通过基于均值和中位数的检验进行通用离群假设检验

Bernhard C. Geiger*†¶、Tobias Koch‡§、Josipa Mihaljevi´c† 和 Maximilian Toller†
*信号处理与语音通信实验室，格拉茨技术大学，奥地利格拉茨 8010
†Know Center Research GmbH，奥地利格拉茨 8010
¶格拉茨机器学习中心，奥地利格拉茨 8010
‡信号理论与通信系，马德里卡洛斯三世大学，西班牙莱加内斯 28911§格雷戈里奥·马拉尼翁健康研究所，西班牙马德里 28007 电子邮箱：geiger@ieee.org， {jmihaljevic,mtoller}@know-center.at，tkoch@ing.uc3m.es

arXiv:2601.00712v1 [cs.IT] 2 Jan 2026

摘要—通用离群假设检验指这样一种假设检验问题：人们观测大量长度为n 的序列，其中大多数服从典型分布 π ，少数服从离群分布 μ；人们希望在不了解 π 和 μ的情况下，判断哪些序列是异常值。与以往研究不同，本文假设观测序列的数量和异常序列的数量都随序列长度增长。在这种情况下，只要异常序列的数量相对于序列总数是次线性的，就可以通过对所有观测序列求均值来估计典型分布 π 。研究表明，在这种情况下，可以达到能够获取 π 和 μ的最大似然检验的错误指数。然而，当异常序列的数量与序列总数成正比时，这种基于均值的检验表现不佳。针对这种情况，本文提出一种基于中位数的检验，通过将所有观测序列的中位数作为估计量来估计 π 。研究表明，这种基于中位数的检验同样能够达到能够获取 π 和 μ的最大似然检验的错误指数，但其成立概率趋近于一。为了形式化地描述这种情况，本文提出典型错误指数，类似于信道编码中随机编码背景下引入的典型随机编码指数。

I. 引言

Li、Nitinawarat 和 Veeravalli [1], 提出的离群假设检验问题，旨在从大量序列中判定哪些是异常值。更确切地说，在离群假设检验中，我们观测到 M 个相互独立、长度为 n 的序列$\mathbf{Y}_1,\dots,\mathbf{Y}_{\mathsf{M}}$，其中 M −T 个序列按照典型分布 π独立同分布（i.i.d.），另外 T $<$ M/2个序列按照离群分布 $\mu \neq \pi$独立同分布；两种分布的取值均来自同一个有限字母表 $\mathcal{Y}$。目标是判定哪些序列服从 μ，即识别异常值序列。对于任意固定的 M，这可以表述为在搜索空间 $[\mathsf{M}] \triangleq \{1,\dots,\mathsf{M}\}$ 中寻找某个泛函的最优解的问题。

我们用函数$\delta\colon \mathcal{Y}^{\mathsf{M}n} \to \bar{\mathcal{S}}$来描述离群假设检验，其中 $\bar{\mathcal{S}}$ 表示 [M] 的所有大小为 T 的子集所构成的集合。换言之，函数 $\delta(\cdot)$以 M 个序列 $\mathbf{Y}_1$，.....，$\mathbf{Y}_{\mathsf{M}}$ 作为输入，并输出索引集合 $\mathcal{S}$，以指出异常值序列。该检验的性能由最大错误概率衡量

$$\epsilon(\delta) = \max_{\mathcal{S}\in\bar{\mathcal{S}}} \sum_{\mathbf{y}_1^{\mathsf{M}}\colon \delta(\mathbf{y}_1^{\mathsf{M}})\neq\mathcal{S}} P_{\mathcal{S}}(\mathbf{y}_1^{\mathsf{M}}) \tag{1}$$

其中， $P_{\mathcal{S}}$表示异常序列的索引为 $\mathcal{S}$ 时$\mathbf{Y}_1,\dots,\mathbf{Y}_{\mathsf{M}}$的联合分布；我们用记号$\mathbf{y}_1^{\mathsf{M}}$表示序列$\mathbf{y}_1,\dots,\mathbf{y}_{\mathsf{M}}$。更确切地说，我们通过考察相应的错误指数，研究 $\epsilon(\delta)$ 趋于零的衰减速度。

$$\alpha(\delta) \triangleq \overline{\lim_{n\to\infty}} -\frac{1}{n}\log\epsilon(\delta) \tag{2}$$

其中，log(·) 是二进制对数函数，lim 表示上极限。

如果已知两个分布 π 和 μ，且已知异常序列的数量 T，那么最大似然检验的错误指数为 $2B(\pi,\ \mu)$ [1, 命题 7], ，其中

$$B(\pi,\mu) \triangleq -\log\left(\sum_{y\in\mathcal{Y}} \pi(y)^{\frac{1}{2}}\mu(y)^{\frac{1}{2}}\right) \tag{3}$$

是 π 和 μ之间的巴氏距离。仅能访问 π 、但不了解 μ [1, 定理 8]的检验也能达到相同的错误指数。

在通用情形下，由于既不知道 π 也不知道 μ，最大似然检验不适用。相反，Li、Nitinawarat 和 Veeravalli 考虑了广义似然比检验（GLRT），其定义见 [1, 式（37）]

$$\delta_{\mathrm{Li}}(\mathbf{Y}_1,\dots,\mathbf{Y}_{\mathsf{M}}) = \operatorname*{argmin}_{\mathcal{S}\in\bar{\mathcal{S}}} \sum_{j\notin\mathcal{S}} D\left(\mathsf{P}_{\mathbf{Y}_j}\middle\|\frac{\sum_{k\notin\mathcal{S}}\mathsf{P}_{\mathbf{Y}_k}}{\mathsf{M}-\mathsf{T}}\right) \tag{4}$$

其中

$$\mathsf{P}_{\mathbf{Y}_k}(y) \triangleq \frac{1}{n}\sum_{\ell=1}^{n}\mathbf{1}\{Y_{k,\ell}=y\}, \quad y\in\mathcal{Y} \tag{5}$$

表示$\mathbf{Y}_k = (Y_{k,\ 1},\ \dots,\ Y_{k,\ n})$的经验分布（型），而 $\mathbf{1}\{\cdot\}$ 表示指示函数

作者姓名按字母顺序排列。

。对于广义似然比检验（GLRT），他们推导出了错误指数的一个下界，该下界严格小于 $2B(\pi,\ \mu)$，但当观测序列的数量Mtends toinfinity [1,Th. 10]时，会收敛于 $2B(\pi,\ \mu)$。这与我们的直觉一致：随着观测序列数量趋于无穷，我们可以通过对所有观测值取平均，准确地估计典型分布。

Li、Nitinawarat 和 Veeravalli 的工作 [1] 随后从多个方向得到了拓展。例如，文献 [2]研究了一种带有拒绝选项的 T 检验 ≤ 1，即该检验可以判定所有观测序列都不是异常值。Nitinawarat 和 Veeravalli [3] 考虑了恰好有一个序列在时间 $1 \leq \lambda \leq n$ 发生变点的情形，并提出了一种用于识别该序列以及 λ的检验。Bu 等人 [4] 针对连续分布 π 和 μ提出了检验。文献 [5]–[7]研究了序贯检测，即提前停止。Acharya 等人 [8] 研究了样本量 n 小于离散分布的字母表大小 $|\mathcal{Y}|$的情形。

上述研究均关注样本量 n 远大于观测序列数量 M 的情形。然而，有时相反的情形也值得关注。一个实际例子是密集传感器网络：由于能量限制，每个传感器都只能在较短的时间间隔内进行感测，或以较低的采样率工作。另一个例子出现在医疗领域，即在相对较短的时间内记录大量患者的数据，例如心电图/脑电图数据或语音录音。大量观测序列的一个优点是，可以准确地估计典型分布，因为

$$\hat{\pi}(y) = \frac{1}{\mathsf{M}} \sum_{k=1}^{\mathsf{M}} \mathsf{P}_{\mathbf{Y}_k}(y), \quad y \in \mathcal{Y} \tag{6}$$

前提是异常序列的数量 T 关于 M 是次线性的。从负面来看，当观测序列的数量较大时，广义似然比检验可能变得难以实施。事实上，该检验会在 [M] 的所有大小为 T 的子集上进行搜索，将其作为候选离群序列集合。因此，其计算复杂度为 $\mathcal{O}(\mathsf{M}^{\mathsf{T}})$；当 M 和 T 较大时，这一复杂度将变得难以承受。在数学上，我们通过令参数 M（观测序列的数量）和 T（异常序列的数量）依赖于序列长度 n，来刻画观测序列数量较大的情形。因此，Li、Nitinawarat 和 Veeravalli [1] 所采用的方法——应用 Sanov 定理，将错误指数表示为相对熵的最小化问题，然后对该最小值进行界定——无法用于广义似然比检验，因为错误概率的次指数项以指数形式依赖于序列的字母表大小 $\mathcal{Y}^{\mathsf{M}}$（$\mathbf{Y}_1$，.....，$\mathbf{Y}_{\mathsf{M}}$），并且对于依赖于 n 的 M，这些项可能变得不可忽略。

为避开这一问题，我们提出基于均值的检验

$$\delta_{\text{mean}}(\mathbf{Y}_1, \ldots, \mathbf{Y}_{\mathsf{M}}) = \underset{\mathcal{S} \in \bar{\mathcal{S}}}{\operatorname{argmax}} \sum_{j \in \mathcal{S}} D\left(\mathsf{P}_{\mathbf{Y}_j} \middle\| \hat{\pi}\right) \tag{7}$$

其错误概率在给定 $\hat{}\pi$ 的条件下可以进行解析分析。特别地，我们证明，如果 $\mathsf{T} = o(\mathsf{M})$ 且 M 关于 n 超线性增长但次指数增长，那么基于均值的检验可达到最优错误指数 $2B(\pi,\ \mu)$。

如果异常序列占所有序列的比例为正，即 $\mathsf{T} = c\mathsf{M}$，$0 < c < \frac{1}{2}$，那么所有类型的均值会产生偏差，从而对基于均值的检验 (7) 的错误指数产生不利影响。在这种情况下，所有类型的中位数

$$\bar{\pi}(y) = \frac{\operatorname{median}\{\mathsf{P}_{\mathbf{Y}_1}(y), \ldots, \mathsf{P}_{\mathbf{Y}_{\mathsf{M}}}(y)\}}{\sum_{y' \in \mathcal{Y}} \operatorname{median}\{\mathsf{P}_{\mathbf{Y}_1}(y'), \ldots, \mathsf{P}_{\mathbf{Y}_{\mathsf{M}}}(y')\}}, \quad y \in \mathcal{Y} \tag{8}$$

可能会对 π提供更好的估计量，因为它对异常值更具鲁棒性。正如我们将要说明的，随着序列长度 n 趋于无穷，中位数会变得准确；但与均值不同，其准确性不会随着 M 的增加而提高。这意味着，尽管中位数对异常值具有鲁棒性，且数值结果似乎表明情况并非如此，但若在 (7) 中以基于中位数的估计 $\bar{\pi}$替换基于均值的估计量 $\hat{}\pi$，所得基于中位数的检验的错误指数仍然较差。因此，我们提出一种两步中位数检验：先从部分序列$\mathbf{Y}_1, \ldots, \mathbf{Y}_{\mathsf{M}}$中计算基于中位数的估计 $\bar{\pi}$，再利用其余部分进行异常值检验。作为一种不那么悲观的性能度量，我们进而提出典型错误指数，将其定义为对 π的所有实现取平均后的错误指数。典型错误指数类似于信道编码随机编码背景下提出的典型随机编码指数；参见例如[9]–[12]。我们证明，两步中位数检验的典型错误指数等于最优错误指数$2B(\pi,\ \mu)$。因此，该检验以趋近于一的概率达到最优指数。

在计算复杂度方面，基于均值的检验和两步基于中位数的检验的计算复杂度为$\mathcal{O}(\mathsf{M}\log\mathsf{M})$，而像广义似然比检验 (4) 那样搜索一个大小为$c\mathsf{M}$ 的子集，其复杂度为 $\underline{\mathcal{O}}(2^{2\mathsf{M}}/\sqrt{\mathsf{M}})$。

本文其余部分安排如下。第二节介绍所考虑的问题设定。第三节讨论基于均值的检验的错误指数。第四节研究两步基于中位数的检验，并给出其典型错误指数。第五节通过数值示例，将所提出的检验与广义似然比检验的性能进行比较。第六节总结全文，并讨论所考虑检验的计算复杂度。

II. 问题设定

假设我们观测到 M_n 个相互独立的长度为-n序列 $\mathbf{Y}_1, \ldots, \mathbf{Y}^{\mathsf{M}_n}$，其中 $\mathsf{M}_n - \mathsf{T}_n$ 个序列独立同分布地服从典型分布 π，而 $\mathsf{T}_n < \mathsf{M}_n/2$ 个序列独立同分布地服从离群分布 $\mu \neq \pi$。在本文中，我们均在 M 和 T 的下标中添加“n”，以表明在符号表示中，这两个参数可能依赖于序列长度 n。我们假设 μ和 π 具有相同的字母表 $\mathcal{Y}$，并且该字母表是有限的。我们进一步假设

$$\pi_{\min} \triangleq \min_{y \in \mathcal{Y}} \pi(y) > 0. \tag{9}$$

根据问题的对称性，不失一般性，我们可以假设前 T_n 个序列是异常值。在这一假设下，序列 $\mathbf{Y}_1,\ldots,\mathbf{Y}^{\mathsf{M}_n}$ 具有联合分布

$$P_{[\mathsf{T}_n]}(\mathbf{y}_1,\ldots,\mathbf{y}_{\mathsf{M}_n})=\prod_{k=1}^{n}\prod_{i=1}^{\mathsf{T}_n}\mu(y_{i,k})\prod_{j=\mathsf{T}_n+1}^{\mathsf{M}_n}\pi(y_{j,k}) \quad (10)$$

其中 $\mathbf{y}_j=(y_{j,1},\ldots\ldots,y_{j,n})^T$，$j=1,\ldots\ldots,\mathsf{M}_n$。我们假设异常序列的数量 T_n 已知。因此，任何离群检验都可以写成函数$\delta(\mathbf{Y}_1,\ldots\ldots,\mathbf{Y}^{\mathsf{M}_n})$：$\mathcal{Y}^{n\mathsf{M}_n}\to\bar{\mathcal{S}}$，其中 $\bar{\mathcal{S}}$表示 $[\mathsf{M}_n]$的所有大小为 T_n 的子集所构成的集合，也就是幂集 $2^{[\mathsf{M}_n]}$ 中所有基数为 T_n 的集合。于是，最大错误概率可写为

$$\epsilon(\delta)=\sum_{\mathbf{y}_1^{\mathsf{M}_n}:\,\delta(\mathbf{y}_1^{\mathsf{M}_n})\neq[\mathsf{T}_n]}P_{[\mathsf{T}_n]}(\mathbf{y}_1^{\mathsf{M}_n}) \quad (11)$$

相应的错误指数如式（2）所示。

对于两步中位数检验，我们进一步引入典型错误指数。具体而言，假设我们使用每个向量 $\mathbf{Y}_1,\ldots\ldots,\mathbf{Y}^{\mathsf{M}_n}$ 的前 $\lceil\rho n\rceil$, $\rho\in(0,1)$ 个样本，记为 $\mathbf{Y}_1^{(1)},\ldots\ldots,\mathbf{Y}^{(1)\,\mathsf{M}_n}$，来生成 π 的基于中位数的估计 $\bar{\pi}$（参见式（8））， 并使用剩余样本 $\mathbf{Y}_1^{(2)},\ldots\ldots,\mathbf{Y}^{(2)\,\mathsf{M}_n}$ 来判定哪些序列是异常值。因此，在第二步中，离群检验由以下函数给出

$$\delta_{\text{median}}(\mathbf{Y}_1^{(2)},\ldots,\mathbf{Y}_{\mathsf{M}_n}^{(2)}|\bar{\pi})=\operatorname*{argmax}_{\mathcal{S}\in\bar{\mathcal{S}}}\sum_{j\in\mathcal{S}}D\left(\mathsf{P}_{\mathbf{Y}_j^{(2)}}\middle\|\bar{\pi}\right). \quad (12)$$

第二步中的最大错误概率可写为

$$\epsilon(\delta_{\text{median}}|\bar{\pi})=\sum_{\mathbf{y}_1^{\mathsf{M}_n}:\,\delta_{\text{median}}(\mathbf{y}_1^{\mathsf{M}_n}|\bar{\pi})\neq[\mathsf{T}_n]}P^{(2)}_{[\mathsf{T}_n]}(\mathbf{y}_1^{\mathsf{M}_n}) \quad (13)$$

其中，$P^{(2)}_{[\mathsf{T}_n]}$ 与式（10）中的定义相同，但索引 k 从 $\lceil\rho n\rceil+1$ 变化到 n。典型错误指数定义为

$$\bar{\alpha}(\delta_{\text{median}})\triangleq\overline{\lim_{n\to\infty}}-\frac{1}{n}\mathsf{E}\left[\log\left(\epsilon(\delta_{\text{median}}|\bar{\pi})\right)\right] \quad (14)$$

其中，期望是对第一步中用于生成 $\bar{\pi}$ 的所有观测序列 $\mathbf{Y}_1^{(1)},\ldots,\mathbf{Y}^{(1)\,\mathsf{M}_n}$ 求取的。

詹森不等式意味着，典型错误指数$\bar{\alpha}(\delta_{\text{median}})$始终不小于错误指数 $\alpha(\delta_{\text{median}})$。直观而言，要获得最优典型错误指数，只需估计量 $\bar{\pi}$ 以趋近于一的概率准确无误。相比之下，错误指数要求 $\bar{\pi}$ 偏离 π 的概率随 n呈指数衰减，且错误指数足够大。我们在第五节中的数值结果表明，这一要求可能过于悲观。

III. 基于均值的检验

我们首先研究式 (7) 中基于均值的检验。下面的定理刻画了当异常值序列的数量相对于 M_n 次线性增长时，该检验的性能。

定理 1：假设序列数 M_n 满足 $\lim_{n\to\infty}\mathsf{M}_n/n=\infty$ 且 $\lim_{n\to\infty}\log(\mathsf{M}_n)/n=0$。进一步假设已知异常值序列的数量，且满足 $\mathsf{T}_n=o(\mathsf{M}_n)$。那么，对于任意一对分布 $\mu\neq\pi$，基于均值的检验 (7) 的错误指数为

$$\alpha_{\text{mean}}=2B(\pi,\mu). \quad (15)$$

证明：参见附录A。 ■

定理 1 依赖于假设：T_n 相对于 M_n是次线性的，因此随着 $n\to\infty$，π 的估计量 $\hat{}\pi$ 会变得准确。当 T_n 与 M_n成正比时，情况便不再如此。特别是，如果 $\mathsf{T}_n=c\mathsf{M}_n$，$c>0$，那么 $\hat{}\pi$ 会收敛到分布 $\nu=(1-c)\pi+c\mu$。沿用定理 1 证明中的相同步骤，可以证明错误指数的下界为

$$\alpha_{\text{mean}}=\min_{Q_1,Q_2}\{D(Q_1\|\mu)+D(Q_2\|\pi)\} \quad (16a)$$

其中，最小值是在满足以下条件的所有分布 $(Q_1,\ Q_2)$ 上取得:

$$D(Q_2|\nu)-D(Q_1\|\nu)\geq 0 \quad (16b)$$

这严格小于 $2B(\pi,\ \mu)$，参见图 2。因此，当异常值序列的数量与序列总数 M_n成正比时，在 π 和 μ已知的情况下，基于均值的检验无法达到最大似然检验的指数。

IV. 基于中位数的检验

由于当 $\mathsf{T}_n=c\mathsf{M}_n$ 时，基于均值的检验表现不佳，因此本节考虑基于中位数的估计量 $\bar{\pi}$（参见式 (8)）。只要 $c<\frac{1}{2}$，它本质上就能抵御异常值。然而，与基于均值的估计量 $\hat{}\pi$ 不同，基于中位数的估计量 (8) 收敛到 π 的速度取决于 n 的增长，而非 M_n。因此，概率 $\bar{\pi}\notin\mathcal{B}_{\varepsilon,\,\infty}(\pi)$，其中

$$\mathcal{B}_{\varepsilon,\infty}(\pi)\triangleq\{Q\in\mathcal{P}(\mathcal{Y})\colon Q(y)\in[\pi(y)-\varepsilon,\pi(y)+\varepsilon],y\in\mathcal{Y}\} \quad (17)$$

不会随 n 超指数衰减，即使 M_n 关于 n 超线性增长。事实上，此概率可界定如下:

引理 1：考虑式 (8) 中给出的 π 的基于中位数的估计量 $\bar{\pi}$，并假设 $\mathsf{T}_n<\mathsf{M}_n/2$。那么，对于每个$\varepsilon>0$,

$$P_{[\mathsf{T}_n]}\left(\bar{\pi}\notin\mathcal{B}_{\varepsilon,\infty}(\pi)\right)\leq 2|\mathcal{Y}|(\mathsf{M}_n-\mathsf{T}_n)\exp(-2n\varepsilon'^2) \quad (18)$$

其中 $\varepsilon'=\varepsilon/(1+|\mathcal{Y}|)$。

证明：参见附录 B。 ■

注意到，(18) 的右端项随 n 指数衰减，但当 ε' 趋于零时，错误指数也趋于零。因此，如果我们对基于中位数的估计量重复定理 1 证明中的步骤，就会得到一个趋于零的错误指数。这与第五节中的数值结果相矛盾；数值结果表明，对于有限的 n，基于中位数的检验优于基于均值的检验；参见图 3。造成这种不一致的一种可能解释是，错误指数过于

悲观，因为它受 $\bar{\pi}$ 收敛到 π 的缓慢速度所主导。因此，我们转而研究式 (14) 中定义的典型错误指数。

为简化分析，我们将刻画第二节末尾所述两步中位数检验的典型错误指数。具体而言，我们把每个序列$\mathbf{Y}_i$ 分成两个子序列$\mathbf{Y}_i^{(1)}$ 和$\mathbf{Y}_i^{(2)}$ ，它们分别由$\mathbf{Y}_i$的前$\lceil \rho n \rceil$ 个样本和剩余的 $n - \lceil \rho n \rceil$ 个样本构成。随后，前一个子序列用于估计 π，而后一个子序列用于检测异常值。由于这些观测序列是独立同分布的，估计量 $\bar{\pi}$将与子序列$\mathbf{Y}_1^{(2)}, \ldots, \mathbf{Y}_{\mathrm{M}_n}^{(2)}$相互独立，因此更易于分析。以下定理刻画了该检验的典型错误指数。

定理 *2*： 考虑第二节末尾所述的两步中位数检验，并假设 $\lim_{n\to\infty} \log(\mathrm{M}_n)/n = 0$。进一步假设异常序列的数量已知且满足 $\mathrm{T}_n < \mathrm{M}_n/2$。那么，对于每一对分布 $\mu \neq \pi$，检验(12)的典型错误指数为

$$\bar{\alpha}_{\text{median}} = 2B(\pi, \mu). \tag{19}$$

证明： 参见附录 C。 ■

定理 2 只要求 M_n 在 n中呈次指数增长。因此，与定理 1 不同，该定理也适用于 M_n 不随 n增长的情形。由此，我们可以将定理 2 得到的典型错误指数与 [1]中得到的结果进行比较。有趣的是，当 π 和 μ 已知时，典型错误指数等于最大似然检验的指数；而错误指数则严格小于 [1, 定理 9]。

V. 数值结果

为验证我们的理论发现，我们进行了三项实验：

1) 数值展示广义似然比检验 (4) 的错误指数确实往往小于两步中位数检验（定理 2）的典型错误指数。2) 对 (16) 进行数值评估，以确认当 $\mathrm{T}_n = c\mathrm{M}_n$时，基于均值的检验并非最优。3) 比较基于均值的检验、单步中位数检验、[1] 以及两步中位数检验，其中 $\mathrm{T}_n = c\mathrm{M}_n$ 时 c 逐渐增大，且 M_n逐渐增大。

A. 广义似然比检验的次优性

我们选取两个伯努利分布 $\mu = B(0.7)$ 和$\pi = B(\theta_\pi)$，并在集合 $\{0.50, 0.55, \ldots, 1.00\}$ 中改变概率 $\pi(0) = 1 - \theta_\pi$。我们选取 $\mathrm{M}_n = 4$，并通过数值计算近似广义似然比检验 [1, Th. 2] 及其下界 [1, Th. 3] 的错误指数。我们通过计算目标函数，近似求解相应的优化问题以及网格上的约束。更具体地说，对于每个 $\pi(0)$，我们评估 [1, 中从 q_1 到 q_4 的伯努利分布

[1]单步中位数检验使用相同的观测序列来生成基于中位数的估计量 $\bar{\pi}$ of π，并检测异常值序列。

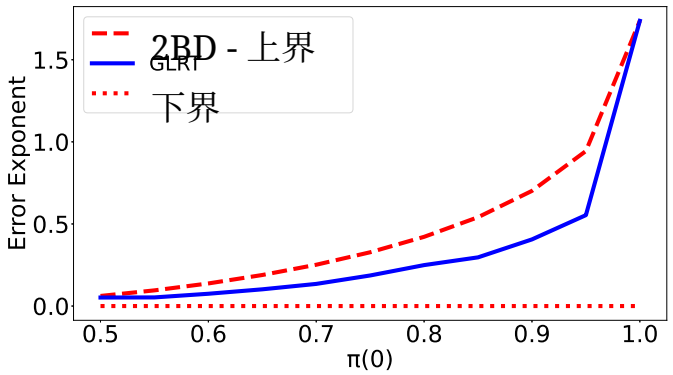

图 1. 数值演示表明，广义似然比检验的错误指数通常小于 $2B(\pi, \mu)$。

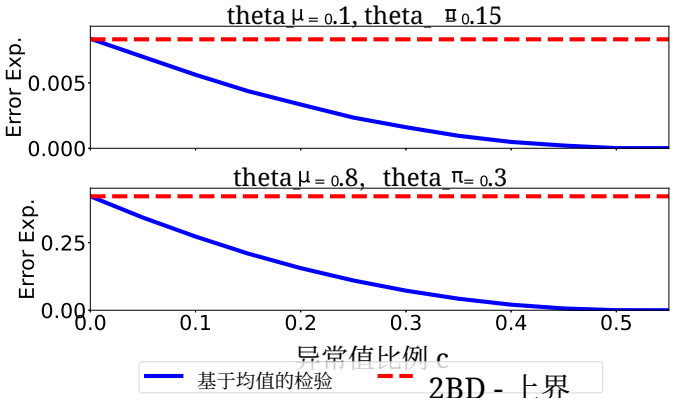

图 2. 随着异常值比例增加时，基于均值的检验的错误指数 α_{mean} c在两种设置下。

式（20）–（21）]给出了从集合 $\{0.00, 0.05, \ldots, 1\}$ 中抽取的成功概率 θ_{q_i}, $i = 1, \ldots, 4$。我们采用类似的方法来评估定理 [1, Th. 3] 中的下界。这次模拟的结果如图 1 所示。我们观察到，广义似然比检验的错误指数仅在 $\pi(0) = 0.5$ 和 $\pi(0) = 1$ 时达到 $2B(\pi, \mu)$。根据定理 [1, Th. 3],，当 M_n 较大时，广义似然比检验的性能会更好；但请注意，对于较小的 M_n，该定理给出的下界将变得无效。事实上，当 $\mathrm{M}_n = 4$ 时，我们对下界的近似值为零；参见 [1, 图 1]。

B. 当$\mathrm{T}_n = c\mathrm{M}_n$时的基于均值的检验

当异常值比例 c 为正时，观测集合 $\mathbf{Y}_1^{\mathrm{M}_n}$ 是从混合分布 $\nu = c\mu + (1-c)\pi$中抽取的。我们针对伯努利分布（即 $\mathcal{Y} = \{0, 1\}$）以及集合$\{0.00, 0.05, \ldots, 0.55\}$中的异常值比例 c，评估基于均值的检验 (7) 所对应的错误指数。我们分别从数对 (0.10,0.15) 和 (0.8, 0.3) 中选取离群分布和典型分布的成功概率 θ_μ 和 θ_π。因此，我们评估这两个分布分别相似和非相似的情形。假设集合$\{0.001, 0.006, 0.011, \ldots, 0.996\}$中的所有成功概率均满足相应条件，我们对伯努利分布 Q_1 和 Q_2 数值计算了 (16)。图 2 展示了

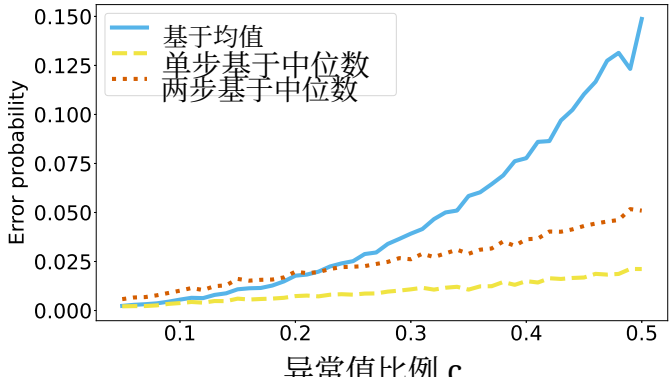

图 3. 对比基于均值、单步基于中位数和两步基于中位数的检验在不断增加的异常值比例 c 下的表现。

两组 θ_μ, θ_π 的结果。正如预期，非相似分布 π 和 μ 的错误指数更大；仅当 $c = 0$ 时，基于均值的检验的错误指数才与 $2B(\pi, \mu)$ 的上界重合。（注意，在这种情况下，我们假设 M_n 随 n 超线性增长。）我们还观察到，随着 c 增大，α_{mean} 逐渐减小；当 $c \geq 0.5$ 时，其值为零，因为此时均值 $\hat{}\pi$ 对离群分布 μ 的估计优于对典型分布 π的估计。

C. 不同检验的非渐近性能

最后，我们在非渐近情形下评估基于均值的检验、单步中位数检验和两步中位数检验的性能，即在 M_n 和 n 有限时的性能。具体而言，我们设定 $\mathrm{M}_n = 500$，并令异常值比例 c在集合 $\{0.05, 0.06, \ldots, 0.5\}$中变化。对于每个 c，我们选取 μ和 π作为分类分布，其中 $|\mathcal{Y}| = 5$，各个符号的概率从均匀分布中抽取，随后归一化为 1。接下来，我们分别从 μ 和π中抽取 $\mathrm{T}_n = \lfloor c\mathrm{M}_n \rfloor$和 $\mathrm{M}_n - \mathrm{T}_n$ 个长度为 $n = 250$ 的序列。基于这些序列，我们计算基于均值的检验、单步中位数检验和两步中位数检验的检验统计量，并对后者设定 $\rho = 0.5$。图 3 展示了每种检验的平均错误概率，该概率取自实验的 200 次独立运行的平均值。我们观察到，基于均值的检验表现最差，除非 c 较小；这可能是因为当 T_n 与 M_n 成正比时，对 π 的估计量存在偏差（参见图 2）。此外，单步中位数检验通常优于我们在定理 2 中分析的两步检验。我们认为，这是因为两步检验的样本效率较低。然而，对于单步中位数检验，我们尚未能够得到典型错误指数的表达式，因为在这种情况下，$\bar{\pi}$ 以及用于检测异常值的观测序列 $\mathbf{Y}_1, \ldots, \mathbf{Y}_{\mathrm{M}_n}$ 彼此相关。

VI. 计算复杂度

基于均值和两步中位数的检验将相应检验统计量 $D\left(\mathrm{P}_{\mathbf{Y}_j} \| \hat{}\pi\right)$ 和 $D\left(\mathrm{P}_{\mathbf{Y}_i^{(2)}} \| \bar{\pi}\right)$ 最大的那些 T_n 序列判定为异常值。因此，这些检验需要计算每个序列的类型（计算复杂度为 $\mathcal{O}(n|\mathcal{Y}|)$），对所有类型取平均值或中位数（计算复杂度分别为$\mathcal{O}(\mathrm{M}_n|\mathcal{Y}|)$或 $\mathcal{O}(|\mathcal{Y}|\mathrm{M}_n \log \mathrm{M}_n)$），计算 M_n 个序列中每个序列的检验统计量（计算复杂度：$\mathcal{O}(\mathrm{M}_n|\mathcal{Y}|)$），并对 M_n 个检验统计量进行排序（计算复杂度：$\mathcal{O}(\mathrm{M}_n \log \mathrm{M}_n)$），以确定其中最大的 T_n 个。固定 $|\mathcal{Y}|$ 后，$2B(\pi, \mu)$ 的最优错误指数因此可以通过 $\mathcal{O}(n) + \mathcal{O}(\mathrm{M}_n \log \mathrm{M}_n)$的复杂度实现。相比之下，[1]中提出的广义似然比检验（GLRT）仅在 $\mathrm{M}_n \to \infty$时渐近地达到该错误指数，但其计算复杂度为 $\mathcal{O}(\mathrm{M}_n^{\mathrm{T}_n})$。将该检验表述为组合聚类问题 [13,第 4 节2], , 并通过 K-均值 [14], [15], 以次优方式求解，可将计算复杂度降至 $\mathcal{O}(\mathrm{M}_n)$，但其错误指数小于 $2B(\pi, \mu)$ [15, 定理 1 和 3]。

附录A 定理 1 的证明

由于通用情形下的错误指数不可能大于已知 π 和 μ 时的错误指数 $2B(\pi, \mu)$，因此只需给出错误指数的下界。

我们首先证明，根据 $\mathrm{T}_n = o(\mathrm{M}_n)$ 这一假设，估计量 $\hat{}\pi$ 与 π 的差异超过一个 $\varepsilon > 0$ 的概率会随 M_n 指数级衰减。又由于假设 $\lim_{n\to\infty} \mathrm{M}_n/n = \infty$，这意味着该概率会随 n 超指数级衰减。具体而言，定义

$$\mathcal{B}_{\varepsilon,\infty}(\pi) \triangleq \{Q \in \mathcal{P}(\mathcal{Y}): \ Q(y) \in [\pi(y) - \varepsilon, \pi(y) + \varepsilon], y \in \mathcal{Y}\} \tag{20}$$

我们有

$$\begin{aligned} & P_{[\mathsf{T}_n]}(\hat{\pi} \notin \mathcal{B}_{\varepsilon,\infty}(\pi)) \\ & \leq \sum_{y\in\mathcal{Y}} P_{[\mathsf{T}_n]}\left(\left|\pi(y) - \frac{1}{\mathsf{M}_n}\sum_{\ell=1}^{\mathsf{M}_n} \mathsf{P}_{\mathbf{Y}_\ell}(y)\right| > \varepsilon\right) \\ & = \sum_{y\in\mathcal{Y}} P_{[\mathsf{T}_n]}\left(\left|\sum_{\ell=1}^{\mathsf{M}_n} (\mathsf{P}_{\mathbf{Y}_\ell}(y) - \pi(y))\right| > \varepsilon \mathsf{M}_n\right) \end{aligned} \tag{21}$$

其中，不等式由并集界以及 $\mathcal{B}_{\varepsilon,\infty}(\pi)$ 和 $\hat{}\pi$ 的定义得出。接下来注意到，在 $P_{[\mathsf{T}_n]}$ 下，

$$\mathsf{E}\left[\mathsf{P}_{\mathbf{Y}_\ell}(y)\right] = \begin{cases} \mu(y), & \ell = 1, \ldots, \mathsf{T}_n \\ \pi(y), & \ell = \mathsf{T}_n + 1, \ldots, \mathsf{M}_n \end{cases} \tag{22}$$

对于 $y \in \mathcal{Y}$。因此，我们可以写成

$$\begin{aligned} & \left|\sum_{\ell=1}^{\mathsf{M}_n} (\mathsf{P}_{\mathbf{Y}_\ell}(y) - \pi(y))\right| \\ & = \left|\sum_{\ell=1}^{\mathsf{M}_n} (\mathsf{P}_{\mathbf{Y}_\ell}(y) - \mathsf{E}\left[\mathsf{P}_{\mathbf{Y}_\ell}(y)\right]) + \mathsf{T}_n\mu(y) - \mathsf{T}_n\pi(y)\right| \\ & \leq \left|\sum_{\ell=1}^{\mathsf{M}_n} (\mathsf{P}_{\mathbf{Y}_\ell}(y) - \mathsf{E}\left[\mathsf{P}_{\mathbf{Y}_\ell}(y)\right])\right| + \mathsf{T}_n|\mu(y) - \pi(y)| \\ & \leq \left|\sum_{\ell=1}^{\mathsf{M}_n} (\mathsf{P}_{\mathbf{Y}_\ell}(y) - \mathsf{E}\left[\mathsf{P}_{\mathbf{Y}_\ell}(y)\right])\right| + \mathsf{T}_n \end{aligned} \tag{23}$$

其中，第一个不等式由三角不等式得出，第二个不等式则由以下事实得出：$\pi(y)$ $\mu(y) \in [0,1]$。因此，根据霍夫丁不等式 [16]，我们得到

$$\begin{aligned}
&P_{[\mathsf{T}_n]}(\hat{\pi} \notin \mathcal{B}_{\varepsilon,\infty}(\pi)) \\
&\quad \le \sum_{y\in\mathcal{Y}} P_{[\mathsf{T}_n]}\left(\left|\sum_{\ell=1}^{\mathsf{M}_n}(\mathsf{P}_{\mathbf{Y}_\ell}(y) - \mathsf{E}\left[\mathsf{P}_{\mathbf{Y}_\ell}(y)\right])\right| > \varepsilon\mathsf{M}_n - \mathsf{T}_n\right) \\
&\quad \le 2|\mathcal{Y}|\exp\left(-2\frac{(\varepsilon\mathsf{M}_n - \mathsf{T}_n)^2}{\mathsf{M}_n}\right). \qquad (24)
\end{aligned}$$

在 $\mathsf{T}_n = o(\mathsf{M}_n)$ 的假设下，(24) 右端项中的指数在 $\mathsf{M}_n \to \infty$ 时渐近地等于 $\varepsilon^2\mathsf{M}_n$。因此，$P_{[\mathsf{T}_n]}(\hat{\pi} \notin \mathcal{B}_{\varepsilon,\infty}(\pi))$ 随 M_n 指数衰减，并随 n超指数衰减。

接下来，我们对错误概率$\epsilon(\delta_{\text{mean}})$给出上界。事实上，我们有

$$\begin{aligned}
\epsilon(\delta_{\text{mean}}) &= \text{Prob}\left(\delta_{\text{mean}}(\mathbf{Y}_1^{\mathsf{M}_n}) \neq [\mathsf{T}_n], \hat{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)\right) \\
&\quad + \text{Prob}\left(\delta_{\text{mean}}(\mathbf{Y}_1^{\mathsf{M}_n}) \neq [\mathsf{T}_n], \hat{\pi} \notin \mathcal{B}_{\varepsilon,\infty}(\pi)\right) \\
&\le \text{Prob}\left(\delta_{\text{mean}}(\mathbf{Y}_1^{\mathsf{M}_n}) \neq [\mathsf{T}_n], \hat{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)\right) \\
&\quad + 2|\mathcal{Y}|\exp\left(-2\frac{(\varepsilon\mathsf{M}_n - \mathsf{T}_n)^2}{\mathsf{M}_n}\right) \qquad (25)
\end{aligned}$$

其中不等式由（$\hat{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)$）和 Hoeffding 不等式（$\pi \notin \mathcal{B}_{\varepsilon,\infty}(\pi)$）得出。由于 $\text{Prob}(\hat{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi))$ 随 n 超指数衰减，$\epsilon(\delta_{\text{mean}})$ 的错误指数由第一项的错误指数决定，下面我们将对该项进行界定。

事实上，对于离散分布，相对熵是连续的。令

$$U_i(\hat{\pi}) \triangleq D\left(\mathsf{P}_{\mathbf{Y}_i}\middle\|\hat{\pi}\right) \qquad (26)$$

由此，对于 $\hat{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)$，$U_i(\hat{\pi})$ 位于 $[U_i(\pi) - \varepsilon', U_i(\pi) + \varepsilon']$ 中，其中 ε' 随 $\varepsilon \to 0$ 趋于零。因此，(25) 右端项的第一项可上界为

$$\begin{aligned}
&\text{Prob}\left(\delta_{\text{mean}}(\mathbf{Y}_1^{\mathsf{M}_n}) \neq [\mathsf{T}_n], \hat{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)\right) \\
&\quad \le P_{[\mathsf{T}_n]}\left(\min_{i\le\mathsf{T}_n} U_i(\pi) \le \max_{j>\mathsf{T}_n} U_j(\pi) + 2\varepsilon'\right) \\
&\quad = P_{[\mathsf{T}_n]}\left(\bigcup_{i\le\mathsf{T}_n}\bigcup_{j>\mathsf{T}_n}\{U_i(\pi) < U_j(\pi) + 2\varepsilon'\}\right) \\
&\quad \le \sum_{i=1}^{\mathsf{T}_n}\sum_{j=\mathsf{T}_n+1}^{\mathsf{M}_n} P_{[\mathsf{T}_n]}\left(U_i(\pi) < U_j(\pi) + 2\varepsilon'\right) \\
&\quad = \mathsf{T}_n(\mathsf{M}_n - \mathsf{T}_n)P_{[\mathsf{T}_n]}\left(U_1(\pi) \le U_{\mathsf{T}_n+1}(\pi) + 2\varepsilon'\right) \qquad (27)
\end{aligned}$$

其中，第一个不等式成立是因为，对于 $\hat{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)$，我们可以用 $U_i(\pi) \pm \varepsilon'$ 近似 $U_i(\hat{\pi})$，然后去掉条件 $\hat{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)$；第二个不等式由并集界得出；最后一个等式则源于问题的对称性。

由于 $\mathsf{T}_n < \mathsf{M}_n/2$，并且根据定理的假设，$\mathsf{M}_n$ 在 n中是次指数的，因此 (27) 右端项的误差指数由以下表达式的指数决定：

$$P_{[\mathsf{T}_n]}(U_1(\pi) \le U_{\mathsf{T}_n+1}(\pi) + 2\varepsilon')$$

。因此，对于已知 π 的情况，我们按照 [1]中的步骤来研究该指数。事实上，我们可以应用[1, 中的 Sanov 定理版本，引理 1]，得到

$$\begin{aligned}
&-\lim_{n\to\infty}\frac{1}{n}\log P_{[\mathsf{T}_n]}(U_{\mathsf{T}_n+1}(\pi) - U_1(\pi) + 2\varepsilon' > 0) \\
&\quad = \min_{Q_1,Q_2}\{D(Q_1\|\mu) + D(Q_2\|\pi)\} \qquad (28)
\end{aligned}$$

其中，最小值是在所有满足 $D(Q_2\|\pi) - D(Q_1\|\pi) + 2\varepsilon' \ge 0$ 的分布 (Q_1, Q_2) 上取到的。利用该条件将 $D(Q_2\|\pi)$ 下界为 $D(Q_1\|\pi) - 2\varepsilon'$，并注意到 ε' 不依赖于 Q_1，我们可以将式 (28) 中的最小值下界为

$$\begin{aligned}
&\min_{Q_1,Q_2}\{D(Q_1\|\mu) + D(Q_2\|\pi)\} \\
&\quad \ge \min_{Q_1}\{D(Q_1\|\mu) + D(Q_1\|\pi)\} - 2\varepsilon' \\
&\quad = 2B(\mu,\pi) - 2\varepsilon' \qquad (29)
\end{aligned}$$

其中，第二行中的最小值是在所有概率分布 Q_1 上取到的，且根据引理 2]，等于 $2B(\mu,\pi)$ [1, 。结合式 (25)–(29)，我们得到

$$\alpha_{\text{mean}} \triangleq \varliminf_{n\to\infty} -\frac{1}{n}\log\epsilon(\delta_{\text{mean}}) \ge 2B(\mu,\pi) \qquad (30)$$

令 ε' 从右侧趋近于零。由于如证明开头所述，α_{mean} 不能超过 $2B(\mu, \pi)$，因此定理 1 得证。

APPENDIX B
PROOF OF LEMMA 1

Let

$$m(y) \triangleq \text{median}\{\mathsf{P}_{\mathbf{Y}_1}(y), \ldots, \mathsf{P}_{\mathbf{Y}_{\mathsf{M}_n}}(y)\} \qquad (31)$$

从而

$$\overline{\pi}(y) = \frac{m(y)}{\sum_{y'\in\mathcal{Y}} m(y')}. \qquad (32)$$

可以证明，若 $m(y) \in [\pi(y) - \varepsilon', \pi(y) + \varepsilon']$，则我们也有 $\overline{\pi}(y) \in [\pi(y) - \varepsilon, \pi(y) + \varepsilon]$，其中 $\varepsilon = \varepsilon'(1 + |\mathcal{Y}|)$。事实上，$m(y) \in [\pi(y) - \varepsilon', \pi(y) + \varepsilon']$ 意味着

$$1 - |\mathcal{Y}|\varepsilon' \le \sum_{y\in\mathcal{Y}} m(y) \le 1 + |\mathcal{Y}|\varepsilon' \qquad (33)$$

因此

$$\frac{\pi(y) - \varepsilon'}{1 + |\mathcal{Y}|\varepsilon'} \le \overline{\pi}(y) \le \frac{\pi(y) + \varepsilon'}{1 - |\mathcal{Y}|\varepsilon'}. \qquad (34)$$

重写下界后，可将其放宽为

$$\begin{aligned}
\overline{\pi}(y) &\ge \frac{\pi(y) - \varepsilon'}{1 + |\mathcal{Y}|\varepsilon'} \\
&= \pi(y) - \varepsilon'\frac{1 + |\mathcal{Y}|\pi(y)}{1 + |\mathcal{Y}|\varepsilon'} \\
&\ge \pi(y) - \varepsilon'\frac{1 + |\mathcal{Y}|}{1 + |\mathcal{Y}|\varepsilon'} \\
&\ge \pi(y) - \varepsilon'(1 + |\mathcal{Y}|) \qquad (35)
\end{aligned}$$

其中，第三步是因为 $\pi(y) \le 1$，第四步则是由 $|\mathcal{Y}|\varepsilon'$ 的非负性得出。类似地，我们可以将式（34）中的上界放宽为

$$\overline{\pi}(y) \le \frac{\pi(y)+\varepsilon'}{1-|\mathcal{Y}|\varepsilon'} \le \pi(y) + \varepsilon'(1+|\mathcal{Y}|). \tag{36}$$

因此，我们可以给出如下上界

$$\begin{aligned}
&P_{[\mathsf{T}_n]}(\bar{\pi} \notin \mathcal{B}_{\varepsilon,\infty}(\pi)) \\
&\quad = P_{[\mathsf{T}_n]}\left(\bigcup_{y\in\mathcal{Y}} \{\overline{\pi}(y) \notin [\pi(y)-\varepsilon, \pi(y)+\varepsilon]\}\right) \\
&\quad \le \sum_{y\in\mathcal{Y}} P_{[\mathsf{T}_n]}\left(\overline{\pi}(y) \notin [\pi(y)-\varepsilon, \pi(y)+\varepsilon]\right) \\
&\quad \le |\mathcal{Y}| \max_{y\in\mathcal{Y}} P_{[\mathsf{T}_n]}\big(m(y) \notin [\pi(y)-\varepsilon', \pi(y)+\varepsilon']\big)
\end{aligned} \tag{37}$$

其中，第一个不等式来自并集界，第二个不等式则基于这样一个事实：对于每个 $y \in \mathcal{Y}$，事件 $\{\pi(y) \notin [\pi(y)-\varepsilon, \pi(y)+\varepsilon]\}$ 包含于事件$\{m(y) \notin [\pi(y)-\varepsilon', \pi(y)+\varepsilon']\}$。

接下来，我们证明数集的中位数具有以下性质：

引理 *2*：设 $\mathcal{K} = \{k_1, \ldots, k_{\mathsf{L}}\}$ 为一个包含 L 个数的集合，并设 $\mathcal{K}_{\mathcal{S}} = \{k_i\}_{i\in\mathcal{S}}$ 表示 $\mathcal{K}$ 中由 $\mathcal{S} \subseteq$ [L]索引的子集。对于任意满足 $|\mathcal{S}| >$L/2的 $\mathcal{S}$，有

$$\min_{i\in\mathcal{S}} k_i \le \text{median}\{k_1, \ldots, k_{\mathsf{L}}\} \le \max_{i\in\mathcal{S}} k_i. \tag{38}$$

证明：证明方法是将$\mathcal{K}$ 中的元素排序，并注意到中位数是 $\mathcal{K}$ 中排序索引为 $(\mathrm{L}+1)/2$ 的元素（当 L 为奇数时），或排序索引为 $\mathrm{L}/2$ 和 $\mathrm{L}/2+1$ 的元素的平均值（当 L 为偶数时）。只有当集合 $\mathcal{S}$ 所含元素的排序索引全都小于或全都大于中位数的排序索引时，(38) 中的不等式才可能不成立。然而，由于集合 $\mathcal{S}$ 的基数严格大于 $\mathrm{L}/2$，这两种情况都不可能发生。证明完毕。 ■

由于异常序列的数量满足 $\mathsf{T}_n < \mathsf{M}_n/2$，我们可以令 $\mathcal{S}$ 为从典型分布中抽取的序列的类型，即 $\mathbf{P}_{\mathbf{Y}_i}$(*y*),$i = \mathsf{T}_n + 1$，…，$\mathsf{M}_n$，从而应用引理 2。因此，我们可以将中位数 $m(y)$ 界定为

$$\min_{i>\mathsf{T}_n} \mathsf{P}_{\mathbf{Y}_i} \le m(y) \le \max_{i>\mathsf{T}_n} \mathsf{P}_{\mathbf{Y}_i}. \tag{39}$$

由并集界可得

$$\begin{aligned}
&P_{[\mathsf{T}_n]}\big(m(y) \notin [\pi-\varepsilon', \pi+\varepsilon']\big) \\
&\quad \le P_{[\mathsf{T}_n]}\left(\min_{i>\mathsf{T}_n} \mathsf{P}_{\mathbf{Y}_i}(y) < \pi(y)-\varepsilon'\right) \\
&\qquad + P_{[\mathsf{T}_n]}\left(\max_{i>\mathsf{T}_n} \mathsf{P}_{\mathbf{Y}_i}(y) > \pi(y)+\varepsilon'\right) \\
&\quad = P_{[\mathsf{T}_n]}\left(\bigcup_{i>\mathsf{T}_n} \{\mathsf{P}_{\mathbf{Y}_i}(y) < \pi(y)-\varepsilon'\}\right) \\
&\qquad + P_{[\mathsf{T}_n]}\left(\bigcup_{i>\mathsf{T}_n} \{\mathsf{P}_{\mathbf{Y}_i}(y) > \pi(y)+\varepsilon'\}\right) \\
&\quad \le (\mathsf{M}_n - \mathsf{T}_n) P_{[\mathsf{T}_n]}(\mathsf{P}_{\mathbf{Y}_{\mathsf{T}_n+1}}(y) < \pi(y)-\varepsilon') \\
&\qquad + (\mathsf{M}_n - \mathsf{T}_n) P_{[\mathsf{T}_n]}(\mathsf{P}_{\mathbf{Y}_{\mathsf{T}_n+1}}(y) > \pi(y)+\varepsilon').
\end{aligned} \tag{40}$$

由于 $n\mathbf{P}_{\mathbf{Y}_{\mathsf{T}_n+1}}$(*y*) 是一个二项随机变量，具有 n次尝试和成功概率 π(*y*)，我们可以使用 Hoeffding 不等式证明

$$P_{[\mathsf{T}_n]}(\mathsf{P}_{\mathbf{Y}_{\mathsf{T}_n+1}}(y) < \pi(y)-\varepsilon') \le \exp\left(-2n\varepsilon'^2\right) \tag{41a}$$
$$P_{[\mathsf{T}_n]}(\mathsf{P}_{\mathbf{Y}_{\mathsf{T}_n+1}}(y) > \pi(y)+\varepsilon') \le \exp\left(-2n\varepsilon'^2\right). \tag{41b}$$

因此，

$$\begin{aligned}
&P_{[\mathsf{T}_n]}\big(m(y) \notin [\pi-\varepsilon', \pi+\varepsilon']\big) \\
&\quad \le 2(\mathsf{M}_n - \mathsf{T}_n)\exp\left(-2n\varepsilon'^2\right)
\end{aligned} \tag{42}$$

结合式 (37)，即可证明该引理。 ■

附录 C 定理 2 的证明

由于通用情形下的典型错误指数不可能大于已知 π和 μ 时的错误指数，因此只需对典型错误指数给出下界。

根据全期望定律，我们将式 (14) 右端项上的期望写为

$$\begin{aligned}
&\mathsf{E}\left[\log\big(\epsilon(\delta_{\text{median}}|\bar{\pi})\big)\right] \\
&= \mathsf{E}[\log\big(\epsilon(\delta_{\text{median}}|\bar{\pi})\big)\big|\bar{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)] P_{[\mathsf{T}_n]}\big(\bar{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)\big) \\
&\quad + \mathsf{E}[\log\big(\epsilon(\delta_{\text{median}}|\bar{\pi})\big)\big|\bar{\pi} \notin \mathcal{B}_{\varepsilon,\infty}(\pi)] P_{[\mathsf{T}_n]}\big(\bar{\pi} \notin \mathcal{B}_{\varepsilon,\infty}(\pi)\big) \\
&\le \mathsf{E}[\log\big(\epsilon(\delta_{\text{median}}|\bar{\pi})\big)\big|\bar{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)] P_{[\mathsf{T}_n]}\big(\bar{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)\big)
\end{aligned} \tag{43}$$

其中，该不等式成立是因为错误概率$\epsilon(\delta_{\text{median}}|\bar{\pi})$不可能超过 1。

接下来回顾

$$\delta_{\text{median}}(\mathbf{Y}_1^{(2)}, \ldots, \mathbf{Y}_{\mathsf{M}}^{(2)}|\bar{\pi}) = \operatorname*{argmax}_{\mathcal{S}\in\bar{\mathcal{S}}} \sum_{j\in\mathcal{S}} U_j(\bar{\pi}) \tag{44}$$

其中

$$U_i(\bar{\pi}) \triangleq D\Big(\mathsf{P}_{\mathbf{Y}_i^{(2)}} \,\Big\|\, \bar{\pi}\Big). \tag{45}$$

如果任意典型序列的检验统计量 $U_j(\bar{\pi})$ 大于任意异常序列的检验统计量，就会发生错误。利用这一点，我们得到

$$\begin{aligned}
&\epsilon(\delta_{\text{median}}|\overline{\pi}) \\
&\quad = P_{[\mathsf{T}_n]}\left(\min_{\ell\le\mathsf{T}_n} U_\ell(\bar{\pi}) < \max_{j>\mathsf{T}_n} U_j(\bar{\pi})\right) \\
&\quad = P_{[\mathsf{T}_n]}\left(\bigcup_{\ell\le\mathsf{T}_n}\bigcup_{j>\mathsf{T}_n} \{U_\ell(\bar{\pi}) < U_j(\bar{\pi})\}\right) \\
&\quad \le \sum_{\ell=1}^{\mathsf{T}_n}\sum_{j=\mathsf{T}_n+1}^{\mathsf{M}_n} P_{[\mathsf{T}_n]}\left(U_\ell(\bar{\pi}) < U_j(\bar{\pi})\right) \\
&\quad = \mathsf{T}_n(\mathsf{M}_n - \mathsf{T}_n) P_{[\mathsf{T}_n]}(U_{\mathsf{T}_n+1}(\bar{\pi}) - U_1(\bar{\pi}) > 0) \\
&\quad \le \mathsf{M}_n^2 P_{[\mathsf{T}_n]}(U_{\mathsf{T}_n+1}(\bar{\pi}) - U_1(\bar{\pi}) > 0)
\end{aligned} \tag{46}$$

其中，第一个不等式由并集界得出，后续等式则源于问题的对称性。由于根据假设 $\bar{\pi} \in \mathcal{B}_{\varepsilon,\ \infty}(\pi)$，我们得到

$$
\begin{aligned}
&U_{\mathsf{T}_n+1}(\bar{\pi}) - U_1(\bar{\pi}) \\
&= D(\mathsf{P}_{\mathbf{Y}^{(2)}_{\mathsf{T}_n+1}} \| \bar{\pi}) - D(\mathsf{P}_{\mathbf{Y}^{(2)}_1} \| \bar{\pi}) \\
&= \sum_{y \in \mathcal{Y}} \mathsf{P}_{\mathbf{Y}^{(2)}_{\mathsf{T}_n+1}}(y) \log \frac{\mathsf{P}_{\mathbf{Y}^{(2)}_{\mathsf{T}_n+1}}(y)}{\bar{\pi}(y)} - \mathsf{P}_{\mathbf{Y}^{(2)}_1}(y) \log \frac{\mathsf{P}_{\mathbf{Y}^{(2)}_1}(y)}{\bar{\pi}(y)} \\
&\le \sum_{y \in \mathcal{Y}} \mathsf{P}_{\mathbf{Y}^{(2)}_{\mathsf{T}_n+1}}(y) \log \frac{\mathsf{P}_{\mathbf{Y}^{(2)}_{\mathsf{T}_n+1}}(y)}{\pi(y) - \varepsilon} - \mathsf{P}_{\mathbf{Y}^{(2)}_1}(y) \log \frac{\mathsf{P}_{\mathbf{Y}^{(2)}_1}(y)}{\pi(y) + \varepsilon} \\
&\le D(\mathsf{P}_{\mathbf{Y}^{(2)}_{\mathsf{T}_n+1}} \| \pi) - D(\mathsf{P}_{\mathbf{Y}^{(2)}_1} \| \pi) + \tilde{\varepsilon} \qquad (47)
\end{aligned}
$$

其中

$$
\tilde{\varepsilon} \triangleq 2 \frac{\varepsilon}{\pi_{\min} - \varepsilon}. \qquad (48)
$$

在 (47) 中，最后一个不等式是将前两项在 π 处展开一阶泰勒级数得到的。由于 (47) 的右端项不依赖于 $\bar{\pi}$，且 $\mathbf{Y}^{(2)}_1, \ldots, \mathbf{Y}^{(2)}_{\mathsf{M}}$ 与估计量 π 相互独立，因此由 (46)、(47) 可得

$$
\begin{aligned}
&\mathsf{E}[\log\big(\epsilon(\delta_{\text{median}}|\bar{\pi})\big) \big| \bar{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)] \\
&\quad \le \log P_{[\mathsf{T}_n]}(D(\mathsf{P}_{\mathbf{Y}^{(2)}_{\mathsf{T}_n+1}} \| \pi) - D(\mathsf{P}_{\mathbf{Y}^{(2)}_1} \| \pi) + \tilde{\varepsilon} > 0) \\
&\qquad + 2 \log \mathsf{M}_n. \qquad (49)
\end{aligned}
$$

利用 $\mathbf{Y}^{(2)}_1, \ldots, \mathbf{Y}^{(2)}_{\mathsf{M}}$ 分别服从典型分布和异常分布且相互独立同分布这一事实，我们可以应用 [1, 引理 1] 中给出的 Sanov 定理版本，得到

$$
\begin{aligned}
&-\frac{1}{n - \lceil \rho n \rceil} \log P_{[\mathsf{T}_n]}(D(\mathsf{P}_{\mathbf{Y}^{(2)}_{\mathsf{T}_n+1}} \| \pi) - D(\mathsf{P}_{\mathbf{Y}^{(2)}_1} \| \pi) + \tilde{\varepsilon} > 0) \\
&\qquad \to \min_{Q_1, Q_2} \{D(Q_1 \| \mu) + D(Q_2 \| \pi)\} \qquad (50)
\end{aligned}
$$

当 n 趋于无穷时。在 (50) 中，最小值取遍所有满足 $D(Q_2\|\pi) - D(Q_1\|\pi) + \tilde{\varepsilon} \ge 0$ 的分布 (Q_1, Q_2)。按照与 (29) 中类似的步骤，可得

$$
\min_{Q_1, Q_2} \{D(Q_1 \| \mu) + D(Q_2 \| \pi)\} \ge 2B(\mu, \pi) - \tilde{\varepsilon} \qquad (51)
$$

其中，对 Q_1 取最小值时遍历所有概率分布 Q_1。由于

$$
\frac{1}{n} = \frac{n - \lceil \rho n \rceil}{n} \frac{1}{n - \lceil \rho n \rceil} \qquad (52)
$$

并且根据假设，$\lim_{n\to\infty} \log \mathsf{M}_n / n = 0$，由此可从 (46)–(51) 得到

$$
\begin{aligned}
&\varliminf_{n\to\infty} -\frac{1}{n} \mathsf{E}\left[\log\big(\epsilon(\delta_{\text{median}}|\bar{\pi})\big) \big| \bar{\pi} \in \mathcal{B}_{\varepsilon,\infty}(\pi)\right] \\
&\quad \ge (1 - \rho)\big(2B(\mu, \pi) - \tilde{\varepsilon}\big) \qquad (53)
\end{aligned}
$$

（其中 $\varliminf$ 表示下极限）。这里，因子 $(1-\rho)$ 是 $(n - \lceil \rho n \rceil)/n$ 当 $n \to \infty$ 时的极限。

将引理 1 应用于长度为 $\lceil \rho n \rceil$ 的序列 $\mathbf{Y}^{(1)}_1$、……、$\mathbf{Y}^{(1)}_{\mathsf{M}_n}$，可得概率 $P_{[\mathsf{T}_n]}\big(\bar{\pi} \in \mathcal{B}_{\varepsilon,\ \infty}(\pi)\big)$ 趋于 1（当 $n \to \infty$ 时）。因此，令 ρ 和 ε 趋于零，由 (43) 和 (53) 即可得出该定理。 ■

致谢

本研究部分得到了欧盟 HORIZON研究与创新计划的资助，资助协议编号为 101120657，项目为 ENFIELD（European Lighthouse to Manifest Trustworthy and Green AI）；西班牙科学、创新与大学部通过资助项目 PID2024-159557OB-C21（MI-CIU/AEI/10.13039/501100011033 和 ERDF/UE）提供资助；马德里自治区通过资助项目 IDEA-CM（TEC-2024/COM-89）提供资助。

参考文献

[1] Y. Li、S. Nitinawarat 和 V. V. Veeravalli，“通用离群假设检验”，*IEEE*信息论汇刊，第 60 卷，第 7 期，第 4066–4082 页，2014 年 7 月。[2] L. Zhou、A. Hero 和 Y. Wei，“带拒绝选项的二阶渐近最优离群序列检测”，载于*IEEE*信息论研讨会（*ITW*）论文集，2021 年，第 1–6 页。[3] S. Nitinawarat 和 V. V. Veeravalli，“通用最快离群检测与隔离”，载于*2015* 年 *IEEE*国际信息论研讨会（*ISIT*）论文集，2015 年 6 月，第 770–774 页。[4] Y. Bu、S. Zou、Y. Liang 和 V. V. Veeravalli，“连续观测的通用离群序列检测”，载于*2016* 年 *IEEE*国际声学、语音与信号处理会议（*ICASSP*）论文集，2016 年 3 月，第 4254–4258 页。[5] Y. Li、S. Nitinawarat 和 V. V. Veeravalli，“通用序贯离群假设检验”，载于*2014* 年 *IEEE*国际信息论研讨会（*ISIT*）论文集，2014 年 6 月，第 3205–3209 页。[6] ——，“通用序贯离群假设检验”，载于*2014* 年第 *48* 届阿西洛马会议：信号、系统与计算机论文集，2014 年 11 月，第 281–285 页。[7] S. C. Sreenivasan 和 S. Bhashyam，“异常数据流的序贯非参数检测”，*IEEE*信号处理快报，第 28 卷，第 932–936 页，2021 年。[8] J. Acharya、A. Jafarpour、A. Orlitsky 和 A. T. Suresh，“离群检测与广义接近性检验的次线性算法”，载于*2014* 年 *IEEE*国际信息论研讨会（*ISIT*）论文集，2014 年 6 月，第 3200–3204 页。[9] A. Barg 和 G. Forney，“随机码：最小距离与错误指数”，*IEEE*信息论汇刊，第 48 卷，第 9 期，第 2568–2573 页，2002 年 9 月。[10] A. Nazari、A. Anastasopoulos 和 S. S. Pradhan，“多址信道的错误指数：下界”，*IEEE*信息论汇刊，第 60 卷，第 9 期，第 5095–5115 页，2014 年 9 月。[11] N. Merhav，“典型随机码的错误指数”，*IEEE*信息论汇刊，第 64 卷，第 9 期，第 6223–6235 页，2018 年 9 月。[12] L. V. Truong 和 A. Guillén i Fàbregas，“广义随机 Gilbert-Varshamov 码：典型错误指数与集中性质”，*IEEE*信息论汇刊，第 70 卷，第 2 期，第 820–853 页，2024 年 2 月。[13] Y. Li 和 V. V. Veeravalli，“大型数据集中的离群序列检测：通用假设检验与聚类的比较”，载于*2016* 年 *IEEE*国际声学、语音与信号处理会议（*ICASSP*）论文集，2016 年 3 月，第 6180–6184 页。[14] Y. Bu、S. Zou 和 V. V. Veeravalli，“用于通用离群序列检测的线性复杂度指数一致检验”，载于*2017* 年 *IEEE*国际信息论研讨会（*ISIT*）论文集，2017 年 6 月，第 988–992 页。[15] ——，“用于通用离群序列检测的线性复杂度指数一致检验”，*IEEE*信号处理汇刊，第 67 卷，第 8 期，第 2115–2128 页，2019 年。[16] W. Hoeffding，“有界随机变量之和的概率不等式”，美国统计学会会刊，第 58 卷，第 301 期，第 13–30 页，1963 年 3 月。