Changepoint detection: theory, algorithms, and application in intracellular transport

An Abstract

Submitted On the Twenty-first day of April, 2025

To The Department of Mathematics

In Partial Fulfillment Of The Requirements

Of The School of Sciences and Engineering

Of Tulane University

For The Degree Of

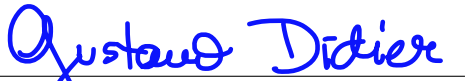
Doctor of Philosophy

By

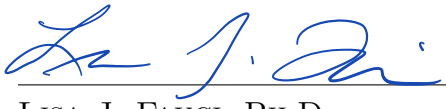
Thuy Linh Do

Approved: 

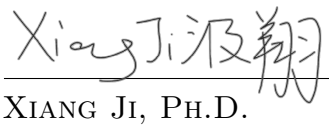
Scott A. McKinley, Ph.D.
Chairman



Gustavo Didier, Ph.D.



Lisa J. Fauci, Ph.D.



Xiang Ji, Ph.D.

Christine K. Payne, Ph.D.

Abstract

Changepoint detection plays a vital role in statistical analysis, with applications in biological data and beyond. In this thesis, we explore methods for detecting changes in mean, variance, and velocity, addressing challenges in computational efficiency and statistical reliability.

Motivated by intracellular transport with the challenge of detecting velocity changes in multidimensional data, we first introduce CPLASS, an algorithm using a Markov Chain Monte Carlo (MCMC)-based approach with a special proposal for navigating the parameter space. With the mathematical proving techniques from the Empirical Process theory, we show the consistency and convergence rates in estimating parameters using our proposed method. To deal with the small sample sizes that are common in molecular motor data, we introduce a speed penalty that improves small sample size power and performance while not compromising the large sample consistency. In the study of lysosomal transport, a statistical test for stationary states is proposed using a piecewise linear continuous model. Its effectiveness is examined under different conditions.

A second algorithm introduced is the Dendrogram Pruning and Merging (DPM) algorithm, an agglomeration approach that reduces computational cost by constructing a hierarchical structure of changepoint locations. This algorithm is developed to detect changes in mean and variance. We also propose a new selection criterion for model selection based on the proposed pruning and merging procedure, called the Dendrogram Selection Criterion (DSC). Finally, we establish theoretical guarantees for the DPM algorithm, proving its convergence and consistency. Through simulations and real-world applications, we evaluate the strengths and limitations of our methods, aiming to contribute useful tools for changepoint detection research.

Changepoint detection: theory, algorithms, and application in intracellular transport

A Dissertation

Submitted On the Twenty-first day of April, 2025

To The Department of Mathematics

In Partial Fulfillment Of The Requirements

Of The School of Sciences and Engineering

Of Tulane University

For The Degree Of

Doctor of Philosophy

By

Thuy Linh Do

Approved:

Scott A. McKinley, Ph.D.
Chairman

Gustavo Didier, Ph.D.

Lisa J. Fauci, Ph.D.

Xiang Ji, Ph.D.

Christine K. Payne, Ph.D.

© Copyright by Thuy Linh Do, 2025

All Rights Reserved

Acknowledgments

As I reach the final pages of this thesis, I find myself reflecting on the journey that brought me here—a journey filled with curiosity, challenges, and the unwavering support of many people to whom I owe my deepest gratitude.

First and foremost, I am profoundly grateful to my advisor, Prof. Scott A. McKinley, whose wisdom, patience, and encouragement have guided me through this research. He is the warmest person I have ever met—someone who not only teaches me how to ask the necessary questions to further my research but also how to balance life and work. His unwavering support and kindness have made this journey not just intellectually rewarding but personally enriching. Words cannot fully express my gratitude to him. I feel incredibly fortunate to have had him as my advisor.

I would also like to extend my deepest gratitude to Prof. XuanLong Nguyen from the Department of Statistics at the University of Michigan, as well as my committee members, Prof. Gustavo Didier, Prof. Lisa Fauci, Prof. Xiang Ji, and Prof. Christine K. Payne, for their invaluable insights and thoughtful feedback. Their guidance has pushed me to refine my ideas and has helped shape this work into what it is today.

I am sincerely grateful to the Department of Mathematics at Tulane University for providing the resources and environment that fostered my growth as a researcher. A special thank you to Prof. Tai Ha, the Department Chair, for his tireless care and support in ensuring we have the best resources for our research. I also deeply appreciate my colleagues and friends—John, Irene, Lan, Yuwei, Will, and Vinh—for all the laughter and good times. Whether through late-night discussions or simply sharing the joys and struggles of academic life, your companionship has meant the world to me.

I would like to express my heartfelt gratitude to the exceptional teachers and advisors who inspired me and guided me through my initial steps in the journey of studying mathematics and statistics. My thanks go to Prof. Hung Thang Dang, Prof. Minh Ha Le, Prof. Quoc Anh

Trinh, Prof. Thac Dung Nguyen, Prof. Duc Tai Pho, and Prof. Phuong Bac Dao from Hanoi University of Sciences, as well as Prof. Phu Hoang Lan Nguyen, Prof. Thi Hoa Mai Dao from the University of Education.

To my collaborator, best friend, husband, and the love of my life, Dat Do—thank you for sharing with me all the joys, difficulties, and challenges along this journey. Your unwavering support, patience, and encouragement have been my greatest source of strength. I am grateful for the stimulating discussions and shared enthusiasm that made our joint work both exciting and fulfilling. But more than that, thank you for always believing in me, for walking this journey by my side, and for making every moment—both in research and in life—so much more meaningful.

And to our little sweet baby boy, Daniel—you may not yet understand how much inspiration and motivation you have given me throughout this journey. Every hug, every tiny laugh, and every moment with you fills my heart with endless love. One day, when you are old enough to read this, I want you to know how deeply I love you, my sweetheart. You are and always will be my greatest joy.

I dedicate this thesis to the most important woman in my life—my mom, Phuong Thanh. Her unwavering support and tireless encouragement have shaped every step of my journey. I would not be here today, writing these words and presenting this work, without her endless love and sacrifices. She has given everything to ensure her children grow up healthy, happy, and full of possibilities. For that, and for so much more, I am forever grateful.

To the memory of my late father, Xuan Tuc—I believe that somewhere in the sky, he is watching his little girl grow up, surrounded by love and support.

To my brother, Viet Cuong—my role model—thank you for sharing every moment of my life and for your quiet, unwavering support. Your presence has been a constant source of strength and inspiration, even in moments when words were not needed.

Last but not least, I want to express my heartfelt thanks to my parents-in-law, Thanh Giang and Van Thi, for their kindness, generosity, and belief in me. I am so lucky to have you all in my life.

This journey has been one of learning, growth, and resilience, and I am deeply grateful to everyone who has been a part of it. This thesis may have my name on it, but it is a testament to the collective wisdom, kindness, and generosity of those around me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Changepoint detection is a fundamental problem in statistical analysis, with wide-ranging applications in fields such as finance [4, 43, 81], genomics [49, 103], and signal processing [57, 70, 81]. In this thesis, the primary motivation stems from a specific and biologically significant domain: intracellular transport. This process, where molecular motors move along cytoskeletal filaments to transport cargo within cells, is inherently stochastic and often exhibits abrupt transitions in motion patterns, such as changes in velocity, direction, or diffusive behavior. Understanding these transitions is essential for characterizing underlying biophysical mechanisms and pathological conditions.

Detecting such changes presents unique challenges that go beyond traditional changepoint problems. Unlike typical formulations that focus on mean shifts in time series, intracellular transport data are multidimensional, continuous, and governed by physical constraints such as positional continuity. These complexities demand new statistical methodologies that can handle parameter dependencies, small-sample robustness, and high-dimensional dynamics.

In this thesis, we address these challenges through a series of methodological and theoretical advancements in changepoint detection. Each contribution is directly inspired by open problems in intracellular transport, while also offering tools applicable to a broader class of changepoint problems. The work spans novel algorithm design,

statistical testing, model selection, and theoretical analysis.

1.1 Overview of the Thesis

This dissertation presents a unified framework for changepoint detection, with a central focus on applications to intracellular transport. The core contributions include:

- A specialized Markov Chain Monte Carlo (MCMC)-based algorithm for detecting velocity changes in multidimensional data, tailored to molecular motor trajectory analysis.
- Asymptotic theory for changepoint estimators, offering insights into consistency and convergence.
- A statistical test for identifying stationary states in particle motion, grounded in continuous piecewise linear models.
- A hierarchical segmentation approach—Dendrogram Pruning and Merging (DPM)—for efficient detection of multiple changepoints in mean and variance.
- Theoretical guarantees for DPM, including convergence rates derived via empirical process theory.

These contributions are further elaborated across the chapters of the thesis, as summarized below.

1.2 Summary of Chapters

Chapter 2: Detecting Velocity Changes in Multidimensional Data. The first methodological contribution of this thesis is the development of CPLASS, an algorithm designed for detecting changes in velocity within multidimensional data. Traditional changepoint detection methods struggle with such problems due to the dependencies

between adjacent segments and the need for continuity constraints. CPLASS overcomes these challenges by utilizing a Markov Chain Monte Carlo (MCMC)-based approach with tailored proposal mechanisms, improving efficiency in exploring parameter space. This method is particularly well-suited for analyzing intracellular transport data, where molecular motor trajectories undergo complex, multidimensional transitions. Furthermore, a speed penalty is introduced to enhance robustness in small-sample settings, ensuring statistical power while maintaining consistency as sample size increases.

Chapter 3: Asymptotic Properties of Changepoint Estimation. To further understand the statistical properties of the changepoint detection method, this chapter examines the asymptotic behavior of changepoint estimators obtained by maximizing the criterion function with the strengthened Schwarz Information Criterion (sSIC) penalty. This analysis provides theoretical insights into the consistency and convergence rates of the proposed algorithms, offering a deeper understanding of their reliability in large-sample settings. Chapters 2 and 3 will be turned into a manuscript which is co-authored by Dat Do (University of Michigan), Keisha J. Cook (Clemson University), Nathan Rayens, Christine K. Payne (Duke University), and Scott A. McKinley.

Chapter 4: Testing for Stationary States in Intracellular Transport. Intracellular transport is governed by complex interactions between molecular motors, cytoskeletal structures, and cytosolic crowding. In this chapter, we establish a statistical test to determine whether a particle trajectory exhibits stationary behavior as defined in a mathematical sense. Unlike previous studies that define stationarity based on speed thresholds, our approach leverages a continuous piecewise linear model to provide a more rigorous characterization. The test is developed specifically for two-dimensional datasets, but the methodology can be extended to higher dimensions. Additionally, we explore how the presence of anchor diffusion, caused by a secondary

noise source, affects the performance of our test. This chapter will be turned into a manuscript, which is co-authored by Scott A. McKinley.

Chapter 5: Dendrogram Pruning and Merging (DPM) for Multiple Changepoint Detection. This chapter introduces the Dendrogram Pruning and Merging (DPM) algorithm, a novel method for detecting multiple changepoints in mean and variance. DPM constructs a hierarchical structure (dendrogram) of changepoint locations, requires fitting a model only once at an overfitted level, and employs an optimal rule to prune and merge, significantly reducing computational costs. We also present the Dendrogram Information Criterion (DIC), a new model selection criterion that incorporates segment-wise parameter distances and lengths. Through extensive simulations, we demonstrate that DIC outperforms existing criteria in certain scenarios by providing greater robustness and interpretability.

Chapter 6: Theoretical Guarantees for the DPM Algorithm. The final chapter establishes the theoretical foundations of the DPM algorithm by proving its convergence rates and consistency properties. Using empirical process theory—an approach not previously applied in this context—we provide rigorous proofs that validate the effectiveness of our method. These results contribute to the broader literature on changepoint detection by offering new insights into the theoretical guarantees of hierarchical segmentation approaches. Chapters 5 and 6 will be turned into a manuscript, which is co-authored by Dat Do (University of Michigan) and Scott A. McKinley.

Chapter 7: Conclusion and future investigation.

Chapter 2

Detecting Velocity Changes in Multidimensional Data

This chapter introduces CPLASS, an algorithm for detecting changes in velocity within multidimensional data, addressing fundamental challenges in probability structure and search methodology. This problem in one dimension is known as the change in slope problem (see [9, 39]). Unlike traditional changepoint detection methods that focus on mean shifts, detecting changes in velocity requires specialized approaches due to continuity constraints and parameter dependencies. Existing algorithms, including binary segmentation and simple dynamic programming methods, struggle with these complexities. To overcome these limitations, we introduce a tailored 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 molecular motor trajectories exhibit complex, multidimensional transitions. To enhance robustness in small-sample regimes, we introduce a speed penalty that improves statistical power while maintaining consistency in the large-sample limit. Additionally, we demonstrate that comparing the proportion of time spent at different speeds provides a more stable metric for trajectory characterization.

This work is a collaboration with my advisor, Scott A. McKinley from the Department of Mathematics at Tulane University, Dat Do from the Department of Statistics at the University of Michigan, Keisha J. Cook from the School of Mathematical and Statistical Sciences at Clemson University, and Nathan Rayens and Christine K. Payne from the Department of Mechanical Engineering and Materials Science at Duke University.

2.1 Introduction

Change point detection problems have been researched for over sixty years with applications in many fields, such as signal processing [57, 70, 81], speech processing [3, 56, 122], financial analysis [4, 43, 81], bio-informatics and genomics [49, 103], environmental science [76], and many others. Consequently, this subject is deeply studied in theoretical and applied literature. Various methods and corresponding theories have been developed for the changes in mean problems, including Bayesian and frequentist approaches. Our paper focuses on detecting changes in velocity: the multidimensional data is a discretization of the velocity integral, which imposes a constraint on what would be a change-in-mean problem for the position increments, since a physical position process should be continuous in time. We model the data as a sequence of independent Gaussian fluctuations about a fictional anchor whose trajectory is continuous and piecewise-linear. The goal of the change point method is to identify the most likely trajectory if such an anchor existed. We propose the CPLASS algorithm (Continuous Piecewise Linear Approximation using Stochastic Searching methods) for this problem. In the following, we only focus on a retrospective change point framework closely related to our work. For recent reviews of change point methods, we refer to Aminikhanghahi and Cook (2017) [1], Truong, Oudre, and Vayatis (2020) [128], Fearnhead and Rigaill (2020) [41], and Shi, Gallagher, Lund, and

Killick (2022) [123].

Related work. Most change point detection schemes suggest changes and their positions and then assess a likelihood score using a probability model. However, there is a possibility of overfitting the model, since proposing more change points results in a better-fit model. Therefore, a penalty in the score for the number of change points proposed is necessary to avoid overfitting. 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) [139, 141]; ℓ_1 penalty such as Lasso, fused Lasso, elastic net, the group Lasso, the mono Lasso [55, 134]; more complex penalties such as the modified BIC criterion (mBIC) [144] which maximizing the asymptotic posterior probability of the data. Regarding finding optimal solutions for detecting change point problems with a given criterion function, some popular searching methods include binary segmentation [120] and its variants - Circular Binary Segmentation [104, 133], Wild Binary Segmentation [44], bottom-up [23], window sliding [19], Pruned Exact Linear Time (PELT) [77], dynamic programming [68] and stochastic optimization [53, 79, 80]. Another well-known procedure for detecting change points is based on a likelihood ratio test. In this procedure, a hypothesis test is first constructed for a single change point model, and then the test is applied to find multiple change points via searching algorithms [6–8].

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 score 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 the biological knowledge of researchers. These methods will not restrict the searching parameter space. It allows for exploring the space with some prior constraints (prior distributions of parameters) provided at the initial step. The

first Bayesian method for detecting multiple change points with an unknown number was provided by Barry and Hartigan [10, 11]. They construct a model called the Product Partition Model (PPM) that performs well in detecting sharp, short-lived changes in the means of independently normally distributed observations. This model has been implemented as an `R` package `bcp` by Erdman et al. [38].

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 [39]. For example, the most common generic approach to detecting multiple change points—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 change point locations might lie between the true change points, making it difficult for binary segmentation to correct these errors. Figure 2.1 illustrates an example of this issue. This problem has also been reported by Baranowski, Chen, and Fryzlewicz [9] and Fearnhead, Maidstone, and Letchford [39]. An alternative method involves a dynamic programming algorithm that minimizes an ℓ_0 penalized cost like Optimal Partitioning [68] or PELT [77]. 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 change point. 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 [39] pointed out that this can perform poorly due to removing information in the data under the transforming process. To overcome these difficulties, there are some methods for detecting changes in slope problems: Trend-filtering (2014) [78], which minimizes the RSS plus an ℓ_1 penalty on changes in slope; Narrowest-Over-Threshold (NOT) (2019) [9], which repeated a

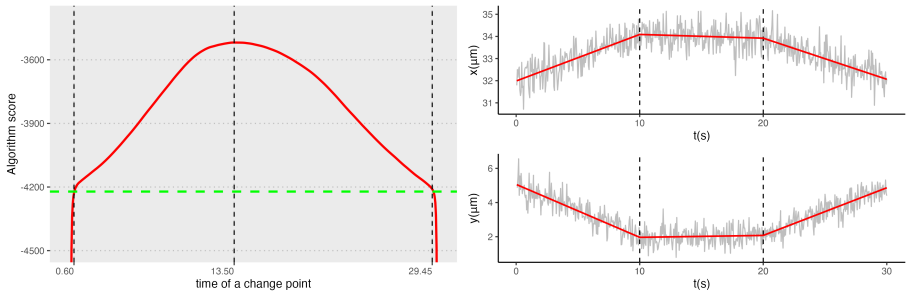


Figure 2.1: Example of a failure of using Binary segmentation for a change-in-speed problem, which tends to add a change point to the two actual changes. The left-hand side image, a 2D simulation of the lysosomal movement trajectory at 20Hz for a duration of 30 seconds, and the two actual change points, 10 seconds and 20 seconds, are represented as the t-vs-x and t-vs-y time series. The dashed lines represent the real change points. The corresponding segmentation is overlaid in red. In the right-hand side image, the CPLASS algorithm scores are computed on the simulation path at each possible addition of a changepoint model. The green dashed line is the score of a null model with no changes. The red solid curve contains the scores of one-change point models. The higher the score, the better the model is.

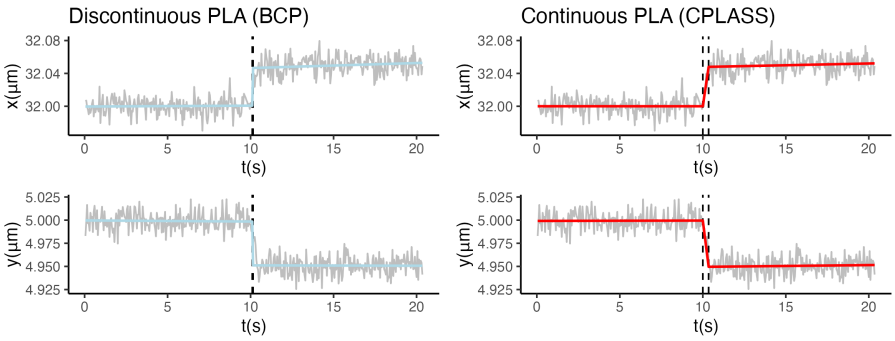
test for a single change in slope on subset data and used the narrowest-over-threshold to combine the results; CPOP (2019) [40] which based on a variant of dynamic programming to minimize the RSS plus an ℓ_0 penalty, i.e. a constant penalty for adding each change; the Narrowest Significance Pursuit (NSP) [47] method, 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.

Particle tracking. The study of intracellular transport relies on analyzing particle trajectories to reveal underlying biophysical states. Traditional methods, such as mean squared displacement (MSD) analysis, are widely used to quantify transport dynamics [98, 100]. However, MSD-based approaches can miss short-lived state transitions, as they average 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 [73, 107], while others are focused on velocity shifts [71, 101], or simultaneous changes in velocity and diffusivity [143].

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 [36, 126, 143]. In previous studies, our group [71, 111] 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 score function with a combination of the strengthened Schwarz Information Criterion (sSIC) [9, 48] 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 2.2 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.

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 [111] (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 [71] with intracellular transport by microtubule-based molecular motors; (3) Additional simulated data sets imitating the trajectories in live cell data.



Segment	1	2
Duration	10.15 s	10.20 s
Speed	2e-04 μm/s	6e-04 μm/s
State	Stationary	Stationary

Segment	1	2	3
Duration	9.95 s	0.35 s	10.00 s
Speed	0 μm/s	0.1978 μm/s	5e-04 μm/s
State	Stationary	Motile	Stationary

Figure 2.2: Comparison between discontinuous and continuous piecewise linear approximations. The simulated lysosomal movement trajectory in 2D at 20Hz for a duration of 20 seconds and the two actual change points, 9 seconds and 10.35 seconds, is represented by the t-vs-x and t-vs-y time series. The dashed lines represent the detected change points. The corresponding segmentation is overlaid in red.

2.2 The CPLASS algorithm

This section introduces the statistical model and algorithm used in this work. In Section 2.2.1, we introduce the statistical model. In Section 2.2.2, we construct a piecewise linear approximation of the data given an assumption that the number of change points is known. In Section 2.2.3 and Section 2.2.4, we propose a changepoints detection method where a criterion score with a penalty is provided, and a stochastic search method is used to find the maximum of the defined criterion score.

2.2.1 Statistical model

Assuming 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 treated as being Gaussian fluctuations around a sequence of unobserved anchor locations, which are denoted $\{a_i\}_{i=1}^n \subset \mathbb{R}^d$. We write

$$Y_i = a_i + \sigma\varepsilon_i, \tag{2.1}$$

where $\{\varepsilon_i\}_{i=1}^n$ is a sequence of independent and identically distributed d-dimensional standard normal random variables with noise magnitude σ^2. 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 ($1 \leq j \leq k-1$) associates with the jth changing time. 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 $i = 1, ..., n$, and the jth change point 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 with respect to the jth segment, the speed is defined as $s_j = \|V_j\|_2$ (for $j \in [k]$). Within each segment, we have

$$a_i = a_{M_{j-1}} + V_j(t_i - \tau_{j-1}), \tag{2.2}$$

where $i = M_{j-1} + 1, ..., M_j$, $j \in [k]$ is the index of the segment. Form the recursive formula (2.2), 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

$$f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}(t) = \left(\underline{a} - \sum_{j=1}^{i}(V_j - V_{j-1})\tau_{j-1}\right) + V_i t \qquad \forall t \in [\tau_{i-1}, \tau_i], i \in [k], \tag{2.3}$$

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

We now represent the model n multivariate observation $(Y_i)_{i=1}^n \subset \mathbb{R}^d$ on $\mathcal{T} =$

$\{t_1, \ldots, t_n\}$ according to a true signal function and Gaussian noises:

$$Y_i \overset{ind.}{\sim} \mathcal{N}(f^0(t_i), \sigma_0^2 I_d), \tag{2.4}$$

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 $\mathbf{V} = (V_0^0, \ldots, V_{k_0-1}^0) \subset \mathbb{R}^d$, 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 observation $(Y_i)_{i=1}^n$, 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 .

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

$$(\widehat{f}_n, \widehat{\sigma}_n^2, \widehat{k}_n) = \underset{f \in \mathcal{F}_k, \sigma^2 \in \Omega, k \leq \overline{k}}{\arg\max} \sum_{i=1}^n \log \mathcal{N}(y_i | f_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}}(t_i), \sigma^2 I_d) - \text{pen}_k, \tag{2.5}$$

with pen_k is the penalty term that preventing the overfitting issue, and get the MLE $\widehat{f}_n := f_{\widehat{\boldsymbol{\tau}}^n, \widehat{\mathbf{V}}^n, \widehat{\underline{a}}^n}$ of $\widehat{k}_n$ pieces.

We note that (2.5) is equivalent to finding the MLE with each $k \in [\overline{k}]$

$$(\widehat{f}_n^{(k)}, \widehat{\sigma}_{n,k}^2) = \underset{f \in \mathcal{F}_k, \sigma^2 \in \Omega}{\arg\max} \sum_{i=1}^n \log \mathcal{N}(y_i | f_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}}(t_i), \sigma^2 I_d), \tag{2.6}$$

and then find

$$\widehat{k}_n = \underset{k \in [\overline{k}]}{\arg\max} \sum_{i=1}^n \log \mathcal{N}(y_i | \widehat{f}_n^{(k)}(t_i), \widehat{\sigma}_{n,k}^2 I_d) - \text{pen}_k. \tag{2.7}$$

2.2.2 Continuous piecewise linear approximation given the changepoints

As discussed in the previous section, finding the MLE of (2.5) can be separated into two steps. In this section, we focus on solving (2.6), 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 $i = 1, \ldots n$, $l = 1, \ldots, d$, and $j = 1, \ldots, k$. 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 $Y^{(l)} = (y_{1l}, y_{2l}, \ldots, y_{nl}) \in \mathbb{R}^n$ $(l = 1, \ldots, d)$ 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

$$Y^{(l)} = \mathbb{T}\underline{W}^{(l)} + \sigma\varepsilon^{(l)}, \tag{2.8}$$

where

$$\mathbb{T} = \begin{bmatrix} 1 & t_1 & (t_1-\tau_1)\mathbb{1}_{t_1>\tau_1} & \cdots & (t_1-\tau_{k-1})\mathbb{1}_{t_1>\tau_{k-1}} \\ 1 & t_2 & (t_2-\tau_1)\mathbb{1}_{t_2>\tau_1} & \cdots & (t_2-\tau_{k-1})\mathbb{1}_{t_2>\tau_{k-1}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_n & (t_n-\tau_1)\mathbb{1}_{t_n>\tau_1} & \cdots & (t_n-\tau_{k-1})\mathbb{1}_{t_n>\tau_{k-1}} \end{bmatrix} \text{ is a } n \times (k+1) \text{ matrix,} \tag{2.9}$$

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

$$\text{RSS}(Y, t; \underline{W}) := \sum_{l=1}^{d} \left\| Y^{(l)} - \mathbb{T}\underline{W}^{(l)} \right\|_2^2. \tag{2.10}$$

The log-likelihood associated with the model is

$$\mathcal{L}(f, \sigma) = \sum_{i=1}^{n} \log \mathcal{N}(y_i | f_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}}(t_i), \sigma^2 I_d) = -\frac{nd}{2}\left[\log(2\pi) + \log(\sigma^2)\right] - \frac{1}{2\sigma^2}\text{RSS}(Y, t; \underline{W}) \tag{2.11}$$

The resulting MLE are

$$\widehat{\underline{W}}_{n,k}^{(l)} = \left(\mathbb{T}^\top \mathbb{T}\right)^{-1} \mathbb{T}^\top Y^{(l)} \quad \text{for } l \in [k], \tag{2.12}$$

$$\hat{\sigma}_{n,k}^2 = \frac{\text{RSS}(Y, t; \widehat{\underline{W}}_{n,k})}{dn}. \tag{2.13}$$

Note that as long as the changepoints τ_j are distinct and $n \geq k+1$, 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 $j = 1, ..., k$. The speed of the associated segment is $\hat{s}_j := \|\widehat{V}_j\|_2$.

2.2.3 Criterion function

The typical representation of the model contains switch-times $\boldsymbol{\tau}$, initial intercepts (or initial anchor locations) $\underline{a}$, segment velocities $\mathbf{V}$, and noise σ. In the following, we introduce the reduced representation of the model in the CPLASS using a switch point vector r.

In the second step, which is to find the changepoints that maximize the object

function (2.7), we utilized a switch point process to infer the number of change points, $k-1$. The switch point process was defined as an $(n-1)$-dimensional sequence of independent identically distributed (i.i.d) Bernoulli random variables [79], denoted $r=(r_1,...,r_{n-1})$, where $r_i=1$ indicated a change point occurred at time observation i with some probability that is known a prior. The number of change points from now is denoted by $|r|$ ($|r|=k-1$). Since a specific $(n-1)$-dimensional vector r corresponds to a fixed list of $|r|$ change points, we then can find an associated piecewise linear approximation whose $\mathrm{RSS}(Y,t;\widehat{\underline{a}},\widehat{W}),\tau_j,\hat{s}_j$ $(j=1,...,|r|+1)$ are determined as discussed in Section 2.2.2. From now on, we use the subscript r to indicate these relationships, i.e., $\widehat{\mathrm{RSS}}_r=\mathrm{RSS}(Y,t;\widehat{\underline{a}}_r,\widehat{W}_r),\tau_j,\hat{s}_j,\widehat{\underline{a}}_r,\tau_{j,r},\hat{s}_{j,r},\widehat{V}_{j,r}$.

For a given list of change points being r, according to the derivation in Section 2.2.2, the piecewise linear model provides a maximized value of the log-likelihood function

$$\widehat{\mathcal{L}}_n=\log(L(Y,t;\widehat{\underline{a}}_r,\widehat{W}_r,\hat{\sigma}_r))=\log\left(\frac{1}{(2\pi)^{nd/2}}\left(\frac{dn}{\widehat{\mathrm{RSS}}_r}\right)^{nd/2}\exp\left(-\frac{nd}{2}\right)\right) \tag{2.14}$$

$$=-\frac{nd}{2}\log\left(\widehat{\mathrm{RSS}}_r\right)+C, \tag{2.15}$$

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

Definition 2.1: Criterion Function

$$\Phi(r)=-\frac{nd}{2}\log\left(\widehat{\mathrm{RSS}}_r\right)-\mathrm{pen}(r). \tag{2.16}$$

Here, $\mathrm{pen}(r)$ refers to the penalty term designed to prevent overfitting. We utilized a strengthened Schwarz Information Criterion (sSIC) penalty expressed as $\log(n)^\gamma|r|$ for $\gamma>1$ (refer to [9, 44]) and a speed-control penalty function to mitigate the occurrence

of unrealistic speed values.

Definition 2.2: Penalty function

$$\text{pen}(r) = \log(n)^{\gamma}|r| + \sum_{j=1}^{|r|+1} h(\hat{s}_{j,r} - s_{cap}), \tag{2.17}$$

where $\gamma > 1$, $|r|$ is the number of changepoints, $\hat{s}_{j,r} = \left\|\widehat{V}_j\right\|_2$ $(j = 1, ..., |r| + 1)$ 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 under the prior knowledge of the scientist on the realistic speed of the trajectories. If there is no information about the speed limit, we often set $h(s) = 0$. The penalty term, then, is the linear penalty ℓ_0 in the form of a strengthened Schwarz information criterion (sSIC) [9, 44]. Remark that $\gamma = 1$ corresponds to the standard SIC penalty considered by Yao [141] in the context of multiple changepoint detection. Under our construction, we require $\gamma > 1$ to provide the consistency theorem support in Chapter 3. This requirement has also been used and discussed by Fryzlewicz in proposing the wild binary segmentation (WBS) [44] for change-in-mean problem and in the Narrowest-over-Threshold (NOT) algorithm for change-in-mean and change-in-slope problem [9]. Based on empirical experiments, we suggest choosing $\gamma = 1.2$.

2.2.4 Stochastic search of 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, it does not work for us since there are local maxima

of the score function. The gradient ascent could get stuck at local maxima and could not return the correct answer for the global maximum of the function (see Figure 2.3 for an example). We chose a stochastic search approach, using a Metropolis-Hastings algorithm as the search algorithm.

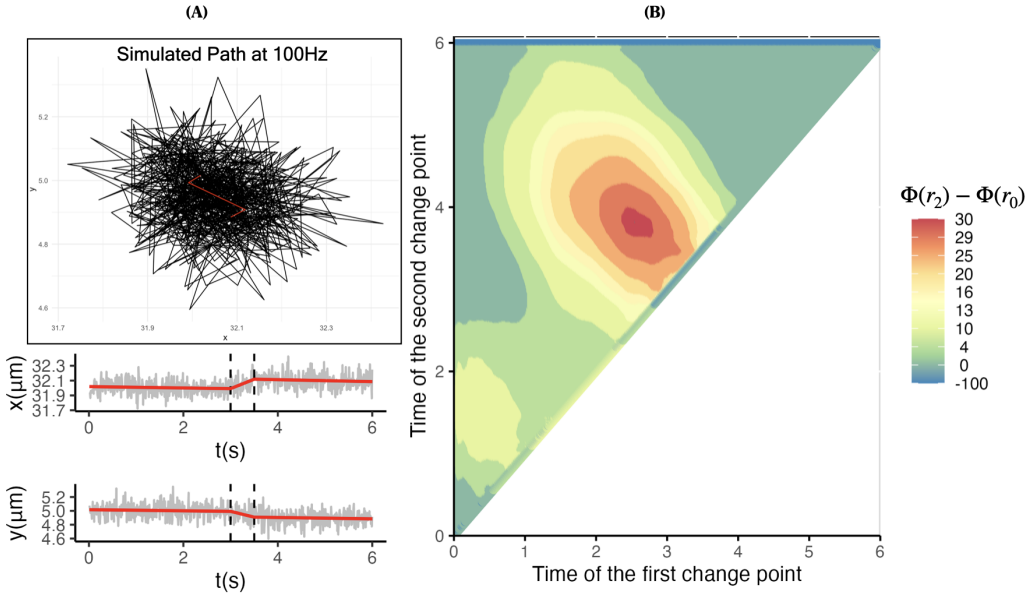


Figure 2.3: An example of gradient ascent could not be used for the change-in-speed problem. The contour plot for the difference in algorithm scores between models of two change points ($\Phi(r_2)$) and the model of no change points ($\Phi(r_0)$). Panel (A), the simulated lysosomal movement trajectory in 2D at 100Hz for a duration of 6 seconds and the two actual change points, 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 change points. The corresponding segmentation is overlaid in red. Panel (B), we calculated CPLASS scores for all possible combinations of two change points on this path and compared them to the log-likelihood of the no-change model.

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 $\exp(\Phi(r))$. This algorithm aims to generate an ergodic Markov chain $\{r^{(t)}\}_{t\geq 0}$ that has $\exp(\Phi(r))$ as its stationary distribution. The maximum of the $\exp(\Phi(r))$ function can then be approximated by the maximum of the resulting $\exp(\Phi(r^{(t)}))$'s. It is important to keep in mind that if the proposal function is irre-

ducible, then the Markov chain attained after running an MH algorithm is irreducible. Under the MH algorithm's rules, the chain's aperiodicity is ensured. Moreover, the chain $\{r^{(t)}\}_{t\geq 0}$ takes its values in a finite space. Therefore, it is uniformly ergodic ([80]). The detailed balance condition holds under our proposed rules (more discussion in Section 2.2.4).

In our algorithm, the proposal function for the switch point process takes its values on the set $\{0,1\}^{n-1}$, where n is the number of observations. There are four types of change point vector proposals: (1) an independent switch point process; (2) the creation or extinction of a change point; (3) the creation or extinction of a segment; or (4) a location shift of a single change point. Let r^{prop} and r^{cur} denote the proposed and current switch point 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 change points to vary by proposing an independent switch point process, r^{prop} following the distribution of Bernoulli random variables with the probability of a change points 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 $i = 1, ..., n-1$. 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 switch point process r^{prop}, the second type proposal, q_{bd}, provides another way for the number of change points to vary from iteration to iteration given the current switch point process r^{cur}. Instead of using the proposal function q_{bd} mentioned

in [53], we modify it such that the new version of q_{bd} will add one or delete one change point on the current list of change points with equal probabilities. In particular, let $\mathbf{M}_{r^{\text{cur}}} = \{M_1, ..., M_{|r^{\text{cur}}|}\}$ be the set of all current change point indices and $r^{\text{prop}} = r^{\text{cur}}$. There is 50% chance that a component is randomly sampled from the current switch 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 switch 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

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

- **Type 3.** Notation q_{bd_2}. The third type of proposal, q_{bd_2}, allows for adding or removing two nearby switches in the current list of changepoints. Like type 2, this proposal also provides a way to vary the number of change points in each iteration. In section 2.2.4, we provide motivation for including and removing consecutive switches. 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 change points for a fixed number of change points. We obtain the proposed switch point process by randomly sampling two components of the current switch point process as follows:

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

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:

$$q_r(r^{\text{prop}}|r^{\text{cur}}; u_r) = \begin{cases} q_{\text{new}}(r^{\text{prop}}), & \text{if } 0 \leq u_r \leq u_1 \\ q_{\text{bd}}(r^{\text{prop}}|r^{\text{cur}}), & \text{if } u_1 < u_r \leq u_2 \\ q_{\text{bd}_2}(r^{\text{prop}}|r^{\text{cur}}), & \text{if } u_2 < u_r \leq u_3 \\ q_{\text{shift}}(r^{\text{prop}}|r^{\text{cur}}), & \text{if } u_3 < u_r \leq 1, \end{cases} \tag{2.18}$$

where $u_1, u_2 - u_1, u_3 - u_2, 1 - u_3$ are probabilities that the proposal type 1, type 2, type 3, and type 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 change points

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

The number of iterations (T_{max}).

Output: A list contains T_{max} switch point processes $\{r^{(t)}\}_{t=0}^{T_{max}}$ 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 $k = 1, ..., K_{r^{(t)}}$.

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 $k = 1, ..., K_{r^{(0)}}$) by using the piecewise linear approximations.

2: **for** $t = 1, 2, \ldots, T_{max}$ **do**

3: Draw $u_r \sim \text{Uniform}(0, 1)$; $r^{\text{prop}} = q_r(.|r^{(t-1)}; u_r)$ (2.18)

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 $k = 1, ..., K_{r^{(t)}}$) 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 $\{\Phi(r^{(t)})\}_{t=1,...,T_{max}}$ 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.

Checking the detailed balance condition in the MH algorithm

Given the proposal function with four proposal types in Section 2.2.4, to prove detailed balance for the MH algorithm, we need to show that the transition kernel

satisfies:

$$\pi(r^{\text{cur}})q_r(r^{\text{prop}}|r^{\text{cur}})\alpha(r^{\text{prop}}|r^{\text{cur}}) = \pi(r^{\text{prop}})q_r(r^{\text{cur}}|r^{\text{prop}})\alpha(r^{\text{cur}}|r^{\text{prop}}), \tag{2.19}$$

where

- $\pi(r) = \exp(\Phi(r))$ is the target posterior distribution of change points.
- $q_r(r^{\text{prop}}|r^{\text{cur}})$ 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 Switch Point Process 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 $q_{\text{bd}}(r^{\text{prop}}|r^{\text{cur}}) = \frac{1}{2|r^{\text{cur}}|}$ then $q_{\text{bd}}(r^{\text{cur}}|r^{\text{prop}}) = \frac{1}{2(n-|r^{\text{prop}}|-1)}$ 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 $q_{\text{bd}_2}(r^{\text{prop}}|r^{\text{cur}}) = \frac{1}{|r^{\text{cur}}|}\mathbb{1}_{\{|r^{\text{cur}}|\geq 2\}}$ then

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

and vice versa. The acceptance rate is then

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

or

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

Type 4: $q_{\mathbf{shift}}$ (Shift proposal) This proposal shifts the position of one of the change points in the current list of change points. The proposal is symmetric as defined.

Therefore, the acceptance rate in this case is: $\alpha(r^{\text{cur}}|r^{\text{prop}}) = \min\left(1, \frac{\pi(r^{\text{prop}})}{\pi(r^{\text{cur}})}\right)$.

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 (2.19), 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 (2.19), we get that both sides of this equation are equal to $\pi(r^{\text{prop}})q_{\text{new}}(r^{\text{cur}}|r^{\text{prop}})$.

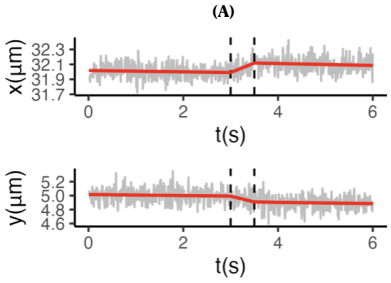
We finish checking the detailed balance condition.

Necessity of adding or removing a segment

In this part, we explain the reason for adding proposal type 3 in the MH algorithm 1 via the following Numerical Experiment.

A 6-second simulated path at 100Hz with changes at $t = 3$s and $t = 3.5$s (Figure 2.4(A)) is introduced. The corresponding segment speeds are $(0, 0.2, 0)$ μm/s. We assume that the current switch point process is $r^{\text{cur}} = \mathbf{0}$, which means that there are no change points. Figure 2.4(B) provides the corresponding log-likelihood and score.

Let's consider the case where the proposal only allows for adding one more change point at each iteration. Ideally, we expect that the algorithm can find one of the true change points, keep it, and find another true change point in the next update. However, it can happen that the proposal, r^{prop}, captures one of the true change



(B)

Changepoints	r	Log-likelihood	Criterion Score	Decision
No changepoints	..00000...00000...	1028.784	-1522.536	
At t = 3	..00010...00000...	1031.056	-1529.536	✖ Reject
At t=3.5	..00000...01000...	1028.846	-1531.745	✖ Reject
At t=3, t=3.5	..00010...01000...	1063.836	-1506.028	✔ Accept

Figure 2.4: *Necessity of the new proposal function, summaries of results from Numerical Experiment 2.2.4.* Panel (A), the simulated lysosomal movement trajectory in 2D at 100Hz for a duration of 6 seconds and the two actual change points, 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 change points. The corresponding segmentation is overlaid in red. Panel (B) shows different models with change points and their corresponding log-likelihood, the algorithm score, and the decision in the MH algorithm 1.

points, i.e., $r_{t=3}^{\text{prop}} = 1$ or $r_{t=3.5}^{\text{prop}}$, the score does not good enough for being accepted by the MH algorithm. Therefore, these naive random walk algorithms can get stuck, and we cannot find a good result for the change points list. To prevent this situation, we introduce the type 3 proposal function, which allows adding two consecutive new change points to the current r. To preserve the irreducibility of the proposal kernel, a proposal for removing a segment from the current path is indispensable. Now, this new update can provide a good r as a proposal switching process, which will be accepted by the MH algorithm (see Panel (B) in Figure 2.4).

2.3 Results

This section investigates the choice of the penalty through experiments and evaluates the performance of the CPLASS algorithm on both simulated and real datasets.

2.3.1 Speed penalty improves estimation without losing power

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

circumstance the adding of speed penalty is necessary.

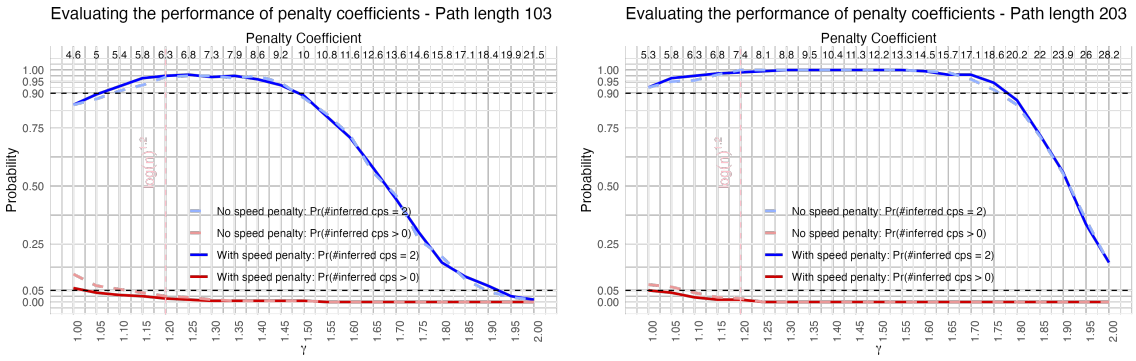
Performance of CPLASS with different values of the linear penalty term

In this experiment, we ran CPLASS with different values of γ and recommend using $\gamma = 1.2$ to ensure good performance of CPLASS for sample sizes $n \geq 100$ (see Figure 2.5). Particularly, we considered the simulation paths under the two following setups:

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

The simulation setups were designed to challenge the algorithm so that the distance between two actual changepoints under the alternative model remains small (10 and 3 time steps for sample sizes $n = 103$ and $n = 203$, respectively), and the speed between the two changing times is slow ($0.15\mu m/s$). Specifically, for $n = 103$, two sets of simulation paths were generated over a time period from 0 to 5.15 seconds with a frequency of 20Hz. The first set consisted of 200 paths with no changes, simulated under the null hypothesis with $\sigma = 0.1$ and $s = 0\mu m/s$. The second set contained 200 paths simulated under the alternative hypothesis, with two actual changes occurring at specific times, namely $t = 2.5s$ and $t = 2.65s$, $\sigma = 0.1$, and $(s_1, s_2, s_3) = (0, 0.15, 0)\mu m/s$. We then ran the CPLASS with different γ values ranging from 1 to 2 (see Panel (A)). We also ran the CPLASS with and without the speed-control penalty. At each γ value, we reported 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), for $n = 203$, we repeated the above procedure with simulated trajectories observed at 20Hz over 10.15 seconds with two actual changes at $t = 5$s



(A) (B)

Figure 2.5: *Varying penalty coefficient values.* Panel (A) 200 simulation paths over 5.15 seconds at 20Hz with $s = 0$, $\sigma = 1$ for the null hypothesis. 200 simulation paths over 5.15 seconds at 20Hz with two actual changes at $t = 2.5$s and $t = 2.65$s, three segments speeds $(s_1, s_2, s_3) = (0, 0.15, 0)$, $\sigma = 0.01$ for the alternative hypothesis. Panel (B) 200 simulation paths over 10.15 seconds at 20Hz with $s = 0$, $\sigma = 1$ 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 $(s_1, s_2, s_3) = (0, 0.15, 0)$, $\sigma = 0.01$ 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.

and $t = 5.15$s.

As shown in Figure 2.5, 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 true number of changepoints under the alternative hypothesis and keeping the probability of failing to detect a changepoint under the null model below 0.05 compared to the version without it.

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 to maintain reasonable speed. To leverage the speed penalty based on the knowledge of the dataset, we conducted CPLASS and BCP on a collection of 250 simulated trajectories derived

from the base parameter sets in Table 1 from [29] at 25Hz. We then evaluated the outputs using the Cumulative Speed Allocation statistic introduced in [29] and the Cumulative Distribution Function of the inferred maximum segment speeds (see Figure 2.6). For every speed $s \geq 0$, the CSA is the inferred proportion of time spent at speeds less than or equal to s. This can be understood as the time-weighted version of the cumulative distribution function of the speed. We refer to Cook et al. (2024) [29] for a more detailed discussion on the CSA.

In Panel (A) of Figure 2.6, 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, 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 $0.5\mu m/s$ 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 in the issue of missing short-fast segment in the use the dicontinous piecewise linear model. When comparing between CPLASS with or without speed penalty, the robust of CSA with respect to segmentation algorithm mentioned in [29] is confirmed.

The right panel of Figure 2.6 illustrates the empirical cumulative distribution (ECDF) of the maximum segment speeds after running the BCP and the two versions of CPLASS. It clearly to see that CPLASS with the speed penalty can follow the curve CDF of the true maximum speed better than BCP and CPLASS without the speed penalty.

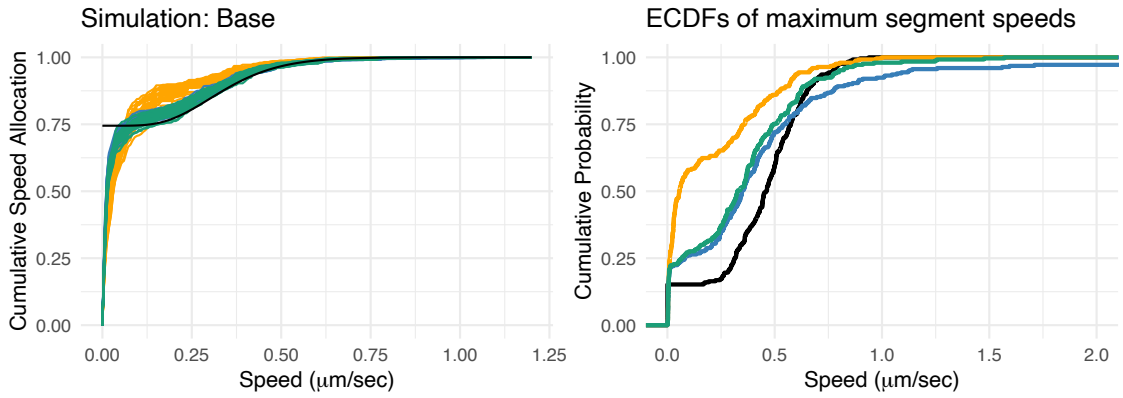


Figure 2.6: *Example of the necessity of the speed penalty.* We simulated a set of 250 simulated trajectories, each of the parameter sets in Table 1 from [29], observed at 25Hz. (**Left**) Cumulative Speed Allocation (CSA) computation for the simulated trajectories. The ***black*** line denotes the theoretical 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 [111]. (**Right**) Empirical cumulative distributions for the collection of maximum segment speeds of 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.

2.3.2 BCP versus CPLASS

In this section, we compared the discontinuous (using BCP) versus the continuous piecewise linear models (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 change points varied due to differences in the duration and speed of the middle segment. The 20 values for the middle segment duration were in the interval from 0.05 second to 1 second, with an increment of 0.05 second. The middle segment speed values (26 different values in μm/s) ranged from 0.005 μm/s to 0.1 μm/s, increasing by 0.005 μm/s. We fixed the first and third segments' duration (2 seconds) and speed (0 μm/s). The sample sizes then varied from $n = 81$ to 100. For each variation of the pair of speed and duration, there were

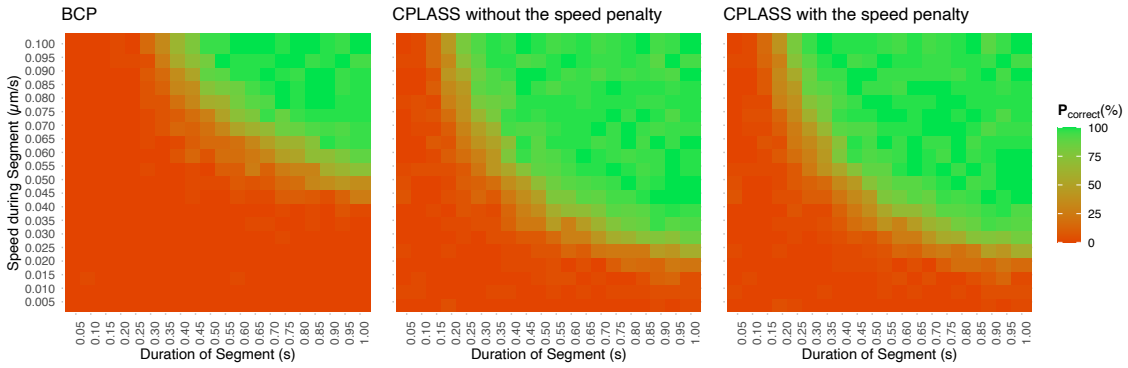


Figure 2.7: *Power analysis comparing CPLASS and BCP*. There are 20000 simulated paths at 20Hz with two actual changes ($\sigma = 0.01$). In each path, the first and the third segment durations are two seconds. We then varied the middle segment's durations (20 different values in seconds) and speeds (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.

50 corresponding simulated paths. The correctly detected percentage $\mathbf{P}_{\text{correct}}$ was then computed as a percentage of finding the correct number of changes for each case of the duration and speed. 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. Figure 2.7 illustrates the comparison. CPLASS effective regions (green region) are expanded compared to BCP, which indicates how CPLASS can detect the correct number of changepoints under the case of the location between two changes and the corresponding speed being small. 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.

2.3.3 In vivo experimental data - lysosomal transport

In this section, we revisit the data sets in [111]. 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, and the TrackMate macro was used to track lysosomes. They measured the lysosome diameter and defined the perinuclear region. We refer to Rayen et al. (2022) [111] for more details about data sets and data processing.

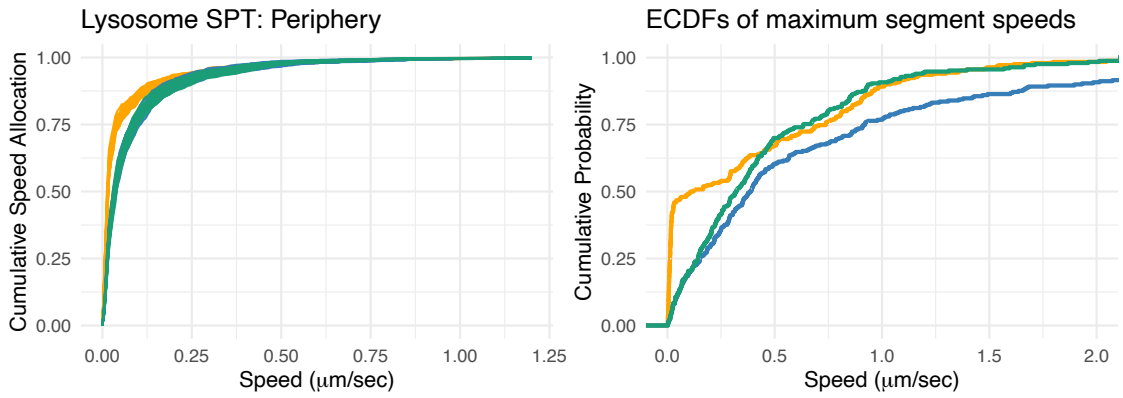


Figure 2.8: *In vivo experimental data analysis.* CSA and CDF computations for a set of 250 experimental lysosome trajectories in the periphery region of the cell from [111]. (**Left**) CSA computation for the trajectories. The ***green*** lines denote the inferred CSA computed after running CPLASS with 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 [111]. (**Right**) Empirical cumulative distributions for the collection of maximum segment speeds of 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.

Figures 2.8 and 2.9 display the results of running both versions of CPLASS and BCP for lysosomal transport in the periphery of cells, as observed by Rayens et al. [111]. The findings consistently demonstrate the robustness of CSA concerning the

segmentation algorithm, as discussed in the 25Hz simulated dataset. The histograms depicting the number of inferred changepoints (Figure 2.9) illustrate the similarity in estimating the number of changepoints across both versions of CPLASS, with the mean estimated changepoints around 6.5 and a standard deviation of 4. Conversely, BCP detected fewer changepoints, with a mean estimated changepoints of approximately 4 and a standard deviation of 4.4. This can be explained by the issue of missing short-fast segments when using the discontinuous piecewise linear approximation. When missing changepoints, more paths are labeled stationary with slow speed (see the left panel of Figure 2.8).

We reassessed the queries regarding how lysosomal transport varies with lysosome size and location. Figure 2.10 indicates that intracellular location, rather than diameter, is a crucial factor in lysosomal motion. This aligns with the findings from Rayens et al.[111]. Analyzing the CSA plot (the left panel of Figure 2.10), 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 2.10 confirms that large lysosomes are slower in transport than small lysosomes; however, overall, there is not a significant difference between these two groups. In [111], 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.

2.3.4 In vitro experimental data - quantum dot transport

In this section, we revisited the data sets used in [71] in which quantum dots were transported by a single kinesin-1 (kin-1) motor, a single dynein-dynactin-BicD2 (DDB) motor, and by a Kinesin-1/DDB pair. In [71], Jensen et al. developed a protocol

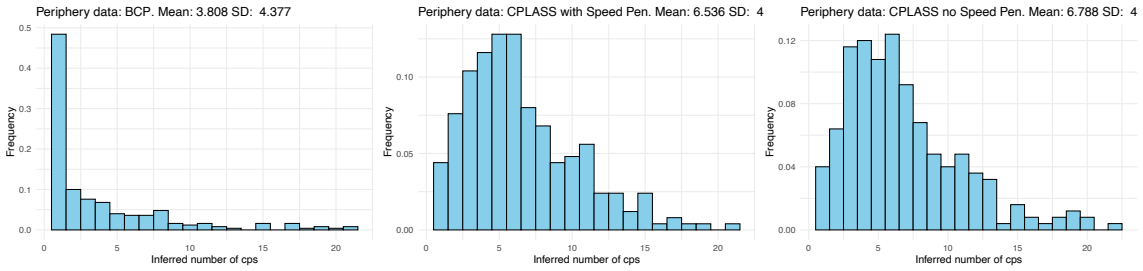


Figure 2.9: Histogram on inferred number of changepoints resulting from running CPLASS and BCP on the lysosomal transport in periphery region observed by Rayens et al. [111].

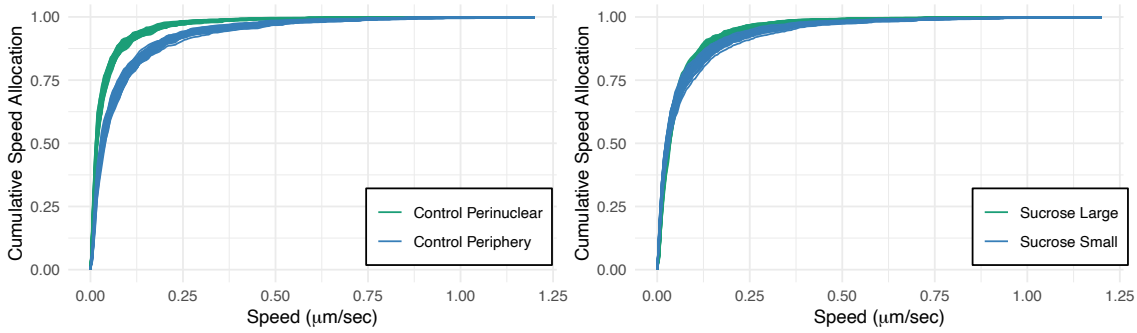


Figure 2.10: *CSA plots for different group comparisons.* (**Left**) The CSA bootstrap ensemble curves compare lysosomal transport in perinuclear and periphery regions. (**Right**) The CSA bootstrap ensemble curves compare the sucrose-treated groups restricted to the periphery region of the cell from [111].

for finding changepoints in cargo trajectories that were projected along the length of a straight microtubule and reporting velocity distributions. The differences in velocity distributions and run lengths revealed the differences for different molecular motor families. Since our proposed CPLASS can handle multidimensional data sets, we applied it directly to these quantum dot data sets without projecting the two-dimensional data into a one-dimensional format. We then calculated the CSA bootstrap ensemble curves (see Figure 2.11) 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

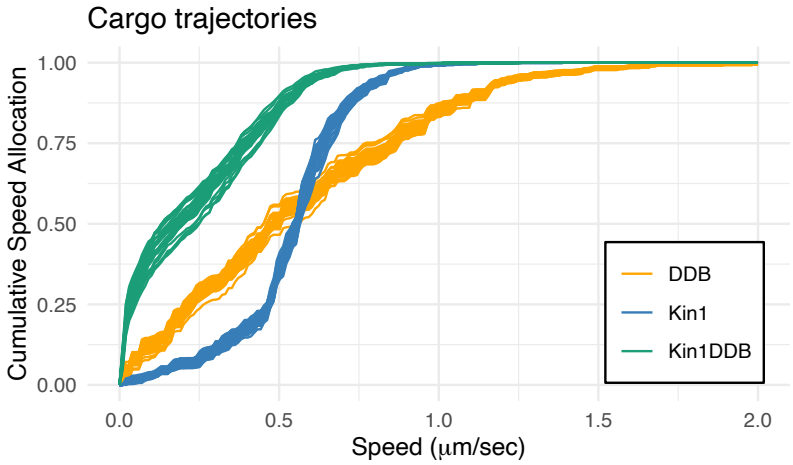


Figure 2.11: *CSA plot for different motor families.* The CSA bootstrap ensemble curves compare cargo trajectories among three groups of motor transports: kinesin-1 (Kin1), Dynein-Dynactin-BicD2 (DDB), and Kinesin-1/DDB pairs.

steps processively with consistent behavior, while DDB (orange curves) exhibits a broader range of speeds. When both motors are present (green curves), the speed is generally lower, reflecting the tug-of-war state. This confirms the observations made about the data sets but offers a more refined and robust characterization.

2.4 CPLASS under diffusing anchor case

In this section, we studied the effect of anchor diffusing cases on the detection of changes in CPLASS, where a second source of noise appears in the anchor location. In particular, within each segment, we have

$$a_i = a_{M_{j-1}} + V_j(t_i - \tau_{j-1}) + \xi\sqrt{\Delta}\sum_{l=M_{j-1}+1}^{i}\eta_l, \tag{2.20}$$

where $i = M_{j-1}+1, ..., M_j$, $j \in [k]$ is the index of the segment, $\eta_l \sim \mathcal{N}(0, I_d)$. When a segment speed equals 0, the anchor in the considered segment is a Brownian motion.

We simulated trajectories from Equation (2.20) with the standard deviation of the

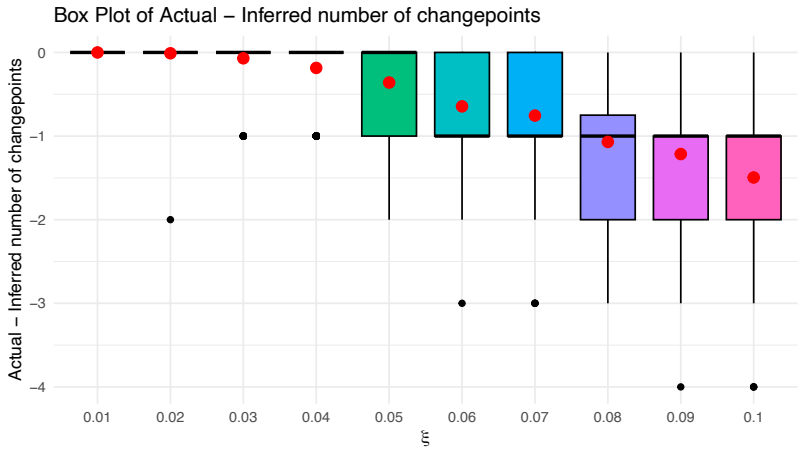


Figure 2.12: *Anchor Diffusing.* The differences between the actual and inferred number of changepoints for 200 simulated trajectories using the anchor diffusing model in Equation (2.20) at a framerate of 20Hz and ξ vary from 0.01 to 0.1 are depicted. The box plots illustrate the variance in these differences at various values of ξ related to the second noise source. The red dots present the mean values accordingly to these differences in each ξ.

second noise source ξ values ranging from 0.01 to 0.1. For each ξ, 200 trajectories were simulated at 20Hz ($\Delta = 0.05$) and $\sigma = 0.2$ with three actual segments. The changepoint times were at 2 seconds and 12 seconds. The corresponding segment times were $(2, 10, 2)$ seconds, actual segment velocities are $V_1 = (0.5, -0.5)^\top, V_2 = (0, 0)^\top, V_3 = (0.5, 0.5)^\top$, total time is 14 seconds ($n = 280$). We then ran CPLASS on the simulated data. Figure 2.12 shows the box plot that summarizes differences between the actual and inferred changepoints. It is evident that CPLASS can effectively detect the true number of changepoints when $\xi = (0.01, 0.02, 0.03, 0.04)$ with a small variance in the differences. Overfitting issues begin when ξ increases from 0.05 to 0.1. It can be explained that as the noise in the anchor increases, Brownian motion can cause the path to have more false changepoints (see Figure 2.14 for an example). CSA plots in Figure 2.13 also confirm this phenomenon. When $\xi = 0.01$, the ensemble-inferred CSA curves follow the ensemble-actual CSA curves well. We observe a clear difference between the inferred and actual CSA curves at $\xi = 0.05$ and $\xi = 0.1$.

We see that despite the challenge of anchor diffusion, CPLASS continues to effectively provide a good continuous piecewise linear approximation in the mean that follows the trajectories well (see Figure 2.14). The main difficulty now lies in classifying the states of segments where the diffusing anchor leads to more false positives in labeling the states; this means that the path seems to be moving, but this movement is caused by diffusion rather than drift. Chapter 4 presents a hypothesis test on the *Stationary/Motile* segments, and we will revisit this anchor diffusion model with further discussion.

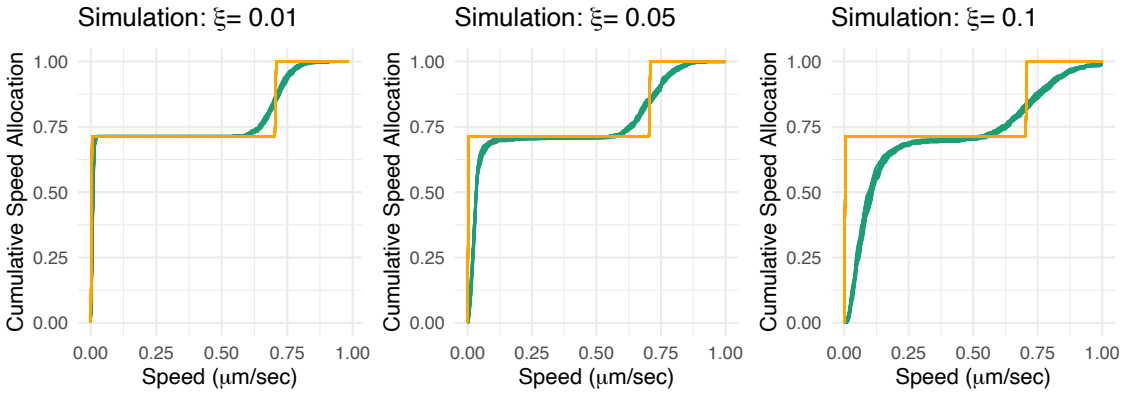


Figure 2.13: *Anchor Diffusing - CSA comparisons.* We simulated three sets of trajectories at a frame rate of 20Hz and ξ values of 0.01, 0.05, and 0.1. The inferred parameters are collected after running CPLASS. The ***green*** lines in each plot denote the inferred CSA computations for each set of simulated trajectories. The ***orange*** lines denote CSA bootstrap samples of the true simulated trajectories.

2.5 Discussion

In this chapter, 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 to detecting multiple changepoints do not

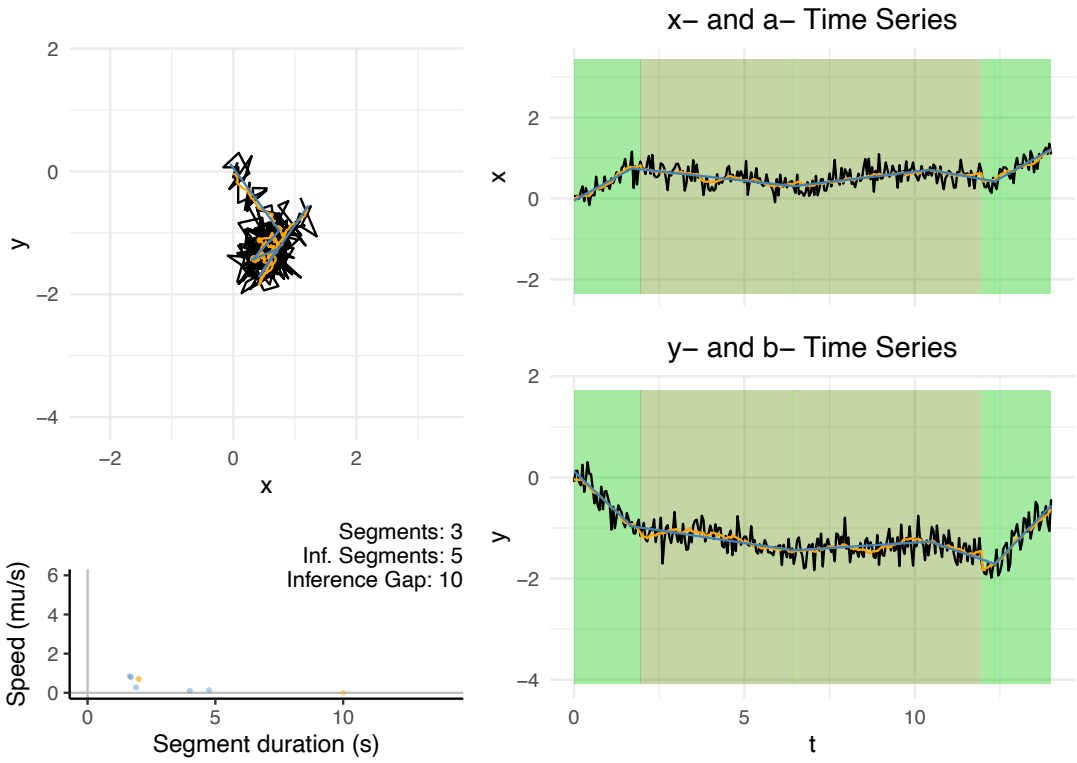


Figure 2.14: *Anchor Diffusing example.* An example simulated path under the anchor diffusing model at $\xi = 0.1$ with two actual changepoint times at 2 and 12 seconds, actual segment velocities are $V_1 = (0.5, -0.5)^\top$, $V_2 = (0, 0)^\top$, $V_3 = (0.5, 0.5)^\top$. In the middle segment, the anchor is Brownian motion. The blue line represents the inferred anchor position resulting from CPLASS, and the orange line denotes the actual anchor locations. The right panel shows the state segments for each time series. Each ***green*** panel denotes a *Motile* segment, and each ***red*** panel denotes a *Stationary* segment. The shaded ***gray*** panel denotes the differentiating overlap of the inferred vs. actual segment panels, i.e., the inference gap detailed in [29] and will be revisited in Section 4.3. In the segment duration versus speed plot, the ***blue*** points denote the inferred segments, and ***orange*** points denote the actual simulated segments. With $\xi = 0.1$, the noisy trajectory makes CPLASS detect more changepoints than the truth.

work for detecting changes in velocity. For example, binary segmentation detects a single changepoint but struggles with initial errors in changepoint locations. Existing dynamic programming algorithms like PELT and optimal partitioning cannot handle change-in-velocity due to continuity assumptions that create parameter dependencies,

breaching independence structures. To address these issues, Baranowski et al. [9] proposed the Narrowest-Over-Threshold (NOT) algorithm, while Fearnhead et al. [39] introduced a variant of dynamic programming, and Kim et al. [78] 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 (in Chapter 3), real-world applications—such as molecular motor data—often involve small sample sizes (n). To address this, we introduced a speed penalty that enhances statistical power and robustness for small n while preserving consistency in the large-sample limit. Furthermore, we demonstrated that comparing the new tool - Cumulative Speed Allocation (CSA) [29] and Cumulative Distribution Function (CDF) reveals that the proportion of time spent at different speeds offers a more stable performance metric than segment velocity counts, making it less sensitive to algorithmic variations.

Crucially, our method is inherently multidimensional, allowing it to capture complex structures in diverse datasets. Nevertheless, computational efficiency poses a challenge, as MCMC search is inherently slow. Future efforts will concentrate on optimizing the search process to enhance convergence speed and maximize likelihood estimation. Furthermore, being mindful of the challenges posed by anchor diffusion, we will require a more robust statistical tool to study the state of the segment (i.e., Motile states vs. Stationary states - see Chapter 4). In addition, ensuring the consistency of CSA inference remains an open theoretical question that merits further investigation.

Finally, our dataset analysis provides a refined and more robust characterization of molecular motor behavior, reinforcing known biological findings with greater precision. This improved resolution opens avenues for deeper insights into biophysical processes,

demonstrating the potential of CPLASS in both theoretical and applied contexts.

Chapter 3

Consistent estimate of changepoints with sSIC

In this chapter, we consider the asymptotic properties of estimating changepoints by maximizing (2.16) under the choice of the linear penalty $|r|\log(n)^\gamma$ (with $\gamma > 1$) (which equivalent to obtain the penalized MLE $\hat{f}_n$, σ_n^2 and $\hat{k}_n$ of (2.5) with pen $= k\log(n)^\gamma$).

3.1 Statement of consistency theorem

Our model aims to learn the continuous and piecewise linear function of particles' movement from noisy data. A particle is observed in a period from 0 to T with frame rate Δ, i.e., observed at time $\mathcal{T} := \{t_1, \ldots, t_n\}$, where $t_i = i\Delta$ for $i = 1, \ldots, n$, and $n = T/\Delta$ is assumed to be an integer. The final goal of this theoretical section is to understand the consistency of the estimated function as the frame rate $\Delta \to 0$ (i.e., $n \to \infty$). Because T is held fixed in our argument, without loss of generality, we can assume $T = 1$, so that $n = 1/\Delta$ and $t_i = i/n$ for all i.

We start by recalling notations for the class of piecewise linear model from $\mathcal{T} \to \mathbb{R}^d$.

On $t \in \mathcal{T}$, a multivariate continuous piecewise linear 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 $0 =: \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_{k-1} < \tau_k := t_n = 1$, sets of slopes $\mathbf{V} = (V_0, \ldots, V_{k-1}) \subset \mathcal{V} \subset \mathbb{R}^d$, and initial intercept

$\underline{a} \in \mathcal{A} \subset \mathbb{R}^d$, is defined as

$$f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}(t) = \left(\underline{a} - \sum_{j=1}^{i}(V_j - V_{j-1})\tau_{j-1}\right) + V_i t \qquad \forall t \in (\tau_i, \tau_{i+1}], i \in [0,k], \tag{3.1}$$

where $\mathcal{V}$ and $\mathcal{A}$ are two compact subsets of $\mathbb{R}^d$. In the following, we assume those compact sets are known and fixed. When $V_j \neq V_{j-1}$ for all $j \in [1, k-1]$, the signal function $f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}$ is said to have k segments (or pieces) and $(k-1)$ changepoints. Denote $\mathcal{F}_k$ by the collection of such signal functions with k pieces.

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

$$y_i \overset{ind.}{\sim} \mathcal{N}(f^0(t_i), \sigma_0^2 I_d), \tag{3.2}$$

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 $\mathbf{V} = (V_0^0, \ldots, V_{k_0-1}^0) \subset \mathbb{R}^d$, 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 observation $(Y_i)_{i=1}^n$, 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.

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

$$(\widehat{f}_n, \widehat{\sigma}_n^2, \widehat{k}_n) = \underset{f \in \mathcal{F}_k, \sigma^2 \in \Omega, k \leq \overline{k}}{\arg\max} \sum_{i=1}^{n} \log \mathcal{N}(y_i | f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}(t_i), \sigma^2 I) - (k-1)(\log(n))^\gamma, \tag{3.3}$$

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 $\widehat{k}_n$ pieces. As discussed in the

main text, given the optimal signal function $\widehat{f}_n$, the optimal variance $\widehat{\sigma}_n^2$ can be shown as the average RSS:

$$\widehat{\sigma}_n^2 = \frac{\sum_{i=1}^n \|y_i - \widehat{f}_n(t_i)\|^2}{nd}, \tag{3.4}$$

so that problem (3.3) is equivalent to maximizing the criterion function:

$$(\widehat{f}_n, \widehat{k}_n) = \underset{f \in \mathcal{F}_k, k \leq \overline{k}}{\arg\max} -\frac{nd}{2} \log\left(\sum_{i=1}^n \|y_i - f_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}}(t_i)\|^2\right) - (k-1)(\log(n))^\gamma. \tag{3.5}$$

Theorem 3.1: Consistency Theorem

Suppose that data is generated according to the true model (2.4) with true signal function $f^0 = f_{\boldsymbol{\tau}^0, \mathbf{V}^0, \underline{a}^0}$ of k_0 pieces, $\min_i |\tau_i^0 - \tau_{i-1}^0| > \underline{C}_1 > 0$ 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 (2.5), where pen $= k \log(n)^\gamma$ (with $\gamma > 1$), satisfies

$$\mathbb{P}_0\left(\widehat{k}_n = k_0, \max_{i=1,\ldots,k_0-1} |\widehat{\tau}_i^n - \tau_i^0| \leq C\sqrt{\frac{\log n}{n}}\right) \to 1, \tag{3.6}$$

as $\Delta \to 0$ ($n \to \infty$), where $\mathbb{P}_0$ is the probability associated with the true model, C is a contant depending on $\underline{C}_1$, $\underline{C}_2$, and $\overline{k}$.

Theorem 3.1 established the consistency of the changepoints and number of changepoints in penalized MLE problem (2.5) with the pen $= k \log(n)^\gamma$ (for $\gamma > 1$). Note that we allow the true signal function f^0 to vary as n varies. As long as the condition $\min_i |\tau_i^0 - \tau_{i-1}^0| > \underline{C}_1 > 0$ and $\min_i \|V_i^0 - V_{i-1}^0\| > \underline{C}_2 > 0$ hold, then the consistency result (3.6) holds.

The findings closely resemble those from the Narrowest-over-Threshold method used by Baranowski, Chen, and Fryzlewicz [9], as well as Fearnhead, Maidstone, and Letchford [39], 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 [93] due to the continuous requirement of the signal function. Our proof technique relies on empirical process theory [129], 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 the lysosome cannot exceed $2\mu m/s$. 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 2.3.1, 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.

We note that (3.3) is equivalent to finding the MLE with each $k \in [\overline{k}]$

$$(\widehat{f}_n^{(k)}, \widehat{\sigma}_{n,k}^2) = \underset{f \in \mathcal{F}_k, \sigma^2 \in \Omega}{\arg\max} \sum_{i=1}^{n} \log \mathcal{N}(y_i | f_{\boldsymbol{\tau}, \mathbf{v}, \underline{a}}(t_i), \sigma^2 I), \tag{3.7}$$

and then find

$$\widehat{k}_n = \underset{k \in [\overline{k}]}{\arg\max} \sum_{i=1}^{n} \log \mathcal{N}(y_i | \widehat{f}_n^{(k)}(t_i), \widehat{\sigma}_{n,k}^2 I) - (k-1)(\log(n))^{\gamma}. \tag{3.8}$$

Hence, to prove Theorem 3.1, we aim to understand the convergence of the parameters and likelihood in (2.6) for each k first. Central to our theoretical development is the empirical process theory [129], which provides uniform convergence of empirical average log-likelihood to 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

$$\overline{\mathcal{L}}_n(f,\sigma^2) = \frac{1}{n}\sum_{i=1}^{n} \log \mathcal{N}(y_i|f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}(t_i),\sigma^2). \tag{3.9}$$

and

$$\overline{\mathcal{L}}_0(f,\sigma^2) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Y_i\sim N(f^0(t_i),\sigma_0^2)} \log \mathcal{N}(Y_i|f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}(t_i),\sigma^2), \tag{3.10}$$

respectively.

3.2 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 $\mathbb{R}^d$ is our problem), define the Hellinger process distance [52, 129] between product densities $p = \otimes_{i=1}^n p_i$ and $q = \otimes_{i=1}^n q_i$ and its average to be

$$h_n^2(p,q) := \frac{1}{2}\sum_{i=1}^{n} h^2(p_i,q_i), \quad \overline{h}_n^2(p,q) := \frac{1}{n} h_n^2(p,q). \tag{3.11}$$

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 function $f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}$ and noise level σ^2, let $p^{(n)}_{\boldsymbol{\tau},\mathbf{V},\underline{a},\sigma^2}$ denote the product density $\otimes_{i=1}^n \mathcal{N}(y_i|f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}(t_i),\sigma^2)$ on $\mathcal{Y}^n$. Let $p_0^{(n)}$ denote the product true density $\otimes_{i=1}^n \mathcal{N}(y_i|f^0(t_i),\sigma_0^2)$ on $\mathcal{Y}^n$. The 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 calculate (or sufficiently provide an upper bound for) the smallest number of balls to cover this space in the Hellinger process distance. This number is often referred to as the "covering number", which we will now define.

Definition 3.1: 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) $\overline{h}_n \left(\dfrac{\mathbf{p}_j^L + p_0^{(n)}}{2}, \dfrac{\mathbf{p}_j^U + p_0^{(n)}}{2} \right) \leq \delta$ 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 [129], define the entropy integral:

$$J_B(\delta) := \int_{\delta^2/c_0}^{\delta} H_B^{1/2}(u) du \vee \delta, \quad 0 < \delta \leq 1. \tag{3.12}$$

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

Theorem 3.2: Theorem 8.14 in [129]

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

$$\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n), \tag{3.13}$$

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

$$\mathbb{P}_0 \left(\overline{h}_n \left(p^{(n)}_{\widehat{\boldsymbol{\tau}}, \widehat{\mathbf{V}}, \widehat{\underline{a}}, \widehat{\sigma^2}}, p_0^{(n)} \right) \geq \delta \right) \leq c \exp\left(-\frac{n\delta^2}{c^2} \right). \tag{3.14}$$

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

Theorem 3.3: Theorem 8.13 in [129]

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:

$$\overline{h}_n(\bar{p}^{(n)}_{\boldsymbol{\tau},\mathbf{V},\underline{a},\sigma^2}, p_0^{(n)}) \leq R \quad \forall(\tau, \mathbf{V}, \underline{a}, \sigma^2) \in \Theta, \tag{3.15}$$

$$b \leq C_1\sqrt{n}R^2 \wedge 8\sqrt{n}R, \tag{3.16}$$

and

$$b \geq \sqrt{D^2(C_1+1)}\left(\int_{b/(2^6\sqrt{n})}^{R} H_B^{1/2}\left(\frac{u}{\sqrt{2}}, \Theta\right) du \vee R\right), \tag{3.17}$$

then

$$\mathbb{P}_0\Big(\sup_{(\tau,\mathbf{V},\underline{a},\sigma^2)\in\Theta} \sqrt{n}|\bar{Z}_{\boldsymbol{\tau},\mathbf{V},\underline{a},\sigma^2} - \bar{A}_{\boldsymbol{\tau},\mathbf{V},\underline{a},\sigma^2}| \geq b\Big) \leq D\exp\left[-\frac{b^2}{D^2(C_1+1)R^2}\right], \tag{3.18}$$

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 3.2 and 3.3, we refer to Vandegeer (2020) [129].

3.3 Convergence of latent piecewise functions and likelihood functions

There are some notations that we used during the proofs of this thesis. 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 $b_n \lesssim a_n$. 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*

$$h^2(\mathcal{N}(\mu, \sigma^2 I_d), \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2 I_d)) \asymp \|\mu - \tilde{\mu}\|^2 + (\sigma^2 - \tilde{\sigma}^2)^2, \tag{3.19}$$

and

$$\sup_{y \in \mathbb{R}^d} \left| \mathcal{N}(y|\mu, \sigma^2 I_d) - \mathcal{N}(y|\tilde{\mu}, \tilde{\sigma}^2 I_d)) \right| \lesssim \|\mu - \tilde{\mu}\| + |\sigma^2 - \tilde{\sigma}^2|. \tag{3.20}$$

Proof of Lemma 1. **1. Proof of (3.19):**

Recall the Hellinger distance between two location-scale Gaussian:

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

Firstly, we notice

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

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

We also have that

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

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

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

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 (3.20): 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

$$\frac{\partial}{\partial \mu}\mathcal{N}(y|\mu, \sigma^2 I_d) = \mathcal{N}(y|\mu, \sigma^2 I_d) \cdot \frac{y-\mu}{\sigma^2}, \tag{3.21}$$

$$\frac{\partial}{\partial \sigma^2}\mathcal{N}(y|\mu, \sigma^2 I_d) = \mathcal{N}(y|\mu, \sigma^2 I_d) \cdot \left(-\frac{d}{2\sigma^2} + \frac{\|y-\mu\|^2}{\sigma^4}\right). \tag{3.22}$$

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

$$\left|\frac{\partial}{\partial \mu}\mathcal{N}(y|\mu, \sigma^2 I_d)\right| \leq \mathcal{N}(y|\mu, \sigma^2 I_d) \cdot \frac{\|y-\mu\|}{\sigma^2}, \tag{3.23}$$

$$\left|\frac{\partial}{\partial \sigma^2}\mathcal{N}(y|\mu, \sigma^2 I_d)\right| \leq \mathcal{N}(y|\mu, \sigma^2 I_d) \cdot \left(\frac{d}{2\sigma^2} + \frac{\|y-\mu\|^2}{\sigma^4}\right). \tag{3.24}$$

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

$$\sup_{y\in\mathbb{R}^d} \left|\mathcal{N}(y|\mu, \sigma^2 I_d) - \mathcal{N}(y|\tilde{\mu}, \tilde{\sigma}^2 I_d))\right| \leq L_\mu \|\mu - \tilde{\mu}\| + L_{\sigma^2}|\sigma^2 - \tilde{\sigma}^2|, \tag{3.25}$$

where

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

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

$$
\|f-g\|_n=\left(\frac{1}{n}\sum_{i=1}^{n}\|f(i/n)-g(i/n)\|^2\right)^{1/2}. \tag{3.26}
$$

Theorem 3.4: Convergence rates of parameters

Given the same condition 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 $1-c_1n^{-c_2}$ we have

$$
\|\widehat{f}_n^{(k)}-f^0\|_n^2\le C\left(\frac{k\log n}{n}\right),\qquad |\widehat{\sigma}_{n,k}^2-\sigma_0^2|\le C\left(\frac{k\log n}{n}\right)^{1/2}, \tag{3.27}
$$

and

$$
0\le\overline{\mathcal{L}}_n(\widehat{f}_n^{(k)})-\overline{\mathcal{L}}_n(f^0)\le Ck\frac{\log(n)}{n}, \tag{3.28}
$$

where C only depends on d, $\mathcal{V},\mathcal{A}$ and Ω (but not k and n).

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 3.2 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 pieces. We generalize the technique in [51] for our model.

Step 1.1. Covering the space of changepoint models with fixed changes under ℓ_∞ norm. Suppose that the set of changes $\boldsymbol{\tau} = (\tau_1, \dots, \tau_{k-1}) \subset (0,1)$ 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

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

We have that $|B| \asymp (1/\varepsilon)^{(k+1)d+1}$, 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

$$\begin{aligned}
\left\| f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}(t) - f_{\boldsymbol{\tau},\tilde{\mathbf{V}},\tilde{\underline{a}}}(t) \right\|_\infty &= \left\| (\underline{a} - \tilde{\underline{a}}) + \sum_{j=1}^{i} (V_j - \tilde{V}_j - (V_{j-1} - \tilde{V}_{j-1}))\tau_{j-1} + (V_i - \tilde{V}_i)t \right\|_\infty \\
&\le \|\underline{a} - \tilde{\underline{a}}\|_\infty + \left(\sum_{j=1}^{i} \|V_j - \tilde{V}_j\|_\infty + \sum_{j=1}^{i-1} \|V_j - \tilde{V}_j\|_\infty \right) \max_j \tau_{j-1} + \|V_i - \tilde{V}_{i-1}\|_\infty t \le (2k+2)\varepsilon.
\end{aligned}$$

Hence,

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

Combining with Lemma 1, it implies

$$\sup_{y_i \in \mathbb{R}^d} |\mathcal{N}(y_i | f_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}}(t_i), \sigma^2 I) - \mathcal{N}(y_i | f_{\boldsymbol{\tau}, \tilde{\mathbf{V}}, \tilde{\underline{a}}}(t_i), \tilde{\sigma}^2 I)| \lesssim k\varepsilon, \quad \forall i \in [n]. \tag{3.29}$$

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 $N \asymp (k/\delta)^{(k+1)d+1}$ 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

$$\sup_{y_i \in \mathbb{R}^d} |\mathcal{N}(y_i | f_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}}(t_i), \sigma^2 I) - p_{ji}(y_i)| \leq \delta, \quad \forall i \in [1, n]. \tag{3.30}$$

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)

$$H(y) = \begin{cases} b_1 \exp(-b_2 \|y\|^2), & \|y\| \geq B, \\ (\sqrt{2\pi}\underline{\sigma}^2)^{-d}, & \text{otherwise} \end{cases} \tag{3.31}$$

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 following:

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

With this construction, (3.30) implies

$$p_{ji}^L(y_i) \leq \mathcal{N}(y_i | f_{\boldsymbol{\tau}, \mathbf{V}, \underline{a}}(t_i), \sigma^2 I) \leq p_{ji}^U(y_i) \quad \forall y_i \in \mathbb{R}^d, i \in [1, n]. \tag{3.32}$$

Hence, this collection of brackets satisfies condition (ii) in the definition of covering with bracketing (Defintion 3.1). Now, we are checking condition (i). For any j, i and $\overline{B} \geq B$,

$$\begin{aligned}\int_{\mathbb{R}^d}(p_{ji}^U - p_{ji}^L)dy &\leq \int_{\|y\|\leq\overline{B}} 2\delta dy + \int_{\|y\|\geq\overline{B}} H(y)dy \\ &\lesssim \delta\overline{B}^d + \overline{B}^d \exp\left(-b_2\overline{B}^2\right), \end{aligned} \tag{3.33}$$

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 (3.33) gives

$$\int_{\mathbb{R}^d}(p_{ji}^U - p_{ji}^L)dy \lesssim \delta\left(\log\left(\frac{1}{\delta}\right)\right)^{d/2}. \tag{3.34}$$

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

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

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

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

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

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

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

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

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

Step 2. Convergence rate of parameter estimation Consequence of Theorem 3.2.

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

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

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

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

Substitute $\varepsilon = \varepsilon_n$ to the conclusion of Theorem 3.2, we have

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

where C_B depends on d, $\mathcal{V}, \mathcal{A}$, 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

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

therefore,

$$\|\widehat{f}_n^{(k)} - f^0\|_n^2 \leq C \left(\frac{k \log n}{n} \right), \qquad |\widehat{\sigma}_{n,k}^2 - \sigma_0^2| \leq C \left(\frac{k \log n}{n} \right)^{1/2}, \quad \text{for } k \geq k_0 \tag{3.35}$$

where C depends on d, $\mathcal{V}, \mathcal{A}$, 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 3.3 to show the convergence of the likelihood functions. All of the following arguments hold for $k \geq k_0$, and we work with $\dfrac{\mathcal{N}(y_i | f_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}}}(t_i), \widehat{\sigma}_{n,k}^2) + \mathcal{N}(y_i | f^0(t_i), \sigma_0^2)}{2\mathcal{N}(y_i | f^0(t_i), \sigma_0^2)}$ instead of $\dfrac{\mathcal{N}(y_i | f_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}}}(t_i), \widehat{\sigma}_{n,k}^2)}{\mathcal{N}(y_i | f^0(t_i), \sigma_0^2)}$ because the former is always bounded below by $1/2$, but the latter is not.

By the concavity of the log function, we have

$$\frac{1}{2} \log \frac{\mathcal{N}(y_i | f_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}}}(t_i), \widehat{\sigma}_{n,k}^2)}{\mathcal{N}(y_i | f^0(t_i), \sigma_0^2)} \leq \log \left(\frac{\mathcal{N}(y_i | f_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}}}(t_i), \widehat{\sigma}_{n,k}^2) + \mathcal{N}(y_i | f^0(t_i), \sigma_0^2)}{2\mathcal{N}(y_i | f^0(t_i), \sigma_0^2)} \right) \tag{3.36}$$

Recall that we have proved

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

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

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

Substitute $R = C\left(\frac{k\log n}{n}\right)^{1/2}, b = C\frac{k\log n}{n^{1/2}}$ in Theorem 3.3, we have $b \le \sqrt{n}R^2 \le \sqrt{n}R$, and

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

Hence,

$$\mathbb{P}_0\left(\sup_{\bar{h}_n\left(\bar{p}^{(n)}_{\boldsymbol{\tau},\mathbf{V},\underline{a},\sigma}, p_0^{(n)}\right) \le C\left(\frac{k\log n}{n}\right)^{1/2}} \left|\sqrt{n}(P_n - P_0)\log\left(\frac{\mathcal{N}(y_i|f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}(t_i), \widehat{\sigma}^2_{n,k}) + \mathcal{N}(y_i|f^0(t_i), \sigma_0^2)}{2\mathcal{N}(y_i|f^0(t_i), \sigma_0^2)}\right)\right| \ge C\frac{k\log n}{n^{1/2}}\right)$$
$$\le c_1 n^{-c_2}, \tag{3.37}$$

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

$$\mathbb{P}_0\left(\left|(P_n - P_0)\log\left(\frac{\mathcal{N}(y_i|f_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}}}(t_i), \widehat{\sigma}^2_{n,k}) + \mathcal{N}(y_i|f^0(t_i), \sigma_0^2)}{2\mathcal{N}(y_i|f^0(t_i), \sigma_0^2)}\right)\right| \ge C\frac{k\log n}{n}\right) \le 2c_1 n^{-c_2}, \tag{3.38}$$

Furthermore,

$$P_0\log\left(\frac{\mathcal{N}(y_i|f_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}}}(t_i), \widehat{\sigma}^2_{n,k}) + \mathcal{N}(y_i|f^0(t_i), \sigma_0^2)}{2\mathcal{N}(y_i|f^0(t_i), \sigma_0^2)}\right)$$
$$= -\frac{1}{n}\sum_{i=1}^{n} KL\left(\mathcal{N}(y_i|f^0(t_i), \sigma_0^2)||\frac{\mathcal{N}(y_i|f_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}}}(t_i), \widehat{\sigma}^2_{n,k}) + \mathcal{N}(y_i|f^0(t_i), \sigma_0^2)}{2}\right) \le 0. \tag{3.39}$$

This implies that

$$\mathbb{P}_0\left(P_n\log\left(\frac{\mathcal{N}(y_i|f_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}}}(t_i), \widehat{\sigma}^2_{n,k}) + \mathcal{N}(y_i|f^0(t_i), \sigma_0^2)}{2\mathcal{N}(y_i|f^0(t_i), \sigma_0^2)}\right) \le C\frac{k\log n}{n}\right) \le 1 - 2c_1 n^{-c_2}. \tag{3.40}$$

Together with (3.36), we have that

$$\mathbb{P}_0\left(P_n \log\left(\frac{\mathcal{N}(y_i|f_{\widehat{\boldsymbol{\tau}},\widehat{\mathbf{V}},\widehat{\underline{a}}}(t_i),\widehat{\sigma}^2_{n,k})}{\mathcal{N}(y_i|f^0(t_i),\sigma_0^2)}\right) \leq C\frac{k\log n}{n}\right) \leq 1 - 2c_1 n^{-c_2}. \tag{3.41}$$

In other words, for $k \geq k_0$, we have shown that $\overline{\mathcal{L}}_n(\widehat{f}_n^{(k)},\widehat{\sigma}^2_{n,k}) - \overline{\mathcal{L}}_n(f^0,\widehat{\sigma}_0^2) \leq C\frac{k\log(n)}{n}$ for some constant C depends on $d,\mathcal{V},\mathcal{A}$ 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. □

3.4 Proof of consistency of sSIC

Proof of Theorem 3.1. The optimal number of changepoint is estimated as

$$\widehat{k}_n = \arg\max_{k\leq\overline{k}}\left[\overline{\mathcal{L}}_n(\widehat{f}_n^{(k)},\widehat{\sigma}^2_{n,k}) - (k-1)\frac{(\log(n))^\gamma}{n}\right].$$

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 $n \to \infty$.

sSIC is not over-fitted. From Theorem 3.4, we have with probability tending to 1, for all $k \geq k_0$,

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

for some positive constant C depending on $\overline{k}, d$ and parameters' spaces. Therefore, for all $k > k_0$ (and $k < \overline{k}$), the increase in average log-likelihood when over-fitting can

be characterized as

$$\begin{aligned}\overline{\mathcal{L}}_n(\widehat{f}_n^{(k)},\widehat{\sigma}^2_{n,k})-\overline{\mathcal{L}}_n(\widehat{f}_n^{(k_0)},\widehat{\sigma}^2_{n,k_0}) &\leq \overline{\mathcal{L}}_n(\widehat{f}_n^{(k)},\widehat{\sigma}^2_{n,k})-\overline{\mathcal{L}}_n(f_0,\sigma_0^2)\\ &\leq Ck\frac{(\log(n))}{n}\\ &< (k-k_0)\frac{(\log(n))^\gamma}{n},\end{aligned}$$

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

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

for all sufficiently large n, implies that

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

as $n\to\infty$.

sSIC is not under-fitted. Notice that

$$\widehat{k}_n=\underset{k\leq\overline{k}}{\arg\max}\left[\overline{\mathcal{L}}_n(\widehat{f}_n^{(k)},\widehat{\sigma}^2_{n,k})-(k-1)\frac{(\log(n))^\gamma}{n}\right]=\underset{k\leq\overline{k}}{\arg\max}\left[-\log\left(\hat{\sigma}^2_{n,k}\right)-(k-1)\frac{(\log(n))^\gamma}{n}\right].$$

In this part, it is more convenient to use an equivalent formula from (3.5) as

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

From Theorem 3.4, 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 $k < k_0$, [93] (Lemma 5.4) showed that there exists a positive constant C depends on $\underline{C}_1$ and $\underline{C}_2$ such that

$$\frac{\sum_{i=1}^{n} \left\| y_i - \widehat{f}_n^{(k)}(t_i) \right\|^2}{dn} > \sigma_0^2 + C, \tag{3.42}$$

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 ([93], 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. As $C' > (k_0 - k)\dfrac{(\log(n))^\gamma}{n}$ for all sufficiently large n, we have

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

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

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

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

$$\mathbb{P}\left(\max_{i=1,\dots,k_0-1} |\widehat{\tau}_i - \tau_i^0| \le C\left(\frac{\log n}{n}\right)^{1/2} \middle| \widehat{k}_n = k_0\right) \to 1. \tag{3.43}$$

Recall from Theorem 3.4 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 $|\widehat{\tau}_i - \tau_i^0| \le C\left(\frac{\log n}{n}\right)^{1/2}$ $\forall i = 1, \dots, k_0$, then (3.43) will be proved.

We separate the proof into two steps.

Step 1. Prove consistency with rate $(\log n/n)^{1/3}$**.** We will do this by proving by contradiction. Indeed, assume that there exists $i \in [k_0]$ such that $\widehat{\tau}_j \notin [\tau_i^0 - \epsilon_n, \tau_i^0 + \epsilon_n]$, with $\epsilon_n \gg (\log n/n)^{1/3}$ for all $j \in [k_0]$. Without loss of generality, we can further assume that $\tau_{i-1}^0 < \tau_i^0 - \epsilon_n < \tau_i^0 + \epsilon_n < \tau_{i+1}^0$. Then,

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

In the interval $[\lfloor n(\tau_i^0 - \epsilon_n)\rfloor, \lfloor n(\tau_i^0 + \epsilon_n)\rfloor]$, the function $\widehat{f}_n^{(k_0)}$ is a linear function, meanwhile f^0 is a piecewise linear function with 2 pieces having slope V_{i-1}^0 and V_i^0,

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

$$\frac{1}{n}\sum_{j=\lfloor n(\tau_i^0-\epsilon_n)\rfloor}^{\lfloor n(\tau_i^0+\epsilon_n)\rfloor}\left\|\widehat{f}_n^{(k_0)}(j/n)-f^0(j/n)\right\|^2 \tag{3.44}$$

$$=\int_{\tau_i^0-\epsilon_n}^{\tau_i^0+\epsilon_n}\left\|\widehat{f}^{(k_0)}(t)-f^0(t)\right\|^2 dt+o(1/n) \tag{3.45}$$

Let $\tilde{f}$ be the best linear approximate of f^0 in the interval $[\tau_i^0-\epsilon_n,\tau_i^0+\epsilon_n]$. We then have that $\int_{\tau_i^0-\epsilon_n}^{\tau_i^0+\epsilon_n}\left\|\widehat{f}^{(k_0)}(t)-f^0(t)\right\|^2 dt \geq \int_{\tau_i^0-\epsilon_n}^{\tau_i^0+\epsilon_n}\left\|\tilde{f}(t)-f^0(t)\right\|^2 dt$.

$\tilde{f}$ minimizes the following loss function

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

where $g(t)$ is a linear function in the interval $[\tau_i^0-\epsilon_n,\tau_i^0+\epsilon_n]$.

Let $\Delta g=g-f^0$. We can rewrite the loss function as

$$\begin{aligned}\text{Loss}&=\int_{\tau_i^0-\epsilon_n}^{\tau_i^0}\left\|g(t)-f^0(t)\right\|^2 dt+\int_{\tau_i^0}^{\tau_i^0+\epsilon_n}\left\|g(t)-f^0(t)\right\|^2 dt\\&=\epsilon_n\left(\left\|\Delta g(\tau_i^0)\right\|^2+\langle\Delta g(\tau_i^0),\Delta g(\tau_i^0-\epsilon_n)\rangle+\left\|\Delta g(\tau_i^0-\epsilon_n)\right\|^2\right)\\&+\epsilon_n\left(\left\|\Delta g(\tau_i^0)\right\|^2+\langle\Delta g(\tau_i^0),\Delta g(\tau_i^0+\epsilon_n)\rangle+\left\|\Delta g(\tau_i^0+\epsilon_n)\right\|^2\right).\end{aligned} \tag{3.46}$$

Notice that $\Delta g(\tau_i^0)=\dfrac{\epsilon_n g(\tau_i^0-\epsilon_n)+\epsilon_n f(\tau_i^0+\epsilon_n)}{2\epsilon_n}-f^0(\tau_i^0)$, so we can consider $g(\tau_i^0-\epsilon_n)$ and $g(\tau_i^0+\epsilon_n)$ as two "free" parameters in Equation 3.46. Set the derivatives of Loss with respect to those free parameters to 0, we have

$$0\overset{\text{set}}{=}\frac{\partial\text{Loss}}{\partial g(\tau_i^0-\epsilon_n)}=\epsilon_n\left(2\Delta g(\tau_i^0)+\frac{5}{2}\Delta g(\tau_i^0-\epsilon_n)\right), \tag{3.47}$$

and

$$0 \overset{\text{set}}{=} \frac{\partial \text{Loss}}{\partial g(\tau_i^0 + \epsilon_n)} = \epsilon_n \left(2\Delta g(\tau_i^0) + \frac{5}{2} \Delta g(\tau_i^0 + \epsilon_n) \right), \tag{3.48}$$

From Equations (3.47) and (3.48), we have that $\Delta g(\tau_i^0 - \epsilon_n) = \Delta g(\tau_i^0 + \epsilon_n)$. Since $\tilde{f}$ minimize the loss function, it satisfies $\Delta\tilde{f}(\tau_i^0 - \epsilon_n) = \Delta\tilde{f}(\tau_i^0 + \epsilon_n)$ 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:

$$\begin{aligned} \Delta\tilde{f}(\tau_i^0) - \Delta\tilde{f}(\tau_i^0 - \epsilon_n) &= \tilde{f}(\tau_i^0) - f^0(\tau_i^0) - \tilde{f}(\tau_i^0 - \epsilon_n) + f^0(\tau_i^0 - \epsilon_n) \\ &= \epsilon_n \tilde{V} - \epsilon_n V_{i-1}^0 = \epsilon_n (\tilde{V} - V_{i-1}^0) \end{aligned} \tag{3.49}$$

$$\begin{aligned} \Delta\tilde{f}(\tau_i^0 + \epsilon_n) - \Delta\tilde{f}(\tau_i^0) &= \tilde{f}(\tau_i^0 + \epsilon_n) - f^0(\tau_i^0 + \epsilon_n) - \tilde{f}(\tau_i^0) + f^0(\tau_i^0) \\ &= \epsilon_n \tilde{V} - \epsilon_n V_i^0 = \epsilon_n (\tilde{V} - V_i^0) = \epsilon_n (\tilde{V} - V_{i-1}^0) + \epsilon_n (V_{i-1}^0 - V_i^0). \end{aligned} \tag{3.50}$$

These implies that $2\Delta\tilde{f}(\tau_i^0 + \epsilon_n) - 2\Delta\tilde{f}(\tau_i^0) = \epsilon_n (V_{i-1}^0 - V_i^0)$. Combining with Equation 3.48, we got $\Delta\tilde{f}(\tau_i^0 + \epsilon_n) = \Delta\tilde{f}(\tau_i^0 - \epsilon_n) = \frac{2}{9}\epsilon_n (V_{i-1}^0 - V_i^0)$. Then

$$\begin{aligned} \int_{\tau_i^0 - \epsilon_n}^{\tau_i^0 + \epsilon_n} \left\| \widehat{f}^{(k_0)}(t) - f^0(t) \right\|^2 dt &\geq \int_{\tau_i^0 - \epsilon_n}^{\tau_i^0 + \epsilon_n} \left\| \tilde{f}(t) - f^0(t) \right\|^2 dt \\ &\geq \epsilon_n \left\| \Delta\tilde{f}(\tau_i^0) + \frac{1}{2}\Delta\tilde{f}(\tau_i^0 + \epsilon_n) \right\|^2 + \frac{3}{4}\epsilon_n \left\| \Delta\tilde{f}(\tau_i^0 + \epsilon_n) \right\|^2 \\ &\geq \frac{1}{6}\epsilon_n^3 \left\| V_{i-1}^0 - V_i^0 \right\|^2 > \frac{1}{6}\epsilon_n^3 \underline{C}_2^2. \end{aligned} \tag{3.51}$$

This implies $C\frac{\log n}{n} \geq \frac{1}{6}\epsilon_n^3 \underline{C}_2^2$ which contradicts the assumption that $\epsilon_n \gg (\log n/n)^{1/3}$.

We finish showing

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

Step 2. Improve the rate to $(\log n/n)^{1/2}$**.** In the previous step, we proved that for every true changepoint τ_i^0, there is a $\hat{\tau}_i^n$ (for $i\in\{1,\dots,k_0\}$) will converge to it with the rate $(\log n/n)^{1/3}$. In this step, we will show that we can achieve a better rate for this convergence.

For any $i\in\{1,\dots,k_0\}$, 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 $\hat{f}$ each have two segments within 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

$$\frac{1}{n}\sum_{j=\lfloor nt_1\rfloor}^{\lfloor nt_2\rfloor}\left\|\widehat{f}_n^{(k_0)}(j/n)-f^0(j/n)\right\|^2=\int_{t_1}^{t_2}\left\|\hat{f}^{k_0}(t)-f^0(t)\right\|^2dt+o(1/n).\tag{3.52}$$

Let $\tilde{f}$ have a changepoint at τ_i^n, being the best two-piecewise linear approximation of f^0 in the interval $[t_1,t_2]$. We then have that $\int_{t_1}^{t_2}\left\|\hat{f}^{(k_0)}(t)-f^0(t)\right\|^2dt\geq\int_{t_1}^{t_2}\left\|\tilde{f}(t)-f^0(t)\right\|^2dt$. Using the same strategy as in Step 1, $\tilde{f}$ minimizes the following loss function

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

where $g(t)$ is a 2-piecewise linear in the interval $[t_1,t_2]$.

Let $\Delta g = g - f^0$. We can write the loss function as

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

Takes $g(t_1), g(\widehat{\tau}_i^n), g(t_2)$ as "free" parameters. Note that $\Delta g(\tau_i^0)$ depends on those parameters in the sense that $\Delta g(\tau_i^0) = \dfrac{(\tau_i^0 - \widehat{\tau}_i^n)g(\widehat{\tau}_i^n) + (t_2 - \tau_i^0)g(t_2)}{t_2 - \widehat{\tau}_i^n} - f^0(\tau_i^0)$, and that

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

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

Set the derivatives of Loss with respect to those free parameters to 0, we have

$$0 \overset{\text{set}}{=} \frac{\partial \text{Loss}}{\partial \widehat{f}(t_1)} = (\widehat{\tau}_i^n - t_1)[2(\Delta g(t_1)) + \Delta g(\widehat{\tau}_i^n)], \tag{3.53}$$

and

$$\begin{aligned}
0 \overset{\text{set}}{=} \frac{\partial \text{Loss}}{\partial \widehat{f}(\widehat{\tau}_i^n)} &= (\widehat{\tau}_i^n - t_1)[\Delta g(t_1) + 2(\Delta g(\widehat{\tau}_i^n))] \\
&+ (\tau_i^0 - \widehat{\tau}_i^n)\left[\left(2\frac{\tau_i^0 - \widehat{\tau}_i^n}{t_2 - \widehat{\tau}_i^n} + 1\right)(\Delta g(\tau_i^0)) + \left(\frac{\tau_i^0 - \widehat{\tau}_i^n}{t_2 - \widehat{\tau}_i^n} + 2\right)(\Delta g(\widehat{\tau}_i^n))\right] \\
&+ (t_2 - \tau_i^0)\left[2\frac{\tau_i^0 - \widehat{\tau}_i^n}{t_2 - \widehat{\tau}_i^n}(\Delta g(\tau_i^0)) + \frac{\tau_i^0 - \widehat{\tau}_i^n}{t_2 - \widehat{\tau}_i^n}(\Delta g(t_2))\right] \\
&= \left[\frac{3}{2}(\widehat{\tau}_i^n - t_1) + \left(\frac{\tau_i^0 - \widehat{\tau}_i^n}{t_2 - \widehat{\tau}_i^n} + 2\right)(\tau_i^0 - \widehat{\tau}_i^n)\right](\Delta g(\widehat{\tau}_i^n)) + 3(\tau_i^0 - \widehat{\tau}_i^n)(\Delta g(\tau_i^0)) \\
&+ \frac{(t_2 - \tau_i^0)(\tau_i^0 - \widehat{\tau}_i^n)}{t_2 - \widehat{\tau}_i^n}(\Delta g(t_2)),
\end{aligned} \tag{3.54}$$

and

$$
\begin{aligned}
0 \overset{\text{set}}{=} \frac{\partial \text{Loss}}{\partial \widehat{f}(t_2)} &= (\tau_i^0 - \widehat{\tau}_i^n)\left[\frac{t_2-\tau_i^0}{t_2-\widehat{\tau}_i^n}(\Delta g(\widehat{\tau}_i^n)) + 2\frac{t_2-\tau_i^0}{t_2-\widehat{\tau}_i^n}(\Delta g(\tau_i^0))\right] \\
&\quad + (t_2-\tau_i^0)\left[\left(2\frac{t_2-\tau_i^0}{t_2-\widehat{\tau}_i^n}+1\right)(\Delta g(\tau_i^0)) + \left(\frac{t_2-\tau_i^0}{t_2-\widehat{\tau}_i^n}+2\right)(\Delta g(t_2))\right] \\
&= \frac{(\tau_i^0-\widehat{\tau}_i^n)(t_2-\tau_i^0)}{t_2-\widehat{\tau}_i^n}(\Delta g(\widehat{\tau}_i^n)) + 3(t_2-\tau_i^0)(\Delta g(\tau_i^0)) \\
&\quad + (t_2-\tau_i^0)\left(\frac{\tau_i^0-\widehat{\tau}_i^n}{t_2-\widehat{\tau}_i^n}+2\right)(\Delta g(t_2)). \qquad (3.55)
\end{aligned}
$$

From Equation (3.54), we have

$$
\left[\frac{3}{2}\frac{(\widehat{\tau}_i^n - t_1)}{\tau_i^0-\widehat{\tau}_i^n} + \frac{\tau_i^0-\widehat{\tau}_i^n}{t_2-\widehat{\tau}_i^n} + 2\right](\Delta g(\widehat{\tau}_i^n)) + 3(\Delta g(\tau_i^0)) + \frac{t_2-\tau_i^0}{t_2-\widehat{\tau}_i^n}(\Delta g(t_2)) = 0. \qquad (3.56)
$$

Equation (3.55) implies

$$
\frac{t_2-\tau_i^0}{t_2-\widehat{\tau}_i^n}(\Delta g(\widehat{\tau}_i^n)) + 3(\Delta g(\tau_i^0)) + \left(\frac{\tau_i^0-\widehat{\tau}_i^n}{t_2-\widehat{\tau}_i^n}+2\right)(\Delta g(t_2)) = 0. \qquad (3.57)
$$

Add them up, we have

$$
\left[\frac{3}{2}\frac{(\widehat{\tau}_i^n - t_1)}{\tau_i^0-\widehat{\tau}_i^n} + 3\right](\Delta g(\widehat{\tau}_i^n)) + 6(\Delta g(\tau_i^0)) + 3(\Delta g(t_2)) = 0. \qquad (3.58)
$$

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 (3.53), (3.57) and (3.58) 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 finally get

$$\begin{aligned}
\tilde{V}_i - V_i^0 &= \frac{14(4t_2 - \tau_i^0 - \hat{\tau}_i^n)}{2(6\tau_i^0 - 9\hat{\tau}_i^n + 3t_2)(4t_2 - \tau_i^0 - \hat{\tau}_i^n) - 9(4\tau_i^0 - t_1 - 3\hat{\tau}_i^n)(t_2 - \hat{\tau}_i^n)}(\tau_i^0 - \hat{\tau}_i^n)(V_{i-1}^0 - V_i^0)\\
&\asymp (\tau_i^0 - \hat{\tau}_i^n)(V_{i-1}^0 - V_i^0). \qquad (3.59)
\end{aligned}$$

We have that

$$\begin{aligned}
C\frac{\log n}{n} \geq \int_{t_1}^{t_2} \left\|\Delta\tilde{f}(t)\right\|^2 dt &\geq (t_2 - \tau_i^0)\left[||\Delta\tilde{f}(\tau_i^0)||^2 + \langle\Delta\tilde{f}(\tau_i^0), \Delta\tilde{f}(t_2)\rangle + ||\Delta\tilde{f}(t_2)||^2\right] =\\
&(t_2 - \tau_i^0)\left[\left\|\Delta\tilde{f}(\tau_i^0) + \frac{1}{2}\Delta\tilde{f}(t_2)\right\|^2 + \frac{3}{4}||\Delta\tilde{f}(t_2)||^2\right]\\
&\gtrsim (\tau_i^0 - \hat{\tau}_i^n)^2||V_{i-1}^0 - V_i^0||^2 > (\tau_i^0 - \hat{\tau}_i^n)^2\underline{C}_2^2, \quad \text{for } i \in [k_0 - 1]. \qquad (3.60)
\end{aligned}$$

This implies that

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

□

Chapter 4

Testing for Stationary States in Intracellular Transport

While studying lysosomal transport and considering that the underlying motor dynamics, cytoskeleton, and cytosolic crowding are unknown, it is beneficial to examine the two "states": *motile* and *stationary* [29, 111]. The true stationary state in our study here is understood in the Mathematical sense that the particle's distribution within this state remains invariant over time. Meanwhile, in the work of Rayens et al. [111], the authors use the stationary term to suggest that the particles are not moving excessively, defined by a threshold of under 100nm/s of the considered segment speed. In this chapter, we establish a test for stationary states of a particle trajectory under the assumption that the data is generated according to the piecewise linear continuous model. Motivated by the 2D datasets we are using, the tests will be provided with details for $d = 2$. However, the idea can be easily extended to higher dimensions.

We then explore the effectiveness of the test under various circumstances, particularly when applying it to the model with inferred changepoints from CPLASS and in cases where the true anchor locations are not linear but diffuse due to a second source of noise, which we refer to as the anchor diffusing model.

4.1 The matrix form of the model

We revisit Section 2.2.2, where the matrix presentation of the model is introduced for each dimension (see Equation (2.8)). For $d = 2$ and $\{Y_i\}_{i=1}^n \subset \mathbb{R}^2$ denote n observed data, given $k-1$ known time of changepoints τ_j (for $j = 1, \ldots, k-1$), we then have

$$Y^{(1)} = \mathbb{T}\underline{W}^{(1)} + \sigma\varepsilon^{(1)}, \tag{4.1}$$

$$Y^{(2)} = \mathbb{T}\underline{W}^{(2)} + \sigma\varepsilon^{(2)}, \tag{4.2}$$

where $\mathbb{T}$ is a $n \times (k+1)$ matrix (2.9), $\varepsilon^{(1)}, \varepsilon^{(2)} \sim \mathcal{N}(0, I_n)$, $Y^{(1)} = (y_{11}, \ldots, y_{n1})^\top$, $Y^{(2)} = (y_{12}, \ldots, y_{n2})^\top$ are $n \times 1$ vectors of observations in each corresponding dimension, $\underline{W}^{(1)} = \left(\underline{a}_1, w_{11}, ..., w_{k1}\right)^\top$, and $\underline{W}^{(2)} = \left(\underline{a}_2, w_{12}, ..., w_{k2}\right)^\top$ are $(k+1) \times 1$ vectors of parameters of the model in each corresponding dimension. The velocity vector associated to the jth segment is $V_j = \left(\sum_{i=1}^j w_{i1}, \sum_{i=1}^j w_{i2}\right)^\top$, for $j = 1, ..., k$.

For simplicity, we can write the two-dimensional model as the following combined-matrix form

$$\mathscr{Y} = \mathscr{T}\mathbf{w} + \sigma\epsilon, \tag{4.3}$$

where $\mathscr{Y} = (y_{11}, \ldots, y_{n1}, y_{12}, \ldots, y_{n2})^\top$ is a $2n \times 1$ vector of observations, $\mathscr{T} = \begin{bmatrix} \mathbb{T} & \mathbf{0} \\ \mathbf{0} & \mathbb{T} \end{bmatrix}$ is a $2n \times (2k+2)$ matrix containing four block matrices, a $(2k+2) \times 1$ vector of unknown regression parameters $\mathbf{w} = (\underline{a}_1, w_{11}, \ldots, w_{k1}, \underline{a}_2, w_{12}, \ldots, w_{k2})^\top$, and $\epsilon = (\epsilon_1, \ldots, \epsilon_{2n})^\top \sim \mathcal{N}(0, \sigma^2 I_{2n})$ is a $2n \times 1$ vector of random errors. As discussed in Section 2.2.2, as long as those conditions hold, $\mathscr{T}$ is full rank with the rank of $2k+2$.

The residual sum-of-squares (RSS) is written as

$$\mathrm{RSS}(Y,t;\mathbf{w}) := \|\mathscr{Y} - \mathscr{T}\mathbf{w}\|_2^2 = (\mathscr{Y} - \mathscr{T}\mathbf{w})^\top(\mathscr{Y} - \mathscr{T}\mathbf{w}). \tag{4.4}$$

The log-likelihood associated with the model defined in equations (4.3) is

$$\mathcal{L}(f,\sigma) = \sum_{i=1}^{n} \log \mathcal{N}(y_i | f_{\boldsymbol{\tau},\mathbf{V},\underline{a}}(t_i), \sigma^2 I_d) = -\left[\log(2\pi) + \log(\sigma^2)\right] - \frac{1}{2\sigma^2}\mathrm{RSS}(Y,t;\mathbf{w}) \tag{4.5}$$

The resulting MLEs are

$$\hat{\mathbf{w}} = \left(\mathscr{T}^\top\mathscr{T}\right)^{-1}\mathscr{T}^\top\mathscr{Y}, \tag{4.6}$$

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}(Y,t;\hat{\mathbf{w}})}{2n} = \frac{1}{2n}\mathscr{Y}^\top\left[I - \mathscr{T}\left(\mathscr{T}^\top\mathscr{T}\right)^{-1}\mathscr{T}^\top\right]\mathscr{Y}. \tag{4.7}$$

The velocity vector corresponding to the kth segment is $\widehat{V}_j = \left(\sum_{i=1}^{j}\widehat{w}_{i1}, \sum_{i=1}^{j}\widehat{w}_{i2}\right)^\top$, for $j = 1, ..., k$. The speed of the associated segment is $\hat{s}_j := \|\widehat{V}_j\|_2$.

4.2 The test for stationary segments

We want to determine whether a segment is stationary, which means the hypothesis test is $H_0 : s_{j^*} = 0$ against $H_1 : s_{j^*} \neq 0$, where j^* represents the considered segment. Assume that there are n observations $\{Y_i\}_{i=1}^n \subset \mathbb{R}^2$ of the considering k segments. Let the corresponding segments velocities be $\{V_j\}_{j=1}^k \subset \mathbb{R}^2$ and τ_j $(j = 1, ..., k-1)$ be the time of the change points. Since $s_j = \|V_j\|_2 = \sqrt{\left(\sum_{i=1}^{j} w_{i1}\right)^2 + \left(\sum_{i=1}^{j} w_{i2}\right)^2}$, we have that $s_j = 0$ equivalent to $\left(\sum_{i=1}^{j} w_{i1}\right)^2 = \left(\sum_{i=1}^{j} w_{i2}\right)^2 = 0$. Hence, the test is then

$$H_0 : \left(\sum_{i=1}^{j^*} w_{i1}\right)^2 = \left(\sum_{i=1}^{j^*} w_{i2}\right)^2 = 0 \quad \text{vs.} \quad H_1 : \left(\sum_{i=1}^{j^*} w_{i1}\right)^2 \neq 0 \vee \left(\sum_{i=1}^{j^*} w_{i2}\right)^2 \neq 0.$$

According to the defined model (4.3), this test can be rewritten to test for m linear parametric functions of parameters which is in the form:

$$H_0 : \mathbf{Lw} = \mathbf{0} \quad \text{vs.} \quad H_1 : \mathbf{Lw} \neq \mathbf{0}, \tag{4.8}$$

where $\mathbf{w} = (\underline{a}_1, w_{11}, \ldots, w_{k1}, \underline{a}_2, w_{12}, \ldots, w_{k2})^\top$, $\mathbf{L} = (L_1, L_2)^\top$ is a $2 \times (2k+2)$ matrix, and a $(2k+2) \times 1$ vector of known constants $L_j = (l_{j1}, l_{j2}, \ldots, l_{j(2k+2)})$ $(j = 1, 2)$. Specifically, $\mathbf{L} = (l_{ij})$ with $l_{1i} = l_{2i} = 1$ for $i = 1, \ldots, j^*$ and $l_{ij} = 0$ elsewhere.

For example,

- For the first segment, $H_0 : w_{11} = w_{12} = 0$. This is the same as $H_0 : \mathbf{Lw} = \mathbf{0}$, where $\mathbf{L} = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 1 & 0 & \ldots & 0 \end{bmatrix}$.
- For the second segment, $H_0 : w_{11} + w_{21} = w_{12} + w_{22} = 0$. This is the same as $H_0 : \mathbf{Lw} = \mathbf{0}$, where $\mathbf{L} = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 & \ldots & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 1 & 1 & \ldots & 0 \end{bmatrix}$.

With this presentation, transfer the test for $s^{j^*} = 0$ to test the model parameters' linear functions.

We recall some basic concepts to develop tools for solving this test.

4.2.1 Estimable function

Testable hypotheses are closely related to estimable functions [121]. Since we want to test (4.8), it is important to be able to estimate $L_j\mathbf{w}$ (for $j = 1, 2$). We say a linear function $L\mathbf{w}$ of the parameters in $\mathbf{w}$ is an estimable function if there exists some linear combination of the observations Y whose expected value is $L\mathbf{w}$. If $\mathscr{T}$ is full rank as we consider here, then all linear parametric function in $\mathbf{w}$ are estimable, and it implies $L_j\mathbf{w}$ for $j = 1, 2$ are estimable. Specifically, consider $L_j = (l_{j1}, l_{j2}, \ldots, l_{j(2k+2)})$ is a non-null vector, then $L_j\widehat{\mathbf{w}}$ is the best linear unbiased estimator of $L_j\mathbf{w}$ in the sense of

having minium variance as well as maximum likelihood under the normal distribution assumption of the residual (Gauss - Markov theorem).

In general cases where the design matrix is not of full rank, some linear parametric functions do not admit the unbiased linear estimator, and nothing can be inferred about them. The linear parametric functions that are not estimable are said to be *confounded*. In this case, $L\mathbf{w}$ need to be a linear combination of the rows of $\mathscr{T}$ to be estimable.

4.2.2 Theorems on quadratic forms in normal variables.

We will use the following theorems to develop the hypothesis test. We refer to [33, 34, 54, 62, 63, 110, 117] for more details on the analysis of linear models. Here, we present helpful theorems along with the proofs necessary for our test development.

Theorem 4.1

If the $n \times 1$ vector $y \sim \mathcal{N}(0, \sigma^2 I)$ and M is an $n \times n$ symmetric idempotent matrix of rank m then

$$y^\top \frac{M}{\sigma^2} y \sim \chi^2(m)$$

Proof of Theorem 4.1. Since M is symmetric, it can be diagonalized with an orthogonal matrix U. Furthermore, since M is idempotent, it has eigenvalues that are either 0 or 1. Hence, we can choose U so that

$$U^\top M U = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}. \tag{4.9}$$

The dimension of the identity matrix will be equal to the rank of M, since the number of non-zero roots in the rank of the matrix. Since the sum of the roots is equal to the trace, the dimension is also equal to the trace of M. Now let $z = \frac{1}{\sigma} U^\top y$. Compute

the moments of $z = \frac{1}{\sigma} U^\top y$

$$\begin{aligned}
\mathbb{E}(z) &= \frac{1}{\sigma} U^\top \mathbb{E}(y) = 0, \\
\mathrm{Var}(v) &= \frac{1}{\sigma^2} U^\top \sigma^2 I U = U^\top U = I \quad \text{since } U \text{ is orthogonal,}
\end{aligned}$$

Together with the assumption that y is normally distributed, we conclude that $z \sim \mathcal{N}(0, I)$.

Now consider the distribution of $y^\top \frac{M}{\sigma^2} y$ using the transformation z. Since U is orthogonal, its inverse is equal to its transpose. This means that $y = \sigma \left(U^\top\right)^{-1} z = \sigma U v$. Now, write the quadratic form as follows

$$\frac{y^\top M y}{\sigma^2} = z^\top U^\top M U z = z^\top \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} z = \sum_{i=1}^{m} z_i^2$$

This is the sum of squares of m standard normal variables and so is a χ^2 variable with m degrees of freedom.

□

Theorem 4.2

If $y \sim \mathcal{N}(0, \sigma^2 I)$, M is a symmetric idempotent matrix of order n, and L is a $k \times n$ matrix, then Ly and $y^\top M y$ are independent distributed if $LM = 0$

Proof of Theorem 4.2. Define the matrix U as before so that

$$U^\top M U = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}. \tag{4.10}$$

Let $rank(M) = m$, $z = U^\top y$ and partition z as follows $z = (\underbrace{z_1, z_2, \ldots, z_m}_{z^{(1)}}, \underbrace{z_{m+1}, \ldots, z_n}_{z^{(2)}})$. The number of elements of $z^{(1)}$ is m, and z_2 contains $n - m$ elements. We have $z^{(1)}$ and $z^{(2)}$ are independent since they are independent standard normals.

In the next, we will show that $y^\top My$ depends only on $z^{(1)}$ and Ly depends only on $z^{(2)}$. Given that z_i are independent, $y^\top My$ and Ly will be independent. We have that $y^\top My = z^\top \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} z = (z^{(1)})^\top z^{(1)}$.

Considering $LU(U^\top MU) = LMU = 0$ since $(LM = 0)$, and $LU(U^\top MU) = (C_1, C_2)\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} = 0$, where LU is partition as (C_1, C_2), C_1 has k rows and m columns. C_2 has k rows and $n-m$ columns. This implies that $C_1 = 0$, then $LU = (0, C_2)$. Now $Ly = LUU^\top y = LUz = C_2 z^{(2)}$ (since U is orthogonal and $C_1 = 0$).

We proved that Ly depends only on $z^{(2)}$, and $y^\top My$ depends only on $z^{(1)}$. Since $z^{(1)}$ and $z^{(2)}$ are independent, so are Ly and $y^\top My$. □

Theorem 4.3: Craig's Theorem

If $y \sim \mathcal{N}(\mu, \Sigma)$ where Σ is positive definite, then $q_1 = y^\top Ay$ and $q_2 = y^\top By$ are independently distributed if $A\Sigma B = 0$.

Proof of Theorem 4.3. We refer to [33, 34] for details. □

Theorem 4.4

Let $\mathscr{Y} = (y_{11}, \ldots, y_{n1}, y_{12}, \ldots, y_{n2})^\top$ follows a multivariate normal distribution $\mathcal{N}(\mu, \sigma^2 I_{2n})$, then the maximum likelihood estimator $L\hat{\mathbf{w}}$ of estimable linear parametric function is independently distributed of $\hat{\sigma}^2$; $L\hat{\mathbf{w}} \sim \mathcal{N}\left(L\mathbf{w}, \sigma^2 L(\mathscr{T}^\top \mathscr{T})^{-1} L^\top\right)$ and $\dfrac{2n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(2n - 2k - 2)$ where $rank(\mathscr{T}) = 2k + 2$.

Proof of Theorem 4.4. We have that

$$\hat{\mathbf{w}} = \left(\mathscr{T}^\top \mathscr{T}\right)^{-1} \mathscr{T}^\top \mathscr{Y} = \left(\mathscr{T}^\top \mathscr{T}\right)^{-1} \mathscr{T}^\top (\mathscr{T}\mathbf{w} + \epsilon) = \mathbf{w} + \left(\mathscr{T}^\top \mathscr{T}\right)^{-1} \mathscr{T}^\top \epsilon,$$

then

$$\begin{aligned}
\mathbb{E}(L\hat{\mathbf{w}}) &= L\left(\mathscr{T}^\top\mathscr{T}\right)^{-1}\mathscr{T}^\top\mathbb{E}(\mathscr{Y}) = L\left(\mathscr{T}^\top\mathscr{T}\right)^{-1}\mathscr{T}^\top\mathscr{T}\mathbf{w} = L\mathbf{w},\\
\mathrm{Var}(L\hat{\mathbf{w}}) &= L\mathrm{Var}(\hat{\mathbf{w}})L^\top = L\mathrm{Var}\left[\left(\mathscr{T}^\top\mathscr{T}\right)^{-1}\mathscr{T}^\top\epsilon\right]L^\top\\
&= L\left(\mathscr{T}^\top\mathscr{T}\right)^{-1}\mathscr{T}^\top\mathrm{Var}(\epsilon)\mathscr{T}\left(\mathscr{T}^\top\mathscr{T}\right)^{-1}L^\top\\
&= \sigma^2 L\left(\mathscr{T}^\top\mathscr{T}\right)^{-1}\mathscr{T}^\top I_{2n}\mathscr{T}\left(\mathscr{T}^\top\mathscr{T}\right)^{-1}L^\top\\
&= \sigma^2 L\left(\mathscr{T}^\top\mathscr{T}\right)^{-1}L^\top.
\end{aligned}$$

Since $\hat{\mathbf{w}}$ is a linear function of Y and $L\hat{\mathbf{w}}$ is a linear function of $\hat{\mathbf{w}}$, so $L\hat{\mathbf{w}} \sim \mathcal{N}\left(L\mathbf{w}, \sigma^2 L(\mathscr{T}^\top\mathscr{T})^{-1}L^\top\right)$.

Let $A = I - \mathscr{T}\left(\mathscr{T}^\top\mathscr{T}\right)^{-1}\mathscr{T}^\top$ and $B = L\left(\mathscr{T}^\top\mathscr{T}\right)^{-1}\mathscr{T}^\top$, then

$$\begin{aligned}
L\hat{\mathbf{w}} &= L\left(\mathscr{T}^\top\mathscr{T}\right)^{-1}\mathscr{T}^\top\mathscr{Y} = B\mathscr{Y},\\
\frac{2n\tilde{\sigma}^2}{\sigma^2} &= \mathscr{Y}^\top\left[I - \mathscr{T}\left(\mathscr{T}^\top\mathscr{T}\right)^{-1}\mathscr{T}^\top\right]\mathscr{Y} = \mathscr{Y}^\top\frac{A}{\sigma^2}\mathscr{Y}.
\end{aligned}$$

Notice that A is a symmetric, idempotent matrix of rank $2n - 2k - 2$. Using Theorem 4.1, we have $\dfrac{2n\tilde{\sigma}^2}{\sigma^2} \sim \chi^2(2n - 2k - 2)$.

Also, $BA = L\left(\mathscr{T}^\top\mathscr{T}\right)^{-1}\mathscr{T}^\top\left[I - \mathscr{T}\left(\mathscr{T}^\top\mathscr{T}\right)^{-1}\mathscr{T}^\top\right] = 0$. So using Theorem 4.2, we conclude that $\mathscr{Y}^\top A\mathscr{Y}$ and $B\mathscr{Y}$ are independently distributed.

□

4.2.3 Returning to the hypothesis test

We aim to develop a hypothesis test (4.8) concerning the two linear functions $L_1\mathbf{w}$ and $L_2\mathbf{w}$, and we find that they are estimable based on the previous argument. We have the maximum likelihood estimator $L_j\hat{\mathbf{w}}$ $(j = 1, 2)$ where $\hat{\mathbf{w}} = \left(\mathscr{T}^\top\mathscr{T}\right)^{-1}\mathscr{T}^\top\mathscr{Y}$ $\mathbf{L}\hat{\mathbf{w}} = (L_1\hat{\mathbf{w}}, L_2\hat{\mathbf{w}})^\top$ is a 2×1 vector. Theorem 4.4 implies that $\mathbf{L}\hat{\mathbf{w}}$ is a 2-dimensional

Gaussian random vector with mean vector $\mathbb{E}(\mathbf{L}\widehat{\mathbf{w}}) = \mathbf{L}\mathbf{w}$ and rank 2 covariance matrix $\text{Cov}(\mathbf{L}\widehat{\mathbf{w}}) = \sigma^2\mathbb{V}$ where $\mathbb{V} = (v_{ij})_{i,j=1,2}$ with $v_{ij} = L_i(\mathscr{T}^\top\mathscr{T})^{-1}L_j^\top$ is the $(i,j)^{\text{th}}$ element of $\mathbb{V}$. From here, under the null, we have $\dfrac{(\mathbf{L}\widehat{\mathbf{w}}-\mathbf{0})^\top\mathbb{V}^{-1}(\mathbf{L}\widehat{\mathbf{w}}-\mathbf{0})}{\sigma^2}$ follows a χ^2 - distribution with 2 degree of freedom. Also, by Theorem 4.4, $\dfrac{2n\widehat{\sigma}^2}{\sigma^2}$ follows a χ^2 - distribution with $(2n-(2k+2))$ degrees of freedom where $\widehat{\sigma}^2 = \text{RSS}/(2n)$ is the maximum likelihood estimator of σ^2. And by Theorem 4.3, $\dfrac{(\mathbf{L}\widehat{\mathbf{w}}-\mathbf{0})^\top\mathbb{V}^{-1}(\mathbf{L}\widehat{\mathbf{w}}-\mathbf{0})}{\sigma^2}$ and $\dfrac{2n\widehat{\sigma}^2}{\sigma^2}$ are independently distributed.

Under the null hypothesis $H_0 : \mathbf{L}\mathbf{w} = \mathbf{0}$, we have the statistic

$$F = \frac{2n-2k-2}{2}\frac{(\mathbf{L}\widehat{\mathbf{w}}-\mathbf{0})^\top\mathbb{V}^{-1}(\mathbf{L}\widehat{\mathbf{w}}-\mathbf{0})}{2n\widehat{\sigma}^2}$$

follows F-distribution with 2 and $(2n-2k-2)$ degrees of freedom.

So the null hypothesis is rejected whenever $F \geq F_{1-\alpha}(2, 2n-2k-2)$ where $F_{1-\alpha}(2, 2n-2k-2)$ denotes the $100\alpha\%$ points on F-distribution with 2 and $(2n-2k-2)$ degrees of freedom. The test also returns the p-value, which is $P(F_{2,2n-2k-2} \geq F)$.

4.3 Effectiveness of the test

In this section, we apply our test for stationary segments to both simulation and real data. Note that the test is constructed under the assumption that the true changepoint locations are known and that the data follow the continuous piecewise linear regression model. However, in reality, we do not know the location of the changepoint, and the data typically do not perfectly adhere to the continuous piecewise linear model. We aim to investigate how the test performs with the inferred changepoints from CPLASS, especially when the actual anchor locations are not straight lines. Given the consistency theorem of CPLASS discussed in Chapter 3, we have mathematical

support that when the sample size is large enough, the inferred number of changepoints and their locations will be close to the truth. Thus, the test can still operate effectively with the output from CPLASS without compromising its power. However, with the diffusion of the anchor locations discussed in Section 2.4, CPLASS faces challenges with false changepoints. Will the *Motile/Stationary* test also be impacted? The subsequent sections will analyze the simulation data to validate these considerations.

4.3.1 On simulation data

We utilize simulation to investigate the following two cases: (1) at what value of the segment speed will the test indicate motility, and (2) at what level of anchor noise in the anchor diffusing scenario will false positives increase.

We use the idea in [29] of *Inference Gap* - the percentage of mislabeled time steps. For simulated data, we can assess what proportion of the time the inference protocol yields mislabeled states. Let $\{J_i\}_{i=1}^n$ and $\{\widehat{J}_i\}_{i=1}^n$ be the true labeled and inferred labeled of *Stationary/Motile* in each observation time t_i (0 for Stationary/ 1 for Motile), respectively. The Inference Gap of a single path is defined as

$$\text{Inference gap (Pathwise)} := \frac{100}{n}\sum_{i=1}^{n}\mathbb{1}\{\widehat{J}_i = J_i\}. \tag{4.11}$$

In the following simulations, we used the same generating rules: (a) For each varied value of segment speed/ second source of noise ξ, 200 trajectories were simulated at 20Hz ($\Delta = 0.05$) with 14 seconds ($n = 280$), $\sigma = 0.2$, (b) there were two actual changepoint times at $2s$ and $12s$, the corresponding segment times are $(2, 10, 2)$ seconds, (c) the actual segment velocities for the first and the third segments were $V_1 = (0.5, -0.5)^\top$ and $V_3 = (0.5, 0.5)^\top$. For scenario (1), we varied the middle velocities V_2, while in the second scenario, we kept $V_2 = (0, 0)^\top$ and varied the value of ξ.

Simulation under the continuous piecewise linear model - Vary the middle segment speed

In this experiment, we adjusted the speed of the middle segment from 0 to $0.1\mu m/s$. Based on the simulation rules previously outlined, the first and third segments were motile. When $s_2 = 0\mu m/s$, the middle segment is classified as stationary; in all other cases, it is considered motile. After running CPLASS to detect changepoints, we used the test to label *Stationary/Motile* segments. The left panel of Figure 4.1 displays the estimated time spent motile resulting from the test on the segmentation by CPLASS. The red dashed line shows the true proportion of time spent motile. For $s_2 = 0$, the test labels are correct, nearly 100% for all motile and stationary segments. For speeds s_2 greater than 0.025 seconds, the true proportion lies within the 95% bootstrap confidence interval, even aligning with the true proportion. In the center panel of Figure 4.1, the false positive rate decreases while the false negative rate remains at 0 as the speed s_2 increases from $0.01\mu m/s$ to $0.1\mu m/s$. This shows the test can detect the actual stationary ($s_2 = 0$) and motile segments with the corresponding speed of over $0.025\mu m/s$ accurately. The left panel displays that the average pathwise Inference Gap (red dots) decreases as the speed increases from $0.01\mu m/s$ to $0.1\mu m/s$, which once again confirms the effectiveness of the test.

We will now return to the concern regarding how the test performs when running on the inferred segments from CPLASS compared to the actual segments. The same summarized procedure is presented in Figure 4.2, where we conduct the test on the actual segments. We can see how it aligns with the test output running on the inferred segments from CPLASS. This confirms the thought discussed at the beginning of the section.

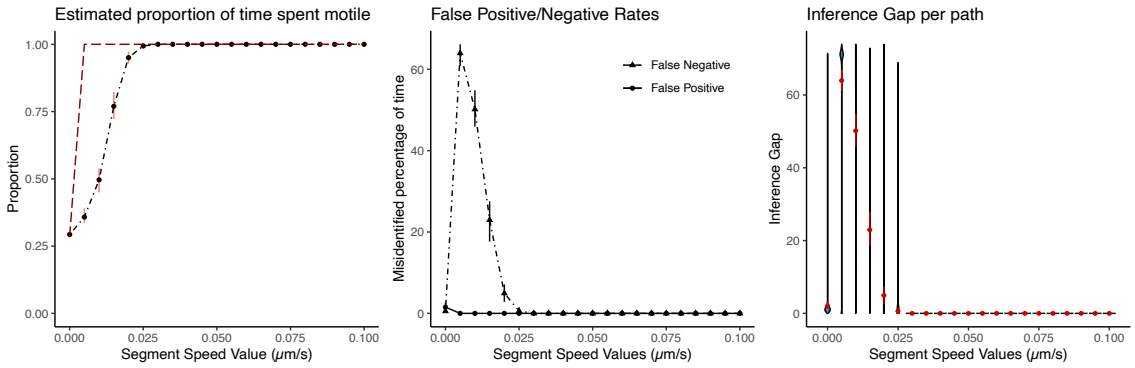


Figure 4.1: *Simulated data sets - Vary the middle segment speed. Applying the test to the inferred segments from CPLASS.* (**Left**) Estimates for the proportion of time spent *Motile* across different values of the middle segment speed values from 0 to $0.1\mu m/s$ using the *Stationary/Motile* test. The red dashed line is for the true proportion of time spent *Motile*. (**Center**) Ensemble averages for the percentage of mislabeled time steps, broken down into False Positives and False Negatives. Error bars computed from bootstrap resampling of paths. (**Right**) The path-by-path distribution of the percentage of time spent misidentified in the *Stationary/Motile* dichotomy with the test.

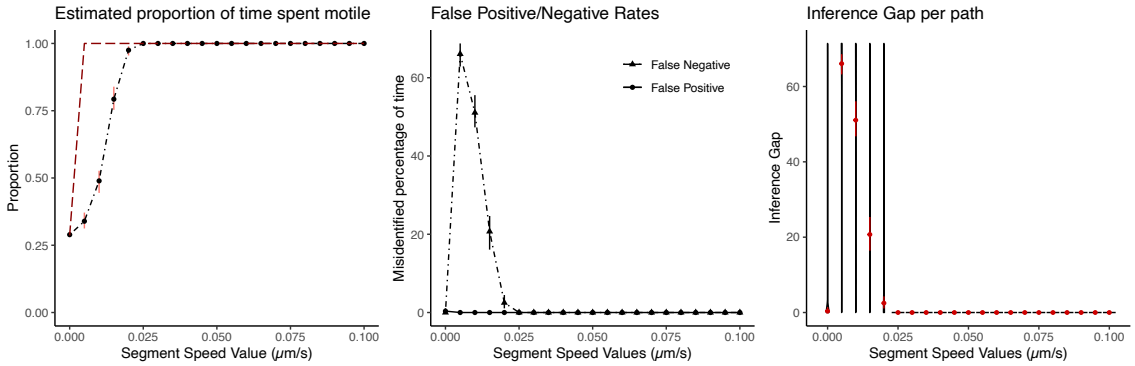


Figure 4.2: *Simulated data sets - Vary the middle segment speed. Applying the test to the actual segments .* (**Left**) Estimates for the proportion of time spent *Motile* across different values of the middle segment speed values from 0 to $0.1\mu m/s$ using the *Stationary/Motile* test. The red dashed line is for the true proportion of time spent *Motile*. (**Center**) Ensemble averages for the percentage of mislabeled time steps, broken down into False Positives and False Negatives. Error bars computed from bootstrap resampling of paths. (**Right**) The path-by-path distribution of the percentage of time spent misidentified in the *Stationary/Motile* dichotomy with the test.

False positives from Anchor diffusing - Vary the ξ value

Since the anchor location will not ideally be straight in the real data, as discussed in Section 2.4, when the second source of noise is present in the anchor locations (see Equation (2.20)), the actual stationary segment ($s = 0$) exhibits Brownian motion. This leads CPLASS to detect more changepoints due to increased fluctuations around the anchor locations. We aim to examine the level of this second source of noise ξ, at which the test labels the segment as motile, even though the true speed is $0\mu m/s$.

In this experiment, we varied the value of ξ from 0.01 to 0.1 while keeping the middle segment stationary (i.e., $s_2 = 0\mu m/s$). For each value of ξ, we simulated 200 trajectories with the same rule discussed at the beginning of the section. We then ran CPLASS to detect changepoints and apply the test to label *Stationary/Motile* segments. The same summary of the segmentation/classification protocol discussed in the vary speed experiment is applied here and reported in Figure 4.3. All three panels in Figure 4.3 show that the noisier this second source is, the bigger the Inference Gap. When $s_2 = 0$, the particle's movement during this period follows Brownian motion. When ξ increases from 0.01 to 0.1, the path will look more like moving than stationary, increasing the *false positive* rates shown in the center panel of the figure. When it comes to detecting actual motile segments, the test still does a good job of keeping the *false negative* rates to 0.

We returned to assess the test's performance on the actual segments. We are aware of a slightly overfitted number of changepoints issue with CPLASS when ξ increases from 0.05 to 0.1, as reported in Figure 2.12 in Section 2.4. However, this does not significantly impact the labeling results, as shown in Figure 4.3 and Figure 4.4. This confirms that running the Motile/Stationary test on CPLASS output will not compromise its effectiveness. The main challenge arises from the Brownian motion aspect, where increasing ξ can move the anchor away from the horizontal straight line

and lead the test to conclude that the segment is motile mistakenly.

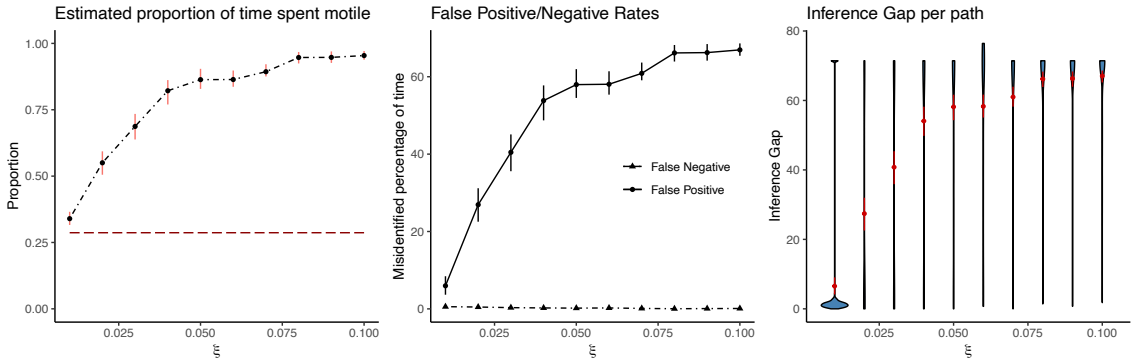


Figure 4.3: *Simulated data sets - Vary the values of ξ. Applying the test to the inferred segments from CPLASS.* (**Left**) Estimates for the proportion of time spent *Motile* across different values of ξ from 0.01 to 0.1 using the *Stationary/Motile* test. The red dashed line is for the true proportion of time spent *Motile*. (**Center**) Ensemble averages for the percentage of mislabeled time steps, broken down into False Positives and False Negatives. Error bars computed from bootstrap resampling of paths. (**Right**) The path-by-path distribution of the percentage of time spent misidentified in the *Stationary/Motile* dichotomy with the test.

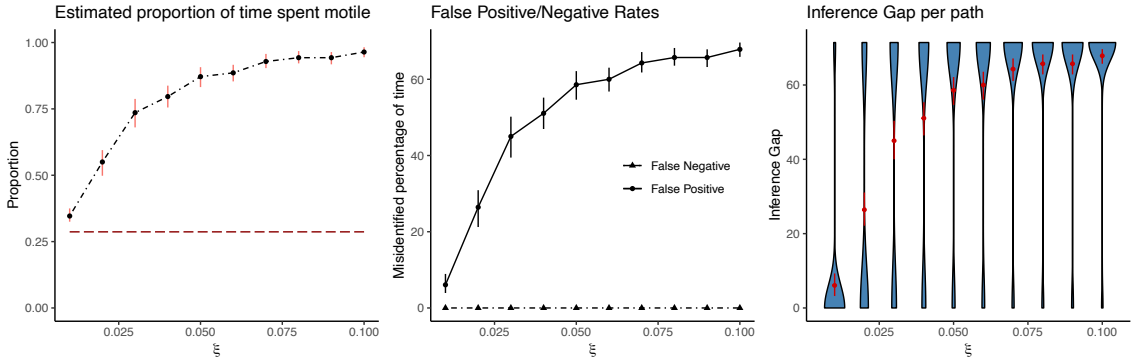


Figure 4.4: *Simulated data sets - Vary the values of ξ. Applying the test to the actual segments.* (**Left**) Estimates for the proportion of time spent *Motile* across different values of ξ from 0.01 to 0.1 using the *Stationary/Motile* test. The red dashed line is for the true proportion of time spent *Motile*. (**Center**) Ensemble averages for the percentage of mislabeled time steps, broken down into False Positives and False Negatives. Error bars computed from bootstrap resampling of paths. (**Right**) The path-by-path distribution of the percentage of time spent misidentified in the *Stationary/Motile* dichotomy with the test.

4.3.2 In vitro experimental data - quantum dot transport

In this section, we ran the test on 101 CPLASS segmented trajectories of quantum dots transported by a Kinesin-1/DDB (Kin1DDB) pair ([71]). As discussed in Section 2.3.4, the Kin1DDB data associated with the tug-of-war scenario, where both motors are present, keeps the dots moving more slowly. There are *long-slow* segments in this dataset. We want to investigate how the test works on returning labels for these *long-slow* segments, i.e., segment duration lasts more than $10s$ and segment speed is close to $0\mu m/s$. The test output demonstrates the active behavior of the quantum dots transported by the Kinesin-1/DDB pair, with 93.3% of the collected segments labeled as motile. Figure 4.5 indicates that the test labels some *long-slow* segments as *motile* (see segment durations ranging from longer than 5 seconds to 25 seconds, with corresponding speeds lower than $0.1\mu m/s$). These segments will be labeled as *stationary* using the $0.1\mu m/s$ cutoff employed in Rayens et al.'s paper ([111]). We zoom in on this area where all the segment speeds are under $0.1\mu m/s$ in Panel (B) of the figure; we also report the error bars associated with estimated segment speeds, which are calculated based on the Standard Errors of estimating $\mathbf{w}$ and under the assumption of normally distributed on the error. Panel (B) explains why the test labels the long, slow segments in this data as motile. A shorter error bar indicates a more precise measurement of the coefficient. When the error bar does not intersect the $s = 0$ line, the test will definitively reject the stationary hypothesis.

We extracted the trajectories with long, slow segments for closer examination. Figure 4.6 illustrates one of these paths. The noise associated with this path is relatively low, leading to a more accurate velocity estimation. This explains why the test shows strong agreement (with very small p-values) that the quantum dots in these segments are motile. While the cut-off threshold of $100nm/s$ consistently indicates that these segments are *Stationary*, without accounting for the noise information or

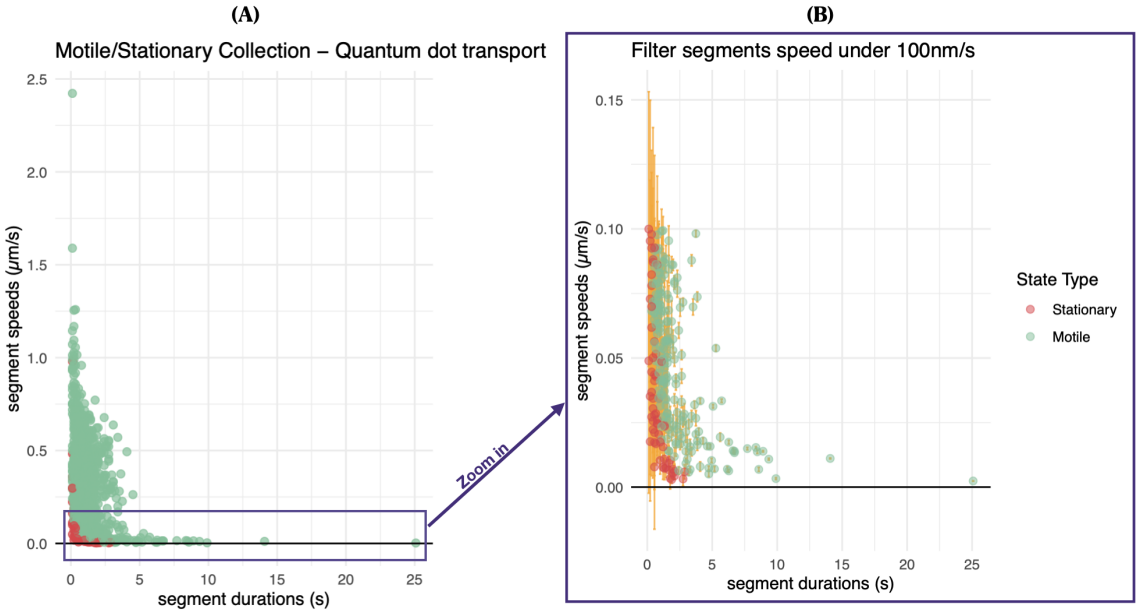


Figure 4.5: *Quantum dot transported by Kinesin-1/DDB pairs.* The *Motile/Stationary* test is applied to the segmented quantum dot trajectories (motor Kin1/DDB). (**Panel A**) After conducting the test at the significance level $\alpha = 0.01$, we collected all segment durations and speeds. The ***green*** dots represent Motile segments, while the ***red*** dots indicate Stationary segments. (**Panel B**) A zoomed-in area where all segment speeds are below $100nm/s$; the ***orange*** lines show the error bars for estimating segment speeds, calculated based on the Standard Errors of estimating $\mathbf{w}$ in the regression model.

the precision in estimating the model's parameters, the *Motile/Stationary* test does incorporate this information for labeling segments. However, there remains a question of whether these motile segments identified by the test are truly motile in the sense of drifting or if they are a false positive resulting from the diffusion of the anchor. This requires further consideration regarding future work incorporating the test for the null hypothesis that the anchor exhibits Brownian motion.

4.4 Discussion

This chapter discusses the Stationary/Motile hypothesis test designed for 2D data, based on the assumption that the changepoint locations are predetermined and the

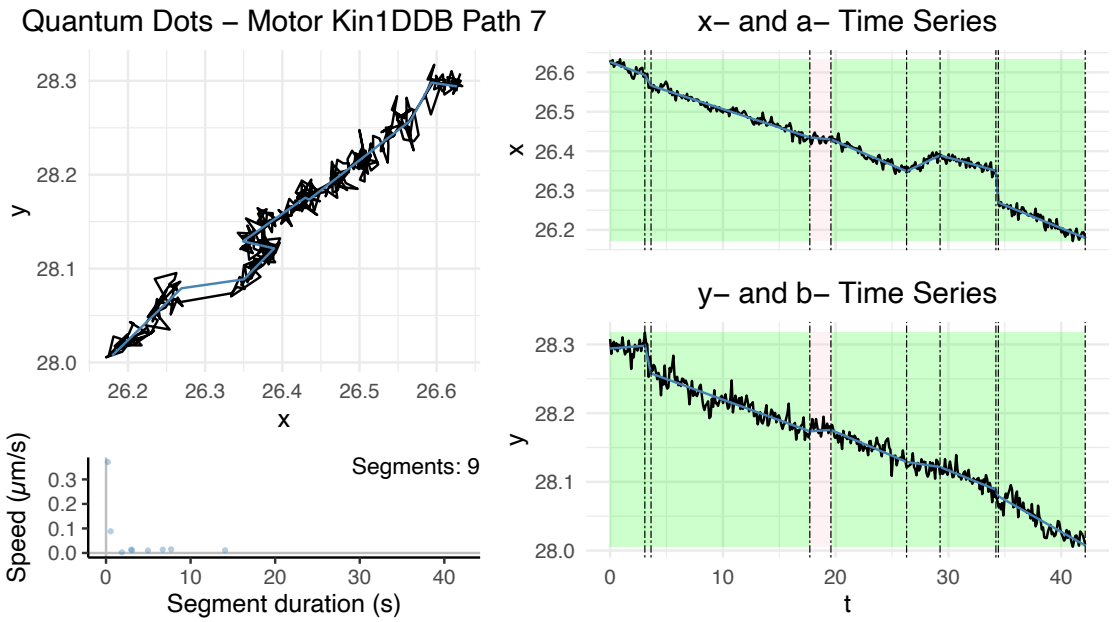


Figure 4.6: *Quantum dot transported by Kinesin-1/DDB pairs path 7.* An example of a path in the data whose long-slow segments are the ***third***, ***fifth***, ***seventh***, and ***ninth*** segments with estimated segment durations and speeds are $(14.1s, 0.0111\mu m/s)$, $(6.71s, 0.0139\mu m/s)$, $(4.95s, 0.0103\mu m/s)$ and $(7.7s, 0.0149\mu m/s)$, respectively. The right panel shows the state segments for each time series. Resulting from the *Motile/Stationary* test, each ***green*** panel denotes a *Motile* segment, each ***red*** panel denotes a *Stationary* segment, and the ***black*** dashed lines represent the changepoint times. In the segment duration versus speed plot, the ***blue*** points denote the inferred segments.

data is generated under a continuous piecewise linear model. We reformulate the test for segment speed into a general linear hypothesis testing framework, where the test statistic is shown to follow an F distribution. Through simulation experiments, we illustrate the effectiveness of the test in identifying stationary/motile segments and its capability to detect the motile segment at a speed as low as $0.025\mu m/s$, associated with a duration of $10s$ and a noise level of σ at 0.2. We also demonstrate that the test operates effectively on the inferred changepoints from CPLASS and support the notion that considering the proportion of time spent at various speeds will yield a more stable metric for trajectory characterization, as noted in Chapter 2. We point

out the limitation of the test when applied to the anchor diffusing model, where the actual anchor's location follows a Brownian motion with drift instead of linear paths. Subsequently, we apply the test to the quantum dot dataset and discuss the significance of considering noise information in the path while classifying the states of the segments.

The ideas presented in this chapter mark an initial step toward developing a robust statistical test for segment states in intracellular transport. Real-world challenges arise from both the diffusivity of the anchor and the small sample size. Several key concerns remain for future work:

- How can we clarify whether the motile segment identified by the current test is due to drifting or just diffusion?
- How can we address the double-dipping issue in small-sample scenarios where selection effects might still introduce bias?

Regarding the latter concern, the double-dipping issue arises because the same data is used both to detect changepoints and to test properties of the identified segments. This can lead to inflated false positive rates [15, 90], as the hypothesis test is not truly independent of the selection process. However, CPLASS demonstrates strong changepoint detection performance even for small sample sizes (e.g., $n \geq 50$), reducing the severity of this issue. While this mitigates the concern to some extent, the statistical inference problem remains: current selective inference techniques, which adjust for data-driven selection, focus primarily on mean-change detection [67, 87] and do not directly apply to velocity changepoint detection methods like CPLASS. Furthermore, existing selective inference research in changepoint problems mainly addresses uncertainty quantification of the detected changepoints rather than statistical testing of segment properties.

One potential direction for addressing this issue is to develop post-selection inference methods tailored to CPLASS, ensuring valid statistical tests for segment classification. An alternative approach is post-inference selection, as introduced in the Narrowest Significance Pursuit (NSP) method ([47]), which provides confidence intervals for changepoints rather than relying on post-hoc hypothesis testing. While NSP is currently available to detect slope changes in one-dimensional data, its principles suggest promising future directions for higher-dimensional velocity-based changepoint inference.

Chapter 5

Dendrogram Pruning and Merging (DPM) for Multiple Changepoint Detection

As discussed in Chapter 3, one limitation of CPLASS is its computational efficiency. The stochastic search relies on randomness to explore the search space, potentially spending time investigating unpromising regions before locating an optimal or near-optimal solution. One way to reduce the search workload is by initially overfitting the data with $k-1$ changepoints ($k > k_0$) and then applying a rule to eliminate the unnecessary ones. With this approach, we only need to search for r with fixed k segments and subsequently use a straightforward computational process to prune redundant changepoints and merge the corresponding segments optimally until we receive the final solution. This motivation led to the development of the Dendrogram Pruning and Merging (DPM) segmentation algorithm, designed to detect multiple changepoints, specifically in mean, variance, and both mean and variance. This work marks the first step in our journey toward constructing DPM for intracellular transport.

This is a joint work with Dat Do from the Department of Statistics at the University

of Michigan and my advisor, Scott A. McKinley. First, we establish its theoretical foundations, highlighting its rapid convergence rates and versatility in handling various kernels and multi-dimensional data. Next, we explain how DPM reduces computational costs by incorporating an efficient search method, such as Binary Segmentation (BinSeg), to create a dendrogram of changepoint locations without the necessity of repeated model fitting. Additionally, we introduce the Dendrogram Selection Criterion (DSC) for model selection, which uses segment-wise parameter distances and segment lengths rather than relying solely on the number of free parameters. Through simulation studies, we demonstrate that DSC is competitive with existing information criteria and offers improved robustness in specific scenarios.

5.1 Introduction

Introduction of the model and its applications. Changepoint detection is a fundamental problem in statistical analysis and signal processing, aimed at identifying points where the statistical properties of a sequence of observations shift. This problem has broad applications across various fields, including quality control [95, 105, 137], finance [6, 7, 86], climate studies [13, 92, 112, 131], and genomics [49, 103]. The roots of changepoint detection trace back to early work in the mid-20th century, particularly in industrial quality control, where the focus was on detecting shifts in manufacturing processes. The seminal Cumulative Sum (CUSUM) method, introduced by Page in 1954 [105], was one of the first systematic approaches for changepoint detection, designed to detect shifts in the mean of a sequential process. In finance, changepoint methods have been employed to detect regime changes in market conditions, while in climate science, they help in identifying shifts in temperature or precipitation patterns. Genomics has also seen widespread use of changepoint detection to identify significant changes in DNA sequences [49, 64, 65, 103], such as copy number variations.

Inference in changepoint detection models. Making inferences in changepoint detection involves modeling the homogeneous signal in each segment and the number of changepoints, either *a priori* or via some model selection methods [128]. Here, we focus on the setting where data within each segment is modeled as i.i.d. from a pre-specified kernel, and parameters are estimated using the Maximum Likelihood Estimation (MLE) method. This approach is flexible and applicable for several data types, such as discrete data (by modeling them using, e.g., Binomial or Poisson kernel) and continuous data (using Gaussian kernel). In this chapter, we aim to study the large-sample limit of the MLE for the changepoint model (with a possibly *misspecified* number of changes) and, from that theoretical insight, propose a practically appealing method for making inferences and selecting changepoint models.

Specifically, assume that we obtain n multivariate observations $(y_i)_{i=1}^n \subset \mathbb{R}^d$ in a equally spaced grid $(t_i)_{i=1}^n$ in $[0,1]$ according to the true model with k_0 segments:

$$y_i \overset{\text{ind.}}{\sim} p(y_i|f^0(t_i)); \qquad f^0 = \sum_{i=1}^{k_0} 1_{[\tau_{i-1}^0, \tau_i^0)} \theta_i^0, \tag{5.1}$$

where $\{p(y|\theta) : \theta \in \Theta\}$ is a pre-specified kernel, $\boldsymbol{\tau}^0 = (\tau_1^0, \ldots, \tau_{k_0-1}^0) \subset (0,1)$ are the true changepoints with $\tau_0^0 = 0, \tau_{k_0}^0 = 1$, and $(\theta_i^0)_{i=1}^{k_0}$ are true parameters for k_0 homogeneous segments. Those parameters are summarized using the signal function $f^0 : [0,1] \to \Theta$ with k_0 pieces.

Because k_0 is often unknown in practice, we may over-fit the system with a large number of segments k (possibly $k > k_0$):

$$(\widehat{\boldsymbol{\tau}}^n, \widehat{\boldsymbol{\theta}}^n) = \arg\max \sum_{i=1}^n \log p(y_i|f_{\boldsymbol{\tau},\boldsymbol{\theta}}(t_i)); \qquad f_{\boldsymbol{\tau},\boldsymbol{\theta}} = \sum_{i=1}^k 1_{[\tau_{i-1},\tau_i)} \theta_i, \tag{5.2}$$

and get the MLE $\widehat{f}^n := f_{\widehat{\boldsymbol{\tau}}^n, \widehat{\boldsymbol{\theta}}^n} = \sum_{i=1}^k 1_{[\widehat{\tau}_{i-1}^n, \widehat{\tau}_i^n)} \widehat{\theta}_i^n$ of k pieces. We then study the convergence rate of this over-fitted signal function and find that although the model

is overfitted, the estimated signal $\widehat{f}^n$ still converges to f^0 at a fast, parametric root n rate. Hence, there are several redundant estimated changepoints that need to be pruned (i.e., merge nearby segments) to make inferences from this over-fitted system. To this end, we suggest a merging procedure to find the true number of changepoints. Aesthetically, the procedure comes with a binary tree representing the hierarchy of changes in the detected changepoints.

Literature review. The body of literature concerning multiple changepoint detection problems can be divided into two primary categories. The first category focuses on minimizing a criterion function consisting of a fit measure (likelihood or least-squares) and a penalty to avoid overfitting. Regarding fixed changepoint numbers, we refer to [139] using least-squares estimation for iid noise model, [18] with an extension of the previous work to cases where variance depends on the mean, [106] and [144] use a likelihood criterion with a penalty depending on both the number and location of changepoints to favor a more uniformly spread estimated changepoint distribution. For an unknown but bounded number of change points problems, the choices for the penalty terms can be found in [141] with the Schwarz criterion, and [89] uses a more general criterion but is still linear in the number of changepoints. The discussion of changepoint problems for dependent observation can be found in [84] where they use penalized least-squares estimation, with a penalty linear in the number of changepoints; more related work can be found in [82, 83]. [88] uses model selection penalties inspired by [16]. [17] and others (e.g., [27, 28, 138]) study Schwarz-like penalties in more general forms. [113] propose using the Minimum Description Length (MDL) criterion. [55] use a least-squares criterion with a total variation penalty. However, the total variation penalty is suboptimal for balancing type-I and type-II errors, as noted by [20] and [24]. [115] and [116] also consider this penalty as part of the fused lasso penalty proposed by [127]. The SMUCE estimator ([43]) and the empirical Bayes method by

[35] also fit into the penalized framework. When it comes to the practical question of how it is feasible to minimize a criterion function, there are dynamic programming algorithms for doing this, including optimal partitioning [69], the Pruned Exact Linear Time (PELT) algorithm [109], the pruned dynamic programming by [114], and FPOP [97]. The PELT, pruned dynamic programming, and FPOP reduce the computational cost of dynamic programming, which is $O(n^2)$ (see [69]) to linear in best-case scenarios (while remaining quadratic speed in worst-case ones). Another attempt to reduce the computational cost includes a genetic algorithm [113]. [136] utilizes the fast discrete wavelet transform for changepoint detection, while [66] and [37] introduce the "moving sum" (MOSUM) technique, which involves an additional bandwidth parameter.

The second category contains the class of methods based on Binary Segmentation (BinSeg) and its variants. BinSeg operates by recursively detecting changepoints and splitting data into subsegments until a stopping criterion is met. It is a "greedy" and sequential approach, depending on prior stages, with each stage involving one-dimensional optimization. Originally proposed by [135], BinSeg has been shown to be consistent under various settings, including those addressed by [132], [5], and [22]. It has been applied in both univariate and multivariate time series segmentation (e.g., [25, 26, 48]). Although BinSeg has low computational complexity and is easy to implement, its "top-down" strategy can struggle with challenging data structures, particularly when segments contain multiple changepoints in complex configurations (see [44]). Modifications like Circular Binary Segmentation [104, 133], Wild Binary Segmentation [44], and the Narrowest-Over-Threshold [9] method aim to improve performance, albeit with increased computational demands.

Because of its broad applications, the literature on changepoint detection is abundant; we refer to [1, 41, 123, 128] for general contexts. Here, we only focus on the theory and methods for fitting changepoint models with MLE. For the single change-

point model, [61] studied the asymptotic normality of the changepoint given other parameters are known. For the Gaussian model, we refer to [140].[59, 60] explore the asymptotic theory for binomial changepoint cases. [30] studied the asymptotic distribution of usual likelihood ratio test statistics in a single change point case for an exponential family. However, the resulting test statistics do not have simple null limiting distributions. In addition, we are not aware of any results in the literature on the null limiting distribution of the usual likelihood ratio test statistic in multiple change point problems. [82, 85] extended the results for dependent noise and high-dimensional data. [58] showed the convergence of estimated changepoints and the asymptotic distribution of parameters when the true number of changes is known.

The agglomerative approach receives relatively scant attention in the changepoint literature. In nonparametric changepoint analysis, [99] (Section 6) proposed a heuristic algorithm for merging pre-specified segmentation using a divergence based on U-statistics. For detecting changes in mean for one-dimensional signal, [45] considered using the Unbalanced Haar wavelet to construct a (potentially over-fitted) estimator for the underlying signal, then post-process it by merging down using the learned wavelet coefficients. Compared to the mentioned method, ours is motivated by the convergence rate of an over-fitted signal, which is learned using the flexible MLE framework. However, we assume the user uses some algorithms to get the MLE.

Contributions. This work presents an agglomeration algorithm called *Dendrogram Pruning and Merging* (DPM) for solving multiple change point detection problems. It is supported by a theory of fast convergence rates and can work with different kernels and multi-dimensional data. DPM also offers a low computational cost when combined with a fast search method such as BinSeg, since we only need to fit the model once at an overfitted level and then construct a binary tree (dendrogram) of changepoint locations (e.g., see Figure 5.3). This dendrogram visualizes the pruning and merging

procedure of those changepoints without refitting the model (Figure 5.2), where we can observe the positions of changepoints and the associated distances, referred to as dendrogram heights, at each level of the tree (see Section 5.3 for details on the distance mentioned here). We also introduce a selection criterion called the *Dendrogram Selection Criterion* (DSC)[1] for model selection based on the constructed dendrogram of changepoint locations. Using empirical process theory [129], the consistency in estimating the number of changepoints using the DPM with DSC is guaranteed. In simulation studies, we demonstrate that DSC competes well with popular information criteria used in changepoint detection problems, and it proves to be more robust in certain scenarios discussed later in the paper. One reason for this is that the formulation of DSC incorporates the relative distances between the parameters of fitted adjacent segments and their lengths, instead of relying solely on the number of free parameters, as is the case with other popular information criteria such as the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC).

Notations. Let $p(y|\theta)$ be a pre-specified density/mass function with data $y \in \mathcal{Y} \subset \mathbb{R}^d$ and parameter θ belongs to a compact parameter space $\Theta \subset \mathbb{R}^m$. Let $\mathcal{F}_k(\Theta)$ be the space of "piecewise-constant functions with k pieces" from $\mathcal{T} = [0, 1]$ to Θ. Denote $\mathcal{T}_{\uparrow}^{k-1} = \{\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{k-1}) : 0 < \tau_1 < \tau_2 < \cdots < \tau_{k-1} < 1\}$ by the set of $(k-1)$ possible changepoints. Each function $f_{\boldsymbol{\tau},\boldsymbol{\theta}} \in \mathcal{F}_k$ is parametrized by $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{k-1}) \in \mathcal{T}_{\uparrow}^{k-1}$ and k parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_{k+1}) \subset \Theta$ satisfying $\theta_i \neq \theta_{i+1} \forall i \in [k]$ (where $[k] = \{1, ..., k\}$) as

$$f_{\boldsymbol{\tau},\boldsymbol{\theta}} = \theta_1 1_{[0,\tau_1)} + \theta_2 1_{[\tau_1,\tau_2)} + \cdots + \theta_{k-1} 1_{[\tau_{k-2},\tau_{k-1})} + \theta_k 1_{[\tau_{k-1},1]}. \tag{5.3}$$

[1] which was first presented in [31] with the name DIC (Dendrogram Information Criterion)

Henceforth, such a function has exactly $(k-1)$ changepoints. Denote $\mathcal{F}_{\leq k}(\Theta) = \cup_{k' \leq k} \mathcal{F}_{k'}(\Theta)$ by the space of piecewise-constant functions with at most k pieces. Let $\|\cdot\|$ be the Euclidean norm on Θ, and $L_2 := \{f : [0,1] \to \Theta : \|f\|_{L_2} < \infty\}$ be the space of square integral functions, where $\|f\|_{L_2}^2 := \int_0^1 \|f(t)\|^2 \, dt$.

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 $a_n \leq C b_n$ 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 $b_n \lesssim a_n$. 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 $V(p,q) \leq \sqrt{2 KL(p\|q)}$.

5.2 Asymptotic behavior of over-fitted signal functions and its implications

In this section, we show that the over-fitted signal function still converges to the true signal function with the fast root-n rate, which is just as fast as the exact-fitted estimated signal. As a consequence, for each true changepoint, there is at least an estimated changepoint that consistently estimates it. This convergence behavior motivates a post-processing procedure of merging nearby segments with similar distribution, which will be useful for summarization and model selection.

5.2.1 Convergence rate of over-fitted signal functions

Assume that we observed data $y_1, \ldots, y_n$ on the fixed, equally spaced design point $(t_i)_{i=1}^n = (1/n, 2/n, \ldots, 1)$ as

$$y_i \overset{ind.}{\sim} p(y | f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0}(t_i)), \quad \text{for } i = 1, 2, \ldots, n, \tag{5.4}$$

where $\boldsymbol{\tau}^0 = (\tau_1^0, \ldots, \tau_{k_0-1}^0)$ are true changepoints and $\boldsymbol{\theta}^0 = (\theta_1^0, \ldots, \theta_{k_0}^0)$ are true parameters in segments. We fit the data with an over-fitted system with $k \geq k_0$ segments to get the MLE

$$\widehat{\boldsymbol{\tau}}^n, \widehat{\boldsymbol{\theta}}^n = \arg\max_{\boldsymbol{\tau},\boldsymbol{\theta}} \sum_{i=1}^{n} \log p(y_i | f_{\boldsymbol{\tau},\boldsymbol{\theta}}(t_i)), \tag{5.5}$$

where $\widehat{\boldsymbol{\tau}}^n = (\hat{\tau}_1^n, \ldots, \hat{\tau}_{k-1}^n)$, $\widehat{\boldsymbol{\theta}}^n = (\widehat{\theta}_1^n, \ldots, \widehat{\theta}_k^n)$, and we denote $\widehat{f}^n = f_{\widehat{\boldsymbol{\tau}}^n, \widehat{\boldsymbol{\theta}}^n}$ for short. A natural distance for evaluating convergence for functions is the average empirical L_2 distance [130]: For two functions $f, g : [0,1] \to \Theta$, define

$$\|f - g\|_n := \left(\frac{1}{n} \sum_{i=1}^{n} \|f(i/n) - g(i/n)\|^2 \right)^{1/2}. \tag{5.6}$$

This is also known as the empirical L_2 risk used in [32, 45]. Our strategy to examine the convergence rate of $\widehat{f}^n$ to f^0 in $\|\cdot\|_n$ contains two steps. We first use the application of empirical process theory in M-estimation [129] to show the convergence of the MLE's densities to the true signal's densities. Then, we relate it to the convergence of the estimated signals by investigating the relationship between the geometry of density and functional space. Indeed, the empirical process theory provides us with a useful tool for establishing the concentration behavior of MLE of a sequence of independent observations like in the Changepoint detection model, in which we will concisely present the main required results with notations adapted to our specific problem below.

Given two sequences of densities $(p_1, \ldots, p_n)$ and $(q_1, \ldots, q_n)$ on $\mathcal{Y}$, define the Hellinger process [52, 129] between product densities $p = \otimes_{i=1}^n p_i$ and $q = \otimes_{i=1}^n q_i$ and its average to be

$$h_n^2(p,q) := \frac{1}{2} \sum_{i=1}^{n} h^2(p_i, q_i), \quad \overline{h}_n^2(p,q) := \frac{1}{n} h_n^2(p,q). \tag{5.7}$$

For function $f_{\boldsymbol{\tau},\boldsymbol{\theta}}$, denote $p^{(n)}_{\boldsymbol{\tau},\boldsymbol{\theta}}$ by the product density $\otimes_{i=1}^{n} p(\cdot|f_{\boldsymbol{\tau},\boldsymbol{\theta}}(t_i))$ on $\mathcal{Y}^n$. The Empirical process theory provides many useful concentration inequalities uniformly over balls in the density space $\{p^{(n)}_{\boldsymbol{\tau},\boldsymbol{\theta}} : (\boldsymbol{\tau},\boldsymbol{\theta}) \in \mathcal{T}^{k-1}_{\uparrow} \times \Theta^k\}$ so that the convergence rate of MLE $p^{(n)}_{\widehat{\boldsymbol{\tau}},\widehat{\boldsymbol{\theta}}}$ boils down to calculate (or sufficiently provide an upper bound for) the smallest number of balls to cover this space in the Hellinger process distance. This number is often referred to as the "covering number," which we will now define.

Definition 5.1: Entropy number with bracketing

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

(i) $\overline{h}_n\left(\dfrac{p_j^L + p^0}{2}, \dfrac{p_j^U + p^0}{2}\right) \leq \delta$ and

(ii) $p_{ji}^L(y) \leq p(y|f_{\boldsymbol{\tau},\boldsymbol{\theta}}(i/n)) \leq p_{ji}^U(y)$ for all $y \in \mathcal{Y}$ and $i \in [n]$.

Then $N_B(\delta)$ and $H_B(\delta) = \log N_B(\delta)$ are called the Hellinger covering number and entropy number with bracketing, respectively.

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

$$J_B(\delta) := \int_{\delta^2/c_0}^{\delta} H_B^{1/2}(u)du \vee \delta, \quad 0 < \delta \leq 1. \tag{5.8}$$

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

Theorem 5.1: Theorem 8.14 in [129]

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

$$\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n), \tag{5.9}$$

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

$$\mathbb{P}\left(\overline{h}_n\left(p^{(n)}_{\widehat{\boldsymbol{\tau}}^n,\widehat{\boldsymbol{\theta}}^n}, p^{(n)}_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right) \geq \delta\right) \leq c\exp\left(-\frac{n\delta^2}{c^2}\right). \tag{5.10}$$

In general, the entropy number $H_B(\delta) \asymp \log(1/\delta)$ as in usual parametric model leads to convergence rate $\delta_n \asymp n^{-1/2}$ and $H_B(\delta) \asymp (1/\delta)^\alpha$ leads to $\delta_n \asymp n^{-\alpha/(2\alpha+1)}$ [129].

In the following, we introduce two conditions for the kernel $p(y|\theta)$ that will be needed for later theorems on the convergence rate of signal functions.

Condition (K1). *The kernel $p(x|\theta)$ satisfies*

$$\underline{c}\,\|\theta - \theta'\| \leq h(p(\cdot|\theta), p(\cdot|\theta')) \leq \overline{c}\,\|\theta - \theta'\| \quad \forall \theta, \theta' \in \Theta, \tag{5.11}$$

for some constant $\underline{c}, \overline{c}$ only depend on p and Θ.

Condition (K2). *Suppose that $\sup_{\theta\in\Theta}\|p(\cdot|\theta)\|_\infty$ is bounded, $\|p(\cdot|\theta) - p(\cdot|\theta')\|_\infty \lesssim \|\theta - \theta'\|$ for all $\theta, \theta' \in \Theta$, and $p(y|\theta)$ has uniformly light tails, i.e., there exist constants D, d_1, and d_2 so that $p(y|\theta) \leq d_1\exp(-d_2\|y\|^{d_3})\forall\,\|y\| \geq D, \theta \in \Theta$.*

Condition (K1) is useful to show the equivalence of $\|\cdot\|_n$ and $\overline{h}_n$, which connects the convergence rate in densities to that of signal functions, as shown in Lemma 2. This condition is satisfied for popular kernels such as Gaussian and Poisson kernels

with bounded parameter space. Proof of checking this condition is deferred to the appendix.

Lemma 2. *Under condition (K1), for all* $(\boldsymbol{\tau}, \boldsymbol{\theta}), (\boldsymbol{\tau}', \boldsymbol{\theta}') \in \mathcal{T}_{\uparrow}^{k-1} \times \Theta^k$ *we have*

$$\left\| f_{\boldsymbol{\tau},\boldsymbol{\theta}} - f_{\boldsymbol{\tau}',\boldsymbol{\theta}'} \right\|_n \asymp \overline{h}_n(p^{(n)}_{\boldsymbol{\tau},\boldsymbol{\theta}}, p^{(n)}_{\boldsymbol{\tau}',\boldsymbol{\theta}'}). \tag{5.12}$$

Condition (K2) is used to establish the bound on the entropy number with bracketing of kernel p as $H_B(\delta) \asymp k \log(n/\delta)$, which will be used to show the desired convergence rate.

Theorem 5.2: Convergence rate of the MLE overfitted signal function

Under assumption (K1) and (K2), there exists universal constants c_1, c_2 so that with probability at least $1 - c_1 n^{-c_2}$, we have

$$\left\| f_{\widehat{\boldsymbol{\tau}}^n,\widehat{\boldsymbol{\theta}}^n} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0} \right\|_n \leq C \left(\frac{\log n}{n} \right)^{1/2}, \tag{5.13}$$

where C is a constant that only depends on kernel p, parameter space Θ and k.

A simulation study to confirm this convergence rate is carried out in Section 5.5.

Before discussing the Dendrogram-based approach with signal functions, we present the following table containing all notations used throughout the chapter to enhance the flow of the discussion.

---

[2]After pruning and merging from the starting MLE overfitted function $\hat{f}_n$, the signal functions $f_{\hat{\boldsymbol{\tau}}^{n,(\kappa)},\hat{\boldsymbol{\theta}}^{n,(\kappa)}}$ $(\kappa \in [k-1])$are no longer MLE.

True signal function	**MLE signal function**	**Signal function (Not necessarily MLE)**
$f_0 = f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0} = f_{\boldsymbol{\tau}^{0,(k_0)},\boldsymbol{\theta}^{0,(k_0)}}$	$\hat{f}_n = \hat{f}_{\hat{\boldsymbol{\tau}}^n,\hat{\boldsymbol{\theta}}^n} = f_{\hat{\boldsymbol{\tau}}^{n,(k)},\hat{\boldsymbol{\theta}}^{n,(k)}}$	$f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n} = f_{\boldsymbol{\tau}^{n,(k)},\boldsymbol{\theta}^{n,(k)}}$
Associated sequence of signal functions after pruning and merging		
$\{f_{\boldsymbol{\tau}^{0,(\kappa)},\boldsymbol{\theta}^{0,(\kappa)}}\}_{\kappa=1}^{k_0}$	$\{f_{\hat{\boldsymbol{\tau}}^{n,(\kappa)},\hat{\boldsymbol{\theta}}^{n,(\kappa)}}\}_{\kappa=1}^{k}$ [2] (for $k \geq k_0$)	$\{f_{\boldsymbol{\tau}^{n,(\kappa)},\boldsymbol{\theta}^{n,(\kappa)}}\}_{\kappa=1}^{k}$ (for $k \geq k_0$)
Associated parameters at the level κ-th of the Dendrogram		
$\boldsymbol{\tau}^{0,(\kappa)} = \left(\tau_1^{0,(\kappa)},\ldots,\tau_\kappa^{0,(\kappa)}\right)$ $\boldsymbol{\theta}^{0,(\kappa)} = \left(\theta_1^{0,(\kappa)},\ldots,\theta_\kappa^{0,(\kappa)}\right)$	$\hat{\boldsymbol{\tau}}^{n,(\kappa)} = \left(\hat{\tau}_1^{n,(\kappa)},\ldots,\hat{\tau}_\kappa^{n,(\kappa)}\right)$ $\hat{\boldsymbol{\theta}}^{n,(\kappa)} = \left(\hat{\theta}_1^{n,(\kappa)},\ldots,\hat{\theta}_\kappa^{n,(\kappa)}\right)$	$\boldsymbol{\tau}^{n,(\kappa)} = \left(\tau_1^{n,(\kappa)},\ldots,\tau_\kappa^{n,(\kappa)}\right)$ $\boldsymbol{\theta}^{n,(\kappa)} = \left(\theta_1^{n,(\kappa)},\ldots,\theta_\kappa^{n,(\kappa)}\right)$

Table 5.1: Notation Table

5.2.2 Consequence on the asymptotic behavior of over-fitted parameters

Now, we examine the implications of the convergence rate of signal functions for parameters $\widehat{\boldsymbol{\tau}}^n$ and $\widehat{\boldsymbol{\theta}}^n$, which are the main concerns when reporting results from the changepoint model. Let T_{ij} be the number of design points in $[\tau_{i-1}^0, \tau_i^0) \cap [\hat{\tau}_{j-1}^n, \hat{\tau}_j^n)$ and $p_{ij} = T_{ij}/n$, where we drop the dependence of n on T and p's for ease of notation, we have

$$\left\| f_{\widehat{\boldsymbol{\tau}}^n\widehat{\boldsymbol{\theta}}^n} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0} \right\|_n^2 = \frac{1}{n}\sum_{i=1}^{k_0}\sum_{j=1}^{k} T_{ij}\left\|\widehat{\theta}_j^n - \theta_i^0\right\|^2 = \sum_{i,j=1}^{k_0,k} p_{ij}\left\|\widehat{\theta}_j^n - \theta_i^0\right\|^2 \leq C\left(\frac{\log n}{n}\right), \tag{5.14}$$

and $\sum_{j=1}^{k} p_{ij} = p_i^0$, $\sum_{i=1}^{k_0} p_{ij} = \widehat{p}_j^n$, where $p_i^0 = \tau_i^0 - \tau_{i-1}^0$ and $\widehat{p}_j^n = \hat{\tau}_j^n - \hat{\tau}_{j-1}^n$. We take $\tau_0^0 = \widehat{\tau}_0^n = 0$ and $\tau_{k_0}^0 = \widehat{\tau}_k^n = 0$ as convention. This implies:

(i) For segment-wise parameters θ , if the intersection p_{ij} is non-vanishing and $\widehat{p}_j \to p_i^0$ (e.g., case $i = j = 1$ Figure 5.1), we have $\widehat{\theta}_j^n \to \theta_i^0$ in the fast $\sqrt{n}$ rate (up to a logarithmic factor);

(ii) For each true τ_i^0, there exists at least an estimated changepoint $\widehat{\tau}_j^n$ tends to it

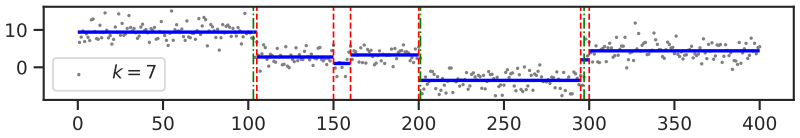


Figure 5.1: Over-fitted signal with $k = 7$ segments (in blue) and estimated changepoints (in red) is plotted against true changepoints (in green) with $k_0 = 4$ for simulated data with $n = 400$.

with $1/n$ rate (up to a logarithmic factor), as will be shown in Lemma 3.

Hence, there will be a subset of k_0 over-fitted parameters $(\widehat{\tau}_j^n, \widehat{\theta}_j^n)_{j=1}^k$ that match the rate of exact-fitted signal in the setting $k = k_0$ is known (see also [58]). The behavior of the $(k - k_0)$ redundant estimated changepoint is more complicated and can be grouped into two categories:

(i) Changepoint $\widehat{\tau}_j^n$ is put in the middle of a long homogeneous segment (e.g., case $j = 2, 3$ in Figure 5.1) with two consecutive $\widehat{\theta}_j^n$ and $\widehat{\theta}_{j+1}^n$ approximate θ_i^0 well;

(ii) Several changepoints $\widehat{\tau}_j^n$ try to estimate the same true changepoint (e.g., case $j = 5, 6$ in Figure 5.1) resulting in short segment $[\hat{\tau}_{j-1}^n, \hat{\tau}_j^n)$ and arbitrary $\widehat{\theta}_j^n$ which does not match θ_i^0.

Instead of trying to characterize each category individually, which is challenging, we use the fact that convergence rate (5.14) leads to a simple observation:

$$p_{ij} \left\| \widehat{\theta}_j^n - \theta_i^0 \right\|^2 \leq C \left(\frac{\log n}{n} \right), \quad \text{and} \quad p_{i(j+1)} \left\| \widehat{\theta}_{j+1}^n - \theta_i^0 \right\|^2 \leq C \left(\frac{\log n}{n} \right),$$

for all $i \in [k_0]$ and $j \in [k]$. Combining with pigeonhole-type argument to show the existence of index i and two consecutive indices j and $j + 1$ such that $p_{ij} \asymp \widehat{p}_j^n$ and $p_{i(j+1)} \asymp \widehat{p}_{j+1}^n$, we have

$$\widehat{p}_j^n \left\| \widehat{\theta}_j^n - \theta_i^0 \right\|^2 + \widehat{p}_{j+1}^n \left\| \widehat{\theta}_{j+1}^n - \theta_i^0 \right\|^2 \lesssim p_{ij} \left\| \widehat{\theta}_j^n - \theta_i^0 \right\|^2 + p_{i(j+1)} \left\| \widehat{\theta}_{j+1}^n - \theta_i^0 \right\|^2 \lesssim \left(\frac{\log n}{n} \right)$$

Hence, an application of Cauchy-Schwarz inequality implies

$$\frac{\widehat{p}_j^n \widehat{p}_{j+1}^n}{\widehat{p}_j^n + \widehat{p}_{j+1}^n} \left\| \widehat{\theta}_j^n - \widehat{\theta}_{j+1}^n \right\|^2 \leq \widehat{p}_j^n \left\| \widehat{\theta}_j^n - \theta_i^0 \right\|^2 + \widehat{p}_{j+1}^n \left\| \widehat{\theta}_{j+1}^n - \theta_i^0 \right\|^2 \lesssim \left(\frac{\log n}{n} \right). \tag{5.15}$$

Notably, the left-hand side of this inequality does not depend on the true parameters $(\boldsymbol{\tau}^0, \boldsymbol{\theta}^0)$ anymore, but only on the estimated parameters. This motivates us to post-process the over-fitted solution to prune such redundant changepoint $\hat{\tau}_j^n$ and merge similar $\{\widehat{\theta}_j^n, \widehat{\theta}_{j+1}^n\}$ to obtain a more sparse summary of the model parameters. Because k_0 is unknown, we prune those changepoints one by one until no changepoint is left, resulting in a dendrogram (tree) of segmentation in a bottom-up fashion, which will be useful for selecting model and hierarchy discovery simultaneously. We are presenting this procedure in detail in the next section.

5.3 Dendrogram Pruning and Merging (DPM)

5.3.1 Dendrogram Pruning and Merging Algorithm

Now, we describe an algorithm to sequentially project a given signal function of k pieces to spaces of functions with fewer pieces. The treatment in this section is not exclusive to the MLE but should be considered a general strategy for performing dimension reduction for piecewise functions. Interestingly, it is also provable to be optimal (in L_2 sense), as will be shown in Proposition 1. Later, we will apply this method to the MLE and show that it has good asymptotic properties.

Given a signal function $f_{\boldsymbol{\tau},\boldsymbol{\theta}} \in \mathcal{F}_k(\Theta)$ with parameters $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{k-1}) \in \mathcal{T}_\uparrow^k \subset (0,1)^{k-1}$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta^k$. In the following, we write $\tau_0 = 0$ and $\tau_k = 1$ as convention. For each $j \in [k-1]$, calculate the dissimilarity between the j-th segment

and the $(j+1)$-th segment by

$$d_j := \left(\frac{(\tau_j - \tau_{j-1})(\tau_{j+1} - \tau_j)}{\tau_{j+1} - \tau_{j-1}} \left\| \theta_j - \theta_{j+1} \right\|^2 \right)^{1/2}, \quad \forall j \in [k-1]. \tag{5.16}$$

Note that d_j is small if either $\|\theta_j - \theta_{j+1}\|$ or $(\tau_j - \tau_{j-1})(\tau_{j+1} - \tau_j)$ is small. We choose $j^* = \arg\min_{j \in [k-1]} d_j$ and define

$$\theta^*_{j^*} = \frac{\tau_{j^*} - \tau_{j^*-1}}{\tau_{j^*+1} - \tau_{j^*-1}} \theta_{j^*} + \frac{\tau_{j^*+1} - \tau_{j^*}}{\tau_{j^*+1} - \tau_{j^*-1}} \theta_{j^*+1}. \tag{5.17}$$

We then prune τ_{j^*} and merge the j^* and (j^*+1)-th segments to obtain $f_{\tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{\theta}}} \in \mathcal{F}_{k-1}(\Theta)$ with

$$\tilde{\boldsymbol{\tau}} = (\tau_1, \ldots, \tau_{j^*-1}, \tau_{j^*+1}, \ldots, \tau_{k-1}) \quad \text{and} \quad \tilde{\boldsymbol{\theta}} = (\theta_1, \ldots, \theta_{j^*-1}, \theta^*_{j^*}, \theta_{j^*+2}, \ldots, \theta_k). \tag{5.18}$$

Finally, we can perform this sequentially to get a sequence of piecewise functions with decreasing numbers of segments, as summarized in Algorithm 3. We call this algorithm "Dendrogram Pruning and Merging" as we sequentially merge consecutive segments of the given function, resulting in a tree (dendrogram) of nested segmentation.

Recall the L_2 norm of the difference between two functions f and $g : [0,1] \to \Theta$ is

$$\|f - g\|_{L_2} = \left(\int_0^1 \|f(t) - g(t)\|^2 \, dt \right)^{1/2}.$$

This L_2 norm is the limit version of the average empirical L_2 distance given by equation (5.5), i.e., $\|f - g\|_n \xrightarrow{n \to \infty} \|f - g\|_{L_2}$ for $f, g \in L_2$. Furthermore, for $\boldsymbol{\tau}, \boldsymbol{\tau}' \subset \{1/n, 2/n, \ldots, 1\}$, we have $\left\| f_{\boldsymbol{\tau}, \boldsymbol{\theta}} - f_{\boldsymbol{\tau}', \boldsymbol{\theta}'} \right\|_n = \left\| f_{\boldsymbol{\tau}, \boldsymbol{\theta}} - f_{\boldsymbol{\tau}', \boldsymbol{\theta}'} \right\|_{L_2}$. The following proposition explains the optimality of the proposed merging procedure in L_2 sense.

Proposition 1. *Given $f_{\boldsymbol{\tau}, \boldsymbol{\theta}} \in \mathcal{F}_k(\Theta)$, the function $f_{\tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{\theta}}}$ obtained from the pruning and*

Algorithm 3 Dendrogram Pruning and Merging (DPM) segmentation algorithm

Require: A piecewise signal function $f_{\boldsymbol{\tau}^{(k)},\boldsymbol{\theta}^{(k)}} = \sum_{i=1}^{k} 1_{[\tau_{j-1},\tau_j)}\theta_j$ of k pieces.

1: **for** κ runs backward from k to 2 **do**
2: calculate $d_j^{(\kappa)}$ from $f_{\boldsymbol{\tau}^{(\kappa)},\boldsymbol{\theta}^{(\kappa)}}$ as in Equation (5.16);
3: find the optimal index $j^* = \arg\min d_j^{(\kappa)}$ to prune;
4: obtain $(\boldsymbol{\tau}^{(\kappa-1)}, \boldsymbol{\theta}^{(\kappa-1)}) = (\tilde{\boldsymbol{\tau}}^{(\kappa)}, \tilde{\boldsymbol{\theta}}^{(\kappa)})$ as in Equation (5.17) and (5.18).
5: **end for**
6: **return** A sequence of signal functions $(f_{\boldsymbol{\tau}^{(\kappa)},\boldsymbol{\theta}^{(\kappa)}})_{\kappa=1}^{k}$ with $\kappa = 1, \ldots, k$ pieces, respectively.

merging procedure above is the optimal projection of $f_{\boldsymbol{\tau},\boldsymbol{\theta}}$ *into the space of functions with at most* $(k-1)$ *pieces in the* L_2 *sense:*

$$f_{\tilde{\boldsymbol{\tau}},\tilde{\boldsymbol{\theta}}} = \underset{f \in \mathcal{F}_{\leq k-1}(\Theta)}{\arg\min} \|f_{\boldsymbol{\tau},\boldsymbol{\theta}} - f\|_{L_2}^2 . \tag{5.19}$$

Furthermore,

$$d_{j^*} = \left\| f_{\boldsymbol{\tau},\boldsymbol{\theta}} - f_{\tilde{\boldsymbol{\tau}},\tilde{\boldsymbol{\theta}}} \right\|_{L_2} . \tag{5.20}$$

Convergent properties of signal functions on the dendrogram. Because DPM minimizes the dissimilarity d_j as in Equation (5.16) in every step, it is a useful device to sequentially eliminate similar consecutive parameters ($\theta_j \approx \theta_{j+1}$) and short segments ($\tau_j \approx \tau_{j+1}$), which exactly aligns with the behavior of over-fitted parameters in changepoint models (as discussed in Section 5.2.2). Now, suppose that we have a sequence of estimate $f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n} \in \mathcal{F}_{\leq k}(\Theta)$ (not necessarily the MLE) to the true signal function $f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0} \in \mathcal{F}_{k_0}(\Theta)$ such that

$$\left\| f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0} \right\|_n \leq C\epsilon_n \quad \forall n, \tag{5.21}$$

for $k \geq k_0$, C is a constant (not depend on n) and $\epsilon_n \to 0$ (as $n \to \infty$). We now examine the effect of DPM on those fast-convergent sequences of functions. In particular, all the signal functions on the dendrogram of $f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n}$ with at least k_0 segments will be shown

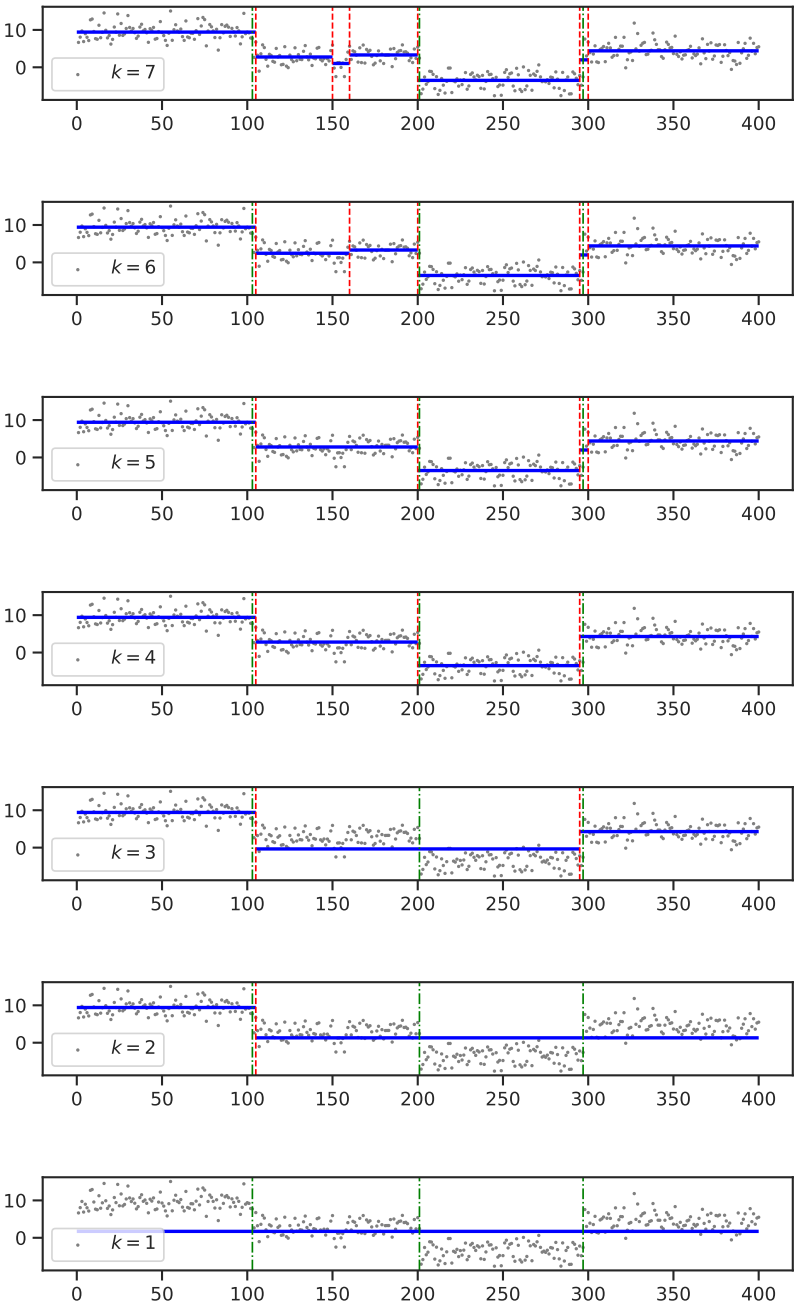


Figure 5.2: Illustration for the pruning and merging procedure with simulated data from a normal mean multiple changepoint model with the actual number of segments being $k_0 = 4$, and a constant standard deviation being $\sigma = 2$. The sample size $n = 400$. The data is overfitted by $k = 7$. The green vertical lines represent the actual locations of the changepoints, while the red lines indicate the detected changepoints in each pruning and merging step.

to inherit the same convergence rate to $f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}$, albeit possibly having less segment than the original estimator $f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n}$. Because we perform the pruning and merging operations sequentially, the following two lemmas on the stability of this operation will be useful in establishing the convergence of the whole sequence. First, we show that such a sequence of signal functions $\{f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n}\}_n \subset \mathcal{F}_{\leq k}$ will not miss any true changepoint.

Lemma 3. *Given a signal function $f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0} \in \mathcal{F}_{k_0}(\Theta)$. Let C be a constant that does not depend on n, and $k \geq k_0$. For any sequence of signal functions $f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n} \in \mathcal{F}_k(\Theta)$ that satisfies (5.21) with constant C and tolerance ϵ_n, we have that for every τ_i^0 ($i \in [k_0-1]$), there exist at least a sequence $\tau_{j_n}^n$ ($j_n \in [k-1]$) satisfying*

$$\mathbb{P}\left(|\tau_{j_n}^n - \tau_i^0| \leq C'\epsilon_n^2\right) \to 1 \quad i \in [k_0 - 1]$$

as $n \to \infty$, where the constant C' depends on C, $f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}$, Θ and k.

Using this result, we can show that the convergence rate to the true signal function will not change when we prune the redundant changepoints in the following lemma. We refer to Table 5.1 for the notation conventions.

Lemma 4. *Assume the same constant C and sequence $f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n}$ as in Lemma 3. Further assume that $k \geq k_0 + 1$, then*

$$\left\|f_{\boldsymbol{\tau}^{n,(\kappa-1)},\boldsymbol{\theta}^{n,(\kappa-1)}} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 \leq \left\|f_{\boldsymbol{\tau}^{n,(\kappa)},\boldsymbol{\theta}^{n,(\kappa)}} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 + w_\kappa\epsilon_n^2 \lesssim \epsilon_n^2,$$

for all $\kappa \in [k_0 + 1, k]$, where the constants w_κ's depend only on $C, f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}, \Theta$, and k.

5.3.2 Asymptotic property of the Dendrogram Pruning and Merging

Together with Theorem 5.2, Lemma 3 and Lemma 4 pave us the way to study the asymptotic properties of the signal functions obtained by applying DPM to the over-

fitted MLE, which we present in the following. Given data $(t_i, y_i)_{i=1}^n$ generated from true model (5.1) with the true signal function $f^0 = f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}$. Let $\hat{f}^n = f_{\widehat{\boldsymbol{\tau}}^n,\widehat{\boldsymbol{\theta}}^n}$ be the MLE of the changepoint model (5.2) with at most k segments ($k \geq k_0$) and $(f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa)},\widehat{\boldsymbol{\theta}}^{n,(\kappa)}})_{\kappa=1}^k$ be the sequence of signal functions resulting from the DPM algorithm applying for $\widehat{f}^n$. Let $(f_{\boldsymbol{\tau}^{0,(\kappa)},\boldsymbol{\theta}^{0,(\kappa)}})_{\kappa=1}^{k_0}$ denote the sequence of signal functions when applying DPM to the true signal function. The under-fitted levels $\kappa < k_0$ of the true signal functions inform us the hierarchy of changes' magnitude. Moreover, the convergence of $f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa)},\widehat{\boldsymbol{\theta}}^{n,(\kappa)}}$ to $f_{\boldsymbol{\tau}^{0,(\kappa)},\boldsymbol{\theta}^{0,(\kappa)}}$ for every κ shows the stability of DPM algorithm and will be useful for model selection.

Now, we state our main result on the convergence of each function on the dendrogram to the true functions, where the convergence rate ϵ_n is made explicit as $\epsilon_n = (\log n/n)^{1/2} \to 0$ from now on.

Theorem 5.3: Convergence of the DPM signal functions

Assuming conditions (K1) and (K2). There exists universal constants c_1, c_2 so that with probability at least $1 - c_1 n^{-c_2}$, we have

$$\left\| f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa)},\widehat{\boldsymbol{\theta}}^{n,(\kappa)}} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0} \right\|_n \leq C \left(\frac{\log n}{n} \right)^{1/2}, \quad \forall k_0 \leq \kappa \leq k, \text{ and} \tag{5.22}$$

$$\left\| f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa)},\widehat{\boldsymbol{\theta}}^{n,(\kappa)}} - f_{\boldsymbol{\tau}^{0,(\kappa)},\boldsymbol{\theta}^{0,(\kappa)}} \right\|_n \leq C \left(\frac{\log n}{n} \right)^{1/2}, \quad \forall 1 \leq \kappa \leq k_0, \tag{5.23}$$

where C is a constant that only depends on p, $f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}, \Theta$, and k.

Corollary 1. *With the same assumption and probability as Theorem 5.3, we have*

$$|\widehat{\tau}_j^{n,(\kappa)} - \tau_j^{0,(\kappa)}| \leq C \left(\frac{\log n}{n} \right), \quad \|\widehat{\theta}_j^{n,(\kappa)} - \theta_j^{0,(\kappa)}\| \leq C \left(\frac{\log n}{n} \right)^{1/2}$$

for all $j \in [\kappa]$ and $\kappa \in [k_0]$, where C is a constant that only depends on p, $f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}, \Theta$,

and k.

Theorem 5.3 demonstrates that as we move from over- to exact-fitted levels, the signal functions produced by the DPM algorithm converge to the true signal function with a high probability. This implies that the differences between $f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa)},\widehat{\boldsymbol{\theta}}^{n,(\kappa)}}$ and $f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa-1)},\widehat{\boldsymbol{\theta}}^{n,(\kappa-1)}}$ for $k_0 < \kappa \leq k$ in the empirical L_2 norm will tend to 0 with a high probability. We will turn to study these differences, which will be called the height of the dendrogram.

Definition 5.2: Dendrogram's Heights

The height of the dendrogram of the MLE at the κ-th level is denoted by

$$d_n^{(\kappa)} = \left\| f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa)},\widehat{\boldsymbol{\theta}}^{n,(\kappa)}} - f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa-1)},\widehat{\boldsymbol{\theta}}^{n,(\kappa-1)}} \right\|_n, \quad \text{for } 1 < \kappa \leq k, \tag{5.24}$$

and height of the dendrogram of the true function at the κ-th level is denoted by

$$d_0^{(\kappa)} = \left\| f_{\boldsymbol{\tau}^{0,(\kappa)},\boldsymbol{\theta}^{0,(\kappa)}} - f_{\boldsymbol{\tau}^{0,(\kappa-1)},\boldsymbol{\theta}^{0,(\kappa-1)}} \right\|_n, \quad \text{for } 1 < \kappa \leq k_0, \tag{5.25}$$

The following theorem will discuss the convergence rate of these heights at the under-fit, exact-fit, and over-fit levels.

Theorem 5.4: Convergence of the dendrogram heights

With the same assumption and probability as Theorem 5.3,

$$d_n^{(\kappa)} \leq C \left(\frac{\log n}{n} \right)^{1/2}, \quad \forall k_0 < \kappa \leq k, \tag{5.26}$$

and

$$|d_n^{(\kappa)} - d_0^{(\kappa)}| \leq C \left(\frac{\log n}{n} \right)^{1/2}, \quad \forall 1 < \kappa \leq k_0, \tag{5.27}$$

where C is a constant that only depends on p, $f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}, \Theta$, and k.

Theorem 5.4 motivates the use of the Dendrogram in model selection, resulting in the Dendrogram Information Criteria (DSC) discussed in the next section. The theorem states that as the sample size (n) becomes large ($n \to \infty$), if the model is over-fitted ($\kappa > k_0$), the height of the dendrogram tends to 0. Conversely, when transitioning from the exact fitted level to the under-fitted level, the height of the dendrogram $d_n^{(\kappa)}$ tends to $d_0^{(\kappa)} > 0$ with a high probability. An illustration is given in Figure 5.3.

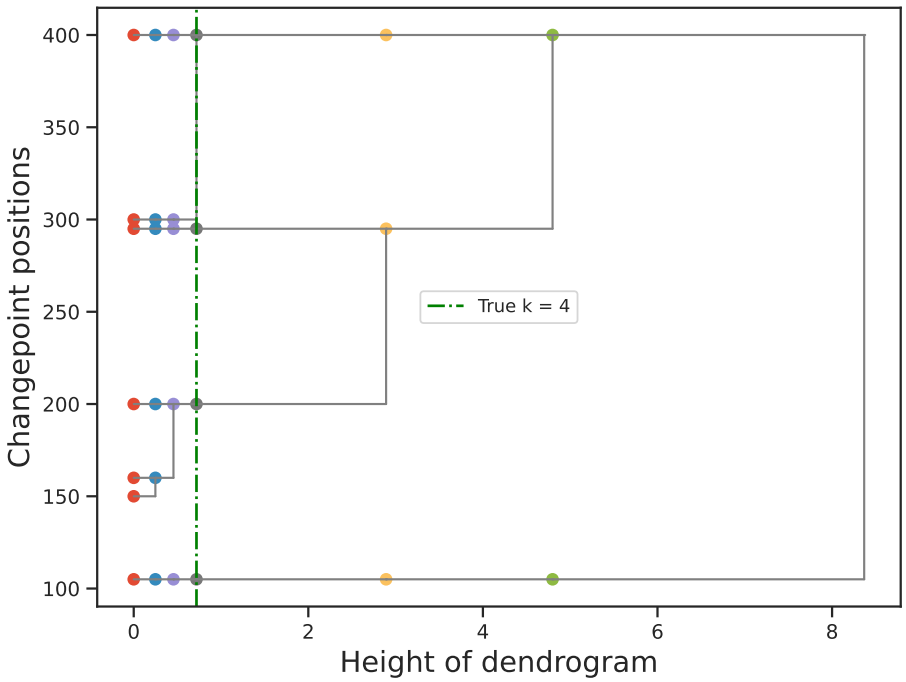


Figure 5.3: Dendrogram for the overfitted signal function. Running on the same simulated data as mentioned in Figure 5.2.

5.3.3 Relations with Cumulative Sum Test (CUSUM) test in detecting Change-in-mean

Under the change-in-mean problem, we make a remark on the relationship between the height of the dendrogram $d^{(\kappa)}$ and the CUSUM test in the literature [12, 42, 105].

The CUSUM test statistic is widely recognized for identifying changes in the mean of a monitored process sequence. It can be intuitively explained that the less distinction between the sample means of the data collected before and after a time t, the less evident a changepoint is. This is consistent with the construction of the dissimilarity Equation (5.16) and the height of the dendrogram. We will go into detail with the simplest case: detect a change-in-mean in one-dimensional Gaussian data.

Known variance Suppose that each y_i is normally distributed with mean μ_i $(i = 1, ..., n)$ and common variance σ^2, we test the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_n = \mu.$$

versus

$$H_1 : \mu_1 = \cdots = \mu_c \neq \mu_{c+1} = \cdots = \mu_n,$$

where c is the unknown location of the single changepoint. The likelihood-ratio test compares the maximum of the likelihood for a model with a change at c to the maximum of the likelihood for a model with no change is determined as:

$$LR_c = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(y_i - \bar{y}_{1:n})^2 - \sum_{i=1}^{c}(y_i - \bar{y}_{1:c})^2 - \sum_{i=c+1}^{n}(y_i - \bar{y}_{c+1:n})^2\right],$$

where the notation $\bar{y}_{s:t}$ for $t \geq s$ is used for the sample mean of $y_{s:t}$:

$$\bar{y}_{s:t} = \frac{1}{t-s+1}\sum_{i=s}^{t} y_i.$$

This likelihood-ratio test statistic can be rewritten as $LR_c = C_c^2/\sigma^2$, where C_c is the so-called CUSUM statistic

$$C_c = \sqrt{\frac{c(n-c)}{n}}|\bar{y}_{1:c} - \bar{y}_{(c+1):n}|.$$

Under the null hypothesis, the CUSUM statistic is the absolute value of a normal random with mean 0 and variance σ^2, and thus the likelihood-ratio statistic has a chi-squared distribution with 1 degree of freedom. Since the position of the changepoint c is unknown, the likelihood-ratio test is $LR = \max_{c\in\{1,\dots,n-1\}} LR_c$. This test is equivalent to one that is based on the maximum of the CUSUM statistics, C_c. We will detect a changepoint if $LR > a$ for some suitably chosen value a, and the choice of a will determine the significance level of the test. It has been proven that the asymptotic distribution for $\max_c C_c$ is the Gumbel distribution [140].

In terms of testing zero versus one changepoint problem, recall the height of the dendrogram at the second level for this specific change-in-mean problem

$$\begin{aligned} d_n^{(2)} = \left\| f_{\hat{\boldsymbol{\tau}}^{n,(2)},\hat{\boldsymbol{\theta}}^{n,(2)}} \right\|_n &= \left(\frac{(\tau_1-\tau_0)(\tau_n-\tau_1)}{\tau_n-\tau_0} |\hat{\mu}_1 - \hat{\mu}_n|^2 \right)^{1/2} \\ &= \left(\frac{(n\tau_1-n\tau_0)(n\tau_n-n\tau_1)}{n\tau_n-n\tau_0} |\hat{\mu}_1 - \hat{\mu}_n|^2 \right)^{1/2} \\ &= \left(\frac{c(n-c)}{n} |\bar{y}_{1:c} - \bar{y}_{(c+1):n}|^2 \right)^{1/2} = C_c, \end{aligned} \tag{5.28}$$

So, the height of the dendrogram coincides with the CUSUM statistic form at the pruning changepoint location $\tau_1 = c/n$.

Unknown variance When the noise variance, σ^2, is unknown, the likelihood ratio test becomes

$$LR_c = n \log \left(\frac{\sum_{i=1}^{n} (y_i - \bar{y}_{1:n})^2}{\sum_{i=1}^{c} (y_i - \bar{y}_{1:c})^2 + \sum_{i=c+1}^{n} (y_i - \bar{y}_{c+1:n})^2} \right).$$

We can still re-write this test as a monotonic function of the CUSUM statistic, i.e.,

$$LR_c = n \log \left(\frac{S^2}{S^2 - C_c^2} \right),$$

where $S^2 = \sum_{i=1}^{n}(y_i - \bar{y}_{1:n})^2$ is the residual sum of squares under a model with no change, and C_c is the CUSUM statistic. Hence, the likelihood-ratio test for a change is still equivalent to one based on $\max_c C_c$. Notice that the distribution of the max-CUSUM statistics under the null will differ from the known variance one. We refer to [12, 42] for further discussion on the use of CUSUM.

In both instances, whether the variance is known or unknown, the CUSUM test is capable of identifying a shift in the mean. Despite the CUSUM test form and the height of the dendrogram under the change-in-mean setup being similar, the CUSUM test looks for the location $c \in \{1, ..., n-1\}$ where C_c achieves the maximum value, indicating evidence of a strong difference between the two neighboring means. In contrast, we find the position where the dissimilarity $d_j = C_{nj}$ $(j \in [k-1])$ is smallest, suggesting weak evidence of a changepoint present there, so the changepoint needs to be pruned.

5.4 Dendrogram Selection Criterion (DSC)

As mentioned in the previous section, Theorem 5.4 provides useful information for building a criterion for model selection given the constructed dendrogram after running the DPM algorithm. This section discusses the formation of the Dendrogram Selection Criterion (DSC) and its consistency theory in choosing the number of changepoints. The construction of DSC is notable for considering the information on dendrogram heights. Particularly, suppose we obtain over-fitted MLE $\hat{f}^n = f_{\widehat{\boldsymbol{\tau}}^{n,(k)}, \widehat{\boldsymbol{\theta}}^{n,(k)}}$ from n data $(t_i, y_i)_{i=1}^n$ with equally spaced (t_i). Let $d_n^{(\kappa)}$ be the height of the $\kappa-$th level of the dendrogram of this function, with $\kappa = 2, \ldots, k$. Consider

$$\mathrm{DSC}_n^{(\kappa)} := -\left[\left(d_n^{(\kappa)}\right)^{\beta} + \omega_n \bar{\ell}_n\left(f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa)}, \widehat{\boldsymbol{\theta}}^{n,(\kappa)}}\right)\right], \tag{5.29}$$

where $\overline{\ell}_n(f) := \frac{1}{n}\sum_{i=1}^n \log p(y_i|f(t_i))$ is the empirical average log-likelihood, and ω_n and $\beta > 0$ are tuning hyper-parameter. A lower DSC value indicates a better fit. Let $k_n = \arg\min_{\kappa=2,\ldots,k} \mathrm{DSC}_n^{(\kappa)}$. The following lemma establishes the convergence behavior of the log-likelihood function $\overline{\ell}_n$. Because we compare log-likelihoods now, it requires some mild conditions on how the log-likelihood behaves across models (i.e., relative conditions of $p(y|\theta)$ and $p(y|\theta')$ when $\theta \approx \theta'$).

Condition (K3). *There exist positive constants c_α and c_β such that for all sufficiently small ϵ and $\theta_0, \theta \in \Theta$ such that $\|\theta - \theta_0\| \le \epsilon$, we have*

$$(1 - c_\beta\epsilon)\log p(y|\theta_0) + c_\alpha\epsilon \ge \log p(y|\theta) \ge (1 + c_\beta\epsilon)\log p(y|\theta_0) - c_\alpha\epsilon \quad \forall y.$$

Besides, there exists constant $\gamma_1, \gamma_2 > 0$ so that $\mathbb{P}_{y\sim p(\theta^0_{j'})}(|\log p(y|\theta^0_j)| \ge z) \le e^{-\gamma_1 z^{\gamma_2}}$ for all $z \ge 0$ and $j, j' \in [k_0]$.

We can show that condition (K3) satisfies many popular kernels p belonging to the exponential family.

Lemma 5. *Assume conditions (K1-K3). Let*

$$\overline{\mathcal{L}}_0(f) = \frac{1}{n}\sum_{i=1}^{k_0}\sum_{j:\tau^0_{i-1}\le t_j<\tau^0_i} \mathbb{E}_{Y_j\sim p(\theta^0_i)}\log p(Y_j|f(t_j))$$

be the population average log-likelihood of the model under signal function f, then with probability increasing to 1 as $n \to \infty$, we have:

1. *(Over-fitted levels) there exists constant $C_o > 0$ only depends on k, p, Θ such that*

$$\overline{\ell}_n\left(f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa)},\widehat{\boldsymbol{\theta}}^{n,(\kappa)}}\right) - \overline{\mathcal{L}}_0(f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}) < C_o\left(\frac{\log(n)}{n}\right)^{1/2} \quad \forall\kappa \in [k_0, k]; \tag{5.30}$$

2. *(Exact-fitted level) there exists constant* $C_e > 0$ *only depends on* $k, p, \boldsymbol{\tau}^0, \boldsymbol{\theta}^0$ *such that*

$$\left|\overline{\ell}_n\left(f_{\widehat{\boldsymbol{\tau}}^{n,(k_0)},\widehat{\boldsymbol{\theta}}^{n,(k_0)}}\right) - \overline{\mathcal{L}}_0\left(f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right)\right| \leq C_e\left(\frac{\log(n)}{n}\right)^{1/2}; \tag{5.31}$$

3. *(Under-fitted levels) there exists constant* $C_u > 0$ *only depends on* $k, p, \boldsymbol{\tau}^0, \boldsymbol{\theta}^0$ *such that*

$$\left|\overline{\ell}_n\left(f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa)},\widehat{\boldsymbol{\theta}}^{n,(\kappa)}}\right) - \overline{\mathcal{L}}_0\left(f_{\boldsymbol{\tau}^{0,(\kappa)},\boldsymbol{\theta}^{0,(\kappa)}}\right)\right| \leq C_u\left(\frac{\log(n)}{n}\right)^{1/2}, \quad \forall \kappa \in [k_0 - 1]. \tag{5.32}$$

As described in Equation (5.29), DSC is the combination of the two main ingredients: log-likelihood and dendrogram information. For an intuitive explanation of how DSC works, consider the following: at overfitted levels ($\kappa \in [k_0 + 1, k]$), the dendrogram heights tend to 0 (as shown in Theorem 5.4), while the empirical average log-likelihood remains nearly constant, approaching $\overline{\mathcal{L}}_0(f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0})$ (as seen in Equation (5.30) - Lemma 5). At the overfitted level, $d_n^{(k_0)} \neq 0$ (Theorem 5.4), while the empirical average log-likelihood is not significantly different from those at the overfitted levels (as seen in Equation (5.30) and Equation (5.31) - Lemma 5). When it comes to the under-fitted levels ($\kappa \in [k_0 - 1]$), there is a significant drop in the log-likelihood function (as seen in Equation (5.32) combining with the fact that $\overline{\mathcal{L}}_0(f_{\boldsymbol{\tau}^{0,(\kappa)},\boldsymbol{\theta}^{0,(\kappa)}}) < \overline{\mathcal{L}}_0(f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0})$), while $d_n^{(\kappa)}$ is getting larger (as in Theorem 5.4). By choosing appropriate tuning parameters, DSC can capture this information and minimize the value at the exact fitted level.

We are now prepared to demonstrate the consistency of DSC in correctly selecting the number of changepoints in the following theorem.

Theorem 5.5: Consistency theorem for DSC

Assume conditions (K1-K3). For any $\beta > 0$ and ω_n such that $1 \ll \omega_n \ll (n/\log n)^{1/2}$, we have $k_n \to k_0$ in probability, i.e, DSC is consistent.

Data scaling problem and suggestion on choosing the tuning parameters of DSC. When the data is on different scales, the log-likelihood remains the same, but the parameters $\boldsymbol{\theta}$ will take on the scale of the data; this affects the DSC score because its penalty term $d_n^{(\kappa)}$ considers the differences between the parameters. For robustness in computation DSC, the formula of DSC at the model with κ segments is defined as follows.

$$\mathrm{DSC}_n^{(\kappa)} := -\left[\frac{\left(d_n^{(\kappa)}\right)^{\beta}}{\max_{\kappa' \in [k]}\left(d_n^{(\kappa')}\right)^{\beta}} + \omega_n \frac{\overline{\ell}_n\left(f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa)},\widehat{\boldsymbol{\theta}}^{n,(\kappa)}}\right)}{\left|\overline{\ell}_n\left(f_{\widehat{\boldsymbol{\tau}}^{n,(k)},\widehat{\boldsymbol{\theta}}^{n,(k)}}\right)\right|}\right], \tag{5.33}$$

where $\beta = 1/2$, $\omega_n = \log(n)$ based on our experience in conducting experiments (also as a suggestion in [31]). The maximal dendrogram height, $\max_{\kappa' \in [k]}\left(d_n^{(\kappa')}\right)^{1/2}$, and $\left|\overline{\ell}_n\left(f_{\widehat{\boldsymbol{\tau}}^{n,(k)},\widehat{\boldsymbol{\theta}}^{n,(k)}}\right)\right|$ (which tends to $\left|\overline{\mathcal{L}}_0(f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0})\right|$ a.s.) are two rescaled terms for the square root of the dendrogram height and the empirical average log-likelihood, respectively. Theorem 5.5 still holds in this case since these two terms tend to be constants with high probability as the sample size gets large (Theorem 5.4 and Lemma 5).

Remark 1. *Since DSC uses the dendrogram heights information, it cannot be separated from the DPM algorithm. Therefore, whenever we mention DSC, we mean the DPM algorithm is run beforehand. The* ***DPM-DSC*** *is used to indicate that we will initially run the DPM algorithm for a maximum likelihood estimate (MLE) overfitted signal function. This will give us a dendrogram containing all pruned and merged functions and the associated dendrogram heights. We will then calculate the Dendrogram Selection Criterion (DSC) at each pruned and merged step and select the segmentations associated with the Dendrogram level that has the lowest DSC value.*

5.5 Experiments

In this section, we conducted experiments to test how the theory holds up in different situations. We also compared our proposed DSC in Equation (5.33) to popular criteria used in the penalized approach and state-of-the-art methods in changepoint detection problems. Finally, we demonstrated a practical application of our proposed method using a real dataset.

All the experiments below were implemented in Python. We used the package `ruptures` (see [128]) with either the dynamic programming (Dynp) or the Binary Segmentation (BinSeg) methods to seek the MLE of parameters of the model. Recall that the computational cost of the Dynp method is heavy ($O(n^2)$). However, this algorithm is well-known for guaranteeing global optimal solutions [8, 14, 125, 128]. With this in mind, when it comes to the exact solution for the optimal problems, we used the Dynp method with a limited sample size range. For large sample-size problems, we used the BinSeg method, which is well-known for its linear computational cost ($O(n \log n)$).

5.5.1 Synthetic data

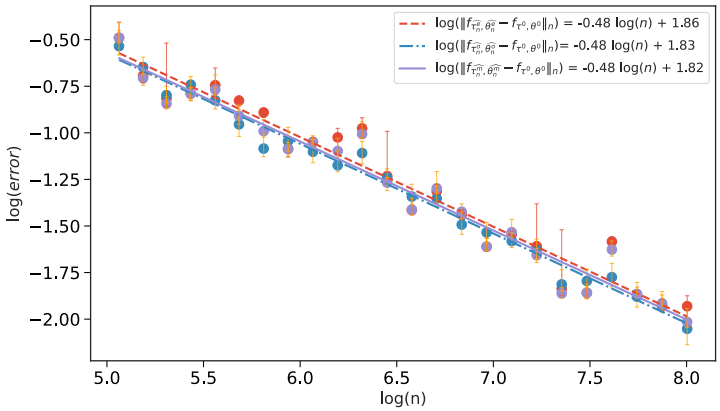
Convergence rates

In this experiment, we illustrated the convergence rate of the signal function of the estimated exact-fitted, over-fitted, and merged changepoint function to the truth (refer to Theorem 5.2 and Theorem 5.3). The two-dimensional data was generated from a normal mean multiple changepoint models with constant standard deviation being $\sigma = 1$, the true parameters $\boldsymbol{\tau}^0 = (0.25, 0.5, 0.75)$ and $\boldsymbol{\theta}^0 = ([0, 0], [2, -1], [-1, 1], [3, 2])$. We considered the natural logarithm of the sample size $\log(n)$ ranging from 5 to 8 (so that n ranges from 148 to 2980) and generated n samples from the true model. We then overfitted the data by $k = 6$ segments (5 changepoints) models. The Dynp

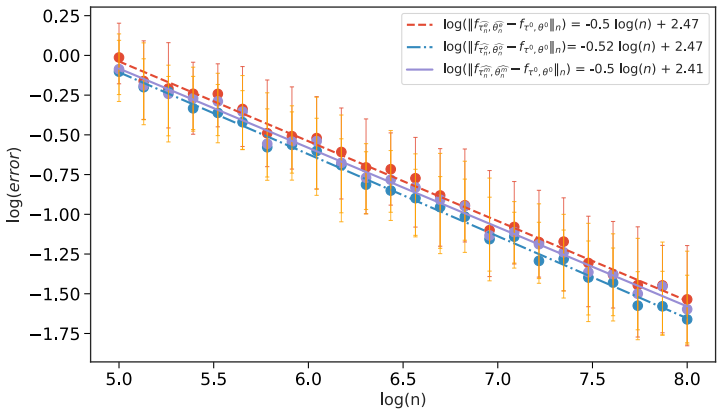
method was adopted to search for the MLE under both exact-fitted and over-fitted setups. In the `ruptures` package, the associated cost function is called $c_{i.i.d}$, and the sum of this cost function is equal to the negative log-likelihood. We applied the DPM algorithm (Algorithm 3) for the MLE overfitted function $f_{\hat{\boldsymbol{\tau}}_n^o,\hat{\boldsymbol{\theta}}_n^o} = \hat{f}_n$ to get the pruned and merged function at the k_0-th level $f_{\hat{\boldsymbol{\tau}}_n^m,\hat{\boldsymbol{\theta}}_n^m} = f_{\hat{\boldsymbol{\tau}}^{n,(k_0)},\hat{\boldsymbol{\theta}}^{n,(k_0)}}$ in the dendrogram, then measured the error of all estimators $f_{\hat{\boldsymbol{\tau}}_n^o,\hat{\boldsymbol{\theta}}_n^o}$, $f_{\hat{\boldsymbol{\tau}}_n^e,\hat{\boldsymbol{\theta}}_n^e}$, $f_{\hat{\boldsymbol{\tau}}_n^m,\hat{\boldsymbol{\theta}}_n^m}$ to the true function $f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}$ in the distances defined in Equation (5.14). Each experiment was repeated 64 times, and we plotted the average and quartile bar of the logarithm of the error in Figure 5.4(a).

The experiment showed that all cases shared the same fast convergence rate $n^{-1/2}$. It is noticeable that the pruned and merged cases and the over-fitted cases achieved faster convergence than the exact-fitted cases.

Furthermore, we designed the experiment to confirm the uniform convergent rates for any true parameters of the models (only providing the actual number of changepoints instead of fixing the true parameters of the model, i.e., $\boldsymbol{\tau}^0$ and $\boldsymbol{\theta}^0$). This has been supported by the results of Theorem 5.2 for the overfitted level where the constant C does not depend on the true signal function $f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}$. However, we have not yet had the theoretical results of these strong uniform convergence rates for the exact fitted and the pruned and merged functions. Under this extra experiment, the two-dimensional data was generated from a normal mean multiple changepoint model with the actual number of segments being $k_0 = 4$ (3 changepoints) and the constant standard deviation being $\sigma = 1$. The simulated procedure was as follows: $(p_1^0, p_2^0, p_3^0, p_4^0)$ followed a uniform Dirichlet distribution with a constant concentration parameter $\alpha = 1$, $\tau_j^0 = \sum_{i=1}^{j} p_i$ $(\tau_4^0 = 1)$, θ_j^0 followed an uniform distribution on an interval $[-5, 5]$ for $j = 1, .., 4$. Similar to the previous experiment, we considered the natural logarithm of the sample size $\log(n)$ ranging from 5 to 8 (so that n ranges from 148 to 2980) and generated n samples from the true model. We then overfitted the data by $k = 6$



(a) Convergence rates under the first setup with fixed $\boldsymbol{\tau}^0 = (0.25, 0.5, 0.75)$, and $\boldsymbol{\theta}^0 = ([0,0],[2,-1],[-1,1],[3,2])$



(b) Convergence rates under the second setup with a fixed number of changepoints being 3 and vary $\boldsymbol{\tau}^0$, $\boldsymbol{\theta}^0$ at each replication and each sample size.

Figure 5.4: Rates of convergence of overfitted ($k = 6$), exact-fitted ($k = k_0 = 4$), and pruned and merged ($\kappa = k_0 = 4$) signal functions.

segments (5 changepoints) models and applied the DPM algorithm (Algorithm 3) for the MLE overfitted function $f_{\hat{\boldsymbol{\tau}}_n^o, \hat{\boldsymbol{\theta}}_n^o} = \hat{f}_n$ to get the pruned and merged function at the k_0-th level $f_{\hat{\boldsymbol{\tau}}_n^m, \hat{\boldsymbol{\theta}}_n^m} = f_{\hat{\boldsymbol{\tau}}^{n,(k_0)}, \hat{\boldsymbol{\theta}}^{n,(k_0)}}$ in the dendrogram, then measure the error of all estimators $f_{\hat{\boldsymbol{\tau}}_n^o, \hat{\boldsymbol{\theta}}_n^o}$, $f_{\hat{\boldsymbol{\tau}}_n^e, \hat{\boldsymbol{\theta}}_n^e}$, $f_{\hat{\boldsymbol{\tau}}_n^m, \hat{\boldsymbol{\theta}}_n^m}$ to the true function $f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0}$ in the distances defined in Equation (5.14). Each experiment was repeated 64 times, and we plotted the average and quartile bar of the logarithm of the error in Figure 5.4(b).

The experiment showed that we still have the fast convergence rate $n^{-1/2}$ for overfitted, exactfitted, pruned and merged signal functions. This suggests stronger results on the convergence rate theory for Dendrogram, which should be considered in future work.

Due to the heavy computation of the dynamic programming algorithm ($O(n^2)$), in this section, we reported the convergence rates when the sample size varies from 148 to 2980. For a wider range of sample sizes, we conducted an experiment detailed in Appendix A, with sample sizes ranging from 100 to 10000, using the binary segmentation (BinSeg) method to find MLEs. This method is well-known for its linear computation cost. The results remain the same as reported in this section.

Comparision of DSC to different information criteria

In this section, we set up two situations to compare DSC and some popular information criteria used for the changepoint detection problem. The following criteria were chosen as competitors to DSC:

Criterion	**Formula**
Bayesian Information Criterion (BIC) [141]	$\text{BIC} = -2\hat{\ell}_k + [kr + k - 1]\log n$
Akaike Information Criterion (AIC) [72]	$\text{AIC} = -2\hat{\ell}_k + 2[kr + k - 1]$
Modified Akaike Information Criterion (mAIC) [102]	$\text{mAIC} = -2\hat{\ell}_k + 2[kr + 3(k-1)]$
Modified Bayesian Information Criteria ·mBIC_1: [106],[21] ·mBIC_2: [144]	 ·$\text{mBIC}_1 = -2\hat{\ell}_k + \left[kr + 2\sum_{i=1}^{k}\left(\tau_i - \tau_{i-1} - \frac{1}{k}\right)^2\right]\log n$ ·$\text{mBIC}_2 = -2\hat{\ell}_k + 3(k-1)\log n + \sum_{i=1}^{k}\log(\tau_i - \tau_{i-1})$
Minimum Description Length (MDL) [113],[94], [96]	$\text{MDL} = -2\hat{\ell}_k + 2\log(k-1) + 2(k-1)\log n + r\sum_{i=1}^{k}\log(n\tau_i - n\tau_{i-1})$

where $-2\hat{\ell}_k = -2\sum_{\kappa=1}^{k}\left[\sum_{j=n\tau_{i-1}+1}^{n\tau_i}\log p\left(y_j | f_{\widehat{\boldsymbol{\tau}}_n^{(\kappa)}, \widehat{\boldsymbol{\theta}}_n^{(\kappa)}}(t_j)\right)\right]$ is the negative of the twice maximum valued of the log-likelihood function, r is the number of free parameters in each segment, and k is the number of segments. All these criteria were discussed in

Section 3 of [50].

A. Simulation on different sample sizes. Under this setup, the two-dimensional data was generated from a normal multiple mean-shifted model with the actual number of segments being 7 (6 changepoints) and the constant standard deviation being $\sigma = 2$. We varied the decimal logarithm of the sample size $\log_{10}(n)$ from 2 to 4 (so that n ranges from 100 to 10000).

For the DSC approach, we overfitted the data by $k = 11$ and ran the BinSeg with the $c_{i.i.d}$ cost function (similar to the previous experiment) to find the associated MLE. For each sample size, the experiment was repeated 100 times. We then ran the DPM algorithm for the MLE overfitted function $\hat{f}_n$ to get the pruned and merged functions $f_{\hat{\boldsymbol{\tau}}^{n,(\kappa)},\hat{\boldsymbol{\theta}}^{n,(\kappa)}}$ $(\kappa \in [k])$ in the dendrogram. The DSC scores were calculated at each pruned and merged step, and the segmentations with the lowest score value are reported.

For all of the penalized cost function problems with the 6 mentioned criteria, we ran BinSeg to find the minimum solutions for each of the corresponding penalized cost functions listed at the beginning of Section 5.5.1 while considering the maximum number of change points to search was 10.

Figure 5.5 shows that the proportion of correct choices (the number of times the correct number of actual changepoints are detected out of 100 replications for each sample size) for DSC increase from 57% to over 80% as the sample size n increases from 300 ($\log_{10}(n) = 2.5$) to 10000 ($\log_{10}(n) = 4$). In contrast, other criteria have this proportion below 60% for all considered sample sizes. The average choice number of changepoints of DSC is around 5.75 (close to the number of true changepoints 6) when $\log_{10}(n)$ ranges from 2.5 to 4. Meanwhile, all six remaining criteria attempt to overfit the model. Among these six criteria, BIC and mBIC_1 demonstrate the best performance; following that, MDL and mBIC_2, and finally, mAIC and AIC.

Notice that the BIC does not behave well even with a large sample size. It can be explained since the classical BIC of Schwarz (1978) [119] setting does not theoretically fit segmentation-related problems as mentioned in [144].

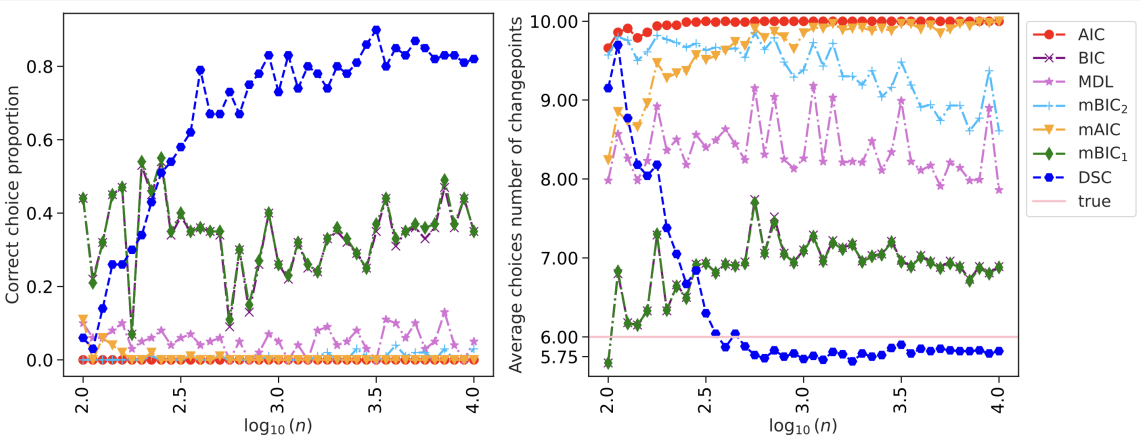


Figure 5.5: *Simulations results on the different sample sizes - Section 5.5.1A.*: Comparison between DSC and information criteria AIC, BIC, mAIC, mBIC_1, mBIC_2, mAIC, MDL in correct choice proportion and the average choices number of segments. The number of actual changepoints is 6 (in a solid light pink line). We considered the maximum value of the number of changepoints 10. Reran the experiments using the binary segmentation method.

We chose the BinSeg method for running the experiment because of its linear computational cost. Dynamic programming (Dynp) and PELT will be the preferred choices for finding the exact solution for the optimization problems. However, Dynp has $O(n^2)$ in computational cost. So, it is not a good option when running with a large sample size (says $n > 3000$). To compare all criteria using the Dynp, we refer to Appendix B, where we rerun the whole experiment but for a smaller sample size range. The results remained the same as we report in this section. PELT is a well-known method providing fast computation with the linear cost in best-case scenarios [109]. In the next experiment, we will use PELT to run the comparison.

B. Simulation on the different number of changepoints. We reproduced one of the experiments in [50] where, under the setup, the segment length generation ensured at least 50 for each segment. In the previous experiment, the number of actual

changepoints was fixed to be 6 while varying the changepoint locations and the sample sizes. Allowing for cases where the segment length is small may discourage the usage of BIC and its variants since the theoretical development of information-criterion-based approaches was based on assuming the number of changepoints is fixed and the spacing between consecutive changepoints is large (see [141], [9]). Aware of this phenomenon, we followed the setup in [50] to ensure that we compare all criteria in their appropriate context. Let the number of changepoints rise from 1 to 15, and the length of the k segments was generated by

$$(L_1, L_2, ..., L_k) \sim 50 + \text{Multinomial}(50 \cdot k, p_1, ..., p_k),$$

where $(p_1, ..., p_k)$ follows a uniform Dirichlet distribution with a constant concentration parameter $\alpha = 1$. The locations of changepoints were $n\tau_j = \sum_{i=1}^{j} L_i$ for $j = 1, ..., k$ ($n = \sum_{i=1}^{k} L_i$, $\tau_k = 1$). The mean shift pattern for the normally distributed time series was as follows

$$(\mu, \sigma^2) \overset{n\tau_1}{\to} (\mu + \Delta\mu, \sigma^2) \overset{n\tau_2}{\to} (\mu, \sigma^2) \overset{n\tau_3}{\to} (\mu + \Delta\mu, \sigma^2) \overset{n\tau_3}{\to} (\mu, \sigma^2) \cdots .$$

As the paper suggests, we choose the fixed mean shift of $\Delta = 1.25$, the initial mean $\mu = 1$, and $\sigma = 1$.

To measure the performance, we used three criteria: the precision rate, the recall rate, and the ratio of changepoint numbers. Let $\hat{k} - 1$, $k - 1$, and $\tilde{k} - 1$ be the number of detected changepoints, true changepoints, and correctly detected changepoints. If the detected changepoint is located in the interval $[n\tau - 5, n\tau + 5]$, it is said to be a correctly detected changepoint. The precision score is defined as

$$\text{Precision Rate} = \frac{\tilde{k} - 1}{\hat{k} - 1}.$$

The recall score is defined as

$$\text{Recall rate} = \frac{\tilde{k} - 1}{k - 1},$$

and the ratio of changepoint numbers is

$$\text{Ratio} = \frac{\hat{k} - 1}{k - 1}.$$

We used the Pruned Exact Linear Time (PELT) [109] algorithm implemented in the `ruptures` package with the customized cost functions for running the penalized problem with six criteria AIC, BIC, mAIC, mBIC_1, mBIC_2 and MDL. The PELT algorithm and Dynp are optimal detection methods for returning the exact solution for changepoint detection problems [128]. In DPM-DSC, we overfit the model and search for the associated MLE solution via the Dynp algorithm. Then, we apply the pruning and merging procedure (DPM algorithm) and calculate DSC scores in each pruning and merging step. Finally, we return the best segmentations with the lowest DSC scores. For the choice of the overfitted number of segments, to save the computational cost in running dynamic programming, we suggest using the number of detected changepoints after running AIC or BIC as the suggestion for the overfitted values. Since it is well known that both criteria have the same overestimation issue [74, 75, 124, 144]. In this experiment, we used the number of detected changepoints after running AIC as the overfitted value in running DPM-DSC.

Figure 5.6 shows the results of the Monte Carlo simulation with 100 times replication in the precision rate, recall rate, and the ratio of changepoint numbers. The results agree with the reported results in [50] with the six considered criteria. The precision rates provided by BIC, mBIC_1, mBIC_2, mAIC, and DSC closely align with over 80% of the accurately identified points, exceeding the precision offered by AIC and

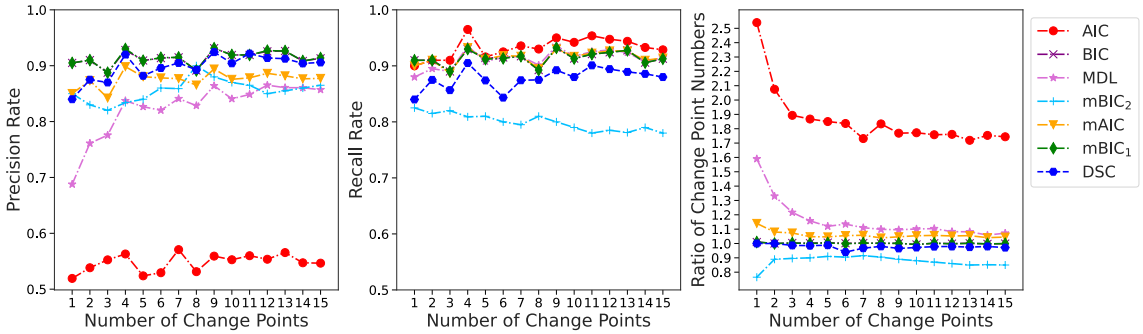


Figure 5.6: *Simulation results on the different number of changepoints - Section 5.5.1B.* Comparison between DSC to information criteria AIC, BIC, mAIC, mBIC_1, mBIC_2, mAIC, MDL in the precision rate (**left**), the recall rate (**center**), and the ratio of changepoint numbers (**right**).

MDL. We can see the improvement in the precision rate of MDL when the number of changepoints is increased. DSC has a competitive precision rate range from 0.86 to 0.9 when the number of changepoints rises from 1 to 15.

In terms of the recall rate, mBIC_2 consistently yields the lowest value in all cases. Its heavy penalty term of $3 \log n$ on newly emerging changepoints can explain it. For all other criteria, including DSC, the recall rates are mostly over 0.85.

Finally, by looking at the plot of the ratio of changepoint numbers, we can see how AIC overfitted and MDL slightly underfitted the model while other criteria, including DSC, accurately estimated the correct number of changepoints. It is evident that DSC competes with BIC and mBIC_1; the three criteria mostly align throughout the increasing number of changepoints.

Exploring DPM-DSC with a different kernel

This section will compare our proposed method to the Tail-Greedy Unbalanced Haar (TGUH) method by [45], which belongs to the same class of bottom-up ideas as our approach. Both methods are comparable in working on a univariate data sequence, which is modeled as a piecewise-constant function plus i.i.d Gaussian noise (refer to an experiment in Appendix C). To make another challenge to our algorithm, we generated

data under the Poisson kernel. Poisson models are well-known for counting data, where observations are non-negative integers (e.g., detecting changes in traffic counts [142], call rates, or failure rates in a system). In this experiment, the one-dimensional data was simulated from a multiple changepoints model, with the number of segments being 4 and the sample size $n = 400$. In each segment, the data points came from an i.i.d Poisson(λ). The lambda values in each segment were $1.14, 9.04, 1.26$, and 4.36, respectively. We then ran DPM-DSC with the overfitted number of segments being 11 and used the BinSeg method to find the minimum of the $c_{i.i.d}$ cost function customized to work for the Poisson kernel. The computational costs for DPM-DSC are then competitive compared to the TGUH method since we only run BinSeg once for $k = 11$ segments, then apply our pruning and merging procedure with DSC to find the best changepoint locations.

Regarding the Tail-Greedy Unbalanced Haar method implemented in `breakfast` package in R, we run the `breakfast()`function on the simulated data. Figure 5.7 shows the results where DPM-DSC successfully detects the correct number and locations of changepoints. The Tail-Greedy Unbalanced Haar method returns 7 changepoints, including 3 true changepoints. This can be explained since the author mentioned that their current methods specifically support changes in the mean model with i.i.d. Gaussian noise, assuming constant variance. Applying the change in the mean model to the multiple changepoints with Poisson kernel data, where the variance varies accordingly as the mean change, may reduce the efficiency of detecting the true changepoint locations.

Additionally, the `breakfast` package offers many different state-of-the-art methods for detecting change-in-mean problem, including Isolated Detection method [2], Narrowest-Over-Threshold (NOT) [9], Wild Binary Segmentation [44], and Wild Binary Segmentation 2 [46].

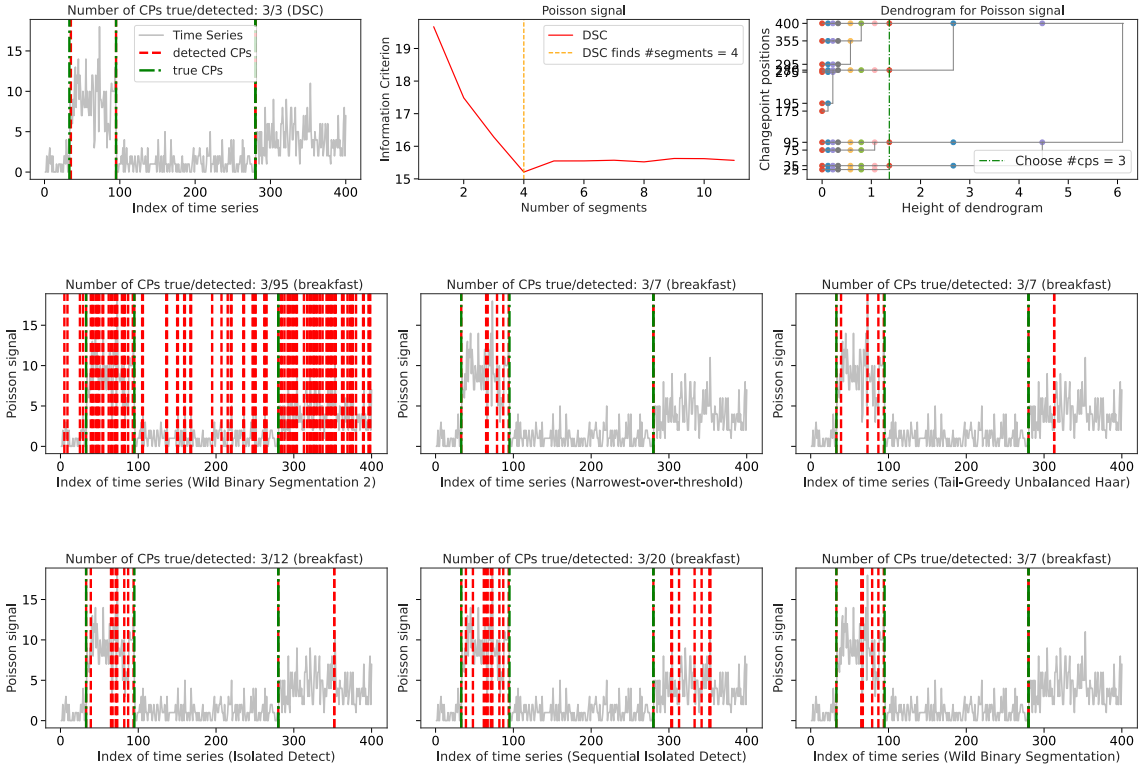


Figure 5.7: Comparision between DPM-DSC and methods provided by `breakfast` package for a one-dimensional simulated data from a multiple changepoints model with Poisson kernel. The green dashed lines represent the true changepoints, and the red dashed lines represent the detected changepoints in each method.

Figure 5.7 shows reported results for all the listed methods. We can see the best results belong to NOT, Tail-Greedy Unbalanced Haar, and Wild Binary Segmentation with the 3 correctly detected changepoints over 7 detected changepoints.

This experiment demonstrates the importance of being able to choose a suitable kernel in MLE methods when dealing with diverse types of data.

5.5.2 Real data experiment

In this section, we compared the detection performance of the information criteria from the study by [50] with our approach using DPM-DSC when analyzing real datasets. We examined the SCADA signals of wind turbines, which were the same datasets used

in the previous study. Recognizing changes in wind turbine operation is essential for identifying and addressing potential issues before they become serious. Additionally, it provides valuable insights for planning routine inspections and maintenance. The dataset was explored by [91] and [50] in detail.

There were 11 proposed SCADA signals of a wind turbine. The labeled changepoints were included in the dataset for each of the signals. The original data was sampled at a typical 10-minute resolution, after which the signals were averaged on a daily basis.

For comparison purposes, we chose one signal discussed in [50], which was the nacelle temperature signal (collected from 1 January 2017 to 12 September 2018, totaling 620 days), and another signal, the gear bearing temperature (collected from 1 January 2017 to 18 May 2019, totaling 868 days). We used the same assumption that the time series followed a normal distribution and that both mean and variance could change in the two signals.

Regarding the nacelle temperature signals, the positions of labeled changepoints were at the 357th and 493rd observations. For the gear-bearing temperature, the 682nd was labeled as the changepoint.

According to the findings of Gao (2024), the changepoint detection algorithm overfitted the signals when evaluated using six information criteria compared to the labeled changepoints provided in the datasets.

In the nacelle temperature signals, the ratios of true changepoints over the total detected changepoints were reported as 2/6 using AIC, 2/4 using BIC, 2/6 using mAIC, 2/4 using mBIC_1, 2/3 using mBIC_2, and 2/4 using MDL. We reproduced the experiment and compared it to the DPM-DSC approach (the overfit setup at $k = 5$ segments, which was the output from BIC) (Figure 5.8), where DSC returned the correct number of changepoints and their estimated locations.

In the gear-bearing temperature signals with one labeled changepoint, the rates were

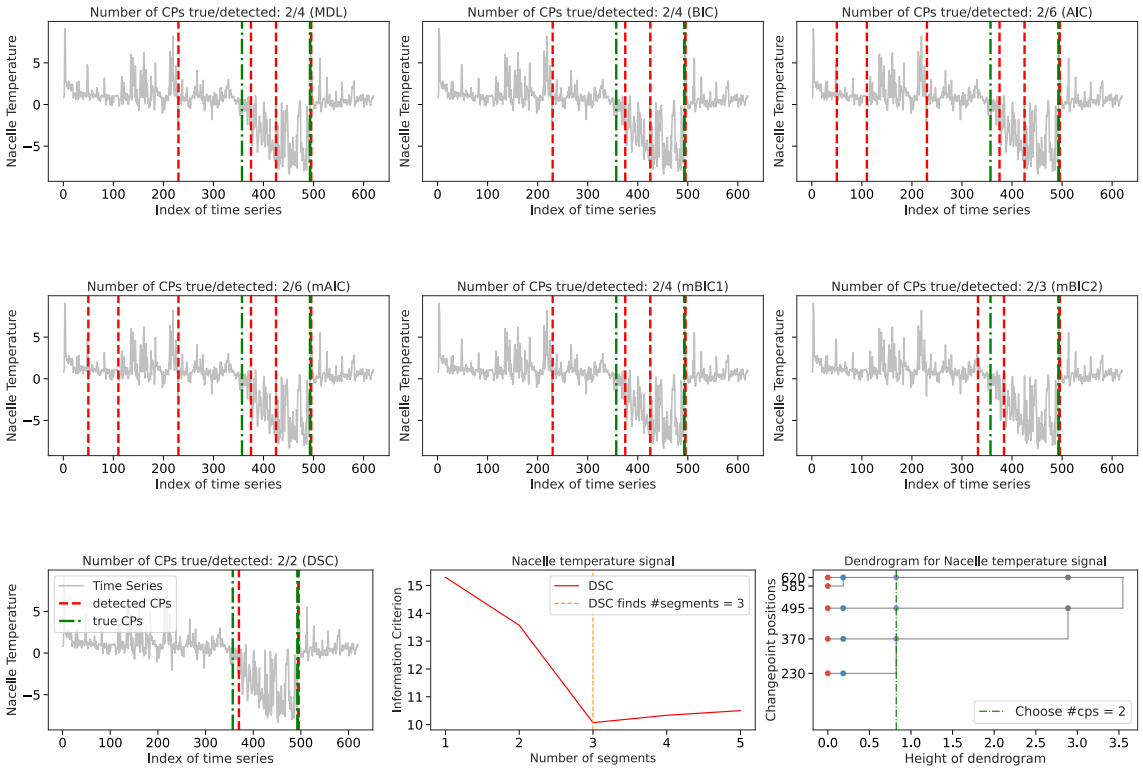


Figure 5.8: Nacelle temperature signals with two labeled changepoints (in vertical dashed green lines). The red dashed lines represent the estimated changepoint locations associated with each information criterion.

1/12 using AIC, 1/7 using BIC, 1/12 using mAIC, 1/7 using mBIC_1, 1/6 using mBIC_2, and 1/8 using MDL. We then ran the DPM-DSC with the overfitted model at $k = 8$ segments, as suggested by running BIC (Figure 5.9). It returned one changepoint, and the associated location was close to the labeled changepoint.

5.6 Discussion

This chapter introduces the *Dendrogram Pruning and Merging* (DPM) algorithm, a novel approach for multiple changepoint detection that is both computationally efficient and theoretically grounded. Without requiring repeated model fitting, DPM constructs a binary tree of candidate changepoints, allowing for effective pruning and

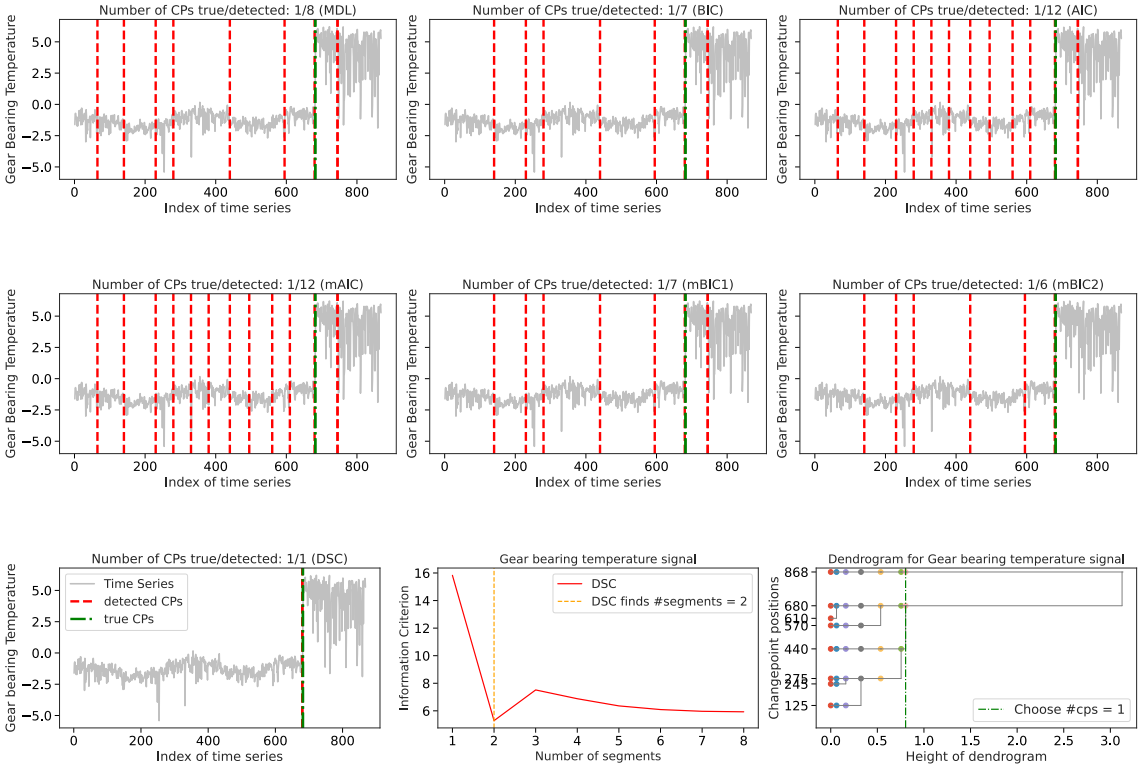


Figure 5.9: Gear bearing temperature signals with one labeled changepoint (in vertical dashed green lines). The red dashed lines represent the estimated changepoint locations associated with each information criterion.

merging in a single model pass. This significantly reduces computational cost while maintaining high accuracy.

A key advantage of DPM is its flexibility—it seamlessly integrates with different kernel setups and extends to multi-dimensional data, making it a versatile tool for complex segmentation tasks. Additionally, we propose the *Dendrogram Selection Criterion* (DSC), which leverages hierarchical structure and parameter distances for model selection. Theoretical guarantees ensure the consistency of our approach, and empirical results demonstrate its competitive performance against widely used information criteria. Details of the proofs are provided in Chapter 6.

These contributions offer a scalable and adaptable solution for changepoint detection

across various applications. Future work may explore extensions to dependent data structures or different changepoint detection frameworks, such as detecting changes in velocity discussed in Chapter 2, refine theoretical guarantees, and investigate hybrid methods that further enhance efficiency and robustness.

Appendix A: Rerun the convergence rate experiment by using BinSeg algorithm

As discussed at the end of Section 5.5.1, we used the BinSeg method with the $c_{i.i.d}$ cost function to find the MLE for over- and exact-fitted models. The two-dimensional data were generated from a normal multiple mean-shifted model, with the actual number of segments being 6 (5 changepoints) and the constant standard deviation set to $\sigma = 2$. We then overfitted the data with $k = 12$. We considered the decimal logarithm of the sample size, $\log_1 0(n)$, ranging from 2 to 4 (so that n ranged from 100 to $10,000$). For each sample size, the experiment was repeated 64 times, and we plotted the average and quartile bar of the decimal logarithm of the error in Figure 5.10. We still observe the same convergence rate as reported in Section 5.5.1

Appendix B: Rerun an experiment on different sample sizes by using the dynamic programming algorithm

This experiment aims to compare information criteria using the dynamic programming method. Since the computational cost for using Dynp is $O(n^2)$, we limited the range of the sample size, such as the natural logarithm of it $\log(n)$, from 6 to 7.9 (so that n ranges from 403 to 2697). Under this setup, the two-dimensional data was generated from a normal multiple mean-shifted model with the actual number of

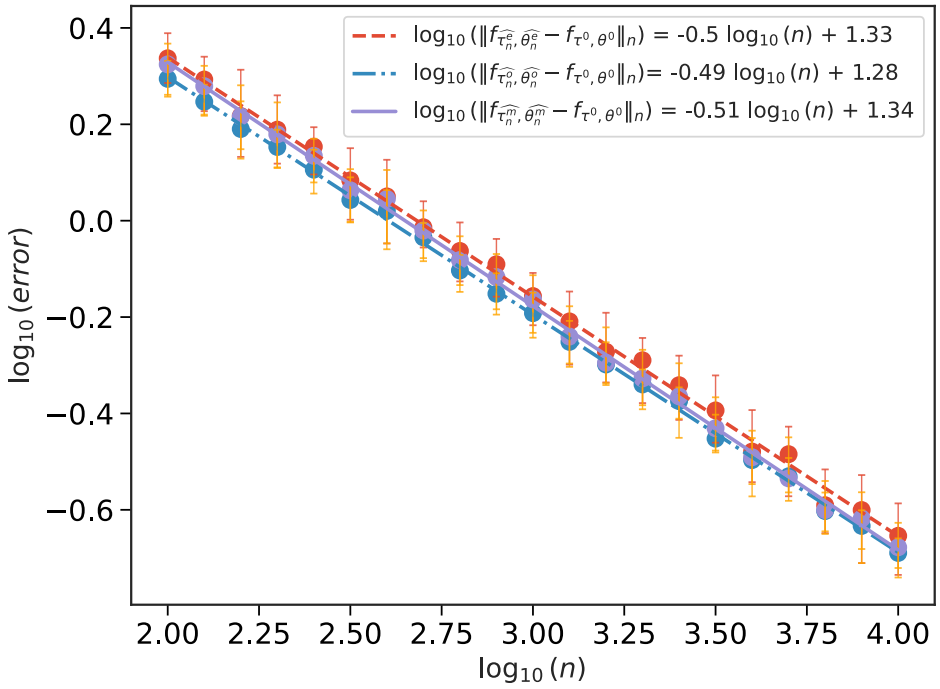


Figure 5.10: Rates of convergence of overfitted, exact-fitted, and merge signal functions (experiment used the BinSeg method), $k_0 = 6$, $k = 12$.

segments being 7 (6 changepoints) models and the constant standard deviation being $\sigma = 2$. We then overfitted the data by $k = 11$ and chose the dynamic programming with the $c_{i.i.d}$ cost function to find the associated MLE. For each sample size, the experiment was repeated 100 times. Applying the merging procedure in Section 5.3, we pruned and merged the overfitted function $\hat{f}_n$ to get the pruned and merged functions $f_{\hat{\boldsymbol{\tau}}^{n,(\kappa)},\hat{\boldsymbol{\theta}}^{n,(\kappa)}}$ ($\kappa \in [k]$) in the dendrogram. We then computed the DSC scores in each pruning and merging step and report the segmentations with the lowest score value. After the dynamic programming was executed with $k = 11$, it stored the best solutions for the MLEs for all $k = 1, ..., 11$. We then chose among computed segmentations the one that minimizes the penalized problem under each considered criterion. Figure 5.11 shows the reported results. The left panel shows the proportion of correct choices for each criterion, representing the ratio of the times the correct number of changepoints

was detected over 100 repetitions in each experiment. The right panel displays the average number of changepoints chosen over 100 repetitions in each experiment. It can be observed that while most of the criteria considered overfit the model, DSC tends to select the correct number of changepoints as the sample size increases.

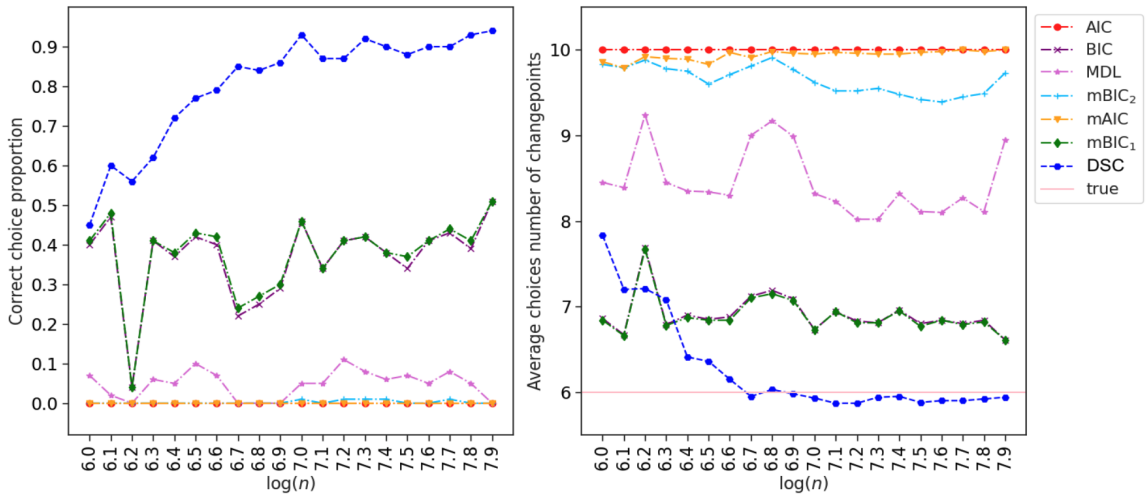


Figure 5.11: *Simulations results on the different number of changepoints*: Comparison between DSC and information criteria AIC, BIC, mAIC, mBIC_1, mBIC_2, mAIC, MDL in correct choice proportion and the average choices number of segments. The number of actual changepoints is 6 (in a solid light pink line). We consider the maximum value of the number of changepoints to be 10. The experiments use the dynamic programming method.

Appendix C: Comparision DPM-DSC to `breakfast` options

In this section, we provided an example comparing our proposed method with DSC to the algorithms available in the `breakfast` package, as discussed in Section 5.5.1, in a context where all methods in the package were claimed to work effectively. We simulated a univariate data sequence from a mean-shifted model with i.i.d. Gaussian noise, containing 10 actual changepoints, a sample size of $n = 400$, and a constant $\sigma = 2$. Figure 5.12 confirms the good performance of all algorithms, including the Isolated De-

tection method, Narrowest-Over-Threshold (NOT), Wild Binary Segmentation, Wild Binary Segmentation 2, and Tail-Greedy Unbalanced Haar. With fast computation, they all correctly detected the number of changepoints and their locations.

Regarding DPM with DSC, we also obtained competitive results, correctly detecting the number of changepoints and their locations. The computational cost was linear since we used a one-time BinSeg method at the overfitted level $k = 15$ and then merged the resulting dendrogram using our proposed procedure.

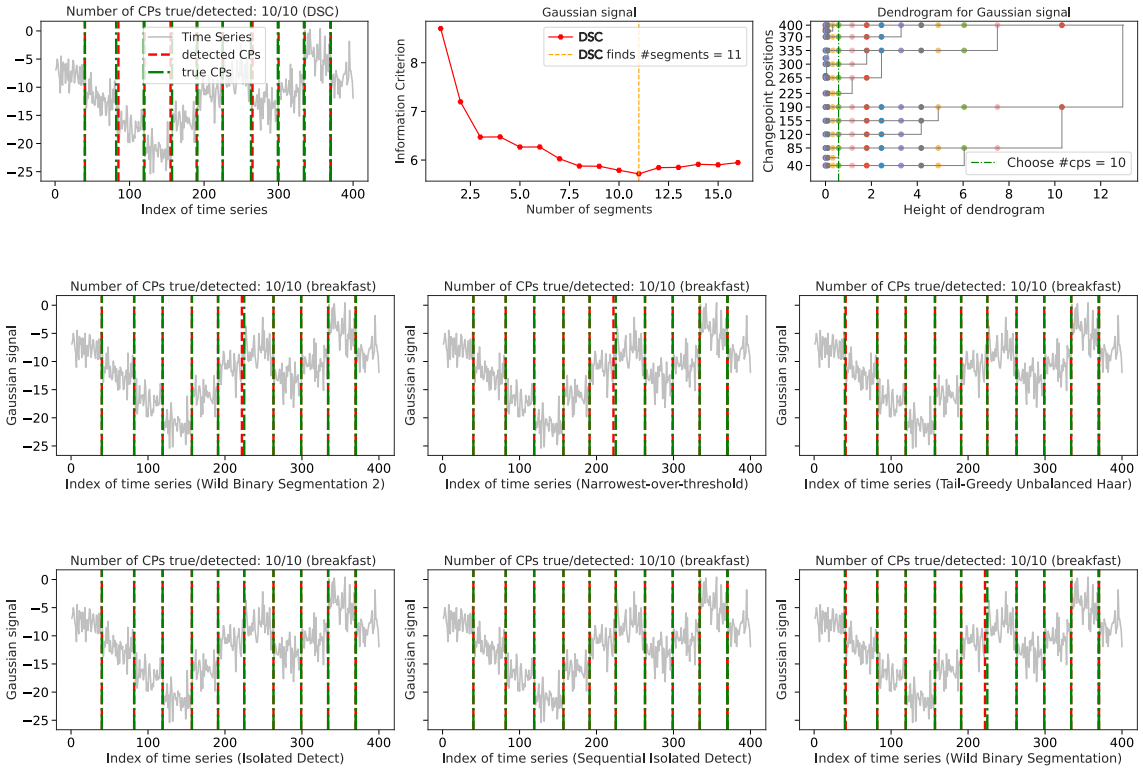


Figure 5.12: Comparision between DPM-DSC and methods provided by `breakfast` package for a one-dimensional simulated data from a multiple changepoints model with Gaussian kernel. The green dashed lines represent the true changepoints, and the red dashed lines represent the detected changepoints in each method.

Chapter 6

Theoretical Guarantees for the DPM Algorithm

6.1 Proof of Section 5.2: Convergence rate of exact- and over-fitted signal functions

The main goal of this section is to prove Theorem 5.2. We do it by applying the empirical process theory [129] (Theorem 5.1) with the same idea that we prove the similar result for CPLASS in Chapter 3. The key is to bound the Hellinger entropy number of the family of the changepoint model distributions.

6.1.1 Proof of Lemma 2

Proof of Lemma 2. This is a direct consequence of condition (K1). □

6.1.2 Proof of Theorem 5.2

For a true set of parameter $(\boldsymbol{\tau}^0, \boldsymbol{\theta}^0)$, denote for short $p^0 := (p(\cdot|f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}(i/n)))_{i=1}^n$. To derive the convergence rate of the model under the Hellinger process, we aim to bound the bracketing entropy of the model space and then use an application of empirical process theory to yield the result.

Proof of Theorem 5.2. The proof is divided into a few small steps.

Step 1. Showing convergence for densities in the Hellinger process.

We aim to bound the entropy of the class of densities corresponding to $k-$pieces piecewise constant functions.

$$H_B(\delta) \lesssim \log(n/\delta). \tag{6.1}$$

In the following, we continue to use the technique [51] for our model.

Step 1.1. Covering the space of changepoint models with fixed changes under ℓ_∞ norm. Suppose that the set of changes $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{k-1}) \in \mathcal{T}_\uparrow^{k-1} \cap \{1/n, 2/n, \ldots, 1\}$ and $\boldsymbol{\theta} \in \Theta^k$. Given $\delta > 0$, because Θ is compact in $\mathbb{R}^m$, for each $\theta_j \in \mathbb{R}^m$ $(j \in [k])$, we can choose an δ-net $(\theta_{i_j})_{i_j=1}^N$ of Θ under ℓ_∞ norm with the cardinality $N \asymp (1/\delta)^m$. Consider the net

$$B = \{(\theta_{i_1}, \ldots, \theta_{i_k}) : i_1, \ldots, i_k \in [N]\} \subset \Theta^k.$$

We have that $|B| \asymp (1/\delta)^{km}$, and for every $\boldsymbol{\theta} \in \Theta^k$, there exists an element $\tilde{\boldsymbol{\theta}}$ in B that is δ-close to it under ℓ_∞ norm. For all $j \in [k]$ and $t \in [\tau_j, \tau_{j+1})$, we have

$$\left\| f_{\boldsymbol{\tau},\boldsymbol{\theta}}(t) - f_{\boldsymbol{\tau},\tilde{\boldsymbol{\theta}}}(t) \right\|_\infty = \left\| \theta_j - \tilde{\theta}_j \right\|_\infty \leq \delta.$$

Hence, $\left\| f_{\boldsymbol{\tau},\boldsymbol{\theta}}(t) - f_{\boldsymbol{\tau},\tilde{\boldsymbol{\theta}}}(t) \right\|_\infty \leq k\delta \quad \forall t \in [0,1]$. Using condition $(K2)$, it implies

$$\sup_{y_i \in \mathbb{R}^d} |p(y_i | f_{\boldsymbol{\tau},\boldsymbol{\theta}}(i/n) - p(y_i | f_{\boldsymbol{\tau},\tilde{\boldsymbol{\theta}}}(i/n)| \lesssim k\delta, \quad \forall i \in [n].$$

Step 1.2. Covering the space of changepoint models with fixed changes under the Hellinger distance (with bracketing). For every $\varepsilon > 0$, from the previous step, we have collection of densities $\{p_j\}_{j=1}^N$ with $p_j = (p_{j1}, \ldots, p_{jn})$ on $\mathcal{Y}^n$,

the associated parameters are in B and $N \asymp (k/\varepsilon)^{km}$ such that for every $\boldsymbol{\theta} \in \Theta^k$, there exists a p_j satisfying

$$\sup_{y_i \in \mathbb{R}^d} |p(y_i|f_{\boldsymbol{\tau},\boldsymbol{\theta}}(i/n)) - p_{ji}(y_i)| \lesssim \varepsilon, \quad \forall i \in [n]. \tag{6.2}$$

Moreover, we have the parameters of p_{ij} are in a compact space $\Theta \subset \mathbb{R}^m$ for all $i \in [n], j \in [N]$. Hence, we can find an upper bound (envelop)

$$H(y) = \begin{cases} d_1 \exp(-d_2 \|y\|^{d_3}), & \|y\| \geq D, \\ \sup_\theta \|p(\cdot|\theta)\|_\infty, & \text{otherwise} \end{cases} \tag{6.3}$$

of $p_{ij}(y)$ for all $j \in [N]$ and $i \in [n]$, for some constants $d_1, d_2, D > 0$. We can construct brackets $[p_j^L, p_j^U]$ with $p_j^U = (p_{j1}^U, \ldots, p_{jn}^U)$ and $p_j^L = (p_{j1}^L, \ldots, p_{jn}^L)$ as following"

$$\begin{aligned} p_{ji}^L(y) &= \max\{p_{ji}(y) - \varepsilon, 0\}, \\ p_{ji}^U(y) &= \min\{p_{ji}(y) + \varepsilon, H(y)\} \end{aligned}$$

With this construction, (6.2) implies

$$p_{ji}^L(y_i) \leq p(y_i|f_{\boldsymbol{\tau},\boldsymbol{\theta}}(t_i)) \leq p_{ji}^U(y_i) \quad \forall y_i \in \mathbb{R}^d, i \in [n]. \tag{6.4}$$

Therefore, this collection of brackets satisfies condition (ii) in the Definition 5.1 of covering with bracketing. We now check condition (i). For any j, i and $\bar{D} \geq D$,

$$\begin{aligned} \int_{\mathbb{R}^d} (p_{ji}^U - p_{ji}^L) dy &\leq \int_{\|y\| \leq \overline{D}} 2\varepsilon dy + \int_{\|y\| \geq \overline{D}} H(y) dy \\ &\lesssim \varepsilon \overline{D}^d + \overline{D}^d \exp\left(-b_2 \overline{D}^{d_3}\right), \end{aligned} \tag{6.5}$$

where we use spherical coordinates to have

$$\int_{\|y\|\le\overline{D}} dx = \frac{\pi^{d/2}}{\Gamma(d/2+1)}\overline{D}^d \lesssim \overline{D}^d,$$

and

$$\begin{aligned}\int_{\|y\|\ge\overline{D}} \exp\left(-d_2\|y\|^{d_3}\right) &\lesssim \int_{r\ge\overline{D}} r^{d-1}\exp\left(-d_2 r^{d_3}\right)dr \\ &= \frac{1}{d_3 d_2^{1/d_3}}\int_{\overline{D}^{d_3}}^{\infty} u^{d/d_3-1}\exp(-u)du \quad (\text{with } u = d_2 r^{d_3}) \\ &\le \frac{1}{d_3 d_2^{1/d_3}}\overline{D}^{d-d_3}\exp(-\overline{D}^{d_3}).\end{aligned}$$

Hence, in (6.5), choosing $\overline{D} = D(\log(1/\varepsilon))^{1/d_3}$ gives

$$\int_{\mathbb{R}^d}(p_{ji}^U - p_{ji}^L)dy \lesssim \varepsilon\left(\log\left(\frac{1}{\varepsilon}\right)\right)^{d/d_3}. \tag{6.6}$$

Moreover, denote $p_i^0 = p(y_i|f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}(i/n))$ the density of y_i under the true model. Because $p_{ji}^U \ge p_{ji}^L$, we have

$$\begin{aligned}h^2\left(\frac{p_{ji}^U+p_i^0}{2},\frac{p_{ji}^L+p_i^0}{2}\right) &= \int_{\mathbb{R}^d}\left(\sqrt{\frac{p_{ji}^U+p_i^0}{2}}-\sqrt{\frac{p_{ji}^L+p_i^0}{2}}\right)^2 dy \\ &\le \int_{\mathbb{R}^d}\left(\frac{p_{ji}^U+p_i^0}{2}-\frac{p_{ji}^L+p_i^0}{2}\right)dy \\ &= \frac{1}{2}\int_{\mathbb{R}^d}(p_{ji}^U-p_{ji}^L)dy \\ &\lesssim \varepsilon\left(\log\left(\frac{1}{\varepsilon}\right)\right)^{d/d_3}.\end{aligned}$$

Therefore,

$$\overline{h}_n\left(\frac{p_j^U+p_0^{(n)}}{2},\frac{p_j^L+p_0^{(n)}}{2}\right) \lesssim \varepsilon^{1/2}(\log(1/\varepsilon))^{\frac{d}{2d_3}}.$$

Therefore, there exists a positive constant c which does not depend on ε such that

$$H_B(c\varepsilon^{1/2}\log(1/\varepsilon)^{\frac{d}{2d_3}}) \leq \log N \lesssim k\log(1/\varepsilon).$$

Let $\delta = c\varepsilon^{1/2}(\log(1/\varepsilon))^{\frac{d}{2d_3}}$, we have $\log(1/\delta) \asymp \log(1/\varepsilon)$, which leads to

$$H_B(\delta) \lesssim k\log(1/\delta),$$

for all δ 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

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

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

$$H_B(\delta) \lesssim k\log\left(\frac{n}{\delta}\right).$$

We finished showing that $H_B(\delta) \lesssim \log(n/\delta)$, where the constant in this inequality can depend on Θ, p, and k (but not n and δ).

Step 2. Application of Theorem 5.1. Since $\log(n/u)$ is a non-increasing function of u, we have

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

for all δ small enough, for some constant w_k only depends on Θ and k. Hence, for $\Psi(\delta) = w_k\delta(\log(n/\delta))^{1/2}$, we have $\Psi(\delta)/\delta^2$ is a non-increasing function, and let $\delta_n = \max\{1, 2cw_k\}(\log n/n)^{1/2}$, we have

$$c\Psi(\delta_n) = cw_k\delta_n(\log(n/\delta_n))^{1/2} \leq \delta_n \times (2cw_k(\log(n))^{1/2}) \leq \delta_n^2\sqrt{n}.$$

Substitute $\delta = \delta_n$ to the conclusion of Theorem 5.1, we have

$$\begin{aligned}\mathbb{P}\left(\overline{h}_n\left(p^{(n)}_{\widehat{\boldsymbol{\tau}}^n,\widehat{\boldsymbol{\theta}}^n}, p^{(n)}_{\tau^0,\theta^0}\right) \geq \max\{1, 2cw_k\}(\log n/n)^{1/2}\right) &\leq c\exp\left(-(\max\{1, 2cw_k\})^2\log(n)/c^2\right) \\ &\leq c_1 n^{-c_2},\end{aligned}$$

where w_k depends on Θ and k only, and $c_1 = c$ and $c_2 = 1/c^2$ are universal constants.

Step 3. Derive convergence for parameters in the empirical average distance.

From Lemma 2, we have that

$$\left\|f_{\hat{\tau}^n,\widehat{\theta}^n} - f_{\tau^0,\theta^0}\right\|_n \asymp \overline{h}_n\left(p^{(n)}_{\widehat{\boldsymbol{\tau}}^n,\widehat{\boldsymbol{\theta}}^n}, p^{(n)}_{\tau^0,\theta^0}\right).$$

Therefore,

$$\mathbb{P}\left(\left\|f_{\hat{\tau}^n,\widehat{\theta}^n} - f_{\tau^0,\theta^0}\right\|_n \leq C\left(\frac{\log n}{n}\right)^{1/2}\right) \geq 1 - c_1 n^{-c_2},$$

where C is a constant that only depends on kernel p, parameter space Θ and k, and c_1 and c_2 are universal constants. □

Proof of Lemma 2. This is a direct consequence of condition (K1). □

6.2 Proof of Section 5.3: Convergence rates of parameters arising from DPM of over-fitted signal functions

This section provides proofs for Proportion 1 (see Section 6.2.1), Lemma 3, Lemma 4, Theorem 5.3 (see Section 6.2.4), and Theorem 5.4 (see Section 6.2.5).

In the proofs of Lemma 3 and Lemma 4, we will work with a sequence of signal functions $f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n} \in \mathcal{F}_k(\Theta)$ whose the empirical L_2 risk satisfies the following condition:

$$\left\|f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 = \sum_{i,j=1}^{k_0,k} p_{ij} \left\|\theta_j^n - \theta_i^0\right\|^2 \lesssim \epsilon_n^2, \tag{6.7}$$

where $f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0} \in \mathcal{F}_{k_0}(\Theta)$, $k > k_0$, $\epsilon_n \xrightarrow{n\to\infty} 0$, $p_{ij} = \left|[\tau_{i-1}^0, \tau_i^0) \cap [\tau_{j-1}^n, \tau_j^n)\right|$[1], $i \in [k_0]$, $j \in [k], \tau_0^0 = \tau_0^n = 0$, and $\tau_{k_0}^0 = \tau_k^n = 1$. We have $\sum_{j=1}^k p_{ij} = p_i^0$, $\sum_{i=1}^{k_0} p_{ij} = p_j^n$, where $p_i^0 = \tau_i^0 - \tau_{i-1}^0$ and $p_j^n = \tau_j^n - \tau_{j-1}^n$.

Equation (6.7) implies that

$$p_{ij} \left\|\theta_j^n - \theta_i^0\right\|^2 \lesssim \epsilon_n^2, \quad \text{and} \quad p_{i(j+1)} \left\|\theta_{j+1}^n - \theta_i^0\right\|^2 \lesssim \epsilon_n^2. \tag{6.8}$$

For visualization purposes, we will use the plots illustrated in Figure 6.1 in proofs where the locations of $k_0 - 1$ changepoints associated with $f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}$, and $k - 1$ changepoints associated with $f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n}$ are determined. In particular, when the changepoint positions are specified as cases, these plots are helpful for calculating the empirical L_2 risk, as shown in the proof of Lemma 3 and Lemma 4. Figure 6.1 gives an example of how we read these plots. The red line contains $k_0 - 1$ changepoint locations associated with $f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}$ with indices $i^0 \in [k_0]$. The blue line contains $k - 1$ changepoint locations

[1] the notation $|[a, b)| = b - a$

associated with $f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n}$ with indices $j \in [k]$, and j^* is the index of the changepoint that needs to be pruned. The green line contains the $k-2$ changepoint locations after applying the pruning and merging procedure on $f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n}$ with indices $j^{(k-1)} \in [k-1]$. An interval between two dashed arrows indicates the intersection between a true segment and an overfitted or pruned and merged segment. Between each intersection, there are corresponding true parameters $\boldsymbol{\theta}^0$ and parameters $\boldsymbol{\theta}^n$. For example, if $\tau^n_{j^*-1} < \tau^0_{i^0} < \tau^n_{j^*} < \tau^0_{i^0+1}$ (Figure 6.1), then $p_{i^0+1,j^*} = \left|[\tau^0_{i^0}, \tau^0_{i^0+1}) \cap [\tau^n_{j^*-1}, \tau^n_{j^*})\right| = \left|[\tau^0_{i^0}, \tau^n_{j^*})\right| = \tau^n_{j^*} - \tau^0_{i^0}$. When $t \in [\tau^0_{i^0}, \tau^n_{j^*})$, the associated true parameter is $f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}(t) = \theta^0_{i^0+1}$ and the estimated parameter is $f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n}(t) = \theta^n_{j^*}$. We then have the term $p_{i^0+1,j^*}\left\|\theta^n_{j^*} - \theta^0_{i^0+1}\right\|^2$ is part of the calculation of $\left\|f_{\boldsymbol{\tau}^n\boldsymbol{\theta}^n} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2$.

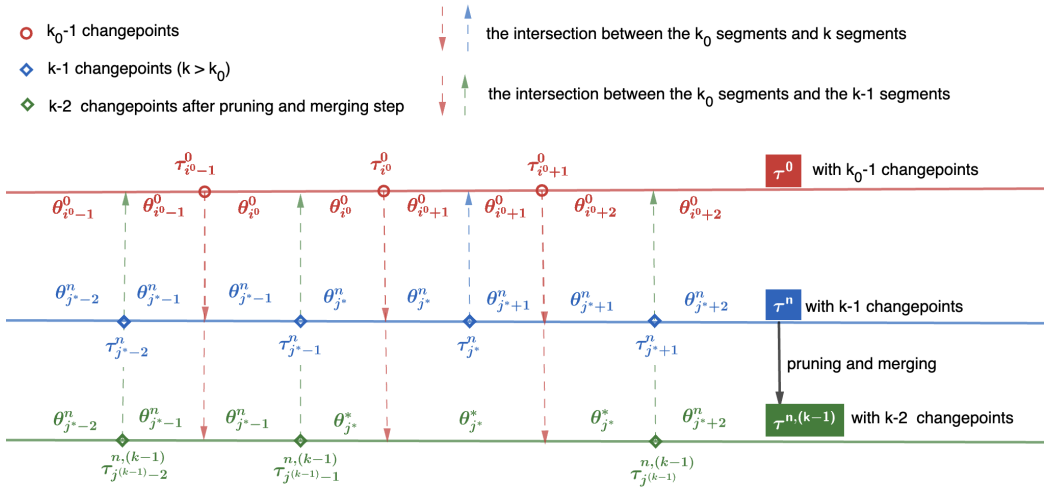


Figure 6.1: Visualization plot for computing the empirical L_2 risk.

6.2.1 Proof of Proposition 1

Proof of Proposition 1. We separate this proof into three steps.

Step 1. Proof by construction. We first start with an arbitrary $f_{\boldsymbol{\tau}',\boldsymbol{\theta}'} \in \mathcal{F}_{\leq(k-1)}(\Theta)$ and iterative show that there exist a better solutions of Equation (5.19) and stop at $f_{\tilde{\boldsymbol{\tau}},\tilde{\boldsymbol{\theta}}}$. Denote $\boldsymbol{\tau}' = (\tau'_1, \ldots, \tau'_{k-2})$, $\tau'_0 = 0, \tau'_{k-1} = 1$, and $\boldsymbol{\theta}' = (\theta'_1, \ldots, \theta'_{k-1})$

Step 2. Optimal $\boldsymbol{\tau}'$. We will separate this step into proving the two following claims. Firstly, we will prove the claim that $\boldsymbol{\tau}'$ must be a subset of $\boldsymbol{\tau}$ to achieve the optimal solution in Equation (5.19). Then, we will show that this optimal $\boldsymbol{\tau}'$ must contain $k-2$ distinct changepoints.

Claim 1: $\boldsymbol{\tau}'$ must be a subset of $\boldsymbol{\tau}$ in order to achieve the optimal solution in Equation (5.19)

Proof of the claim

Indeed, if there exists $\tau_i' \in (\tau_j, \tau_{j+1})$, we consider the positions of τ_{i+1}' and τ_{i-1}'. If τ_{i+1}' also belongs to (τ_j, τ_{j+1}), a better solution can be achieved by letting $\theta_{i+1}' = \theta_{j+1}$, $\tau_i' = \tau_j$, and $\tau_{i+1}' = \tau_{j+1}$. Similarly, a better solution also exists if $\tau_{i-1}' \in (\tau_j, \tau_{j+1})$. Hence, we are left with the case $\tau_{i-1}' \leq \tau_j < \tau_i' < \tau_{j+1} \leq \tau_{i+1}'$. The loss function (5.19), as an integral, when restricting to the interval $[\tau_j, \tau_{j+1}]$, equals

$$\int_{\tau_j}^{\tau_{j+1}} \left\| f_{\boldsymbol{\tau},\boldsymbol{\theta}}(t) - f_{\boldsymbol{\tau}',\boldsymbol{\theta}'}(t) \right\|^2 dt = (\tau_i' - \tau_j) \left\| \theta_i' - \theta_{j+1} \right\|^2 + (\tau_{j+1} - \tau_i') \left\| \theta_{i+1}' - \theta_{j+1} \right\|^2. \quad (6.9)$$

Hence, if $\left\| \theta_i' - \theta_{j+1} \right\|^2 \geq \left\| \theta_{i+1}' - \theta_{j+1} \right\|^2$, a better solution can be achieved by letting $\tau_i' = \tau_j$, and otherwise by letting $\tau_i' = \tau_{j+1}$. It finishes the proof of the first claim.

Claim 2: The optimal $\boldsymbol{\tau}'$ must contain $k-2$ distinct changepoints.

Proof of the claim

We will prove this claim by contradiction. Assuming there is an optimal $\boldsymbol{\tau}'$ with $k-3$ distinct changepoints, we proved that $\boldsymbol{\tau}'$ must be a subset of $\boldsymbol{\tau}$ (as in Claim 1). Without loss of generality (WLOG), we assume that $\tau_1' = \tau_1, \ldots, \tau_{j-1}' = \tau_{j-1}, \tau_j' = \tau_{j+1}, \ldots, \tau_{jj-2}' = \tau_{jj-1}, \tau_{jj-1}' = \tau_{jj+1}, \ldots, \tau_{k-3}' = \tau_{k-1}'$, and $\tau_0' = \tau_0 = 0, \tau_{k-2}' = \tau_k = 1$

(see Figure 6.2). The loss function (5.19) equals:

$$\int_0^1 \left\| f_{\boldsymbol{\tau},\boldsymbol{\theta}}(t) - f_{\boldsymbol{\tau}',\boldsymbol{\theta}'}(t) \right\|^2 dt = (\tau_j - \tau_{j-1}) \left\| \theta'_j - \theta_j \right\|^2 + (\tau_{j+1} - \tau_j) \left\| \theta'_j - \theta_{j+1} \right\|^2 + (\tau_{jj} - \tau_{jj-1}) \left\| \theta'_{jj-1} - \theta_{jj} \right\|^2 + (\tau_{jj+1} - \tau_{jj}) \left\| \theta'_{jj-1} - \theta_{jj+1} \right\|^2 . \tag{6.10}$$

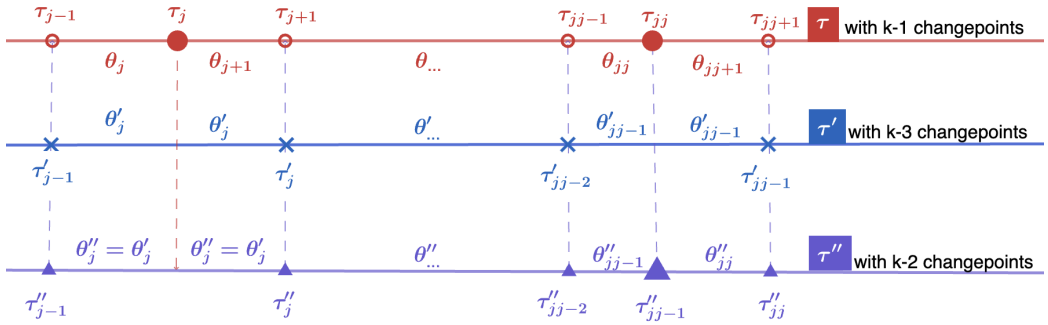


Figure 6.2: Illustration for the proof of Claim 2.

There always exists a τ'' with $k-2$ changepoints that is better than τ'. Indeed, let $\tau''_1 = \tau'_1 = \tau_1, \ldots, \tau''_{j-1} = \tau'_{j-1} = \tau_{j-1}, \tau''_j = \tau'_j = \tau_{j+1}, \ldots, \tau''_{jj-2} = \tau'_{jj-2} = \tau_{jj-1}, \tau''_{jj-1} = \tau_{jj}, \tau''_{jj} = \tau'_{jj-1} = \tau_{jj+1}, \ldots, \tau''_{k-2} = \tau'_{k-3} = \tau_{k-1}$, $\tau''_0 = 0$, $\tau''_{k-1} = 1$, and $\theta''_j = \theta'_j$. The loss function (5.19) are now:

$$\int_0^1 \left\| f_{\boldsymbol{\tau},\boldsymbol{\theta}}(t) - f_{\boldsymbol{\tau}'',\boldsymbol{\theta}''}(t) \right\|^2 dt = (\tau_j - \tau_{j-1}) \left\| \theta'_j - \theta_j \right\|^2 + (\tau_{j+1} - \tau_j) \left\| \theta'_j - \theta_{j+1} \right\|^2 . \tag{6.11}$$

It is clear that $\left\| f_{\boldsymbol{\tau},\boldsymbol{\theta}} - f_{\boldsymbol{\tau}',\boldsymbol{\theta}'} \right\|_{L_2}^2 > \left\| f_{\boldsymbol{\tau},\boldsymbol{\theta}} - f_{\boldsymbol{\tau}'',\boldsymbol{\theta}''} \right\|_{L_2}^2$. Therefore, we cannot say τ' is an optimal solution in this case, which contradicts the assumption we make at the beginning of the proof of this claim.

We can use the same argument to show the contradiction to any case with a number of changepoints less than $k-2$. This finishes the proof of the second claim.

Step 3. Optimal $\boldsymbol{\theta}'$. Hence, $(k-2)$ elements in $\boldsymbol{\tau}'$ all belong to $\boldsymbol{\tau}$. WLOG, assume $\tau'_1 = \tau_1, \ldots, \tau'_{j-1} = \tau_{j-1}, \tau'_j = \tau_{j+1}, \ldots, \tau'_{k-2} = \tau_{k-1}$, and we also denote

$\tau_0 = \tau'_0 = 0, \tau'_{k-1} = \tau_k = 1$. From here, we deduce that the optimal $\boldsymbol{\theta}'$ is

$$\theta'_1 = \theta_1, \ldots, \theta'_{j-1} = \theta_{j-1}, \theta'_{j+1} = \theta_{j+2}, \ldots, \theta'_{k-1} = \theta_k.$$

It leaves with finding optimal θ'_j. The objective function in (5.19) becomes

$$(\tau_j - \tau_{j-1}) \left\|\theta'_j - \theta_j\right\|^2 + (\tau_{j+1} - \tau_j) \left\|\theta'_j - \theta_{j+1}\right\|^2.$$

Hence, the optimal solution is

$$\theta'_j = \frac{\tau_j - \tau_{j-1}}{\tau_{j+1} - \tau_{j-1}}\theta_j + \frac{\tau_{j+1} - \tau_j}{\tau_{j+1} - \tau_{j-1}}\theta_{j+1} = \theta^*_j,$$

and the choice of optimal j is specified as the merging procedure, as j^*.

Prove $d_{j^*} = \left\|f_{\boldsymbol{\tau},\boldsymbol{\theta}} - f_{\tilde{\boldsymbol{\tau}},\tilde{\boldsymbol{\theta}}}\right\|_{L_2}$**.** From the previous steps, the optimal $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\tau}}$ are

$$\tilde{\boldsymbol{\tau}} = (\tau_1, \ldots, \tau_{j^*-1}, \tau_{j^*+1}, \ldots, \tau_{k-1}) \quad \text{and} \quad \tilde{\boldsymbol{\theta}} = (\theta_1, \ldots, \theta_{j^*-1}, \theta^*_{j^*}, \theta_{j^*+2}, \ldots, \theta_k), \tag{6.12}$$

where $\theta^*_{j^*} = \frac{\tau_{j^*} - \tau_{j^*-1}}{\tau_{j^*+1} - \tau_{j^*-1}}\theta_{j^*} + \frac{\tau_{j^*+1} - \tau_{j^*}}{\tau_{j^*+1} - \tau_{j^*-1}}\theta_{j^*+1}$. The L_2 norm of the difference between the true signal functions becomes:

$$\begin{aligned}
&\int_0^1 \left\|f_{\boldsymbol{\tau},\boldsymbol{\theta}}(t) - f_{\tilde{\boldsymbol{\tau}},\tilde{\boldsymbol{\theta}}}(t)\right\|^2 dt = \int_{\tau_{j^*-1}}^{\tau_{j^*+1}} \left\|f_{\boldsymbol{\tau},\boldsymbol{\theta}}(t) - f_{\tilde{\boldsymbol{\tau}},\tilde{\boldsymbol{\theta}}}(t)\right\|^2 dt \\
&= (\tau_{j^*} - \tau_{j^*-1}) \left\|\theta^*_{j^*} - \theta_{j^*}\right\|^2 + (\tau_{j^*+1} - \tau_{j^*}) \left\|\theta^*_{j^*} - \theta_{j^*+1}\right\|^2, \\
&= (\tau_{j^*} - \tau_{j^*-1}) \left\|\frac{\tau_{j^*} - \tau_{j^*+1}}{\tau_{j^*+1} - \tau_{j^*-1}}\theta_{j^*} + \frac{\tau_{j^*+1} - \tau_{j^*}}{\tau_{j^*+1} - \tau_{j^*-1}}\theta_{j^*+1}\right\|^2 \\
&+ (\tau_{j^*+1} - \tau_{j^*}) \left\|\frac{\tau_{j^*} - \tau_{j^*-1}}{\tau_{j^*+1} - \tau_{j^*-1}}\theta_{j^*} + \frac{\tau_{j^*-1} - \tau_{j^*}}{\tau_{j^*+1} - \tau_{j^*-1}}\theta_{j^*+1}\right\|^2 \\
&= \frac{(\tau_{j^*} - \tau_{j^*-1})(\tau_{j^*+1} - \tau_{j^*})^2}{(\tau_{j^*+1} - \tau_{j^*-1})^2} \left\|\theta_{j^*} - \theta_{j^*+1}\right\|^2 + \frac{(\tau_{j^*+1} - \tau_{j^*})(\tau_{j^*} - \tau_{j^*-1})^2}{(\tau_{j^*+1} - \tau_{j^*-1})^2} \left\|\theta_{j^*} - \theta_{j^*+1}\right\|^2 \\
&= \frac{(\tau_{j^*} - \tau_{j^*-1})(\tau_{j^*+1} - \tau_{j^*})}{\tau_{j^*+1} - \tau_{j^*-1}} \left\|\theta_{j^*} - \theta_{j^*+1}\right\|^2 = d^2_{j^*}.
\end{aligned} \tag{6.13}$$

□

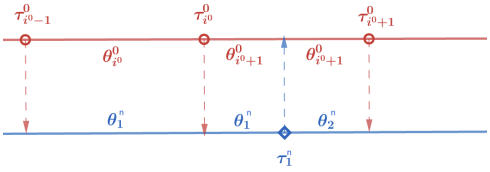
6.2.2 Proof of Lemma 3

Proof of Lemma 3. We will prove this lemma by contradiction. For simplicity, we will use the notation τ_j^n instead of $\tau_{j_n}^n$. Let us assume that there exist $i^0 \in \{1, ..., k_0 - 1\}$ such that there does not exist any sequence of τ_j^n (for $j \in [k-1]$, $\tau_0^n = 0, \tau_k^n = 1$) converging to $\tau_{i^0}^0$ with rate ϵ_n^2, i.e., $|\tau_j^n - \tau_{i^0}^0|/\epsilon_n^2 \xrightarrow{n\to\infty} \infty$ ($i^0 \in [k_0 - 1]$, $\tau_0^0 = 0, \tau_{k_0}^0 = 1$) for all $j \in [k-1]$. We would like to show that this assumption will never happen under the setup of the Lemma 3. For simplicity, we call this assumption (C1).

Indeed, we consider the following three cases.

Case 1: All $\tau_j^n > \tau_{i^0}^0$ for $j \in [k]$. This implies that $\tau_1^n > \tau_{i^0}^0$. Refer to (6.8), we have that $|\tau_{i^0}^0 - \tau_{i^0-1}^0| \|\theta_1^n - \theta_{i^0}^0\|^2 \lesssim \epsilon_n^2$.

On the one hand, since $|\tau_{i^0}^0 - \tau_{i^0-1}^0|$ is a constant, the above equation implies $\|\theta_1^n - \theta_{i^0}^0\|^2 \lesssim \epsilon_n^2$.



It means that $\|\theta_1^n - \theta_{i^0}^0\| \xrightarrow{n\to\infty} 0$. Together with the assumption that $\theta_{i^0}^0 \neq \theta_{i^0+1}^0$, this concludes

$$\|\theta_1^n - \theta_{i^0+1}^0\| \not\to 0, \tag{6.14}$$

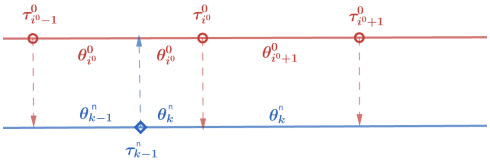
and thereby, $\tau_{i^0}^0 < \tau_1^n < \tau_{i^0+1}^0$.

On the other hand, $|\tau_1^n - \tau_{i^0}^0|/\epsilon_n^2 \xrightarrow{n\to\infty} \infty$ (by (C1)) and $|\tau_1^n - \tau_{i^0}^0| \|\theta_1^n - \theta_{i^0+1}^0\|^2 \lesssim \epsilon_n^2$ (refer to (6.8)) imply $\|\theta_1^n - \theta_{i^0+1}^0\| \xrightarrow{n\to\infty} 0$. This implication contradicts the conclusion in (6.14). Therefore, under this circumstance, it has been proven that (C1) will not occur.

Case 2: All $\tau_j^n < \tau_{i^0}^0$ for $j \in [k]$. This implies that $\tau_{k-1}^n < \tau_{i^0}^0$.

We have $|\tau^0_{i^0+1} - \tau^0_{i^0}| \left\|\theta^0_{i^0+1} - \theta^n_k\right\|^2 \lesssim \epsilon_n^2$ (refer to (6.8)) implying $\left\|\theta^0_{i^0+1} - \theta^n_k\right\|^2 \lesssim \epsilon_n^2$. This means

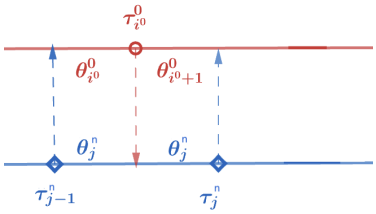
$$\left\|\theta^n_k - \theta^0_{i^0}\right\| \not\to 0, \tag{6.15}$$



therefore, $\tau^0_{i^0-1} < \tau^n_{k-1} < \tau^0_{i^0}$.

However, the assumption (C1) $|\tau^0_{i^0} - \tau^n_{k-1}|/\epsilon_n^2 \xrightarrow{n\to\infty} \infty$ and $|\tau^0_{i^0} - \tau^n_{k-1}| \left\|\theta^n_k - \theta^0_{i^0}\right\|^2 \lesssim \epsilon_n^2$ (refer to (6.8)) imply $\|\theta^n_k - \theta^0_{i^0}\| \xrightarrow{n\to\infty} 0$ (contradict (6.15)). Hence, we also show that (C1) will not occur in this case.

Case 3. There exists $j \in [k]$ such that $\tau^n_{j-1} < \tau^0_{i^0} < \tau^n_j$. On the one hand, we have $|\tau^0_{i^0} - \tau^n_{j-1}|/\epsilon_n^2 \xrightarrow{n\to\infty} \infty$ and $|\tau^0_{i^0} - \tau^n_{j-1}| \left\|\theta^n_j - \theta^0_{i^0}\right\|^2 \lesssim \epsilon_n^2$ (refer to (6.8)) imply $\left\|\theta^n_j - \theta^0_{i^0}\right\| \xrightarrow{n\to\infty} 0$. On the other hand, $|\tau^n_j - \tau^0_{i^0}|/\epsilon_n^2 \xrightarrow{n\to\infty} \infty$ and $|\tau^n_j - \tau^0_{i^0}| \|\theta^n_j - \theta^0_{i^0+1}\|^2 \lesssim \epsilon_n^2$ (refer to (6.8)) imply $\|\theta^n_j - \theta^0_{i^0+1}\| \xrightarrow{n\to\infty} 0$. Since $\theta^0_{i^0} \neq \theta^0_{i^0+1}$, we clearly see a contradiction in this case. Therefore, (C1) is not possible to occur.



We finish proving Lemma 3. □

6.2.3 Proof of Lemma 4

The next lemma (Lemma 4) shows the stability of the convergence rate of signal functions under merging: At all the κ levels where $k \geq \kappa \geq k_0$, the merged function of an estimate must have the same convergence rate as itself.

Proof of Lemma 4. We only need to show that for all $k \geq k_0 + 1$ and the associated signal function $f_{\boldsymbol{\tau}^{n,(k)},\boldsymbol{\theta}^{n,(k)}}$ such that $\left\|f_{\boldsymbol{\tau}^{n,(k)},\boldsymbol{\theta}^{n,(k)}} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n \lesssim \epsilon_n$, then

$$\left\|f_{\boldsymbol{\tau}^{n,(k-1)},\boldsymbol{\theta}^{n,(k-1)}} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 \leq \left\|f_{\boldsymbol{\tau}^{n,(k)},\boldsymbol{\theta}^{n,(k)}} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 + w_k\epsilon_n^2. \tag{6.16}$$

The rest follows from the induction argument. Since the difference between each merging step mainly occurs at the merged changepoint location $\tau^n_{j^*}$, we will focus on the three consecutive changepoints as following $\tau^n_{j^*-1}, \tau^n_{j^*}, \tau^n_{j^*+1}$ and address all of possible cases among these locations while doing the merging procedure.

From now on, in terms of the indices, i^0 is used for the index of the changepoints associated with $f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}$ $(1 \le i^0 \le k_0 - 1)$, j is used for the index of changepoints associated with $f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n}$ $(1 \le j \le k-1)$, j^* is the index associated to the merging position, $j^{(k-1)}$ is for the index of the merged changepoints at the $(k-1)$th-level of the dendrogram $(1 \le j^{(k-1)} \le k-2)$.

Case 1: $|\tau^n_{j^*} - \tau^0_{i^0}| \lesssim \epsilon_n^2$ **(for some** $i^0 \in [k_0 - 1]$**).** There are three sub-cases under this circumstance.

Case 1.1: $|\tau^n_{j^*-1} - \tau^0_{i^0}| \lesssim \epsilon_n^2$ **and** $|\tau^n_{j^*} - \tau^0_{i^0}| \lesssim \epsilon_n^2$**.** WLOG, we assume $\tau^n_{j^*-1} < \tau^0_{i^0} < \tau^n_{j^*} < \tau^n_{j^*+1} \le \tau^0_{i^0+1}$ (see Figure 6.3). In this case, we have

$$\left\|f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 = \{\cdot\} + p_{i^0 j^*}\left\|\theta^n_{j^*} - \theta^0_{i^0}\right\|^2 + p_{(i^0+1)j^*}\left\|\theta^n_{j^*} - \theta^0_{i^0+1}\right\|^2 + p_{(i^0+1)(j^*+1)}\left\|\theta^n_{j^*+1} - \theta^0_{i^0+1}\right\|^2, \tag{6.17}$$

$$\left\|f_{\boldsymbol{\tau}^{n,(k-1)},\boldsymbol{\theta}^{n,(k-1)}} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 = \{\cdot\} + p_{i^0 j^*}\left\|\theta^*_{j^*} - \theta^0_{i^0}\right\|^2 + (p_{(i^0+1)j^*} + p_{(i^0+1)(j^*+1)})\left\|\theta^*_{j^*} - \theta^0_{i^0+1}\right\|^2, \tag{6.18}$$

where $\theta^*_{j^*} = \dfrac{(p_{i^0 j^*} + p_{(i^0+1)j^*})\theta^n_{j^*} + p_{(i^0+1)(j^*+1)}\theta^n_{j^*+1}}{p_{i^0 j^*} + p_{(i^0+1)j^*} + p_{(i^0+1)(j^*+1)}}$ and $\{\cdot\}$ for representing the same terms in both equations (6.17) and (6.18).

Due to the convexity of $\|\cdot\|^2$, we have

$$\begin{aligned}&\left(p_{i^0 j^*} + p_{(i^0+1)j^*} + p_{(i^0+1)(j^*+1)}\right)\left\|\theta^*_{j^*} - \theta^0_{i^0+1}\right\|^2 \\ &\le p_{i^0 j^*}\left\|\theta^n_{j^*} - \theta^0_{i^0+1}\right\|^2 + p_{(i^0+1)j^*}\left\|\theta^n_{j^*} - \theta^0_{i^0+1}\right\|^2 + p_{(i^0+1)(j^*+1)}\left\|\theta^n_{j^*+1} - \theta^0_{i^0+1}\right\|^2\end{aligned} \tag{6.19}$$

In the inequality (6.19), we $\pm p_{i^0 j^*}\left\|\theta^*_{j^*} - \theta^0_{i^0}\right\|^2 + \{\cdot\}$ to the left hand side of it and

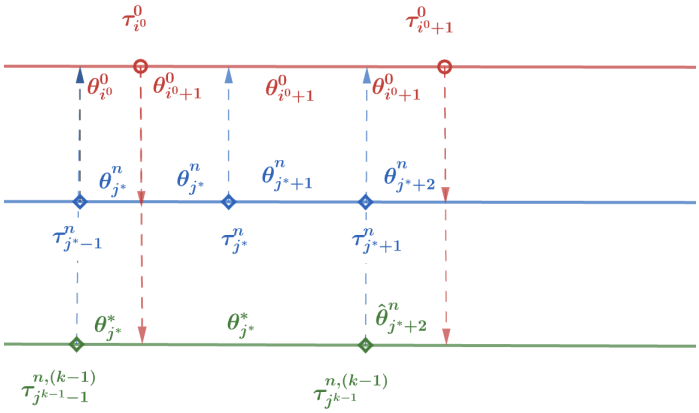


Figure 6.3: Illustration for Case 1.1.

$\pm p_{i^0 j^*} \left\|\theta^n_{j^*} - \theta^0_{i^0}\right\|^2 + \{\cdot\}$ to the right hand side of it, we have

$$\begin{aligned} & p_{i^0 j^*} \left(\left\|\theta^*_{j^*} - \theta^0_{i^0+1}\right\|^2 - \left\|\theta^*_{j^*} - \theta^0_{i^0}\right\|^2 \right) + \left\| f_{\boldsymbol{\tau}^{n,(k-1)}, \boldsymbol{\theta}^{n,(k-1)}} - f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0} \right\|_n^2 \\ & \leq p_{i^0 j^*} \left(\left\|\theta^n_{j^*} - \theta^0_{i^0+1}\right\|^2 - \left\|\theta^n_{j^*} - \theta^0_{i^0}\right\|^2 \right) + \left\| f_{\boldsymbol{\tau}^n, \boldsymbol{\theta}^n} - f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0} \right\|_n^2 \end{aligned} \tag{6.20}$$

Using the assumption that $p_{i^0 j^*} \lesssim \epsilon_n^2$, and the compactness of Θ, we have

$$p_{i^0 j^*} \left(\left\|\theta^n_{j^*} - \theta^0_{i^0+1}\right\|^2 - \left\|\theta^n_{j^*} - \theta^0_{i^0}\right\|^2 - \left\|\theta^*_{j^*} - \theta^0_{i^0+1}\right\|^2 + \left\|\theta^*_{j^*} - \theta^0_{i^0}\right\|^2 \right) \leq w_k \epsilon_n^2$$

for some constant w_k depending on $f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0}, k$ and Θ. This implies

$$\left\| f_{\boldsymbol{\tau}^n, \boldsymbol{\theta}^n} - f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0} \right\|_n^2 \geq \left\| f_{\boldsymbol{\tau}^{n,(k-1)}, \boldsymbol{\theta}^{n,(k-1)}} - f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0} \right\|_n^2 - w_k \epsilon_n^2. \tag{6.21}$$

A similar argument can be used to show the same conclusion for the cases where $\tau^n_{j^*-1} < \tau^n_{j^*} < \tau^0_{i^0} < \tau^n_{j^*+1} \leq \tau^0_{i^0+1}$, $\tau^0_{i^0} < \tau^n_{j^*-1} < \tau^n_{j^*} < \tau^n_{j^*+1} \leq \tau^0_{i^0+1}$, $\tau^n_{j^*-1} < \tau^n_{j^*} < \tau^0_{i^0} < \tau^0_{i^0+1} \leq \tau^n_{j^*+1}$ and $\tau^0_{i^0} < \tau^n_{j^*-1} < \tau^n_{j^*} << \tau^0_{i^0+1} \leq \tau^n_{j^*+1}$.

Case 1.2: $|\tau^n_{j^*} - \tau^0_{i^0+1}| \lesssim \epsilon_n^2$ and $|\tau^n_{j^*+1} - \tau^0_{i^0+1}| \lesssim \epsilon_n^2$. WLOG, we assume that $\tau^0_{i^0} \leq \tau^n_{j^*-1} < \tau^n_{j^*} < \tau^0_{i^0+1} < \tau^n_{j^*+1}$ (see Figure 6.4).

In this case, we have

$$\left\|f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n}-f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2=\{\cdot\}+p_{(i^0+1)j^*}\left\|\theta^n_{j^*}-\theta^0_{i^0+1}\right\|^2+p_{(i^0+1)(j^*+1)}\left\|\theta^n_{j^*+1}-\theta^0_{i^0+1}\right\|^2 + p_{(i^0+2)(j^*+1)}\left\|\theta^n_{j^*+1}-\theta^0_{i^0+2}\right\|^2, \tag{6.22}$$

$$\left\|f_{\boldsymbol{\tau}^{n,(k-1)},\boldsymbol{\theta}^{n,(k-1)}}-f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2=\{\cdot\}+\left(p_{(i^0+1)j^*}+p_{(i^0+1)(j^*+1)}\right)\left\|\theta^*_{j^*}-\theta^0_{i^0+1}\right\|^2 + p_{(i^0+2)(j^*+1)}\left\|\theta^*_{j^*}-\theta^0_{i^0+2}\right\|^2, \tag{6.23}$$

where $\theta^*_{j^*}=\dfrac{p_{(i^0+1)j^*}\theta^n_{j^*}+\left(p_{(i^0+1)(j^*+1)}+p_{(i^0+2)(j^*+1)}\right)\theta^n_{j^*+1}}{p_{(i^0+1)j^*}+p_{(i^0+1)(j^*+1)}+p_{(i^0+2)(j^*+1)}}$ and $\{\cdot\}$ for representing the same terms in both equations (6.22) and (6.23).

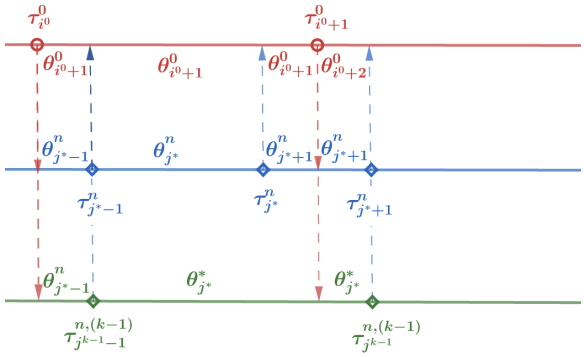


Figure 6.4: Illustration for Case 1.2.

Due to the convexity of $\|\cdot\|^2$, we have

$$\begin{aligned}&\left(p_{(i^0+1)j^*}+p_{(i^0+1)(j^*+1)}+p_{(i^0+2)(j^*+1)}\right)\left\|\theta^*_{j^*}-\theta^0_{i^0+1}\right\|^2\\ &\le p_{(i^0+1)j^*}\left\|\theta^n_{j^*}-\theta^0_{i^0+1}\right\|^2+p_{(i^0+1)(j^*+1)}\left\|\theta^n_{j^*+1}-\theta^0_{i^0+1}\right\|^2+p_{(i^0+2)(j^*+1)}\left\|\theta^n_{j^*+1}-\theta^0_{i^0+1}\right\|^2\end{aligned} \tag{6.24}$$

In the inequality (6.24), we $\pm p_{(i^0+2)(j^*+1)}\left\|\theta^*_{j^*}-\theta^0_{i^0+2}\right\|^2+\{\cdot\}$ to the left hand side

of it and $\pm p_{(i^0+2)(j^*+1)} \left\|\theta^n_{j^*+1} - \theta^0_{i^0+2}\right\|^2 + \{\cdot\}$ to the right hand side of it, we have

$$\begin{aligned} & p_{(i^0+2)(j^*+1)} \left(\left\|\theta^*_{j^*} - \theta^0_{i^0+1}\right\|^2 - \left\|\theta^*_{j^*} - \theta^0_{i^0+2}\right\|^2 \right) + \left\| f_{\boldsymbol{\tau}^{n,(k-1)},\boldsymbol{\theta}^{n,(k-1)}} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0} \right\|_n^2 \\ & \leq p_{(i^0+2)(j^*+1)} \left(\left\|\theta^n_{j^*+1} - \theta^0_{i^0+1}\right\|^2 - \left\|\theta^n_{j^*+1} - \theta^0_{i^0+2}\right\|^2 \right) + \left\| f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0} \right\|_n^2 \end{aligned} \tag{6.25}$$

Using the assumption that $p_{(i^0+2)(j^*+1)} \lesssim \epsilon_n^2$, and the compactness of Θ, we have

$$p_{(i^0+2)(j^*+1)} \left(\left\|\theta^n_{j^*+1} - \theta^0_{i^0+1}\right\|^2 - \left\|\theta^n_{j^*+1} - \theta^0_{i^0+2}\right\|^2 - \left\|\theta^*_{j^*} - \theta^0_{i^0+1}\right\|^2 + \left\|\theta^*_{j^*} - \theta^0_{i^0+2}\right\|^2 \right) \leq w_k \epsilon_n^2$$

for some constant w_k depending on $f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}, k$ and Θ. This implies

$$\left\| f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0} \right\|_n^2 \geq \left\| f_{\boldsymbol{\tau}^{n,(k-1)},\boldsymbol{\theta}^{n,(k-1)}} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0} \right\|_n^2 - w_k \epsilon_n^2. \tag{6.26}$$

A similar argument can be used to show the same conclusion for the cases where $\tau^0_{i^0} \leq \tau^n_{j^*-1} < \tau^n_{j^*} < \tau^n_{j^*+1} < \tau^0_{i^0+1}$, $\tau^0_{i^0} \leq \tau^n_{j^*-1} < \tau^0_{i^0+1} < \tau^n_{j^*} < \tau^n_{j^*+1}$,$\tau^n_{j^*-1} \leq \tau^0_{i^0} < \tau^0_{i^0+1} < \tau^n_{j^*} < \tau^n_{j^*+1}$ and $\tau^n_{j^*-1} \leq \tau^0_{i^0} < \tau^n_{j^*} < \tau^n_{j^*+1} < \tau^0_{i^0+1}$.

Case 1.3: $\tau^n_{j^*}$ is the only changepoint that converges to $\tau^0_{i^0}$ with the rate ϵ_n^2. We want to show that this case is not possible. We will prove this by the contradiction method. Let's assume that there is only one changepoint in $\{\tau^n_j\}_{j=1}^{k-1}$, say $\tau^n_{j^*}$, that converges to $\tau^0_{i^0}$ with the rate ϵ_n^2. WLOG, we assume that $\tau^n_{j^*} < \tau^0_{i^0}$ (see Figure 6.5). We have $|\tau^n_{j^*} - \tau^0_{i^0}| \lesssim \epsilon_n^2$, $|\tau^n_{j^*-1} - \tau^0_{i^0}| \gg \epsilon_n^2$ and $|\tau^n_{j^*+1} - \tau^0_{i^0}| \gg \epsilon_n^2$. These imply $p_{j^*} \gg \epsilon_n^2$ and $p_{j^*+1} \gg \epsilon_n^2$. It leads to $\left\|\theta^n_{j^*+1} - \theta^0_{i^0}\right\| \xrightarrow{n\to\infty} 0$ and $\left\|\theta^n_{j^*+1} - \theta^0_{i^0+1}\right\| \xrightarrow{n\to\infty} 0$. Since $\theta^0_{i^0} \neq \theta^0_{i^0+1}$, we have $\left\|\theta^n_{j^*} - \theta^n_{j^*+1}\right\| \not\to 0$. By the pruning and merging procedure rule and Theorem 5.2, we always have that

$$\left\| f_{\boldsymbol{\tau}^{n,(k-1)},\boldsymbol{\theta}^{n,(k-1)}} - f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n} \right\|_n^2 \leq \left\| f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0} \right\|_n^2 \lesssim \epsilon_n^2, \tag{6.27}$$

where $\left\|f_{\boldsymbol{\tau}^{n,(k-1)},\boldsymbol{\theta}^{n,(k-1)}} - f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n}\right\|_n^2 = \dfrac{p_{j^*}p_{j^*+1}}{p_{j^*}+p_{j^*+1}}\left\|\theta^n_{j^*} - \theta^n_{j^*+1}\right\|^2$ (by Proposition 1). Since $\dfrac{p_{j^*}p_{j^*+1}}{p_{j^*}+p_{j^*+1}} > \dfrac{1}{2}\min\{p_{j^*}, p_{j^*+1}\} \gg \epsilon_n^2$ and $\left\|\theta^n_{j^*} - \theta^n_{j^*+1}\right\| \not\to 0$ (by the assumption), we then have

$$\left\|f_{\boldsymbol{\tau}^{n,(k-1)},\boldsymbol{\theta}^{n,(k-1)}} - f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n}\right\|_n^2 \gg \epsilon_n^2 \text{ (contradict (6.27))}.$$

Therefore, it is impossible that there is only $\tau^n_{j^*}$ that converges to $\tau^0_{i^0}$.

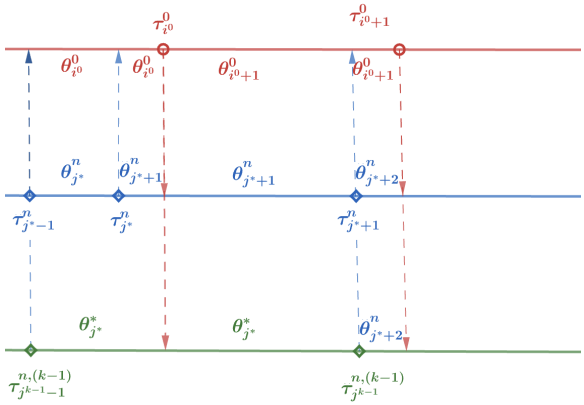


Figure 6.5: Illustration for Case 1.3.

Case 2: $|\tau^n_{j^*} - \tau^0_{i^0}| \gg \epsilon_n^2$**.** There are three main subcases associated with this circumstance.

Case 2.1: $|\tau^n_{j^*-1} - \tau^0_{i^0}| \lesssim \epsilon_n^2$**,** $|\tau^n_{j^*} - \tau^0_{i^0}| \gg \epsilon_n^2$**,** $|\tau^n_{j^*} - \tau^0_{i^0+1}| \gg \epsilon_n^2$**, and** $|\tau^n_{j^*+1} - \tau^0_{i^0+1}| \lesssim \epsilon_n^2$**.**

WLOG, we assume that $\tau^n_{j^*-1} < \tau^0_{i^0} < \tau^n_{j^*} < \tau^0_{i^0+1} < \tau^n_{j^*+1}$ (see Figure 6.6) In this case, we have $p_{i^0j^*} \lesssim \epsilon_n^2$, $p_{(i^0+2)(j^*+1)} \lesssim \epsilon_n^2$, and

$$\begin{aligned}\left\|f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 = \{\cdot\} &+ p_{i^0j^*}\left\|\theta^n_{j^*} - \theta^0_{i^0}\right\|^2 + p_{(i^0+1)j^*}\left\|\theta^n_{j^*} - \theta^0_{i^0+1}\right\|^2 \\ &+ p_{(i^0+1)(j^*+1)}\left\|\theta^n_{j^*+1} - \theta^0_{i^0+1}\right\|^2 \\ &+ p_{(i^0+2)(j^*+1)}\left\|\theta^n_{j^*+1} - \theta^0_{i^0+2}\right\|^2, \qquad (6.28)\\ \left\|f_{\boldsymbol{\tau}^{n,(k-1)},\boldsymbol{\theta}^{n,(k-1)}} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 = \{\cdot\} &+ p_{i^0j^*}\left\|\theta^*_{j^*} - \theta^0_{i^0}\right\|^2 \\ &+ \left(p_{(i^0+1)j^*} + p_{(i^0+1)(j^*+1)}\right)\left\|\theta^*_{j^*} - \theta^0_{i^0+1}\right\|^2 \\ &+ p_{(i^0+2)(j^*+1)}\left\|\theta^*_{j^*} - \theta^0_{i^0+2}\right\|^2, \qquad (6.29)\end{aligned}$$

where $\theta^*_{j^*} = \dfrac{\left(p_{i^0j^*} + p_{(i^0+1)j^*}\right)\theta^n_{j^*} + \left(p_{(i^0+1)(j^*+1)} + p_{(i^0+2)(j^*+1)}\right)\theta^n_{j^*+1}}{p_{i^0j^*} + p_{(i^0+1)j^*} + p_{(i^0+1)(j^*+1)} + p_{(i^0+2)(j^*+1)}}$ and $\{\cdot\}$ for representing the same terms in both equations (6.28) and (6.29).

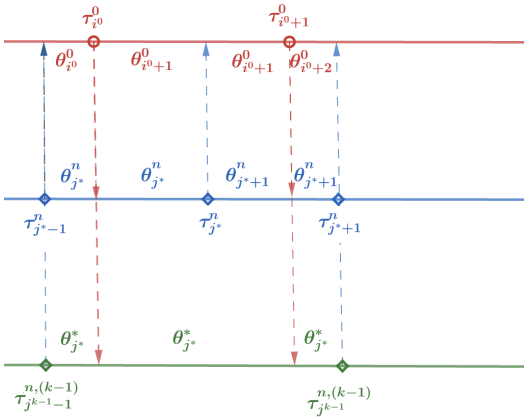


Figure 6.6: Illustration for Case 2.1.

Due to the convexity of $\|\cdot\|^2$, we have

$$\begin{aligned}
&\left(p_{i^0j^*} + p_{(i^0+1)j^*} + p_{(i^0+1)(j^*+1)} + p_{(i^0+2)(j^*+1)}\right) \left\|\theta^*_{j^*} - \theta^0_{i^0+1}\right\|^2 \\
&\leq \left(p_{i^0j^*} + p_{(i^0+1)j^*}\right) \left\|\theta^n_{j^*} - \theta^0_{i^0+1}\right\|^2 + \left(p_{(i^0+1)(j^*+1)} + p_{(i^0+2)(j^*+1)}\right) \left\|\theta^n_{j^*+1} - \theta^0_{i^0+1}\right\|^2
\end{aligned} \tag{6.30}$$

In the inequality (6.30), we $\pm p_{i^0j^*} \left\|\theta^*_{j^*} - \theta^0_{i^0}\right\|^2 \pm p_{(i^0+2)(j^*+1)} \left\|\theta^*_{j^*} - \theta^0_{i^0+2}\right\|^2 + \{\cdot\}$ to the left hand side of it and $\pm p_{i^0j^*} \left\|\theta^n_{j^*} - \theta^0_{i^0}\right\|^2 \pm p_{(i^0+2)(j^*+1)} \left\|\theta^n_{j^*+1} - \theta^0_{i^0+2}\right\|^2 + \{\cdot\}$ to the right hand side of it, we have

$$\begin{aligned}
&\left\|f_{\boldsymbol{\tau}^{n,(k-1)},\boldsymbol{\theta}^{n,(k-1)}} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 + p_{i^0j^*} \left(\left\|\theta^*_{j^*} - \theta^0_{i^0+1}\right\|^2 - \left\|\theta^*_{j^*} - \theta^0_{i^0}\right\|^2\right) \\
&+ p_{(i^0+2)(j^*+1)} \left(\left\|\theta^*_{j^*} - \theta^0_{i^0+1}\right\|^2 - \left\|\theta^*_{j^*} - \theta^0_{i^0+2}\right\|^2\right) \\
&\leq \left\|f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 + p_{i^0j^*} \left(\left\|\theta^n_{j^*} - \theta^0_{i^0+1}\right\|^2 - \left\|\theta^n_{j^*} - \theta^0_{i^0}\right\|^2\right) \\
&+ p_{(i^0+2)(j^*+1)} \left(\left\|\theta^n_{j^*+1} - \theta^0_{i^0+1}\right\|^2 - \left\|\theta^n_{j^*+1} - \theta^0_{i^0+2}\right\|^2\right)
\end{aligned} \tag{6.31}$$

Using the assumption that $p_{i^0j^*} \lesssim \epsilon_n^2$ and $p_{(i^0+2)(j^*+1)} \lesssim \epsilon_n^2$, and the compactness of Θ,

we have

$$\left\|f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 \geq \left\|f_{\boldsymbol{\tau}^{n,(k-1)},\boldsymbol{\theta}^{n,(k-1)}} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 - w_k\epsilon_n^2, \tag{6.32}$$

for some constant w_k depending on $f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}, k$ and Θ.

Case 2.2: $|\tau_{j^*-1}^n - \tau_{i^0}^0| \gg \epsilon_n^2, |\tau_{j^*}^n - \tau_{i^0}^0| \gg \epsilon_n^2$ **and** $|\tau_{j^*+1}^n - \tau_{i^0+1}^0| \lesssim \epsilon_n^2$**.** It implies $\tau_{i^0}^0 < \tau_{j^*-1}^n < \tau_{j^*}^n < \tau_{i^0+1}^0$ (see Figure 6.7). If either $\tau_{j^*-1}^n$ or $\tau_{j^*}^n$ are not in the interval $(\tau_{i^0}^0, \tau_{i^0+1}^0)$, then $\nexists j$ $(1 \leq j \leq k-1)$ such that $\tau_j^n \xrightarrow{n\to\infty} \tau_{i^0}^0$ which violate Lemma 3. WLOG, we consider $\tau_{i^0}^0 < \tau_{j^*-1}^n < \tau_{j^*}^n < \tau_{i^0+1}^0 < \tau_{j^*+1}^n$. In this case, we have $p_{i^0j^*} \lesssim \epsilon_n^2$, $p_{(i^0+2)(j^*+1)} \lesssim \epsilon_n^2$, and

$$\begin{aligned}\left\|f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 = \{\cdot\} &+ p_{(i^0+1)j^*}\left\|\theta_{j^*}^n - \theta_{i^0+1}^0\right\|^2 + p_{(i^0+1)(j^*+1)}\left\|\theta_{j^*+1}^n - \theta_{i^0+1}^0\right\|^2 \\ &+ p_{(i^0+2)(j^*+1)}\left\|\theta_{j^*+1}^n - \theta_{i^0+2}^0\right\|^2, \end{aligned} \tag{6.33}$$

$$\begin{aligned}\left\|f_{\boldsymbol{\tau}^{n,(k-1)},\boldsymbol{\theta}^{n,(k-1)}} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 = \{\cdot\} &+ (p_{(i^0+1)j^*} + p_{(i^0+1)(j^*+1)})\left\|\theta_{j^*}^* - \theta_{i^0+1}^0\right\|^2 \\ &+ p_{(i^0+2)(j^*+1)}\left\|\theta_{j^*}^* - \theta_{i^0+2}^0\right\|^2, \end{aligned} \tag{6.34}$$

where $\theta_{j^*}^* = \dfrac{p_{(i^0+1)j^*}\theta_{j^*}^n + \left(p_{(i^0+1)(j^*+1)} + p_{(i^0+2)(j^*+1)}\right)\theta_{j^*+1}^n}{p_{(i^0+1)j^*} + p_{(i^0+1)(j^*+1)} + p_{(i^0+2)(j^*+1)}}$ and $\{\cdot\}$ for representing the same terms in both equations (6.33) and (6.34).

Due to the convexity of $\|\cdot\|^2$, we have

$$\begin{aligned}&\left(p_{(i^0+1)j^*} + p_{(i^0+1)(j^*+1)} + p_{(i^0+2)(j^*+1)}\right)\left\|\theta_{j^*}^* - \theta_{i^0+1}^0\right\|^2 \\ &\leq p_{(i^0+1)j^*}\left\|\theta_{j^*}^n - \theta_{i^0+1}^0\right\|^2 + \left(p_{(i^0+1)(j^*+1)} + p_{(i^0+2)(j^*+1)}\right)\left\|\theta_{j^*+1}^n - \theta_{i^0+1}^0\right\|^2 \end{aligned} \tag{6.35}$$

In the inequality (6.35), we $\pm p_{(i^0+2)(j^*+1)}\left\|\theta_{j^*}^* - \theta_{i^0+2}^0\right\|^2 + \{\cdot\}$ to the left hand side

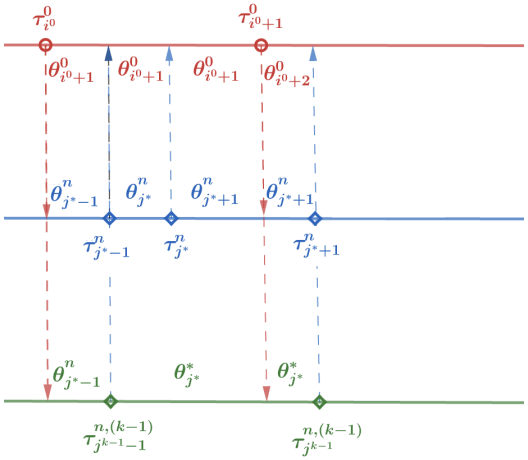


Figure 6.7: Illustration for Case 2.2.

of it and $\pm p_{(i^0+2)(j^*+1)} \left\|\theta^n_{j^*+1} - \theta^0_{i^0+2}\right\|^2 + \{\cdot\}$ to the right hand side of it, we have

$$\begin{aligned}
&\left\|f_{\boldsymbol{\tau}^{n,(k-1)},\boldsymbol{\theta}^{n,(k-1)}} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 + p_{(i^0+2)(j^*+1)} \left(\left\|\theta^*_{j^*} - \theta^0_{i^0+1}\right\|^2 - \left\|\theta^*_{j^*} - \theta^0_{i^0+2}\right\|^2\right) \\
&\leq \left\|f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 + p_{(i^0+2)(j^*+1)} \left(\left\|\theta^n_{j^*+1} - \theta^0_{i^0+1}\right\|^2 - \left\|\theta^n_{j^*+1} - \theta^0_{i^0+2}\right\|^2\right) \qquad (6.36)
\end{aligned}$$

Using the assumption that $p_{(i^0+2)(j^*+1)} \lesssim \epsilon_n^2$, and the compactness of Θ, we have

$$\left\|f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 \geq \left\|f_{\boldsymbol{\tau}^{n,(k-1)},\boldsymbol{\theta}^{n,(k-1)}} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2 - w_k \epsilon_n^2, \qquad (6.37)$$

for some constant w_k depending on $f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}, k$ and Θ.

Case 2.3: $|\tau^n_{j^*-1} - \tau^0_{i^0}| \gg \epsilon_n^2, |\tau^n_{j^*} - \tau^0_{i^0}| \gg \epsilon_n^2, |\tau^n_{j^*+1} - \tau^0_{i^0}| \gg \epsilon_n^2$ **and** $|\tau^n_{j^*-1} - \tau^0_{i^0+1}| \gg \epsilon_n^2, |\tau^n_{j^*} - \tau^0_{i^0+1}| \gg \epsilon_n^2, |\tau^n_{j^*+1} - \tau^0_{i^0+1}| \gg \epsilon_n^2$**.** It implies $\tau^0_{i^0} < \tau^n_{j^*-1} < \tau^n_{j^*} < \tau^n_{j^*+1} < \tau^0_{i^0+1}$ (see Figure 6.8). Notice that any other cases will lead to a contradiction to Lemma 3.

In this case, we have

$$\left\|f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n}-f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2=\{\cdot\}+p_{(i^0+1)j^*}\left\|\theta_{j^*}^n-\theta_{i^0+1}^0\right\|^2+p_{(i^0+1)(j^*+1)}\left\|\theta_{j^*+1}^n-\theta_{i^0+1}^0\right\|^2 \tag{6.38}$$

$$\left\|f_{\boldsymbol{\tau}^{n,(k-1)},\boldsymbol{\theta}^{n,(k-1)}}-f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2=\{\cdot\}+\left(p_{(i^0+1)j^*}+p_{(i^0+1)(j^*+1)}\right)\left\|\theta_{j^*}^*-\theta_{i^0+1}^0\right\|^2, \tag{6.39}$$

where $\theta_{j^*}^*=\dfrac{p_{(i^0+1)j^*}\theta_{j^*}^n+p_{(i^0+1)(j^*+1)}\theta_{j^*+1}^n}{p_{(i^0+1)j^*}+p_{(i^0+1)(j^*+1)}}$ and $\{\cdot\}$ for representing the same terms in both equations (6.38) and (6.39).

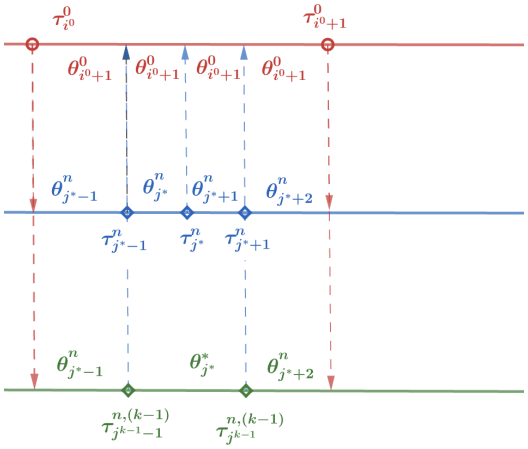


Figure 6.8: Illustration for Case 2.3.

Due to the convexity of $\|\cdot\|^2$, we have

$$\left(p_{i^0j^*}+p_{i^0(j^*+1)}\right)\left\|\theta_{j^*}^*-\theta_{i^0}^0\right\|^2\leq p_{i^0j^*}\left\|\theta_{j^*}^n-\theta_{i^0}^0\right\|^2+p_{i^0(j^*+1)}\left\|\theta_{j^*+1}^n-\theta_{i^0}^0\right\|^2 \tag{6.40}$$

In the inequality (6.40), we add $\{\cdot\}$ to the left, and the right-hand side of it, we have

$$\left\|f_{\boldsymbol{\tau}^{n,(k-1)},\boldsymbol{\theta}^{n,(k-1)}}-f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2\leq\left\|f_{\boldsymbol{\tau}^n,\boldsymbol{\theta}^n}-f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}\right\|_n^2. \tag{6.41}$$

Finally, from (6.21), (6.26), (6.32), (6.37), (6.41), we finish proving (6.16). □

6.2.4 Proof of Theorem 5.3: Asymptotic behavior of the signal functions in the dendrogram

Proof of Theorem 5.3. We divide the proof into two parts: overfitted levels and under-fitted levels.

Part 1: Convergence rate on overfitted levels. From Theorem 5.2, there exists a constant C depending on Θ and k so that under the assumption (K1) with probability of at least $1 - c_1 n^{-c_2}$, we have

$$\left\| \hat{f}_{\widehat{\boldsymbol{\tau}}^n, \widehat{\boldsymbol{\theta}}^n} - f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0} \right\|_n \leq C \left(\frac{\log n}{n} \right)^{1/2}. \tag{6.42}$$

Using Lemma 4, there exists constants $w_k, \ldots, w_{k_0+1}$ depending on $f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0}, k$ and Θ such that for every $\kappa \in [k_0 + 1, k]$, we have

$$\left\| f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa)}, \widehat{\boldsymbol{\theta}}^{n,(\kappa)}} - f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0} \right\|_n^2 \leq \left\| f_{\widehat{\boldsymbol{\tau}}^n, \widehat{\boldsymbol{\theta}}^n} - f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0} \right\|_n^2 + \frac{\log n}{n} \sum_{m=\kappa}^{k} w_m. \tag{6.43}$$

From (6.43) and (6.42), we get the convergence rate for all level $\kappa \geq k_0 + 1$:

$$\left\| f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa)}, \widehat{\boldsymbol{\theta}}^{n,(\kappa)}} - f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0} \right\|_n \lesssim \left(\frac{\log n}{n} \right)^{1/2}. \tag{6.44}$$

When $\kappa = k_0$, we want to show that $\left\| f_{\widehat{\boldsymbol{\tau}}^{n,(k_0)}, \widehat{\boldsymbol{\theta}}^{n,(k_0)}} - f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0} \right\|_n \lesssim \left(\frac{\log n}{n} \right)^{1/2}$. This is still achieved by using Lemma 4, since we have

$$\left\| f_{\widehat{\boldsymbol{\tau}}^{n,(k_0+1)}, \widehat{\boldsymbol{\theta}}^{n,(k_0+1)}} - f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0} \right\|_n^2 \geq \left\| f_{\widehat{\boldsymbol{\tau}}^{n,(k_0)}, \widehat{\boldsymbol{\theta}}^{n,(k_0)}} - f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0} \right\|_n^2 + w_{k_0+1} \frac{\log n}{n},$$

and $\left\| f_{\widehat{\boldsymbol{\tau}}^{n,(k_0+1)}, \widehat{\boldsymbol{\theta}}^{n,(k_0+1)}} - f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0} \right\|_n^2 \lesssim \frac{\log n}{n}$ has been proven in (6.44).

Part 2: Convergence rate on under-fitted levels. At the exact fitted level, by Lemma 3, $k_0 - 1$ changepoints indexed by j $(1 \leq j \leq k_0 - 1)$ will converge to $k_0 - 1$

actual changepoints indexed by i^0 ($1 \le i^0 \le k_0 - 1$) with rate $\dfrac{\log n}{n}$. Also by the result in Part 1, we have $\left\| f_{\widehat{\boldsymbol{\tau}}^{n,(k_0)},\widehat{\boldsymbol{\theta}}^{n,(k_0)}} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0} \right\|_n^2 \lesssim \dfrac{\log n}{n}$, it leads to

$$|\hat{\tau}_i^{n,(k_0)} - \tau_i^0| \lesssim \frac{\log n}{n}, \quad \left\|\hat{\theta}_i^{n,(k_0)} - \theta_i^0\right\|^2 \lesssim \frac{\log n}{n} \quad \forall i \in [k_0], \quad \hat{\tau}_{k_0}^n = \tau_{k_0}^0 = 1. \tag{6.45}$$

It is straightforward to show that for every $i \in [k_0]$ ($\tau_0^0 = \hat{\tau}_0^{n,(k_0)} = 0$, $\tau_{k_0}^0 = \hat{\tau}_{k_0}^{n,(k_0)} = 1$), we have

$$\left| \frac{(\hat{\tau}_i^{n,(k_0)} - \hat{\tau}_{i-1}^{n,(k_0)})(\hat{\tau}_{i+1}^{n,(k_0)} - \hat{\tau}_i^{n,(k_0)})}{\hat{\tau}_{i+1}^{n,(k_0)} - \hat{\tau}_{i-1}^{n,(k_0)}} \left\|\hat{\theta}_i^{n,(k_0)} - \hat{\theta}_{i+1}^{n,(k_0)}\right\|^2 - \frac{(\tau_i^0 - \tau_{i-1}^0)(\tau_{i+1}^0 - \tau_i^0)}{\tau_{i+1}^0 - \tau_{i-1}^0} \left\|\theta_i^0 - \theta_{i+1}^0\right\|^2 \right| \lesssim \frac{\log n}{n} \tag{6.46}$$

Hence, the optimal choice of i to merge for $f_{\hat{\boldsymbol{\tau}}^{n,(k_0)},\hat{\boldsymbol{\theta}}^{n,(k_0)}}$ will be the same as $f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}$ for every n large enough. After merging, we also have

$$\left|(\hat{\tau}_{i+1}^{n,(k_0)} - \hat{\tau}_{i-1}^{n,(k_0)}) - (\tau_{i+1}^0 - \tau_{i-1}^0)\right| \lesssim \frac{\log n}{n},$$

and

$$\left\| \left(\frac{\hat{\tau}_i^{n,(k_0)} - \hat{\tau}_{i-1}^{n,(k_0)}}{\hat{\tau}_{i+1}^{n,(k_0)} - \hat{\tau}_{i-1}^{n,(k_0)}} \hat{\theta}_i^{n,(k_0)} + \frac{\hat{\tau}_{i+1}^{n,(k_0)} - \hat{\tau}_i^{n,(k_0)}}{\hat{\tau}_{i+1}^{n,(k_0)} - \hat{\tau}_{i-1}^{n,(k_0)}} \hat{\theta}_{i+1}^{n,(k_0)} \right) - \left(\frac{\tau_i^0 - \tau_{i-1}^0}{\tau_{i+1}^0 - \tau_{i-1}^0} \theta_i^0 + \frac{\tau_{i+1}^0 - \tau_i^0}{\tau_{i+1}^0 - \tau_{i-1}^0} \theta_{i+1}^0 \right) \right\|^2 \lesssim \frac{\log n}{n}.$$

Hence, $\left\| f_{\widehat{\boldsymbol{\tau}}^{n,(k_0-1)},\widehat{\boldsymbol{\theta}}^{n,(k_0-1)}} - f_{\boldsymbol{\tau}^{0,(k_0-1)},\boldsymbol{\theta}^{0,(k_0-1)}} \right\|_n^2 \lesssim \dfrac{\log n}{n}$. The rest of the proof follows using induction. □

6.2.5 Proof of Theorem 5.4: Asymptotic behavior of the heights

Proof of Theorem 5.4. Continue from the proof of Theorem 5.3 , for every $\kappa \geq k_0 + 1$, we have that

$$\left\| f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa)},\widehat{\boldsymbol{\theta}}^{n,(\kappa)}} - f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa-1)},\widehat{\boldsymbol{\theta}}^{\kappa-1}} \right\|_n^2 \leq 2 \left(\left\| f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa)},\widehat{\boldsymbol{\theta}}^{n,(\kappa)}} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0} \right\|_n^2 + \left\| f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa-1)},\widehat{\boldsymbol{\theta}}^{n,(\kappa-1)}} - f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0} \right\|_n^2 \right) \lesssim \frac{\log n}{n}, \tag{6.47}$$

where $\left\| f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa)},\widehat{\boldsymbol{\theta}}^{n,(\kappa)}} - f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa-1)},\widehat{\boldsymbol{\theta}}^{n,(\kappa-1)}} \right\|_n = d_n^{(\kappa)}$ (by Proposition 1). Therefore, $d_n^{(\kappa)} \lesssim \left(\frac{\log n}{n} \right)^{1/2}$.

When $\kappa \leq k_0$, the conclusion follows from inequality (6.46) in the proof of Theorem 5.3.

□

6.3 Proof of Section 5.4: Consistency of DSC

This section contains the primary work of Dat Do on proving the consistency in estimating the number of changepoints using DSC. The proving techniques are developed based on the work on the Dendrogram of mixing measures paper [31].

We first prove Lemma 5, which provides upper bounds for the log-likelihood at the over-fitted and under-fitted levels and a lower bound for the exact-fitted level. These results will be useful for building model selection criteria.

Proof of Lemma 5. 1. *(Over-fitted levels)* If the conclusion (5.30) is correct for $\kappa = k$, then because $(\boldsymbol{\tau}^n, \boldsymbol{\theta}^n)$ is the MLE, we have

$$\overline{\ell}_n(f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa)},\widehat{\boldsymbol{\theta}}^{n,(\kappa)}}) \leq \overline{\ell}_n(f_{\widehat{\boldsymbol{\tau}}^n,\widehat{\boldsymbol{\theta}}^n}) \quad \forall \kappa \leq k,$$

which leads to the same conclusion for all $\kappa \leq k$. Hence, it is sufficient to prove (5.30) for $\kappa = k$ only, and we proceed to do it. Recall that we have proved

$$\overline{h}_n(p_{\widehat{\boldsymbol{\tau}}^n,\widehat{\boldsymbol{\theta}}^n}, p_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}) \leq C\left(\frac{\log n}{n}\right)^{1/2},$$

with the probability at least $1 - c_1 n^{-c_2}$, where C only depends on p, f and Θ. By invoking the proof of Theorem 8.14 in [129] (the first step)[2], we can have a bound for the empirical process:

$$\sqrt{n}(P_n - P_0)\log\frac{\widehat{p}_{n,i}(y_i) + p_{0,i}(y_i)}{2p_{0,i}(y_i)} \leq C\left(\frac{\log n}{n}\right), \tag{6.48}$$

where $\widehat{p}_{n,i}(y_i) = p(y_i|\widehat{f}^n(t_i))$ is the MLE density, $p_{0,i}(y_i) = p(y_i|f^0(t_i))$ is the true density at t_i, and the expectation $(P_n - P_0)$ corresponds to

$$P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}, \quad P_0 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{y_i\sim p(f_{\tau^0,\theta^0}(t_i))}.$$

Using the concavity of the logarithm function, we have

$$P_n \log\frac{\widehat{p}_{n,i}(y_i) + p_{0,i}(y_i)}{2p_{0,i}(y_i)} \geq \frac{1}{2}P_n\log\frac{\widehat{p}_{n,i}(y_i)}{p_{0,i}(y_i)}. \tag{6.49}$$

Furthermore,

$$P_0\log\frac{\widehat{p}_{n,i}(y_i) + p_{0,i}(y_i)}{2p_{0,i}(y_i)} = -\frac{1}{n}\sum_{i=1}^{n}KL\left(p(f_{\tau^0,\theta^0}(t_i))\|\frac{p(\widehat{f}^n(t_i)) + p(f_{\tau^0,\theta^0}(t_i))}{2}\right) \leq 0. \tag{6.50}$$

---

[2]similar to the argument in the proof of convergence rate of likelihood function in Theorem 3.4, Chapter 3.

Combine these two inequalities with (6.48), we have

$$\overline{\ell}_n(\widehat{f}^n) - P_n \log p(y_i|f_{\tau^0,\theta^0}(t_i)) = P_n \log \frac{\widehat{p}_{n,i}(y_i)}{p_{0,i}(y_i)} \leq C\frac{\log n}{n}, \tag{6.51}$$

with probability at least $1 - 2c_1 n^{-c_2}$. Besides, using Chebyshev's inequality, we have

$$\begin{aligned}
&\mathbb{P}_0(|(P_n - P_0)\log p(y_i|f_{\tau^0,\theta^0}(t_i))| \geq r) \\
&\leq \sum_{j=1}^{k_0} \mathbb{P}_0\left(\frac{1}{n}\left|\sum_{t_i\in[\tau_{j-1}^0,\tau_j^0)} \log p(y_i|f) - \mathbb{E}\log p(y_i|f_{\tau^0,\theta^0}(t_i))\right| \geq \frac{r}{k_0}\right) \\
&\leq k_0^2 \sum_{j=1}^{k_0} \frac{(\tau_j^0 - \tau_{j-1}^0)^2 \mathrm{Var}(\log p(y|(\theta_j^0))}{nr^2},
\end{aligned}$$

for all $r > 0$. By plugging in $r = (\log n/n)^{1/2}$, we have

$$\mathbb{P}_0\left(|(P_n - P_0)\log p(y_i|f_{\tau^0,\theta^0}(t_i))| \geq \left(\frac{\log n}{n}\right)^{1/2}\right) \leq \frac{C'}{\log(n)},$$

where C' only depends on f^0. Aggregate this with (6.51), we have with a probability at least of $1 - \max\{C'/\log(n), c_1 n^{-c_2}\}$ that

$$\overline{\ell}_n(\widehat{f}^n) - P_0 \log p(y_i|f_{\tau^0,\theta^0}(t_i)) = \overline{\ell}_n(\widehat{f}^n) - \overline{\mathcal{L}}_0(f_{\tau^0,\theta^0}) \leq C_o\left(\frac{\log n}{n}\right)^{1/2}.$$

for $C_o = \max\{0, C + C'\}$ only depends on $p, k, \boldsymbol{\tau}^0, \boldsymbol{\theta}^0$.

2. (Exact-fitted level) From Lemma 3 and Theorem 5.3, there exists an event A_n with $\mathbb{P}_0(A_n) \geq 1 - c_1 n^{-c_2}$, where $\mathbb{P}_0$ is the law of $(y_1, \ldots, y_n)$ under the true model, such that under A_n, the exact-fitted signal function on the dendrogram

possesses the changepoints $\widehat{\boldsymbol{\tau}}^{n,(k_0)} \in \mathcal{T}_{\uparrow}^{k_0}$ and parameters $\widehat{\boldsymbol{\theta}}^{n,(k_0)} \in \Theta^{k_0}$ satisfying

$$\|\widehat{\theta}_j^{n,(k_0)} - \theta_j^0\| \leq \delta_n = C\left(\frac{\log n}{n}\right)^{1/2}, \quad |\widehat{\tau}_j^{n,(k_0)} - \tau_j^0| \leq \delta_n^2, \quad \forall j \in [k_0], \quad (6.52)$$

where C depends on $(\boldsymbol{\tau}^0, \boldsymbol{\theta}^0), \Theta, f$ and k. We write $\widehat{\tau}_j^{n,(k_0)} = \tilde{\tau}_j^n$ and $\widehat{\theta}_j^{n,(k_0)} = \tilde{\theta}_j^n$ for ease of notation. Recall that we aim to prove that

$$\overline{\ell}_n(f_{\tilde{\boldsymbol{\tau}}^n, \tilde{\boldsymbol{\theta}}^n}) = \frac{1}{n}\sum_{j=1}^{k_0} \sum_{\substack{i \in [n] \\ t_i \in [\tilde{\tau}_{j-1}^n, \tilde{\tau}_j^n)}} \log p(y_i | \tilde{\theta}_j^n),$$

satisfies

$$\left|\overline{\ell}_n(f_{\tilde{\boldsymbol{\tau}}^n, \tilde{\boldsymbol{\theta}}^n}) - \overline{\mathcal{L}}_0(f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0})\right| \leq C_e(\log n/n)^{1/2},$$

with high probability (tends to 1), where C_e is a positive constant and

$$\overline{\mathcal{L}}_0(f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0}) = \frac{1}{n}\sum_{j=1}^{k_0} \sum_{\substack{i \in [n] \\ t_i \in [\tau_{j-1}^0, \tau_j^0)}} \mathbb{E}_{y_i \sim p(\tilde{\theta}_j^0)} \log p(y_i | \tilde{\theta}_j^0).$$

We will show that

$$\overline{\ell}_n(f_{\tilde{\boldsymbol{\tau}}^n, \tilde{\boldsymbol{\theta}}^n}) - \overline{\mathcal{L}}_0(f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0}) \geq -C_e(\log n/n)^{1/2}, \quad (6.53)$$

with high probability, and the part to show

$$\overline{\ell}_n(f_{\tilde{\boldsymbol{\tau}}^n, \tilde{\boldsymbol{\theta}}^n}) - \overline{\mathcal{L}}_0(f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0}) \leq C_e(\log n/n)^{1/2},$$

is similar. Indeed, it can be seen that those two terms in the LHS of (6.53) are different in both the changepoints ($\tilde{\boldsymbol{\tau}}^n$ and $\boldsymbol{\tau}^0$) and parameters ($\tilde{\boldsymbol{\theta}}^n$ and $\boldsymbol{\theta}^0$). We first use condition (K3) to approximate all $\tilde{\theta}_j^n$ by θ_j^0. Indeed, for all

$(y_1, \ldots, y_n) \in A_n$, use (K3) and combine with inequalities (6.52), we have

$$
\begin{aligned}
\bar{\ell}_n(f_{\tilde{\boldsymbol{\tau}}^n, \tilde{\boldsymbol{\theta}}^n}) &= \frac{1}{n} \sum_{j=1}^{k_0} \sum_{\substack{i \in [n] \\ t_i \in [\tilde{\tau}_{j-1}^n, \tilde{\tau}_j^n)}} \log p(y_i | \tilde{\theta}_j^n) \\
&\geq \frac{1}{n} \sum_{j=1}^{k_0} \left(\sum_{\substack{i \in [n] \\ t_i \in [\tilde{\tau}_{j-1}^n, \tilde{\tau}_j^n)}} \log p(y_i | \theta_j^0) \right) (1 + c_\beta \delta_n) - c_\alpha \delta_n \\
&\geq \frac{1}{n} (1 + c_\beta \delta_n) \sum_{j=1}^{k_0} \left(\sum_{\substack{i \in [n] \\ t_i \in [\tau_{j-1}^0, \tau_j^0)}} \log p(y_i | \theta_j^0) - \sum_{t_i \in [\tilde{\tau}_{j-1}^n, \tilde{\tau}_j^n) \setminus [\tau_{j-1}^0, \tau_j^0)} \left| \log p(y_i | \theta_j^0) \right| \right. \\
&\qquad \left. - \sum_{t_i \in [\tau_{j-1}^0, \tau_j^0) \setminus [\tilde{\tau}_{j-1}^n, \tilde{\tau}_j^n)} \left| \log p(y_i | \theta_j^0) \right| \right) - c_\alpha \delta_n \\
&\geq \frac{1}{n} (1 + c_\beta \delta_n) \sum_{j=1}^{k_0} \left(\sum_{\substack{i \in [n] \\ t_i \in [\tau_{j-1}^0, \tau_j^0)}} \log p(y_i | \theta_j^0) - \sum_{t_i \in [\tau_j^0 - \delta_n^2, \tau_j^0) \cup [\tau_j^0, \tau_j^0 + \delta_n^2)} \left| \log p(y_i | \theta_j^0) \right| \right. \\
&\qquad \left. - \sum_{t_i \in [\tau_{j-1}^0, \tau_{j-1}^0 + \delta_n^2) \cup [\tau_j^0 - \delta_n^2, \tau_j^0)} \left| \log p(y_i | \theta_j^0) \right| \right) - c_\alpha \delta_n \\
&= (1 + c_\beta \delta_n) \sum_{j=1}^{k_0} (D_{nj} - E_{nj} - F_{nj}) - c_\alpha \delta_n, \qquad (6.54)
\end{aligned}
$$

where

$$
D_{nj} := \frac{1}{n} \sum_{\substack{i \in [n] \\ t_i \in [\tau_{j-1}^0, \tau_j^0)}} \log p(y_i | \theta_j^0), \quad E_{nj} = \frac{1}{n} \sum_{t_i \in [\tau_j^0 - \delta_n^2, \tau_j^0) \cup [\tau_j^0, \tau_j^0 + \delta_n^2)} \left| \log p(y_i | \theta_j^0) \right|,
$$

and

$$
F_{nj} = \frac{1}{n} \sum_{t_i \in [\tau_{j-1}^0, \tau_{j-1}^0 + \delta_n^2) \cup [\tau_j^0 - \delta_n^2, \tau_j^0)} \left| \log p(y_i | \theta_j^0) \right|.
$$

Notice that there is no $\tilde{\boldsymbol{\tau}}^n$ and $\tilde{\boldsymbol{\theta}}^n$ in the definitions of D_{nj}, E_{nj}, F_{nj}'s anymore. We can simply bound D_{nj} using Chebyshev's inequality. Denote $T_i = \#\{i : t_i \in$

$[\tau_{j-1}^0, \tau_j^0)\} = \lceil n\tau_j^0 \rceil - \lceil n\tau_{j-1}^0 \rceil \asymp n(\tau_j^0 - \tau_{j-1}^0)$, we have

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n} \left| \sum_{\substack{i\in[n] \\ t_i \in [\tau_{j-1}^0, \tau_j^0)}} (\log p(y_i|\theta_j^0) - \mathbb{E}_{p(\theta_j^0)} \log p(y_i|\theta_j^0) \right| \geq \frac{r}{k_0} \right) \\
&= \mathbb{P}\left(\frac{1}{T_i} \left| \sum_{\substack{i\in[n] \\ t_i \in [\tau_{j-1}^0, \tau_j^0)}} (\log p(y_i|\theta_j^0) - \mathbb{E}_{p(\theta_j^0)} \log p(y_i|\theta_j^0) \right| \geq \frac{nr}{T_i k_0} \right) \\
&\leq \frac{k_0^2 T_j \mathrm{Var}_{p(\theta_j^0)}(\log p(y|\theta_j^0))}{n^2 r^2}
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{P}\left(|\sum_j D_{nj} - \overline{\mathcal{L}}_0(f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0})| \geq r \right) &\leq \sum_{j=1}^{k_0} \mathbb{P}\left(\frac{1}{n} \left| \sum_{\substack{i\in[n] \\ t_i \in [\tau_{j-1}^0, \tau_j^0)}} (\log p(y_i|\theta_j^0) - \mathbb{E}_{p(\theta_j^0)} \log p(y_i|\theta_j^0) \right| \geq \frac{r}{k_0} \right) \\
&\leq \sum_{j=1}^{k_0} \frac{k_0^2 T_j \mathrm{Var}_{p(\theta_j^0)}(\log p(y|\theta_j^0))}{n^2 r^2} \\
&\leq \frac{k_0^2 \max_j \mathrm{Var}_{p(\theta_j^0))}(\log p(y|\theta_j^0)}{nr^2}.
\end{aligned}$$

For E_n, we can use the second condition in (K3) to have

$$\begin{aligned}
\mathbb{P}(E_{nj} \geq (\log n/n)^{1/2}) &\leq \sum_{2\log(n)\text{terms}} \mathbb{P}(|\log p(y_i|\theta_j^0)| \geq n(\log n/n)^{1/2})/\log(n)) \\
&\leq \log(n)\gamma_1 e^{-(\log n/n)^{\gamma_2/2}}.
\end{aligned}$$

Similar for F_{nj}. Hence, there exists a sequence of event B_n with $\mathbb{P}(B_n) \geq 1 - C/\log(n)$ such that

$$(1 + c_\beta \delta_n) \sum_{j=1}^{k_0} (D_{nj} - E_{nj} - F_{nj}) - c_\alpha \delta_n \geq \overline{\mathcal{L}}_0(f_{\boldsymbol{\tau}^0, \boldsymbol{\theta}^0}) - C_e (\log n/n)^{1/2}.$$

Combine with A_n, we have

$$\mathbb{P}(\overline{\ell}_n(f_{\tilde{\boldsymbol{\tau}}^n,\tilde{\boldsymbol{\theta}}^n}) \geq \overline{\mathcal{L}}_0(f_{\tau^0,\theta^0}) - C_e(\log n/n)^{1/2}) \geq \mathbb{P}(A_n \cap B_n) \geq 1 - C/\log(n).$$

By using the inequality $\log p(y_i|\tilde{\theta}_j^n) \leq (1 - c_\beta\delta_n)\log p(y_i|\tilde{\theta}_j^n) + c_\alpha\delta_n$ (instead of the upper bound) in inequality (6.54) and proceed similarly, we also have the bound in the reverse direction

$$\mathbb{P}(\overline{\ell}_n(f_{\tilde{\boldsymbol{\tau}}^n,\tilde{\boldsymbol{\theta}}^n}) \geq \overline{\mathcal{L}}_0(f_{\tau^0,\theta^0}) + C_e(\log n/n)^{1/2}) \geq 1 - C/\log(n).$$

Hence,

$$\mathbb{P}(|\overline{\ell}_n(f_{\tilde{\boldsymbol{\tau}}^n,\tilde{\boldsymbol{\theta}}^n}) - \overline{\mathcal{L}}_0(f_{\tau^0,\theta^0})| \leq C_e(\log n/n)^{1/2}) \geq 1 - 2C/\log(n) \to 1$$

as $n \to \infty$.

3. (Under-fitted level) We proceed similarly to the exact-fitted level, given that we have already shown that

$$\left\| f_{\widehat{\boldsymbol{\tau}}^{n,(\kappa)},\widehat{\boldsymbol{\theta}}^{n,(\kappa)}} - f_{\boldsymbol{\tau}^{0,(\kappa)},\boldsymbol{\theta}^{0,(\kappa)}} \right\|_n \leq C\left(\frac{\log n}{n}\right)^{1/2}, \quad \forall 1 \leq \kappa \leq k_0, \tag{6.55}$$

as in Theorem 5.3.

□

Proof of Theorem 5.5. By combining Theorem 5.4 and Lemma 5, there exists a set A_n with $\mathbb{P}_0(A_n) \to 1$ such that on A_n, we have

$$d_n^{(\kappa)} = \begin{cases} O((\log n/n)^{1/2}) & \text{if } \kappa > k_0 \\ d_0^{(\kappa)} + O((\log n/n)^{1/2}) & \text{if } \kappa \leq k_0, \end{cases} \tag{6.56}$$

and

$$\overline{\ell}_n^{(\kappa)} = \begin{cases} \overline{\mathcal{L}}_0(f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}) + O((\log n/n)^{1/2}) & \text{if } \kappa \geq k_0 \\ \overline{\mathcal{L}}_0(f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}) + O((\log n/n)^{1/2}) & \text{if } \kappa = k_0 \\ \overline{\mathcal{L}}_0(f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}) - L^{(\kappa)} + o(1) & \text{if } \kappa < k_0, \end{cases} \tag{6.57}$$

where $L^{(\kappa)} = KL(p_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0} \| p_{\boldsymbol{\tau}^{0,(\kappa)},\boldsymbol{\theta}^{0,(\kappa)}}) > 0$ for all $\kappa < k_0$, and the constants in the big O and small o notions only depends on $(\boldsymbol{\tau}^0, \boldsymbol{\theta}^0), \Theta$, and k. Hence, we have

$$-\mathrm{DSC}_n^{(\kappa)} = \begin{cases} \omega_n \overline{\mathcal{L}}_0(f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}) + O(\omega_n(\log n/n)^{1/2}) & \text{if } \kappa > k_0, \\ \omega_n \overline{\mathcal{L}}_0(f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}) + d_0^{(k_0)} - O(\omega_n(\log n/n)^{1/2}) & \text{if } \kappa = k_0, \\ \omega_n \overline{\mathcal{L}}_0(f_{\boldsymbol{\tau}^0,\boldsymbol{\theta}^0}) + d_0^{(\kappa)} - \omega_n(L^{(\kappa)} + o(1)) & \text{if } \kappa < k_0. \end{cases} \tag{6.58}$$

Because $\omega_n \to \infty$, $\omega_n(\log n/n)^{1/2} \to 0$, and $L^{(\kappa)}$ is strictly positive, this implies that $\mathrm{DSC}^{(k_0)}$ is the smallest number for all n large enough. Hence, $\mathbb{P}_0(k_n = k_0) \geq \mathbb{P}_0(A_n) \to 1$, which means $k_n \to k_0$ in probability. □

6.4 Checking conditions

We spend this section checking the conditions $(K1), (K2), (K3)$ on popular kernels.

Condition (K1). *The kernel $p(x|\theta)$ satisfies*

$$\underline{c} \|\theta - \theta'\| \leq h(p(\cdot|\theta), p(\cdot|\theta')) \leq \overline{c} \|\theta - \theta'\| \quad \forall \theta, \theta' \in \Theta, \tag{6.59}$$

for some constant $\underline{c}, \overline{c}$ only depend on p and Θ.

Condition (K2). *Suppose that $\sup_{\theta\in\Theta} \|p(\cdot|\theta)\|_\infty$ is bounded, $\|p(\cdot|\theta) - p(\cdot|\theta')\|_\infty \lesssim \|\theta - \theta'\|$ for all $\theta, \theta' \in \Theta$, and $p(y|\theta)$ has uniformly light tails, i.e., there exist constants D, d_1, d_2, and d_2 so that $p(y|\theta) \leq d_1 \exp(-d_2 \|y\|^{d_3}) \forall \|y\| \geq D, \theta \in \Theta$.*

Condition (K3). *There exist positive constants c_α and c_β such that for all sufficiently small ϵ and $\theta_0, \theta \in \Theta$ such that $\|\theta - \theta_0\| \le \epsilon$, we have*

$$(1 - c_\beta\epsilon)\log p(y|\theta_0) + c_\alpha\epsilon \ge \log p(y|\theta) \ge (1 + c_\beta\epsilon)\log p(y|\theta_0) - c_\alpha\epsilon \quad \forall y.$$

Besides, there exists constant $\gamma_1, \gamma_2 > 0$ so that $\mathbb{P}_{y\sim p(\theta^0_{j'})}(|\log p(y|\theta^0_j)| \ge z) \le e^{-\gamma_1 z^{\gamma_2}}$ for all $z \ge 0$ and $j, j' \in [k_0]$.

Checking condition (K1)

Proposition 2. *Condition (K1) satisfied for:*

1. *Location Gaussian family $\{\mathcal{N}(\theta, \Sigma) : \theta \in \Theta\}$ for a compact set $\Theta \subset \mathbb{R}^d$ and a fixed, positive-definite covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$;*

2. *Location-scale Gaussian family $\{\mathcal{N}(\theta, \Sigma) : \theta \in \Theta, \Sigma \in S\}$ for a compact subspace $\Theta \subset \mathbb{R}^d$, and S is a compact set of positive-definite covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$.*

3. *Univariate Poisson family $\{\mathcal{P}(\theta) : \theta \in \Theta\}$ for a compact subspace $\Theta \subset \mathbb{R}$.*

Proof. 1. Recall the Hellinger distance between two Gaussian with the same covariance matrix Σ:

$$h^2(\mathcal{N}(\theta_1, \Sigma), \mathcal{N}(\theta_2, \Sigma)) = 1 - \exp\left\{-\frac{1}{8}(\theta_1 - \theta_2)^\top \Sigma^{-1}(\theta_1 - \theta_2)\right\}.$$

We aim to show that $h^2(\mathcal{N}(\theta_1, \Sigma), \mathcal{N}(\theta_2, \Sigma)) \asymp \|\theta_1 - \theta_2\|^2$. Indeed, first notice that

$$\frac{1}{\sigma_{\max}}\|\theta_1 - \theta_2\|^2 \le (\theta_1 - \theta_2)^\top \Sigma^{-1}(\theta_1 - \theta_2) \le \frac{1}{\sigma_{\min}}\|\theta_1 - \theta_2\|^2,$$

for all $\theta_1, \theta_2 \in \Theta$, where $\sigma_{\min}$ and $\sigma_{\max}$ are the smallest and largest eigenvalue of Σ. Besides, we have

$$cx \le 1 - \exp(-x) \le x \tag{6.60}$$

for all $x \in [0, C]$ where c depends on C. Because the parameter space Θ is compact, we have $0 \leq (\theta_1 - \theta_2)^\top \Sigma^{-1} (\theta_1 - \theta_2) \leq \frac{1}{\sigma_{\min}} \|\theta_1 - \theta_2\|^2 \leq \frac{1}{\sigma_{\min}} \text{diam}(\Theta)^2$. Hence,

$$\|\theta_1 - \theta_2\|^2 \asymp \frac{1}{8}(\theta_1 - \theta_2)^\top \Sigma^{-1} (\theta_1 - \theta_2) \asymp h^2(\mathcal{N}(\theta_1, \Sigma), \mathcal{N}(\theta_2, \Sigma)),$$

for all $\theta_1, \theta_2 \in \Theta$.

2. Recall the Hellinger distance between two location-scale Gaussian:

$$h^2(\mathcal{N}(\theta_1, \Sigma_1), \mathcal{N}(\theta_2, \Sigma_2)) = 1 - \frac{|\Sigma_1|^{1/4} |\Sigma_2|^{1/4}}{|(\Sigma_1 + \Sigma_2)/2|^{1/2}} \exp\left\{ -\frac{1}{8}(\theta_1 - \theta_2)^\top \left(\frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (\theta_1 - \theta_2) \right\}.$$

Using the same technique as in the previous proof, we have

$$h^2(\mathcal{N}(\theta_1, \Sigma_1), \mathcal{N}(\theta_2, \Sigma_2)) \asymp \log(|\Sigma_1|) + \log(|\Sigma_2|) - 2 \log\left(\left| \frac{\Sigma_1 + \Sigma_2}{2} \right| \right) + \|\theta_1 - \theta_2\|^2.$$

By the second order Taylor expansion, we have that

$$\log(|\Sigma_1|) + \log(|\Sigma_2|) - 2 \log\left(\left| \frac{\Sigma_1 + \Sigma_2}{2} \right| \right) \asymp \|\Sigma_1 - \Sigma_2\|^2.$$

Hence,

$$h^2(\mathcal{N}(\theta_1, \Sigma_1), \mathcal{N}(\theta_2, \Sigma_2)) \asymp \|\Sigma_1 - \Sigma_2\|^2 + \|\theta_1 - \theta_2\|^2.$$

3. The squared Hellinger distance between two Poisson distributions with parameters θ and θ' is given by:

$$h^2(\mathcal{P}(\theta), \mathcal{P}(\theta')) = 1 - e^{-\frac{1}{2}(\sqrt{\theta} - \sqrt{\theta'})^2}.$$

We then have that $h(\mathcal{P}(\theta), \mathcal{P}(\theta')) \approx C|\sqrt{\theta} - \sqrt{\theta'}|$. Since $|\sqrt{\theta} - \sqrt{\theta'}|$ is equivalent

to $\dfrac{|\theta - \theta'|}{\sqrt{\theta} + \sqrt{\theta'}}$, we can find bounds:

$$\underline{c}|\theta - \theta'| \leq C|\sqrt{\theta} - \sqrt{\theta'}| \leq \overline{c}|\theta - \theta'|$$

for some constants $\underline{c}, \overline{c}$ depending on Θ. Thus, the Poisson kernel satisfies the given condition (K1).

□

Checking condition (K2)

Proposition 3. *Condition (K2) satisfied for:*

1. *Location Gaussian family* $\{\mathcal{N}(\theta, \Sigma) : \theta \in \Theta\}$ *for a compact set* $\Theta \subset \mathbb{R}^d$ *and a fixed, positive-definite covariance matrix* $\Sigma \in \mathbb{R}^{d \times d}$*;*

2. *Location-scale Gaussian family* $\{\mathcal{N}(\theta, \Sigma) : \theta \in \Theta, \Sigma \in S\}$ *for a compact subspace* $\Theta \subset \mathbb{R}^d$*, and* S *is a compact set of positive-definite covariance matrix* $\Sigma \in \mathbb{R}^{d \times d}$*.*

3. *Univariate Poisson family* $\{\mathcal{P}(\theta) : \theta \in \Theta\}$ *for a compact subspace* $\Theta \subset \mathbb{R}$*.*

Proof. 1. For the location Gaussian family the density is

$$p(y|\theta) = (2\pi)^{-m/2}(\det \Sigma)^{-1/2} \exp\left(-\frac{1}{2}(y - \theta)^\top \Sigma^{-1}(y - \theta)\right),$$

- **Boundedness of** $\sup_{\theta \in \Theta} \|p(\cdot|\theta)\|_\infty$.

  The maximum of the density occurs at $y = \theta$, giving

  $$p(y|\theta) \leq (2\pi)^{-m/2}(\det \Sigma)^{-1/2}.$$

  This upper bound is independent of θ, so we conclude that

  $$\sup_{\theta \in \Theta} \|p(\cdot|\theta)\|_\infty < \infty.$$

This condition is satisfied.

- **Continuity condition:** $\|p(\cdot|\theta) - p(\cdot|\theta')\|_\infty \lesssim \|\theta - \theta'\|$

  Taking the derivative with respect to θ, we have

$$\begin{aligned}\frac{\partial}{\partial\theta}p(y|\theta) &= p(y|\theta)\cdot\|\Sigma^{-1}(y-\theta)\| \\ &\le p(y|\theta)\cdot \underbrace{\|\Sigma^{-1}\|}_{\text{operator norm}}\cdot\|y-\theta\| \quad \text{(by Cauchy-Schwarz inequality)} \\ &\le p(y|\theta)\frac{\|y-\theta\|}{\sigma_{\min}}. \quad (\sigma_{\min} \text{ is the smallest eigenvalue of } \Sigma)\end{aligned}$$

  From the Mean Value Theorem, the difference $|p(y|\theta) - p(y|\theta')|$ can be expressed in terms of the gradients with respect to θ:

$$\sup_{y\in\mathbb{R}^d}|p(y|\theta) - p(y|\theta')| \le L_\theta\|\theta-\theta'\|,$$

  where

$$\begin{aligned}L_\theta &= \sup_{y\in\mathbb{R}^d} p(y|\theta)\frac{\|y-\theta\|}{\sigma_{\min}} \\ &= \sup_{y\in\mathbb{R}^d} \underbrace{\frac{(2\pi)^{-m/2}(\det\Sigma)^{-1/2}}{\sigma_{\min}}}_{\text{a fixed number}} \underbrace{\|\theta-\theta'\|\exp\left(-\frac{1}{2}(y-\theta)^\top\Sigma^{-1}(y-\theta)\right)}_{\text{is bounded}} \le C\end{aligned}$$

  is Lipschitz constant.

  This implies $\|p(\cdot|\theta) - p(\cdot|\theta')\|_\infty \lesssim \|\theta - \theta'\|$.

- **Light tails condition:** $p(y|\theta) \le d_1\exp\left(-d_2\|y\|^{d_3}\right)$ for large $\|y\|$.

  The Gaussian density has an exponential tail:

$$p(y|\theta) = (2\pi)^{-m/2}(\det\Sigma)^{-1/2}\exp\left(-\frac{1}{2}(y-\theta)^\top\Sigma^{-1}(y-\theta)\right).$$

First, we notice that

$$\frac{1}{\sigma_{\max}}\|y-\theta\|^2 \leq (y-\theta)^\top \Sigma^{-1}(y-\theta) \leq \frac{1}{\sigma_{\min}}\|y-\theta\|^2,$$

for all $\theta \in \Theta$, where $\sigma_{\min}$ and $\sigma_{\max}$ are the smallest and largest eigenvalue of Σ. Additionally, for large $\|y\|$, we approximate $\|y-\theta\| \approx \|y\|$ (since Θ is compact and θ is bounded).

Thus, we obtain:

$$p(y|\theta) \leq d_1 \exp(-d_2\|y\|^2),$$

where $d_1 = (2\pi)^{-m/2}(\det\Sigma)^{-1/2}$ and $d_2 = \dfrac{1}{2\sigma_{\max}}$. Since this satisfies the required exponential decay condition with $d_3 = 2$, the Gaussian density has uniformly light tails.

2. For the location-scale Gaussian family, where both mean θ and the covariance matrix Σ vary, the Gaussian density is given by:

$$p(y|\theta,\Sigma) = (2\pi)^{-m/2}(\det\Sigma)^{-1/2}\exp\left(-\frac{1}{2}(y-\theta)^\top\Sigma^{-1}(y-\theta)\right).$$

- **Boundedness of** $\sup_{\theta\in\Theta,\Sigma\in S}\|p(\cdot|\theta,\Sigma)\|_\infty$.

  The density is maximized at $y=\theta$, giving

$$p(y|\theta,\Sigma) \leq (2\pi)^{-m/2}(\det\Sigma)^{-1/2}.$$

  Since S is compact and contains only positive-definite matrices, both $\det\Sigma$ and its inverse are bounded away from zero and infinity. Thus, there exists

a constant C such that

$$\sup_{\theta\in\Theta,\Sigma\in S} \|p(\cdot|\theta,\Sigma)\|_\infty \leq C < \infty.$$

- **Continuity condition:** $\|p(\cdot|\theta,\Sigma) - p(\cdot|\theta',\Sigma')\|_\infty \lesssim \|\theta-\theta'\| + \|\Sigma-\Sigma'\|$.

  Using the same technique as the location Guassian case, we analyze the difference: $|p(y|\theta,\Sigma) - p(y|\theta',\Sigma')|$.

  Since the density is smooth in both θ and Σ, we approximate:

$$p(y|\theta,\Sigma) - p(y|\theta',\Sigma') \approx \frac{\partial p}{\partial\theta}(\theta-\theta') + \frac{\partial p}{\partial\Sigma}(\Sigma-\Sigma').$$

  The derivatives are bounded because Θ and S are compact. Thus, there exists a constant C such that:

$$\|p(\cdot|\theta,\Sigma) - p(\cdot|\theta',\Sigma')\|_\infty \leq C(\|\theta-\theta'\| + \|\Sigma-\Sigma'\|).$$

- **Light Tails Condition:** $p(y|\theta,\Sigma) \leq d_1\exp(-d_2\|y\|^{d_3})$ for large $\|y\|$

  Similar to the location Gaussian case.

3. For the Poisson probability mass function, we have

$$p(y|\theta) = \frac{\theta^y e^{-\theta}}{y!}, \quad y\in\mathbb{N}, \quad \theta > 0.$$

- **Boundedness of** $\sup_{\theta\in\Theta}\|p(\cdot|\theta)\|_\infty$.

  The probability mass function is always between 0 and 1 for every $\theta\in\Theta$. Thus, $\sup_{\theta\in\Theta}\|p(\cdot|\theta)\|_\infty$ is finite for a bounded set Θ.

- **Continuity Condition** $\|p(\cdot|\theta) - p(\cdot|\theta')\|_\infty \lesssim \|\theta-\theta'\|$

Taking the derivative of $p(y|\theta)$ with respect to θ, we have

$$\frac{\partial p}{\partial \theta} = p(y|\theta)\left(\frac{y}{\theta} - 1\right).$$

From the Mean Value Theorem, the difference $|p(y|\theta) - p(y|\theta')|$ can be experess as the following:

$$\sup_{y\in\mathbb{R}^d} |p(y|\theta) - p(y|\theta')| \leq L_\theta |\theta - \theta'|.$$

where $L_\theta = \sup_{y\in\mathbb{R}^d} p(y|\theta)\left|\frac{y}{\theta} - 1\right| \leq C$ because of the compactness of Θ and the properties of the Poisson mass function.

- **Light tails condition** $p(y|\theta) \leq d_1 \exp(-d_2\|y\|^{d_3})$
  For large y, Stirling's approximation $y! \approx \sqrt{2\pi y}\left(\frac{y}{e}\right)^y$ gives:

$$p(y|\theta) \approx \frac{\theta^y e^{-\theta}}{\sqrt{2\pi y}(y/e)^y} = \frac{1}{\sqrt{2\pi y}} e^{y\ln\theta - \theta - y\ln y + y} = e^{-y(\ln y - \ln\theta - 1) - \theta} \cdot \frac{1}{\sqrt{2\pi y}}$$

For large y, the dominant term in the exponent is $-y\ln y$, ensuring exponential decay. Thus there exists constants d_1, d_2, d_3 such that: $p(y|\theta) \leq d_1 \exp(-d_2 y^{d_3})$ for sufficiently large y, confirming light tails.

□

Checking condition (K3)

Proposition 4. *Condition (K3) satisfied for:*

1. *Location Gaussian family* $\{\mathcal{N}(\theta, \Sigma) : \theta \in \Theta\}$ *for a compact set* $\Theta \subset \mathbb{R}^d$ *and a fixed, positive-definite covariance matrix* $\Sigma \in \mathbb{R}^{d\times d}$*;*

2. *Location-scale Gaussian family* $\{\mathcal{N}(\theta, \Sigma) : \theta \in \Theta, \Sigma \in S\}$ *for a compact subspace* $\Theta \subset \mathbb{R}^d$*, and* S *is a compact set of positive-definite covariance matrix* $\Sigma \in \mathbb{R}^{d\times d}$*.*

3. Univariate Poisson family $\{\mathcal{P}(\theta) : \theta \in \Theta\}$ *for a compact subspace* $\Theta \subset \mathbb{R}$.

Proof. 1,2. We will check (K3) for the location-scale Gaussian family (also see [31]). For the location Gaussian family, the technique is similar.

Let $\sigma_{\min}$ and $\sigma_{\max}$ be the smallest and largest eigenvalue of Σ.

Consider parameters (θ_0, Σ_0) and (θ, Σ) such that $\|\theta - \theta_0\|, \|\Sigma - \Sigma_0\| \leq \epsilon$ sufficiently small. Plugging the pdf of the Gaussian distribution into condition (K3), we aim to show that there exist c_α and $c_\beta > 0$ such that

$$(1+c_\beta\epsilon)\log(|\Sigma_0|)-\log(|\Sigma|)+(1+c_\beta\epsilon)(x-\theta_0)^\top\Sigma_0^{-1}(x-\theta_0)-(x-\theta)^\top\Sigma^{-1}(x-\theta)+c_\alpha\epsilon \geq 0.$$

Firstly, we note that

$$\frac{d\log(|\Sigma|)}{d\Sigma} = \Sigma^{-1}$$

being bounded above and below (as positive-definite matrices). So that the map $\Sigma \mapsto \log(|\Sigma|)$ is Lipschitz, i.e., there exists constant c_σ such that

$$|\log(|\Sigma_0|) - \log(|\Sigma|)| \leq c_\sigma \|\Sigma_0 - \Sigma\| \leq c_\sigma\epsilon.$$

Hence, for all $c_\beta > c_\sigma/(d\log(\sigma_{\min}))$, we have

$$c_\beta\epsilon\log(|\Sigma_0|) \geq c_\sigma\epsilon \geq |\log(|\Sigma|) - \log(|\Sigma_0|)|.$$

So that

$$(1 + c_\beta\epsilon)\log(|\Sigma_0|) - \log(|\Sigma|) > 0.$$

Next, denote $u = x - \theta_0$ and $\Delta u = \theta_0 - \theta$. Using the boundedness of Σ_0 again, we have

$$\Sigma_0^{-1} \geq (1 - c_\Sigma)\Sigma^{-1},$$

where c_Σ only depends on $\sigma_{\min}$ and $\sigma_{\max}$. Hence, to show that

$$(1 + c_\beta\epsilon)(x - \theta_0)^\top \Sigma_0^{-1}(x - \theta_0) - (x - \theta)^\top \Sigma^{-1}(x - \theta) + c_\alpha\epsilon \geq 0,$$

we only need to show

$$(1 + c_\beta\epsilon)(1 - c_\Sigma\epsilon)u^\top \Sigma^{-1}u - (u + \Delta u)^\top \Sigma^{-1}(u + \Delta u) + c_\alpha\epsilon \geq 0.$$

This is equivalent to

$$\begin{aligned}
&(c_\beta - c_\Sigma)\epsilon \left(u - \frac{\Delta u}{(c_\beta - c_\Sigma)\epsilon}\right)^\top \Sigma^{-1} \left(u - \frac{\Delta u}{(c_\beta - c_\Sigma)\epsilon}\right) + c_\alpha\epsilon \\
&\geq c_\beta c_\Sigma \epsilon^2 + \frac{1}{(c_\beta - c_\Sigma)\epsilon}(\Delta u)^\top \Sigma^{-1}(\Delta u),
\end{aligned}$$

which is correct by noticing that the first term of LHS is non-negative, the second term is greater than the whole RHS when choosing $c_\beta > c_\Sigma$ and choose

$$c_\alpha \geq c_\beta c_\Sigma \epsilon + \frac{1}{(c_\beta - c_\Sigma)} \frac{1}{\sigma_{\min}},$$

as $(\Delta u)^\top \Sigma^{-1}(\Delta u) \leq \|\Delta u\|^2 / \sigma_{\min} \leq \epsilon^2 / \sigma_{\min}$.

For the second part of the condition (K3), we need to check whether the tail bound condition holds:

$$\mathbb{P}_{y \sim p(\mu, \Sigma)} \left(|\log p(y|\mu, \Sigma)| \geq z \right) \leq e^{-\gamma_1 z^{\gamma_2}},$$

for some $\gamma_1, \gamma_2 > 0$.

We have that for large z, the left-hand side of the condition simplifies to:

$$\mathbb{P}\left(\frac{1}{2}(y-\theta)^\top \Sigma^{-1}(y-\theta) \geq z\right) = \mathbb{P}\left((y-\theta)^\top \Sigma^{-1}(y-\theta) \geq 2z\right).$$

Since $(y-\theta) \sim \mathcal{N}(0,\Sigma)$, the quadratic form $Q = (y-\theta)^\top \Sigma^{-1}(y-\theta)$ follows a chi-square distribution with d degrees of freedom: $Q \sim \chi^2_d$. Using standard chi-square tail bounds, for $t \geq d$, we have that $\mathbb{P}(\chi^2_d \geq t) \leq e^{-t/4}$.

Setting $t = 2z$, we obtain:

$$\mathbb{P}\left((y-\theta)^\top \Sigma^{-1}(y-\theta) \geq 2z\right) \leq e^{-z/2}.$$

Therefore, $\gamma_1 = 1/2$ and $\gamma_2 = 1$ in this case.

3. For the Poisson kernel, recall that the probability mass function is:

$$p(y|\theta) = \frac{\theta^y e^\theta}{y!}.$$

Taking the logarithm, the log-likelihood function is $\log p(y|\theta) = y\log\theta - \theta - \log(y!)$.

Considering the inequality in condition (K3), we want to show that there exists $c_\beta, c_\alpha > 0$ such that

$$\begin{aligned} y(1-c_\beta\epsilon)\log\theta_0 - (1-c_\beta\epsilon)\theta_0 + c_\alpha\epsilon &\geq y\log\theta - \theta \\ &\geq y(1+c_\beta\epsilon)\log\theta_0 - (1+c_\beta\epsilon)\theta_0 - c_\alpha\epsilon \end{aligned} \tag{6.61}$$

Since $|\theta - \theta_0| \leq \epsilon$, we assume that

$$\theta = \theta_0 + \delta, \quad \text{with } |\delta| \leq \epsilon.$$

Using first-order Taylor expansion:

$$\log \theta \approx \log \theta_0 + \frac{\delta}{\theta_0}.$$

So,

$$\begin{aligned} y \log \theta - \theta &\approx y \left(\log \theta_0 + \frac{\delta}{\theta_0} \right) - (\theta_0 + \delta) \\ &= y \log \theta_0 + y \frac{\delta}{\theta_0} - \theta_0 - \delta. \end{aligned}$$

Comparing to the left-hand side of (6.61): $y \log \theta_0 - y c_\beta \epsilon \log \theta_0 - \theta_0 + c_\beta \theta_0 + c_\alpha \epsilon$, for small ϵ, this bound holds if we set $\delta = c_\beta \epsilon \theta_0$.

Comparing to the right-hand side of (6.61): $y \log \theta_0 + y c_\beta \epsilon \log \theta_0 - \theta_0 - c_\beta \theta_0 - c_\alpha \epsilon$, for small ϵ, this bound holds if we set $\delta = -c_\beta \epsilon \theta_0$.

Thus, the bound in (6.61) holds for the Poisson distribution for the appropriate choices of c_α, c_β.

In terms of the second part of the condition (K3), we want to check if there exist constants $\gamma_1, \gamma_2 \geq 0$ such that: $\mathbb{P}_{y \sim p(\theta_0)} \left(|\log p(y|\theta_0)| \geq z \right) \leq e^{-\gamma_1 z^{\gamma_2}}$.

Since $\log p(y|\theta_0) = y \log \theta_0 - \theta_0 - \log y!$, we need to control the probability: $\mathbb{P}(|y \log \theta_0 - \theta_0| \geq z)$. Using concentration inequalities for Poisson distributions, we know $\mathbb{P}(|Y - \theta_0| \geq t) \leq 2e^{-t^2/(2\theta_0)}$. Since the log-likelihood depends linearly

on Y, we substitute $t = \dfrac{z}{\log \theta_0}$, leading to:

$$\mathbb{P}\left(|y \log \theta_0 - \theta_0| \geq z\right) \leq 2e^{-\dfrac{z^2}{2\theta_0 \log^2 \theta_0}}.$$

We can choose $\gamma_1 = \dfrac{1}{2\theta_0 \log^2 \theta_0}$ and $\gamma_2 = 2$, and the second part of Condition (K3) is satisfied.

□

Chapter 7

Conclusion and future investigation

In this thesis, we introduced and explored novel statistical methodologies for change-point detection in multidimensional data, with a particular focus on intracellular transport analysis. Our contributions span theoretical advancements, algorithmic developments, and practical applications, offering robust solutions to challenges in detecting velocity changes and segment classification.

The development of the CPLASS algorithm in Chapters 2 and 3 addressed critical gaps in existing change-in-velocity detection methods. Traditional approaches, including binary segmentation and dynamic programming, were inadequate due to the continuity constraint of the model. In the framework for detecting velocity changes, there are limitations in handling multidimensional data within existing methods. For preventing the overfitted issue, we introduced a tailored penalty function with the present of the strengthened Schwarz Information Criterion and a customized speed penalty. By using an MCMC-based framework with customized proposal mechanisms, CPLASS effectively explores the complex parameter space and delivers reliable segmentation results. While the consistency theorem ensures long-term robustness, real-world applications with small sample sizes necessitated additional methodological refinements. The introduction of a speed penalty and the comparison of different trajectory characterization methods, such as Cumulative Speed Allocation (CSA)

and Cumulative Distribution Function (CDF), further enhanced the stability and reliability of speed inference. However, computational efficiency remains a challenge, and future efforts should focus on optimizing the MCMC search process. Additionally, ensuring the mathematical consistency of CSA-based inference and addressing anchor diffusion effects remain open problems for further exploration.

Chapter 4 introduced a hypothesis testing framework for segment classification into stationary and motile states under a continuous piecewise linear model assumption. By reformulating the problem within the general linear hypothesis testing framework, we demonstrated the efficacy of an F-distributed test statistic in detecting motile segments, even at low speeds. The test was validated through simulation studies and real-data applications, underscoring its effectiveness in trajectory characterization. However, significant challenges persist, particularly in handling anchor diffusion effects and mitigating the double-dipping problem that arises from using the same data to infer and test changepoints. While existing literature on selective inference offers potential solutions, current methods focus primarily on mean-change problems and are not directly applicable to velocity changes. Future research should aim to adapt post-selection inference techniques to stochastic search-based changepoint detection methods like CPLASS, ensuring valid statistical inferences under these conditions. Additionally, alternative approaches such as *post-inference selection*, introduced by Fryzlewicz (2023) [47], could provide a promising direction for establishing confidence intervals around detected changepoints rather than post-hoc hypothesis testing.

In Chapters 5 and 6, we introduced the Dendrogram Pruning and Merging (DPM) algorithm as a computationally efficient alternative for multiple changepoint detection. Unlike conventional methods requiring repeated model fitting, DPM constructs a hierarchical binary tree structure for changepoint selection in a single pass, significantly reducing computational costs. The associated Dendrogram Selection Criterion (DSC)

further enhances model selection by leveraging hierarchical relationships and parameter distances. Theoretical guarantees affirm the consistency of this approach, while empirical evaluations demonstrate its competitive performance. Given its adaptability to various kernel setups and multi-dimensional data, DPM presents a promising direction for future changepoint detection frameworks. Future research could explore integrating DPM with velocity change detection techniques, extending its applicability to dependent data structures, and refining theoretical guarantees to enhance its robustness.

A combination of CPLASS and DPM presents a promising direction for future research, enabling the detection of both changes in velocity and changes in diffusivity. One potential application lies in studying the movement of bacteria, which, unlike motor-driven vesicles, are not transported along microtubules but instead move through extracellular environments such as the body's mucus layers. These layers act as the first line of defense against pathogens and present a heterogeneous medium that can significantly affect bacterial motion. In this thesis, we focus on intracellular transport involving dynein- and kinesin-driven vesicles (see Chapter 2). In contrast, bacteria such as Salmonella utilize flagellar propulsion, which drives a well-known behavior called run-and-tumble motion [108, 118]. Each Salmonella cell is equipped with multiple flagella that rotate in unison to generate directed motion. When flagella become unsynchronized, this leads to tumbling—a brief period of randomized movement—before the next directed run. The nature of bacterial motion can be influenced by various biological conditions, including the viscosity and composition of mucus, the presence of immune factors like antibodies, and other environmental stresses (see Fig. 6 in [118]). Developing a changepoint detection framework that integrates CPLASS for detecting velocity transitions and DPM for identifying changes in diffusivity could offer new insights into how bacterial locomotion adapts under such

conditions.

Overall, this thesis advances the field of change-point detection by developing novel methods tailored for velocity changes in multidimensional data. The CPLASS algorithm provides a statistically rigorous and flexible approach for segmenting intracellular transport trajectories, while the hypothesis testing framework for segment classification offers a reliable tool for distinguishing motile and stationary states. The DPM algorithm contributes a computationally efficient alternative for changepoint detection, paving the way for future methodological advancements. Despite these contributions, challenges such as computational efficiency, selective inference for stochastic search methods, and anchor diffusion effects remain areas for further investigation. Addressing these challenges will be crucial in refining these methodologies and expanding their applicability in biophysical and broader statistical contexts. Through continued development and interdisciplinary collaboration, the insights gained from this work hold the potential to significantly enhance the precision and interpretability of complex dynamic processes in intracellular transport and beyond.

References

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

[2] Andreas Anastasiou and Piotr Fryzlewicz. Detecting multiple generalized change-points by isolating single ones. *Metrika*, 85:141–174, 2022.

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

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

[5] Jushan Bai. Estimating multiple breaks one at a time. *Econometric Theory*, 13:315–352, 1997.

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

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

[8] Jushan Bai and Pierre Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18:1–22, 2003.

[9] Rafal Baranowski, Yaakov Chen, and Piotr Fryzlewicz. 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, 2019.

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

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

[12] Michèle Basseville and Igor V. Nikiforov. Detection of abrupt changes: theory and application. *Technometrics*, 36:550, 1993.

[13] C. Beaulieu, J. Chen, and J.L. Sarmiento. Change-point analysis as a tool to detect abrupt climate variations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1962):1228–1249, 2012.

[14] R. Bellman. On a routing problem. *Quarterly of Applied Mathematics*, 16:87–90, 1958.

[15] Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

[16] Lucien Birgé and Pascal Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, 2001.

[17] Lars Boysen, Andre Kempe, Volkmar Liebscher, Axel Munk, and Olaf Wittich. Consistencies and rates of convergence of jump-penalized least squares estimators. *The Annals of Statistics*, 37(1):157–183, 2009.

[18] Jon V. Braun, R. J. Braun, and Hans-Georg Mueller. Multiple Changepoint Fitting via Quasi-Likelihood, with Application to DNA Sequence Segmentation. *Biometrika*, 87(2):301–314, 2000.

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

[20] B. Brodsky and B. Darkhovsky. *Nonparametric Methods in Change-Point Problems*. Kluwer Academic Publishers, 1993.

[21] Jiahua Chen, Arjun Gupta, and Jianmin Pan. Information criterion and change point problem for regular models. *Sankhyā: The Indian Journal of Statistics*, 68(2):252–282, 05 2006.

[22] K. M. Chen, A. Cohen, and H. Sackrowitz. Consistent multiple testing for change points. *Journal of Multivariate Analysis*, 102:1339–1343, 2011.

[23] S. S. Chen and P. S. Gopalakrishnan. 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, 1998.

[24] Haeran Cho and Piotr Fryzlewicz. Multiscale interpretation of taut string estimation and its connection to unbalanced haar wavelets. *Statistics and Computing*, 21:671–681, 2011.

[25] Haeran Cho and Piotr Fryzlewicz. Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statistica Sinica*, 22:207–229, 2012.

[26] Haeran Cho and Piotr Fryzlewicz. Multiple change-point detection for high-dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society Series B*, 77:475–507, 2015.

[27] Gabriela Ciuperca. A general criterion to determine the number of change-points. *Statistics & Probability Letters*, 81:1267–1275, 2011.

[28] Gabriela Ciuperca. Model selection by lasso methods in a change-point model. *Statistical Papers*, 55:349–374, 2014.

[29] Keisha J. Cook, Nathan Rayens, Linh Do, Christine K. Payne, and Scott A. McKinley. Considering experimental frame rates and robust segmentation analysis of piecewise-linear microparticle trajectories, 2024.

[30] Miklós Csörgő and Lajos Horváth. *Limit Theorems in Change-Point Analysis*. John Wiley & Sons, 1997.

[31] Dat Do, Linh Do, Scott A McKinley, Jonathan Terhorst, and Xuanlong Nguyen. Dendrogram of mixing measures: Learning latent hierarchy and model selection for finite mixture models. *arXiv [stat.ME]*, March 2024.

[32] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[33] M. F. Driscoll and W. R. Grundberg. A History of the Development of Craig's Theorem. *The American Statistician*, 40:65–69, 1986.

[34] M. F. Driscoll and B. Krasnicka. An Accessible Proof of Craig's Theorem in the General Case. *The American Statistician*, 49:59–61, 1995.

[35] C. Du, C. L. Kao, and S. Kou. Stepwise signal extraction via marginal likelihood. *Journal of the American Statistical Association*, 111:314–330, 2016.

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

[37] M. Eichinger and C. Kirch. MOSUM Control Charts for Monitoring Time Series. *Quality and Reliability Engineering International*, 34:1263–1278, 2018.

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

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

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

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

[42] Paul Fearnhead and Piotr Fryzlewicz. Detecting a single change-point. *arXiv [stat.ME]*, October 2022.

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

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

[45] Piotr Fryzlewicz. Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *The Annals of Statistics*, 46(6B):3390–3421, 2018.

[46] Piotr Fryzlewicz. Detecting possibly frequent change-points: Wild binary segmentation 2 and steepest-drop model selection. *Journal of the Korean Statistical Society*, 49:1–44, 03 2020.

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

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

[49] Andreas Futschik, Thomas Hotz, Axel Munk, and Hannes Sieling. Multiscale DNA partitioning: statistical evidence for segments. *Bioinformatics*, 30(16):2255–2262, 04 2014.

[50] Zhanzhongyu Gao, Xun Xiao, Yi-Ping Fang, Jing Rao, and Huadong Mo. A selective review on information criteria in multiple change point detection. *Entropy*, 2024.

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

[52] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

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

[54] Arne Hallam. Some theorems on quadratic forms and normal variables. Lecture note, 2008.

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

[56] Z. Harchaoui, F. Vallet, A. Lung-Yut-Fong, and O. Cappe. 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, 2009.

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

[58] Heping He and Thomas A. Severini. Asymptotic properties of maximum likelihood estimators in models with multiple change points. *Bernoulli*, 16(3):759 – 779, 2010.

[59] David V Hinkley. Time-ordered classification. *Biometrika*, 59(3):509–523, 1972.

[60] David V Hinkley and Elizabeth A Hinkley. Inference about the change-point in a sequence of binomial variables. *Biometrika*, 57(3):477–488, 1970.

[61] D.V. Hinkley. Inference about the change-point in a sequence of random variables. *Biometrika*, 57(1):1–17, 1970.

[62] R. R. Hocking. *The Analysis of Linear Models*. Brooks/Cole, Monterey, 1985.

[63] R. R. Hocking. *Methods and Applications of Linear Models*. Wiley, New York, 1996.

[64] P.W. Holland and C. E. P. M. Statistical methods for identifying copy number changes in high-throughput sequencing data. *Statistical Applications in Genetics and Molecular Biology*, 11(1):1–28, 2012.

[65] Y. Huang et al. A comprehensive framework for the detection of copy number variations from next-generation sequencing data. *Nature Protocols*, 8(1):145–158, 2013.

[66] M. Huskova and M. Slaby. Moving sum control charts for change detection. *Statistics in Medicine*, 20:2041–2051, 2001.

[67] Sangwon Hyun, Kevin Z. Lin, Max G'Sell, and Ryan J. Tibshirani. Post-selection inference for changepoint detection algorithms with application to copy number variation data. *Biometrics*, 77(3):1037–1049, 2021.

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

[69] B. Jackson, J.D. Scargle, D. Barnes, et al. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2):105–108, 2005.

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

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

[72] Richard H. Jones and Indranil Dey. Determining one or more change points. *Chemistry and Physics of Lipids*, 76(1):1–6, 1995.

[73] Joshua D. Karslake, Eric D. Donarski, Sarah A. Shelby, Lisa M. Demey, Victor J. DiRita, Sarah L. Veatch, and Julie S. Biteen. Smaug: Analyzing single-molecule tracks with nonparametric bayesian statistics. *Methods*, 193:16–26, 2021.

[74] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[75] Richard W. Katz. On some criteria for estimating the order of a markov chain. *Technometrics*, 23(3):243–249, 1981.

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

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

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

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

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

[81] M. Lavielle and G. Teyssiere. Adaptive detection of multiple change-points in

asset price volatility. In *Long-Memory in Economics*, pages 129–156. Springer Verlag, Berlin, Germany, 2007.

[82] Marc Lavielle. Detection of multiple changes in a sequence of dependent variables. *Stochastic Processes and their Applications*, 83(1):79–102, 1999.

[83] Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal Processing*, 85(8):1501–1510, 2005.

[84] Marc Lavielle and Eric Moulines. Least-squares estimation of an unknown number of shifts in a time series. *Journal of Time Series Analysis*, 21(1):33–59, 2000.

[85] Marc Lavielle and Gilles Teyssière. *Detection of multiple change-points in multivariate time series*, volume 46. Springer, 2006.

[86] Marc Lavielle and Gilles Teyssière. Adaptive detection of multiple change-points in asset price volatility. *Springer*, pages 129–156, 2007.

[87] Vo Nguyen Le Duy, Hiroki Toda, Ryota Sugiyama, and Ichiro Takeuchi. Computing valid p-values for optimal changepoints by selective inference using dynamic programming. In *Advances in Neural Information Processing Systems*, volume 33, pages 14434–14445, 2020.

[88] Edith Lebarbier. Detecting multiple change-points in the mean of gaussian process by model selection. *Signal Processing*, 85(4):717–736, 2005.

[89] C. B. Lee. Estimating the number of change points in a sequence of independent normal random variables. *Statistics and Probability Letters*, 25:241–248, 1995.

[90] Jason Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-

selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[91] S. Letzgus. Change-point detection in wind turbine scada data for robust condition monitoring with normal behaviour models. *Wind Energy Science*, 5(4):1375–1397, 2020.

[92] J. Li, R.B. Lund, and Y. Bai. Multiple changepoint detection in climate data series using a hierarchical bayesian approach. *Journal of Climate*, 28(19):7724–7740, 2015.

[93] Jian Liu, Shiying Wu, and J Zidek. On segmented multivariate regression. *Statistica Sinica*, pages 497–525, August 1997.

[94] QiQi Lu, Robert Lund, and Thomas C. M. Lee. An MDL approach to the climate segmentation problem. *The Annals of Applied Statistics*, 4(1):299 – 319, 2010.

[95] J.M. Lucas and R.B. Crosier. Fast initial response for CUSUM quality control schemes: Give your CUSUM a head start. *Technometrics*, 24(3):199–205, 1982.

[96] Lijing Ma and Georgy Sofronov. Change-point detection in autoregressive processes via the cross-entropy method. *Algorithms*, 13(5), 2020.

[97] Robert Maidstone, Toby Hocking, Guillem Rigaill, and Paul Fearnhead. On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, 27(2):519–533, 2017.

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

[99] David S Matteson and Nicholas A James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345, 2014.

[100] Nilah Monnier, Syuan-Ming Guo, Masashi Mori, Jun He, Péter Lénárt, and Mark Bathe. Bayesian Approach to MSD-Based Analysis of Particle Motion in Live Cells. *Biophysical Journal*, 103(3):616–626, 2012.

[101] Susanne Neumann, Romain Chassefeyre, George E. Campbell, and Sandra E. Encalada. Kymoanalyzer: A software tool for the quantitative analysis of intracellular transport in neurons. *Traffic*, 18(1):71–88, 2017.

[102] Yoshiyuki Ninomiya. Change-point model selection via AIC. *Annals of the Institute of Statistical Mathematics*, 67(5):943–961, October 2015.

[103] Yue S. Niu and Heping Zhang. The screening and ranking algorithm to detect dna copy number variations. *Annals of Applied Statistics*, 6(3):1306–1326, September 2012.

[104] Adam B Olshen, Eswaran S Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, 2004.

[105] E. S. Page. Continuous inspection schemes. *Biometrika*, 41:100–115, 1954.

[106] Jianmin Pan and Jiahua Chen. Application of modified information criterion to multiple change point problems. *Journal of Multivariate Analysis*, 97(10):2221–2241, 2006.

[107] Fredrik Persson, Magnus Lindén, Carl Unoson, and Johan Elf. Extracting intracellular diffusive states and transition rates from single-molecule tracking data. *Nature Methods*, 10:265–269, 2013.

[108] E. M. Purcell. Life at low reynolds number. *American Journal of Physics*, 45(1):3–11, 1977.

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

[110] C. R. Rao. *Linear Statistical Inference and its Applications*. Wiley, New York, 2nd edition, 1973.

[111] Nathan T. Rayens, Keisha J. Cook, Scott A. McKinley, and Christine K. Payne. Transport of lysosomes decreases in the perinuclear region: Insights from change-point analysis. *Biophysical Journal*, 121(7):1205–1218, 2022.

[112] J. Reeves, J. Chen, X.L. Wang, R. Lund, and Q. Lu. A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46(6):900–915, 2007.

[113] Thomas C. M Lee Richard A Davis and Gabriel A Rodriguez-Yam. Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101(473):223–239, 2006.

[114] Guillaume Rigaill. A pruned dynamic programming algorithm to recover the best segmentations with 1 to k_{max} change-points. *Journal de la Societe Francaise de Statistique*, 156:180–205, 2015.

[115] Alessandro Rinaldo. Properties and refinements of the fused lasso. *Annals of Statistics*, 37:2922–2952, 2009.

[116] C. Rojas and B. Wahlberg. Multiple change-point detection in time series with the fused lasso. *IEEE Transactions on Signal Processing*, 62:360–370, 2014.

[117] J. R. Schott. *Matrix Analysis for Statistics*. Wiley, New York, 1997.

[118] Holly A. Schroeder, Jay Newby, Alison Schaefer, Babu Subramani, Alan Tubbs, M. Gregory Forest, Ed Miao, and Samuel K. Lai. LPS-binding IgG arrests actively motile Salmonella Typhimurium in gastrointestinal mucus. *Mucosal Immunology*, 13(5):814–823, 2020.

[119] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464, 1978.

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

[121] S.R. Searle. Estimable functions and testable hypotheses in linear models. Technical report, Cornell University, 1966.

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

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

[124] Ritei Shibata. Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, 63(1):117–126, 04 1976.

[125] S.M.Kay and A.V.Oppenheim. *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*. Prentice Hall, 1993.

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

[127] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:91–108, 2005.

[128] Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

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

[130] Aad van der Vaart and Harry van Zanten. Information rates of nonparametric gaussian process methods. *J. Mach. Learn. Res.*, 12(60):2095–2119, 2011.

[131] V.K.C. Venema, O. Mestre, E. Aguilar, I. Auer, J. Guijarro, P. Domonkos, et al. Benchmarking homogenization algorithms for monthly data. *Climate of the Past*, 8(1):89–115, 2012.

[132] Eswaran S. Venkatraman. Consistency results in multiple change-point problems. Technical Report Technical Report No. 24, Department of Statistics, Stanford University, 1992.

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

[134] Jean-Philippe Vert and Kevin Bleakley. 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, 2010.

[135] Liudmila Vostrikova. Detection of 'disorder' in multidimensional random processes. *Soviet Mathematics Doklady*, 24:55–59, 1981.

[136] Y. Wang. Jump and Sharp Cusp Detection by Wavelets. *Biometrika*, 82:385–397, 1995.

[137] M. West and P.J. Harrison. *Bayesian forecasting and dynamic models*. Springer-Verlag, 1989.

[138] Y. Wu. Simultaneous change point analysis and variable selection in a regression problem. *Journal of Multivariate Analysis*, 99:2154–2171, 2008.

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

[140] Qiwei Yao and Richard A. Davis. The asymptotic behavior of the likelihood ratio statistic for testing a shift in mean in a sequence of independent normal variates. *Sankhyā: The Indian Journal of Statistics, Series A*, 48(3):339–353, 1986.

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

[142] Yuan-Chuan Yao. Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *The Annals of Statistics*, 12(4):1434–1447, 1984.

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

[144] Nancy R. Zhang and David O. Siegmund. A Modified Bayes Information Crite-

rion with Applications to the Analysis of Comparative Genomic Hybridization Data. *Biometrics*, 63(1):22–32, 04 2007.

Biography

The author was born and raised in Hanoi, Vietnam, and graduated from Vietnam National University - University of Education with a Bachelor in Mathematics Teacher Education in 2017. The author started the PhD program at Tulane University - Department of Mathematics in August 2019, eventually completing the program in May 2025.