arXiv:2603.01989v1 [math.ST] 2 Mar 2026

基于 Wasserstein 距离的时间序列数据亚稳态识别：通过变点检测和分段聚类

David Gentile
塔夫茨大学
david.gentile@tufts.edu

Joshua Huang塔夫茨大学
joshua.huang@tufts.edu

James M. Murphy塔夫茨大学
jm.murphy@tufts.edu

2026年3月3日

摘要

时间序列分析中的变点检测是一项困难且重要的应用统计问题，过去几十年来，人们为此开发了多种方法。本文将 Wasserstein度量用作多维时间序列数据中的变点识别工具，以无监督的方式识别时间序列中的簇。我们利用一维情形下最优传输代价的简洁性，在片段数远少于数据点数时，快速识别数据的分段（即一条轨迹上的一组变点）和聚类，且不对涉及的具体分布作任何参数假设。我们提出的变点检测方法在数据规模和样本维度上均呈线性扩展。我们在理想化的合成数据轨迹，以及来自分子动力学模拟和水下声学领域的真实轨迹上测试了该方法。我们发现，通过估计 Wasserstein度量导数得到变点并据此对这些时间序列进行分段，再将识别出的分段视为测度、依据 Wasserstein度量衡量相似性并进行聚类，能够成功识别过程分布律中的亚稳态。

1 引言

1.1 研究动机

变点检测（CPD）旨在识别时间序列的分段，将数据划分为少量内部连贯的序列。这是时间序列数据分析中的一项重要任务，已有许多涵盖广泛方法的不同技术来解决这一问题 [7]。作为一种统计工具，它应用于金融 [39, 30, 47], 气候建模 [53, 10, 27], 语音分析 [57, 18, 29], 边缘检测[58, 33, 32], 和健康监测 [40, 3, 69]。此类算法的任务是通过识别发生显著定性变化的特定观测值，对时间序列进行分段——例如，血压变得过高的时刻，或股票市场出现异常交易的时刻。在本研究中，我们主要关注利用概率测度上的Wasserstein度量来辅助变点检测，并结合经典的数据聚类技术，为大型且复杂的时间序列数据集生成精简描述，特别是分子动力学模拟（MDS）产生的轨迹。这种方法与一些CPD流程 [2, 70, 52], 中观察到的模式恰好相反：后者先进行聚类，再利用聚类结果识别数据分段；而我们则通过将分段视为原始信息的降维表示，利用变点检测来辅助数据聚类。

CPD 是研究时间序列数据的有力工具，因为尽管给定轨迹可能很长，即观测值数量很大，可能包含数百万个样本，但如果时间序列倾向于在少数几个亚稳态之间振荡，且观测值之间的方差较低，那么轨迹也可能存在冗余。因此，如果亚稳态的数量以及轨迹在这些状态之间转换的速率都相对较小，

将轨迹视为一系列表示其片段的经验分布，便可大幅减少应用基于距离的聚类技术所需的成对比较次数。因此，如何搜索这些变点是这一方法的关键问题。

本文的研究动机源于对 MDS 轨迹数据的研究。研究这些轨迹时，一个常见的假设是，在固定环境中，给定粒子的构型空间可以划分为若干区域，这些区域代表粒子的“亚稳”状态及其间的跃迁；空间中的亚稳区域是粒子倾向于长时间停留的区域，例如能量阱的盆地，而跃迁则是亚稳集合的补集。由此自然引出一个问题：我们如何根据粒子的轨迹数据构建其行为的降阶模型？解决这个问题的一种方法是基于分段的聚类，这是一种在 [20]中探讨过的技术。通过在时间序列数据中寻找变点，可以将数量极大的数据点缩减为相对较少的片段；同时利用数据的时间序列特性，减少计算相似度矩阵所需的成对距离数量，再通过密度峰值聚类算法 [55, 65]，根据该矩阵识别这些片段的聚类。随后，各个数据点继承其父片段的标签。

最初启发本工作的CPD方法之一是BarT [19]，，它是对SIMPLE算法的一种基于输运的改进 [35]，，二者都依赖于最大似然估计 [66]，因而必须对分布具有统计先验 [35] [19]。这两种方法都从面向优化的观点看待CPD问题：建立一个损失函数，该函数基于在给定一组变点下观测轨迹的对数似然，并加上一个正则化项；该正则化项通过对每个变点施加惩罚来调节算法产生的变点数量，而对多个可观测量中同时检测到的变化给予较低惩罚。这些优化程序代价极高，这意味着为正则化项寻找最优惩罚参数并不现实。通过将我们的统计观点从参数化转向非参数化，我们放弃了关于变点最优性的理论保证，但获得了一种能够更快分析轨迹的方法。

最优传输为具有空间属性的数据提供了一套特别适用的工具。具体而言，它能够量化两个概率测度之间的差异，同时编码这些测度所支撑空间的底层几何结构。当我们尝试通过观测数据辨别轨迹在何时离开或进入构型空间中的不同区域，并在其中停留较长时间时，这一点非常重要。还要注意，当分布的支撑集不同时，传输方法相较于其他常用的分布比较方法具有这一优势——例如，对于两个支撑集不相交的测度，Kullback-Leibler散度甚至没有定义 [66]。我们将时间序列的分段视为一系列经验概率测度，并运用这套传输工具解决 CPD 问题。我们使用移动窗口沿时间序列扫描，以最优传输代价衡量相邻片段之间的距离，然后将代表两个相邻片段的经验测度之间距离异常大的时刻识别为变点。

我们将时间序列视为某个具有隐含规律的随机过程的一次实现。如果该随机过程的规律连续演化，那么我们可以将其视为测度空间中的一条参数化曲线，即某种类似流形的空间中的一条粒子轨迹。如果这条粒子轨迹在一段时间内集中于流形的某个小区域，那么该随机过程就处于亚稳态。如果粒子正在亚稳态之间移动，则处于过渡态。通过识别变点和亚稳态之间的跃迁，对轨迹进行分段后，我们将采用经典的无监督聚类方法，把经验观测到的亚稳态片段归为一组，从而得到轨迹的降阶描述。

借鉴流形中粒子轨迹的类比，我们将采用以下启发式方法检测亚稳态：如果轨迹的速度相对较小，则粒子没有移动（因而处于亚稳态）。我们的方法是数据驱动的；一般来说，我们无法获得例如控制该轨迹的解析随机微分方程，因此必须根据可用数据近似曲线的微分。

为此，我们将采用 2-Wasserstein度量作为一种非参数统计工具，来分析含噪时间序列数据。具体而言，我们的假设如下：假设我们观测到一个定义在紧区间$[0,T]$上的随机过程 X_t，且存在某个时刻 $t_0 \in (0,T)$，此时过程的分布律发生变化。那么，如果 $\hat{\mu}_{t_0^-}$ 是由样本 $\{X_t\}_{t\le t_0}$ 生成的过程的经验分布，而 $\hat{\mu}_{t_0^+}$ 是由样本 $\{X_t\}_{t_0\ge t}$生成的经验分布，即，

$$\hat{\mu}_{t_0^-} = \sum_{t=0}^{t_0} \frac{1}{t_0}\delta_{X_t}, \quad \hat{\mu}_{t_0^+} = \sum_{t=t_0+1}^{T} \frac{1}{T-t_0}\delta_{X_t},$$

那么我们预期 t_0 应满足

$$\underset{0\le t\le T}{\operatorname{argmax}}\, W_2^2(\hat{\mu}_{t^-}, \hat{\mu}_{t^+}).$$

自然地，我们不会期望在给定轨迹中只出现一个变点——我们不再寻找 Wasserstein距离的全局极大值，而是考察局部极大值。为找到这些极大值，我们采用滑动窗口方法，固定窗口大小为 w，并针对每个 $t = w,$ ……，$T - w$ 计算由前后各 w 个样本定义的经验测度之间的 Wasserstein距离。随后应用截断值，识别出相邻窗口之间距离异常大的离群点。

识别出一系列变点后，我们便确定了轨迹的一种分段方式，将其划分为两两不相交且连续的子集。接下来，我们希望以无监督方式对这些片段进行分类。为此，我们再次将片段视为经验测度，并通过 2-Wasserstein度量衡量它们两两之间的相似性。我们采用的聚类方法是高级密度峰值聚类（ADPC）[55, 65]。ADPC 的优势在于它完全无监督，并能通过统计准则自动确定聚类中心的数量。

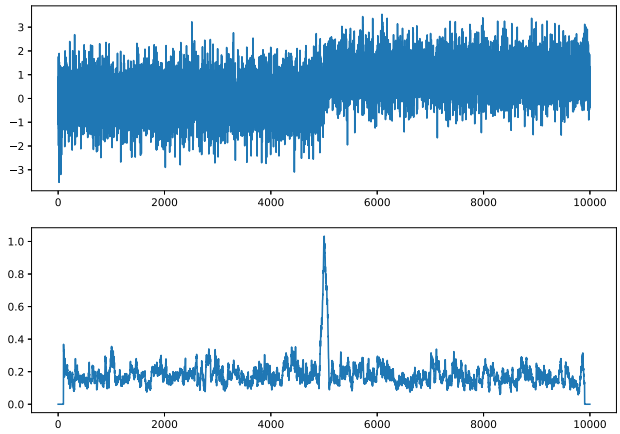


图 1：变点最简单的情形：过程分布律在时间 t=5000 发生突变。过程分布律的变化反映在 Wasserstein 度量导数中。上图：一段人工生成的时间序列，前 5000 个样本来自 $\mathcal{N}(0,\ 0.85)$ 分布，后 5000 个样本来自 $\mathcal{N}(1,\ 0.75)$ 分布。下图：时间序列 Wasserstein 度量导数的近似值，其极值出现在$t = 5000$，对应于分布的突变。

1.2 贡献

在本研究中，我们提出了一种非参数变点检测算法，并将其与无监督聚类方法相结合，以识别时间序列数据中的亚稳态。我们发现，对于理想化的合成数据集、分子动力学模拟数据和水下声学数据，该方法能够在概率分布空间中识别相干簇，从而为数据提供紧凑的模型。我们的 CPD 方法在数据规模上呈线性扩展，因此非常适用于大规模数据集；此外，只要数据不会在状态之间过快振荡，该方法还能显著减少簇识别所需计算的成对距离数量。

1.3 论文结构

第 2 节概述理解变点检测方法和所用聚类算法所需的背景知识。第 3 节介绍我们用于获得数值结果的算法。第 4 节展示实验的数值结果，并通过一系列易于理解的低维示例，证明我们的方法行之有效。第 5 节总结并讨论研究结果，同时提出拓展和延伸本文工作的若干未来方向。

2 背景

2.1 移动窗口变点检测的既有研究及其与相关文献的关系

CPD 涵盖了应用统计学领域的大量方法。若要全面了解该领域，请读者参阅综述 [7, 53, 42, 59] 以及教科书 [9]。一般而言，CPD 方法具有若干用于对其进行分类的特征，下面将对此进行概述。

CPD方法可分为在线方法和离线方法。在线方法会与其试图推断的过程同步运行，而离线方法则事后识别轨迹中的变化，一次性检查所有数据 [45, 49]。广受认可的在线方法包括 [2,36, 8], ，随着硬件性能提升使实时变点检测更为可行，这些方法也日益流行。我们的方法属于离线方法，因为我们使用基于整个轨迹的统计量，回溯性地对时间序列进行分段，然后在数据中寻找模式。

区分CPD方法时，还需考虑它们是否基于对底层数据作出的某种参数假设。参数化方法是最早一批用于解决CPD问题的方法之一，可追溯至沃尔德的序贯概率比检验 [63]。参数化方法和非参数方法各有优势与权衡：前者通过明确假设分布的底层性质（例如，过程平稳且服从高斯分布）来运作，而后者不作此类假设。

参数化模型在优化模型选择过程中可能会产生较高的计算成本。例如，SIMPLE [35], 是启发我们开展这项工作的初始算法之一，它是一种参数化 CPD 方案，通过优化一个损失函数来进行分析，而该损失函数必须基于某种统计先验推导出来，例如，假设数据来自某个分布族，如高斯分布、拉普拉斯分布、指数分布等。相比之下，非参数方法避免了模型选择问题，但往往需要在推理时使用更多数据 [7]。此前用于非参数 CPD 的工具包括但不限于柯尔莫哥洛夫-斯米尔诺夫检验 [49, 44, 46], 图 [14,13], ，以及我们尤其感兴趣的最优传输；相关研究考察了基于无分布方法 [15, 16,25], 、基于核的方法 [64], 以及 Wasserstein距离熵正则化的方法 [67]。我们的方法与 [17, 15], 中的方法相似，但也存在一些关键差异 —— 尤其是，我们使用度量导数，而非瓦瑟斯坦双样本检验。瓦瑟斯坦双样本检验 [21, 51] 是一种无分布的非参数检验，适用于实数线上的一维分布；它用于判断将一个分布单调重排为另一个分布后，所得结果是否与恒等函数存在统计显著差异。若不存在这种显著差异，则接受原假设（两个样本来自同一分布）。这与我们的方法有所不同：我们将在下文介绍的方法直接处理相邻分布之间的Wasserstein距离，并在这些数据中寻找离群点。此外，我们关注的是取值于非欧几里得域的数据集，例如三维环面。近期已有研究考察了平坦环面上的瓦瑟斯坦双样本检验 [28], ，这可能为今后扩展我们的工作提供一条途径。

对各种方法进行最后一项分类时，CPD方法可分为监督式或无监督式。前者所处的情形是变点预先已标注，训练的目标是开发一个模型，使其能够准确预测未纳入训练数据的标签。另一方面，无监督式CPD旨在不掌握真实标注的情况下识别数据中的结构——我们的方法正是在这种情形下运行。无监督方法的一种常见做法是将数据聚类为一组状态，并寻找簇标签发生变化的点。这种方法的经典例子包括滑动窗口与自底向上（SWAB）算法 [38] 和最小描述长度（MDL）算法 [50]。我们的方法则颠倒了这一范式，尝试根据识别出的变点为数据确定合适的标签。

移动窗口方案面临的一项挑战，尤其是我们的方法所面临的突出挑战，是如何选择窗口大小 w。由于窗口本质上是瞬时分布的统计估计量，因此窗口中自然是样本越多越好，这样便能更准确地估计真实分布。另一方面，在实际操作中，窗口大小也很容易设置得过大 w ——不妨假设存在一组真实变点：如果平均真实片段长度远小于窗口大小，那么一个典型窗口就会包含多个分段，因此在窗口大小较大的 w 情形下，我们的方法将无法检测到变点。因此，我们的算法将 w 设为可调参数，用户可以根据具体使用场景进行调整。

启发我们工作的主要前身是 SIMPLE 和 BarT 算法；这两种算法都是为分子动力学轨迹数据的特定应用而开发的 [19], ，后者是前者基于最优传输和 Wasserstein距离的扩展。SIMPLE 适用于一个在 J 维空间中 $\{\{Y_{j,t}\}_{t=1}^{T}\}_{j=1}^{J}$的过程，它首先选择一个对数似然函数 l ，用于表示在给定条件下观测到数据的合理性

变点集合 变点 $\{\tau\}_{i=1}^{K}$，然后通过动态规划算法进行最小化 m

$$\sum_{j=1}^{J}\sum_{i=0}^{K}\hat{l}(Y_{j,\tau_{j,i}+1},\ldots,Y_{j,\tau_{j,i+1}})-\lambda\sum_{i=1}^{K}q(S_i)$$

其中， S_i 是 $\{1,\ldots,J\}$ 的子集，对应于在时间 τ_i发生变化的维度，而 q： $2^{\{1,\ldots,J\}}\to$R 是满足以下条件的严格递增惩罚函数：当 A 和 B 为非空不相交子集时， $q(A\cup B)\leq q(A)+q(B)$；例如，在 SIMPLE 的原始实现中，作者取 $q(\cdot)=|\cdot|^{\alpha}$，适用于 $\alpha\in(0,1]$。这会促使算法选择许多维度同时发生变化的变点——任何变点都会产生惩罚，但仅在数据集的少数维度上表现出变化的点会受到更重的惩罚。

BarT 采用基于最优传输的观点（有关最优传输的详细信息，请参见第 2.2 节），对ˆ损失函数 l 进行调整，以考虑某个点属于亚稳态之间重心过渡的可能性。这一ˆ调整通过取 $l(\cdot)=\max\{L_{ms}(\cdot)\ L_{tr}(\cdot)\}$实现，其中 L_{ms} 和 L_{tr} 分别是对应于亚稳区域和过渡区域的对数似然函数。具体而言，该算法假设亚稳态服从拉普拉斯分布，并将 (μ_j,σ_j) 作为 j阶分段 $\{X_t\}_{t=\tau_j}^{\tau_{j+1}}$的参数。于是有

$$L_{ms}(\{X_t\}_{t=\tau_j}^{\tau_{j+1}})=\log\prod_{t=\tau_j}^{\tau_{j+1}}\frac{1}{2\sigma_j}\exp\left(-\frac{|X_t-\mu_j|}{\sigma_j}\right),\quad L_{tr}(\{X_t\}_{t=\tau_j}^{\tau_{j+1}})=\log\prod_{t=\tau_j}^{\tau_{j+1}}\frac{1}{2\sigma_{j,t}}\exp\left(-\frac{|X_t-\mu_{j,t}|}{\sigma_{j,t}}\right),$$

其中，$(\mu_{j,\ t},\ \sigma_{j,\ t})$ 对应于第 $(j-1)$ 个片段与第 $(j+1)$ 个片段之间 Wasserstein测地线的参数。这些参数很容易计算，因为两个拉普拉斯分布 γ_1 和γ_2 的 Wasserstein测地线本身也是拉普拉斯分布，而（单位时间长度）插值曲线的参数可直接由 γ_1 和 γ_2 [19] 的估计参数得出。我们的工作受到这种由输运驱动的方法启发，但本质上有所不同：我们放弃了显式损失函数（其优化成本高昂），转而利用 Wasserstein度量赋予经验分布集合的几何结构。

变点检测和时间序列数据的片段聚类问题也在 [17], 中得到探讨，尽管我们的工作采用了略有不同的方法来处理一维时间序列问题，并进一步将该方法拓展到多维非欧几里得时间序列数据，包括取向量值的数据以及满足流形假设的数据。例如，我们将把方法应用于分子动力学模拟数据，这些数据可表示为环面上的时间序列。

2.2 最优传输作为概率分布上的几何

最优传输是一个广泛的领域。若要全面了解这一主题，请参阅 [56,5, 61, 48]中的任一文献。下文将简要概述最必要的内容。接下来，我们将遵循 [56]中的论述。

设 $(\Omega\ d)$ 是一个完备可分度量空间；通常取 $\Omega=\mathrm{R}^d$，并采用标准欧几里得范数。设$\mathcal{P}_2(\Omega)$ 为 Ω 上有限方差概率测度构成的空间，设 $\mu,\nu\in\mathcal{P}_2(\Omega)$ 中的元素满足 $c:\Omega\times\Omega\to\mathrm{R}$，最后设 $\Pi(\mu,\nu)$ 为 $\Omega\times\Omega$ 上边缘分布分别为 μ 和 ν的概率分布集合。那么， μ和 ν 之间以 c为代价的最优传输代价为

$$OT(\mu,\nu):=\inf\left\{\int_{\Omega\times\Omega}c(x,y)\ \mathrm{d}\gamma(x,y)\ :\ \gamma\in\Pi(\mu,\nu)\right\}.$$

若取 $c(x,y)=d^2(x,y)$，则最优传输代价的平方根定义了 $\mathcal{P}_2(\Omega)$ 上的一个度量，称为 2-Wasserstein度量，记作 W_2。此外，在这种情况下，可以将 $(\mathcal{P}_2(\Omega)\ W_2)$ 正式解释为一个黎曼流形 [37, 11]。概率分布空间的流形解释赋予该空间一种动力学，这对我们方法的动机至关重要。当测度支撑在实数线上且相对于勒贝格测度绝对连续时，以下便捷恒等式成立：

$$W_2^2(\mu,\nu)=\int_0^1|F_\mu^-(x)-F_\nu^-(x)|^2\ \mathrm{d}x,\tag{1}$$

其中，F_μ^{-1}是广义逆累积分布函数，定义为

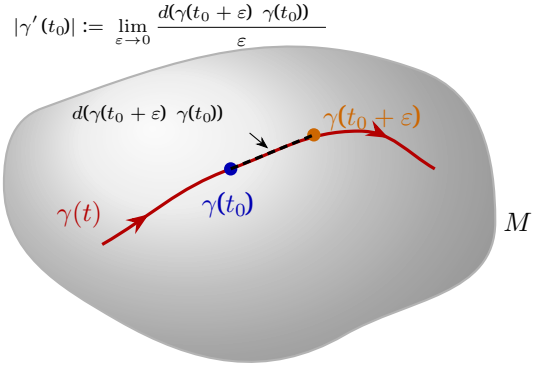


图 2：流形上的一条曲线；度量导数定义为差商的极限，其中分子是邻近点之间的差。

$$F_\mu^{-1}(x) = \inf\{t : \mu([-\infty, t) \geq x\}.$$

该公式在处理一维经验测度时如此便利，是因为它将计算最优传输代价的问题转化为对两个列表进行排序的问题，而这比求解一般的线性规划容易得多。

尽管 Wasserstein几何 [56], 中存在“梯度”的概念，但对于我们的目的，采用如下定义的 Wasserstein度量导数 [6], 就足够了：设 $\gamma\colon [0,1] \to \mathcal{P}_2(\Omega)$，且 $t_0 \in (0,\ 1)$。若以下极限存在，则称 γ 在 t_0 处具有度量导数 $|\dot{\gamma}|$：

$$|\dot{\gamma}| = \lim_{t \to t_0} \frac{W_2(\gamma(t), \gamma(t_0))}{t - t_0}. \tag{2}$$

对于任何可定义绝对连续性概念的度量空间，都可以定义度量导数。文献 [4],表明，在测度 $\gamma(t)$ 满足非常一般的条件时，式 (2) 中的极限存在；具体而言， $\gamma(t)$ 关于 $[0,1]$ 上的勒贝格测度在时间上绝对连续，且对几乎所有 t均成立。

度量导数的直观含义是，它表示过程的分布在概率分布空间中的瞬时速度。度量导数是备受赞誉的度量空间上梯度流理论的起点 [6], ，未来或可为拓展本研究提供良好的基础。

Wasserstein距离以一种能够编码底空间几何的方式为概率分布空间赋予度量，因此非常适合用于分析物理轨迹，也便于使用狄拉克云作为相关测度的估计。这使我们能够量化两个相邻窗口中点的差异，并近似计算轨迹在Wasserstein空间中的度量导数。借助这些信息，我们可以预测过程的分布律是否处于亚稳状态，或正处于亚稳态之间的跃迁过程中。

从较高层面来看，对我们的工作至关重要的是以下几点。设 $\{X_t\}_{t=0}^T$ 为定义在某个度量空间Ω上的随机过程；在实践中，我们将数据视为过程 [43], 的一个实现，并试图根据给定数据对该特定实现作出推断。对每一个 $t \in [0,T]$，我们可以用 γ_t 表示随机变量 X_t的分布。我们将假定这一参数化测度族位于 $\mathcal{P}_2(\Omega)$——即Ω上具有有限二阶矩的概率测度之集合——中，于是至少可以在形式上通过将Wasserstein度量解释为源于某种黎曼结构，而把 $\mathcal{P}_2(\Omega)$视为一个无穷维黎曼流形。若我们的随机过程 X_t 的定律以连续方式演化，则可将参数化测度族 γ_t 视为该流形中的一条曲线，亦即一条粒子轨迹。若这条粒子轨迹在某段时间内集中于流形的一个小区域内，则随机过程 X_t 处于一个亚稳态。若该粒子正在亚稳态之间移动，则它处于一个过渡态。先验地，我们通常并不知道一个过程的亚稳态将会是什么——因此，我们的目标是利用由2-瓦瑟斯坦距离所给出的 $\mathcal{P}_2(\Omega)$的度量结构，来识别粒子处于亚稳状态的那些时间区间（以下称为“片段”）。然后，

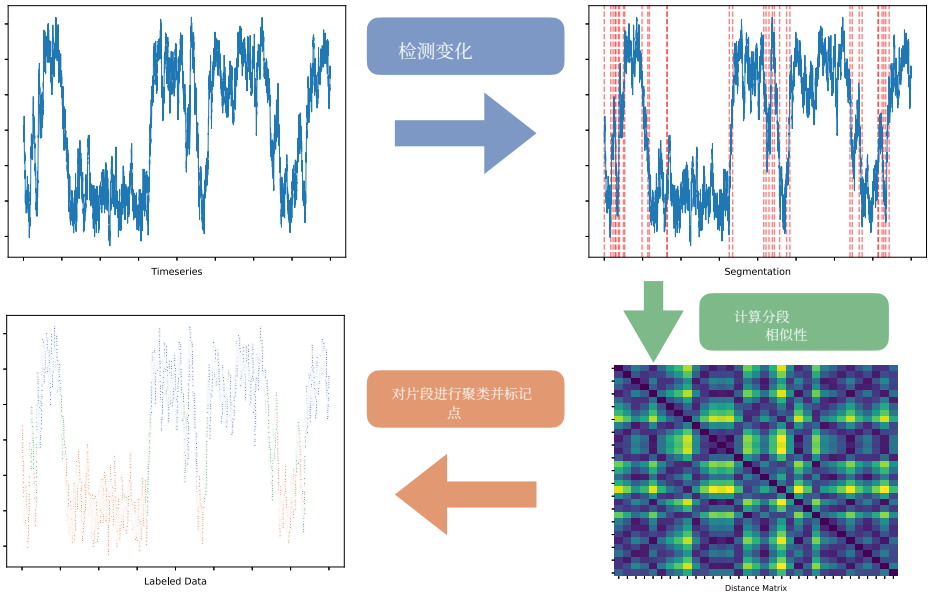


图 3：程序流程：首先，将时间序列拆分为一组互不重叠的片段；然后，根据 Wasserstein距离计算成对相似度矩阵；最后，对相似度矩阵应用聚类算法，学习各片段的标签，并由此推得时间序列中各数据点的标签。

使用经典的聚类方法，我们将通过把经验上观测到的亚稳态片段归并在一起，得到该轨迹的降阶描述。

2.3 基于相似度矩阵的聚类技术

将轨迹拆分为一系列片段后，接下来要做的是以连贯的方式将它们分组。换言之，我们要在降维数据集中识别簇。每个片段都由一个经验分布表示。延续 CPD方法的思路，我们将使用 2-Wasserstein度量量化片段之间的差异。由于 Wasserstein空间中的每个元素都是一个测度，因此适用于为片段赋予标签的聚类方法受到限制。例如，文献中曾研究过基于 Wasserstein度量的 K-均值方法，但关于其扩展性的理论结果仍在发展中，实证研究也表明其扩展性可能较差 [71]。为解决这一难题，我们将采用一种仅基于片段数据之间成对距离的聚类技术。另一方面，使用成对距离矩阵也进一步凸显了对时间序列进行分段可能有益的原因。总而言之，我们的数据集可能包含数百万个单独的数据点，因此在数据点层面计算成对距离矩阵的代价高得难以承受；然而，如果能将数据拆分为数量远少于原始数据点的片段，就可以减轻这一计算负担。

令 S 表示时间序列中的分段集合，$\hat{\nu}_i$ 表示与每个分段相关联的经验测度，我们按以下规则构造相似度矩阵

$$A = \left[\exp\left(-\frac{1}{2\sigma}W_2^2(\hat{\nu}_i, \hat{\nu}_j)\right)\right]_{i,j=1}^{|S|}$$

因此，仅构造相似度矩阵就需要 $\mathcal{O}(|S|^2)$ 次运算。在本文中，我们始终取 $\sigma = 1$。

高级密度峰值聚类是一种旨在识别多模态数据集 [55, 23] 拓扑结构的方法。该方案由若干启发式规则驱动，优点是完全无监督。首先，通过寻找周围点局部密度达到最大值的点，来确定数据集的“候选中心”。确定一组中心后，便可据此对环境空间进行 Voronoi划分。随后检查该划分的边界，以识别数据集拓扑结构中的鞍点，进而识别数据中的峰（众数）。所有这些步骤均由统计方法驱动，即对局部密度等量采用合适的估计量。在我们的场景中，使用 ADPC 的两大优势是

它不受维数限制，也就是说，它会尝试仅根据距离矩阵来估计由分段集合所隐含的内在维数；此外，它无需预先输入数据中簇数量的猜测。因此，我们的流程适用于复杂的轨迹，这些轨迹无法通过“目测”来判断：要么是因为数据本质上具有高维特征，要么是因为各簇的众数彼此过于接近，难以凭人的判断加以区分。

3 算法

众所周知，对于实数线上的一维经验测度，最优传输代价很容易计算n，且等价于对两个列表进行排序。对于两个经验测度 $\hat{\mu} = \frac{1}{n}\sum \delta_{x_i}$ 和$i=1$，根据 [48],

$\underline{n}\hat{\nu} = \frac{1}{n}\sum \delta_{y_i}$，我们假设它们已排序，使得对于每个 i，$x_i \le x_{i+1}, y_i \le y_{i+1}$，于是有$i=1$

$$W_2^2(\hat{\mu},\hat{\nu}) = \sum_{i=1}^{n} |x_i - y_i|^2.$$

使用式 (1) 处理原子数量不同的测度并不会明显更困难，因为经验累积分布函数的伪逆可通过对观测样本进行排序轻松获得。在我们的研究中，我们还将研究单位圆上的一维传输代价，相关细节见 [22]。本研究使用的所有传输算法均在 [26] 中实现。

除一维传输代价的计算方面外，我们的 CPD 还取决于两个超参数，w 和 q。窗口大小 w 指定了我们在每个时间点用于近似过程的分布律的经验分布样本量，而 q 是一个介于 0 和 1 之间的实数，它指定了相对于近似 Wasserstein度量导数分布的分位数截断值，表明某个数据点可能会被标记为变点。实质上，w 控制我们在任一给定时刻为近似时间序列分布的瞬时变化所要求的样本数量，而 q 控制算法对时间序列变化的敏感程度。例如，如果 q 取值较小，则会有更多点被列为候选点；反之，如果 q 取值接近 1，则识别出的候选点会很少。

我们的方法如下：对于固定的数据集 $\{X_t\}_{t=1}^{T}$ 和参数 w，q 我们通过以下方式近似度量 Wasserstein 导数

$$\left|\hat{\dot{\gamma}}\right| = W_2(\mu_t^-, \mu_t^+),$$

其中 $\mu_t^- = \{X_t\}_{t-w}^{t}$ $\mu_t^+ = \{X_t\}_{t}^{t+w}$。这一近似让我们得以了解过程的分布律在概率分布空间中移动的速度。然后，利用这些信息，我们寻找分布出现异常加速的区域。

为了更清楚地说明我们究竟要寻找什么，不妨考虑一种理想的物理情形：假设我们观察一个粒子在时间区间 $[0,1]$ 内于平面上的运动轨迹。在某个时刻 t_0，粒子开始从初始位置匀速移动到终点位置，并于时刻 t_1 到达，之后便停止运动。于是，粒子的速度可用方波表示，即示性函数 $\mathbb{1}_{[t_0,t_1)}(t)$。

我们的整体变点检测方案中隐含着一种滤波选择：先验地，如果我们设定分位数截断值 q，那么直接返回高于该阈值的点，就会将数据集的 $(1-q)$% 返回为变点候选；由于轨迹的连续性，其中许多候选点彼此相邻。换句话说，经过方案的前两个步骤，我们会得到一组变化区域，现在希望从中提取出精细筛选的变化点。

最后，我们只需逐分量处理，即可将我们的方法扩展到多维数据。具体来说，我们将时间序列的每个分量视为一个独立的过程，并针对每个维度执行分段和聚类程序。形式上，我们可以将整个流程视为一种变换 $F = L \circ S$ where$S$，它将输入时间序列 $\{\vec{X}_t\} \subset \mathrm{R}^D$ 映射为相应的一组片段 $\mathcal{S} = \{S_i\}_{i=1}^{k}$；而 $L: S_i \mapsto l_i$ 则将片段映射为标签，因此 F：$\{\{X_{t,d}\}_{t=1}^{T}\}_{d=1}^{D} \mapsto \{\{l_{t,d}\}_{t=1}^{T}\}_{d=1}^{D}$。随后，通过拼接标签 $(l_{t_0,\ 1} l_{t_0,\ 2} \ldots l_{t_0,\ d})$，为 $\vec{X}_{t_0}$ 赋予一个总体状态。

我们用以下命题总结上述内容：

算法1 变点检测算法（ComputeChangePoints）

输入： 时间序列数据 X_t，数据点总数 T，窗口大小 w，分位数阈值 q
输出： 变点时间集合

步骤1： 计算导数幅值数组
初始化 $\left|\hat{\dot{\gamma}}\right| \leftarrow$ 大小为（$T-2w$）的空数组
对于 $t = w$ **到** $T-w$，**执行**
    计算 $|\hat{\dot{\gamma}}|[t] \leftarrow X_t$的 $|\dot{\gamma}(t)|$ 估计值
结束循环

步骤 2: 基于分位数阈值识别候选时刻
计算 $\tau_q \leftarrow q$ 个分位数的 $|\hat{\dot{\gamma}}|$
$\mathcal{C} \leftarrow \{t : |\hat{\dot{\gamma}}|[t] \geq \tau_q,\ t = w, w+1, \ldots, T-w\}$ ▷ 候选集

步骤 3： 筛选连续片段中的拐点
将 $\mathcal{C}$划分为连续子序列 $\mathcal{S}_1$，$\mathcal{S}_2$，……，$\mathcal{S}_k$
初始化 $\mathcal{I} \leftarrow \emptyset$ ▷ 拐点集合
对每个分段 $\mathcal{S}_i$ **执行**
    识别分段 $\mathcal{S}_i$内的拐点
    将检测到的拐点添加到 $\mathcal{I}$
结束循环
返回 $\mathcal{I}$

算法 2 变点检测子程序：识别拐点

要求： 度量导数数据 $\{D_t\}_{t=1}^{T}$，分位数阈值 q
确保： 度量导数数据的拐点
步骤 1： 识别 $|\hat{\dot{\gamma}}|$中的连续子序列
$\mathcal{C} \leftarrow \{t : D_t >$ **分位数**$(D, q)\}$
$\mathcal{S} \leftarrow \{\{D_t\}_{t=t_i}^{t_{i+1}} : \{t_i, t_i+1, \ldots, t_i+k\} \subset \mathcal{C}$是一个最长子序列$\}$
步骤 2： 报告各子序列的极值
$\mathcal{I} \leftarrow \{\}$
对每个子序列 $S \in \mathcal{S}$ **执行**
    $\mathcal{I} \leftarrow \mathcal{I} \cup \{\text{argmin}\, S, \text{argmax}\ S\}$
结束循环
返回 $\mathcal{I}$

算法 3 亚稳态识别过程 (IdentifyStates)

要求：时间序列数据 $\{\{X_{t,i}\}_{t=1}^{T}\}_{i=1}^{d}$，窗口大小 $w \in \mathbb{N}$，分位数阈值 $q \in (0,1)$，聚类方法 Cluster 确保：点标签 $\{l_t\}_{t=1}^{T}$for $i = 1 : D$ do 步骤 1：识别时间序列的分段 3：$\mathcal{C} \leftarrow$ ComputeChangePoints(X_t, w, q) ▷ 识别变点$\mathcal{S} \leftarrow \{\{X_t\}_{t=t_i}^{t_{i+1}} : t_i \in \mathcal{C}\}$ ▷将数据划分为片段 步骤 2：计算片段之间的成对 W_2相似度 6：$D \leftarrow \left[\exp\left(-\frac{1}{2}W_2^2(S_i, S_j)\right)\right.$ for S_i, $\left.S_j \in \mathcal{S}\right]$ ▷计算相似度矩阵 步骤 3：根据相似度矩阵对片段进行聚类$L \leftarrow$ Cluster(D) 9：步骤 4：根据父片段的标签为各点赋标签$l_i \leftarrow [L(S(X_{t,i}))$ for $t = 1 : T]$end for 12：PointLabels $\leftarrow [l_1(t) l_2(t) \ldots l_d(t)$ for $t = 1 : T]$ ▷拼接标签以生成最终分配结果 return PointLabels

命题1。 对于固定的 w 和 q，对于输入 $\{\vec{X}_t\}_{t=1}^T \subset \mathbb{R}^D$，算法*1*的运行时间复杂度为 $\mathcal{O}(TD)$。固定 $w \in \{1, \dots \lfloor T/2 \rfloor\}, q \in (0,\ 1)$，并假设片段数量 N与输入数据的长度 $X = \{X_t\}_{t=1}^T$成正比。

证明。 该结论可直接由算法的构造得出：对于每个 $t = w, \dots, T-w$，我们执行两次排序操作，其复杂度为 $\mathcal{O}(w \log w)$，但相对于输入数据的大小和维度而言为 $\mathcal{O}(1)$。同样，用于识别已识别分段拐点的子程序，其复杂度不超过对每个分段的经验分布进行排序。由于假设分段数量随数据点数量线性增长，我们通过考虑每个分段的期望长度来进行均摊，该长度为 T/N。对大小为 T/N的列表执行 $N \propto T$次排序的复杂度为 $\mathcal{O}(N \frac{T}{N} \log(\frac{T}{N}) = \mathcal{O}(T)$。由于该程序会针对数据的每个 D维执行，因此算法的总渐近时间复杂度为 $\mathcal{O}(TD)$。□

4 方法与数值结果

对于接下来要考察的每个特定时间序列，我们都会应用本方法，并检查由此生成的点标签。为了便于可视化，我们将考察取值于 R 和三维环面上的时间序列。虽然几乎所有数据集都是计算机生成的，但其中一些比其他数据集更具人工性——例如，“玩具”数据集就是专门构建的，用作测试我们程序的试验场 [1]。分子动力学模拟取自 Deeptime Python 库 [31]以及用于化学模拟的 GROMACS 库[1]。我们的数据集集合中还包括一条由美国海军研究实验室提供的、通过物理实验生成的轨迹[60]。

4.1 生成的数据集

4.1.1 ToyData数据集

我们的第一个数据集来自一系列人工构造的轨迹，旨在为我们的程序提供理想化的运行环境。我们固定两个均值 $m_1,\ m_2 \in$ R 和一个方差 $\sigma > 0$，以确定两个彼此分离的拉普拉斯分布 $\mu_1,\ \mu_2$；同时固定过渡长度 L、样本总数 T 以及变点集合$\{t_i\} \subset \{1 \dots T\}$。随后，通过对 $X_t \sim \mu_1$进行采样（采样范围为 $t = 0,$ ……，t_1），再对 $X_t \sim [\mu_1, \mu_2]_t$的 Wasserstein插值进行采样（采样范围为 $t = t_1 + 1,$ ……，t_2），接着对 $X_t \sim \mu_2$进行采样（采样范围为 $t = t_2 + 1,$ ……，t_3），如此重复这一模式，直到时间 T，从而生成该轨迹。

玩具拉普拉斯 精确率与召回率			
# 真实CPS	# 检测CPS	精确率	召回率
38	38	0.89	0.89

图4：玩具数据集的精确率和召回率得分表。精确率和召回率是对$tol = 0, \dots, 100$ 的容差取平均值后得到的。

我们发现，该程序能够识别轨迹中的变点，并准确地对由此产生的片段进行分类；我们在表4中将输出与真实数据进行了比较总结。由于时间序列的方差较大，我们并不期望算法输出的变点集合与真实标注完全相同，而是期望每个候选变点都接近一个真实标注点。因此，我们对一系列容差下的精确率和召回率得分取平均值。

4.1.2 朗之万轨迹

接下来，我们研究将我们的算法应用于一条样本朗之万轨迹。支配粒子轨迹的能量景观有两个明显的势阱，中心分别位于 $+1$ 和 -1，两者之间有一个势垒。我们可以将其视为调控粒子从一个状态跃迁到另一个状态难易程度的因素。轨迹使用 OpenMM [24] 进行模拟，粒子质量设为 1amu，温度设为 500K，阻尼系数设为 10ps^{-1}.

我们观察到，使用我们的算法对轨迹进行分段，并通过 ADPC 进行聚类后，识别出三个不同的状态：其中两个分别对应一个势阱，第三个则捕捉了两个亚稳态之间的跃迁。

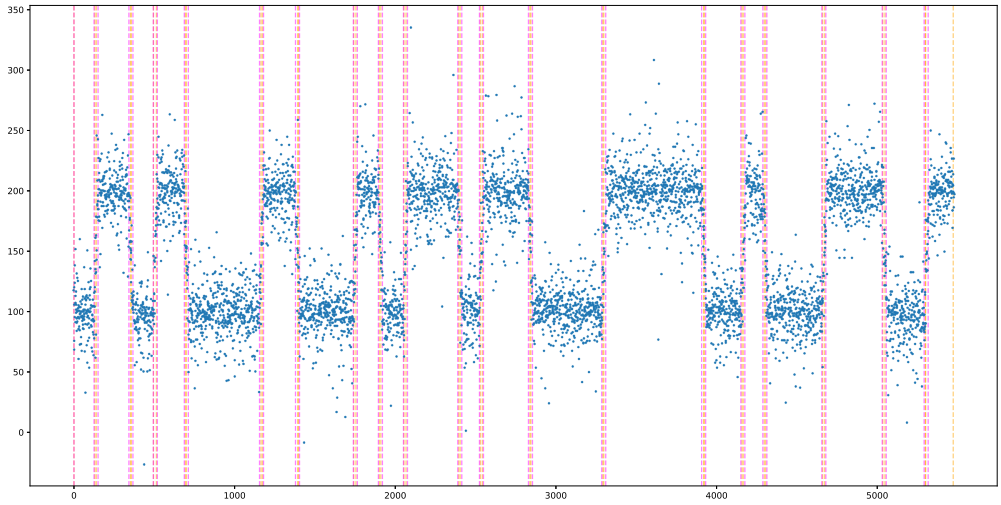


图5：在理想化场景中检测变点。我们从均值为100、方差为20的固定拉普拉斯分布，以及均值为200、方差为20的固定拉普拉斯分布中交替采样，生成一个合成数据集。我们通过从两个状态之间的Wasserstein测地线（McCann插值）采样来模拟状态间的跃迁；轨迹上的变点以红色虚线标出。

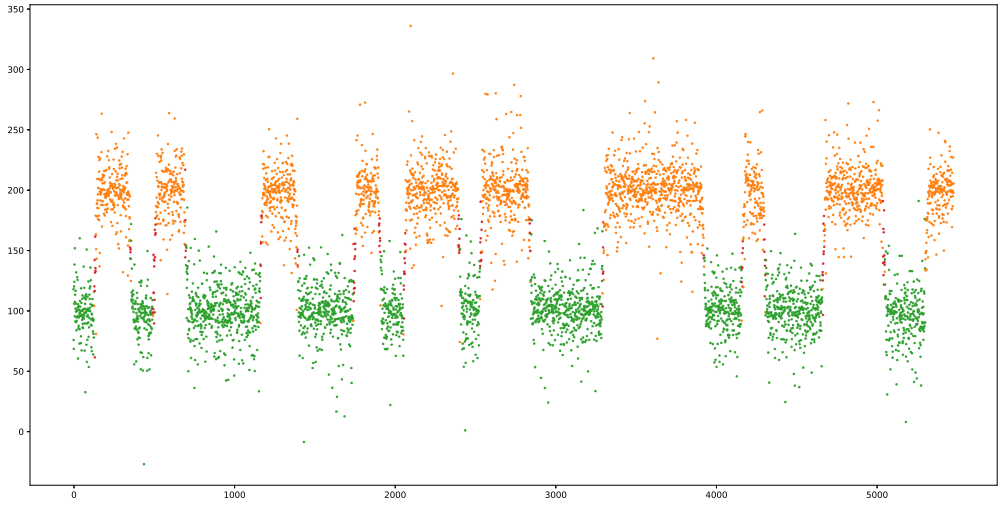


图6：使用高级密度峰值聚类算法以无监督方式学习片段簇标签。在本例中，聚类方案识别出两个亚稳态，并为过渡区域单独分配一个标签。获得此聚类结果所用的参数为 $w = 25$，$q = 0.95$.

图7：基于拉普拉斯分布的玩具轨迹分段与聚类



图8：为将我们的结果与标准的“开箱即用”聚类方法进行比较，我们使用 K均值对数据集进行聚类。所得聚类未能反映底层真实分布的方差，并呈现出硬划分超平面。

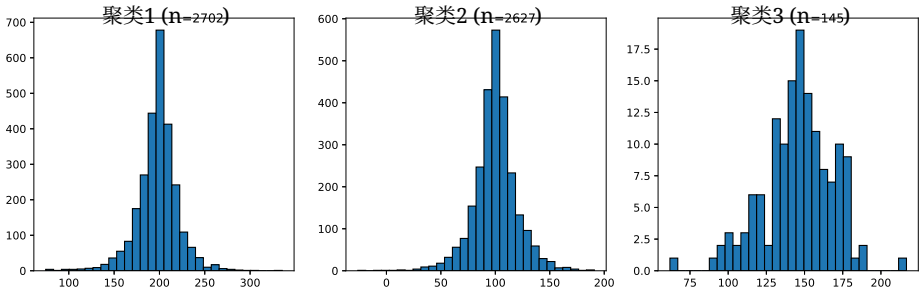


图9：使用 ADPC 的基于分段的聚类所学习到的玩具数据集各簇的经验分布。

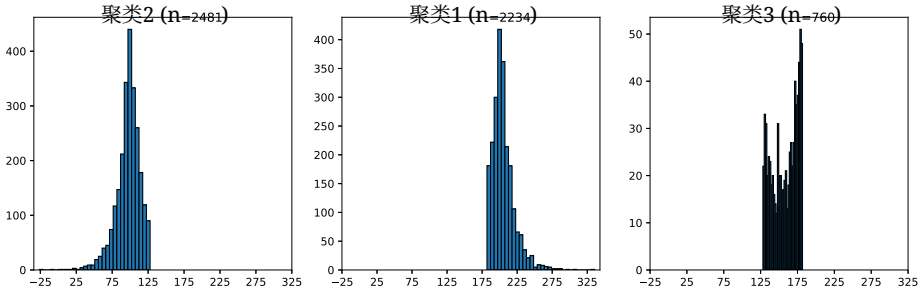
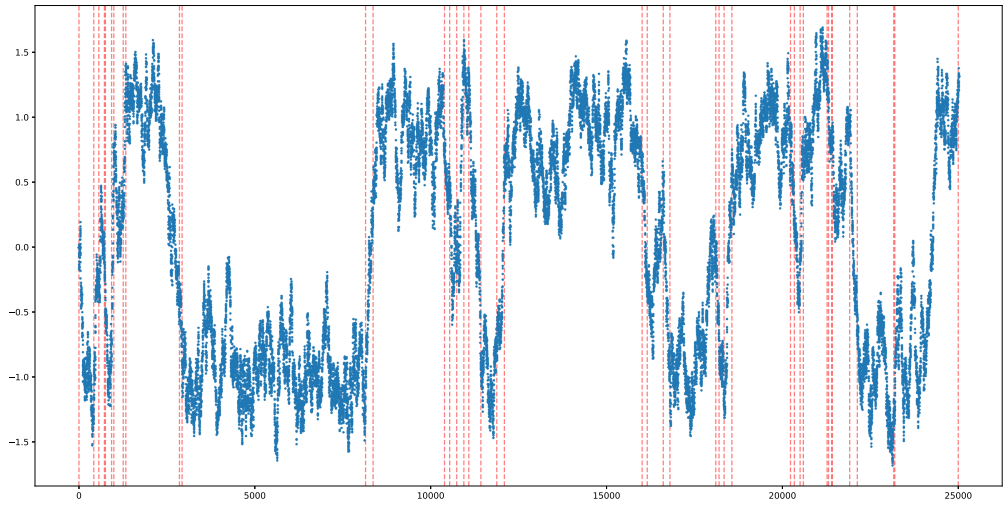


图10：使用 K均值所学习到的玩具数据集各簇的经验分布。



Fi图11：朗之万轨迹中的前25,000步，并标注检测到的变点。

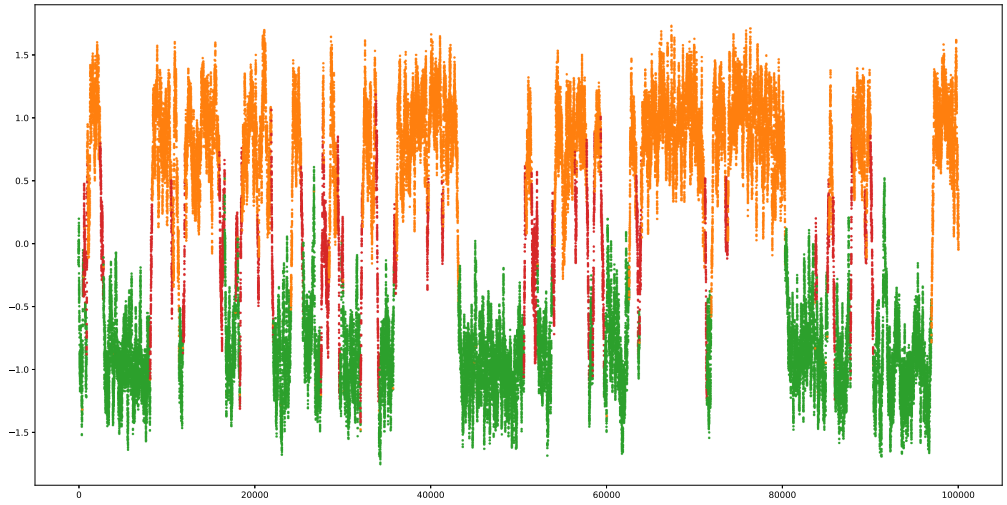


图12：使用高级密度峰值聚类对朗之万轨迹进行基于分段的聚类。用于获得此聚类的参数为 $w = 215,\ q = 0.86.$

图13：朗之万轨迹分段与聚类

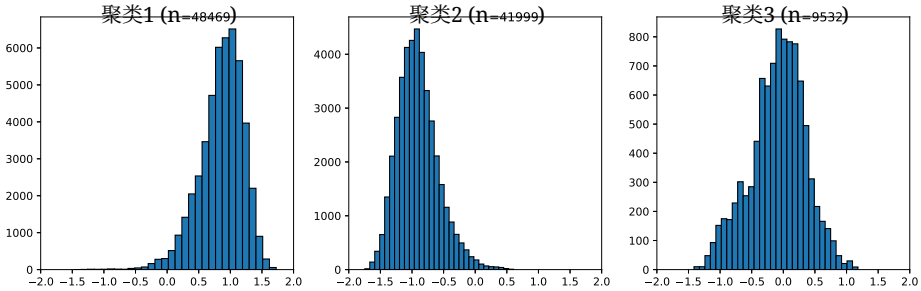


图14：通过基于分段的聚类与高级密度峰值聚类识别出的朗之万轨迹数据集中的簇的经验分布

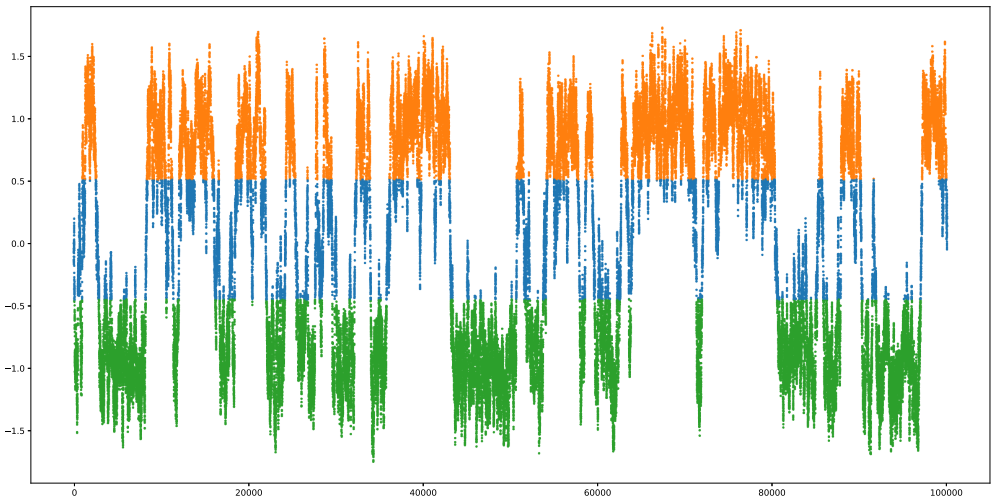


图15：为了将我们的结果与标准的“开箱即用”聚类方法进行比较，我们使用 K 均值对数据集进行聚类。所得聚类未能反映真实底层分布的方差，并呈现出硬划分超平面。

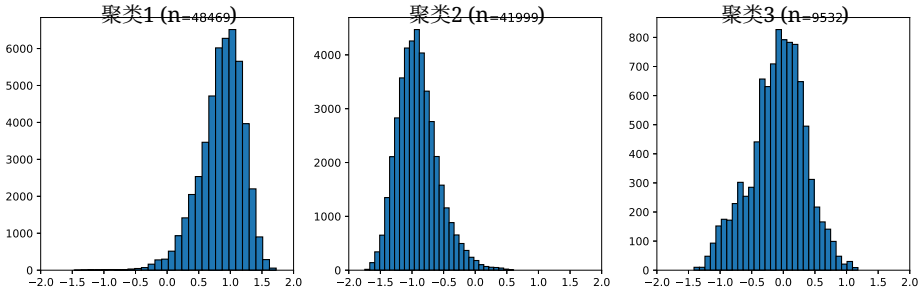


图16：通过基于分段的聚类与 ADPC 学得的朗之万数据集中的簇的经验分布。

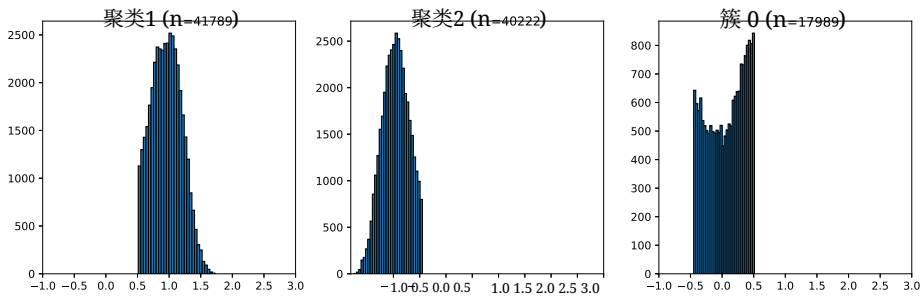


图 17：通过 K 均值学习得到的朗之万数据集各簇的经验分布 s.

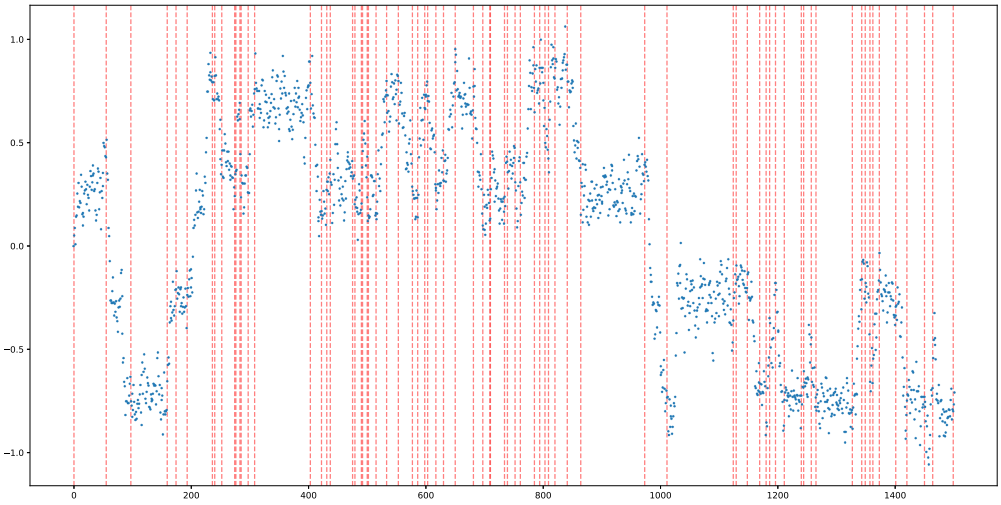


图 18：Prinz 势中的粒子轨迹，标注了检测到的变点 .

4.1.3 Prinz 势中的粒子

现在我们进一步增加示例的复杂性，将我们的方法应用于 Prinz 势中的一条粒子轨迹。这条轨迹在定性上与前一条 Prinz 轨迹相似——能量景观中有四个势阱，粒子可以在其中移动，各相邻势阱之间势垒的高度略有不同。

该轨迹由欧拉–丸山型积分器生成，其势能为

$$V(x) = 4\left(x^8 + 0.8e^{-80x^2} + 0.2e^{-80(x-0.5)^2} + 0.5e^{-40(x+0.5)^2}\right),$$

并由以下方程驱动

$$x_{t+1} = x_t - \frac{h\nabla V(x_t)}{md} + \sqrt{\frac{2hkT}{md}}\eta_t$$

其中 $\eta_t \sim \mathcal{N}(0,\ 1)$ [31]。

使用我们的算法对轨迹进行分段，并通过 ADPC 对这些片段进行聚类后，我们发现识别出了五种状态：其中一种对应势能景观中的每个不同势阱，另有一种似乎对应于两个跃迁势垒最低的势阱之间的过渡区域。

[1]我们的代码 和数据可在 https://github.com/dcgentile / mfpy

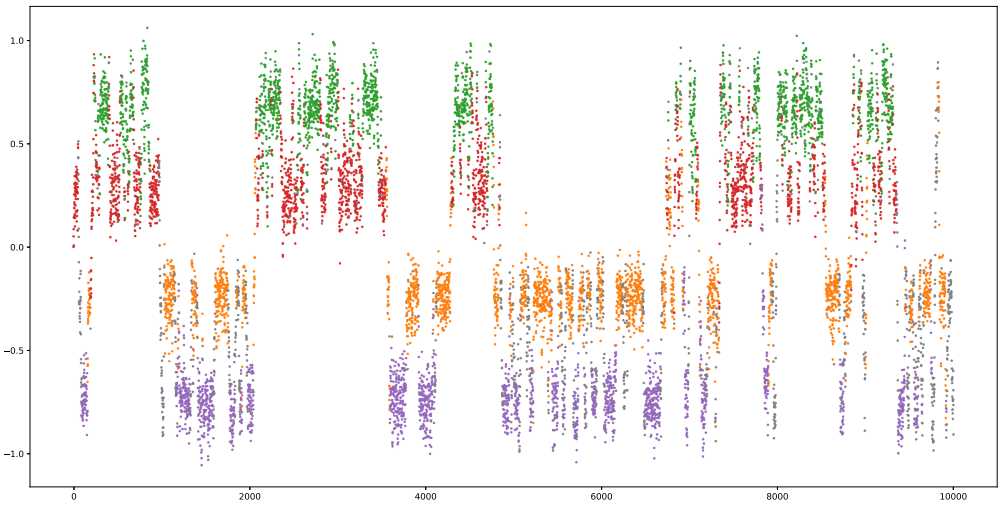


图19：Prinz 势中粒子的轨迹，采用基于分段的聚类，并结合高级基于密度峰值的聚类进行聚类。按数量排名前四的簇以颜色标示，而这些簇之外的所有点均显示为灰色。获得该聚类所用的参数为 $w = 16, q = 0.5$.

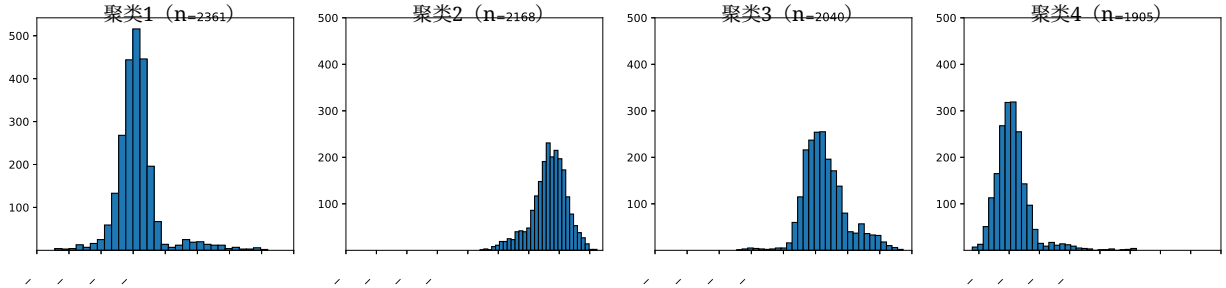


图20：通过基于分段的聚类并使用高级基于密度峰值的聚类，在Prinz势数据集中识别出的前四个簇的经验分布。

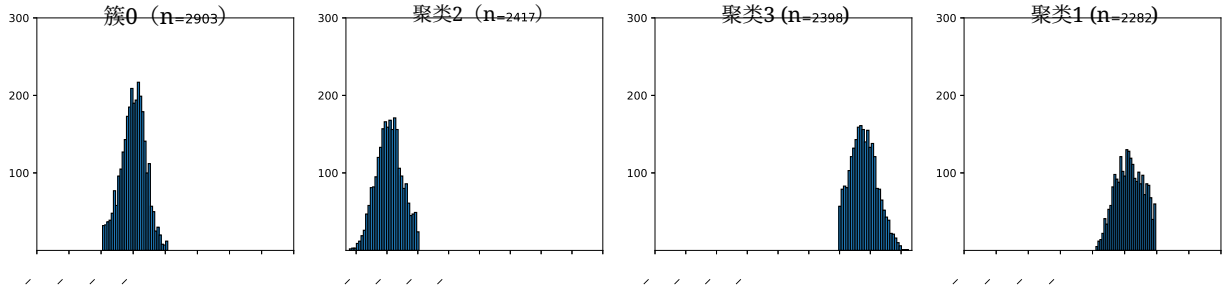
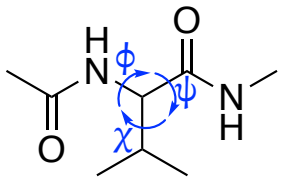


图21：由K均值学习得到的Prinz数据集中各簇的经验分布。



2-乙酰氨基-*N*,3-二甲基丁酰胺
化学式：$C_8H_{16}N_2O_2$

图 22：Ace-Val-NMe 的结构，并标注了主链二面角。

4.2 分子动力学轨迹

现在，我们将注意力转向 Ace-Val-NMe（缬氨酸二肽，或 VDP）的模拟分子轨迹，这些轨迹由 GROMACS 分子模拟软件包 [1] 生成。通常，VDP 轨迹通过马尔可夫状态模型技术 [12]，建模，其中数据使用 PCCA++ [54] 等技术进行聚类，然后根据簇的识别结果估计可逆转移矩阵。这类轨迹在计算化学领域已得到充分研究，并常用于该领域算法的基准测试 [20]。现代机器学习技术也被用于对这些轨迹建模，通过深度学习找到用于描述模型的最优非线性函数基 [41], [68]。尽管这些轨迹是在数十个空间维度上模拟的（VDP 含有 28 个原子，因此原则上需要 84 个维度来表示每个原子位置的坐标），对应于构成该分子的原子，但众所周知，通过追踪所谓的“二面角”，即 ϕ、ψ 和 χ，VDP 轨迹的维度可以降至 $d = 3$。这些角度具有周期性，因此轨迹可以嵌入三维环面。

因此，这意味着现在必须将我们的方法应用于非欧几里得情形。原则上，这不会改变我们的方法，但意味着我们不能再依赖 (1) 来提升计算效率，而必须处理圆上的传输代价 [34]。我们的分段方法使 ADPC 能够识别出 9 个明显分离的不同簇。

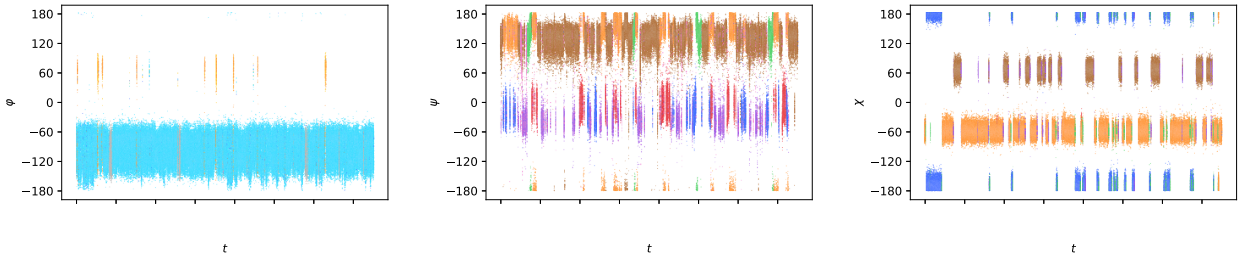
另一个值得关注的数据集是 Ace-Ala-NMe（丙氨酸二肽，或 ADP）；与 VDP 一样，它也常用于多维尺度算法的基准测试。与之前的数据集不同，ADP 给我们的方法带来了困难，因为它在亚稳态之间跃迁得很快，从而导致欠采样问题。具体来说，在我们关注的亚稳态中，有些状态出现得并不频繁，且每次持续时间都很短；即使尝试手动识别轨迹中较短部分的变点，我们也发现有些片段仅包含十几个点。正如前文所述，这意味着必须采用非常小的窗口大小 w，以避免窗口中包含某个角度下多个不同的状态；但当然，选择较小的 w 意味着我们看到的分段往往无法很好地代表任何累积分布函数。

值得庆幸的是，VDP是一条相对更易于处理的分子动力学轨迹，这一点可以从其各条角轨迹的相对稳定性中观察到。我们的流程识别出的9个占据最显著的簇彼此清晰分离，约占全部数据的82%，其余部分则属于更小的簇。由于VDP不会像ADP那样经历缓慢而频繁的跃迁，因此在VDP数据集上得到的簇分离要干净得多，并且在分量级上考察时也一目了然。观察VDP数据的各个分量，我们发现该流程识别出的状态之间差异相对细微：这些状态在外观上保持一致，但支撑集存在重叠。这一点在 ψ 角上最为明显，算法在 $[-60, 60]$ 附近识别出3个特征鲜明的簇，并在 $[60, 210]$附近又识别出3个特征鲜明的簇。

4.3 水下声学

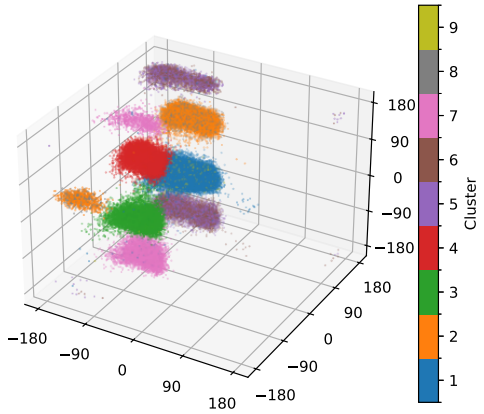
我们的真实世界数据集来自美国海军研究实验室开展的一项实验 [60]。在该实验中，两台水下扬声器被拖曳在一艘船后，在9秒内发射四种不同的信号，每种信号重复两次。其中三个样本为线性调频（LFM），即信号从某一频率开始，线性增加到另一个终止频率；另一个样本则为连续波形。

我们发现，将算法稍作修改并应用于数据的声谱图，便能够为样本识别出一组合适的变点。在这种情况下，我们的方法之所以得以简化，是因为计算

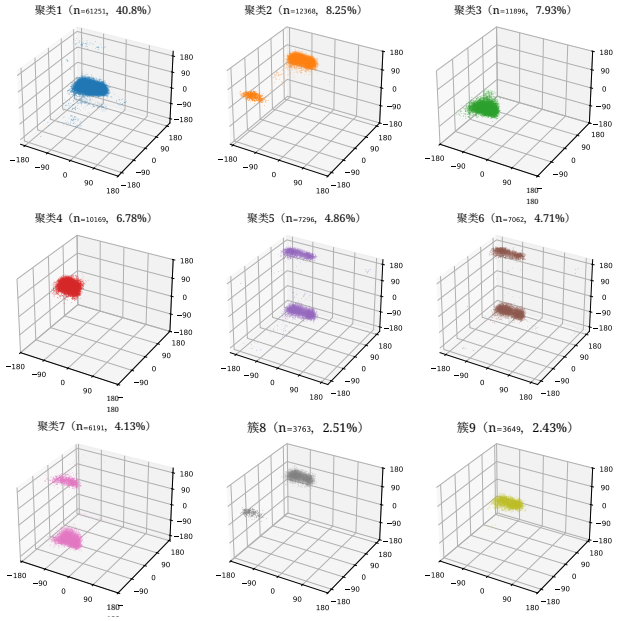


（a） ϕ 角　　（b） ψ 角　　（c） χ 角

图23：通过基于分段的聚类与圆周最优传输对VDP二面角进行标记。每个角度均经过CPD和片段聚类处理。随后按分量合并各簇，生成最终标记。



（a）使用高级密度峰值进行基于分段的聚类后，缬氨酸二肽轨迹的最终标签分配。每个角度使用的参数为$w = 50$ $q = 0.95$



(b) 使用基于分段的高级密度峰值聚类对缬氨酸二肽轨迹进行聚类。注意：由于仅统计了数量最多的前 9 个簇，各百分比之和不为 1。这些簇合计占数据集的 82.43%。

图 24：缬氨酸二肽聚类

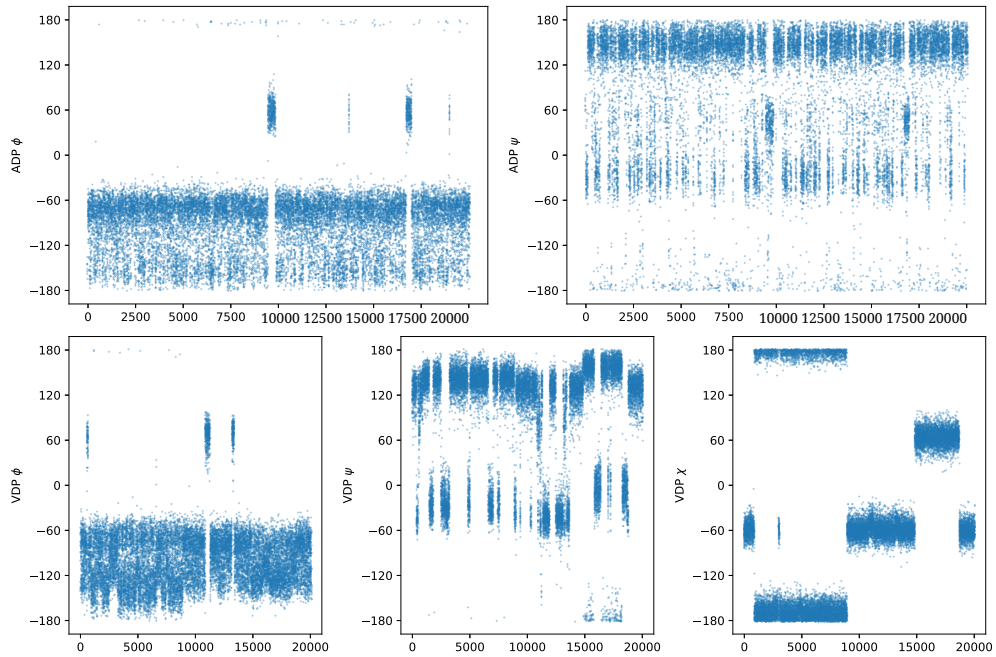


图25：ADP与VDP的分量角比较。我们的算法难以处理ADP，因为它在状态之间来回快速跳变，状态之间存在大量重叠，这一点在ADP的 ψ 角（右上）中尤为明显。VDP对我们而言是更易处理的数据集，因为其变化更清晰且持续时间更长。

水下声学数据精确率与召回率					
# 真实CPS	# 检测CPS	精确率	召回率	最大精确率	最大召回率
31	36	0.513	0.597	0.694	0.806

图 26：玩具数据集的精确率和召回率得分表。精确率和召回率是对容差 $tol = 1536$、.....、12800 取平均值。

对此类信号进行声谱图分析，自然会得到一组按顺序排列的功率谱密度。这与将VDP的高维原子坐标表示映射到三维环面时所发生的降维在性质上有所不同。在后一种情况下，以完整原子坐标表示的样本与以主链二面角表示的样本之间存在一一对应关系。而在我们的水下声学实验中，发生的是另一种压缩。我们不计算模数转换器读取的采样电压之间的Wasserstein距离，而是使用短时傅里叶变换（STFT）来考察频域内容随时间的变化。使用形状参数为0.25 [62], 的Tukey窗，对原始电压数据进行STFT计算后得到的声谱图是一个数组：其列表示信号在某个时间窗内的频率内容分布，而其行表示某一频带的频率内容随时间的集中程度。声谱图的分辨率取决于STFT中每个窗所用的样本数；在本例中，我们使用 $n = 512$。总体而言，这显著减少了所计算的Wasserstein距离数量：我们不再逐个滑过原始数据的每个样本，而只需计算声谱图中相邻列之间的Wasserstein距离。在4.3节中，时域样本总数为421，848，但计算声谱图后需要比较的功率谱密度仅有941个。由于有真实数据，我们还像处理人工生成的玩具数据集那样，为算法的输出生成精确率和召回率评分表，只是稍作调整：容差现在随每个傅里叶窗的样本数缩放，本例中该数取为 $n = 512$。数据的采样率为每秒11718个样本，因此，相应的容差范围可理解为能够检测到大约在 $0.1 - 1.0$ 秒内发生的变化。

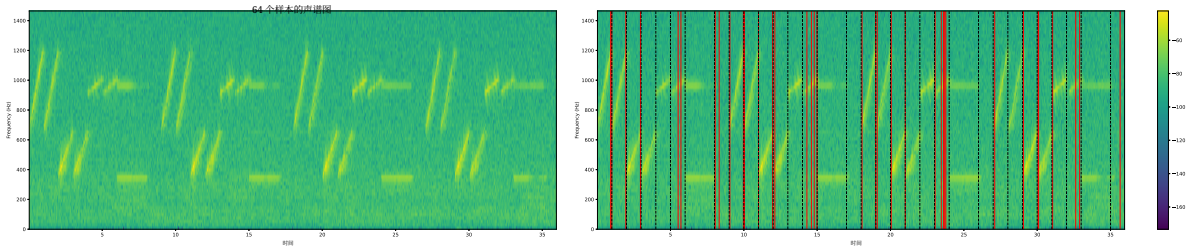


图 27：标注了检测到的变点的水下声学数据。此算法修改版本的参数为：短时傅里叶变换的窗口大小为 $n = 512$，以及 $q = 0.96$.

5 讨论与结论

在这项工作中，我们研究了将Wasserstein度量用作识别时间序列数据中变点的工具，并通过将问题提升到分布空间，将大型数据集聚类，使数据点能够被视为较大分段的一部分，从而减少构建相似度矩阵所需的成对计算次数。我们的工作拓展了对这一方法的先前研究：我们去除了对数据统计结构的参数假设，并扩大了适用于基于传输的变点检测的相关数据集范围，使其涵盖多维数据。我们通过对数据集的每个分量分别执行变点检测和聚类，规避了维数灾难。

未来，可以通过采用一种与用于分段的传输驱动框架更加协调的聚类方法，来扩展本研究。具体而言，可以利用 Wasserstein重心的几何结构，实现一种 K-均值类算法，或者将基于密度峰值的聚类的某种变体应用于 Wasserstein空间。由于计算 Wasserstein重心所涉及的计算复杂度，这种方法可能需要对Wasserstein距离进行熵正则化。此外，参数选择 w 和 q 的优化也值得研究。文献中的其他研究已经探讨了如何为基于柯尔莫哥洛夫-斯米尔诺夫检验的 CPD算法选择最优窗口大小；在基于 Wasserstein 的算法背景下，也可以考虑开展类似研究。

此外，如何将这种方法与计算化学领域广泛采用的 Markov状态建模方法相协调，也值得研究。

致谢： 作者衷心感谢 NSF DMS-2309519 和 NSF DMS-2318894 提供的支持。我们还要感谢美国海军研究实验室的 Roger M. Oba 和 Laurie T. Fialkowski 同事，他们就相关问题进行了有益交流，并提供了水下声学数据集；同时感谢 Hannah J. Bates 博士凭借其在化学绘图软件方面的知识，积极协助创建图表。

参考文献

[1] Mark James Abraham 等。《GROMACS：通过从笔记本电脑到超级计算机的多级并行实现高性能分子模拟》。载于：*SoftwareX* 1–2（2015），第19–25页。ISSN：2352-7110。数字对象标识符：10.1016/j.softx.2015.06.001.[2] Ryan Prescott Adams 和 David J. C. MacKay。贝叶斯在线变点检测。2007。数字对象标识符：10.48550/ arXiv.0710.3742。arXiv：0710.3742 [stat].[3] Abu Al Hassan 和 Phong Ba Dao。《连接数据与诊断：风力涡轮机趋势监测与变点检测集成的系统综述和案例研究》。载于：*Energies* 18.19（2025），第5166页。ISSN：1996-1073。数字对象标识符：10.3390/en18195166.
[4] Luigi Ambrosio。《度量空间值有界变差函数》。载于：*Annali della Scuola Normale Superiore di Pisa - Classe di Scienze* 17.3（1990），第439–478页。ISSN：2036-2145.[5] Luigi Ambrosio、Elia Bru´e 和 Daniele Semola。最优传输讲义。第130卷。UNITEXT。Cham：Springer国际出版公司，2021。ISBN：978-3-030-72161-9 978-3-030-72162-6。数字对象标识符：10.1007/978-3-030-72162-6.
[6] Luigi Ambrosio、Nicola Gigli 和 Giuseppe Savare。梯度流：度量空间与概率测度空间中的梯度流。Springer Science & Business Media，2005。ISBN：978-3-7643-2428-5。

[7] Samaneh Aminikhanghahi 和 Diane J. Cook。《时间序列变点检测方法综述》。载于：知识与信息系统 51.2（2017），第339–367页。ISSN：0219-3116。数字对象标识符：10.1007/s10115-016-0987-z.[8] 时间序列分段在线算法 | IEEE会议出版物 | IEEE Xplore。https://ieeexplore.ieee.org/abstract/document/989531.[9] Basseville 和 Nikiforov 著作：《突变检测》。https://people.irisa.fr/Michele.Basseville/kniga/.[10] Claudie Beaulieu、Jie Chen 和 Jorge L. Sarmiento。《变点分析作为检测气候突变的工具》。载于：英国皇家学会哲学汇刊 A辑：数学、物理与工程科学 370.1962（2012），第1228–1249页。ISSN：1364-503X。数字对象标识符：10.1098/rsta.2011.0383.[11] Jean-David Benamou 和 Yann Brenier。《蒙日-康托罗维奇质量传输问题的计算流体力学解法》。载于：数值数学 84.3（2000），第375–393页。ISSN：0029-599X，0945-3245。数字对象标识符：10.1007/s002110050002.[12] Gregory R. Bowman、Vijay S. Pande 和 Frank No´e 编。《马尔可夫状态模型及其在长时间尺度分子模拟中的应用导论》。第797卷。实验医学与生物学进展。Dordrecht：施普林格·荷兰，2014。ISBN：978-94-007-7605-0 978-94-007-7606-7。数字对象标识符：10.1007/978-94-007-7606-7.[13] Hao Chen 和 Lynna Chu。《基于图的变点分析》。载于：统计学及其应用年度评论 10.第10卷，2023（2023），第475–499页。ISSN：2326-8298，2326-831X。数字对象标识符：10.1146/annurev-statistics-122121-033817.[14] Hao Chen 和 Nancy Zhang。《基于图的变点检测》。载于：统计学年刊 43.1（2015），第139–176页。ISSN：0090-5364，2168-8966。数字对象标识符：10.1214/14-AOS1269.[15] Kevin C. Cheng。《基于最优传输的概率时间序列分析模型与方法》。博士论文。美国马萨诸塞州：塔夫茨大学，2023。ISBN：9798374400946.[16] Kevin C. Cheng 等。《统计变点检测中的匹配滤波》。载于：IEEE开放信号处理期刊 1（2020），第159–176页。ISSN：2644-1322。数字对象标识符：10.1109/OJSP.2020.3035070.[17] Kevin C. Cheng 等。《基于最优传输的变点检测与时间序列分段聚类》。载于：ICASSP 2020——2020年IEEE国际声学、语音与信号处理会议（ICASSP）。2020，第6034–6038页。数字对象标识符：10.1109/ICASSP40776.2020.9053344.[18] Roman Cmejla 等。《用于言语流利度和构音障碍自动评估的贝叶斯变点检测》。载于：语音通信 55.1（2013），第178–189页。ISSN：0167-6393。数字对象标识符：10.1016/j.specom.2012.08.003.[19] Jovan Damjanovic、Yu-Shan Lin 和 James M. Murphy。《将分子动力学时间序列中的变化建模为瓦瑟斯坦重心插值》。载于：第十四届国际抽样理论与应用会议。2023.[20] Jovan Damjanovic、James M. Murphy 和 Yu-Shan Lin。《CATBOSS：基于分段拆分的轨迹聚类分析》。载于：化学信息与建模期刊 61.10（2021），第5066–5081页。ISSN：1549-9596。数字对象标识符：10.1021/acs.jcim.1c00598.[21] Eustasio del Barrio 等。《基于L2-瓦瑟斯坦距离的拟合优度检验》。载于：统计学年刊 27.4（1999），第1230–1239页。ISSN：0090-5364。JSTOR：120161.[22] Julie Delon、Julien Salomon 和 Andrei Sobolevski。《圆上的蒙日代价快速传输优化》。载于：应用数学学会期刊 70.7（2010），第2239–2258页。ISSN：0036-1399。数字对象标识符：10. 1137/090772708.[23] Maria d'Errico 等。《通过非参数密度峰值聚类自动构建高维数据集地形》。载于：信息科学 560（2021），第476–492页。ISSN：0020-0255。数字对象标识符：10.1016/j.ins.2021.01.010.[24] Peter Eastman 等。《OpenMM 7：用于分子动力学的高性能算法快速开发》。载于：PLOS计算生物学 13.7（2017）。Robert Gentleman 编，e1005659。ISSN：1553-7358。数字对象标识符：10.1371/journal.pcbi.1005659.[25] Kamil Faber 等。《WATCH：面向高维时间序列数据的瓦瑟斯坦变点检测》。载于：2021年IEEE国际大数据会议（大数据）。2021，第4450–4459页。数字对象标识符：10 . 1109 / BigData52589.2021.9671962。arXiv：2201.07125 [cs].[26] R´emi Flamary 等。《POT：Python最优传输》。载于：机器学习研究期刊 22.78（2021），第1–8页。ISSN：1533-7928.[27] Colin Gallagher、Robert Lund 和 Michael Robbins。《具有长期趋势的气候时间序列中的变点检测》。载于：气候期刊 26.14（2013），第4994–5006页。ISSN：0894-8755，1520-0442。数字对象标识符：10.1175/JCLI-D-12-00704.1。

[28] Javier Gonz´ález-Delgado 等：《基于平坦环面上的 Wasserstein 距离的两样本拟合优度检验及其与结构生物学的相关性》。载于：《电子统计学杂志》17.1（2023），第1547–1586页。ISSN：1935-7524, 1935-7524。数字对象标识符：10.1214/23-EJS2135。[29] Vishwa Gupta：《使用深度神经网络进行说话人变点检测》。载于：2015年IEEE声学、语音与信号处理国际会议（ICASSP）。2015年，第4420–4424页。数字对象标识符：10.1109/ICASSP. 2015.7178806。[30] Reza Habibi：《金融领域的贝叶斯在线变点检测》。载于：《金融互联网季刊》17.4（2021），第27–33页。ISSN：2719-3454。数字对象标识符：10.2478/fiqf-2021-0025。[31] Moritz Hoffmann 等：《Deeptime：用于从时间序列数据中学习动力学模型的 Python 机器学习库》。载于：《机器学习：科学与技术》3.1（2021），第015009页。ISSN：2632-2153。数字对象标识符：10.1088/2632-2153/ac3de0。[32] Jun S Huang 和 Dong H Tseng：《边缘检测的统计理论》。载于：《计算机视觉、图形学与图像处理》43.3（1988），第337–346页。ISSN：0734-189X。数字对象标识符：10.1016/0734-189X(88)90087-4。[33] Shenyang Huang 等：《动态图库的拉普拉斯变点检测》。载于：第26届ACM SIGKDD知识发现与数据挖掘国际会议论文集。KDD ’20。美国纽约：计算机协会，2020年，第349–358页。ISBN：978-1-4503-7998-4。数字对象标识符：10. 1145/ 3394486.3403077。[34] Shayan Hundrieser、Marcel Klatt 和 Axel Munk：《圆周最优传输的统计学》。载于：《创新应用的方向统计学：纪念弗洛伦斯·南丁格尔诞辰二百周年文集》。Ashis SenGupta 和 Barry C. Arnold 编。新加坡：Springer Nature，2022年，第57–82页。ISBN：978-981-19-1044-9。数字对象标识符：10.1007/978-981-19-1044-9 4。[35] 《识别大型系统中的局部变化：生物分子模拟的变点检测》。https://www.pnas.org/doi/epdf/10.1073/pnas.1415846112。数字对象标识符：10.1073/pnas.1415846112。[36] B. Jackson 等：《区间数据最优划分算法》。载于：《IEEE信号处理快报》12.2（2005），第105–108页。ISSN：1558-2361。数字对象标识符：10.1109/LSP.2001.838216。[37] Richard Jordan、David Kinderlehrer 和 Felix Otto：《福克–普朗克方程的变分表述》。载于：《SIAM数学分析杂志》29.1（1998），第1–17页。ISSN：0036-1410。数字对象标识符：10 . 1137 / S0036141096303359。[38] E. Keogh 等：《时间序列分段在线算法》。载于：2001年IEEE国际数据挖掘会议论文集。2001年，第289–296页。数字对象标识符：10.1109/ICDM.2001.989531。[39] Kyungwon Kim 等：《基于新型累积和方法的金融时间序列无监督变点检测与趋势预测》。载于：《IEEE Access》10（2022），第34690–34705页。ISSN：2169-3536。数字对象标识符：10.1109/ACCESS.2022.3162399。[40] Siqi Liu、Adam Wright 和 Milos Hauskrecht：《临床决策支持系统规则监测的变点检测方法》。载于：《医学人工智能》91（2018），第49–56页。ISSN：0933-3657。数字对象标识符：10.1016/j.artmed.2018.06.003。[41] Andreas Mardt 等：《用于分子动力学深度学习的 VAMPnets》。载于：《Nature Communications》9.1（2018），第5页。ISSN：2041-1723。数字对象标识符：10.1038/s41467-017-02388-1。[42] Yue S. Niu、Ning Hao 和 Heping Zhang：《多变点检测：选择性综述》。载于：《统计科学》31.4（2016），第611–623页。ISSN：0883-4237。JSTOR：26408091。[43] Bernt Øksendal。《随机微分方程：导论及其应用》。Springer Science & Business Media，2010年。ISBN：978-3-642-14394-6。[44] Oscar Hernan Madrid Padilla 等。《最优非参数变点检测与定位》。2019年。arXiv：1905。10019 [计算机科学、统计学]。[45] E. S. PAGE：《连续检验方案》。载于：《Biometrika》41.1-2（1954），第100–115页。ISSN：0006-3444。数字对象标识符：10.1093/biomet/41.1-2.100。[46] Jan Peter Patist：《最优窗口变化检测》。载于：第七届IEEE国际数据挖掘会议研讨会（ICDMW 2007）。2007年，第557–562页。数字对象标识符：10.1109/ICDMW.2007.9。[47] Andrey Pepelyshev 和 Alexey Polunchenko：《通过最快变点检测方法进行实时金融监测》。载于：《统计学及其界面》10.1（2017），第93–106页。ISSN：1938-7989。数字对象标识符：10.4310/ SII.2017.v10.n1.a9。[48] Gabriel Peyr´e 和 Marco Cuturi。《计算最优传输》。2020年。数字对象标识符：10.48550/arXiv.1803.00567。arXiv：1803.00567 [stat]。[49] Dominique Picard：《时间序列中的变点检验与估计》。载于：《应用概率进展》17.4（1985），第841–867页。ISSN：0001-8678，1475-6064。数字对象标识符：10.2307/1427090。

[50] Thanawin Rakthanmanon 等：《时间序列插补：时间序列流聚类需要忽略部分数据》。载于：2011年IEEE第11届国际数据挖掘会议。2011年，第547–556页。数字对象标识符：10.1109/ ICDM.2011.146。[51] Aaditya Ramdas、Nicol´as Garc´ıa Trillos 和 Marco Cuturi：《关于 Wasserstein 双样本检验及相关非参数检验族》。载于：《Entropy》19.2（2017）。ISSN：1099-4300。数字对象标识符：10.3390/e19020047。[52] Sasank Reddy 等：《利用手机确定交通方式》。载于：《ACM传感器网络汇刊》6.2（2010），13:1–13:27。ISSN：1550-4859。数字对象标识符：10.1145/1689239.1689243。[53] Jaxk Reeves 等：《气候数据变点检测技术综述与比较》。载于：《应用气象与气候学杂志》46.6（2007），第900–915页。ISSN：1558-8424，1558-8432。数字对象标识符：10.1175/JAM2493.1。[54] Susanna R¨oblitz 和 Marcus Weber：《基于 PCCA+ 的模糊谱聚类：在马尔可夫状态模型和数据分类中的应用》。载于：《数据分析与分类进展》7.2（2013），第147–179页。ISSN：1862-5355。数字对象标识符：10.1007/s11634-013-0134-6。[55] Alex Rodriguez 和 Alessandro Laio：《通过快速搜索和发现密度峰进行聚类》。载于：《Science》344.6191（2014），第1492–1496页。数字对象标识符：10.1126/science.1242072。[56] Filippo Santambrogio。《应用数学家的最优传输：变分法、偏微分方程与建模》。第87卷。《非线性微分方程及其应用进展》。Cham：Springer International Publishing，2015年。ISBN：978-3-319-20827-5 978-3-319-20828-2。数字对象标识符：10.1007/978-3-319-20828-2。[57] Rasool Tahmasbi 和 Sadegh Rezaei：《用于语音活动检测的 GARCH 模型变点检测》。载于：《IEEE音频、语音与语言处理汇刊》16.5（2008），第1038–1046页。ISSN：1558-7924。数字对象标识符：10.1109/TASL.2008.922468。[58] Jean-Yves Tourneret、Michel Doisy 和 Marc Lavielle：《受乘性噪声干扰的多个变点贝叶斯离线检测：在SAR图像边缘检测中的应用》。载于：《信号处理》83.9（2003），第1871–1887页。ISSN：0165-1684。数字对象标识符：10.1016/S0165-1684(03)00106-3。[59] Charles Truong、Laurent Oudre 和 Nicolas Vayatis：《离线变点检测方法的选择性综述》。载于：《信号处理》167（2020），第107299页。ISSN：0165-1684。数字对象标识符：10.1016/j.sigpro.2019.107299。[60] Altan Turgut、Laurie T. Fialkowski 和 Jeffrey A. Schindall：《浅水环境夏季剖面下波导不变量深度依赖性的实测》。载于：《美国声学学会杂志》139.6（2016），EL184–EL189。ISSN：0001-4966。数字对象标识符：10.1121/1.4953581。[61] C´edric Villani。《最优传输》。M. Berger 等编。第338卷。《数学科学基础丛书》。柏林、海德堡：Springer，2009年。ISBN：978-3-540-71049-3 978-3-540-71050-9。数字对象标识符：10.1007/ 978-3-540-71050-9。[62] Pauli Virtanen 等：《SciPy 1.0：Python科学计算的基础算法》。载于：《Nature Methods》17.3（2020），第261–272页。ISSN：1548-7105。数字对象标识符：10.1038/s41592-019-0686-2。[63] A. Wald：《统计假设的序贯检验》。载于：《数学统计年刊》16.2（1945），第117–186页。ISSN：0003-4851，2168-8990。数字对象标识符：10.1214/aoms/1177731118。[64] Jie Wang、Rui Gao 和 Yao Xie：《采用核投影 Wasserstein 距离的双样本检验》。载于：第25届国际人工智能与统计会议论文集。PMLR，2022年，第8022–8055页。[65] Yizhang Wang 等：《密度峰值聚类算法：2014—2023十年回顾》。载于：《专家系统与应用》238（2024），第121860页。ISSN：0957-4174。数字对象标识符：10.1016/j.eswa.2023.121860。[66] Larry Wasserman。《统计学概要：统计推断简明课程》。Springer统计学丛书。美国纽约州纽约：Springer，2004年。ISBN：978-1-4419-2322-6 978-0-387-21736-9。数字对象标识符：10.1007/978-0-387-21736-9。[67] Matthew Werenski 等：《通过最优传输研究秩能量统计量：连续性、收敛与变点检测》。载于：《IEEE信息论汇刊》70.7（2024），第5168–5190页。ISSN：0018-9448，1557-9654。数字对象标识符：10.1109/TIT.2024.3367182。[68] Hao Wu 和 Frank No´e：《从时间序列数据中学习马尔可夫过程的变分方法》。载于：《非线性科学杂志》30.1（2020），第23–66页。ISSN：1432-1467。数字对象标识符：10.1007/s00332-019-09567-y。[69] Rui Wu、Chao Liu 和 Dongxiang Jiang：《用于复杂机械系统可靠预测与健康管理的无监督贝叶斯变点检测方法》。载于：《可靠性工程与系统安全》245（2024），第110037页。ISSN：0951-8320。数字对象标识符：10.1016/j.ress.2024.110037。[70] Yu Zheng 等：《利用原始 GPS 数据学习交通方式，服务于万维网上的地理应用》。载于：第17届万维网国际会议论文集。WWW ’08。美国纽约：计算机协会，2008年，第247–256页。ISBN：978-1-60558-085-2。数字对象标识符：10.1145/1367497.1367532。

[71] Yubo Zhuang、Xiaohui Chen和Yun Yang。“用于聚类概率分布的Wasserstein K-均值”。载于：神经信息处理系统进展 35（2022），第11382–11395页。