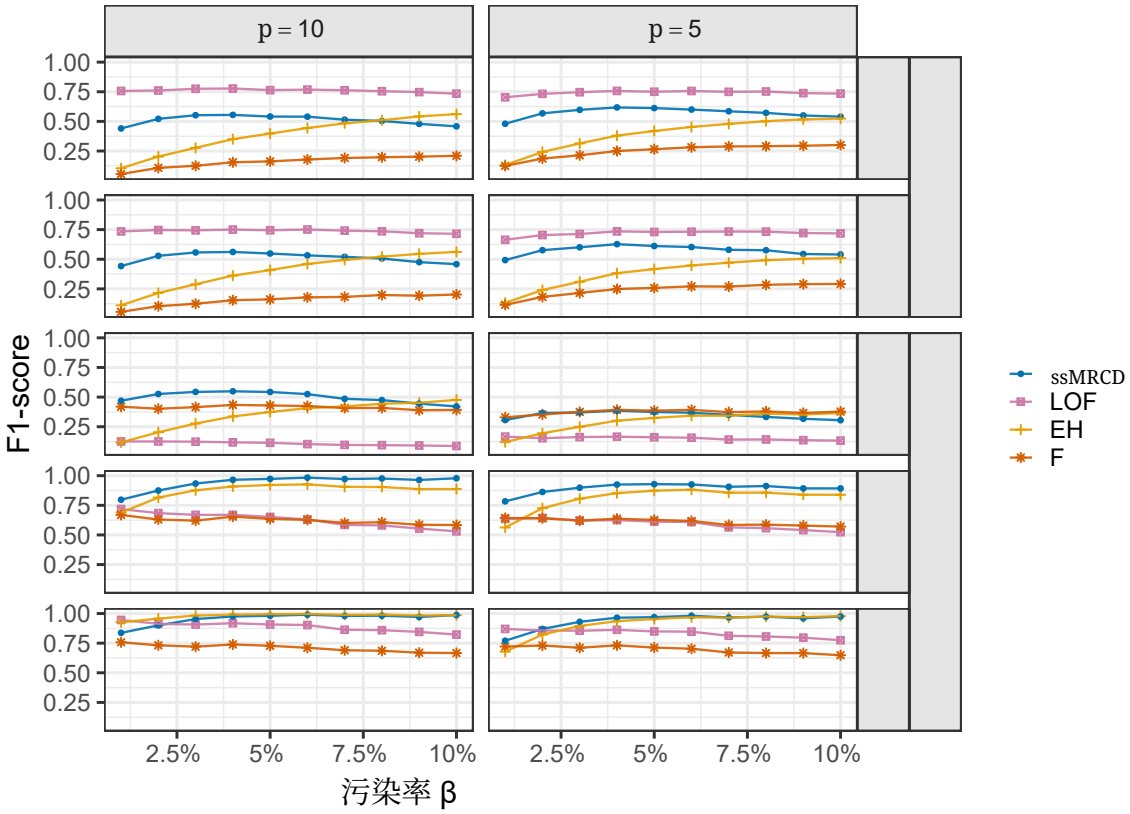

图 7：通过完全随机切换实现污染时，四种离群点检测方法在不同情景下的 F1 分数。每个点代表 100 次重复的均值。

5 示例

本节使用由 GeoSphere Austria (2022) 提供的数据集。该数据集包含奥地利气象站的月度天气数据，用于测试和比较不同的方法。数据集包含以下测量值：气压 [hPa] (p)、相对空气湿度 [%] (rel)、日照时数月总量 [h] (s)、风速 [m/s](vv)、离地面 2 米处的气温 [°C] (t)，以及日均降水量 [mm] (rsum)，以上数据均为 2021 年所有月份的平均值，涉及 $n = 183$分布于奥地利各地的气象站（另见图11或图14）。所有方法使用的坐标均为纬度和经度。

我们将 $k = 10$ 设为所有方法的统一参数，即无论采用何种方法，我们都希望将一个观测值与其十个最近邻进行比较。尽管对于 Ernst 和 Haesbroeck (2016) 的方法，可以剔除邻居间异质性水平相对较高的观测值，但我们希望纳入所有观测值，因此将$\beta = 1$设为该值。即使邻居间存在较强的异质性，某个观测值仍可能是一个有趣的离群点，用肉眼即可清楚辨认。此外，如前一节所述，Ernst 和 Haesbroeck (2016) 的模拟结果显示，在较高的 β下表现最佳。对于 Filzmoser 等人 (2013) 的方法，按照模拟设置，我们允许十个邻居中有一个与局部离群点相似。截断值再次设为 0.2。对于局部离群因子，我们采用与模拟相同的截断值 1.5 。

对于 ssMRCD 局部异常值检测方法，我们在协方差估计中采用基于网格的邻域结构。考虑到阿尔卑斯山地景观，尤其是奥地利西部地区的地形特点，我们希望采用较为局部的协方差结构，因此选择较精细的网格，其邻域数为 $N = 21$，每个邻域平均包含 $n_i \approx 8.7$ 个观测值。其他可行方案可以基于底层结构，例如历史或政治原因，也可以采用其他分类方法，例如对空间坐标进行聚类。此外，我们根据各邻域中心之间的距离，使用反距离权重构建加权矩阵 $\boldsymbol{W}$，并选取默认平滑程度 $\lambda = 0.5$，以实现充分平滑，同时保留精细网格结构的局部性。

作为使用默认值 $\lambda = 0.5$ 的替代方案，我们可以设置一个模拟过程。假设真实数据未受污染，我们可以像第 4 节的模拟研究那样交换观测值，并将其定义为

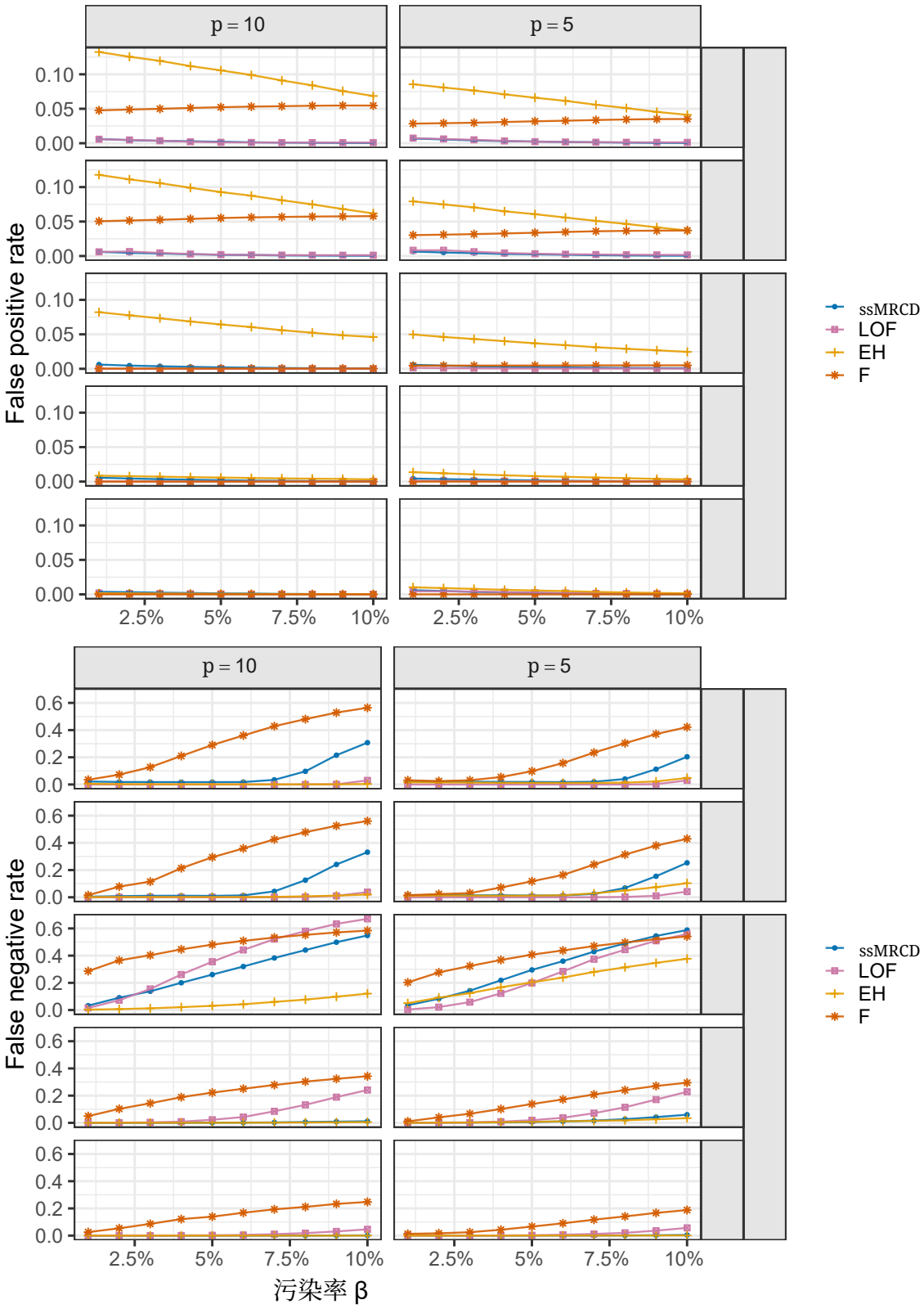

图8：通过 Ernst 和 Haesbroeck (2016) 的切换方法实现污染时，四种离群点检测方法在不同情景下的假阳性率和假阴性率。每个点代表100次重复的均值。

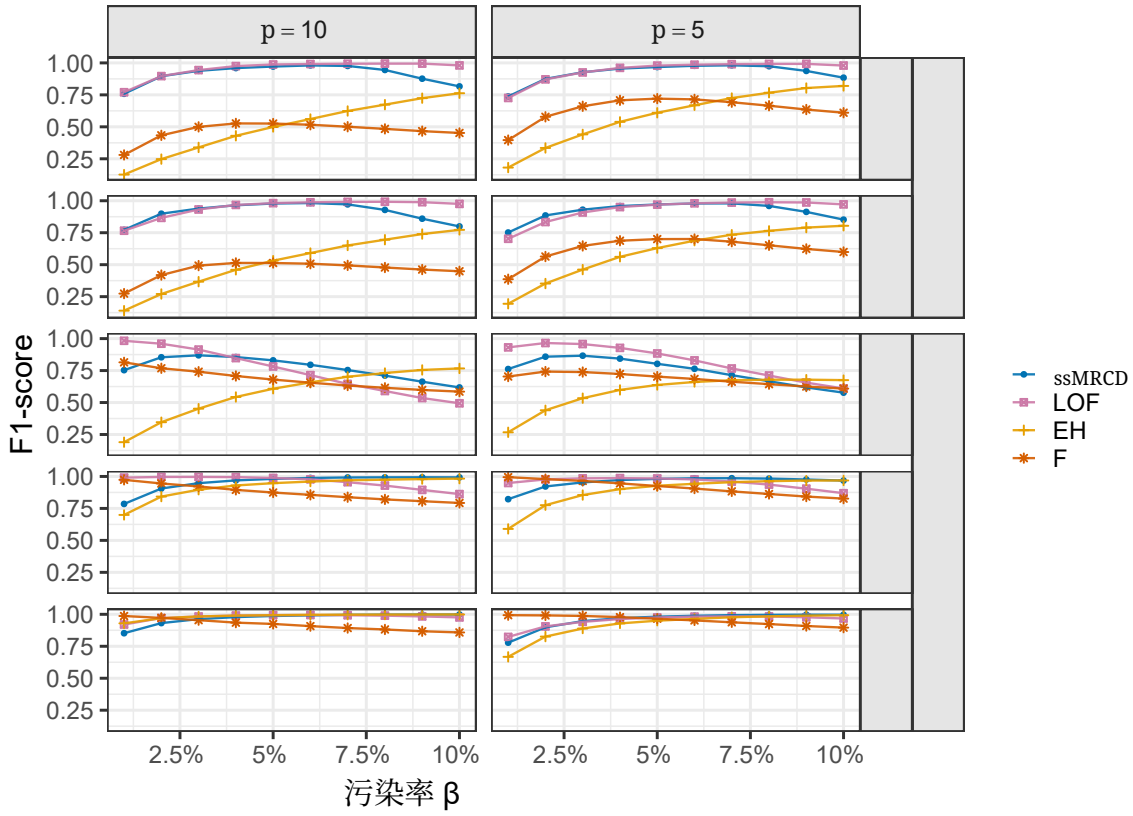

图 9：F1 分数：通过切换方法进行污染的四种离群点检测方法 f Ernst 和 Haesbroeck (2016) 对于不同的情景。每个点表示 100 次重复的均值。

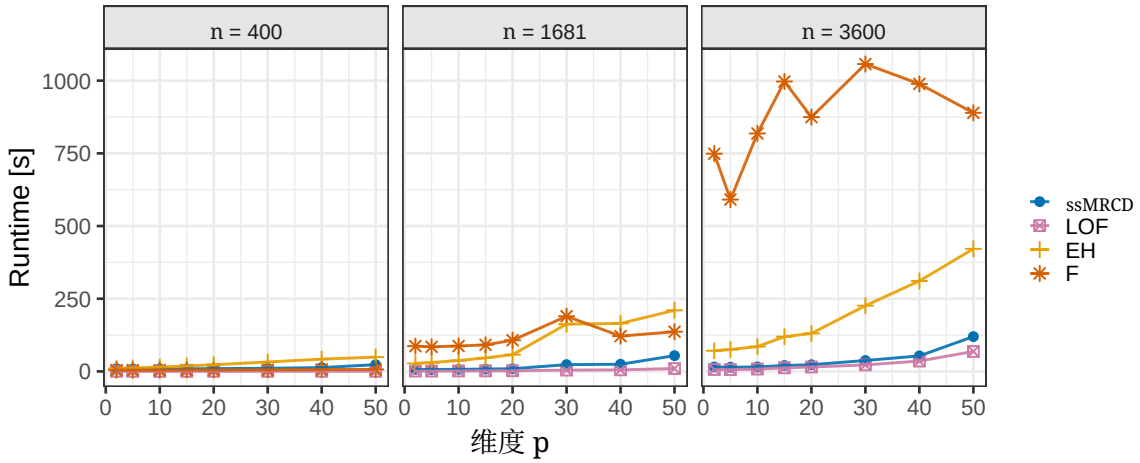

图 10：使用模拟研究设置 1 比较异常值检测算法的平均运行时间，观测值数量为 n，维度为 p。使用了前述异常值检测性能模拟研究中的参数设置。

局部异常值。我们可以使用 ssMRCD 和不同的 λ 选择来应用异常值检测技术（这也可以应用于 ssMRCD 的其他参数设置），然后分析发现的离群点比例和离群点总数。由于只关注已知的假阴性率会导致发现的离群点出现更高的假阳性率，因此将发现的离群点总数也纳入考虑是合理的。λ 的良好取值需要在较低的假阴性率和总体上相对较少的发现离群点之间取得平衡。有趣的是，该过程支持在此数据集中选择 $\lambda = 0.5$。

每个邻域对应的 ssMRCD 相关矩阵如图11所示。观测值和 ssMRCD 相关矩阵的容忍椭圆均按照其所属邻域着色。由于维度为 $p = 6$，我们将维数降至全局最小协方差行列式相关矩阵 $\boldsymbol{T}$ 的前两个特征向量 v_1 和 v_2，该矩阵显示在左上角，希望借此呈现大部分相关方差。此外，右侧还添加了 $\boldsymbol{T}$ 的双标图，以便将相关矩阵与天气数据联系起来。

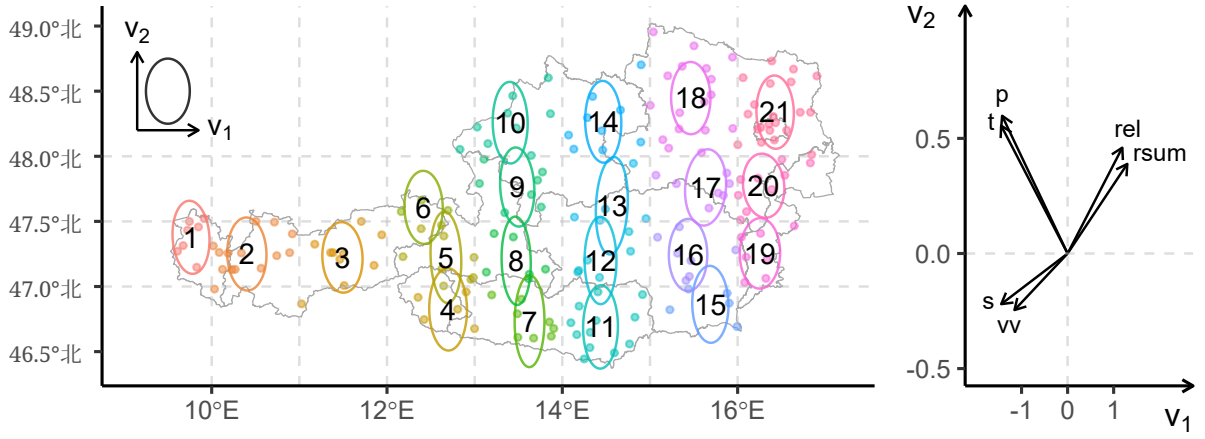

图11：左侧展示了奥地利及其各地区的轮廓，以及为 ssMRCD 邻域设置的网格结构。在此，奇异点被分配给相邻的邻域。对于每个邻域 a_i，索引 i 被置于中心位置，并沿 $\boldsymbol{T}$ 的第一和第二特征向量坐标绘制相应相关矩阵的容差椭圆；$\boldsymbol{T}$ 可在左上角作为参考看到。右侧展示了 $\boldsymbol{T}$ 的双标图。观测值和椭圆按照其所属邻域着色。

应用全部四种离群点检测方法后，共有 24 个观测值被归类为离群值。其中，最多的离群值（21 个）由 Ernst 和 Haesbroeck (2016) 的方法识别，最少的（3 个）由 Filzmoser 等人 (2013) 识别；这与关于假阳性率（FPR）和假阴性率（FNR）的模拟结果一致，尤其是在随机场设置下。图12展示了各方法针对每个观测值用于异常值检测的距离。为进一步比较结果，图13的上半部分展示了全部 24 个被归类为离群值的观测值，以及相应的距离值与截断值之比。比值大于 1 的为离群值。可以看到，有多个气象站仅被 Ernst 和 Haesbroeck (2016) 的方法归类为离群值，这与模拟结果一致，表明这些气象站中可能存在一些假阳性。图13下半部分的面板b)可能就是一个假阳性的例子。气象站 Feuerkogel（面板a)）未被 Filzmoser 等人(2013) 的方法检测到，这同样与随机场设置下的模拟结果以及总体较高的假阴性率（FNR）一致。有趣的是，局部离群因子（LOF）似乎也存在缺陷，例如未能检测出气象站 Patscherkofel（面板c)）是离群值。不过，气象站 Schoeckl（面板d)）被所有方法检测到。

查看图14，我们可以对三个局部离群站点中的两个为何具有局部离群性，以及面板b)为何可能并非离群点，找到一些解释。Schoeckl 和 Feuerkogel 两个站点位于较高海拔，且比周围大多数 k近邻站点更为暴露，这很容易导致天气模式有所不同；而站点 Linz-Stadt（面板b)）则位于相当平坦的区域，与周边环境相似。Patscherkofel 已深入阿尔卑斯山区，周围既有山谷中的站点，也有山地上的站点。尽管从图13的面板c)可以看出，Patscherkofel 的风速明显不同，但尚不清楚为何它与海拔和暴露度相似的站点也相差如此之大。

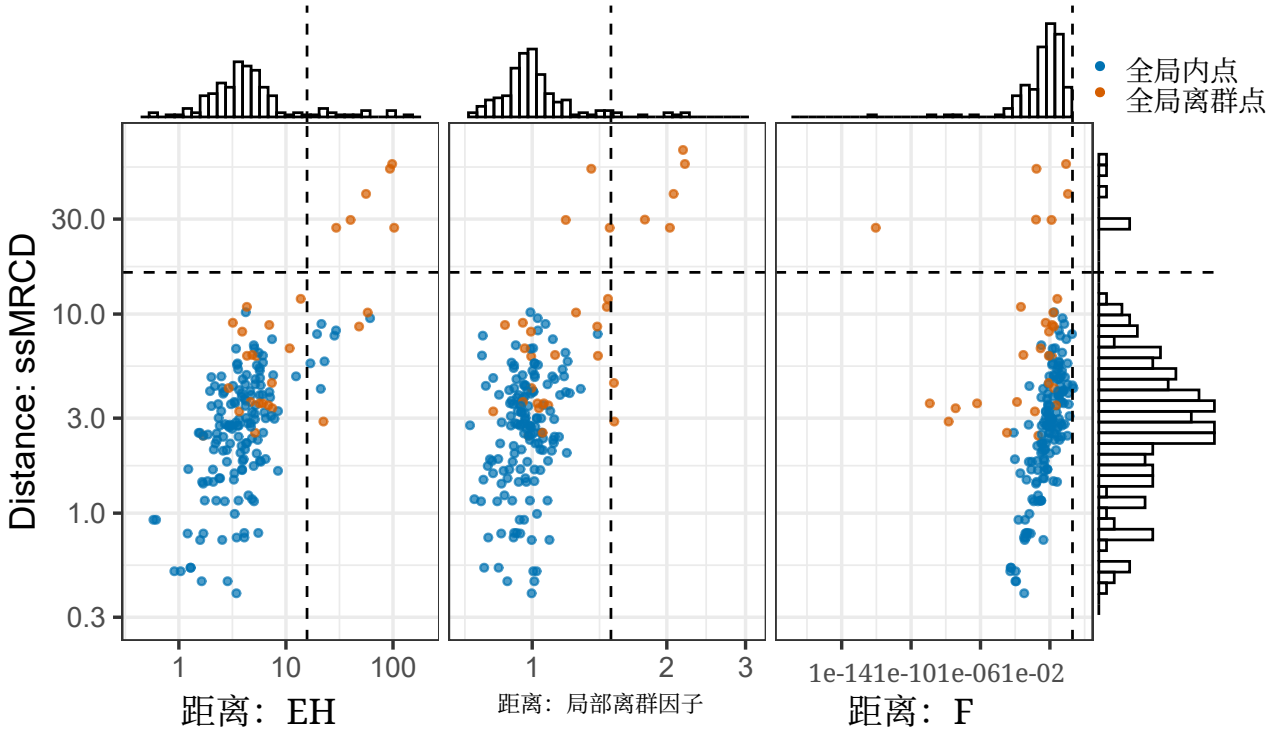

图12：距离-距离图，展示了 EH（次近距离）、LOF（局部离群因子）和 ofF（隔离度）的离群度评分与 ssMRCD 的次近距离之间的关系。橙色观测值是根据以 MCD 作为协方差估计量计算的稳健 MD 判定的全局异常值。边缘处的直方图展示了不同离群度评分的分布。

6 结论

本文通过扩展Filzmoser 等人 (2013 年) 和 Ernst 和 Haesbroeck (2016) 的方法，丰富了多变量局部异常值检测这一有限工具箱。我们开发的 ssMRCD 基于 MRCD（Boudt 等人，2020），通过采用空间平滑估计，弥合了成对马氏距离所用的完全局部与完全全局协方差矩阵之间的差距，取代了Ernst 和 Haesbroeck (2016) 所用的极端局部协方差矩阵。

我们通过最小化问题定义了 ssMRCD，并证明了该估计量的理论性质，例如等变性和崩溃点。对于所提出的算法在底层协方差矩阵发生合理空间变化时具有稳定收敛这一性质，我们给出了一个启发式论证。此外，我们比较了Filzmoser 等人(2013 年)、Ernst 和 Haesbroeck(2016) 的方法以及 ssMRCD 异常值检测方法与针对局部离群点调整的局部离群因子（Schubert 等人，2012），考察它们在模拟数据和来自奥地利气象站的真实世界数据上的异常值检测性能和计算效率。
尽管我们赞同Ernst 和 Haesbroeck (2016) 的结论，即很难为异常值检测技术选出“最佳”方法，但基于 ssMRCD 的异常值检测技术似乎是唯一一种在所有模拟情景中都能提供可靠（但仍有改进空间）结果的方法。此外，尽管计算过程相当复杂，它在运行时间方面仍能与其他方法相竞争。不过，要对真实数据进行全面分析，仍然最好采用不同的离群点检测方法并比较结果，以充分利用现有方法的所有潜在优势。比较多种方法的结果能让我们更深入地了解数据，也能更加可靠地识别出显著的局部异常值。

ssMRCD 的协方差结构也可用于局部异常值检测以外的领域。所有适用于空间数据的方法，都可以通过改用 ssMRCD 加以扩展，例如空间主成分分析。应用 ssMRCD 的一个特殊情形，可能是分析中需要考虑存在结构突变的空间数据。最后，还可以将这些思路迁移到时间序列领域：此时，空间依赖性由多元时间序列的时间依赖性取代，且依赖结构可能随时间变化。这类情形通常会给异常值检测带来相当大的挑战。

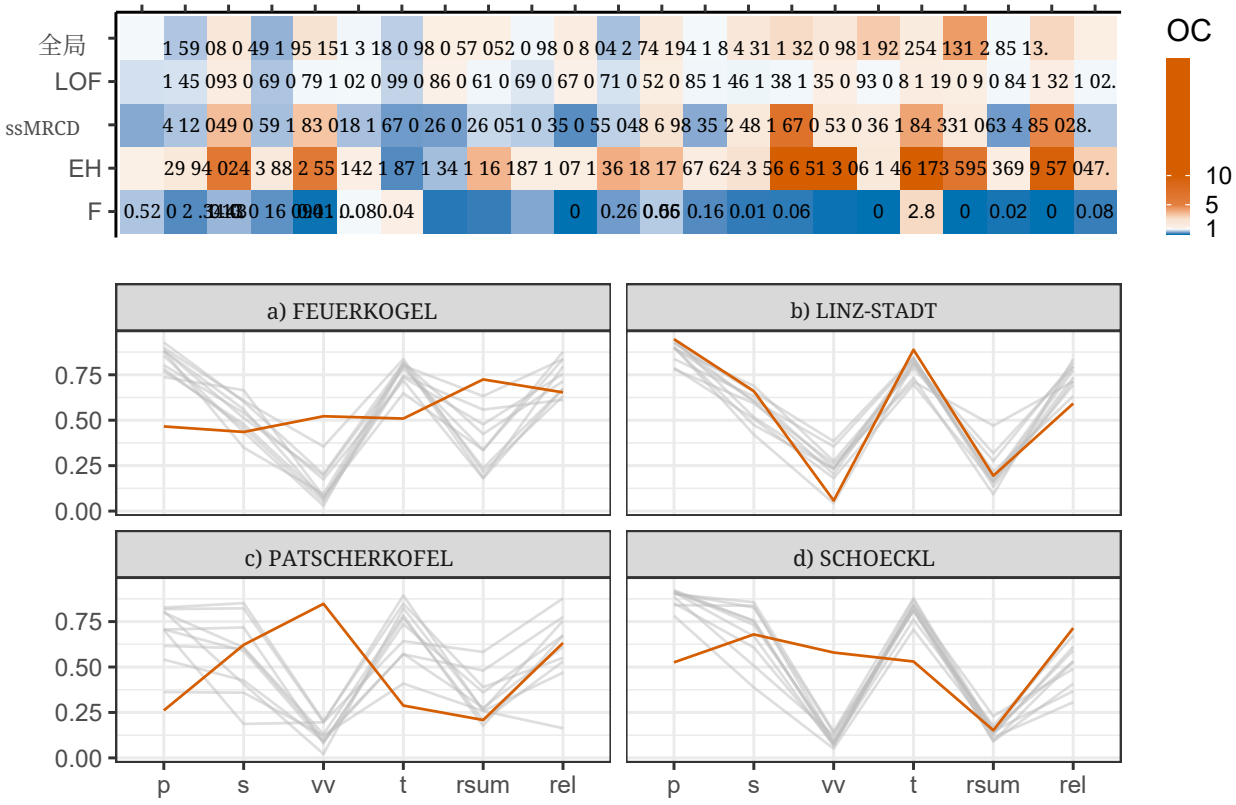

图13：上半部分：对于所有四种方法，根据气象站相对于截断值（OC）的离群度评分及其全局离群度，将判定为离群点的气象站着色。下半部分：在面板 a)-d) 中，选取了四个具有代表性的气象站，以展示不同方法之间的差异。每个变量均缩放至范围 $[0, 1]$。选定离群点的数值以橙色标示，相应的 10 个最近气象站以灰色标示。

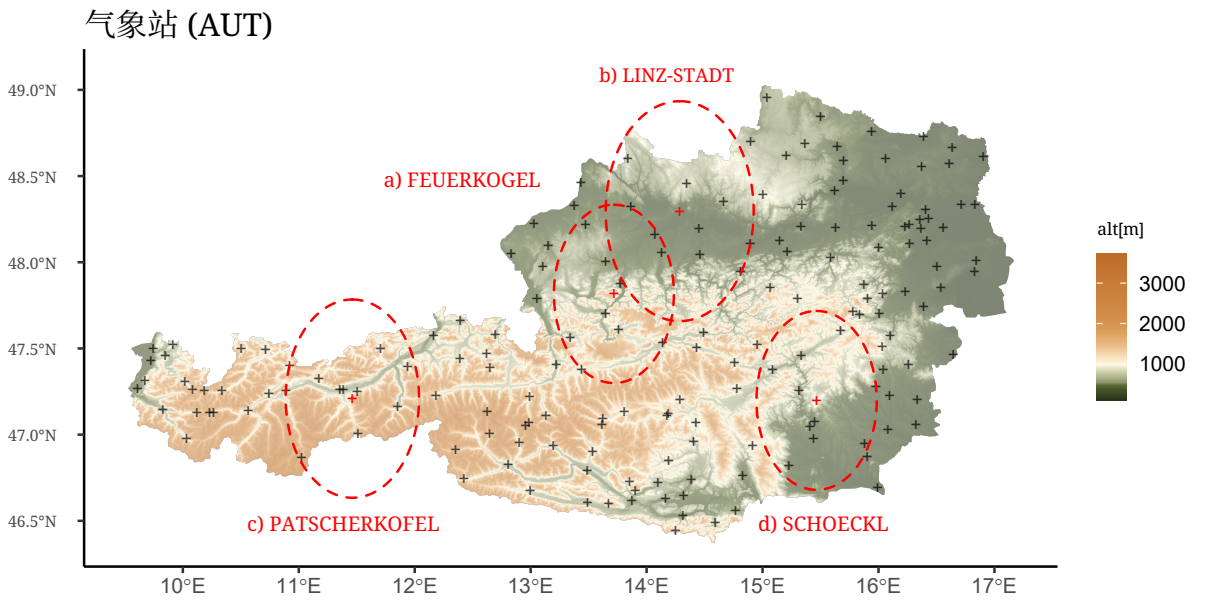

图14：奥地利的海拔地图，标示了所有气象站以及以红色标出的四个选定离群值。每个红色虚线椭圆表示与该离群点进行比较的 k 个最近邻。

致谢

由欧盟（SEMACRET，资助协议编号 101057741）和英国研究与创新署（UKRI）共同资助。

参考文献

Bellino, A., Alfani, A., Riso, L. D., and Baldantoni, D. (2019). 河流生态系统中关键问题及其潜在成因的多元空间分析。环境科学与污染研究, 27(25):30969–30976. Boudt, K., Rousseeuw, P. J., Vanduffel, S., and Verdonck, T. (2020). 最小正则化协方差行列式估计量。统计与计算, 30(1):113–128. Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). 局部离群因子。载于*2000* 年 *ACM SIGMOD* 数据管理国际会议论文集 - *SIGMOD '00*。ACM Press. Croux, C. and Haesbroeck, G. (1999). 最小协方差行列式散布矩阵估计量的影响函数与效率。多元分析杂志, 71(2):161–190. Ernst, M. and Haesbroeck, G. (2016). 空间多元数据中局部异常值检测技术的比较。数据挖掘与知识发现, 31(2):371–399. Filzmoser, P., Ruiz-Gazen, A., and Thomas-Agnan, C. (2013). 局部多元离群点的识别。统计论文, 55(1):29–47. Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J.-B., and Thirion, B. (2012). 使用稳健协方差估计量检测高维神经影像数据集中的离群点。医学图像分析, 16(7):1359–1370. GeoSphere Austria (2022). *https://data.hub.zamg.ac.at*. Gneiting, T., Kleiber, W., and Schlather, M. (2010). 多元随机场的 Matérn 交叉协方差函数。美国统计学会会刊, 105(491):1167–1177. Harris, P., Brunsdon, C., Charlton, M., Juggins, S., and Clarke, A. (2013). 使用稳健地理加权方法进行多元空间异常值检测。数学地球科学, 46(1):1–31. Hubert, M., Rousseeuw, P. J., and Verdonck, T. (2012). 一种用于稳健位置与散布估计的确定性算法。计算与图形统计学杂志, 21(3):618–637。

Hubert, M. and Vandervieren, E. (2008). 一种适用于偏态分布的调整箱线图。计算统计与数据分析, 52(12):5186–5201. Maronna, R., Martin, D., and Yohai, V. (2006). 稳健统计：理论与方法。John Wiley & Sons, 奇切斯特。Rousseeuw, P. J. (1984). 最小二乘中位数回归。美国统计学会会刊, 79(388):871–880. Rousseeuw, P. J. (1985). 具有高崩溃点的多元估计。数理统计与应用, 8(283-297):37. Rousseeuw, P. J. and Driessen, K. V. (1999). 一种用于最小协方差行列式估计量的快速算法。技术计量学, 41(3):212–223. Schubert, E., Zimek, A., and Kriegel, H.-P. (2012). 重新审视局部异常值检测：关于局部性的一种广义观点及其在空间、视频和网络异常值检测中的应用。数据挖掘与知识发现, 28(1):190–237。

附录 A 第2节（方法论）的证明

定理1（仿射等变性）。 设 $\boldsymbol{T}$ 是数据集 $\boldsymbol{X}$ 的任意稳健、正则且仿射等变的协方差估计量，在此记为 $\boldsymbol{T}(\boldsymbol{X})$。那么，以目标矩阵 $\boldsymbol{T}(\boldsymbol{X})$ 为目标的空间平滑 *MRCD* 位置和协方差估计量具有仿射等变性。

证明。 设 i 为固定值， $\boldsymbol{T}(\boldsymbol{X})$ 为上述协方差估计量， $\boldsymbol{Y}:=\boldsymbol{X}\boldsymbol{A}'+\mathbf{1}_n\boldsymbol{b}'$ 为变换后的数据矩阵。那么，对于任意子集组合 $\mathcal{H}$ 以及所有 $j=1,\ldots,N$，均有

$$\begin{aligned}\boldsymbol{K}_j^Y(\mathcal{H}) &= \rho_j\boldsymbol{T}(\boldsymbol{Y})+(1-\rho_j)Cov(\boldsymbol{Y}_{H_j})\\ &= \rho_j\boldsymbol{T}(\boldsymbol{X}\boldsymbol{A}'+\mathbf{1}_n\boldsymbol{b})+(1-\rho_j)Cov(\boldsymbol{X}_{H_j}\boldsymbol{A}'+\mathbf{1}_n\boldsymbol{b})\\ &= \rho_j\boldsymbol{A}\boldsymbol{T}(\boldsymbol{X})\boldsymbol{A}'+(1-\rho_j)\boldsymbol{A}Cov(\boldsymbol{X}_{H_j})\boldsymbol{A}'\\ &= \boldsymbol{A}\left[\rho_j\boldsymbol{T}(\boldsymbol{X})+(1-\rho_j)Cov(\boldsymbol{X}_{H_j})\right]\boldsymbol{A}'\\ &= \boldsymbol{A}\boldsymbol{K}_j^X(\mathcal{H})\boldsymbol{A}'.\end{aligned}$$

由此可得

$$\begin{aligned}&\left((1-\lambda)\boldsymbol{K}_i^Y(\mathcal{H})+\lambda\sum_{j=1,j\neq i}^{N}\omega_{ij}\boldsymbol{K}_j^Y(\mathcal{H})\right)=\\ &=\left((1-\lambda)\boldsymbol{A}\boldsymbol{K}_i^X(\mathcal{H})\boldsymbol{A}'+\lambda\sum_{j=1,j\neq i}^{N}\omega_{ij}\boldsymbol{A}\boldsymbol{K}_j^X(\mathcal{H})\boldsymbol{A}'\right)\\ &=\boldsymbol{A}\left((1-\lambda)\boldsymbol{K}_i^X(\mathcal{H})+\lambda\sum_{j=1,j\neq i}^{N}\omega_{ij}\boldsymbol{K}_j^X(\mathcal{H})\right)\boldsymbol{A}'. \qquad (9)\end{aligned}$$

利用行列式的乘法性质以及 $\det(\boldsymbol{A})\neq 0$，可知 $\boldsymbol{A}$ 在最小化问题中只是一个常数，不会影响最优子集组合的选择，

$$\begin{aligned}f(\mathcal{H}) &= \sum_{i=1}^{N}\det\left((1-\lambda)\boldsymbol{K}_i^Y(\mathcal{H})+\lambda\sum_{j=1,j\neq i}^{N}\omega_{ij}\boldsymbol{K}_j^Y(\mathcal{H})\right)\\ &= \det(\boldsymbol{A})^2\sum_{i=1}^{N}\det\left((1-\lambda)\boldsymbol{K}_i^X(\mathcal{H})+\lambda\sum_{j=1,j\neq i}^{N}\omega_{ij}\boldsymbol{K}_j^X(\mathcal{H})\right).\end{aligned}$$

结合公式 (9)，证明了协方差估计量具有仿射等变性。由于位置估计量被定义为算术平均数，而算术平均数具有仿射等变性，因此公式 (5) 中所述的性质也得以满足。 □

定理 2。 考虑固定的 $\rho>0$、正则且固定的 $\boldsymbol{T}=\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$ 以及数据矩阵 $\boldsymbol{X}_n$下的原始*MRCD*估计量。则以下结论成立：

a. MRCD 位置估计量 $\hat{}\boldsymbol{\mu}_n$ 的有限样本崩溃点为 $\min(h, n-h+1)/n$。 *b.* 原始*MRCD* 协方差估计量 $\hat{}\boldsymbol{\Sigma}_n$ 的有限样本爆炸崩溃点为 $(n-h+1)/n$.

$\hat{}$*c. MRCD* 协方差估计量 $\boldsymbol{\Sigma}_n$ 的有限样本内爆崩溃点为1。

证明。c. 关于符号，另请参见 Boudt等人（2020）。变换后的协方差矩阵$\boldsymbol{K_W}=\rho\boldsymbol{I}+(1-\rho)c_\alpha\boldsymbol{S_W}(H)$的特征值下界为$\rho>0$，其中 $\boldsymbol{S_W}(H)$是 H 的一个子集 $\boldsymbol{X}$的协方差矩阵。因此， $\lambda_i(\boldsymbol{K_W}^{-1})=\lambda_i(\boldsymbol{K_W})^{-1}\leq 1/\rho$ 以及 $\left\|\boldsymbol{K_W}^{-1}\right\|_2\leq 1/\rho$，其中 $\|.\|_2$ 表示谱矩阵范数。对于协方差矩阵，谱范数等于其最大特征值。由此可得

$$\begin{aligned}\left\|\left(\boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}\boldsymbol{K_W}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}'\right)^{-1}\right\|_2 &= \left\|\boldsymbol{Q}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{K_W}^{-1}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q}'\right\|_2\\ &\leq \left\|\boldsymbol{Q}'\boldsymbol{\Lambda}^{-1/2}\right\|_2^2\left\|\boldsymbol{K_W}^{-1}\right\|_2\\ &\leq c/\rho,\end{aligned}$$

对于 $c>0$，因为 $\boldsymbol{Q}'\boldsymbol{\Lambda}^{-1/2}$ 也是正则且固定的。这意味着

$$\lambda_i\left(\boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}\boldsymbol{K_W}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}'\right)^{-1}=\lambda_i\left((\boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}\boldsymbol{K_W}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}')^{-1}\right)\leq c/\rho$$

对于所有 $i=1,\ldots\ldots,p$。由此可得

$$\lambda_i(\boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}\boldsymbol{K_W}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}')\geq\rho/c>0\ \ \forall i=1,\ldots,p,$$

对于所有子集 H，特别是对于最优子集 H^*。因此，$\hat{}\boldsymbol{\Sigma}_n=\boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}\boldsymbol{K_W}^*\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}'$ 的特征值也远离零，从而可知内爆崩溃点为 1。

b. 显然，$\epsilon_n^*(\hat{}\boldsymbol{\Sigma}_n;\ \boldsymbol{X}_n)\leq(n-h+1)/n$，因为在这种情况下，无论其取值如何，$\hat{}$选定的子集中始终至少存在一个观测值，从而破坏估计。我们需要证明 $\epsilon_n^*(\boldsymbol{\Sigma}_n;\boldsymbol{X}_n)>(n-h)/n$。

假设 $\epsilon_n^*(\hat{}\boldsymbol{\Sigma}_n;\ \boldsymbol{X}_n)\leq(n-h)/n$。我们可以任意更改 $m\leq n-h$ 个观测值，并将所得矩阵记为 $\boldsymbol{X}_{n,m}=(\boldsymbol{x}_1^*,\ldots,\boldsymbol{x}_m^*,\boldsymbol{x}_{m+1},\ldots,\boldsymbol{x}_n)'$，其中 $\boldsymbol{x}_1^*,\ldots,\boldsymbol{x}_m^*$ 是被替换的观测值（不失一般性，将其放在前 m 行）。上确界为无穷大等价于

$$\forall C>0\ \exists\boldsymbol{X}_{n,m}:\ \ |\ln(\lambda_1(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_{n,m})))-\ln(\lambda_1(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_n)))|>C. \tag{10}$$

由于 $\ln(\lambda_1(\hat{}\boldsymbol{\Sigma}_n(\boldsymbol{X}_n))$ 是常数，且 ln 单调递增且无界，因此不失一般性，我们可以假设绝对值内的 $\hat{}$ 值非负。此外，将常数 $\ln(\lambda_1(\boldsymbol{\Sigma}_n(\boldsymbol{X}_n)))$ 移至等式右侧，公式（10）等价于最大特征值 $\lambda_1(\boldsymbol{\Sigma}_n(\boldsymbol{X}_{n,\ m}))$ 无界，

$$\forall C>0\ \exists\ \boldsymbol{x}_1^*,\ldots,\boldsymbol{x}_m^*:\ \ \lambda_1(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_{n,m}))>C. \tag{11}$$

注意，对于任意 -维矩阵，$\det(\quad\boldsymbol{X}\quad=\prod_{i=1}^p\lambda_i\ \boldsymbol{X}\quad\quad p\quad\quad\boldsymbol{X}\quad\quad\boldsymbol{\Sigma}_n\ \boldsymbol{X}_{n\ m}=1-\rho\ c_\alpha Cov\ \boldsymbol{X}_{n\ m;H^*}+)()$，$\rho\boldsymbol{T}\quad H^*\quad\boldsymbol{X}_{n\ m}\quad 1-\rho\ c_\alpha Cov\ \boldsymbol{X}_{n\ m;H}+\rho\boldsymbol{T}$

且 $\hat{}(.)()(.)$，其中 是 . 的观测值子集，该子集使 $\det(()(.))$ 在所有子集H中取最小值。如上所示，对于所有 $\boldsymbol{X}$ 和 $\rho>0$，$(1-\rho)c_\alpha Cov(\boldsymbol{X})+\rho\boldsymbol{T}$的特征值都与零保持正距离，

$$\lambda_i((1-\rho)c_\alpha Cov(\boldsymbol{X})+\rho\boldsymbol{T})\geq c>0,\ \ \forall i=1,\ldots,p.$$

对于矩阵$\boldsymbol{X}_{n,m;H^*}$，可得

$$\prod_{i=2}^p\lambda_i(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_{n,m}))\geq c^{p-1}>0.$$

˜定义常数 C 为

$$\tilde{C}=\frac{\det((1-\rho)c_\alpha Cov(\boldsymbol{X}_{n,m;\tilde{H}})+\rho\boldsymbol{T})}{c^{p-1}},$$

˜其中，$H=m+1,\ldots,m+h$ 表示 h 个固定且未改变的观测值的索引，这些观测值属于$\boldsymbol{X}_{n,m}$，其存在性由

˜ ˆ˜$m\leq n-h$保证。然后，由 C的条件 (11)，存在 $\boldsymbol{x}_1^*,\ldots,\boldsymbol{x}_m^*$ 使得 $\lambda_1(\boldsymbol{\Sigma}_n(\boldsymbol{X}_{n,m}))>C$，从而得到

$$\det((1-\rho)c_\alpha Cov(\boldsymbol{X}_{n,m;H^*})+\rho\boldsymbol{T})>\det((1-\rho)c_\alpha Cov(\boldsymbol{X}_{n,m;\tilde{H}})+\rho\boldsymbol{T}).$$

这与所选子集 H^*的行列式最小化性质相矛盾。因此，$\epsilon_n^*(\hat{}\boldsymbol{\Sigma}_n;\ \boldsymbol{X}_n)\ >(n-h)/n$。

a. 使用与前面相同的论证，显然 $\epsilon_n^*(\hat{}\boldsymbol{\mu}_n;\ \boldsymbol{X}_n)\leq(n-h+1)/n$。下面证明 $\epsilon_n^*(\hat{}\boldsymbol{\mu}_n;\ \boldsymbol{X}_n)\leq h/n$。同样，我们可以论证，$\sup\ ||\hat{}\boldsymbol{\mu}_n(\boldsymbol{X}_{n,m})\ -\hat{}\boldsymbol{\mu}_n(\boldsymbol{X}_n)||=+\infty$ 等价于

$$\forall C>0\ \exists\ \boldsymbol{x}_1^*,\ldots,\boldsymbol{x}_m^*:\ \|\hat{\boldsymbol{\mu}}_n(\boldsymbol{X}_{n,m})\|>C. \tag{12}$$

我们需要找到 $m=h$ 个被替换的数据点，使位置估计量的范数无界，同时协方差矩阵的行列式仍保持最小。对于固定的数据集 $\boldsymbol{X}_n$，我们得到观测值的最优子集 H^*，并将一个固定但可任意大的数 $L>0$ 加到这 $m=h$ 个观测值的第一个坐标上，

$$\forall i\in H^*:\tilde{x}_{i1}=x_{i1}+L\text{ and }\tilde{x}_{ij}=x_{ij}\ \forall j=2,\ldots,p.$$

因此，所选子集$\boldsymbol{X}_{n,\ m;H^*}$ 中第一坐标的样本均值等于 $\boldsymbol{X}_{n;H^*}$ 中第一坐标的原始均值加上 L。同样，由于我们取的是同一个子集，样本协方差也与之前相同

H^*，因为它不受施加于所有已使用观测值的常数平移影响。这意味着正则化协方差及其行列式也与之前相同，并且该行列式对于 $\boldsymbol{X}_n$的所有其他子集而言都是最小的。为了证明新数据矩阵$\boldsymbol{X}_{n,\ m}$ 中子集 H^* 的最小性，我们只需考虑同时包含原始观测值和交换后（任意大的）观测值的子集。

˜关于这些子集之一的第一个坐标的样本均值 H，它是

$$\tilde{M}_1 = \frac{1}{h}\left(\sum_{j=1}^{k} x_{i_j 1} + \sum_{j=k+1}^{h} \tilde{x}_{i_j 1}\right) = \frac{1}{h}\left(\sum_{j=1}^{k} x_{i_j 1} + \sum_{j=k+1}^{h} (x_{i_j 1} + L)\right) = M_1 + L - \frac{k+1}{h}L,$$

其中， M_1 是子集 $\boldsymbol{X}_{n;\tilde{H}}$ 第一个坐标的（固定）均值，且两个求和都非空。关于第一个坐标的方差（即样本协方差矩阵的第一个对角元素），我们看到

$$\begin{aligned} Cov(\boldsymbol{X}_{n,m;\tilde{H}})_{11} &= \frac{1}{h-1}\left(\sum_{j=1}^{k}(x_{i_j 1} - \tilde{M}_1)^2 + \sum_{j=k+1}^{h}(\tilde{x}_{i_j 1} - \tilde{M}_1)^2\right) \\ &= \frac{1}{h-1}\left(\sum_{j=1}^{k}\underbrace{(x_{i_j 1} - M_1 - L + \frac{k+1}{h}L)^2}_{\mathcal{O}(L^2)} + \sum_{j=k+1}^{h}\underbrace{(x_{i_j 1} + L - M_1 - L + \frac{k+1}{h}L)^2}_{\mathcal{O}(L^2)}\right) \\ &= \mathcal{O}(L^2) \end{aligned}$$

因此， $Cov(\boldsymbol{X}_{n,m;\tilde{H}})$的弗罗贝尼乌斯范数及其正则化分别为 $\mathcal{O}(L^2)$，由此可知，对于某个常数$\beta > 0$,

$$\begin{aligned} \mathcal{O}(L^2) &= \left\|(1-\rho)c_\alpha Cov(\boldsymbol{X}_{n,m;\tilde{H}}) + \rho\boldsymbol{T})\right\|_F \\ &\leq \beta\left\|(1-\rho)c_\alpha Cov(\boldsymbol{X}_{n,m;\tilde{H}}) + \rho\boldsymbol{T})\right\|_2 \\ &= \beta\lambda_1((1-\rho)c_\alpha Cov(\boldsymbol{X}_{n,m;\tilde{H}}) + \rho\boldsymbol{T}) \\ &\leq \beta\frac{\det((1-\rho)c_\alpha Cov(\boldsymbol{X}_{n,m;\tilde{H}})}{c^{p-1}}, \end{aligned}$$

这是由于有限维空间中矩阵范数的等价性，且 c 是上文中的常数。任意增大 L，我们便看到，混合子集对应的行列式大于仅由交换后的观测值构成的最优子集 H^* 的行列式。

现在假设 $\epsilon_n^*(\hat{\boldsymbol{\mu}}_n;\ \boldsymbol{X}_n) = m/n < \min(h, n-h+1)/n$，并从公式（12）出发，

$$\forall C > 0\ \exists\, \boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_m^* : \|\hat{\boldsymbol{\mu}}_n(\boldsymbol{X}_{n,m})\| > C.$$

这意味着

$$\forall C > 0\ \exists\, \boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_m^* : \sum_{j=1}^{k}\left\|\boldsymbol{x}_{i_j}\right\| + \sum_{j=k+1}^{h}\left\|\boldsymbol{x}_{i_j}^*\right\| \geq \left\|\left(\sum_{j=1}^{k}\boldsymbol{x}_{i_j} + \sum_{j=k+1}^{h}\boldsymbol{x}_{i_j}^*\right)\right\| > C,$$

其中 $i_j \in H^*, j = 1, \ldots, h$。因此，对于所有 $C > 0$，都存在某个 $\boldsymbol{x}_{i_j}^*$，其范数大于 C。不失一般性，假设它是 $\boldsymbol{x}_1^*$，且第一个坐标起决定作用，

$$\forall C > 0\ \exists\, \boldsymbol{x}_1^* \in H^* : |x_{11}^*| \geq C.$$

对于 $m < h < n-h+1$，不可能只将交换点纳入子集中，而且总是可以得到一个包含 h 个原始观测值的子集。对于 $m < n-h+1 < h$，情况也是如此。因此， H^*中至少有一个交换点 $\boldsymbol{x}_1^*$ 和一个原始点。但运用与之前相同的论证，由原始点和任意大的点组成的混合子集的行列式最终会与最优性相矛盾，因为某个点的行列式会大到足以使得存在一个包含 h个原始观测值的子集，其行列式更小。

□

定理3。 空间平滑*MRCD*的位置估计量 $\left(\hat{\boldsymbol{\mu}}_{SSM,n,i}\right)_{i=1}^N$ 的有限样本崩溃点为

$$\epsilon_n^*(\left(\hat{\boldsymbol{\mu}}_{SSM,n,i}\right)_{i=1}^N; \boldsymbol{X}_n) = \min_{i=1,\ldots,N} \min(n_i - h_i + 1, h_i)/n_i.$$

证明。 对于 $i = 1, \ldots, N$，位置估计是从邻域a_i 中选出的 h_i 个观测值的标准样本均值，选取方式旨在使目标函数（3）最小。将一个邻域 a_i中的观测值替换为任意大的值，同时保持其他邻域不变（使矩阵 $\boldsymbol{K}_j$ 保持有界），我们可以将定理2关于MRCD结构化协方差矩阵 $\boldsymbol{K}_i$的结果应用于此。 $\boldsymbol{K}_i$ 的崩溃点为 $\min(n_i - h_i + 1, h_i)/n_i$。 因此，要使至少一个位置估计量失效，我们需要将一个邻域中的观测值替换掉 $\min_{i=1,\ldots,N} \min(n_i - h_i + 1, h_i)/n_i$ 这一比例。 □

定理 4。 给定一个固定且正则的目标矩阵 $\boldsymbol{T}$，空间平滑 *MRCD* 协方差估计量 $\left(\hat{}\boldsymbol{\Sigma}_{SSM,\ n,\ i}\right)_{i=1}^{N}$ 的有限样本内爆崩溃点等于

$$\epsilon_n^*(\left(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}\right)_{i=1}^{N}; \boldsymbol{X}_n) = 1.$$

有限样本爆炸崩溃点是

$$\epsilon_n^*(\left(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}\right)_{i=1}^{N}; \boldsymbol{X}_n) = \min_{i=1,\ldots,N}(n_i - h_i + 1)/n_i.$$

证明。 由于空间平滑 MRCD 协方差估计量 $\boldsymbol{K}_i$ 按照 MRCD 方法在每个邻域上进行正则化，只要目标矩阵 $\boldsymbol{T}$ 是正则的，所有特征值就都为正且远离零（见定理2）。因此，协方差估计量永远不会奇异，内爆崩溃点为1。

ˆ 对于第二部分，我们固定邻域 a_i。注意，协方差估计量定义为 $\boldsymbol{\Sigma}_{SSM,n,i} = (1-\lambda)\boldsymbol{K}_i(\mathcal{H}^*) + \lambda\sum_{j=1,\ j\neq i}^{N}\omega_{ij}\boldsymbol{K}_j(\mathcal{H}^*)$，其中最优子集为 $\mathcal{H}^*$。矩阵 $\boldsymbol{K}_i = \boldsymbol{K}_i(\mathcal{H}^*)$ 按照 MRCD 的方式构造，基于子集 H_i^* 的样本协方差矩阵和目标矩阵。由于我们假设 $\boldsymbol{T}$ 固定，对于 $\boldsymbol{K}_i$，只有在与 MRCD 相同的条件下，才能得到任意大的特征值（见定理2）。因此，将 $(n_i - h_i + 1)/n_i$ 替换为任意值，可导致 $\boldsymbol{K}_i$ 的特征值任意大。对于单个邻域协方差估计量 $\boldsymbol{\Sigma}_{SSM,n,i}$ 的爆炸崩溃点（在 $\boldsymbol{W}$ 和 λ 的一般设置下），只需至少一个 $\boldsymbol{K}_j$ 达到其崩溃点即可。因此，$\boldsymbol{\Sigma}_{SSM,n,i}$ 的有限样本爆炸崩溃点为 is

$$\epsilon_n^*(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}; \boldsymbol{X}_n) = \min_{i=1,\ldots,N}\{(n_i - h_i + 1)/n_i\}. \tag{13}$$

由于 $\epsilon_n^*(\hat{}\boldsymbol{\Sigma}_{SSM,\ n,\ i};\ \boldsymbol{X}_n)$ 已经独立于 i，因此空间平滑 ˆ MRCD 协方差估计量的总体爆炸崩溃点等于 $\epsilon_n^*(\boldsymbol{\Sigma}_{SSM,\ n,\ i};\boldsymbol{X}_n)$。 □

B 第 3 节证明（算法和 C步）

定理 *5.* 对于每个 $j = 1, \ldots, N$，令 $\rho_j \in (0, 1)$，并令 H_j^0 为 a_j 的任意初始子集，其大小分别为 $h_j \in (n_j/2, n_j)$。令 α_j 为所用观测值的相应比例， $\alpha_j = h_j/n_j$。令 $\mathcal{H}^0 = (H_1^0, \ldots, H_N^0)$ 为这些子集的组合，且 $\lambda \in [0, 1)$ 固定。假设目标矩阵 $\boldsymbol{T}(\boldsymbol{X})$ 正定、对称且固定，并按照方程 *(2)* 定义 $\boldsymbol{K}_j(\mathcal{H}^0)$，其中 $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$。根据各自的子集，计算每个邻域 $a_j, j = 1, \ldots, N$ 的样本均值， $\boldsymbol{m}_j(\mathcal{H}^0) = \frac{1}{h_j}\sum_{k\in H_j^0}\boldsymbol{x}_k$。

固定邻域 a_i。对于 $\boldsymbol{x} \in a_i$，定义由子集 $\mathcal{H}^0$ 给出的基于马氏距离的度量为

$$d(\boldsymbol{x}; \mathcal{H}^0) = (\boldsymbol{x} - \boldsymbol{m}_i(\mathcal{H}^0))' \left[(1-\lambda)\boldsymbol{K}_i(\mathcal{H}^0) + \lambda\sum_{j=1,j\neq i}^{N}\omega_{ij}\boldsymbol{K}_j(\mathcal{H}^0)\right]^{-1}(\boldsymbol{x} - \boldsymbol{m}_i(\mathcal{H}^0)).$$

对于大小为 h_i 的 a_i 的新子集 H_i^1，满足

$$\sum_{k\in H_i^1} d(\boldsymbol{x}_k; \mathcal{H}^0) \leq \sum_{k\in H_i^0} d(\boldsymbol{x}_k; \mathcal{H}^0),$$

记为 $\tilde{\mathcal{H}} = (H_1^0, \ldots, H_i^1, \ldots, H_N^0)$ （注意，对于 $j \neq i$， $\boldsymbol{K}_j(\tilde{\mathcal{H}}) = \boldsymbol{K}_j(\mathcal{H}^0)$）。那么，

$$\det\left((1-\lambda)\boldsymbol{K}_i(\tilde{\mathcal{H}}) + \lambda\sum_{j=1,j\neq i}^{N}\omega_{ij}\boldsymbol{K}_j(\tilde{\mathcal{H}})\right) \leq \det\left((1-\lambda)\boldsymbol{K}_i(\mathcal{H}^0) + \lambda\sum_{j=1,j\neq i}^{N}\omega_{ij}\boldsymbol{K}_j(\mathcal{H}^0)\right)$$

当且仅当 $\boldsymbol{K}_i(\tilde{\mathcal{H}}) = \boldsymbol{K}_i(\mathcal{H}^0)$ 且 $\boldsymbol{m}_i(\tilde{\mathcal{H}}) = \boldsymbol{m}_i(\mathcal{H}^0)$。

证明。 该证明基本沿袭了Boudt 等人 (2020）的思路。邻域 a_i 是固定的。因此，关于 i 需要最小化其行列式的矩阵为

$$
\begin{aligned}
\boldsymbol{A}_1 &:= \left((1-\lambda)\boldsymbol{K}_i(\mathcal{H}^0) + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}^0) \right) \\
&= \left((1-\lambda)[(1-\rho_i)c_{\alpha_i} Cov(\boldsymbol{X}_{H_i^0}) + \rho_i \boldsymbol{T}] + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}^0) \right) \\
&= \left(\underbrace{(1-\lambda)(1-\rho_i)c_{\alpha_i}}_{:=\tilde{\rho}} Cov(\boldsymbol{X}_{H_i^0}) + \underbrace{(1-\lambda)\rho_i \boldsymbol{T} + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}^0)}_{:=\boldsymbol{\Omega}} \right) \\
&= \tilde{\rho}\, Cov(\boldsymbol{X}_{H_i^0}) + \boldsymbol{\Omega}, \\
\boldsymbol{A}_2 &:= \left((1-\lambda)\boldsymbol{K}_i(\tilde{\mathcal{H}}) + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\tilde{\mathcal{H}}) \right) \\
&= \tilde{\rho}\, Cov(\boldsymbol{X}_{H_i^1}) + \boldsymbol{\Omega},
\end{aligned}
$$

其中$\boldsymbol{\Omega}$ 是一个固定的正定协方差矩阵。

由于原始证明并未限定为凸线性组合，因此我们可以采用相同的证明方法，以矩阵$\boldsymbol{A}_1$、 $\boldsymbol{A}_2$ 和$\boldsymbol{\Omega}$ 替代 $\boldsymbol{K}_1$、 $\boldsymbol{K}_2$ 和 $\boldsymbol{\Lambda}$，并使用相应的因子

$$
\begin{aligned}
w_j &= \sqrt{k\tilde{\rho}/h_i}, && j = 1, \ldots, h_i \\
&= \sqrt{k/(p+1)}, && j = h_i + 1, \ldots, k.
\end{aligned}
$$

其中 $k = h_i + p + 1$.[1] □

[1]注意，这个证明可以推广到任何形式的线性组合，其中包含固定矩阵和某个子集的样本协方差矩阵。