arXiv:2603.09318v1 [stat.ME] 10 Mar 2026

基于惊讶度的异常检测

Rob J Hyndman，莫纳什大学计量经济学与商业统计系；
David T Frazier，莫纳什大学计量经济学与商业统计系

2026年3月11日

摘要

异常检测方法应用广泛，但往往依赖临时规则或强假设，而且通常关注尾部事件，因而会漏掉发生在众数之间低密度间隙中的“类内”异常值。我们提出一个统一框架，将异常定义为在某个（可能设定错误的）模型下概率异常低的观测值。对于每个观测值，我们计算其惊异度（广义密度的负对数），并将异常分数定义为惊异度至少与该观测值一样大的概率。这样一来，复杂单变量或多变量数据的异常检测就转化为估计单变量惊异度分布的上尾。我们提出两种对模型具有稳健性的尾部概率估计量：一种是基于观测惊异度分布的经验估计量，另一种是对高阈值以上的数据拟合广义帕累托分布的极值估计量。对于经验法，我们给出尾部排序得以保留的条件，并通过德沃雷茨基–基弗–沃尔福维茨不等式推导有限样本置信保证。对于广义帕累托分布方法，我们确立了可确保经典极值行为的一般性尾部条件。模拟以及对法国死亡率和测试赛板球数据的应用表明，即使存在严重的模型设定错误，该方法仍然有效。

关键词：极值理论、广义帕累托分布、离群点检测、尾界

1 引言

异常检测方法往往相当临时随意，缺乏坚实的理论基础，或者依赖可能不符合实际情况的强分布假设。我们旨在通过将异常定义为在给定模型下概率较低的观测值，并提出一种通用的异常检测框架来解决这一问题。该框架估计与密度相同或更低（因而更异常）的观测值相关联的概率质量。我们的方法能够以统一的方式处理异常检测，并适用于任何能够在观测值空间上定义概率分布的异常检测问题。它适用于单变量和多变量情形，也可用于无条件分布和条件分布。尽管该方法要求指定概率分布，但对于该分布存在的严重模型设定错误仍具有稳健性。

基本思想是计算每个观测值的惊异度（等于负对数密度），然后根据惊异度分布估计尾概率。这使得我们能够估计获得至少与该观测值的惊异度一样大的惊异度的概率，进而可用于标记异常值。异常检测的惊奇度方法通过在估计的尾部概率上设定阈值，为在标记异常值时控制假阳性率提供了一种有原则的方法。

令 $\boldsymbol{y}_1$、……、 $\boldsymbol{y}_n$ 为定义在适当样本空间上的概率分布的独立同分布观测值，并令 f 为 F的广义概率密度函数。也就是说， f 通过在 F的跳跃点处使用狄拉克δ函数来定义；因此，当 F 为连续型时，它表示密度函数；当 F 为离散型时，它表示概率质量函数；当 F 为连续分量和离散分量的混合分布时，它表示二者的组合。

观测值的惊异度 定义为 $\boldsymbol{y}_i$ （Stone2022）

$$s_i = -\log f(\boldsymbol{y}_i).$$

虽然在统计学中它们更常被称为“对数得分”（Good1952），但在此语境下，我们更倾向于使用“惊异度”一词（由 Tribus 1961创造），因为这个词体现了这样一种观点：低密度观测值比高密度观测值更“令人惊讶”（因而可能是异常）。较大的 s_i 表示 $\boldsymbol{y}_i$ 是不太可能出现的值，因此可能是异常。另一方面，典型值的惊讶度较低。在信息论中，平均惊奇度称为随机变量的熵（Cover &Thomas2006、Stone 2022），可用于衡量不确定性。如果 f 是由参数模型得到的，那么惊异度之和就等于数据对数似然的负值。

令 $G(s) = \Pr\{S \leq s\}$ 为惊异度值的分布函数，其中 $S = -\log f(\boldsymbol{Y})$，且 $\boldsymbol{Y} \sim F$。那么，观测到惊异度至少与 s 一样大的值的概率为 $\Pr(S \geq s) = 1 - G(s^-)$，其中 $G(s^-) = \lim_{t \uparrow s} G(t)$。由此可自然地为每个观测值 $\boldsymbol{y}_i$赋予一个异常分数：

$$p_i = \Pr(S \geq s_i) = 1 - G(s_i^-),$$

其中 $s_i = -$对数 $f(\boldsymbol{y}_i)$。我们将惊异度为 $p_i < \alpha$ 的观测值定义为异常值，其假阳性率为 α，前提是所有观测值均来自 F。相较于直接计算 F 尾部的概率，这种做法更可取，因为它允许异常值出现在 F 分布的任何位置，而不只是在其尾部。例如，异常值可能位于多峰分布的众数之间，也可能出现在分布 F 的其他低密度区域。与基于距离的异常检测方法（例如 Wilkinson 2018）相比，这种做法也更可取，因为它能够自然地处理偏斜分布和重尾分布。

惊异度方法　　这种方法将处理 F（一种可能是多变量的

分布，甚至是定义在流形或某种非欧几里得空间上的分布）的问题，转化为处理定义在 $\mathbb{R}$ 上的单变量分布 G 。我们可以写成

$$G(s) = \Pr\{-\log f(\boldsymbol{Y}) \leq s\} = \Pr\{f(\boldsymbol{Y}) \geq e^{-s}\}.$$

这清楚地表明，惊异概率与分布的最高密度区域有关（Hyndman 1996年）。事实上，每个满足 $p_i < \alpha$ 的观测值都位于$100(1-\alpha)$% 最高密度区域之外，该区域属于 F。

如果 $F = N(\mu, \ \sigma^2)$，那么 $p_i = 2[1-\Phi(|y_i-\mu|/\sigma)]$, 其中 Φ 是标准正态分布函数。在这种情况下，将满足 $p_i < \alpha$ 的观测值识别为异常值，等价于采用阈值为 $\Phi^{-1}(1-\alpha/2)$ 的 Z 分数检验。因此，我们的方法可以看作是基于 Z 分数的异常检测方法（包括 Grubbs 检验）的推广（Grubbs 1950）。

它也可以看作是其他异常检测方法的推广。例如，Frank Hampel 提出的一种方法，即 Hampel 识别器（Davies 和 Gather 未注明日期），旨在根据某个观测值与相邻观测值是否相差很大来识别时间序列中的异常值。如果 $|y_t - m_t|/a_t > \tau$，则将该观测值标记为异常，其中

$$m_t = \text{median}(y_{t-h}, \ldots, y_{t+h}),$$

$$\text{and} \qquad a_t = \text{median}(|y_{t-h} - m_t|, \ldots, |y_{t+h} - m_t|).$$

阈值 τ 通常是在正态分布的假设下设定的。当我们从惊异度的角度来看时，这相当于设定

$$F = N(m_t, a_t^2/\Phi^{-2}(0.75)).$$

假异常率为 $\alpha = 2[1-\Phi(\tau/\Phi^{-1}(0.75))]$。

我们通常不知道 F，因此需要根据数据或理论对其进行估计。在低密度区域，这类估计尤其脆弱或不确定，而这恰恰是我们寻找异常值的地方。此外，我们还需要一种不依赖强分布假设的方法。

我们的方法从一个假定的 F开始，该分布可以根据数据估计得出，也可以由理论推导得出。我们使用这个分布计算惊异度值$\{s_1, \ldots, s_n\}$。然后，我们使用三种方法之一，根据这些惊异度值估计尾部概率$\{p_1, \ldots, p_n\}$。

(1) 使用假定分布 F 计算尾部概率，即 $p_i = \int I\{f(\boldsymbol{y}) \leq e^{-s_i}\} d\boldsymbol{y}$，其中 I 是示性函数；(2) 使用观测到的惊异度$\{s_1, \ldots, s_n\}$的经验分布函数，因此 p_i 是大于或等于 s_i的惊异度所占的比例；(3) 对最大的惊异度值拟合广义帕累托分布（GPD），并用它来估计尾部概率。

显然，方法（1）作出了一个强假设，即 F 能够准确表示数据。方法（2）和（3）允许用于计算惊异度的分布存在相当大的模型设定错误。方法（2）只要求尾部惊异度的排序与真实分布下可能得到的排序相似，而方法（3）则要求极值理论适用于惊异度分布。

在第2节中，我们将阐述支持使用经验惊异度概率所需的理论，并确定在哪些条件下，这种方法能够很好地估计真实意外度概率。同样，第3 节将给出使用广义帕累托分布近似惊异度分布尾部所需的理论结果。

虽然我们主要以独立同分布数据的形式呈现主要结果，但第4 节表明，这些方法同样适用于条件分布。我们在第5节通过若干模拟实验展示这些方法，并在第6节介绍两个应用。随后在第7节进行讨论。附录包含第2节和第3节结果所需的证明。

2 经验惊异度概率

根据上述方法（2），我们将 $p_i = \Pr(S \geq s_i)$ 估计为观测到的惊异度中不小于 s_i的比例。使用该方法，阈值为 α 时的异常边界对应于$100(1-\alpha)\%$ 的最高密度区域，该区域使用 Hyndman (1996年) 的分位数密度方法计算得出。

令 $F(\cdot)$ 表示 $\boldsymbol{Y}$ 的真实分布函数，并假设我们使用 F 对 $\boldsymbol{Y}$ 进行建模。相应的广义密度函数分别记为 $f(\cdot)$ 和 $f(\cdot)$。要使经验方法准确，拟合模型下较大惊异度值的排序必须与真实分布下的排序一致。这等价于要求真实惊异度与拟合惊异度在上尾通过严格递增变换相关联。

回顾一下，真实分布下的惊异度定义为 $S = -\log f(\boldsymbol{Y})$，并且记得 $G(s) = \Pr(S \leq s)$ 表示在 $s \in \mathbb{S} \subseteq \mathbb{R}$ 处的 S 累积分布函数，其中 $\mathbb{S}$ 表示惊讶度的支撑集。同样，令 $S = -\log f(\boldsymbol{Y})$ 表示拟合分布下的惊异度，并令 $\widehat{\mathbb{P}}_n(s) := n^{-1}\sum_{i=1}^{n} 1[\widehat{S}_i \leq s]$ 表示拟合分布样本 $S_1, \ldots, S_n$ 的经验累积分布函数。要使经验估计量 $\mathbb{P}_n(s)$ 能够在任意给定样本中准确估计 $G(s)$ 的上尾，以下条件最终证明既是

必要条件，也是充分条件。

假设 2.1。 存在一个 $s_\star$ 和一个函数 $h(\cdot)$，该函数在 $\succ$ $[s_\star, \infty)$ 上严格递增，使得对所有 $S \geq s_\star$， $S = h(S)$ 几乎必然成立。

引理 2.1。 设 $S_1, \ldots, S_n$ 独立同分布，且其分布函数为 G。那么，对于任意 $\alpha \in (0,1)$，以及 $\epsilon = \sqrt{\log(2/\alpha)/(2n)}$，对所有 $s \geq s_\star$， $\widehat{\mathbb{P}}_n(s) - \epsilon \leq G(s) \leq \widehat{\mathbb{P}}_n + \epsilon$ 以至少 $1-\alpha$ 的概率成立，当且仅当假设 *2.1* 成立。

引理 2.1 表明，如果拟合分布的尾部与实际数据分布的尾部位置相近，那么拟合分布的经验惊异度概率就是对真实惊异度概率 $G(s)$的非常准确的度量。关键在于，这并不要求拟合分布的尾部与实际数据分布的尾部相同；只要求超过阈值 $s_\star$ 之后，两者之间存在单调变换关系。也就是说，要使这种方法准确，我们必须确信，拟合分布的低密度区域与真实数据分布的低密度区域位于相同位置。

需要特别指出的是，假设2.1并不适用于所有分布对：如果真实数据存在偏度，但拟合模型假设其具有对称性，或者观测数据中存在未被模型刻画的双峰性，那么假设2.1 可能不成立。

3 惊异度分布的极值理论

我们的第三种方法是将极值理论应用于（单变量）惊异度值。我们将证明，只要惊异度分布的尾部满足较弱的条件，最大惊奇度就会收敛到三种极值分布之一。这为对最大的惊异度值拟合广义帕累托分布（GPD），并用其估计尾部概率提供了依据。

这项研究的一个重要先驱是Kandanaarachchi 与 Hyndman (2022)，他们提出通过留一法核密度估计来估计 F，然后对最极端的惊异度值拟合广义帕累托分布（GPD）。其基本思想是，即使 F 的估计不够准确，惊异度分布的尾部仍应能由 GPD 良好近似。Hyndman 等人(2025) 在核密度估计的背景下，为这一方法提供了一些理论依据。

在此，我们将这一思路扩展为允许 F 为任意分布，包括条件分布，只要其二阶矩有限即可。具体而言，我们将广义帕累托分布拟合到 $\{s_i\}_{i=1}^n$中最大的 β 个观测值上，从而估计位置、尺度和形状参数（μ、σ、ξ）。对于每个观测值，我们计算惊异度至少与其相同的概率，$p_i = \beta\left[1 - P\left(s_i \mid \hat{}\mu、\hat{}\sigma、\hat{}\xi\right)\right]$，其中 P为广义帕累托分布的分布函数。未包含在 $\{s_i\}_{i=1}^n$ 中最大的 β 个观测值中的观测值，其概率设为 $p_i = \beta$。为了论证这一方法的合理性，我们寻求关于 $S = -\log f(\boldsymbol{Y})$ 的惊异度分布的条件，其中 $\boldsymbol{Y} \sim F$，而 f 是相应的广义密度函数；这些条件类似于 Fisher-Tippett-Gnedenko（FTG）定理（Coles 2001）。

我们考虑关于 $S = -\log f(\boldsymbol{Y})$ 的惊异度分布的三种可能条件，它们描述了尾部逐渐变重的情形：

1. 其中 S 是次高斯的（例如，当 f 有上下界时）；2. 其中 S 是次指数的（例如，当 f 服从高斯分布时）；3. 其中 S 具有多项式矩（例如，当 f 服从学生 t 分布时）。

我们将这些表述为三个可选假设。

假设 3.1 （次高斯）。随机变量 $S = -\log f(\boldsymbol{Y}) \in \mathbb{R}$，并满足：

对于所有 $\lambda \in \mathbb{R}$，以及某个 $\nu > 0$,

$$\mathbb{E}\exp\{\lambda(S - \mathbb{E}[S])\} \leq \exp\{\lambda^2\nu^2/2\}.$$

假设 3.2（次指数）。 随机变量 S 在零点的某个邻域内存在指数矩；即，对于参数 ν 和 b,

$$\mathbb{E}\exp\{\lambda(S - \mathbb{E}[S])\} \leq \exp\{\lambda^2\nu^2/2\} \quad \text{for all} \quad |\lambda| < 1/b.$$

假设3.3（多项式）。 随机变量 $|S|$ 存在阶数为 $\alpha \geq 1$的多项式矩；即， $\mathbb{E}[|S|^p] \leq C^p$ 对于某个 $C > 0$，使得 $C^p - 1 > 0$。

假设 3.1 在密度 f 在 f的支撑集上有界且有下界时成立，例如有限集上的质量函数。但如果对数密度在 f的支撑集上没有下界，则该假设可能不成立。

假设 3.2在 f 在 $\mathbb{R}^d$ 上具有全支撑且对数密度无下界时成立，例如高斯分布。

假设 3.3 在 f 具有多项式尾部时成立，例如学生 t 分布。

在每个假设 3.1–3.3下，我们寻求关于惊讶度最大值的极限分布的结果 $M_n = \max\{s_1, \ldots, s_n\}$。定理 3.1给出了这一结果，证明见附录。

定理 3.1。 设 $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ 是来自 F 的独立同分布序列，并令 $S = -\log f(\boldsymbol{Y})$，其中 $\boldsymbol{Y} \sim F$。此外，令 $M_n = \max\{s_1, \ldots, s_n\}$，其中 $s_i = -\log f(y_i)$。

1. 如果惊异度满足假设 *3.1*，那么

$$\sup_{s:s>0} \left| \Pr\left\{ \left| M_n - \mathbb{E}[S] \right| \geq \sqrt{2\nu^2 s} + \sqrt{2\nu^2 \log(2n)} \right\} - e^{-s} \right| = o(1).$$

2. 如果惊异度满足假设 *3.2*，那么

$$\sup_{s:s>1/b}\left|\Pr\left\{\left|M_n-\mathbb{E}[S]\right|\geq(2b)s+(2b)\log(2n)\right\}-e^{-e^{-s}}\right|=o(1).$$

3. 如果惊异度满足假设 *3.3*，那么

$$\sup_{s:s>C}\left|\Pr\left\{\left|M_n-\mathbb{E}[S]\right|\geq(Csn^{1/p})\right\}-e^{-s^{-p}}\right|=o(1).$$

定理3.1 表明，在这三种假设下，M_n 的极限分布均属于极值分布类。如果惊异度具有类似高斯分布的尾部，那么最大惊异度服从反韦布尔分布；如果它仅具有指数矩，那么最大惊异度将服从冈贝尔分布；如果最大惊异度仅具有多项式矩，那么它服从弗雷歇分布。由于这些分布都属于广义极值分布族，因此可以用广义帕累托分布（Coles2001，定理 4.1）近似该分布的上尾。

特别是，即使假定分布 F 不正确，只要惊异度分布的尾部满足三种假设之一，我们仍可以对最大的惊异度值拟合广义帕累托分布，并用它来估计尾部概率。

定理 3.1 也有助于阐明错误指定用于计算惊异度值的分布所带来的后果。如果将数据建模为轻尾（即次高斯），但实际上具有重尾（次指数或多项式），则最大惊讶度服从弗雷歇分布。然而，反之并不成立：若在数据实际上为轻尾（例如次高斯）时假定重尾模型（例如次指数或多项式），则所得最大惊讶度的分布仍与较轻尾部一致（即其为次高斯）。随着惊异度分布的尾部变得更重，收敛变慢，并且需要更多数据才能可靠地检测惊讶度。实际上，低估尾部厚重程度会招致重大惩罚——表现为缓慢的收敛

国内生产总值的，并因此导致不准确的异常检测——而高估尾部厚重程度的代价相对较小。这表明，在实践中，宁可错在假定重尾这一侧更为稳妥。

4 条件分布的适用性 S

引理2.1 和定理3.1 中的理论结果假设，惊异度基于从某个未知分布 $\boldsymbol{Y}$生成的独立同分布数据。这样，这些结果不受 $\boldsymbol{Y}$维度的影响；所得惊异度 $S_i = -\log f(\boldsymbol{Y})$ 始终是单变量的，并且仍然独立同分布，因为 $\boldsymbol{Y}$ 的任何函数也都是独立同分布的。

然而，在许多建模情境中，例如回归情境，我们未必关心衡量数据联合分布所产生的惊异度，而是关注给定一组固定条件变量（例如年龄或性别）时的惊异度。在这类情形下，我们可以将观测数据视为变量 $\boldsymbol{Y} = (Z,\ \boldsymbol{X})$，其中 Z 是我们感兴趣的某个结果变量，而 $\boldsymbol{X}$ 是用于预测或解释 Z的变量。在这类情形下，我们可以直接对形式为 $f_Z(z \mid \boldsymbol{X} = \boldsymbol{x})$ 的条件模型应用惊奇度方法，而惊异度 $S = -\log f_Z(Z \mid \boldsymbol{X} = \boldsymbol{x})$ 可以理解为是在“给定某个 $\boldsymbol{X} = \boldsymbol{x}$ 的条件下”计算得出的。

对于固定的条件变量向量，不难看出我们的理论结果为何仍然成立：通过固定 $\boldsymbol{X} = \boldsymbol{x}$，由于我们观测到独立同分布的 $\boldsymbol{Y} = (Z_i,\ \boldsymbol{X}_i)$，$Z \mid \boldsymbol{X} = \boldsymbol{x}$ 的分布在给定 $\boldsymbol{X} = \boldsymbol{x}$的条件下也独立同分布，因此由此得到的理论仍然适用。也就是说，例如，如果我们希望在协变量 $\boldsymbol{x}$的固定取值处计算惊讶度，那么可以在这组固定协变量的条件下应用引理 2.1 和定理 3.1 。然而，重要的是要认识到，如果我们希望同时针对

结果变量和协变量（即允许 Z 和 $\boldsymbol{X}$ 变化）计算惊讶度，那么惊讶度的含义可能因实例而异。为了解其中的原因，考虑一个简单的例子：我们希望使用一个简单的高斯回归模型来检测 $Z \mid X$的惊讶度，因此

$$S_i = -\log f_Z(Z_i \mid X_i) \propto (Z_i - \beta_0 - \beta_1 X_i)^2.$$

现在， S_i 可能因两个既不同又相互关联的原因而变大：要么是相对于 $\beta_0 + \beta_1 X$的观测值而言， Z 本身较大；要么是因为 X 并不典型，导致 $Z - \beta_0 - \beta_1 X$较大。

也就是说，在一般情况下，当我们在 $\boldsymbol{X}$ 未固定的条件模型中寻找惊讶度时， S_i 的行为由联合分布决定：

$$\Pr(S_i > s) = \int \mathbb{1}(-\log f_Z(z \mid \boldsymbol{x}) > s) f_Z(z \mid \boldsymbol{x}) f_X(\boldsymbol{x})\, dz\, d\boldsymbol{x}.$$

这说明， S_i 中的极端行为可能源于 Z、 $\boldsymbol{X}$的极值，也可能仅仅是因为 Z 和 $\boldsymbol{X}$ 取值时，$f_Z(Z \mid \boldsymbol{X})$非常小。关键在于，即使 Z 和 $\boldsymbol{X}$ 取的是对于其边际分布而言典型的值，后一种情况仍然可能发生；但这些值并未被假定的条件模型很好地捕捉。

5 实验

5.1 标准单变量正态值

虽然计算惊讶度时需要假设某个密度函数 f ，但我们希望得到的尾部概率对错误设定的分布具有稳健性。假设我们从 $N(0, 1)$ 分布中生成 1000 个观测值，但使用 $t(4)$ 分布来计算惊异度值。经验尾部概率以及使用

广义帕累托分布方法得到的尾部概率，尽管计算惊讶度时使用的分布不正确，仍应能给出合理的尾部概率估计。

为研究假定分布的影响，我们分别使用正确的正态密度和错误的 $t(4)$ 密度来计算惊异度值。然后，对于每组惊异度值，我们通过三种方式估计惊异度尾部概率：(a) 使用计算惊异度值时所用的同一分布；(b) 使用经验尾部概率；(c) 使用拟合于最大的 100 个惊异度值的广义帕累托分布。

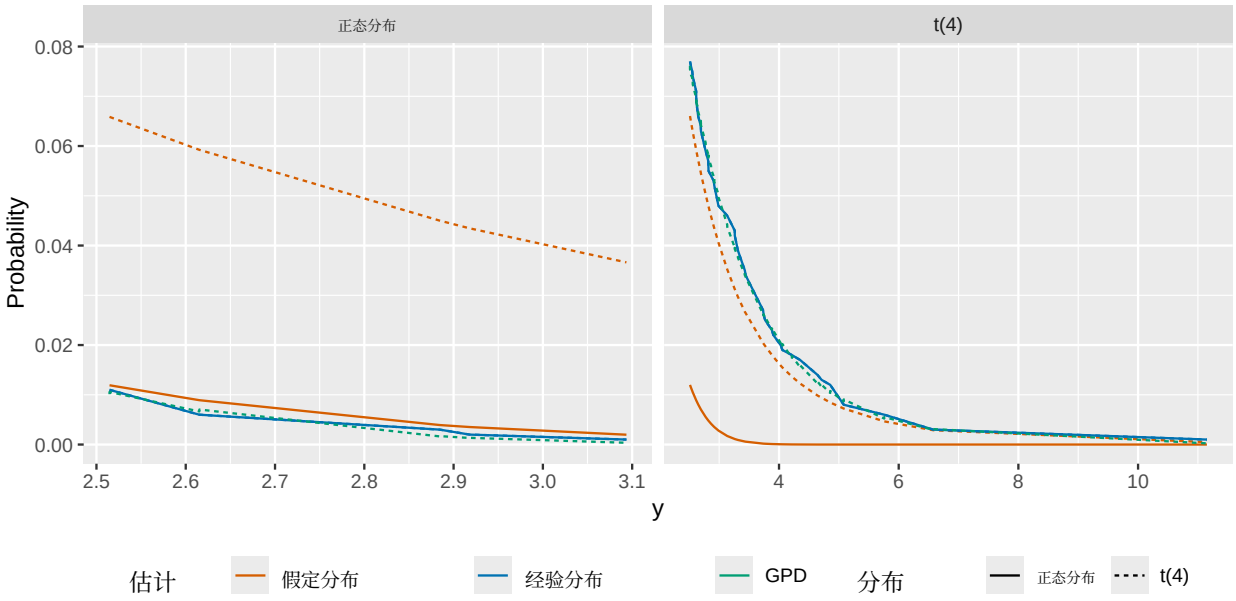


图1：左：观测值根据 $N(0,1)$ 分布计算得出。右：观测值根据 $t(4)$ 分布计算得出。分布表示用于计算惊异度值的分布。估计方法表示尾部概率的计算方式。这表明，即使使用错误的分布计算惊异度值，基于经验分布或广义帕累托分布计算的惊异度概率仍相对准确。

各种估计值显示在图1 的左侧面板中，针对大于 2.5 的观测值，表明经验尾部概率和广义帕累托分布估计仍然准确，

无论使用哪种分布来计算惊讶度。实线显示使用真实 $N(0,1)$ 分布计算惊讶度所得的结果，而虚线显示使用误设的 $t(4)$ 分布计算惊讶度所得的结果。直接根据 $t(4)$（“假定分布”）计算的估计非常不准确，但所有其他估计都接近真实情况。

总之，即使我们不知道数据的分布，也可以使用某种不正确的假定分布来计算惊异度，然后应用广义帕累托分布或经验分布，以获得对惊异度概率的良好估计。

相反的情况如图 1 右侧面板所示，其中数据由 $t(4)$ 分布生成，而惊异度值分别使用正确的$t(4)$ 密度和不正确的 $N(0,\ 1)$ 密度计算。即使使用错误的分布计算惊异度值，经验尾部概率和广义帕累托分布估计仍然相当准确。

5.2 双变量伽马值

所用近似方法的质量取决于可用数据量，因此在接下来的实验中，我们比较不同方法在各种样本量下的尾部概率。令 $\boldsymbol{Y} = (Y_{i,\ 1},\ Y_{i,\ 2})'$，其中 $Y_{i,\ j}$ 为相互独立的 Gamma(2,2)随机变量。我们从该二元分布中生成 n 个观测值，并考察随着 n 从 100 增至 10，000，概率估计的质量如何变化。惊异度值分别使用正确的分布以及错误的二元正态分布计算；后者的均值为 $(1,\ 1)'$，协方差矩阵为 $0.5\boldsymbol{I}$。对于每组惊异度值，我们通过两种方式估计尾部概率：(a) 使用经验尾部概率；(b) 使用根据惊异度值最大的 10% 拟合的广义帕累托分布。对于每种样本量，我们重复该实验 1000 次

。图中显示的异常率是对不同样本量下检测到的异常值比例进行的 loess平滑，并附有95%置信区间。在真实分布下，假异常率当然为 0.01。

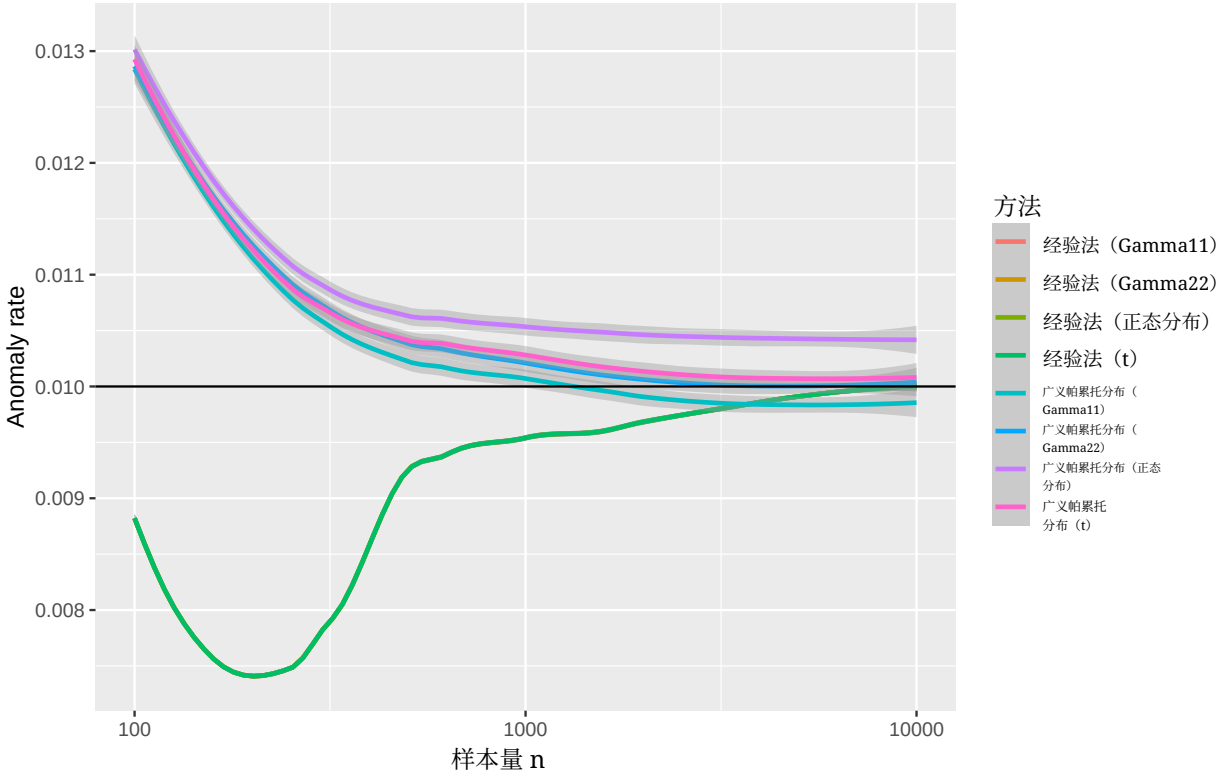


图 2：当 $\alpha = 0.01$ 时，不同近似方法下估计的假异常率。

请注意，在正确的伽马分布和错误的正态分布下，经验尾部概率完全相同，因为尾部惊讶度的排序并未改变。显然，所有方法都存在一定的小样本偏差，而广义帕累托分布比经验近似收敛得更快。在广义帕累托分布拟合中，使用学生t参考分布的表现几乎与使用真实模型相当，而使用正态参考分布的效果明显较差，即使样本量很大也是如此。原因在于尾部不匹配：正态分布比真实的伽马分布尾部更轻，而学生 t 分布的尾部则更重。这与我们先前的评论一致：假设尾部更重是更稳妥的做法，因为低估尾部厚重程度会导致

收敛变慢，而高估尾部厚重程度造成的代价相对较小。

6 应用

6.1 法国死亡率

人类死亡率数据库中提供了按年龄（单岁）和性别细分的法国死亡率人类死亡率数据库 (2025)。我们将考察1816年至1999年间0至85岁男性和女性的数据，共有172条独立时间序列，每条长度为184。

图 3显示了两种性别和所有年龄的数据。所有年龄的死亡率都随着时间推移而下降，尤其是自1950年以来。婴儿死亡率（0岁）远高于儿童期其他年龄段，所有年份的死亡率都在约10岁时降至最低。最高死亡率出现在最高年龄组。两次战争的影响在男性死亡率中尤为明显。

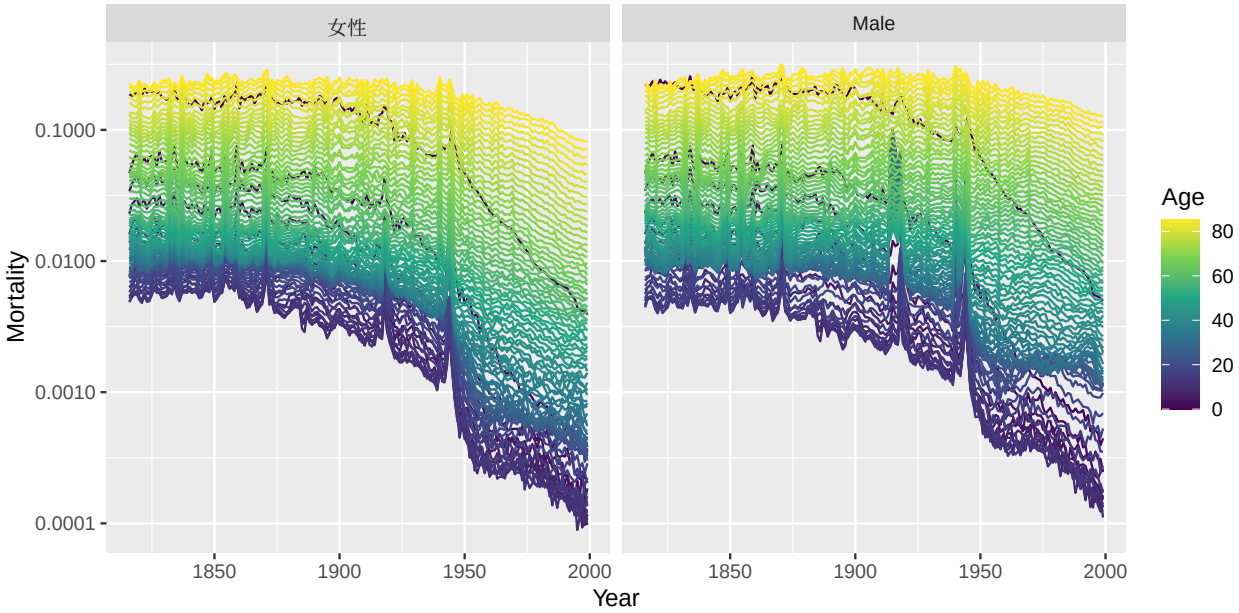


图3：1816年至1999年法国按性别和年龄划分的死亡率。由于不同年龄组的死亡率差异很大，我们采用对数尺度。

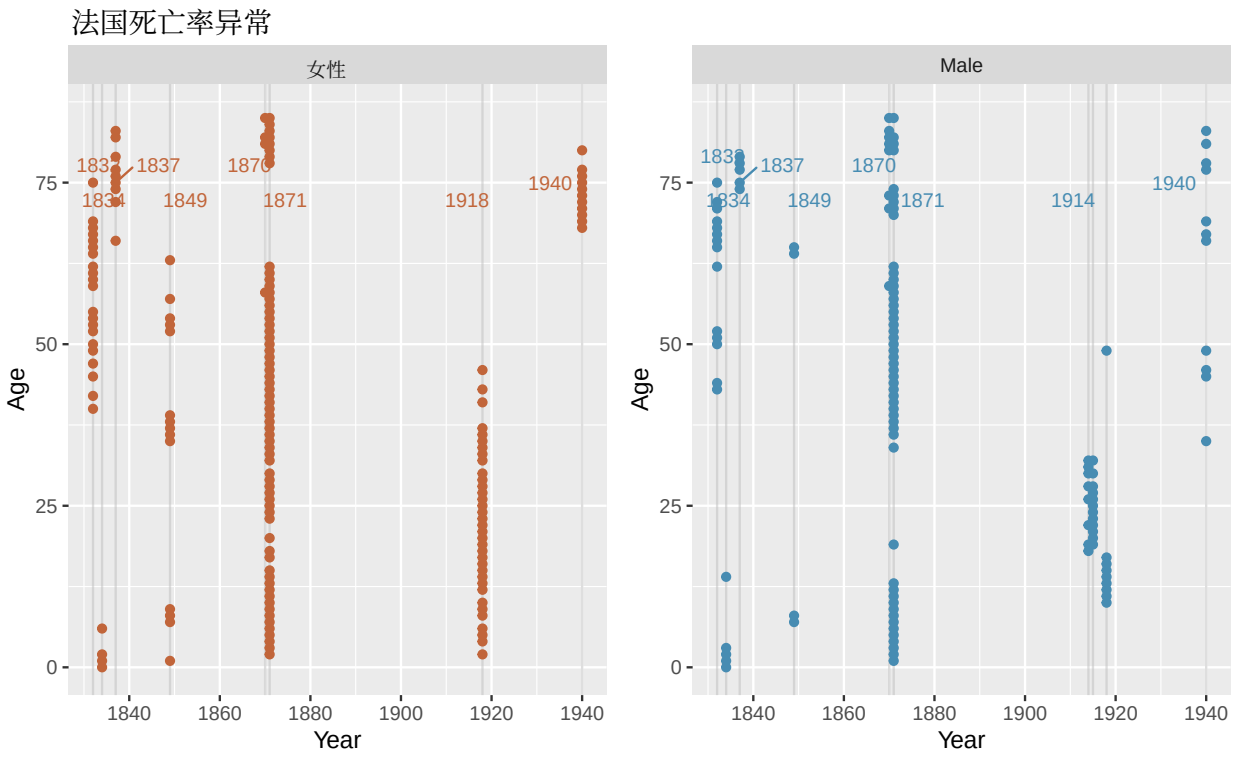


图4：按年份和年龄识别出的法国死亡率数据中的异常值。由此可见法国历史上的战争和流行病。

我们使用条件分布（1）计算惊讶度，然后在参数为 $\alpha = 0.01$ 的广义帕累托分布下计算惊异度概率。由于观测值很多，出现大量假阳性的可能性很高，因此我们只保留同一年、同性别中至少有三个年龄组出现异常的情况。最终，在31648个观测值中发现了289个异常值。图4 展示了这些异常值，它们与法国历史上发生过的各种战争和流行病相吻合：

- 1832，1849年：霍乱暴发
- 1870：普法战争
- 1871：对“巴黎公社”的镇压
- 1914-1918年：第一次世界大战
- 1918：西班牙流感暴发
- 1940：第二次世界大战

此外，1834年的婴儿死亡率似乎异常偏高，而1837年的老年人死亡率似乎异常偏高。造成这种情况的原因尚不清楚。

6.2 板球测试赛未出局次数

1834年至2025年10月6日期间参加板球测试赛的所有男性和女性的职业击球数据来自Hyndman(2026）。我们将探究是否有些击球手未被淘汰的局数比例异常高，这种情况称为“未出局”。该数据集包含97649局比赛结果，其中12695局未出局。因此，未出局的总体比例为 $12695/97649 = 0.130$。

图 5 显示了每位击球手未出局次数所占比例与其参加局数之间的关系。每张图左侧出现的异常模式，是由数据的离散性造成的——未出局次数和参加局数都必须是

整数。由于击球手的未出局次数和参加局数可能相同，因此图中存在一些点重叠，使相应数据点的颜色密度更高。只参加过少数几局的击球手，其未出局比例平均而言往往更高，方差也更大；而参加过许多局的击球手则不然。

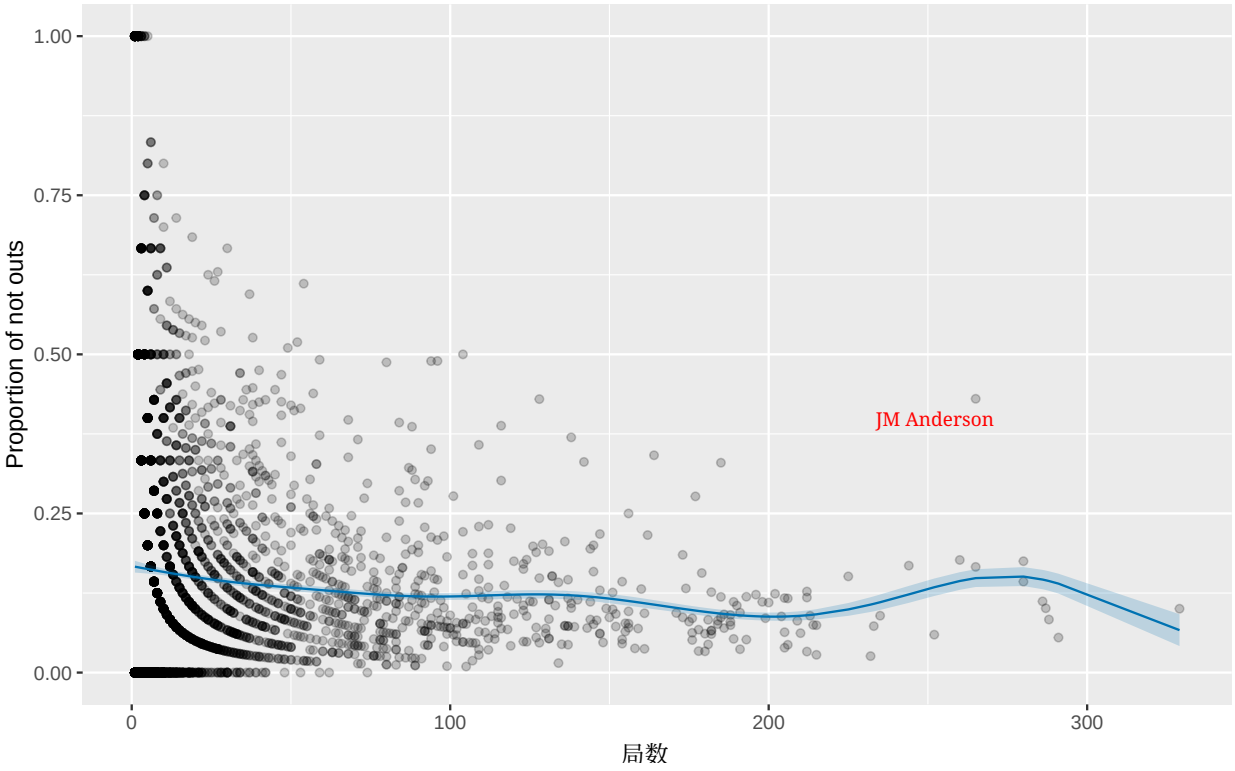


图 5：每位击球手的未出局次数比例与其所打局数之间的关系。蓝线及其对应的 95%置信区间显示了击球手未被淘汰的概率与其所打局数之间的关系。

这表明，我们可以构建一个广义加性模型（Wood 2017），以每位击球手的未出局次数为因变量，以其所打局数为自变量。自然可以使用带有 logit链接函数的二项分布：

$$\text{未出局数} \mid \text{局数} \sim \text{二项分布}(n = \text{局数}, \quad p),$$

其中 p 表示击球手在一局中未被判出局的概率，

$$\log(p/(1-p)) = g(\text{Innings}),$$

以及 $g()$ 是一个需要以非参数方式估计的光滑函数。使用带有惩罚回归样条的 mgcv 软件包（Wood 2025）得到的拟合结果以蓝色显示在图5中。

现在，我们可以使用拟合模型根据二项分布计算惊讶度，并找出最异常的击球手。虽然二项分布可能是一个很好的初步近似，但它确实假设击球手在整个职业生涯中被淘汰的概率不会发生变化；这是一个不太可能成立的假设，因为它没有考虑技能的发展、经验的价值或年龄增长的影响。因此，我们将使用广义帕累托分布近似来计算惊异度概率。

最异常的击球手全都是“尾端击球手”（即技术并不出色的击球手），而且他们打了很长时间（因此有很多局）。由于他们在球队击球顺序中排在最后或接近最后，因此更有可能在球队一局结束时保持未出局。

惊异度最大的是英格兰的击球手 Jimmy Anderson，他在 265 局中有 114 次未出局，远高于预期的未出局次数 $265 \times 0.148 = 39.3$。图 5中也能看到这一异常：对于数据的这一部分而言，它显得有些反常。虽然 Jimmy Anderson 并非出色的击球手，但他擅长防守，能够长时间击球而不被淘汰，从而为另一位击球手争取了得分时间。我们发现，这个异常无论是在未出局比例还是局数方面都不异常；与其他一些数值相比，实际比例与预测比例之间的差异也并不异常。然而，由于采用了统计模型，我们能够考虑到

该数据集的特征，例如数据的离散性和变化方差，从而识别出在模型背景下异常的观测值。

7 讨论

基于意外度的框架提供了一种灵活且有原则的方法，用于量化在任何指定分布模型下某个观测值有多么不寻常。我们方法的一个关键特点是，即使假定分布存在设定错误，它仍然有效。这与许多传统异常检测方法形成对比：在那些方法中，尾部的模型设定错误可能导致异常分数严重失真。通过将多变量（或其他复杂）建模问题转化为估计惊异度分布尾部的一元问题，我们将问题简化到一个可以轻松应用经验法和极值技术的领域。

我们的理论结果明确指出，哪些类型的模型误设可以容忍。异常值是相对于假定分布下低密度区域定义的 F。要使经验尾概率估计量可靠，假定模型必须正确识别这些低密度区域的位置，即使它不能准确捕捉尾部的形状。假设2.1将这一要求形式化：真实惊异度与拟合惊异度只需在极端尾部通过严格递增变换相一致。满足这一条件时，引理2.1 保证经验惊异度概率在上尾构成真实尾部概率的有效一致置信带。值得注意的是，这并不要求 F 具有正确的尾部行为，甚至不要求其具有正确的分布形式，只需具有正确的尾部排序。如果这一条件不成立，例如假定模型在数据呈偏态时强加对称性，或在数据呈多峰时强加单峰性，经验估计量便不再提供一致保证。实际应用中，

如此严重的误设-规格应当能够通过探索性数据分析加以识别

。广义帕累托分布方法为估计惊异度尾部概率提供了互补途径。我们的极值结果表明，在广泛的条件下（惊异度尾部呈次高斯、次指数或多项式形式），最大惊奇度会收敛到广义极值分布族中的某个分布。这些条件涵盖了几乎所有常见的统计模型。因此，即使假定分布与真实分布只有较松散的联系，广义帕累托分布近似仍有望提供稳健的尾部估计。也就是说，只要所选阈值深入尾部，且在该点之后，假定分布与真实分布的尾部排序相同，广义帕累托分布方法就应能得出准确的尾部概率。这种情况与经验方法略有不同，因为广义帕累托分布的依据是惊讶度的渐近行为；这意味着，如果 n 足够大，使得该近似足够准确，那么阈值以上的尾部排序也应当正确。

广义帕累托分布的阈值选择仍是一个尚未解决的方法学问题；极值理论文献中的成熟技术（如稳定性图或最近次序统计量启发式方法）可以适用于惊异度数据，但针对这一情境的系统性研究仍有待未来开展。

在实践中，选择经验方法还是广义帕累托分布方法，取决于样本量 n 和所需的显著性水平 α。经验估计量无需调参或估计模型，并能提供有限样本一致误差控制，但当 $n\alpha$ 较小时，结果可能会有较大波动。当 $n\alpha$ 较小时，广义帕累托分布估计量最为有用，因为它可以降低方差，并允许外推至观测最大值之外；但它需要选择阈值并对尾部拟合参数模型，这会增加复杂性，也可能导致模型设定错误。当惊异度分布呈现离散性、尖峰，

或存在其他可能不希望通过参数化尾部模型平滑掉的非平滑特征时，经验方法可能更为合适。在许多情况下，同时使用这两种方法进行稳健性检验或许是明智之举。可以预期，在模型设定错误程度适中的情况下，这两种方法都能表现良好；但如果模型设定错误严重到真实分布与拟合分布的尾部排序存在显著差异，则两种方法都无法保证有效。

意外度框架还具有可解释性和可扩展性方面的优势。惊讶度提供了一种基于模型的单一不太可能性度量，适用于离散、连续和混合分布。这避免了基于距离的方法固有的困难，尤其是在高维或高度偏斜的情形下。在计算方面，该方法效率很高：惊讶度的计算与样本量呈线性关系，经验尾部估计只需排序，而即使对于非常大的数据集，广义帕累托分布拟合也很快。

总之，意外度框架为异常检测提供了一种稳健、可解释且有理论依据的方法。通过将工作模型的设定与尾部概率的估计分开，并结合经验法和极值方法，我们构建了一个灵活的系统，即使在模型设定错误相当严重的情况下也能表现良好。这些应用展示了如何利用这些理念，在复杂的现实世界数据中检测出有意义的异常值；理论结果则阐明了该方法何时以及为何有效。

我们的方法通过 `weird` R 软件包（Hyndman 2026）实现。该软件包提供了易于使用的界面，可通过经验法和广义帕累托分布方法计算惊讶度并估计尾部概率。

软件与可复现性

本文所呈现的所有结果均可使用以下网址提供的数据和代码复现：https://github.com/robjhyndman/Surprisal_理论。分析使用 R 4.5.2 版（R 核心团队 2025）进行，使用了以下 R 软件包：weird（Hyndman 2026）、tsibble（Wang 等，2026）、targets（Landau 2026、2021）、ggplot2（Wickham 等，2026、Wickham2016）以及其他 tidyverse（Wickham 等，2019）软件包。

资助

两位作者均获得澳大利亚研究委员会探索项目 DP250100702 的资助。Rob Hyndman 是澳大利亚研究委员会工业转型培训中心“优化技术、集成方法与应用”（OPTIMA）的成员，该项目编号为 IC200100009。

参考文献

Coles, S.（2001），极值统计建模导论，选自*Springer* 统计学系列，Springer，英国伦敦。

Cover, T. M. 与 Thomas, J. A.（2006），信息论要素，第 2 版，Wiley .

Davies, L. & Gather, U.（n.d.），《多个离群值的识别》，88，782–792。

Good, I. J.（1952），《理性决策》，皇家统计学会会刊：*B*辑，统计方法论14，107–114。

Grubbs, F. E.（1950），《用于检验离群观测值的样本准则》，数学统计年刊21(1)，27–58。

网址：*https://projecteuclid.org/euclid.aoms/1177729885*

人类死亡率数据库（2025），‘加州大学伯克利分校（美国）；马克斯·普朗克人口研究所（德国）’。访问于2025年10月6日。**网址：** *www.mortality.org*

Hyndman, R. J.（1996年），“计算和绘制最高密度区域”，《美国统计学家》50(2), 120–126。

Hyndman, R. J.（2026），*weird*：《使用 *R* 进行异常检测》中的函数和数据集，作者 *Rob J Hyndman*。

URL： *https://pkg.robjhyndman.com/weird/*

Hyndman, R. J.、Kandanaarachchi, S. & Turner, K.（2025），“当 lookout 发现 crackle 时：通过核密度估计进行异常检测”。

网址： *https://robjhyndman.com/publications/lookout2.html*

Kandanaarachchi, S. & Hyndman, R. J.（2022），“用于离群值检测的留一法核密度估计”，《计算与图形统计学杂志》31，586–599。

Landau, W. M. (2021), ‘targets R 包：一种用于可复现性和高性能计算的动态、面向函数的类 Make 声明式流水线工具包’, 《开放源代码软件期刊》6(57), 2959.

网址： *https://doi.org/10.21105/joss.02959*

Landau, W.M.(2026), *targets*：动态、面向函数的类 *Make* 声明式流水线。

网址： *https://cran.r-project.org/package=targets*

R 核心团队（2025），*R*：用于统计计算的语言和环境，R 统计计算基金会，维也纳，奥地利。**网址：** *https://www.R-project.org/*

Stone, J. V.（2022），信息论：教程式导论，Sebtel Press，谢菲尔德，英格兰。

Tribus,M. (1961),《热力学与热静力学：能量、信息和物质状态导论，并附工程应用》，D. Van Nostrand，纽约，美国。

Wang, E., Cook, D., Hyndman, R. J., O'Hara-Wild, M., Smith, T. & Davis, W. (2026),*tsibble*：整洁时序数据框与工具。**网址：** *https://tsibble.tidyverts.org*

Wickham, H. (2016), *ggplot2*：优雅的数据分析图形，Springer-Verlag，纽约，美国。

网址： *https://ggplot2-book.org/*

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. & Yutani, H. (2019),《欢迎来到tidyverse》，《开放源代码软件期刊》4(43)，1686。

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., van den Brand, T. & Posit, PBC (2026), *ggplot2*：使用图形语法创建优雅的数据可视化。**网址：** *https://cran.r-project.org/package=ggplot2*

Wilkinson, L. (2018), '通过分布式聚合可视化大数据离群值'，《*IEEE*可视化与计算机图形学汇刊》24(1),256–266.

Wood, S. N.（2017），广义加性模型：基于*R*的入门，CRC出版社 s.

Wood, S. N.（2025），具有自动平滑估计的混合广义加性模型计算工具。

网址： *https://cran.r-project.org/package=mgcv*

引理的证明2.1

证明。定义 $\mathbb{P}_n(s) = n^{-1}\sum_{i=1}^{n} 1[S_i \leq s]$。暂且假设，在集合 $s \geq s_\star$ 上，$\widehat{\mathbb{P}}_n(s) = \mathbb{P}_n(s)$ 几乎处处成立。如果满足这一条件，那么

$$\begin{aligned}
\Pr\left\{\sup_{s\geq s_\star}\left|\widehat{\mathbb{P}}_n(s) - G(s)\right| > \epsilon\right\} &= \Pr\left\{\sup_{s\geq s_\star}\left|\mathbb{P}_n(s) - G(s)\right| > \epsilon\right\} \\
&\leq \Pr\left\{\sup_{s\in\mathbb{S}}\left|\mathbb{P}_n(s) - G(s)\right| > \epsilon\right\} \\
&= 2e^{-2n\epsilon^2}, \qquad (1)
\end{aligned}$$

其中，最后一个不等式由德沃雷茨基–凯弗–沃尔福威茨定理得出，对每个$\epsilon > 0$都成立。取 $\epsilon = \sqrt{\log(2/\alpha)/(2n)}$，可得

$$\begin{aligned}
\Pr\left\{\sup_{s\in\mathbb{S}}\left|\mathbb{P}_n(s) - G(s)\right| > \epsilon\right\} &\leq 2e^{-2n\epsilon^2} \\
&= 2e^{-2n\left(\frac{\log(2/\alpha)}{2n}\right)} \\
&= 2e^{-\log(2/\alpha)} \\
&= \alpha.
\end{aligned}$$

因此，以至少 $1-\alpha$的概率，$\sup_{s\geq s_\star}\left|\mathbb{P}_n(s) - G(s)\right| \leq \epsilon$。固定一个 $s \geq s_\star$ 并重新整理各项，即得所述不等式

$$\widehat{\mathbb{P}}_n(s) - \epsilon \leq G(s) \leq \widehat{\mathbb{P}}_n(s) + \epsilon.$$

要使上述论证成立，必须在集合 $s \geq s_\star$, $\mathbb{P}_n(s) = \mathbb{P}_n(s)$ 上几乎必然成立。现在我们证明，当且仅当假设2.1 成立时，这一条件才成立。

情况 1： $\Rightarrow$。取 $s \geq s_\star$。考虑随机变量 $1[S_i \leq s] = 1$，并注意到，根据假设假设2.1，必有 $1[h(S_i) \leq h(s)] = 1[S_i \leq s]$ 几乎处处成立。然而，由于 $h(\cdot)$ 严格递增，因此其逆函数也严格递增，所以

（几乎处处）

$$1[h(\widehat{S}_i) \le h(s)] = 1[\widehat{S}_i \le s].$$

因此，对于 $s \ge s_\star$， $\mathbb{P}_n(s) = \mathbb{P}_n(s)$ 且（1）成立。

情况 2： $\Leftarrow$。现在，假设对所有 $s \ge s_\star$ 都有 $\mathbb{P}_n(s) = \mathbb{P}_n(s)$。则集合

$$\mathbb{S} := \{S_1, \ldots, S_n : S_i > s_\star\} \text{ and } \widehat{\mathbb{S}} := \{\widehat{S}_1, \ldots, \widehat{S}_n : \widehat{S}_i > s_\star\}$$

必须满足 $|\mathbb{S}| = |\widehat{\mathbb{S}}|$，因为 $\mathbb{P}_n(s) = \mathbb{P}_n(s)$。因此， $\mathbb{S}$ 与 $\widehat{\mathbb{S}}$ 中的取值必须相同（在元素重新排序的意义下）。因此，存在一个双射，使得对所有$s > s_\star$都有 $S_i = h(S_i)$。

□

B 定理的证明3.1

我们依次证明这三种情况。

B.1 S 具有次高斯尾部

引理 B.1. 假设 $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ 是来自密度 f 的独立同分布样本，并令 $M_n =$最大值$\{s_1, \ldots, s_n\}$，其中 $s_i = -$ 对数 $f(\boldsymbol{y}_i)$。若 $S = -$ 对数 $f(\boldsymbol{Y}) \in \mathbb{R}$，且满足对所有 $\lambda \in \mathbb{R}$ 以及某个 $\nu > 0$,

$$\mathbb{E}\exp\{\lambda(S - \mathbb{E}[S])\} \le \exp\{\lambda^2\nu^2/2\},$$

那么，对于 $s \in \mathbb{R}_+$,

$$\Pr\{M_n - \mathbb{E}[S] \ge s\} \le 1 - \left(1 - e^{-\frac{1}{2}s^2/\nu^2}\right)^n.$$

证明。对于任意 $\lambda > 0$,

$$\begin{aligned}\Pr\{S \ge s\} = \Pr\left\{e^{\lambda(S-\mathbb{E}[S])} \ge e^{\lambda(s-\mathbb{E}[S])}\right\} &\le e^{-\lambda(s-\mu)}\mathbb{E}e^{\lambda(S-\mu)} \\ &\le e^{-\lambda(s-\mu)}e^{\lambda^2\nu^2/2}.\end{aligned}$$

现在，我们可以在 λ 中最大化上述界的右侧，从而得到最优值 $\lambda = (s-\mu)/\{M^2\}$。代入后，我们得到以下界

$$\Pr\{S \geq s\} \leq e^{-\frac{1}{2}(s-\mu)^2/\nu^2}.$$

所述结果现由以下观察即可得出：

$$\begin{aligned}\Pr\left(M_n - \mathbb{E}[S] \geq s\right) = 1 - \Pr\{M_n \leq s\} &= 1 - \Pr\{S \leq s\}^n \\ &= 1 - (1 - \Pr\{S \geq s\})^n \\ &\leq 1 - \left(1 - e^{-\frac{1}{2}s^2/\nu^2}\right)^n.\end{aligned}$$

□

推论 B.1。 若满足假设 *3.1* ，则对所有 $s > 0$,

$$\Pr\left(|M_n - \mathbb{E}[S]| \geq \sqrt{2\nu^2 s} + \sqrt{2\nu^2 \log(2n)}\right) \leq e^{-s}.$$

证明。 这直接由并集界得到：

$$\begin{aligned}\Pr\left(|M_n| \geq \sqrt{2\nu^2 s} + \sqrt{2\nu^2 \log(n)}\right) &\leq n \Pr\left\{|S| \geq \sqrt{2}\nu^2 s + \sqrt{2\nu^2 \log(n)}\right\} \\ &\leq 2ne^{-[\sqrt{2\nu^2 s} + \sqrt{2\nu^2 \log(n)}]^2/(2\nu^2)} \\ &= 2ne^{-\frac{1}{2\nu^2}(2s\nu^2) - \frac{2\nu^2 \log(2n)}{2\nu^2} - \frac{\sqrt{2\nu^2 s}\sqrt{2\nu^2 \log(2n)}}{2\nu^2}} \\ &\leq e^{-\sqrt{\log(2n)s} - s} \\ &\leq e^{-s}\end{aligned}$$

第一个结果由 S 的独立同分布假设和并集界得出，而第二个结果则基于 S 是次高斯这一假设。其余不等式通过重新整理各项得出。

□

推论 B.1 很有用，因为它明确指出，如果惊异度是次高斯的，那么最大惊讶度将服从反韦布尔分布，因此属于极值分布类。

B.2 S 具有指数矩

引理 B.2. 设 $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ 是来自密度 f的独立同分布样本，并令 $M_n = \max\{s_1, \ldots, s_n\}$其中 $s_i = -\log f(\boldsymbol{y}_i)$。若满足假设 *3.2*，则对于 $s \in \mathbb{R}_+$,

$$\Pr\left(M_n - \mathbb{E}[S] \geq s\right) \leq 1 - \begin{cases} (1 - e^{-\frac{1}{2}s^2/\nu^2})^n & \text{if } 0 \leq s \leq \nu^2/b \\ (1 - e^{-s/2b})^n & \text{if } s > \nu^2/b \end{cases}.$$

证明。 结论可通过与引理 B.1类似的论证得到，直到我们对尾界取最大值为止。具体而言，对于任意 $\lambda > 0$,

$$\begin{aligned} \Pr\left\{S - \mathbb{E}[S] \geq s\right\} = \Pr\left\{e^{\lambda(S-\mathbb{E}[S])} \geq e^{\lambda s}\right\} &\leq e^{-\lambda s}\mathbb{E}e^{\lambda(S-\mathbb{E}[S])} \\ &\leq e^{-\lambda s}e^{\lambda^2\nu^2/2}. \end{aligned}$$

现在，我们必须考虑约束优化，以确保右侧存在。具体而言，我们从引理 B.1 的证明中得知，无约束最优值为$\lambda = s/\nu^2$，这对应于集合 $0 \leq t < \nu^2/b$上的约束最优值。然而，当 $s \geq \nu^2/b$ 时，最大值在 $\lambda = 1/b$ 处取得。综合以上结果可得

$$\Pr\left\{S - \mathbb{E}[S] \geq s\right\} \leq \begin{cases} e^{-\frac{1}{2}s^2/\nu^2} & \text{if } 0 \leq s \leq \nu^2/b; \\ e^{-s/2b} & \text{if } s > \nu^2/b. \end{cases}$$

再次注意到以下事实，即可得到所述结果：

$$\begin{aligned}\Pr\{M_n - \mathbb{E}[S] \geq s\} &= 1 - \Pr\{M_n \leq s\} \\ &= 1 - \Pr\{S \leq s\}^n \\ &\leq 1 - \begin{cases} (1 - e^{-\frac{1}{2}s^2/\nu^2})^n & \text{if } 0 \leq s \leq \nu^2/b; \\ (1 - e^{-s/2b})^n & \text{if } s > \nu^2/b. \end{cases}\end{aligned}$$

□

与引理 B.2 类似的结果$\Pr\{M_n - \mathbb{E}[S] \leq s\}$以及 $\Pr\{|M_n - \mathbb{E}[S]| \geq s\}$也同样成立，但需额外乘以因子 2，即：

$$\Pr\{|M_n - \mathbb{E}[S]| \geq s\} \leq 1 - \begin{cases} (1 - 2e^{-\frac{1}{2}s^2/\nu^2})^n & \text{if } 0 \leq s \leq \nu^2/b \\ (1 - 2e^{-s/2b})^n & \text{if } s > \nu^2/b. \end{cases}$$

推论 B.2。 在假设 *3.2*下，对于 $s > \nu^2/b$，当 $n \to \infty$

$$\Pr\left(|M_n - \mathbb{E}[S]| \geq (2b)s + (2b)\log(2n)\right) \leq 1 - e^{-e^{-s}} + o(1/n).$$

证明。 应用假设 3.2 可知

$$\begin{aligned}\Pr\left(|M_n - \mathbb{E}[S]| \geq (2b)s + (2b)\log(2n)\right) &\leq 1 - \left\{1 - 2e^{-[2bs + 2b\log(2n)]/2b}\right\}^n \\ &= 1 - \left(1 - e^{-s}/n\right)^n \\ &= 1 - e^{-e^{-s}} - \left[\left(1 - e^{-s}/n\right)^n - e^{e^{-s}}\right].\end{aligned}$$

注意到，对于所有 $s > 0$，现在即可得出结果。

$$\left[\left(1 - e^{-s}/n\right)^n - e^{-e^{-s}}\right] = o(1) \leq 0.$$

□

推论B.2 很有用，因为它表明，如果惊异度具有次指数尾，那么由此得到的最大惊异度尾概率将收敛于某个极值分布的尾概率。特别地，推论B.2 意味着

$$\sup_{s:s>\nu^2/b}\left|\Pr\left\{|M_n-\mu|\geq\left[2bs+2b\log(2n)\right]\right\}-\left(1-e^{e^{-s}}\right)\right|=o(1).$$

为了说明这意味着 $M_n-\mathbb{E}[S]$ 经过适当标准化后依分布收敛到某个极值分布，我们先回顾:

$$\Pr\{X_n>x\}-\Pr\{X>x\}=1-\Pr\{X_n\leq x\}-[1-\Pr\{X\leq x\}]=\Pr\{X\leq x\}-\Pr\{X_n\leq x\}.$$

因此,

$$\sup_{x\in\mathbb{X}}|\Pr\{X_n\leq x\}-\Pr\{X\leq x\}|=\sup_{x\in\mathbb{X}}|\Pr\{X_n>x\}-\Pr\{X>x\}|,$$

因此，当 X_n 和 X 均为连续变量时，我们看到推论 B.2 表明 $(M_n-\mathbb{E}[S]-2b\log(2n))/2b$ 的累积分布函数收敛到 I 型极值分布（即冈贝尔分布）的累积分布函数；也就是说，利用推论 B.2 及上述结果，可得

$$\sup_{s:s>\nu^2/b}\left|\Pr\left(|M_n-\mu|\leq[2bs+2b\log(2n)]\right\}-e^{e^{-s}}\right|=o(1).$$

这是依分布收敛的一种概念。

B.3 S 具有多项式矩

假设 $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$ 是来自密度 f的独立同分布样本，并令 $M_n=\max\{s_1,\ldots,s_n\}$ 其中 $s_i=-\log f(\boldsymbol{y}_i)$。若满足假设 3.3，则对于 $s>0$ 且满足 $s>C>0$,

$$\Pr\left(|M_n-\mathbb{E}[S]|\geq s\right)\leq 1-(1-C^p/s^p)^n.$$

因此,

$$\sup_{s:s>C}\left|\Pr\left(|M_n-\mathbb{E}[S]|\leq(s\cdot n^{1/p})\right)-e^{-s^{-p}}\right|=o(1).$$

证明。根据马尔可夫不等式，

$$\Pr\{|S-\mathbb{E}[S]|\geq s\}\leq\frac{\mathbb{E}|S|^p}{s^p}\leq\frac{C^p}{s^p}.$$

为推导第一个所述结果，再次注意到

$$\begin{aligned}\Pr\{|M_n-\mathbb{E}[S]|\geq s\}&=1-\Pr\{|M_n-\mathbb{E}[S]|\leq s\}\\&=1-\Pr\{|S-\mathbb{E}[S]|\leq s\}^n\\&=1-(1-\Pr\{|S-\mathbb{E}[S]|\geq s\})^n\\&\leq 1-(1-C^p/s^p)^n.\end{aligned}$$

为推导第二个所述结果，将第一个结果应用于 $t=sn^{1/p}C>1$。显然，当 n 足够大时，该结果成立，由此可得

$$\Pr\{|M_n-\mathbb{E}[S]|\geq t\}\leq 1-\left(1-\frac{C^p}{(Csn^{1/p})^p}\right)^n=1-\left(1-s^{-p}/n\right)^n.$$

现在，回想一下，对于 $x\geq 0$,

$$\lim_{n\to\infty}(1-x/n)^n=e^{-x}.$$

由于对于所有 $x\geq 0$，$x\mapsto e^{-x}$ 都有界，因此可得

$$\sup_{s:s>C}\left|\Pr\{|M_n-\mathbb{E}[S]|\geq t\}-e^{-s^{-p}}\right|=o(1).$$

□