Change-in-velocity detection for multidimensional data

Linh Do[1], Dat Do[2], Keisha J. Cook[3], Scott A. McKinley[1]

Department of Mathematics, Tulane University, New Orleans, LA 70118, USA[1];

Department of Statistics, University of Chicago, Chicago, IL 60637, USA[2];

School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC 29634, USA[3];

October 31, 2025

Abstract

In this work, we introduce CPLASS[1] (Continuous Piecewise-Linear Approximation via Stochastic Search), an algorithm for detecting changes in velocity within multidimensional data. The one-dimensional version of this problem is known as the change-in-slope problem (see [6, 14]). Unlike traditional changepoint detection methods that focus on changes in mean, detecting changes in velocity requires a specialized approach due to continuity constraints and parameter dependencies, which frustrate popular algorithms like binary segmentation and dynamic programming. To overcome these difficulties, we introduce a specialized penalty function to balance improvements in likelihood due to model complexity, and a Markov Chain Monte Carlo (MCMC)-based approach with tailored proposal mechanisms for efficient parameter exploration. Our method is particularly suited for analyzing intracellular transport data, where the multidimensional trajectories of microscale cargo are driven by teams of molecular motors that undergo complex biophysical transitions. To ensure biophysical realism in the results, we introduce a speed penalty that discourages overfitted of short noisy segments while maintaining consistency in the large-sample limit. Additionally, we introduce a summary statistic called the Cumulative Speed Allocation, which is robust with respect to idiosyncracies of changepoint detection while maintaining the ability to discriminate between biophysically distinct populations.

Keywords: Changepoint detection, velocity change, Markov Chain Monte Carlo (MCMC), stochastic search, empirical process theory, intracellular transport, multidimensional time series, trajectory analysis.

1 Introduction

Changepoint detection problems have been studied for over sixty years with applications in many fields, such as signal processing [29, 31, 40], speech processing [2, 28, 50], financial analysis [4, 17, 40], bio-informatics and genomics [21, 45], environmental science [34], and many others. One of the canonical challenges in the field is to detect changes in mean for time series consisting of independent observations. The movement of microparticles transported by molecular motors presents a distinct challenge. The motion of organelles and other intracellular cargo is marked by persistent thermal fluctuations, periods of directed motion along microtubule filaments, and paused, almost stationary, states [25, 26, 37]. While the increment process for microparticle locations can be thought of as a change-in-mean process, the fact that the cargo

[1]The code is available at the following GitHub repository: [GitHub].

positions are continuous in time presents a novel challenge. In fact, it is more fruitful to think in terms of changes in velocity, a multidimensional generalization of what is commonly called the change-in-slope model. For data of this kind, we model particle locations as a sequence of independent Gaussian fluctuations about a phenomenological anchor whose trajectory is continuous and piecewise-linear. The goal of the changepoint analysis is to identify the most likely trajectory if such an anchor existed. We propose the CPLASS algorithm (Continuous Piecewise Linear Approximation via Stochastic Search) for this problem. In the following, we only focus on a retrospective changepoint framework closely related to data collected in the study of intracellular transport, but the method applies more generally. For recent reviews of changepoint methods, we refer to Aminikhanghahi and Cook (2017) [1], Truong, Oudre, and Vayatis (2020) [53], Fearnhead and Rigaill (2020) [16], and Shi, Gallagher, Lund, and Killick (2022) [51].

Related work. Most changepoint detection schemes involve proposing a vectors of changepoint times and then evaluating a likelihood value that arises from a stated probability model. However, more complex models can produce higher likelihood values and so a penalty must be introduced based on the number of changepoints in the proposal. There are several choices for the penalty, for instance, an ℓ_0 penalty used by the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) [57, 58]; ℓ_1 penalty such as Lasso, fused Lasso, elastic net, the group Lasso, the mono Lasso [27, 56]; more complex penalties such as the modified BIC criterion (mBIC) [60] which maximizes the asymptotic posterior probability of the data. Regarding finding optimal solutions for detecting changepoint problems with a given criterion function, some popular searching methods include binary segmentation [49] and its variants - Circular Binary Segmentation [46, 55], Wild Binary Segmentation [18], bottom-up [10], window sliding [9], Pruned Exact Linear Time (PELT) [35], dynamic programming [30] and stochastic optimization [24, 38, 39]. Another well-known procedure for detecting changepoints is based on a likelihood ratio test. In this procedure, a hypothesis test is first constructed for a single changepoint model, and then the test is applied to find multiple changepoints via search algorithms [3–5].

In Bayesian methods, a random sampling approach is taken to find suitable parameter sets. After proposing a sequence of location vectors and associated means, a likelihood value is evaluated at each step, and the proposed parameter set is accepted or rejected via the Metropolis-Hastings algorithm. The penalty is encoded in prior information based on subject-matter knowledge of experts. To our knowledge, the first Bayesian method for detecting an unknown number of changepoints was developed by Barry and Hartigan [7, 8]. They constructed a model called the Product Partition Model (PPM), which performs well in detecting sharp, short-lived changes in the means of independent normally distributed observations. This method has been implemented as an `R` package `bcp` by Erdman et al. [13].

While the problem of changes in mean has been studied extensively, few methods are available for studying changes in velocity (or changes in slope in the one-dimensional case). These problems are fundamentally more challenging [14]. For example, the most common generic approach to detecting multiple changepoints – binary segmentation – does not effectively identify changes in velocity. This method iteratively applies a method for detecting a single changepoint. In a velocity change problem, initial estimates for changepoint locations might lie between the true changepoints, making it difficult for binary segmentation to correct these errors. In Figure 1, we illustrate an example of this issue. This problem has also been reported by Baranowski, Chen, and Fryzlewicz (2019) [6] and Fearnhead, Maidstone, and Letchford (2019) [15]. An alternative method involves a dynamic programming algorithm that minimizes

an ℓ_0 penalized cost like Optimal Partitioning [30] or PELT [35]. However, this approach is also unsuitable for addressing changes in velocity due to model dependencies that arise from the continuity of the location process at each changepoint [15]. One can come up with an idea to take first differences in the data; from here, a change-in-velocity is transformed into a change-in-mean, and then one of the methods for detecting changes in the mean can be applied. However, Fearnhead and Grose [14] pointed out that this can perform poorly due to removing information in the data under the transformation process. To overcome these difficulties, there are some methods for detecting change-in-slope problems: Trend-filtering (2014) [36], which minimizes the RSS plus an ℓ_1 penalty on changes in slope; Narrowest-Over-Threshold (NOT) (2019) [6], which repeated a test for a single change in slope on subset data and used the narrowest-over-threshold to combine the results; CPOP (2019) [15] which based on a variant of dynamic programming to minimize the RSS plus an ℓ_0 penalty, i.e. a constant penalty for adding each change; the Narrowest Significance Pursuit (NSP) [19] method, which, unlike others that identify the location of changepoints, offers the shortest intervals guaranteed to contain at least one changepoint at a specified confidence level. All current versions of these methods deal with one-dimensional data, while our challenge comes from two-dimensional particle trajectories.

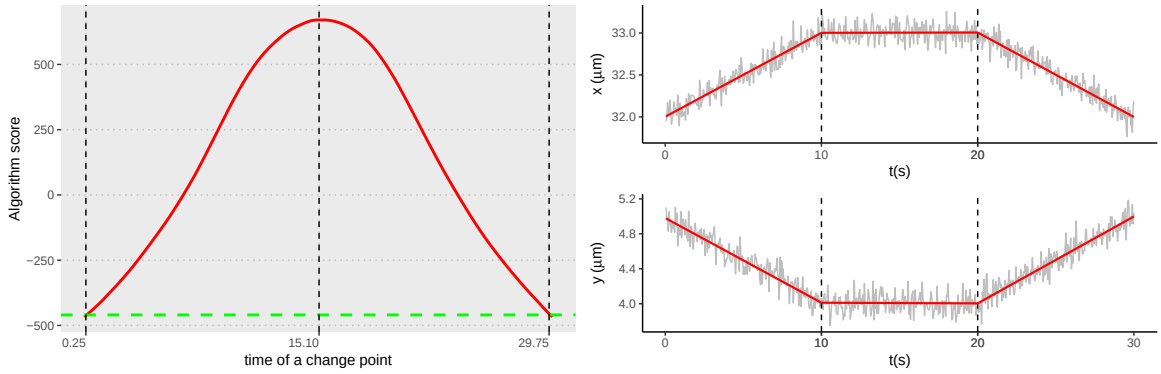
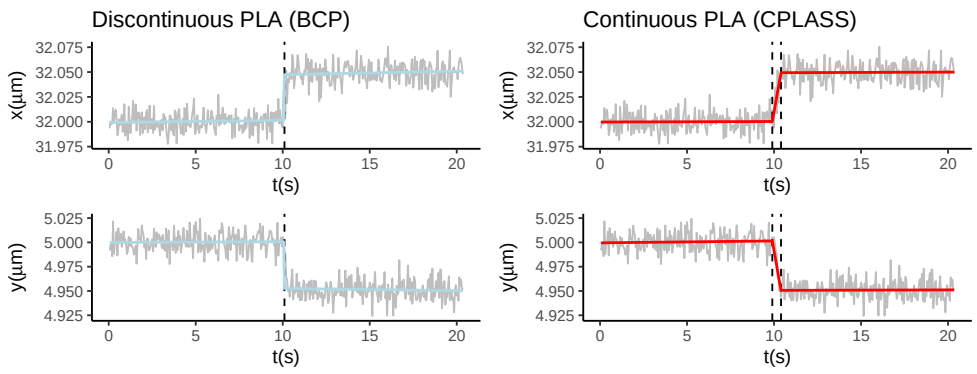


Figure 1: *Binary segmentation failure in a change-in-velocity setting.* A 2D trajectory was simulated at 20Hz (the position was observed every $1/20 = 0.05$ seconds) for 30 s with Gaussian noise ($\sigma = 0.1$). Two true changepoints occur at 10 s and 20 s, generated with velocity vectors $v_x = (0.1, 0, -0.1), \mu$m/s and $v_y = (-0.1, 0, 0.1), \mu$m/s. **Left panel**: CPLASS criterion values, with the null model (green dashed) and one-changepoint models (red solid). Higher values indicate a better fit. The figure illustrates how Binary segmentation incorrectly introduces an extra changepoint between the true ones. **Right panel**: The data on which the criterion function was calculated. Time series for x and y positions (gray), true changepoint times (dashed), and segmentation fit (red).

Particle tracking. The study of intracellular transport commoly involves analyzing individual particle trajectories that might reveal underlying biophysical states. Traditional summary statistics, such as mean squared displacement (MSD) analysis, are widely used to quantify transport dynamics [42, 43]. However, MSD-based approaches can fail to characterize short-lived state transitions, since the method involves averaging over entire trajectories. Therefore, some more recent methods have been developed to identify changes of state within individual particle trajectories. Some focus on capturing abrupt changes in diffusivity [33, 47], while others



Segment	1	2
Duration (s)	10.0500	10.2500
Speed (μm/s)	0.0002	0.0004

Segment	1	2	3
Duration (s)	9.8500	0.5000	9.9500
Speed (μm/s)	0.0002	0.1413	0.0001

Figure 2: *Comparison between discontinuous and continuous piecewise linear approximations.* The simulated lysosomal movement trajectory in 2D at 20Hz for a duration of 20.3 seconds and the two actual changepoints, 9 seconds and 10.35 seconds, are represented by the *t-vs-x* and *t-vs-y* time series. The dashed lines represent the detected changepoints. The corresponding segmentation is overlaid in red. The red column in the CPLASS output table indicates a significant active segment that BCP missed.

are focused on velocity shifts [32, 44], or simultaneous changes in velocity and diffusivity [59].

Although most changepoint algorithms developed for the segmentation of biophysical data were meant to detect changes in the mean value or the variance in time series, our algorithm deals with the detection of changepoints in a multiple linear regression model [12, 52, 59]. In previous studies, our group [32, 48] used the Bayesian Changepoint (`bcp`) algorithm to partition the paths into segments and then modeled the segmented paths as discontinuous piecewise-linear plus stationary noise. Here, we construct a *continuous* piecewise-linear function, design a criterion function with a combination of the strengthened Schwarz Information Criterion (sSIC) [6, 20] and a customized speed penalty, and use a stochastic searching method to find a suitable approximation for the data. The importance of using a continuous version comes from the issue of missing short-fast segments in the discontinuous models. Figure 2 shows a comparison between the discontinuous piecewise linear model and the continuous version of it. We can see a significant motile segment in the path; however, the discontinuous method missed it, while our CPLASS successfully returned the motile segment.

In this work we articulate the CPLASS algorithm for two-dimensional change-in-velocity detection, providing motivation for each of its components in Section 2; demonstrating its effectiveness on *in vivo* particle tracking data and simulations thereof in Section 3; and proving statistical consistency of the algorithm in the high-frequency limit in Appendix C.

Data sets. We use the following data sets to validate the proposed method: (1) BS-C-1 monkey kidney epithelial cells and A549 human lung epithelial cells [48] (obtained from the Duke University Cell Culture Facility). For these data sets, intracellular transport is a requirement of cellular functions related to lysosomes; (2) Data sets featuring quantum dots being transported by a single kinesin-1 motor, a single dynein-dynactin-BicD2 (DDB) motor, and a kin1-DDB pair [32] with intracellular transport by microtubule-based molecular motors;

(3) Additional simulated data sets imitating the trajectories in live cell data [11].

2 The CPLASS algorithm

In this section, we introduce the statistical model and algorithm used in this work. In Section 2.1, we introduce the statistical model. In Section 2.2, we construct a continuous piecewise-linear approximation of the data given the assumption that the number of changepoints is known. In Section 2.3 and Section 2.4, we propose a changepoint detection method where a criterion value with a penalty is provided, and a stochastic search method is used to find the maximum of the defined criterion function.

2.1 Statistical model

We assume the d-dimensional data observations $\{Y_i\}_{i=1}^n \subset \mathbb{R}^d$ at time $\mathcal{T} := \{t_1, \ldots, t_n\} \subseteq [0, T]$ are independent and identically distributed (iid) Gaussian fluctuations around a sequence of unobserved anchor locations, which are denoted $\{a_i\}_{i=1}^n \subset \mathbb{R}^d$. We write

$$Y_i = a_i + \sigma\varepsilon_i, \tag{1}$$

where $\{\varepsilon_i\}_{i=1}^n$ is a sequence of iid d-dimensional standard normal random variables with noise magnitude σ. We employ the convention that $t_0 = 0$ and $a_0 = \underline{a} \in \mathbb{R}^d$. Assume that there are k segments ($k + 1 \leq n$) and τ_j ($1 \leq j \leq k - 1$) is the changepoint time associated with the j change of velocity. For a natural number k, let $[k]$ denote the set $\{1, \ldots, k\}$. We further assume that observations are made on a uniform grid of size $\Delta = t_i - t_{i-1} = T/n$ for $i \in [n]$, and the jth changepoint can be approximated by an observation time, $\tau_j = t_{M_j}$, where $M_j := \lfloor \tau_j/\Delta \rfloor > 0$, for all $j \in [k-1]$, with the convention that $\tau_0 = 0$, $M_0 = 0$, $\tau_k = n\Delta = T$ and $M_k = n$.

Let $V_j \in \mathbb{R}^d$ denote the velocity vector of the jth segment, with speed defined as $s_j = \|V_j\|_2$ (for $j \in [k]$), where $\|\cdot\|_2$ is the Euclidean norm. Within each segment, we have

$$a_i = a_{M_{j-1}} + V_j(t_i - \tau_{j-1}), \tag{2}$$

where $i = M_{j-1} + 1, ..., M_j$, $j \in [k]$ is the index of the segment. From the recursive formula (2), we construct a multivariate continuous piecewise linear function (signal function) $f_{\boldsymbol{\tau},\mathbf{V},\underline{a}} : \mathcal{T} \to \mathbb{R}^d$, parametrized by changepoint $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{k-1})$ with $0 =: \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_{k-1} < \tau_k := T$, sets of velocities $\mathbf{V} = (V_0, \ldots, V_{k-1}) \subset \mathbb{R}^d$, and initial intercept (initial anchor position) $\underline{a} \in \mathbb{R}^d$, is defined as

$$\text{Signal function:} \quad f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}(t) = \left(\underline{a} - \sum_{j=1}^{i}(V_j - V_{j-1})\tau_{j-1}\right) + V_i t, \quad \forall t \in [\tau_{i-1}, \tau_i], i \in [k], \tag{3}$$

When $V_j \neq V_{j-1}$ for all $j \in [k]$, the signal function $f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}$ is said to have k segments and $(k-1)$ changepoints. Let $\mathcal{F}_k$ denote the collection of signal functions with k segments.

We assume that n multivariate observation $(Y_i)_{i=1}^n \subset \mathbb{R}^d$ on $\mathcal{T} = \{t_1, \ldots, t_n\}$ is generated according to a true signal function and Gaussian noises:

$$Y_i \overset{ind.}{\sim} \mathcal{N}(f^0(t_i), \sigma_0^2 I_d), \tag{4}$$

where $f^0(t) := f_{\boldsymbol{\tau}^0,\mathbf{V}^0,\underline{a}^0}$ is the true signal function of k_0 segments with true changepoints $\boldsymbol{\tau}^0 = (\tau_1^0, \ldots, \tau_{k_0-1}^0)$, sets of velocities $\mathbf{V}^0 = (V_0^0, \ldots, V_{k_0-1}^0) \subset \mathbb{R}^d$, and initial intercept

$\underline{a}^0 \in \mathbb{R}^d$. σ_0^2 is the true variance and I_d is the $d-$dimensional identity matrix. Given the set of observations $(Y_i)_{i=1}^n$, our goal is to infer the true number of segments k_0, parameters $\boldsymbol{\tau}^0, \mathbf{V}^0, \underline{a}^0$ of the true signal function and the noise level σ_0^2 .

Given a pre-specified upper bound $\overline{k}$ of the number of segments, we aim to estimate the parameters by maximizing the penalized likelihood:

$$(\widehat{f}_n, \widehat{\sigma}_n^2, \widehat{k}_n) = \underset{f \in \mathcal{F}_k, \sigma^2 \in \Omega, k \leq \overline{k}}{\arg\max} \sum_{i=1}^{n} \log \mathcal{N}(y_i | f_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}}(t_i), \sigma^2 I_d) - \text{pen}_k, \tag{5}$$

where pen_k is a penalty term to be defined later. We note that (5) is equivalent to finding the MLE for each $k \in [\overline{k}]$

$$(\widehat{f}_n^{(k)}, \widehat{\sigma}_{n,k}^2) = \underset{f \in \mathcal{F}_k, \sigma^2 \in \Omega}{\arg\max} \sum_{i=1}^{n} \log \mathcal{N}(y_i | f_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}}(t_i), \sigma^2 I_d), \tag{6}$$

and then identifying

$$\widehat{k}_n = \underset{k \in [\overline{k}]}{\arg\max} \sum_{i=1}^{n} \log \mathcal{N}(y_i | \widehat{f}_n^{(k)}(t_i), \widehat{\sigma}_{n,k}^2 I_d) - \text{pen}_k. \tag{7}$$

2.2 Continuous piecewise linear approximation given the changepoints

As noted above, finding the MLE of (5) can be separated into two steps. In this section, we focus on solving (6), which is given fixed $k-1$ changepoints, what is the MLE for parameters under considering the continuous piecewise linear model.

Consider the following matrices: (1) $\mathbb{Y} = (y_{il})$ is a $n \times d$ matrix represents the observed data; (2) $\mathbb{V} = (v_{jl})$ is a $k \times d$ matrix containing all segment velocities; (3) $\mathbb{W} = (w_{jl})$ is a $k \times d$ matrix represents the k differences between two consecutive velocities (i.e., $\mathbb{W}[1, .] = V_1, \mathbb{W}[j, .] = V_j - V_{j-1}$ for $j = 2, \ldots, k$), for $i \in [n]$, $l \in [d]$, and $j \in [k]$. We aim to find the MLEs of w_{jl}, $\underline{a} = (\underline{a}_1, \ldots, \underline{a}_d)$ and σ.

Since d dimensions are considered independent, given the changepoints, we can find the MLEs of the model independently in each dimension. With this in mind, let $Y^{(l)} = (y_{1l}, y_{2l}, \ldots, y_{nl}) \in \mathbb{R}^n$ $(l = [d])$ be the l-th column of matrix $\mathbb{Y}$, $\underline{W}^{(l)} = (\underline{a}_l, w_{1l}, \ldots, w_{kl})$ be the vector contains the l-th initial intercept and l-th column of matrix $\mathbb{W}$. We can then introduce the following matrix form associated with the l-th dimension

$$Y^{(l)} = \mathbb{T}\underline{W}^{(l)} + \sigma\varepsilon^{(l)}, \tag{8}$$

where

$$\mathbb{T} = \begin{bmatrix} 1 & t_1 & (t_1 - \tau_1)\mathbb{1}_{t_1 > \tau_1} & \cdots & (t_1 - \tau_{k-1})\mathbb{1}_{t_1 > \tau_{k-1}} \\ 1 & t_2 & (t_2 - \tau_1)\mathbb{1}_{t_2 > \tau_1} & \cdots & (t_2 - \tau_{k-1})\mathbb{1}_{t_2 > \tau_{k-1}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_n & (t_n - \tau_1)\mathbb{1}_{t_n > \tau_1} & \cdots & (t_n - \tau_{k-1})\mathbb{1}_{t_n > \tau_{k-1}} \end{bmatrix} \text{ is a } n \times (k+1) \text{ matrix}, \tag{9}$$

$\mathbb{1}_{t_i>\tau_j} = \begin{cases} 0 \text{ if } t_i \le \tau_j \\ 1 \text{ if } t_i > \tau_j, \end{cases}$, and $\varepsilon^{(l)} \sim \mathcal{N}(0, I_n)$ is an $n \times 1$ error vector.
The residual sum-of-squares (RSS) is written

$$\text{RSS}(Y, t; \underline{W}) := \sum_{l=1}^{d} \left\| Y^{(l)} - \mathbb{T}\underline{W}^{(l)} \right\|_2^2. \tag{10}$$

The log-likelihood associated with the model is

$$\mathcal{L}(f, \sigma) = \sum_{i=1}^{n} \log \mathcal{N}(y_i | f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}(t_i), \sigma^2 I_d) = -\frac{nd}{2}\left[\log(2\pi) + \log(\sigma^2)\right] - \frac{1}{2\sigma^2}\text{RSS}(Y, t; \underline{W}) \tag{11}$$

The resulting MLE are

$$\widehat{\underline{W}}^{(l)}_{n,k} = \left(\mathbb{T}^\top \mathbb{T}\right)^{-1} \mathbb{T}^\top Y^{(l)} \quad \text{for } l \in [k], \tag{12}$$

$$\hat{\sigma}^2_{n,k} = \frac{\text{RSS}(Y, t; \widehat{\underline{W}}_{n,k})}{dn}. \tag{13}$$

Note that as long as the changepoints τ_j are distinct and $n \ge k+1$, then $\mathbb{T}$ is full rank $k+1$, and we can make sure that $\mathbb{T}^\top\mathbb{T}$ has a unique inverse. From now on, we drop the subscript (n, k) in the MLE representations for ease of notation. The velocity vector corresponding to the jth segment is $\widehat{V}_j = \left(\sum_{i=1}^{j} \widehat{w}_{i1}, \ldots, \sum_{i=1}^{j} \widehat{w}_{id}\right)^\top$, for $j \in [k]$. The speed of the associated segment is $\hat{s}_j := \|\widehat{V}_j\|_2$.

2.3 Criterion function

The typical representation of the model contains changepoint times $\boldsymbol{\tau}$, initial intercepts (or initial anchor locations) $\underline{a}$, segment velocities $\mathbf{V}$, and noise σ. In the following, we introduce a reduced representation of the model that uses changepoint vector $r = (r_1, ..., r_{n-1})$ of ones and zeros, in the spirit of Laveille [38]. When $r_i = 1$, this indicates that a changepoint has occured during the ith time step. Henceforth we use the notation $|r|$ to denote the number of changepoints (hence in the notation introduced in the last section, $|r| = k-1$. For any given a particular changepoint vector r, we can find an associated piecewise linear approximation whose $\text{RSS}(Y, t; \widehat{\underline{a}}, \widehat{W}), \tau_j, \hat{s}_j$ ($j \in [|r|+1]$) are determined as discussed in Section 2.2. to this end, we will use the subscript r to indicate these relationships, i.e., $\widehat{\text{RSS}}_r = \text{RSS}(Y, t; \widehat{\underline{a}}_r, \widehat{W}_r), \tau_j, \hat{s}_j, \widehat{\underline{a}}_r, \tau_{j,r}, \hat{s}_{j,r}, \widehat{V}_{j,r}$. Altogether, for a given r, according to the derivation in Section 2.2, the piecewise linear model yields a maximized value of the log-likelihood function when it has the form

$$\widehat{\mathcal{L}}_n = \log(L(Y, t; \widehat{\underline{a}}_r, \widehat{W}_r, \hat{\sigma}_r)) = \log\left(\frac{1}{(2\pi)^{nd/2}}\left(\frac{dn}{\widehat{\text{RSS}}_r}\right)^{nd/2} \exp\left(-\frac{nd}{2}\right)\right) \tag{14}$$

$$= -\frac{nd}{2}\log\left(\widehat{\text{RSS}}_r\right) + C, \tag{15}$$

where C is a constant. We define the criterion function of the algorithm as follows.

Definition 2.1: Criterion Function

$$\Phi(r) = -\frac{nd}{2}\log\left(\widehat{\mathrm{RSS}}_r\right) - \mathrm{pen}(r). \tag{16}$$

Here, pen(r) refers to the penalty term designed to prevent overfitting. We used a strengthened Schwarz Information Criterion (sSIC) penalty expressed as $\log(n)^\gamma\rho$ where ρ is the number of parameter in the model and require $\gamma > 1$ (refer to [6, 18]), and a speed-control penalty function to mitigate the occurrence of unrealistic speed values.

Definition 2.2: Penalty function

$$\mathrm{pen}(r) = \log(n)^\gamma\rho + \sum_{j=1}^{|r|+1} h(\hat{s}_{j,r} - s_{cap}), \tag{17}$$

where $\gamma > 1$, $\rho = d(|r|+2)+1 = d(k+1)+1$ is the total number of parameters of the model, $|r|$ is the number of changepoints, $k = |r|+1$ is the number of segments $\hat{s}_{j,r} = \left\|\widehat{V}_j\right\|_2$ $(j = 1, ..., |r|+1)$ is the estimated segment speed, s_{cap}, decided by the practitioner, is the maximum speed that has no penalty, and $h(s) = \max\{0, s\}$.

The speed control function is added based on the prior knowledge of the scientist on the range of realistic particle speeds. If there is no information about the speed limit, one can set $h(s) = 0$. The penalty term, then, is the linear ℓ_0 penalty in the form of a strengthened Schwarz information criterion (sSIC) [6, 18]. Remark that $\gamma = 1$ corresponds to the standard SIC penalty considered by Yao [58] in the context of multiple changepoint detection. Under our construction, we require $\gamma > 1$ to provide the consistency theorem support (see Section 3.1). This requirement has also been used and discussed by Fryzlewicz in proposing the wild binary segmentation (WBS) [18] for the change-in-mean problem and in the Narrowest-over-Threshold (NOT) algorithm for the change-in-mean and change-in-slope problem [6]. Based on empirical experiments, we suggest choosing $\gamma = 1.01$ (see Section 3.2.1 and Figure 5). This choice guarantees that the sSIC criterion remains valid for the broadest possible class of signals, as aligns with the argument in [18]. Unless otherwise stated, the results reported hereafter correspond to sSIC with $\gamma = 1.01$, a setting that yields outcomes highly comparable to those from SIC.

2.4 Stochastic search of the changepoint space

When it comes to the search methods, the issue with using popular methods such as binary segmentation, PELT, or optimal partitioning has been discussed in the introduction. Another popular approach, called gradient ascent, has been used in optimization problems. However, this is not an ideal strategy because our criterion function commonly has multiple local maxima (see Figure 3 for an example). We chose a stochastic search approach, using a Metropolis-Hastings algorithm as the search algorithm.

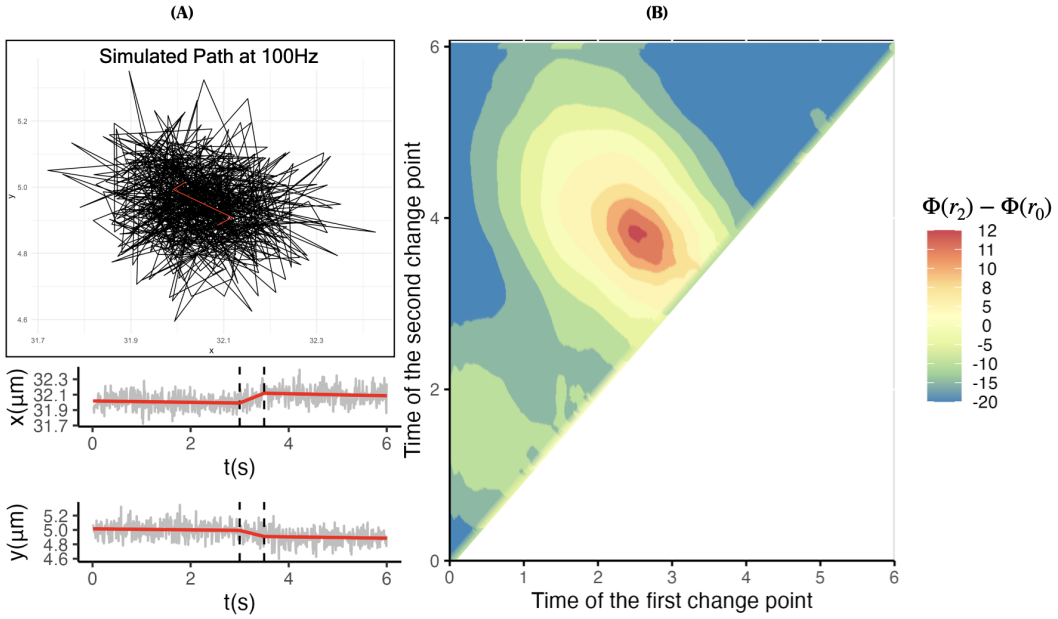


Figure 3: *Example showing why gradient ascent is unsuitable for the change-in-speed problem.* **Panel (A)** shows a simulated 2D lysosomal trajectory at 100 Hz for 6 s, containing two true changepoints at 3 s and 3.5 s. The corresponding segment speeds are $(0, 0.2, 0)$ μm/s. The noisy t–x and t–y time series are plotted in gray, with the true changepoints marked by dashed lines and the true segmentation overlaid in red. **Panel (B)** shows a contour plot of CPLASS criterion value differences between two-changepoint models, $\Phi(r_2)$, and the no-changepoint model, $\Phi(r_0)$. The global maximum corresponds to the true changepoints, but a gradient ascent search may get trapped in local maxima and fail to recover them.

2.4.1 Metropolis-Hastings algorithm as the searching algorithm

Notice that finding the maximum of the $\Phi(r)$ function is equivalent to finding the maximum of the $\exp(\Phi(r))$. Using Metropolis-Hastings, we can generate an ergodic Markov chain $\{r^{(t)}\}_{t\geq 0}$ that has $C\exp(\Phi(r))$ as its stationary distribution. The maximum of the $\exp(\Phi(r))$ function can then be approximated by the maximum of the sequence $\{\exp(\Phi(r^{(t)}))\}$. It is important to keep in mind that if the proposal function is irreducible, then the Markov chain attained after running an MH algorithm is both irreducible and aperiodic. Moreover, since the chain $\{r^{(t)}\}_{t\geq 0}$ takes its values in a finite space, it is uniformly ergodic ([39]). In Appendix C, we show that the detailed balance condition holds for our proposed rules.

To be specific, in our algorithm, the proposal function for the change point process takes its values on the set $\{0, 1\}^{n-1}$, where n is the number of observations. There are four types of changepoint vector proposals: (1) an independent changepoint vector; (2) the creation or extinction of a changepoint; (3) the creation or extinction of a segment; or (4) a location shift of a single changepoint. Let r^{prop} and r^{cur} denote the proposed and current changepoint process, respectively. The following summary describes each type of proposal.

- **Type 1.** Notation q_{new}. It allows for escaping local maxima and for the number of changepoints to vary by proposing an independent change point vector, r^{prop} following the distribution of Bernoulli random variables with the probability of a changepoints is $1 - \exp(-\lambda\Delta)$, where Δ is the time between observations: $r_i^{\text{prop}} \overset{iid}{\sim} \text{Bernoulli}(1 -$

$e^{-\lambda\Delta}$), for $i \in [n-1]$. We have that:

$$\begin{aligned} q_{\text{new}}(r^{\text{prop}}|r^{\text{cur}}) &= q_{\text{new}}(r^{\text{prop}}) \\ &= (1-\exp(-\lambda\Delta))^{|r^{\text{prop}}|} \times \exp(-\lambda\Delta)^{n-|r^{\text{prop}}|-1}, \end{aligned}$$

where λ is chosen by the practitioner.

- **Type 2.** Notation q_{bd}. While type 1 allows to independently draw a new changepoint vector r^{prop}, the second type proposal, q_{bd}, provides another way for the number of changepoints to vary from iteration to iteration given the current changepoint vector r^{cur}. Instead of using the proposal function q_{bd} mentioned in [24], we modify it such that the new version of q_{bd} will add one or delete one changepoint on the current list of changepoints with equal probabilities. In particular, let $\mathbf{M}_{r^{\text{cur}}} = \{M_1, ..., M_{|r^{\text{cur}}|}\}$ be the set of all current changepoint indices and $r^{\text{prop}} = r^{\text{cur}}$. There is 50% chance that a component is randomly sampled from the current change point indices, $s \sim \text{Uniform}\,(\mathbf{M}_{r^{\text{cur}}})$, then let $r_s^{\text{prop}} = 0$. Otherwise, a component s is drawn from the complement of the current change point indices set, $s \sim \text{Uniform}\,(\mathbf{M}^c_{r^{\text{cur}}})$, then $r_s^{\text{prop}} = 1$. For such a proposal, we have that

$$q_{\text{bd}}(r^{\text{prop}}|r^{\text{cur}}) = \begin{cases} \frac{1}{2|r^{\text{cur}}|}\mathbb{1}_{\{r^{\text{cur}} \neq \mathbf{0}\}}, & \text{if we remove a changepoint} \\ \frac{1}{2(n-1-|r^{\text{cur}}|)}\mathbb{1}_{\{r^{\text{cur}} \neq \mathbf{1}\}}, & \text{if we add a changepoint} \end{cases}$$

- **Type 3.** Notation q_{bd_2}. The third type of proposal, q_{bd_2}, allows for adding or removing two nearby changees in the current list of changepoints. Like type 2, this proposal also provides a way to vary the number of changepoints in each iteration. In section 2.4.2, we provide motivation for including and removing consecutive changees. Let $r^{\text{prop}} = r^{\text{cur}}$. We set the chances of adding or deleting a segment to be equal. In the case where we add a segment, a set $\{s, s'\} \subseteq \mathbf{M}^c_{r^{\text{cur}}}$ is randomly drawn from one of the $|r^{\text{cur}}|+1$ segments $[M_{j-1}, M_j]$ ($\text{d}_j := M_j - M_{j-1}$) where $M_0 = 0, M_{|r^{\text{cur}}|+1} = n$, and $M_1, ..., M_{|r^{\text{cur}}|} \in \mathbf{M}_{r^{\text{cur}}}$, then let $r_s^{\text{prop}} = r_{s'}^{\text{prop}} = 1$. In the case where we remove a segment, a set of two consecutive indices in the set of changepoints indices, i.e., $\{s, s+1\} \subseteq \mathbf{M}_{r^{\text{cur}}}$ ($|r^{\text{cur}}| \geq 2$) is randomly chosen, we then let $r_s^{\text{prop}} = r_{s+1}^{\text{prop}} = 0$. For this type of proposal, we have that

$$q_{\text{bd}_2}(r^{\text{prop}}|r^{\text{cur}}) = \begin{cases} \frac{1}{|r^{\text{cur}}|}\mathbb{1}_{\{|r^{\text{cur}}| \geq 2\}}, & \text{if we delete a segment} \\ \frac{1}{2}\sum_{j=1}^{|r^{\text{cur}}|+1} \frac{(\text{d}_j - 1)(\text{d}_j - 2)}{(n - |r^{\text{cur}}| - 1)(n - |r^{\text{cur}}| - 2)}, & \text{if we insert a new segment} \end{cases}$$

where d_j is the length of the jth segment and $\sum_j^{|r^{\text{cur}}|+1} \text{d}_j = n$.

- **Type 4.** Notation q_{shift}. This type of proposal allows exploration of the best combination of changepoints for a fixed number of changepoints. We obtain the proposed change point vector by randomly sampling two components of the current change point vector as follows:

$$\begin{aligned} s &\sim \text{Uniform}\left(\mathbf{M}_{|r^{\text{cur}}|}\right), \\ s' &\sim \text{Uniform}\left(\mathbf{M}^c_{|r^{\text{cur}}|}\right). \end{aligned}$$

The proposal of this type, which is symmetric, is defined as

$$r_i^{\text{prop}} = \begin{cases} r_i^{\text{cur}}, & \text{if } i \neq s, s' \\ 1 - r_i^{\text{cur}}, & \text{otherwise.} \end{cases}$$

We have that

$$q_{\text{shift}}(r^{\text{prop}}|r^{\text{cur}}) = \frac{1}{|r^{\text{cur}}|} \times \frac{1}{n-1-|r^{\text{cur}}|}.$$

Finally, we combine all these proposal types to become one final proposal function:

$$q_r(r^{\text{prop}}|r^{\text{cur}}; u_r) = \begin{cases} q_{\text{new}}(r^{\text{prop}}), & \text{if } 0 \leq u_r \leq u_1 \\ q_{\text{bd}}(r^{\text{prop}}|r^{\text{cur}}), & \text{if } u_1 < u_r \leq u_2 \\ q_{\text{bd}_2}(r^{\text{prop}}|r^{\text{cur}}), & \text{if } u_2 < u_r \leq u_3 \\ q_{\text{shift}}(r^{\text{prop}}|r^{\text{cur}}), & \text{if } u_3 < u_r \leq 1, \end{cases} \tag{18}$$

where $u_1, u_2 - u_1, u_3 - u_2, 1 - u_3$ are probabilities that the proposal types 1, 2, 3, and 4 are chosen, respectively. The sampling algorithm is then described in Algorithm 1. We set $u_1 = 1/4, u_2 = 3/8, u_3 = 1/2$ as default in the algorithm. We then introduce the CPLASS algorithm as in Algorithm 2.

Algorithm 1 MH algorithm: Unknown number of changepoints

Input: The observed data $(\mathbf{x}, \mathbf{y}, \mathbf{t})$, the rate of change point processes (λ), a time rate (Δ). The number of iterations (T_{max}).

Output: A list contains T_{max} change point vectors $\{r^{(t)}\}_{t=0}^{T_{max}}$ with their corresponding $\widehat{\text{RSS}}_r(\mathbf{x}, \mathbf{y}, \mathbf{t}; r^{(t)})$, $\tau_{k,r^{(t)}}, \hat{s}_{k,r^{(t)}}$ for $k \in [K_{r^{(t)}}]$.

1: $t = 0$. Draw randomly $r^{(0)}$ from a $n-1$ i.i.d Bernoulli$(1 - e^{-\lambda\Delta})$, then compute $\widehat{\text{RSS}}_r(\mathbf{x}, \mathbf{y}, \mathbf{t}; r^{(0)})$, $\tau_{k,r^{(0)}}, \hat{s}_{k,r^{(0)}}$ (for $k \in [K_{r^{(0)}}]$) by using the piecewise linear approximations.
2: **for** $t = 1, 2, \ldots, T_{max}$ **do**
3: Draw $u_r \sim \text{Uniform}(0, 1)$; $r^{\text{prop}} = q_r(.|r^{(t-1)}; u_r)$ (18)
4: Compute the acceptance probability

$$\log\left(\alpha\left(r^{(t-1)}, r^{\text{prop}}\right)\right) = \begin{cases} \min\left\{0, \Phi(r^{\text{prop}}) - \Phi(r^{(t-1)}) + \log\left(\frac{q_r(r^{(t-1)}|r^{\text{prop}}; u_r)}{q_r(r^{\text{prop}}|r^{(t-1)}; u_r)}\right)\right\}, & e^{\Phi(r^{(t-1)})} q_r(r^{\text{prop}}|r^{(t-1)}; u_r) > 0 \\ 0, & e^{\Phi(r^{(t-1)})} q_r(r^{\text{prop}}|r^{(t-1)}; u_r) = 0. \end{cases}$$

5: **if** $\alpha\left(r^{(t-1)}, r^{\text{prop}}\right) \geq \text{Uniform}(0, 1)$ **then**
6: Set $r^{(t)} = r^{\text{prop}}$
7: **else**
8: Set $r^{(t)} = r^{(t-1)}$
9: **end if**
10: Compute $\widehat{\text{RSS}}_r(\mathbf{x}, \mathbf{y}, \mathbf{t}; r^{(t)})$, $\tau_{k,r^{(t)}}, \hat{s}_{k,r^{(t)}}$ (for $k \in [K_{r^{(t)}}]$) by using the piecewise linear approximations.
11: **end for**

Algorithm 2 CPLASS algorithm

Input: The output from running Algorithm 1.

Output: A list contains the continuous piecewise linear approximation of the data in terms of $\mathbf{x}$ and $\mathbf{y}$, changes in time, segment durations, segment speeds.

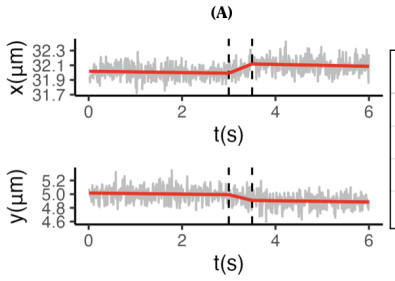
1: Finding the maximum of the collected $\{\Phi(r^{(t)})\}_{t=1,\ldots,T_{max}}$ and returning the corresponding $r^{(t*)}$.

2: Using the continuous piecewise linear approximation with the finding $r^{(t*)}$ and returning the final output of the algorithm.

2.4.2 Necessity of Type 3 proposals

While it is natural to add, subtract, or shift individual points, the specific structure of the change-in-velocity model – which may involve short, fast segments – also requires the proposals for adding or subtracting pairs of changepoints that would mark the endpoints of short segments. Certainly, MH will still succeed in the absence of such proposals, but finding short, fast segments can take much longer. This can happen because it is unlikely in long paths for the algorthim to randomly propose two positions that are close to each other, but there is a deeper issue that also impedes convergence. Similar to what is seen in Figure 1, adding one endpoint of a short segment, but not the other, can lead to a decrease in the criterion function. So it becomes unlikely that a correct proposal of a short segment endpoint will be accepted.

We demonstrate the issue explicitly with the following numerical experiment. We simulated a 6-second path at 100Hz with changes at $t = 3$s and $t = 3.5$s. See Figure 4(A). The segment speeds are $(0, 0.2, 0)$ μm/s, and we assume that the current changepoint vector is $r^{\text{cur}} = \mathbf{0}$, i. e., there are no changepoints. Figure 4(B) provides the corresponding log-likelihood and criterion values for three proposals: two that involve adding one of the true endpoints of the active segment, and one that adds both simultaneously. Given the displayed criterion function evaluations, a progression from zero inferred changepoints to one and then two is prohibitively unlikely.



(B)

Changepoints	r	Log_likelihood	Criterion Value	Decision
No changepoints	..00000...00000...	1028.784	996.1998	
At t = 3	..00010...00000...	1031.056	985.4383	✖ Reject
At t=3.5	..00000...01000...	1028.846	983.2289	✖ Reject
At t=3, t=3.5	..00010...01000...	1063.836	1005.1851	✔ Accept

Figure 4: *Necessity of the new proposal function, summaries of results from Numerical Experiment 2.4.2.* **Panel (A)**, the simulated lysosomal movement trajectory in 2D at 100Hz for a duration of 6 seconds and the two actual changepoints, 3 seconds and 3.5 seconds, is represented by the t-vs-x and t-vs-y time series. The dashed lines represent the real changepoints. The corresponding segmentation is overlaid in red. **Panel (B)** shows different models with changepoints and their corresponding log-likelihood, the algorithm criterion value, and the decision in the MH algorithm 1.

3 Results

This section introduces the theorem of consistency in estimating changepoints and their locations, investigates the choice of the penalty through experiments, and evaluates the performance of the CPLASS algorithm on both simulated and real datasets.

3.1 Consistent estimation of changepoints with the linear penalty

We consider the asymptotic properties of estimating changepoints by maximizing (16) under the choice of the linear penalty $\rho \log(n)^\gamma = [d(|r|+2)+1]\log(n)^\gamma = [d(k+1)+1]\log(n)^\gamma$ (with $\gamma > 1$) (which equivalent to obtain the penalized MLE $\hat{f}_n$, σ_n^2 and $\hat{k}_n$ of (5) with pen $= \rho \log(n)^\gamma$). We further assume that the sets of velocities $\mathbf{V} \subset \mathcal{V} \subset \mathbb{R}^d$, initial intercept $\underline{a} \in \mathcal{A} \subset \mathbb{R}^d$, and $\sigma^2 \in \Omega = [\underline{\sigma}^2, \overline{\sigma}^2] \subset (0, \infty)$, where $\mathcal{V}, \mathcal{A}$ are compact subsets of $\mathbb{R}^d$ and Ω is the compact subset of $\mathbb{R}$.

Theorem 3.1: Consistency Theorem

Suppose that data is generated according to the true model (4) with true signal function $f^0 = f_{\boldsymbol{\tau}^0, \mathbf{V}^0, \underline{a}^0}$ of k_0 segments, $\min_i |\tau_i^0 - \tau_{i-1}^0| > \underline{C}_1 > 0$ and $\min_i \|V_i^0 - V_{i-1}^0\| > \underline{C}_2 > 0$. Then the penalized MLE solution $\widehat{f}_n, \widehat{\sigma}_n^2, \widehat{k}_n$ obtained from (5), where pen $= \rho \log(n)^\gamma$ (with $\gamma > 1$), satisfies

$$\mathbb{P}\left(\widehat{k}_n = k_0, \max_{i \in [k_0 - 1]} |\widehat{\tau}_i^n - \tau_i^0| \leq C\sqrt{\frac{\log n}{n}}\right) \to 1, \tag{19}$$

as $\Delta \to 0$ ($n \to \infty$), where $\rho = d(|r|+2)+1 = d(k+1)+1$ is the total number of parameters of the model, $|r|$ is the number of changepoints, k is the number of segments, C is a contant depending on $\underline{C}_1$, $\underline{C}_2$, and $\overline{k}$.

The proof of the theorem is in the Appendix D. Theorem 3.1 established the consistency of the changepoints and number of changepoints in penalized MLE problem (5) with the pen $= \rho \log(n)^\gamma$ (for $\gamma > 1$).

The findings closely resemble those from the Narrowest-over-Threshold method used by Baranowski, Chen, and Fryzlewicz [6], as well as Fearnhead, Maidstone, and Letchford [14], designed to identify slope changes. Here, specifically consider the penalized MLE framework and show that the estimator is also consistent. The theoretical justification of our model is somewhat more challenging than [41] due to the continuous requirement of the signal function. Our proof technique relies on empirical process theory [54], which is a popular framework for showing the consistency of M-estimations such as MLE.

Notice that we provide the consistency theorem for the linear penalty term. There is an assumption about the compactness of the space of segment speeds associated with a trajectory. In practice, this is a reasonable assumption; for example, we know that the speed of molecular motors cannot exceed 5μm/s in most cases. Adding the speed penalty will not reduce the consistency of CPLASS, as it only reflects the practitioner's prior knowledge of the upper limit of speed supporting the compactness assumptions. In Section 3.2, we conducted numerical experiments to confirm the necessity of incorporating the speed penalty term, ensuring it does not reduce the consistency in estimating the number of changepoints.

3.2 Speed penalization improves estimation without losing power

We revisit the penalty construction through the following numerical experiments: (1) performance of CPLASS when varying the value of γ in the linear penalty term, (2) the effect of adding the speed penalty on the output of CPLASS, and (3) under what circumstances adding the speed penalty is necessary.

3.2.1 Performance of CPLASS with different values of the linear penalty term

In this experiment, we ran CPLASS with different values of γ and recommend using $\gamma = 1.01$ to ensure good performance of CPLASS for sample sizes $n \geq 50$ in detecting *short motile* segments (see Figure 5). Particularly, we consider the simulation paths under the two following setups:

$$H_0 : \mathcal{M}_0, \text{ model with no changes,} \qquad H_1 : \mathcal{M}_1, \text{ model with two changes.}$$

The simulation setups are designed to challenge the algorithm so that the distance between two actual changepoints under the alternative model remains small (9 and 3 time steps for sample sizes $n = 53$ and $n = 203$, respectively). The speed between the two changing times is 0.1μm/s for case $n = 53$ and 0.15μm/s for case $n = 203$.

Specifically, for plotting Panel (A), two sets of simulation paths were generated over a time period from 0s to 2.65s ($n = 53$) with a frequency of 20Hz. The first set consists of 200 paths with no changes, simulated under the null hypothesis with $\sigma = 0.01$ and $s = 0\mu$m/s. The second set contains 200 paths simulated under the alternative hypothesis, with two actual changes occurring at specific times, namely $t = 1.1s$ and $t = 1.55s$, $\sigma = 0.01$, and $(s_1, s_2, s_3) = (0, 0.1, 0)\mu$m/s. We then run the CPLASS with different γ values ranging from 1 to 2. We also run the CPLASS with and without the speed-control penalty. At each γ value, we report the probability of the algorithm returning the correct number of changes—specifically two—in 200 alternative paths, as well as the probability of the algorithm returning a different number of changes from zero in the 200 null paths.

In Panel (B), we repeat the above procedure with simulated trajectories observed at 20Hz, with the noise $\sigma = 0.01$ over 10.15 seconds ($n = 203$). Under the null, there are no changepoints. Under the alternative, the two actual changes are at $t = 5$s and $t = 5.15$s, with the actual speed vector being $(0, 0.15, 0)\mu$m/s.

As shown in Figure 5, CPLASS effectively detects the true number of changepoints in 95% of the trajectories across both models, even when the distance between two changepoints and the corresponding segment speed is minimal. The speed penalty function maintains a high probability of identifying the actual number of changepoints under the alternative hypothesis and keeping the probability of failing to detect a changepoint under the null model nearly 0. In the next section, we will discuss the reason for proposing the speed penalty.

3.2.2 Necessity of the speed penalty

The previous experiment indicates that adding the speed penalty does not negatively affect the algorithm's output. In other words, it does not impact the number of detected changepoints, but only leads to slight alterations in the locations of changes. This has the effect of tempering inferred segment speeds when sharper transitions would result in biophysically implausible results (see Figure 10 in Appendix A). To leverage the speed penalty based on the knowledge of the dataset, we implemented CPLASS and BCP on a collection of 250 simulated trajectories

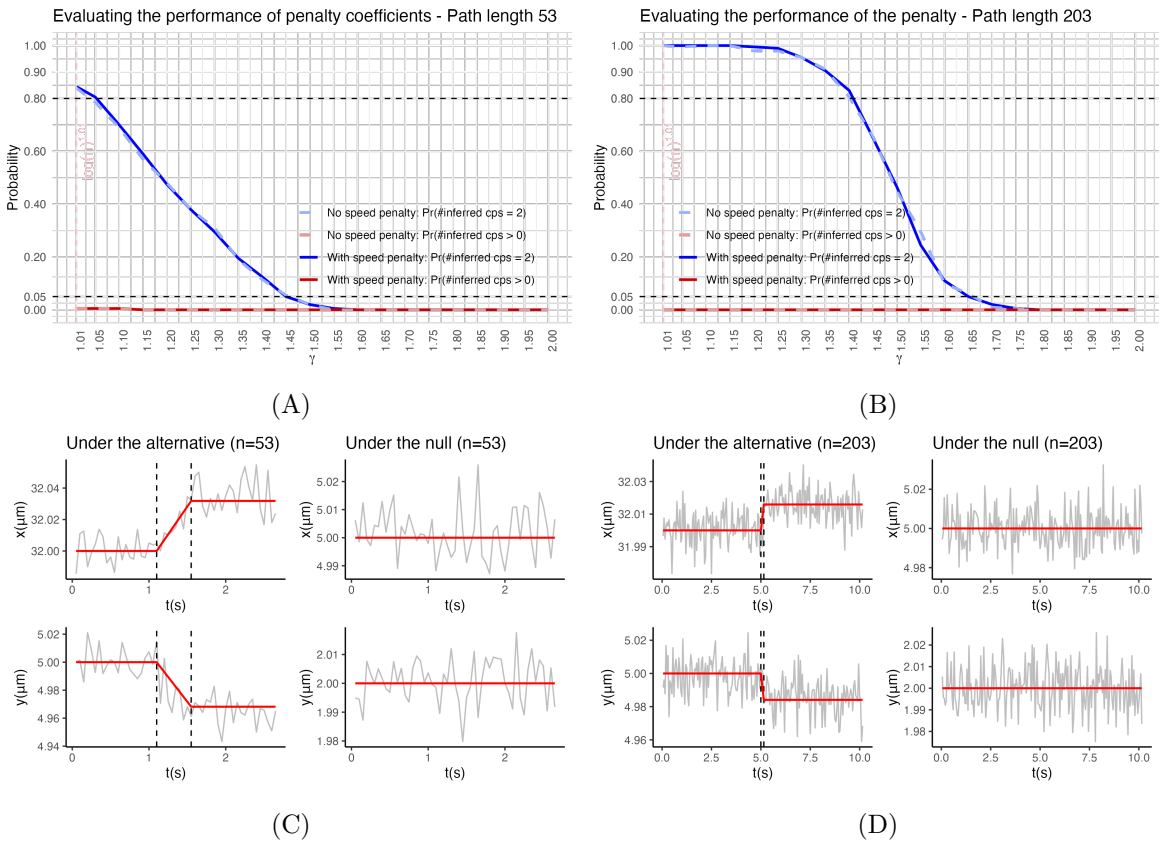


Figure 5: *Investigating CPLASS on detecting motile short segments while varying the γ values (Section 3.2.1).* **Panel (A)** 200 simulation paths over 2.65 seconds at 20Hz with $s = 0$, $\sigma = 0.01$ for the null hypothesis. 200 simulation paths over 2.65 seconds at 20Hz ($n = 53$) with two actual changes at $t = 1.1$s and $t = 1.55$s, three segments speeds $(s_1, s_2, s_3) = (0, 0.1, 0)\mu$m/s, $\sigma = 0.01$ for the alternative hypothesis. **Panel (B)** 200 simulation paths over 10.15 seconds at 20Hz ($n = 203$) with $s = 0$, $\sigma = 0.01$ for the null hypothesis. 200 simulation paths over 10.15 seconds at 20Hz with two actual changes at $t = 5$s and $t = 5.15$s, three segments speeds $(s_1, s_2, s_3) = (0, 0.15, 0)\mu$m/s, $\sigma = 0.01$ for the alternative hypothesis. The red colors show the simulation results under the null hypotheses, and the blue colors show the simulation results under the alternative hypotheses. **Panel (C) and (D)** The simulated paths under the null model with no changepoints and the alternative model with two stationary segments and the motile short segment at the middle.

derived from the base parameter sets in Table 1 from [11] (also see Table 1 in Appendix B) at 25Hz. We then compared the inferred segmentation outputs by BCP, CPLASS (with and without speed penalties), and the truth by using two summary statistics, namely the Cumulative Speed Allocation statistic introduced in [11] and the Cumulative Distribution Function of the inferred maximum segment speeds (see Figure 6). For every speed $s \geq 0$, the CSA is the inferred proportion of time spent at speeds less than or equal to s. We refer to Cook et al. (2024) [11] for a more detailed discussion on the CSA.

In Panel (A) of Figure 6, we display the result of applying the CPLASS (with and without the speed penalty) and BCP algorithms to 250 simulated trajectories. Each member of the CSA curve ensembles, orange for the BCP output, blue for CPLASS without the speed penalty,

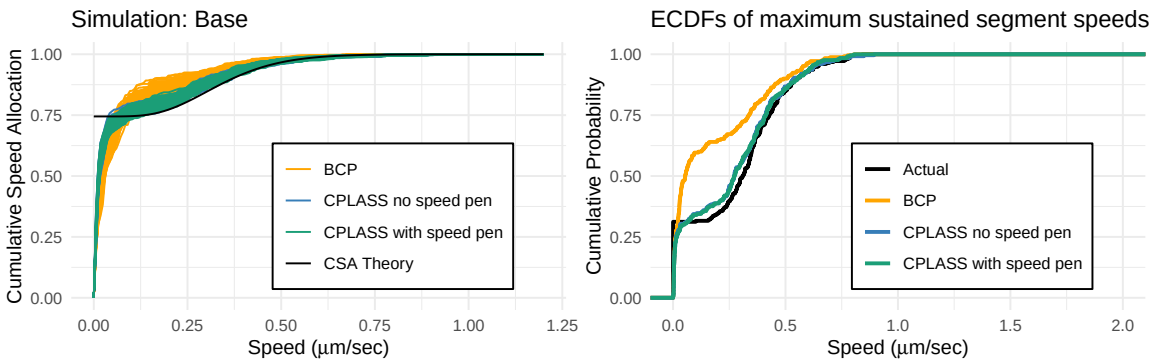


Figure 6: *Example of the necessity of the speed penalty (Section 3.2.2).* We simulated a set of 250 simulated trajectories, each of the parameter sets in Table 1 from [11], observed at 25Hz. (**Left**) Cumulative Speed Allocation (CSA) computation for the simulated trajectories. The ***black*** line denotes the theoretical CSA of each parameter set. The ***green*** lines denote the inferred CSA computed after running CPLASS with the speed penalty. The ***blue*** lines denote the inferred CSA computed after running CPLASS without the speed penalty. The ***orange*** lines denote the inferred CSA computed after running BCP in [48]. (**Right**) Empirical cumulative distributions for the collection of maximum sustained segment speeds (segment durations of least 0.6s) the simulated trajectories after running BCP (***orange***), CPLASS with speed penalty (***green***), CPLASS without speed penalty (***blue***) are compared to the actual maximum segment speeds represented in ***black*** color.

and green for CPLASS with the speed penalty, is the inferred CSA calculated from bootstrap resampling of the 250 paths. The evident gap between the CSA ensembles highlights the distinction between BCP and CPLASS, particularly in the proportion of time that the simulated particles are moving at speeds of 0.5μm/s or slower. Meanwhile, both versions of CPLASS (with the speed penalty activated or deactivated) closely follow the theoretical CSA curve (in black), which was used to simulate the data. This is consistent with the argument we discussed earlier in this paper regarding the issue of missing short-fast segments when using the discontinuous piecewise linear model.

The right panel of Figure 6 illustrates the empirical cumulative distribution (ECDF) of the *maximum sustained segment speeds* after running the BCP and the two versions of CPLASS. Here we present the ECDF of the maximum segment speeds where the associated durations are at least 0.6s, since very short durations make it difficult to detect changepoints. We observed that CPLASS, with or without the speed penalty, more accurately matches the CDF of the true maximum speed than the results produced by BCP.

3.3 Change-in-mean versus change-in-velocity analysis

In this section, we report the results of a direct comparison between change-in-mean analysis of location increments (using BCP) versus continuous piecewise linear approximation (CPLASS with and without speed penalty) by simulating 20000 paths under the same conditions. All paths had two actual changes observed at 20Hz with $\sigma = 0.01$, but the positions of the changepoints varied due to differences in the duration and speed of the middle segment. The 20 values for the middle segment duration were selected over the interval from 0.05 seconds to 1 second, with an increment of 0.05 seconds. The middle segment speed values (20 different

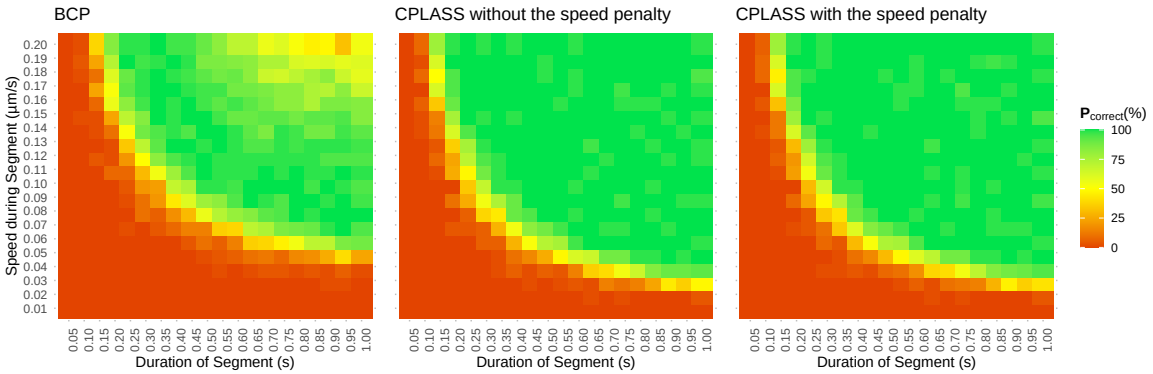


Figure 7: *Power analysis comparing CPLASS and BCP (Section 3.3).* There are 20000 simulated paths at 20Hz with two actual changes ($\sigma = 0.01$). In each path, the first and the third segment durations are two seconds. We then varied the middle segment's durations (20 different values in seconds) and speeds (20 different values in μm/s). For each variation of the pair of speed and duration, there are 50 corresponding paths. The correctly detected percentage $\mathbf{P}_{\text{correct}}$ is then computed as percentages of finding the correct number of changes for each case of the duration and speed. The paths are simulated under the same rule as in Figure 5 Panel (C), (D).

values in μm/s) ranged from 0.01 μm/s to 0.2 μm/s, increasing by 0.01 μm/s. We fixed the durations of the first and third segments (2 seconds) as well as those segment speeds (0 μm/s). Therefore, the sample sizes (number of location observations) varied from $n = 81$ to 100. For each variation of the pair of speed and duration, there were 50 corresponding simulated paths. The correctly detected percentage $\mathbf{P}_{\text{correct}}$ was then computed for each speed/duration pair as the proportion of paths in which the segmentation algorithm reported the correct number of changes. We ran CPLASS and BCP for these 20000 simulation paths; the results showed that CPLASS (with both versions) was better than BCP in detecting short segments and slow segments, which BCP treated as having no movement (e.g., see Appendix A.3 Figure 13). Figure 7 illustrates the comparison. Regions in which the algorithms are considered "effective" are colored in green (see legend for color gradient). The region in which CPLASS (both with and without the speed penalty) is effective is notably larger than that of BCP. CPLASS also achieved better performance than BCP in detecting the true number of changepoints with $\mathbf{P}_{\text{correct}} \geq 90\%$ for all cases where the middle segment had a duration varying from 0.45s to 1s and the speed varying from 0.08μm/s to 0.2μm/s. Additionally, this experiment again confirms that adding the speed penalty function to the penalty function maintains the correctly detected percentage compared to using only the linear penalty term.

3.4 In vivo experimental data - lysosomal transport

In this section, we revisit the data sets in [48]. Two cell lines, monkey kidney epithelial cells, and human lung epithelial cells were cultured in different media but with identical conditions. Cells were supplemented with fetal bovine serum and incubated at 37°C and 5% carbon dioxide. For imaging experiments, cells were transduced with CellLight Lysosomes-green emerald fluorescent protein to label lysosomes fluorescently. Transduction was carried out according to the manufacturer's instructions. The study used live cell imaging and single-particle tracking to observe and characterize lysosome motion. A confocal microscope was used to collect images,

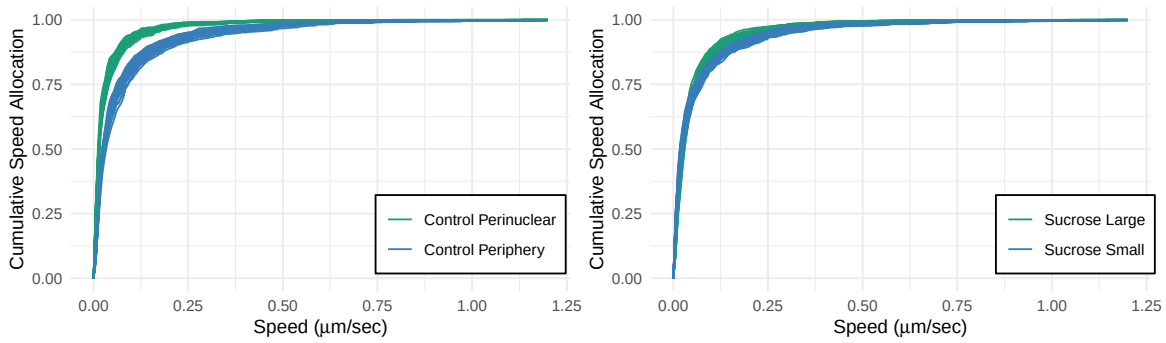


Figure 8: *CSA plots for different group comparisons (Section 3.4).* (**Left**) The CSA bootstrap ensemble curves compare lysosomal transport in perinuclear and periphery regions. (**Right**) The CSA bootstrap ensemble curves compare the sucrose-treated groups restricted to the periphery region of the cell from [48].

and the TrackMate macro was used to track lysosomes. Lysosome trajectories were labeled as being in the perinuclear or peripheral regions of the cells, and were sorted by size. We refer to Rayen et al. (2022) [48] for more details about the data sets and data processing.

We reassessed the queries regarding how lysosomal transport varies with lysosome size and location. Figure 8 indicates that intracellular location, rather than diameter, is a crucial factor in lysosomal motion. This aligns with the findings from Rayens et al.[48]. Analyzing the CSA plot (the left panel of Figure 8), we observe that the lysosome in the perinuclear region spends more time moving slowly compared to that in the peripheral region. The right panel of Figure 8 confirms that large lysosomes are slower in transport than small lysosomes; however, overall, there is not a significant difference between these two groups. In [48], in order to study the differences between these group comparisons, the authors first classify the segment speeds into groups of motile ($s > 0.1\mu$m/s) and stationary ($s < 0.1\mu$m/s), then analyze the empirical cumulative distributions for the motile group. Meanwhile, using CPLASS and CSA, we can analyze the entire collection of segmented speeds and durations without establishing a threshold for the motile segment group.

3.5 In vitro experimental data - quantum dot transport

This section revisited the data sets used in [32] in which quantum dots are transported by a single kinesin-1 (kin-1) motor, a single dynein-dynactin-BicD2 (DDB) motor, and by a Kinesin-1/DDB pair. In [32], Jensen et al. developed a protocol for finding changepoints in cargo trajectories that can be projected along the length of a straight microtubule and reporting velocity distributions. The differences in velocity distributions and run lengths revealed the difference for different molecular motor families. Since our proposed CPLASS can handle multidimensional data sets, we apply it directly to these quantum dot data sets without projecting the two-dimensional data into a one-dimensional format. We then calculate the CSA bootstrap ensemble curves (see the left panel of Figure 9) based on the collection of estimated segment speeds and durations in each motor family group after running CPLASS. The CSA plot illustrates the differences among the motor experiments that correspond with what one might expect in a molecular motor "folklore". In other words, Kinesin-1 steps processively with consistent behavior, while DDB (orange curves) exhibits a broader range of speeds. When

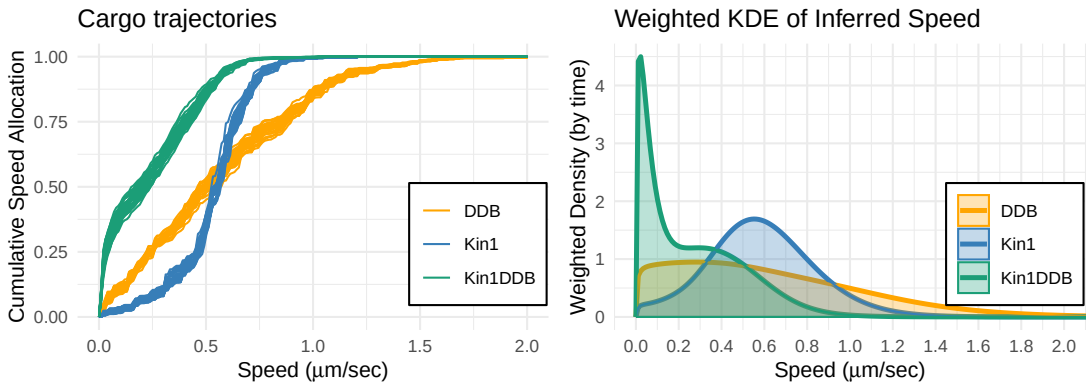


Figure 9: *CSA plot for different motor families (Section 3.5).* (**Left**) The CSA bootstrap ensemble curves compare cargo trajectories among three groups of motor transports: kinesin-1 (Kin1), Dynein-Dynactin-BicD2 (DDB), and Kinesin-1/DDB pairs. (**Right**) Weighted kernel density estimation of inferred segment speeds, duration-weighted, for the three groups.

both motors are present (green curves), the speed is generally lower, reflecting the tug-of-war state. The right panel of Figure 9 - weighted kernel density estimation (WKDE) of inferred segment speeds with weights given by segment durations across the three groups is a good tool for visualizing our argument. The tug-of-war phenomenon is more evident in this plot with more slow-segments - one mode around 0 (Kin1DDB), and the second mode is evident for cases when one of the motors wins. The multimodal we observe in the WKDE here is also reported in the paper of [32]. All of these confirm the observations made about the data sets but offer a more refined and robust characterization.

4 Discussion

In this work, we introduced the CPLASS algorithm for detecting changes in velocity within multidimensional data, addressing key challenges in both probability structure and search methodology. While detecting changes in velocity seems to be a similar statistical problem to detecting changes in mean, it is fundamentally more challenging. Popular generic approaches for detecting multiple changepoints do not work for detecting changes in velocity. For example, popular "bottom-up" strategies, like binary segmentation, are initiated by assuming a single changepoint and finding the most likely location. However, as we discuss in the main text, the changepoint time inferred under the assumption that there is one change may not appear among (or near) the changepoints detected assuming there are two. This is a marked difference from the change-in-mean detection problem. Moreover, existing dynamic programming algorithms like PELT and optimal partitioning cannot handle change-in-slope due to continuity assumptions that create parameter dependencies, breaching essential independence structures. To address these issues, Baranowski et al. [6] proposed the Narrowest-Over-Threshold (NOT) algorithm, while Fearnhead et al. Fearnhead et al. [15] introduced a variant of dynamic programming, and Kim et al. [36] offered trend-filtering methods. These work well for one-dimensional slope changes, but our challenge arises from analyzing multidimensional intracellular transport data. These challenges motivated our development of an MCMC-based approach, which includes specialized proposal mechanisms tailored to efficiently navigate the parameter space.

While we established a consistency theorem for our method, real-world applications, such as molecular motor data, pose additional challenges because practitioners often have prior knowledge of a practical upper bound on attainable speeds. To incorporate this information, we introduced a speed penalty that adjusts the placement of change points to discourage unrealistically short segments that would imply implausibly high speeds. This acts as a domain-informed prior, improving realism in the inference of speeds while preserving consistency in the large sample limit. Furthermore, by comparing Cumulative Speed Allocation (CSA) [11] with the Cumulative Distribution Function (CDF), we showed that quantifying the proportion of time spent at different speeds yields a more stable performance metric than segment velocity counts, making the approach less sensitive to algorithmic variation.

Crucially, our method is inherently multidimensional, allowing it to capture complex structures in diverse datasets. However, computational efficiency remains a challenge, because MCMC search is inherently slow. In future work we will explore alternative methods for improving the search process for maximize likelihood estimation. Additionally, ensuring the consistency of CSA inference via segmentation analysis remains an open theoretical question, warranting further exploration.

Acknowledgements

This research was supported by the NSF-Simons Southeast Center for Mathematics and Biology (SCMB) through the grants National Science Foundation DMS1764406 and Simons Foundation/SFARI 594594. We thank Christine K. Payne and William O. Hancock for many substantive discussions and sharing insights from their labs.

References

[1] Aminikhanghahi, S. and Cook, D. (2017). A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367.

[2] Angelosante, D. and Giannakis, G. B. (2012). Group lassoing change-points piece-constant AR processes. *EURASIP Journal on Advances in Signal Processing*, 70.

[3] Bai, J. (1999). Likelihood ratio tests for multiple structural changes. *Journal of Econometrics*, 91(2):299–323.

[4] Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78.

[5] Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18:1–22.

[6] Baranowski, R., Chen, Y., and Fryzlewicz, P. (2019). Narrowest-over-threshold detection of multiple change points and change-point-like features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):649–672.

[7] Barry, D. and Hartigan, J. A. (1992). Product partition models for change point problems. *Annals of Statistics*, pages 260–279.

[8] Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88:309–319.

[9] Braverman, V. (2016). Sliding window algorithms. In Kao, M. Y., editor, *Encyclopedia of Algorithms*. Springer.

[10] Chen, S. S. and Gopalakrishnan, P. S. (1998). Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, page 8, Landsdowne, VA.

[11] Cook, K. J., Rayens, N., Do, L., Payne, C. K., and McKinley, S. A. (2025). Considering experimental frame rates and robust segmentation analysis of piecewise-linear microparticle trajectories. *Mathematical Biosciences and Engineering*, 22(10):2595–2626.

[12] Duderstadt, K. E., Geertsema, H. J., and van Oijen, A. M. (2016). Simultaneous real-time imaging of leading and lagging strand synthesis reveals the coordination dynamics of single replisomes. *Molecular Cell*, 64(6):1035–1047.

[13] Erdman, C. and Emerson, J. W. (2008). bcp: An R package for performing a Bayesian analysis of change point problems. *Journal of Statistical Software*, 23.

[14] Fearnhead, P. and Grose, D. (2022). cpop: Detecting changes in piecewise-linear signals. *arXiv*, 2208:11009.

[15] Fearnhead, P., Maidstone, R., and Letchford, A. (2019). Detecting Changes in Slope with an ℓ_0 Penalty. *Journal of Computational and Graphical Statistics*, 28(2):265–275.

[16] Fearnhead, P. and Rigaill, G. (2020). Relating and comparing methods for detecting changes in mean. *Stat*, 9(1):e291.

[17] Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change-point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580.

[18] Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243 – 2281.

[19] Fryzlewicz, P. (2023). Narrowest significance pursuit: Inference for multiple change-points in linear models. *Journal of the American Statistical Association*.

[20] Fryzlewicz, P. and Subba Rao, S. (2014). Multiple change-point detection for auto-regressive conditional heteroscedastic processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(5):903–924.

[21] Futschik, A., Hotz, T., Munk, A., and Sieling, H. (2014). Multiscale DNA partitioning: statistical evidence for segments. *Bioinformatics*, 30(16):2255–2262.

[22] Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5):1233 – 1263.

[23] Giné, E. and Nickl, R. (2021). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press.

[24] Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

[25] Gross, S. P. (2004). Hither and yon: a review of bi-directional microtubule-based transport. *Physical biology*, 1(2):R1.

[26] Hancock, W. O. (2014). Bidirectional cargo transport: Moving beyond tug of war. *Nature Reviews Molecular Cell Biology*, 15(9):615–628.

[27] Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(490):1480–1493.

[28] Harchaoui, Z., Vallet, F., Lung-Yut-Fong, A., and Cappe, O. (2009). A regularized kernel-based approach to unsupervised audio segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1665–1668, Taipei, Taiwan.

[29] Haynes, K., Eckley, I. A., and Fearnhead, P. (2017). Computationally efficient changepoint detection for a range of penalties. *Journal of Computational and Graphical Statistics*, 26(1):134–143.

[30] Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tsai, T. T. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2):105–108.

[31] Jandhyala, V., Fotopoulos, S., Macneill, I., and Liu, P. (2013). Inference for single and multiple change-points in time series. *Journal of Time Series Analysis*, 34(4):423–446.

[32] Jensen, M. A., Feng, Q., Hancock, W. O., and McKinley, S. A. (2021). A change point analysis protocol for comparing intracellular transport by different molecular motor combinations. *Mathematical Biosciences and Engineering*, 18(6):8962–8996.

[33] Karslake, J. D., Donarski, E. D., Shelby, S. A., Demey, L. M., DiRita, V. J., Veatch, S. L., and Biteen, J. S. (2021). SMAUG: Analyzing Single-Molecule Tracks with Nonparametric Bayesian Statistics. *Methods*, 193:16–26.

[34] Killick, R., Eckley, I. A., Ewans, K., and Jonathan, P. (2010). Detection of changes in variance of oceanographic time-series using changepoint analysis. *Ocean Engineering*, 37:1120–1126.

[35] Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

[36] Kim, S. J., Koh, K., Boyd, S., and Gorinevsky, D. (2009). ℓ_1 trend filtering. *SIAM Review*, 51(2):339–360.

[37] Kunwar, A., Tripathy, S. K., Xu, J., Mattson, M. K., Anand, P., Sigua, R., Vershinin, M., McKenney, R. J., Yu, C. C., Mogilner, A., et al. (2011). Mechanical stochastic tug-of-war models cannot explain bidirectional lipid-droplet transport. *Proceedings of the National Academy of Sciences*, 108(47):18960–18965.

[38] Lavielle, M. (1998). Optimal segmentation of random processes. *IEEE Transactions on Signal Processing*, 46:1365–1373.

[39] Lavielle, M. and Lebarbier, E. (2001). An application of MCMC methods for the multiple change-points problem. *Signal Processing*, 81(1):39–53.

[40] Lavielle, M. and Teyssiere, G. (2007). Adaptive detection of multiple change-points in asset price volatility. In *Long-Memory in Economics*, pages 129–156. Springer Verlag, Berlin, Germany.

[41] Liu, J., Wu, S., and Zidek, J. (1997). On segmented multivariate regression. *Statistica Sinica*, pages 497–525.

[42] Manzo, C. and Garcia-Parajo, M. F. (2015). A review of progress in single particle tracking: From methods to biophysical insights. *Reports on Progress in Physics*, 78(12):124601.

[43] Monnier, N., Guo, S.-M., Mori, M., He, J., Lénárt, P., and Bathe, M. (2012). Bayesian Approach to MSD-Based Analysis of Particle Motion in Live Cells. *Biophysical Journal*, 103(3):616–626.

[44] Neumann, S., Chassefeyre, R., Campbell, G. E., and Encalada, S. E. (2017). Kymoanalyzer: A software tool for the quantitative analysis of intracellular transport in neurons. *Traffic*, 18(1):71–88.

[45] Niu, Y. and Zhang, H. (2012). The screening and ranking algorithm to detect DNA copy number variations. *Annals of Applied Statistics*, 6(3):1306–1326.

[46] Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572.

[47] Persson, F., Lindén, M., Unoson, C., and Elf, J. (2013). Extracting intracellular diffusive states and transition rates from single-molecule tracking data. *Nature Methods*, 10:265–269.

[48] Rayens, N. T., Cook, K. J., McKinley, S. A., and Payne, C. K. (2022). Transport of lysosomes decreases in the perinuclear region: Insights from changepoint analysis. *Biophysical Journal*, 121(7):1205–1218.

[49] Scott, A. J. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3):507–512.

[50] Seichepine, N., Essid, S., Fevotte, C., and Cappe, O. (2014). Piecewise constant nonnegative matrix factorization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6721–6725, Florence, Italy.

[51] Shi, X., Gallagher, C., Lund, R., and Killick, R. (2022). A comparison of single and multiple change-point techniques for time series data. *Computational Statistics and Data Analysis*, 170:107433.

[52] Sosa-Costa, A., Manzano, D., Molina, R. J., Portillo, J. R., and Parrondo, J. M. R. (2018). PLANT: A method for detecting changes of slope in noisy trajectories. *Biophysical Journal*, 114(9):2044–2051.

[53] Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

[54] van de Geer, S. (2000). *Empirical Processes in M-estimation*. Cambridge University Press.

[55] Venkatraman, E. S. and Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657–663.

[56] Vert, J.-P. and Bleakley, K. (2010). Fast detection of multiple change-points shared by many signals using group LARS. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, pages 2343–2351.

[57] Yao, Q. and Au, S.-K. (1989). Least squares estimation of a step function. *Sankhyā: The Indian Journal of Statistics, Series A*, 51(3):370–381.

[58] Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters*, 6(3):181–189.

[59] Yin, S., Song, N., and Yang, H. (2018). Detection of velocity and diffusion coefficient change points in single-particle trajectories. *Biophysical Journal*, 115(2):217–229.

[60] Zhang, N. R. and Siegmund, D. O. (2007). A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomic Hybridization Data. *Biometrics*, 63(1):22–32.

Appendix

A Additional figures

In this section, we have provided the output plots from CPLASS for simulated and real data. Note that for coloring the stationary (pink area) and motile (green area) in the following plots, we use a threshold of 0.1μm/s.

A.1 CPLASS on the simulation dataset

Figure 10 shows two paths from the base simulation dataset mentioned in Section 3.2.2 under both versions of CPLASS (no speed penalty vs. with speed penalty). We can see that the number of detected changepoints remains the same; however, the speed penalty helps prevent the unrealistic speed in the output of CPLASS.

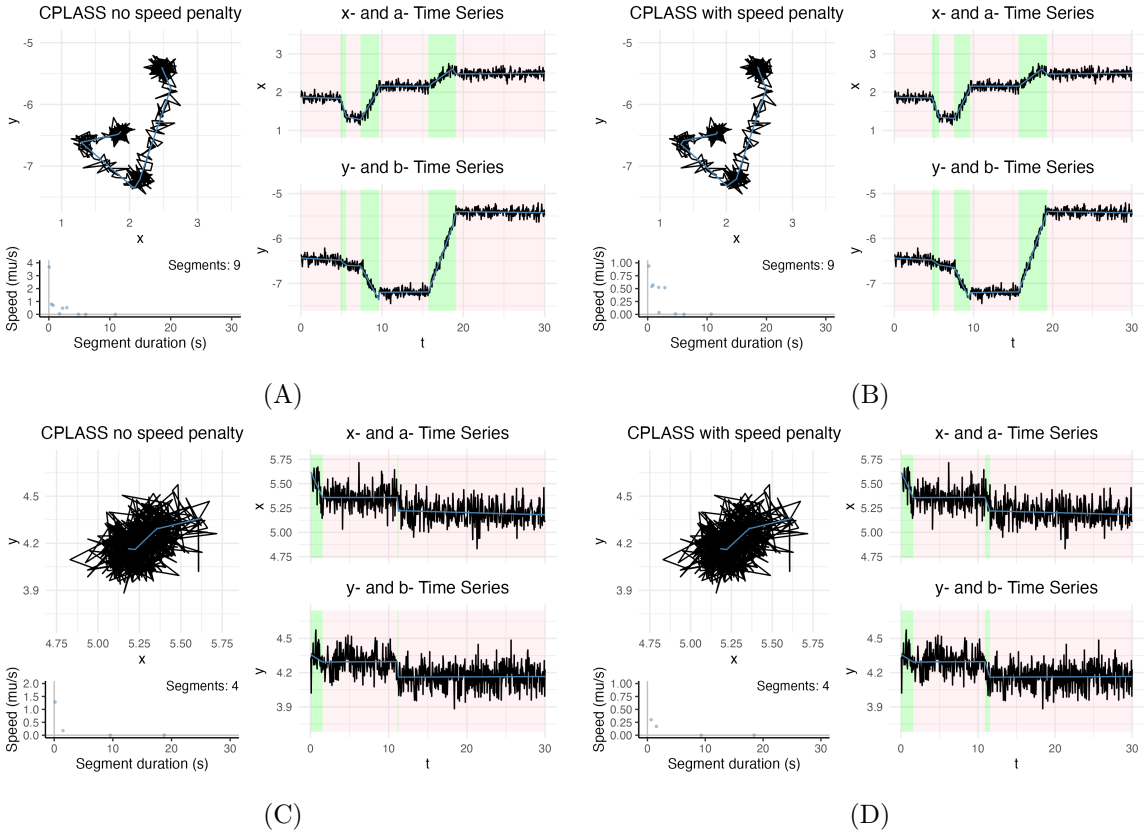


Figure 10: *Examples of the necessity of the speed penalty.* Two trajectories were simulated at 20Hz using the Base parameter set. These examples show the benefit of adding the speed penalty to prevent unrealistic speed.

A.2 Examples of CPLASS performance for experimental observations

In this section, we provided some real paths and output from CPLASS, Figure 11 shows four paths coming from different datasets, the first two panel (A) and (B) are paths from the lysosomal transport dataset in the periphery region of the cell, the last two panel (C) and

(D) are paths from the quantum dot transport datasets under two circumstances - one with a single Kinesin-1 motor (panel (C)) and one when the two motors Kin1 and DDB present at the same time. Figure 12 is another path in the group Kin1DDB. We can observe bidirectional, multi-state transport, possibly due to tug-of-war dynamics between the motors. For the Kin1 path, the motor-cargo complex appears to be always in active transport; the detected changes in this case are when it changes the speed or direction during the movement.

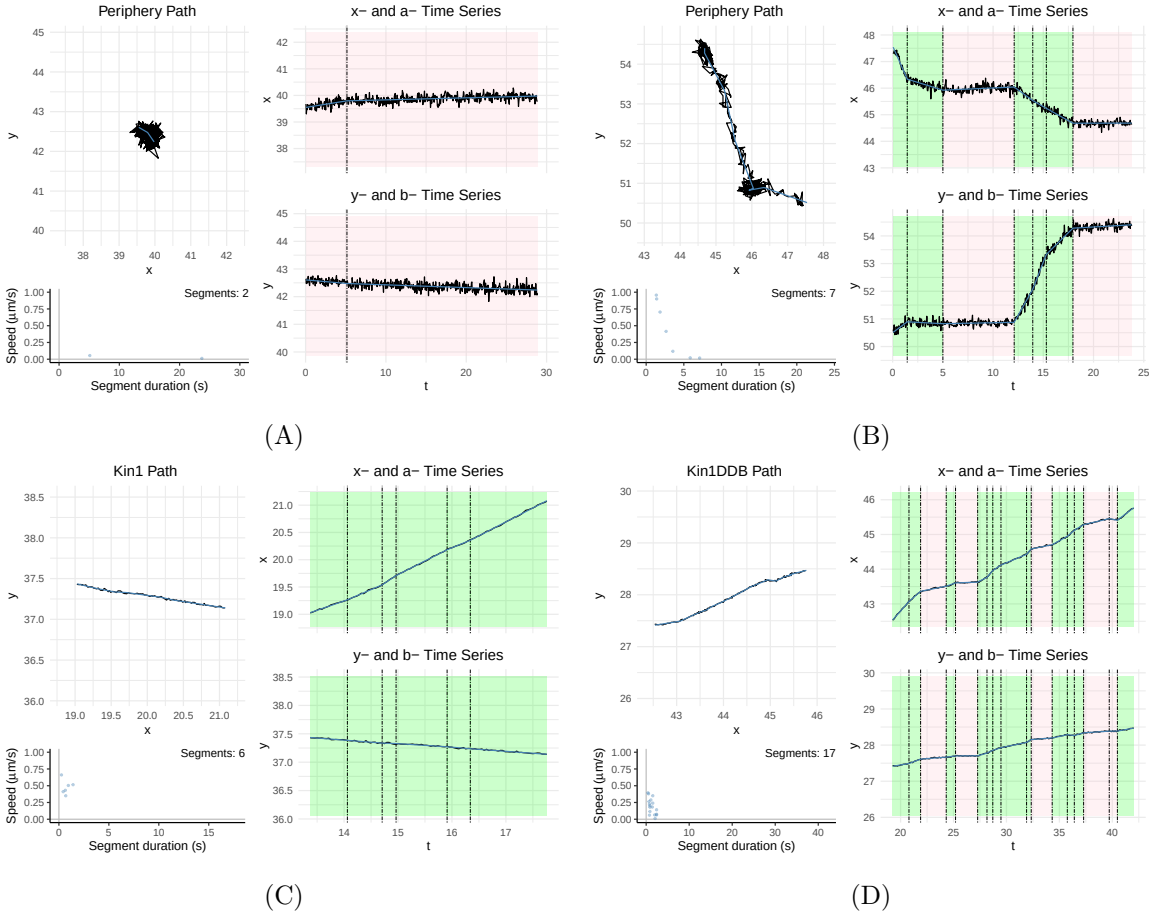


Figure 11: *CPLASS performance for intracellular transport trajectories* **Panel (A) and (B)** - *in vivo* lysosomal transport - examples of a non-active path (Panel A) and an active path (Panel B). Panel (C) and (D) - *in vitro* quantum dot transport - are examples of an active path with a kin1 motor (Panel C) and a path moving and pausing while the two motors kin1 and DDB are present at the same time.

A.3 BCP and CPLASS on detecting short segments

In this section, we provide an example where BCP struggles with detecting short segments as discussed in Section 3.3 and Figure 7. With this issue, the motile segment was missed, and the whole path was labeled stationary. Meanwhile, CPLASS detected the short fast segment and labeled the middle segment active (see Figure 13).

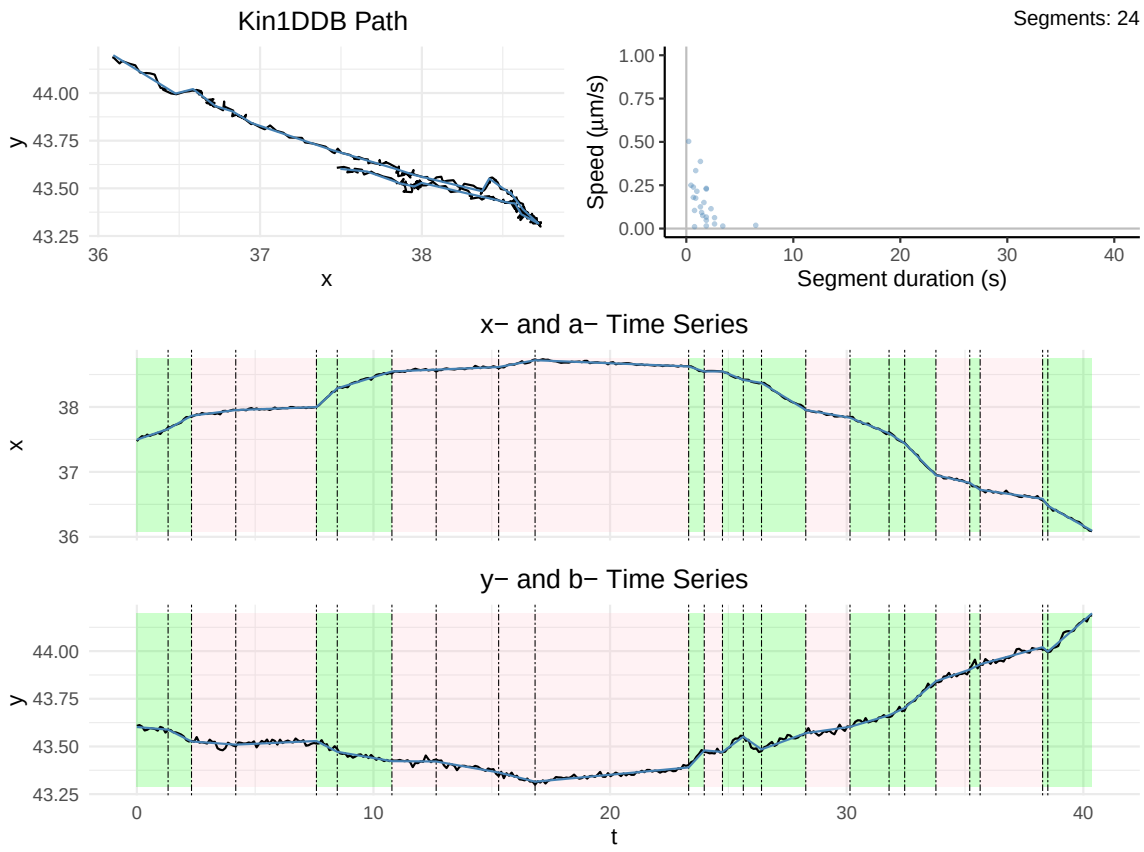


Figure 12: A Kin1DDB path reflecting the tug-of-war phenomenon.The green and pink shaded regions indicate segments that are estimated to have a speed of over or under 100 nm/s respectively.

B Table of the parameter sets for simulating the dataset in Section 3.2.2

This section contains the parameter sets (Table 1) used to simulate 250 trajectories at 25Hz. We refer to [11] for more discussion on the datasets.

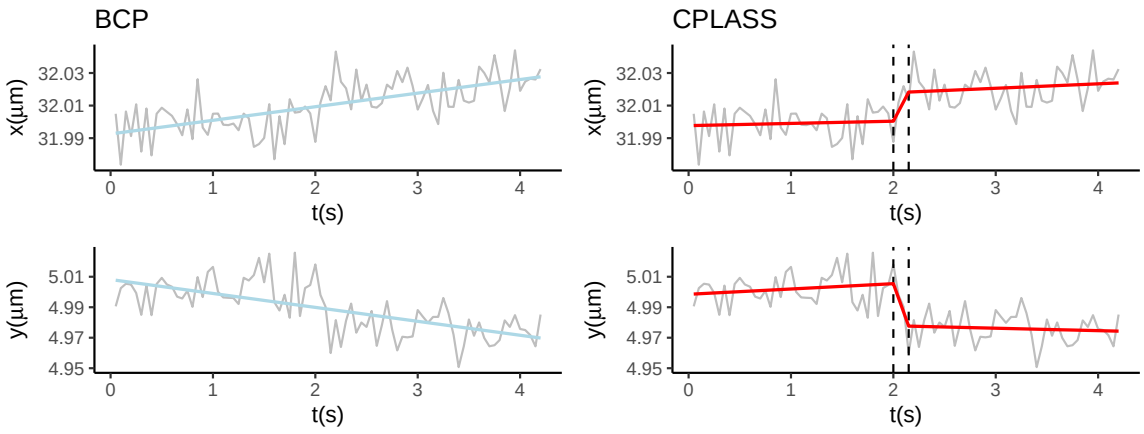
Here, we summarize the model and data generation (Section 2.1 in [11]) .

The simulated trajectories in the referenced study are generated using a two-state stochastic model representing cargo motion that alternates between *stationary* and *motile* states. The model is parameterized by transition probabilities between these states, motile speed distributions, and additive noise terms.

During simulation, the cargo switches between states at each discrete time step Δ according to transition probabilities:

- p: probability of switching from stationary $\rightarrow$ motile,
- q: probability of switching from motile $\rightarrow$ stationary.

When in the motile state, the cargo speed is sampled from a Gamma distribution with shape α and rate β. The direction of motion may reverse or persist with probabilities P_{reverse} and



Segment	1
Duration	4.2 s
Speed	0.0122 μm/s
State	Stationary

Segment	1	2	3
Duration	1.95 s	0.15 s	2.05 s
Speed	0.0037 μm/s	0.2207 μm/s	0.0032 μm/s
State	Stationary	Motile	Stationary

Figure 13: *BCP and CPLASS on detecting short segments.* This path is one of 50 simulated paths mentioned at Section 3.3 at 20Hz, $n = 84$, $\sigma = 0.01$, two actual time changes created three segments with the corresponding duration $(2, 0.2, 2)s$ and speed $(0, 0.17, 0)\mu$m/s. While BCP failed to detect the changes, CPLASS can return the two changepoints with the estimated locations and speed close to the actual locations and speed.

P_{continue}, respectively. A Gaussian noise term with standard deviation σ_{cargo} is added to model cargo localization uncertainty. The total number of simulated points is n, and the final simulation time is $t_f = n\Delta$.

The authors define three parameter groups ("Base", "Contrast", and "Mimic"). The "Base" group serves as the reference setting used to simulate synthetic lysosome trajectories at multiple frame rates. The corresponding parameters are summarized in Table 1 below.

To generate data, the simulation proceeds as follows:

1. Initialize the trajectory in either state.

2. For each time step Δ, update the state based on p and q.

3. If motile, draw a speed from Gamma(α, β) and move the particle accordingly, allowing reversals or continuations based on P_{reverse} and P_{continue}.

4. Add Gaussian noise with magnitude σ_{cargo}.

5. Because the segments are not exponentially distributed and the process is non-Markovian, a burn-in period is required and the process is repeated until a total time t_f is reached.

This process yields a set of synthetic lysosome trajectories that serve as ground truth for evaluating segmentation and motion-state inference algorithms.

Parameter	Base	Description
n	200	Number of Observations
p	1	Probability *Stationary* to *Motile*
q	0.5	Probability *Motile* to *Stationary*
α	8	Speed Shape Parameter
β	0.02	Speed Rate Parameter
$\bar{D}$	300 nm	Average Distance Traveled
σ	5 s	Average *Stationary* Duration
P_{reverse}	0.3	Probability of reversal
P_{continue}	0.3	Probability of same direction
σ_{cargo}	0.1	Noise magnitude of Cargo
Δ	0.04 s	Time Step
t_f	$n\Delta$	Final simulation time

Table 1: Parameters used to model simulated lysosome trajectories. "Base" denotes parameters for simulated lysosome trajectories at any frame rate. In this paper, we use the simulated data at 25Hz, i.e., $\Delta = 0.04$.

C Checking the detailed balance condition in the MH algorithm

Given the proposal function with four proposal types in Section 2.4, to prove detailed balance for the MH algorithm, we need to show that the transition kernel satisfies:

$$\pi(r^{\text{cur}})q_r(r^{\text{prop}}|r^{\text{cur}})\alpha(r^{\text{prop}}|r^{\text{cur}}) = \pi(r^{\text{prop}})q_r(r^{\text{cur}}|r^{\text{prop}})\alpha(r^{\text{cur}}|r^{\text{prop}}), \tag{20}$$

where

- $\pi(r) = \exp(\Phi(r))$ is the target posterior distribution of changepoints.
- $q_r(r^{\text{prop}}|r^{\text{cur}})$ is the overall proposal function, combining four different proposal types with predefined probabilities, and
- $\alpha(r^{\text{prop}}|r^{\text{cur}})$ is the MH acceptance probability:

$$\alpha\left(r^{\text{prop}}|r^{\text{cur}}\right) = \min\left(1, \frac{\pi(r^{\text{prop}})q_r\left(r^{\text{cur}}|r^{\text{prop}}\right)}{\pi(r^{\text{cur}})q_r\left(r^{\text{prop}}|r^{\text{cur}}\right)}\right).$$

In order to verify the detailed balance condition in the MH algorithm, we need to analyze whether the proposal function $q_r(r^{\text{prop}}|r^{\text{cur}})$ satisfies symmetry, meaning that the probability of proposing r^{prop} given r^{cur} is equal to the probability of proposing r^{cur} given r^{prop}, or if any asymmetry exists, it is properly accounted for in the acceptance probability.

We analyze the acceptance probability of each type of proposal in detail:

Type 1: q_{new} (Independent changepoint vector proposal) In this proposal, we generate a completely new set of changepoints independently of the current state r^{cur}. The new changepoints are generated from a Bernoulli process with probability $1 - \exp(-\lambda\Delta)$.

We have

$$q_{\text{new}}(r^{\text{prop}}|r^{\text{cur}}) = \left(\frac{1-e^{-\lambda\Delta}}{e^{-\lambda\Delta}}\right)^{|r^{\text{prop}}|} \times (e^{-\lambda\Delta})^{n-1},$$
$$q_{\text{new}}(r^{\text{cur}}|r^{\text{prop}}) = \left(\frac{1-e^{-\lambda\Delta}}{e^{-\lambda\Delta}}\right)^{|r^{\text{cur}}|} \times (e^{-\lambda\Delta})^{n-1}.$$

Therefore,

$$\alpha(r^{\text{cur}}|r^{\text{prop}}) = \min\left(1, \frac{\pi(r^{\text{prop}})}{\pi(r^{\text{cur}})} \times \left(\frac{e^{-\lambda\Delta}}{1-e^{-\lambda\Delta}}\right)^{|r^{\text{prop}}|-|r^{\text{cur}}|}\right),$$

Type 2: $q_{\mathbf{bd}}$ (Birth/death proposal) This proposal adds or removes a single changepoints at random. We have that if $q_{\text{bd}}(r^{\text{prop}}|r^{\text{cur}}) = \dfrac{1}{2|r^{\text{cur}}|}$ then $q_{\text{bd}}(r^{\text{cur}}|r^{\text{prop}}) = \dfrac{1}{2(n-|r^{\text{prop}}|-1)}$ and vice versa. The acceptance rate is then

$$\alpha(r^{\text{cur}}|r^{\text{prop}}) = \min\left(1, \frac{\pi(r^{\text{prop}})}{\pi(r^{\text{cur}})} \times \frac{|r^{\text{cur}}|}{n-|r^{\text{prop}}|-1}\right) \text{ or}$$
$$\alpha(r^{\text{cur}}|r^{\text{prop}}) = \min\left(1, \frac{\pi(r^{\text{prop}})}{\pi(r^{\text{cur}})} \times \frac{n-|r^{\text{cur}}|-1}{|r^{\text{prop}}|}\right),$$

respective.

Type 3: $q_{\mathbf{bd}_2}$ (Segment insertion/deletion proposal) For this type of proposal, if $q_{\text{bd}_2}(r^{\text{prop}}|r^{\text{cur}}) = \dfrac{1}{|r^{\text{cur}}|}\mathbb{1}_{\{|r^{\text{cur}}|\geq 2\}}$ then

$$q_{\text{bd}_2}(r^{\text{cur}}|r^{\text{prop}}) = \frac{1}{2}\sum_{j=1}^{|r^{\text{prop}}|+1} \frac{(\mathrm{d}_j-1)(\mathrm{d}_j-2)}{(n-|r^{\text{prop}}|-1)(n-|r^{\text{prop}}|-2)}$$

and vice versa. The acceptance rate is then

$$\alpha(r^{\text{cur}}|r^{\text{prop}}) = \min\left(1, \frac{\pi(r^{\text{prop}})}{\pi(r^{\text{cur}})} \times |r^{\text{cur}}| \times \mathbb{1}_{\{|r^{\text{cur}}|\geq 2\}} \times \frac{1}{2}\sum_{j=1}^{|r^{\text{prop}}|+1} \frac{(\mathrm{d}_j-1)(\mathrm{d}_j-2)}{(n-|r^{\text{prop}}|-1)(n-|r^{\text{prop}}|-2)}\right)$$

or

$$\alpha(r^{\text{cur}}|r^{\text{prop}}) = \min\left(1, \frac{\pi(r^{\text{prop}})}{\pi(r^{\text{cur}})} \times \frac{\frac{1}{|r^{\text{prop}}|}\mathbb{1}_{\{|r^{\text{prop}}|\geq 2\}}}{\frac{1}{2}\sum_{j=1}^{|r^{\text{cur}}|+1} \frac{(\mathrm{d}_j-1)(\mathrm{d}_j-2)}{(n-|r^{\text{cur}}|-1)(n-|r^{\text{cur}}|-2)}}\right), \text{ respectively.}$$

Type 4: $q_{\mathbf{shift}}$ (Shift proposal) This proposal shifts the position of one of the changepoints in the current list of changepoints. The proposal is symmetric as defined. Therefore, the acceptance rate in this case is: $\alpha(r^{\text{cur}}|r^{\text{prop}}) = \min\left(1, \dfrac{\pi(r^{\text{prop}})}{\pi(r^{\text{cur}})}\right)$.

Given all the acceptance rates for each type of proposal, there are two cases which are

Case 1: $\pi(r^{\mathbf{prop}})q_{\mathbf{new}}(r^{\mathbf{cur}}|r^{\mathbf{prop}}) \geq \pi(r^{\mathbf{cur}})q_{\mathbf{new}}(r^{\mathbf{prop}}|r^{\mathbf{cur}})$**.** We then have

$$\alpha(r^{\text{prop}}|r^{\text{cur}}) = 1; \quad \alpha(r^{\text{cur}}|r^{\text{prop}}) = \frac{\pi(r^{\text{cur}})q_{\text{new}}\left(r^{\text{prop}}|r^{\text{cur}}\right)}{\pi(r^{\text{prop}})q_{\text{new}}\left(r^{\text{cur}}|r^{\text{prop}}\right)}.$$

Plugging into the two sides of Equation (20), we get that both sides of this equation are equal to $\pi(r^{\text{cur}})q_{\text{new}}(r^{\text{prop}}|r^{\text{cur}})$.

Case 2: $\pi(r^{\mathbf{prop}})q_{\mathbf{new}}(r^{\mathbf{cur}}|r^{\mathbf{prop}}) < \pi(r^{\mathbf{cur}})q_{\mathbf{new}}(r^{\mathbf{prop}}|r^{\mathbf{cur}})$**.** We then have

$$\alpha(r^{\text{cur}}|r^{\text{prop}}) = 1; \quad \alpha(r^{\text{prop}}|r^{\text{cur}}) = \frac{\pi(r^{\text{prop}})q_{\text{new}}\left(r^{\text{cur}}|r^{\text{prop}}\right)}{\pi(r^{\text{cur}})q_{\text{new}}\left(r^{\text{prop}}|r^{\text{cur}}\right)}.$$

Plugging into the two sides of Equation (20), we get that both sides of this equation are equal to $\pi(r^{\text{prop}})q_{\text{new}}(r^{\text{cur}}|r^{\text{prop}})$.

We finish checking the detailed balance condition.

D Consistency theorems

D.1 Continuous piecewise linear model

Recall from Section 2.1 of the main text, we assume n multivariate observations $(Y_i)_{i=1}^n \subset \mathbb{R}^d$ of a particle's locations on $\mathcal{T} = \{t_1, \ldots, t_n\}$ are generated according to a true signal function and Gaussian noises:

$$y_i \overset{ind.}{\sim} \mathcal{N}(f^0(t_i), \sigma_0^2 I_d), \tag{21}$$

where $f^0(t) := f_{\boldsymbol{\tau}^0, \mathbf{V}^0, \underline{a}^0}(t)$ is the true signal function of k_0 segments with true changepoints $\boldsymbol{\tau}^0 = (\tau_1^0, \ldots, \tau_{k_0-1}^0)$, sets of slopes $\mathbf{V}^0 = (V_0^0, \ldots, V_{k_0-1}^0) \subset \mathbb{R}^d$, and initial intercept $\underline{a}^0 \in \mathbb{R}^d$. I_d is the $d-$dimensional identity matrix. σ_0^2 is the true variance, which is assumed to belong to a known compact set $\Omega = [\underline{\sigma}^2, \overline{\sigma}^2] \subset (0, \infty)$. As discussed in the main text, given the set of observations $(Y_i)_{i=1}^n$, our goal is to infer the true number of segments k_0, the parameters $\boldsymbol{\tau}^0, \mathbf{V}^0, \underline{a}^0$ of the true signal function, and the noise level σ_0^2.

CPLASS is designed to learn these parameters by maximizing a penalized likelihood of changepoint models with at most $\overline{k}$ segments in the data

$$(\widehat{f}_n, \widehat{\sigma}_n^2, \widehat{k}_n) = \underset{f \in \mathcal{F}_k, \sigma^2 \in \Omega, k \leq \overline{k}}{\arg\max} \sum_{i=1}^n \log \mathcal{N}(y_i | f_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}}(t_i), \sigma^2 I) - (d(k+1)+1)(\log(n))^\gamma, \tag{22}$$

where $1 < \gamma < \infty$, to get the MLE $\widehat{f}_n := f_{\widehat{\boldsymbol{\tau}}^n, \widehat{\mathbf{V}}^n, \widehat{\underline{a}}^n}$ of $\widehat{k}_n$ segments. As discussed in the main text, given the optimal signal function $\widehat{f}_n$, the optimal variance $\widehat{\sigma}_n^2$ can be shown to be the average RSS:

$$\widehat{\sigma}_n^2 = \frac{\sum_{i=1}^n \|y_i - \widehat{f}_n(t_i)\|^2}{nd}. \tag{23}$$

It follows that (22) is equivalent to maximizing the criterion function:

$$(\widehat{f}_n, \widehat{k}_n) = \underset{f \in \mathcal{F}_k, k \leq \overline{k}}{\arg\max} -\frac{nd}{2} \log\left(\sum_{i=1}^n \|y_i - f_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}}(t_i)\|^2\right) - (d(k+1)+1)(\log(n))^\gamma. \tag{24}$$

We note that (22) is equivalent to finding the MLE with each $k \in [\overline{k}]$

$$(\widehat{f}_n^{(k)}, \widehat{\sigma}_{n,k}^2) = \underset{f \in \mathcal{F}_k, \sigma^2 \in \Omega}{\arg\max} \sum_{i=1}^{n} \log \mathcal{N}(y_i | f_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}}(t_i), \sigma^2 I), \tag{25}$$

and then find

$$\widehat{k}_n = \underset{k \in [\overline{k}]}{\arg\max} \sum_{i=1}^{n} \log \mathcal{N}(y_i | \widehat{f}_n^{(k)}(t_i), \widehat{\sigma}_{n,k}^2 I) - (d(k+1)+1)(\log(n))^{\gamma}. \tag{26}$$

Hence, to prove Theorem 3.1, we aim to understand the convergence of the parameters and likelihood in (25) for each k first. Central to our theoretical development is the empirical process theory [54], which provides uniform convergence of the empirical average log-likelihood to the population average log-likelihood. For a function $f : \mathcal{T} \to \mathbb{R}^d$ and $\sigma > 0$, denote the empirical and population average log-likelihood with respect to parameter f and σ^2 as

$$\overline{\mathcal{L}}_n(f, \sigma^2) = \frac{1}{n} \sum_{i=1}^{n} \log \mathcal{N}(y_i | f_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}}(t_i), \sigma^2). \tag{27}$$

and

$$\overline{\mathcal{L}}_0(f, \sigma^2) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{Y_i \sim N(f^0(t_i), \sigma_0^2)} \log \mathcal{N}(Y_i | f_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}}(t_i), \sigma^2), \tag{28}$$

respectively.

D.2 Preliminaries on empirical process theory

Given two sequences of non-negative functions $(p_1, \ldots, p_n)$ and $(q_1, \ldots, q_n)$ on $\mathcal{Y}$ (support of data, which is $\mathbb{R}^d$ in our problem), define the Hellinger process distance [23, 54] between product densities $p = \otimes_{i=1}^n p_i$ and $q = \otimes_{i=1}^n q_i$ and its average to be

$$h_n^2(p, q) := \frac{1}{2} \sum_{i=1}^{n} h^2(p_i, q_i), \quad \overline{h}_n^2(p, q) := \frac{1}{n} h_n^2(p, q). \tag{29}$$

Note that for the arguments below, for ease of notation, we write $\hat{f}_n^{(k)}$ for $f_{\widehat{\boldsymbol{\tau}}^{n,k}, \widehat{\mathbf{V}}^{n,k}, \widehat{\underline{a}}^{n,k}}$, and f^0 for $f_{\boldsymbol{\tau}^0, \mathbf{V}^0, \underline{a}^0}$.

For a function $f_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}}$ and noise level σ^2, we define the product density

$$p^{(n)}_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}, \sigma^2} = \otimes_{i=1}^n \mathcal{N}(y_i | f_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}}(t_i), \sigma^2)$$

on $\mathcal{Y}^n$. Let $p_0^{(n)}$ denote the true product density $\otimes_{i=1}^n \mathcal{N}(y_i | f^0(t_i), \sigma_0^2)$ on $\mathcal{Y}^n$. Empirical process theory provides many useful concentration inequalities uniformly over balls in the density space $\{p^{(n)}_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}, \sigma^2} : \boldsymbol{\tau} \subset \mathcal{T}, \mathbf{V} \subset \mathcal{V}, \underline{a} \in \mathcal{A}, \sigma^2 \in \Omega\}$ so that the convergence rate of MLE boils down to calculating (or sufficiently providing an upper bound for) the smallest number of balls needed to cover this space in the Hellinger process distance. This number is often referred to as the "covering number," which we now define.

Definition 4.1: Entropy number with bracketing

For $\delta > 0$ and a set $\Theta \subset \mathcal{T}^{k-1} \times \mathcal{V}^k \times \mathcal{A} \times \Sigma$, let $N_B(\delta, \Theta)$ be the smallest integer N such that there exists a collection of non-negative functions $\{\mathbf{p}_j^L, \mathbf{p}_j^U\}_{j=1}^N$ with $\mathbf{p}_j^L = (p_{j1}^L, \ldots, p_{jn}^L)$ and $\mathbf{p}_j^U = (p_{j1}^U, \ldots, p_{jn}^U)$ such that for every $(\boldsymbol{\tau}, \mathbf{V}, \underline{a}, \sigma^2) \in \Theta$, there is a j such that

(i) $\overline{h}_n \left(\frac{\mathbf{p}_j^L + p_0^{(n)}}{2}, \frac{\mathbf{p}_j^U + p_0^{(n)}}{2} \right) \leq \delta$ and

(ii) $p_{ji}^L(y_i) \leq \mathcal{N}(y_i | f_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}}(t_i), \sigma^2) \leq p_{ji}^U(y_i)$ for all $y_i \in \mathcal{Y}$ and $i \in [n]$.

Then $N_B(\delta, \Theta)$ and $H_B(\delta, \Theta) = \log N_B(\delta, \Theta)$ are called the Hellinger covering number and entropy number with bracketing, respectively. When $\Theta = \mathcal{T}^{k-1} \times \mathcal{V}^k \times \mathcal{A} \times \Sigma$, we write those numbers as $N_B(\delta)$ and $H_B(\delta)$ for short.

For c_0 being a suitable universal constant [54], define the entropy integral:

$$J_B(\delta) := \int_{\delta^2/c_0}^{\delta} H_B^{1/2}(u) du \vee \delta, \quad 0 < \delta \leq 1. \tag{30}$$

The main result, for which the notation is adapted to our model, is stated as follows.

Theorem 4.1: Theorem 8.14 in [54]

Suppose there exists a function $\Psi(\delta) \geq J_B(\delta)$, and $\Psi(\delta)/\delta^2$ is a non-increasing function of δ. Then for a given sequence (δ_n) and a universal constant $c > 0$ satisfying

$$\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n), \tag{31}$$

we have that for all $\delta \geq \delta_n$,

$$\mathbb{P}_0 \left(\overline{h}_n \left(p^{(n)}_{\widehat{\boldsymbol{\tau}}, \widehat{\mathbf{V}}, \widehat{\underline{a}}, \widehat{\sigma}^2}, p_0^{(n)} \right) \geq \delta \right) \leq c \exp\left(-\frac{n\delta^2}{c^2} \right). \tag{32}$$

We also need a uniform concentration bound of the empirical process:

Theorem 4.2: Theorem 8.13 in [54]

Let positive numbers R, D, C_1, b, and a subset of parameter space $\Theta \subset \mathcal{T}^{k-1} \times \mathcal{V}^k \times \mathcal{A} \times \Sigma$ satisfy:

$$\overline{h}_n(\bar{p}^{(n)}_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}, \sigma^2}, p_0^{(n)}) \leq R \quad \forall (\tau, \mathbf{V}, \underline{a}, \sigma^2) \in \Theta, \tag{33}$$

$$b \leq C_1 \sqrt{n} R^2 \wedge 8\sqrt{n} R, \tag{34}$$

and

$$b \geq \sqrt{D^2(C_1 + 1)} \left(\int_{b/(2^6\sqrt{n})}^{R} H_B^{1/2}\left(\frac{u}{\sqrt{2}}, \Theta \right) du \vee R \right), \tag{35}$$

then

$$\mathbb{P}_0\Big(\sup_{(\tau,\mathbf{V},\underline{a},\sigma^2)\in\Theta} \sqrt{n}|\bar{Z}_{\boldsymbol{\tau},\mathbf{V},\underline{a},\sigma^2} - \bar{A}_{\boldsymbol{\tau},\mathbf{V},\underline{a},\sigma^2}| \geq b\Big) \leq D\exp\left[-\frac{b^2}{D^2(C_1+1)R^2}\right], \quad (36)$$

where $\mathbf{Z}_{\boldsymbol{\tau},\mathbf{V},\underline{a},\sigma^2} = (Z_{1,\boldsymbol{\tau},\mathbf{V},\underline{a},\sigma^2}, \cdots, Z_{n,\boldsymbol{\tau},\mathbf{V},\underline{a},\sigma^2})$ with

$$Z_{i,\boldsymbol{\tau},\mathbf{V},\underline{a},\sigma^2} = \frac{1}{2}\log\left(\frac{\mathcal{N}(y_i|f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}(t_i),\sigma^2)+\mathcal{N}(y_i|f^0(t_i),\sigma_0^2)}{2\mathcal{N}(y_i|f^0(t_i),\sigma_0^2)}\right),$$

$\bar{Z}_{\boldsymbol{\tau},\mathbf{V},\underline{a},\sigma^2} = \frac{1}{n}\sum_i^n Z_{i,\boldsymbol{\tau},\mathbf{V},\underline{a},\sigma^2}$, $\bar{A}_{\boldsymbol{\tau},\mathbf{V},\underline{a},\sigma^2} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{Y_i\sim\mathcal{N}(f^0(t_i),\sigma_0^2)} Z_{i,\boldsymbol{\tau},\mathbf{V},\underline{a},\sigma^2}$, and $\bar{p}^{(n)}_{\boldsymbol{\tau},\mathbf{V},\underline{a},\sigma^2} = \frac{p^{(n)}_{\boldsymbol{\tau},\mathbf{V},\underline{a},\sigma^2}+p_0^{(n)}}{2}$.

For the proof of Theorems 4.1 and 4.2, we refer to van de Geer [54].

D.3 Convergence of latent piecewise functions and likelihood functions

We used some notations in the following proofs. We present them as follows.

Notation. For two sequences $(a_n)_{n=1}^\infty$ and $(b_n)_{n=1}^\infty$, we write $a_n \lesssim b_n$ (or $a_n = O(b_n)$) if $a_n \leq Cb_n$ where C is a constant not depending on n. We write $a_n \gtrsim b_n$ when $b_n \lesssim a_n$, and $a_n \asymp b_n$ if $a_n \gtrsim b_n$ and $a_n \lesssim b_n$. We write $a_n \ll b_n$ (or $a_n = o(b_n)$) if $a_n/b_n \to 0$ as $n \to \infty$. For two density functions p and q, denote $h^2(p,q) = \frac{1}{2}\int(p^{1/2}(y) - q^{1/2}(y))^2 dy$ by the square Hellinger distance, $V(p,q) = \frac{1}{2}\int|p(y)-q(y)|dy$ the Total Variation distance, and $KL(p\|q) = \int p(y)\log\frac{p(y)}{q(y)}dy$ by the Kullback-Leibler divergence between p and q. They are related by $V^2 \leq \sqrt{2}h \leq V$ and $V(p,q) \leq \sqrt{2KL(p\|q)}$.

Lemma 1. *Suppose that $\Omega \subset [\underline{c},\bar{c}]$ with $0 < \underline{c} < \bar{c} < \infty$. For all $\mu,\tilde{\mu} \in \mathbb{R}^d$ and $\sigma^2,\tilde{\sigma}^2 \in \Omega$, we have*

$$h^2(\mathcal{N}(\mu,\sigma^2 I_d),\mathcal{N}(\tilde{\mu},\tilde{\sigma}^2 I_d)) \asymp \|\mu-\tilde{\mu}\|^2 + (\sigma^2-\tilde{\sigma}^2)^2, \quad (37)$$

and

$$\sup_{y\in\mathbb{R}^d}\left|\mathcal{N}(y|\mu,\sigma^2 I_d) - \mathcal{N}(y|\tilde{\mu},\tilde{\sigma}^2 I_d))\right| \lesssim \|\mu-\tilde{\mu}\| + |\sigma^2-\tilde{\sigma}^2|. \quad (38)$$

Proof of Lemma 1. **1. Proof of (37):**

Recall the Hellinger distance between two location-scale Gaussian densities:

$$h^2(\mathcal{N}(\mu,\sigma^2 I_d),\mathcal{N}(\tilde{\mu},\tilde{\sigma}^2 I_d)) = 1 - \frac{\sigma^{1/2}\tilde{\sigma}^{1/2}}{((\sigma^2+\tilde{\sigma}^2)/2)^{1/2}}\exp\left\{-\frac{1}{8}\left(\frac{2}{\sigma^2+\tilde{\sigma}^2}\right)\|\mu-\tilde{\mu})\|^2\right\}.$$

Firstly, we notice

$$\frac{1}{2\max\{\sigma^2,\tilde{\sigma}^2\}}\|\mu-\tilde{\mu}\|^2 \leq \frac{1}{\sigma^2+\tilde{\sigma}^2}\|\mu-\tilde{\mu}\|^2 \leq \frac{1}{2\min\{\sigma^2,\tilde{\sigma}^2\}}\|\mu-\tilde{\mu}\|^2.$$

Therefore, $\frac{1}{\sigma^2+\tilde{\sigma}^2}\|\mu-\tilde{\mu}\|^2 \asymp \|\mu-\tilde{\mu}\|^2$.

We also have that

$$cx \leq 1-\exp(-x) \leq x,$$

for all $x \in [0, C]$ where c depends on C. Hence,

$$h^2(\mathcal{N}(\mu, \sigma^2 I_d), \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2 I_d)) \asymp \log(\sigma^2) + \log(\tilde{\sigma}^2) - 2\log\left(\frac{\sigma^2 + \tilde{\sigma}^2}{2}\right) + \|\mu - \tilde{\mu}\|^2.$$

Let $\delta = \sigma^2 - \tilde{\sigma}^2$. We have

$$\begin{aligned}\log(\sigma^2) + \log(\tilde{\sigma}^2) - 2\log\left(\frac{\sigma^2 + \tilde{\sigma}^2}{2}\right) &= \log\left(\frac{4\tilde{\sigma}^4 + 4\delta\tilde{\sigma}^2}{4\tilde{\sigma}^4 + 4\tilde{\sigma}^2\delta + \delta^2}\right) \\ &= \log\left(\frac{1 + \delta/\tilde{\sigma}^2}{1 + \delta/\tilde{\sigma}^2 + \delta^2/(4\tilde{\sigma}^4)}\right).\end{aligned}$$

For small δ, use the approximation that $\frac{1+x}{1+y} \approx 1 + (x - y)$ and $\log(1 - x) \approx -x$ we have

$$\log(\sigma^2) + \log(\tilde{\sigma}^2) - 2\log\left(\frac{\sigma^2 + \tilde{\sigma}^2}{2}\right) \approx \log\left(1 - \frac{\delta^2}{4\tilde{\sigma}^4}\right) \approx -\frac{\delta^2}{4\tilde{\sigma}^4}$$

This implies that

$$\log(\sigma^2) + \log(\tilde{\sigma}^2) - 2\log\left(\frac{\sigma^2 + \tilde{\sigma}^2}{2}\right) \asymp (\sigma^2 - \tilde{\sigma}^2)^2.$$

We finish proving

$$h^2(\mathcal{N}(\mu, \sigma^2 I_d), \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2 I_d)) \asymp \|\mu - \tilde{\mu}\|^2 + (\sigma^2 - \tilde{\sigma}^2)^2.$$

2. Proof of (38): We have that

$$\mathcal{N}(y|\mu, \sigma^2 I_d) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|y - \mu\|^2}{2\sigma^2}\right)$$

The derivative with respect to μ and σ^2 are then

$$\frac{\partial}{\partial \mu}\mathcal{N}(y|\mu, \sigma^2 I_d) = \mathcal{N}(y|\mu, \sigma^2 I_d) \cdot \frac{y - \mu}{\sigma^2}, \tag{39}$$

$$\frac{\partial}{\partial \sigma^2}\mathcal{N}(y|\mu, \sigma^2 I_d) = \mathcal{N}(y|\mu, \sigma^2 I_d) \cdot \left(-\frac{d}{2\sigma^2} + \frac{\|y - \mu\|^2}{\sigma^4}\right). \tag{40}$$

By Cauchy-Schwarz inequality and triangle inequality, we have that:

$$\left|\frac{\partial}{\partial \mu}\mathcal{N}(y|\mu, \sigma^2 I_d)\right| \leq \mathcal{N}(y|\mu, \sigma^2 I_d) \cdot \frac{\|y - \mu\|}{\sigma^2}, \tag{41}$$

$$\left|\frac{\partial}{\partial \sigma^2}\mathcal{N}(y|\mu, \sigma^2 I_d)\right| \leq \mathcal{N}(y|\mu, \sigma^2 I_d) \cdot \left(\frac{d}{2\sigma^2} + \frac{\|y - \mu\|^2}{\sigma^4}\right). \tag{42}$$

From the Mean Value Theorem, the difference $\left|\mathcal{N}(y|\mu, \sigma^2 I_d) - \mathcal{N}(y|\tilde{\mu}, \tilde{\sigma}^2 I_d))\right|$ can be expressed in terms of the gradients with respect to μ and σ^2:

$$\sup_{y \in \mathbb{R}^d} \left|\mathcal{N}(y|\mu, \sigma^2 I_d) - \mathcal{N}(y|\tilde{\mu}, \tilde{\sigma}^2 I_d))\right| \leq L_\mu \|\mu - \tilde{\mu}\| + L_{\sigma^2}|\sigma^2 - \tilde{\sigma}^2|, \tag{43}$$

where

$$L_\mu = \sup_{y\in\mathbb{R}^d} \frac{\|y-\mu\|}{\sigma^2}\mathcal{N}(y|\mu,\sigma^2 I_d) = \sup_{y\in\mathbb{R}^d} \underbrace{\frac{1}{(2\pi)^{d/2}\sigma^{2(d/2+1)}}}_{\text{is bounded since } \sigma^2\in[\underline{\sigma}^2,\overline{\sigma}^2]} \underbrace{\|y-\mu\|\exp\left(-\frac{\|y-\mu\|^2}{2\sigma^2}\right)}_{\text{is bounded}} \le C',$$

and

$$\begin{aligned}
L_{\sigma^2} &= \sup_{y\in\mathbb{R}^d}\left(\frac{d}{2\sigma^2}+\frac{\|y-\mu\|^2}{\sigma^4}\right)\mathcal{N}(y|\mu,\sigma^2 I_d)\\
&= \sup_{y\in\mathbb{R}^d}\left(\frac{d}{2\sigma^2}+\frac{\|y-\mu\|^2}{\sigma^4}\right)\frac{1}{(2\pi\sigma^2)^{d/2}}\exp\left(-\frac{\|y-\mu\|^2}{2\sigma^2}\right)\\
&= \sup_{y\in\mathbb{R}^d} \underbrace{\frac{d}{2(2\pi)^{d/2}\sigma^{2(d/2+1)}}}_{\text{is bounded since } \sigma^2\in[\underline{\sigma}^2,\overline{\sigma}^2]} \underbrace{\exp\left(-\frac{\|y-\mu\|^2}{2\sigma^2}\right)}_{\text{is bounded}} + \underbrace{\frac{1}{(2\pi)^{d/2}\sigma^{2(d/2+2)}}}_{\text{is bounded since } \sigma^2\in[\underline{\sigma}^2,\overline{\sigma}^2]} \underbrace{\|y-\mu\|^2\exp\left(-\frac{\|y-\mu\|^2}{2\sigma^2}\right)}_{\text{is bounded}} \le C''
\end{aligned}$$

are Lipschitz constants.

We then obtain

$$\sup_{y\in\mathbb{R}^d}\left|\mathcal{N}(y|\mu,\sigma^2 I_d)-\mathcal{N}(y|\tilde{\mu},\tilde{\sigma}^2 I_d))\right| \lesssim \|\mu-\tilde{\mu}\|+|\sigma^2-\tilde{\sigma}^2|.$$

□

To prepare for the next theorem, we define the average empirical L_2 distance between two functions $f, g : [0,1] \to \mathbb{R}^d$ as follows

$$\|f-g\|_n = \left(\frac{1}{n}\sum_{i=1}^n \|f(i/n)-g(i/n)\|^2\right)^{1/2}. \tag{44}$$

Theorem 4.3: Convergence rates of parameters

Given the same conditions as in Theorem 3.1, for all $k \ge k_0$, there exist universal constants $c_1, c_2 > 0$ such that with at least probability $1 - c_1 n^{-c_2}$ we have

$$\|\widehat{f}_n^{(k)} - f^0\|_n^2 \le C\left(\frac{\rho\log n}{n}\right), \qquad |\widehat{\sigma}_{n,k}^2 - \sigma_0^2| \le C\left(\frac{\rho\log n}{n}\right)^{1/2}, \tag{45}$$

and

$$0 \le \overline{\mathcal{L}}_n(\widehat{f}_n^{(k)}) - \overline{\mathcal{L}}_n(f^0) \le C\rho\frac{\log(n)}{n}, \tag{46}$$

where $\rho = d(k+1)+1$ is the number of parameters of the model, and C only depends on $\mathcal{V}, \mathcal{A}$ and Ω (but not k, n and d).

Proof. The proof is divided into a few small steps.

Step 1. Bound the covering number of the space of changepoint models To apply Theorem 4.1 to provide the convergence of parameters and likelihood, we need to bound the covering number with bracketing for the space of changepoint models with k segments. We generalize the technique in [22] for our model.

Step 1.1. Covering the space of changepoint models with fixed changes under ℓ_∞ norm. Suppose that the set of changes $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{k-1}) \subset (0,1)$ is fixed. Given $\varepsilon > 0$, because $\mathcal{A}$ is compact in $\mathbb{R}^d$, we can find a set $\{\underline{a}_i\}_{i=1}^{N_1} \subset \mathcal{A}$ with $N_1 \asymp (1/\varepsilon)^d$ such that for every $\underline{a} \in \mathcal{A}$, there exists an $\underline{a}_i$ so that $\|\underline{a} - \underline{a}_i\|_\infty < \varepsilon$. We say that $\{\underline{a}_i\}_{i=1}^{N_1}$ is an ε-net of $\mathcal{A}$ under ℓ_∞ norm. Similarly, we can find an ε-net $(V_i)_{i=1}^{N_2}$ of $\mathcal{V}$ and $(\sigma_i^2)_{i=1}^{N_3}$ of Ω with the cardinality $N_2 \asymp (1/\varepsilon)^d$ and $N_3 \asymp (1/\varepsilon)$. Consider the net

$$B = \{(\underline{a}_{i_0}, V_{i_1}, \ldots, V_{i_k}, \sigma^2_{i_{k+1}}) : i_0 \in [N_1], i_1, \ldots, i_k \in [N_2], i_{k+1} \in [N_3]\} \subset \mathcal{A} \times \mathcal{V}^k \times \Omega.$$

We have that $|B| \asymp (1/\varepsilon)^\rho$ ($\rho = d(k+1)+1$), and for every tuple $(\underline{a}, \mathbf{V}, \sigma^2) \in \mathcal{A} \times \mathcal{V}^k \times \Omega$, there exists an element $(\tilde{\underline{a}}, \tilde{\mathbf{V}}, \tilde{\sigma}^2)$ in B that is $\varepsilon-$close to it under ℓ_∞ norm. For all $i \in [k]$ and $t \in [\tau_i, \tau_{i+1})$, we have

$$\left\| f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}(t) - f_{\boldsymbol{\tau},\tilde{\mathbf{V}},\tilde{\underline{a}}}(t) \right\|_\infty = \left\| (\underline{a} - \tilde{\underline{a}}) + \sum_{j=1}^{i} (V_j - \tilde{V}_j - (V_{j-1} - \tilde{V}_{j-1}))\tau_{j-1} + (V_i - \tilde{V}_i)t \right\|_\infty$$

$$\leq \|\underline{a} - \tilde{\underline{a}}\|_\infty + \left(\sum_{j=1}^{i} \|V_j - \tilde{V}_j\|_\infty + \sum_{j=1}^{i-1} \|V_j - \tilde{V}_j\|_\infty \right) \max_j \tau_{j-1} + \|V_i - \tilde{V}_{i-1}\|_\infty t \leq (2k+2)\varepsilon.$$

Hence,

$$\|f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}(t) - f_{\boldsymbol{\tau},\tilde{\mathbf{V}},\tilde{\underline{a}}}(t)\|_\infty \leq (2k+2)\varepsilon \quad \forall t \in [0,1].$$

Combining with Lemma 1, it implies

$$\sup_{y_i \in \mathbb{R}^d} |\mathcal{N}(y_i | f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}(t_i), \sigma^2 I) - \mathcal{N}(y_i | f_{\boldsymbol{\tau},\tilde{\mathbf{V}},\tilde{\underline{a}}}(t_i), \tilde{\sigma}^2 I)| \lesssim k\varepsilon, \quad \forall i \in [n]. \tag{47}$$

Step 1.2. Covering the space of changepoint models with fixed changes under Hellinger distance (with bracketing). For every $\delta > 0$, from the previous step, we have a collection of product normal densities $\{\mathbf{p}_j\}_{j=1}^N$ with $\mathbf{p}_j = (p_{j1}, \ldots, p_{jn})$ on $\mathcal{Y}^n$ and $N \asymp (k/\delta)^\rho$ such that for every tuple $(\underline{a}, \mathbf{V}, \sigma^2) \in \mathcal{A} \times \mathcal{V}^k \times \Omega$, there exists a $\mathbf{p}_j$ satisfying

$$\sup_{y_i \in \mathbb{R}^d} |\mathcal{N}(y_i | f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}(t_i), \sigma^2 I) - p_{ji}(y_i)| \leq \delta, \quad \forall i \in [n]. \tag{48}$$

Moreover, we have the mean and variance of p_{ji} are in a compact space $\mathcal{M} \subset \mathbb{R}^d$ and $\Omega = [\underline{\sigma}^2, \overline{\sigma}^2] \subset (0, \infty)$ for all $i \in [n], j \in [N]$. Hence, we can find an upper bound (envelop)

$$H(y) = \begin{cases} b_1 \exp(-b_2 \|y\|^2), & \|y\| \geq B, \\ (\sqrt{2\pi}\underline{\sigma}^2)^{-d}, & \text{otherwise} \end{cases} \tag{49}$$

of $p_{ji}(y)$ for all $j \in [1, N]$ and $i \in [1, n]$, for some constants $b_1, b_2, B > 0$. We can construct brackets $[\mathbf{p}_j^L, \mathbf{p}_j^U]$ with $\mathbf{p}_j^U = (p_{j1}^U, \ldots, p_{jn}^U)$ and $\mathbf{p}_j^L = (p_{j1}^L, \ldots, p_{jn}^L)$ as follows:

$$p_{ji}^L(y) = \max\{p_{ji}(y) - \delta, 0\},$$
$$p_{ji}^U(y) = \min\{p_{ji}(y) + \delta, H(y)\}.$$

With this construction, (48) implies

$$p_{ji}^L(y_i) \leq \mathcal{N}(y_i | f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}(t_i), \sigma^2 I) \leq p_{ji}^U(y_i) \quad \forall y_i \in \mathbb{R}^d, i \in [n]. \tag{50}$$

Hence, this collection of brackets satisfies condition (ii) in the definition of covering with bracketing (Definition 4.1). Now, we are checking condition (i). For any j, i and $\overline{B} \geq B$,

$$\begin{aligned}\int_{\mathbb{R}^d} (p_{ji}^U - p_{ji}^L) dy &\leq \int_{\|y\| \leq \overline{B}} 2\delta dy + \int_{\|y\| \geq \overline{B}} H(y) dy \\ &\lesssim \delta \overline{B}^d + \overline{B}^d \exp\left(-b_2 \overline{B}^2\right), \end{aligned} \tag{51}$$

where we use spherical coordinates to have

$$\int_{\|y\| \leq \overline{B}} dy = \frac{\pi^{d/2}}{\Gamma(d/2+1)} \overline{B}^d \lesssim \overline{B}^d,$$

and

$$\begin{aligned}\int_{\|y\| \geq \overline{B}} \exp\left(-b_2 \|y\|^2\right) dy &\lesssim \int_{r \geq \overline{B}} r^{d-1} \exp\left(-b_2 r^2\right) dr \\ &= \frac{1}{2b_2^{1/2}} \int_{\overline{B}^2}^{\infty} u^{d/2-1} \exp(-u) du \quad (\text{with } u = b_2 r^2) \\ &\leq \frac{1}{2b_2^{1/2}} \overline{B}^{d-2} \exp(-\overline{B}^2).\end{aligned}$$

Hence, choosing $\overline{B} = B(\log(1/\delta))^{1/2}$ in (51) gives

$$\int_{\mathbb{R}^d} (p_{ji}^U - p_{ji}^L) dy \lesssim \delta \left(\log\left(\frac{1}{\delta}\right)\right)^{d/2}. \tag{52}$$

Moreover, denote $p_i^0 = \mathcal{N}(y_i | f^0(t_i), \sigma_0^2)$ the density of y_i under the true model. Because $p_{ji}^U \geq p_{ji}^L$, we have

$$\begin{aligned}h^2\left(\frac{p_{ji}^U + p_i^0}{2}, \frac{p_{ji}^L + p_i^0}{2}\right) &= \int_{\mathbb{R}^d} \left(\sqrt{\frac{p_{ji}^U + p_i^0}{2}} - \sqrt{\frac{p_{ji}^L + p_i^0}{2}}\right)^2 dy \\ &\leq \int_{\mathbb{R}^d} \left(\frac{p_{ji}^U + p_i^0}{2} - \frac{p_{ji}^L + p_i^0}{2}\right) dy \\ &= \frac{1}{2} \int_{\mathbb{R}^d} (p_{ji}^U - p_{ji}^L) dy \\ &\lesssim \delta \left(\log\left(\frac{1}{\delta}\right)\right)^{d/2}.\end{aligned}$$

Therefore,

$$\overline{h}_n\left(\frac{\mathbf{p}_j^U + p_0^{(n)}}{2}, \frac{\mathbf{p}_j^L + p_0^{(n)}}{2}\right) \lesssim \delta^{1/2} (\log(1/\delta))^{d/4}.$$

Hence, there exists a positive constant c which does not depend on δ such that

$$H_B(c\delta^{1/2} \log(1/\delta)^{d/4}) \leq \log N \lesssim \rho \log(1/\delta).$$

Let $\epsilon = c\delta^{1/2}(\log(1/\delta))^{d/4}$, we have $\log(1/\epsilon) \asymp \log(1/\delta)$, which yields

$$H_B(\epsilon) \lesssim \rho \log(1/\epsilon),$$

for all ϵ sufficiently small.

Step 1.3. Aggregate changepoints Because there are $\binom{n}{k}$ ways to choose k changepoints among n data points, the covering number with bracketing of the whole model can be bounded as follows:

$$N_B(\epsilon) \lesssim \binom{n}{k}(1/\epsilon)^\rho \leq \left(\frac{n}{\epsilon}\right)^\rho.$$

Hence, the entropy number with bracketing of the whole model can be bounded as

$$H_B(\epsilon) \lesssim \rho \log\left(\frac{n}{\epsilon}\right).$$

In particular, there exists C_B that only depends on $\mathcal{V}, \mathcal{A}$ and Ω such that for n sufficiently large, $H_B(\epsilon) \leq C_B \rho \log(n/\epsilon)$.

Step 2. Convergence rate of parameter estimation This is a consequence of Theorem 4.1.

Since $\log(n/u)$ is a non-increasing function of u, we have

$$\begin{aligned} J_B(\varepsilon) &\leq \int_{\varepsilon^2/c_0}^{\varepsilon} (C_B \rho \log(n/u))^{1/2} du \vee \varepsilon \\ &\leq C_B^{1/2} \varepsilon \left(\rho \log \frac{n}{(\varepsilon^2/c_0)}\right)^{1/2} \\ &\leq C_B^{1/2} \varepsilon (\rho \log(n/\varepsilon))^{1/2}, \end{aligned}$$

for all ε small enough. Hence, for $\Psi(\varepsilon) = C_B^{1/2}\varepsilon(\rho \log(n/\varepsilon))^{1/2}$, we have $\Psi(\varepsilon)/\varepsilon^2$ is a non-increasing function, and let $\varepsilon_n = \max\{1, 2cC_B^{1/2}\}(\rho \log n/n)^{1/2}$ ($c > 0$ is a given universal constant), we have

$$c\Psi(\varepsilon_n) = cC_B^{1/2}\varepsilon_n(\rho \log(n/\varepsilon_n))^{1/2} \leq \varepsilon_n \times (2cC_B^{1/2}(\rho \log(n))^{1/2}) \leq \varepsilon_n^2\sqrt{n}.$$

Substitute $\varepsilon = \varepsilon_n$ to the conclusion of Theorem 4.1, we have

$$\begin{aligned} \mathbb{P}_0\left(\overline{h}_n\left(p^{(n)}_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\underline{\widehat{a}},\widehat{\sigma}^2}, p_0^{(n)}\right) \geq \max\{1, 2cC_B^{1/2}\}\left(\frac{\rho \log n}{n}\right)^{1/2}\right) &\leq c\exp\left(-\left(\max\{1, 2cC_B^{1/2}\}\right)^2 \rho \log(n)/c^2\right) \\ &\leq c_1 n^{-c_2}, (\text{since } \rho \geq 1 \text{ and} (\max\{1, 2cC_B^{1/2}\})^2 \geq 1) \end{aligned}$$

where C_B depends on $\mathcal{V}, \mathcal{A}$, and Ω only, and $c_1 = c$ and $c_2 = \dfrac{1}{c^2}$ are universal constants.

As a consequence of Lemma 1, we have that

$$\|\widehat{f}_n^{(k)} - f^0\|_n^2 + |\widehat{\sigma}_{n,k}^2 - \sigma_0^2|^2 \asymp \overline{h}_n^2\left(p^{(n)}_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\underline{\widehat{a}},\widehat{\sigma}^2}, p_0^{(n)}\right),$$

therefore,

$$\|\widehat{f}_n^{(k)} - f^0\|_n^2 \leq C\left(\frac{\rho \log n}{n}\right), \qquad |\widehat{\sigma}_{n,k}^2 - \sigma_0^2| \leq C\left(\frac{\rho \log n}{n}\right)^{1/2}, \quad \text{for } k \geq k_0 \tag{53}$$

where C depends on $\mathcal{V}, \mathcal{A}$, and Ω only, and c_1 and c_2 are universal constants.

Step 3. Convergence rate of likelihood functions We denote $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ and $P_0 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{Y_i\sim\mathcal{N}(f^0(t_i),\sigma_0^2)}$. We aim to apply Theorem 4.2 to show the convergence of the likelihood functions. All of the following arguments hold for $k \geq k_0$, and we work with $\frac{\mathcal{N}(y_i|f_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}}}(t_i),\widehat{\sigma}^2_{n,k}) + \mathcal{N}(y_i|f^0(t_i),\sigma_0^2)}{2\mathcal{N}(y_i|f^0(t_i),\sigma_0^2)}$ instead of $\frac{\mathcal{N}(y_i|f_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}}}(t_i),\widehat{\sigma}^2_{n,k})}{\mathcal{N}(y_i|f^0(t_i),\sigma_0^2)}$ because the former is always bounded below by $1/2$, but the latter is not.

By the concavity of the log function, we have

$$\frac{1}{2}\log\frac{\mathcal{N}(y_i|f_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}}}(t_i),\widehat{\sigma}^2_{n,k})}{\mathcal{N}(y_i|f^0(t_i),\sigma_0^2)} \leq \log\left(\frac{\mathcal{N}(y_i|f_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}}}(t_i),\widehat{\sigma}^2_{n,k}) + \mathcal{N}(y_i|f^0(t_i),\sigma_0^2)}{2\mathcal{N}(y_i|f^0(t_i),\sigma_0^2)}\right) \tag{54}$$

Recall that we have proved

$$\mathbb{P}_0\left(\overline{h}_n\left(p^{(n)}_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}},\widehat{\sigma}^2_{n,k}}, p_0^{(n)}\right) \leq C\left(\frac{\rho\log n}{n}\right)^{1/2}\right) \geq 1 - c_1 n^{-c_2},$$

for some constant $C > 0$. And for any density p we have that $h\left(\frac{p+p_0}{2}, p_0\right) \leq h(p,p_0) \leq 4h\left(\frac{p+p_0}{2}, p_0\right)$. Therefore,

$$\mathbb{P}_0\left(\overline{h}_n\left(\frac{p^{(n)}_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}},\widehat{\sigma}^2_{n,k}} + p_0^{(n)}}{2}, p_0^{(n)}\right) \leq C\left(\frac{\rho\log n}{n}\right)^{1/2}\right) \geq 1 - c_1 n^{-c_2},$$

Substitute $R = C\left(\frac{\rho\log n}{n}\right)^{1/2}, b = C\frac{\rho\log n}{n^{1/2}}$ in Theorem 4.2, we have $b \leq \sqrt{n}R^2 \leq \sqrt{n}R$, and

$$b \geq R\left(\log\left(\frac{2^6\sqrt{n}}{b}\right)\right)^{1/2} \geq \int_{b/(2^6\sqrt{n})}^R H_B^{1/2}\left(\frac{u}{\sqrt{2}}, \Theta\right) du \vee R.$$

Hence,

$$\mathbb{P}_0\left(\sup_{\overline{h}_n\left(\overline{p}^{(n)}_{\boldsymbol{\tau},\mathbf{V},\underline{a},\sigma}, p_0^{(n)}\right) \leq C\left(\frac{\rho\log n}{n}\right)^{1/2}} \left|\sqrt{n}(P_n - P_0)\log\left(\frac{\mathcal{N}(y_i|f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}(t_i),\widehat{\sigma}^2_{n,k}) + \mathcal{N}(y_i|f^0(t_i),\sigma_0^2)}{2\mathcal{N}(y_i|f^0(t_i),\sigma_0^2)}\right)\right| \geq C\frac{\rho\log n}{n^{1/2}}\right)$$
$$\leq c_1 n^{-c_2}, \tag{55}$$

for some universal constants c_1, c_2. Combining with the bound on the Hellinger distance,

$$\mathbb{P}_0\left(\left|(P_n - P_0)\log\left(\frac{\mathcal{N}(y_i|f_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}}}(t_i),\widehat{\sigma}^2_{n,k}) + \mathcal{N}(y_i|f^0(t_i),\sigma_0^2)}{2\mathcal{N}(y_i|f^0(t_i),\sigma_0^2)}\right)\right| \geq C\frac{\rho\log n}{n}\right) \leq 2c_1 n^{-c_2}, \tag{56}$$

Furthermore,

$$P_0\log\left(\frac{\mathcal{N}(y_i|f_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}}}(t_i),\widehat{\sigma}^2_{n,k}) + \mathcal{N}(y_i|f^0(t_i),\sigma_0^2)}{2\mathcal{N}(y_i|f^0(t_i),\sigma_0^2)}\right)$$
$$= -\frac{1}{n}\sum_{i=1}^n KL\left(\mathcal{N}(y_i|f^0(t_i),\sigma_0^2)||\frac{\mathcal{N}(y_i|f_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}}}(t_i),\widehat{\sigma}^2_{n,k}) + \mathcal{N}(y_i|f^0(t_i),\sigma_0^2)}{2}\right) \leq 0. \tag{57}$$

This implies that

$$\mathbb{P}_0\left(P_n \log\left(\frac{\mathcal{N}(y_i|f_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}}}(t_i),\widehat{\sigma}^2_{n,k})+\mathcal{N}(y_i|f^0(t_i),\sigma_0^2)}{2\mathcal{N}(y_i|f^0(t_i),\sigma_0^2)}\right)\leq C\frac{\rho\log n}{n}\right)\leq 1-2c_1n^{-c_2}. \quad (58)$$

Together with (54), we have that

$$\mathbb{P}_0\left(P_n \log\left(\frac{\mathcal{N}(y_i|f_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}}}(t_i),\widehat{\sigma}^2_{n,k})}{\mathcal{N}(y_i|f^0(t_i),\sigma_0^2)}\right)\leq C\frac{\rho\log n}{n}\right)\leq 1-2c_1n^{-c_2}. \quad (59)$$

In other words, for $k\geq k_0$, we have shown that $\overline{\mathcal{L}}_n(\widehat{f}_n^{(k)},\widehat{\sigma}^2_{n,k})-\overline{\mathcal{L}}_n(f^0,\widehat{\sigma}_0^2)\leq C\frac{\rho\log(n)}{n}$ for some constant C depends on $\mathcal{V},\mathcal{A}$ and Ω only with a high probability. The part $\overline{\mathcal{L}}_n(\widehat{f}_n^{(k)},\widehat{\sigma}^2_{n,k})-\overline{\mathcal{L}}_n(f^0,\widehat{\sigma}_0^2)\geq 0$ (for $k\geq k_0$) is done by the MLE property. □

D.4 Proof of consistency of sSIC

Proof of Theorem 3.1. Our estimate for the number of changepoints is is given by

$$\widehat{k}_n=\underset{k\leq\overline{k}}{\arg\max}\left[\overline{\mathcal{L}}_n(\widehat{f}_n^{(k)},\widehat{\sigma}^2_{n,k})-\rho\frac{(\log(n))^\gamma}{n}\right],$$

where $\rho=d(k+1)+1$ is the number of parameters of the model.

We aim to prove that

$$\mathbb{P}(\widehat{k}_n=k_0)\to 1$$

by first showing that $\mathbb{P}(\widehat{k}_n>k_0)\to 0$ and then $\mathbb{P}(\widehat{k}_n<k_0)\to 0$ as $n\to\infty$.

Asymptotically, sSIC does not over-fit. From Theorem 4.3, we have with probability tending to 1, for all $k\geq k_0$,

$$0\leq\overline{\mathcal{L}}_n(\widehat{f}_n^{(k)},\widehat{\sigma}^2_{n,k})-\overline{\mathcal{L}}_n(f^0,\sigma_0^2)\leq C\rho\frac{\log(n)}{n},$$

for some positive constant C depending on $\overline{k}$ and the parameter space. Therefore, for all $k>k_0$ (and $k<\overline{k}$), the increase in average log-likelihood when over-fitting can be characterized by

$$\begin{aligned}\overline{\mathcal{L}}_n(\widehat{f}_n^{(k)},\widehat{\sigma}^2_{n,k})-\overline{\mathcal{L}}_n(\widehat{f}_n^{(k_0)},\widehat{\sigma}^2_{n,k_0})&\leq\overline{\mathcal{L}}_n(\widehat{f}_n^{(k)},\widehat{\sigma}^2_{n,k})-\overline{\mathcal{L}}_n(f_0,\sigma_0^2)\\&\leq C\rho\frac{\log(n)}{n}=C[d(k+1)+1]\frac{\log(n)}{n}\\&<d(k-k_0)\frac{(\log(n))^\gamma}{n},\end{aligned}$$

for all n large enough, since $\gamma>1$ and C does not depend on k,d and n. Hence,

$$\overline{\mathcal{L}}_n(\widehat{f}_n^{(k_0)},\widehat{\sigma}^2_{n,k_0})-(d(k_0+1)+1)\frac{(\log(n))^\gamma}{n}>\overline{\mathcal{L}}_n(\widehat{f}_n^{(k)},\widehat{\sigma}^2_{n,k})-(d(k+1)+1)\frac{(\log(n))^\gamma}{n}\quad\forall k\in[k_0+1,\overline{k}],$$

for all sufficiently large n, implying that

$$\mathbb{P}(\widehat{k}_n>k_0)\to 0,$$

as $n\to\infty$.

Asymptotically, sSIC does not under-fit. Notice that

$$\widehat{k}_n = \arg\max_{k\leq \overline{k}} \left[\overline{\mathcal{L}}_n(\widehat{f}_n^{(k)}, \widehat{\sigma}^2_{n,k}) - \rho\frac{(\log(n))^\gamma}{n}\right] = \arg\max_{k\leq \overline{k}} \left[-\log\left(\hat{\sigma}^2_{n,k}\right) - \rho\frac{(\log(n))^\gamma}{n}\right].$$

In this part, it is more convenient to use an equivalent formula from (24) as

$$\widehat{k}_n = \arg\min_{k\leq \overline{k}} \left[\log\left(\frac{\sum_{i=1}^n \left\|y_i - \widehat{f}_n^{(k)}(t_i)\right\|^2}{dn}\right) + \rho\frac{(\log(n))^\gamma}{n}\right].$$

From Theorem 4.3, we have the convergence of RSS at $k = k_0$ as

$$\frac{\sum_{i=1}^n \left\|y_i - \widehat{f}_n^{(k_0)}(t_i)\right\|^2}{dn} \to \sigma_0^2$$

in probability. However, for all $k < k_0$, [41] (Lemma 5.4) showed that there exists a positive constant C depending on $\underline{C}_1$ and $\underline{C}_2$ such that

$$\frac{\sum_{i=1}^n \left\|y_i - \widehat{f}_n^{(k)}(t_i)\right\|^2}{dn} > \sigma_0^2 + C, \tag{60}$$

with probability tending to 1. The reason behind this inequality is that the under-fitted signal function $\widehat{f}_n^{(k)}$ (when $k < k_0$) always misses at least one true changepoint, i.e., there exists τ_r^0 (for $r \in [1, k_0 - 1]$) so that $\widehat{f}_n^{(k)}$ put no changepoint in $[\tau_r^0 - \underline{C}_1/4, \tau_r^0 + \underline{C}_1/4]$. As a consequence, the RSS in this segment is asymptotically greater than σ_0^2 ([41], Lemma 5.3). From this result, we have a positive constant C' depending on C and σ_0^2 so that for all $k < k_0$,

$$\log\left(\frac{\sum_{i=1}^n \left\|y_i - \widehat{f}_n^{(k)}(t_i)\right\|^2}{dn}\right) > \log\left(\frac{\sum_{i=1}^n \left\|y_i - \widehat{f}_n^{(k_0)}(t_i)\right\|^2}{dn}\right) + C',$$

with probability tending to 1. Since $C' > d(k_0 - k)\frac{(\log(n))^\gamma}{n}$ for all sufficiently large n, we have

$$\log\left(\frac{\sum_{i=1}^n \left\|y_i - \widehat{f}_n^{(k)}(t_i)\right\|^2}{dn}\right) + \rho\frac{(\log(n))^\gamma}{n} > \log\left(\frac{\sum_{i=1}^n \left\|y_i - \widehat{f}_n^{(k_0)}(t_i)\right\|^2}{dn}\right) + (d(k_0+1)+1)\frac{(\log(n))^\gamma}{n},$$

as $n \to \infty$, under those events. Hence,

$$\mathbb{P}(\widehat{k}_n < k_0) \to 0.$$

Combining with the previous part, we conclude that

$$\mathbb{P}(\widehat{k}_n = k_0) \to 1,$$

as $n \to \infty$.

Convergence of changepoints. We have

$$\mathbb{P}\left(\widehat{k}_n = k_0, \max_{i=1,\dots,k_0-1} |\widehat{\tau}_i^n - \tau_i^0| \le C\left(\frac{\log n}{n}\right)^{1/2}\right)$$
$$= \mathbb{P}\left(\max_{i=1,\dots,k_0-1} |\widehat{\tau}_i^n - \tau_i^0| \le C\left(\frac{\log n}{n}\right)^{1/2} \middle| \widehat{k}_n = k_0\right) \mathbb{P}(\widehat{k}_n = k_0).$$

Because we have shown that $\mathbb{P}(\widehat{k}_n = k_0) \to 1$, it suffices to prove

$$\mathbb{P}\left(\max_{i=1,\dots,k_0-1} |\widehat{\tau}_i^n - \tau_i^0| \le C\left(\frac{\log n}{n}\right)^{1/2} \middle| \widehat{k}_n = k_0\right) \to 1. \tag{61}$$

Recall from Theorem 4.3 that

$$\mathbb{P}\left(\|\widehat{f}_n^{(k_0)} - f^0\|_n^2 \le C\frac{\log n}{n}\right) \to 1.$$

Therefore, if we can show that the event $\{\|\widehat{f}_n^{(k_0)} - f^0\|_n^2 \le C\frac{\log n}{n}\}$ implies $\{|\widehat{\tau}_i^n - \tau_i^0| \le C\left(\frac{\log n}{n}\right)^{1/2} \ \forall i \in [k_0]\}$, then (61) will be proved.

We separate the proof into two steps, where we first show that $\widehat{\tau}_i \to \tau_i^0$ for all i (consistency), then prove the rate of convergence $(\log n/n)^{1/2}$. Figure 14 illustrates the two steps in this proof, where the first step aims to show the consistency of inferred changepoints converging to the true changepoints with rate $(\log n/n)^{1/3}$ using local information around the truth, and then improve the rate to $(\log n/n)^{1/2}$ by incorporating more information about the nearby true changepoints.

Before discussing the details of the two steps, we would like to point out two simple points that help clarify our argument.

- Consider two linear functions $\hat{f}(t) \in \mathbb{R}^d$ and $f_0(t) \in \mathbb{R}^d$ in an interval $t \in [t_1, t_2]$, let $\Delta\hat{f}(t) = \hat{f}(t) - f_0(t)$, then we have:

$$\int_{t_1}^{t_2} \left\|\hat{f}(t) - f_0(t)\right\|^2 dt = (t_2 - t_1)\left[\left\|\Delta\hat{f}(t_1)\right\|^2 + \langle\Delta\hat{f}(t_1), \Delta\hat{f}(t_2)\rangle + \left\|\Delta\hat{f}(t_2)\right\|^2\right]. \tag{62}$$

- Given a linear function $f(t) \in \mathbb{R}^d$ in an interval $[t_1, t_2]$, if $t_0 \in [t_1, t_2]$, we have that

$$f(t_0) = \frac{(t_0 - t_1)f(t_1) + (t_2 - t_0)f(t_2)}{t_2 - t_1}. \tag{63}$$

Step 1. Prove consistency $\widehat{\tau}_i^n \to \tau_i^0$ for all i. Given that

$$\|\widehat{f}_n^{(k_0)} - f^0\|_n^2 \le C\frac{\log n}{n}, \quad \forall n,$$

for some constant C, we first show that $\widehat{\tau}_i^n \to \tau_i^0$ for all $i \in [k_0]$ with a rate no slower than $(\log n/n)^{1/3}$. Note that this rate is not optimal, but it shows the consistency, and we will improve the convergence rate in Step 2.

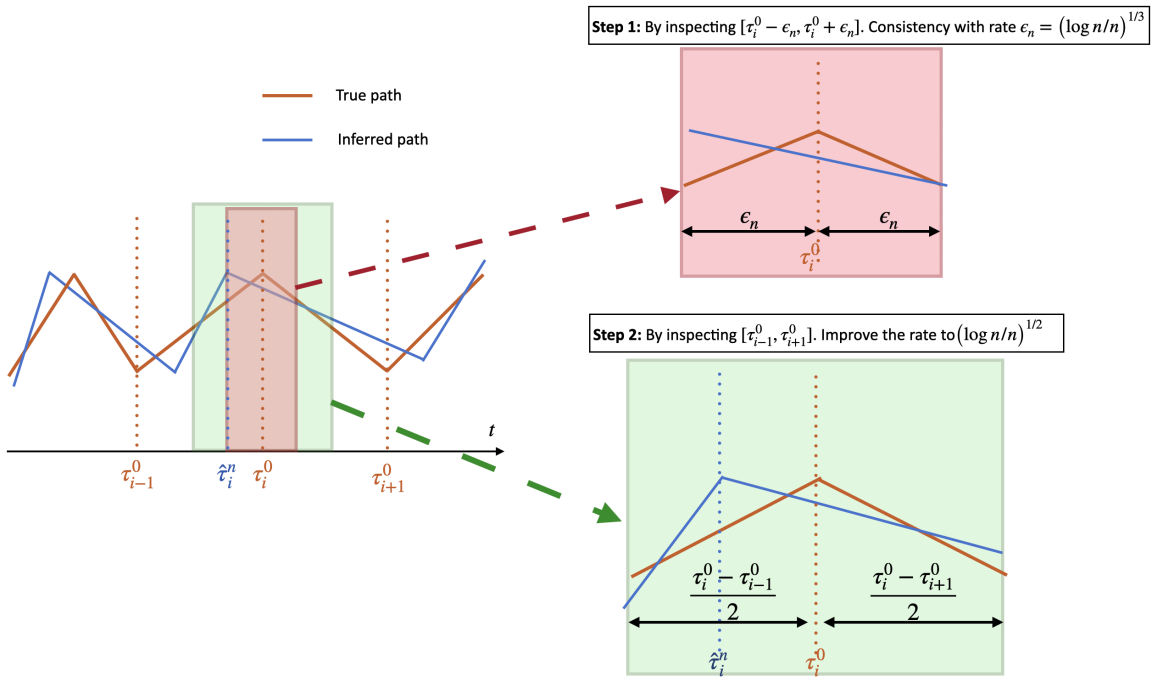


Figure 14: An illustration for the two steps in proving the convergence rate of changepoints.

Step 1.1. Setup. The main goal of Step 1 is to show

$$|\widehat{\tau}_i^n - \tau_i^0| \lesssim \left(\frac{\log n}{n}\right)^{1/3} \quad \forall i \in [k_0], \tag{64}$$

as $n \to \infty$, where the multiplicative constant in the inequality above does not depend on n. We proceed with a proof by contradiction. Assume that this limit does not hold, then there exists an index $i \in [k_0]$ and an increasing subsequence $(m_n)_{n=1}^{\infty} \subset \mathbb{N}$ such that $\widehat{\tau}_j^{m_n} \notin [\tau_i^0 - \epsilon_{m_n}, \tau_i^0 + \epsilon_{m_n}]$ for all $j \in [k_0]$, with $\epsilon_{m_n} \gg (\log m_n/m_n)^{1/3}$, i.e., $\lim_{n\to\infty} \dfrac{\epsilon_{m_n}}{(\log(m_n)/m_n)^{1/3}} \to \infty$ as $n \to \infty$. Without loss of generality, we can further assume that $\tau_{i-1}^0 < \tau_i^0 - \epsilon_{m_n} < \tau_i^0 + \epsilon_{m_n} < \tau_{i+1}^0$. Then,

$$\sum_{j=\lfloor m_n(\tau_i^0 - \epsilon_{m_n})\rfloor}^{\lfloor m_n(\tau_i^0+\epsilon_{m_n})\rfloor} \left\| \widehat{f}_{m_n}^{(k_0)}(j/m_n) - f^0(j/m_n) \right\|^2 \le m_n \left\| \widehat{f}_{m_n}^{(k_0)} - f^0 \right\|_{m_n}^2 \le C \log(m_n).$$

We are inspecting the LHS to show the contradiction. By the assumption, in the interval $[\lfloor m_n(\tau_i^0 - \epsilon_{m_n})\rfloor, \lfloor m_n(\tau_i^0 + \epsilon_{m_n})\rfloor]$, the function $\widehat{f}_{m_n}^{(k_0)}$ is a linear function, meanwhile f^0 is a piecewise linear function with 2 segments having slope V_{i-1}^0 and V_i^0, respectively. When n is large enough, the LHS can be approximated as

$$\begin{aligned} &\frac{1}{m_n} \sum_{j=\lfloor m_n(\tau_i^0 - \epsilon_{m_n})\rfloor}^{\lfloor m_n(\tau_i^0+\epsilon_{m_n})\rfloor} \left\| \widehat{f}_{m_n}^{(k_0)}(j/m_n) - f^0(j/m_n) \right\|^2 \\ &\qquad = \int_{\tau_i^0 - \epsilon_{m_n}}^{\tau_i^0+\epsilon_{m_n}} \left\| \widehat{f}_{m_n}^{(k_0)}(t) - f^0(t) \right\|^2 dt + o(1/m_n) \end{aligned} \tag{65}$$

Hence, we have

$$\int_{\tau_i^0-\epsilon_{m_n}}^{\tau_i^0+\epsilon_{m_n}} \left\| \widehat{f}_{m_n}^{(k_0)}(t) - f^0(t) \right\|^2 dt \leq C \frac{\log(m_n)}{m_n}. \tag{66}$$

Step 1.2. Finding the best linear approximation. Our aim now is to show that the inequality (66) can not hold since $\widehat{f}_{m_n}^{(k_0)}$ is a linear function while f^0 has two segments. Let $\tilde{f}$ be the best linear approximation of f^0 in the interval $[\tau_i^0 - \epsilon_{m_n}, \tau_i^0 + \epsilon_{m_n}]$. We then have that

$$\int_{\tau_i^0-\epsilon_{m_n}}^{\tau_i^0+\epsilon_{m_n}} \left\| \widehat{f}_{m_n}^{(k_0)}(t) - f^0(t) \right\|^2 dt \geq \int_{\tau_i^0-\epsilon_{m_n}}^{\tau_i^0+\epsilon_{m_n}} \left\| \tilde{f}(t) - f^0(t) \right\|^2 dt.$$

The optimal function $\tilde{f}$ and the expression on the RHS can be calculated exactly based on ϵ_{m_n} and $(V_i^0 - V_{i-1}^0)$, and we are going to find it in this step.

By definition, $\tilde{f}$ minimizes the following loss function

$$\text{Loss}(g) = \int_{\tau_i^0-\epsilon_{m_n}}^{\tau_i^0+\epsilon_{m_n}} \left\| g(t) - f^0(t) \right\|^2 dt,$$

over all linear function $g(t)$ in the interval $[\tau_i^0 - \epsilon_{m_n}, \tau_i^0 + \epsilon_{m_n}]$.

Let $\Delta g = g - f^0$. We can rewrite the loss function (as we mentioned earlier in (62)):

$$\begin{aligned}
\text{Loss}(g) &= \int_{\tau_i^0-\epsilon_{m_n}}^{\tau_i^0} \left\| g(t) - f^0(t) \right\|^2 dt + \int_{\tau_i^0}^{\tau_i^0+\epsilon_{m_n}} \left\| g(t) - f^0(t) \right\|^2 dt \\
&= \epsilon_{m_n} \left(\left\| \Delta g(\tau_i^0) \right\|^2 + \langle \Delta g(\tau_i^0), \Delta g(\tau_i^0 - \epsilon_{m_n}) \rangle + \left\| \Delta g(\tau_i^0 - \epsilon_{m_n}) \right\|^2 \right) \\
&\quad + \epsilon_{m_n} \left(\left\| \Delta g(\tau_i^0) \right\|^2 + \langle \Delta g(\tau_i^0), \Delta g(\tau_i^0 + \epsilon_{m_n}) \rangle + \left\| \Delta g(\tau_i^0 + \epsilon_{m_n}) \right\|^2 \right).
\end{aligned} \tag{67}$$

Notice that $\Delta g(\tau_i^0) = g(\tau_i^0) - f^0(\tau_i^0) = \dfrac{\epsilon_{m_n} g(\tau_i^0 - \epsilon_{m_n}) + \epsilon_{m_n} g(\tau_i^0 + \epsilon_{m_n})}{2\epsilon_{m_n}} - f^0(\tau_i^0)$ (as in (63)). Let $a = g(\tau_i^0 - \epsilon_{m_n})$ and $b = g(\tau_i^0 + \epsilon_{m_n})$, then $\Delta g(\tau_i^0) = \dfrac{a+b}{2} - f^0(\tau_i^0)$. Since the linear function g is completely identified by a and b, so is Loss(g). We can consider a and b as two "free" parameters here to optimize Loss, thereby, $\dfrac{\partial \Delta g(\tau_i^0)}{\partial a} = \dfrac{1}{2} I_d$, $\dfrac{\partial \Delta g(\tau_i^0)}{\partial b} = \dfrac{1}{2} I_d$, $\dfrac{\partial \left\| \Delta g(\tau_i^0) \right\|^2}{\partial a} = \Delta g(\tau_i^0)$, $\dfrac{\partial \left\| \Delta g(\tau_i^0) \right\|^2}{\partial b} = \Delta g(\tau_i^0)$. Set the derivatives of Loss with respect to those free parameters to 0, we have

$$0 \overset{\text{set}}{=} \frac{\partial \text{Loss}}{\partial a} = \epsilon_{m_n} \left(2\Delta g(\tau_i^0) + \frac{5}{2} \Delta g(\tau_i^0 - \epsilon_{m_n}) \right), \tag{68}$$

and

$$0 \overset{\text{set}}{=} \frac{\partial \text{Loss}}{\partial b} = \epsilon_{m_n} \left(2\Delta g(\tau_i^0) + \frac{5}{2} \Delta g(\tau_i^0 + \epsilon_{m_n}) \right), \tag{69}$$

From Equations (68) and (69), we have that $\Delta g(\tau_i^0 - \epsilon_{m_n}) = \Delta g(\tau_i^0 + \epsilon_{m_n})$. Since $\tilde{f}$ minimize the loss function, it satisfies $\Delta \tilde{f}(\tau_i^0 - \epsilon_{m_n}) = \Delta \tilde{f}(\tau_i^0 + \epsilon_{m_n})$ where $\Delta \tilde{f} = \tilde{f} - f^0$.

On the other hand, for every linear function $\tilde{f}$ with the slope vector $\tilde{V} \in \mathbb{R}^d$ in the considered interval, we always have:

$$\begin{aligned}\Delta\tilde{f}(\tau_i^0) - \Delta\tilde{f}(\tau_i^0 - \epsilon_{m_n}) &= \tilde{f}(\tau_i^0) - f^0(\tau_i^0) - \tilde{f}(\tau_i^0 - \epsilon_{m_n}) + f^0(\tau_i^0 - \epsilon_{m_n}) \\ &= \epsilon_{m_n}\tilde{V} - \epsilon_{m_n}V_{i-1}^0 = \epsilon_{m_n}(\tilde{V} - V_{i-1}^0) \qquad (70) \\ \Delta\tilde{f}(\tau_i^0 + \epsilon_{m_n}) - \Delta\tilde{f}(\tau_i^0) &= \tilde{f}(\tau_i^0 + \epsilon_{m_n}) - f^0(\tau_i^0 + \epsilon_{m_n}) - \tilde{f}(\tau_i^0) + f^0(\tau_i^0) \\ &= \epsilon_{m_n}\tilde{V} - \epsilon_{m_n}V_i^0 = \epsilon_{m_n}(\tilde{V} - V_i^0) = \epsilon_{m_n}(\tilde{V} - V_{i-1}^0) + \epsilon_{m_n}(V_{i-1}^0 - V_i^0). \\ &\qquad (71)\end{aligned}$$

These implies that $2\Delta\tilde{f}(\tau_i^0+\epsilon_{m_n}) - 2\Delta\tilde{f}(\tau_i^0) = \epsilon_{m_n}(V_{i-1}^0 - V_i^0)$. Combining with $2\Delta\tilde{f}(\tau_i^0) + \frac{5}{2}\Delta\tilde{f}(\tau_i^0 + \epsilon_{m_n}) = 0$ (Equation (69)), we get

$$\Delta\tilde{f}(\tau_i^0 + \epsilon_{m_n}) = \Delta\tilde{f}(\tau_i^0 - \epsilon_{m_n}) = \frac{2}{9}\epsilon_{m_n}(V_{i-1}^0 - V_i^0). \qquad (72)$$

Step 1.3. Derive the contradiction from the explicit expression of $\tilde{f}$. We have

$$\begin{aligned}\int_{\tau_i^0-\epsilon_{m_n}}^{\tau_i^0+\epsilon_{m_n}} \left\|\widehat{f}^{(k_0)}(t) - f^0(t)\right\|^2 dt &\geq \int_{\tau_i^0-\epsilon_{m_n}}^{\tau_i^0+\epsilon_{m_n}} \left\|\tilde{f}(t) - f^0(t)\right\|^2 dt \\ &\geq \int_{\tau_i^0}^{\tau_i^0+\epsilon_{m_n}} \left\|\tilde{f}(t) - f^0(t)\right\|^2 dt \\ &\overset{(62)}{=} \epsilon_{m_n}\left[\left\|\Delta\tilde{f}(\tau_i^0)\right\|^2 + \langle\Delta\tilde{f}(\tau_i^0), \Delta\tilde{f}(\tau_i^0+\epsilon_{m_n})\rangle + \left\|\Delta\tilde{f}(\tau_i^0+\epsilon_{m_n})\right\|^2\right] \\ &= \epsilon_{m_n}\left\|\Delta\tilde{f}(\tau_i^0) + \frac{1}{2}\Delta\tilde{f}(\tau_i^0+\epsilon_{m_n})\right\|^2 + \frac{3}{4}\epsilon_{m_n}\left\|\Delta\tilde{f}(\tau_i^0+\epsilon_{m_n})\right\|^2 \\ &\geq \frac{3}{4}\epsilon_{m_n}\left\|\Delta\tilde{f}(\tau_i^0+\epsilon_{m_n})\right\|^2 \\ &\overset{(72)}{=} \frac{1}{27}\epsilon_{m_n}^3\left\|V_{i-1}^0 - V_i^0\right\|^2 \overset{\text{Theorem 3.1}}{\geq} \frac{1}{27}\epsilon_{m_n}^3\underline{C}_2^2. \qquad (73)\end{aligned}$$

This implies $C\frac{\log m_n}{m_n} \geq \frac{1}{27}\epsilon_{m_n}^3\underline{C}_2^2$ which contradicts the assumption that $\epsilon_{m_n} \gg (\log m_n/m_n)^{1/3}$.

Hence, we finish showing

$$\mathbb{P}\left(\max_{i=1,\ldots,k_0-1}|\widehat{\tau}_i - \tau_i^0| \leq C\left(\frac{\log n}{n}\right)^{1/3}\middle|\widehat{k}_n = k_0\right) \to 1.$$

Step 2. Improve the rate to $(\log n/n)^{1/2}$. In the previous step, we established that for each true changepoint τ_i^0, there exists an estimator τ_i^n converging to it at rate $(\log n/n)^{1/3}$. This rate was obtained by exploiting only local information around τ_i^0, where the analysis relied on a linear approximation. In this step, we refine the argument: given the established consistency of τ_i^n, we can now enlarge the interval around τ_i^0 and employ a two-piecewise linear approximation. By incorporating this additional information, we can improve the convergence rate further.

Step 2.1. Setup. Consider any index $i \in \{1, \ldots, k_0\}$, there exists $t_1, t_2 \in (0,1)$ where $t_1 > \dfrac{\tau_{i-1}^0 + \tau_i^0}{2}$ and $t_2 < \dfrac{\tau_i^0 + \tau_{i+1}^0}{2}$. With this defined interval $[t_1, t_2]$ and given the rate we have proven in Step 1, we can ensure that f^0 and $\hat{f}$ are both piecewise linear functions with exactly two segments in this interval. WLOG, we assume that $\hat{\tau}_i^n \leq \tau_i^0$. Based on the rate proved in Step 1, it is clear to see that $\hat{\tau}_i^n \in [t_1, t_2]$.

We have that

$$\sum_{j=\lfloor nt_1 \rfloor}^{\lfloor nt_2 \rfloor} \left\| \widehat{f}_n^{(k_0)}(j/n) - f^0(j/n) \right\|^2 \leq n \left\| \widehat{f}_n^{(k_0)} - f^0 \right\|_n^2 \leq C \log(n).$$

When n is large enough, the LHS can be approximated as

$$\frac{1}{n} \sum_{j=\lfloor nt_1 \rfloor}^{\lfloor nt_2 \rfloor} \left\| \widehat{f}_n^{(k_0)}(j/n) - f^0(j/n) \right\|^2 = \int_{t_1}^{t_2} \left\| \hat{f}_n^{(k_0)}(t) - f^0(t) \right\|^2 dt + o(1/n). \tag{74}$$

Hence,

$$\int_{t_1}^{t_2} \left\| \tilde{f}(t) - f^0(t) \right\|^2 dt \leq \int_{t_1}^{t_2} \left\| \hat{f}_n^{(k_0)}(t) - f^0(t) \right\|^2 dt \leq C \frac{\log n}{n}, \tag{75}$$

where

$$\tilde{f} = \arg\min_g \text{Loss}(g), \quad \text{Loss}(g) = \int_{t_1}^{t_2} \left\| g(t) - f^0(t) \right\|^2 dt, \tag{76}$$

over the space of all piecewise functions g having two segments in $[t_1, t_2]$ and a changepoint at $\hat{\tau}_i^n$.

Step 2.2. Find the best approximation $\tilde{f}$. Similar to Step 1, we can also find an explicit solution for $\tilde{f}$. Let $\Delta g = g - f^0$. We can write the loss function as

$$\begin{aligned}
\text{Loss}(g) &= \int_{t_1}^{\widehat{\tau}_i^n} \left\| g(t) - f^0(t) \right\|^2 dt + \int_{\widehat{\tau}_i^n}^{\tau_i^0} \left\| g(t) - f^0(t) \right\|^2 dt + \int_{\tau_i^0}^{t_2} \left\| g(t) - f^0(t) \right\|^2 dt \\
&= (\widehat{\tau}_i^n - t_1) \left[\|\Delta g(t_1)\|^2 + \langle \Delta g(t_1), \Delta g(\widehat{\tau}_i^n) \rangle + \|\Delta g(\widehat{\tau}_i^n)\|^2 \right] \\
&+ (\tau_i^0 - \widehat{\tau}_i^n) \left[\|\Delta g(\widehat{\tau}_i^n)\|^2 + \langle \Delta g(\widehat{\tau}_i^n), \Delta g(\tau_i^0) \rangle + \left\|\Delta g(\tau_i^0)\right\|^2 \right] \\
&+ (t_2 - \tau_i^0) \left[\left\|\Delta g(\tau_i^0)\right\|^2 + \langle \Delta g(\tau_i^0), \Delta g(t_2) \rangle + \|\Delta g(t_2)\|^2 \right].
\end{aligned}$$

Because of its condition, g is completely identified by three values $g(t_1), g(\widehat{\tau}_i^n)$, and $g(t_2)$, which we will treat as parameters of the loss function to optimize over. Let $c = g(t_1), d = g(\widehat{\tau}_i^n), e = g(t_2) \in \mathbb{R}^d$. Note that $\Delta g(\tau_i^0)$ depends on those parameters as

$$\Delta g(\tau_i^0) = \frac{(\tau_i^0 - \widehat{\tau}_i^n) g(\widehat{\tau}_i^n) + (t_2 - \tau_i^0) g(t_2)}{t_2 - \widehat{\tau}_i^n} - f^0(\tau_i^0) = \frac{(\tau_i^0 - \widehat{\tau}_i^n) d + (t_2 - \tau_i^0) e}{t_2 - \widehat{\tau}_i^n} - f^0(\tau_i^0),$$

and

$$\frac{\partial (\Delta g(\tau_i^0))}{\partial d} = \frac{\tau_i^0 - \widehat{\tau}_i^n}{t_2 - \widehat{\tau}_i^n}, \quad \frac{\partial (\Delta g(\tau_i^0))}{\partial e} = \frac{t_2 - \tau_i^0}{t_2 - \widehat{\tau}_i^n},$$

$$\frac{\partial \left\|\Delta g(\tau_i^0)\right\|^2}{\partial d} = 2 \frac{\tau_i^0 - \widehat{\tau}_i^n}{t_2 - \widehat{\tau}_i^n} (\Delta g(\tau_i^0)), \quad \frac{\partial \left\|\Delta g(\tau_i^0)\right\|^2}{\partial e} = 2 \frac{t_2 - \tau_i^0}{t_2 - \widehat{\tau}_i^n} (\Delta g(\tau_i^0)).$$

Setting the derivatives of Loss with respect to c, d, e to 0, we have

$$0 \overset{\text{set}}{=} \frac{\partial \text{Loss}}{\partial c} = (\widehat{\tau}_i^n - t_1)[2(\Delta g(t_1)) + \Delta g(\widehat{\tau}_i^n)], \tag{77}$$

and

$$\begin{aligned}
0 \overset{\text{set}}{=} \frac{\partial \text{Loss}}{\partial d} &= (\widehat{\tau}_i^n - t_1)[\Delta g(t_1) + 2(\Delta g(\widehat{\tau}_i^n))] \\
&+ (\tau_i^0 - \widehat{\tau}_i^n)\left[\left(2\frac{\tau_i^0 - \widehat{\tau}_i^n}{t_2 - \widehat{\tau}_i^n} + 1\right)(\Delta g(\tau_i^0)) + \left(\frac{\tau_i^0 - \widehat{\tau}_i^n}{t_2 - \widehat{\tau}_i^n} + 2\right)(\Delta g(\widehat{\tau}_i^n))\right] \\
&+ (t_2 - \tau_i^0)\left[2\frac{\tau_i^0 - \widehat{\tau}_i^n}{t_2 - \widehat{\tau}_i^n}(\Delta g(\tau_i^0)) + \frac{\tau_i^0 - \widehat{\tau}_i^n}{t_2 - \widehat{\tau}_i^n}(\Delta g(t_2))\right] \\
&= \left[\frac{3}{2}(\widehat{\tau}_i^n - t_1) + \left(\frac{\tau_i^0 - \widehat{\tau}_i^n}{t_2 - \widehat{\tau}_i^n} + 2\right)(\tau_i^0 - \widehat{\tau}_i^n)\right](\Delta g(\widehat{\tau}_i^n)) + 3(\tau_i^0 - \widehat{\tau}_i^n)(\Delta g(\tau_i^0)) \\
&+ \frac{(t_2 - \tau_i^0)(\tau_i^0 - \widehat{\tau}_i^n)}{t_2 - \widehat{\tau}_i^n}(\Delta g(t_2)),
\end{aligned} \tag{78}$$

and

$$\begin{aligned}
0 \overset{\text{set}}{=} \frac{\partial \text{Loss}}{\partial e} &= (\tau_i^0 - \widehat{\tau}_i^n)\left[\frac{t_2 - \tau_i^0}{t_2 - \widehat{\tau}_i^n}(\Delta g(\widehat{\tau}_i^n)) + 2\frac{t_2 - \tau_i^0}{t_2 - \widehat{\tau}_i^n}(\Delta g(\tau_i^0))\right] \\
&+ (t_2 - \tau_i^0)\left[\left(2\frac{t_2 - \tau_i^0}{t_2 - \widehat{\tau}_i^n} + 1\right)(\Delta g(\tau_i^0)) + \left(\frac{t_2 - \tau_i^0}{t_2 - \widehat{\tau}_i^n} + 2\right)(\Delta g(t_2))\right] \\
&= \frac{(\tau_i^0 - \widehat{\tau}_i^n)(t_2 - \tau_i^0)}{t_2 - \widehat{\tau}_i^n}(\Delta g(\widehat{\tau}_i^n)) + 3(t_2 - \tau_i^0)(\Delta g(\tau_i^0)) \\
&+ (t_2 - \tau_i^0)\left(\frac{\tau_i^0 - \widehat{\tau}_i^n}{t_2 - \widehat{\tau}_i^n} + 2\right)(\Delta g(t_2)).
\end{aligned} \tag{79}$$

From Equation (78), we have

$$\left[\frac{3}{2}\frac{(\widehat{\tau}_i^n - t_1)}{\tau_i^0 - \widehat{\tau}_i^n} + \frac{\tau_i^0 - \widehat{\tau}_i^n}{t_2 - \widehat{\tau}_i^n} + 2\right](\Delta g(\widehat{\tau}_i^n)) + 3(\Delta g(\tau_i^0)) + \frac{t_2 - \tau_i^0}{t_2 - \widehat{\tau}_i^n}(\Delta g(t_2)) = 0. \tag{80}$$

Equation (79) implies

$$\frac{t_2 - \tau_i^0}{t_2 - \widehat{\tau}_i^n}(\Delta g(\widehat{\tau}_i^n)) + 3(\Delta g(\tau_i^0)) + \left(\frac{\tau_i^0 - \widehat{\tau}_i^n}{t_2 - \widehat{\tau}_i^n} + 2\right)(\Delta g(t_2)) = 0. \tag{81}$$

Add them up, we have

$$\left[\frac{3}{2}\frac{(\widehat{\tau}_i^n - t_1)}{\tau_i^0 - \widehat{\tau}_i^n} + 3\right](\Delta g(\widehat{\tau}_i^n)) + 6(\Delta g(\tau_i^0)) + 3(\Delta g(t_2)) = 0. \tag{82}$$

Notice that

$$\begin{aligned}
\Delta g(\widehat{\tau}_i^n) &= \Delta g(t_1) + (t_1 - \widehat{\tau}_i^n)(V_{i-1} - V_{i-1}^0), \\
\Delta g(\tau_i^0) &= \Delta g(\widehat{\tau}_i^n) + (\tau_i^0 - \widehat{\tau}_i^n)(V_i - V_i^0) + (\tau_i^0 - \widehat{\tau}_i^n)(V_i^0 - V_{i-1}^0), \\
\Delta g(t_2) &= \Delta g(\tau_i^0) + (t_2 - \tau_i^0)(V_i - V_i^0).
\end{aligned}$$

From these three equations, we have

$$\begin{aligned}
\Delta g(\widehat{\tau}_i^n) &= \Delta g(t_1) + (t_1 - \hat{\tau}_i^n)(V_{i-1} - V_{i-1}^0),\\
\Delta g(\tau_i^0) &= \Delta g(t_1) + (t_1 - \hat{\tau}_i^n)(V_{i-1} - V_{i-1}^0) + (\tau_i^0 - \widehat{\tau}_i^n)(V_i - V_i^0) + (\tau_i^0 - \widehat{\tau}_i^n)(V_i^0 - V_{i-1}^0),\\
\Delta g(t_2) &= \Delta g(t_1) + (t_1 - \hat{\tau}_i^n)(V_{i-1} - V_{i-1}^0) + (t_2 - \hat{\tau}_i^n)(V_i - V_i^0) + (\tau_i^0 - \hat{\tau}_i^n)(V_i^0 - V_{i-1}^0).
\end{aligned}$$

$\tilde{f}$ satisfies all of the above equations. Based on these expressions, we rewrite the equations (77), (81) and (82) (replacing g by $\tilde{f}$) to be a system of three equations with three variables $\Delta\tilde{f}(t_1)$, $(t_1 - \hat{\tau}_i^n)(\tilde{V}_{i-1} - V_{i-1}^0)$, and $\tilde{V}_i - V_i^0$ as following

$$\begin{aligned}
&3\Delta\tilde{f}(t_1) + (t_1 - \hat{\tau}_i^n)(\tilde{V}_{i-1} - V_{i-1}^0) = 0,\\
&\frac{4t_2 - \tau_i^0 - \hat{\tau}_i^n}{t_2 - \hat{\tau}_i^n}\Delta\tilde{f}(t_1) + \frac{4t_2 - \tau_i^0 - \hat{\tau}_i^n}{t_2 - \hat{\tau}_i^n}(t_1 - \hat{\tau}_i^n)(\tilde{V}_{i-1} - V_{i-1}^0) + 3(\tau_i^0 - \hat{\tau}_i^n)(\tilde{V}_i - V_i^0) = 3(\tau_i^0 - \hat{\tau}_i^n)(V_{i-1}^0 - V_i^0),\\
&\frac{12\tau_i^0 - 3t_1 - 9\hat{\tau}_i^n}{2(\tau_i^0 - \hat{\tau}_i^n)}\Delta\tilde{f}(t_1) + \frac{12\tau_i^0 - 3t_1 - 9\hat{\tau}_i^n}{2(\tau_i^0 - \hat{\tau}_i^n)}(t_1 - \hat{\tau}_i^n)(\tilde{V}_{i-1} - V_{i-1}^0) + (6\tau_i^0 - 9\hat{\tau}_i^0 + 3t_2)(\tilde{V}_i - V_i^0) = 7(\tau_i^0 - \hat{\tau}_i^n)(V_{i-1}^0 - V_i^0).
\end{aligned}$$

Using the Gaussian Elimination method, we arrive at

$$\begin{aligned}
&\Delta\tilde{f}(t_1) + \frac{1}{3}(t_1 - \hat{\tau}_i^n)(\tilde{V}_{i-1} - V_{i-1}^0) = 0,\\
&(t_1 - \hat{\tau}_i^n)(\tilde{V}_{i-1} - V_{i-1}^0) + B_n(\tau_i^0 - \hat{\tau}_i^n)(\tilde{V}_i - V_i^0) = B_n(\tau_i^0 - \hat{\tau}_i^n)(V_{i-1}^0 - V_i^0),\\
&(\tilde{V}_i - V_i^0) = C_n(\tau_i^0 - \hat{\tau}_i^n)(V_{i-1}^0 - V_i^0).
\end{aligned}$$

where

$$B_n = \frac{9(t_2 - \hat{\tau}_i^n)}{2(4t_2 - \tau_i^0 - \hat{\tau}_i^n)},$$

and

$$C_n = \frac{14(4t_2 - \tau_i^0 - \hat{\tau}_i^n)}{2(6\tau_i^0 - 9\hat{\tau}_i^n + 3t_2)(4t_2 - \tau_i^0 - \hat{\tau}_i^n) - 9(4\tau_i^0 - t_1 - 3\hat{\tau}_i^n)(t_2 - \hat{\tau}_i^n)}.$$

Because $\lim_{n\to\infty}\widehat{\tau}_i^n = \tau_i^0$, we have

$$B_n \to B_0 = \frac{9(t_2 - \tau_i^0)}{4(2t_2 - \tau_i^0)} \in (0, \infty),$$

and

$$C_n \to C_0 = \frac{28(2t_2 - \tau_i^0)}{24t_2 - 21\tau_i^0 + 9t_1} \in (0, \infty),$$

which are two positive constants. Hence,

$$(\tilde{V}_i - V_i^0) \asymp C_0(\tau_i^0 - \hat{\tau}_i^n)(V_{i-1}^0 - V_i^0), \tag{83}$$

$$(t_1 - \hat{\tau}_i^0)(\tilde{V}_{i-1} - V_{i-1}^0) \asymp B_0(\tau_i^0 - \hat{\tau}_i^n)(V_{i-1}^0 - V_i^0) \tag{84}$$

Plugging in the definition of $\Delta\tilde{f}(t_2)$ (and notice that $\Delta\tilde{f}(t_1) + \frac{1}{3}(t_1 - \hat{\tau}_i^n)(\tilde{V}_{i-1} - V_{i-1}^0) = 0$), we have

$$\Delta\tilde{f}(t_2) \asymp \underbrace{\left(\frac{2}{3}B_0 + (t_2 - \tau_i^0)C_0 - 1\right)}_{E}(\tau_i^0 - \hat{\tau}_i^n)(V_{i-1}^0 - V_i^0) \asymp (\tau_i^0 - \hat{\tau}_i^n)(V_{i-1}^0 - V_i^0). \tag{85}$$

Note: (85) holds as long as $E \neq 0$, which is equivalent to

$$t_1 \neq \frac{224t_2^3 - 448t_2^2\tau_i^0 - 24t_2^2 + 280t_2(\tau_i^0)^2 - 3t_2\tau_i^0 - 56(\tau_i^0)^3 + 21(\tau_i^0)^2}{9(t_2 - \tau_i^0)}.$$

Therefore, generically the coefficient is nonzero; it only vanished for the special choice of t_1 displayed above.

Step 2.3. Derive the estimation rate from the explicit expression of $\tilde{f}$. We have that

$$C\frac{\log n}{n} \geq \int_{t_1}^{t_2} \left\|\Delta\tilde{f}(t)\right\|^2 dt \geq \int_{\tau_i^0}^{t_2} \left\|\Delta\tilde{f}(t)\right\|^2 \tag{86}$$

$$\begin{aligned}
&\overset{(62)}{=} (t_2 - \tau_i^0)\left[||\Delta\tilde{f}(\tau_i^0)||^2 + \langle\Delta\tilde{f}(\tau_i^0), \Delta\tilde{f}(t_2)\rangle + ||\Delta\tilde{f}(t_2)||^2\right] \\
&= (t_2 - \tau_i^0)\left[\left\|\Delta\tilde{f}(\tau_i^0) + \frac{1}{2}\Delta\tilde{f}(t_2)\right\|^2 + \frac{3}{4}||\Delta\tilde{f}(t_2)||^2\right] \\
&\geq (t_2 - \tau_i^0)\frac{3}{4}\left\|\Delta\tilde{f}(t_2)\right\|^2 \\
&\overset{(85)}{\gtrsim} (\tau_i^0 - \hat{\tau}_i^n)^2||V_{i-1}^0 - V_i^0||^2 > (\tau_i^0 - \hat{\tau}_i^n)^2\underline{C}_2^2, \quad \text{for } i \in [k_0 - 1].
\end{aligned} \tag{87}$$

This implies that

$$\mathbb{P}\left(\max_{i=1,\ldots,k_0-1} |\widehat{\tau}_i^n - \tau_i^0| \leq C\left(\frac{\log n}{n}\right)^{1/2} \middle| \widehat{k}_n = k_0\right) \to 1.$$

□