自行车运动员的心脏难题

Andrew Nugent[a]、**Yi Ting Loo**[a]和 **Jack Buckingham**[a]

[a]现实世界系统数学 II 博士培训中心，华威大学

本文稿已向公众开放，采用知识共享署名许可协议：https://creativecommons.org/licens es/by/4.0/

心律失常是心律异常，由传导系统的问题引起，并导致心搏不规则。越来越多的证据表明，频繁进行耐力运动训练会增加心律失常风险。心律失常通过心电图（ECG）进行诊断，但运动员在锻炼时通常无法进行心电图检查。Crickles 先前的研究探讨了常见心率数据在检测心律失常迹象方面的效用。研究假设，一种称为“间断性”的特征可能是心律失常的表现；它指运动员承受体力负荷时心率出现跳变。研究发现，“有间隙”活动所占比例与有关心律问题的问卷回答之间存在相关性。我们在这一指标的基础上，探索检测心率数据中尖峰的各种方法，从而通过尖峰率描述活动中的不规则程度。我们首先比较了这些方法在模拟数据上的表现，发现使用移动平均进行平滑，并对残差设置固定阈值，效果最好。随后，我们将该方法应用于 Crickles 提供的 168 名运动员的实数数据，但未发现尖峰率与问卷回答之间存在显著相关性。然而，只考虑心率超过每分钟 160 次（bpm）时出现的尖峰，则发现了显著相关性。这支持了以下假设：只有高心率时出现的跳变才有助于识别心律失常，也表明需要进一步研究更好的指标，以刻画心率数据的特征。

心律失常 | 心率数据 |尖峰检测 |泊松过程

arXiv:2601.02011v1 [stat.AP] 5 Jan 2026

1. 引言

心律失常是指心脏节律异常，其中最常见的类型是心房颤动（AF）（1）。通过在心电图（ECG）检查期间观察到心房颤动发作，即可确诊（2）；患者佩戴便携式心电图记录仪有助于诊断（1、3）。适度的体育锻炼与心房颤动风险降低有关，但有一些证据表明，剧烈的耐力运动会增加

个人风险（4、5）。因此，人们开始尝试使用运动员佩戴的智能手表中的心电图，记录其运动信息并检测心房颤动（6、7）。由于配备心电图功能的设备可能价格昂贵，若能改为判断常见的心率数据是否包含心房颤动的相关迹象，或许会更有益。

Crickles 项目（8）旨在为运动员提供心脏负荷分析，该分析利用其 Strava 活动数据计算得出，并告知他们相较于其他运动员的训练负荷。这些数据已被 Ian Green 和 Mark Dayer 用于研究能否利用心率数据检测心律问题（9）的存在。他们假设，高心率下的跳变（例如从 180 次/分钟跳至 195 次/分钟后立即再次下降），或跳至不切实际的高值，可能具有信息价值。由于运动员的心率

监测器会持续对心肌中的当前活动进行采样，因此它可能会受到心律失常的影响。Green 等人并未声称此类跳变直接表明一次心房颤动发作，也未声称它们能准确反映运动员的心率；他们认为，这些观测到的跳变是心率监测器产生的错误。他们的核心假设是，此类错误的存在和/或频率仍可能具有信息价值。

在（9）中，每项活动根据两个关键特征被归类为“规则”“不明确”“检查_绑带”或“不规则”。最受关注的特征是图1所示的“间断性”：如果一项活动的观测心率读数范围内存在超过阈值的间隙，则该活动被视为有间隙。此阈值设定为

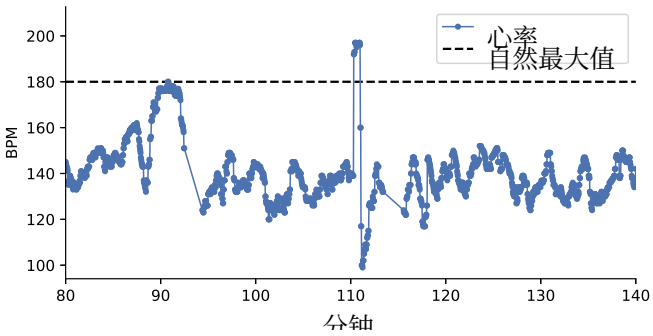


图 1。时间序列显示了有间隙的心率，因为心率超过了自然最大值（虚线）。据推测，这些间隙是由异常心脏电信号导致心率监测器出现的瞬时异常，因此，运动员出现这些间隙的频率升高可能与心律失常有关。

运动员的模态心率与乳酸阈值心率（LTHR）中的较高值，用于指示运动员的心率何时被认为达到

重要性声明

近年来越来越多的证据表明，经常参加高强度耐力训练的人患心房颤动（AF）的风险较高，这是一种心律不规则的类型。尽管可以通过心电图（ECG）准确诊断心房颤动，但运动员佩戴的常见运动设备只能在活动期间收集以每分钟心跳次数（BPM）表示的心率数据，通常每秒采样一次，从而生成时间序列。本项目分析这些时间序列的特征及其与心房颤动诊断之间的相关性，旨在为运动员提供能够提示心律不规则预警信号的有用指标，从而促使其更早诊断并随后接受心房颤动治疗。

未声明存在利益冲突。

较高的范围。可以通过在整个活动过程中将心率标记为“已访问”或“未访问”来检测间断性。如果存在一个高于某个未访问心率的已访问心率，并且二者之间超过阈值，则可观察到一个间隙。由于检测间断性的过程使用了整个时间序列中的心率分布，因此它能够简单地以布尔指标表示整个活动是否存在间断性。心率读数的较低端也可能存在间断性，这可以采用上述过程的镜像方法进行检测；但由于运动员此时并未承受体力负荷，因此假定这些间隙并不能表明存在心律失常。按照设计，间断性检测侧重于心率较高端发生的跳跃，因此可能忽略心率处于正常或相对较低水平时出现的任何有信息价值的行为。

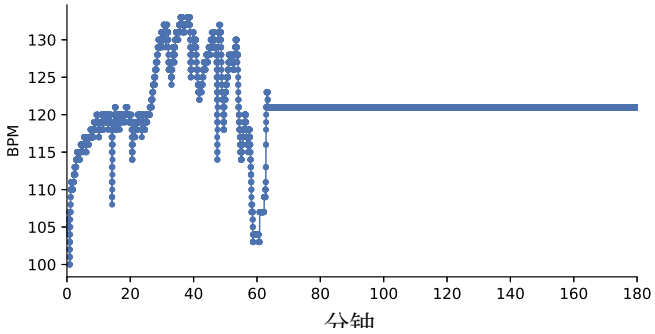


图2. 时间序列显示心率滞留，即心率在很长一段时间内保持在每分钟120次；这在生理上不太可能，尤其是在耐力训练期间，表明心率监测器可能存在错误。

对于第二个特征，如果心率至少连续90秒保持不变（如图2所示）则称该活动为“滞留”。这表明监测器未能正确采集心率。随后根据表1对活动进行分类。

表1. 活动分类

		非滞留	滞留
有间隙	仅高端	不规则	检查_带
	仅低端	不明确	
	两端均有	不明确	
非间断		规则	

被归类为“检查_带”或“不明确”的活动将不纳入考虑。由于运动员通常会记录大量活动，因此这并未造成问题。对于每位运动员，不规则比例被定义为剩余活动中被归类为不规则活动的百分比。

文献（9）中的研究旨在调查运动员的不规则比例与其心脏健康之间的关系。Crickles 用户受邀参与一项问卷调查，其中的关键问题是

> 您目前或过去是否有过心律问题?

此问题的“是/否”回答用于将运动员分为确诊患有心律失常和未确诊患有心律失常两类。Crickles 数据集仅保留了以下运动员：提供了调查回答的运动员。随后进一步筛选数据集，仅考虑自行车运动员，并假设他们会使用胸带监测心率。之后发现，不规则比例与心脏健康调查的回答之间存在 0.24 的统计显著相关性。

迄今为止，针对这一问题的研究通过将整段活动标记为“有间隙”或“非间断”，来描述时间序列中的跳变。我们旨在进一步拓展这项研究，更详细地刻画每项活动／每位运动员的不规则程度，以加强此前观察到的相关性。我们首先开发“尖峰检测”算法，以在心率数据中定位各处跳变，不受心率读数范围的影响，并提取尖峰高度、尖峰出现时的心率等可能重要的特征。然后，我们推断每位运动员的尖峰发生率，并检验其与报告的心律问题之间是否存在相关性。

2. 方法

本节首先介绍多种检测心率数据中尖峰的方法，包括平滑、小波变换和阈值处理。

我们接着介绍一种模拟心率数据的方法。由于针对 Crickles 可用数据的隐私控制严格限制了我们自行访问数据，因此创建模拟数据至关重要。进行探索性分析时，我们可以访问少量样本时间序列；但要利用更大的数据集，则需要 Crickles 团队运行分析并返回结果。这一限制必然影响了我们的方法：我们先开发了多种尖峰检测方法并调整其参数，随后在模拟数据上比较这些方法的表现，以选出一种应用于真实数据集的方法。下文将概述我们的方法及比较过程，最后介绍我们对检测到的尖峰所进行的推断。

A. 尖峰检测方法。一种可用于检测时间序列数据中尖峰的方法，是先对时间序列进行去趋势处理（10），例如，从原始数据中减去平滑后的数据。然后，可以对残差设定阈值，残差超过该阈值的时刻即记为尖峰。可根据平滑后的心率时间序列，找出尖峰出现时对应的心率。还可以通过从尖峰出现时的原始心率数据中减去平滑后的心率，计算尖峰的高度。

A.1. *平滑方法。*为了平滑心率时间序列，一种方法是使用简单移动平均。设 $\{X(t):\ t=1,\ldots,T\}$ 为心率时间序列，将平滑后的时间序列 $X(t)$ 定义为

$$\overline{X}(t) := \frac{1}{w}\sum_{k=-i}^{i} X(t+k) \qquad [1]$$

其中，$w = 2i+1$ 称为宽度。

一种更复杂的平滑方法是在离散小波基中寻找稀疏表示——这是一种广泛用于信号去噪的技术（11、12）。该方法的数学理论在补充材料的第S1节C 中进行了说明。

A.2. 阈值法。一种简单的阈值法是通过计算残差的第95百分位数并乘以一个常数来设定固定阈值。

另一种方法是基于滚动窗口对残差设置自适应阈值（13），而不是根据全部残差计算固定阈值。对于残差序列中的每个时间 t，考虑一个宽度为 w、以时间 t 为中心的窗口，并通过以下方式计算阈值 Thresh(t):

$$\text{Thresh}(t) = \text{offset} + \mu(t) + \text{constant} * \sigma(t) \qquad [2]$$

其中，$\mu(t)$ 和 $\sigma(t)$ 分别是窗口内残差的均值和标准差。引入偏移量并将其设为较小的值 0.01，是为了防止在心率时间序列存在缺失数据的区域检测到尖峰。事实上，这些区域中的标准差 $\sigma(t)$ 应为零，但有限数值精度导致的波动意味着这里的残差并不完全为零。

A.3. 连续小波变换。前面讨论的方法是在对时间序列进行去趋势后，通过对残差进行阈值处理来检测尖峰。另一种方法则使用连续小波变换（CWT）。宽泛地说，CWT 衡量信号在不同时间尺度和位置上与给定“母小波”形状的相似度。通过对适当的小时间尺度下的这些系数进行阈值处理，我们可以识别时间序列中类似尖峰的位置。同样，该阈值可以是固定的，也可以是自适应的。CWT 与前文提到的离散小波变换密切相关，这两种变换以及由其衍生的尖峰检测方法，都在补充材料的第 S1 节中有更详细的介绍。

B. 模拟算法。我们的模拟模型借鉴了现有的电价模型，因为已知电价也会以类似的方式出现尖峰（14，15）。这些模型将这一过程拆分为多个组成部分，每个部分旨在捕捉数据的不同特征，其中之一是短暂的突发尖峰。采用类似的方法，我们考虑三个组成部分，并按顺序进行模拟：

1. 基础心率（X_t） 2. 额外的噪声过程（Y_t） 3. 尖峰过程（Z_t）

所有变量均为一维，单位为每分钟心跳次数。

基础心率决定时间序列的总体形态，也可用于纳入间歇训练等不同活动。 X_t 代表该

运动员。我们使用形式如下的随机微分方程对 X_t 建模

$$dX_t = -V'(X_t, t)\,dt + \sigma\,dW_t^{(1)} \qquad [3]$$

其中，$V(x,\ t)$ 是一个非对称势能面，σ 表示噪声量，$dW_t^{(1)}$ 是标准布朗运动。V 可以随时间变化，从而让我们能够模拟复杂活动，例如通过移动 V 的最低点来模拟间歇训练；而 σ 则保持恒定。之所以将 V 设为非对称，是因为运动员在运动期间的心率往往接近自然上限心率，但不会经常超过它。

然而，在我们的示例数据中，心率降至较低水平（100–160 次/分钟）并不少见，因此非对称势能面对这种情况的惩罚较轻。我们根据现有数据目测拟合 V 的具体形状以及 σ 的取值；它们的作用是提供合理的背景过程，以便我们尝试检测尖峰。图 S2.1 中展示了 V。

额外的噪声过程使我们能够纳入异方差性，因为在活动过程中有时会观察到噪声水平发生阶跃变化。图中展示了一个真实数据示例。3。在某些情况下，活动开始后的前十分钟左右比其余时间噪声更大。对此现象的一种可能解释（9）是，活动开始时汗液尚未积聚，导致心率监测器发生接触故障，进而产生噪声数据。由于这种噪声被认为是监测器造成的，而非运动员心率波动性增加所致，因此我们将其纳入一个独立的过程。

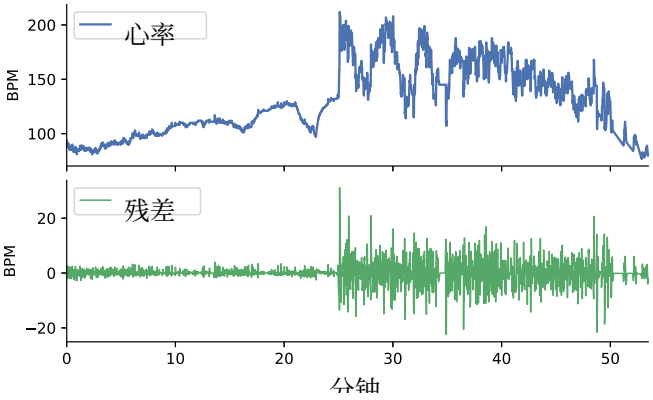


图 3. 显示真实心率样本中异方差性的时间序列。图中展示了记录的心率（上）以及去除十秒窗口移动平均后的残差（下）。可以清楚地看到，大约在 25 分钟后，点方差显著增加。

额外的噪声过程是一个简单的均值回复型奥恩斯坦–乌伦贝克过程（16），使用以下随机微分方程建模

$$dY_t = -\alpha(t)\,Y_t + \beta(t)\,dW_t^{(2)} \qquad [4]$$

其中，$\alpha > 0$ 描述过程以何种速率回复至其均值 0，β 是扩散系数，$dW_t^{(2)}$ 是标准布朗运动，且与 $dW_t^{(1)}$ 相互独立。α 和 β随时间变化，但不取决于过程的状态；其取值旨在为活动开始时增加额外噪声。

最后，尖峰过程 Z_t在最后进行模拟。由于尖峰期间观察到的高心率并不被视为个体心率的真实反映（9），因此将其视为一个独立过程是合理的。考虑到心律失常相关的尖峰也被认为更可能出现在心率较高、心脏负荷较重时（9），Z_t 将取决于 X_t。

尖峰过程由非齐次跳跃过程 J_t 的增量给出。设 $r : \mathrm{R} \to \mathrm{R}$ 为一个函数，用于描述尖峰发生速率随心率变化的关系。

随后使用该函数计算泊松强度

$$\lambda(t) = r(X_t) \quad [5]$$

它用于模拟非齐次泊松过程 P_t，该过程决定尖峰出现的时间。随后从对数正态分布中抽取尖峰的大小，得到跳跃过程 J_t。该跳跃过程的增量即为尖峰 Y_t，因此 Y_t 除在尖峰出现时取正值外，其余时候均为零。

模拟的最终输出是合成心率（H_t），其计算方法是先对 $X_t + Y_t + Z_t$ 进行移动平均平滑，再将结果四舍五入到最接近的整数。Crickles 对该模拟方法进行了检验，发现它能够生成逼真的心率时间序列。

我们使用具有多种特征的模拟数据来检验所述尖峰检测方法，包括间歇训练、活动水平的阶跃变化和异方差性，以及活动水平和噪声保持恒定的简单情形。模拟时间序列示例见 S2。对于每个时间序列，模拟算法还会输出一份尖峰列表，其中包含尖峰出现的时间和高度，以及尖峰出现时的基础心率。我们将这些数据与各方法检测到的尖峰进行比较。

C. 比较指标。为了比较上述各种尖峰检测方法在模拟数据上的性能，我们采用两种误差指标。两者都使用一组取自模拟数据的真实尖峰，以及一组由所选尖峰检测方法输出的检测到的尖峰。

对于第一种误差指标，我们设定严格的 5 秒时间窗口，并逐一扫描检测到的尖峰。对于每个检测到的尖峰，如果能在该时间窗口内与一个真实尖峰配对，就将其计为“正确尖峰”。若成功配对，则将这两个尖峰分别从各自的列表中移除。扫描完检测到的尖峰后，我们计算精确率和召回率

$$\text{precision} = \frac{\#\text{ correct spikes}}{\#\text{ detected spikes}}$$
$$\text{recall} = \frac{\#\text{ correct spikes}}{\#\text{ true spikes}}$$

并据此计算每种活动的 F_1 分数

$$F_1\text{ score} = \frac{2}{\text{precision}^{-1} + \text{recall}^{-1}}$$

，然后对所有模拟活动取平均值。F_1分数是召回率与精确率的调和平均数，是一种公认的性能指标（17–19）。该分数的取值范围为 0（最差）至 1（最好）。

第二个误差指标称为“尖峰密度误差”，它将尖峰的位置从单个点转换为一组可能点的分布。这样做的原因是，尽管尖峰在某一点达到峰值，但它们确实有一定宽度，部分是由于对真实和模拟心率时间序列进行的平滑处理。因此，尖峰事件并非发生在某个唯一时刻，所以将真实和检测到的尖峰时刻表示为分布而非点是合理的。该指标还允许我们纳入尖峰的高度。

我们定义一个核密度 $f : \mathrm{R} \to \mathrm{R}$，此处选择以 0 为中心、标准差为 5 的高斯分布（与前文严格阈值相对应）。该核密度将经过平移和缩放，以替代每个尖峰对应的单点值。也可以采用任何其他核，只要满足以下条件:

1. f 为正且对称。.
2. f 在 0 处达到全局最大值
3. f 可积，且积分为 1。

将平移并重新归一化后的函数记为 $f_s(t)$

$$f_s(t) = \frac{f(t-s)}{\int_0^T f(\tau - s)\, d\tau} \quad [6]$$

其中 T 为时间序列的长度。

给定 n 个发生在时刻 $0 \le s_1 < \cdots < s_n \le T$的尖峰，其对应高度为 $h_1, .., h_n$，定义尖峰密度ρ： $[0, T] \to \mathrm{R}^+$ 为

$$\rho(t) = \sum_{i=1}^{n} h_i\, f_{s_i}(t) \quad [7]$$

给定实数时刻和检测到的尖峰时刻列表，分别将对应的密度记为 ρ_r 和 ρ_d。尖峰密度误差定义为

$$\epsilon = \frac{1}{T}\int_0^T \left|\rho_r(t) - \rho_d(t)\right| dt \quad [8]$$

其误差单位为每小时 BPM。尖峰密度误差的特性将在S4中讨论。我们还将该误差与零尖峰检测模型的误差进行比较；在该模型中，始终不会检测到任何尖峰。零模型的误差就是真实尖峰密度的积分，因此我们定义误差影响

$$\epsilon^* = \frac{1}{T}\int_0^T \rho_r(t) - \left|\rho_r(t) - \rho_d(t)\right| dt \quad [9]$$

其单位同样为每小时 BPM。如果 ϵ^* 为负值，则意味着总体上检测到了错误的尖峰，因此不检测任何尖峰反而会更准确。如果 ϵ^* 为正值，则意味着尖峰检测方法通过正确识别尖峰降低了误差（即产生了正向影响），因此较大的 ϵ^* 表明尖峰检测效果良好。

D. 比较结果。为了比较第2.A小节中描述的尖峰检测方法，我们模拟了 45 组不同活动类型的心率时间序列训练集（活动类型列于第2.B小节最后一段）。这些训练数据用于调整尖峰检测方法中的所有相关参数，例如移动平均的宽度以及阈值算法中的常数乘数。我们对参数进行了网格搜索，使每种尖峰检测方法在所选参数下的 F1分数和误差影响达到最大。

在调整所有参数，使所有尖峰检测方法在训练数据集上均达到最佳分数后，随后将这些方法应用于测试数据集。该测试数据集包含900个模拟心率时间序列，且同样涵盖各种活动类型。根据尖峰检测的结果，

对所有模拟活动的平均F1分数和平均误差影响进行了评估，如图4所示。此外，还对每个模拟活动类型子集的平均F1分数和平均误差影响进行了评估，以开展更详细的分析。

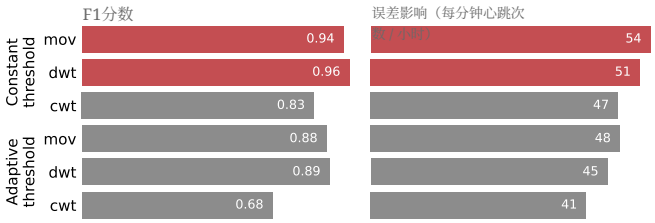


图4. 柱状图比较了对所有模拟活动取平均后的 F1分数（左）和误差影响（右）。两项指标均显示，基于移动平均（mov）和采用固定阈值的离散小波变换平滑的方法表现最佳（均以红色突出显示）。两者之中，考虑到方法简单，最终选择移动平均方法对实测数据进行尖峰检测。

从图4可以看出，采用固定阈值的移动平均和离散小波变换平滑方法都能准确检测出真实的模拟尖峰，不受活动类型影响，且两项性能指标的结果相近。由于移动平均方法能够更准确地检测尖峰高度，因此其误差影响更大。图5展示了在一个模拟心率时间序列示例中，该尖峰检测方法的具体表现。根据这些结果，并考虑到该方法相对简单，最终选择采用移动平均和固定阈值的方法作为后续使用的尖峰检测方法。

随后，将此尖峰检测方法应用于 Crickles 提供的 168 名运动员的实测数据，使用与比较时相同的一组参数，即移动平均平滑的宽度为 $w = 200$ 秒，阈值处理的常数乘数为 2.5。

E. 推断。将此方法应用于实测数据后，还需描述每位运动员的尖峰频率。我们重点探讨以下两个问题：

1. 尖峰总体频率与心律问题报告之间是否存在相关性？2. 尖峰频率与尖峰发生时的心率之间是否存在相关性？此外，对于有心律问题和没有心律问题的个体，这种关系是否相同？

为回答第一个问题，我们计算尖峰事件的发生率。我们将某个个体所有时间序列中的尖峰总数记为 N，时间序列的总时长记为 T。定义尖峰频率

$$\lambda = \frac{N}{T} \qquad [10]$$

如果尖峰事件的累计数量服从速率恒定的泊松过程，那么这实际上就是最大似然估计（MLE） λ （见S3）。我们为所有个体计算该值。与（9）一样，随后我们进行点二列相关检验和逻辑回归，以探究脉冲率与心律问题之间的关联。

为回答我们的第二个问题，即尖峰频率与心率之间的关系，我们考虑两种方法。首先，我们考虑尖峰频率的更详细模型。我们再次假设尖峰发生的时间由泊松过程决定，但现在允许该过程为随时间变化的速率 $\lambda(t)$的非齐次过程。令 $X(t)$表示移除尖峰后的平滑值心率时间序列，我们假设 $\lambda(t)$是 $X(t)$的分段常数函数。基于这两个模型，我们考虑以下假设：

- H_0：尖峰按照强度恒定的泊松过程发生。
- H_1：尖峰按照泊松过程发生，其强度是心率的分段常数函数。

我们采用似然比检验，判断更复杂的、依赖心率的模型对观测到的尖峰位置是否拟合得显著更好。该假设检验的完整细节见S3。我们对每位运动员以及所有运动员的总体数据分别进行该假设检验。如果零假设被拒绝，则表明心率与脉冲率之间存在需要进一步研究的关系。也可能观察到非均匀性与心律失常之间的关系。

我们还考虑过拟合速率为时间或心率的连续函数的非齐次泊松过程，但样本数据中观测到的尖峰数量太少，无法通过这种方法得出有意义的结果。

3. 结果与讨论

我们对 168 名运动员的数据进行了尖峰检测和推断。所有人都记录了至少 100 次时长不等的活动，其中 45 人在 Crickles 调查中报告有心律问题。我们从数据集中剔除了时长不足 5 分钟以及标记为“滞留”的活动，最终保留约 125,000 次活动。平均而言，每位运动员有 814 小时的活动数据。

A. 初步结果。图6 左侧显示的是按有无心律失常分组后，各运动员脉冲的泊松率箱线图。右侧则是相应的箱线图，展示被标记为不规则的激活所占百分比，称为不规则比例；这是文献（9）中最初用于衡量间断性的指标。

大多数运动员的尖峰率都很低，通常每小时少于 0.5 个尖峰。我们进行点二列相关检验，以评估泊松尖峰率与心律失常之间的相关性。我们还展示了对不规则比例进行同一检验的结果。

表 2. 泊松尖峰率和不规则比例与心脏健康调查的点二列相关性检验结果。

	相关性	p 值
泊松尖峰率	0.06	0.44
不规则比例	0.28	0.00027

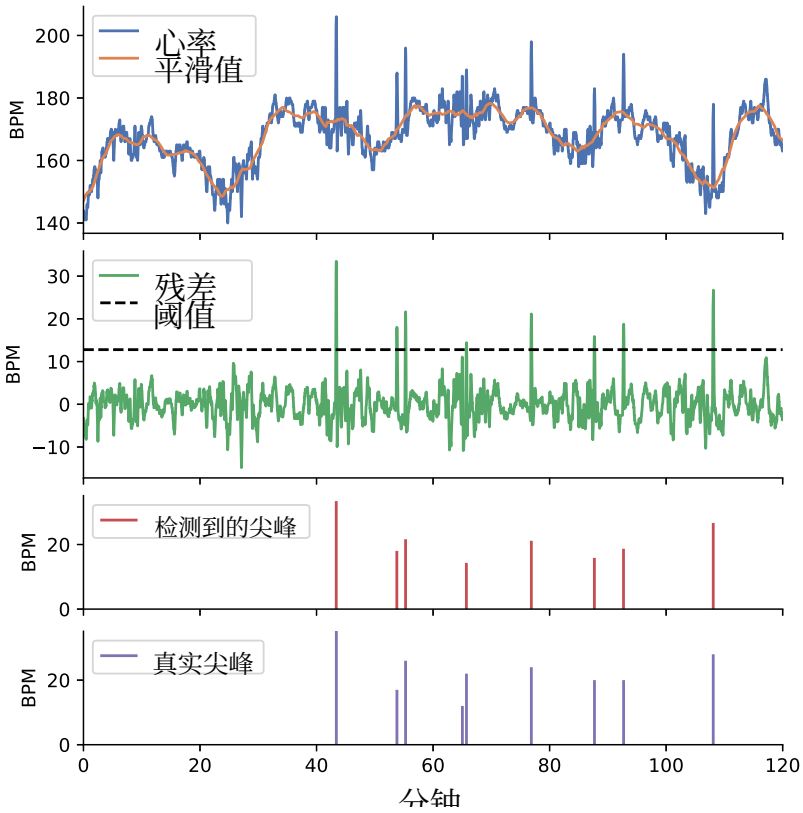


图 5. 使用移动平均法在模拟心率数据上进行尖峰检测的示例_来自图4。顶部图显示经移动平均（橙色）平滑处理的模拟时间序列。第二幅图显示由此得到的残差（绿色）以及评估得到的固定阈值（虚线）。第三幅和第四幅图分别显示检测到的尖峰和真实模拟尖峰，展示了该方法恢复真实尖峰的准确性。

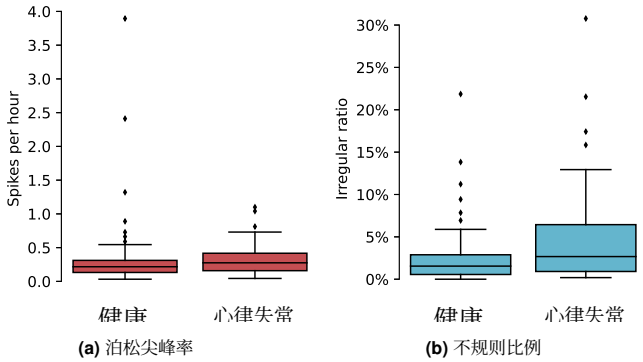


图6。这些箱线图展示了患有和未患有心律失常的运动员的泊松尖峰率（a）和不规则比例（b）的分布。可以清楚地观察到泊松尖峰率和不规则比例的一些异常值，并将在第3节B部分进行分析。不规则性比率与心律失常存在相关性，但泊松尖峰率则没有。

根据这些数值，我们无法得出泊松尖峰率与报告的心律问题之间存在统计显著相关性的结论。相反，不规则比例确实呈现显著相关性。接下来，我们将探讨造成这一差异的几种可能原因。

B. 异常值分析。在图6a 中，我们观察到少数没有心律失常的个体，其泊松尖峰率显著高于预期。为了考察它们对相关性的影响，可以将异常值

从数据集中剔除，即移除高于数据某一阈值百分位数的运动员。为确保比较的是同一组运动员的泊松尖峰率和不规则性比率，我们要求运动员在这两项指标中均不是异常值。剔除异常值对相关性和 p 值的影响如图所示。7。

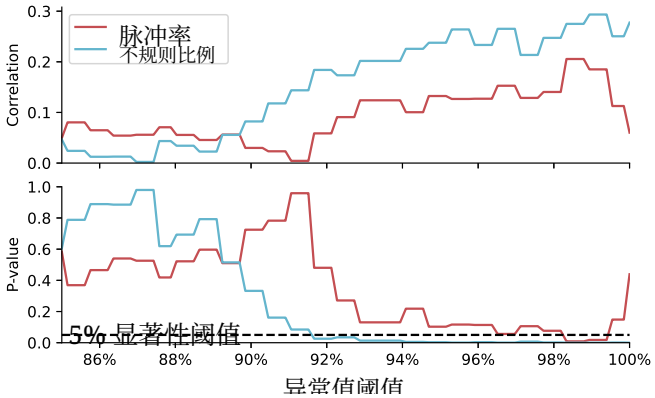


图7。图中显示了泊松尖峰率（红色）和不规则比例（蓝色）与心律失常之间的相关系数（上图）及相应的 p 值（下图），并以异常值阈值为横轴；高于数据中该百分位数阈值的泊松尖峰率或不规则比例的数据集将被剔除。剔除高于第98百分位数的数据后，泊松尖峰率显示出显著相关性，这表明它容易受到数据中某些异常值的影响。

随着从数据集中移除的运动员越来越多，不规则性比率与报告的心律问题之间的相关性逐渐减弱，

p 值则逐渐增大。在考察泊松尖峰率时，我们发现，移除少数离群运动员（即高于第 98 百分位数的运动员）会增强相关性并降低 p 值，表明结果具有显著性。然而，继续移除其他运动员会降低相关性和显著性。这表明，泊松尖峰率与报告的心律问题之间的相关性很容易受到这些异常值的影响。

遗憾的是，由于我们获取数据的权限有限，无法确定这些异常值产生的原因。这些运动员的活动时间序列中可能包含某些未出现在我们样本中的特征，导致我们的尖峰检测方法出现意外表现。泊松率是根据一名运动员所有活动中的尖峰计算得出的，因此，某项活动中产生的错误可能会影响整体比率。对单项活动的这种敏感性可能是这些异常结果的成因之一。相比之下，文献（9）中用于间隙度指标的“规则／不规则”活动布尔分类，限制了任何单项活动对运动员不规则比例的影响，从而提高了稳健性，但代价是信息量减少。

另一种可能是，尖峰检测方法的表现符合预期，但这些异常值由于其心率监测器故障或其他与心律失常无关的因素，尖峰率较高。不规则心律转化为有间隙时间序列的具体机制尚不明确，而且我们观察到的尖峰也可能由其他原因造成。因此，即使尖峰检测方法运行正常，推断出的尖峰率仍可能与心律失常缺乏明确的相关性。

最后，个别异常值可能对应于尚未确诊的心脏疾病，这意味着调查结果并不能准确反映实际的心律问题。此外，也可能有其他人表示自己曾有过心律问题，但目前正在接受治疗。由于此处提出的方法并非诊断工具，我们不会对所分析数据中的任何运动员的心脏健康状况妄加推测。

由于这些异常值的成因不明确，我们认为没有充分理由将其排除在分析之外，因此也承认，目前的尖峰检测和推断方法确实并未改进（9）中提出的原始间隙度指标。

C. 心率依赖性分析。如第 2.E 节所述，对于所有运动员的脉冲率进行非齐次性假设检验，结果显示，在 168 名运动员中，有 123 名存在足够证据拒绝零假设，这表明尖峰按照速率依赖于心率的泊松过程发生。图 S5.1 显示了尖峰发生时的心率与尖峰频率之间的直方图，其中可以看到，大量尖峰被检测到的范围为 $105-140$每分钟心跳次数。

根据 Crickles 的论文（9），研究还假设，只有运动员心率超过乳酸阈值心率（LTHR）的高心率区间内的心率数据特征，才能提供有关心律失常的信息，因为运动员在该心率区间承受着体力负荷。因此，为了检验我们的结果是否符合这一假设，我们将所有运动员低于某个心率阈值的尖峰 h_{thresh} 移除。随后重新计算每位运动员对应的泊松尖峰率，并同样评估新的泊松尖峰率与心律失常之间的点二列相关。图8 展示了相关系数和 p 值随心率阈值升高的变化，h_{thresh}。

在图中 $h_{\text{thresh}} \leq 75$ 范围的初始部分，相关系数和 p 值都相对稳定，保持在滤除任何尖峰之前的数值。这是因为在心率较低的范围内只有极少量尖峰，因此相关性检验产生的结果与小节 3.A 中的结果相近。

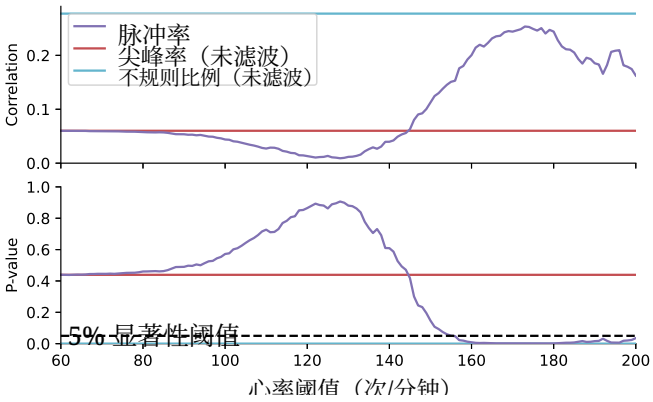


图 8. 心律失常与心率阈值的相关性，其中上图显示相关系数，下图显示相应的 p 值。蓝色和红色水平线分别表示 Crickles 的不规则性比率和未滤除任何尖峰时的泊松尖峰率结果。紫色曲线表示移除低于心率阈值的尖峰后的泊松尖峰率。

在心率阈值范围内 $75 \leq h_{\text{thresh}} \leq 130$，泊松尖峰率与心律失常之间的相关性有所下降，这是因为该范围处于一个模糊地带：对一些运动员而言，这一心率范围被视为偏高，而对另一些运动员而言则偏低。从图S5.1的直方图可以看出，出现尖峰时心率的众数也落在此范围内。因此，超过此范围后，大量尖峰会相对随机地从所有运动员的数据中移除，从而影响检测到的尖峰的显著性，并降低其与心律失常的相关性。

在范围 $130 \leq h_{\text{thresh}} \leq 175$内，然而，可以观察到相关系数稳步上升，p 值持续下降。脉冲率与心律失常的相关性在 $h_{\text{thresh}} = 175$ 处达到峰值。在此范围内，所有运动员的心率都可能被认为偏高。由于此时只剩下高心率下的尖峰，相关性上升说明，只有心率处于较高范围时出现的尖峰，才显著表明受试者报告存在心律问题。这不仅与前文所述的假设一致，也符合关于心律失常的医学认识：运动员在体力负荷下可能会导致心率出现不规则模式（20）。使用泊松尖峰率计算出的心律失常相关性，始终不及使用 Crickles 的不规则性比率所得的相关性高，因为此处使用的是全局心率尖峰阈值，而不是像 Crickles 的研究那样，为每位运动员单独设定乳酸阈值心率。因此，进一步检索以获取乳酸阈值心率数据，并将其应用于尖峰检测方法，可能会进一步提高泊松尖峰率与心律失常之间的相关性。

进一步检索以获取乳酸阈值心率数据，并将其应用于尖峰检测方法，可能会进一步提高泊松尖峰率与心律失常之间的相关性。

最后，当 $h_{\mathrm{thresh}} \geq 175$ 时，相关性和 p 值都会降低，因为此时大多数尖峰都已被移除。还需要更多在如此高心率下出现尖峰的心率数据，才能获得更好的相关性并得出有意义的结论。

4. 后续工作

该项目后续工作的第一步，如第C节倒数第二段所述，应将每位运动员的乳酸阈值心率纳入尖峰检测方法，作为下限，这样只会考虑心率范围达到每位运动员个人高心率标准的尖峰。这需要获取该运动员更多的历史活动数据，以准确估算其乳酸阈值心率；但由于数据隐私，这很可能会构成挑战。

采用当前的尖峰检测方法时还发现，活动数据中很大一部分尖峰都出现在最初的10 分钟内，如图S6.1中的直方图所示。这很可能主要是由心率监测器的缺陷导致的：运动员刚开始活动时，监测器难以获取准确读数。例如，在汗液形成之前，胸带容易与皮肤接触不良，导致心率读数不准确且含有噪声。还需要进一步分析如何处理这类来自心率监测器的噪声数据，以便尖峰检测方法能够正确区分由外部因素导致的噪声和由心律不规则导致的尖峰。处理开始阶段的噪声读数时，一种方法是在实施尖峰检测方法之前，舍弃活动最初的5 至10 分钟的读数。

除了明确检测心率数据中的尖峰外，还可以开展后续工作，刻画心率时间序列的其他特征及其与心律失常的相关性。实测数据表明，心率的波动性不仅在不同活动之间有所差异，在运动员单次活动所记录的心率数据内部也会发生变化。因此，分析和测量心率数据的波动性，也可能有助于发现心律问题。

除了心率测量值外，运动设备通常还会在活动期间记录其他数据，例如速度、距离、海拔、运动员的功率输出等。对这些额外数据进行分析，或许可以用来排除心率数据中的某些特征，例如波动性的变化或心率的剧烈趋势，因为这些可能是运动员锻炼时突然增加体力负荷所致。

总之，除了探索不同的数据驱动方法外，仍有大量后续工作可用于改进当前的尖峰检测方法，并对心率时间序列的其他特征进行建模，以增强这些指标与心律失常之间的相关性。

致谢。 作者谨向 Colm Connaughton 教授致以谢意，感谢他出色的指导；也感谢 Ian Green 提供的宝贵见解，并不辞辛劳地在 Crickles 数据上运行我们的实验。本研究由英国工程与物理科学研究委员会通过华威大学系统数学 II 博士培训中心资助（项目编号 EP/S022244/1）。

参考文献

1. 英国国家医疗服务体系（2021）心律失常（ https://www.nhs.uk/conditions/arrhythmia/）。访问日期：03/06/2021。2. 在欧洲心律协会（EHRA）的特别贡献下，D 等（2010）心房颤动管理指南：欧洲心脏病学会（ESC）心房颤动管理工作组。《欧洲心脏杂志》31(19):2369–2429。3. Camm AJ、Corbucci G、Padeletti L（2012）连续心电监测对心房颤动的用途。《美国心脏病学杂志》110(2):270–276。4. Aizer A 等（2009）剧烈运动与心房颤动风险的关系。《美国心脏病学杂志》103(11):1572–1577。5. Centurión OA 等（2019）心房颤动与耐力身体活动之间的关联：多少才算过量？《心房颤动杂志》12(3)。6. McManus DD 等（2013）在心房颤动患者中使用 iPhone 4s 检测不规则脉搏的一种新方法。《心律》10(3):315–319。7. Li KHC 等（2019）用于监测心率、心率变异性和心房颤动的手机应用程序现状：叙述性综述。《*JMIR*移动健康与泛在健康》7(2):e11606。8.（年份？） Crickles（https://crickles.casa/）。访问日期：03/06/2021。9. Green I、Dayer M（2021）从运动数据中检测心律失常。未发表。10. Mills TC（2003）《经济时间序列中的趋势与周期建模》中的趋势与周期建模“经典”技术。（Palgrave计量经济学丛书。帕尔格雷夫·麦克米伦，伦敦）。11. Graps A（1995）小波简介。《*IEEE*计算科学与工程》2(2)。12. Mallat S（2009）《信号处理小波之旅》。（学术出版社），第三版。13. Van Brakel JPG（2014）使用 z分数的稳健峰值检测算法（https: // st ackoverflow.com/questions/22583391/peak-signal-detection-in-realtime-timeseries-data/ 22640362#22640362）。访问日期：26/05/2021。14. Mayer K、Schmid T、Weber F（2015）电力现货价格建模：结合均值回归、尖峰和随机波动率。《欧洲金融学杂志》21(4):292–315。15. De Jong C（2006）电力价格尖峰的特性：一种机制转换方法。《非线性动力学与计量经济学研究》10(3)。16. Gardiner CW 等（1985）随机方法手册。（施普林格，柏林）第3卷。17. Dice LR（1945）物种间生态关联程度的测量方法。《生态学》26(3):297–302。18. Sørensen TJ（1948）一种基于物种组成相似性确立植物群落学中等振幅群组的方法，以及其在丹麦公地植被分析中的应用。（由 E. Munksgaard 受托出版）。19. Chinchor N、Sundheim BM（1993）第五届信息理解会议（*MUC-5*）：*1993*年*8*月*25*日至*27*日在马里兰州巴尔的摩举行的会议论文集中的 MUC-5 评估指标。20. Guasch E、Mont L（2017）运动诱发性心律失常的诊断、病理生理学与管理。《自然综述：心脏病学》14:88–101。21. Daubechies I（1992）小波十讲。（工业与应用数学学会）。22. Kaiser G（2011）小波入门指南。（Birkhäuser Boston）。23. Lee GR、Gommers R、Waselewski F、Wohlfahrt K、O'Leary A（2019）PyWavelets：用于小波分析的 Python 软件包。《开放源代码软件杂志》4(36):1237。24. Gao HY（1998）使用非负割草机进行小波收缩去噪。《计算与图形统计学杂志》7(4):469–488。

补充信息

S1. 尖峰检测方法

A. 小波简介。以下将以极快的节奏概述本项目所涉及的小波相关理论。若想阅读更易理解的入门介绍，读者可参阅 Daubechies 的优秀著作（21）和 Kaiser 的著作（22）。尤其是，Daubechies 的著作为本节奠定了基础。

傅里叶变换将信号分解为频谱。低频分量描述信号中变化较慢的方面，例如总体趋势；而尖峰等特征则需要更高频率的分量。然而，在傅里叶空间中表示一个尖峰需要所有频率的贡献。事实上，尖峰最极端的情况是以点 $c \in \mathbb{R}$, $\delta_c(t)$ 为中心的狄拉克 δ 函数，其傅里叶变换为 $\hat{\delta}_c(\omega) = e^{-i\omega c}$。这使得难以利用傅里叶变换从信号中分离出尖峰，因此，理想的做法是寻找一种能够在频率和时间两个维度上对“内容”进行局部化的信号分解。

小波理论（12、21）涵盖了许多用于执行这种时频分解的相关变换，其中最重要的是连续小波变换和离散小波变换。它们都从一个母小波开始，$\psi \in L^2(\mathbb{R})$，例如图S1.1中的小波，它定义了一种短暂的振荡。通过平移和缩放 ψ，我们可以生成一系列（子）小波

$$\psi^{a,b}(t) = |a|^{-1/2}\psi\left(\frac{t-b}{a}\right). \qquad [11]$$

这里 $a \in \mathbb{R} \setminus \{0\}$ 是尺度参数，而 $b \in \mathbb{R}$ 是平移参数。通过按 $|a|^{-1/2}$ 的倍数进行重新缩放，我们保留了小波的 L^2 范数，$\|\psi^{a,b}\| = \|\psi\|$。

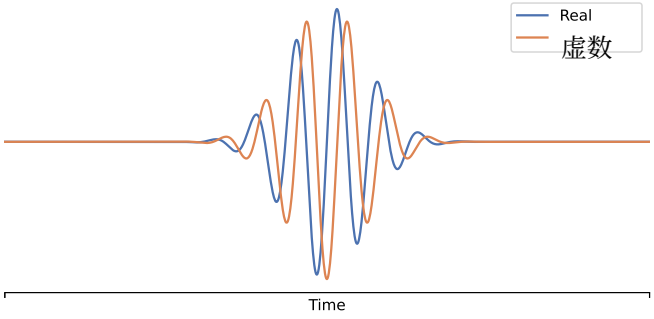


图 S1.1。母小波示例：带宽为2s^2、中心频率为 1 Hz的复 Morlet 小波，$\psi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} e^{2\pi i t}$。

我们必须对可接受的母小波施加一些限制。只需要求 $\psi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$满足

$$\int_{\mathbb{R}} \psi(t)\,dt = 0, \qquad \|\psi\| = \int_{\mathbb{R}} |\psi(t)|^2\,dt = 1. \qquad [12]$$

要求 ψ 的积分为零，这样我们就能对小波变换进行反变换；满足这一条件的母小波称为可容许。关于 L^2-范数的条件只是一种归一化约定。可以放宽这些限制，使 ψ 不必属于 $L^1(\mathbb{R})$，积分也不必为 0，但仍然可容许，不过这在实际应用中意义不大。

信号 $f(t)$的连续小波变换（连续小波变换）为

$$(\mathcal{W}f)(a,b) = \langle f, \psi^{a,b}\rangle = |a|^{-1/2}\int_{\mathbb{R}} f(t)\overline{\psi}\left(\frac{t-b}{a}\right)dt. \qquad [13]$$

与傅里叶变换一样，它使用内积来定义。
然而，与傅里叶变换不同，CWT 包含大量冗余。这意味着不存在用于反演 CWT 的唯一公式。不过，确实存在一种称为恒等分解 的规范公式

$$f(t) = C_\psi^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\mathcal{W}f)(a,b)\,\psi^{a,b}(t)\,\frac{1}{a^2}\,da\,db \qquad [14]$$

其中

$$C_\psi = 2\pi\int_{-\infty}^{\infty}|\hat{\psi}(\xi)|^2\frac{1}{|\xi|}d\xi.$$

此处$\hat{}$ ψ 表示母小波的傅里叶变换。可以证明，对于可容许小波（公式（12）），定义常数 C_ψ 的积分存在且为有限值。

在离散小波变换（*DWT*）中，尺度参数和平移参数仅取离散值

$$a_m = 2^m, \quad b_n = n2^m \qquad [15]$$

从而得到子小波

$$\psi_{m,n}(t) = \frac{1}{\sqrt{2^m}}\psi\left(\frac{t}{2^m} - n\right). \qquad [16]$$

因此，DWT 可以视为 CWT 的离散采样，如图S1.2所示。在较高频率处，DWT 具有更高的时间分辨率，可以捕捉更快速的变化；但其频率分辨率较低，因为对于较低频率分量而言，频率分辨率更为重要。请注意，此时我们仍处于原始信号被连续采样的状态。

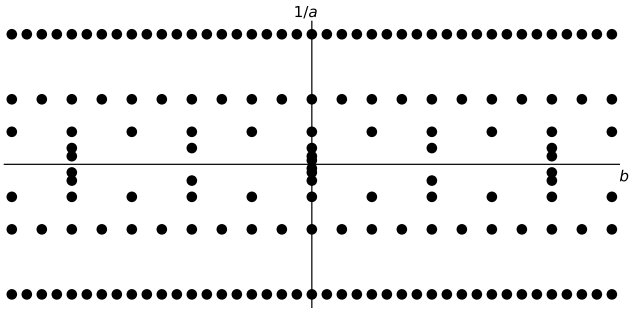


图 S1.2。离散小波变换可以视为连续小波变换的离散采样。图中的点表示这些离散位置。请注意，y轴显示的是 $1/a$，其与小波的中心频率成正比。x轴绘制的是平移参数 b。

对于母小波的某些特定选择，ψ，小波 $\psi_{m,n}$ 构成 $L^2(\mathbb{R})$ 的一组基。对于更具体的选择，这组基是正交归一的，并且只需与相应的小波取内积，就能轻松计算出分解中的系数

$$f(t) = \sum_{m,n}\langle f, \psi_{m,n}\rangle\psi_{m,n}. \qquad [17]$$

如果这组基不是正交归一的，我们仍然可以利用对偶基的Riesz表示元 $\tilde{\psi}_{m,\,n}$来计算系数。~这些函数满足 $\langle \psi_{m,\,n},\ \psi_{m',\,n'} \rangle = \delta_{m,\,m'}\delta_{n,\,n'}$，其中 $\delta_{j,\,k}$ 是 Kronecker δ。 f 在小波基中的系数可通过与这组对偶基取内积得到，

$$f(t) = \sum_{m,n} \langle f, \tilde{\psi}_{m,n} \rangle \psi_{m,n}. \qquad [18]$$

式（15）中所采用的离散化方案，源于即将介绍的多分辨率分析；不过，也可以采用其他离散化方案。母小波与离散化方案相结合后若不能构成基，所得到的便是冗余框架，这超出了本引言的讨论范围。

为了深入理解小波，有必要介绍多分辨率分析。这是离散小波变换最常见的实现方式，例如，本项目所使用的库 `pywavelets` (23)就采用了这种方式。除了计算效率之外，其主要原因之一是它提供了一种构造小波的方法，使其构成正交归一基。

多分辨率分析并不显式定义母小波 ψ，而是从一个（闭）逼近空间的嵌套序列开始

$$\cdots \subset V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \subset \cdots$$

满足

$$\overline{\bigcup_{m\in\mathbb{Z}} V_m} = L^2(\mathbb{R}), \qquad \bigcap_{m\in\mathbb{Z}} V_m = \{0\}.$$

我们还要求

$$\forall n \in \mathbb{Z}, \quad f \in V_0 \ \Leftrightarrow\ t \mapsto f(t-n) \in V_0,$$
$$\forall m \in \mathbb{Z}, \quad f \in V_m \ \Leftrightarrow\ t \mapsto f(2^m t) \in V_0.$$

后一条件正是多分辨率分析这一名称的由来。最后，我们要求存在一个函数 $\phi \in V_0$，使得 $\{\phi_{0,n}\colon\ n \in \mathbb{Z}\}$ 构成 V_0 的正交归一基，其中对所有 $m, n \in \mathbb{Z}$, $\phi_{m,n} = 2^{-m/2}\phi(2^{-m}t - n)$ 均成立。这个 ϕ 称为该多分辨率分析的尺度函数。

通过这一构造，对于 $f \in L^2(\mathbb{R})$，我们可以通过正交投影到 V_m 上来构造连续的近似 $f_m \in V_m$。根据构造，当 $m \to -\infty$ 时，我们逐点有 $f_m \to f$。此外，很明显，尺度函数 ϕ 足以重构完整的多分辨率分析框架。

现在我们可以为 $L^2(\mathbb{R})$ 构造一个正交归一小波基。对于每个 $m \in \mathbb{Z}$，令 W_m 为 V_{m-1} 中 V_m 的正交补，从而有

$$V_{m-1} = V_m \oplus W_m, \quad \text{and} \quad W_m \perp W_{m'} \text{ whenever } m \neq m'.$$

Define

$$\psi = \sum_{n\in\mathbb{Z}} g_n \phi_{-1,n} \ \in W_0 \subset V_{-1} \qquad [19]$$

其中，对于每个 $n \in \mathbb{Z}$, $g_n = (-1)^n \overline{\langle \phi, \phi_{-1,-n+1} \rangle}$。已在（21，定理 5.1.1）中证明，$\psi$ 是生成小波正交归一基的母小波 $\{\psi_{m,n}\colon\ m, n \in \mathbb{Z}\}$，其对应于 $L^2(\mathbb{R})$，其中每个 $\psi_{m,n}(t) = 2^{-m/2}\psi(2^{-m}t - n)$。此外，对于每个 $m \in \mathbb{Z}$，$\{\psi_{m,n}\colon\ n \in \mathbb{Z}\}$ 是子空间 W_m的正交归一基。但需要注意的是，ψ 的这种选择并不唯一。

现在我们描述一种用于计算这些系数的高效算法，这是所谓子带滤波方案的一个示例。假设我们通过正交投影得到了某个函数 f_* 的细尺度近似 $f \in V_m$，并且已经计算出由尺度函数生成的基中的系数（$\langle f, \phi_{m,n} \rangle\colon\ n \in \mathbb{Z}$）。通过重新调整尺度函数，我们不失一般性地假设 $m = 0$。因此，我们有 $f = \sum_{n\in\mathbb{Z}} \langle f, \phi_{0,n} \rangle \phi_{0,n} \in V_0$。我们的目标是将 f 的这一表示替换为使用 $V_0 = V_M \oplus W_M \oplus \cdots \oplus W_1$ 的自然基的表示，其中给定整数 $M \geq 1$ 称为分解的 层级，

$$f = \sum_{n\in\mathbb{Z}} a_{M,n}\phi_{M,n} + \sum_{m=1}^{M} \sum_{n\in\mathbb{Z}} d_{m,n}\psi_{m,n}. \qquad [20]$$

基中尺度函数部分的系数 $a_{M,n}$ 称为近似系数，而基中小波部分的系数 $d_{m,n}$ 称为细节系数。

我们已经定义了系数 g_n，并且可以定义系数 h_n，使得

$$g_n = \langle \psi, \phi_{-1,n} \rangle, \quad h_n = \langle \phi, \phi_{-1,n} \rangle. \qquad [21]$$

在（21，第 5.6 节）中证明，对于 $f \in L^2(\mathbb{R})$ 以及所有$m, n \in \mathbb{Z}$,

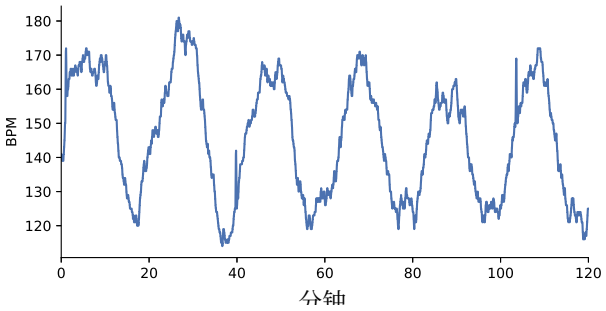
$$d_{m,n} = \langle f, \psi_{m,n} \rangle = \sum_{k\in\mathbb{Z}} \overline{g_{k-2n}} \langle f, \phi_{m-1,k} \rangle, \qquad [22a]$$

$$a_{m,n} = \langle f, \phi_{m,n} \rangle = \sum_{k\in\mathbb{Z}} \overline{h_{k-2n}} \langle f, \phi_{m-1,k} \rangle. \qquad [22b]$$

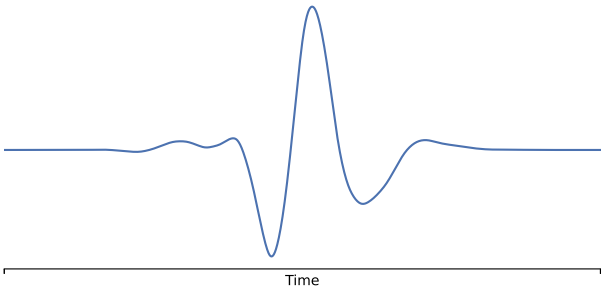
这为我们提供了一种递归算法，用于计算 f 的任意层级 $M \geq 0$ 的近似系数和细节系数，起始于我们已经计算出的内积 $\{f, \phi_{0,n}\colon\ n \in \mathbb{Z}\}$。事实上，利用公式（22），我们可以将 m 层的近似系数分别与滤波器（g_{-n}）和（h_{-n}）进行卷积，并舍弃结果中的奇数项，从而生成 $m+1$ 层的系数。这种对系数进行的子采样称为抽取。在信号处理文献中，滤波器（g_{-n}）和（h_{-n}）这一对称为滤波器组。

当然，在实际应用中，我们并没有连续采样信号，而是假设信号以规则间隔进行采样。这实际上对我们有利。首先，我们可以使用快速傅里叶变换，高效地完成所有卷积。其次，在多分辨率分析开始时，我们不需要进一步近似 f。相反，我们可以通过卷积精确计算 $k \in \mathbb{Z}$ 的第一组近似系数 $a_{0,\,k} = \langle f,\ \phi_{0,\,k} \rangle$。在这种情况下，无论采用哪一级 M 分解，都可以精确恢复 f 的原始采样。不过，时间序列的长度限制了 M所能达到的最大值。实际上，每一步都会将系数抽取，数量减半，因此这一过程不可能无限持续下去。

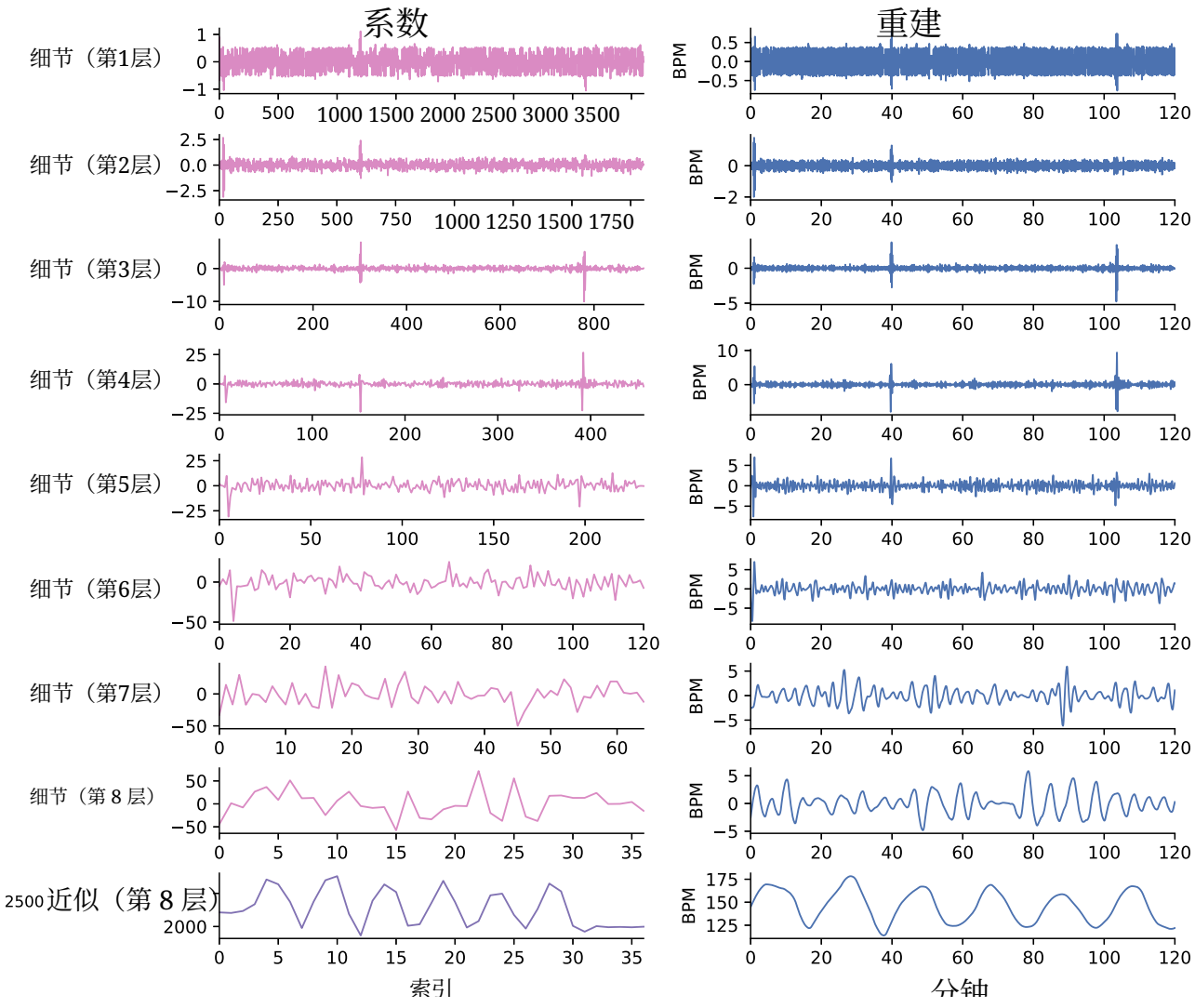
更一般地，当我们有离散采样信号时，必须谨慎，不要轻信与之对应的小波系数



(a) 原始时间序列



(b)五阶Symlet小波



(c) DWT 分解

图 S1.3. 使用 DWT 进行多分辨率分解的示例。图（a）显示原始的模拟时间序列；图（b）显示由多分辨率分析隐式定义的五阶对称小波；图（c）显示前八个分解层级的细节系数和近似系数。为减少系数中的端点效应，分解前对信号两端进行了对称延拓。图（c）中，系数（左侧）在每个层级进行抽取，因此每个后续层级的系数数量都是前一层的一半。随着分解层级增加，其幅度也会增大，因为每个层级的小波基都按 $\sqrt{2}$ 的比例缩小。重构结果（右侧）显示了各层级对原始时间序列的贡献，将这些结果相加即可恢复原始时间序列。

高于奈奎斯特频率的频率。对于连续小波变换而言，这一点尤为重要，因为该算法不会自然地阻止我们超出这一范围。不过，与傅里叶分析不同，这只是一条经验法则，因为对于某些母小波族而言，单一频率这一概念比对于其他母小波族更为明确。

在实际应用中，通常还会以某种方式延拓信号，以避免端点效应。默认情况下，我们可以简单地将采样区间之外的信号解释为零。不过，通过对信号进行对称延拓，我们可以减小各层系数中的端点效应。

图 S1.3 展示了使用一种称为 5 阶“对称小波”的正交小波对模拟心率进行的分解。

最后，我们注意到，可以放宽对基 $\{\phi_{-1,n}:\ n\in\mathbb{Z}\}$ 为正交归一的条件（12，第 7.4 节）。如果它们仅构成 V_0 的里斯基基，那么我们为 $L^2(\mathbb{R})$ 得到的小波基就不是正交归一的，但与之前一样，可以使用对偶基计算该基下的系数。这类小波基通常称为双正交小波基。相应多分辨率分析的滤波器组包含四个滤波器：一对用于分析（分解），另一对用于合成（重建）。

B. 使用连续小波变换进行尖峰检测。
我们可以使用连续小波变换对心率时间序列进行尖峰检测。如图 S1.4 所示，尖峰表现为延伸至较小尺度（较高频率）的脊线。选择适当的尺度后，我们可以得到一个仅关于平移参数的函数，并对其应用阈值，以识别尖峰。与对心率时间序列的残差设置阈值的方法一样，这里也可以使用固定阈值或自适应阈值。

选择小波对于使用 CWT 进行尖峰检测非常重要。实值小波会使每个尖峰呈现振荡模式，从而难以确定尖峰的数量及其位置。因此，在本报告中，我们使用复 Morlet（或 Gabor）小波族中的一种复小波。其形式为

$$\psi(t)=\frac{1}{\sqrt{b\pi}}e^{-t^2/b}e^{2\pi ict} \qquad [23]$$

其中，带宽 b 和中心频率 c 为参数。示例如图 S1.1 所示。在本项目中，我们使用了 $b=1\mathrm{s}^2$ 和 $c=1\,\mathrm{Hz}$。使用复小波的优点在于，小波系数并不是在零点两侧振荡，而是在复数

平面。因此，对系数取绝对值后，每个尖峰都会形成一个平滑的单峰。

对连续小波变换系数进行阈值处理后，会得到若干个包含尖峰的连续区域。我们取每个区域的中点，将其指定为一个尖峰。

C. 使用离散小波变换进行尖峰检测。我们在正文中介绍了主要的尖峰检测方法，其步骤为

1. 平滑原始信号，2. 从原始信号中减去平滑后的信号，

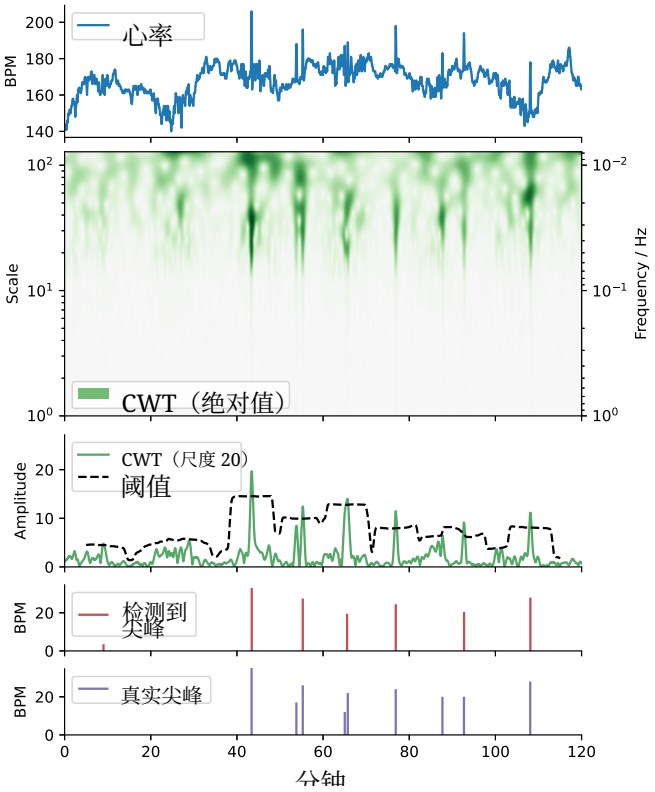


图 S1.4。 使用连续小波变换进行尖峰检测的示例。首先，对上方图中的模拟心率进行连续小波变换，所用小波为带宽$2\mathrm{s}^2$、中心频率为1 Hz 的复 Morlet 小波（见图 S1.1）。第二幅图显示其绝对值。随后，在尺度为20s（0.05 Hz）处对连续小波变换结果采用自适应阈值处理，结果见第三幅图。底部两幅图显示，在此示例中检测到了大多数模拟尖峰，但漏掉了三个。此外，在时间序列开头还检测到一个额外尖峰，不过其高度被赋予了一个很小的值，因此不会对误差影响造成显著影响。

3. 使用阈值（固定或自适应）识别残差中的尖峰。

离散小波变换为这一过程提供了另一种平滑方法。

在信号去噪（12，第 11 章）中，我们尝试为信号寻找一种稀疏表示。正如我们已经看到的，使用多分辨率分析可以利用一组近似系数和细节系数获得一种表示。如图S1.3c所示，并非所有这些细节系数都具有重要意义，其中许多可以合理地视为噪声。通过将较小的系数收缩为零，我们可以获得原始时间序列的稀疏近似。

获得稀疏表示时，有几种可选的阈值处理方式。最常用的是硬阈值、 $T_\lambda(x)$和软阈值、 $S_\lambda(x)$，因为它们最适用于统计风险分析。不过，本项目采用了非负伽罗特(24)、 $G_\lambda(x)$，它兼具两者的优点。这三种方法的定义如下：

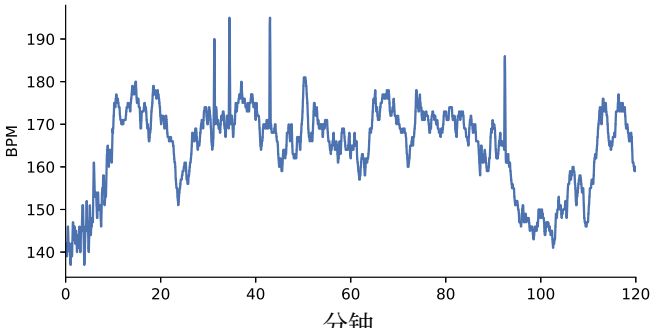
$$T_\lambda(x)=\begin{cases}x & \text{if } |x|\geq\lambda,\\ 0 & \text{if } |x|<\lambda;\end{cases} \qquad [24a]$$

$$S_\lambda(x) = \begin{cases} x - \lambda & \text{if } x \geq \lambda, \\ x + \lambda & \text{if } x \leq -\lambda, \\ 0 & \text{if } -\lambda < x < \lambda; \end{cases} \quad [24b]$$

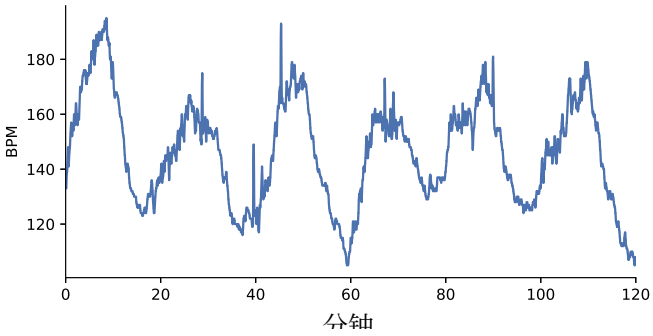
$$G_\lambda(x) = \begin{cases} x - \lambda^2/x & \text{if } |x| \geq \lambda, \\ 0 & \text{if } |x| < \lambda. \end{cases} \quad [24c]$$

D. 缺失数据处理。在将尖峰检测方法应用于 Crickles 提供的大量心率数据时，我们遇到的一个问题是，某些记录的时间序列中存在缺失数据片段，在单次记录的活动中，这些片段的时长从几分钟到几小时不等，导致结果不切实际，例如检测到数量极不合理的尖峰。

为处理这些问题，我们通过在心率最后一次记录点和下一次记录点之间进行线性插值，在缺失数据区域填入虚拟心率读数，从而使时间序列中现在每秒都有心率读数。此操作在使用移动平均平滑心率时间序列之前完成。对插值后的时间序列进行平滑后，我们移除了为插值而补入心率读数的区域，使最终的平滑版本与原始心率数据对齐。随后，我们计算残差，并根据前文所述，在残差上设定阈值以检测尖峰。我们采用这种方法，而非简单地从数据集中移除空白区域，因为它确保了被缺失区域隔开的已记录心率片段彼此独立，不会被视为同一段连续活动。



(a) 随机活动模式



(b) 间歇训练

图 S2.2. 具有不频繁尖峰的模拟时间序列示例。(a) 中使用恒定势能面，并在前 10 分钟内加入额外噪声。(b) 中使用势能面的最小值生成间歇训练模式。从前 10 分钟开始，每隔一个 10 分钟时段都会加入额外噪声。

S2. 示例模拟时间序列

此处展示用于模拟方程中基础心率的势能面。3。接下来我们给出两个模拟时间序列的示例。

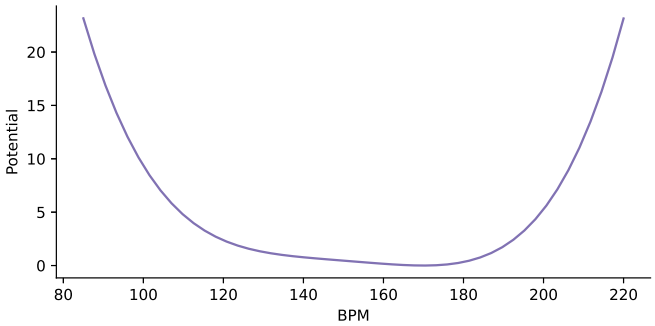


图 S2.1. 心率函数形式下固定势能面的示例。势能面的不对称性确保心率不会频繁超过自然上限，同时允许心率降至更低水平。

S3. 泊松推断

这里介绍用于考察以下假设的似然比检验：

- H_0：尖峰按照强度恒定的泊松过程发生。
- H_1 尖峰按照强度为心率分段常数函数的泊松过程发生。

我们首先计算 H_0 下的最大似然估计。设某个个体所有时间序列中的尖峰总数为 N，时间序列的总长度为 T。在 H_0 下，对数似然函数为

$$\ell_0(\lambda_0) = N \log(\lambda_0) - \lambda_0 T \quad [25]$$

其最大值在以下情况下取得

$$\hat{\lambda_0} = \frac{N}{T} \quad [26]$$

当然，前提是 $T > 0$。

现在我们考虑备择假设下的模型。用 $X(t)$ 表示移除了尖峰的平滑心率时间序列。将可能的心率范围划分为 k 个区间 $[X_0, X_1)$ $[X_1, X_2) \ldots, [X_{k-1}, X_k]$，其中 $X_0 < X_1, \ldots, X_k$。定义 N_i 为在 $X_t \in [X_{i-1}, X_i)$ 时，整个时间序列中检测到的尖峰事件数量。此外，定义 T_i 为 $X_t \in [X_{i-1}, X_i)$ 所持续的时间。

$$N_i = \sum_{j=1}^{N} \mathbb{1}\{X_t \in [X_{i-1}, X_i)\} \quad [27]$$

$$T_i = \int_0^T \mathbb{1}\{X_t \in [X_{i-1}, X_i)\}\,\mathrm{dt} \quad [28]$$

对数似然函数为

$$\ell(\lambda_1, \ldots, \lambda_k) = \sum_{i=1}^{k} N_i \log(T_i) - T_i\lambda_i \quad [29]$$

通过分别最大化求和中的各项即可使其最大化，由此得到最大似然估计

$$(\hat{\lambda_1}, \ldots, \hat{\lambda_k}) = \left(\frac{N_1}{T_1}, \ldots, \frac{N_k}{T_k}\right) \quad [30]$$

由于齐次泊松过程是这一更复杂模型的一个受限情形，因此更复杂的模型总能对数据提供“更好”的拟合。我们希望通过判断这一更复杂的模型是否能显著改善拟合效果，检验泊松过程的强度与心率之间是否存在显著关系。备注。理论上，可以在生理上不现实、永远无法达到的心率范围内纳入任意数量的区间。为避免不必要地增加 k，并避免在定义最大似然估计时遇到问题，我们假设所有 $T_i > 0$，并舍弃或合并不满足此条件的区间。

检验统计量为

$$\begin{aligned} D &= 2\left[\ell(\hat{\lambda}) - \ell_0(\hat{\lambda_0})\right] \\ &= 2\left[\sum_{i=1}^{k}\left(N_i \log\left(\frac{N_i}{T_i}\right) - 1\right) - N\left(\log\left(\frac{N}{T}\right) - 1\right)\right] \end{aligned}$$

将其与自由度为 $k-1$ 的卡方分布进行比较，以检验显著性。

S4. 尖峰密度误差

尖峰密度误差的一些基本性质：

- 由于它是若干非负函数之和的积分，因此始终非负。
- 由于尖峰的大小或数量没有理论上的限制，因此误差没有上界。
- 漏检或误检相对较大的尖峰时，误差会增大。

下面给出一个命题，说明准确度与误差之间的关系符合预期。

命题1. 考虑这样一种情况：只有一个真实尖峰和一个检测到的尖峰，二者大小相同（为简单起见，假设大小为 1），且检测到的尖峰与真实尖峰相距 h。令 s 为真实尖峰的时间，$s+h$ 为检测到的尖峰的时间。假设密度核f 的支撑有限，并且尖峰的位置距离时间序列的开头或结尾足够远，使得 f_s 和 f_{s+h} 的支撑完全位于 $[0, T]$内。在这种情况下，尖峰密度误差会随着 h减小而单调下降，并且由下式给出

$$\epsilon(h) = \frac{2}{T}\int_{-\frac{h}{2}}^{\frac{h}{2}} f(t)\,dt = \frac{4}{T}\int_0^{\frac{h}{2}} f(t)\,dt \quad [31]$$

证明。假设 $h > 0$。回顾平移和缩放的密度核的定义（公式6）

$$f_s(t) = \frac{f(t-s)}{\int_0^T f(\tau - s)\,d\tau}$$

利用关于尖峰和密度核的假设，可将其化简为

$$f_s(t) = f(t-s)$$

接着定义

$$\begin{aligned} g(t) &= f_s(t) - f_{s+h}(t) \\ &= f_s(t) - f_s(t-h) \end{aligned}$$

由于密度核 f 在 0 以下递增，

$$f_s(t) \geq f_s(t-h)\,\forall\, t < s$$

因此

$$g(t) \geq 0\,\forall\, t < s.$$

类似地，由于 f 在 0 以上递减，

$$g(t) \leq 0\,\forall\, t > s + h.$$

在区域 $s < t < s + h$内，$f_s(t)$递减，而 $f_{s+h}(t)$递增，因此存在一点 $t^* \in (s, s+h)$，使得 $g(t^*) = 0$、$g(t) \geq 0\,\forall\, t < t^*$ 且 $g(t) \leq 0\,\forall\, t > t^*$。由于 f不必严格递增，这样的点可能不唯一。由于

$$\begin{aligned} f_s\left(s + \frac{h}{2}\right) &= f_s\left(s - \frac{h}{2}\right) \\ &= f_{s+h}\left(s - \frac{h}{2} - h\right) \\ &= f_{s+h}\left(s + \frac{h}{2}\right) \end{aligned}$$

$t^* = s + \frac{h}{2}$ 是一个合适的点。

回顾尖峰密度误差的定义（公式8）

$$\epsilon = \frac{1}{T}\int_0^T \left|\rho_r(t) - \rho_d(t)\right| dt$$

我们有

$$\left|\rho_r(t) - \rho_d(t)\right| = \begin{cases} f_s(t) - f_{s+h}(t) & \text{if } t < t^* \\ 0 & \text{if } t = t^* \\ f_{s+h}(t) - f_s(t) & \text{if } t > t^* \end{cases}$$

因此

$$
\begin{aligned}
\epsilon =& \frac{1}{T}\int_0^T \left|\rho_r(t)-\rho_d(t)\right| dt \\
=& \frac{1}{T}\int_0^{t^*} f_s(t)-f_{s+h}(t)\,dt + \frac{1}{T}\int_{t^*}^{T} f_{s+h}(t)-f_s(t)\,dt \\
=& \frac{1}{T}\int_0^{t^*} f_s(t)\,dt - \frac{1}{T}\int_0^{t^*} f_s(t-h)\,dt \\
&+ \frac{1}{T}\int_{t^*}^{T} f_s(t-h)\,dt - \frac{1}{T}\int_{t^*}^{T} f_s(t)\,dt \\
=& \frac{1}{T}\int_0^{t^*} f_s(t)\,dt - \frac{1}{T}\int_0^{t^*-h} f_s(t)\,dt \\
&+ \frac{1}{T}\int_{t^*-h}^{T} f_s(t)\,dt - \frac{1}{T}\int_{t^*}^{T} f_s(t)\,dt \\
=& \frac{2}{T}\int_{t^*-h}^{t^*+h} f_s(t)\,dt
\end{aligned}
$$

使用 $t^* = s + \frac{h}{2}$即可得到方程31。由于 $f \geq 0$ 减小 h 会降低误差。利用 $h < 0$ 的对称性，可以对 f作出同样的论证。

□

S5. 心率尖峰的分布

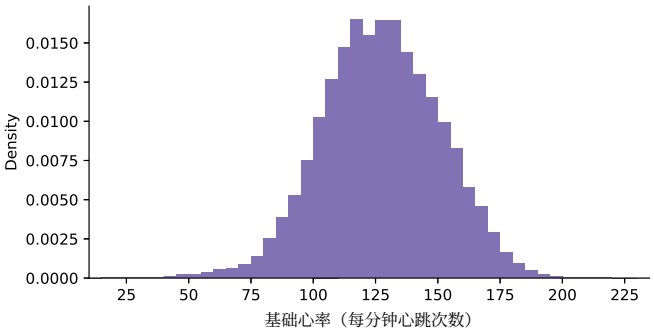


图 S5.1. 直方图显示了从 Crickles 的实数心率数据中检测到尖峰时的心率分布。为考虑不同运动员的活动次数和尖峰数量差异，我们为每位运动员分别生成了直方图，再将这些直方图取平均，得到此处所示的直方图。该分布呈现出类似正态分布的曲线，大多数尖峰检测到的心率范围为 $105-140$ bpm。

S6. 尖峰时间分布

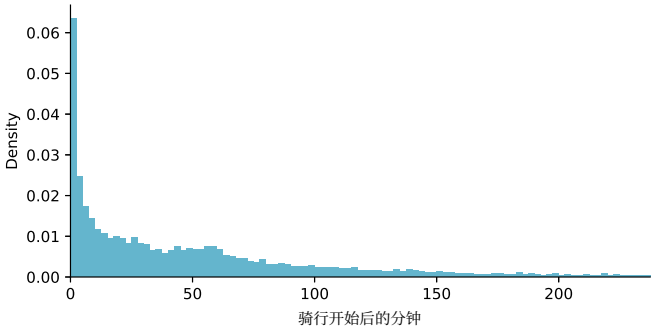


图 S6.1。 直方图显示了从 Crickles 的实数心率数据中检测到尖峰的时间分布。为考虑不同运动员的活动和尖峰数量差异，我们为每位运动员分别生成了直方图，然后对这些直方图取平均，得到此处的直方图。这表明，大多数尖峰均在时间序列开始后的前 10 分钟内被检测到，这可能是由于运动设备在活动开始时无法记录准确的心率，正如第 4 节所述。