arXiv:2303.02826v1 [stat.ME] 6 Mar 2023

统计周期过程中变化后分布未知时的最快变化检测

Yousef Oleyaeimotlagh[a]、Taposh Banerjee[a]、Ahmad Taha[b] 和 Eugene John[ca]匹兹堡大学，[b]范德堡大学，[c]德克萨斯大学圣安东尼奥分校

文章历史
编于2023年3月7日

摘要
本文针对统计周期过程中的变化最快检测问题提出了若干算法。这类过程的统计特性虽非平稳，但会以固定时间间隔重复。假设决策者已知变化前分布，但未知变化后分布。在此框架下，研究了三类问题：鲁棒最快变化检测、联合最快变化检测与分类，以及多时隙最快变化检测。在多时隙问题中，变化可能发生在一个周期内的具体时隙未知。本文针对每个问题提出了算法，并证明这些算法要么具有精确最优性，要么在低虚警率区间内渐近最优。随后，将所提出的算法用于交通数据异常检测，以及心电图（ECG）数据中的心律失常检测与识别。还通过模拟数据展示了这些算法的有效性。

关键词
鲁棒变化检测、联合变化检测与故障隔离、多槽变点检测、异常检测、交通数据、心律失常检测与识别。

1. 引言

在经典的最快变化检测问题中（见 [21], [25], [27]），决策者观测一个分布已知的随机过程。在某个时刻，该过程的分布发生变化。该问题的目标是在满足虚警率约束的前提下，尽可能快速地检测出这一分布变化，并将延迟降至最低。该问题在统计过程控制（[23]）、传感器网络（[6]）、信息物理系统监测（[11]）、神经数据中的状态变化（[1]）、交通监测（[8]）以及一般的异常检测（[7, 8]）等领域都有应用。

在许多异常检测应用中，观测到的过程具有统计周期性行为。以下是一些示例：

(1) 心电图数据中的心律失常检测：心电描记术（ECG）数据具有近似周期性的波形模式，由一系列P波、QRS波群和ST段组成。心律失常可能导致这种规律模式发生变化（[14]）。

联系人：Taposh Banerjee。电子邮件：taposh.banerjee@pitt.edu

(2) 神经尖峰数据中的变化检测：在某些脑机接口（BCI）研究中（[29]），对动物进行一系列相同的实验试验，使每次试验产生相似的放电模式。某个事件或触发信号（实验的一部分）可能在某次试验后改变放电模式。(3) 城市交通数据中的异常检测：研究发现，纽约市（NYC）某街道交叉口的车辆数量呈现繁忙时段和空闲时段交替的规律模式（[2–4, 7, 8]）。拥堵或事故可能导致车辆数量下降或增加。(4) 社交网络数据：研究还发现，纽约市某闭路电视摄像机附近发布的Instagram消息数量也呈现近似周期性行为（[2–4, 7, 8]）。(5) 高速公路上的拥堵模式检测：在交通密度估计问题中，人们通常希望先检测交通状况的模式（拥堵或不拥堵），然后再决定用于估计的模型（[28]）。受纽约市数据行为的启发，该应用中的交通强度也可以建模为具有统计周期性。

在 [5], 中，引入了一类新的随机过程，称为独立且周期性同分布（i.p.i.d.）过程，用于对统计周期性数据建模。在这一过程中，随机变量序列相互独立，且这些变量的分布以给定的周期 T 为周期变化。

统计周期过程也可以用循环平稳过程（[13]）建模。然而，使用 i.p.i.d. 过程建模可实现样本级检测，并发展出有力的最优性理论。

在 [5], 中，针对 i.p.i.d. 过程中的最快变化检测，建立了贝叶斯理论。研究表明，与 i.i.d. 情形类似，最优方法是使用 Shiryaev统计量进行变化检测，也就是说，根据数据计算变化已经发生的后验概率。然而，在 i.p.i.d. 情形中，只有当 Shiryaev统计量序列越过一系列随时间变化但呈周期性的阈值时，才会宣告发生变化。研究还表明，当虚警概率约束趋于零时，单阈值检验渐近最优。所提出的算法还可以递归方式实现，并且只需有限记忆。因此，i.p.i.d. 过程这一设定给出了一个非 i.i.d. 情形的例子，其中精确最优算法能够高效实现。 [5] 中的结果在变化前和变化后的分布均已知时有效。

本文研究变化后分布未知时独立非同分布过程中的最快变化检测问题。我们考虑该问题在极小极大和贝叶斯设定下的三种不同表述：

(1) 鲁棒最快变化检测：在第2节中，我们首先研究独立非同分布过程中的鲁棒最快变化检测问题。在此问题中，我们假设变化后分布族未知，但属于某个不确定类。我们还假设变化后分布族中存在一个最不利分布。随后，我们证明，基于最不利分布设计的算法对于贝叶斯延迟度量而言是极小极大鲁棒的。(2) 最快检测与故障识别：在第3节中，我们考虑这样一个问题：变化后分布未知，但属于一个有限分布类。针对这一设置，我们解决了独立非同分布过程中的最快变化检测与隔离联合问题。我们还将所提出的算法应用于真实心电图数据，以检测心律失常。

(3) 多时隙最快变化检测：在第4节中，我们研究独立非同分布过程中的多时隙最快变化检测问题。在此问题中，给定周期内变化可能发生的确切时隙未知。我们证明，基于混合的检验是渐近最优的。

我们工作的一个显著特点是，除了为所提出的算法建立最优性理论外，我们还将其应用于真实数据或仿真数据，以展示其有效性。具体而言，在第5.1节中，我们研究交通数据中的异常检测。在第5.4节中，我们将开发的算法应用于心电图数据中的心律失常检测与隔离。在第5.3节、第5.2节和第5.4.5节中，我们还将算法应用于模拟数据，以展示其有效性。

2. 鲁棒最快变化检测

2.1. 模型与问题表述

我们首先定义本文将用于对统计周期性随机过程建模的过程。

定义2.1 ([5])。 若随机过程 $\{X_n\}$ 满足以下条件，则称其为独立且周期同分布（独立非同分布）：

(1) 随机变量 $\{X_n\}$ 相互独立。(2) 若 X_n 的密度为 f_n，其中 $n \geq 1$，则存在一个正整数 T，使得密度序列 $\{f_n\}$ 以 T为周期:

$$f_{n+T} = f_n, \quad \forall n \geq 1.$$

i.p.i.d. 过程的分布完全由（X_1，.....，X_T）的有限维乘积分布或密度集合（f_1，$\cdots$，f_T）刻画，我们称该过程是分布为（f_1，$\cdots$，f_T）的 i.p.i.d. 过程。我们关注的变点问题如下。在正常状态下，数据被建模为一个分布为（f_1，$\cdots$，f_T）的 i.p.i.d. 过程。在某个时刻，由于发生了某个事件，该 i.p.i.d. 过程的分布偏离（f_1，$\cdots$，f_T）。具体而言，考虑另一个周期性密度序列 $\{g_n\}$，使得

$$g_{n+T} = g_n, \quad \forall n \geq 1.$$

假设在变化点 ν，i.p.i.d. 过程的分布从（f_1，$\cdots$，f_T）切换为（g_1，$\cdots$，g_T）：

$$X_n \sim \begin{cases} f_n, & \forall n < \nu, \\ g_n & \forall n \geq \nu. \end{cases} \tag{1}$$

密度（$g_1, \cdots, g_T$）不必与密度集合（$f_1, \cdots, f_T$）中的密度全部不同，但我们假设至少存在一个 i 使它们不同——

:

$$g_i \neq f_i, \quad \text{for some } i = 1, 2, \cdots, T. \tag{2}$$

本文假设变化后分布 (g_1, $\cdots$, g_T) 未知。此外，存在 T 个分布族 $\{\mathcal{P}_i\}_{i=1}^T$，使得

$$g_i \in \mathcal{P}_i, \quad i = 1, 2, \ldots, T.$$

决策者已知这些分布族 $\{\mathcal{P}_i\}_{i=1}^T$。下文中，我们使用如下记号

$$G = (g_1, g_2, \ldots, g_T)$$

表示变化后的独立非同分布分布。

设 τ 是过程 $\{X_n\}$ 的停止时刻，即一个取正整数值的随机变量，使得事件 $\{\tau \leq n\}$ 属于由 X_1, $\cdots$, X_n 生成的 σ-代数。换言之，是否达到 $\tau \leq n$ 完全由前 n 个观测值决定。我们宣告变化发生在停止时刻 τ。为了找到检测分布变化的最佳停止规则，我们需要一个性能准则。为此，我们将变化点 ν 建模为一个随机变量，其先验分布为

$$\pi_n = \mathsf{P}(\nu = n), \quad \text{for } n = 1, 2, \cdots.$$

对于每个 $n \in \mathbb{N}$，我们用 P_n^G 表示变化发生在 $\nu = n$ 且变化后分布为 G 时观测过程 $\{X_n\}$ 的分布。我们用 E_n^G 表示相应的期望。利用此记号，我们定义平均概率测度

$$\mathsf{P}^{\pi,G} = \sum_{n=1}^{\infty} \pi_n \, \mathsf{P}_n^G.$$

为体现停止时刻在变化发生前到达时所产生的虚警惩罚，我们定义虚警概率如下:

$$\mathsf{P}^{\pi,G}(\tau < \nu).$$

注意，虚警概率 $\mathsf{P}^{\pi,G}(\tau < \nu)$ 与变化后的分布 G 无关。因此，在下文中，我们省略对 G 的提及，仅以如下形式表示虚警概率:

$$\mathsf{P}^{\pi}(\tau < \nu).$$

为惩罚检测延迟，我们采用如下平均检测延迟:

$$\mathsf{E}^{\pi,G}\left[(\tau - \nu)^+\right],$$

其中 $x^+ =$ 最大值$\{x, 0\}$。

我们感兴趣并希望求解的优化问题是

$$\inf_{\tau\in\mathbf{C}_\alpha}\ \sup_{G:g_i\in\mathcal{P}_i,i\le T}\ \mathsf{E}^{\pi,G}\left[(\tau-\nu)^+\right], \tag{3}$$

其中

$$\mathbf{C}_\alpha=\{\tau:\mathsf{P}^\pi(\tau<\nu)\le\alpha\},$$

且 α 是对虚警概率给定的约束。

当分布族 $\{\mathcal{P}_i\}_{i=1}^T$ 是单元素集合时，即变化后分布已知且固定为 G 时，文献 [5] 研究了该问题的拉格朗日松弛。理解文献 [5]中给出的解，是求解式 (3) 中鲁棒问题的基础。下一节将讨论文献 [5] 给出的解，以及该解对式 (3) 中约束版本的启示。

2.2. 已知变化后分布下的精确最优解与渐近最优解

对于已知的变化后分布 $G=(g_1,\ .....,\ g_T)$，且变化点服从几何分布时，文献 [5] 证明，式 (3) 松弛版本的精确最优解是基于周期性阈值序列的停止规则。文献还证明，在虚警约束$\alpha\to 0$的渐近情形下，只需使用一个阈值。此外，在渐近情形下，可以放宽变化点服从几何分布这一假设。本节余下部分中，我们假设 G 已知且固定。

2.2.1. 精确最优算法

设变化点 ν 服从几何分布:

$$\mathsf{P}(\nu=n)=(1-\rho)^{n-1}\rho,\qquad \text{for } n=1,2,\cdots.$$

(3) 的松弛版本（已知 G）为

$$\inf_{\tau}\ \mathsf{E}^{\pi,G}\left[(\tau-\nu)^+\right]+\lambda_f\ \mathsf{P}^\pi(\tau<\nu), \tag{4}$$

其中 $\lambda_f>0$ 是虚警代价的惩罚项。现在，定义 $p_0=0$ 和

$$p_n=\mathsf{P}^{\pi,G}(\nu\le n|X_1,\cdots,X_n),\ \text{for } n\ge 1. \tag{5}$$

那么，(4) 等价于求解

$$\inf_{\tau}\ \mathsf{E}^{\pi,G}\left[\sum_{n=0}^{\tau-1}p_n+\lambda_f(1-p_\tau)\right]. \tag{6}$$

更新后的信念 p_n 可以使用以下方程递归计算：$p_0 = 0$ 对于 $n \geq 1$,

$$p_n = \frac{\tilde{p}_{n-1}\, g_n(X_n)}{\tilde{p}_{n-1}\, g_n(X_n) + (1 - \tilde{p}_{n-1}) f_n(X_n)}, \tag{7}$$

其中

$$\tilde{p}_{n-1} = p_{n-1} + (1 - p_{n-1})\rho.$$

由于这些更新不是平稳的，因此无法使用经典最优停止理论 [22] 或动态规划 [9]来解决该问题。然而，式（7）中的结构在每经过固定的时间 T后便会重复。受此启发，本文在 [5], 中针对具有周期性转移和成本结构的马尔可夫决策过程建立了一套控制理论。随后利用这一新的控制理论求解式(6)中的问题。

定理 2.2 （[5]）**：** 存在阈值 A_1、 A_2、...、 A_T、 $A_i \geq 0$、$\forall i$，使得停止规则

$$\tau^* = \inf\{n \geq 1 : p_n \geq A_{(n \bmod T)}\}, \tag{8}$$

其中 （n 模 T） 表示 n 对 T取模，在（6）中的问题上是最优的。这些阈值取决于 λ_f 的选择。

事实上， [5] 中给出的解对于更一般的变点问题同样有效，在该问题中，每个时隙分别采用独立的延迟惩罚项和虚警惩罚项。本文不对此进行讨论。

2.2.2. 渐近最优算法

对于较大的 T，在某些应用中其值甚至可能超过一百万，存储 T 种不同阈值在计算上并不可行。因此，我们有必要考察单阈值算法是否最优。 [5] 证明了周期性阈值算法严格最优。然而， [5]证明了在虚警概率较低的情形下，单阈值检验渐近最优。我们将在下文讨论这一结果。

设存在 $d \geq 0$ ，使得

$$\lim_{n\to\infty} \frac{\log \mathsf{P}(\nu > n)}{n} = -d. \tag{9}$$

若 $\pi =$ 几何分布(ρ)，则 $d = |\log(1-\rho)|$。此外，令

$$I = \frac{1}{T}\sum_{i=1}^{T} D(g_i \parallel f_i), \tag{10}$$

其中 $D(g_i \parallel f_i)$ 是密度 g_i 与 f_i之间的 Kullback-Leibler 散度:

$$D(g_i \parallel f_i) = \int g_i(x) \log \frac{g_i(x)}{f_i(x)} dx.$$

下述定理在 [5]中得到证明。

定理 2.3 ([5])。设信息数 I 如(10) 中所定义，且满足$0 < I < \infty$。此外，设 d 如(9)中所示。则，有

$$A_1 = A_2 = \cdots = A_T = 1 - \alpha,$$

$\tau^* \in \mathbf{C}\alpha$，且

$$\begin{aligned} \mathsf{E}^{\pi,G}\left[(\tau^* - \nu)^+\right] &= \inf_{\tau \in \mathbf{C}_\alpha} \mathsf{E}^{\pi,G}\left[(\tau - \nu)^+\right](1 + o(1)) \\ &= \frac{|\log \alpha|}{I + d}(1 + o(1)), \qquad as\ \alpha \to 0. \end{aligned} \tag{11}$$

此处，当 $\alpha \to 0$时， $o(1) \to 0$ 。

2.2.3. 该问题约束版本的解

现在我们论证，与经典情形一样，问题 (6) 的松弛版本可用于为问题 (3) 的约束版本提供解。为完整起见，我们给出证明。

引理 2.4。 如果 α 是最优停止规则 τ^* 在 （6） 中可达到的虚警概率，那么对于该 α， τ^* 对约束问题 （3） 也是最优的。

证明。 根据定理 2.2，我们有

$$\mathsf{E}^{\pi,G}\left[(\tau^* - \nu)^+\right] + \lambda_f\, \mathsf{P}^{\pi}(\tau^* < \nu) \le \mathsf{E}^{\pi,G}\left[(\tau - \nu)^+\right] + \lambda_f\, \mathsf{P}^{\pi}(\tau < \nu). \tag{12}$$

如果 $\mathsf{P}^{\pi}(\tau^* < \nu) = \alpha$ 且 $\mathsf{P}^{\pi}(\tau < \nu) \le \alpha$，那么

$$\begin{aligned} \mathsf{E}^{\pi,G}\left[(\tau^* - \nu)^+\right] + \lambda_f\, \mathsf{P}^{\pi}(\tau^* < \nu) &= \mathsf{E}^{\pi,G}\left[(\tau^* - \nu)^+\right] + \lambda_f\, \alpha \\ \le \mathsf{E}^{\pi,G}\left[(\tau - \nu)^+\right] + \lambda_f\, \mathsf{P}^{\pi}(\tau < \nu) &\le \mathsf{E}^{\pi,G}\left[(\tau - \nu)^+\right] + \lambda_f\, \alpha. \end{aligned} \tag{13}$$

从两边约去 $\lambda_f\, \alpha$ ，得到

$$\mathsf{E}^{\pi,G}\left[(\tau^* - \nu)^+\right] \le \mathsf{E}^{\pi,G}\left[(\tau - \nu)^+\right].$$

□

以下引理保证，通过最优停止规则 τ^* 可以实现广泛的虚警概率 α。

引理 2.5。 随着式(6)中的 $\lambda_f \to \infty$ 增大，最优停止规则 τ^* 所达到的虚警概率趋于零。

证明。 当 $\lambda_f \to \infty$时，若 τ^* 的虚警概率始终有别于零的下界，则贝叶斯风险

$$\mathsf{E}^{\pi,G}\left[(\tau^* - \nu)^+\right] + \lambda_f \, \mathsf{P}^{\pi}(\tau^* < \nu)$$

将发散至无穷大。这将与 τ^* 是最优的事实相矛盾，因为在足够大的 λ_f 下，通过在足够大的确定性时间停止，我们可以获得更小的风险。

□

2.3. 未知变化后分布的最优鲁棒算法

现假设变化后分布 G 未知，并在对变化后分布族 $\{\mathcal{P}_i\}_{i=1}^{T}$作出假设的条件下，给出式(3)的最优解。具体而言，我们将 [26] 中针对独立同分布过程得到的结果扩展到独立非同分布过程。以下将假设所有涉及的密度彼此等价（相互绝对连续）。此外，假设变化点 ν 是一个服从几何分布的随机变量。

为了陈述关于 $\{\mathcal{P}_i\}_{i=1}^{T}$ 的假设，我们需要先给出一些定义。若随机变量 Z_2 随机占优于另一个随机变量 Z_1，则称

$$\mathsf{P}(Z_2 \geq t) \geq \mathsf{P}(Z_1 \geq t), \quad \text{for all } t \in \mathbb{R}.$$

我们使用记号

$$Z_2 \succ Z_1.$$

若 $\mathcal{L}_{Z_2}$ 和 $\mathcal{L}_{Z_1}$ 分别是 Z_2 和 Z_1 的概率分布，我们也使用记号

$$\mathcal{L}_{Z_2} \succ \mathcal{L}_{Z_1}.$$

现在我们引入独立非同分布过程中随机有界性的概念。下文中，我们用

$$\mathcal{L}(\phi(X), g)$$

表示随机变量 X 的某个函数 $\phi(X)$ 的分布，其中变量 X 的密度为 g。

定义 2.6（独立非同分布过程中的随机有界性；最不利分布）。

我们称族 $\{\mathcal{P}_i\}_{i=1}^T$ 被独立非同分布分布律随机有界

$$\bar{G} = (\bar{g}_1, \bar{g}_2, \ldots, \bar{g}_T),$$

并称 G 为最不利律（LFL），如果

$$\bar{g}_i \in \mathcal{P}_i, \quad i = 1, 2, \ldots, T,$$

and

$$\mathcal{L}\left(\log \frac{\bar{g}_i(X_i)}{f_i(X_i)}, g_i\right) \succ \mathcal{L}\left(\log \frac{\bar{g}_i(X_i)}{f_i(X_i)}, \bar{g}_i\right), \quad \text{for all } \ g_i \in \mathcal{P}_i, \quad i = 1, 2, \ldots, T. \tag{14}$$

考虑使用 LFL $G = (\bar{g}_1, \ \bar{g}_2, \ \ldots, \ \bar{g}_T)$ 设计的停止规则 τ^*：

$$\bar{\tau}^* = \inf\{n \geq 1 : \bar{p}_n \geq A_{(n \bmod T)}\}, \tag{15}$$

其中 $\bar{p}_0 = 0$，且

$$\bar{p}_n = \frac{\tilde{p}_{n-1}\ \bar{g}_n(X_n)}{\tilde{p}_{n-1}\ \bar{g}_n(X_n) + (1 - \tilde{p}_{n-1}) f_n(X_n)}, \tag{16}$$

其中

$$\tilde{p}_{n-1} = \bar{p}_{n-1} + (1 - \bar{p}_{n-1})\rho.$$

现陈述我们关于独立非同分布过程中的鲁棒最快变化检测的主要结果。

定理 2.7。假设以下条件成立：

(1) 族 $\{\mathcal{P}_i\}_{i=1}^T$ 受独立非同分布分布律随机有界

$$\bar{G} = (\bar{g}_1, \bar{g}_2, \ldots, \bar{g}_T).$$

(2) 令 $\alpha \in [0, 1]$ 为一个约束，使得

$$\mathsf{P}^{\pi}(\bar{\tau}^* < \nu) = \alpha,$$

其中 $\bar{\tau}^*$ 是采用 *LFL* (15)设计的最优规则。

(3) 所涉及的所有似然比函数均连续。

(4) 变化点 ν 服从几何分布。

因此，利用 *LFL* 设计的停止规则 $\bar{\tau}^*$ *in* (15) 对于 (3) 中的鲁棒约束问题是最优的。

证明。 证明的关键步骤是证明，对于每个 $k \in \mathbb{N}$,

$$\begin{aligned} \mathsf{E}_k^{\bar{G}}\left[(\bar{\tau}^* - k)^+ | \mathcal{F}_{k-1}\right] \succ \mathsf{E}_k^{G}\left[(\bar{\tau}^* - k)^+ | \mathcal{F}_{k-1}\right], \\ \text{for all } G = (g_1, \dots g_T) : g_i \in \mathcal{P}_i,\ i \le T, \end{aligned} \tag{17}$$

其中 $\mathcal{F}_{k-1}$ 是由观测值 X_1、.....、 X_{k-1} 生成的σ代数。若（17）成立，则对每个 $k \in \mathbb{N}$, 有

$$\begin{aligned} \mathsf{E}_k^{\bar{G}}\left[(\bar{\tau}^* - k)^+\right] \ge \mathsf{E}_k^{G}\left[(\bar{\tau}^* - k)^+\right] \\ \text{for all } G = (g_1, \dots g_T) : g_i \in \mathcal{P}_i,\ i \le T. \end{aligned} \tag{18}$$

对变化点的先验分布取平均后，得到

$$\begin{aligned} &\mathsf{E}^{\pi,\bar{G}}\left[(\bar{\tau}^* - \nu)^+\right] = \sum_k \pi_k \mathsf{E}_k^{\bar{G}}\left[(\bar{\tau}^* - k)^+\right] \ge \sum_k \pi_k \mathsf{E}_k^{G}\left[(\bar{\tau}^* - k)^+\right] = \mathsf{E}^{\pi,G}\left[(\bar{\tau}^* - \nu)^+\right], \\ &\text{for all } G = (g_1, \dots g_T) : g_i \in \mathcal{P}_i,\ i \le T. \end{aligned} \tag{19}$$

最后一个等式给出

$$\begin{aligned} \mathsf{E}^{\pi,\bar{G}}\left[(\bar{\tau}^* - \nu)^+\right] \ge \mathsf{E}^{\pi,G}\left[(\bar{\tau}^* - \nu)^+\right], \\ \text{for all } G = (g_1, \dots g_T) : g_i \in \mathcal{P}_i,\ i \le T. \end{aligned} \tag{20}$$

这意味着

$$\mathsf{E}^{\pi,\bar{G}}\left[(\bar{\tau}^* - \nu)^+\right] = \sup_{G: g_i \in \mathcal{P}_i, i \le T} \mathsf{E}^{\pi,G}\left[(\bar{\tau}^* - \nu)^+\right], \tag{21}$$

¯这里之所以取等号，是因为分布 G 属于右侧所考虑的族。现在，若 τ 是满足虚警概率约束¯（其值为 α）的任意停止规则，那么由于 $\bar{\tau}^*$ 是 LFL 的最优检验 G, 我们有（参见定理 2.2 和引理 2.4）

$$\begin{aligned} \sup_{G: g_i \in \mathcal{P}_i, i \le T} \mathsf{E}^{\pi,G}\left[(\tau - \nu)^+\right] &\ge \mathsf{E}^{\pi,\bar{G}}\left[(\tau - \nu)^+\right] \ge \mathsf{E}^{\pi,\bar{G}}\left[(\bar{\tau}^* - \nu)^+\right] \\ &= \sup_{G: g_i \in \mathcal{P}_i, i \le T} \mathsf{E}^{\pi,G}\left[(\bar{\tau}^* - \nu)^+\right]. \end{aligned} \tag{22}$$

最后一个等式证明了停止规则 $\bar{\tau}^*$ 对问题（3）的稳健最优性。

现在我们证明关键步骤 (17)。为此，我们证明对于每个整数$N \ge 0$,

$$\begin{aligned} \mathsf{P}_k^{\bar{G}}\left[(\bar{\tau}^* - k)^+ > N | \mathcal{F}_{k-1}\right] \ge \mathsf{P}_k^{G}\left[(\bar{\tau}^* - k)^+ > N | \mathcal{F}_{k-1}\right], \\ \text{for all } G = (g_1, \dots g_T) : g_i \in \mathcal{P}_i,\ i \le T. \end{aligned} \tag{23}$$

由于事件 $N = 0$ $\{(\bar{\tau}^* - k)^+ > 0\}$ 是 $\mathcal{F}_{k-1}$-可测的，这显然成立。因此我们

只需证明 $N \geq 1$ 的情形。为此，我们首先有

$$\begin{aligned} \mathsf{P}_k^{\bar{G}}\left[(\bar{\tau}^* - k)^+ \leq N | \mathcal{F}_{k-1}\right] &= \mathsf{P}_k^{\bar{G}}\left[\bar{\tau}^* \leq k+N | \mathcal{F}_{k-1}\right] \\ &= \mathsf{P}_k^{\bar{G}}\left[f(h_1(X_1), h_2(X_2), \ldots, h_{k+N}(X_{k+N})) \ \geq \ 0 \mid \mathcal{F}_{k-1}\right], \end{aligned} \tag{24}$$

其中

$$h_i(x) = \log \frac{\bar{g}_i(x)}{f_i(x)},$$

函数 $f(z_1, \ z_2, \ \ldots\ldots, \ z_N)$ 定义为

$$f(z_1, z_2, \ldots, z_N) = \max_{1 \leq n \leq N} \left(\sum_{i=1}^{n} (1-\rho)^{k-1} \rho \ \exp\left(\sum_{i=k}^{n} z_i \right) - B_n \right), \tag{25}$$

and

$$B_n = \frac{A_n}{1 - A_n} (1-\rho)^n.$$

现在回顾公式（14），可知

$$\mathcal{L}\left(h_i(X_i), g_i\right) \succ \mathcal{L}\left(h_i(X_i), \bar{g}_i\right), \quad \text{for all} \ \ g_i \in \mathcal{P}_i, \quad i = 1, 2, \ldots, T. \tag{26}$$

由于函数 f 连续（因为它是连续函数的最大值），且关于每个自变量单调不减，因此 [26] 中的引理 III.1 表明

$$\begin{aligned} &\mathsf{P}_k^{\bar{G}}\left[f(h_1(X_1), h_2(X_2), \ldots, h_{k+N}(X_{k+N})) \ \geq \ 0 \mid \mathcal{F}_{k-1}\right] \\ &\quad \leq \mathsf{P}_k^{G}\left[f(h_1(X_1), h_2(X_2), \ldots, h_{k+N}(X_{k+N})) \ \geq \ 0 \mid \mathcal{F}_{k-1}\right], \\ &\qquad \text{for all } G = (g_1, \ldots g_T) : g_i \in \mathcal{P}_i, \ i \leq T. \end{aligned} \tag{27}$$

结合公式（24）和（27）可得

$$\begin{aligned} \mathsf{P}_k^{\bar{G}}\left[(\bar{\tau}^* - k)^+ \leq N \mid \mathcal{F}_{k-1}\right] &= \mathsf{P}_k^{\bar{G}}\left[\bar{\tau}^* \leq k+N \mid \mathcal{F}_{k-1}\right] \\ &= \mathsf{P}_k^{\bar{G}}\left[f(h_1(X_1), h_2(X_2), \ldots, h_{k+N}(X_{k+N})) \geq 0 \mid \mathcal{F}_{k-1}\right] \\ &\leq \mathsf{P}_k^{G}\left[f(h_1(X_1), h_2(X_2), \ldots, h_{k+N}(X_{k+N})) \geq 0 \mid \mathcal{F}_{k-1}\right] \\ &= \mathsf{P}_k^{G}\left[\bar{\tau}^* \leq k+N \mid \mathcal{F}_{k-1}\right] \\ &= \mathsf{P}_k^{G}\left[(\bar{\tau}^* - k)^+ \leq N \mid \mathcal{F}_{k-1}\right], \\ &\qquad \text{for all } G = (g_1, \ldots g_T) : g_i \in \mathcal{P}_i, \ i \leq T. \end{aligned} \tag{28}$$

这证明了 (23)，从而也证明了 (17)。 □

3. 最快联合检测与分类

3.1. 联合检测与分类问题表述

我们假设在正常状态下，数据可以建模为一个分布为 $(g_1^{(0)}, \cdots, g_T^{(0)})$ 的独立非同分布过程。在某个时刻 ν，该独立非同分布过程不再由密度 $(g_1^{(0)}, \cdots, g_T^{(0)})$ 描述，而是由密度 $(g_1^{(\ell)}, \cdots, g_T^{(\ell)})$、$\ell = 1$、2、.....、$M$ 中的一个描述，其中

$$g_{n+T}^{(\ell)} = g_n^{(\ell)}, \quad \forall n \geq 1, \quad \ell = 1, 2, \ldots, M.$$

具体而言，在时间点 ν，随机变量的分布从$\{g_n^{(0)}\}$变为 $\{g_n^{(\ell)}\}$：对于某个 $\ell = 1$，2，.....，M，

$$X_n \sim \begin{cases} g_n^{(0)}, & \forall n < \nu, \\ g_n^{(\ell)} & \forall n \geq \nu. \end{cases} \tag{29}$$

我们希望尽可能快地检测到 (29) 中描述的变化，同时满足关于虚警率和误分类概率的约束。从数学上说，我们要寻找一对 (τ, δ)，其中 τ 是停止时刻，即

$$\{\tau \leq n\} \in \sigma(X_1, X_2, \ldots, X_n),$$

而 δ 是决策规则，即一个满足以下条件的映射

$$\delta(X_1, X_2, \ldots, X_\tau) \in \{1, 2, \ldots, M\}.$$

令 $\mathsf{P}_\nu^{(\ell)}$ 表示过程 $\{X_n\}$ 在变化发生于时间 ν 且变化后分布为 $(g_1^{(\ell)}, \cdots, g_T^{(\ell)})$ 时的概率律。令 $\mathsf{E}_\nu^{(\ell)}$ 表示相应的期望。无变化时，我们使用记号 E_∞。我们关心的问题如下 [17]:

$$\begin{aligned} \min_{\tau,\delta} \quad & \max_{1\leq\ell\leq M} \sup_{\nu\geq 1} \operatorname{ess\,sup} \mathsf{E}_\nu^{(\ell)}[(\tau - \nu + 1)^+ | X_1, \cdots, X_{\nu-1}], \\ \text{subj. to} \quad & \mathsf{E}_\infty[\tau] \geq \beta, \\ \text{and} \quad & \mathsf{P}_1^{(\ell)}[\tau < \infty, \delta \neq \ell] \leq a_\beta \, \mathsf{E}_1^{(\ell)}[\tau], \quad \ell = 1, 2, \ldots, M, \\ & \text{where } \log a_\beta^{-1} \sim \log \beta, \text{ as } \beta \to \infty. \end{aligned} \tag{30}$$

此处，esssup 表示随机变量 $\mathsf{E}_\nu^{(\ell)}[(\tau - \nu + 1)^+ | X_1, \cdots, X_{\nu-1}]$ 的本质上确界，即以概率 1 支配该随机变量的最小常数。在下文中，对于 β 的两个函数 $h(\beta)$ 和 $f(\beta)$，我们使用$f(\beta) \sim h(\beta)$（如 $\beta \to \infty$）表示这两个函数之比在极限情况下趋于 1。关于变点检测与隔离的这一问题表述及其他问题表述的更多动机，可参见文献 [25], [17], [20]。

$M = 1$时的检测算法

当 $M = 1$ 时，即只有一个变化后独立非同分布分布时，对于检测分布变化而言，渐近最优的算法是文献 [3] 提出的周期性累积和（CUSUM）算法。在该算法中，我们计算统计量序列

$$W_{n+1} = W_n^+ + \log \frac{g_{n+1}^{(1)}(X_{n+1})}{g_{n+1}^{(0)}(X_{n+1})} \tag{31}$$

并在统计量超过阈值 A 时立即发出警报：

$$\tau_c = \inf\{n \geq 1 : W_n \geq A\}. \tag{32}$$

定义

$$I_{10} = \frac{1}{T}\sum_{i=1}^{T} D(g_i^{(1)} \parallel g_i^{(0)}), \tag{33}$$

其中， $D(g_i^{(1)} \parallel g_i^{(0)})$ 是密度 $g_i^{(1)}$ 和 $g_i^{(0)}$ 之间的 Kullback-Leibler 散度。随后，在 [3] 中证明了以下结果。

定理 3.1 ([3])：令（33）中定义的信息数 I_{10} 满足 $0 < I_{10} < \infty$。则，当 $A = \log\beta$ 时，

$$\mathsf{E}_\infty[\tau_c] \geq \beta,$$

并且当 $\beta \to \infty$ 时，

$$\begin{aligned} &\sup_{\nu\geq 1} \ ess\sup \mathsf{E}_\nu[(\tau_c - \nu + 1^+|X_1, \cdots, X_{\nu-1}] \\ &\sim \inf_{\tau:\mathsf{E}_\infty[\tau]\geq\beta} \ \sup_{\nu\geq 1} \ ess\sup \mathsf{E}_\nu[(\tau - \nu + 1)^+|X_1, \cdots, X_{\nu-1}] \\ &\sim \frac{\log\beta}{I_{10}}. \end{aligned} \tag{34}$$

因此，当虚警约束 $\beta \to \infty$ 时，周期性 CUSUM 算法在检测分布变化方面是渐近最优的。此外，由于变化前和变化后密度函数的集合 $(g_1^{(0)}, \cdots, g_T^{(0)})$ 和 $(g_1^{(1)}, \cdots, g_T^{(1)})$ 都是有限的，因此可以通过有限记忆来存储这些 $2T$密度函数，从而计算式 (31) 中的递推。

3.3. 联合检测与分类算法

当可能的变化后分布数量为 $M > 1$，且我们也希望准确分类真实的变化后分布时，周期性累积和（CUSUM）算法就不够用了。现在，我们提出一种能够同时进行检测和分类的算法。

对于 $\ell = 1$、.....、 M，定义停止时刻

$$\tau_\ell = \inf\left\{ n \geq 1 : \max_{1\leq k\leq n} \min_{0\leq m\leq M, m\neq \ell} \sum_{i=k}^{n} \log \frac{g_i^{(\ell)}(X_i)}{g_i^{(m)}(X_i)} \geq A \right\}. \tag{35}$$

我们检测-分类问题的停止时刻和决策规则定义如下:

$$\begin{aligned} \tau_{dc} &= \min_{1\leq \ell\leq M} \tau_\ell, \\ \delta_{dc} &= \arg\min_{1\leq \ell\leq M} \tau_\ell. \end{aligned} \tag{36}$$

将式 (35) 中的每个 τ_ℓ 替换为

$$\tilde{\tau}_\ell = \inf\left\{ n : \max_{n-L_\beta\leq k\leq n} \min_{0\leq m\leq M, m\neq \ell} \sum_{i=k}^{n} \log \frac{g_i^{(\ell)}(X_i)}{g_i^{(m)}(X_i)} \geq A \right\} \tag{37}$$

对于适当选取的窗口 L_β （将在下文的定理中说明），即可得到上述算法的窗口受限版本。

对于 $1 \leq \ell \leq M$ 和 $0 \leq m \leq M,\ \ m \neq \ell$，定义

$$I_{\ell m} = \frac{1}{T} \sum_{i=1}^{T} D(g_i^{(\ell)} \parallel g_i^{(m)}), \tag{38}$$

and

$$I^* = \min_{1\leq \ell\leq M} \min_{0\leq m\leq M, m\neq \ell} I_{\ell m}. \tag{39}$$

回忆一下，我们要寻找满足以下条件的 $(\tau,\ \delta)$:

$$\mathsf{E}_\infty[\tau] \geq \beta(1 + o(1)),\ \text{as } \beta \to \infty, \tag{40}$$

and

$$\begin{gathered} \mathsf{P}_1^{(\ell)}[\tau < \infty, \delta \neq \ell] \leq a_\beta\, \mathsf{E}_1^{(\ell)}[\tau], \quad \ell = 1, 2, \ldots, M, \\ \text{where } \log a_\beta^{-1} \sim \log \beta, \ \text{ as } \beta \to \infty. \end{gathered} \tag{41}$$

Let

$$C_\beta = \{(\tau, \delta) : \text{conditions in (40) and (41) hold}\}. \tag{42}$$

定理 3.2。 设信息数 I^* 如（39）中定义，并满足 $0 < I^* < \infty$。那么，当 $A =$ 对数 $4M\beta$ 时，

$$(\tau_{dc}, \delta_{dc}) \in C_\beta.$$

此外，

$$\begin{aligned}
&\max_{1\le\ell\le M}\sup_{\nu\ge 1} \operatorname{ess\,sup} \mathsf{E}_\nu^{(\ell)}[(\tau_{dc}-\nu+1)^+|X_1,\cdots,X_{\nu-1}] \\
&\sim \inf_{(\tau,\delta)\in C_\beta}\max_{1\le\ell\le M}\sup_{\nu\ge 1} \operatorname{ess\,sup} \mathsf{E}_\nu^{(\ell)}[(\tau-\nu+1)^+|X_1,\cdots,X_{\nu-1}] \\
&\sim \frac{\log\beta}{I^*},\quad as\ \beta\to\infty.
\end{aligned} \tag{43}$$

最后，只要满足以下条件，检验 (37) 的窗口受限版本也具有相同的渐近最优性：

$$\liminf_{\beta\to\infty}\frac{L_\beta}{\log\beta} > \frac{1}{I^*}.$$

例如，以下条件满足

$$L_\beta = \frac{\log\beta}{I^*}(1+\epsilon)$$

对任意固定的 $\epsilon>0$ 均成立。

证明。对于 $1\le\ell\le M$ 和 $0\le m\le M,\ m\ne\ell$，定义

$$Z_i(\ell,m)=\log\frac{g_i^{(\ell)}(X_i)}{g_i^{(m)}(X_i)}$$

为时刻 i 的对数似然比。在证明的其余部分，为了使公式简洁，我们用 $X_1^{\nu-1}$ 表示向量

$$X_1^{\nu-1}=(X_1,X_2,\ldots,X_{\nu-1}).$$

对于每个 $1\le\ell\le M$ 和 $0\le m\le M,\ m\ne\ell$，我们首先证明序列$\{Z_i(\ell,m)\}$ 满足以下陈述：

$$\begin{aligned}
\sup_{\nu\ge1}\operatorname{ess\,sup}\mathsf{P}_\nu^{(\ell)}\left(\max_{t\le n}\sum_{i=\nu}^{\nu+t}Z_i(\ell,m)\ge I_{\ell m}(1+\delta)n\ \middle|\ X_1^{\nu-1}\right) \\
\to 0,\ \text{as } n\to\infty,\quad \forall\delta>0,
\end{aligned} \tag{44}$$

其中 $I_{\ell m}$ 如式 (38) 所定义。

为证明式 (44)，注意到当 $n\to\infty$

$$\frac{1}{n}\sum_{i=\nu}^{\nu+n}Z_i(\ell,m)\to I_{\ell m},\quad \text{a.s. } \mathsf{P}_\nu^{(\ell)},\ \forall\nu\ge1. \tag{45}$$

上述显示成立是因为观测过程具有独立非同分布的性质。这

意味着，当 $n \to \infty$

$$\max_{t \le n} \frac{1}{n} \sum_{i=\nu}^{\nu+t} Z_i(\ell, m) \to I_{\ell m}, \quad \text{a.s. } \mathsf{P}_\nu^{(\ell)}, \ \forall \nu \ge 1. \tag{46}$$

为证明这一点，注意到

$$\max_{t \le n} \frac{1}{n} \sum_{i=\nu}^{\nu+t} Z_i(\ell, m) = \max \left\{ \max_{t \le n-1} \frac{1}{n} \sum_{i=\nu}^{\nu+t} Z_i(\ell, m), \ \frac{1}{n} \sum_{i=\nu}^{\nu+n} Z_i(\ell, m) \right\}. \tag{47}$$

对于固定的 $\epsilon > 0$，由 (45) 可知，当 n 足够大时，(46) 中的左侧大于 $I_{\ell m}(1-\epsilon)$。此外，设左侧的最大值在点 k_n 处取得，则

$$\max_{t \le n} \frac{1}{n} \sum_{i=\nu}^{\nu+t} Z_i(\ell, m) = \frac{1}{n} \sum_{i=\nu}^{\nu+k_n} Z_i(\ell, m) = \frac{k_n}{n} \frac{1}{k_n} \sum_{i=\nu}^{\nu+k_n} Z_i(\ell, m).$$

现在， k_n 不可能有界，因为上述等式中的左侧以下界 $I_{\ell m}(1-\epsilon)$ 为界，而且上述等式右侧的分母中含有 n。这意味着，对于任意固定的 i，都有 $k_n > i$，且 $k_n \to \infty$。因此，$\frac{1}{k_n} \sum_{i=\nu}^{\nu+k_n} Z_i(\ell, m) \to I_{\ell m}$。由于 $k_n/n \le 1$，当 n 足够大时，(46) 中的左侧小于 $I_{\ell m}(1+\epsilon)$。这就证明了 (46)。为证明 (44)，注意到由于这些过程具有 i.p.i.d. 性质

$$\begin{aligned} &\sup_{\nu \ge 1} \operatorname{ess\,sup} \mathsf{P}_\nu^{(\ell)} \left(\max_{t \le n} \sum_{i=\nu}^{\nu+t} Z_i(\ell, m) \ge I_{\ell m}(1+\delta) n \,\middle|\, X_1^{\nu-1} \right) \\ &\quad = \sup_{1 \le \nu \le T} \mathsf{P}_\nu^{(\ell)} \left(\max_{t \le n} \sum_{i=\nu}^{\nu+t} Z_i(\ell, m) \ge I_{\ell m}(1+\delta) n \right). \end{aligned} \tag{48}$$

由于式（46）的结论，以及式（48）右侧的最大值仅涉及有限多个项，因此右侧趋于零。

接下来，我们证明序列 $\{Z_i(\ell, m)\}$ 对每个 $1 \le \ell \le M$ 和 $0 \le m \le M$, $m \ne \ell$ 都满足以下结论:

$$\begin{aligned} \lim_{n \to \infty} \sup_{k \ge \nu \ge 1} \operatorname{ess\,sup} \mathsf{P}_\nu^{(\ell)} \left(\frac{1}{n} \sum_{i=k}^{k+n} Z_i(\ell, m) \le I_{\ell m} - \delta \,\middle|\, X_1^{k-1} \right) \\ = 0, \quad \forall \delta > 0. \end{aligned} \tag{49}$$

为证明 (49)，注意到由于该过程具有独立非同分布特性，我们有

$$\begin{aligned}
&\sup_{k\geq\nu\geq 1} \operatorname{ess\,sup}\ \mathsf{P}_{\nu}^{(\ell)}\left(\frac{1}{n}\sum_{i=k}^{k+n} Z_i(\ell,m) \leq I_{\ell m}-\delta \,\middle|\, X_1^{k-1}\right)\\
&\quad= \sup_{k\geq\nu\geq 1} \mathsf{P}_{\nu}^{(\ell)}\left(\frac{1}{n}\sum_{i=k}^{k+n} Z_i(\ell,m) \leq I_{\ell m}-\delta\right)\\
&\quad= \sup_{\nu+T\geq k\geq\nu\geq 1} \mathsf{P}_{\nu}^{(\ell)}\left(\frac{1}{n}\sum_{i=k}^{k+n} Z_i(\ell,m) \leq I_{\ell m}-\delta\right)\\
&\quad= \max_{1\leq\nu\leq T}\ \max_{\nu\leq k\leq\nu+T} \mathsf{P}_{\nu}^{(\ell)}\left(\frac{1}{n}\sum_{i=k}^{k+n} Z_i(\ell,m) \leq I_{\ell m}-\delta\right).
\end{aligned} \tag{50}$$

由于式 (45) 且最大化次数有限，上述等式右侧在 $n\to\infty$ 时趋于零，对任意 δ 均成立。由于满足 [17] 中的条件 A1 和 A2，结论由 [17] 的定理 4 得出。

□

注 1. 该算法及其最优性也很容易推广到多流数据的情形：此时观测流的数量有限，且变化后只有一个流受到影响。目标是检测变化，同时以较低的概率识别受影响的流。

4. 多时隙最快变化检测

设 $(f_1,\cdots,f_T)$ 和 $(g_1,\cdots,g_T)$ 分别表示两个独立非同分布过程的概率分布，其中 $f_i\neq g_i, \forall i$。我们假设对数似然比的二阶矩有限且为正：

$$\begin{aligned}
&\text{(M1)}\quad 0<\mathsf{E}_1\left(\left|\log\frac{g_i(X_i)}{f_i(X_i)}\right|\right)<\infty,\ i=1,2,\cdots,T\\
&\text{(V1)}\quad 0<\mathsf{E}_1\left(\log\frac{g_i(X_i)}{f_i(X_i)}\right)^2<\infty,\ i=1,2,\cdots,T.
\end{aligned} \tag{51}$$

此处 E_1 表示变化在时刻 $\nu=1$发生时的期望。

在多槽变点检测问题中，每个周期内仅有 T 时间槽的一个子集发生变化。为刻画这一点，我们现在为变化后分布引入一种新记号，以强调密度发生变化的时隙。对于一个子集$S\subset\{1,2,\ldots,T\}$，将一种可能的变化后独立非同分布分布律定义为

$$g_S=(g_{S,1},g_{S,2},\ldots,g_{S,T}), \tag{52}$$

with

$$g_{S,i}=\begin{cases} g_i, & \text{if } i\in S\\ f_i, & \text{if } i\notin S.\end{cases} \tag{53}$$

注意，

$$g_S = (g_1, \cdots, g_T), \text{ if } S = \{1, \ldots, T\}.$$

因此，集合 S 表示发生变化的时隙：

$$X_n \sim \begin{cases} f_n, & \forall n < \nu \\ g_{S,n} & \forall n \geq \nu, \end{cases} \tag{54}$$

其中 $g_{S,\ n+T} = g_{S,\ n}$, $\forall n$。决策者并不知道这个集合。不过，已知

$$S \in \mathcal{S} \subset 2^{\{1,\ldots,T\}},$$

即， S 属于 $\{1, \ldots, T\}$ 的幂集的子集所构成的族。例如，

$$\mathcal{S} = \{S : |S| \leq m\},$$

其中 $|S|$ 表示集合 S 的大小。我们将提出的变化检测算法在 $m \ll T$ 时尤其有用。

设 τ 是过程 $\{X_n\}$ 的一个停止时刻，即一个取正整数值的随机变量，并且事件 $\{\tau \leq n\}$ 属于由 $\{X_1, \cdots, X_n\}$ 生成的 σ- 代数。我们将变化点 ν 建模为具有先验分布 π 的随机变量：

$$\pi_n = \mathsf{P}(\nu = n), \quad \text{for } n = 1, 2, \cdots.$$

令 P_n^S 表示变化在时刻 $\nu = n$ 发生于时隙 S 时观测过程 $\{X_n\}$ 的分布，并定义

$$\mathsf{P}_\pi^S = \sum_{n=1}^{\infty} \pi_n \, \mathsf{P}_n^S.$$

我们用 E_n^S 和 E_π^S 表示相应的期望。对于每个 $S \in \mathcal{S}$，我们寻求以下问题的解：

$$\min_{\tau \in \mathbf{C}_\alpha} \mathsf{E}_\pi^S \left[\tau - \nu | \tau \geq \nu\right], \tag{55}$$

其中

$$\mathbf{C}_\alpha = \{\tau : \mathsf{P}_\pi^S(\tau < \nu) \leq \alpha\}, \tag{56}$$

且 α 是对虚警概率给定的约束。事实上，我们要寻找一种算法，能够对每个 S统一解决上述问题。

4.1. 多槽变点检测算法

考虑集合 $\mathcal{S}$ 上的一个混合分布或概率质量函数:

$$p_S \geq 0,\ \forall S \in \mathcal{S}, \quad \text{and} \quad \sum_{S\in\mathcal{S}} p_S = 1.$$

定义混合统计量

$$R_n = \frac{1}{\Pi_n}\sum_{k=1}^{n}\pi_k \sum_{S\in\mathcal{S}} p_S \prod_{i=k}^{n}\frac{g_{S,i}(X_i)}{f_i(X_i)}, \tag{57}$$

其中 $\Pi_n = \mathsf{P}(\nu > n)$，停止规则为

$$\tau_{mps} = \inf\{n \geq 1 : R_n > A\}. \tag{58}$$

下文将该算法称为混合周期性 Shiryaev 算法或 MPS 算法。注意,

$$R_n = \sum_{S\in\mathcal{S}} p_S \frac{1}{\Pi_n}\sum_{k=1}^{n}\pi_k \prod_{i=k}^{n}\frac{g_{S,i}(X_i)}{f_i(X_i)}.$$

因此，统计量 R_n 是 $|\mathcal{S}|$ 周期性 Shiryaev 统计量的混合（参见 [5]和第 2.2 节），每个 $S \in \mathcal{S}$对应一个统计量。因此，对于几何先验 $\pi = \text{Geom}(\rho)$，统计量 R_n 可以递归计算，并使用有限记忆（参见 [5]中的引理 5.1）。

4.2. 检测延迟下界

本节将针对满足虚警概率约束（56）的任意停止时刻，给出平均检测延迟的下界。我们作出以下假设。

（A1）设存在 $d \geq 0$，使得

$$\lim_{n\to\infty}\frac{\log \mathsf{P}(\nu > n)}{n} = -d. \tag{59}$$

（A2）此外，设

$$\sum_{n=1}^{\infty}\pi_n|\log\pi_n| < \infty. \tag{60}$$

如果 $\pi =$几何分布(ρ)，那么

$$\frac{\log \mathsf{P}(\nu > n)}{n} = \frac{\log(1-\rho)^n}{n} = \frac{n\log(1-\rho)}{n} = \log(1-\rho).$$

因此， $d = |$对数$(1-\rho)|$。此外，

$$\sum_{n=1}^{\infty} \pi_n |\log \pi_n| = \frac{1-\rho}{\rho} \log \frac{1}{(1-\rho)} + \log \frac{1}{\rho} < \infty.$$

因此，上述条件 (A1) 和 (A2) 均由几何先验满足。

我们先从一个证明很简单的引理开始。定义

$$Z_i = \log \frac{g_{S,i}(X_i)}{f_i(X_i)}, \tag{61}$$

and

$$I_S = \frac{1}{T} \sum_{i \in S} D(g_{S,i} \parallel f_i). \tag{62}$$

引理 4.1。对于(61) 中定义的 Z_i 和(62)中定义的 I_S，对每个 $S \in \mathcal{S}$ 且当$n \to \infty$时，我们有

$$\frac{1}{n} \sum_{i=k}^{k+n-1} Z_i \;\; \to \;\; I_S, \quad \mathsf{P}_k^S \;\; a.s., \; \forall k \geq 1. \tag{63}$$

下界由以下定理给出。

定理 4.2。设信息数 I_S 如(62)中所定义。此外，设先验分布 π 满足(59)中的条件 ***(A1)***。那么，对于任意停止时刻 $\tau \in \mathbf{C}\alpha$，我们有

$$\mathsf{E}_\pi^S [\tau - \nu | \tau \geq \nu] \geq \frac{|\log \alpha|}{I_S + d}(1 + o(1)), \qquad as \; \alpha \to 0. \tag{64}$$

这里 $o(1) \to 0$ 当 $\alpha \to 0$ 时。

证明。该结果由上述引理 4.1 和定理 5.1 得出 [5] （另见第 2.2 节）。

□

4.3. MPS 算法的最优性

现在我们证明，对于每个变化后时隙 $S \in \mathcal{S}$，当虚警约束 $\alpha \to 0$ 时，MPS 算法 (58) 对问题 (55) 是渐近最优的。我们先证明一个重要引理。

定义

$$\gamma_k(\epsilon) = \sum_{n=1}^{\infty} \mathsf{P}_k^S \left(\left| \frac{1}{n} \sum_{i=k}^{k+n-1} Z_i - I_S \right| > \epsilon \right), \tag{65}$$

其中 Z_i 如式（61）所定义。

引理 4.3。对于每个 $\epsilon > 0$,

$$\sum_{k=1}^{\infty} \pi_k \gamma_k(\epsilon) < \infty, \tag{66}$$

其中， $\gamma_k(\epsilon)$ 在(65)中定义。

证明。 序列

$$\{\gamma_k(\epsilon)\}_{k=1}^{\infty}$$

(65) 中的序列以 T 为周期，因此上述序列中至多有 T 个不同值：

$$\gamma_1(\epsilon), \cdots, \gamma_T(\epsilon).$$

因此，如果我们证明这 T 个数对于任意 $\epsilon > 0$ 都是有限的，就可以自动得到

$$\sum_{k=1}^{\infty} \pi_k \gamma_k(\epsilon) \leq \max_{1 \leq k \leq T} \gamma_k(\epsilon) < \infty, \quad \forall \epsilon > 0.$$

此外，由于我们没有对密度 (f_1， $\cdots$， f_T) 和 (g_1， $\cdots$， g_T) 所取的实际值作出任何明确假设，因此可以利用这些过程的 i.p.i.d. 特性，只需证明

$$\gamma_1(\epsilon) < \infty, \quad \forall \epsilon > 0.$$

回顾 $\gamma_1(\epsilon)$ 的定义：

$$\gamma_1(\epsilon) = \sum_{n=1}^{\infty} \mathsf{P}_1^S \left(\left| \frac{1}{n} \sum_{i=1}^{n} Z_i - I_S \right| > \epsilon \right).$$

对于 $\ell = 1$， 2， $\cdots$， T，定义

$$Z_i^{(\ell)} = \begin{cases} Z_i & \text{if } i = mT + \ell \text{ for } m = 0, 1, 2, \cdots \\ 0 & \text{otherwise .} \end{cases}$$

注意

$$Z_i^{(\ell)} = 0, \text{ if } \ell \notin S.$$

另外请记住,

$$I_S = \frac{1}{T} \sum_{\ell \in S} I_\ell,$$

其中

$$I_\ell = D(g_\ell \parallel f_\ell).$$

利用这些定义，我们可以写出

$$\frac{1}{n}\sum_{i=1}^{n} Z_i - I_S = \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell\in S} Z_i^{(\ell)} - \frac{1}{T}\sum_{\ell\in S} I_\ell = \sum_{\ell\in S}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i^{(\ell)} - \frac{I_\ell}{T}\right).$$

这意味着

$$\mathsf{P}_1^S\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - I_S\right| > \epsilon\right) \le \sum_{\ell\in S}\mathsf{P}_1^S\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i^{(\ell)} - \frac{I_\ell}{T}\right| > \frac{\epsilon}{|S|}\right).$$

因此，要证明上述方程中左侧的可求和性，我们需要证明右侧每个 $|S|$ 项都可求和。同样，由于这些过程具有 i.p.i.d. 性质，而且我们没有对密度（f_1，$\cdots$，f_T）和（g_1，$\cdots$，g_T）作出任何明确假设，因此只需证明右侧任意一项可求和即可。也就是说，对于 $\ell \in S$，我们要证明

$$\sum_{n=1}^{\infty}\mathsf{P}_1^S\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i^{(\ell)} - \frac{I_\ell}{T}\right| > \frac{\epsilon}{|S|}\right) < \infty.$$

定义 $\ell = 1$，2，$\cdots$，T，

$$I_i^{(\ell)} = \begin{cases} I_\ell & \text{if } i = mT + \ell \text{ for } m = 0, 1, 2, \cdots, \ell \in S \\ 0 & \text{otherwise.} \end{cases}.$$

利用这个定义，对于 $\ell \in S$，我们写作

$$\frac{1}{n}\sum_{i=1}^{n} Z_i^{(\ell)} - \frac{I_\ell}{T} = \frac{1}{n}\sum_{i=1}^{n}(Z_i^{(\ell)} - I_i^{(\ell)}) + \frac{1}{n}\sum_{i=1}^{n} I_i^{(\ell)} - \frac{I_\ell}{T}.$$

因此，对于 $\tilde{Z}_i^{(\ell)} = Z_i^{(\ell)} - I_i^{(\ell)}$，我们有

$$\begin{aligned}\mathsf{P}_1^S\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i^{(\ell)} - \frac{I_\ell}{T}\right| > \frac{\epsilon}{|S|}\right) \le{}& \mathsf{P}_1^S\left(\left|\frac{1}{n}\sum_{i=1}^{n}\tilde{Z}_i^{(\ell)}\right| > \frac{\epsilon}{2|S|}\right) \\ &+ \mathsf{P}_1^S\left(\left|\frac{1}{n}\sum_{i=1}^{n} I_i^{(\ell)} - \frac{I_\ell}{T}\right| > \frac{\epsilon}{2|S|}\right).\end{aligned} \tag{67}$$

现在，$-\frac{1}{n}\sum_{i=1}^{n} I_i^{(\ell)} = \frac{1}{n} I_\ell \lfloor \frac{n}{T} \rfloor \to \frac{I_\ell}{T}$，由于 $k \to \infty$。因此，当 n 足够大时，(67) 右侧的第二项恒为零。因此，我们只需证明

$$\sum_{n=1}^{\infty}\mathsf{P}_1^S\left(\left|\frac{1}{n}\sum_{i=1}^{n}\tilde{Z}_i^{(\ell)}\right| > \frac{\epsilon}{2|S|}\right) < \infty, \quad \ell \in S.$$

为此，注意在项 $\mathsf{P}_1^S\left(\left|\frac{1}{n}\sum_{i=1}^n \tilde{Z}_i^{(\ell)}\right| > \frac{\epsilon}{2|S|}\right)$ 中，和式$\sum_{i=1}^n \tilde{Z}_i^{(\ell)}$ 每隔 T 个时间步才更新一次。因此，作为 n 的函数，该概率在 $kT+\bar{\ell}$ 和 $(k+1)T+\overline{\ell}-1$ 之间单调递减，对每个$k=0$，1，2，.....均如此。利用这一事实，我们可以写出

$$\begin{aligned}\sum_{n=1}^{\infty}\mathsf{P}_1^S\left(\left|\frac{1}{n}\sum_{i=1}^{n}\tilde{Z}_i^{(\ell)}\right|>\frac{\epsilon}{2|S|}\right) &\le T+T\sum_{j=1}^{\infty}\mathsf{P}_1^S\left(\left|\frac{1}{jT+\ell}\sum_{i=1}^{jT+\ell}\tilde{Z}_i^{(\ell)}\right|>\frac{\epsilon}{2|S|}\right)\\ &\le T+T\sum_{j=1}^{\infty}\mathsf{P}_1^S\left(\left|\frac{1}{j}\sum_{i=1}^{jT+\ell}\tilde{Z}_i^{(\ell)}\right|>\frac{\epsilon}{2|S|}\right).\end{aligned}$$

最右侧的求和是有限的，因为其中的求和是由一组分布与 Z_ℓ 在 P_1 [25]下相同的 j 独立同分布随机变量构成的。对于方差有限的独立同分布随机变量，该求和是有限的。另见 [24]。□

换言之，上述引理指出，i.p.i.d. 过程满足完全收敛条件 [25], [24]。

定理 4.4。 设先验分布满足条件 *(A1)* 和 *(A2)*。对于 (58) 中的 $A=\frac{1-\alpha}{\alpha}$，对每个 $S\in\mathcal{S}$,

$$\mathsf{P}_\pi^S(\tau_{mps}<\nu)\le\alpha$$

and

$$\mathsf{E}_\pi^S\left[\tau_{mps}-\nu|\tau_{mps}\ge\nu\right]\le\frac{|\log\alpha|}{I_S+d}(1+o(1)),\qquad as\ \alpha\to 0. \tag{68}$$

证明。 结论可直接由引理 4.3 及 [24]中给出的论证得出。不过，为完整起见，我们仍详细给出证明。

回顾一下，混合统计量定义为

$$R_n=\frac{1}{\Pi_n}\sum_{k=1}^{n}\pi_k\sum_{S\in\mathcal{S}}p_S\prod_{i=k}^{n}\frac{g_{S,i}(X_i)}{f_i(X_i)}, \tag{69}$$

其中 Π_n =P($\nu>n$)，停止规则定义为

$$\tau_{mps}=\inf\{n\ge 1: R_n>A\}. \tag{70}$$

我们首先注意到，对于任意离散整值随机变量（例如停止时刻） τ,

$$\mathsf{E}[\tau]=\sum_{n=0}^{\infty}\mathsf{P}(\tau>n)\le N+\sum_{n=N}^{\infty}\mathsf{P}(\tau>n), \tag{71}$$

其中 N 是任意正整数。通过谨慎选择上述N 的值，并对 P($\tau>n$) 求得上界，即可得到该定理。

对于 $0 < \epsilon < I_S + d$，令

$$N = N_\alpha = 1 + \left\lfloor \frac{\log(A_\alpha/\pi_k)}{I_S + d - \epsilon} \right\rfloor,$$

其中

$$A_\alpha = \frac{1-\alpha}{\alpha}.$$

在式 (71) 中使用 $N = N_\alpha$ 和 $\tau = (\tau_{mps} - k)^+$，可得

$$\begin{aligned} \mathsf{E}_k^S[(\tau_{mps} - k)^+] &\le N_\alpha + \sum_{n \ge N_\alpha} \mathsf{P}_k^S(\tau_{mps} > k + n) \le N_\alpha + \sum_{n \ge N_\alpha} \mathsf{P}_k^S(R_{k+n} < A_\alpha) \\ &= N_\alpha + \sum_{n \ge N_\alpha} \mathsf{P}_k^S(\log R_{k+n} < \log A_\alpha). \end{aligned} \tag{72}$$

Now,

$$R_{k+n} = \frac{1}{\Pi_{k+n}} \sum_{t=1}^{k+n} \pi_t \sum_{S \in \mathcal{S}} p_S \prod_{i=t}^{k+n} \frac{g_{S,i}(X_i)}{f_i(X_i)}.$$

MPS统计量的下界为

$$R_{k+n} \ge \frac{1}{\Pi_{k+n}} \pi_k \; p_S \prod_{i=k}^{k+n} \frac{g_{S,i}(X_i)}{f_i(X_i)}.$$

这里，右侧的 S 是真实的变化后多图集。对两边取对数，可得

$$\log(R_{k+n}) \ge |\log \Pi_{k+n}| + \log(\pi_k) + \log(p_S) + \sum_{i=k}^{k+n} Z_i. \tag{73}$$

利用（73），我们可以对 $n \ge N_\alpha$ 时（72）中的概率进行界定，

$$\begin{aligned} \mathsf{P}_k^S(\log R_{k+n} < \log A_\alpha) &\le \mathsf{P}_k^S\left(|\log \Pi_{k+n}| + \log(p_S) + \sum_{i=k}^{k+n} Z_i < \log(A_\alpha/\pi_k) \right) \\ &= \mathsf{P}_k^S\left(\frac{1}{n} \sum_{i=k}^{k+n} Z_i + \frac{|\log \Pi_{k+n}|}{n} + \frac{\log(p_S)}{n} < \frac{\log(A_\alpha/\pi_k)}{n} \right) \\ &\le \mathsf{P}_k^S\left(\frac{1}{n} \sum_{i=k}^{k+n} Z_i + \frac{|\log \Pi_{k+n}|}{n} + \frac{\log(p_S)}{n} < I_S + d - \epsilon \right) \\ &= \mathsf{P}_k^S\left(\frac{1}{n} \sum_{i=k}^{k+n} Z_i < I_S + d - \frac{|\log \Pi_{k+n}|}{n} - \frac{\log(p_S)}{n} - \epsilon \right). \end{aligned} \tag{74}$$

现在选择足够小的 α，使得对于每个 $n \geq N_\alpha$

$$\left| d - \frac{|\log \Pi_{k+n}|}{n} \right| < \frac{\epsilon}{4},$$
$$\left| \frac{\log(p_S)}{n} \right| < \frac{\epsilon}{4}.$$

具体而言，选取足够小的 α，使得上面显示的陈述对所有满足以下条件的 n 都成立

$$n \geq 1 + \left\lfloor \frac{\log(A_\alpha)}{I_S + d - \epsilon} \right\rfloor.$$

这确保所选取的足够小的 α 不依赖于指标 k。由此可得

$$\begin{aligned} &\mathsf{P}_k^S(\log R_{k+n} < \log A_\alpha) \\ &\leq \mathsf{P}_k^S \left(\frac{1}{n} \sum_{i=k}^{k+n} Z_i < I_S + d - \frac{|\log \Pi_{k+n}|}{n} - \frac{\log(p_S)}{n} - \epsilon \right) \\ &\leq \mathsf{P}_k^S \left(\frac{1}{n} \sum_{i=k}^{k+n} Z_i < I_S - \frac{\epsilon}{2} \right). \end{aligned} \tag{75}$$

将其代入式（72），当 α 足够小时，对 k一致地有:

$$\begin{aligned} \mathsf{E}_k^S[(\tau_{mps} - k)^+] &\leq N_\alpha + \sum_{n \geq N_\alpha} \mathsf{P}_k^S(\log R_{k+n} < \log A_\alpha) \\ &\leq N_\alpha + \sum_{n \geq N_\alpha} \mathsf{P}_k^S \left(\frac{1}{n} \sum_{i=k}^{k+n} Z_i < I_S - \frac{\epsilon}{2} \right) \\ &\leq N_\alpha + \sum_{n=1}^{\infty} \mathsf{P}_k^S \left(\left| \frac{1}{n} \sum_{i=k}^{k+n} Z_i - I_S \right| > \frac{\epsilon}{2} \right) \\ &= N_\alpha + \gamma_k(\epsilon/2). \end{aligned} \tag{76}$$

由此可得

$$\begin{aligned} \mathsf{E}_\pi^S[(\tau_{mps} - \nu)^+] &= \sum_{k=1}^{\infty} \pi_k \mathsf{E}_k^S[(\tau_{mps} - k)^+] \\ &\leq \sum_{k=1}^{\infty} \pi_k N_\alpha + \sum_{k=1}^{\infty} \pi_k \gamma_k(\epsilon/2) \\ &\leq \sum_{k=1}^{\infty} \pi_k \left(1 + \frac{\log(A_\alpha/\pi_k)}{I_S + d - \epsilon} \right) + \sum_{k=1}^{\infty} \pi_k \gamma_k(\epsilon/2) \\ &= \frac{\log(A_\alpha)}{I_S + d - \epsilon} + \text{ constant}. \end{aligned} \tag{77}$$

该结果剩余项是一个常数（不是 α 的函数），这是由所作的假设决定的

这是在定理陈述中作出的假设所致，也是引理 4.3 的结果。最后，

$$\begin{aligned}\mathsf{E}_\pi^S[\tau_{mps}-\nu|\tau_{mps}\geq\nu] &= \frac{\mathsf{E}_\pi^S[(\tau_{mps}-\nu)^+]}{\mathsf{P}_\pi^S(\tau_{mps}\geq\nu)} \\ &\leq \frac{\frac{\log(A_\alpha)}{I_S+d-\epsilon}+\text{ constant}}{1-\alpha} \\ &= \frac{|\log\alpha|}{I_S+d-\epsilon}(1+o(1)), \quad \text{ as } \alpha\to 0.\end{aligned} \tag{78}$$

结果现已得出，因为 ϵ 可以变得任意小。

将 $A=A_\alpha$ 设定为可确保满足虚警约束这一事实，可由 [24]推出。

□

因此，MPS 算法达到了定理 4.2 中给出的渐近下界，并且在 S上是一致渐近最优的。

5. 数值结果

5.1. 将周期性 *CUSUM* 算法应用于洛杉矶交通数据

本节将展示如何在洛杉矶（LA）高速公路的交通流指标数据上训练并应用周期性 CUSUM 算法（31）。为便于参考，现将该算法重述如下。

$$W_{n+1}=W_n^+ + \log\frac{g_{n+1}^{(1)}(X_{n+1})}{g_{n+1}^{(0)}(X_{n+1})};\ \ W_n=0. \tag{79}$$

$$\tau_c=\inf\{n\geq 1: W_n\geq A\}. \tag{80}$$

我们使用工作日交通数据训练变化前模型，并使用周末或节假日数据训练变化后模型。随后，我们将算法应用于交通数据，以检测周末或节假日交通流。下面将详细讨论其应用。

我们通过纳入多个历史交通流数据集，将算法应用于洛杉矶县 I-10 西向高速公路沿线选定的站点（见图 1）。这些数据集在 2020年8月—2021年9月期间可通过交通性能测量系统（PeMS 网站 https://pems.dot.ca.gov/）免费获取。PeMS 报告的交通流量计数可作为每个站点交通流的代理指标；如图 1 所示，这十个站点之间的间距约为 0.33 英里。PeMS 以存档格式每日发布数据。

下载的洛杉矶某个站点的数据采用逗号分隔的文本格式，包含多个字段。每个字段都包含有用的交通属性，例如时间戳、站点标识符、每个 5 分钟时间段的车辆数等（样例数据见图 2）。此外，我们还能够使用站点编号（图 2 中的第二列）作为 PeMS 中的键，在洛杉矶地图上精确定位每个站点，并将这些

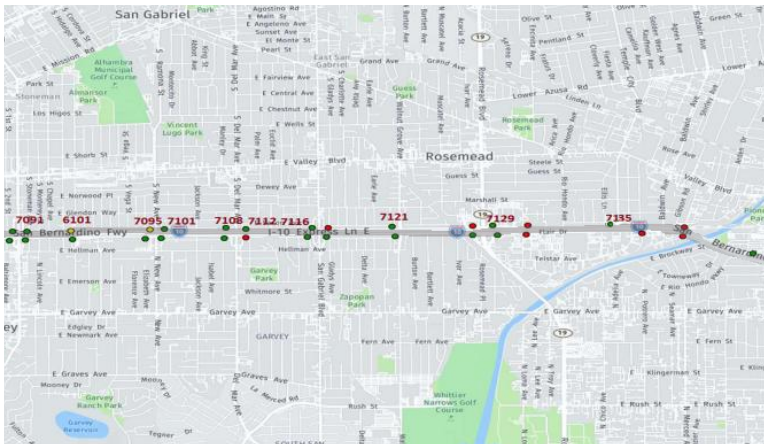


图 1：所选站点地图。西向交通的站点编号以数字标示：图中展示了主要特征，浅灰色区域突出显示了 I-10 公路（克里斯托弗·哥伦布横贯大陆公路）位于路易斯安那州、加利福尼亚州的路段。

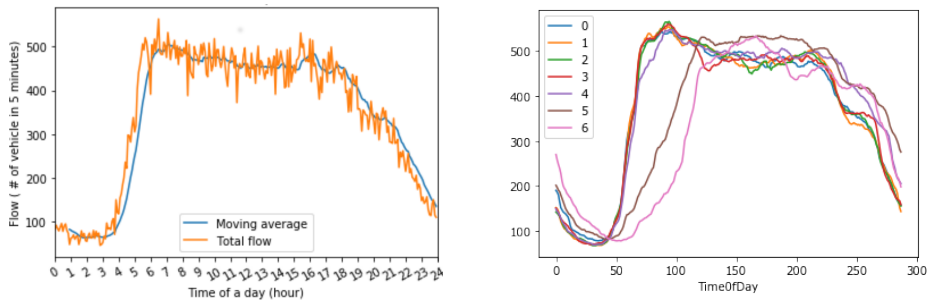
	date	station_id	district_id	FreeWay	dir_travel	lane_type	station_len	sample_cnt	percent_obsv	total_flow	avg_occp	avg_spd*
0	04/24/2021 00:00:00	768202	7	2	E	ML	0.335	40	100	40.0	0.0092	70.7
1	04/24/2021 00:05:00	768202	7	2	E	ML	0.335	40	100	31.0	0.0077	67.9
2	04/24/2021 00:10:00	768202	7	2	E	ML	0.335	36	100	32.0	0.0075	70.8
3	04/24/2021 00:15:00	768202	7	2	E	ML	0.335	40	100	26.0	0.0054	69.3
4	04/24/2021 00:20:00	768202	7	2	E	ML	0.335	40	100	34.0	0.0066	69.

* Other columns are disregarded that might include Lane N Samples, Lane N Flow, Lane N Avg Occ, Lane N Avg Speed, Lane N Observed

图 2：东向车辆检测站（VDS）的原始数据示例，其“站点编号”（第二列）以 8202 结尾。请注意，标记为“总流量”的列（第十列）定义为每五分钟的车辆数。 -

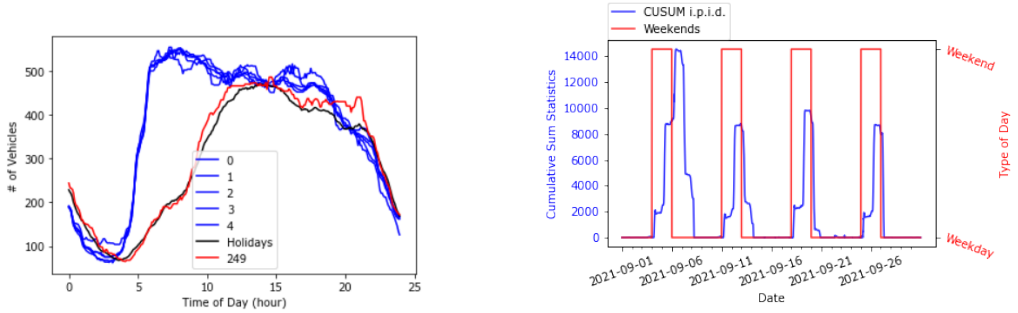
另附一张表，其中包含所选交通站点的地理坐标 [lat, lon]，如图1所示。

我们在一周中的不同日期，以及同一目标路段内的多个站点上，都观察到了相似的模式（图3）。我们选取了2021年8月某个随机工作日中 ID 以 7095 结尾的传感器交通流量数据用于训练，并将2021年9月的最后一个月留作测试集。图4展示了采用平滑技术得到的样本路径对比：测试集中2021年的劳动节假日数据，与2020年劳动节假日（标签249）及8月各周一的数据进行比较。由于读数存在噪声，我们采用了该站点前一小时样本的中位数移动平均（MMA）（为便于后续分析，我们删除了重复项，并在站点数据集中保留最后一条记录）。



(a) (b)

图3：(a)：索引为3（7095）的站点一天内的样本路径示意图（288个时间区间）。(b)：2021年8月不同日期（周一=0）的平均交通流量计数。



(a) (b)

图 4：(a) 正常时期（8 月的星期一）的部分训练样本路径与 2021 年劳动节及 2020年劳动节（标签 249）的比较。(b)：2021年9月6日劳动节周的周期性累积和（CUSUM）检验统计量和事件标签。

应用周期性累积和算法时，我们假设 $T = 288$ (12×24)（区间数量），并对其进行建模

$$\begin{aligned} g_i^{(0)} &= \text{Poisson}(\lambda_i^{(0)}), \quad i = 1, 2, \ldots, T \\ g_i^{(1)} &= \text{Poisson}(\lambda_i^{(1)}), \quad i = 1, 2, \ldots, T. \end{aligned} \tag{81}$$

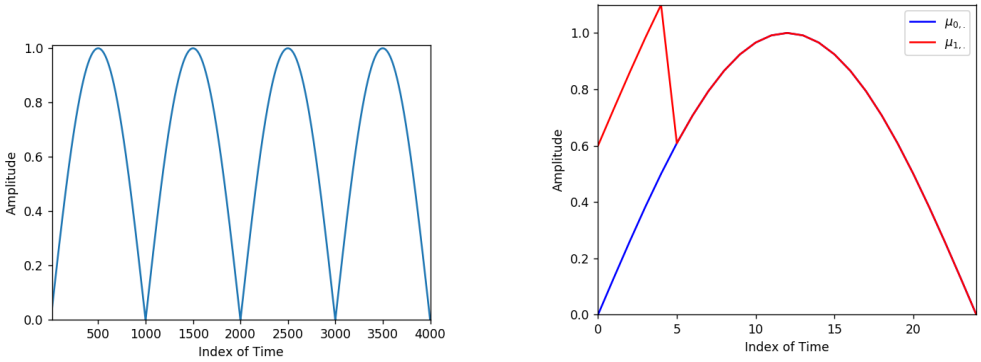


图 5：(a)：以 1 kHz 的频率采样的四个正半弦波示意图（每周期采样数）。(b)：变化前和实际变化后概率密度函数的参数示意图；实际变化后的概率密度函数在第一个时隙 ($[0,\ 4]$) 向上移动。

随后，我们根据训练数据学习泊松参数。在图 4b 中，我们绘制了测试数据的周期性 CUSUM 统计量。红色矩形块表示周末的位置，蓝色曲线表示检验统计量。如图所示，检验统计量在周末前后急剧上升，表明检测到了交通流的变化。

5.2. 多时隙最快变化检测的数值结果

本节将 MPS 算法 (57) 应用于模拟含噪正弦数据。为便于参考，我们在此重述该算法。

$$R_n = \frac{1}{\Pi_n} \sum_{k=1}^{n} \pi_k \sum_{S \in \mathcal{S}} p_S \prod_{i=k}^{n} \frac{g_{S,i}(X_i)}{f_i(X_i)}, \tag{82}$$

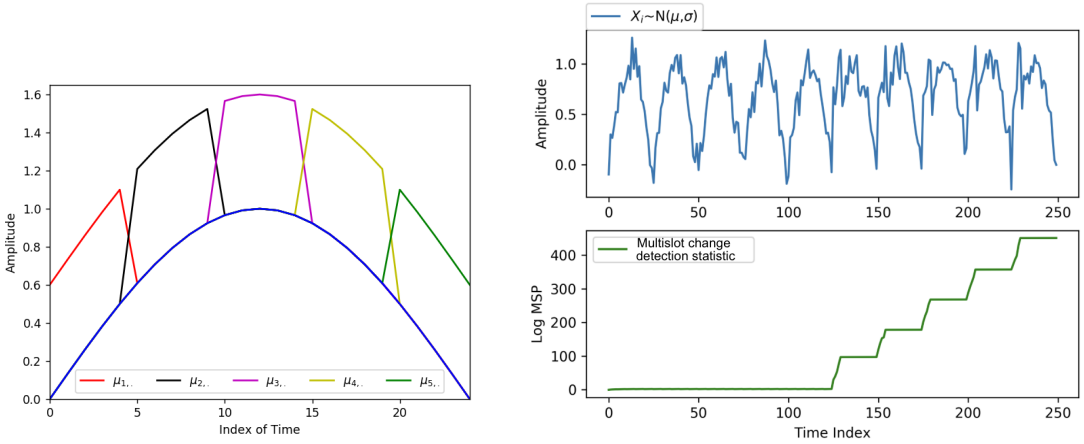
其中 $\Pi_n = \mathrm{P}(\nu > n)$，停止规则

$$\tau_{mps} = \inf\{n \geq 1 : R_n > A\}. \tag{83}$$

具体而言，我们假设观测到如图 5a 所示的一系列正弦波形的含噪版本。在变化点处，正弦信号的形状发生失真，如图 5b 所示。目标是实时检测这种失真。我们假设已知失真的类型，但不知道失真的确切位置。因此，实际失真可能是图 6a 所示的五种情况中的任意一种。设计 MPS 算法 (82) 时，我们假设每种失真出现的概率相同。

令 $h(t)$ 为图 5b 所示的蓝色正弦信号。然后，我们假设 $T = 25$并使用以下方式生成模拟数据：

$$f_i = \mathcal{N}\left(\mu_{0,i}, 0.01\right), \quad i = 1, 2, \ldots, 25,$$



(a) (b)

图6：(a)：所有可能的变化后波形示意图。(b)：变点位于时间索引125处（第五个周期结束时）时，高斯分布变化前/后的检验统计量和样本路径。

其中， μ_0 是一个长度为 25 的向量，定义为

$$\mu_{0,i} = h(i), \quad i = 1, 2, \ldots, 25.$$

变化后数据由以下方式生成

$$g_i = \mathcal{N}\left(\mu_{1,i}, 0.01\right), \quad i = 1, 2, \ldots, 25,$$

其中

$$\mu_{1,i} = \begin{cases} \mu_{0,i}, & \text{for i} \notin [0,4] := [0,1,2,3,4] \\ \mu_{0,i} + 0.6, & \text{for i} \in [0,4]. \end{cases}$$

因此，假设真实的变化后时隙为以下内容，从而生成变化后数据

$$S = [0, 4].$$

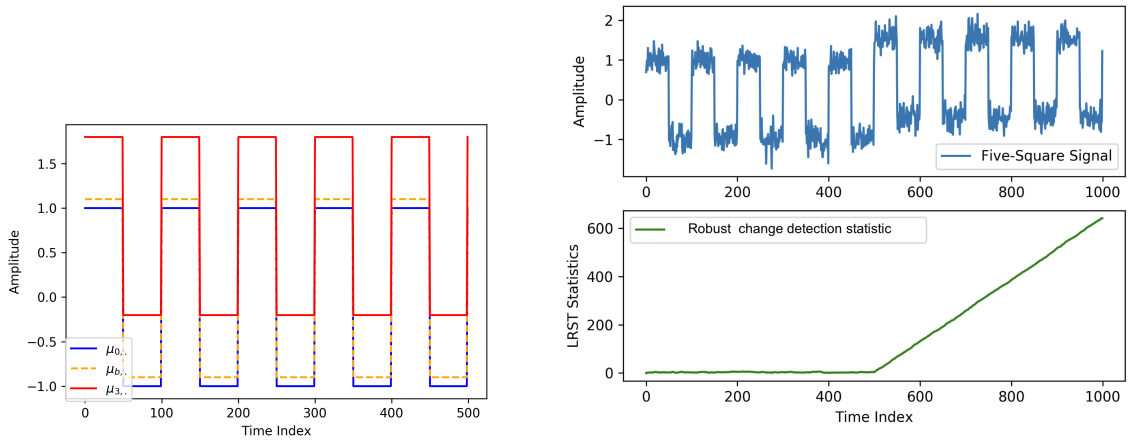
但是，我们假设多时隙族 $\mathcal{S}$ 为

$$\mathcal{S} = \{[0,4], [5,9], [10,14], [15,19], [20,24]\}$$

with

$$p_S = \frac{1}{5}, \quad \text{for all } S \in \mathcal{S}.$$

我们还假设变化点服从参数为 $\rho = 0.01$ 的几何先验，即 $\pi_k = (1-\rho)^{k-1}\rho$。我们在图 6b 中绘制了生成的数据和 MPS 算法统计量。由图可见，该算法能够相当



(a) (b)

图7: （a）：鲁棒分布变化检测中涉及的不同矩形波形示意图。（b）：五方波信号的样本路径示意图，其均值在变点 ν=500 处上移 0.8，以及计算得到的鲁棒检验统计量。

有效地检测出变化。我们针对不同的变化时隙重复进行了仿真。无论时隙索引如何，MPS 算法都能够检测出变化，且没有虚警。

5.3. 鲁棒最快变化检测算法的数值结果 m

本节将式（15）中定义的鲁棒算法应用于模拟数据。为便于参考，我们在此重述该算法。

$$\bar{\tau}^* = \inf\{n \geq 1 : \bar{p}_n \geq A_{(n \bmod T)}\}, \tag{84}$$

其中 $\bar{p}_0 = 0$，且

$$\bar{p}_n = \frac{\tilde{p}_{n-1}\, \bar{g}_n(X_n)}{\tilde{p}_{n-1}\, \bar{g}_n(X_n) + (1 - \tilde{p}_{n-1}) f_n(X_n)}, \tag{85}$$

with

$$\tilde{p}_{n-1} = \bar{p}_{n-1} + (1 - \bar{p}_{n-1})\rho.$$

请注意，此处 ($\bar{g}_1$，……，$\bar{g}_T$) 是最不利 i.p.i.d.分布，而 (f_1，……，f_T) 是变点前独立非同分布律。

在此数值实验中，我们假设观测到矩形波形的含噪版本；参见图 7a。在 500 时刻的变化点之前，矩形波形每 50 个时隙在 $+1$ 和 -1 之间交替（图 7a 中的蓝色曲线）。变化点之后，波形在 $+1.8$ 和 -0.2 之间切换（图 7a 中的红色波形）。我们假设决策者不知道变化后的确切波形，但知道偏差至少为 0.1（图 7a 中的橙色虚线波形）。我们假设观测到的波形是在加入方差为 0.01 的高斯零均值随机变量后得到的。在这种设置下，可以

证明均值水平为橙色的高斯独立非同分布过程是最不利的。生成的观测序列和鲁棒变化检测统计量 (85) 如图 7b 所示。我们使用 $\rho = 0.01$ 来生成该统计量。从图中可以看出，鲁棒变化检测算法能够有效检测波形模式的变化。

5.4. 心电图心律失常检测与故障隔离的数值结果

由于每次心跳过程中（如图8所示的一个 PQRS 周期）心肌细胞内的电活动（内部动力学），心电图数据大多具有周期性。鉴于其在诊断方面的应用，近年来已有大量论文从机器学习和统计模式识别的角度，研究如何开发算法来自动检测并分类心电图数据中的心律失常 [10, 12, 16]。本节将第 3 节中开发的适用于独立非同分布过程的最快变化检测与故障隔离算法应用于真实心电图数据和模拟小波数据。为便于查阅，我们再次列出该算法。

对于 $\ell = 1, \ldots, M$，定义停止时

$$\tau_\ell = \inf\left\{ n \geq 1 : \max_{1\leq k\leq n} \min_{0\leq m\leq M, m\neq \ell} \sum_{i=k}^{n} \log \frac{g_i^{(\ell)}(X_i)}{g_i^{(m)}(X_i)} \geq A \right\}. \tag{86}$$

我们的检测-分类问题的停止时刻和决策规则定义如下:

$$\begin{aligned} \tau_{dc} &= \min_{1\leq \ell\leq M} \tau_\ell, \\ \delta_{dc} &= \arg\min_{1\leq \ell\leq M} \tau_\ell. \end{aligned} \tag{87}$$

将上述算法中的每个 τ_ℓ （35）替换为，即可得到其窗口受限版本

$$\tilde{\tau}_\ell = \inf\left\{ n : \max_{n-L_\beta\leq k\leq n} \min_{0\leq m\leq M, m\neq \ell} \sum_{i=k}^{n} \log \frac{g_i^{(\ell)}(X_i)}{g_i^{(m)}(X_i)} \geq A \right\} \tag{88}$$

对于窗口 L_β 的适当选择。回顾一下，这里 $(g_1^{(0)}, \ldots, g_T^{(0)})$ 是正常独立非同分布分布，而对于 $\ell \neq 0$，$(g_1^{(\ell)}, \ldots, g_T^{(\ell)})$ 是表示 ℓ 类型异常或变化的变化后独立非同分布分布。

*5.4.1. MIT-BIH*数据集

本文使用从复杂生理信号研究资源（PhysioNet）网站下载的 MIT 波士顿 BethIsrael 大学医院（MIT-BIH）数据集。所获取的数据集包含来自 47 名人类受试者的 48 条记录，每位受试者的数据记录时长约为半小时 [19]。数据包含以下信息: 双通道信号的二维数组、用于标注心律失常类型的专家标注一维数组，以及用于标记位置的一维数组

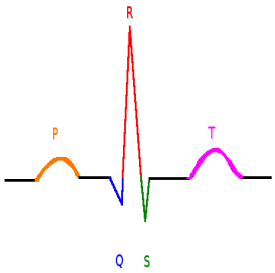


图8：正常心搏形态特征示意图：已识别出峰值。

表1：MIT-BIH数据集中 AAMI 对正常心搏和异常心搏的等价类别。

id	类别代表符号	符号
0	N	‘N’, ‘e’, ‘j’, ‘L’, ‘R’
1	V	‘V’,‘E’
2	S	‘S’, ‘A’,‘a’,‘J’
3	F	‘F’

R波峰可为每条心电图数据的解读提供充足信息。该二维数组信号由双通道正弦波构成，在10毫伏（mV）范围内具有11位分辨率，来源于12导联标准心电图设备；MIT-BIH数据库中的所有心电图均以恒定的360样本/秒（Hz）采样率采集 [18]。

由于主导联II（mlII）通道是所有患者通用的心电图记录，因此12个导联中仅提供了该导联的标注。因此，我们以与 [10]类似的方式分析了该数组。

我们采用医疗仪器促进协会（AAMI）标准中的四类表示法，将 MIT-BIH 中不同的标注重新聚类为较小的簇。该标准包含四个较大的类别，分别为“N”（即 MIT-BIH 中表示正常心搏的“N”“e”“j”“L”或“R”）、“S”（室上性异位搏动）、“V”（室性异位搏动）和“F”（融合搏动），并已用于将 MIT-BIH 中观测到的 12 类标注重新聚类为表 1 中经过验证的四类。基于该表，我们将每个标签归入 AAMI 中具有代表性的类别。

我们选取编号（ID）为 208 的患者作为患者特异性分析对象。如表2所示，该患者对应的标注数组大小约为 3,000，占 MIT-BIH 数据集中全部 112,000 条标注的一部分（其中包括 ID 为=208的患者）。如表2所示，我们从心电图波形中移除了室上性心律失常（“S”簇）。

表2：MIT-BIH数据库中，ID为=208的患者与所有标签对应的重新聚类标注数量。

	'N'	'V'	'S'	'F'	'Q'*	总计
ID为=208的患者	1,585	992	2	373	86	3,039
MIT-BIH	90,631	7,236	2,781	803	11,196	112,647

* 表示一个簇，其中包含所有未归入前四个簇的标注。

表3：从ID为=208的患者中选取的训练集和测试集中各簇不同心搏的数量。

	'N'	'V'	'F'	总计
训练	817	488	171	1,476
Test	789	477	177	1,443

5.4.2. 数据中心化与标准化

如图8所示，对心电图的主导联II（mlII）数据进行分段的一种方法是从心搏中获取 R-R 间期的中点索引。对 ID 为=208的患者应用上述分区方法后，得到 2,951 条心搏数据，其中包含主导联II数据的 R标注；同时，剔除"S"和"Q"波后，标注数量也相同。由此，结果中排除了全部人工标注心搏中的 88 条标注心搏。受心电图设备采样率的限制，心搏长度上限为 360。所有获取的心搏时长各不相同。对于长度小于 360 的心搏，我们应用基于快速傅里叶变换（FFT）的重采样函数（该实现示例见图9下方的图）。

5.4.3. 训练集与测试集划分

我们从 AAMI 标准中"N""V"和"F"三个簇的所有心搏中随机抽取了 50%，用于训练（每类心搏的数量如表 3 所示），约占全部心搏的 48%。为展示算法的有效性，我们使用了由 10 个心搏组成的序列。由于在实时检验场景中，"S"和"Q"类心搏可能以任意顺序出现，因此我们仅使用原始心电图进行测试，其中所有标注均属于所研究的三个类别之一；为便于可视化，通常会以每批 10 个心搏的形式展示，类似于图9。

为了训练每个类别的 i.p.i.d. 模型，我们假设心电图波形是受到高斯噪声干扰的确定性波形。我们使用训练数据学习高斯独立同分布过程的均值和方差。我们使用了重采样后得到的时间 $T = 360$。学习得到的均值和方差参数如图10所示。粗线表示期望值，虚线表示与均值线相差一个标准差的位置。如图10所示，只有少数时隙能够区分两个分布。因此，我们只关注 $[130, 155]$ 或 $[200, 220]$ 的离散时间区间，以提高预测的准确性。

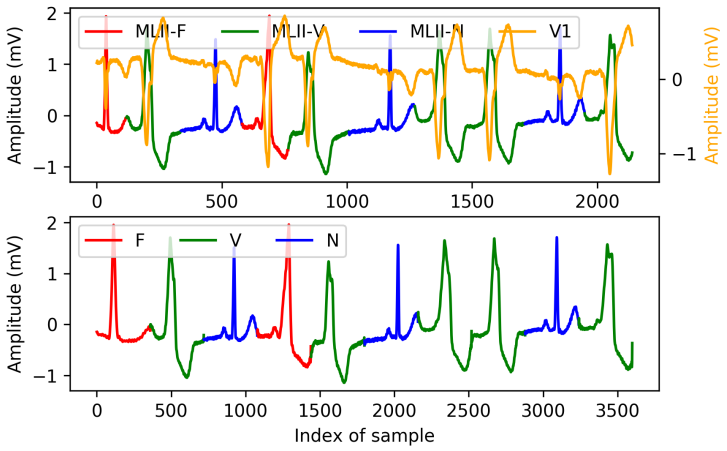


图9：ID为=208的患者双通道心电图MLII数据中前十个原始心搏与重采样分段数据的对比示意图。

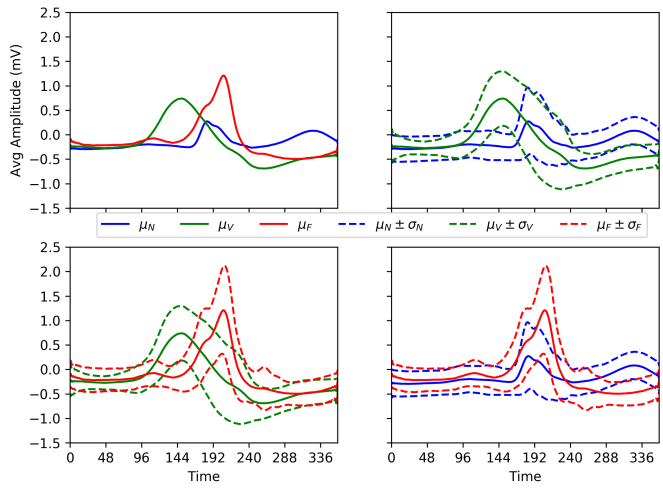


图10：ID为=208的患者不同心搏类型的变化前/后参数示意图。

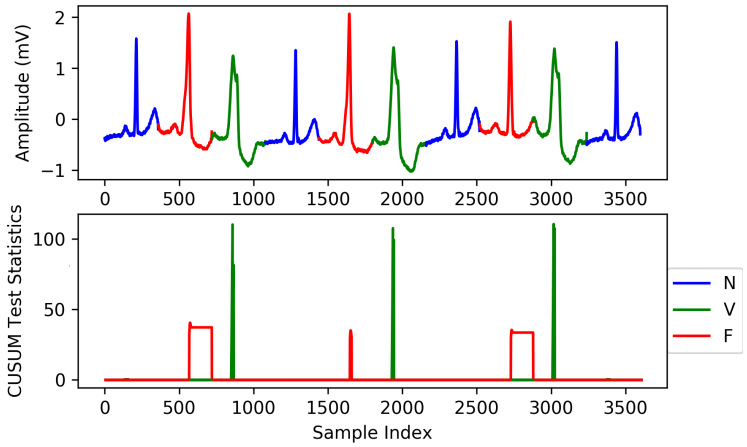


图11：从索引为80的正常心搏开始的包含十个心搏的心电图样本路径示例及计算得到的i.p.i.d.故障隔离检验统计量。

5.4.4. 将 *I.P.I.D.* 最快检测与隔离算法应用于心电图数据的结果

在图11中，我们绘制了从测试集中的第80个心搏开始的、包含十个心搏的心电图数据所得到的检验统计量。具体而言，我们针对每个类别绘制公式（88）中的统计量。红色统计量表示F型心律失常，绿色统计量表示V型心律失常。这些统计量的数值出现尖峰，表明已检测到相应类型的心律失常。如图所示，该算法在检测心律失常方面相当准确。需要指出的是，每当检验统计量越过阈值时，我们都会将其重置为零。

接下来，我们考虑起始心搏索引为1373的片段，如图12所示。从图中可以看出，其中既有虚警，也有错误的故障隔离。最后，我们将该算法应用于图13所示的另一个片段，该片段以索引235处的V型心搏开始，并以索引244处的V型心搏结束。该片段仅出现了一次误分类错误，结果是将其隔离为F型，而不是V型。

5.4.5. 将 *I.P.I.D.* 最快检测与隔离算法应用于小波数据的结果

由于 MIT-BIH数据集中的多类别数据有限，我们使用小波模拟数据来展示算法在三类检测与分类方面的有效性。Ricker 小波或 Marr 小波是 ECG 上一种抗噪的小波变换，在美洲也称为墨西哥帽小波或 Marr 小波 [15]。为了进行模拟，我们使用了墨西哥帽小波，它与 ECG 心搏的形态特征相似，且其变化前后分布的参数已知。用数学术语来说，Marr 小波的公式如下。

$$\psi(t) = \frac{2}{\sqrt[4]{9\pi}} \left(1 - t^2\right) e^{-t^2/2}$$

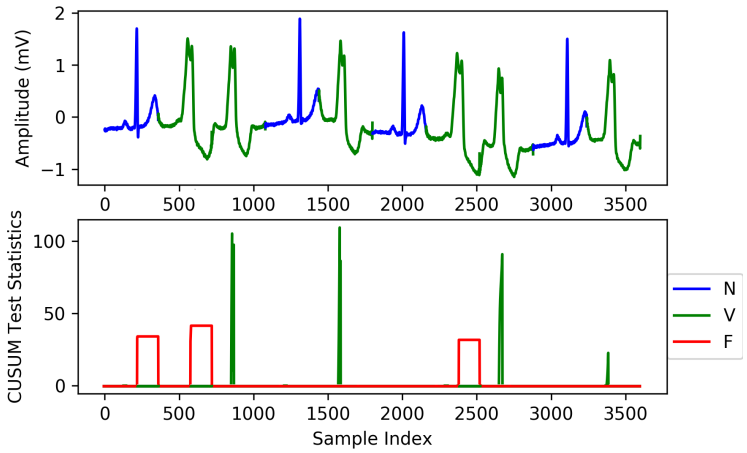


图 12：i.p.i.d.故障隔离检验统计量的变化过程：一种心律失常在索引 1374 处发生，随后出现四种与 V 型心律失常相同的心律失常，最后以 V 型心律失常结束。

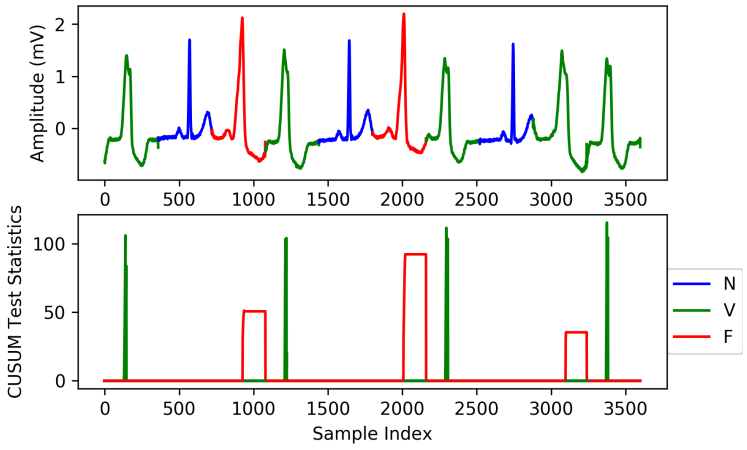


图 13：测试集中从索引 235 处的心律失常开始、包含十个心搏的心电图片段示意图。

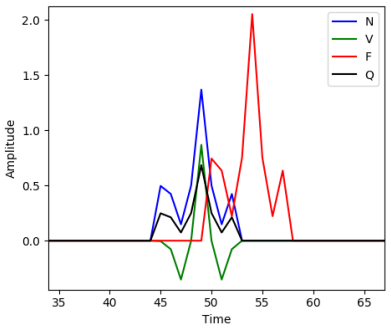


图14：通过对一个小波应用一组变换，展示四种不同小波在特定时间索引处的放大图。

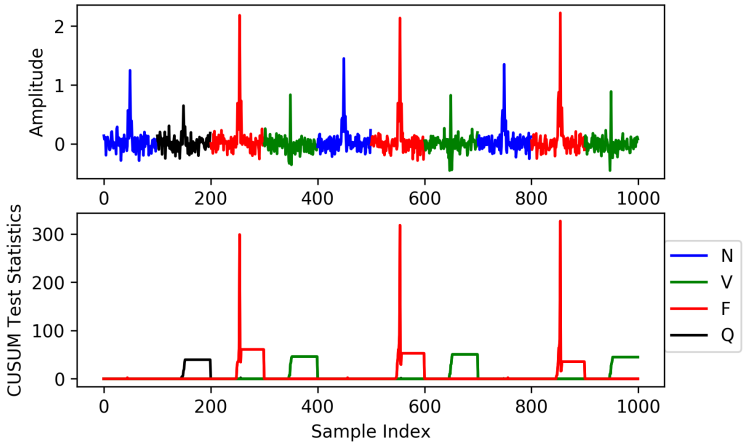


图15：具有不同类型异常的序列中，样本路径及相应的i.p.i.d.故障隔离检验统计量示意图。

对于离散时间仿真，我们从 SciPy 的 Ricker 小波生成函数中重新采样了一个以零为中心、长度为 100 的波形。墨西哥帽小波的不同函数变化形式，例如上移、缩放、时间延迟，或对两种扰动进行积分，产生了三种异常类型。如图 14 所示，总计有四个类别。实际数据通过向小波添加均值为零、方差为 0.01 的高斯噪声生成，然后将带噪波形串联起来，形成类心电图波形模式。结果绘于图 15 和图 16。如图所示，我们的算法能够实时准确地检测并识别故障。

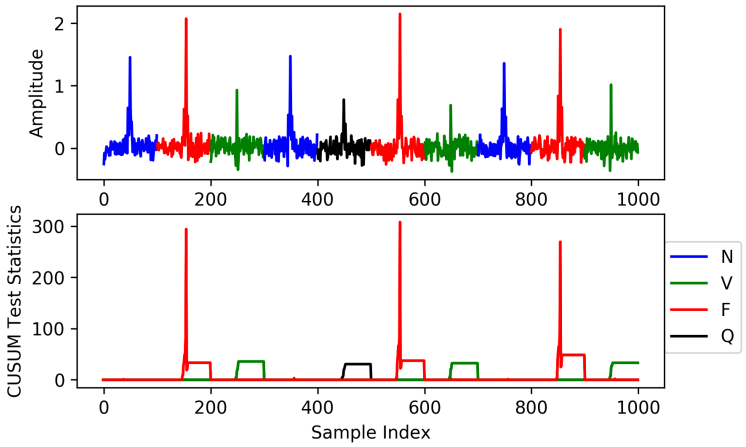


图16：不同样本路径及其对应的独立非同分布故障隔离检验统计量示意图。

6. 结论

我们开发了用于变化后独立非同分布分布律未知时，在独立非同分布过程中进行最快变化检测的算法。我们引入了最不利 i.p.i.d.分布的概念，并证明，基于最不利 i.p.i.d.分布设计的多阈值 Shiryaev 算法具有鲁棒最优性。随后，我们提出了一种用于独立非同分布情形下最快变化检测和故障隔离的算法，并证明当虚警率和误分类率趋于零时，该算法是渐近最优的。我们还证明，对于多时隙最快变点检测问题，基于混合的检验是渐近最优的。我们证明，所开发的算法可成功用于检测真实流量数据和真实心电图数据中的异常。

7. 致谢

Yousef Oleyaeimotlagh、Taposh Banerjee 和 Ahmad Taha 的研究工作部分得到美国国家科学基金会第 1917164 号拨款的资助。Yousef Oleyaeimotlagh、Taposh Banerjee 和 Eugene John 的研究工作也部分得到美国国家科学基金会第 2041327 号拨款的资助。

参考文献

[1] T. Banerjee、S. Allsop、K.M. Tye、D. Ba 和 V. Tarokh，神经数据中状态变化的序贯检测，载于第*9*届*IEEE*国际生物医学工程学会神经工程会议论文集，2019年3月。[2] T. Banerjee、P. Gurram 和 G. Whipps，统计周期性过程变化的贝叶斯最快检测，载于*2019*年*IEEE*国际信息论研讨会（*ISIT*）。IEEE，2019年，第2204–2208页。[3] T. Banerjee、P. Gurram 和 G. Whipps，周期性统计行为偏离的最快检测，载于*ICASSP 2019-2019*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2019年，第5351–5355页。[4] T. Banerjee、P. Gurram 和 G. Whipps，统计

周期随机过程的序贯检测理论，载于2019年第57届Allerton通信、控制与计算年会（Allerton）。IEEE，2019年，第290–297页。[5] T. Banerjee、P. Gurram 和 G.T. Whipps，统计周期性随机过程中变化检测的贝叶斯理论，《IEEE信息论汇刊》67（2021年），第2562–2580页。[6] T. Banerjee 和 V.V. Veeravalli，传感器网络中数据高效的最快变化检测，《IEEE信号处理汇刊》63（2015年），第3727–3735页。[7] T. Banerjee、G. Whipps、P. Gurram 和 V. Tarokh，利用多模态数据进行异常检测的循环平稳统计模型与算法，载于第6届IEEE全球信号与信息处理大会论文集，2018年11月。[8] T. Banerjee、G. Whipps、P. Gurram 和 V. Tarokh，非平稳环境中利用多模态数据进行序贯事件检测，载于第21届国际信息融合会议论文集，2018年7月。[9] D. Bertsekas，《动态规划与最优控制》，第II卷，Athena Scientific，贝尔蒙特，马萨诸塞州，2017年。[10] Y. Can、B.V.K.V. Kumar 和 M.T. Coimbra，利用心电图信号的形态与动态特征进行心搏分类，《IEEE生物医学工程汇刊》59（2012年），第2930–2941页。可访问：https://dx.doi.org/10.1109/tbme.2012.2213253。[11] Y.C. Chen、T. Banerjee、A.D. Dom´ınguez-Garc´ıa 和 V.V. Veeravalli，最快线路停运检测与识别，《IEEE电力系统汇刊》31（2016年），第749–758页。[12] P. Dechazal、M. O'Dwyer 和 R. Reilly，利用心电图形态和心搏间期特征进行心搏自动分类，《IEEE生物医学工程汇刊》51（2004年），第1196–1206页。[13] W.A. Gardner、A. Napolitano 和 L. Paura，循环平稳性：半个世纪的研究，《信号处理》86（2006年），第639–697页。[14]A.L. Goldberger、L.A.N. Amaral、L. Glass、J.M. Hausdorff、P.C. Ivanov、R.G. Mark、J.E. Mietus、G.B. Moody、C.K. Peng 和 H.E. Stanley，PhysioBank、PhysioToolkit和PhysioNet：复杂生理信号研究资源的新组成部分，《循环》101（2000年6月13日），第e215–e220页。《循环》电子版页面：http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218；doi：10.1161/01.CIR.101.23.e215。[15] J.A. Guti´errez-Gnecchi、R. Morfin-Maga˜na、D. Lorias-Espinoza、A. del Carmen Tellez-Anguiano、E. Reyes-Archundia、A. M´endez-Pati˜no 和 R. Casta˜neda-Miranda，基于数字信号处理的心律失常分类：小波变换与概率神经网络，《生物医学信号处理与控制》32（2017年），第44–56页。可访问：https: //www.sciencedirect.com/science/article/pii/S1746809416301677。[16] A.Y. Hannun、P. Rajpurkar、M. Haghpanahi、G.H. Tison、C. Bourn、M.P. Turakhia 和 A.Y. Ng，利用深度神经网络对动态心电图进行心脏病学家水平的心律失常检测与分类，《自然-医学》25（2019年），第65–69页。可访问：https://doi.org/10.1038/s41591-018-0268-3。[17] T.L. Lai，随机系统中的序贯多重假设检验与高效故障检测与隔离，《IEEE信息论汇刊》46（2000年），第595–608页。[18] G.B. Moody 和 R.G. Mark，MIT-BIH心律失常数据库的影响，《IEEE医学与生物学工程杂志》20（2001年），第45–50页。[19] G.B. Moody 和 R.G. Mark，MIT-BIH心律失常数据库（2005年）。可访问：https: //physionet.org/content/mitdb/1.0.0/，访问日期：2021年1月18日。[20] I.V. Nikiforov，一类序贯检验中检测/隔离延迟的下界，《IEEE信息论汇刊》49（2003年），第3037–3046页。[21] H.V. Poor 和 O. Hadjiliadis，《最快检测》，剑桥大学出版社，2009年。[22] A.N. Shiryayev，《最优停止规则》，施普林格出版社，纽约，1978年。[23] G. Tagaras，关于自适应控制图设计最新进展的综述，《质量技术杂志》30（1998年），第212–231页。[24] A. Tartakovsky，《序贯变化检测与假设检验：一般非独立同分布

随机模型与渐近最优规则，Chapman and Hall/CRC，2019年。[25] A.G. Tartakovsky、I.V. Nikiforov 和 M. Basseville，序贯分析：假设检验与变点检测，统计学，CRC出版社，2014年。[26] J. Unnikrishnan、V.V. Veeravalli 和 S.P. Meyn，极小极大鲁棒最快变化检测，IEEE《信息论汇刊》57（2011），第1604–1614页。[27] V.V. Veeravalli 和 T. Banerjee，最快变化检测，《Academic Press信号处理文库：第3卷——阵列与统计信号处理》，2014年。[28] S.C. Vishnoi、S.A. Nugroho、A.F. Taha、C. Claudel 和 T. Banerjee，基于非对称元胞传输模型、匝道连接且在有界扰动下的鲁棒交通密度估计，载于*2020*年美国控制会议（*ACC*）。IEEE，2020年，第1197–1202页。[29] Y. Zhang、N. Malem-Shinitski、S.A. Allsop、K. Tye 和 D. Ba，根据二元观测估计可分离马尔可夫随机场（*SMURF*），《神经计算》30（2018），第1046–1079页。