存在未知本底时利用间距设定上限

Lolian Shtembari* 和 Allen Caldwell
Max Planck Institute for Physics, Munich, DE 80805

在存在未知本底的情况下，为拟议过程产生的事件率设定上限，是搜寻稀有过程时经常遇到的问题。基于事件分布中异常大的“间隙”（或间距）的方法，可以为拟议信号分布的率设定上限。本文提出两种新颖的基于间距的方法：“排序间距和”和“互补间隔乘积”，将它们作为检验，并利用合成数据和已发表数据集与现有检验方法进行比较。

arXiv:2303.09520v1 [physics.data-an] 16 Mar 2023

引言

许多旨在发现理论上的稀有过程的实验，可能会遇到所收集的数据不足以宣称发现信号的情况。在这种情况下，数据用于为所考虑的稀有过程产生的事件数设定上限，进而可用于为拟议模型的物理量设定上限。例如，确定弱相互作用大质量粒子（WIMPs）在探测器中与原子碰撞反冲的截面上限，如 CRESST [1] 或 CDMS [2] 实验。相关实验常受到了解不足的背景污染，此时必须依据观测到的事件分布特性，在不进行任何本底扣除的情况下设定信号强度上限。间距统计是超越单纯事件计数、设定信号强度上限的一种方法。

由于目标过程产生的事件分布的预期形状已知，因此可以利用观测到的事件分布与预期分布之间的差异，估算该过程所对应的事件数，并达到所需的置信水平。此类分析通过拟合优度检验进行，并假设在分析窗口内收集到的观测事件数只是一个随机变量的实现，该随机变量服从率 μ未知的泊松分布。对于选定的拟合优度检验，目标是确定与所需置信水平相对应的事件率 μ。

下面简要回顾如何在假设目标过程服从任意事件分布的情况下使用拟合优度检验，以及如何在定义p值时考虑观测事件数的随机性。

接下来，我们将讨论如何利用基于间距的检验给出上限：首先简要回顾最大间隙法和最优区间法 [3]，然后介绍两种新检验，分别基于排序后的间距列表和间距乘积。

* lolian@mpp.mpg.de

使用检验统计量设定上限

已有多种非参数拟合优度检验可用于检验单变量分布。概率积分变换 [4, 5], 使得检验各种截然不同的单变量分布成为可能；它基本上将拟合优度检验简化为简单的均匀性检验。

概率积分变换

假设我们有 n 个样本 $\{y_i\}$，并希望定量检验这些样本是否为已知分布 $f(y)$ 的随机变量，并且是否按照 $f(y)$ 独立同分布。只考虑累积分布函数为$F(y)$的连续分布 $f(y)$ 时，可以通过 $x_i = F(y_i)$ 将样本变换到单位区间 $[0, 1]$ 上。这样一来，问题就简化为检验变换后的样本 $\{x_i\}$ 是否服从标准均匀分布 $\mathcal{U}(0,1)$。

泊松分布与 p值

给定一个由 n 均匀分布的样本 $\{x_i\}$组成的数据集、一个检验统计量 T 以及一个数据函数 $t = g(\{x_i\})$，p值可直接根据 $F_T(t|n)$计算，其中 F_T 是恰有 n 个事件时检验 T 的累积分布函数。此处，p值将分析窗口中的事件数视为固定参数。如果我们将观测到的事件数视为一个具有相应泊松分布的随机变量，就可以通过对所有可能的事件数取平均来修正 p值的定义。对于率为 μ的泊松分布，以及观测到的检验统计量值 $t_{obs} = g(\{x_i\})$，泊松平均 p 值计算如下：

$$1 - p = F_{T,Pois}(t_{obs}|\mu) = \sum_{n=1}^{\infty} F_T(t_{obs}|n) \cdot \frac{\mu^n e^{-\mu}}{n!} \quad (1)$$

其中求和从 $n = 1$ 开始，因为检验统计量通常对 $n = 0$ 没有定义。若未观测到事件，$n = 0$，则无法改进简单的泊松统计量。

这便是在数据集为空的极限情况下对该方法的扩展。

设定上限

给定检验 T 及其分布 $F_{T,\ Pois}$，如上所述，对于一个由 n 个事件组成的给定数据集 $\{x_i\}$，在选定事件率 μ 的情况下，我们可以计算其p值。我们不必进行拟合优度检验，评估事件分布在给定 μ时与均匀分布的拟合程度；而是可以确定使p值达到所需水平的率 μ ，从而确定在达到所需置信水平（CL）时，均匀分布的事件子集所对应的事件率。

例如，对于给定的数据集 $\{x_i\}$ ，在置信水平 CL 下，事件率上限 μ_U满足：

$$1-p=F_{T,Pois}(t_{obs}|\mu_U)=CL \tag{2}$$

间隔统计

给定区间 $[0,1]$, 内均匀分布且有序的 n 个事件$\{x_i\}$（其中 $x_i<x_j\,\forall i<j$），我们可以将 $n+1$ 间距 s 定义为 $s_{j,\ 1}=x_j-x_{j-1}$，其中$x_0=0$且 $x_{n+1}=1$。这些间距是相邻事件之间的间隙，但一般来说，我们也可以考虑高阶间隔： k 阶间距为 $s_{j,\ k}=x_j-x_{j-k}$。

基于这些间距，可以构造检验统计量，从而得到比简单计数（泊松）检验更有竞争力的事件率上限，因为它们不仅考虑分析窗口内数据的总数，还考虑数据的分布，并利用事件密度相对较低的区域来估计事件分布中潜在的均匀成分。

最大间隙法

最大间隙检验已在物理学文献 [3] （更早时也见于统计学文献 [6]）中提出。该检验考察其中最大的间距，以确定事件率上限。检验统计量定义如下：

$$T_{MG}(\{x_i\})=s_{max,1}=\max_i(s_{i,1}) \tag{3}$$

给定事件数量时，$s_{max,\ 1}$ 的分布是已知的 [3, 6]。假设事件均匀分布，较大的假设事件率 μ 会导致观察到较大 $s_{max,\ 1}$ 值的概率较小。如果观测值 $s_{max,\ 1}$ 确实相对于我们的预期较大，则可据此在指定的置信水平下排除 μ 的某些取值。

这种检验定义尤其有用，因为它基本不受单位区间内事件可能聚集的影响；这种聚集会使事件分布偏离标准的均匀分布。当实验受到未知背景的影响，并且概率积分变换是相对于所研究信号的分布进行时，就会出现这种情况：变换后，可能来自信号源的事件将呈均匀分布，而来自未知背景源的事件则不会呈均匀分布，从而在单位区间内形成事件过度密集的区域，以及间距较大的区域。

泊松平均累积分布$F_{MG,\ Pois}$可以通过解析方法轻松计算，Yellin [3]给出了其表达式：

$$F_{MG,Pois}(x|\mu)=\sum_{t=0}^{m}\frac{(tx-\mu)^te^{-tx}}{t!}\left(1+\frac{t}{\mu-tx}\right) \tag{4}$$

其中， m 是 $\le \mu/x$的最大整数。事件率 μ_{MG}的 CL 上限满足：

$$F_{MG,Pois}(s_{max,1}|\mu_{MG})=CL \tag{5}$$

最优区间法

除最大间隙法外，Yellin [3] 还提出了最优区间法。该方法不考察最大的间距，而是考察有序间距之和，即高阶间隔。

最大间隙法将最大一阶间隔的大小与给定事件率 μ 下该间隔内不含事件的预期进行比较。同样，对于高阶间隔，例如 $k=2$，我们可以找出最大的二阶间隔$s_{max,\ 2}$ 并将其大小与给定候选事件率 μ下该间隔内仅含一个事件的预期进行比较。对于数据所允许的任意阶间隔（$k\le n$，其中 n 是事件数），均可进行此类分析；对于给定事件率 μ，将得到 n 个不同的泊松平均 p 值，每种间隔阶数对应一个：

$$s_{max,k}=\max_i(s_{i,k}) \tag{6}$$

$$1-p_k=F_{O,k,Pois}(s_{max,k}|\mu) \tag{7}$$

其中，$F_{O,k}$ 是给定事件数时$s_{max,k}$的累积分布。对于 n 个事件，$F_{O,k}$ ($k>1$)的解析公式尚不清楚，但 Yellin 通过大规模蒙特卡洛计算得出了数值近似，并为较大的 n [7]值给出了近似渐近分布。

为了排除所提出的事件率 μ，可以考察可用的最小 p值:

$$p_{min} = \min_k(p_k) \tag{8}$$

由于 p_{min} 不服从均匀分布，因此需要知道给定事件率 μ 时它的累积分布， $F_{O,\ min}$，最终 p 值为:

$$p_{fin} = F_{O,min}(p_{min}|\mu) \tag{9}$$

$F_{O,\ min}$ 的解析公式尚不清楚，因此通过蒙特卡洛模拟得出数值近似。对该公式求逆，得到事件率的上限 μ_{OI}，使其对应的值不超过给定的 CL:

$$F_{O,min}(p_{min}|\mu_{OI}) = 1 - CL \tag{10}$$

排序间隔和方法

迄今讨论的检验旨在探测事件之间是否存在任何异常大的间隙。对于数量相对较少的待分析事件，可能只有一个这样的异常大间隙，此时最大间隙法或许已经能给出最严格的上限。如果存在多个这样的异常大间隙，且这些间隙彼此相邻，那么就可以将它们合并为一个更高阶间隔，此时最优区间法能给出更有竞争力的上限。

然而，如果这些异常大间隙之间夹杂着许多小间隙，那么最优区间法的灵敏度就会降低，与最大间隙法相比也不会有明显改进：在这些情况下，有序间隔和方法会趋于饱和，并受单个低阶间隔主导。

理想情况下，我们希望将异常大的间隙分别合并，而不包含它们之间的间隙。这种方法可以提高检验的灵敏度，并有望设定更具竞争力的上限。下面介绍这种方法。

给定单位区间$[0,1]$, 中的一组 n 事件 $\{x_i\}$ ，先考虑一阶间隔 $\{s_{i,1}\}$，然后将其按从大到小排序，构成由 $n+1$ 个元素 $\{g_i\}$ 组成的集合（其中 $g_1 > g_2 > ..g_n > g_{n+1}$）。给定这样一组排序后的间距 $\{g_i\}$，可以对其元素求和，以考察高阶间隔。排序间距的第 k阶和 G_k，就是最大的 k 个一阶间隔之和:

$$G_k = \sum_{i=1}^{k} g_i \tag{11}$$

给定 n 个事件 G_k ，最多共有 n 个非平凡的 $n+1$ 一阶间隔，因为所有一阶间隔之和受约束为 1。

现在可以计算给定事件率 μ 下各阶排序间距和的p值:

$$1 - p_k = F_{S,k,Pois}(G_k|\mu) \tag{12}$$

其中，$F_{S,\ k}$ 是给定事件数时 G_k 的累积分布。 n 个事件对应的$F_{S,\ k}$解析公式已知：该公式最早由 Mauldon[8]推导得出。我们采用附录 A 中介绍的另一种方法，独立地重新推导了该公式。

与最优区间法一样，可以将最小p值 p_{min} 用作检验统计量，以排除过大的 μ 值:

$$p_{min} = \min_k(p_k)$$

由于 p_{min} 不再是有效的p值，因此需要知道给定事件率 μ 下它的累积分布，$F_{S,\ min}$，在这种情况下，最终p值的计算方式为:

$$p_{fin} = F_{S,min}(p_{min}|\mu) \tag{13}$$

其解析公式$F_{S,\ min}$ 未知，因此通过蒙特卡洛模拟获得数值近似。对公式求逆，需要找到满足以下条件的 CL 上限 μ_S：

$$F_{S,min}(p_{min}|\mu_S) = 1 - CL \tag{14}$$

虽然针对 n 的$F_{S,\ k}$ 解析公式已知，但其表现不佳，因为它依赖于极大数值之间的迭代差值（随着 n 增大），在计算机上使用时容易发生灾难性数值消去。为避免这些问题，我们在适当的网格上以高数值精度计算函数值，从而构建可靠的单调三次样条插值[9] ，可在默认的 64 位浮点运算下使用。目前，这些插值已制表至 $n = 700$ ，可用于估算最高达 $\mu \lesssim 580$的事件率，因此也相应地限制了可分析的事件数量。用于计算检验统计量和上限的加速表及代码，已在 Julia 的 SpacingStatistics.jl [10]软件包中提供。

互补间隔乘积法

最后，我们提出另一种检验统计量，它将事件之间的间距结合起来，而不考虑它们的相对位置。我们这一提议的基础是一个最初由

Moran [11] 提出的检验，该检验由所有相邻事件之间间距的乘积构成：

$$T(\{x_i\}) = -\sum_{i=1}^{n+1} \log(s_{i,1}) \tag{15}$$

其中 $s_{i,1}$ 为一阶间隔， n 是单位区间内的事件数。该检验最初被提出作为一种拟合优度检验，用于检测数据簇，以对抗均匀分布的零假设：小间距的存在会使间距的乘积整体趋向更极端的值。

为了使该检验能够检测大间距的存在，我们考虑每个一阶间隔的补数，并将它们相乘：

$$C(\{x_i\}) = -\sum_{i=1}^{n+1} \log(1 - s_{i,1}) \tag{16}$$

对于固定事件数 n， F_C（$C|n$），该量的分布尚无解析形式，但我们基于蒙特卡洛模拟推导了一个数值近似，并将其制成表格，适用于 $n \leq 10^4$。此外，由于该检验的定义形式为$\sum g$（$s_{i,1}$），通常可以推导其渐近分布，正如 Darling [12]和 LeCam[13], 所述；我们采用他们的方法，求得了 C 的极限分布。 C 在 $n \to \infty$ 时的渐近分布为：

$$f_C(C|n \to \infty) = \mathcal{N}(n \cdot \mu_\infty, n \cdot \sigma_\infty) \tag{17}$$

其中， μ_∞ 和 σ_∞ 的公式见附录 B。根据这一结果，我们可以用它来估算任意较大 n 值对应的检验统计量，从而有效地将该检验及其上限计算的适用范围扩展到大量事件的情形。

对于给定事件率 μ ，该检验的泊松平均 p 值为：

$$F_{C,Pois}(C|\mu) = \sum_{n=1}^{\infty} F_C(C|n) \cdot \frac{\mu^n e^{-\mu}}{n!} \tag{18}$$

对该公式求逆，可得 CL上限 μ_C，满足：

$$F_{C,Pois}(C|\mu_C) = CL \tag{19}$$

表列的分布以及用于计算检验统计量和上限的代码，可在 Julia 的 SpacingStatistics.jl [10]程序包中获取。

性能比较

我们通过模拟示例对所提出方法的性能进行了广泛研究。在检验中，我们引入了形状和强度各异的背景。

下文将我们的方法与标准泊松检验以及最优区间法进行比较。最优区间法被认为是在受未知背景影响的实验中设定上限的最先进方法。

无背景实验

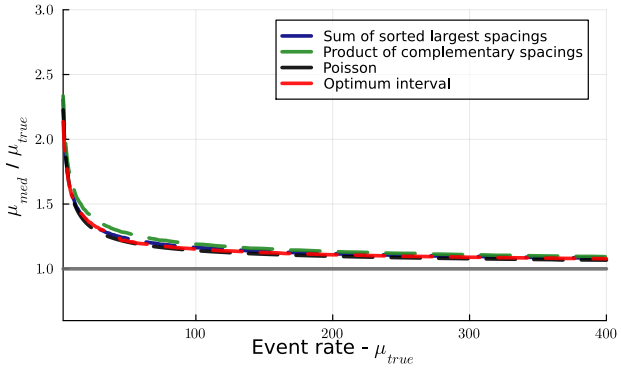

图1. 中位数 $CL = 0.90$ 上限按模拟中使用的事件率归一化；数据根据均匀分布生成（无背景）。

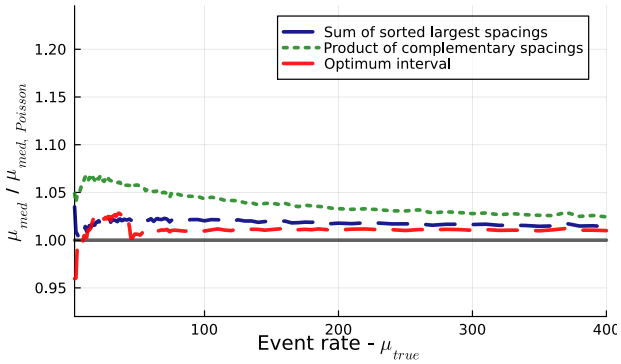

图2. 中位数 $CL = 0.90$ 上限按泊松检验的结果归一化；数据根据纯均匀分布生成（无背景）。

我们首先考虑实验中没有额外背景污染的情况，以估算不同方法的基线。为简便起见，我们选择均匀分布来生成事件；该分布与所有检验的零假设一致，并改变数据生成时使用的事件率。图1展示了使用不同方法得到的事件率 $CL = 0.90$ 上限的中位数。为了更清楚地辨别各方法效率之间的差异

我们可以查看相对于泊松极限归一化后的结果，如图2所示。我们注意到，在这一基线情景下，泊松检验表现最佳，这与预期一致。尽管如此，泊松检验的结果并未显著优于其他任何方法。

指数背景单独实验

接下来，我们研究模拟中存在背景、且信号强度与之相比可以忽略的情况：这模拟了信号可能不存在的稀有过程搜索。在实验中，我们直接在累积空间中生成数据（因此始终假设信号分布是平坦的，即零假设），并首先考虑单位区间 $[0,1]$ 上截断的、率为 0.1 的指数背景。在暗物质搜索实验中，事件分布经过累积空间变换后，往往会在分析窗口的一端形成峰值，并带有迅速衰减的尾部。图3报告了测得事件率中位数 $CL = 0.90$上限相对于注入的背景事件率的归一化结果（下图），以及我们的方法与最优区间检验结果之比（上图）。我们省略了泊松极限，因为它只会随事件总数缩放，且竞争力不足。分析结果可见，面对峰值相对明显的事件分布时，所有方法的表现都相近，与无背景情形一样。所有方法都能滤除大部分背景贡献，并重建较小的总体事件率。对于较小的注入背景率 (≤ 100)，非局域方法（排序间距和与互补间隔乘积）能够给出严格 $5-10\%$ 的上限。随着背景率增大 (≥ 200)，排序间距和检验的表现与最优区间检验相当，而互补间距乘积的结果则最差可达 $5-10\%$。

背景与信号混合

现在我们研究可检测的信号分布受到背景污染的情况。在模拟中，我们将考虑本底信号比为 5，即背景事件率是信号事件率的 5 倍。由于我们直接在累积空间中进行分析，因此始终假设信号分布均匀。背景事件分布的支撑集覆盖分析窗口的四分之一（也就是说，背景事件占据单位区间的四分之一），我们考虑形状各异的分布：具体而言，背景分布是由一个或多个均匀分布混合而成，其总宽度之和为 0.25。

图4左栏展示了不同的背景分布选择，以及由此得到的中位数 $CL = 0.90$

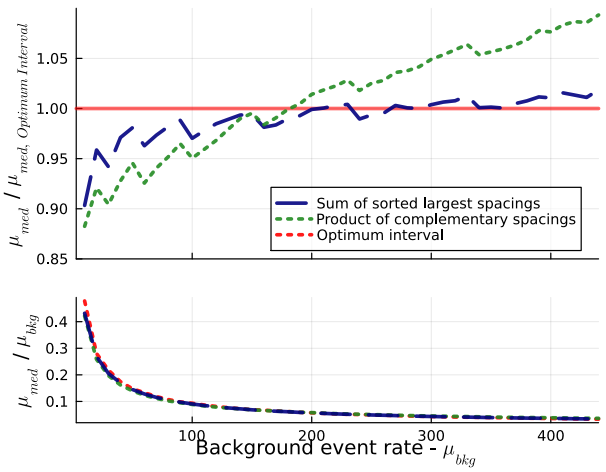

图3. 中位数 $CL = 0.90$ 上限相对于最优区间结果的归一化值；数据按照率为 0.1 的指数分布生成（仅背景）。

中栏展示了不同方法得到的上限，并以最优区间的结果进行归一化。

如果背景分布完全集中在一个区域，且该区域位于分析窗口的一端（如图4第一行所示），那么生成的事件分布中就会出现一段连续的低密度区域。正如前文所述，这对最优区间法来说是最理想的情况。结果也印证了这一预期：最优区间法的结果最多比其他方法好 20%。

当我们将背景分布移至分析窗口中间，甚至将其拆分成两个或更多峰时，会发现我们提出的检验更加灵敏，能够给出更有竞争力的上限。对于双峰背景分布，平均而言，上限可以比最优区间法低20%；对于五峰背景分布，这一优势可提升至40%。我们提出的检验（图4中间列）的表现，源于它们对事件密度较低的大区域十分敏感，无论这些区域的数量或位置如何。仅有两个背景峰时，就有可能得到比最优区间法好20%甚至40%的上限，因为事件密度较低的区域还可以进一步细分。

考虑信号极弱或不存在的情况，我们分析了若事件仅按背景分布分布时所得的限制。模拟结果显示在图4右侧一列中。可以看到，无论背景分布的形状如何，排序间距和与互补间距乘积法所得的90%置信水平中位数结果始终小于最优区间法的对应结果；随着无事件区域数量增加，限制的改进幅度最高可达3倍

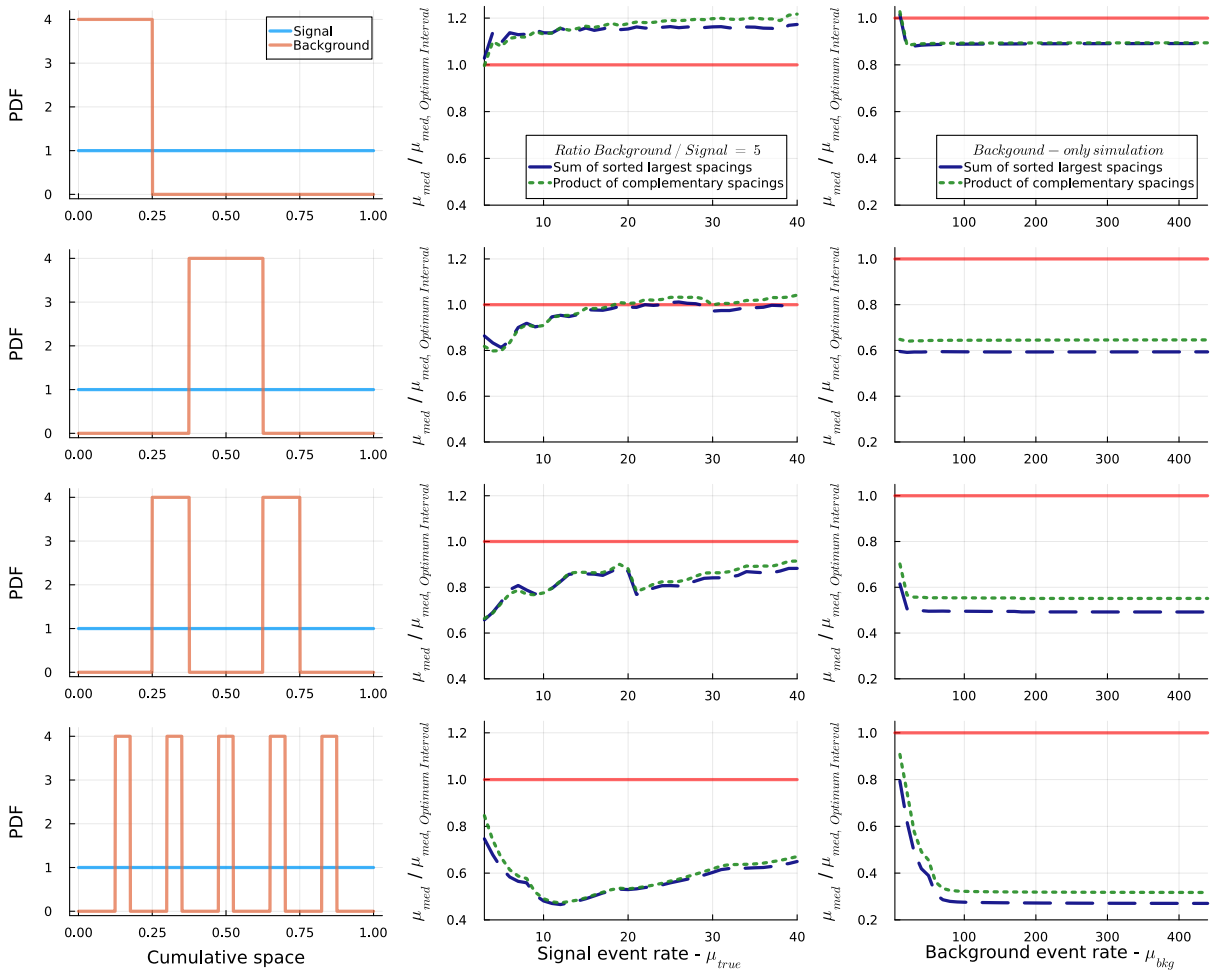

图4. 累积空间中不同事件分布下限值设定方法的比较：（左列）背景和信号的事件分布；（中列）对于背景与信号混合比例为 $\mu_{bkg}/\mu_{sig} = 5$ 的模拟事件分布， $CL = 0.90$ 上限相对于最优区间结果的中位数；（右列）对于纯背景型事件分布（ $\mu_{bkg}/\mu_{sig} = \infty$），90%置信水平上限相对于最优区间结果的中位数。

增加。

多峰背景分布的情况很值得关注，尤其是当其具有明确且相对狭窄的峰时，因为这类似于一种实验情形：人们试图测量三体或多体衰变的事件率。此类衰变的预期能谱相对平坦，可能受到峰状背景分布的污染，而这些分布代表了具有标准模型对应过程的反应。我们提出的方法非常适合解决这些问题，因为它们能够滤除来自这些“峰”的贡献，并估计底层的“平坦”事件率，而无需在分析中引入额外参数（从而使结果产生偏差），以修改、限制或分割感兴趣区域来排除峰状背景。

与似然比检验的比较

作为进一步的示例，我们比较了迄今讨论过的非参数检验与似然比（LR）检验在具有峰状背景时的效率。我们考虑一种高斯本底，其事件率是信号事件率的十倍。将背景的原始分布族输入似然比方法，同时将位置、宽度和本底信号比设为自由参数。图5展示了所用的背景分布，以及在不同信号事件率下得到的中位数 $CL = 0.90$ 上限。查看结果可见，似然比方法给出的上限最严格；这并不令人意外，因为分析中纳入了部分背景信息。基于间距的非参数检验所设定的上限虽然更保守，但仍与似然比方法的结果相当接近；我们的方法所得结果

在低事件率时仅比似然比方法高出不超过20%，在高事件率时上限最多高出10%。此外，非参数检验无需对背景形状作任何假设。

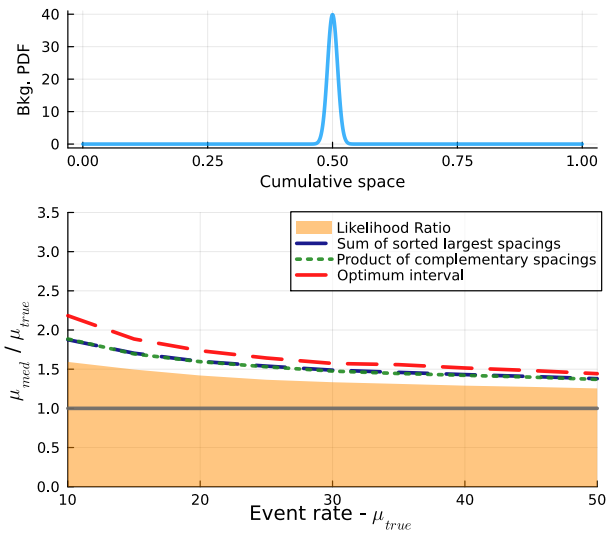

图5.（上）累积空间中的背景分布：高斯分布，参数为 $\mu = 0.5$ 和 $\sigma = 0.01$；（下）归一化至信号事件率的中位数 $CL = 0.90$ 上限。

CRESST数据示例

我们现在复现 CRESST 合作组 [1]的分析，以确定弱相互作用大质量粒子（WIMP）的截面上限。我们使用 CRESST 合作组最近发布的公开数据集 [14]，该数据集 [14] 附带了有关其 $CaWO_4$ 靶材装置的能量分辨率和探测效率的信息。我们分析的数据如图6（上）所示。

给定特定的弱相互作用大质量粒子质量，并假设一种 WIMP 速度分布，即可计算感兴趣能量范围 $[E_{min}, E_{max}]$内的微分率 $\frac{dN}{dE}$。将整个能量范围内微分率的积分记为 $\Lambda = \int_{E_{min}}^{E_{max}} \frac{dN}{dE} dE$，对于一组有序事件 $\{E_i\}$，概率积分变换为：

$$x_i = \frac{1}{\Lambda} \cdot \int_{E_{min}}^{E_i} \frac{dN}{dE} dE \tag{20}$$

从而得到单位区间$[0, 1]$内的一组有序事件 $\{x_i\}$。

图 6（下）展示了使用两种不同弱相互作用大质量粒子质量下计算得到的信号分布进行变换后，累积空间中的事件分布。除细微差异外，这两个数据集非常

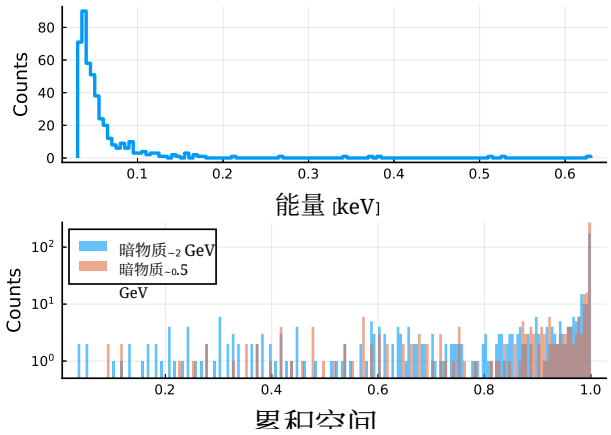

图 6。（上）CRESST 数据的直方图 [14]，由粒子与 $CaWO_4$ 晶体相互作用产生的能量沉积组成；（下）使用两种假设的弱相互作用大质量粒子质量对应的信号分布变换后的数据直方图：0.5 和 2Gev/c^2

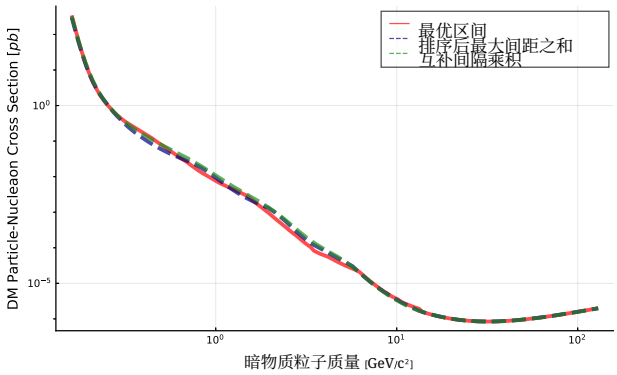

图 7. $CL = 0.90$ 用不同检验计算得到的弱相互作用大质量粒子-核子截面上限随弱相互作用大质量粒子质量的变化。

相似，显示出一个在 1 附近极度尖锐的分布，而在单位区间的其余部分，事件分布几乎呈线性。

图 7 展示了我们使用自身方法计算的截面上限 $CL = 0.90$，以及使用最优区间法计算的上限；该方法由 CRESST 合作组正式采用，所得结果与正式发表的上限相符。

比较我们的计算结果后可以发现，它们之间并无显著差异。为了更清楚地了解各项结果之间的差别，我们将使用这些方法得到的上限归一化为官方上限（使用最优区间法得到），如图 8 所示。我们发现，对于给定数据，互补间距乘积法得到的上限平均高出 25% 至 50%；而排序间距和法则呈现振荡变化：在弱相互作用大质量粒子质量较低时，上限最多可低 30%；在质量约为 ~ 5 Gev/c^2 时，上限则最多可高 40%。

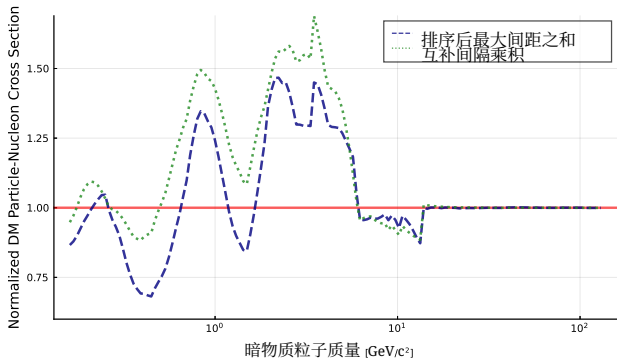

图 8. $CL = 0.90$ 采用我们提出的方法计算得到的弱相互作用大质量粒子—核子截面上限，相对于最优区间结果的归一化值，随弱相互作用大质量粒子质量的变化。

最后，对于弱相互作用大质量粒子质量为 $\geq$ 20 Gev/c^2，的情形，所有方法均达到饱和并给出相同结果，重建出的信号事件率为 $\sim$ 2.3 个事件，对应于分析窗口为空时泊松检验的 $CL = 0.90$ 上限。

这个基于已发表数据集的示例以及我们的性能比较结果表明，在存在未知背景的情况下设定上限时，通常并不存在“最佳”检验统计量，但结果高度依赖于事件的实际分布。

结论

本文提出了两种在存在未知本底时设定上限的新方法，即排序间距和法和互补间距乘积法，并探讨它们相对于最优区间法的性能。所提出的检验利用分析窗口中事件密度较低的区域，无论这些区域的数量多少或彼此之间的位置如何，以此估计累积空间中的底层均匀事件分布。这些特点使我们提出的检验成为稀有过程搜索分析中可行的替代方法，有助于为感兴趣的参数设定具有竞争力的上限，尤其是在面对峰状多峰本底时。

致谢

我们感谢 Heerak Banerjee 博士、Oliver Schulz 博士和 Nahuel Ferreiro Iachellini 博士，他们提出的宝贵讨论意见和建议促成了本工作的最终完成。

附录 A

给定区间 $[0, \mu]$ 内的 $m+2$ 有序样本 $\{x_{(i)}\}$，其中 $x_{(0)} = 0$ 且 $x_{(m+1)} = \mu$，我们可以将 $m+1$间距 $\{s_j\}$ 定义为 $s_j = x_{(j)} - x_{(j-1)}$，并定义间距的有序集合 $\{s_{(j)}\}$，使得 $s_{(j)} < s_{(k)} \forall j < k$。

$s_{(1)}$ 的分布已知:

$$P(s_{(1)} = x|m, \mu) = \frac{m(m+1)}{\mu}\left[1 - \frac{x(m+1)}{\mu}\right]^{m-1} \tag{21}$$

$s_{(1)}$ 和 $s_{(2)}$ 的联合分布可写为:

$$P(s_{(1)} = x, s_{(2)} = y|m, \mu) = \\ = P(s_{(1)} = x|m, \mu) \cdot P(s_{(2)} = y|m, \mu, s_{(1)} = x) \tag{22}$$

给定有序集合 $\{s_{(j)}\}$，从所有间距中减去 $s_{(1)}$，即可将集合缩减为仅含有 m 个元素 $s_{(2)} - s_{(1)}$、.....、$s_{(m+1)} - s_{(1)}$，其总和为 $\mu - (m+1) \cdot s_{(1)}$。对于这个缩减后的集合，$s_{(2)} - s_{(1)}$ 是新的最小元素，其分布由方程（21）给出。因此，我们可以将 $P(s_{(1)} = x, \ s_{(2)} = y)$ 改写如下:

$$P(s_{(1)} = x, s_{(2)} = y|m, \mu) = \\ = P(s_{(1)} = x|m, \mu) \cdot P(s_{(1)} = y - x|m - 1, \mu - (m+1)x) \tag{23}$$

给定 $P(s_{(1)} = x, \ s_{(2)} = y)$，我们可以在其支撑集上积分，以边缘化 $P(s_{(1)} + s_{(2)} = z)$。

我们将最小的 k 个间距之和定义为 S_k:$S_k = \sum_{j=1}^{k} s_{(j)}$。$P(s_{(1)} = x, \ g_k = y)$ 的联合分布可以表示为:

$$P(s_{(1)} = x, S_k = y|m, \mu) = \\ = P(s_{(1)} = x|m, \mu) \cdot P(S_k = y|m, \mu, s_{(1)} = x) = \\ = P(s_{(1)} = x|m, \mu) \cdot P(S_{k-1} = y - kx|m - 1, \mu - (m+1)x) \tag{24}$$

假设我们知道 $P(s_{k-1} = x)$，就可以对$s_{(1)}$ 进行边缘化，从而得到 g_k的分布。这表明 S_{k-1} 和 S_k的分布之间存在递归关系，使我们能够提出关于 S_k 分布的假设，并通过归纳法加以证明。我们得到的 S_k 分布为:

$$P(S_k = x|n, 1) = \\ = A(k, n) \cdot \sum_{j=1}^{k} a(j, k)\left[1 - \left(\frac{n+2-j}{k+1-j}\right)x\right]^{n-1} h(x, j, k, n) \tag{25}$$

其中:

$$A(k,n)=\frac{n(n+1)!}{(n+1-k)^{k-1}(n+1-k)!} \tag{26}$$

$$a(j,k)=\frac{(-1)^{j-1}(k+1-j)^{k-2}}{(k-j)!(j-1)!} \tag{27}$$

$$h(x,j,k,n)=H(x)-H\left(x-\frac{k+1-j}{n+2-j}\right) \tag{28}$$

其中 $H(x)$ 是赫维赛德阶跃函数。

最后，由于单位区间内的所有间距 $s_{(j)}$ 之和为 1， $m+1-k$ 个最小间距之和就是 1 减去 k $G_k=$ 个最大间距之和。因此，其分布为: $\sum_{j=m+2-k}^{m+1} s_{(j)}$ g_k

$$P(G_k=x|m,1)=P(S_{m+1-k}=1-x|m,1) \tag{29}$$

附录 B

利用勒卡姆定理 [13], ，我们能够推导出 PCS 检验统计量的渐近分布。PCS 在样本数$n\to\infty$ 趋于无穷时的渐近分布为:

$$f_C(C|n\to\infty)=\mathcal{N}(n\cdot\mu_\infty,\,n\cdot\sigma_\infty) \tag{30}$$

其中参数如下:

$$\mu_\infty(x)=e^{-x}\left[\mathrm{E}_1(-x)-2i\pi\right] \tag{31}$$

$$\sigma_\infty^2(x)=e^{-x}\left[2A(x)-4i\pi B(x)-2xe^{-x}\mathrm{E}_1(-x)\right]- \\ -1-(x^2+1)e^{-2x}C(x) \tag{32}$$

其中:

$$A(x)=xF\left(\frac{[1,1,1]}{[2,2,2]};x\right)+ \\ +\frac{\ln^2(-x)}{2}+\gamma\ln(-x)+\frac{\pi^2}{12}+\frac{\gamma^2}{2} \tag{33}$$

$$B(x)=\gamma+\ln(x)+x \tag{34}$$

$$C(x)=\left[e^{-x}\mathrm{E}_1(-x)\right]^2+4i\pi e^{-x}\mathrm{E}_1(-x)-4\pi \tag{35}$$

其中， F 是超几何函数， γ 是欧拉–马歇罗尼常数。

[1] CRESST，M. Bravin *et al.*，Astropart. Phys. **12**，107（1999）。[2] CDMS，D. Abrams *et al.*，Phys. Rev. D **66**，122003（2002）。[3] S. Yellin，Phys. Rev. D **66**，032005（2002）。[4] K. Pearson，Biometrika **1**，390（1933）。[5] K. Pearson，Biometrika **25**，379（1933）。[6] R. A. Fisher，伦敦皇家学会会刊：A辑 **125**，54（1929）。[7] S. Yellin，arXiv（2008），0709.2701。[8] J. G. Mauldon，剑桥哲学学会数学会刊 **47**，331–336（1951）。

[9] F. N. Fritsch 和 R. E. Carlson，《SIAM 数值分析杂志》**17**，238（1980）。[10] L. Shtembari，Spacingstatistics.jl，github.com/bat/ SpacingStatistics.jl/tree/dev。[11] P. A. P. Moran，皇家统计学会杂志：B辑（方法论）**13**，147（2018）。[12] D. A. Darling，数学统计年刊**24**，239（1953）。[13] L. L. Cam，巴黎大学统计学院出版物**7**，7（1958）。[14] CRESST 合作组，A. H. Abdelhameed *et al.*，Phys. Rev. D **100**，102002（2019）。