arXiv:2302.02156v1 [stat.ME] 4 Feb 2023

单元格异常值的挑战

Jakob Raymaekers
荷兰马斯特里赫特大学定量经济学系

Peter J. Rousseeuw 鲁汶大学统计学与数据科学系，比利时

2023年2月4日

摘要

众所周知，现实世界数据通常包含离群值。“离群值”一词通常指一个观测，即 $n \times d$ 数据矩阵中的一行。近年来，另一类离群值逐渐受到关注，即单元格异常值。这些可疑单元格（条目）可能出现在数据矩阵的任何位置。即使离群单元格所占比例相对较小，也可能污染半数以上的行，这对按行稳健方法而言是一个问题。本文讨论单元格异常值带来的挑战，以及迄今为止为应对这些挑战而提出的一些方法。文章针对位置、协方差和回归的按单元崩溃值给出了新的结果。本文提出了一种用于对应分析的按单元格稳健方法，并通过现实世界数据进行了说明。最后，本文提出若干值得讨论的问题。

关键词：异常检测、崩溃值、对应分析、多元统计、稳健性。

1 引言

现实世界数据通常包含一些不符合大多数数据所呈现模式的元素。这些离群值可能是严重错误，足以严重扭曲统计分析；但它们也可能是具有价值的信息，值得进一步检查。因此，无论离群值是由什么原因造成的，我们都希望能够检测出它们。

稳健统计的方法（例如 Huber（1964）和 Hampel等人（1986）的形式化方法）是先进行稳健拟合，即构建一个能够很好地拟合大多数数据的模型，然后寻找偏离该拟合的观测。这种方法的一个重要基础假设是按案例污染：一个样本要么是异常值，要么完全未受污染。

这也称为按行污染，因为样本通常以数据矩阵的行来表示。按案例污染有多种形式化定义，但最常见的定义是假设观测数据以概率 $1-\varepsilon > 0.5$ 来自干净分布F，并以概率 $\varepsilon < 0.5$ 来自任意分布 H。换句话说，我们观察到 $\boldsymbol{X}$ ，定义如下

$$\boldsymbol{X} = (1-B)\boldsymbol{Y} + B\boldsymbol{Z} \tag{1}$$

其中， $B \sim$ 伯努利(ε)取值为 0 或 1， $\boldsymbol{Y} \sim F$， $\boldsymbol{Z} \sim H$，且 B 与 $\boldsymbol{Y}$和 $\boldsymbol{Z}$相互独立。这称为*Tukey-Huber* 污染模型 （THCM）。一种常用的等价表述为 $\boldsymbol{X} \sim F_\varepsilon := (1-\varepsilon)$ $F + \varepsilon H$。总体目标是在只观测到 F_ε ，且重要的是，不对 H作任何假设的情况下，估计 F 的特征。

基于这一模型，人们开发了种类繁多的稳健统计方法，应用于位置与协方差估计、回归模型拟合、主成分分析与判别分析，以及时间序列分析等领域，参见 Hampel 等人（1986）、Rousseeuw 和 Leroy（1987年）以及 Maronna 等人（2019年）的著作。

在 THCM 模型下，假设一个观测要么来自一个任意分布 H ，该分布可能与 F毫无关系；要么是从 F中抽取的一个完全符合分布的样本。因此，基于该模型开发的方法往往要么信任某个观测的所有坐标，要么同时降低其所有坐标的权重。但也存在这样的情况：某个观测的一个或多个坐标（条目、单元格）可能有问题，而其他坐标则正常。若将这类观测的整体权重降低，我们可能会丢失其未受污染单元格中蕴含的宝贵信息。

这解释了为什么如今还在考虑另一种污染模型，即单元格异常值模型。该模型最早由 Alqallaf 等人（2009）发表。他们假设我们观测到的随机变量 $\boldsymbol{X}$ 由下式给出

$$\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{B})\boldsymbol{Y} + \boldsymbol{B}\boldsymbol{Z}$$

其中， $\boldsymbol{B}$ 是一个对角矩阵，其对角元素只能取零或一。这使得观测到的随机向量 $\boldsymbol{X}$ 的某些分量可能受到污染，而其他分量则保持清洁。因此，该模型假设数据最初由目标分布 F 生成，但随后通过将某些单元格替换为其他值而受到污染。在给定的一行中，可能没有单元格被替换，也可能有少数单元格被替换，甚至所有单元格都被替换。图1展示了左图中的按观测（按行）模型与右图中的按单元格模型之间的差异。

单元格异常值体现了从行异常值到按单元格范式的转变。可以说，最显著的变化是，少量污染单元就可能污染很大比例的行。例如，假设每个单元格被污染的概率为ε ，且所有单元格是否被污染相互独立。在这种情况下，

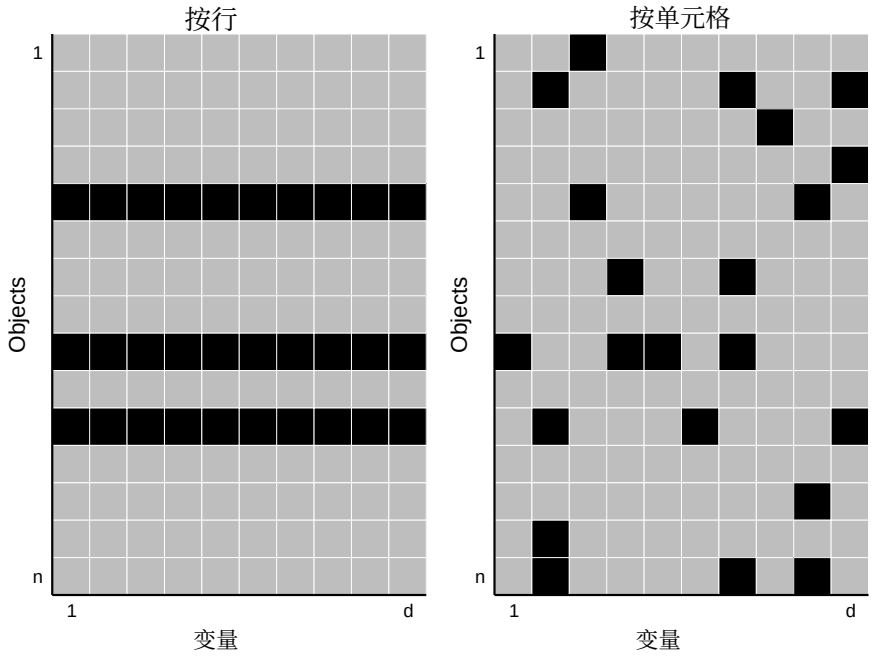


图1：按行离群值模型（左）与按单元格离群值模型（右）的对比。黑色表示离群。少量单元格异常值就可能污染许多行。

一个样本（行）至少有一个单元格受到污染的概率等于

$$P[\text{row is contaminated}] = 1 - (1 - \varepsilon)^d \tag{2}$$

它会随着维度 d 的增加而迅速增大。例如，在 $d = 15$ 维度中，比例为$\varepsilon = 0.05$ 的污染单元就足以使平均超过 50% 的行受到污染，此时为 THCM模型开发的方法将不再可靠。

早在单元格异常值这一概念于 Alqallaf 等人（2009）发表之前，就已经有人非正式地讨论过这个问题。我们中的一位清楚地记得，20 世纪 80 年代初，Alfio Marazzi 和 Werner Stahel 在苏黎世联邦理工学院的研讨会上提出过这一话题。当时它被视为一个尚待解决的问题，其中包括公式（2）。当时普遍认为，现有工具不足以应对这一挑战。

本文其余部分安排如下。第 2 节概述了针对各种情形下按单元格异常值所发展的方法。第 3 节证明了关于按单元崩溃值的一些新结果。接着，第 4 节提出了首个用于对应分析的按单元格稳健方法，并通过两个真实数据示例加以说明。随后，第 5 节评估按单元格稳健性的当前研究现状，并提出若干值得讨论的问题。第 6 节总结全文。

2 处理单元格异常值的方法 s

本节概述了为处理按单元格异常值而发展的方法。我们首先讨论“点云”，即单类别多变量数值数据。随后，我们讨论其他模型，包括回归、分类、主成分分析和聚类。

2.1 检测单元格异常值

检测单元格异常值是一项棘手的任务。最直接的检测方法是分别检查每个变量中的离群值。一种简单的方法是对每个变量进行稳健标准化，得到稳健 z分数，然后将这些 z分数与参考分布（例如标准高斯分布）的分位数进行比较，以判断是否应将某个单元标记为异常。检测边际离群值的一种更复杂的方法是使用单变量滤波器，例如 Gervini 和 Yohai（2002）提出的方法。无论采用哪种方法，为了涵盖范围更广、可能不对称的分布，都可以先像 Raymaekers 和 Rousseeuw（2021c年）那样，对变量进行稳健变换，使其具有中心正态性。

虽然标记边际离群值很快，但只有相当极端的单元格异常值才能被可靠地识别出来，因为这种方法没有考虑变量之间的关系。例如，考虑图2中的双变量情形。观测2有一个边际离群值$x_{2,\ 2}$，而观测3有一个边际离群值$x_{3,\ 1}$。但对于观测4，情况就不那么明朗了。尽管相对于大多数数据点所呈现的模式而言，它显然是一个离群观测——在这些数据点中，X_1 和 X_2 呈负相关——但边际方法并未发现其$x_{4,\ 1}$ 或$x_{4,\ 2}$ 有什么异常。这说明，一般而言，仅标记边际离群值是不够的。

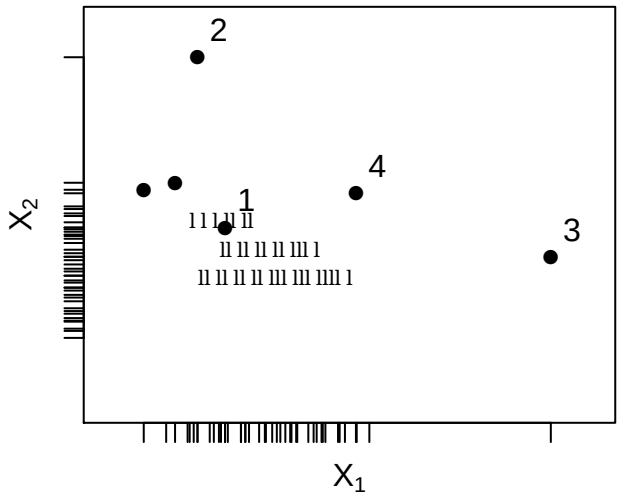


图2：单元格异常值未必是边际离群值。

此外，仅凭图2无法判断观测4是否在 X_1 或 X_2 中存在单元格异常值，或者两者中都存在。但请注意，如果$x_{4、1}$或$x_{4、2}$

是单元格异常值，那么如果数据中还有更多与 X_1 和 X_2相关的变量，或许就能作出判断。这表明，在这种特定情况下，维数灾难或许反而是一种福气。

这一思想促成了检测单元格异常值的方法，这些方法考虑的不仅仅是边缘分布。其中一个方向是扩展 Gervini 和 Yohai（2002）的单变量 GY 滤波器。Leung 等人（2017）将其扩展为双变量滤波器，用于位置和协方差的估计。最近，Saraceno 和 Agostinelli（2021）进一步将其推广到了两个以上的维度。

另一种方法是 Rousseeuw 和 Van den Bossche（2018）提出的偏离单元格检测（**DDC**）算法。该方法使用稳健简单线性回归，将每个变量对数据中的其他变量进行回归。综合这些信息，可以得到每个单元格的预测值。原始数据集与其预测值之间的差异为每个单元格提供了一个残差。随后，可以通过较大的标准化单元格残差来标记单元格异常值。这种方法受到了一定关注；事实上，滤波方法最成功的版本都将 **DDC** 纳入其滤波器中。Walach 等人（2020）通过汇总成对对数比上的偏离度，针对代谢组学数据采用了类似的方法。

最近，一些方法开始尝试超越双变量关系的组合，将数据中的所有关系纳入考量。Raymaekers 和 Rousseeuw（2021a年）依次考察每个观测，并寻找一个向量参数，以捕捉将该观测纳入整体所需的偏移。通过 L^1 正则化，他们对该偏移施加稀疏性约束，使得实际发生偏移的仅是该观测中少数几个单元格。这种算法称为**cellHandler**，已被证明能有效识别复杂的对抗性单元格异常值。**cellHandler** 的输入是数据矩阵，以及下一小节所述的该矩阵协方差矩阵的稳健估计值之一。

d	单变量 GY	双变量 GY	DDC	cellHandler
5	0.0007	0.0027	0.0005	0.0076
10	0.0012	0.0087	0.0012	0.0186
20	0.0037	0.0304	0.0033	0.0674
50	0.0059	0.1672	0.0150	0.5594

表 1：标记单元格异常值：数据集的计算时间（秒），其中 $n = 1000$且维度为 d。

表 1 比较了几种检测单元格异常值的方法的计算时间。样本量始终为 $n = 1000$，维度 d 从 5 到 50 不等。数据来自均值为零、协方差矩阵为$\Sigma_{jk} =(-0.9)^{|j-k|}$ 的高斯分布，其中 10% 的单元格被替换为数值 5。所有计算均在 R 中完成，报告的时间是 10 次重复实验的平均值。正如预期，单变量滤波器速度最快。双变量 GY 滤波器耗时更长，与.....相比也是如此

DDC。**cellHandler** 方法相对而言最慢，但仍然相当快 st.

由于检测单元格异常值的方法有多种，因此在具体情形下如何选择，了解一些指导原则会有所帮助。如果有理由认为所有单元格异常值在边际上都属于离群值，那么使用单变量滤波器就足够了。如果预计它们会在双变量图中显得突出，例如可以使用双变量 GY 滤波器（Leung 等人，2017）。不过，为了减少计算量，也可以运行 **DDC**，它综合了多个变量的信息，并为标记单元格提供预测值。**cellHandler**也能做到这一点。要在后两者之间做出选择，需要注意，**cellHan-dler** 需要稳健协方差估计，该估计假设 $n > p$，并且还假设数据形状大致呈椭圆形。**DDC** 不作这些假设，速度也更快；但在适用的情况下，**cellHandler** 检测复杂异常值的准确度更高。

一种有些相关的方法是 Debruyne 等人（2019）提出的 **SPADIMO** 算法。它并非纯粹的按单元格方法，因为它首先根据稳健距离较大这一特征，从按案例稳健位置和协方差矩阵中识别离群案例。接下来，它尝试找出导致这些案例偏离的单元格。为此，它使用正则化回归来识别偏离度最大的稀疏方向。

2.2 位置与协方差的估计

研究单元格异常值时，如果说有什么令人安心的发现，那大概就是从稳健性的角度来看，通过结合边际分布上的稳健位置估计，仍然能够可靠地估计多元位置。一个典型例子是逐坐标中位数，定义如下：

$$\hat{\boldsymbol{\mu}}(\boldsymbol{X}) = (\underset{i=1}{\overset{n}{\text{median}}}\, x_{i1}, \ldots, \underset{i=1}{\overset{n}{\text{median}}}\, x_{id}) \tag{3}$$

对于包含 n 行和 d 列的数据集 $\boldsymbol{X}$。虽然这种逐分量位置估计未必总是具有最理想的性质，但至少在一般条件下是一致的，而且非常稳健。变量尺度的估计也是如此，换言之，就是协方差矩阵的对角线元素。

然而，一旦我们离开估计多元位置和逐分量尺度这一舒适区，事情就会变得更具挑战性。对于协方差矩阵的单元格级稳健估计，Van Aelst 等人（2011年）提出了逐单元 Stahel-Donoho 估计量。这是对按案例 Stahel-Donoho 协方差估计量的扩展（Stahel, 1981年；Donoho，1982）。该方法在按案例污染下表现尚可，但在逐单元污染情形下，如 Agostinelli 等人（2015年）的图1和 Rousseeuw 和 Van den Bossche（2018年）的图9所示，其表现已不及近期提出的方法。

Danilov（2010）深入研究了在估计协方差矩阵之前先对数据进行筛选的方法。他考察了一个三步流程：（i）检测受污染的单元格；（ii）以某种方式处理这些单元格，以减弱其影响；（iii）估计

位置和协方差矩阵。在检测方面，他考察了基于稳健 Z 分数的边际方法，以及基于部分马氏距离的多变量方法。最终，他提出采用后者，但指出，若要在所有可能的子空间中识别按单元格的异常值，计算时间会随维度 d 呈指数级增长。对于离群单元的处理，他比较了温莎化、删失和将其设为缺失值这几种方法。最终方案采用多变量单元格检测方法，随后将离群单元设为缺失值，再对所得不完整数据集进行最大似然估计。沿着这一思路，Agostinelli 等人（2015年）提出了两步广义 S 估计量（2SGS），Leung 等人（2017）对其进行了扩展。在此后数年间，它一直是表现最佳的方法。2SGS方法结合了双变量滤波器与缺失数据广义 S 估计量（Danilov 等人，2012），并已在 R 软件包 `GSE`中实现（Leung 等人，2019年）。

近年来，针对对抗性单元格异常值条件下协方差矩阵的估计，出现了一些新的尝试。Raymaekers 和 Rousseeuw（2021a年）提出了一种迭代方法：（i）运行**cellHandler**标记离群单元，（ii）将这些单元设为缺失值，以及（iii）应用期望最大化算法重新估计协方差矩阵。尽管这种方法表现相当不错，但它本质上纯粹是一种算法，没有易于处理的目标函数，因此其性质难以分析。不过，基于这些思路，Raymaekers 和 Rousseeuw（2022b）提出了 Rousseeuw（1984）广为人知的案例稳健最小协方差行列式（MCD）估计量的单元格稳健版本。新的 cellMCD 方法确实会优化一个目标函数，并且已证明在逐单元污染下具有较高的崩溃值。其算法会反复执行某种 C 步骤，每一步都会降低目标函数（与案例稳健 MCD 的 C 步骤一样）。因此，该算法保证会收敛到一个解，但该解可能是局部极小值。初始估计量结合了 Raymaekers 和 Rousseeuw（2021b年）的稳健相关系数估计量与**DDC**，不过也可以采用其他初始估计。本文段所述的方法已在 R软件包 `cellWise` （Raymaekers 和 Rousseeuw，2022a）中实现。

d	经典	MCD	S	2SGS	cellMCD
5	0.0000	0.0423	0.1448	0.5730	0.2403
10	0.0002	0.0950	0.3145	1.6476	0.3733
20	0.0002	0.3065	1.0548	15.206	1.3153
50	0.0015	2.0049	8.0532	213.02	24.770

表 2：位置和协方差的估计：含有 $n = 1000$ 且维度 d 变化的数据集的计算时间（秒）。

表 2 列出了经典均值和协方差矩阵的计算时间，以及案例稳健 MCD 和 S 估计量的计算时间，并将它们与按单元格稳健方法 2SGS 和 cellMCD 进行了比较。设置与表 1 相同，所有方法

均使用 R 的默认版本运行。值得注意的是，按单元格稳健方法耗时最长，但仍在可行范围内。2SGS 方法最慢，部分原因是其默认初始估计量可以替换为其他估计量。

Rousseeuw（2023年）最近提出了一种利用单元级权重对数据进行统计分析的方法，并将其用于位置和协方差的估计。一种可能的应用是，先根据单元格的离群程度（例如依据标准化残差）为其赋予权重，然后通过单元级加权最大似然法计算重新加权位置和协方差矩阵。

2.3 精度矩阵的估计

许多应用主要关注的是协方差矩阵的逆矩阵，也称为精度矩阵。在低维情况下，当然可以先估计协方差矩阵，再对估计结果求逆；但在高维情况下，这种做法就行不通了，因为估计得到的协方差矩阵往往是奇异的。此时，与其对协方差矩阵进行正则化，不如直接估计精度矩阵并同时对其进行正则化。这样会使精度矩阵中出现零值，对应于偏相关为零，而这又对应于高斯图模型中的条件独立。高维精度矩阵最常用的估计量之一是 Friedman 等人（2008）的图形套索。该方法需要输入一个协方差矩阵的估计值$\widehat{\Sigma}$，通常是经典的经验协方差矩阵。随后，精度矩阵 Θ 的估计方式如下

$$\hat{\Theta} = \underset{\Theta \succcurlyeq 0}{\operatorname{argmin}} \left(\operatorname{trace}\left(\widehat{\Sigma}\Theta\right) - \log(\det(\Theta)) + \lambda \sum_{j \neq \ell} |\Theta_{j\ell}| \right) . \tag{4}$$

一个自然的问题是，对 Σ 采用稳健版本是否会得到稳健的精度矩阵估计量，答案确实是肯定的。Ollerer 和 Croux (2015)、Croux 和 Ollerer (2016)、Tarr 等人 (2016)、Loh 和 Tan (2018) 以及 Katayama 等人 (2018) 使用基于稳健成对相关的不同 Σ，从经验和理论层面研究了这一思路。这些研究中的 Σ通常具有以下形式

$$\widehat{\Sigma}_{j\ell} = \mathrm{s}(\boldsymbol{X}_{\cdot j})\, \mathrm{s}(\boldsymbol{X}_{\cdot \ell})\, \mathrm{r}(\boldsymbol{X}_{\cdot j}, \boldsymbol{X}_{\cdot \ell}) \tag{5}$$

其中，s(·)表示稳健尺度估计量，通常为Rousseeuw和Croux（1993）的 Q_n 估计量；r(·)表示稳健相关系数估计量，例如Spearman秩相关。

请注意， Σ 必须是半正定的，式（4）才能成立。当 Σ 不是半正定时，首先要对其进行修正，例如加上单位矩阵的一个较小倍数。由此得到的精度矩阵估计量在非对抗性逐单元污染下表现良好，计算速度快，并且由于目标函数（4）定义明确，因此可以进行理论分析。但由于它们关注的是成对相关，因此在面对更具对抗性的单元格异常值时，其性能可能会下降。

2.4 回归

含有单元格异常值时的回归，主要是在数值型预测变量和响应变量的线性模型中进行研究的。更具体地说，我们将观测到的配对值（$\boldsymbol{x}_1$，y_1）.....，（$\boldsymbol{x}_n$，y_n）建模，其中 $\boldsymbol{x}_i$ 是一个 $p \times 1$ 向量，适用于 $i = 1$，.....，n，如下：

$$y_i = \alpha + \boldsymbol{\beta}\boldsymbol{x}_i + \text{noise}_i \tag{6}$$

其中需要估计回归系数 α 和 $\boldsymbol{\beta} = (\beta_1、.....、\beta_p)$。

在按案例稳健线性回归中，观测到的一对变量 $(\boldsymbol{x}_i,\ y_i)$ 被视为要么离群，要么完全干净。单元格稳健回归的设定则有所不同。由于响应变量 y_i 是单变量的，只有一个单元，因此它仍然是要么离群，要么干净。主要区别在于预测变量的污染。我们不再将 $\boldsymbol{x}_i = (x_{i1}、.....、x_{ip})$ 整体视为离群或完全干净，而是允许其中一些单元格 x_{ij} 受到污染，其余单元格保持干净。我们希望在估计 α 和 $\boldsymbol{\beta}$时使用 $\boldsymbol{x}_i$ 中的干净单元格，同时限制 $\boldsymbol{x}_i$中受污染单元格的不利影响。

最早提出的单元格稳健回归方法之一是¨Ollerer 等人（2016年）的射击 S 估计量。它使用坐标下降算法（Bezdek 等人，1987年），也称为“射击算法”。如果我们只拟合第 j个斜率系数 β_j，同时将其他斜率固定为其先前的值，则可以

$$y_i^{(j)} \sim \hat{\alpha}_j^{\text{new}} + \hat{\beta}_j^{\text{new}} x_{ij} \qquad \text{where} \qquad y_i^{(j)} := y_i - \hat{\boldsymbol{\beta}}_{-j}^{\text{old}} \boldsymbol{x}_{i,-j} \ . \tag{7}$$

这可以看作新响应变量 $\boldsymbol{y}^{(j)}$ 对第 j 个预测变量进行的简单回归。射击 S 估计量对坐标下降算法做了两项改动。第一，在简单回归（7）中，用按案例稳健 S 估计量替代普通最小二乘回归。第二，在每次迭代中对响应变量 $\boldsymbol{y}^{(j)}$ 进行“稳健化”，以确保按单元格异常值不会在迭代步骤中传播到新的响应值。Bottmer 等人（2022年）最近提出了一个适用于高维情形的射击 S 回归版本，该方法采用硬阈值化，从而得到稀疏的回归系数估计。

Leung 等人（2016年）提出了另一种方案。他们的模型假设，在单元格受到污染之前，配对变量 $(\boldsymbol{x}_i,\ y_i)$ 服从联合多元正态分布。他们首先使用 Agostinelli 等人（2015年）提出的 2SGS 估计量，并在第一步采用修正滤波器，估计配对变量 $(\boldsymbol{x}_i,\ y_i)$ 联合分布的位置 $\hat{}\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{\Sigma}$。然后，根据 $\hat{}\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$，通过常用公式估计斜率系数和截距

$$\hat{\boldsymbol{\beta}} := \widehat{\boldsymbol{\Sigma}}_{xx}^{-1} \widehat{\boldsymbol{\Sigma}}_{xy} \quad \text{followed by} \quad \hat{\alpha} = \hat{\mu}_y - \hat{\boldsymbol{\mu}}_x^\top \hat{\boldsymbol{\beta}} \ . \tag{8}$$

该估计量的稳健性源于 2SGS 方法，该方法已知在许多情况下表现相当出色。另一方面，这种方法要求协方差矩阵可逆，因此不适用于更高维的情形；Bottmer 等人（2022年）提出的射击 S 版本则能够处理这类情形。

p	经典	LTS	S	射击 S
5	0.0008	0.0442	0.0440	0.5525
10	0.0011	0.0720	0.0947	1.1577
20	0.0018	0.1558	0.3650	3.2820
50	0.0034	0.8009	1.7363	12.417

表 3：线性回归：含有 $n = 1000$ 和 p个回归变量的数据集的计算时间（秒）。

表 3 展示了若干回归方法的计算时间，采用的设置与表 1 和表 2 相同。我们将经典最小二乘回归与案例稳健最小截尾平方回归（Rousseeuw，1984年）和 S 回归（Rousseeuw 和 Yohai，¨1984年）进行比较，并与 Ollerer 等人（2016年）提出的逐单元稳健射击 S 方法进行比较。后者速度最慢，但仍然相当可行，尽管其计算完全使用 R。未展示 Leung 等人（2016年）提出的方法，因为其计算时间与表 2 中的 2SGS 方法相同，且额外计算非常快。

Agostinelli 和 Yohai（2016年）针对更一般的线性混合模型稳健估计量问题提出了一种有趣的方法。他们提出了一种复合估计量，通过最小化变量对上的马氏距离尺度来进行估计。使用成对马氏距离，使这种方法比逐案例 S 估计量更能抵御单元格异常值。评估这种方法对抗对抗性单元格异常值的稳健性将会很有意义。

Filzmoser 等人（2020）提出的方法以逐案例稳健方法（如 MM 回归）为起点。随后，将SPADIMO （Debruyne 等人，2019年）应用于标记案例，以期识别导致其偏离度的单元格。接着，使用未标记单元格的最近邻对这些标记单元格进行插补，并利用插补后的数据更新回归残差。最后，根据标准化残差计算逐案例权重，并将其用于加权最小二乘回归，以获得更新后的回归系数。该三步过程反复迭代。由于该算法依赖逐案例稳健的初始估计量，因此一般而言无法实现按单元格稳健性。不过，替换初始估计量或许能够实现按单元格稳健性。

最后，我们提及一些近期相关的方案。Su 等人（2021）首先计算配对变量 $(\boldsymbol{x}_i、\ y_i)$ 联合分布的稳健协方差矩阵。为此，他们采用一种基于上述成对相关的方法，使用 Gnanadesikan 和 Kettenring（1972）提出的方法，或 Boudt 等人（2012）研究的高斯秩相关。随后，他们将该协方差矩阵的平方根代入自适应套索目标函数，以进行变量选择。Toka 等人（2021）提出一个三步流程：(i) 使用稳健 Z 分数识别边际逐单元离群值，并将其设为缺失值；(ii) 基于完整案例计算按案例稳健的 MCD 位置和协方差估计量；

然后按照期望最大化算法的 E 步对缺失值进行插补；(iii) 对插补后的数据进行稳健套索回归。Saraceno等人（2021）使用多个单变量MM回归进行稳健的似不相关回归，并将 Agostinelli 等人（2015）的 2SGS方法应用于残差矩阵。针对含组成型协变量的回归，ˇ Stefelov´a 等人（2021）提出一个四步流程：(i) 通过DDC检测离群单元；(ii) 对不在离群行中的离群单元进行插补；(iii) 对插补后的数据进行逐行稳健回归；(iv) 基于多重插补进行推断。

我们以简短讨论作为本节的结尾。尽管针对单元格稳健回归已有多种方法，但仍有很大的进一步研究空间。首先，迄今为止，人们对异常值检测的关注并不多。对于样本级稳健性，根据 $\boldsymbol{x}_i$ 或 y_i 的残差是否离群，或两者是否都离群，可以对异常值类型进行分类，参见 Rousseeuw和van Zomeren（1990）。检测这类案例级离群值，关键在于获得可靠的稳健拟合（$\hat{}\alpha$，$\hat{}\boldsymbol{\beta}$）。但在许多现有方案中，单元格异常值是在某种预处理步骤中检测的，尚未利用稳健拟合来检测它们。其次，除少数例外，大多数方案尚未具备可证明的稳健性保证，例如在逐单元污染下具有较高的崩溃值。这可能是因为缺少适用于逐案例稳健回归的自然框架，也可能是因为处理单元格异常值总体上十分复杂。

2.5 主成分分析

主成分分析（PCA）是多元统计中的一项重要工具。经典主成分分析有多种表述方式，例如基于协方差矩阵的谱分解，或基于中心化数据矩阵的奇异值分解。一种理解方式是，PCA将含有 d 列的数据集 $\boldsymbol{X}$ 近似为一个新数据集 $\boldsymbol{X}$，该数据集位于一个 q维仿射子空间$\widehat{}$中，其维度为 $q < d$，使弗罗贝尼乌斯范数 $||\boldsymbol{X} - \boldsymbol{X}||_F$最小。逐案例稳健主成分分析有多种方法，包括球面主成分分析（Locantore等人，1999）、基于最小截尾平方正交回归的主成分分析（Maronna，2005）、PCAgrid投影追踪方法（Croux等人，2007），以及混合ROBPCA方法（Hubert等人，2005）。此外，还有逐案例稳健稀疏主成分分析方法，参见Croux等人（2013）、Hubert等人（2016）和Wang和Van Aelst（2020）。

对于单元格异常值，情况会更加棘手。其中一个难点在于，经典主成分分析会将数据正交投影到一个较低维度的子空间上。将含有单元格异常值的样本（行）进行投影会带来问题，因为离群单元可能会影响所有单元格。因此，每当算法进行投影时，都应处理这些离群单元。

Croux等人（2003年）较早提出了针对相关问题——单元格稳健因子分析的方案。首个单元格稳健主成分分析方法由Maronna和$\widehat{}$Yohai（2008）提出。他们不再最小化 $||\boldsymbol{X} - \boldsymbol{X}||_F$的弗罗贝尼乌斯范数，而是对 $\boldsymbol{X} - \boldsymbol{X}$ 的每个单元应用有界函数 ρ，并将所得误差求和，以此最小化稳健的$\widehat{}$近似误差

。该最小化过程通过迭代求解加权最小二乘问题来实现。有界 ρ 函数可确保单个单元不会主导近似误差。She等人（2016年）提出了另一种单元格稳健主成分分析方案。Hubert等人（2019年）提出的MacroPCA方法旨在提供一种一体化方法，既能处理缺失值，又能抵御单元格和按行离群值。该方法以**DDC** 为基础，并融入了ROBPCA的部分元素。它会对导致数据点远离拟合子空间的单元格进行插补，并生成可供绘图的单元格残差。

d	经典 ROBPCA		PCAgrid	MacroPCA
10	0.0043	0.1091	0.2216	0.1551
50	0.0161	0.2414	1.8707	0.4163
100	0.0413	0.3641	4.0087	0.9460
500	1.0227	2.9865	22.727	12.142
1000	3.3084	8.6023	56.335	13.778

表 4：主成分分析：数据集维度为 d、含有 $n = 1000$ 的计算时间（秒）。

表 4 展示了其中若干种主成分分析方法的计算时间。实验设置与表 1 至表 3 相同，但将维度提高到了 d。任务是计算前 10 个分量。案例稳健 ROBPCA 和 PCAgrid 方法比经典主成分分析耗时更长，但仍然足够快。MacroPCA 方法比 ROBPCA 慢，但比 PCAgrid 快，这表明单元格稳健主成分分析在计算上是可行的。

2.6 聚类

目前已有多种用于对含有单元格异常值的数据进行聚类的方法。最早的方法是Farcomeni（2014）提出的，他使用混合模型，并加入一个由零和一组成的 $n \times d$ 参数矩阵，用于指示哪些单元格应被标记并视为缺失。其目标是最大化未标记单元格的观测似然。该方法已在 R软件包 `snipEM`（Farcomeni和Leung，2019）中实现，但该软件包目前已不再维护。Garc´ıa-Escudero 等人（2021）考虑了一种簇呈线性结构的模型，即各簇位于低维子空间附近。随后，按照上文讨论的 Maronna和Yohai（2008）的方法，每个簇都通过按单元格稳健主成分分析进行拟合。他们像最小截尾平方回归一样，最小化逐坐标残差的稳健尺度。该算法迭代执行以下三个步骤：(i) 通过稳健主成分分析更新子空间参数；(ii) 更新用于指示哪些单元格被视为离群的权重；(iii) 更新组成员关系，其方式类似于通常的k均值聚类算法。

2.7 时间序列

据我们所知，目前尚无将逐单元稳健方法应用于时间序列的相关研究，但我们将介绍一种此类方法证明有用的情形，即自回归模型的拟合。假设我们有一个单变量时间序列 y_t，其形式为 $t = 1$、.....、 n。阶数为 p 的自回归模型 AR(p) 随后假设

$$y_t = \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + \varepsilon_t$$

对于 $t = p+1, \ldots, n$，其中 ε_t 独立同分布，服从均值为零、标准差为 σ 的高斯分布。为简单起见，我们假设模型不含截距项。注意，有 $n-p$ 组完整数据（$y_t, y_{t-1}, \ldots, y_{t-p}$），可以将它们作为连续的行组合成一个 $(n-p) \times (p+1)$ 矩阵 $\boldsymbol{Z}$。因此，AR(p) 模型可以看作一个线性模型，其中 $\boldsymbol{Z}$ 的第一列是响应变量，其余列是回归变量；已有多种经典方法可用于估计 $\boldsymbol{Z}$中的参数向量 $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$。其中最简单的方法就是最小二乘回归估计量，在此语境下有时称为“朴素普通最小二乘”。

现在考虑一个离群值 y_t。在 AR(p) 模型中，它会出现 $p+1$ 次：一次作为响应变量，并在每个 p 滞后项中各出现一次。换言之，一个离群的 y_t 会影响 $p+1$ 的行，每次影响其中一个单元。由于任何仿射等变回归方法的崩溃值至多为 0.5，在这种情况下使用此类方法，其崩溃值至多为 $1/(2p+2)$，可能低得令人难以接受。（如果先对原始序列进行差分，例如为消除季节效应，则这个上界还会再减半。）因此，要对 $\boldsymbol{\beta}$ 进行稳健估计困难重重。关于这一问题的讨论，以及对 AR(p) 模型中各类离群值相关文献的综述，见 Rousseeuw 和 Leroy（1987年）第 7.2 节。

这是逐单元稳健方法大显身手的情形。作为示例，我们生成了一个参数为 $\boldsymbol{\beta} = (0.5, 0.2, 0.2)$、长度为 $n = 1000$ 的 AR(3)时间序列 $\sigma = 1$ 。随后，我们通过将每 7 个观测中的一个 y_t 替换为离群值 10，制造了一种“星期效应”（生成数据并进行后续分析的 R脚本可从网页获取 `https://wis.kuleuven.be/statdatascience/robust/papers-2020-2029`）。因此，原始时间序列中有 14% 的离群值，但它们污染了 $\boldsymbol{Z}$ 中 997 行里的 569 行，也就是 57% 的行。经典估计得到的 $\boldsymbol{\beta}$ 为 (0.11, 0.11, 0.08)，与 (0.5, 0.2, 0.2) 相差甚远，而误差标准差的估计 $\sigma = 1$ 为 3.79。逐案例稳健回归在这里也不起作用：最小截尾平方得到 $\hat{\boldsymbol{\beta}} = (0.73, 0.02, 0.01)$ 和 $\hat{\sigma} = 1.85$，R 中默认的 MM回归则得到 $\boldsymbol{\beta} = (0.14, 0.14, 0.12)$ 和 $\hat{\sigma} = 1.72$。幸运的是，逐单元稳健方法确实能给出良好的结果。Leung 等人（2016）的回归方法首先通过 2SGS 计算 $\boldsymbol{Z}$ 的稳健协方差矩阵 $\boldsymbol{\Sigma}$，然后应用（8），得到 $\hat{\boldsymbol{\beta}} = (0.52, 0.21, 0.17)$。此外，由 $\hat{\sigma}^2 = \widehat{\boldsymbol{\Sigma}}_{yy} - \hat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\Sigma}}_{xx} \hat{\boldsymbol{\beta}}$ 可得 $\hat{\sigma} = 0.96$。将相同公式应用于 cellMCD协方差矩阵，结果也相近：$\hat{\boldsymbol{\beta}} = (0.53, 0.18, 0.19)$，且 $\hat{\sigma} = 0.92$。

总而言之，t 自回归模型本身就可能产生许多离群单元格，而 h

残障观测　　逐单元稳健方法则可以处理这些离群单元格。

2.8 其他各种设置

Aerts 和 Wilms（2017）构建了线性判别分析和二次判别分析的单元格稳健版本。为此，他们按照¨Croux 和 Ollerer（2016）的方法，为每个类别估计一个逐单元稳健精度矩阵，并将其代入分类规则。后续研究可以处理这样一种情况：待分类的新数据点也可能包含离群单元。

在函数型数据分析中，观测不是矩阵的一行，而是一个函数，例如一条曲线。在这种情形下，单元格异常值对应的是曲线中的局部偏离（如尖峰），而非整条曲线的整体偏离度。在 Hubert 等人（2015）对函数型离群值的分类体系中，这些被称为孤立离群值，并且已有方法用于检测它们。Garc´ıa-Escudero 等人（2021）将逐单元稳健方法应用于函数型数据。

3 崩溃值

崩溃值在案例级稳健方法的研究中发挥着重要作用，因为崩溃值为正是稳健性的必要条件。在若干模型中，自然的等变性意味着崩溃值的上限为 50%，而已有方法能够使崩溃值接近这一上限。

在按单元格稳健性的框架下，Alqallaf 等人（2009）定义了位置估计量的渐近按单元崩溃值。本文将聚焦于 Donoho 和 Huber（1983）以及 Lopuha¨a 和 Rousseeuw（1991）所定义的有限样本崩溃值。下文将研究两个具有实际重要性的统计模型，并说明在一些乍看之下很自然的条件下，不可能获得较高的按单元崩溃值。

3.1 位置和协方差估计量的崩溃

考虑一个由 n 个位于 d 维空间中的数据点构成的有限样本，其数据矩阵记为 $\boldsymbol{X}$。我们希望通过估计量 $\hat{}\boldsymbol{\mu}$ 估计其未知的位置向量。接着，我们将 $\hat{}\boldsymbol{\mu}$ 在 $\boldsymbol{X}$ 处的有限样本逐单元崩溃值定义为：每列中需要替换的单元格的最小比例，替换后可使估计值超出所有界限。形式化地，设 $\boldsymbol{X}^m$ 为通过将 $\boldsymbol{X}$ 的每列中至多 m 个单元格替换为任意值而得到的任一污染样本。那么，位置估计量 $\hat{}\boldsymbol{\mu}$ 在 $\boldsymbol{X}$ 处的 有限样本逐单元崩溃值 定义为

$$\varepsilon_n^*(\hat{\boldsymbol{\mu}}, \boldsymbol{X}) = \min\left\{\frac{m}{n} : \sup_{\boldsymbol{X}^m} ||\hat{\boldsymbol{\mu}}(\boldsymbol{X}^m) - \hat{\boldsymbol{\mu}}(\boldsymbol{X})|| = \infty\right\}. \tag{9}$$

这是 Alqallaf 等人（2009）定义的渐近按单元崩溃值的一个自然有限样本版本。

我们先从一个显而易见的说明开始。如果某个估计量的按案例崩溃值存在上界，那么这个上界也适用于其按单元崩溃值。事实上，通过用其他观测替换 m 个样本所得到的污染数据集，也可以通过用替换观测中的单元格替换被替换观测的单元格来构造；这会影响每列中的 m 个单元格。因此，根据 Lopuha¨a 和 Rousseeuw（1991）的定理 2.1，对于任何平移等变位置估计量，(9) 至多为 $\lfloor (n+1)/2 \rfloor / n$ 。

我们将看到，一些看似自然的条件会导致相当低的按单元崩溃值。其中一个条件是

$$\begin{gathered}\text{if all points of } \boldsymbol{X} \text{ lie in a lower-dimensional} \\ \text{affine subspace, then } \hat{\boldsymbol{\mu}}(\boldsymbol{X}) \text{ lies in that subspace as well.}\end{gathered} \tag{10}$$

这是一个非常弱的精确拟合性质，但如下结果表明，它仍然会限制崩溃值。

命题 1. 任何满足 (10) 的 位置估计量$\hat{}\boldsymbol{\mu}$ ，其按单元崩溃值都有如下界：

$$\varepsilon_n^*(\hat{\boldsymbol{\mu}}, \boldsymbol{X}) = \left\lceil \frac{n}{d} \right\rceil / n \approx \frac{1}{d} \tag{11}$$

对于任意数据集 $\boldsymbol{X}$，

证明利用了这样一个事实：只需在每个变量中替换不超过 $\lceil n/d \rceil$ 个单元格，就能将数据移至某个超平面。本节所有证明均见附录。

Alqallaf 等人（2009）证明，位置估计量的渐近按单元崩溃值存在一个上界，该上界也会像 $1/d$ 一样趋于零。为此，他们要求满足两个条件。其中之一是估计量满足所谓的 δ 一致性。另一个条件是$\hat{}\boldsymbol{\mu}$具有仿射等变性，即对于所有 $d \times 1$向量 $\boldsymbol{v}$ 以及所有 $d \times d$非奇异矩阵 $\boldsymbol{A}$，都有

$$\hat{\boldsymbol{\mu}}(\boldsymbol{v}\mathbf{1}_d^\top + \boldsymbol{X}\boldsymbol{A}) = \boldsymbol{v} + \boldsymbol{A}^\top \hat{\boldsymbol{\mu}}(\boldsymbol{X}) \tag{12}$$

其中 $\mathbf{1}_d$ 是全 1 向量。换言之，该估计量对平移和线性变换具有等变性。(12) 中转置的 $\boldsymbol{A}^\top$ 是因为 $\boldsymbol{X}$ 将数据点按行而非按列排列。

我们的结果更为一般，因为我们不需要 δ-一致性，而且可以使用正交等变性，而非仿射等变性。正交等变性是一种较弱的等变形式，只要求正交矩阵 $\boldsymbol{A}$ 满足 (12)。

推论1。 任何正交等变的位置估计量都满足性质 (10)，因此命题*1*的上界适用。

这些都表明，对于希望具备按单元格稳健性的估计量而言，正交等变性，甚至较弱的条件（10），都是不理想的性质。但对于位置而言，通过结合逐坐标稳健位置估计，可以轻松避免这一问题。

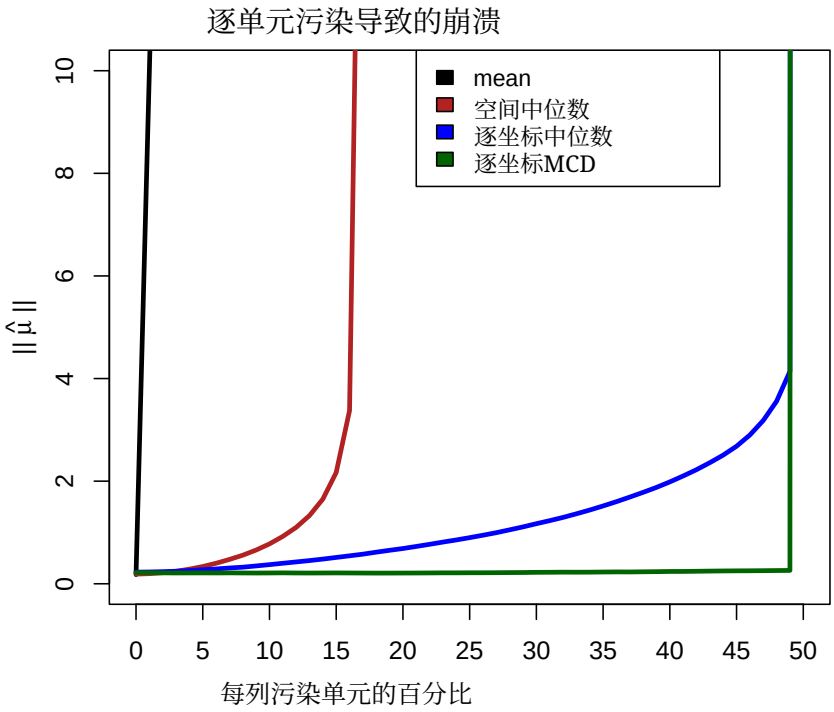


图3：逐单元污染导致崩溃的经验示例，适用于 $n = 100$个样本和 $d = 4$ 个维度。

我们提供一个简单的经验示例。实验过程如下。首先，从 $d = 4$ 维标准高斯分布中生成了 $n = 100$个点。这些点每列有 $k = 0$ 个离群单元。接下来，每列离群单元的数量以1为步长递增，使具有离群单元的样本尽可能少地重叠，因此当 $k = 25$ 时，所有样本都有一个离群单元。所有离群单元的值均设为500。在每个污染数据集上运行以下位置估计量：算术平均数、使 $\sum_{i=1}^{n} ||\boldsymbol{x}_i - \hat{}\boldsymbol{\mu}||$最小的空间中位数、逐坐标中位数（3）以及逐坐标单变量MCD估计量。图3绘制了每个估计量的欧几里得范数 $||\hat{}\boldsymbol{\mu}||$与每列污染单元百分比之间的关系。图中显示了200次重复实验的平均范数。横轴终止于50%，这对应于平移等变估计量崩溃值的上界。

在图3中，我们可以看到均值最先远离原点。事实上，如果让 n增大，并使每个离群单元的值从 500 增至无穷大，均值所在的直线就会变成垂直线，因为每列只需有 1 个单元格异常值就足以导致崩溃。相比之下，空间中位数的范数在很长一段时间内都保持较低水平，之后才开始增大。（事实上，当每列离群单元所占比例恰为 $1/d = 25\%$ 时，空间中位数的范数达到 250，因为此时所有数据点都位于超平面 $x_1 + x_2 + x_3 + x_4 = 500$ 附近。）请注意，空间中位数并非仿射等变，但它是正交等变的，因此推论1适用。

坐标　　逐坐标中位数超出了上界 $1/d$ ，因为它不满足　　e

条件 (10)。事实上，Rousseeuw 和 Leroy（1987年）已经给出了一个简单的反例（见第 250 页）：当 $d > 2$ 时，逐坐标中位数位于超平面之外。（对于 $d \leqslant 2$ ，该上界显然成立。）每个变量的中位数只有在该变量中离群值达到或超过 50% 时才会失效，因此将这些单独的中位数组合起来用于 (3) 时，情况也相同。对每个坐标的单变量MCD估计进行组合时，推理也相同。

^按单元格稳健协方差估计比位置估计更难。协方差（散布）矩阵的估计量可能以两种方式崩溃。与逐案例情形类似，我们可以将协方差估计量的逐单元爆炸崩溃值 定义为 $\boldsymbol{\Sigma}$

$$\varepsilon_n^+(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{X}) = \min\left\{\frac{m}{n} : \sup_{\boldsymbol{X}^m} \lambda_1(\widehat{\boldsymbol{\Sigma}}) = \infty\right\} \tag{13}$$

其中， λ_1 表示最大特征值。此外，我们将 $\boldsymbol{\Sigma}$ 的按单元格内爆崩溃^值 定义为

$$\varepsilon_n^-(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{X}) = \min\left\{\frac{m}{n} : \inf_{\boldsymbol{X}^m} \lambda_d(\widehat{\boldsymbol{\Sigma}}) = 0\right\} \tag{14}$$

其中， λ_d 是最小特征值。内爆与爆炸同样糟糕，因为奇异协方差矩阵不可逆，因此无法计算马氏距离、进行判别分析等。

在这种情形下，获得良好的按单元格崩溃值也比获得其逐案例对应值更难。例如，在典型数据集中，经典协方差矩阵的逐案例内爆崩溃值非常高，实际上为 $(n-d)/n \approx 1$。这是因为只要保留原始数据点中的$d+1$ ，经典协方差矩阵就仍然非奇异。与此形成鲜明对比的是，其按单元格 内爆崩溃值相当低。即使是满足以下直观条件（与 (10) 类似）的所有协方差估计量，也同样如此：

$$\begin{gathered}\text{if all points of } \boldsymbol{X} \text{ lie in a lower-dimensional} \\ \text{affine subspace, then } \widehat{\boldsymbol{\Sigma}}(\boldsymbol{X}) \text{ is singular.}\end{gathered} \tag{15}$$

^**命题2**。任何满足(15)的协方差估计量 $\boldsymbol{\Sigma}$的逐单元内陷崩溃值均有如下界：

$$\varepsilon_n^-(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{X}) = \left\lceil \frac{n-1}{d} \right\rceil / n \approx \frac{1}{d} \tag{16}$$

对任意数据集 $\boldsymbol{X}$均成立。

这一结果似乎是新的，而且也可由等变性推出。回顾一下，如果对于所有 $d \times 1$ 向量 $\boldsymbol{v}$ 和所有 $d \times d$ 非奇异矩阵 $\boldsymbol{A}$ 均有如下关系，则称^协方差估计量 $\boldsymbol{\Sigma}$ 具有仿射等变性：

$$\widehat{\boldsymbol{\Sigma}}(\boldsymbol{v}\mathbf{1}_d^\top + \boldsymbol{X}\boldsymbol{A}) = \boldsymbol{A}^\top \widehat{\boldsymbol{\Sigma}}(\boldsymbol{X})\boldsymbol{A} \,. \tag{17}$$

推论 2。 任何仿射等变协方差估计量 $\boldsymbol{\Sigma}$ 都满足性质 （15），因此命题*2*的上界适用。

推论 2 中的上界 $\lceil (n-1)/d \rceil / n \approx 1/d$远低于从仿射等变协方差估计量的按案例崩溃值继承而来的上界，后者为 $\lfloor (n-d+1)/2 \rfloor / n \approx 0.5$。此前文献并未提及，在逐单元设置中，内爆崩溃可能轻易发生。对于 Ollerer 和 Croux（2015年）的逐单元稳健协方差估计量¨，文献只给出了其爆炸崩溃值。上述^结果表明，可以通过施加类似 $\lambda_d(\boldsymbol{\Sigma}) \geqslant a$ 的先验约束来排除内爆，也可以采用一种^形式，其中 $\boldsymbol{\Sigma}$是两个矩阵的凸组合，而其中一个矩阵是系数较小的单位矩阵。这两种方法都不符合仿射等变性，但这里讨论的是逐单元设置。在不具备仿射等变性的逐案例情形中，也采用过这种机制，例如参见 Ollila 和 Tyler（2014）。

请注意，尽管上述结果并不乐观，仍然可以构造出具有更高按单元崩溃值的协方差估计量，例如 Raymaekers 和 Rousseeuw（2022b）提出的方法。他们提出的逐单元稳健 MCD 估计量的崩溃值为 $(n-h+1)/n$ ，其中 h 最低可取 $\lfloor n/2 \rfloor + 1$ ，但建议设为 $h = 0.75n$。因此，其渐近崩溃值可达到 0.5，这是可能达到的最高值。

3.2 回归中的崩溃

对于线性回归，我们使用模型 (6)，但将其稍作不同的表述 a s

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\gamma} + \text{noise} \tag{18}$$

其中 $\boldsymbol{y} = (y_1, \ldots, y_n)$ 是响应值的 $n \times 1$ 列向量，而 $\boldsymbol{X}$ 的各行形式为 $(1, \boldsymbol{x}_i^\top)$。$(p+1)$ 向量 $\times 1$ $\boldsymbol{\gamma} = (\alpha, \beta_1, \ldots, \beta_p)^\top$ 将截距和斜率组合在一起。我们还将实际数据（即除作为常数的第一列之外的所有数据）组合成 $n \times (p+1)$ 矩阵 $\boldsymbol{Z}$，其第 i 行为 $(x_{i1}, \ldots, x_{ip}, y_i)$。

现在我们对数据集 $\boldsymbol{Z}$ 进行污染。我们将 $\boldsymbol{Z}^m$ 定义为：通过在 $\boldsymbol{Z}$ 的每一列中至多替换 m 个单元格为任意值而得到的任意受污染协变量集合。我们将回归估计量 $\widehat{\boldsymbol{\gamma}}$ 在数据集 $\boldsymbol{Z}$ 上的有限样本逐单元崩溃值定义为

$$\varepsilon_n^*(\widehat{\boldsymbol{\gamma}}, \boldsymbol{Z}) = \min \left\{ \frac{m}{n} : \sup_{\boldsymbol{X}^m} ||\widehat{\boldsymbol{\gamma}}(\boldsymbol{Z}^m) - \widehat{\boldsymbol{\gamma}}(\boldsymbol{Z})|| = \infty \right\}. \tag{19}$$

许多现有的回归方法 $\widehat{\boldsymbol{\gamma}}$ 满足以下直观性质：

$$\begin{gathered} \text{if } y_i = \boldsymbol{x}_i^\top \boldsymbol{\gamma}_0 \text{ for all } i = 1, \ldots, n \text{ and } \boldsymbol{X} \text{ is not} \\ \text{in a lower-dimensional subspace, then } \widehat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}_0 \, . \end{gathered} \tag{20}$$

要求 $\boldsymbol{X}$ 不属于低维向量子空间，是为了避免多重共线性，因为多重共线性会给许多回归方法带来问题。

命题3. 任何满足条件的回归估计量 $\widehat{\gamma}$，其按单元崩溃值都有界 (20)

$$\varepsilon_n^*(\widehat{\gamma}, \boldsymbol{X}, \boldsymbol{y}) \leqslant \left\lceil \frac{n-1}{p+1} \right\rceil / n \approx \frac{1}{p+1} \tag{21}$$

对于任意数据集 $\boldsymbol{Z}$。

此外，在自然的等变性条件下，弱条件 (20) 也成立。其中第一种是回归等变性，它要求对于所有 $(p+1) \times 1$ 向量 $\boldsymbol{v}$，均有

$$\widehat{\boldsymbol{\gamma}}(\boldsymbol{X}, \boldsymbol{y} + \boldsymbol{X}\boldsymbol{v}) = \widehat{\boldsymbol{\gamma}}(\boldsymbol{X}, \boldsymbol{y}) + \boldsymbol{v} \tag{22}$$

对于所有数据集 $(\boldsymbol{X},\ \boldsymbol{y})$。这一符合直觉的条件类似于位置估计量的平移等变性。第二种性质是尺度等变性，这意味着对于任意常数c，都有

$$\widehat{\boldsymbol{\gamma}}(\boldsymbol{X}, c\,\boldsymbol{y}) = c\widehat{\boldsymbol{\gamma}}(\boldsymbol{X}, \boldsymbol{y})\,. \tag{23}$$

我们甚至不需要仿射等变性；仿射等变性表示，对于所有非奇异矩阵 $\boldsymbol{A}$，$\widehat{\boldsymbol{\gamma}}(\boldsymbol{X}\boldsymbol{A},\ \boldsymbol{y}) = \boldsymbol{A}^{-1}\widehat{\boldsymbol{\gamma}}(\boldsymbol{X},\ \boldsymbol{y})$。

推论 3. 任何具有回归等变性(22)和尺度等变性的估计量 $\widehat{\gamma}$ 都满足性质(20)，因此命题 (23) *3*的上界适用。

我们也可以轻易写出关于噪声项尺度估计量内爆的类似结果。

总之，若要获得按单元崩溃值更高的回归估计量，就必须放弃等变性条件（22）和（23）的自然组合，甚至还要放弃直观条件（20）。处理高维数据时，我们已经习惯使用会使回归过程舍弃部分协变量的惩罚项；但即使面对较低维度的数据，也可能需要采用这类技术或全新的方法。

4 按单元格稳健对应分析

本节介绍首个用于对应分析的按单元格稳健方法。

4.1 经典对应分析

对应分析（CA）由 Hirschfeld（1935年）提出，随后由 Benz´ecri 等人（1973年）进一步发展，作为一种分析分类数据的方法，这些数据可以用列联表表示。其主要目的是构建所谓的双标图，即一种概括列联表部分结构的二元图。该双标图可以理解为类似于绘制连续数值数据的前两个主成分。

设 $\boldsymbol{X}$为一个由计数 x_{ij}构成的 $n \times d$列联表，其中有 $i = 1$，.....，n以及 $j = 1$，.....，d。记 $N = \sum_{i=1}^{n} \sum_{j=1}^{d} \boldsymbol{X}_{ij}$为 $\boldsymbol{X}$中所有计数之和，并令 $\boldsymbol{P} = \boldsymbol{X}/N$为相对频率矩阵。记 $\boldsymbol{r}$为包含$\boldsymbol{P}$各行之和的 $n \times 1$列向量，记 $\boldsymbol{c}$为包含其各列之和的 $d \times 1$向量。另记 $\boldsymbol{D_r} = \text{diag}(\boldsymbol{r})$和$\boldsymbol{D_c}$ =diag($\boldsymbol{c}$)。最后，令 $\boldsymbol{R} = \boldsymbol{D_r}^{-1}\boldsymbol{P}$为标准化行剖面矩阵，其中每一行的和均为 1。对应分析基于加权中心化行剖面矩阵

$$\boldsymbol{S} := \boldsymbol{D_r}^{1/2} \left(\boldsymbol{R} - \mathbf{1}_d \boldsymbol{c}^\top \right) \boldsymbol{D_c}^{-1/2} \tag{24}$$

其中，$\mathbf{1}_d$是由 1 组成的 $d \times 1$ 列向量。注意，$\mathbf{1}_d \boldsymbol{c}^\top$ 的每一行之和也为 1，因此 $\boldsymbol{R} - \mathbf{1}_d \boldsymbol{c}^\top$ 的行总和均为零。这意味着 $\boldsymbol{R} - \mathbf{1}_d \boldsymbol{c}^\top$ 的列线性相关，因此 $\boldsymbol{S}$ 也不是满秩的。

从这里开始，我们将 $\boldsymbol{S}$ 视为一个包含连续变量（列）的数据集。由于 $\boldsymbol{S}$ 在构造上是奇异的，因此对其进行奇异值分解很有用，结果如下：

$$\boldsymbol{S} = \boldsymbol{U} \boldsymbol{\Gamma} \boldsymbol{V}^\top . \tag{25}$$

这里，$\boldsymbol{U}$ 的正交列包含左奇异向量，$\boldsymbol{\Gamma}$ 是一个对角线上为奇异值的对角矩阵，而 $\boldsymbol{V}$ 的正交列包含右奇异向量。若将 $\boldsymbol{S}$的秩记为 k，则矩阵 $\boldsymbol{U}$为 $n \times k$，$\boldsymbol{\Gamma}$ 为 $k \times k$，载荷矩阵 $\boldsymbol{V}$ 为 $d \times k$。利用这个奇异值分解，我们现在可以表示数据集 $\boldsymbol{S}$的行和列。行的主坐标为 $\boldsymbol{D_r}^{-1/2} \boldsymbol{U} \boldsymbol{\Gamma}$，列的主坐标为 $\boldsymbol{D_c}^{-1/2} \boldsymbol{V} \boldsymbol{\Gamma}$。在双标图中，我们只显示两者的前两个坐标。

4.2 单元格稳健对应分析

为了构建一种单元级稳健对应分析方法，我们需要对 $\boldsymbol{S}$ 的奇异值分解（25）进行稳健化。为此，我们对 Hubert等人（2019年）的逐单元稳健 MacroPCA 方法进行修改。其原始版本拟合主成分分析模型

$$\boldsymbol{S} = \mathbf{1}_n \boldsymbol{\mu}^\top + \boldsymbol{T} \boldsymbol{V}^\top + \text{noise} \tag{26}$$

其中，$\boldsymbol{\mu}$是中心，得分矩阵 $\boldsymbol{T}$为 $n \times k$，载荷矩阵 $\boldsymbol{V}$为 $d \times k$。噪声与 $\boldsymbol{S}$的维度相同。在此情境下，k（即主成分的数量）的选择很重要。这是因为，当某个点到拟合子空间的正交距离较大时，该点会被视为离群点；因此，若选择最大的 k，所有点都会处于子空间中，也就没有点会被视为离群点。这也意味着载荷并非嵌套的，也就是说，$k = 2$ 的载荷列不是 $k = 3$载荷列的子集。MacroPCA 选择能够解释至少 80% 变异性的最小分量数作为 k。

由于我们的目标是拟合模型 (25)，我们在 MacroPCA 中添加了一个选项，可预先固定中心 $\boldsymbol{\mu}$，而不是根据数据对其进行估计。就当前目的而言，我们将 $\boldsymbol{\mu}$ 固定为零，因此 (26) 的第一项消失。我们将估计载荷 $\widehat{\quad}$ $\widehat{\quad}$ $\widehat{\quad}$和得分分别记为 $\boldsymbol{V}$和 $\boldsymbol{T}$。对于对角矩阵 $\boldsymbol{\Gamma}$ 中的奇异值，我们取 MacroPCA 估计的特征值的 n 倍平方根。最后，我们计算 $\widehat{\boldsymbol{U}} = \widehat{\boldsymbol{T}}\widehat{\boldsymbol{\Gamma}}^{-1}$。$\widehat{\quad}$ $\widehat{\quad}$ $\widehat{\quad}$将所得矩阵 $\boldsymbol{U}$、$\boldsymbol{\Gamma}$ 和 $\boldsymbol{V}$ 组合起来，可得到分解 (25) 的稳健版本。使用这些矩阵，可以像之前一样绘制双标图。

为了说明这种按单元格稳健对应分析方法，我们考虑两个例子。这些例子此前由 Riani 等人（2022）在案例稳健对应分析的背景下进行过研究。

4.3 服装数据的对应分析

第一个例子是一个列联表，其中包含从欧盟以外地区流入其 28 个成员国的服装贸易流（数据来自 2020 年之前，因此当时英国仍包括在内）。Riani 等人（2022）的列联表中，各列对应五个不同的价格区间，从最低到最高。因此，列联表 $\boldsymbol{X}$ 的大小为28×5，矩阵 $\boldsymbol{P}$和 $\boldsymbol{S}$也具有相同的大小。

作为对数据的初步探索，我们首先将 DDC 应用于 $\boldsymbol{S}$。（请注意，我们不应将 DDC 应用于原始列联表 $\boldsymbol{X}$，因为其各行的尺度不可比较，这是由于这些尺度与各国的人口规模相关。）$\widehat{\quad}$ $\widehat{\quad}$DDC 算法生成预测数据矩阵 $\boldsymbol{S}$，进而得到单元格残差 $\boldsymbol{S} - \boldsymbol{S}$。图 4 左侧面板展示了服装数据的 DDC 单元图，以图形方式呈现标准化单元格残差。红色单元格表示标准化残差高于$\sqrt{\chi^2_{1,\ 0.99}} \approx 2.57$，颜色越深，实际残差越大。这些单元格的值高于预测值。蓝色单元格表示标准化残差低于$-\sqrt{\chi^2_{1,\ 0.99}}$，即其值低于预测值。多数单元格以黄色显示，其值大致符合预测值。

单元图让我们注意到几种异常的贸易流。具体而言，与根据其他贸易流所预期的情况相比，英国（国家代码 GB）、斯洛伐克（SK）、希腊（GR）、罗马尼亚（RO）、马耳他（MT）和拉脱维亚（LV）进口了大量最低价格区间的衣物。这些结果是可以解释的。例如，众所周知，英国从孟加拉国、印度和中国进口大量非常便宜的衣物。斯洛伐克也从这些国家进口了很多衣物；从其蓝色单元格可以看出，它进口的较高三个价格区间的衣物相对较少。

接下来，我们将 MacroPCA 的零中心版本应用于 $\boldsymbol{S}$，得到稳健载荷矩阵 $\widehat{\boldsymbol{V}}$、得分 $\widehat{\boldsymbol{T}}$ 和估计特征值。根据这些结果，我们推导出 $\widehat{\boldsymbol{\Gamma}}$ 和 $\widehat{\boldsymbol{U}} = \widehat{\boldsymbol{T}}\widehat{\boldsymbol{\Gamma}}^{-1}$。$\widehat{\quad}$ $\widehat{\quad}$ $\widehat{\quad}$使用 $\boldsymbol{U}$、$\boldsymbol{\Gamma}$ 和 $\boldsymbol{V}$ 得到图 5 中的双标图。我们看到，SK 在左侧离群，并且与最低价格区间的类别 X_1 有很强的关联，而它位于

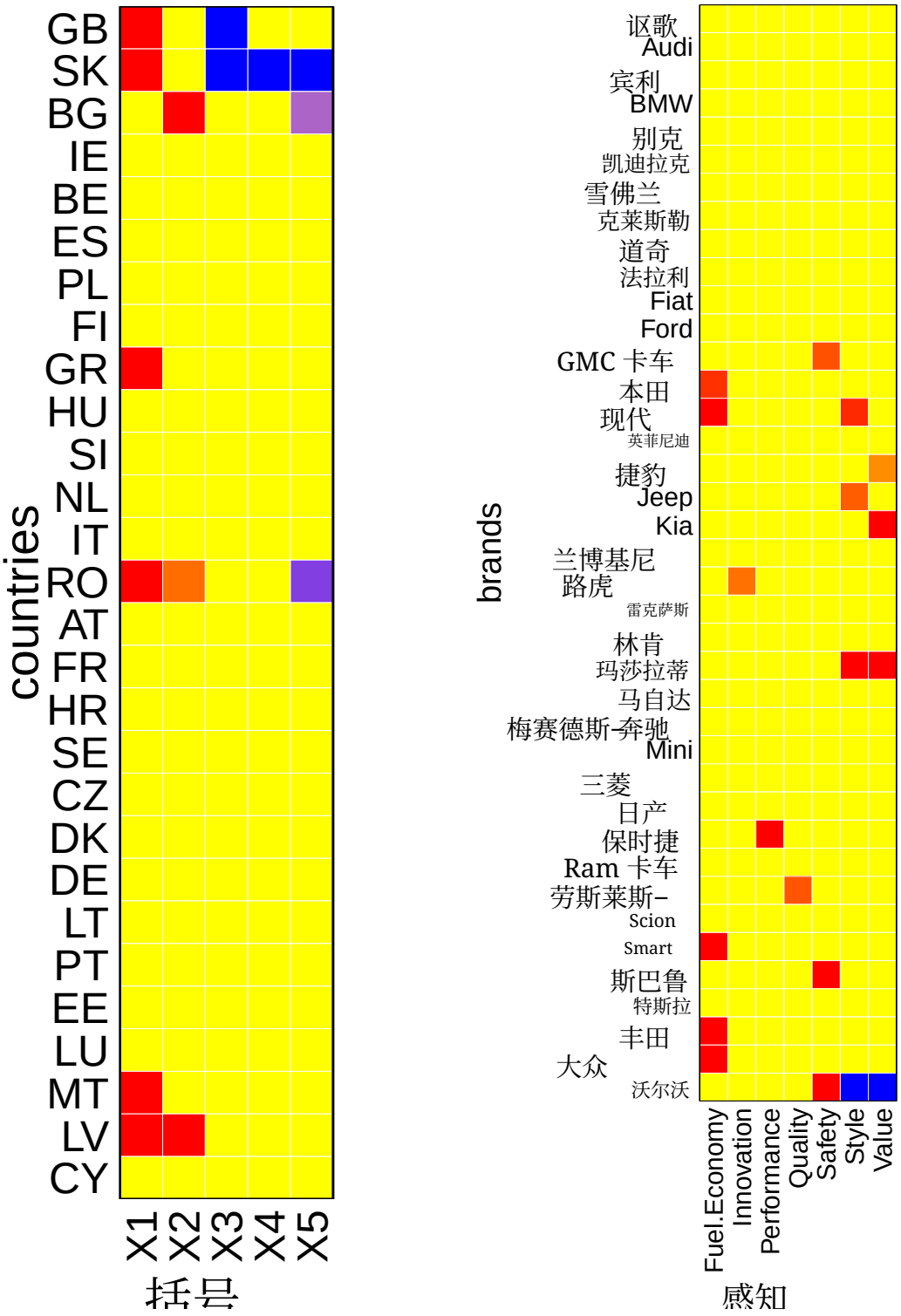
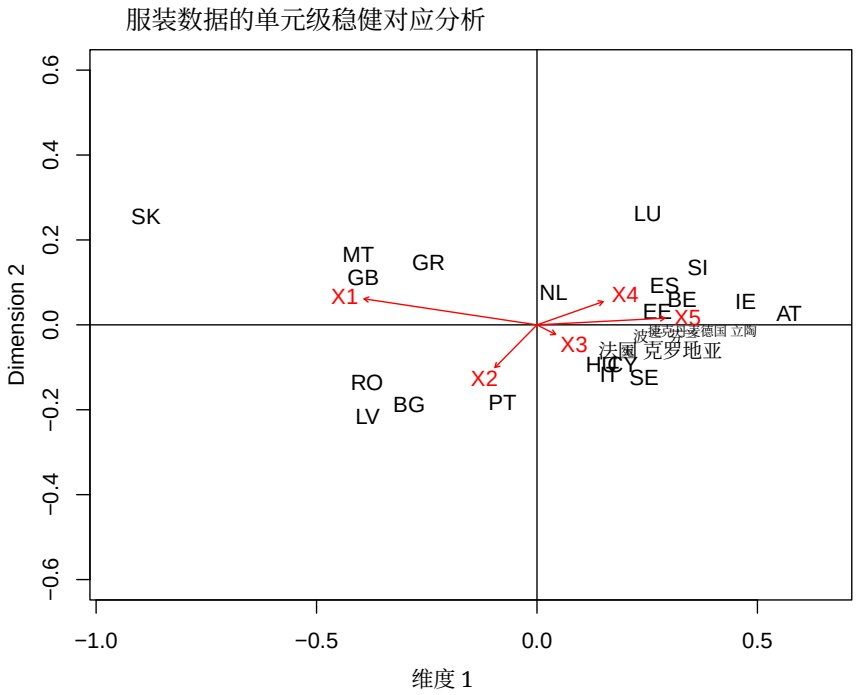


图 4：检测偏离单元格（DDC）在服装数据（左）和品牌数据（右）上的结果单元图。



Fi图 5：基于所提出的按单元格稳健方法绘制的服装数据双标图 t.

与三个最高价格区间对应的变量 X_3、X_4 和 X_5 方向相反。我们还注意到，一小组国家 MT、GB、GR 位于最低价格区间 X_1的方向，而国家 RO、LV、BG 则位于两个最低价格区间 X_1 和 X_2的方向。右侧的国家在较高价格区间上的支出更多。总体而言，这幅双标图与 Riani 等人（2022）图 4 中的逐案例稳健双标图颇为相似。

4.4 品牌数据的对应分析

第二个示例是《消费者报告》（美国）开展的 2014年汽车品牌认知调查的汇总列联表，该表可在 `https://boraberan.wordpress.com/2016/09/22/` 上公开获取。调查询问了 1578 名参与者，他们认为 39 个不同汽车品牌具备哪些属性。可能的属性包括燃油经济性、创新、性能、质量、安全性、风格和价值。受访者必须从列表中选出他们认为适用于某个品牌的所有特征。因此，所得列联表 $\boldsymbol{X}$ 的大小为 39×7，矩阵 $\boldsymbol{S}$ 的维度也相同。

我们再次首先对 $\boldsymbol{S}$ 应用 **DDC**，以标记离群单元格。结果单元图显示在图 4 的右侧面板中，并揭示了一些有趣的见解。我们看到，沃尔沃相当特殊，竟有多达三个单元格被标记。其安全性认知高于预测值，而风格和价值则低于根据其其他特征所预期的水平。现代和玛莎拉蒂各有两个单元格被标记。现代的表现相当

出色，燃油经济性和风格相较其其他特征而言相当突出；玛莎拉蒂的风格和价值似乎都异常高。（不禁让人怀疑，有时“价值”是否被理解为“昂贵”，而非“物有所值”。）此外，还有另外 12 个观测各有一个按单元格异常值。例如，GMC 卡车的安全性认知较高，保时捷的性能较高，丰田和大众的燃油经济性较高。在 39 个品牌中，共有 15 个至少有一个单元格被 DDC 标记。标记单元格总数为 19 个，仅占所有单元格的 7%，但涉及近 40% 的样本。尽管如此，采用高崩溃点方法进行案例稳健分析仍然可行，因为超过 50% 的样本没有离群单元格。

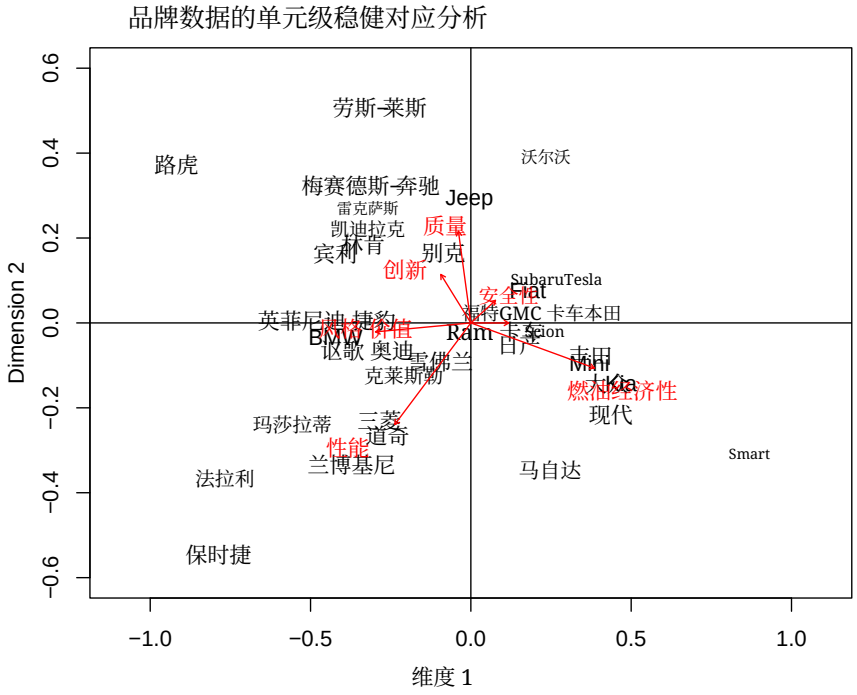


图6：基于所提出的逐单元稳健拟合绘制的品牌数据双标图。

图6展示了基于逐单元稳健 MacroPCA 方法得到的双标图。它与 Riani 等人（2022）得到的结果相当相似，而这两种稳健拟合都与其图6所示的经典对应分析结果大不相同。我们可以在图6中看到一些直观上容易解释的结构。沃尔沃通常被认为是最安全的汽车品牌之一，确实位于安全性属性所指的方向上。有一组跑车（法拉利、玛莎拉蒂、兰博基尼、保时捷）远远位于性能箭头所指的方向上。劳斯莱斯和梅赛德斯-奔驰位于创新和质量所指的方向上，而 Smart、现代和大众在燃油经济性方面的载荷较大。

5 讨论与挑战

本节提出一些值得探讨的问题，并讨论发展理论和方法以应对单元格异常值所面临的开放性挑战。

5.1 将离群值解释为按单元格或逐案例离群值

逐案例离群值和单元格异常值的概念引发了一些解释方面的问题。一方面，按案例污染和逐单元污染作为生成模型有着明确的含义，也就是说，它们是生成离群值的不同方式。许多研究者都使用这些模型来设计模拟研究。

另一方面，面对新收到的数据时，要判断一个异常行究竟是逐案例离群值（即由不同于大多数样本的机制生成），还是一个大体干净、但含有一些离群单元的行，往往是一个不可识别的问题。举例来说，考虑 Raymaekers 和 Rousseeuw（2021a年）中的一个玩具示例。假设干净数据是独立同分布的，来自四维标准高斯分布 $F = N(\mathbf{0},\ \boldsymbol{I})$，并考虑行 (10，0，0，0)。从逐案例的角度来看，这个观测与行 $(\sqrt{50},\ \sqrt{50},\ 0,\ 0)$ 以及行 (5，5，5，5)“等价”，因为二者都可以通过对数据进行正交变换得到，而在这种变换下，分布 $N(\mathbf{0},\ \boldsymbol{I})$ 保持不变。但从按单元格的角度来看，人们会认为第一行有一个单元格异常值，第二行有两个，最后一行有四个。因此，结论取决于我们采用逐案例范式还是单元格范式。

虽然可能无法基于数据判断离群值属于案例级还是单元格级，但从估计的角度来看，或许没有必要作出这种区分。在第 4.1 节中，Danilov（2010）指出：“如果协方差结构已知得足够准确，能够识别逐单元污染，那么用于检测逐单元污染的方法也可以用来消除案例级离群值的影响。如果我们从一个案例级离群值中移除若干单元，使剩余子向量不再离群，那么它对协方差估计的影响就会显著减小，从而可以容忍。”我们认为这种观点颇为合理。不过，为了有效消除对抗性案例级异常值的影响，在寻找危害最大的单元格时，需要相当准确的拟合（协方差估计）。成对相关方法很可能会受到对抗性案例级异常值的严重影响，因此可能需要与其他针对案例级离群值的方法结合使用，例如 2SGS。“解除武装”这一方法似乎确实可行。将单元格 MCD 估计量应用于 `robustbase`R软件包（Maechler 等，2022）中所有此前通过案例级稳健方法分析过的基准数据集时，所得拟合结果相当相似（Raymaekers 和 Rousseeuw，2022b）。

另一种观点认为，在逐单元格情境中，如果某个样本中有太多单元格被标记，或必须更改单元格才能使该样本回归正常，那么就可以将其视为逐案例离群值。这一思路被用作算法工具，例如在2SGS方法中，样本会在GSE步骤中整体降低权重；MacroPCA方法中也采用了这一思路。作为

一种解释结果的方式，这一思路也已应用于**DDC**，其中有一条规则用于标记包含大量单元格离群程度的行。在这种情况下，它属于一种后处理步骤。不过，我们仍应意识到，这种观点与按行和按单元格离群值的生成模型并不完全相符，除非假设存在一个同时生成这两类离群值的模型。确实有几位作者生成了同时包含这两类离群值的数据。

5.2 稀疏性的启示

近年来，计算机科学和电气工程领域对稳健性问题的关注日益增加。一种尤其流行的处理观测级稳健性的方法，是在现有目标函数中加入一个关于用于编码离群值的参数向量的惩罚项。由此，估计得到的参数向量会变得稀疏，对应于数量较少的标记离群值。据我们所知，这一思路最早见于 Sardy 等人（2001）：他们指出，在回归中使用 Huber 损失函数，与对位移参数 $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$施加$L^1$ 惩罚，两者之间存在对应关系。具体而言，他们证明，最小化

$$\sum_{i=1}^{n} \left(y_i - \boldsymbol{\beta}\boldsymbol{x}_i - \delta_i\right)^2 + \lambda \sum_{i=1}^{n} |\delta_i| \tag{27}$$

ˆover $\boldsymbol{\beta}$ 和 $\boldsymbol{\delta}$ 得到的结果，与最小化所得的 $\boldsymbol{\beta}$ 相同

$$\sum_{i=1}^{n} \rho_{\mathrm{Huber},\lambda}\left(y_i - \boldsymbol{\beta}\boldsymbol{x}_i\right) \ . \tag{28}$$

McCann 和 Welsch（2007）、Laska 等人（2009）以及 Li（2013）也采用了这种方法，随后 She 和 Owen（2011）进一步阐述了它与稳健统计中 M 估计的一般联系。他们证明，对 $\boldsymbol{\delta}$ 施加非凸惩罚会带来更稳健的估计量，这与采用非凸损失函数的 M 估计量可能比采用凸损失函数的 M 估计量更稳健的道理相同。尽管如此，针对凸惩罚和凸损失函数发表的研究仍多于针对非凸惩罚和非凸损失函数的研究，因为凸函数更便于证明定理和编写算法。

类似的方法也可用于处理单元格异常值。与其使用一个 n 维的位移向量 $\boldsymbol{\delta}$，不如使用一个由位移构成的 $n \times d$ 矩阵 $\boldsymbol{\Delta}$ 来描述单元格异常值，并对 $\boldsymbol{\Delta}$施加惩罚，使其变得稀疏。在低秩矩阵补全的背景下，Zhou 等人（2010）和 Cand`es 等人（2011）采用了这种方法。我们的研究背景与他们不同，主要有两个原因。首先，他们的方法是针对主成分分析提出的，但对数据矩阵的转置具有等变性；而在主成分分析的稳健统计文献中，可以存在按行异常值，例如完全污染的行，但不允许存在完全污染的列，因为这会导致载荷无法识别。一个更根本的区别是，他们的方法

并非旨在抵御对抗性污染，而是假设污染单元按照均匀分布出现在随机位置，且彼此独立，也独立于干净数据。正因这一假设，他们不需要使用非凸惩罚；事实上，他们使用的是 L^1 惩罚。他们的方法受到了广泛关注；在异常值与真实低秩子空间之间的正交距离不大时，该方法在其假设下表现良好（She 等人，2016年）。

对于回归，可以通过最小化采用类似的方法

$$\underset{\boldsymbol{\beta},\boldsymbol{\Delta}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} (y_i - \boldsymbol{\beta}(\boldsymbol{x}_i - \boldsymbol{\Delta}_{i,\cdot}))^2 + \lambda \mathcal{P}(\boldsymbol{\Delta}) \right) \tag{29}$$

其中，$\mathcal{P}(\cdot)$ 表示对 $\boldsymbol{\Delta}$ 中单元格的惩罚。Zhu 等人（2011年）提出了类似的思想，使用凸弗罗贝尼乌斯范数作为惩罚项，即 $\mathcal{P}(\cdot) = ||\cdot||_F$。这种方法旨在处理位移矩阵稠密但有界的情况，但对于潜在的严重单元格异常值构成的稀疏集合而言则不太适用。Chen 等人（2013年）沿着这一方向进行了进一步研究。与 She 和 Owen（2011年）一样，他们得出结论：基于凸优化的方法无法抵抗协变量中的观测异常值或单元格异常值。相反，他们使用截尾内积。

对于协方差矩阵的估计，也可以尝试一种类似于（29）的方法。自然的做法是最小化位移数据的高斯负对数似然，如下

$$\underset{\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\Delta}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(\log|\boldsymbol{\Sigma}| + (\boldsymbol{x}_i - \boldsymbol{\Delta}_{i,\cdot} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\Delta}_{i,\cdot} - \boldsymbol{\mu}) \right) + \lambda \mathcal{P}(\boldsymbol{\Delta}) \right) \tag{30}$$

其中，$\mathcal{P}(\cdot)$ 同样表示对 $\boldsymbol{\Delta}$ 中元素的惩罚。尽管这种优化可以相当高效地进行，但我们发现它会产生一种使估计协方差矩阵收缩的偏差。Raymaekers 和 Rousseeuw（2022b）的单元级最小协方差行列式是对这一方法的进一步发展。它并非对位移矩阵 $\boldsymbol{\Delta}$ 进行惩罚，而是对一个$n \times d$ 矩阵 $\boldsymbol{W}$ 进行惩罚，该矩阵由零和一组成，用于指示在估计中应包含哪些单元格。此时，目标函数变为由所包含单元格构成的不完整数据集的观测似然。这避免了估计协方差矩阵中的偏差。

5.3 单元格异常值模型

Alqallaf 等人（2009）将单元格异常值的概念形式化，旨在取代式（1）中的按观测 Tukey-Huber 污染模型（THCM）。更确切地说，他们假设我们观测到一个由以下过程生成的随机变量

$$\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{B})\boldsymbol{Y} + \boldsymbol{B}\boldsymbol{Z} \tag{31}$$

其中，$\boldsymbol{B}$ 是一个对角矩阵，其对角元素只能取零或一。具体而言，他们假设 $\boldsymbol{B} = \text{diag}(B_1, \ldots, B_d)$，其中 $B_j \sim$服从伯努利分布 Bernoulli(ε)。Alqallaf 等人（ 2009 ）将其称为独立污染模型（ICM）

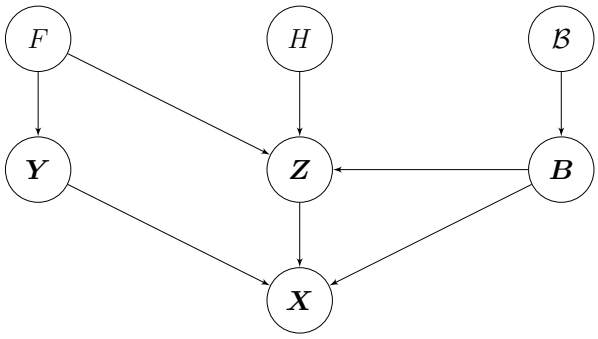
不过，这个名称需要稍加斟酌，因为“独立”一词并非指统计独立性，而只是意味着 $\boldsymbol{B}$ 的对角元素可以彼此不同。正如 Alqallaf 等人（2009）所指出的，对 (31) 施加等式 $B_1 = \cdots = B_d$，即可得到 THCM；此时， $B_1, \ldots, B_d$ 并非统计独立。另一个极端是 $B_1, \ldots, B_d$ 实际上彼此统计独立的情形。如果除此之外，我们还假设 $\boldsymbol{B}$ 独立于 $\boldsymbol{Y}$ 和 $\boldsymbol{Z}$，那么就得到他们所称的完全独立污染模型（FICM）。在 FICM 下，每个单元受到污染的概率为 ε，而单元是否受到污染，既不取决于干净 $\boldsymbol{Y}$ 的取值，也不取决于污染 $\boldsymbol{Z}$ 的取值。

尽管开发适用于 FICM 模型的方法已经很困难，但我们想指出，正如 Alqallaf 等人（2009年）已经提到的，FICM 是一种限制性模型，可以进行大幅扩展。回顾一下，遵循 FICM 的随机向量 $\boldsymbol{X}$ 可按式（31）表示，其中 $\boldsymbol{Y}$ 服从某种干净分布 F，而 $\boldsymbol{Z}$ 服从某种未指定分布 H。重要的是， B_1、.....、B_d、$\boldsymbol{Y}$、$\boldsymbol{Z}$ 被假定为相互独立。有人可能会认为，这一独立性假设相当强。此外，迄今为止，在大多数实证研究中，生成离群单元时还额外假设 $\boldsymbol{Z}$ 的各分量也彼此独立，而且往往甚至是常数。这会产生边缘离群的单元，因此原则上可以通过查看各变量的边缘分布来检测。

在式（1）的逐案例 THCM 模型中，通常假设 B 与 $\boldsymbol{Y}$ 独立，即某个样本是否从污染分布中抽取，与干净数据的取值无关。通常不对 H 作任何假设，但在模拟中，H 有时取决于干净数据的分布 F。例如，在协方差估计中，人们常沿着 F 的真实协方差矩阵的最后一个特征向量方向生成离群案例，而这通常是最具对抗性的选择。

FICM 模型也可以用类似的方式扩展。与 THCM 一样，我们不希望 $\boldsymbol{B}$ 依赖于 $\boldsymbol{Y}$。但是， $\boldsymbol{Z}$ 的取值可以依赖于 $\boldsymbol{B}$ 以及变量 $\boldsymbol{Y}$的分布F。Raymaekers 和 Rousseeuw（2021a，2022b）提出的结构性（对抗性）单元格异常值生成方法就是如此。在该方法中， $\boldsymbol{B}$的对角元素独立地从伯努利(ε)分布中抽取。接下来，根据 $\boldsymbol{Y}$在具有 $B_j = 1$的变量上的分布的协方差矩阵的最后一个特征向量，计算污染单元。这用到了 $\boldsymbol{Y}$的分布，但不涉及 $\boldsymbol{Y}$本身的取值。

这些依赖关系可以用有向无环图来描述



其中，$\mathcal{B}$ 是 $\boldsymbol{B}$的多元分布。最终观测到的 $\boldsymbol{X}$ 位于底部。请注意，存在一支从分布 F 指向 $\boldsymbol{Z}$的箭头，但不存在从价值 $\boldsymbol{Y}$ 指向 $\boldsymbol{Z}$的箭头。我们认为，这一模型最接近 THCM 研究框架中一直占主导地位的理念。与迄今为止主要考虑的情形相比，该模型允许存在更具挑战性的单元格异常值。另一方面，这类离群值在实际中是否真的常见，是一个值得探讨的问题。要回答这个问题，我们需要恰当的方法来检测它们。

5.4 单元格异常值与重尾

我们认为，强调污染模型（31）与重尾这一概念之间的差异十分重要。重尾分布通常被定义为尾部不受指数界约束的概率分布，例如柯西分布。近年来，针对重尾数据的方法受到了广泛关注，例如 Fan 等人（2021）避免了无处不在的次高斯限制。但请注意，在这一语境下，并不存在对抗性污染或结构化离群值的概念。

尽管重尾分布通常可能生成看似能够由（31）生成的样本，但二者之间存在关键差异。若观测随机向量的边缘分布遵循重尾分布，那些看似离群的单元格仍然携带有关底层分布的信息。相反，在模型（31）下，我们不希望对 $\boldsymbol{Z}$ 的分布作出任何假设，因此无法利用 $\boldsymbol{Z}$ 来获取关于 $\boldsymbol{Y}$ 的分布特征的信息，而后者才是我们所关注的。因此，由 $\boldsymbol{Z}$ 生成的单元格不携带关于 $\boldsymbol{Y}$ 的信息。在逐单元污染模型中，我们因此应当努力消除 $\boldsymbol{Z}$ 的影响。若有人告诉我们哪些单元格是由 $\boldsymbol{Z}$ 生成的，那么最好的做法便是将其舍弃，并继续处理其余单元格。

举一个简单的例子，考虑一个包含单个极端值的单变量数据集。如果我们确切知道数据来自柯西分布，那么就应该保留这个异常值，并计算柯西分布的最大似然估计。反之，如果我们知道数据来自高斯分布，只是这个可疑值来自某种其他分布，那么就应该将其舍弃，并用其余数值计算高斯分布的最大似然估计。

尽管两者的模型和目标不同，但融合这两个领域的思路或许仍有价值。一般而言，我们预计，针对重尾数据的估计量在逐单元污染下仍能表现良好。这至少使它们有望成为迭代算法的初始估计量，而这些算法旨在处理可能包含单元格异常值的数据。

6 结论

单元格异常值为统计数据分析带来了引人入胜但棘手的挑战。本文回顾了用于检测单元格异常值、估计位置和协方差，以及执行回归和主成分分析等其他任务的一些方法。根据表1至表4，我们得出结论：计算时间并非采用逐单元稳健方法的障碍。我们指出，若坚持某些直观概念，位置、协方差和回归估计量的单元格崩溃值相当低，这意味着我们必须探索新的方向。随后，我们提出了一种用于对应分析的逐单元稳健方法。最后，我们讨论了若干值得深入思考的问题，这些问题或可引发讨论并推动进一步研究。

软件可用性。 第4节所述的单元级稳健对应分析方法的代码可在 R软件包 `cellWise` (Raymaekers 和 Rousseeuw，2022a)中获取，并附有一个重现图4至图6的示例文档 `Correspondence analysis examples` 。 - -

致谢。 感谢 Mia Hubert 进行了富有启发性的讨论 .

附录

此处给出第 3 节中崩溃结果的证明。

命题 *1* 的证明。 考虑由方程 $\boldsymbol{x}^\top \mathbf{1}_d = c$ 给出的超平面 H，其中 $\mathbf{1}_d$是全 1 向量，且 c 取值足够大，使得所有数据点 $\boldsymbol{x}_i$ 都满足$\boldsymbol{x}_i^\top \mathbf{1}_d < c$。因此， $\boldsymbol{X}$ 位于 H的一侧，而 H 不平行于任何坐标轴。于是，我们可以在 $\boldsymbol{X}$ 的每一行中替换一个单元，使得所有生成的数据点都位于H上，从而得到污染数据集 $\boldsymbol{X}^m$。在每个变量中，我们最多替换 $\lceil n/d \rceil$个单元格，即其 n 个单元格的 $\lceil n/d \rceil / n$ 。由性质 (10) 可知，$\hat{}\boldsymbol{\mu}(\boldsymbol{X}^m)$ 也位于 H 上。由于可以任意增大取值 c，$\hat{}\boldsymbol{\mu}(\boldsymbol{X}^m)$ 会崩溃。

□

推论*1*的证明。考虑一个数据集 $\{\boldsymbol{x}_i;\ i = 1, \ldots, n\}$，它位于超平面 H上。于是， $\boldsymbol{x}_i' = \boldsymbol{x}_i - \boldsymbol{x}_1$ 位于平移后的超平面 H_0 上，该超平面经过原点。现在取一个正交矩阵 $\boldsymbol{U}$，将 H_0 映射到由前 d 个标准正交基向量张成的超平面上。这样便将数据点映射到 $\widetilde{\boldsymbol{x}}_i = \boldsymbol{U}\boldsymbol{x}_i'$。现在取正交矩阵 $\boldsymbol{V} =$

$\mathrm{diag}(1,\dots,1,-1)$。由正交等变性可得 $\hat{\boldsymbol{\mu}}(\{\boldsymbol{V}\widetilde{\boldsymbol{x}}_i\}) = \boldsymbol{V}(\hat{\boldsymbol{\mu}}(\{\widetilde{\boldsymbol{x}}_i\}))$。但是$\{\boldsymbol{V}\widetilde{\boldsymbol{x}}_i\} = \{\widetilde{\boldsymbol{x}}_i\}$，因此 $\hat{\boldsymbol{\mu}}(\{\widetilde{\boldsymbol{x}}_i\}) = \boldsymbol{V}(\hat{\boldsymbol{\mu}}(\{\widetilde{\boldsymbol{x}}_i\}))$，所以其最后一个坐标必为零。因此$\hat{\boldsymbol{\mu}}(\{\boldsymbol{x}_i\}) = \boldsymbol{U}^\top \hat{\boldsymbol{\mu}}(\{\widetilde{\boldsymbol{x}}_i\}) + \boldsymbol{x}_1$位于 H上，这就证明了式（10）。 □

命题*2*的证明。当 $n \leqslant d$ 所有 $\boldsymbol{X}$ 中的点都位于一个低维仿射子空间上时，因此 (15) 表明 $\lambda_d(\boldsymbol{\Sigma}(\boldsymbol{X})) = 0$，所以 $m = 0$ 就足以导致崩溃。接下来，我们可以假设 $n > d$。选取数据集的一行，并考虑一个不平行于任何坐标轴的仿射超平面。于是，我们可以在其余 $n-1$ 行中替换一个单元，使得所有所得点都位于同一个超平面上，从而得到污染数据集 $\boldsymbol{X}^m$。我们可以在每个变量中替换不超过 $\lceil (n-1)/d \rceil$ 个单元格，这占其 n个单元格的 $\lceil (n-1)/d \rceil / n$。根据精确拟合性质 (15)，可得 $\lambda_d(\boldsymbol{\Sigma}(\boldsymbol{X}^m)) = 0$，因此我们得到内爆。

□

推论 *2* 的证明。 考虑一个 $\boldsymbol{X}$ 位于低维子空间中的数据集。我们进行平移并旋转 $\boldsymbol{X}$，使其像 $\boldsymbol{Y}$位于由前$d-1$ 个标准正交基向量张成、经过原点的超平面中。根据 $\boldsymbol{\Sigma}$ 的仿射等变性，可得 $\boldsymbol{C} := \boldsymbol{\Sigma}(\boldsymbol{Y})$与 $\boldsymbol{\Sigma}(\boldsymbol{X})$具有相同的特征值。现在我们将 C划分为大小分别为 $d-1$和 1 的块，并考虑以下矩阵 $\boldsymbol{A}$:

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{21}^\top \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{bmatrix} \quad \text{and} \quad \boldsymbol{A} = \begin{bmatrix} \boldsymbol{I}_{d-1} & \boldsymbol{0} \\ \boldsymbol{0} & 2 \end{bmatrix}.$$

现在，对 线性变换 $\boldsymbol{Y}$为 $\boldsymbol{Y} = \boldsymbol{Y}\boldsymbol{A}$。根据 $\boldsymbol{\Sigma}$ 的仿射等变性，我们得到 in

$$\widehat{\boldsymbol{\Sigma}}(\widetilde{\boldsymbol{Y}}) = \boldsymbol{A}^\top \boldsymbol{C} \boldsymbol{A} = \begin{bmatrix} \boldsymbol{C}_{11} & 2\boldsymbol{C}_{21}^\top \\ 2\boldsymbol{C}_{21} & 4\boldsymbol{C}_{22} \end{bmatrix}.$$

但后一个矩阵必定等于 $\boldsymbol{C}$，因为根据构造，$\boldsymbol{Y} = \boldsymbol{Y}$。因此 $2\boldsymbol{C}_{21} = \boldsymbol{C}_{21}$且 $4\boldsymbol{C}_{22} = \boldsymbol{C}_{22}$，这只有在 $\boldsymbol{C}_{21} = \boldsymbol{0}$ 且 $\boldsymbol{C}_{22} = 0$时才可能，所以 $\boldsymbol{C}$ 是奇异的。 □

命题*3*的证明。 取数据集中的第一个样本 $\boldsymbol{z}_1$，并考虑任意一个非零数β_0。然后计算 $\alpha_0 = y_1 - \beta_0 x_{11} - \ldots - \beta_0 x_{1p}$。将它们合并为向量 $\boldsymbol{\gamma}_0 =(\alpha_0, \beta_0, \ldots\ldots, \beta_0)^\top$。于是，观测1位于方程为 $y = \boldsymbol{x}^\top \boldsymbol{\gamma}_0$的超平面 H 上。注意，H 不平行于任何坐标轴。对于其余的 $n-1$ 个样本，我们可以分别替换 $\boldsymbol{Z}$ 的一个单元格，使所有得到的点都位于 H上，从而得到污染数据集 $\boldsymbol{Z}^m$。对于每个变量（包括响应变量），我们至多替换 $\lceil (n-1)/(p+1) \rceil$ 个单元格，即其 n 个单元格的 $\lceil (n-1)/(p+1) \rceil / n$。由式（20）可得 $\hat{\boldsymbol{\gamma}}(\boldsymbol{Z}^m) = \boldsymbol{\gamma}_0$。由于 β_0 可以任意大，$\hat{\boldsymbol{\gamma}}(\boldsymbol{Z}^m)$会崩溃。

□

推论 3 的证明。考虑数据集 $(\boldsymbol{X}, \boldsymbol{y})$，其中对于所有 $i = 1, \ldots, n$，存在某个 $\boldsymbol{\gamma}_0$，使得 $y_i = \boldsymbol{x}_i^\top \boldsymbol{\gamma}_0$。构造 $\widetilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\gamma}_0 = \boldsymbol{0}$。根据回归等变性，$\hat{\boldsymbol{\gamma}}(\boldsymbol{X}, \widetilde{\boldsymbol{y}}) = \hat{\boldsymbol{\gamma}}(\boldsymbol{X}, \boldsymbol{y}) - \boldsymbol{\gamma}_0$。利用 $\widetilde{\boldsymbol{y}} = 0\widetilde{\boldsymbol{y}}$ 这一事实以及在 $c = 0$ 下的尺度等变性，我们得到 $\hat{\boldsymbol{\gamma}}(\boldsymbol{X}, \widetilde{\boldsymbol{y}}) = \hat{\boldsymbol{\gamma}}(\boldsymbol{X}, 0\widetilde{\boldsymbol{y}}) = 0\hat{\boldsymbol{\gamma}}(\boldsymbol{X}, \widetilde{\boldsymbol{y}}) = \boldsymbol{0}$。因此 $\hat{\boldsymbol{\gamma}}(\boldsymbol{X}, \boldsymbol{y}) = \hat{\boldsymbol{\gamma}}(\boldsymbol{X}, \widetilde{\boldsymbol{y}}) + \boldsymbol{\gamma}_0 = \boldsymbol{\gamma}_0$，由此证明了 (20)。 □

参考文献

Aerts, S., Wilms, I., 2017年。逐单元稳健正则判别分析。Statistical Anal- ysis and Data Mining: The ASA Data Science Journal 10 (6), 436–447。Agostinelli, C., Leu ng, A., Yohai, V. J., Zamar, R. H., 2015年。存在单元格污染和案例污染时多元位置与散布的稳健估计。TEST 24, 441– 461。Agostinelli, C., Yohai, V. J., 2016年。线性混合模型的复合稳健估计量。美国统计学会会刊 111 (516), 1764–1774。Alqallaf, F., Van Aelst, S., Yohai, V. J., Zamar, R. H., 2009年。多元数据中的离群值传播。The Annals of Statistics 37, 311–331。Benz´ecri, J.-P., 等，1973年。L'analyse des donn´ees。第 2 卷。Dunod Paris。Bezdek, J., Hathaway, R., Howard, R., Wilson, C., Windham, M., 1987年。坐标下降分组变量版本的局部收敛分析。《优化理论与应用杂志》54, 471–477。Bottmer, L., Croux, C., Wilms, I., 2022年。含离群值大型数据集的稀疏回归。《欧洲运筹学杂志》297 (2), 782–794。Boudt, K., Cornelissen, J., Croux, C., 2012。高斯秩相关估计量：稳健性性质。《统计与计算》22 (2), 471–483。Cand`es, E. J., Li, X., Ma, Y., Wright, J., 2011年。稳健主成分分析？《美国计算机学会期刊》58 (3), 1–37。Chen, Y., Caramanis, C., Mannor, S., 2013年。对抗性污染下的稳健稀疏回归。载于：国际机器学习会议。机器学习研究会论文集，第 774–782 页。Croux, C., Filzmoser, P., Fritz, H., 2013年。稳健稀疏主成分分析。Technometrics 55, 202–214。Croux, C., Filzmoser, P., Oliveira, M., 2007年。基于投影寻踪的稳健主成分分析算法。《化学计量学与智能实验室系统》87, 218–225。Croux, C., Filzmoser, P., Pison, G., Rousseeuw, P. J., 2003年。通过稳健交替回归拟合乘法模型。《统计与计算》13, 23–36。¨Croux, C., Ollerer, V., 2016年。利用秩相关度量对逆协方差矩阵进行稳健稀疏估计。载于：《稳健统计学最新进展：理论与应用》。施普林格，第 35–55 页。

Danilov, M., 2010年。非仿射等变情形下的多元散布稳健估计。博士论文，不列颠哥伦比亚大学。Danilov, M., Yohai, V. J., Zamar, R. H., 2012。缺失数据存在时多元位置与散布的稳健估计。《美国统计学会会刊》107，1178–1186。Debruyne, M., H¨oppner, S., Serneels, S., Verdonck, T., 2019。偏离度：哪些变量贡献最大？《统计与计算》29，707–723。Donoho, D., Huber, P., 1983。崩溃点的概念。载于：Bickel, P., Doksum, K., Hodges, J.（编），Erich Lehmann纪念文集。Wadsworth，贝尔蒙特，第157–184页。Donoho, D. L., 1982。多元位置估计量的崩溃性质。博士资格论文，哈佛大学。Fan, J., Wang, W., Zhu, Z., 2021。重尾数据的收缩原理：高维稳健低秩矩阵恢复。《统计年鉴》49（3），1239。Farcomeni, A., 2014。存在逐项离群值时的稳健约束聚类。Technometrics 56（1），102–111。Farcomeni, A., Leung, A., 2019。snipEM软件包：用于稳健估计与聚类的剪除方法。CRAN，R软件包1.0.1版。Filzmoser, P., H¨oppner, S., Ortner, I., Serneels, S., Verdonck, T., 2020。逐单元稳健M回归。Computational Statistics & Data Analysis 147，106944。Friedman, J., Hastie, T., Tibshirani, R., 2008。利用图形套索进行稀疏逆协方差估计。《生物统计学》9（3），432–441。Garc´ıa-Escudero, L.-A., Rivera-Garc´ıa, D., Mayo-Iscar, A., Ortega, J., 2021。采用逐单元修剪的聚类分析及其在曲线稳健聚类中的应用。《信息科学》573，100–124。Gervini, D., Yohai, V. J., 2002。一类稳健且完全有效的回归估计量。《统计年鉴》30（2），583–616。Gnanadesikan, R., Kettenring, J. R., 1972。利用多响应数据进行稳健估计、残差分析和异常值检测。《生物测定学》，81–124。Hampel, F. R., Ronchetti, E. M., Rousseeuw, P., Stahel, W. A., 1986。《稳健统计：基于影响函数的方法》。Wiley-Interscience，纽约。

Hirschfeld, H. O., 1935年。相关与列联之间的联系。载于：《剑桥哲学学会数学学报》第31卷。剑桥大学出版社，第520–524页。Huber, P. J., 1964。位置参数的稳健估计。《数学统计年鉴》35（1），73–101。Hubert, M., Reynkens, T., Schmitt, E., Verdonck, T., 2016。含离群值的高维数据稀疏主成分分析。Technometrics 58，424–434。Hubert, M., Rousseeuw, P. J., Segaert, P., 2015。多元函数型离群值检测（附讨论）。《统计方法与应用》24，177–202。Hubert, M., Rousseeuw, P. J., Van den Bossche, W., 2019。MacroPCA：一种一体化主成分分析方法，可处理缺失值以及按单元格和按行的离群值。Technometrics 61（4），459–473。URL https://doi.org/10.1080/00401706.2018.1562989Hubert, M., Rousseeuw, P. J., Vanden Branden, K., 2005。ROBPCA：一种稳健主成分分析的新方法。Technometrics 47，64–79。Katayama, S., Fujisawa, H., Drton, M., 2018。单元格污染下的稳健稀疏高斯图模型。《统计》7（1），e181。Laska, J. N., Davenport, M. A., Baraniuk, R. G., 2009。通过正义追求从稀疏受污染测量中进行精确信号恢复。载于：2009年阿西洛马信号、系统与计算机会议论文集。IEEE，第1556–1560页。Leung, A., Danilov, M., Yohai, V., Zamar, R., 2019。GSE软件包：存在单元格污染、案例污染和缺失数据时的稳健估计。CRAN，R软件包4.2版。URL https://CRAN.R-project.org/package=GSELeung, A., Yohai, V., Zamar, R., 2017。单元格污染和案例污染下的多元位置与散布矩阵估计。Computational Statistics & Data Analysis 111，59–76。Leung, A., Zhang, H., Zamar, R., 2016。单元格污染和案例污染存在时的稳健回归估计与推断。Computational Statistics & Data Analysis 99，1–11。Li, X., 2013。具有恒定污染比例的压缩感知与矩阵补全。《构造逼近》37（1），73–99。Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J., Cohen, K., 1999。函数型数据的稳健主成分分析。TEST 8，1–73。

Loh, P.-L., Tan, X. L., 2018年。高维稳健精度矩阵估计： ϵ污染下的按单元格污染。Electronic Journal of Statistics 12, 1429–1467。

Lopuha¨a, H. P., Rousseeuw, P. J., 1991年。多元位置和协方差矩阵仿射等变估计量的崩溃点。The Annals of Statistics, 229–248。

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., R¨uckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E., di Palma, M., 2022年。robustbase：稳健统计基础。R软件包，CRAN。网址 https://CRAN.R-project.org/package=robustbaseMaronna, R., 2005。基于稳健尺度的主成分与正交回归。Technometrics 47, 264–273。Maronna, R. A., Martin, R. D., Yohai, V. J., Salibi´an-Barrera, M., 2019。稳健统计：理论与方法（附 R）。John Wiley & Sons。Maronna, R. A., Yohai, V. J., 2008。基于逐元素污染的数据矩阵稳健低秩近似。Technometrics 50 (3), 295–304。McCann, L., Welsch, R. E., 2007。使用最小角回归和元素集抽样进行稳健变量选择。Computational Statistics & Data Analysis 52, 249–257。¨Ollerer, V., Alfons, A., Croux, C., 2016。用于稳健回归的射击 S 估计量。Computational Statistics 31, 829–844。¨Ollerer, V., Croux, C., 2015。稳健高维精度矩阵估计。载于：现代非参数、稳健与多变量方法。施普林格，第 325–350 页。Ollila, E., Tyler, D., 2014。散布矩阵的正则化 M 估计量。IEEE信号处理汇刊 62, 6059–6070。Raymaekers, J., Rousseeuw, P., 2021a年。通过稀疏回归和稳健协方差处理单元格异常值。Journal of Data Science, Statistics, and Visualisation 1 (3)。网址 https://doi.org/10.52933/jdssv.v1i3.18Raymaekers, J., Rousseeuw, P., 2022a。cellWise：分析含单元格异常值的数据。R软件包。网址 https://CRAN.R-project.org/package=cellWiseRaymaekers, J., Rousseeuw, P. J., 2021b年。适用于高维数据的快速稳健相关。Technometrics 63, 184–198。网址 https://doi.org/10.1080/00401706.2019.1677270Raymaekers, J., Rousseeuw, P. J., 2021c年。将变量变换为中心正态性。Machine Learning, 1–23。网址 https://link.springer.com/article/10.1007/s10994-021-05960-5

Raymaekers, J., Rousseeuw, P. J., 2022b。单元格级最小协方差行列式估计量。arXiv 预印本。网址 https://arxiv.org/abs/2207.13493Riani, M., Atkinson, A. C., Torti, F., Corbellini, A., 2022年。稳健对应分析。皇家统计学会会刊C辑：应用统计 71, 1381–1401。Rousseeuw, P. J., 1984。最小平方中位数回归。美国统计学会会刊 79, 871–880。Rousseeuw, P. J., 2023年。分析单元格级加权数据。arXiv 预印本。网址 https://arxiv.org/abs/2209.12697。计量经济学与统计学，即将发表。Rousseeuw, P. J., Croux, C., 1993。绝对中位差的替代方法。美国统计学会会刊 88, 1273–1283。Rousseeuw, P. J., Leroy, A. M., 1987年。稳健回归与异常值检测。John Wiley & Sons。Rousseeuw, P. J., Van den Bossche, W., 2018年。检测偏离数据单元。Techno metrics 60, 135–145。网址 https://doi.org/10.1080/00401706.2017.1340909 Rousseeuw, P. J., van Zomeren, B. C., 1990年。揭示多变量异常值和杠杆点。美国统计学会会刊 85, 633–639。Rousseeuw, P. J., Yohai, V., 1984。利用 S 估计量进行稳健回归。载于：稳健与非线性时间序列分析，J. Franke、W. H¨ardle、R.D. Martin 编，统计学讲义第 26 卷，施普林格，纽约，第 256–272 页。Saraceno, G., Agostinelli, C., 2021年。基于统计深度滤波器的稳健多变量估计。TEST, 1–25。Saraceno, G., Alqallaf, F., Agostinelli, C., 2021年。一种适用于行级和单元格级污染的稳健似不相关回归。arXiv 预印本 arXiv:2107.00975。Sardy, S., Tseng, P., Bruce, A., 2001年。稳健小波去噪。IEEE 信号处理汇刊 49 (6), 1146–1152。She, Y., Li, S., Wu, D., 2016年。稳健正交补主成分分析。美国统计学会会刊 111 (514), 763–771。She, Y., Owen, A. B., 2011年。使用非凸惩罚回归进行异常值检测。美国统计学会会刊 106 (494), 626–639。Stahel, W. A., 1981年。稳健估计：无穷小最优性及协方差矩阵估计。博士论文，ETH Z¨urich。

ˇStefelov´a, N., Alfons, A., Palarea-Albaladejo, J., Filzmoser, P., Hron, K., 2021. 含有单元格异常值的组成型协变量稳健回归。《数据分析与分类进展》，1–41。Su, P., Tarr, G., Muller, S., 2021. 逐单元污染下的稳健变量选择。arXiv预印本 arXiv:2110.12406。Tarr, G., M¨uller, S., Weber, N., 2016. 逐单元污染下精度矩阵的稳健估计。Computational Statistics & Data Analysis 93, 404–420。Toka, O., C¸etin, M., Arslan, O., 2021. 存在单元格异常值和案例级离群值时的稳健回归估计与变量选择。《哈切特佩数学与统计学杂志》50, 289–303。Van Aelst, S., Vandervieren, E., Willems, G., 2011. 采用单元级权重的Stahel-Donoho估计量。《统计计算与模拟杂志》81, 1–27。ˇ Walach, J., Filzmoser, P., Kouˇril, S., Friedeck`y, D., Adam, T., 2020. 基于成对对数比的代谢组学单元格异常值检测与生物标志物识别。《化学计量学杂志》34 (1), e3182。Wang, Y., Van Aelst, S., 2020. 基于最小截尾平方的稀疏主成分分析。Technometrics 62 (4), 473–485。Zhou, Z., Li, X., Wright, J., Cand`es, E., Ma, Y., 2010. 稳定主成分追踪。载于：2010年IEEE国际信息论研讨会。IEEE, 第1518–1522页。Zhu, H., Leus, G., Giannakis, G. B., 2011. 用于扰动压缩采样的稀疏性感知总最小二乘。IEEE 信号处理汇刊 59, 2002–2016。