arXiv:2603.05700v1 [stat.ME] 5 Mar 2026

变点检　　测量流式细胞术细胞群体的变点检测

Yik Lun Kei [*1] **Qi Wang** [*1] **Paul A. Parker** [1] **Francois Ribalet** [2] **Sangwon Hyun** [1]

摘要

海洋中遍布浮游植物，其光合作用总量可与所有陆生植物相当，因此对碳循环和气候系统至关重要。流式细胞术的最新进展使海洋学家能够沿科研航次航线测量单个细胞的光学特征，从而获得具有单细胞分辨率的微生物数据。在海洋微生物生态学中，检测细胞计数式浮游生物分布对环境响应发生骤变的位置，是一项重要任务。本文提出一种潜在空间高斯混合专家模型，用于检测重复采样且经过聚类的浮游植物观测中的变点。通过低维表示先验均值的偏移来识别变点，并采用组融合套索惩罚项强制其呈现分段常数结构。随后通过交替方向乘子法求解优化问题。将该方法应用于流式细胞术数据后，结果识别出一个具有重要科学意义的变点，该变点与两个海洋区系之间的过渡带相吻合。

1. 引言

浮游植物是微观的光合生物，构成海洋食物网的基础，并为全球氧气产量作出重要贡献（Field etal.，1998）。分析浮游植物群落的变化对于理解海洋生物化学过程、生态系统动态和长期气候变率至关重要（Rib-alet et al.，2025）。流式细胞术的最新进展使研究人员能够沿着航线连续采集浮游植物的单细胞测量数据（Sosik & Olson，2007；Olson etal.，2017）。在每个小时的时间点，都会测量大量细胞，每个细胞均由三种光学特征表征：红色荧光、橙色荧光，以及细胞直径。这些测量构成了高频时间序列，包含高维观测，能够反映不同环境梯度下浮游植物群落结构的变化（Hyun et al.，2023a）。然而，流式细胞术数据的高维性、细胞群体天然的聚类特征以及大量重复测量，都给统计分析带来了重大挑战。

分析流式细胞术时间序列的一项基本任务，是识别浮游植物物种的分布及其与环境条件之间关系发生突变的位置。这些变化可能对应水团之间的转变、营养盐可利用性的变化，以及其他环境驱动因素。在实践中，流式细胞术数据中的分布变化往往具有异质性，且彼此交织。因此，准确定位这些变点能够以数据为依据划定环境转变，并刻画海洋环境条件发生变化的位置。

近几十年来，针对无重复观测且不含协变量的单变量和多变量时间序列，人们提出了大量变点检测方法。(Bai &Perron，1997) 将问题表述为利用线性模型检测结构突变，并通过对分段常数均值进行最小二乘估计来识别变点。(Killicket al.，2012) 提出了剪枝精确线性时间算法，该算法采用惩罚似然表述，以线性计算复杂度实现多重变点检测。(Matteson &James，2014) 提出了一种非参数分裂程序，通过使用多变量距离度量递归地划分时间序列来检测分布变化。同样，(Chen &Zhang，2015)、(Chu &Chen，2019) 和 (Song & Chen，2022a) 提出了非参数方法，通过刻画不同时期经验分布之间的差异来识别变点。此外，(Garreau& Arlot，2018) 和 (Song &Chen，2022b) 采用基于核的方法提取高维特征，以刻画分布差异。

然而，由于存在若干结构性挑战，许多现有的变点检测方法无法直接应用于流式细胞术数据。首先，细胞自然聚集成不同的群体（即物种），导致每个时间点观测到的混合分布复杂且呈多峰。

[1]加州大学圣克鲁兹分校统计系 [2]华盛顿大学海洋学院。通讯作者：Sangwon Hyun <shyun2@ucsc.edu>。

预印本。*2026*年*3*月*9*日。

其次，测量数据具有高维特征，且每个时间点都有重复测量，而大多数现有方法都假设每个时间点只有一次观测。第三，盐度和光照强度等环境协变量会影响混合组成以及各个群体的特征，因此难以确定哪些因素驱动了变化。由此，迫切需要一种能够处理重复且聚类分布，同时纳入协变量效应的变点检测方法。

为应对这些挑战，我们提出一种仅包含解码器的潜在空间混合模型，用于在高频重复测量的细胞数据中进行变点检测。具体而言，我们的框架使用多元高斯混合专家解码器，对给定环境协变量时单细胞观测的条件分布进行建模。混合权重、成分均值和方差由神经网络参数化，网络以观测到的协变量和一个低维表示作为输入；该低维表示由每个时间点观测到的所有细胞共享。低维表示由高斯先验管理，其均值参数通过流式细胞术数据学习得到。随后，通过检测随时间变化的先验均值中的偏移来识别变点，并利用组融合 LASSO 惩罚强制其呈现分段常数结构。这一建模方式使解码器和先验均值能够联合学习，同时明确考虑协变量的影响。我们通过交替方向乘子法（ADMM）求解由此得到的优化问题。据我们所知，这是首个能够直接对细胞群体层面的多元、重复测量、聚类且依赖协变量的数据进行变点检测的方法。

本文其余部分安排如下。第2 节介绍我们的潜在空间混合模型，并对变点检测问题进行形式化。第3 节介绍交替方向乘子法学习算法和朗之万动力学。第4节说明变点定位和模型选择的流程。第5 节使用真实协变量的合成数据评估我们的方法。第6节展示了该方法在一次海洋科考航次采集的真实流式细胞术数据上的应用。第7 节总结全文，并讨论其局限性和未来研究方向。

2. 潜在空间模型

2.1. 模型设定

我们研究一次海洋考察航次期间采集的流式细胞术数据，其中环境条件和单个细胞的测量值按时间顺序记录。在每个时间点 t，我们采集 $\{\boldsymbol{y}_b^t\}_{b=1}^B \subset \mathbb{R}^3$，它们对应于三项光学测量（红色荧光，橙色荧光和细胞直径），测量对象为 B 个单个细胞。令 $\boldsymbol{x}^t \in \mathbb{R}^p$ 表示时间t观测到的协变量，例如温度、海表盐度和亚硝酸盐水平。所有细胞，无论所属物种为何，都具有相同的协变量 $\boldsymbol{x}^t$，这反映了从海洋中采集细胞时共同的环境条件。此外，我们假设这些细胞可以聚类为 K 个类别，分别对应不同的物种。

对于 $\boldsymbol{y}^t \in \mathbb{R}^3$ 在特定时间 t，我们考虑一个潜在空间高斯混合专家解码器:

$$\boldsymbol{y}^t \sim P(\boldsymbol{y}^t|\boldsymbol{x}^t,\boldsymbol{z}^t) = \sum_{k=1}^{K} \boldsymbol{\pi}_k(\boldsymbol{x}^t,\boldsymbol{z}^t)\times \mathcal{N}^{(k)}\big(\boldsymbol{y}^t;\boldsymbol{\beta}^{(k)}(\boldsymbol{x}^t,\boldsymbol{z}^t),\boldsymbol{\Sigma}^{(k)}(\boldsymbol{x}^t,\boldsymbol{z}^t)\big) \tag{1}$$

其中，$K \in \mathbb{N}$ 表示簇的数量，且假设 K 已知。由于目标是检测变点，而非对单个细胞进行聚类，因此可以根据科学依据提供合理的 K，以完成检测任务。在流式细胞术中，细胞群体通常会被聚类，因此高斯混合分布可作为描述不同物种重复细胞的一种自然且有效的模型。此处，$\boldsymbol{z}^t \in \mathbb{R}^d$ 表示时间 t 的低维表示，用于刻画海洋环境条件与细胞群体之间未观测到的关系，而这种关系会引发变化。本节稍后将进一步讨论相关细节。

该框架包含三个建模组件，并使用神经网络来提高灵活性。具体而言，我们将式（1）中解码器的混合系数参数化为

$$\boldsymbol{\pi}(\boldsymbol{x}^t,\boldsymbol{z}^t) = \mathrm{Softmax}\big(\mathrm{MLP}_{\boldsymbol{\phi}_1}(\boldsymbol{x}^t,\boldsymbol{z}^t)\big) \in [0,1]^K,$$

从而确保 $\sum_{k=1}^K \boldsymbol{\pi}_k(\boldsymbol{x}^t,\ \boldsymbol{z}^t) = 1$。此外，高斯分量的均值建模为

$$\boldsymbol{\beta}(\boldsymbol{x}^t,\boldsymbol{z}^t) = \mathrm{MLP}_{\boldsymbol{\phi}_2}(\boldsymbol{x}^t,\boldsymbol{z}^t) \in \mathbb{R}^{K\times 3}$$

其中，第 k 行对应于 $\boldsymbol{\beta}^{(k)}(\boldsymbol{x}^t,\ \boldsymbol{z}^t) \in \mathbb{R}^3$。最后，对角协方差矩阵为

$$\mathrm{diag}\big(\boldsymbol{\Sigma}^{(k)}(\boldsymbol{x}^t,\boldsymbol{z}^t)\big) = \mathrm{Softplus}\big(\mathrm{MLP}_{\boldsymbol{\phi}_3}(\boldsymbol{x}^t,\boldsymbol{z}^t)\big) \in \mathbb{R}_+^d$$

对每个簇 $k \in [K]$ 均如此。全文中，$\boldsymbol{\phi} = \{\boldsymbol{\phi}_1,\boldsymbol{\phi}_2,\boldsymbol{\phi}_3\}$ 表示来自三个分支的神经网络参数集合。

请注意，我们感兴趣的分布是条件密度 $P(\boldsymbol{y}^t|\boldsymbol{x}^t)$，并将其参数化为

$$\begin{aligned} P(\boldsymbol{y}^t|\boldsymbol{x}^t) &= \int P(\boldsymbol{y}^t,\boldsymbol{z}^t|\boldsymbol{x}^t)d\boldsymbol{z}^t \\ &= \int P(\boldsymbol{y}^t|\boldsymbol{z}^t,\boldsymbol{x}^t)P(\boldsymbol{z}^t)d\boldsymbol{z}^t. \end{aligned}$$

在本研究中，我们令 $P(\boldsymbol{z}^t|\boldsymbol{x}^t) := P(\boldsymbol{z}^t)$ 不依赖协变量（Bowman 等人，2016；Lee 等人，2017；Bhattacharyya 等人，2018），从而确保潜在表示 $\boldsymbol{z}^t \in \mathbb{R}^d$ 从光学特征中抽取海洋信息 $\{\boldsymbol{y}_b^t\}_{b=1}^B \subset \mathbb{R}^3$，并以海洋协变量$\boldsymbol{x}^t \in \mathbb{R}^p$ 在时刻 t为条件。

随后，我们对潜变量施加一个可学习的高斯先验，即

$$\boldsymbol{z}^t \sim P(\boldsymbol{z}^t) = \mathcal{N}(\boldsymbol{z}^t; \boldsymbol{\mu}^t, \boldsymbol{I}_d)$$

其中 $\boldsymbol{\mu}^t \in \mathbb{R}^d$ 为均值向量， $\boldsymbol{I}_d$ 为单位矩阵。借助解码器 $P(\boldsymbol{y}^t|\boldsymbol{x}^t,\ \boldsymbol{z}^t)$，我们将 $\boldsymbol{z}^t \in\mathbb{R}^d$ 视为时刻 t处相应海洋状态的低维表示，而非细胞特异性因子或簇特异性因子。解码器与先验的这一组合使该框架能够检测由潜在先验均值偏移所刻画的突变，详见第3节。尽管许多现有的变点检测方法仅限于检测某一特定的预先参数化的变化类型，所提出的框架仍能通过其潜在表示捕捉广泛的异质性共现分布变化。

图1概述了所提出的框架。解码器 $P_{\boldsymbol{\phi}}(\boldsymbol{y}^t|\boldsymbol{z}^t,\ \boldsymbol{x}^t)$，其神经网络参数为 $\boldsymbol{\phi}$，并在各时间点之间共享。中间的阴影圆表示观测响应，底部的实心圆表示每个细胞共享的共同环境条件（作为协变量）。顶部的虚线圆表示潜变量，其中先验均值通过学习以促进变点检测。我们估计先验参数 $\{\boldsymbol{\mu}^t\}_{t=1}^T$，以便从光学测量 $\{\boldsymbol{y}^t\}_{t=1}^T$ 中识别变点，而施加于先验参数的组融合 LASSO 正则化在第3节中加以描述。值得注意的是，该框架十分简洁，无需编码器。

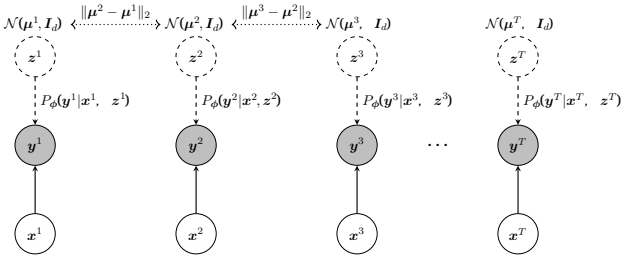


图*1.* 先验分布和潜在空间高斯混合专家解码器概览。

2.2. 变点

在本研究中，目标是检测海洋环境条件与浮游植物物种之间关系中的 changepoint。通过将信息抽象为服从高斯先验的潜在表示，我们可以指定

通过先验参数 $\boldsymbol{\mu}^t \in \mathbb{R}^d$ 为$t = 1,\ \ldots\ldots,\ T$ 指定变点。令 $\{D_r\}_{r=0}^{R+1} \subset \{1, 2, \ldots, T\}$ 表示满足以下条件的一组有序变点

$$1 = D_0 < D_1 < \cdots < D_R < D_{R+1} = T.$$

这些变点将时间跨度划分为 $R + 1$ 个时间段，使得每个时间段内的先验参数保持不变：

$$\boldsymbol{\mu}^{D_r} = \boldsymbol{\mu}^{D_r+1} = \cdots = \boldsymbol{\mu}^{D_{r+1}-1},\ r = 0, \ldots, R,$$

且连续时间段的先验参数存在差异，即均值偏移：

$$\boldsymbol{\mu}^{D_r} \neq \boldsymbol{\mu}^{D_{r+1}},\ \ r = 0, \ldots, R-1,$$

按照约定 $\boldsymbol{\mu}^{D_{R+1}} = \boldsymbol{\mu}^{D_R}$。多重变点检测问题旨在根据海洋学协变量和光学测量恢复变点集合 $\{D_r\}_{r=1}^R$，其中变点数量 R 及其位置均未知。记号上，我们将 $\boldsymbol{\mu} \in \mathbb{R}^{T\times d}$ 表示为一个矩阵，其中第 t行对应于 $\boldsymbol{\mu}^t \in \mathbb{R}^d$。

我们不再从数据空间中重复且经聚类的测量值检测变点，而是通过潜在空间中的均值偏移来定义变点。这一表述反映了这样一种前提：有意义的变化源于潜在海洋条件的扰动及其对物种组成的影响，而非个别观测中的瞬态波动。在潜在空间中操作，使所提出的方法能够汇总重复细胞间的信息，从而提高其对测量噪声、不同聚类间异质性以及混合组成变异的稳健性。

3. 学习与推断

3.1. 从观测序列学习先验

本节介绍一种用于学习变点检测先验参数的优化方法。根据（Levy-Leduc &Harchaoui，2007；Vert & Bleak-ley，2010；Lin etal.，2017）的框架，针对重复细胞群体的变点检测问题被表述为一个组融合 LASSO问题。将观测序列分布的对数似然记为

$$l(\boldsymbol{\phi}, \boldsymbol{\mu}) = \sum_{t=1}^{T}\sum_{b=1}^{B} \log \int P_{\boldsymbol{\phi}}(\boldsymbol{y}_b^t|\boldsymbol{z}^t, \boldsymbol{x}^t) P_{\boldsymbol{\mu}^t}(\boldsymbol{z}^t) d\boldsymbol{z}^t.$$

接下来，我们要解决以下约束优化问题：

$$\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\phi},\boldsymbol{\mu}} -l(\boldsymbol{\phi}, \boldsymbol{\mu}) + \lambda \sum_{t=1}^{T-1} \|\boldsymbol{\nu}^{t+1} - \boldsymbol{\nu}^t\|_2 \quad (2)$$
$$\text{subject to } \boldsymbol{\mu} = \boldsymbol{\nu}.$$

其中，$\lambda > 0$ 是组融合 LASSO 惩罚的调节参数，$\boldsymbol{\nu}^t \in \mathbb{R}^d$ 表示松弛变量 $\boldsymbol{\nu} \in \mathbb{R}^{T\times d}$的第 t行。由于正则化施加于与似然相关的先验均值上，学习得到的先验会将分布变化纳入潜在空间。

具体而言，组融合 LASSO 惩罚通过对多元总变差进行正则化，促使学习到的参数呈现分段常数结构。该惩罚使差异具有稀疏性，同时允许多个坐标同时发生变化。将其应用于先验参数的序列差分后，得到的潜在表示会突出有意义的转变，并抑制细微波动。

为促进参数的有效更新，(Vert 与 Bleakley，2010) 引入了两个变量 $(\boldsymbol{\alpha},\ \boldsymbol{\beta}) \in \mathbb{R}^{1\times d}\times\mathbb{R}^{(T-1)\times d}$用于分解松弛变量，其定义如下

$$\begin{aligned}\boldsymbol{\alpha} &= \boldsymbol{\nu}^1\\ \boldsymbol{\beta}_{t,\cdot} &= \boldsymbol{\nu}^{t+1} - \boldsymbol{\nu}^t \quad \forall\, t = 1,\dots,T-1.\end{aligned}$$

此外，$\boldsymbol{\nu} \in \mathbb{R}^{T\times d}$可以整理为$\boldsymbol{\nu} = \mathbf{1}_{T,\ 1}\boldsymbol{\alpha} + \boldsymbol{D}\boldsymbol{\beta}$其中 $\boldsymbol{D} \in \{0,1\}^{T\times(T-1)}$是一个设计矩阵，当 $i > j$ 时取值为 $\boldsymbol{D}_{ij} = 1$，否则为 0。从本质上说，将松弛变量分解为 $(\boldsymbol{\alpha},\ \boldsymbol{\beta})$，即可对 $\|\boldsymbol{\beta}_{t,\ \cdot}\|_2$进行直接收缩；而直接对 $\|\boldsymbol{\nu}^{t+1} - \boldsymbol{\nu}^t\|_2 +\|\boldsymbol{\nu}^t - \boldsymbol{\nu}^{t-1}\|_2$ 关于 $\boldsymbol{\nu}^t$ 求导，则会导致更新过程相互耦合且复杂。归根结底，松弛变量按 $(\boldsymbol{\alpha},\ \boldsymbol{\beta})$ 的形式更新，这种方式更简单，之后再整理到 $\boldsymbol{\nu} \in \mathbb{R}^{T\times d}$ 中以便后续更新。

根据（Kei 等人，2025），相应的增广拉格朗日函数可表示为

$$\begin{aligned}\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{u}) = -l(\boldsymbol{\phi},\boldsymbol{\mu}) + \lambda\sum_{t=1}^{T-1}\|\boldsymbol{\beta}_{t,\cdot}\|_2 +\\ \frac{\rho}{2}\|\boldsymbol{\mu} - \mathbf{1}_{T,1}\boldsymbol{\alpha} - \boldsymbol{D}\boldsymbol{\beta} + \boldsymbol{u}\|_F^2 - \frac{\rho}{2}\|\boldsymbol{u}\|_F^2\end{aligned} \tag{3}$$

其中 $\boldsymbol{u} \in \mathbb{R}^{T\times d}$ 是缩放对偶变量，而 $\rho \in \mathbb{R}$ 是增广项的另一个惩罚参数。因此，采用交替方向乘子法（Boyd 等人，2011；Zhu，2017；Wang 等人，2019）来求解该约束优化问题：

$$\begin{aligned}\boldsymbol{\phi}_{(a+1)},\boldsymbol{\mu}_{(a+1)} = \underset{\boldsymbol{\phi},\boldsymbol{\mu}}{\arg\min}\ -l(\boldsymbol{\phi},\boldsymbol{\mu}) +\\ \frac{\rho}{2}\|\boldsymbol{\mu} - \boldsymbol{\nu}_{(a)} + \boldsymbol{u}_{(a)}\|_F^2,\end{aligned} \tag{4}$$

$$\begin{aligned}\boldsymbol{\alpha}_{(a+1)},\boldsymbol{\beta}_{(a+1)} = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\arg\min}\ \lambda\sum_{t=1}^{T-1}\|\boldsymbol{\beta}_{t,\cdot}\|_2 +\\ \frac{\rho}{2}\|\boldsymbol{\mu}_{(a+1)} - \mathbf{1}_{T,1}\boldsymbol{\alpha} - \boldsymbol{D}\boldsymbol{\beta} + \boldsymbol{u}_{(a)}\|_F^2,\end{aligned} \tag{5}$$

$$\boldsymbol{u}_{(a+1)} = \boldsymbol{\mu}_{(a+1)} - \boldsymbol{\nu}_{(a+1)} + \boldsymbol{u}_{(a)}, \tag{6}$$

其中，下标 a 表示 ADMM 迭代。重复执行这三个更新，直至满足某个收敛准则。

具体而言，在交替方向乘子法（ADMM）过程中某次迭代时，$\boldsymbol{\mu}^t \in \mathbb{R}^d$ 的解为

$$\boldsymbol{\mu}^t = \frac{1}{1+\rho}\mathbb{E}_{P(\boldsymbol{z}^t|\boldsymbol{y}^t,\boldsymbol{x}^t)}(\boldsymbol{z}^t) + \frac{\rho}{1+\rho}(\boldsymbol{\nu}^t - \boldsymbol{u}^t) \tag{7}$$

此外，高斯混合解码器参数 $\boldsymbol{\phi}$ 的梯度为

$$\nabla_{\boldsymbol{\phi}}\,\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\mu}) = -\sum_{t=1}^{T}\mathbb{E}_{P(\boldsymbol{z}^t|\boldsymbol{y}^t,\boldsymbol{x}^t)}\Big(\nabla_{\boldsymbol{\phi}}\log P(\boldsymbol{y}^t|\boldsymbol{x}^t,\boldsymbol{z}^t)\Big). \tag{8}$$

式（7）中的解和式（8）中的梯度都包含后验密度 $P(\boldsymbol{z}^t|\boldsymbol{y}^t,\boldsymbol{x}^t)$下的条件期望。为从后验分布中采样并近似条件期望，采用朗之万动力学（Xie 等人，2017；Nijkamp 等人，2020；Pang 等人，2020）。具体而言，令下标 k 表示朗之万动力学的时间步，并令 δ 表示一个较小的步长。通过迭代以下方程，利用朗之万动力学从后验分布中抽取样本：

$$\begin{aligned}\boldsymbol{z}_{k+1}^t = \boldsymbol{z}_k^t + \delta\big[\nabla_{\boldsymbol{z}^t}\log P_{\boldsymbol{\phi}}(\boldsymbol{y}^t|\boldsymbol{x}^t,\boldsymbol{z}^t) -\\ (\boldsymbol{z}_k^t - \boldsymbol{\mu}^t)\big] + \sqrt{2\delta}\boldsymbol{\epsilon}\end{aligned} \tag{9}$$

其中 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\ \boldsymbol{I}_d)$。直观而言，通过沿着潜变量的方向在后验密度中移动，即可从后验分布中抽取样本。

此外，$\boldsymbol{\beta} \in \mathbb{R}^{(T-1)\times d}$ 采用分块坐标下降的方式进行更新（Bleakley 和 Vert，2011），方法是对每一行应用以下操作：

$$\boldsymbol{\beta}_{t,\cdot} = \frac{1}{\rho\boldsymbol{D}_{\cdot,t}^\top\boldsymbol{D}_{\cdot,t}}\Big(1 - \frac{\lambda}{\|\boldsymbol{s}_t\|_2}\Big)_+\boldsymbol{s}_t$$

其中 $(\cdot)_+ = \max(\cdot,\ 0)$，并且

$$\boldsymbol{s}_t = \rho\boldsymbol{D}_{\cdot,t}^\top(\boldsymbol{\mu}_{(a+1)} + \boldsymbol{u}_{(a)} - \mathbf{1}_{T,1}\boldsymbol{\alpha} - \boldsymbol{D}_{\cdot,-t}\boldsymbol{\beta}_{-t,\cdot}).$$

$\boldsymbol{\alpha} \in \mathbb{R}^{1\times d}$ 的解为

$$\boldsymbol{\alpha} = (1/T)\mathbf{1}_{1,T}\cdot(\boldsymbol{\mu}_{(a+1)} + \boldsymbol{u}_{(a)} - \boldsymbol{D}\boldsymbol{\beta})^\top.$$

分解 $(\boldsymbol{\alpha},\ \boldsymbol{\beta})$ 更新后，重新构造松弛变量 $\boldsymbol{\nu}$，并更新缩放对偶变量 $\boldsymbol{u}$，从而完成一次交替方向乘子法迭代。

4. 变点定位与模型选择

4.1. 变点定位

在本节中，我们描述一种根据估计的先验参数定位变点的实用方法。

将交替方向乘子法的总迭代次数记为 A，并将第 a次迭代得到的估计先验均值记为$\hat{\boldsymbol{\mu}}_{(a)}$。在训练过程中，我们存储集合 $\{\hat{\boldsymbol{\mu}}_{(a)}\}_{a=1}^{A}$，用以追踪估计先验均值的演变。在每次迭代 a时，我们计算一阶差分的欧几里得范数，即

$$\Delta\hat{\boldsymbol{\mu}}_{(a)}^{t} = \|\hat{\boldsymbol{\mu}}_{(a)}^{t+1} - \hat{\boldsymbol{\mu}}_{(a)}^{t}\|_2 \;\forall\, t \in [2, T].$$

为在每次迭代中评估这些差异的稀疏性程度，我们计算经验峰度:

$$\kappa_{(a)} = \frac{\frac{1}{T-1}\sum_{t=1}^{T-1}\left(\Delta\boldsymbol{\mu}_{(a)}^{t} - \text{mean}(\Delta\boldsymbol{\mu}_{(a)})\right)^4}{\left[\frac{1}{T-1}\sum_{t=1}^{T-1}\left(\Delta\boldsymbol{\mu}_{(a)}^{t} - \text{mean}(\Delta\boldsymbol{\mu}_{(a)})\right)^2\right]^2}.$$

训练后，我们选择经验峰度最大的迭代所对应的估计先验参数。直观而言，较高的峰度表明少数时间点出现显著变化，这与存在突变相一致。为利用所选的$\hat{\boldsymbol{\mu}}$定位变点，我们构建如下定义的数据驱动阈值

$$\mathcal{T}_{\text{thr}} \coloneqq \text{mean}(\Delta\hat{\boldsymbol{\mu}}) + \mathcal{Z}_{\alpha} \times \text{std}(\Delta\hat{\boldsymbol{\mu}}) \tag{10}$$

其中 $\mathcal{Z}_{\alpha}$ 是标准高斯分布的 α%分位数。因此，当$\Delta\hat{\boldsymbol{\mu}}^{D_r} > \mathcal{T}_{\text{thr}}$时，即识别出变点 D_r。

4.2. 模型选择

优化问题（2）涉及一个调节参数λ，其取值变化会导致检测到的变点集合不同。我们采用交叉验证来选择最优的 λ值。具体而言，将原始时间序列数据划分为训练集和测试集。训练数据包含奇数索引时间点上的观测值，测试数据包含偶数索引时间点上的观测值。对于固定的 λ 值，使用训练数据估计模型参数，并根据测试数据评估高斯混合模型的对数似然。使测试集负对数似然最低的惩罚项λ 被选中，随后使用完整数据集和所选 λ重新估计模型参数。

5. 基于真实协变量的模拟研究

在本节中，我们评估所提出的方法在一种由流式细胞术数据中的真实协变量驱动的合成三维时间序列上的性能。聚类观测由一个参数随时间变化的双成分高斯混合模型生成。具体而言，在每个时间点 t，我们生成 $N = 100$ 个独立观测$\boldsymbol{Y}_{t,\,i} \in \mathbb{R}^3$。令

$$\boldsymbol{X}_t = (1,\, p1_t,\, sss_t)^{\top} \in \mathbb{R}^3$$

表示一个包含两个真实协变量的向量，其中 $p1$表示阳光照射量，而 sss 表示海表盐度。我们选择这两个环境变量，是因为它们与浮游植物动态密切相关，并且代表了与海洋微生物相关的两项基本海洋特征——可用于光合作用的光照量每日波动，以及与许多海洋环境条件高度相关的距地球表面的距离（Hyun 等人，2023b）。时间跨度 $T = 296$被划分为三个时间段，分别为 $\mathcal{A}_1 = \{1, \ldots, 100\}$、$\mathcal{A}_2 = \{101, \ldots, 200\}$和 $\mathcal{A}_3 = \{201, \ldots, 296\}$。在此模拟中，真实协变量随时间点而异，并由所有细胞共享。

此外，高斯混合模型的混合系数定义为

$$\pi_{t,k} \propto \exp(\boldsymbol{X}_t^{\top}\boldsymbol{\alpha}_{t,k}), \qquad k = 1, 2,$$

并在各簇之间归一化为1。向量$\boldsymbol{\alpha}_{t,k} \in \mathbb{R}^3$ 是与真实协变量 $\boldsymbol{X}_t$相关联的预先指定系数，旨在引起混合比例随时间发生变化。具体而言，簇1的系数为$(1, 0, 0)^{\top}$（对应于 $t \in \mathcal{A}_1 \cup \mathcal{A}_3$）以及$(2, 0, 0)^{\top}$（对应于$t \in \mathcal{A}_2$）。簇2的系数为$(0, 0, 0.5)^{\top}$（对应于 $t \in \mathcal{A}_1 \cup \mathcal{A}_3$）以及$(0, 0, 0)^{\top}$（对应于 $t \in \mathcal{A}_2$）。

$\boldsymbol{Y}_{t,i}$ 中的每个坐标均具有时变簇特异性均值。令 $\mu_{t\,d}^{(k)}$，表示簇 $k \in \{1, 2\}$ 在时刻 t的维度 $d \in \{1, 2, 3\}$ 的均值。具体而言，对于 $d \in \{1, 2\}$，簇1的均值为

$$\mu_{t,d}^{(k)} = p1_t,$$

而簇2为

$$\mu_{t,d}^{(k)} = \begin{cases} 3 + 5\,p1_t, & t \in \mathcal{A}_1 \cup \mathcal{A}_3, \\ 3, & t \in \mathcal{A}_2. \end{cases}$$

对于 $d = 3$，均值设定为

$$\mu_{t,3}^{(1)} = \boldsymbol{X}_t^{\top}(0,\, 0,\, -0.5), \qquad \mu_{t,3}^{(2)} = \boldsymbol{X}_t^{\top}(6,\, 0,\, 0.5),$$

并且在各时间段之间保持不变。

最后，每个观测值都直接来自一个两成分高斯混合。在每个时间点 t，$\boldsymbol{Y}_{t,\,i} \in \mathbb{R}^3$ 的分布为

$$\pi_{t,1}\,\mathcal{N}\big(\mu_t^{(1)},\, \sigma^2\boldsymbol{I}_3\big) \;+\; \pi_{t,2}\,\mathcal{N}\big(\mu_t^{(2)},\, \sigma^2\boldsymbol{I}_3\big),$$

其中 $\sigma = 0.5$。这产生了一个具有协变量驱动的非平稳性的三维时间序列，其两个变点分别位于 $t = 100$和 $t = 200$。

全文中，我们的方法记为 CPD_{FC}。对于高斯混合专家解码器，我们将数量设为

个成分设为 $K = 2$，潜在维度设为 $d = 3$。解码器神经网络由三个隐藏层组成，每层包含 50 个神经元；前两层由三个分支共享，分别对应混合权重、成分均值和方差。调优参数列表 λ 为 $\{0.01, 0.05, 0.1, 1.0\}$。对于每个 λ，运行 150 次交替方向乘子法（ADMM）迭代。增广项的惩罚参数固定为 $\rho = 0.8$。在每次 ADMM 迭代中，使用 Adam 优化器训练解码器 20 次迭代，学习率为 0.01。松弛变量的块坐标下降运行 20次迭代。对于朗之万动力学，步长为 $\delta = 0.2$，并使用 100 次迭代抽取 100 个样本。训练完成后，使用显著性水平为 $\alpha = 0.99$的阈值 (10) 定位变点。

我们重复数据生成过程 50 次，每次均实施所提出的变点检测方法，并与几种现有方法进行比较：

1. **E-Divisive（James & Matteson，2015）**（`ecp`）：一种非参数、无分布假设的方法，利用能量距离检测多个变点。2. **非参数PELT（Killick &Eckley，2014）**（`changepoint.np`）：一种用于多重变点检测的惩罚似然方法，无需假设参数模型。3.**断点回归（Zeileis等人，2002）**（`strucchange`）：一种结构突变程序，通过动态规划估计回归断点。4. **贝叶斯变点分析（Erdman & Emerson，2008）**（`bcp`）：一种基于马尔可夫链蒙特卡洛的方法，通过均值偏移的后验概率识别变点。

由于目前没有现有方法能够检测重复数据中的变点，因此我们对参与比较的方法进行了调整。对于每个模拟序列，我们通过线性回归剔除协变量效应，并在残差上运行参与比较的方法（Wang，2008）。具体而言，令$\boldsymbol{Y}_{t,i} \in \mathbb{R}^3$ 表示时间t 、重复 i时的三维观测，令 $\boldsymbol{X}_t$ 表示时间t的协变量向量。我们将数据堆叠为

$$\boldsymbol{Y}_{\text{stack}} = \{\boldsymbol{Y}_{t,i}\} \in \mathbb{R}^{TN}, \qquad \boldsymbol{X}_{\text{stack}} = \{\boldsymbol{X}_t\} \in \mathbb{R}^{TN\times 3},$$

并拟合线性回归模型：

$$\boldsymbol{Y}^{(j)}_{\text{stack}} = \boldsymbol{X}_{\text{stack}}\boldsymbol{\beta}^{(j)} + \boldsymbol{\varepsilon}^{(j)}.$$

然后将拟合残差重塑为其原始时间索引，得到一个残差向量 $\boldsymbol{r} = \{r^{(j)}_{t,i}\} \in \mathbb{R}^{TN}$ for$j = 1$，2，3。

对于需要单变量时间序列的方法，我们通过平方范数汇总每个时间点的残差

$$r_t = \sum_{i=1}^{N}\sum_{j=1}^{3}\big(r^{(j)}_{t,i}\big)^2,$$

由此得到长度为 T 的单条时间序列 $\{r_t\}_{t=1}^T$，用于概括剔除协变量效应后残差变异的总量。

四种参与比较的方法按如下方式应用于这些派生特征。strucchange 和 bcp 方法使用单变量序列 $\{r_t\}_{t=1}^T \subset \mathbb{R}$，检测残差中的均值偏移。changepoint.np 中的 PELT 方法则使用非参数代价函数应用于 $\{r_t\}_{t=1}^T$。对于多变量方法 ecp，我们使用每个时间点的三维平均残差向量，

$$\mathbf{m}_t = \frac{1}{N}\sum_{i=1}^{N}\big(\text{resid}^{(1)}_{t,i}, \text{resid}^{(2)}_{t,i}, \text{resid}^{(3)}_{t,i}\big),$$

并将 E-Divisive 应用于序列 $\{\mathbf{m}_t\}_{t=1}^T$。

为评估变点检测方法的性能，我们使用以下指标来评估检测到的位置集合 $\hat{}\mathbf{c} = \{\widehat{c}_1, \ldots, \widehat{c}_K\}$ 与真实集合 $\mathbf{c}^\star = \{c^\star_1, \ldots, c^\star_{K^\star}\}$。

(1) 假阳性（FP）。在容差 τ 范围内与任何真实变点都不匹配的估计变点：

$$\text{FP} = \sum_k \mathbf{1}\left(\min_j |\widehat{c}_k - c^\star_j| > \tau\right).$$

(2) 假阴性（FN）。在 τ 范围内未被检测到的真实变点：

$$\text{FN} = \sum_j \mathbf{1}\left(\min_k |\widehat{c}_k - c^\star_j| > \tau\right).$$

(3) $D_{\text{t}\to\text{e}}$。每个真实变点到其最近估计变点的最坏情况距离：

$$D_{\text{t}\to\text{e}} = \max_j \min_k |\widehat{c}_k - c^\star_j|.$$

(4) $D_{\text{e}\to\text{t}}$。每个估计变点到其最近真实变点的最坏情况距离：

$$D_{\text{e}\to\text{t}} = \max_k \min_j |\widehat{c}_k - c^\star_j|.$$

(5) 计数误差（CE）。检测到的变点数量与真实变点数量之差：

$$\text{CE} = \big|N_{cp} - N^\star_{cp}\big|,$$

其中，N_{cp} 和 $N^\star_{cp}$ 分别表示真实和估计的变点数量。

(6) 覆盖分数 (CS)。一种基于真实时间段与估计时间段之间最大杰卡德重叠度的分割级评分：

$$C(\mathcal{G}, \mathcal{G}') = \frac{1}{T} \sum_{\mathcal{A} \in \mathcal{G}} |\mathcal{A}| \cdot \max_{\mathcal{A}' \in \mathcal{G}'} \frac{|\mathcal{A} \cap \mathcal{A}'|}{|\mathcal{A} \cup \mathcal{A}'|}$$

其中， $\mathcal{G}$ 和 $\mathcal{G}'$ 分别表示真实值和估计的分割结果。

表1 报告了 50 个模拟数据集在六项评估指标上的平均值。总体而言，与所有竞争方法相比，我们的方法具有更低的假阳性（FP）率、假阴性（FN）率和最坏情况距离 $D_{\mathrm{t}\to\mathrm{e}}$。此外，我们的方法获得了更高的覆盖分数（CS），表明其平均而言能够准确恢复真实的变点位置。就最坏情况距离$D_{\mathrm{e}\to\mathrm{t}}$ 和变点数量误差（CE）而言，我们的方法在所有方法中排名第二，这表明其检测到的变点略多于真实值。相比之下，断点回归和 E-divisive 方法往往低估变点数量，体现出一种保守分割策略：它虽能降低虚假检测的风险，却会导致较高的假阴性率。综上，这些结果表明，所提出的方法在各项评估指标上表现良好。

表 1. 不同方法的变点检测性能比较。最佳指标以粗体显示。

方法	FP↓	FN↓	$D_{\mathrm{t}\to\mathrm{e}}$ ↓	$D_{\mathrm{e}\to\mathrm{t}}$ ↓	CE↓	CS↑
$\mathrm{CPD}_{\mathrm{FC}}$	**0.60**	**0.04**	**2.98**	39.72	0.56	**0.94**
BCP	3.92	1.00	95.88	96.64	2.92	0.46
STRUCCHANGE	1.00	1.00	47.12	47.12	**0.00**	0.70
ECP	1.20	0.92	12.10	**18.32**	0.28	0.89
CHANGEPOINT.NP	1.04	0.04	8.80	58.52	1.00	0.80

6. 流式细胞术数据的应用

在本节中，我们将所提出的方法应用于一组真实的流式细胞术数据。图 2展示了研究航次的轨迹图，数据采集于北太平洋——图中以黄色绘出了上行和下行轨迹，并以较小的人工间距将二者分开。研究航次（Gradients 2）从夏威夷出发，先向北航行再向南返航，历时约两周，并穿越了两片不同的海域——南部海域称为“副热带环流”，北部海域称为“亚寒带环流”。我们的实验使用 Seaflow流式细胞仪（Swal- well 等人, 2011）收集了颗粒级数据。在这幅图中，我们用蓝色标出了采用所提出的方法检测到的变点，其纬度约为北纬 33.2 度。我们将在本节末尾进一步讨论这一令人振奋的结果。

我们感兴趣的数据集包含 296 个每小时时间点，时间跨度为 2017 年 6 月 1 日至 2017 年 6 月 14 日。在每个时间

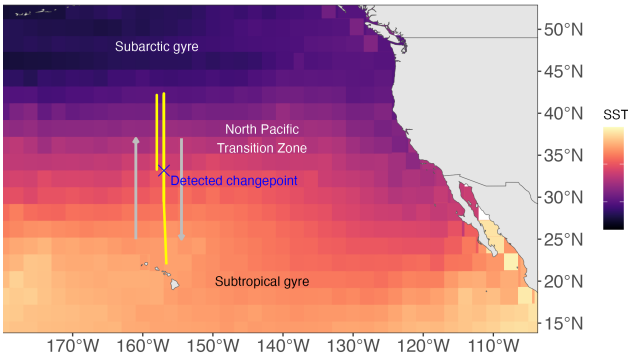


图2. 北太平洋科考航次轨迹图以黄色显示（上行轨迹与下行轨迹略微分开，以便观察）。该科考航次（Gradients 2）穿越了两片不同的海洋水体——南部的“副热带环流”和北部的“亚寒带环流”。将我们提出的方法应用于该数据集后，在约北纬33.2度处检测到一个变点，这与文献中发现的其他若干变点大致一致。

点，测得的浮游植物细胞数量各不相同。共有 39 个协变量，且同一时间点测得的细胞共享同一协变量向量。此外，每个数据点都有一个生物量测量值，用作反映其相对贡献的权重。

在原始数据集中，共有296个时间点，每个时间点包含整整一小时的重复微生物观测（细胞计量颗粒）。各时间点的观测数量约为1,986至20,181。由于数据集规模较大，我们根据生物量值在每个时间点进行加权重采样。具体而言，在每个时间点抽取 $B = 500$ 个观测。选择该样本量是为了在细胞多样性与计算效率之间取得平衡。

具体而言，我们将混合分量的数量设为 $K = 15$，潜在维度设为 $d = 5$。解码器神经网络由三个隐藏层组成，每层有50个神经元。前两个隐藏层由对应于分量权重、均值和方差的三个分支共享。ADMM的惩罚参数 λ通过交叉验证从 $\{0.01, 0.05, 0.1, 1.0\}$ 中选取。对于每个λ，ADMM过程运行200次迭代。使用Adam优化器以0.001的学习率训练解码器，迭代10次。增广项的惩罚参数固定为 $\rho = 0.8$，块坐标下降运行10次迭代。进行朗之万动力学采样时，步长为0.2，共迭代200次并抽取500个样本。变点定位基于估计均值的一阶差分，阈值判定的显著性水平为0.99。图3展示了估计的 $\Delta\boldsymbol{\mu}$ 随时间的变化。

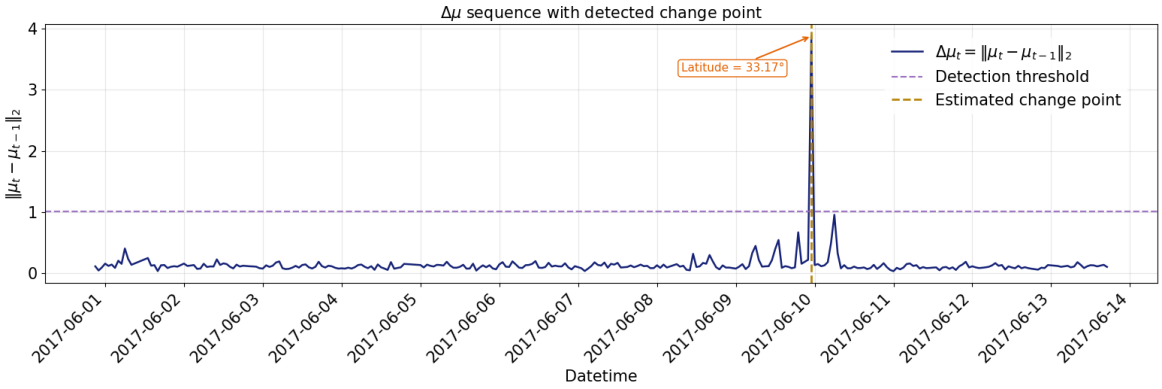


图 *3.* 流式细胞术数据中估计的 $\Delta\mu$ 示意图。检测到的变点以红色竖线表示。检测阈值以水平线表示。

我们的方法在北纬33.2度检测到一个changepoint，这与研究北太平洋断面区域的文献中发现的各种生物和物理变点相当接近。具体而言，该位置与两个生物学changepoint（约在北纬33.1度和北纬33.7度）相差不到1度；它们也由（Jones等人，2021年）（图6左图）检测到。该研究分析了另一次类似研究航次的流式细胞术数据；该航次于约早一年零一个月的2016年4月20日至5月2日出海。此外，著名的Longhurst框架（Longhurst等人，1995年；Longhurst，2007年）传统上用于研究海洋区系，并将北太平洋的两个生物地理省（NPPF，即北太平洋极锋区/带，以及NPTG，即北太平洋热带环流）之间的分界线划定在该航次垂直航迹沿线北纬34.4度左右。此外，（Follett等人，2022年）研究了一种重要的光合微生物物种（称为原绿球藻）的衰退；这种衰退发生在北纬30.4度左右，比我们检测到该点的位置偏南几度。最后，尽管没有明确检测变点，两项研究（Hyun等人，2023a；2025）也分别估计出原绿球藻丰度在同一纬度附近出现上升（一次渐进，一次较为突然），时间约在6月8日伊始，这与我们检测到的变点惊人地相似。

7. 讨论

本文提出一种潜在空间高斯混合专家模型，用于检测带有环境协变量的浮游植物重复观测数据中的变点。据我们所知，现有变点检测方法并非为处理重复且聚类的单细胞测量而设计，而这类测量的分布会随协变量而变化。通过将流式细胞术响应映射为低维表示，所提出的方法利用潜在空间中的均值偏移识别变点，从而捕捉海洋环境条件的变化。模拟研究表明，该方法表现良好，真实数据分析所得结果也与现有科学认知一致。

有几种值得探索的扩展方向。允许混合分量的数量随时间变化，可能会提高模型的灵活性，因为随着航次穿越不同区域，新的细胞群体可能出现，而现有群体可能减少。另一个值得关注的方向是通过采用其他惩罚策略，将模型扩展为检测底层混合分布中的渐变而非突变。最后，纳入结构更复杂或参数化的变点模型，例如线性或受约束的专家分量，可能提高模型的可解释性，并为科学发现提供一个有前景的方向。

影响声明

本文旨在推动含有协变量的重复和聚类细胞群体变点检测领域的发展。我们的研究可能带来多方面的环境影响，包括开发绘制海洋分区的新能力，以及加深对浮游植物种群动态的理解。

参考文献

Bai, J. 和 Perron, P. 含多个结构变化的线性模型的估计与检验。《计量经济学》，65(1)：47–78，1997。

Bhattacharyya, A.、Schiele, B. 和 Fritz, M. 基于“多个样本中择优”目标的序列准确且多样化采样。载于《IEEE计算机视觉与模式识别会议论文集》，第8485–8493页，2018年。

Bleakley, K. 和 Vert, J.-P. 用于多变点检测的组融合套索。arXiv 预印本 arXiv:1106.4199，2011年。

Bowman, S.、Vilnis, L.、Vinyals, O.、Dai, A.、Jozefowicz, R. 和 Bengio, S. 从连续空间生成句子。载于《第20届SIGNLL计算自然语言学习会议论文集》，第10–21页，2016年。

Boyd, S.、Parikh, N.、Chu, E.、Peleato, B.、Eckstein, J. 等。通过交替方向乘子法实现分布式优化与统计学习。《机器学习基础与趋势》，3(1):1–122，2011年。

Chen, H. 和 Zhang, N. 基于图的变点检测。《统计年刊》，43(1):139–176，2015。

Chu, L. 和 Chen, H. 多变量和非欧几里得数据的渐近分布无关变点检测。《统计年刊》，47(1):382–414，2019。

Erdman, C. 和 Emerson, J. W. bcp：用于执行变点问题贝叶斯分析的 R 软件包。《统计软件杂志》，23:1–13，2008。

Field, C. B.、Behrenfeld, M. J.、Randerson, J. T. 和 Falkowski, P. 生物圈的初级生产：整合陆地和海洋组成部分。《科学》，281 (5374):237–240，1998。

Follett, C. L.、Dutkiewicz, S.、Ribalet, F.、Zakem, E.、Caron, D.、Armbrust, E. V. 和 Follows, M. J. 与异养细菌的营养相互作用限制了原绿球藻的分布范围。《美国国家科学院院刊》，119(2)，2022年1月。ISSN 1091-6490。doi: 10.1073/pnas.2110993118。URLhttp://dx.doi.org/10.1073/pnas.2110993118。

Garreau, D. 和 Arlot, S. 基于核的一致变点检测。《电子统计学期刊》，12(2):4440 – 4486，2018年。

Hyun, S., Cape, M. R., Ribalet, F. 和 Bien, J. 使用稀疏回归混合模型，根据协变量对流式细胞术测得的细胞群体进行建模。《应用统计年刊》17(1):357，2023a。

Hyun, S., Rolf Cape, M., Ribalet, F. 和 Bien, J. 使用稀疏回归混合模型，根据协变量对流式细胞术测得的细胞群体进行建模。《应用统计年刊》17(1)，2023年3月。ISSN 1932-6157。doi: 10.1214/22-aoas1631。URLhttp://dx.doi.org/10.1214/22-aoas1631。

Hyun, S., Coleman, T., Ribalet, F. 和 Bien, J. 用于高频流式细胞术数据自动门控的趋势滤波专家混合模型，2025。URL https://arxiv.org/abs/2504.12287。

James, N. A. 和 Matteson, D. S. ecp：用于多元数据非参数多重变点分析的 R 软件包。《统计软件杂志》，62:1–25，2015年。

Jones, C.、Clayton, S.、Ribalet, F.、Armbrust, E. V. 和 Harchaoui, Z. 一种基于核的变化检测方法，用于绘制通过流式细胞术测量的浮游植物群落变化。《生态学与进化方法》，12 (9):1687–1698，2021年6月。ISSN 2041-210X。doi: 10.1111/2041-210x.13647。URLhttp://dx.doi.org/10.1111/2041-210X.13647。

Kei, Y. L., Li, J., Li, H., Chen, Y., and Madrid Padilla, O. H. 基于仅解码器潜在空间模型的动态图变点检测. 机器学习研究汇刊, 2025. ISSN 2835-8856.

Killick, R. and Eckley, I. A. changepoint：用于变点分析的R软件包. 《统计软件杂志》, 58 (1):1–19, 2014.

Killick, R., Fearnhead, P., and Eckley, I. A. 以线性计算成本实现变点的最优检测. 《美国统计协会杂志》, 107 (500):1590–1598, 2012.

Lee, N., Choi, W., Vernaza, P., Choy, C. B., Torr, P. H., and Chandraker, M. DESIRE：含交互智能体的动态场景远期预测. 见IEEE计算机视觉与模式识别会议论文集, pp. 336–345, 2017.

Levy-Leduc, C. 和 Harchaoui, Z. 用套索捕捉变点。《神经信息处理系统进展》，20，2007。

Lin, K.、Sharpnack, J. L.、Rinaldo, A. 和 Tibshirani, R. J. 融合套索的精确误差分析及其在近似变点筛查中的应用。《神经信息处理系统进展》，30，2017。

Longhurst, A.《海洋生态地理学》。学术出版社，2007年1月。doi: 10.1016/B978-0-12-455521-1. X5000-1。

Longhurst, A.、Sathyendranath, S.、Platt, T. 和 Caverhill, C. 根据卫星辐射计数据估算海洋全球初级生产力。《浮游生物研究期刊》，17(6):1245–1271，1995。

Matteson, D. S. 和 James, N. A. 多元数据多重变点分析的一种非参数方法。美国统计学会期刊，109(505):334–345，2014。

Nijkamp, E.、Hill, M.、Han, T.、Zhu, S.-C. 和 Wu, Y. N. 关于基于马尔可夫链蒙特卡洛的基于能量的模型最大似然学习的剖析。载于 AAAI人工智能会议论文集，第34卷，第5272–5280页，2020。

Olson, R. J.、Shalapyonok, A. 和 Sosik, H. M. 通过在线声学聚焦样品颗粒改造流式细胞成像仪，以实现高通量成像。湖沼学与海洋学：方法，15(10):867–874，2017。

Pang, B., Han, T., Nijkamp, E., Zhu, S.-C. 和 Wu, Y. N. 学习潜空间基于能量的先验模型。《神经信息处理系统进展》，33：21994–22008，2020。

Ribalet, F., Dutkiewicz, S., Monier, E. 和 Armbrust, E. V. 未来海洋变暖可能导致原绿球藻生物量和生产力大幅下降。《自然·微生物学》，10(10):2441–2453，2025年10月。doi: 10.1038/s41564-025-02106-4。
URLhttps://doi.org/10.1038/s41564-025-02106-4。

Song, H. 和 Chen, H. 针对重复观测数据的渐近分布无关变点检测。《Biometrika》，109(3):783–798，2022a。

Song, H. 和 Chen, H. 新型基于核的变点检测。arXiv 预印本 arXiv:2206.01853，2022b。

Sosik, H. M. 和 Olson, R. J. 使用流式成像细胞术对采样的浮游植物进行自动分类。湖沼学与海洋学：方法，5 (6):204–216，2007。

Swalwell, J. E.、Ribalet, F. 和 Armbrust, E. V. SeaFlow：一种新型在线流式细胞仪，可连续观测海洋中的浮游植物。湖沼学与海洋学：方法，9(10):466–477，2011年10月。ISSN 1541-5856。doi: 10.4319/lom.2011.9. 466。
URLhttp://dx.doi.org/10.4319/lom.2011.9.466。

Vert, J.-P. 和 Bleakley, K. 使用组LARS快速检测许多信号共有的多个变点。《神经信息处理系统进展》，23，2010。

Wang, X. L. 使用惩罚最大t或F检验检测气候数据序列中的均值偏移时考虑自相关。《应用气象与气候学期刊》，47(9):2423–2444，2008。

Wang, Y.、Yin, W. 和 Zeng, J. 非凸非光滑优化中 ADMM 的全局收敛性。《科学计算期刊》，78：29–63，2019。

Xie, J.、Zhu, S.-C. 和 Nian Wu, Y. 利用时空生成卷积网络合成动态模式。载于 IEEE 计算机视觉与模式识别会议论文集，第 7093–7101 页，2017。

Zeileis, A.、Leisch, F.、Hornik, K. 和 Kleiber, C. struc-change：用于检验线性回归模型中结构变化的R软件包。《统计软件杂志》，7：1–38，2002。

Zhu, Y. 一种应用于广义套索问题的增广ADMM算法。《计算与图形统计学期刊》，26(1)：195–204，2017。