NONPARAMETRIC BAYESIAN STATISTICS HARNESSING THE FORCES OF DATA IN CHANGE-POINT DETECTION AND SURVIVAL ANALYSIS

CHONG ZHONG

PhD

The Hong Kong Polytechnic University

2023

THE HONG KONG POLYTECHNIC UNIVERSITY

DEPARTMENT OF APPLIED MATHEMATICS

NONPARAMETRIC BAYESIAN STATISTICS HARNESSING THE FORCES OF DATA IN CHANGE-POINT DETECTION AND SURVIVAL ANALYSIS

CHONG ZHONG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

July 2023

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

______________________________(Signature)

Chong Zhong (Name of student)

Dedication

This thesis is dedicated to my parents and my fiancée Menyao with appreciation and love.

Abstract

Bayesian nonparametric priors are distributions on functions. In this thesis, we present several novel Bayesian approaches based on the elicitation of a set of nonparametric priors in two problems, change-point detection, and survival analysis. Through our success on each target, we demonstrate the fact that appropriate Bayesian nonparametric priors can harness the power of the data and promote statistical analysis from the perspectives of estimation, inference, prediction, and computation.

In Part I, we propose NOSE and SBPCPM, two jump-size-based Bayesian approaches to solve change-point detection. NOSE globally models the abrupt change process and identifies change-points based on the induced posterior estimates of jump sizes. We establish posterior inferential theories including the minimax optimality of posterior contraction, posterior consistency of both number and locations of change-points, and an asymptotic zero false negative rate in change-point discrimination under a novel Gamma-IBP weighted spike-and-slab type prior. Comprehensive numerical studies demonstrate that NOSE outperforms existing approaches. SBPCPM is extremely useful to detect the imperceptible change-points under a mean-shifted model. We propose a novel Beta process mixture model for the change signal. We establish the pointwisely asymptotic efficiency of the marginal MAP estimates of the change signal under the hypothesis of no change-points. The induced asymptotic normality of the jump size estimators leads to efficient hypothesis testing of change-points.

In Part II, we study the use of nonparametric priors in survival analysis. For right-censored survival outcomes, we propose BuLTM, a novel Bayesian method for prediction under the nonparametric transformation model. Unlike existing methods, we allow the model to be unidentified and assign weakly informative nonparametric priors to the infinite-dimensional parameters to facilitate efficient MCMC sampling. We show that the posterior is proper under the unidentified model. For recurrent event data, we propose a generalized shared frailty model to relax the strict proportional hazard assumption and apply the ANOVA DPP as the prior for baseline survival functions for model estimation.

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisor, Dr. Catherine Liu. Catherine is more than just an academic supervisor to me; she is a guiding light in navigating various aspects of my life. Catherine is always willing to listen to all my opinions and feelings and is open to adjusting her ideas based on mine. While she can be strict with me at times in the course of research, her expectations are cultivating my good habits that span from conducting research to managing tasks effectively. Under her supervision, I have undergone a remarkable transformation from a naïve and unknowledgeable novice in academia to an individual capable of independently completing a thesis. I will forever cherish the moments of thought-provoking discussions and collaboration with Catherine, when we worked together to advance the frontiers of research.

I would like to extend my heartfelt gratitude to my co-supervisor, Dr. Binyan Jiang. It was Binyan that ignited my passion for statistical research during my master's studies. Over the course of three years pursuing my doctorate, Binyan was consistently encouraging me, readily extending his assistance. I have benefited a lot from our discussions.

I would like to deliver my thank to my Ph.D. committee, Prof. Jian Huang, Prof. Xinyuan Song, and Prof. Niansheng Tang, for their profound insights and constructive suggestions that I can follow up to enhance the quality of the thesis.

I would also like to convey my appreciation to my tutor and collaborator in Bayesian statistics, Dr. Junshan Shen. During the formative stages of my doctoral journey, Junshan provided me with unwavering support in the realms of Bayesian computing and Bayesian nonparametric priors. Additionally, I am sincerely grateful to my esteemed collaborators, Drs. Zhihua Ma, Jin Yang, and Xu Zhang, with whom I shared a fruitful and enjoyable collaborative experience.

Furthermore, I extend my gratitude to Dr. Sheng Xu and Dr. Qinyi Zhang, whose profound discussions and experiences they generously shared with me have immensely enriched my knowledge. I would also like to express my profound thanks to Miss Qian Li and Mr. Zicheng Qiu for their guidance in optimization and deep learning studies. Moreover, I am truly grateful to Miss Lulu Zhang and Mr. Shouzheng Chen for their generous assistance and support throughout this journey.

I acknowledge the grants that have contributed to supporting my Ph.D. study. The research included in the thesis is partially supported by the General Research Funding (GRF) 15301519 and 15301123, Research Grants Council (RGC), UGC, Hong Kong.

I pay my tribute and appreciation to Wang Xizhi, Su Shi, and Zhao Mengfu for their wisdom and the most wonderful experience of aesthetics they bring to me.

Deep thanks are delivered to all my families. Particularly, I would like to express my sincere gratitude to my uncle for his great meticulousness and helps. I owe my deepest thanks to my parents who gave me the most invaluable education in my youth and always support me in return for nothing. Finally, no words can express my appreciation and love for my kind, loving, smart, considerate, and humorous fiancée Mengyao. If a Ph.D. degree is the greatest ideal in my past 27 years, falling in love with Mengyao is the greatest gift I obtain in my whole life. It is Mengyao who makes all my gain during this journey meaningful, just like a number one followed by a long series of zeros. I dedicate this thesis to my parents and my fiancée Mengyao.

Contents

List of Figures

List of Tables

Chapter 0

Introduction: Power of Bayesian nonparametrics

Bayesian statistics has gained arising recognition in data science as it sheds light on new approaches by the spirit of updating *prior belief* with the information from newly emerged data. The toolkits equipped with Bayesian statistics have effectively expanded people's knowledge and capability to discover the world. Inspired by the success of Bayesian approaches, in this thesis, in two classical statistical fields, from a Bayesian perspective, we succeed in proposing new models and novel methods that outperform existing approaches in comprehensive comparisons.

It is well known that, in Bayesian analysis, priors play a defining role and have a substantive impact on the final model results of estimation, inference, and prediction. Specifically, we elucidate the philosophy behind the motivation and construction of various Bayesian nonparametric prior processes for four specific questions addressed in the thesis.

0.1 Elicitation of nonparametric priors in the thesis

We give an instant introduction to some fundamental concepts of Bayesian nonparametrics first. The Bayesian nonparametric priors' origin is the need of characterizing the "uncertainty" of the *data-generating distributions* (DGD). In Bayesian statistics, the data are envisioned as random variables drawn from an unknown distribution or DGD. Consequently, one can express the information from data through an induced likelihood function. In many cases, the uncertainty of the DGD cannot be characterized by a finite number of random parameters. Rather, the DGD itself might be an *infinite dimensional parameter* randomly drawn from some functional space. For example, the infinite-dimensional parameter could be an unknown mean function without any parametric assumptions in a regression model. Then one's prior belief is imposed on the infinite-dimensional parameter itself and summarized by a *prior process*, which is a distribution on the space of random functions. Models with infinite-dimensional parameters are called *nonparametric model* and the prior processes for random functions are called *nonparametric priors*.

When people discuss prior elicitation, the following three concerns are never absent.

- **Support.** The support of a prior distribution/process is always the foremost concern. Only with a suitable domain can one properly assign the prior belief to the prior.

- **Posterior.** Priors have a profound impact on the posterior even with a pretty huge data size. In some statistical models, poorly assigned priors may even incur an improper posterior (e.g. examples in (Gelman et al., 2013, pp. 59)). Other impacts include the posterior contraction rate, posterior consistency, and the convergence of the maximum a posteriori (MAP) estimates; among others.

- **Computation.** Elicitation of nonparametric priors is often accompanied by the concern of computational feasibility. The proposed Bayesian nonparametric priors in this thesis are all computationally feasible. The reason is that they borrow strength from the stick-breaking construction (Sethuraman, 1994) so that the nonparametric priors can be well-approximated by a truncated sum of a series of products of independent random variables. Thus, the induced Markov Chain Monte Carlo (MCMC) sampling is straightforward and convenient.

In the following, we briefly introduce the nonparametric priors elicited in Chapter 1 to 4 of the thesis and describes how they meet with the above concerns.

Chapter 1: Gamma-IBP model

For change-point detection in general application scenarios, we globally model the abrupt change scheme through an infinite dimensional parameter instead of modeling a finite vector of segment parameters. Therefore, the support of the proposed nonparametric prior includes a collection of pure jump functions. Meanwhile, inspired by the *horizontal* sparsity of the jump locations (Frick et al., 2014), we observe a *vertically* nearly black (Donoho et al., 1992) nature on the vector of jump sizes and turn the change-point detection into Bayesian model selection. Thus, we call for a shrinkage nonparametric prior. Furthermore, to establish nice posterior inferential theories, we innovatively assign a Gamma hyperprior for the sparsity level within the IBP stick-breaking weights (Teh et al., 2007) for the latent indicator in the spike-and-slab type jump heights, leading to minimax optimal posterior contraction and dual posterior consistency.

Chapter 2: Signed Beta process

To detect shifts of means with imperceptible jump sizes and moderate data size, we propose a signed Beta process (SBP) for the abrupt change signal in a mixture form of two Beta processes since the jumps can either be upward or downward. We derive the stick-breaking representation of the SBP and thus, the posterior sampling is straightforward. The SBP leads to pointwisely asymptotic efficiency of the marginal MAP estimates of the abrupt change signal under the null hypothesis of no change-points.

Chapter 3: Quantile-knots I splines

Under the *unidentified nonparametric transformation model*, we compress the original support of the *transformation function* into a collection of *nonnegative monotonic functions* through an exponential transformation. On the compressed support, we assign a *weakly informative* nonparametric prior for the recast transformation so as to *facilitate posterior sampling* by controlling the posterior variance. We formulate a tuning-free quantile-knot I-splines nonparametric prior based on the empirical quantiles of survival outcomes, leading to convenient and efficient computation. We show that the posterior under the unidentified model is always proper and well-converged.

Chapter 4: ANOVA DDP

We propose a generalized shared frailty model to allow dependence among different treatment groups, where the group-wise baseline survival probability functions are dependent and unknown. The support of the nonparametric prior is composed of a collection of dependent random measures. We apply the ANOVA dependent Dirichlet process (ANOVA DDP, De Iorio et al. (2004)) to the baseline survival probability functions, where the ANOVA type dependence is imposed on the stick-breaking weights of the Dirichlet process.

0.2 Organization of the thesis

The thesis consists of TWO parts of change-point detection and survival analysis. Each part consists of TWO chapters.

Part I: Change-point detection

In Chapter 1, we propose an original and general NOn-SEgmental (NOSE) approach for the detection of multiple change-points. NOSE identifies change-points by the non-negligibility of posterior estimates of the jump heights. Alternatively, under the Bayesian paradigm, NOSE treats the step-wise signal as a *global infinite dimensional parameter* drawn from a proposed process of atomic representation, where the random jump heights determine the locations and the number of change-points simultaneously. The random jump heights are further modeled by a *Gamma-Indian buffet process shrinkage prior* under the form of discrete spike-and-slab. The induced maximum a posteriori estimates of the jump heights are consistent and enjoy a zero-diminishing false negative rate in discrimination under a 3-sigma rule. The success of NOSE is guaranteed by the posterior inferential results such as the *minimax optimality of the posterior contraction rate*, and *posterior consistency of both locations and the number of abrupt changes*. NOSE is applicable and effective to detect scale shifts, mean shifts, and structural changes in regression coefficients under linear or autoregression models. Comprehensive simulations and several real-world examples demonstrate the superiority of NOSE in detecting abrupt changes under various data settings.

Chapter 2 is motivated by the detection of multiple change-points in the *London House Index data*, where existing methods detect inconsistently and diversely owing to the relatively *small magnitudes of jump sizes*. We propose a novel *jump-size-based Bayesian approach* to address the problem, which is distinct from the mainstream methods that were developed based on modeling the locations and/or the number of change-points. We assign a nonparametric

Bayesian prior, named *signed Beta process* to model the change signal process, the maximum a posteriori estimates of which are *pointwise asymptotically efficient*. The induced posterior estimator of the jump size is *asymptotically normal* so that we can conduct a Z-type test to identify the change-point pointwisely. We have a thorough and comprehensive analysis of the proposed method for the detection of change-points in the London House Index data. Our method is not only applicable to data subject to imperceptible structural changes but also to the other common mean-shifted scenarios with either noticeable or imperceptible shifts, demonstrated by abundant experiments.

Part II: Survival analysis

In Chapter 3, we address the Bayesian prediction of survival times under a *nonparametric transformation model* (NTM). Fitting the NTM has been hampered due to the lack of identifiability. Existing approaches typically constrain the parameter space to ensure identifiability, but they incur intractable computation and cannot scale up to complex data; other approaches address the identifiability issue by making strong a *priori* assumptions on either of the nonparametric components and thus are subject to misspecifications. Utilizing a Bayesian workflow, we address the challenge by constructing new *weakly informative nonparametric priors* for infinite-dimensional parameters so as to remedy flat likelihoods associated with *unidentified models*. To facilitate the applicability of these new priors, we subtly impose an exponential transformation on top of NTMs, which compresses the space of infinite-dimensional parameters to positive quadrants while maintaining interpretability. Simulations reveal that our method is robust and outperforms the competing methods. Applications in several real datasets demonstrate the superior predictive capability of the proposed method.

In Chapter 4, we aim to display the latest tendency in Bayesian computing, in the sense of automating the posterior sampling, through Bayesian analysis of survival modeling for multivariate survival outcomes with a complicated data structure. Motivated by relaxing the strong assumption of proportionality and the restriction of a common baseline population, we propose a generalized shared frailty model which includes both parametric and nonparametric frailty random effects so as to incorporate both treatment-wise and temporal variation for multiple events. We develop a survival-function version of ANOVA dependent Dirichlet process to model the dependency among baseline survival functions. The posterior sampling is implemented by the No-U-Turn sampler in Stan, a contemporary Bayesian computing tool, automatically. The proposed model is validated by analysis of the bladder cancer recurrences data. The estimation is consistent with existing results. Our model and Bayesian inference provide evidence that the Bayesian paradigm fosters complex modeling and feasible computing in survival analysis and Stan relaxes the posterior inference.

0.3 Publication status

The project in Chapter 1 is under review at *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*. The preprint is available at arXiv:2209.14995v2 (Chong ZHONG, Zhihua MA, Xu ZHANG, and Catherine LIU). I was the lead investigator, responsible for all major areas of concept formation, methodology, mathematical proofs, and manuscript composition. Zhihua MA contributed the 3-sigma discriminant criterion and the case of scale-shift in real data analysis. Xu ZHANG proved the theorem on truncation asymptotic equivalence. Catherine LIU is the supervisory author on this project and was involved throughout the project.

The project in Chapter 2 was submitted to *Bayesian Analysis* on July 18, 2023. This is joint work with Zhihua MA, Junshan SHEN, and Catherine LIU. Zhihua Ma contributed work in the early stage. Junshan Shen contributed to the idea of the signed Beta process. I contributed to the hypothesis testing procedure, simulations, and real data analysis, and was responsible for the mathematical proof and manuscript composition. Catherine LIU is the supervisory author on this project and was involved throughout the project.

The project in Chapter 3 was once submitted to *Journal of the American Statistical Association* on May 19, 2022, rejected and revised, pending submission again. The preprint is available at arXiv:2205.14504v4 (Chong ZHONG, Junshan SHEN, Jin YANG, Catherine LIU, and Zhaohai LI). I was the lead investigator, responsible for all major areas of Bayesian modeling, methodology, mathematical proof, data analysis, and manuscript composition. Jin YANG contributed equally with me in the data analysis part. Junshan SHEN monitored the nonparametric Bayesian seminar. Zhaohai LI participated in the latter half of the manuscript composition.

Catherine LIU is the supervisory author on this project and was involved throughout the project.

A version of Chapter 4 has been published in IntechOpen *Computational Statistics and Applications* (Chapter 5, link https://www.intechopen.com/chapters/79845, by far reached 152 download, Chong ZHONG, Zhihua MA, Junshan SHEN, and Catherine LIU). I was the lead investigator, responsible for all major areas of idea, modeling, methodology, data analysis, as well as manuscript composition. Zhihua MA contributed to the NIMBLE code demo and participated in the discussion of NIMBLE and Stan. Junshan SHEN monitored related seminars on the dependent Dirichlet Process. Catherine LIU is the supervisory author on this project and was involved throughout the project.

Part I

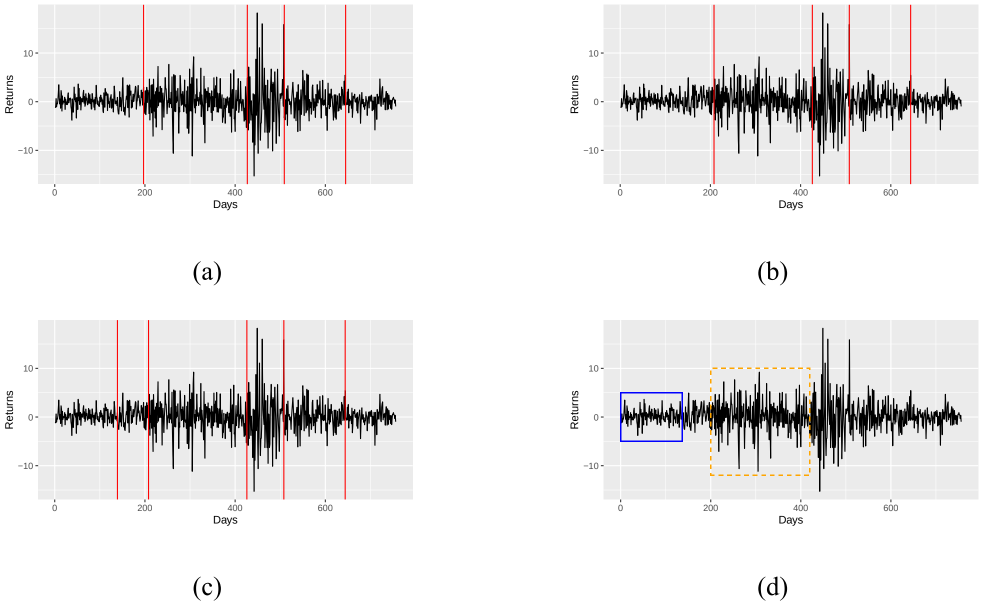
Chapter 1

Non-segmental Bayesian detection of multiple change-points

1.1 Introduction

Detection of multiple change-points has long been an active research topic with a broad range of applications in economics, health study, genetics, and finance, to name a few. The change detection is needy in cases with mean shifts (Frick et al. (2014); Fryzlewicz (2014); Du et al. (2016); Romano et al. (2022); among others), scale shifts (Killick et al. (2012); Haynes et al. (2017); among others), and structural abrupt changes in regression models (Bai and Perron (2003); Korkas and Pryzlewiczv (2017); Baranowski et al. (2019); among others). Since the abrupt change pattern used to be mathematically expressed as a stepwise function or sum of segment-wise functions, existing methods incline to study segmental parameters such as piecewise mean parameters and segment-wise log-likelihood ratios to unveil the changes such as the number, locations, and jump sizes. In this chapter, we attempt to propose an original and general procedure of change-point detection under a novel NOn-SEgmental (NOSE) spirit which models the pure jump process of the change mechanism by a *global infinite-dimensional parameter*.

Our approach is motivated by a suspected change-point under-discrimination case arising from asset pricing and portfolio management. Specifically, we look into the US log daily returns of agriculture industry portfolios (DRAIP) from January 2007 to December 2019, available at http://mba.tuck.dartmouth.edu. Understanding the shifts on the scale of the recast daily return data can help evaluate the risk of investment on these portfolios since the variation of daily returns usually acts as a measure of the risk of a portfolio. The DRAIP dataset is displayed as a black line in Figure 1.1. One can observe noticeably that, i) the data have no shifts on the mean since all data are centered around zero stably; ii) the variations of daily returns have uneven shifts, most of which are modest except the apparent variation on time interval $(400, 500)$. Existing methods such as NOT (Baranowski et al., 2019), SMUCE (Frick et al., 2014), and PELT (Killick et al., 2012) can work on this dataset to detect scale changes, summarized in Figures 1.1(a)-1.1(c). The numbers of change-points detected are 4, 4, and 5, respectively. *Nonetheless*, one may suspect the possibility of under-detection of change-points for areas highlighted in, a) the orange rectangle between (200, 400) that is bouncing-visible and b) the blue rectangle between (0, 200) that is bouncing-mild. Note that the aforementioned methods share the same spirit of modeling the *local segment parameters* directly, and may lose the structural information. Instead, we are driven to formulate a *global process* for the underneath abrupt change mechanism to discover the possible changes. Our approach is introduced in subsections 1.1.1-1.1.3.



(a)

(b)

(c)

(d)

Figure 1.1: Plots of estimated locations of change-points (in red vertical lines) by different methods and DRAIP data (in black lines). (a), SMUCE; (b), NOT; (c), PELT; (d) original data.

1.1.1 Global curve function parameter $\theta(t)$

The abrupt change, in almost all literature, is characterized as a *pure jump process* $\sum_{k=1}^{K+1} \theta_k$ $I(\tau_{k-1} \leq t < \tau_k)$, and have been dealt with by focusing on segment parameters θ_k directly. Here K denotes the unknown total number of change-points, τ_k denotes the k-th change-point, and the argument t is defined on a state space $\mathcal{T}$ that is not limited to a temporal or spatial state. Let $\boldsymbol{\tau}_{1:K} = \{\tau_1, \dots, \tau_K\}$, where τ can be a placeholder. We assume that the adjacent θ_k's are distinguishable in the sense that $\theta_k \neq \theta_{k+1}$ for all $1 \leq k \leq K$. Rather than looking into local segmental parameters θ_k, we globally denote the pure jump process or the stepwise function as $\theta(t)$. Consequently, our approach starts from an atomic representation of the curve function $\theta(t)$ from the perspective of jump sizes and locations of change-points.

Let $(h_1, \xi_1), (h_2, \xi_2), \dots$ be a countably infinite collection of atoms and heights at locations. A draw of an atomic random measure is written as

$$q(\cdot) \equiv \sum_{\ell=1}^{\infty} h_\ell \delta_{\xi_\ell}(\cdot), \tag{1.1}$$

where δ_{ξ_ℓ} is an atom at ξ_ℓ with h_ℓ being its height of the jump in q. Then, we propose a prior process $\boldsymbol{Q}$ for $\theta(t)$ in the form of the cumulative integral of q

$$\theta(t) \sim \boldsymbol{Q} \equiv \int_{-\infty}^{t} q(u) du = \sum_{\ell=1}^{\infty} h_\ell I(\xi_\ell \leq t). \tag{1.2}$$

As the jumps may be downward or upward, the jump sizes $h_\ell \in \mathbb{R}$ are allowed to be *sign-varying* and may be *dependent* rather than being *non-negative* and *independent* in the atomic representation in a completely random measure (Kingman, 1967).

Since those jumps with negligible heights are not considered to be abrupt changes, one may approximate the prior process $\boldsymbol{Q}$ in a truncation form $\boldsymbol{Q}^L$,

$$\boldsymbol{Q}^L = \int_{-\infty}^{t} q^L(u) du = \sum_{\ell=1}^{L} h_\ell I(\xi_\ell \leq t) \text{ with } q^L = \sum_{\ell=1}^{L} h_\ell \delta_{\xi_\ell}. \tag{1.3}$$

In practice, one may assume the number of change-points K is bounded by some sufficiently large number L, say, $L = [n/D]$, the integer part of the ratio between the number of observations

n and D. Here D reflects one's prior belief on the minimum distance between any two adjacent change-points. For example, the PELT method sets the default minimum segment length as $D = 2$ in the R package `changepoint` (Killick and Eckley, 2014). In Theorem 1.4 of Section 1.3, we will state the asymptotic equivalence of the truncation form (1.3) to the atomic expression (1.2) under the Gamma-IBP prior model proposed in (1.5).

1.1.2 Shrinkage prior for $\theta(t)$

Let $\theta(t) \equiv \theta$. The underlying distribution for drawing a sample sequence $\boldsymbol{y} = (y_1, \ldots, y_n)$ is denoted by $f(\cdot|\theta, \boldsymbol{\gamma})$, where θ is the abrupt change parameter that determines the abrupt changes and $\boldsymbol{\gamma}$ is the nuisance parameters that does not contribute to the abrupt change mechanism. Suppose that the n samples $\boldsymbol{y}$ are observed at $\boldsymbol{t}_{1:n}$. Then the likelihood is

$$\boldsymbol{l}(\boldsymbol{y}|\theta, \boldsymbol{\gamma}) = \prod_{i=1}^{n} f(y_i|\theta(t_i), \boldsymbol{\gamma}).$$

This brings us to the posterior estimate of $\theta(t)$ under prior (1.3). Once we obtain a posterior estimate based on the observed data $\boldsymbol{y}$, we immediately have the increments of $\theta(t)$ between t_i and t_{i+1}, denoted as $d_i = \theta(t_{i+1}) - \theta(t_i)$. The increment sequence d_i acts as a KEY signal of change-points in our methodology: clearly, the jump height vector $\boldsymbol{d} = (d_1, \ldots, d_{n-1})$ represents the jump heights/sizes at all states. Thus, those locations with non-negligible jump sizes are naturally segregated from those ignorable and thus, identified as change-points. Consequently, we tend to employ posterior estimates of d_i sequence as the features to discriminate change-points based on some criterion rule that will be presented in subsection 1.1.3.

Note that drawing a random trajectory of $\theta(t)$ is equivalent to randomly drawing vectors $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_L)$ and $\boldsymbol{h} = (h_1, \ldots, h_L)$. Since $\boldsymbol{h}$ are heights of atoms at $\boldsymbol{\xi}$, we sample $\boldsymbol{\xi}$ first and then sample $\boldsymbol{h}$, and randomly assign $\boldsymbol{h}$ to the atoms. Since one can only observe $\boldsymbol{y}$ at discrete states $\boldsymbol{t}_{1:n}$, it is meaningless to assume that the change-points take place between two adjacent data points. Hence, we assume that all jumps of $\theta(t)$ only take place on $t_i, i = 1, \ldots, (n-1)$ without loss of generality (the last data point is omitted as a change-point). Then the prior for atoms ξ_ℓ is naturally defined as

$$\xi_1 \sim U(\boldsymbol{t}_{1:(n-1)}), \quad \xi_\ell | \xi_1, \ldots, \xi_{\ell-1} \sim U(\boldsymbol{t}_{1:(n-1)} \setminus \boldsymbol{\xi}_{1:(\ell-1)}),\ \ell \geq 2, \tag{1.4}$$

where $Z \setminus A$ denotes the complement of set A given the universe Z. In other words, ξ_ℓ are sampled from $\boldsymbol{t}_{1:(n-1)}$ uniformly without replacement. As a result, $\boldsymbol{\xi}$ is just a subset of $\boldsymbol{t}_{1:(n-1)}$ for any $L < (n-1)$.

Note that under prior (1.4), $\boldsymbol{h}$ is a subset of $\boldsymbol{d}$ containing all non-zero entries of $\boldsymbol{d}$. Hence we will discuss the sparseness of the jump height vector $\boldsymbol{d}$ before the prior elicitation of $\boldsymbol{h}$.

Nearly black vector: K_n-sparsity

In general, we allow the number of change-points K to be arbitrarily large but require $K << n$ as $n \to \infty$. One may select a sufficiently large truncation number L so that $K << L$ too. Then the jump height vector $\boldsymbol{d}$ belongs to $l_0[K_n]$, a class of *nearly black vectors* (Donoho et al. (1992); Castillo and van der Vaart (2012)), explicitly expressed as

$$l_0[K_n] = \{\boldsymbol{v} \in \mathbb{R}^p : \sum_{i=1}^{p} I(|v_i| > 0) \leq K_n\},$$

where v_i is the ith entry of $\boldsymbol{v}$ and $K_n (\geq K)$ is a given integer so that $K_n = o(L)$, as $n, L \to \infty$. We call that $\boldsymbol{d}$ possesses K_n-sparsity. Note that $\boldsymbol{h}$ is also K_n-sparse since $\boldsymbol{d}$ and $\boldsymbol{h}$ share the same cardinality.

Under the above K_n sparsity, we transfer change-point detection to searching for a sparse posterior solution to the jump height vector $\boldsymbol{d}$ and $\boldsymbol{h}$. Therefore, we will introduce next a shrinkage prior for the random vector $\boldsymbol{h}$ in model (1.5). Our K_n-sparsity is inspired by the "horizontal" sparsity of the vector of jump locations in Frick et al. (2014, subsection 6.3) under Gaussian linear models, though we take a "vertical" view on the jump heights instead. By penalizing the number of change points, the SMUCE method by Frick, Munk, and Sieling attains a minimax optimal rate up to a logarithm term on the distance between locations of true and estimated change-points; by a constructed shrinkage prior, our proposed NOSE achieves the minimax optimal posterior contraction rate over the $l_0[K_n]$ class within the Bayesian context. Nonetheless, these two different kinds of views on sparsity lead to different estimation procedures and consistency. SMUCE has to estimate the number and locations of change-points sequentially and obtains the consistency of the number of change-points only. In contrast, NOSE estimates the number and the locations of change-points *simultaneously* because, under the jump-size-weighted atomic representation (1.3), a non-negligible jump size certainly indicates a change-point. As a result, NOSE achieves consistency of both the number and locations of change-points.

Prior for h: Gamma-IBP model

The prior for $\boldsymbol{h}$ is expressible as follows.

$$\begin{aligned} &h_\ell | Z_\ell \sim (1 - Z_\ell)\delta_0 + Z_\ell F_0, \ F_0 = \text{Laplace}(0, \lambda), \\ &Z_\ell | \eta_\ell \sim \text{Bernoulli}(\eta_\ell), \ \eta_\ell = \prod_{j=1}^{\ell} p_j, \ p_j | \alpha \sim \text{Beta}(\alpha, 1), \ \alpha \sim \text{Gamma}(a, b), \end{aligned} \tag{1.5}$$

where Z_ℓ are latent binary variables determined by the sparsity parameters η_ℓ, δ_0 denotes the mass function at 0, Laplace$(0, \lambda)$ represents a zero-centered Laplace distribution with precision parameter λ, and Gamma(a, b) represents the Gamma distribution with density $\{\Gamma(a)b^a\}^{-1}x^{a-1}$ exp$(-x/b)$. Prior (1.5) is a special class of *discrete spike-and-slab prior* with a surely-zero spike δ_0 and a Laplace slab F_0. Specifically, the sparsity parameters η_ℓ are exponentially decreasing products of a series of Beta variables with a mass parameter α, which is modeled by a Gamma hyperprior for the purpose of dominating the whole sparsity of prior (1.5). Consequently, $\boldsymbol{Z} = (Z_1, \ldots, Z_L)$ can be viewed as a stick-breaking representation of an L-truncated single row in the Indian buffet process (IBP) (Teh et al., 2007). Therefore, prior $\boldsymbol{h}$ is named as the Gamma-IBP model hereafter.

The nest of the IBP construction and the Gamma hyperprior results in a *strict exponential decrease* on the dimensionality $|\boldsymbol{Z}|$, and maintains sufficient weight on the true sparsity level K_n. Therefore, it suffices to reach the minimax optimal posterior contraction rate (Castillo and van der Vaart, 2012). On the other hand, the IBP construction further controls the tail probability $Pr\{|\boldsymbol{Z}| > k\}$ for any $k > 0$, and hence, obtains consistent posterior model selection with a smaller cut-off compared to Castillo et al. (2015). The detailed justifications and results are summarized in Section 1.2.

1.1.3 Discrimination of change-points

After the prior elicitation in subsection 1.1.2, we propose a change-point discrimination procedure based on the induced posterior. We first obtain posterior estimates of the increments $\boldsymbol{d}$ and then simply compare the value of the estimates with some data-driven threshold. Under the priors (1.4) and (1.5), the posterior of $\boldsymbol{\xi}$ and $\boldsymbol{h}$ are sampled through Markov Chain Monte Carlo (MCMC). Suppose one has drawn N posterior samples of $\boldsymbol{h}$ and $\boldsymbol{\xi}$, denoted as ${}^j h_\ell$ and ${}^j\xi_\ell, j = 1, \ldots, N$. Then for any t_i, the marginal posterior samples of $\theta(t_i)$ are determined as ${}^j\theta(t_i) = \sum_{\ell=1}^{L} {}^j h_\ell I({}^j\xi_\ell \leq t_i)$.

With N marginal posterior samples of $\theta(t_i)$, one can approximate the maximum of a posteriori (MAP) estimate of $\theta(t_i)$ as the mode of sample density of $\{{}^j\theta(t_i)\}_{j=1}^{N}$, denoted as $\hat{\theta}(t_i)^{\mathrm{MAP}}$. Let $\{\zeta_i\}_{i=2}^{n}$ be

$$\zeta_i = \hat{\theta}(t_{i+1})^{\mathrm{MAP}} - \hat{\theta}(t_i)^{\mathrm{MAP}}, i = 1, \ldots, (n-1),$$

the diffed series of $\hat{\theta}(t_i)^{\mathrm{MAP}}$. Note that ζ_i is a posterior estimate of d_i i.e. a posterior estimate of the jump size at t_i. Nevertheless, ζ_i is not the MAP estimate $\hat{d}_i^{\mathrm{MAP}} = \{\widehat{\theta(t_{i+1}) - \theta(t_i)}\}^{\mathrm{MAP}}$ but an approximation to $\hat{d}_i^{\mathrm{MAP}}$ in practice. The reason why we do not employ $\hat{d}_i^{\mathrm{MAP}}$ directly is that the marginal posterior of d_i is poorly approximated by MCMC samples due to high auto-correlation between samples of ${}^j d_i = \{{}^j\theta(t_{i+1}) -^j \theta(t_i)\}, j = 1, \ldots, N$. Therefore, the density of d_i estimated from MCMC samples of $\theta(t_i)$ is useless and so is the mode. Let $\hat{\sigma} \equiv (\mathrm{Var}\{\zeta_i\}_{i=1}^{n-1})^{1/2}$ be the sample standard deviation of $\{\zeta_i\}_{i=2}^{n}$. Then we determine change-point locations $\tau_k, k \in 1, \ldots, K$ based on the following discrimination rule.

Discrimination rule

3-sigma *If at t_i, the absolute posterior estimate of jump size $|\zeta_i| > 3\hat{\sigma}$, then t_i is discriminated as a change-point; otherwise, not a change-point.*

It is intuitive to employ the above 3-sigma rule for change-point discrimination due to the nearly black nature of $\boldsymbol{d}$. The 3-sigma rule has been widely used in outlier detection (Pukelsheim, 1994), where the outliers are considered to be far away from the center of the population. In our case, the nearly black $\boldsymbol{d}$ indicates that the population of ζ_i concentrates at zero except for some outliers. Hence, those points that are sufficiently far away from zero are naturally discriminated as outliers, i.e. change-points.

The threshold for negligibility takes the value $3\hat{\sigma}$. It is a kind of "global" threshold based on all entries of the posterior estimates of vector $\boldsymbol{d}$. In existing approaches, most thresholds for spike-and-slab priors are "local". Some local thresholds shrink those coordinates whose posterior estimates are under some prespecified values to zero (Pati et al. (2014); Ročková and George (2016); Ročková (2018); among others), and the others shrink those coordinates whose posterior non-zero probability is smaller than 0.5 (Barbieri and Berger (2004); Scheipl et al. (2012); Cappello et al. (2023); among others). However, a local threshold may be sensitive to the ratio between jump sizes and within-segment variations in our numerical experience. The 3-sigma global criterion grants us a strong ability to recognize those even small jump sizes since each jump size is compared with the vast majority of zeros on stationary points, regardless of the within-segment variations. Under the 3-sigma rule, we show the near zero false negative rate of discrimination; see Corollary 1.3 in Section 1.2.

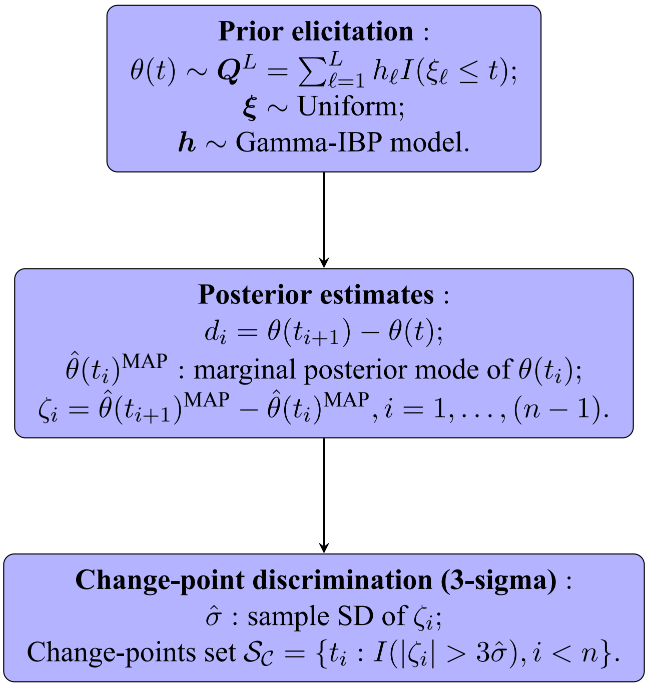

Prior elicitation :
$\theta(t) \sim \boldsymbol{Q}^L = \sum_{\ell=1}^{L} h_\ell I(\xi_\ell \leq t)$;
$\boldsymbol{\xi} \sim$ Uniform;
$\boldsymbol{h} \sim$ Gamma-IBP model.
Posterior estimates :
$d_i = \theta(t_{i+1}) - \theta(t)$;
$\hat{\theta}(t_i)^{\text{MAP}}$: marginal posterior mode of $\theta(t_i)$;
$\zeta_i = \hat{\theta}(t_{i+1})^{\text{MAP}} - \hat{\theta}(t_i)^{\text{MAP}}, i = 1, \ldots, (n-1)$.
Change-point discrimination (3-sigma) :
$\hat{\sigma}$: sample SD of ζ_i;
Change-points set $\mathcal{S}_{\mathcal{C}} = \{t_i : I(|\zeta_i| > 3\hat{\sigma}), i < n\}$.

Figure 1.2: Flowchart of the proposed methodology.

We provide an overview of the workflow of the proposed change-point detection method in Figure 1.2 and summarize it as follows.

Step 1: construct a truncated prior for $\theta(t)$ in the form of (1.3). Assign priors (1.4) and (1.5) to $\boldsymbol{\xi}$ and $\boldsymbol{h}$, respectively.

Step 2: draw N posterior samples of $\boldsymbol{\xi}$ and $\boldsymbol{h}$. Obtain the marginal MAP estimate of $\theta(t)$ as $\hat{\theta}(t_i)^{\text{MAP}} = \arg\max_x f_i(x)$, where f_i is the empirical density of ${}^j\theta(t_i) = \sum_{\ell=1}^{L} {}^j h_\ell I({}^j\xi_\ell \leq t_i)$, $j = 1, \ldots, N$, $i = 1, \ldots, n$.

Step 3: obtain $\zeta_i = \hat{\theta}(t_{i+1})^{\text{MAP}} - \hat{\theta}(t_i)^{\text{MAP}}$ as an estimate of d_i. The set of discriminated change-points is $\mathcal{S}_{\mathcal{C}} = \{t_i : I(|\zeta_i| > 3\hat{\sigma}), i < n\}$.

1.1.4 Application scenarios

We illustrate some application scenarios of the proposed method here. NOSE works in the detection of mean shifts and scale shifts such as,

Scenario 1: shifts in means of Gaussian variables (Gaussian mean-shifted model). We have a series of real observations $y_i \sim N\{\theta(t_i), \sigma^2\}$, for $i = 1, \ldots, n$. The global parameter $\theta(t)$ represents the location parameter.

Scenario 2: shifts in the parameter of Poisson variables. We have a series of integer observations $y_i \sim \text{Poisson}\{\theta(t_i)\}$, for $i = 1, \ldots, n$. The global parameter $\theta(t)$ characterizes the changes in mean and variance simultaneously.

Scenario 3: shifts in the scale parameters of Gaussian variables (Gaussian scale-shifted model). We have a series of real observations $y_i \sim N\{\mu, \exp[\theta(t_i)]\}$, for $i = 1, \ldots, n$. The global parameter $\theta(t)$ represents the scale parameter through an exponential transformation to guarantee the non-negativity.

Meanwhile, NOSE is also applicable to detect structural changes in regression/autoregression models.

Scenario 4: structural changes of an AR(1) model. Data are generated from the model

$$y_t = \phi_0 + \theta(t)y_{t-1} + \epsilon_t,$$

where ϕ_0 is the fixed intercept, $E(\epsilon_t) = 0$ and $E(\epsilon_t\epsilon_s) = \sigma^2 I(t = s)$. The global parameter $\theta(t)$ represents the autocorrelation coefficient.

Scenario 5: structural changes of a linear regression model. Data are recorded as independent pairs of (y_{tj}, X_{tj}), for $j = 1, \ldots, n_t, t = 1, \ldots, T$. The association between y and X is characterized by

$$y_{tj} = \beta_0 + \theta(t)X_{tj} + \epsilon_{tj},$$

where β_0 is a fixed intercept, $E(\epsilon_{tj}) = 0$ and $E(\epsilon_{tj}\epsilon_{sj'}) = \sigma^2 I(t = s)$. The global parameter $\theta(t)$ represents the regression coefficient at time t. Note that by taking $n_t = 1$ for all t and $X_t = y_{t-1}$, this scenario reduces to Scenario 4.

1.1.5 Related work

Review on segmental approaches

As we state at the very beginning, most existing methods of change-point detection are segmental approaches in the sense that they *estimate multiple segment parameters or conduct a series of tests based on segment parameters*. One may summarize them into two main streams.

i) Penalized methods. *Penalized methods optimize an objective function in the sum of segment-specific costs and a penalty*. The cost is versatile and chosen based on types of changes (mean, scale, or autocorrelation for instance) while the penalty term is deterministic to the methodology. For the penalty term, linear l_0 penalization to the *vector of segment parameters/features* to control *the number of change-points* might be the most popular choice (Yao (1984); Killick et al. (2012); Frick et al. (2014); Romano et al. (2022); Jula Vanegas et al. (2021); among others). Alternatively, l_1 penalization to the *vector of segment parameters/features and their jump sizes* is also considered (Tibshirani et al. (2005); Chernozhukov et al. (2017); among others). We note that Bayesian approaches can be attributed to penalized methods in the sense that one employs priors to automatically penalize the number of change-points (Fearnhead (2006); Wyse et al. (2011); Ko et al. (2015); among others), or even cover ratios between observations in segments and total sample size (Du et al., 2016).

ii) Binary-segmentation (BS) variants. The BS procedure involves the sequential partitioning of a given data stream into two distinct subsegments (Vostrikova, 1981). This partitioning is carried out based on the identification of a change-point, which is determined by applying specific testing criteria to the previously split subsegments. Under this spirit, Fryzlewicz (2014) developed the so-called "bottom-up" strategy in the sense that one *determines a change-point from subsets of the data (local ground) and then aggregates local features* as the overall model. Baranowski et al. (2019) further enhanced the "bottom-up" strategy by a narrowest over threshold (NOT) so that they draw the subsample set from the narrowest interval. There are some other BS variants works such as Cho and Fryzlewicz (2015), Fryzlewicz (2018), Fang et al. (2020); among others.

Spike-and-slab prior revisit

The spike-and-slab priors are usually categorized as continuous and discrete priors. The continuous spike-and-slab employs two continuous densities for both spike and slab terms, with one highly concentrated and the other dispersed (Carlstein et al. (1988); Narisetty and He (2014); Hahn and Carvalho (2015); among others). It is convenient in MCMC sampling, while the posterior solution may not provide sparse estimates automatically. The discrete spike-and-slab priors (Yen (2011); Yang et al. (2016); Shin and Liu (2021); Ray and Szabó (2022); among others) have great progress in recent years from the computational aspect. Under a special Gaussian sequence model, Castillo and van der Vaart (2012) establishes the conditions for the minimax optimal posterior contraction rate with discrete spike-and-slab priors while remaining consistent model selection unsolved. Conditions for consistent posterior model selection with discrete spike-and-slab priors are given by Castillo et al. (2015), while the posterior contraction is not optimal. With a data-dependent slab term, Martin et al. (2017) obtains both minimax optimality and model selection consistency under an empirical Bayes approach.

Most of the existing work for discrete spike-and-slab priors considers i.i.d. sparsity parameters. In this chapter, our discrete spike-and-slab prior is coupled with dynamic IBP stick-breaking weights. Such kind of dynamic spike-and-slab prior was first employed by (Williamson et al., 2010) for topic modeling. It has been extended to factor models with possibly infinite many factors (Knowles and Ghahramani (2011); Ročková and George (2016); James (2017); Ma and Liu (2022); Ohn and Kim (2022); among others). We are the first to employ the IBP discrete spike-and-slab to change-point detection, unlike existing work that employs continuous spike-and-slab prior with invariant sparsity parameter (Cappello et al., 2023).

The rest of this chapter is organized as follows. Section 1.2 studies the asymptotic behavior of the posterior and detection performance. Section 1.3 provides technical details of the Bayesian implementation of our method. Sections 1.4 and 1.5 present comprehensive simulations and applications to extensive real-world data examples, followed by a brief discussion in Section 1.6. Mathematical proofs and results of additional simulations are included in Supplementary materials. The companion `R` package NOSE is available online.

1.2 Asymptotic behavior of posterior

In this section, we present the theoretical results of the proposed change-point detection method in the asymptotic regime $n, L \to \infty$. We confine our theoretical results in Scenario 1 in subsection 1.1.4, the Gaussian mean-shifted model with invariant variance. Such a scenario is the most common case studied by existing change-point literature, where the invariant variance assumption is also required (Fryzlewicz (2014); Du et al. (2016); Baranowski et al. (2019); among others).

As we mentioned before, the jump height vector $\boldsymbol{d}$ contains all information about the jump sizes, which are deterministic in our approach. Therefore, we will focus on the posterior of $\boldsymbol{d}$. We study THREE aspects of asymptotic behaviors, $\mathbf{1})$ minimax optimal posterior contraction rate and recovery with under detection, $\mathbf{2})$ posterior consistency of model selection, and $\mathbf{3})$ asymptotic zero false negative rate of change-point discrimination under the 3-sigma rule.

From our insight, given the scale parameter σ in Scenario 1, the Gaussian mean-shifted model can be rewritten as a Gaussian sequence model (Castillo and van der Vaart, 2012). Without loss of generality, we assume $\sigma = (\sqrt{2})^{-1}$. If not, one can simply transform the data and will not change the results. Let $\boldsymbol{y}^*$ be the differed series of $\boldsymbol{y}$, so that $y_i^* = y_{i+1} - y_i$ for $i = 1, \ldots, n-1$. Then we obtain the following Gaussian sequence model

$$y_i^* \sim N\left(d_i, 1\right), i = 1, \ldots, (n-1). \tag{1.6}$$

Our theoretical results are given under model (1.6).

Notation

Let $p = n - 1$ and $\boldsymbol{d}_0 = (d_{01}, \ldots, d_{0p})^T$ be the "true" jump height vector. We shall assume that the $\boldsymbol{d}_0 \in l_0[K_n]$ for some given number K_n such that the number of change-points $K \leq K_n$. Since the specification of L depends on n or p, we use L_n in this section. Hereafter, let $\Pi_{n,L_n}(\mathcal{B}|\boldsymbol{y}^*)$ denotes the posterior probability on a Borel set $\mathcal{B}$ under priors (1.4) and (1.5) given data $\boldsymbol{y}^*$. Let $P_{\boldsymbol{d}_0}$ and $E_{\boldsymbol{d}_0}$ denote the probability measure and the expectation operator under the law $N(\boldsymbol{d}_0, I_p)$, respectively.

1.2.1 Posterior contraction

We first give asymptotic results on the posterior contraction of the jump height vector $\boldsymbol{d}$. This contraction rate evaluates the capability that the posterior recover the true jump height vector $\boldsymbol{d}$. We have the following assumption about $n = p + 1$, L_n, and K_n.

(A1) $L_n < p$; $K_n/L_n \to 0$, as $L_n \to \infty$.

By selecting $L_n = [n/D]$, where $D > 1$ is some fixed constant, Assumption (A1) is satisfied as $K_n/n \to 0$, which is a common setting in both high-dimensional regression and change-point literature.

The posterior contraction rate is the rate that the most mass of the posterior concentrates around a ball of the true vector $\boldsymbol{d}_0$. In this chapter, we define the radius of the ball by the following l^q losses (Castillo and van der Vaart, 2012)

$$d_q(\boldsymbol{d}, \boldsymbol{d}_0) = \sum_{i=1}^{p} |d_i - d_{0i}|^q.$$

For $q \in (0, 2]$, Donoho et al. (1992) shows that the minimax optimal rate over $l_0[K_n]$ is

$$r_n^* = K_n \log^{q/2}(p/K_n).$$

The following theorem gives the posterior contraction rate of $\boldsymbol{d}$, which reaches the minimax optimal rate under l^q metrics.

Theorem 1.1 (Minimax optimal posterior contraction rate)**.** *Let $a = c_1 L_n^{-c_3}, b = c_2 L_n^{c_4}$ for some constants $c_1, c_2 > 0$ and $c_3 > c_4 + 1 \geq 2$ in prior* (1.5)*. Under Assumption (A1), as $n, L_n, K_n \to \infty$, for a sufficiently large constant M, we have*

$$\sup_{\boldsymbol{d}_0 \in l_0[K_n]} E_{\boldsymbol{d}_0} \Pi_{n,L_n}\{\boldsymbol{d} : d_q(\boldsymbol{d}, \boldsymbol{d}_0) > M r_n^q K_n^{1-/q/2} | \boldsymbol{y}^*\} \to 0,$$

where $r_n \geq \sqrt{K_n \log(L_n/K_n)}$.

It clearly finds that for $q \in (0, 2]$, the posterior contraction rate given by Theorem 1.1 is at the same order of the minimax optimal rate r_n^*. This result is similar to Castillo and van der Vaart (2012, Theorem 2.2), though the Gamma-IBP model in (1.5) does not belong to any examples studied by them. Actually, the nest form of the IBP prior and the Gamma hyperprior plays a key role in the establishment of Theorem 1.1. As shown by Teh et al. (2007, subsection 3.1), with a fixed α, as the truncation number $L_n \to \infty$, η_ℓ become the order statistics of Beta$(\alpha/L_n, 1)$ and hence, the distribution of the cardinality of the latent indicator $\boldsymbol{Z}$ converges to Poisson(α). With the Gamma hyperprior for α, the whole prior for $\boldsymbol{d}$ can be approximated by a Poisson-Gamma model and hence has strict exponential decrease (Castillo and van der Vaart, 2012, Example 2.3). The choices of hyperparameter (a, b) are also essential but not too strict. On one hand, the relatively large choice of b in the Gamma hyperprior further grants sufficient weight on the true sparsity level K_n so that the posterior can contract in an optimal rate. On the other hand, the very small choice of a makes the Gamma-IBP model sufficiently close to the approximated Poisson-Gamma model. We defer the detailed proof to Supplement 1.7.1.1. Note that we only require the first moment of the Gamma hyperprior $ab = o(L_n^{-1})$ here. In practice, one may allow $ab^2 \to \infty$ as $n, L_n \to \infty$ and hence obtain a very flat Gamma prior which is nearly “noninformative” or “objective”.

Theorem 1.1 requires that $K_n \to \infty$, which is not a common pattern in change-point problems. In the most existing literature, the number of change-points is assumed to be arbitrarily large but finite (Frick et al. (2014); Du et al. (2016); Baranowski et al. (2019); Romano et al. (2022); among others). To this end, in the following, we study the posterior behavior with a finite K_n and set the true number of change-points $K = K_n$. That is, equivalently, the cardinality of the true jump height vector is $|\boldsymbol{d}_0| = K_n$.

The following theorem tells the posterior contraction rate with under detection of change-points for any $K_n < L_n/2$.

Theorem 1.2 (Recovery with under selection)**.** *Under conditions in Theorem 1.1, for $M \geq 10$ and any fixed $K_n < L_n/2$, as $n, L_n \to \infty$, we have*

$$\sup_{\boldsymbol{d}_0 \in l_0[K_n]} E_{\boldsymbol{d}_0} \Pi_{n,L_n}\{d_1(\boldsymbol{d}, \boldsymbol{d}_0) > Mr_n, |\boldsymbol{d}| \leq K_n | \boldsymbol{y}^*\} \to 0.$$

Theorem 1.2 is a direct result of Proposition 5.1 in Castillo and van der Vaart (2012) by taking $A = 1$. By fact that $\binom{L_n}{K_n} \leq (eL_n/K_n)^{K_n} \leq \exp(cr_n^2)$ for some sufficiently large constant c, the right hand side of Proposition 5.1 in Castillo and van der Vaart (2012) tends to zero and hence, Theorem 1.2 holds. The detailed proof is deferred to Castillo and van der Vaart (2012, Section 5).

1.2.2 Posterior consistency of model selection

From the perspective of change-points detection, the model selection corresponds to the capability of correctly detecting the number of change-points, the foremost concern in change-point detection. As mentioned before, our approach distinguishes non-negligible jumps from those zero or near zero. Actually, those too close to zero jumps cannot be detected by any method. Hence, it is necessary to determine a “sufficiently small ” cut-off of non-negligible jump sizes i.e. the non-negligible entries of the true jump height vector $\boldsymbol{d}_0$. Let $S_0 = \{i : |d_{0i}| \neq 0\}$ be the support of non-zero coordinates of $\boldsymbol{d}_0$ and S_0^c be the support of other zero coordinates. In our change-point context, $S_0 = \boldsymbol{\tau}_{1:K_n}$. Let $S = \{i : |d_i| \neq 0\}$ be the support of non-zero coordinates of $\boldsymbol{d}$. Hence, we will study the model selection result on the following class of jump sizes vectors

$$\tilde{l}_0[K_n] = \{\boldsymbol{v} \in l_0[K_n] : \min_{i \in S_0} |d_{0i}| \geq M\sqrt{K_n \log(L_n/K_n)}\},$$

where M is given by Theorem 1.2. The class $\tilde{l}_0[K_n]$ is similar to those classes with cut-offs for model selection consistency in sparse regression literature. In the change-point setting, it indicates that all the jump sizes on change-points are bounded away from zero. We will show that when K_n is bounded, this cut-off still suffices for model selection consistency. In this sense, our cut-off of order $K_n \log(L_n/K_n)$ is slightly better than those cut-offs of order $O(\sqrt{K_n \log p})$, which are commonly presented in existing Bayesian high-dimensional regression literature (Castillo et al. (2015); Jeong and Ghosal (2021); among others).

Theorem 1.2 guarantees that if $\boldsymbol{d}_0 \in \tilde{l}_0[K_n]$, the posterior dimensionality of $\boldsymbol{d}$ can cover all change-points. Meanwhile, we would expect the risk of over-detection to be as small as possible. The Gamma-IBP model (1.5) provides an exponentially decreasing tail probability for the dimension of $\boldsymbol{d}$, controlling the risk of over-detection of change-points. Besides, we have to carefully select the precision parameter λ of the Laplace slab in prior (1.5). Roughly speaking, we require λ to be sufficiently small so that the slab is dispersed enough to provide sufficient mass to recover the non-zero entries of $\boldsymbol{d}_0$. Strictly, we require a precision λ, so that $\lambda||\boldsymbol{d}_0||_1 < \delta$ for some positive but finite constant δ. However, $||\boldsymbol{d}_0||_1$ is unknown in practice. Therefore, we provide the following adaptive $\lambda_n(\delta)$ as the choice of the precision parameter of the Laplace slab under the Gaussian sequence model (1.6).

Let $|\bar{\boldsymbol{y}}| = p^{-1}\sum_{i=1}^{p} |y_i^*|$. The adaptive $\lambda_n(\delta)$ is given by

$$\lambda_n(\delta) = \frac{\delta}{p|\bar{\boldsymbol{y}}|}. \tag{1.7}$$

With the adaptive $\lambda_n(\delta)$, we obtain the following result of no supersets in model selection.

Theorem 1.3 (No supersets). *Let $a = c_1 L_n^{-c_3}, b = c_2 L_n^{c_4}$ for some constants $c_1, c_2 > 0$ and $c_3 > c_4 + 2 \geq 3$ in prior (1.5). Under Assumption (A1), for any fixed $K_n < L_n$ and δ, with*

$\lambda_n(\delta)$ defined in (1.7), as $n, L_n \to \infty$, we have

$$\sup_{\boldsymbol{d}_0 \in \tilde{l}_0[K_n]} E_{\boldsymbol{d}_0} \Pi_{n,L_n}\{\boldsymbol{d} : |\boldsymbol{d}| > K_n | \boldsymbol{y}\} \to 0.$$

In Theorem 1.3, we take a technical route that is different from the fashions of either Castillo et al. (2015) or Martin et al. (2017), which depends on an extremely fast decreasing speed on the prior for dimensionality and the conjugacy of data-dependent normal slab respectively. If one adopts the conditions by Castillo et al. (2015), the posterior contraction rate may be suboptimal. Although Martin et al. (2017) can reach both minimax optimality and no supersets simultaneously, their empirical Bayes approach may be difficult to be extended to other change-point scenarios. Actually, here we borrow the strength from the bound of the tail probability of IBP weights given by factor model literature Ohn and Kim (2022). However, the prior by Ohn and Kim is non-adaptive in the sense that it requires information about the true sparsity level K_n. In contrast, our choice of hyperparameters here only depends on the data sizes n and the truncation number L, and hence is adaptive. We defer the detailed proof to Supplement 1.7.1.2.

The above theorems indicate the following corollary of the posterior consistency of model selection.

Corollary 1.1 (Consistent model selection)**.** *Under the conditions of Theorem 1.3, as $n, L_n \to \infty$, we have*

$$\inf_{\boldsymbol{d}_0 \in \tilde{l}_0[K_n]} E_{\boldsymbol{d}_0} \Pi_{n,L_n}\{\boldsymbol{d} : S = S_0 | \boldsymbol{y}\} \to 1.$$

Proof. According to Castillo et al. (2015), to prove Corollary 1.1, it suffices to proving the following two assertions

$$\inf_{\boldsymbol{d}_0 \in \tilde{l}_0[K_n]} E_{\boldsymbol{d}_0} \Pi_{n,L_n}\{\boldsymbol{d} : S \supset S_0 | \boldsymbol{y}\} \to 1,$$

$$\sup_{\boldsymbol{d}_0 \in \tilde{l}_0[K_n]} E_{\boldsymbol{d}_0} \Pi_{n,L_n}\{\boldsymbol{d} : S \supset S_0, S \neq S_0 | \boldsymbol{y}\} \to 0.$$

The first assertion is a direct result of Theorem 1.2, and the second assertion is a direct result of Theorem 1.3 since $K = K_n$. □

Note that Corollary 1.1 is about the non-zero coordinates of $\boldsymbol{d}$. In other words, Corollary 1.1 indicates that we obtain posterior consistency of both the number and locations of change-points.

1.2.3 False negative rate of discrimination

As mentioned in subsection 1.1.3, we regard the posterior estimator of $\boldsymbol{d}$ as the feature to discriminate change-points $\boldsymbol{\tau}_{1:K_n}$ from $\boldsymbol{t}_{1:n}$ under the 3-sigma rule. To study the asymptotic performance of the 3-sigma discrimination, we use the marginal MAP estimator $\hat{d}_i^{\text{MAP}}$ as the signal at t_i for the theoretical concern. Note that the 3-sigma criterion in subsection 1.1.3 can be viewed as a data-driven threshold based on series $\{\hat{d}_i^{\text{MAP}}\}_{i=1}^{n-1}$.

The result of consistent model selection enables us to study the asymptotic performance of $\hat{d}_i^{\text{MAP}}$ for $i \in S_0$. Let $\hat{\boldsymbol{d}}_{S_0}$ be the least square estimator of non-zero coordinates of $\boldsymbol{d}_0$ given the correct model selection S_0, that is,

$$\hat{\boldsymbol{d}}_{S_0} = \arg\min_{\boldsymbol{d}_{S_0}} ||\boldsymbol{y}^* - X_{S_0}\boldsymbol{d}_{S_0}||_2^2,$$

where $X_S \in \mathbb{R}^{p\times|S|}$ is the submatrix of I_p with columns on the non-zero coordinates. Clearly $X_{S_0}^T X_{S_0} = I_{|S_0|}$. Let $\hat{\boldsymbol{d}}_{S_0}^{\text{MAP}}$ be the marginal MAP estimators of $\boldsymbol{d}$ on the true non-zero support S_0. Let $\boldsymbol{d}_{0S_0}$ be the true non-zero entries in $\boldsymbol{d}_0$. The follow corollary states the consistency and asymptotic normality of $\hat{\boldsymbol{d}}_{S_0}^{\text{MAP}}$.

Corollary 1.2 (Consistency of MAP under strong model selection)**.** *Under conditions in Corollary 1.1, for $\boldsymbol{d}_0 \in \tilde{l}_0[K_n]$ as $n, L_n \to \infty$, we have*

$$\hat{\boldsymbol{d}}_{S_0}^{MAP} \xrightarrow{p} \hat{\boldsymbol{d}}_{S_0}, \ \sqrt{p}(\hat{\boldsymbol{d}}_{S_0}^{MAP} - \boldsymbol{d}_{0S_0}) \xrightarrow{d} N(0, I_{|S_0|}).$$

The proof of Corollary 1.2 is trivial. Under the correct model selection, the prior for $\boldsymbol{d}_{S_0}$ is reduced to the continuous Laplace slab and hence, the MAP estimator $\hat{\boldsymbol{d}}_{S_0}^{\text{MAP}}$ converges to the maximum likelihood estimator $\hat{\boldsymbol{d}}_{S_0}$ almost surely (Pronzato and Pázman, 2013, Theorem 4.16).

Since the model selection converges to be correct in probability, it suffices showing the weak convergence of the MAP estimator $\hat{\boldsymbol{d}}_{S_0}^{\mathrm{MAP}}$ to $\hat{\boldsymbol{d}}_{S_0}$. Then the second assertion is established by the central limit theorem.

The above distribution approximation about $\hat{\boldsymbol{d}}_{S_0}^{\mathrm{MAP}}$ controls the false negative rate under the 3-sigma rule. Let $\bar{d}_0 = p^{-1}\sum_{i=1}^{p} d_{0i}$, $\bar{d} = p^{-1}\sum_{i=1}^{p} \hat{\boldsymbol{d}}_i^{\mathrm{MAP}}$, $\psi_0 = \sqrt{p^{-1}\sum_{i=1}^{p}(d_{0i} - \bar{d}_0)^2}$, and $\psi = \sqrt{p^{-1}\sum_{i=1}^{p}(\hat{\boldsymbol{d}}_i^{\mathrm{MAP}} - \bar{d})^2}$. The 3-sigma rule acts as a special hard threshold that shrinks all $|\hat{\boldsymbol{d}}_i^{\mathrm{MAP}}| < 3\psi$ to zero. We require an upper bound assumption on the norm of $\boldsymbol{d}_0 \in \tilde{l}_0[K_n]$.

(A2) There exists a universal constant M_0, so that $p^{-1/2}||\boldsymbol{d}_0||_2 < M_0[\sqrt{K_n}\log(L_n/K_n)]$.

Assumption (A2) implies that $3\psi_0$ will not exceed any non-zero entries in $\boldsymbol{d}_0$ and hence the 3-sigma rule is suitable for the true jump sizes vector $\boldsymbol{d}_0$ is The following corollary states that under the 3-sigma rule, the probability that a change-point is falsely discriminated as a stationary point is asymptotically zero. We defer the proof to Supplement 1.3.

Corollary 1.3. *Under the conditions in Corollary 1.1 and Assumption (A2), as $n, L_n \to \infty$, we have*

$$\sup_{\boldsymbol{d}_0 \in \tilde{l}_0[K_n]} E_{\boldsymbol{d}_0} \Pi_{n,L_n}\{|\hat{\boldsymbol{d}}_i^{MAP}| < 3\psi, i \in S_0 | \boldsymbol{y}^*\} \to 0.$$

Corollary 1.3 theoretically justifies the 3-sigma criterion for change-point discrimination. In general, the 3-sigma rule is employed for outlier detection, especially for the Gaussian population. In general, the performance of discriminating the outliers depends on two properties, the variation of the population and the distance between the outliers and the center. The cut-off of the $\tilde{l}_0[K_n]$ class guarantees that those outliers (change-points) differ significantly from the zero-center population (stationary points), while the additional Assumption (A2) avoids those outliers from affecting the variation of all the samples too much. Corollary 1.3 implies that even under a very high precision level (3-sigma criterion usually yields a high precision), the recall of the discrimination is sufficiently large and asymptotically converges to one. This is supported by our finite sample simulations under the Gaussian mean-shifted model of Scenario (i), where NOSE enjoys higher recall than other competing approaches.

1.3 Bayesian implementation

In this section, we introduce technical details for the Bayesian implementation of the proposed method.

Uniform convergence of $\theta(t)$

Recall that our methodology stands on $\theta(t)$, the truncated form of $\theta(t)$. Hence it is necessary to check the convergence of the truncated form as $L \to \infty$. We present the uniform convergence of $\theta(t)$ by the following theorem. We defer the proof to Supplement 1.7.1.4.

Theorem 1.4 (Uniform convergence)**.** *For any continuous density F_0 with support $\mathbb{R}$ in* (1.5)*, given $\boldsymbol{\xi}$ and fixed a, b in the Gamma prior for α, the truncated $\boldsymbol{Q}^L$ in* (1.3) *converges to $\boldsymbol{Q}$ in* (1.2) *uniformly for all $t \in \mathcal{T}$ in probability.*

In practice, the choice of the truncation number L depends on one's prior belief on the minimum distance between change-points. In the case where the number of change-points K is not large, a relatively small L is suggested to simplify MCMC sampling. In our experience, when the truncation number exceeds a sufficiently large L, the detection result is stable with L increasing, numerically demonstrating Theorem 1.4.

MCMC sampling

We approximate the posterior distribution through MCMC sampling. Our computation is facilitated by the `nimble` (de Valpine et al., 2017) package in `R`, which uses `BUGS` type syntax (Lunn et al., 2000) and compiles the code into **C++** to facilitate automatic posterior sampling. Samplers for different parameters are automatically assigned by `nimble`. For conjugate parameters, say, p_ℓ, `nimble` assigns Gibbs samplers; for parameters ξ_ℓ and α, `nimble` assigns the default Metropolis-Hasting sampler; for h_ℓ and the corresponding binary indicator Z_ℓ, we configure a reversible jump MCMC sampler to speed up the sampling. The `R` package NOSE based on `nimble` includes several `R` functions applied to application scenarios mentioned in subsection 1.1.4.

Cauchy slab

Note that Theorem 1.4 holds for any continuous density for the slab term. This implies that the choice of slab density for h_ℓ is not limited to Laplace, but also includes some polynomial-tailed densities such as Student-t or Cauchy which prevent over-shrinkage of the non-negligible entries (Bai et al., 2020). Though we establish the posterior inferential theories in Section 1.2 by specifying a Laplace slab in the Gamm-IBP model, we recommend a standard Cauchy slab instead in practice for numerical concerns. We find that the MCMC efficiency of Cauchy slab is about 20 to 50 times of that of the Laplace slab in `nimble`. That is, to draw the same effective sample size of posterior, the time cost of Cauchy slab is much less than the Laplace slab. Note that the moment of Cauchy slab is infinite and hence we do not need to figure out the adaptive precision parameter λ_n for Laplace slab in subsection 1.2.2.

The reason for the low MCMC efficiency of the Laplace slab may be due to the complicated form of conditional posterior distribution of a Laplace prior. Give $Z_\ell = 1$, $h_\ell \sim \text{Laplace}(0, \lambda)$ can be expressed as $h_\ell | \tau_\ell \sim N(0, \lambda^{-2}\tau_\ell^2)$ with $\tau_\ell \sim \exp(1/2)$. The presence of the auxiliary scale parameter τ_ℓ hinders the use of reversible jump MCMC sampler. Meanwhile, the conditional posterior distribution is in an inverse Gaussian form (Ohn and Kim, 2022), which can hardly be simulated from the default Metropolis-Hasting random walk sampler in `nimble`, incurring a very low acceptance rate and thus, low MCMC efficiency.

In the high-dimensional regression setting, Shin and Liu (2021) numerically showed that both the Laplace and Cauchy slab share a very similar performance, while the Cauchy slab appears to enjoy a lower false positive rate and higher cosine similarity to the true parameter. Their results provide a numerical justification to the use of the Cauchy slab in replace of the Laplce slab.

Continuous ξ_ℓ

To determine a discrete draw from states $\boldsymbol{t}_{1:n}$ without replacement is difficult in `nimble`. Hence, we have to make a continuous adjustment to adopt the programming framework of `nimble`. Note that for any t_i and t_{i+1} with an increment $d_i = \theta(t_{i+1}) - \theta(t_i) > 0$, it is equivalent to either draw an atom ξ_ℓ at t_{i+1} exactly, or to draw an atom $\xi_\ell \in (t_i, t_{i+1})$. This motivates us to consider a continuous prior for ξ_ℓ as an approximation. Without loss of generality, we assume $t_i = i$ for $i = 1, \ldots, n$. Then we sample ξ_ℓ from a continuous uniform distribution $U(0, n)$ in `nimble` as the continuous prior for ξ_ℓ.

A risk of the continuous prior ξ_ℓ is that more than one atoms fall into the same interval (t_i, t_{i+1}), which may lead to an ill posterior of increment d_i. Note that the probability that the minimum distance between L uniform $U(0, n)$ variables exceeds 1 is $(1 - n^{-1})^L$. As n increases to $L/n \to 0$, the probability converges to 1, that is, the probability that an interval (t_i, t_{i+1}) contains more than one atom converges to zero. Therefore, the continuous scheme of ξ_ℓ suffices to approximate prior (1.4) when $n >> L$.

In the finite sample case, too closely located atoms may cause over-detection of change-points by wrongly putting increments to data points that are close to the true change-points. To avoid over-detection, we conduct post-processing of change-point. We use the prior belief in the minimum distance D between change-points as the lower bound of the distance between change-points. For each two consecutive estimated change-points $\hat{\tau}_k, \hat{\tau}_{k+1}$, if $|\hat{\tau}_k - \hat{\tau}_{k+1}| < D$, we only retain the left end-point $\hat{\tau}_k$ as a change-point but remove the rest. Such a kind of post-processing based on the prior belief in the minimum distance between change-points is common in most literature (Matteson and James (2014); Baranowski et al. (2019); Cappello et al. (2023); among others). Our sensitivity analysis shows that NOSE is not sensitive to the choice of D; see supplement for more details. This post-processing is applied throughout all numerical studies in this chapter.

Adjustment of $\hat{\sigma}$

In a finite sample experiment, Assumption (A2) may no longer hold, especially if L is chosen as a relatively small number. For a sequence $\{\zeta_i\}_{i=1}^{n-1}$, those ζ_i whose absolute values exceed three times the sample standard deviation may cause a much larger variation than the variation of the zero-center population. To avoid a too large sample deviation, we adopt an empirically adjusted value of $\tilde{\sigma}$ rather than using the sample standard deviation. Note that in a standard normal case, the 3-sigma rule indicates a tail probability of 0.001. Therefore, we first obtain a trimmed sample of ζ_i by cutting off the two tails of 0.0005 probability. Then we use the trimmed sample standard deviation as an empirical adjustment of $\tilde{\sigma}$. The adjustment $\tilde{\sigma}$ is used throughout the numerical studies in this chapter.

1.4 Simulations

Comprehensive simulations are conducted to evaluate the performance of NOSE by comparing it with other state-of-the-art methods available in R Archive Network. We consider examples in Scenarios 1-5 introduced in subsection 1.1.4. For Scenario 5, since most existing approaches are not available for this scenario when there are multiple responses observed at the same time, we report the results given by NOSE only. Results of additional simulations under model misspecification settings of changes in means with autocorrelated noises, changes in means with heavy-tailed noises, and changes in autocorrelation coefficient with model misspecification are deferred to Supplement 1.7.2.1. All numerical studies included in this chapter are conducted under R version 4.1.0 on a Macbook Air with an M1 CPU and 8GB RAM.

Settings

We consider the following settings. Under each simulation setting, 300 Monte Carlo replicate datasets are generated.

(S.1) Changes in normal means on equal segments (in Scenario 1). We have $n = 400$ independent Gaussian observations with $K = 7$ change-points at $(50, 100, 150, 200, 250, 300, 350)$, leading to 8 segments with segment mean $\mu = (0, 1.5, 3, 1.5, 3, 0.5, 2, 0)$. The common scale parameter is set to be $\sigma = \sqrt{2}$.

(S.2) Changes of normal mean on unequal-length segments with large variations (in Scenario 1). We have $n = 916$ independent Gaussian observations with $K = 11$ change-points at $(81, 134, 178, 267, 346, 413, 528, 577, 636, 741, 822)$, leading to 12 segments with segment mean $\mu = (0, 1.23, -0.248, 0.861, -0.534, 1.057, 0.369, 1.331, 0.483, 1.105, -1.101, 0)$. The common scale parameter is set to be $\sigma = 1$. Some jump sizes are smaller than the within-segment variation, leading to many difficulties in correctly identifying change-points.

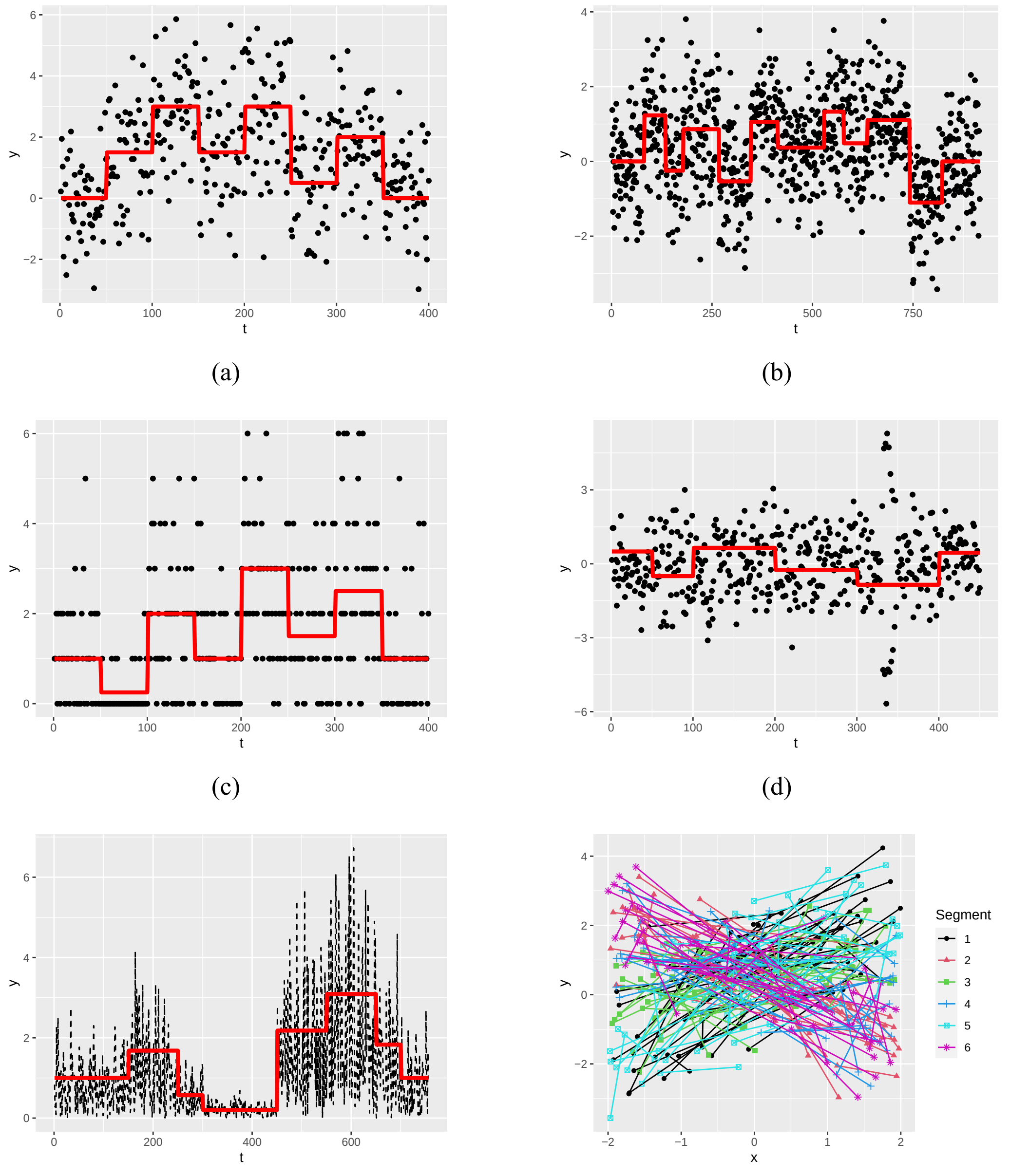
(S.3) Changes of Poisson parameter (in Scenario 2). We have $n = 400$ independent Poisson variables with $K = 7$ change-points at $(50, 100, 150, 200, 250, 300, 350)$, leading to 8 segments with segment parameter $\lambda = (1, 0.25, 2, 1, 3, 1.5, 2.5, 1)$.

(S.4) Changes of normal scale with small variations on the mean (in Scenario 3). The data are generated to simulate the DRAIP data. We have $n = 756$ independent Gaussian observations with $K = 7$ change-points at $(150, 250, 300, 450, 550, 650, 700)$, leading to 8 segments with segment scales $\sigma = (1, 1.68, 0.57, 0.20, 2.18, 3.09, 1.83, 1)$. Meanwhile, we allow small variations on the mean such that the segment mean is $\mu = (0.056, 0.047, -0.034, -0.017, 0.032, 0.068, -0.042, 0.017)$.

(S.5) Changes of autocorrelation coefficient in an AR(1) model (in Scenario 4). The data generating process is $Y_t = \phi Y_{t-1} + \phi_0 + \epsilon_t$. We have $N = 450$ observations with 5 change-points at $t = (50, 100, 200, 300, 400)$, leading to 6 segments with segment autocorrelation coefficient $\phi = (0.5, -0.5, 0.65, -0.25, -0.85, 0.45)$. The model error $\epsilon_t \sim N(0, 1)$.

(S.6) Changes of regression coefficient in a linear regression model (in Scenario 5). Data are generated by $y_{tj} = \beta_0 + \theta(t) X_{tj} + \epsilon_{tj}, j = 1, 2, t = 1, \ldots, 240$, where $\beta_0 = 0.5, X_{tj} \sim U(-2, 2)$, and $\epsilon_{tj} \sim N(0, 1)$. We set $K = 5$ change-points at $t = (40, 80, 120, 160, 200)$, with the segment-wise values $\theta(t) = (1, -1, 0.5, -0.5, 1, -1)$.

Examples of simulated data are presented in Figure 1.3. Figures 1.3(a) to 1.3(c) find that some jump sizes are relatively small and the corresponding change-points are imperceptible in the data stream. Figure 1.3(d) finds that the data with identical signs are clustered in those segments with positive auto-correlation, and opposite signs of data appear alternately in those segments with negative auto-correlation. Figure 1.3(e) presents the centered absolute data $|Y - EY|$ and the true $\theta(t)$ together, where the heights of the centered absolute data reflect the changes in the scale parameters. Figure 1.3(f) presents the covariates and the responses grouped by the state t and labels the curves by the segments at which they are located.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 1.3: Examples of generated data in simulations. (a) to (d), data stream (in points) and $\theta(t)$ (in red lines). (e), centered absolute data stream $|Y_i - E(Y_i)|$ (in dashed line) and $\exp\{\theta(t)\}$ (in red line). (f), data grouped by t (in polylines labeled by segments). (a), **S.1** (Scenario 1); (b), **S.2** (Scenario 1); (c), **S.3** (Scenario 2); (d), **S.5** (Scenario 4); (e), **S.4** (Scenario 3); (f), **S.7** (Scenario 5).

Estimators

In all simulations, we adopt a unified setting of truncation number $L = 25$ and the prior belief on the minimum distance between change-points $D = 15$ for NOSE. We also present the more general choice of $D = 2$ in the supplement to validate the robustness to the choice of D. We run 4 independent parallel MCMC chains and obtain 1200 scans in each chain thinned from a total 15000 after a burn-in period of 3000 iterations. Finally, we get 4800 posterior samples for change-point discrimination. Under such a MCMC setting, in all simulation scenarios, we obtain approximately 1000 average effective sample size of $\theta(t)$ in each replication, guaranteeing the reliability of posterior inference.

Competitors vary among different settings since none of them can be applied to all the above simulation settings. For settings **S.1**, **S.2** and **S.3**, where the mean parameter changes, we compare with the NOT method by Baranowski et al. (2019) in package `not`, the TUGH method by Fryzlewicz (2018) in package `breakfast` (Anastasiou et al., 2022), the MOSUM method by Birte and Claudia (2018) in package `mosum` (Meier et al., 2021), the FDRSeg method by Li et al. (2016) in package `FDRSeg`, the SMUCE method by Frick et al. (2014) in package `StepR`, the WBS method by Fryzlewicz (2014) in package `wbs`, and the PELT method by Killick et al. (2012) in package `changepoint` (Killick and Eckley, 2014), ; for setting **S.4**, where the scale parameter changes, we compare with NOT, SMUCE, and PELT methods; for setting **S.5**, where data are autocorrelated, we compare with the WBSTS method by Korkas and Pryzlewiczv (2017) in pacakge `wbsts` and the B-P method by Bai and Perron (2003) in package `struchchange` (Zeileis et al., 2002). The tuning parameters for the competing methods are set as the default values in the corresponding `R` packages. We do not present results by Bayesian approaches such as `StepSignalMargiLike` (Du et al., 2016) and `solo.cp` (Cappello et al., 2023) here. We find the results of `StepSignalMargiLike` are sensitive to the choices of a maximum number of segments and cannot find a stable estimation of the number; `solo.cp` cannot detect most of change-points in the mean under our simulation settings. We conjecture the reason is that `solo.cp` identifies change-points based on the jump probability, which may fall around $1/2$ when the jump sizes are relatively small, say, our simulation settings.

Assessments and results

Several assessments are employed to measure the accuracy of the detected number of change-points and the accuracy of locations of estimated change-points. We report the frequency table for $\hat{K} - K$, the difference between the number of detected change-points and the true number of change-points to evaluate the accuracy of the detected number of change-points. To measure the accuracy in locations, three assessments are considered, precision, recall, and the scaled

Hausdorff distance (Hausdorff). For all true change-points, we count one true positive (TP) if there is at least one change-point identified within a window of 10 data points and compute the number of false positive (FP) as the number of predicted changes minus TP. Let K be the true number of change-points. Then precision is computed as $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, and recall is computed as TP/K. The scaled Hausdorff distance is computed as

$$d_H = n^{-1}E[\max\{\max_{j=0,\cdots,K+1}\min_{k=0,\cdots,\hat{K}+1}|\tau_j-\hat{\tau}_k|, \min_{k=0,\cdots,\hat{K}+1}\min_{j=0,\cdots,K+1}|\hat{\tau}_k-\tau_j|\}],$$

where $t_0=\tau_0<\cdots<\tau_K<\tau_{K+1}=t_N$ and $t_0=\hat{\tau}_0<\hat{\tau}_1<\ldots<\hat{\tau}_{\hat{K}}<\hat{\tau}_{\hat{K}+1}=t_N$ denotes true and estimated change-points, respectively. The scaled Hausdorff distance takes values in $[0,1]$ and is the smaller the better.

From Table 1.1 we find that NOSE outperforms in the frequency of correctly specifying the number of change-points in all settings. In contrast, other competitors tend to under detect the number of change-points except for the setting **S.3**, where changes take place on both the mean and variance of data. Although the jump sizes under these simulation settings (especially setting **S.2**) are not significant enough to make the changes be identified by eyes, NOSE still enjoys the highest recall in all settings, demonstrating its capability to correctly identify change-points. These results may be evidence that the performances of segmental approaches seem to be less sensitive to small jump sizes than our non-segmental approach, particularly when the nuisance parameter (say, the scale parameter σ in the mean-shifted model) has substantial impacts on the variation of the whole data stream. The precision and Hausdorff distance given by NOSE outperforms under setting **S.3**, and are competitive under other settings. Note that other winners on precision and scaled Hausdorff distance actually underestimate the number of change-points, while a most parsimonious estimator usually brings higher precision and lower Hausdorff distance. Under setting **S.6**, NOSE correctly specifies all change-points in almost all replications, with pretty high precision and recall. In summary, NOSE performs to be the most competitive and robust to correctly specify the number of change-points and estimate their locations accurately.

Table 1.1: Results of change-points detection under settings **S.1** to **S.5** among 300 Monte Carlo replicates. The best results are bold.

Setting	Method	Frequency of $\hat{K} - K$							Precision	Recall	$d_H \times 10^2$
		≤ -3	-2	-1	0	+1	+2	$\geq +3$			
S.1	NOSE	1	1	33	**252**	13	0	0	0.95	**0.94**	**2.1**
	NOT	9	12	31	227	19	2	0	0.93	0.91	2.4
	SMUCE	47	68	130	55	0	0	0	0.85	0.7	3.1
	WBS	16	35	95	138	14	0	2	0.93	0.84	2.5
	FDRSeg	6	16	63	171	29	10	5	0.90	0.88	3.0
	PELT	1	6	12	210	52	16	3	0.91	0.93	2.8
	TUGH	0	0	1	217	51	14	5	0.96	0.93	2.9
	MOSUM	3	3	72	181	41	0	0	**0.98**	0.93	2.6
S.2	NOSE	15	48	77	**144**	15	1	0	0.93	**0.87**	1.5
	NOT	52	91	49	101	7	0	0	0.94	0.82	1.4
	SMUCE	136	113	50	1	0	0	0	0.86	0.67	2.1
	WBS	68	120	74	38	0	0	0	0.95	0.79	**1.2**
	FDRSeg	28	71	74	100	23	2	2	0.88	0.81	2.2
	PELT	38	101	42	107	12	0	0	0.83	0.83	1.4
	TUGH	12	37	53	129	48	17	4	0.97	0.84	2.4
	MOSUM	71	97	98	30	4	0	0	**1**	0.80	**1.2**
S.3	NOSE	4	28	113	**148**	6	1	0	**0.90**	**0.82**	**2.9**
	NOT	37	71	77	90	23	1	1	0.87	0.74	3.2
	SMUCE	10	68	151	69	2	0	0	0.89	0.76	3.0
	WBS	1	5	34	41	65	63	85	0.64	0.76	4.8
	FDRSeg	0	3	6	8	20	22	241	0.47	0.83	5.7
	PELT	25	50	102	61	38	15	9	0.77	0.69	3.5
S.4	NOSE	0	75	71	**150**	4	0	0	0.84	**0.75**	2.3
	NOT	25	221	39	14	0	0	1	**0.91**	0.67	1.5
	SMUCE	40	211	49	0	0	0	0	0.64	0.64	**1.2**
	PELT	1	153	58	83	5	0	0	0.88	0.72	2.0
S.5	NOSE	0	0	98	**154**	46	2	0	0.85	**0.82**	2.6
	WBSTS	4	36	74	122	48	14	2	0.61	0.47	2.8
	B-P	102	68	128	2	0	0	0	**0.89**	0.38	**1.8**
S.6	NOSE	0	0	1	293	6	0	0	0.99	1	0.75

1.5 Applications

1.5.1 DRAIP data: shifts in scale

We report detection results on DRAIP data given by NOSE here. We set $L = 25$ and $D =$

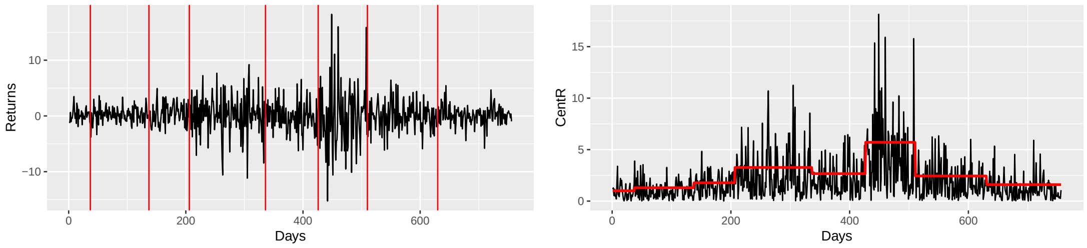


Figure 1.4: DRAIP data and change-point detection results by NOSE. Top, original data and locations of estimated change-points (in vertical lines); bottom, centered absolute data and estimated segment-wise scale parameters (in the horizontal polyline).

15 in this case. As shown by Figure 1.4, NOSE detects 7 change-points. We summarize the piecewise standard deviations and estimated standard deviations given by NOSE on the intervals partitioned by the estimated change-points as well as all jump sizes in Table 1.2. The estimated scale parameters and sample standard deviations are quite close, and both suggest a shift in the estimated change-points, supporting the detection result by NOSE. According to Table 1.2, the first jump size is pretty small, and no wonder why other segmental approaches miss the point. Although the 4th jump size on $t = 336$ is absolute enough to be observed by eyes, it is also missed by other segmental approaches. We conjecture the reason is that the dispersion of the data on the interval $[207, 427]$ is relatively large. As evidence, Figure 1.5 shows the Q-Q plot and the density curve of the data on the interval, where we find the samples on the interval are too dispersed to be Gaussian. It indicates that may hinder the traditional segmental approaches detecting the change-point on the interval. The results of simulations based on the DRAIP data are displayed in Supplement 1.7.2.2 The simulation results demonstrate the difficulty of correctly specifying all the change-points in DRAIP data. Even so, NOSE still outperforms other approaches.

Intervals	Estimated SD	Sample SD	Scale jump sizes
$[1, 37]$	1.000	1.173	-
$[38, 137]$	1.296	1.369	**0.196**
$[138, 206]$	1.778	1.873	0.504
$[207, 336]$	3.266	3.500	1.627
$[337, 426]$	2.666	2.570	**-0.930**
$[427, 510]$	5.708	5.863	3.293
$[511, 630]$	2.437	2.426	-3.437
$[631, 756]$	1.599	1.599	-0.827

Table 1.2: Intervals, intervals partitioned by estimated change-points; Estimated: standard deviation estimated by NOSE; Sample SD: sample SDs on partitioned intervals; Jump sizes, jump sizes calculated from true SDs.

1.5.2 ACGH data: shifts in mean

In the second example, we analyze the public dataset of DNA copy numbers using ACGH for 43 different individuals with a bladder tumor (Stransky et al., 2006), which is available in R package `ecp` (James et al., 2015). For each individual, the copy number is recorded on 2215 locations. We aim to detect the changes in the mean of the copy number. Hence we employ

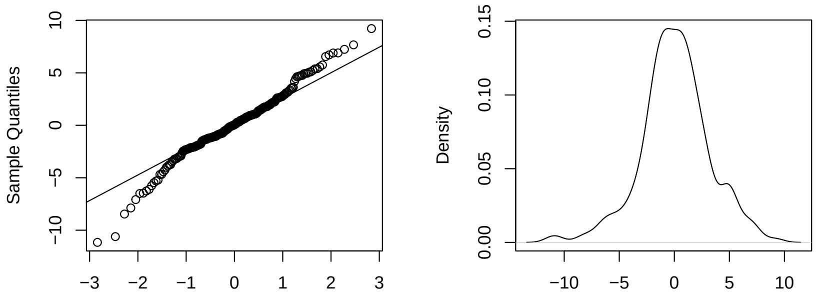
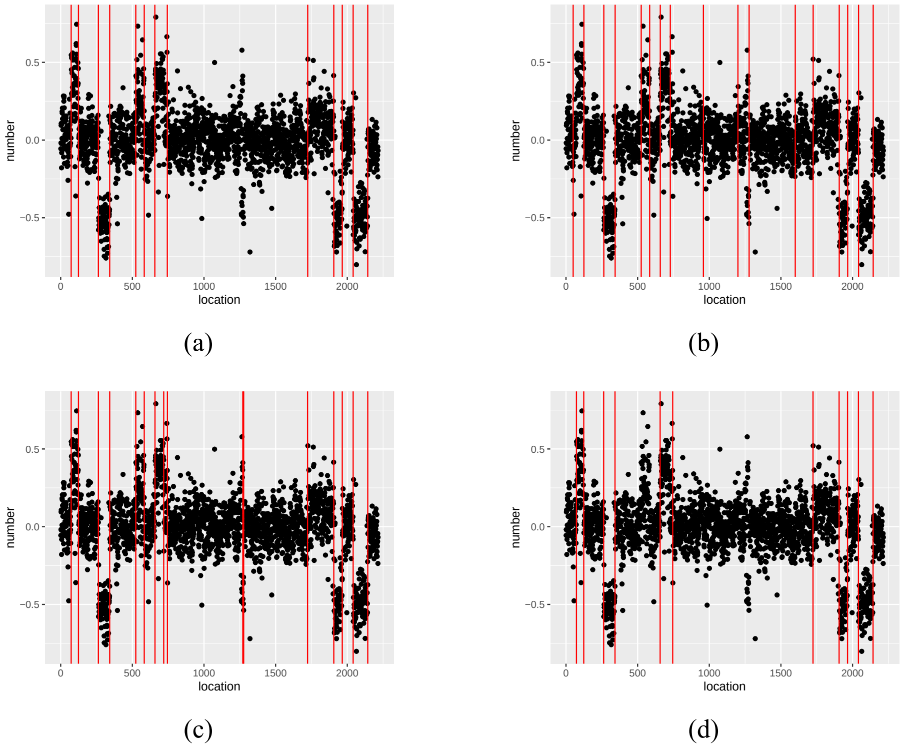


Figure 1.5: Q-Q plot and density plot of DRAIP data on interval $[207, 427]$. Left, Q-Q plot; right, density plot.

NOSE for Gaussian mean changes under scenario (i). As the number of change-points is usually considered to be quite large, we set $L = 55$ to incorporate sufficiently many change-points. The prior belief on the minimum distance between change-points is set as $D = 15$. We display the analysis result of the 37th individual in this chapter.



(a)

(b)

(c)

(d)

Figure 1.6: Plot of ACGH data (in black points) and estimated locations of change-points (in red vertical lines). (a), NOSE; (b), HSMUCE; (c), NOT; (d), R-FPOP.

We display detection results of NOSE, HSMUCE (Pein et al., 2017) and NOT in Figures 1.6(a), 1.6(b) and 1.6(c), where they detect 13, 16, and 15 change-points, respectively. Despite some similarities among them, HSMUCE and NOT are more likely to create short segments gathering several data points that are far away from the means of adjacent segments. We conjecture the points in these short segments are outliers. To eliminate the influence of outliers, we employ the outlier-robust R-FPOP method (Fearnhead and Rigaill, 2019) equipped with the Huber loss and penalized value 1.345 as default; see Figure 1.6(d). We find the data points in those short segments divided by HSMUCE and `NOT` are treated as outliers by R-FPOP. By comparison, NOSE and R-FPOP produce almost the same segmentation, with the only difference being the segment $(524, 583)$, where NOSE creates a new segment while R-FPOP does not. Since this segment contains 60 data points, we feel that it is more appropriate to partition these points into a new segment rather than identifying them as outliers.

We generate simulated data from the estimation results by NOSE in Figure 1.6(a). Since the simulated data are exactly Gaussian without outliers, the results of NOSE, HSMUCE, and R-FPOP are stable and similar to each other, while `NOT` slightly over-detects the change-points. Details are deferred to Supplement 1.7.2.3.

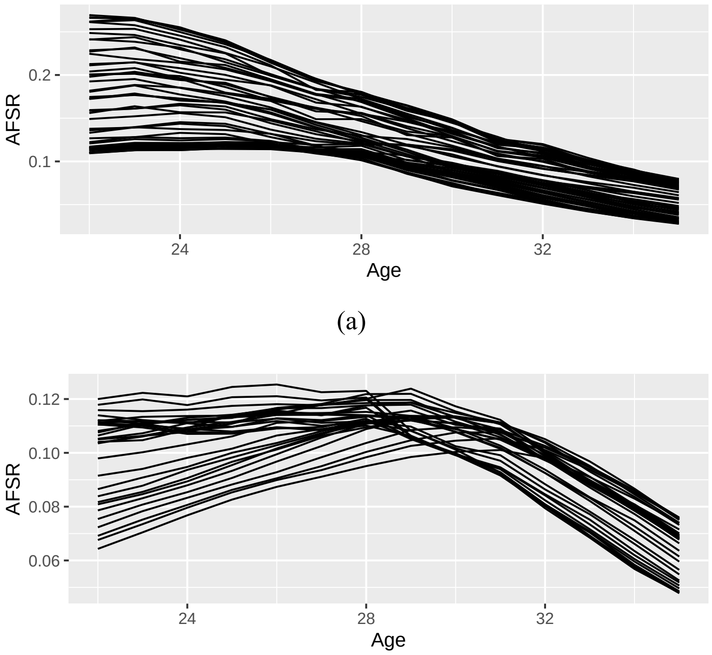
1.5.3 US age-specific fertility rate (ASFR) data: structural changes in linear models

The declining birth rates in many developed countries arouses much interest to the analysis of the annual Age-Specific Fertility Rate (ASFR). Given the year t, let B_{tj} be the number of births during the year to females of a specified age j, and N_{tj} be the number of females of the age j in that reference year. In year t, the ASFR y_{tj} is defined as the ratio between B_{tj} and N_{tj}. We collect ASFR data in the US from 1940 to 2021 at ages 22 to 35, the age period which covers the age with the highest ASFR. Then totally we obtain 1134 responses y_{tj}.

The relationship between the ASFR and specific ages from 22 to 35 seems to be linear. Hence, we consider a linear model with changes in the regression coefficient to characterize their association. We consider following linear models

$$y_{tj} = \beta_0 + \theta(t)X_{tj} + \epsilon_{tj},\ t = 1, \dots, 81,\ j = 1, \dots, 14,$$

where the regressor $X_{\cdot j} = 21 + j$, the regression coefficient $\theta(t)$ may change along with time t, β_0 is a fixed intercept and $\epsilon_{ts} \sim N(0, \sigma^2)$ are i.i.d. model errors. We apply NOSE to detect changes of $\theta(t)$, where the state of data is set to be the year t. We set $L = 25$ and the minimum distance threshold $D = 15$.



(a)

(b)

Figure 1.7: Visualization of the pre- and post-change-points ASFR data in US. (a), relationship between age and ASFR before year 1992; (b), the relationship between age and ASFR after year 1992.

Only one change-point is detected by NOSE at $t = 1992$. To understand the effect of the change-point, we plot the curves of ASFR versus age before and after 1992 in Figure 1.7. From the figure, we can clearly see that before the change point, the ASFR decreases almost linearly with age, and thus the ASFR is highest at age 22. However, after the change point, the association between ASFR and age is non-linear and even non-monotonic, with ASFR first increasing and peaking at age 29 and then decreasing.

1.5.4 House prices in London Borough of Newham: structural changes in AR(1) models

We further explore a real dataset, the average monthly property price P_t in the London Borough of Newham. We take the average of all properties and select the data recorded from January 2010 to November 2020 and we totally have 131 observations. This dataset was once analyzed by Fryzlewicz (2023) to identify the shortest interval of change-points under an AR(1) model.

We adopt the AR(1) model $P_t = \theta(t)P_{t-1} + \theta_0 + \epsilon_t$, where the autocorrelation coefficient $\theta(t)$ is treated as the global parameter that may change, the intercept θ_0 is fixed, and $\epsilon_t \sim N(0, \sigma^2)$ are independent model errors. We set $L = 25$ and $D = 15$.

As shown in Figure 1.8, NOSE detects 1 change-point located in Oct 2016 (location 82). The date of change-point is close to the beginning of the vote of Britain's EU membership referendum, indicating that the structural change may be caused by the event. The WBSTS method cannot detect change-point after processing; the B-P method provides a similar result of change-point detection, where the estimated location is 79. Meanwhile, the estimated confidence interval given by R package `nsp` (Fryzlewicz, 2021) is (24, 97), which covers the change-point estimated by NOSE.

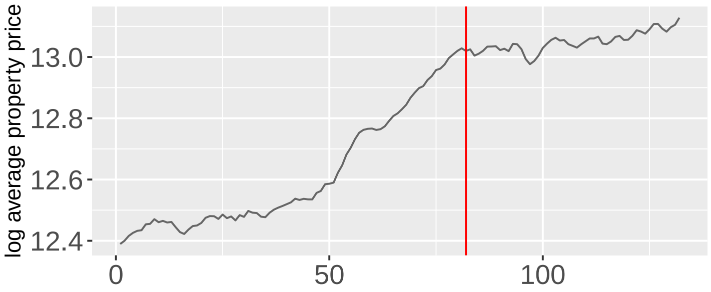


Figure 1.8: House prices in London Borough of Newham and locations of estimated change-points given by NOSE (the red line).

1.6 Discussion

In this chapter, we propose NOSE, a non-segmental change-point detection approach that globally models the abrupt change scheme rather than taking a segment-wise view. NOSE first draws posterior inference to the process of jump sizes and then identifies the change-points based on a 3-sigma threshold. Particularly, the proposed NOSE methodology in this chapter has two pieces of uniqueness.

i.) NOSE models the entire abrupt change process directly through $\theta(t)$ ($\equiv \theta$) rather than aggregating all sets of segment parameters in prevailing methods. In this sense, NOSE can be viewed as an infinite-dimensional extension of `StepSignalMargiLike` (Du et al., 2016), which represents the abrupt change scheme through a finite-dimensional vector $\boldsymbol{\theta}_{1:m}$ with each entry being the latent feature of a segment. Their m is the maximum number of segments and needs to be prespecified. Thus, any misspecification of m is risky to their results of change-point detection. In contrast, the atomic expression of $\theta(t)$ in NOSE looks as if a much "denser" segmentation than `StepSignalMargiLike` so that m can go to infinity. Hence, NOSE is exempted from the sensitivity of the upper bound of the number of segments.

ii.) NOSE may be the first approach that deals with the sparsity of the vector of *jump heights* (vertical), unlike existing penalized approaches that focus on the sparsity of the vector of *jump locations* (horizontal). In detail, NOSE identifies change-points by the posterior estimates (ζ_i) of jump heights/sizes (d_i) on states (i), where any non-negligible jump height/size indicates a change. In the broad sense, NOSE may be viewed as a vertical extension of SMUCE (Frick et al., 2014) in searching for sparse solutions under a high-dimensional regression setting. Different sparsity reviews lead to different theoretical properties: SMUCE reaches minimaxity in the estimation of change locations (up to a logarithm) and consistency of estimation of the number of change-points under the frequentist paradigm; NOSE obtains the posterior minimax optimality in recovering the jump height vector and posterior consistency of *both* the number and the locations of change-points under the Bayesian paradigm.

We may try to explain the success of NOSE from the perspective of cohesion and repulsion in clustering (Natarajan et al., 2023). To some extent, *change-point detection may be viewed as an ordered clustering task on sequential data. Those data points within the same segment can be viewed as a cluster*. Quoting Natarajan et al. (2023), "clusters are composed of objects which have small dissimilarities among themselves (cohesion) and similar dissimilarities to observations in other clusters (repulsion)". Intuitively, jump size may be viewed as a metric of dissimilarity between data points. In our approach, the nearly black jump size vector indicates that there are no dissimilarities within a cluster but significant dissimilarities across different clusters, leading to an ideal clustering under the cohesion-repulsion principle.

The computation cost of NOSE consists of two parts: the time for MCMC sampling and the time for change-point discrimination. The complexity of the latter is obvious $O(n)$ (scanning along the $n-1$ jump sizes), while measuring the complexity of the MCMC procedure is difficult.

Note that the computation of NOSE is feasible, though we admit that the computation cost is affected by the data size n and the truncation number L. As a remedy, we tend to select a not-too-large L in our numerical studies, corresponding to a relatively large minimum distance D (we select $L \approx n/D$). In the future, we plan to develop variational Bayes (VB) procedures to speed up the implementation of the NOSE approach. The VB algorithms in all application scenarios are non-trivial and case-specific, worthwhile for a separate work.

1.7 Supplement

1.7.1 Proofs

1.7.1.1 Proof of Theorem 1.1

Before proving Theorem 1.1, the necessary propositions and a lemma are given as follows.

Proposition 1.1 (Gaussian sequence prior)**.** *Let $S \subset \{1, \dots, p\}$ be the non-zero coordinates of the jump size vector $\mathbf{d}$ of cardinality $|S|$. Let $\mathbf{d}_S$ be the set of non-zero values $\{d_i, i \in S\}$. Let π_{L_n} be a prior selects a dimension s from $\{0, 1, \dots, L\}$. Under the priors for $\boldsymbol{\xi}$ and $\boldsymbol{h}$ in* (1.4) *and* (1.5)*, for a fixed truncation number L, the prior for $\mathbf{d}$ with non-zero coordinates S is in the form of*

$$\pi(\boldsymbol{d}) \propto \frac{1}{\binom{L_n}{|S|}} \pi_{L_n}(|S|) g_S(\boldsymbol{d}_S) \delta_0(\boldsymbol{d}_{S^c}). \tag{1.8}$$

Proof. Drawing a sample of $\boldsymbol{d}$, with non-zero coordinates set S from priors (1.4) and (1.5) can be divided into the following steps

1. Draw $\boldsymbol{\xi}$ so that $S \subset \boldsymbol{\xi}_{1:L_n}$.

2. Given ξ_ℓ, draw indicators Z_ℓ so that $\sum_{\ell=1}^{L_n} Z_\ell = |S|$ and assign those non-zero indicators to locations S.

3. Given the non-zero indicators Z_ℓ, draw $\boldsymbol{d}_S$ from the slab term of h_ℓ and assign zeros to other coordinates.

In terms of step 1, recall that a draw of $\boldsymbol{\xi}_{1:L}$ is a draw of L elements of $\{1, \ldots, p\}$ without replacement. Hence we have

$$Pr\{S \subset \boldsymbol{\xi}_{1:L}\} = \left\{ \binom{p}{L_n} \binom{L_n}{|S|} \right\}^{-1}.$$

In step 2, we immediately have

$$\pi_{L_n}(|S|) = Pr\left\{ |\boldsymbol{Z}| = |S| \right\}.$$

In step 3, we immediately have that

$$g_S(\boldsymbol{d}_S) = \prod_{\ell \in S} F_0$$

becomes the product of Laplace density. Then the prior form in (1.8) is obtained as the product of the above terms. □

Remark 1.1. *Note that in the limiting case $L_n = p$, the prior in the form* (1.8) *takes the same form as the prior (1.2) in Castillo et al. (2015). Similarly, the dimension prior π_{L_n} in* (1.8) *plays the same role of π_p in their seminal work and replaces π_p. Consequently, it suffices to study the properties of $\pi_{L_n}(s)$ with $L_n \to \infty$, and definitely, $p = (n-1) \to \infty$.*

In terms of the properties of dimension prior π_{L_n}, we shall show that π_{L_n} has an exponential decrease by appropriate selection of the hyperparameters (a, b) in the Gamma prior for α, given that L_n is sufficiently large. We start from the following lemma of Poisson approximation.

Lemma 1.1 (Serfling's Poisson approximation)**.** *Let $X_1, \ldots, X_n$ be (possibly dependent) Bernoulli random variables with $p_1 = Pr\{X_1 = 1\}$ and*

$$p_i = Pr\{X_i = 1 | \mathcal{F}_{i-1}\},$$

where $\mathcal{F}_i$ denotes the σ-field generated by $X_1, \ldots, X_i$. Let $W_n = \sum_{i=1}^n X_i$ and Y be Poisson *with mean* $\lambda = \sum_{i=1}^{n} E(p_i)$*. Then*

$$\frac{1}{2}\sum_{k=1}^{n} |Pr\{W_n = k\} - P\{Y = k\}| \leq \sum_{i=1}^{n} E(p_i^2) + \sum_{i=1}^{n} E|p_i - E(p_i)|.$$

The result of Lemma 1.1 will be used to prove the following proposition. Our assertions are given under any fixed L_n.

Proposition 1.2 (Exponential decrease)**.** *Let* $a = c_1 L_n^{-c_3}, b = c_2 L_n^{c_4}$ *for some constants* $c_1, c_2 > 0$ *and* $c_3 > c_4 + 1 \geq 2$ *in prior* (1.5)*. The following assertion holds as* $n, L_n \to \infty$*.*

There exists a constant $C_0 \in (0, 1)$*,*

$$\pi_{L_n}(s) \leq C_0 \pi_{L_n}(s-1), for\ s = 1, \ldots, L_n. \tag{1.9}$$

Proof. We first determine the prior π_{L_n} in Step 2. Obviously, we have

$$\pi_{L_n}(s) = \int_{\mathbb{R}} Pr\{|\boldsymbol{d}| = s|\alpha\}\pi(\alpha)d\alpha.$$

Hence we study the conditional probability $Pr\{|d| = s|\alpha\}$ first, or equivalently, $Pr\ \{|\boldsymbol{Z}| = s|\alpha\}$.

Note that η_ℓ have a Markov structure and for $\ell > 1$,

$$p_\ell^* = Pr\{Z_\ell = 1|\mathcal{F}_{\ell-1}\} = Pr\{Z_\ell = 1|\eta_{\ell-1}\} = \eta_\ell|\eta_{\ell-1}.$$

Following Teh et al. (2007, Eq. 14), given fixed α, for $\ell > 1$,

$$f(\eta_\ell|\eta_{\ell-1}) = \alpha\eta_{\ell-1}^{-\alpha}\eta_\ell^{\alpha-1} I(0 < \eta_\ell < \eta_{\ell-1}).$$

To avoid confusion, we denote $p_1^* = p_1$. Then, one can derive

$$\begin{aligned}
E(p_1^*) &= \int \alpha\eta_1^{\alpha-1} d\eta_1 = \frac{\alpha}{\alpha+1},\\
E(p_2^*) &= \int_0^1 \int_0^{\eta_1} \alpha\eta_1^{-\alpha}\eta_2^{\alpha} d\eta_1 d\eta_2 = \left(\frac{\alpha}{\alpha+1}\right)^2,\\
&\vdots\\
E(p_{L_n}^*) &= \int_{0<\eta_L<\cdots<\eta_1<1} \alpha^{L_n}\eta_{L_n}^{\alpha} \prod_{\ell=1}^{L_n-1} \eta_\ell^{-1} d\eta_1 \dots d\eta_{L_n}\\
&= \left(\frac{\alpha}{\alpha+1}\right)^{L_n}.
\end{aligned}$$

Similarly, we have

$$E(p_1^{*2}) = \frac{\alpha}{\alpha+2};\; E(p_\ell^{*2}) = \left(\frac{\alpha}{\alpha+2}\right)^\ell,\; \ell > 1.$$

We hence obtain the Poisson approximation of the probability $Pr\{|\boldsymbol{d}| = s|\alpha\}$, denoted as π^0_{α,L_n}. As $n, L_n \to \infty$, $\sum_{\ell\geq 1} E(p_\ell^*) = \alpha$. We have $\pi^0_{\alpha,\infty} = \pi^0_\alpha = \text{Pois}(\alpha)$.

By integrating out α under the Gamma prior in (1.5) we obtain the approximated form for π_{L_n}, denoted as π^0. With the hyperprior Gamma(a, b), π^0 becomes a truncated negative binomial distribution

$$\pi^0(s) \propto \frac{\Gamma(s+a)}{s!\Gamma(a)}\left(\frac{b}{b+1}\right)^s\left(\frac{1}{b+1}\right)^a, s = 0, 1, 2, \dots, L_n.$$

For some (a, b) fixed with given L_n,

$$\frac{\pi^0(s+1)}{\pi^0(s)} = \left\{1 - \frac{1-a}{s+1}\right\}\left(\frac{b}{b+1}\right), s = 0, \dots, L_n - 1.$$

And hence it naturally satisfies assertion (1.9) with $C_0 = b/(b+1)$.

By the fact that $\prod_{m=2}^M (1 - 1/m) = M^{-1}$, with $b = c_2 L_n^{c_4}$ with $c_4 \geq 1$ we have

$$\pi^0(s) \geq Q_{n,s}^{-1} s^{-1},\; s \geq 1,$$

where $Q_{n,s}$ acting as the denominator related to L_n to guarantee that $\sum_{s=1}^{L_n} \pi^0(s) = 1$. Since $\log n \leq \sum_{i=1}^{n} i^{-1} \leq 1 + \log n$, we have

$$\pi^0(s) \geq \frac{Q_0}{s(1+\log L_n)} \tag{1.10}$$

for some finite constant Q_0 unrelated to s.

We then show that the approximated distribution π^0 is sufficiently close to the true π_{L_n} and hence assertion (1.9) holds for π_{L_n}. By Jensen's inequality, for $\ell \geq 1$,

$$\begin{aligned}
E|p_\ell^* - E(p_\ell^*)| &\leq \sqrt{\text{Var}(p_\ell^*)} \\
&= \sqrt{\left(\frac{\alpha}{\alpha+2}\right)^\ell - \left(\frac{\alpha}{\alpha+1}\right)^{2\ell}} \\
&< \sqrt{\ell\left(\frac{\alpha}{(\alpha+1)^2(\alpha+2)}\right)\left(\frac{\alpha}{\alpha+2}\right)^\ell} \\
&< \ell\sqrt{\left(\frac{\alpha}{(\alpha+1)^2(\alpha+2)}\right)\left(\frac{\alpha}{\alpha+2}\right)^\ell}
\end{aligned}$$

Hence we have

$$\begin{aligned}
\sum_{\ell=1}^{L_n} E|p_\ell^* - E(p_\ell^*)| &< \sum_{\ell=1}^{\infty} E|p_\ell^* - E(p_\ell^*)| \\
&< \frac{\alpha}{(\alpha+1)(\sqrt{\alpha+2}-\sqrt{\alpha})^2} \\
&< \frac{\alpha}{(\alpha+1)^2}
\end{aligned}$$

Consequently, by Lemma 1.1, for any $s = 0, 1, \ldots, L$, we have

$$|Pr\{|d| = s|\alpha\} - \pi_{\alpha,L}^0(s)| \leq \left(1 + \frac{1}{(\alpha+1)^2}\right)\alpha < 2\alpha$$

The RHS of the above inequality is obtained by taking $L \to \infty$ on the RHS of Lemma (1.1).

Finally, we have

$$|\pi_{L_n}(s) - \pi^0(s)| = \int_0^{+\infty} |Pr\{|d| = s|\alpha\} - \pi_{\alpha,L}^0(s)|\pi(\alpha)d\alpha.$$

Again by Jensen's inequality and (1.10), for $a = c_1 L_n^{-c_3}$, $b = c_2 L^{c_4}$, and $c_3 > c_4 + 1$, we obtain

$$|\pi_{L_n}(s) - \pi^0(s)| \leq 2ab = o[\min_{s \geq 0} \pi^0(s)].$$

Consequently, for all s,

$$\lim_{L_n \to \infty} \frac{\pi_{L_n}(s+1)}{\pi_{L_n}(s)} = \frac{\pi^0(s+1)}{\pi^0(s)}.$$

Since $b/(b+1)$ is bounded away from zero, for sufficiently large L_n, assertion (1.9) always holds. □

Since Theorem 1.1 gives the same assertion as Castillo and van der Vaart (2012, Thereom 2, recovery), we only need to check their conditions.

Proof. For the support of non-zero coordinates of $\boldsymbol{d}$, the density $g_S = \prod_{s=1}^{|S|} F_0$, which is product of $|S|$ univariate densities. Meanwhile, the Laplace density naturally satisfies condition (2.3) in Castillo and van der Vaart (2012) with a finite second moment. The assertion (1.9) implies that the prior π_{L_n} on dimension has a strict exponential decrease. Furthermore, assertion (1.10) implies that

$$K_n \log(L_n/K_n) \geq M \log(\frac{1}{\pi_{L_n}(K_n)})$$

for a universal constant M. Then all conditions required by Castillo and van der Vaart (2012, Thereom 2, recovery) are satisfied. □

1.7.1.2 Proof of Theorem 1.3

We introduce some necessary notations and present some auxiliary lemmas before proving Theorem 1.3.

Under (1.6), for any given data $\boldsymbol{y}$, the difference $\boldsymbol{y}^* \sim N(\boldsymbol{d}_0, I_p)$. Let $f_{p,\boldsymbol{d}}$ be the density of $N(\boldsymbol{d}, I_p)$. For a Borel measurable subset $\mathcal{B}$ of the parameter space, the posterior probability of $\mathcal{B}$ is written as

$$\Pi_{n,L_n}(\mathcal{B}|\boldsymbol{y}^*) = \frac{\int_{\mathcal{B}} \frac{f_{p,\boldsymbol{d}}(\boldsymbol{y}^*)}{f_{p,\boldsymbol{d}_0}(\boldsymbol{y}^*)} d\pi(\boldsymbol{d})}{\int \frac{f_{p,\boldsymbol{d}}(\boldsymbol{y}^*)}{f_{p,\boldsymbol{d}_0}(\boldsymbol{y}^*)} d\pi(\boldsymbol{d})} = \frac{N_n(\mathcal{B})}{R_n}, \tag{1.11}$$

where $\pi(\boldsymbol{d})$ is the prior distribution of $\boldsymbol{d}$ given by (1.8).

We have the following lemma about the lower bound of the denominator R_n.

Lemma 1.2 (Lemma 2 in Castillo et al. (2015))**.** *For sufficiently large p and any $\mathbf{d}_0 \in \mathbb{R}^p$, with support S_0, $K_n = |S_0|$, and g_S being the product of Laplace density with scale parameter λ, we have, almost surely,*

$$R_n \geq \frac{\pi_{L_n}(K_n)}{L_n^{2K_n}} \exp(-\lambda||\boldsymbol{d}_0||_1 - 1).$$

Lemma 1.2 is similar to Lemma 2 in Castillo et al. (2015) by transferring p to L_n. The proof is analogous to theirs.

We also introduce the following lemma to learn about the tail probability of the dimension prior $\pi_{L_n}(s)$.

Lemma 1.3 (Lemma 2.1 in Ohn and Kim (2022))**.** *For any fixed α, for Z_ℓ following the prior distribution in* (1.5)*, we have for any $s \geq 0$*

$$Pr\{|\boldsymbol{Z}| > k|\alpha\} \leq \frac{14\alpha^{k+1}}{3(\alpha+1)^k}.$$

Lemma 1.3 is a special case with $\kappa = 0$ and $p = 1$ of the two-parameter construction of IBP weights in Ohn and Kim (2022). Based on Lemma 1.3, we immediately have the following corollary.

Corollary 1.4 (Tail probability of $\pi_{L_n}(s)$)**.** *Let $a = c_1 L_n^{-c_3}, b = c_2 L_n^{c_4}$ with $c_1, c_2 > 0$, $c_3 > c_4 + 2 \geq 3$ in the Gamma hyperprior in* (1.5)*. For any $k \geq 0$, $S \sim \pi_{L_n}$, as $L_n \to \infty$, we have*

$$Pr\{S > k\} = o(L_n^{-2(k+1)}).$$

Proof.

$$\begin{aligned} Pr\{S > k\} &= \int Pr\{|\boldsymbol{Z}| > k|\alpha\} \text{Gamma}(\alpha; a, b) d\alpha \\ &\leq \frac{14}{3} E\left(\frac{\alpha^{k+1}}{(\alpha+1)^k}\right). \end{aligned}$$

For any $k \geq 1$, x^{k+1}/x^k is concave and thus, by Jensen's inequality we have

$$E\left(\frac{\alpha^{k+1}}{(\alpha+1)^k}\right) \leq \frac{[E(\alpha)]^{k+1}}{[E(\alpha+1)]^k} = o(L^{-2(k+1)}).$$

□

The following lemma provides the property of the adaptive precision parameter $\lambda_n(\delta)$.

Lemma 1.4 (Adaptive $\lambda_n(\delta)$). *Given $\delta > 0$, for $\lambda_n(\delta)$ in (1.7), as $K_n/p \to 0$, $n, p, L_n \to \infty$, we have*

$$\sup_{\boldsymbol{d}_0 \in \tilde{l}_0[K_n]} P_{\boldsymbol{d}_0}\{\lambda_n(\delta)||\boldsymbol{d}_0||_1 \geq \delta\} < \frac{1}{p}.$$

Proof. As $y_i^* \sim N(d_{0i}, 1)$, $|y_i^*|$ follows a folded normal distribution so that

$$E(|y_i^*|) = \sqrt{\frac{2}{\pi}}\exp(-d_{0i}^2) + d_{0i}(1 - 2\Phi(-d_{0i})),$$
$$\text{Var}(|y_i^*|) = d_{0i}^2 + 1 - E^2(|y_i^*|).$$

For $d_{0i} = 0$, $E(|y_i^*|) = \sqrt{2/\pi} \equiv \mu_0, \text{Var}(|y_i^*|) = 1 - \mu_0^2.$; for $d_{0i} \neq 0$, as $L_n \to \infty$, $E(|y_i^*|) \to d_{0i}, \text{Var}(|y_i^*|) \to 1$. Therefore, for sufficiently large p, we have

$$E(|\bar{\boldsymbol{y}}|) \to \mu_0 + \frac{1}{p}||\boldsymbol{d}_0||_1, \text{Var}(|\bar{\boldsymbol{y}}|) \to \frac{1}{p}.$$

Then, by Chebyshev's inequality, we have

$$\begin{aligned} &P_{\boldsymbol{d}_0}\{\lambda_n(\delta)||\boldsymbol{d}_0||_1 \geq \delta\} \\ &\quad = P_{\boldsymbol{d}_0}\{|\bar{\boldsymbol{y}}| \geq \frac{1}{p}||\boldsymbol{d}_0||_1\} \\ &\quad = P_{\boldsymbol{d}_0}\{||\bar{\boldsymbol{y}}| - E(|\bar{\boldsymbol{y}}|)| \geq \mu_0\} \\ &\quad \leq \frac{1}{p\mu_0^2} < \frac{1}{p}. \end{aligned}$$

□

Now we start the proof of Theorem 1.3.

Proof. Let $\sigma(\boldsymbol{y}*)$ be the sigma field generated by the data $\boldsymbol{y}*$. Lemma 1.4 indicates that there exists a Borel set $\mathbb{B}_n \in \sigma(\boldsymbol{y}*)$ so that $P_{\boldsymbol{d}_0}(\mathbb{B}_n^c) < 1/p$ and $\lambda_n(\delta)||\boldsymbol{d}_0||_1 < \delta$ holds on $\mathbb{B}_n$.

Note that

$$\begin{aligned}
E_{\boldsymbol{d}_0}\Pi_{n,L_n}(\mathcal{B}|\boldsymbol{y}^*) &= \int \frac{N_n(\mathcal{B})}{R_n} f_{p,\boldsymbol{d}_0}(\boldsymbol{y}^*)d\boldsymbol{y}^* \\
&= Rn^{-1}\int\int_{\mathcal{B}} f_{p,\boldsymbol{d}}(\boldsymbol{y}^*)d\pi(\boldsymbol{d})d\boldsymbol{y}^* \\
&= Rn^{-1}\int_{\mathcal{B}}\int f_{p,\boldsymbol{d}}(\boldsymbol{y}^*)d\boldsymbol{y}^*d\pi(\boldsymbol{d}) \\
&= Rn^{-1}\pi(\mathcal{B}).
\end{aligned}$$

Hence, by Lemma 1.2 and Corollary 1.4, we have

$$\begin{aligned}
&E_{\boldsymbol{d}_0}\Pi_{n,L_n}\{\boldsymbol{d} : |\boldsymbol{d}| > K_n|\boldsymbol{y}^*\} \\
&\qquad \le P_{\boldsymbol{d}_0}(\mathbb{B}_n^c) + E_{\boldsymbol{d}_0}[\pi(|d| > K_n)\mathbf{1}\mathbb{B}_n] \\
&\qquad < \frac{1}{p} + R_n^{-1}\pi(|d| > K_n) \\
&\qquad \le \frac{1}{p} + Q_1 K_n \log(L_n)L_n^{-2}\exp(\lambda||\boldsymbol{d}_0||_1), \\
&\qquad < \frac{1}{p} + Q_1 K_n \log(L_n)L_n^{-2}\exp(\delta),
\end{aligned}$$

where $Q_1 = (1+\log L_n)(eQ_0\log L_n)^{-1}$ with Q_0 given by (1.10). Obviously, the RHS of the last inequality on the above tends to zero as $n, L_n \to \infty$. □

1.7.1.3 Proof of Corollary 1.3

Proof. Corollary 1.2 implies that d_i^{MAP} is a consistent estimator of d_{0i}. Therefore, with the cut-off of $\tilde{l}_0[K_n]$, it suffices to showing that, for M in Theorem 1.2,

$$\inf_{\boldsymbol{d}_0\in\tilde{l}_0[K_n]} E_{\boldsymbol{d}_0}\Pi_{n,L_n}\left\{\psi < \frac{M}{3}\sqrt{K_n\log(L_n/K_n)}|\boldsymbol{y}^*\right\} \to 1,$$

for as $n, L_n \to \infty$. Since

$$\psi_0 = p^{-1/2}||\boldsymbol{d}_0 - \bar{d}_0\mathbf{1}_p||_2 \le p^{-1/2}||\boldsymbol{d}_0||_2,$$

therefore, $3\psi_0 < M\sqrt{K_n \log(L_n/K_n)}$ by Assumption (A2).

Corollary 1.2 indicates that $\bar{d} \to \bar{d}_0$.

Then by triangle inequality, we have

$$E_{\boldsymbol{d}_0}\Pi_{n,L_n}\left\{\psi < \frac{M}{3}\sqrt{K_n \log(L_n/K_n)}|\boldsymbol{y}^*\right\} \geq$$
$$E_{\boldsymbol{d}_0}\Pi_{n,L_n}\left\{\psi_0 + p^{-1/2}||\boldsymbol{d} - \boldsymbol{d}_0||_2 < \frac{M}{3}\sqrt{K_n \log(L_n/K_n)}|\boldsymbol{y}^*\right\}.$$

Theorem 1.1 indicates that the RHS of the above inequality tends to 1. □

1.7.1.4 Proof of Theorem 1.4

Proof. It is trivial that

$$|\sum_{\ell=1}^{\infty} h_\ell I(\xi_\ell \leq t)| \leq \sum_{\ell=1}^{\infty} |h_\ell|.$$

Then, for any integers $m_1 < m_2$, we have

$$\begin{aligned} P(\sum_{\ell=m_1+1}^{m_2} |h_\ell| > \epsilon) &\leq P\left(\bigcup_{\ell=m_1+1}^{m_2} |h_\ell| > \frac{\epsilon}{m_2 - m_1}\right) \\ &\leq \sum_{\ell=m_1+1}^{m_2} P\Big(|h_\ell| > \frac{\epsilon}{m_2 - m_1}\Big) \\ &\leq \sum_{\ell=m_1+1}^{m_2} \Big[1 - F_0\Big(\frac{\epsilon}{m_2 - m_1}\Big)\Big]\eta_\ell \\ &\quad + F_0\Big(\frac{-\epsilon}{m_2 - m_1}\Big)\eta_\ell \\ &\leq 2\sum_{\ell=m_1+1}^{m_2} \eta_\ell. \end{aligned}$$

This inequality indicates that if $\sum_{\ell=1}^{\infty} \eta_\ell$ is converged, then we have $\sum_{\ell=1}^{\infty} |h_\ell|$ converged according to probability. To prove the convergence of $\sum_{\ell=1}^{\infty} \eta_\ell$, it is equivalent to prove $\sum_{\ell=1}^{\infty} E(\eta_\ell) < \infty$. Firstly, we have

$$E(\eta_\ell) = \prod_{j=1}^{\ell} E\{E(p_j|\alpha)\} = \left\{E\left(\frac{\alpha}{1+\alpha}\right)\right\}^{\ell}.$$

Then by Jensen's inequality, for any fixed a, b in the Gamma prior,

$$\sum_{\ell=1}^{\infty} E(\eta_\ell) \leq \sum_{\ell=1}^{\infty} \left\{ \frac{ab}{1+ab} \right\}^{\ell} = ab < \infty.$$

□

1.7.2 Additional simulations

1.7.2.1 Model misspecification

We conduct additional simulations under the case where our method meets with model misspecification, including heavy-tailed noises in mean-shifted models, auto-correlated noises in mean-shifted models, and an AR(2) model with structural changes. We generate simulated data under the following settings and conduct 300 Monte Carlo replicates under each setting.

(**MS.1**) Changes of means with heavy tailed noises. We generate $n = 400$ $y_i = \mu_i + \epsilon_i$, where $\epsilon_i \sim \sqrt{2}^{-1} t(4)$ are i.i.d. heavy-tailed noises. We set $K = 7$ change-points at $(50, 100, 150, 200, 250, 300, 350)$, leading to 8 segments with segment mean $\mu = (0, 1.5, 3, 1.5, 3, 0.5, 2, 0)$. This setting is similar to setting **S.1** except for the heavy-tailed noise type.

(**MS.2**) Changes of means with auto-correlated noises. We generate $n = 400$ $y_t = \mu_t + \epsilon_t$, where $\epsilon_1 \sim N(0, 1), \epsilon_t = 0.5\epsilon_{t-1} + \alpha_t$, and $\alpha_t \sim N(0, 1)$ are i.i.d. Gaussian noises. We take the same setting on the means μ as in setting **S.1**.

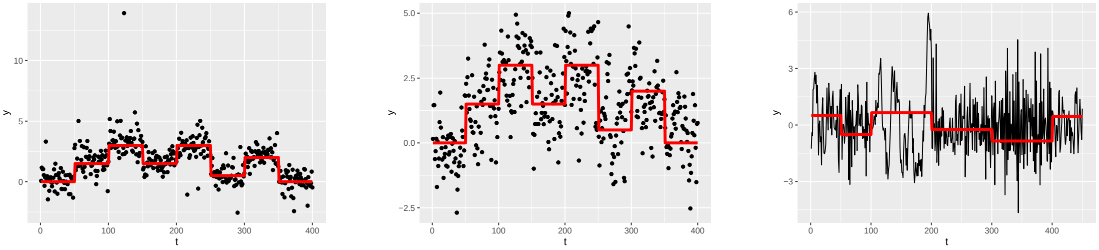
(**MS.3**) Changes of auto-correlation coefficients in mixture of AR(1) and AR(2) model.

We generate $n = 450$ observations and $y_1 \sim N(0, 1)$. For $t \geq 2$,

$$y_t = \begin{cases} 0.5y_{t-1} + \epsilon_t, t \leq 50; \\ -0.5y_{t-1} + \epsilon_t, 50 < t \leq 100; \\ 0.65y_{t-1} + 0.35y_{t-1} + \epsilon_t, 100 < t \leq 200; \\ -0.25y_{t-1} + \epsilon_t, 300 < t \leq 300; \\ -0.85y_{t-1} - 0.35y_{t-2} + \epsilon_t; 300 < t \leq 400; \\ 0.45y_{t-1} + \epsilon_t, 400 < t \leq 450. \end{cases}$$

Here $\epsilon_t \sim N(0, 1)$ are i.i.d. Gaussian noises. Under this setting, $K = 5$ change-points are located at $(50, 100, 200, 300, 400)$.

Examples of the simulated data under cases **MS.1** to **MS.2** are presented in Figures 1.9(a) to 1.9(c). In Figure 1.9(c), the red line denotes the first order auto-correlation coefficient. Note that on the interval $(100, 200)$, both the first and the second order auto-correlation coefficients are positive and hence the signs of the data on the interval are grouped together.



(a)

(b)

(c)

Figure 1.9: Examples of generated data in simulations. (a) to (c), settings **MS.1** to **MS.3**.

Besides competitors under simulation settings **S.1** to **S.5**, we add the heavy-tailed version of package `not` Baranowski et al. (2019) under setting **MS.1**, named NOT-HT; we also include a nonparametric estimator of change-point `changepoint.np` by Haynes et al. (2017) in settings **MS.1** and **MS.2**.

Results are given by Table 1.3. We find that under setting **MS.1**, NOSE is comparable with the best approach even though under model misspecifications. Under setting **MS.2**, MOSUM outperforms since it does not require independent assumptions on the data stream with shifts in the mean. Under setting **MS.3**, although `wbsts` has a higher frequency of correct detection of the number of change-points, their estimation of the locations is poor, leading to much lower precision and recall, and higher Hausdorff distance.

Table 1.3: Results of change-points detection under model mispecification settings **MS.1** to **MS.3** among 300 Monte Carlo replicates. The best results are bold.

Setting	Method	Frequency of $\hat{K} - K$							Precision	Recall	$d_H \times 10^2$
		≤ -3	-2	-1	0	+1	+2	$\geq +3$			
MS.1	NOSE	1	3	4	260	31	1	0	0.97	0.98	1.6
	NOT-HT	0	0	0	**295**	4	1	0	**0.99**	0.98	**0.9**
	SMUCE	0	0	1	107	63	59	70	0.84	0.99	3.8
	WBS	0	0	0	34	18	59	189	0.67	0.99	5.6
	FDRSeg	0	0	0	15	8	22	255	0.55	0.99	6.7
	PELT	0	0	0	73	45	87	95	0.80	0.99	3.8
	PELT-np	0	0	0	227	43	26	4	0.95	0.99	1.8
	TUGH	0	0	0	242	48	9	1	0.97	0.99	1.8
	MOSUM	0	0	3	255	41	1	0	0.98	0.99	1.9
MS.2	NOSE	0	2	19	87	89	65	38	0.70	0.80	5.2
	NOT	1	0	9	57	32	49	153	0.64	0.87	6.1
	SMUCE	0	0	1	2	7	27	264	0.55	0.91	7.5
	WBS	0	0	0	0	4	1	295	0.43	0.94	8.4
	FDRSeg	0	0	0	0	0	1	299	0.28	0.95	9.9
	PELT	4	11	28	126	83	30	18	0.79	0.83	4.6
	PELT-NP	0	1	2	46	76	68	107	0.66	0.84	5.8
	TUGH	0	0	0	1	13	14	272	0.53	0.91	7.1
	MOSUM	0	3	39	**176**	70	12	0	**0.96**	**0.91**	**4.3**
MS.3	NOSE	0	55	144	78	22	7	0	**0.83**	**0.69**	3.8
	WBSTS	14	57	84	**90**	40	15	0	0.54	0.46	7.0
	B-P	191	74	35	0	0	0	0	0.79	0.38	**2.6**

1.7.2.2 Simulations for DRAIP data

We generate a series of independent Gaussian data to simulate the DRAIP data. We generate synthetic data based on the detection result given by NOSE in the real DRAIP data. That is, 7 change-points are set at $(37, 137, 206, 336, 426, 510, 630)$. On each segment divided by these change-points, data are i.i.d. Gaussian variables with means $\mu = (0.141, 0.124, 0.399, 0.214, -0.112, -0.093, -0.053, 0.116)$ (the sample mean of the DRAIP data on each segment) and σ being the sample SDs on those segments divided by NOSE. We conduct 300 Monte Carlo replicates for the simulation. An example is presented in Figure 1.10.

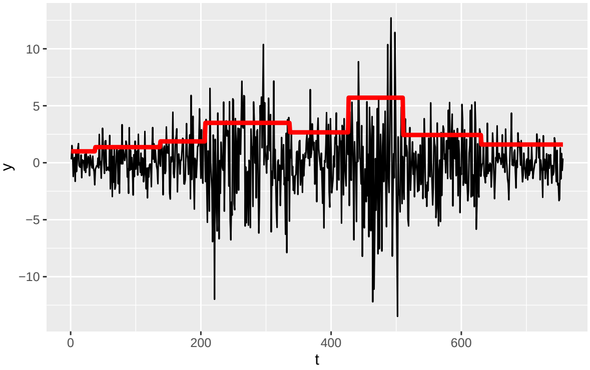


Figure 1.10: Simulated example for the DRAIP data and the true values of scale parameters (in red polyline).

We present the detection results in Table 1.4. As expected, the small jump sizes and varying means lead to serious under-detection of change-points for all approaches. Even so, NOSE performs much better in correctly detecting change-points compared with other approaches. This simulation demonstrates the reliability of detection results given by NOSE on the DRAIP data.

Table 1.4: Results of change-points detection under simulations for the DRAIP data and the ACGH data.

Setting	Method	Frequency of $\hat{K} - K$							Precision	Recall	$d_H \times 10^2$
		≤ -3	-2	-1	0	+1	+2	$\geq +3$			
DRAIP	NOSE	8	147	110	**33**	2	0	0	0.90	**0.71**	3.4
	NOT	224	60	11	5	0	0	0	0.94	0.54	**2.0**
	SMUCE	282	17	1	0	0	0	0	**1**	0.48	19.5
	PELT	95	119	78	8	0	0	0	0.92	0.64	2.6
ACGH	NOSE	0	0	1	108	140	44	7	0.93	0.99	2.5
	HSMUCE	0	0	1	35	131	102	31	0.90	0.93	15.5
	NOT	0	0	0	28	12	107	153	0.81	0.98	18.2
	R-FPOP	0	53	166	21	60	0	0	0.99	0.84	3.25
	SMUCE	0	0	0	0	0	0	300	0.51	0.98	20.9
	WBS	0	0	0	0	0	0	300	0.52	0.98	20.9
	FDRSeg	0	0	0	0	0	0	300	0.30	0.97	21.3
	TUGH	1	0	0	1	0	0	298	0.48	0.96	20.2
	MOSUM	0	0	0	3	5	34	258	0.74	0.94	13.1

1.7.2.3 Simulations for ACGH data

We generate a series of independent Gaussian data to simulate the ACGH data. We use the smooth signal estimated by `DeCAFS` (Romano et al., 2022) as the means of Gaussian variables. The scale parameter is set as the sum of the estimated standard deviations of the drift and the AR(1) noise process. An example is presented in Figure 1.11. As can be found in the figure, such a data-generating process simulates the true data quite well with an average mean square error of 0.0265 (0.001) among the simulated datasets (standard deviation in bracket). The Gaussian scheme naturally avoids most possible outliers. For comparison, we use the detection result on the real ACGH dataset given by NOSE as the golden standard. That is, 13 change-points are set at $(73, 123, 263, 342, 524, 583, 657, 745, 1724, 1906, 1965, 2041, 2143)$. Since the data stream is long, we set the window size for true positive detection as 25 in the simulation. We conduct 300 Monte Carlo replicates for the simulation. The simulation results combined in Table 1.4 show that both NOSE and R-FPOP provide consistent estimation results with that of the real-data experiment in the simulations. By removing most outliers, the results of HSMUCE tend to more similar to that of NOSE. Compared with the real-data experiment, `NOT` seems to be slightly over-detect change-points in simulations. In terms of the remaining methods, they significantly over-detect change-points in both real-data experiments and simulations. We do not incorporate the PELT method here since it fails to detect any change-points in most cases.

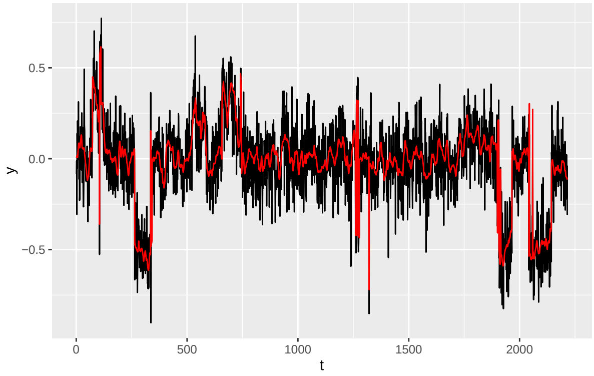


Figure 1.11: Simulated example for the ACGH data and the smooth signal estimated by DeCAFS (in red curves).

Chapter 2

Jump-size-based Bayesian Detection of Multiple Imperceptible Change-points with an Application to London House Index

2.1 Introduction

Detecting the changes of distribution properties in an ordered data series is quite common in practice and has attracted a wealth of research interest in the past several decades (Chernoff and Zacks (1964); Yao (1984); Frick et al. (2014); Fryzlewicz (2023); among others). In this chapter, we consider the following canonical mean-shifted model and its variants (Fryzlewicz (2014); Baranowski et al. (2019); Cappello et al. (2023); among others),

$$Y_t = m(t) + \sigma\epsilon_t,\ t \in [T], \tag{2.1}$$

where $m(t)$ is a *signal function* that characterizes the scheme of abrupt changes, and $[T] = \{1, \dots, T\}$. The *jump size* at each data point t is accordingly represented by the increment of

$m(t)$ as

$$\Delta m(t) = m(t+1) - m(t),\ t \in [T-1]. \tag{2.2}$$

Mathematically, the signal function $m(t)$ used to be expressed as a piecewise constant function

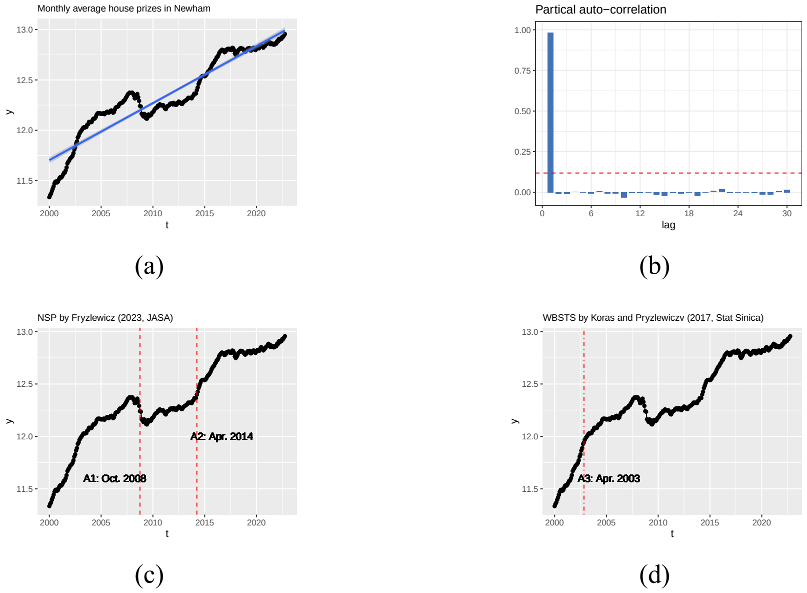
$$m(t) \equiv \sum_{k=0}^{K} \theta_k I(\tau_k \leq t < \tau_{k+1}),\ t \in [0, T], \tag{2.3}$$

where K is an unknown number of change-points, θ_k are called segment parameters, $(\tau_1, \ldots, \tau_K)$ are unknown locations of change-points (we set $\tau_0 = 0$ and $\tau_{K+1} = T$ for notation ease), and ϵ_t are assumed to be model noises with $E(\epsilon_t) = 0$ and $E(\epsilon_t \epsilon_s) = I(t = s)$. Model (2.1) might be the most common pattern characterizing the change scheme with numerous research in the past decade (Killick et al. (2012); Ko et al. (2015); Du et al. (2016); Fryzlewicz (2018); Birte and Claudia (2018); Fearnhead and Rigaill (2019); among others).

The prevailing work above and beyond looks into locations, segment parameters, or both, for change-point detection. Nonetheless, to the best of our knowledge, Cappello et al. (2023) may be the first publication to detect multiple change-points from the perspective of *jump size modeling*. Note that, a change-point is a point with a non-zero jump size. This implies that jump sizes determine the number and the locations of change-points simultaneously. Inheriting the jump-size-based spirit, we are driven to identify change-points from the insight of the *magnitude of jump sizes*. Meanwhile, we are motivated by detecting change-points for London House Index data in the past 22 years. Such index changes usually reflect the association between economics and important public events in a country.

2.1.1 Motivated example

We target to detect possible change-points in Y_t, the average monthly price (in logarithm) of all properties in the London Borough of Newham from January 2000 to October 2022 (the latest published data; 274 months in total). By quick general analytics, one finds there are THREE apparent data characteristics. From Figure 2.1(a), the plot of monthly data Y_t, i) the trend of the data has several segments, and ii) the data are overall ascending even though they have segment-wise variation; from Figure 2.1(b), the plot of partial autocorrelation function, iii) the data have significant order-1 autocorrelations.



(a)

(b)

(c)

(d)

Figure 2.1: (a) Plot of the Newham House Index of monthly average properties from January 2000 to October 2022. (b) Partial autocorrelation function plot of Newham House Index of 274 months. (c) Change-points detected by NSP (Fryzlewicz, 2023, JASA). (d) Change-points detected by WBSTS (Korkas and Pryzlewiczv, 2017, Statistica Sinica).

Figures 2.1(c) and 2.1(d) present the detected change-points given by approaches NSP (Fryzlewicz, 2023) and WBSTS (Korkas and Pryzlewiczv, 2017), respectively. However, their results are surprisingly inconsistent: NSP detects TWO change-points A1 (Oct. 2008) and A2 (Apr. 2014) while WBSTS detects only ONE change-point A3 (Apr. 2003) located at the left side of A1; the method SNCP (Zhao et al., 2022) even reports ZERO change-points. The reason for the diversity might be, based on data characteristics i) and ii), imperceptible jump sizes, or equivalently, subtle structural changes in the autocorrelations. In addition, the detection becomes more tricky owing to the moderate data size (274).

The following questions are the very first to ask. Which one is the real change-point? How to detect change-points when the jump sizes are imperceptible? How to model the imperceptible jump sizes? To address these questions, we propose a signed Beta process to model the signal of abrupt changes $m(t)$ so as to estimate the jump size process $\Delta m(t)$, the posterior estimator of which acts as the testing statistic to identify the change-points out of all data point t. Benefited from the newly constructed prior for the signal function $m(t)$, the posterior estimates of the jump sizes $\Delta m(t)$ enjoy ideal properties that they highly concentrate on zero at unchanged points and tend to be non-zero at change-points.

2.1.2 Related work

Based on the jump size defined in (2.2), one may categorize the existing approaches into two streams, *jump-size-based* or *segment-based*. The `solocp` method by Cappello et al. (2023) models the jump sizes $\Delta m(t)$ for all t directly through a *continuous spike-and-slab prior* coupled with a latent indicator. They identify change-points by comparing the posterior probability of the latent indicator with some threshold. Nevertheless, if the jump sizes are imperceptible compared to the variation of noises, the posterior probability of the latent indicator may be very small, and thus, `solocp` may fail to detect such imperceptible changes. Instead, our proposed method assigns a so-called signed Beta process (SBP) prior to the signal function $m(t)$ directly. Then we implement hypothesis testing taking a posterior estimate of jump size as the testing statistic. We establish the asymptotic efficiency of the marginal maximum a posteriori (MAP) estimates of $m(t)$ and hence asymptotic normality of the testing statistic.

Among the segment-based approaches, frequentists have contributed a large proportion. Frequentist approaches may be categorized into two strategies, model selection, and hypothesis testing. Model selection approaches associate segment-wise cost functions with different penalties on the number of change-points such as l_0 penalties (Boysen et al. (2009); Killick et al. (2012); among others) and l_1 penalties (Tibshirani et al. (2005); Lee et al. (2016); among others). Testing approaches conduct various types of segment-wise statistics including CUSUM (Korkas and Pryzlewiczv (2017); among others), multiscale (Frick et al. (2014); Jula Vanegas et al. (2021); among others), MOSUM (Hušková and Slabỳ (2001); Birte and Claudia (2018); among others), and self-normalization testing statistics (Zhao et al. (2022); among others). These approaches, though enjoy nice theoretical properties such as the minimax optimal localization rate, usually require a large data size or sufficient minimum length of segments to guarantee their performances. For example, on the detection of the structural changes of a non-stationary time series, say, the London House Index data, frequentist approaches may encounter awkward model fitting as shown in our numerical studies.

Segment-based Bayesian approaches for change-point detection may be categorized by the priors employed for change-point modeling. People may assign, i) product partition models to the number of change-points (Barry and Hartigan (1993); Monteiro et al. (2011); Quinlan et al. (2022); among others), ii) Hidden Markov models for the locations of change-points (Chib (1998); Ko et al. (2015); Bardwell and Fearnhead (2017); among others), or iii) empirical Bayes prior for the segment parameters (Du et al. (2016); Liu et al. (2017)). To avoid computation burdens brought by Markov Chain Monte Carlo (MCMC) sampling, some exact posterior computation techniques have been proposed, e.g. Fearnhead (2006) and Wyse et al. (2011). However, in our practice, we find existing Bayesian approaches are sensitive to the specification of the upper bound of the number of change-points. A poorly specified upper bound has a seriously negative impact on detection. In contrast, our proposed method does not need such kind of upper bound.

The rest of the chapter is organized as follows. Section 2.2 outlines the proposed Signed Beta Process Change-Point Modeling (SBPCPM) methodology as well as the asymptotic results. Section 2.3 describes the implementation of the proposed method. Section 2.4 makes a surrogate analytics to the London House Index data by the proposed method. Section 2.5 examines the proposed method under different simulation settings. Section 2.6 concludes the chapter with a brief discussion. Theoretical and empirical proofs are deferred to the Section 2.7.

2.2 Methodology

Suppose one observes a series Y_t, for $t \in [T]$. Under model (2.1), data $\{Y_t\}_{t=1}^{T}$ can be viewed as a series of discrete noisy observations from a continuous signal process $m(t)$. Recall that in equation (2.2), a change-point is defined as a point t where the jump size $\Delta m(t) = m(t+1) - m(t) \neq 0$. Hence, we transfer change-point detection to the following hypothesis testing problem based on the magnitude of jump sizes $\Delta m(t)$,

$$H_0 : \quad \Delta m(t) = 0 \quad \text{v.s.} \quad H_1 : \quad \Delta m(t) \neq 0, \tag{2.4}$$

for $t \in [T-1]$. Clearly, the *null* is equivalent to the statement *"t is an unchanged point"*, and the *alternative* one is equivalent to the statement *"t is a change-point"*. Then the remaining problems are what kind of statistics and tests to be constructed.

In this section, we propose a Z-type test for the testing problem (2.4). The statistic and the test are constructed in the following steps.

Step 1. Prior elicitation and posterior estimates. We formulate a Signed Beta process (SBP) as the prior for the change signal function $m(t)$ and obtain the marginal MAP estimates of $m(t)$, denoted as $\hat{m}^{\text{MAP}}(t)$, for $t \in [T]$.

Step 2. Testing statistic. Let $\zeta_t = \hat{m}^{\text{MAP}}(t+1) - \hat{m}^{\text{MAP}}(t)$ for $t \in [T-1]$ be the posterior estimate of the jump size $\Delta m(t)$. We use ζ_t as the testing statistic based on the result of Corollary 2.1.

Step 3. Z-type test. For a given significance level α, the rejection region is defined as

$$\mathcal{C} = \{|\zeta_t - E(\zeta_t)| > \Phi_{1-\alpha/2}\text{SD}(\zeta_t)\},\ t = 2, \ldots, (T-1),$$

where $\text{SD}(\zeta_t)$ denotes the standard deviation of ζ_t and $\Phi_{1-\alpha/2}$ denotes the $(1-\alpha/2)$ upper quantile of a standard normal distribution.

In the following, we study the above steps in detail. In subsection 2.2.1, we introduce the nonparametric SBP prior for $m(t)$. In subsection 2.2.2, we study the asymptotic distribution of ζ_t under the null hypothesis. In subsection 2.2.3, we discuss the Type I error of the Z-type test.

2.2.1 Signed Beta process

We start from modeling the infinitesimal increments of the function $m(t)$, denoted as $dm(t) = m(t+) - m(t)$ for any $t \in (0, T]$. Note that $dm(t)$ should be *sign-varying* since the jumps at change-points may either be upward or downward. Therefore, inspired by the commonly used Beta process (Hjort, 1990), We formulate a so-called *signed Beta process*, denoted $m(t) \sim \text{SBP}(c_1, c_2, B_1, B_2, p_t)$, such that the infinitesimal increments of $m(t)$ takes the following mixture form

$$dm(t) = Z_t dB_1(t) + (Z_t - 1)dB_2(t), \tag{2.5}$$

where $Z_t \sim \text{Bernolli}(p_t)$ is a process of Bernoulli random variables with probability p_t, and $dB_j(t)$ are defined by two independent Beta processes, denoted as $B_j(t) \sim \text{BP}(c_j, B_j)$. Here c_j is the concentration parameter and B_j is a base measure that is continuous, nonnegative, and monotonic, for $j = 1, 2$. By definition, the distribution of $dB_j(t)$ is

$$dB_j(t) \sim \text{Beta}[c_j \cdot dB_j(t), c_j\{1 - dB_j(t)\}].$$

One may explain (2.5) as follows: the increment $dm(t)$ either takes an upward jump with random height drawn from Be$[c_1 dB_1(t), c_1\{1 - dB_1(t)\}]$ at the rate p_t, or takes a downward jump with random height drawn from Be$[c_2 dB_2(t), c_2\{1 - dB_2(t)\}]$ at the rate $(1 - p_t)$. Therefore, we name such a prior for $m(t)$ as the signed Beta process in the sense that it looks like a Beta process with random signs.

Now we specify the parameters $(c_1, c_2, B_1, B_2, p_t)$ under the change-point detection background. We first specify the probability process $\{p_t\}$ to make sure that the SBP is fully determined by the two Beta processes in (2.5). In other words, we search for a kind of p_t so that given $B_1(t)$ and $B_2(t)$, the indicator process Z_t is not stochastic. Otherwise, the randomness of the indicator process Z_t may lead to a complicated sampling of the SBP.

Recall the atomic form of the Beta process (Paisley and Jordan, 2016)

$$B_j = \sum_{h=1}^{\infty} \pi_{jh} \delta_{\xi_{jh}}, \tag{2.6}$$

where $\xi_{jh} \sim B_j / \int_0^T dB_j(t)dt$ are the i.i.d. atoms of the process $B_j(t)$ and π_{jh} are the weights on ξ_{jh}. Let $\{\xi_{jh}\}$ be the collection of atoms of the process $B_j(t)$ and $\Omega = \{\xi_{1h}\} \cup \{\xi_{2h}\} \subset (0, T)$ be the union of the two collections of atoms of the two Beta processes. Since both B_1 and B_2

are continuous, we have, almost surely,

$$\{\xi_{1h}\} \cap \{\xi_{2h}\} = \emptyset.$$

That is, almost surely, the two Beta processes $B_j(t)$ in (2.5) will not jump simultaneously. Consequently, we set

$$p_t = dB_1(t)/\{dB_1(t) + dB_2(t)\},\ t \in \Omega, \tag{2.7}$$

and set $p_t = 1$otherwise. Under this setting, for $t \in \{\xi_{1h}\}$, $Z_t = 1$ and the SBP takes the same upward jumps as the Beta process $B_1(t)$; for $t \in \{\xi_{2h}\}$, $Z_t = 0$ and the SBP takes the opposite downward jumps as the Beta process $B_2(t)$; for t at other locations, neither $B_1(t)$ nor $B_2(t)$ jumps, so does the SBP.

Accordingly, based on the atomic expression of the Beta process, under setting (2.7), the following atomic expression of the SBP holds almost surely

$$m = \sum_{h=1}^{\infty} \pi_{1h} \delta_{\xi_{1h}} - \sum_{h=1}^{\infty} \pi_{2h} \delta_{\xi_{2h}}, \tag{2.8}$$

where π_{jh} and ξ_{jh} are the same as that in (2.6). Based on the atomic expression (2.8), we obtain the following sampling scheme for $m(t) \sim \text{SBP}(c_1, c_2, B_1, B_2)$.

$$m(t)|\{B_1(t), B_2(t)\} = B_1(t) - B_2(t),\ B_j(t) \sim \text{BP}(c_j, B_j),\ j = 1, 2. \tag{2.9}$$

Under the sampling scheme (2.9), the sampling of the SBP is easy to conduct by introducing the finite approximation and the stick-breaking construction of the Beta process. We defer the sampling details to Section 2.3.

Next, we specify the remaining parameters in the SBP. From the perspective of an objective Bayesian, we do not incorporate any prior information about the jumps of $m(t)$ to the SBP. Therefore, we consider a “neutral” choice of (c_1, c_2, B_1, B_2) so that $c_1 = c_2 = c_0, B_1 = B_2 = B_0$. That is, $B_1(t)$ and $B_2(t)$ in (2.9) are two i.i.d. stochastic processes. Then the prior for $m(t)$

is simplified as

$$m(t) \sim \text{SBP}(c_0, B_0).$$

With the above specification of the SBP, we immediately have the following proposition.

Proposition 2.1. *For $m(t)$ sampled from (2.9) with $c_1 = c_2, B_1 = B_2$, the prior distribution of the jump size $\Delta m(t) = m(t+1) - m(t)$ for any $t \in [T]$ is symmetric and*

$$E\{\Delta m(t)\} = 0, \; \mathit{Var}\{\Delta m(t)\} = 2\mathit{Var}\{B_1(t)\}.$$

The proof of Proposition 2.1 is trivial since $E(\Delta B_1(t)) = E(\Delta B_2(t))$ and $\text{Var}\{B_1(t)\} < \infty$. Since the jump sizes of any Beta process are independent, this proposition indicates that under the SBP, the jump sizes $\Delta m(t)$ are independent continuous variables for all t. These properties enable us to study the asymptotic distribution of the posterior estimates of $m(t)$, the foundation of our test for change-point detection.

2.2.2 Asymptotic results of posterior estimates

In this subsection, we study the distributional approximation to $\hat{m}^{\text{MAP}}(t)$, the marginal MAP estimates of $m(t)$ when the null hypothesis $H_0 : \Delta m(t) = 0$ in test (2.4) holds for all t. That is, there are no change-points along $[0, T]$, that is,

$$m(t) = \theta_0, \; t \in [0, T].$$

In this case, we show that asymptotically, $\hat{m}^{\text{MAP}}(t)$ follows a Gaussian distribution with the same center independently for all t. Hence, the distribution of the statistics ζ_t under the null hypothesis H_0 can be approximated by a Gaussian distribution centered at zero. That explains why we construct a Z-type test for testing (2.4). Our asymptotic results are established under the mean-shifted model (2.1).

We require the following general assumptions to ϵ_t and σ.

(A1) The noises ϵ_t are i.i.d continuous variables.

(A2) $\sigma > 0$ is a known constant.

Assumptions (A1) and (A2) are widely adopted in change-point literature about mean-shifted models, e.g. Frick et al. (2014); Baranowski et al. (2019); Cappello et al. (2023); among others. Under the above two assumptions, we obtain the density of Y_t as

$$f_Y(y) \equiv f_Y(y;\theta_0,\sigma) = f_\epsilon(\frac{y-\theta_0}{\sigma}).$$

We further require the following assumptions on the functional form of $f_Y(y;\theta_0,\sigma)$.

(A3) For any fixed σ, $f_Y(y;\theta_0,\sigma)$ is twice differentible with respect to θ_0; the Fisher information of $f_Y(\theta_0,\sigma)$ exists for all $\theta_0 \in \mathbb{R}$, denoted as $I_Y(\theta_0)$.

(A4) $f_Y(y;\theta_0,\sigma)$ is log-concave with respect to $y \in \mathbb{R}$.

Assumptions (A3) and (A4) are satisfied by a wide range of location-scale families including the Gaussian, Laplace, and logistic families, etc.

Then we obtain the following asymptotic results on the marginal MAP estimates of $m(t)$, denoted as $\hat{m}^{\text{MAP}}(t)$.

Theorem 2.1 (Asymptotic efficiency of MAP estimates)**.** *Under model* (2.1)*, suppose the null hypothesis H_0 in* (2.4) *holds for all $t = 2,\ldots,(T-1)$. That is, there are no change-points. Under Assumptions (A1) to (A4), as $T \to \infty$, for any fixed t, we have*

$$\hat{m}^{MAP}(t) \xrightarrow{p} \theta_0, \quad \sqrt{T}(\hat{m}^{MAP}(t) - \theta_0) \xrightarrow{d} N(0, I_Y^{-1}(\theta_0)).$$

Theorem 2.1 tells that when there are no change-points, the marginal MAP estimates of $m(t)$ are pointwisely consistent and asymptotically normal with efficient variance. This is not surprising based on the results of Proposition 2.1. The detailed proof is deferred to the subsection 2.7.1.

Based on Theorem 2.1, instantly we have the following corollary about the asymptotic distribution of the statistics $\zeta_t = \hat{m}^{\text{MAP}}(t+1) - \hat{m}^{\text{MAP}}(t)$.

Corollary 2.1 (Asymptotic normality of the posterior estimator of jump sizes)**.** *For any fixed* $t \in [T-1]$*, as* $T \to \infty$*,*

$$\zeta_t \xrightarrow{d} N(0, 2T^{-1}I_Y^{-1}(\theta_0)). \tag{2.10}$$

This asymptotic normal result enables us to approximate the distribution of ζ_t through a normal distribution when there are no change-points under the null hypothesis in hypotheses (2.4).

When there exist some change-points, the asymptotic normality of the MAP estimates will not hold. Rather, we empirically observe a mode-shifting phenomenon around change-points.

Proposition 2.2 (Mode-shifting at change-points)**.** *Let* $\mathcal{A}_k = (\tau_k - \delta, \tau_k + \delta)$ *be a* δ *neighborhood of the* k*th change-point* τ_k *for some* $\delta > 0$*. Then for* $t \in \mathcal{A}_k$ *the marginal posterior distributions of* $m(t)$ *are bimodal, and*

$$\hat{m}(t_2)^{MAP} - \hat{m}(t_1)^{MAP} \neq 0,$$

for $t_1, t_2 \in \mathcal{A}_k$ *and* $t_1 < \tau_k < t_2$*.*

We admit we could not provide rigorous proof for the distribution of the testing statistic under the alternative hypothesis. The phenomena of bi-modal posterior and mode-shifting on change-points are observed and summarized in Proposition 2.2 and evidenced by simulations in subsection 2.7.2. From our numerical results, for $t_1 < \tau_k < t_2$, we observe that

$$\hat{m}(t_1)^{\text{MAP}} \approx m(\tau_{k-1}); \quad \hat{m}(t_2)^{\text{MAP}} \approx m(\tau_k)$$

within the δ neighborhood of $\mathcal{A}_k$. In this sense, Proposition 2.2 is direct since the mode of the marginal posterior distribution of $m(t)$ shifts from $m(\tau_{k-1})$ to $m(\tau_k)$ when crossing the change-point τ_k. Compared with Corollary 2.1 which derives the distributional approximation of ζ_t under the null hypothesis, Proposition 2.2 tells the behavior of the testing statistic under the alternative hypothesis. These two results explain the rationale of the aforementioned Z-type test.

2.2.3 Lower Type I error

When there exists at least one change-point i.e. no less than two segments, the distributional approximation in Theorem 2.1 does not hold for all unchanged points since the segment parameters are not identical. However, we still empirically find that, as the minimum length of segments increases, the marginal MAP estimates at unchanged points still tend to zero with similar variance. Thus, the empirical distribution of the difference of marginal MAP estimates ζ_t at unchanged points can still be roughly approximated by a normal distribution. Accordingly, the rejection region of the Z-type test is constructed as

$$\mathcal{C} = \{\zeta_t - \bar{\zeta}_t > \Phi_{1-\alpha/2}\hat{\sigma}_0\}, \tag{2.11}$$

where $\bar{\zeta} = (T-1)^{-1}\sum_{t=1}^{T-1}\zeta_t$ and $\hat{\sigma}_0 = \sqrt{(T-1)^{-1}\sum_{t=1}^{T-1}(\zeta_t - \bar{\zeta})^2}$ denote the sample mean and standard deviation of ζ_t respectively.

Note that the sample standard deviation $\hat{\sigma}_0$ is easily affected by those points where ζ_t are apparently non-zero. As a result, when there are several change-points, $\hat{\sigma}_0$ will exceed the $\mathrm{SD}(\zeta_t)$ at unchanged points. Therefore, the rejection region $\mathcal{C}$ in (2.4) is a kind of “parsimonious” rejection region satisfying the following proposition.

Proposition 2.3 (Lower Type I error). *The Type I error for the hypotheses* (2.4) *with the rejection region* (2.11) *is lower than the nominal significance level* α*.*

The Type I error and the power of the test for hypotheses (2.4) have a clear interpretation in the context of change-point detection. On the one hand, a Type I error indicates that an unchanged point is falsely identified as a false positive (FP) change-point. Hence, controlling the Type I error is equivalent to controlling the risk of over-detection of change-points. In this sense, Proposition 2.3 ensures a lower risk of over-detection than the nominal level. On the other hand, the power of the test is equivalent to the rate of correctly detecting true positive (TP) change-points. That is, a higher power indicates a higher capability of correctly specifying both the number and locations of change-points. In our simulations, we find SBPCPM enjoys sufficient power even under a pretty high significance level (lower Type I error), demonstrating the detection performance.

2.3 Implementation

In this section, we discuss how to implement SBPCPM in change-point detection. Subsection 2.3.1 presents the stick-breaking construction of the SBP and the choice of remaining parameters in the chapter. Subsection 2.3.2 introduces posterior sampling and a post-process on the detected change-points. Subsection 2.3.3 provides the estimation procedure of SBPCPM for segment parameters.

2.3.1 Stick-breaking construction of the SBP

The form of infinite sum in (2.8) is difficult to be implemented in practice. Thus, we consider the finite approximation in implementation. We start from following finite approximation of the Beta process (Paisley et al., 2010). Let $\gamma = B_0(T)$ be the total mass of B_0 on T. The finite approximation of Beta process $B_j \sim \text{BP}(c, B_0)$ is

$$B_j^{(M)} = \sum_{h=1}^{M} \pi_{jh}\delta_{\xi_{jh}},\ \pi_{jh} \sim \text{Beta}(c\gamma/M, c(1-\gamma/M)).$$

According to Paisley et al. (2010), as $M \to \infty$, this approximation converges to the infinite atomic sum expression in (2.6) (Paisley et al., 2010). To simplify the computation, we fix the variation parameter $c = 1$ in this chapter. Then, the stick-breaking construction of the Beta process has a close form (Teh et al., 2007)

$$B = \sum_{h=1}^{M} \pi_h\delta_{\xi_h},\ \pi_1 = v_1 \sim \text{Beta}(\gamma, 1),\ \pi_h = \prod_{\ell=1}^{h-1} v_\ell,\ v_\ell \sim \text{Beta}(\gamma, 1). \tag{2.12}$$

Consequently, based on the sampling scheme (2.9), we obtain the following stick-breaking construction of the SBP$(1, B_0)$. For $j = 1, 2$, we have

$$m^{(M)} = \sum_{h=1}^{M} \pi_{1h}\delta_{\xi_{1h}} - \sum_{h=1}^{M} \pi_{2h}\delta_{\xi_{2h}}, \pi_{j1} = v_{j1} \sim \text{Beta}(\gamma, 1),\ \pi_{jh} = \prod_{\ell=1}^{h-1} v_{j\ell},\ v_{j\ell} \sim \text{Beta}(\gamma, 1), \tag{2.13}$$

where $m^{(M)}$ is a finite truncation of the atomic form (2.8). Under the stick-breaking construction (2.13), we need to specify the base measure B_0. Recall that the atoms ξ_{jh} are i.i.d. samples of the normalized measure $\sim B_0/B_0(T)$. Note that one is expected to assign a uniform prior for the atoms ξ_{jh} since there is no available information about the locations of change-points. Hence, we set the base measure $B_0(t) = \gamma t$ as a linear function on $(0, T]$, where the total mass α is a hyperparameter. We assign a Gamma hyperprior Gamma(a, b) for γ without specifying it.

2.3.2 Posterior inference and post-process on change-points

We conduct MCMC sampling by `nimble` (de Valpine et al., 2017) package in R (R Core Team, 2021), which uses `BUGS` type syntax (Lunn et al., 2000) and compiles the code into `C++` to facilitate automatic posterior sampling. Then the marginal MAP estimates $\hat{m}^{\mathrm{MAP}}(t)$ is obtained by the mode of the empirical density of the marginal MCMC samples.

Note that Proposition 2.2 tells that some marginal MAPs within the δ neighborhood of change-points are non-zero. That brings the risk of over-detection in the sense that testing (2.4) may falsely identify the points around change-points as change-points. In other words, our testing procedure may generate consecutive change-points. To overcome this, we impose a post-process on those change-points identified by testing (2.4). Here we adopt the common practice in the literature that requires a minimum distance between change-points (Matteson and James (2014); Cappello et al. (2023); among others). That is, for two change-points $\hat{C}_1 < \hat{C}_2$ identified by testing (2.4), if $|\hat{C}_1 - \hat{C}_2| < D_0$, where D_0 is a prespecified minimum distance between two change-points, we remove $\hat{C}_2$ so as to avoid consecutive change-points. We set $D_0 = 10$ as the minimum distance between change-points throughout this chapter.

2.3.3 SBPCPM algorithm

We summarize the whole procedure of SBPCPM into Algorithm 1. The hypothesis testing procedure is executed in line 4, the post-process on change-points is executed from lines 5 to 9, and the estimated number and the locations of change-points, $\hat{K}$ and $\hat{\tau}_k$, are output in line 10. If one has further interest in the segment parameters like θ_k, SBPCPM outputs their estimates in line 12.

Algorithm 1 SBPCPM algorithm

```
Input: Data $(Y_1, \dots, Y_T)$; Confidence level $\alpha$; Gamma hyperparameters $(a, b)$; minimum distance $D_0$.
Output: Number of change-points $\hat{K}$; locations of change-points $\mathcal{C}$; segment parameters $\hat{\theta}_k$.
 1: Assign SBP$(1, B_0)$ for $m(t)$, where $B_0(t) = \gamma t$ and $\gamma \sim$ Gamma$(a, b)$.
 2: Draw posterior samples of $m(t)$ by MCMC and obtain marginal MAP estimates $\hat{m}^{\text{MAP}}(t)$.
 3: Compute the series of MAP differences $\zeta_t = \hat{m}(t+1)^{\text{MAP}} - \hat{m}^{\text{MAP}}(t), t = 1, \dots, (T-1)$.
 4: Conduct hypothesis testing (2.4) to all $t$ and obtain the change-point set $\hat{\mathcal{C}} = \{\hat{\tau}_1 < \hat{\tau}_2 < \dots < \hat{\tau}_{K_0}\}$.
 5: for $k = 1, \dots, (k_0 - 1)$, do
 6:     if $|\hat{\tau}_1 - \hat{\tau}_2| < D_0$ then
 7:         Remove $\hat{\tau}_{k+1}$.
 8:     end if
 9: end for
10: $\hat{K} \leftarrow K_0, \tau_0 \leftarrow 0, \tau_{K+1} \leftarrow T, \mathcal{C} \leftarrow \hat{\mathcal{C}}$.
11: Obtain segmentation set $\{S_k = (C_k, C_{k+1})\}_{k=0}^{\hat{K}}$.
12: $\hat{\theta}_k = T_k^{-1} \sum_{t \in S_k} \hat{m}^{\text{MAP}}(t)$, where $T_k = C_{k+1} - C_k$.
```

In our numerical experience, SBPCPM is not sensitive to the choice of the significance level α. In practice, we suggest a pretty high significance level $\alpha = 0.003$. Compared with the commonly used significance level $\alpha = 0.05$, we find the higher significance level does not impact the power too much, especially when the changes are imperceptible.

2.4 Application to London House Index

We analyze the London House Index data in this section. Based on Figure 2.1(b), the London House Index in Newham can be modeled by an AR(1) model. Hence, we focus on a special case of model (2.1), a piecewise AR(1) model as follows,

$$Y_t = a_0 + m(t)Y_{t-1} + \sigma\epsilon_t,\ t = 2, \dots, T. \tag{2.14}$$

In this autoregression model, the signal function $m(t)$ represents the autocorrelation coefficient, or, acts as a regression coefficient in a simple linear model, compared to the "intercept" role in model (2.1). Therefore, in model (2.14), $m(t)$ reflects the structural changes in the autocorrelations, similar to SNCP (Zhao et al., 2022). Both WBSTS (Korkas and Pryzlewiczv, 2017) and NSP (Fryzlewicz, 2023) allow the intercept a_0 to vary along with t, which is beyond the scope of the proposed SBPCPM. Even so, we show that SBPCPM provides a reasonable segmentation and better model fitting with a different interpretation in (2.15).

Under model (2.14), we apply SBPCPM to analyze the London House Index in Newham. We provide the exact segmentation result, estimated autocorrelations, jump sizes, intercept, and the scale of model error estimated by SBPCPM in Table 2.1. According to the segmentation given by SBPCPM, the first segment (Jan 2001 to May 2002) includes the September 11 attacks, the second segment (Jun 2002, Apr 2014) includes the host of the 2012 Olympic Games, the third includes 2014 Scottish independence referendum, and the last includes the time of Britain's EU membership referendum. The COVID-19 pandemic seems to have little impact on the London House Index since no change-points are detected during the pandemic.

Table 2.1: Exact time segments, autocorrelation, jump sizes, intercept, and scale of model error estimated by SBPCPM.

$\hat{\sigma}$	$\hat{a_0}$	Time segment	estimated autocorrelation	Jump size
0.0104	11.279	(Jan 2001, May 2002)	0.0246	-
		(Jun 2002, Apr 2014)	0.0759	0.0513
		(May 2014, Sep 2015)	0.1016	0.0257
		(Oct 2015, Oct 2022)	0.1201	0.0185

2.4.1 Locations of change-points

We present the results of change-point detection given by the prosed SBPCPM in Figure 2.2 and compare the result with the change-points detected by NSP and WBSTS. As the picture shows, SBPCPM detects three change-points, labeled as C1, C2, and C3. The change-point C1 is close to the change-point A3 detected by WBSTS, and change-point C2 is the same as change-point A2 detected by NSP. Besides the above consistent results, SBPCPM brings us two new insights.

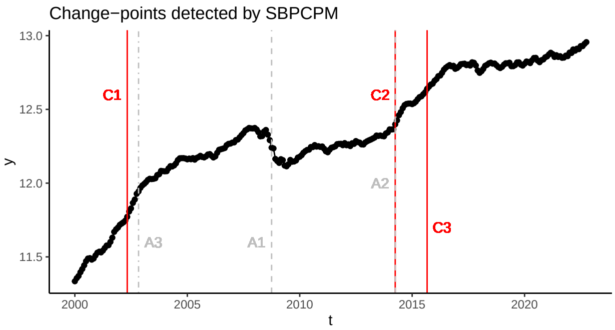


Figure 2.2: Point-line: the original data; vertical lines (red): locations of change-points detected by SBPCPM; vertical lines (grey): locations of change-points detected by other approaches.

Insight 1: three change-points.

The change-points detected by SBPCPM participates the time period into four segments, including a short segment from May 2014 (C2) to Sep 2015 (C3). A natural question is whether it is reasonable to detect three change-points. To some extent, one may find that the data curve becomes less “trending” after Sep 2015 (or sometime later) compared with the curve within the segment between May 2014 and Sep 2015. For further convincing, we consider Bayesian model comparison by comparing the Bayes factor of change-point models (Chib, 1998) implemented by R package `MCMCpack` (Martin et al., 2011). We specify the number of change-points from $K = 1, 2$, to 3 with outer parameters set as default. The matrix of log Bayes factor is given in Table 2.2, which provides strong evidence that the choice of $K = 3$ beats $K = 2$ and anecdotal evidence that $K = 3$ beats $K = 1$. Therefore, we think the number of change-points detected by SBPCPM is appropriate.

Table 2.2: Log Bayes factor matrix given by Chib (1998).

	K=1	K=2	K=3
K=1	0.00	16.50	-1.58
K=2	-16.50	0.00	-18.08
K=3	1.58	18.08	0.00

Insight 2: Oct 2008 may not be a structural change in the autocorrelations.

The other new insight brought by SBPCPM is that the lag in Oct 2008 (point A1 in Figure 2.2) may not be a change-point in the autocorrelations. Indeed, one may conjecture that the lag on the data curve is induced by the shift in the intercept term since the slopes of the curve before and after the time point seems to be similar to each other. The shift in the intercept a_0 in model (2.14) actually changes the mean of Y_t (see the next subsection for details). This may be evidenced by SNCP (Zhao et al., 2022). Although SNCP reports no change-points in the autocorrelations, it detects two change-points in the mean. As shown in Figure 2.3, the two change-points in the mean detected by SNCP are located close to the locations of change-points A1 and A2 detected by NSP. Note that NSP detects both the changes in auto-correlations and the intercept. Hence, we may conclude that the structural break in Oct 2008 is induced by the shift in the intercept.

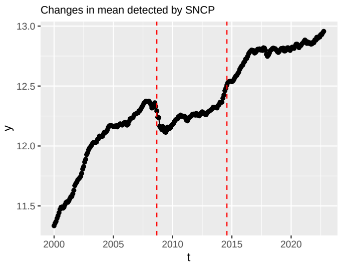


Figure 2.3: Point-line: the original data. Vertical lines (dashed): locations of change-points in mean detected by SNCP.

Unlike NSP, SBPCPM only detects the changes in the autocorrelations while keeping the intercept term fixed under model (2.14). Even though the model setting (2.14) seems to violate this real data scenario, we show that SBPCPM provides a better model fitting than the piecewise OLS fit under NSP segmentation with appropriate interpretation in the following.

2.4.2 Interpretation of estimated autocorrelations and model fitting

Compare to the piecewise OLS estimators under NSP segmentation in Table 2.3, the segment-wise autocorrelations estimated by SBPCPM in Table (2.1) are pretty small while the norm of the fixed intercept term is much greater.

Table 2.3: Piecewise OLS estimators under NSP segmentation.

Time segment	OLS auto-correlation	OLS intercept
(Jan 2000, Oct 2008)	0.979	0.264
(Sep 2008, May 2014)	1.027	-0.331
(Jun 2014, Oct 2022)	0.962	0.489

This is not surprising since the abrupt changes take place on both the intercept a_0 and the autocorrelation signal $m(t)$ in model (2.14) simultaneously, while the SBPCPM addresses the detection of change-point on a univariate parameter only. As a result, the autocorrelations estimated by SBPCPM should have a different interpretation compared to the piecewise OLS estimators.

Let $\mu_t = [1 - m(t)]a_0$. Model (2.14) can be rewritten as its centered form

$$y_t = \mu_t + m(t)(y_{t-1} - \mu_t) + \sigma\epsilon_t,\ t = 2, \ldots, T. \tag{2.15}$$

Since $E(y_t) = \mu_t$, model (2.15) becomes a mean-shifted model with auto-correlated noises. In this sense, with a fixed a_0, the abrupt change signal $m(t)$ estimated by SBPCPM actually reflects the changes in the mean process μ_t. Let $\hat{m}(t)$, $\hat{a}_0$, and $\hat{\mu}_t = [1 - \hat{m}(t)]\hat{a}_0$ denote the estimated change signal, estimated intercept, and estimated mean process given by SBPCPM respectively. Since the $|\hat{m}(t)|$ are pretty small, the corresponding estimated mean process $\hat{\mu}_t$ is indeed highly auto-correlated. Consequently, the noise process in model (2.15) becomes weakly correlated and hence, stationery. That is, most of the autocorrelations are interpreted by $\hat{\mu}_t$ given by SBPCPM. As a result, SBPCPM enjoys better model fitting compared to piecewise OLS estimators under NSP segmentation from the following two aspects.

Residual analysis

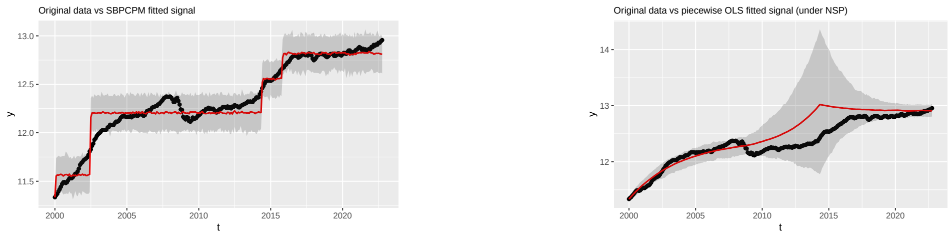
We summarize the model fitting results of SBPCPM and the piecewise OLS estimators under NSP segmentation in Table 2.4. SBPCPM enjoys a lower Root Mean Square Error (RMSE) between the fitted curve and the true data curve, indicating a more accurate model fitting. Meanwhile, we conduct the Augmented Dickey–Fuller test on the residual processes of SBPCPM and OLS estimators. The ADF test on the residual process of SBPCPM rejects the null hypothesis and is in favor of the alternative that the residual process is stationary, while the residual process of OLS estimators is non-stationary. That explains why the OLS estimators poorly fit the trend of the true data.

Table 2.4: Comparison of model fitting residuals between SBPCPM and piecewise OLS (under NSP segmentation).

	SBPCPM	OLS (under NSP)
RMSE	0.096	0.110
Residual ADF p-value	0.025	0.338

Trend fitting and inference

We present the fitted curves given by SBPCPM and the piecewise OLS estimators under NSP segmentation in Figures 2.4(a) and 2.4(b), respectively. One clearly finds that SBPCPM captures the trend of the data well while the piecewise OLS estimators poorly fit the trend. For inference comparison, we sequentially add i.i.d. $N(0, 0.05^2)$ noises to the curves fitted by SBPCPM and the piecewise OLS estimators respectively and obtain 200 synthetic datasets. The 95% empirical confidence bands of the synthetic dataset generated from SBPCPM and the piecewise OLS estimators are presented in the shaded areas of Figures 2.4(a) and 2.4(b), respectively. The empirical confidence band given by SBPCPM covers the true data curve well with a moderate width, while the confidence band given by piecewise OLS estimators is much wider due to the properties of the non-stationary process. Hence, the fitted curve of SBCPM is better from the perspective of inference.



(a)

(b)

Figure 2.4: Synthetic data versus the original data. Point-curve: original data. Polyline (red): pointwise mean of synthetic data. Shaded area: empirical 95% confidence band.

2.5 Simulations

To illustrate the proposed SBPCPM method, we carry out several simulations and compare SBPCPM with existing approaches. For the sake of comparison, we consider the following three assessments:

- The difference between the estimated number $\hat{K}$ and true number of change-points K, denoted as $|\hat{K} - K|$.
- Type I error or False positive rate (FPR) of change-point detection. For each change-point, we consider the window of width 10 around it as the set of change-points. The remaining data points are recognized as the set of unchanged points. We count a false positive (FP) change-point when an unchanged point is falsely identified as a change-point. The FPR of change-point detection is defined as the ratio between the FP number and the number of unchanged points.
- Power or True positive rate (TPR) of change-point detection. For each change-point, we count a true positive (TP) change-point if a detected change-point lies within the window of width 10. The TPR is computed as the ratio between the number of the true positive and the number of change-points K.

In subsection 2.5.1, we validate our analysis results through simulations on the synthetic data generated from the detection results in Section 2.4. In subsection 2.5.2, we present simulation results of detecting noticeable structural changes under piecewise AR(1) models. In subsections

2.5.3 and 2.5.4, we carry out simulations under model (2.1) with noticeable shifts and imperceptible shifts, respectively. Under all simulation settings, we examine the performance of our method under two significance levels, $\alpha = .003$ and $\alpha = .05$.

2.5.1 Setting I: simulations on synthetic London House Index data

The similarity between the synthetic datasets generated from the SBPCPM fit and the real data enables us to carry out simulations for validation purposes. We set the results in Table 2.1 as the ground truth because the synthetic dataset is generated from the SBPCPM fit.

For comparison, we compare the following approaches implemented by corresponding R packages.

- The NSP method (Fryzlewicz, 2023) implemented by CRAN package `nsp` (Fryzlewicz, 2021).
- The SNCP method (Zhao et al., 2022) implemented by package `SNSeg`.
- The WBSTS method (Korkas and Pryzlewiczv, 2017) implemented by CRAN package `wbsts` (Korkas and Fryzlewicz, 2020).

We present the simulation results in Table 2.5. We find that SBPCPM outperforms in the frequency of correctly specifying the number of change-points, supporting our analysis results on the London House Index data. Meanwhile, SBPCPM enjoys the lowest Type I error and the higher power, indicating that the locations of change-points estimated by SBPCPM are precise. In contrast, existing approaches are difficult to correctly identify the change-points in most cases due to the pretty low ratio between the jump sizes and the variation of noises, also called signal-to-noise-ratio.

Table 2.5: Results of change-points detection of structural changes in AR(1) models in 200 replications.

Method	Frequency of $\hat{K} - K$					Type I error/FPR (%)	Power/TPR
	≤ -2	-1	0	+1	$\geq +2$		
SBPCPM($\alpha = 0.003$)	6	30	157	4	3	**0.11**	0.873
SBPCPM($\alpha = 0.05$)	2	13	**165**	10	7	0.19	0.905
NSP	7	154	39	0	0	0.05	0.690
SNCP	185	15	0	0	0	0.19	0.300
WBSTS	94	54	50	2	0	0.28	0.487

2.5.2 Setting II: simulations on data with noticeable structural changes in autocorrelations

Besides the synthetic data that are generated from AR(1) models with imperceptible structural changes, we further validate the proposed SBPCPM method on simulated data with noticeable structural changes in autocorrelations. Under model (2.14), we generate $T = 200$ observations with $K = 3$ change-points located at $t = (50, 100, 150)$. We generate y_1 from $N(0, 0.5^2)$ and set the intercept $a_0 = 1$. The segment parameters $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \theta_3) = (0.5, -0.5, 0.5, -0.5)$. At each change-point, the absolute jump size is 1, and hence, the changes are noticeable according to the plot of a data sample in Figure 2.5. We replicate the Monte Carlo simulations for 200 times.

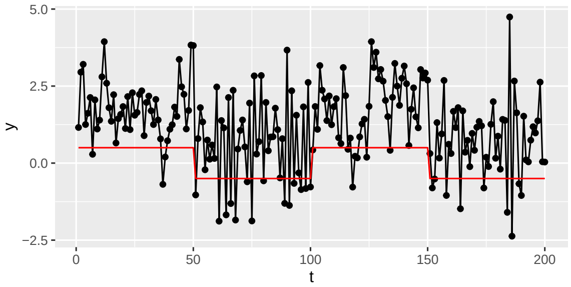


Figure 2.5: An example of simulated data with noticeable structural changes. Point-line: y_t; polyline: the signal function $m(t)$.

We present the simulation results in Table 2.6. We find that SBPCPM still outperforms when the jump sizes are noticeable. We find that neither NSP nor WBSTS performs well under this simulation setting. The reason might be the moderate data size, which induces short segments/intervals and make correct segmentation more difficult.

Table 2.6: Results of change-points detection of structural changes in AR(1) models in 200 replications.

Method	Frequency of $\hat{K} - K$					Type I error/FPR (%)	Power/TPR
	≤ -2	-1	0	+1	$\geq +2$		
SBPCPM($\alpha = 0.003$)	11	30	**138**	20	1	**0.10**	0.873
SBPCPM($\alpha = 0.05$)	7	1	127	49	12	2.19	**0.935**
NSP	193	7	0	0	0	9.53	0.085
SNCP	22	67	104	11	0	0.16	0.721
WBSTS	170	21	9	0	0	9.43	0.117

2.5.3 Setting III: simulations on data with noticeable shifts in mean

Considering that the mean-shifted model (2.1) is the most common practice in change-point detection, we further illustrate our method through simulations under the mean-shifted models. We first present an example with noticeable change-points. We generate $T = 150$ simulated data and set set $K = 2$ change-points at $t = (50, 100)$ with segment parameters $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2) = (0, 5, 10)$, $\sigma^2 = 3$, and $\epsilon_t \sim N(0, 1)$. The jump sizes at two change-points are both 5 and obviously exceed the variation parameter $\sigma = \sqrt{3}$. As a result, the shifts in mean are apparently noticeable as shown in Figure 2.6.

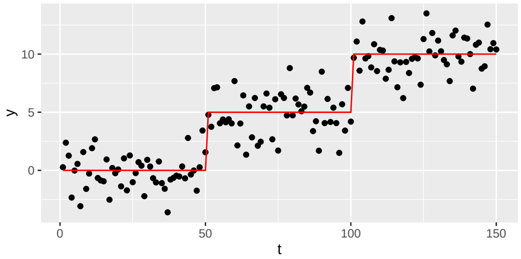


Figure 2.6: An example of simulated data with visible shifts in the mean. Discrete points: data y_t. Polyline: the signal function $m(t)$.

We compare the proposed SBPCPM method with other approaches implemented by the following R packages.

- The `solocp` method (Cappello et al., 2023) implemented by the package `solocp` avalible at https://github.com/lorenzocapp/solocp.

- The SNCP method (Zhao et al., 2022) implemented by their package `SNSeg` supplied by the authors.
- The NOT method (Baranowski et al., 2019) implemented by CRAN package `not` (Baranowski et al., 2022).
- The TUGH method (Fryzlewicz, 2018) implemented by CRAN package `breakfast` (Anastasiou et al., 2022).
- The MOSUM method (Birte and Claudia, 2018) implemented by CRAN package `mosum` (Meier et al., 2021).
- The SMUCE method (Frick et al., 2014) implemented by CRAN package `StepR` (Pein et al., 2022).
- The WBS method (Fryzlewicz, 2014) implemented by CRAN package `wbs` (Baranowski and Fryzlewicz, 2019).

We do not compare with the PELT method (Killick et al., 2010) implemented by R package `changepoint` (Killick and Eckley, 2014) since we find their results are unstable to different choices of penalty types. Table 2.7 presents the detection results under this simulation setting, where we find that SBPCPM correctly detects all change-points and outperforms under significance level $\alpha = 0.003$. Under the significance level $\alpha = 0.05$, SBPCPM is still comparable to other approaches.

Table 2.7: Results of change-points detection for mean-shifted models with noticeable jump sizes in 200 replications.

Method	Frequency of $\hat{K} - K$					Type I error/FPR (%)	Power/TPR
	≤ -2	-1	0	+1	$\geq +2$		
SBPCPM($\alpha = 0.003$)	0	0	**200**	0	0	0	1
SBPCPM($\alpha = 0.05$)	0	5	183	11	1	0.17	1
solo.cp	22	98	80	0	0	0.20	0.645
SNCP	0	0	190	10	0	0.02	1
NOT	0	0	187	11	2	0.03	1
TUGH	0	0	157	33	10	0.14	1
MOSUM	0	0	199	1	0	0.01	1
SMUCE	0	0	193	6	1	0.01	1
WBS	0	0	149	38	13	0.18	1

2.5.4 Setting IV: simulations on data with imperceptible shifts in mean

We here present another simulation example on the mean-shifted model with imperceptible change-points. We again generate $T = 150$ simulated data. We set $K = 2$ change-points at $t = (50, 100)$ with $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2) = (1, 3, 5)$, $\sigma^2 = 3$, and $\epsilon_t \sim N(0, 1)$. Under this setting, the jump sizes at two change-points are both 2, which is close to the within-segment variation. Thus, the two change-points are too imperceptible to be observed by eyes, demonstrated by a plot of a simulated example displayed in Figure 2.7.

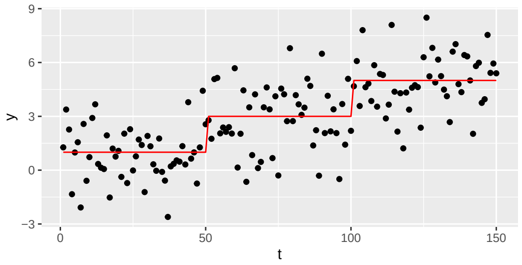


Figure 2.7: An example of simulated data with imperceptible shifts in the mean. Discrete points: data y_t. Polyline: the signal function $m(t)$.

We present the detection result under this setting in Table 2.8. We find `solocp` cannot detect change-points under this setting and thus, omit their results. We find that SBPCPM outperforms under the two significance levels. Compared to the previous noticeable setting, SBPCPM is more robust against imperceptible jump sizes than other approaches. Although the power of NOT and TUGH is slightly higher than SBPCPM, their cost is the apparently higher risk of over detection. This demonstrates the superiority of SBPCPM when facing imperceptible jump sizes.

Table 2.8: Results of change-points detection for mean-shifted models with imperceptible jump sizes in 200 replications.

Method	Frequency of $\hat{K} - K$					Type I error/FPR (%)	Power/TRP
	≤ -2	-1	0	+1	$\geq +2$		
SBPCPM($\alpha = 0.003$)	0	9	**189**	2	0	0.12	0.915
SBPCPM($\alpha = 0.05$)	0	5	183	11	1	0.17	0.928
SNCP	2	51	141	6	0	**0.09**	0.835
NOT	0	7	177	13	3	0.14	**0.935**
TUGH	0	5	156	30	9	0.23	**0.935**
MOSUM	0	14	143	8	1	0.3	0.890
SMUCE	0	1	163	3	1	0.14	0.828
WBS	0	16	146	25	13	0.33	0.870

2.6 Discussion

In this chapter, for multiple change-point detection, we propose SBPCPM, a test-type Bayesian approach, which makes full use of jump sizes, particularly when the magnitudes are imperceptible. Empirically, we consider those jump sizes on change-points τ_k to be imperceptible if the ratio between the minimum absolute jump sizes and the noise variation does not exceed two, expressed as

$$\min_{0\leq k\leq K} \frac{|\theta_k - \theta_{k+1}|}{\sigma} < 2.$$

It is interesting that the ratio in the LHS is consistent with the signal-to-noise ratio defined in Wang et al. (2020, Expression (2) in Section 1), up to a multiplier of the minimum segment length. Note that the segment length measures the contribution of the corresponding segment parameter to the overall likelihood of the data. However, our proposed SBPCPM does not include such a segment parameter since we model the jump size by a "global" form of $\Delta m(t)$. Alternatively, we look into the LHS of (2.2), against the fact that the existing literature investigates the RHS of (2.3), the element in the vector $(\theta_0, \ldots, \theta_K)$ one by one. As a result, SBPCPM does not rely too much on the minimum length of segments (only requires a prespecified lower bound D_0).

Recall that the convergence rate of the MAP estimates $\hat{m}^{\text{MAP}}(t)$ in Theorem 2.1 is at the order of $T^{-1/2}$, the same order of convergence rate of the l_0 jump-size penalized least square (JSPLS) estimator (Boysen et al., 2009, Theorem 1). Nonetheless, the two types of estimates take different types of convergence. The convergence of SBPCPM is pointwise, while the convergence of JSPLS is on the l_2 error of the signal function on the whole support. Since it is trivial to show that the pointwise convergence implies the convergence of l_2 error, one may say that the MAP estimate of $m(t)$ by SBPCPM is more accurate than the JSPLS estimator, especially when the data size T is small. That explains why SBPCPM outperforms other JSPLS variant approaches (e.g. PELT (Killick et al., 2012) and SMUCE (Frick et al., 2014)) under moderate data sizes.

Here come a few remarks on the connection between the NOSE method in Chapter 1 and the current SBPCPM method. Both NOSE and SBPCPM are jump-size-based and thus, they are more powerful to detect imperceptible change-points compared with other methods. Nonetheless, they are different in application. Recall that in Section 1.2, the minimax optimality of NOSE requires that the number of change-points $K_n \to \infty$, while the theoretical results of SBPCPM do not require this condition. In practice, when there are several change-points (e.g. 5 or more), NOSE is preferable. In contrast, when there are sporadic change-points (see the London HPI example), SBPCPM provides a strong supplement to NOSE.

In analyzing the structural changes of the London House Index, we model the data through the piecewise AR(1) model (2.14), a special case of model (2.1), and then employ the proposed SBPCPM method. Nonetheless, the time series community may address the non-stationarity of this dataset and prefer non-time-series models to detect the change-points owing to the strict stationary assumptions. One may consider another alternative, the continuous and piecewise linear mean model (Baranowski et al., 2019, Eq. (b), Section 1), to fit the data by assuming

$$m(t) = \theta_{k,1} + \theta_{k,2}t, \ t \in (\tau_k, \tau_{K+1}], \tag{2.16}$$

with the additional constraint $\theta_{k,1} + \theta_{k,2}\tau_k = \theta_{k+1,1} + \theta_{k+1,2}\tau_k$. The NOT method (Baranowski et al., 2019) is a most general change-point detection approach that can be widely used to detect various kinds of changes including the changes in (2.16). Unfortunately, the change-points presented in Figure 2.8 apparently show an over-detection of change-points. The NOT method is quite sensitive in that it treats any fluctuations of the data curve as change-points. In this sense, our proposed SBPCPM may be considered as an alternative to NOT for the situation of change-points subject to imperceptible jump sizes, where NOT may not apply effectively.

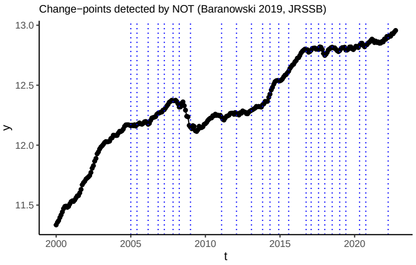


Figure 2.8: Change-points detect by NOT (Baranowski et al., 2019). Point-line: original data; vertical dotted lines: locations of change-points detected by NOT.

Finally, the construction of the signed Beta process plays a key role in SBPCPM. Hence, SBPCPM is mainly used to detect changes in a univariate parameter, such as the mean, autocorrelations, and scale parameters. We leave the detection of changes in multivariate parameters as our future work.

2.7 Supplement

2.7.1 Proof of Theorem 2.1

Proof. The proof starts from the asymptotic normality of $\hat{m}(1)^{\text{MAP}}$ and extends it to all $\hat{m}(t)^{\text{MAP}}$ by the Markov property of the SBP. We separate the likelihood marginal on $m(1)$ from the joint likelihood first. The log likelihood of $(Y_1, \ldots, Y_T)$ given $m(1)$ and $\Delta m(t)$ for $t \leq (T-1)$ is

$$l(Y_1, \ldots, Y_T | m(1), \Delta m(1), \ldots, \Delta m(T{-}1)) = \log f(Y_1 | m(1)) + \sum_{t=1}^{T-1} \log f\{Y_T | m(1) + \sum_{j=1}^{t} \Delta m(t)\}.$$

Thus the posterior distribution of $\{m(1), \Delta m(1), \ldots, \Delta m(T-1)\}$is

$$\begin{aligned}\pi(m(1), \Delta m(1), \ldots, \Delta m(T) | Y_1, \ldots, Y_T) \propto & l\{Y_1, \ldots, Y_T | m(1), \Delta m(1), \ldots, \Delta m(T-1)\} \times \\ & p\{m(1)\} \prod_{t=1}^{T-1} p(\Delta m(t)),\end{aligned}$$

where $p(\cdot)$ denotes the prior density. By integrating out all $\Delta m(t)$ we have the marginal posterior distribution of $m(1)$ as

$$\begin{aligned}\pi\{m(1 | Y_1, \ldots, Y_T)\} \propto p\{m(1)\} \int & l\{Y_1, \ldots, Y_T | m(1), \Delta m(1), \ldots, \Delta m(T-1)\} \times \\ & \prod_{t=1}^{n} p(\Delta m(t)) d\Delta m(1) \ldots d\Delta m(T-1) \equiv p\{m(1)\} l^*\{Y_1, \ldots, Y_T | m(1)\}.\end{aligned}$$

And hence we call $l^*\{Y_1, \ldots, Y_T | m(1)\}$ the likelihood marginal on $m(1)$.

Then we compute the Kullback-Leibler (K-L) divergence and determine its minimum. Let $(Y_1, \ldots, Y_T)$ be a possible realization of $(Y_1, \ldots, Y_T)$. Then the Kullback-Leibler (K-L) divergence of $l^*\{Y_1, \ldots, Y_T | m(1)\}$ relative to $f(Y_1, \ldots, Y_T | \theta_0)$ is defined at any value $m(1) = x$

by

$$\begin{aligned}
\text{K-L}(x) &= E\left\{\sum_{t=1}^{T} \log f(Y_t|\theta_0) - l^*\{Y_1,\dots,Y_T|x\}\right\}\\
&= \int \left\{\sum_{t=1}^{T} \log f(Y_T|\theta_0) - l(Y_1,\dots,Y_T|x,\Delta m(1),\dots,\Delta m(T-1))\right\}\\
&\times \prod_{t-1}^{T} f(Y_t|\theta_0) \prod_{t=1}^{n} p(\Delta m(t)) dY_1\dots dY_T d\Delta m(1)\dots d\Delta m(T-1).
\end{aligned}$$

Since the priors for $\Delta m(t)$s are proper, the minuend term in the first equation can be put into the integral in the second equation directly and $(\Delta m(1),\dots,\Delta m(T-1))$ are integrated out. For the subtraction term in the first equation, the second equation just exchanges its order of integral.

To find out the minimum of the K-L divergence, we first simplify $l^*\{Y_1,\dots,Y_T|x\}$ into a log-likelihood like function. Since f is log-concave according to (A5), we have

$$l(Y_1,\dots,Y_T|x,\Delta m(1),\dots,\Delta m(T-1)) \le \sum_{t=1}^{T} \log f(Y_t|x) + \left.\frac{d\log f}{d\theta}\right|_{\theta=x} \sum_{i=2}^{n}\sum_{t=1}^{i} \Delta m(t).$$

Plug this result in the K-L divergence and we have

$$\begin{aligned}
\text{K-L}(x) \ge & \int \prod_{t-1}^{T} f(Y_t|\theta_0) \sum_{t=1}^{T} \log f(Y_t|\theta_0) dY_1\dots dY_T - \int \prod_{t-1}^{T} f(Y_T|\theta_0) \prod_{t=1}^{n} p(\Delta m(t))\\
&\times \left\{\sum_{t=1}^{T} \log f(Y_t|u) + \left.\frac{d\log f}{d\theta}\right|_{\theta=u} \sum_{i=2}^{n}\sum_{t=1}^{i} \Delta m(t)\right\} dY_1\dots dY_T d\Delta m(1)\dots d\Delta m(T-1)\\
&= \int \prod_{t-1}^{T} f(Y_t|\theta_0) \sum_{t=1}^{T} \log f(Y_t|\theta_0) dY_1\dots dY_T - \int \prod_{t-1}^{T} f(Y_t|\theta_0) \sum_{t=1}^{T} \log f(Y_t|x) dY_1\dots dY_T.
\end{aligned}$$

The second term of the RHS of the second equation holds if $E(\Delta m(t)) = 0$ because $\Delta m(t)$s are integrated out. According to Proposition 1, $E(\Delta m(t)) = 0$ always holds. By information inequality (Murphy, 2012, pp.211), the minimum K-L divergence of $l(Y_1,\dots,Y_T|m(1))$ relative to $f(Y_1,\dots,Y_T|\theta_0)$ is reached if and only if $m(1) = \theta_0$.

Then we need to check conditions for Bernstein–von Mises theorem. It is easily to check that $p\{m(1)\}$ is twice differentiable and $\int_{\mathcal{D}} f(z|\theta_0)\log\{f(z|\theta_0)\}dz < \infty$, where $\mathcal{D}$ is the domain of density f. Therefore, by Theorem 4.16 (Pronzato and Pázman, 2013, pp.98), we immediately have that as $n \to \infty$

$$\hat{m}^{\text{MAP}}(1) \xrightarrow{a.s} \hat{\theta}^{\text{MLE}}, \quad \sqrt{n}\{\hat{m}^{\text{MAP}}(1) - \hat{\theta}^{\text{MLE}}\} \xrightarrow{d} N\{0, J(\theta_0)^{-1}\},$$

where $\hat{\theta}^{\text{MLE}} \xrightarrow{a.s} \theta_0$ is the maximum likelihood estimator of θ_0.

Now we have constructed the asymptotic efficiency of $\hat{m}^{\text{MAP}}(1)$. We then extend it to all $t \geq 2$. Take $t = 2$ for example. We again write the log-likelihood of $(Y_2, \dots, Y_T)$ as

$$l(Y_2, \dots, Y_T|m(2), \Delta m(2), \dots, \Delta m(T-1)) = \log f(Y_1|m(2)) + \sum_{t=2}^{T-1} \log f\{Y_t|m(2) + \sum_{t=2}^{t} \Delta m(t)\}.$$

And the following proof is similar to $m(1)$ and we get as $n \to \infty$

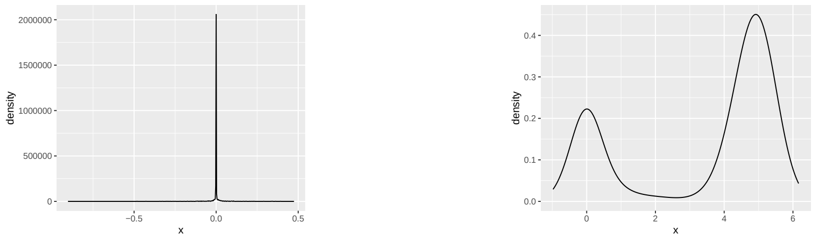
$$\hat{m}^{\text{MAP}}(2) \xrightarrow{a.s} \hat{\theta}^{\text{MLE}}, \quad \sqrt{T-1}\{\hat{m}^{\text{MAP}}(2) - \hat{\theta}^{\text{MLE}}\} \xrightarrow{d} N\{0, J(\theta_0)^{-1}\},$$

Since $\hat{\theta}^{\text{MLE}} \xrightarrow{p} \theta_0$, we obtain $\hat{m}(2)^{\text{MAP}}\ p\ \theta_0$. Meanwhile, by Sluskty's theorem, as $T \to \infty$, we obtain that $\sqrt{T}(\hat{m}^{\text{MAP}}(2) - \theta_0) \xrightarrow{d} N(0, I_Y^{-1}(\theta_0))$.

□

2.7.2 Empirical evidence for Proposition 2.2

In this subsection, we present the empirical evidence for the mode-shifting proposition. We take the simulated data in simulation settings III and IV as examples. Under simulation setting III, we present the density plots of $t = 10$ (unchanged point) and $t = 50$ (change-point) in Figures 2.9(a) and 2.9(b), visualizing the marginal posterior densities of $m(t)$ at unchanged points and change-points, respectively.

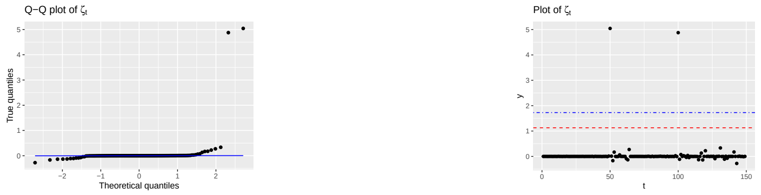


(a) (b)

Figure 2.9: (a) Posterior density of $m(t)$ at an unchanged point $t = 5$ under simulation setting III. (b) Posterior density of $m(t)$ at a change-point $t = 50$ under the same simulation setting.

From Figure 2.9(a), one clearly finds that the posterior is highly concentrated on the true value $\theta_0 = 0$, and thus the MAP estimate is consistent. From Figure 2.9(b), one observed the mode-shifting phenomenon mentioned by Proposition 2.2. That is, the posterior density is bimodal and the two modes locate at the two segment parameters $\theta_0 = 0$ and $\theta_1 = 5$, respectively.

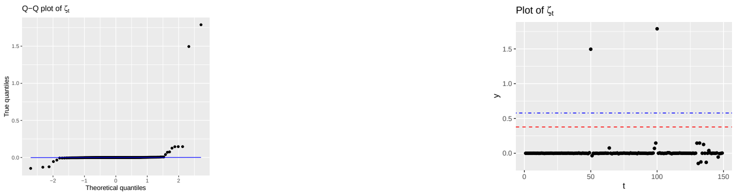
Then we check the distributional approximation to $\zeta_t = \hat{m}(t+1)^{\text{MAP}} - \hat{m}(t)^{\text{MAP}}$, the difference between MAP estimates. Figure 2.10(a) presents the Q-Q plot of ζ_t, where we find most of ζ_t are normally distributed unless two significant outliers at change-points. Figure 2.10(b) presents the original plot of the process of ζ_t, where we find that two change-points fall into the rejection regions while other points are out of the rejection regions under both significance level $\alpha = 0.05$ and $\alpha = 0.003$. These results illustrate the rationale of our proposed testing procedure.



(a) (b)

Figure 2.10: (a) Q-Q plot of ζ_t under simulation setting III. (b) Plot of original process of ζ_t compared with the rejection region under the same setting; dashed horizontal line: rejection upper bound under $\alpha = 0.05$; dotted and dashed horizontal line: rejection upper bound under $\alpha = 0.003$.

The same results of ζ_t are validated under simulation setting IV, where the jump sizes are imperceptible compared to the noise variation; see Figures 2.11(a) and 2.11(b) for evidence.



(a)

(b)

Figure 2.11: (a) Q-Q plot of ζ_t under simulation setting IV. (b) Plot of original process of ζ_t compared with the rejection region under the same setting; dashed horizontal line: rejection upper bound under $\alpha = 0.05$; dotted and dashed horizontal line: rejection upper bound under $\alpha = 0.003$.

Part II

Chapter 3

On Bayesian prediction of survival outcomes through nonparametric transformation models

3.1 Introduction

The traditional linear transformation model raised by Cuzick (1988) is quite flexible, covering whilst not limited to three commonly used survival models, proportional hazards (PH), proportional odds (PO), and accelerated failure time (AFT), and is formulated as

$$h(T) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z} + \epsilon, \tag{3.1}$$

where T is the random censored survival outcome, $\mathbf{Z}$ and $\boldsymbol{\beta}$ are the p-dim predictor vector and the coupled vector of regression coefficients respectively, $h(\cdot)$ is a strictly increasing function that may be *sign varying* on $\mathbb{R}^+$, and ϵ is the model error with distribution function F_ϵ (Cheng et al., 1995). Model (3.1) is called the *nonparametric transformation model* (NTM) when both functional forms of h and F_ϵ are unknown (Horowitz, 1996; Colling and Van Keilegom, 2019).

In predicting survival outcomes, the NTM is apparently preferable because of its model robustness compared to models of PH, PO, AFT, and other survival models assuming either or both of h, F_ϵ specified. However, it also poses the challenge to estimate infinite-dimensional parameters h, S_ϵ in the NTM owning to *model unidentifiability* in the sense that collections of triplet $(h, F_\epsilon, \boldsymbol{\beta})$ generate identical likelihood, called flat likelihood. Nevertheless, estimating such nonparametric components is essential for the prediction of survival outcomes and conditional hazards (Song et al., 2007, pp. 207; Lin et al., 2017, pp. 980), to name a few. This motivates us to overcome the challenging problem of prediction via the NTM.

One may categorize existing approaches of predicting survival outcomes via the NTM into two lines, **i)**, *to make the model identifiable by adding constraints*. Econometricians impose scale normalization to the parametric component (Härdle and Stoker, 1989); and under NTM (3.1), impose location normalization to either h with specified root (Gørgens and Horowitz, 1999; Chen, 2002; among others), or F_ϵ with specified mean or median (Ye and Duan, 1997; Linton et al., 2008; Chiappori et al., 2015; among others). Such approaches focused on establishing theoretical results such as $\sqrt{n}$-convergence, while they did not *touch upon computational feasibility in practice*. As a Bayesian counterpart, Mallick and Walker (2003) evidenced that imposing constrained priors such as the *constrained* Polya tree prior for F_ϵ to identify the NTM is untractable, since an inappropriate center distribution of the Polya tree *incurs slow convergence and poor mixing of posterior* (Müller et al., 2015, pp.39). And **ii)**, *to make strong priori assumptions to circumvent the identifiability issue*. Frequentists either fixed h such as the AFT model (Jin et al., 2003; Ding and Nan, 2011; among others), or made parametric assumptions on F_ϵ such as PH and PO models (Lu and Ying, 2004; Zeng and Lin, 2007a; among others). Alternatively, Bayesian used a two-step procedure to estimate all models and select the "best" (Zhao et al., 2009; de Castro et al., 2014; Zhou and Hanson, 2018). The `R` package `spBayesSurv` (Zhou et al., 2020) based on Zhou and Hanson, as far as we know, may be the optimal tool in prediction provided that it selected the correct model. Despite mathematical or computational convenience, *designating h or F_ϵ is at the risk of misspecification*, leading to inconsistent estimation, invalid statistical inference, and erroneous predictions.

In this chapter, we attempt to seek computationally tractable and robust *Bayesian* prediction under the NTM *without identifying the model*. The spirit of our methodology is based on two concerns.

i. The posterior predictive distribution (PPD) of a future observation is always unique

regardless of model identifiability. Although the parameters in triplet $(h, F_\epsilon, \boldsymbol{\beta})$ under NTM (3.1) are not separately identifiable, they are jointly estimable if their posterior distributions are *proper*. Therefore, the unique PPD can be obtained by integrating all parameters out even though there are multiple solutions of triplet $(h, F_\epsilon, \boldsymbol{\beta})$ that provide the same likelihood; see subsection 3.4.1 for details.

ii. Weakly informative priors make Markov Chain Monte Carlo (MCMC) tractable. In Bayesian analysis, priors play a defining role, have a substantive impact on final model results (Depaoli et al., 2020; van de Schoot et al., 2021), and are analog to constraints that make the model identifiable. Noninformative priors hinder posterior sampling under unidentified models since they cannot control posterior variance to be finite. In contrast, the weakly informative prior is a kind of "stronger" proper prior in the sense that it is able to *control prior variance* moderately on the unconstrained support, and thus is able to *dominate the posterior variance*. Consequently, it facilitates the convergence of posterior sampling by preventing the sampler from running to highly implausible values that are far away from its center (McElreath, 2020, pp.262).

The aforementioned two concerns stand by our methodology. We achieve PPDs of future observations computed from the posterior of $(h, F_\epsilon, \boldsymbol{\beta})$ by assigning two weakly informative priors to the infinite-dimensional parameters, a newly constructed quantile-knots I-splines prior for h, and a common Dirichlet process mixture (DPM) model for F_ϵ, together with a noninformative prior for the parametric component $\boldsymbol{\beta}$. In addition, we obtain an efficient Bayes estimator of identified $\boldsymbol{\beta}$ through posterior projection so as to provide sound relative risks. The predictive capability of our proposed approach is superior to existing methods evaluated by various metrics under different simulated and real data settings.

The contribution of this chapter is tri-folds. Firstly, we solve the standing problem of prediction survival outcomes via the NTM (3.1) efficiently and numerically conveniently. This is realized by the joint strength of two weakly informative priors, quantile-knots I-splines prior for the transformation function, and the DPM model for model error distribution. It is based on I-spline basis functions (Ramsay, 1988) and generates knots from the sample quantiles of censored and uncensored survival times directly. Thus, *a small size of knots enable us to cap-*

ture the major shape of the transformation function well rather than tuning the number of knots in traditional I-spline-based priors that select knots from a long series of equally spaced points (Cai and Dunson, 2007; Wang and Dunson, 2011a; among others). The proposed I-spline type prior is applicable to modeling monotone functions that are differentiable or nondifferentiable by adjusting the smoothness parameter.

Secondly, we provide a new and convenient Bayes estimator for the identified parameter $\boldsymbol{\beta}$ through posterior projection. We impose a unit-norm normalization (Härdle et al., 2004) rather than confining the first entry of the vector parameter to be ± 1 (Gørgens and Horowitz, 1999; Chen, 2002; Song et al., 2007) to avoid specifying the sign of a treatment effect associated with the survival outcomes. The presented posterior modification avoids extra sampling and thus is computationally expedient. In contrast, it is inapplicable to assign constrained priors for $\boldsymbol{\beta}$ directly such as the Polar system prior (Park et al., 2005) or `Stan`'s built-in prior since our prior elicitation has no constraints.

Finally, for practitioners, we provide the `R` package `BuLTM`, which is computationally convenient and efficient to predict survival times and output estimates of predicted survival probability, conditional hazards, and relative risks. For the prediction purpose, simulation studies demonstrate that `BuLTM` outperforms `spBayesSurv` under the PH, PO models, and model misspecification situations, and is comparable to `spBayeSurv` under the AFT model. For the out-sample predictive capability, `BuLTM` is also competitive to `spBayesSurv` in application examples.

The remainder of this chapter is organized as follows. Section 3.2 introduces the recast model of the NTM as the cornerstone of our Bayesian approach. Section 3.3 introduces weakly informative prior elicitation for infinite-dimensional parameters. Section 3.4 introduces the posterior inference procedures including the PPD computation and the posterior projection procedure for $\boldsymbol{\beta}$. Sections 3.5 and 3.6 assess and demonstrate our method compared with existing work by simulations and application examples, respectively. Section 3.7 concludes the chapter with a brief discussion. Related details are collected in the online supplementary materials. The `R` package `BuLTM` is available on GitHub https://github.com/LazyLaker.

3.2 Recast: multiplicative relative risk model

To resolve prediction via the NTM, we first impose the exponential transformation to NTM (3.1) and obtain a recast model

$$H(T) = \xi \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}), \tag{3.2}$$

where the recast transformation $H(\cdot) = \exp\{h(\cdot)\}$ and the model error $\xi = \exp(\epsilon)$ with distribution function F_ξ. The nonparametric transformation model (3.2) with *multiplicative relative risk* $\exp(\boldsymbol{\beta}^T\mathbf{Z})$ is abbreviated as MTM thereafter, where H is *positive* on $\mathbb{R}^+$ and the multiplicative random error ξ is also positive. Let $S_X = 1 - F_X$, where the placeholder X represents ϵ or ξ. MTM (3.2) is equivalent to NTM (3.1) in the sense that they share common parametric component $\boldsymbol{\beta}$, and strictly $S_\epsilon(\cdot) = S_\xi\{\exp(\cdot)\}$ and $h(\cdot) = \log H(\cdot)$.

The above monotonic transformation step plays a critical role in establishing our Bayesian solution. In the Bayesian paradigm, prior elicitation and posterior sampling are two preliminary components of Bayesian inference. Unfortunately, the infinite-dimensional parameter h out of NTM is faced with unprecedented difficulties in both targets.

On one hand, most existing models for *sign-varying* monotone functions are inapplicable to h in that, h may not have an intercept such as the AFT model, preventing usage of approaches that rely on an intercept term in modeling a counterpart of transformation h (Neelon and Dunson, 2004; Shively et al., 2009; Lenk and Choi, 2017, among others); it is also nontrivial to extend to censored observations for those methods that impose a response-based monotonicity shape restriction to the model (Riihimäki and Vehtari, 2010; Lin and Dunson, 2014; Wang and Berger, 2016, among others).

On the other hand, sampling for h often gives rise to trouble if $h(0) \to -\infty$ and lifetimes are close to zero. Take the logit transformed incomplete beta function in Mallick and Walker (2003) for instance. Sampling h may be bothered by *infinity gradient* caused by infinite h under gradient-based samplers such as the Hamilton Monte Carlo and the No-U-Turn Sampler (NUTS) in `Stan` (Carpenter et al., 2017), or by the *poor proposal distributions whose center may disperse to infinity* under Metropolis-type samplers, leading to very slow convergence and a low acceptance rate.

From the insight that it brings huge expedience if one is able to confine the transformation to be *nonnegative*, we are driven to take the recasting as the foremost step to initiate our methodology. Consequently, the exponential transformation compresses the space of infinite-dimensional parameters from $\mathcal{M}_{\mathbb{R}} \times \mathcal{S}_{\mathbb{R}}$ to a reasonable subset of $\mathcal{M}_{\mathbb{R}+} \times \mathcal{S}_{\mathbb{R}+}$, where $\mathcal{M}_{\mathcal{A}}$ denotes the space of monotone functions with range $\mathcal{A}$ and $\mathcal{S}_{\mathcal{A}}$ denotes the space of survival functions with support $\mathcal{A}$. Our spirit has allies in the literature about the transformation model where they rewrote their transformation as the logarithm of a cumulative hazard function (Scheike, 2006; Zeng and Lin, 2006; among others).

Besides its tractability and convenience, the recast MTM (3.2) still maintains interpretability analogous to that of NTM (3.1). Let $\Lambda(\cdot)$ be the cumulative hazard function of a time-to-event. By some simple algebra, for MTM (3.2), the counterpart of expression (1.3) of Cheng et al. (1995) that motivates NTM (3.1) can be represented as

$$G\{\Lambda_{T|\mathbf{Z}}(t)\} = H(t)\exp(-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}), \tag{3.3}$$

where $G^{-1}(\cdot) = -\log\{1 - F_{\xi}(\cdot)\}$ is the link working on the conditional cumulative hazard of the survival time. Specifically, if the link functional forms of $G(s)$ are s and $\{\exp(s) - 1\}$, or, $F_{\xi}(s)$ are $\{1 - \exp(-s)\}$ and $F_{\xi}(s) = (1 + s)^{-1}$ in (3.2), or equivalently, the model error ϵ in (3.1) follows a standard extreme-value distribution and a standard logistic distribution, then the model reduces to PH and PO models respectively.

3.3 Likelihood and prior

3.3.1 Likelihood

For the real survival time T and the random censoring variable C, one denotes the observed time-to-event as $\widetilde{T} = \min(T, C)$. The censoring indicator $\delta = I(T \le C)$. Let S_{ξ} and f_{ξ} be the survival probability and density function of ξ, respectively. In this section, we consider the following quite mild assumptions.

(A1) The exp-transformation H is differentiable.

(A2) The multiplicative random error ξ is continuous.

(A3) The covariate $\mathbf{Z}$ is independent of ξ.

(A4) The censoring variable C is independent of T given $\mathbf{Z}$.

(A1) is required since there is H' functional in the likelihood representation below; (A2) is mild; (A3) is general; (A4) is the general noninformative censorship condition.

With independent triplets of observed data $\{(\widetilde{T}_i, \mathbf{Z}_i, \delta_i)\}_i^n$, one writes the complete data likelihood as

$$\mathcal{L}(\boldsymbol{\beta}, H, S_\xi, f_\xi | \widetilde{T}, \mathbf{Z}, \delta) = \prod_{i=1}^{n} [f_\xi\{H(\widetilde{T}_i)e^{-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i}\}H'(\widetilde{T}_i)e^{-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i}]^{\delta_i}[S_\xi\{H(\widetilde{T}_i)e^{-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i}\}]^{1-\delta_i}. \quad (3.4)$$

For S_ξ and f_ξ, we consider the common DPM models as their priors. Here we employ the truncated stick-breaking construction of the DPM, denoted as

$$S_\xi(\cdot) = 1 - \int F_0(\cdot|\boldsymbol{\theta})dG(\boldsymbol{\theta}),\ f_\xi(\cdot) = \int f_0(\cdot|\boldsymbol{\theta})dG(\boldsymbol{\theta}),\ G = \sum_{l=1}^{L} p_l \delta_{\boldsymbol{\theta}_l},\ \boldsymbol{\theta}_l \sim G_0,$$

where F_0 and f_0 are called kernels from a distribution family parameterized by $\boldsymbol{\theta}$, L is a truncation number of the Dirichlet process, p_l are corresponding sticking-breaking weights, and $\boldsymbol{\theta}_l$ are i.i.d. atoms from the base measure G_0. More justifications about L and choice for G_0 are deferred to *Supplement S.2*. Note that ξ is an arbitrary *continuous positive* random variable. We select the Weibull distributions as kernels out of a positive distribution family since it characterizes a wide range of survival time scenarios (Kottas, 2006; Egleston et al., 2017; Shi et al., 2019, among others). Above consideration is summarized into the expression of priors for S_ξ and f_ξ,

$$S_\xi = 1 - \sum_{l=1}^{L} p_l F_w(\psi_l, \nu_l),\ f_\xi = \sum_{l=1}^{L} p_l f_w(\psi_l, \nu_l), \quad (3.5)$$

where $F_w(\psi, \nu)$ and $f_w(\psi, \nu)$ are the CDF and the pdf of the Weibull distribution, respectively.

To model the differentiable H in (3.4), we propose a quantile-knots I-splines prior. We introduce the details of the proposed prior in the following subsection.

3.3.2 Quantile-knots I-splines prior

Suppose the data $\tilde{T}$ are observed on the interval $D = (0, \tau]$, where τ is the largest survival time in the sample. Note that, H is a nonnegative strictly increasing differentiable function on D based on transformation model (3.2) and the likelihood function (3.4). It is natural to model H and H' by monotone splines,

$$H(t) = \sum_{j=1}^{K} \alpha_j B_j(t), H'(t) = \sum_{j=1}^{K} \alpha_j B_j'(t), \tag{3.6}$$

where $\{\alpha_j\}_{j=1}^K$ are positive coefficients to guarantee nondecreasing monotonicity, $\{B_j(t)\}_{j=1}^K$ are I-spline basis functions (Ramsay, 1988) on D and $\{B_j'(t)\}_{j=1}^K$ are corresponding derivatives. Unlike other I-splines approaches that include an unknown intercept, we simply set the intercept $H(0) = 0$ since it can be derived from assumption (A3) directly, referred to *Supplement S.1*. The number of I-spline basis functions K is the sum of the number of interior knots and the order of smoothness r with $(r-1)$th order derivative existing. Empirically, r may take value from 2 to 4 and we take the default value $r = 3$ in `R` package `splines2`. Interior knots cut the time interval D into $(K - r + 1)$ partitions. Then our concern lies in specifying the number and locations of interior knots for modeling the exp-transformation.

We construct an I-splines type prior based on representations (3.6) by selecting interior knots from empirical quantiles of survival times, namely quantile-knots I-splines prior. First, we fix the initial number of interior knots N_I which is much fewer than that in other typical I-splines type models coupled with the shrinkage prior. Our insight comes from the advantage of quantiles that a small number of quantiles quantify different “locations” of distribution and therefore they can be viewed as alternative measures of the shape of the predictive distribution of T. Meanwhile, the corresponding posterior is not sensitive within the range of a small number of knots, indicating that the proposed prior is free of tuning, referred to *Supplement S.7.1*. It is expedient in implementation compared to those priors requiring tuning, referred to *Supplement S.4.2*.

Next, given the initial number of interior knots N_I, we propose a two-step data-driven procedure to specify their locations using the information of survival times and censoring states.

Let $\hat{F}_X(t) = n^{-1}\sum_{i=1}^{n} I(X_i \leq t)$ be the empirical CDF of X and $\hat{Q}_X(p) = \hat{F}_X^{-1}(p) = \inf\{t : p \leq \hat{F}_X(t)\}$ be the corresponding empirical quantile function, where X is the placeholder for T and $\tilde{T}$, uncensored and observed survival times, respectively. Let $j = 0, \ldots, N_I - 1$.

Step 1: Selects N_I empirical quantiles of uncensored survival times as interior knots $0 < t_0 < \cdots < t_{N_I-1} \leq \tau$, where $t_j = \hat{Q}_T\{j/(N_I - 1)\}$.

Step 2: If $|\hat{F}_T(t_j) - \hat{F}_{\tilde{T}}(t_j)| > z_0 \geq 0.05$, then interpolate a new knot $t_j^* = \hat{Q}_{\tilde{T}}(j/(N_I - 1))$. Output sorted series of $\{t_0, \ldots, t_j, t_j^*, \ldots, t_{N_I-1}\}$ as final interior knots.

In step 1, we choose equally spaced percentiles of uncensored survival times since information about H' is provided by uncensored survival times only. In step 2, we make interpolation in case of high censoring of survival times and insufficient uncensored observations.

Take 5 initial knots for instance i.e. it contains 3 quartiles and 2 endpoints of uncensored survival times. In Figure 3.1, there are apparent deviations between uncensored and observed curves on the first three interior knots. Therefore, we interpolate by three new knots $t_j^* = Q_{\tilde{T}}(j/4)$, for $j = 0, 1, 2$. Finally, we obtain $(t_0^*, t_0, t_1^*, t_1, t_2^*, t_2, t_3, t_4 = \tau)$ as our interior knots. By the above operation, I-spline basis functions $\{B_j(t)\}_{j=1}^{K}$ are specified. We further assign

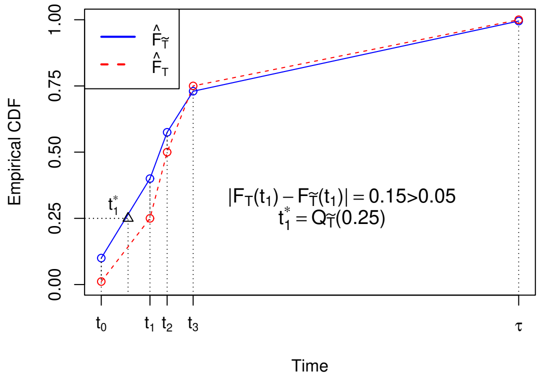


Figure 3.1: Example with 5 initial knots.

an exponential prior for $\{\alpha_j\}_{j=1}^{K}$. Consequently, we have built our quantile-knots I-splines prior for H, which is weakly informative by the fact that, given $\alpha_j \sim \exp(\eta)$, $E\{H(t)\} = \eta^{-1}\sum_{j-1}^{K} B_j(t) < \infty$ and $\mathrm{Var}\{H(t)\} = \sum_{j=1}^{K} \eta^{-2} B_j^2(t) < \infty$ for any $\eta > 0$ and $t < \infty$.

Remark 3.1. *The quantile-knots I-splines prior can also be applied to model nondifferentiable functions. The proposed prior can be viewed as a combination of NII processes, referred to Supplement S.3. Particularly, when $r = 1$, the I-spline function reduces to a straight line on each partition, and the proposed prior reduces to the piecewise exponential prior.*

3.4 Posterior inference

3.4.1 MCMC and posterior prediction

According to above prior settings, nonparametric parameters H and S_ξ in MTM (3.2) are encapsulated in elements of $\boldsymbol{\alpha}$ and $(\boldsymbol{p}, \boldsymbol{\psi}, \boldsymbol{\nu})$, respectively, where $\boldsymbol{\alpha} = \{\alpha_j\}_{j=1}^K$, $\boldsymbol{p} = \{p_l\}_{l=1}^L$, $\boldsymbol{\psi} = \{\psi_l\}_{l=1}^L$,, and $\boldsymbol{\nu} = \{\nu_l\}_{l=1}^L$. Consequently, the nonparametric components (h, S_ϵ) in the original NTM (3.1) are expressed as

$$h(t) = \log\{\sum_{j=1}^K \alpha_j B_j(t)\},\ S_\epsilon(x) = 1 - \sum_{l=1}^L p_l F_w\{\exp(x)|\psi_l, \nu_l\}.$$

Then the estimators of triplet parameters $(h, S_\epsilon, \boldsymbol{\beta})$ can be obtained from the posterior distribution of parameters $\Theta = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{p}, \boldsymbol{\psi}, \boldsymbol{\nu})$. The posterior density of Θ is

$$\pi(\Theta|\widetilde{T}, \mathbf{Z}, \delta) \propto \mathcal{L}(\Theta|\widetilde{T}, Z, \delta)p(\boldsymbol{\alpha})p(\boldsymbol{\beta})p(\boldsymbol{p}) \prod_{l=1}^L G_0(\psi_l, \nu_l),$$

where $\mathcal{L}$ is the likelihood for Θ defined by (3.4) and $p(\cdot)$ represents a prior density. For each parameter in the posterior density, we set their priors as

$$\begin{aligned} &\alpha_j \sim \exp(\eta), p(\boldsymbol{\beta}) \propto 1, G_0(\psi_l, \nu_l) = \text{Gamma}(a, b) \times \text{Gamma}(a, b), \\ &p_l = q_l \prod_{L=1}^{l-1}(1 - q_l), q_l \sim \text{Beta}(1, c), l = 1, \ldots, L-1; p_L = 1 - \sum_{l=1}^{L-1} p_l. \end{aligned} \tag{3.7}$$

Here η is the hyper-parameter of the prior for $\boldsymbol{\alpha}$. The prior for $\boldsymbol{\beta}$ is an improper uniform prior, which is “purely” noninformative. One may either assign a hyperprior for η or fix it to a constant, referred to *Supplement S.7.2* for sensitivity analysis of η. Parameters $\{q_l\}_{l=1}^L$ are stick-breaking weights of the DPM. We fix $c = 1$ as the default total mass parameter in `BuLTM`. For the base measure G_0, we recommend fixing it as that in (3.7) rather than assigning it another hyperprior, referred to *Supplement S.2* for justification.

We note that choices of the prior for $\boldsymbol{\beta}$ can be relaxed to be noninformative. We suggest a purely noninformative prior for $\boldsymbol{\beta}$, such as the improper uniform prior, since it simplifies the form of the posterior and its gradient so as to speed up the MCMC sampler. The following result tells, even though the prior for $\boldsymbol{\beta}$ is improper, under very mild conditions, the posterior in (1.11) is still proper.

Theorem 3.1. *With the improper uniform prior for $\boldsymbol{\beta}$ in (3.7), the posterior distribution in (1.11) is proper under the following conditions: (i) $0 < \widetilde{T}_i < \infty$, for $i = 1, \ldots, n$, (ii) priors for $\{\psi_l, \nu_l\}_{l=1}^{L}$, $\{p_l\}_{l=1}^{L}$ in model (3.5) and $\{\alpha_j\}_{j=1}^{K}$ in model (3.6) are proper, (iii) $0 < K, L < \infty$ in models (3.5) and (3.6), (iv) the kernel f_w in model (3.5) satisfies that $xf_w(x) < \infty$ for all $x > 0$, (v) let $\mathbf{Z}^*$ be the $n_1 \times p$ matrix of the covariates of uncensored observations, where $n_1 = \sum_{i=1}^{n} \delta_i$, and $\mathbf{Z}^*$ is of full rank p.*

This theorem indicates that the impact of the prior for $\boldsymbol{\beta}$ on the prediction is inferior to that of priors for nonparametric components.

Conditions required for Theorem 1 are quite mild. Conditions (i) is a general setting for right censored data. Conditions (ii) and (iii) are general settings for Bayesian analysis. Condition (iv) is naturally satisfied when the Weibull kernel is adopted. Condition (v) is similar to condition (ii) in de Castro et al. (2014), which is a common condition within the survival context. In the right censoring case, condition (v) is also required by Zhou and Hanson (2018) as their algorithm employs a Cholesky decomposition to the covariate matrix. The proof is deferred to *Supplement S.5*.

We implement the NUTS in `Stan` as our MCMC sampler since the domain of Θ is continuous. NUTS is a tuning-free extension of Hamilton Monte Carlo, which is robust and efficient for continuous-variable models. `Stan` has become popular and appealing in recent years since it provides clear automatic posterior sampling procedures. Therefore, users are released from complicated probabilistic deriving and implementation. Our `R` package `BuLTM` is developed based on `Stan`. We approximate the improper uniform prior for $\boldsymbol{\beta}$ through $N(0, 10^6)$ to avoid possible computational issues caused by improper priors in `Stan`.

For prediction purposes, the *posterior predictive survival probability* of a future observation T_0 given covariates $\mathbf{Z}_0$, denoted by $S_{T_0|\mathbf{Z}_0}(t)$, is an average of conditional predictions over the posterior distribution of Θ (Gelman et al., 2013, pp.7). Mathematically, $S_{T_0|\mathbf{Z}_0}(t)$ is the integral of product of conditional survival probability given Θ and $\pi(\Theta|\widetilde{T}, \mathbf{Z}, \delta)$,

$$S_{T_0|\mathbf{Z}_0}(t) = \int S_{T_0|\mathbf{Z}_0}(t|\Theta)\pi(\Theta|\widetilde{T}, \mathbf{Z}, \delta)d\Theta = \int [S_\xi\{H(t)\exp(-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_0)\}]\pi(\Theta|\widetilde{T}, \mathbf{Z}, \delta)d\Theta, \tag{3.8}$$

where S_ξ and H are expressed by elements of Θ as in (3.5) and (3.6), respectively. Note that, alternatively, (3.8) can also be expressed by $(h, S_\epsilon, \boldsymbol{\beta})$. By definition, unidentified MTM (3.2) means that collections of triplets $(\boldsymbol{\beta}, H, S_\xi)$ generate unique likelihood (3.4), which has the same form as $S_{T_0|\mathbf{Z}_0}(t|\Theta)$. The uniqueness of $S_{T_0|\mathbf{Z}_0}(t|\Theta)$ determines the uniqueness of $S_{T_0|\mathbf{Z}_0}(t)$ if the posterior $\pi(\Theta|\widetilde{T}, \mathbf{Z}, \delta)$ is proper. Numerically, the integral (3.8) is approximated by averaging all posterior samples. That is, once samples of $\boldsymbol{\beta}$ and sample paths of H and S_ξ are drawn, denoted by $\boldsymbol{\beta}^{(i)}$, $H^{(i)}$ and $S_\xi^{(i)}$ respectively, for $i = 1, \dots, M$, then the conditional survival probability $S_{T_0|\mathbf{Z}_0}$ and the conditional cumulative hazard $\Lambda_{T|\mathbf{Z}_0}$ are estimated by

$$\hat{S}_{T_0|\mathbf{Z}_0}(t) = N^{-1}\sum_{i=1}^{M} S_\xi^{(i)}\{H^{(i)}(t)\exp(\boldsymbol{\beta}^{(i)\mathrm{T}}Z)\},\ \hat{\Lambda}_{T_0|\mathbf{Z}_0}(t) = -\log(\hat{S}_{T_0|\mathbf{Z}_0}(t)). \tag{3.9}$$

3.4.2 Posterior projection for parametric estimation

Note that without any constraints, we assign two weakly informative priors to nonparametric components (H, S_ξ) or (h, S_ϵ) and a noninformative prior to $\boldsymbol{\beta}$. Then the joint posterior (1.11) of triplet $(h, S_\epsilon, \boldsymbol{\beta})$ is obtained under prior settings in (3.7). Although the posterior of the full set of parameters $(h, S_\epsilon, \boldsymbol{\beta})$ is jointly estimable, the marginal posterior of each component is meaningless. Nonetheless, it is essential for practitioners to have the marginal estimator of the parametric component $\boldsymbol{\beta}$ and related quantities such as relative risks $\exp(-\hat{\boldsymbol{\beta}}^T\mathbf{Z})$. To this end, let $\boldsymbol{\beta}$ be restricted to $||\boldsymbol{\beta}|| = 1$, where $||\cdot||$ is the L_2 norm in the Euclidean space. Our interest focuses on marginal posterior inference and estimation of the *identified unit vector* $\boldsymbol{\beta}/||\boldsymbol{\beta}||$, denoted by $\boldsymbol{\beta}^*$, hereafter.

We obtain a Bayes estimator of $\boldsymbol{\beta}^*$ through posterior modification. This is inspired by a state-of-the-art posterior projection technique. In essence, it is to project the marginal posterior of unconstrained $\boldsymbol{\beta}$ to the constrained parameter space of $\boldsymbol{\beta}^*$. Note that the parameter space of $\boldsymbol{\beta}^*$, the unit hyper-sphere $||\boldsymbol{\beta}^*|| = 1$, is exactly the Stiefel manifold $\text{St}(1, p)$ in $\mathbb{R}^p$. Define a metric projection operator into a set $\mathcal{A}$ as the mapping $m_{\mathcal{A}} : \mathbb{R}^p \to \mathcal{P}(\mathcal{A})$, where $\mathcal{P}(\mathcal{A})$ is the power set of $\mathcal{A}$. Let $\text{dist}(\boldsymbol{x}, \mathcal{A}) = \inf\{||\boldsymbol{x} - x^*||, x^* \in \mathcal{A}\}$ be the distance between $\boldsymbol{x} \in \mathbb{R}^p$ and $\mathcal{A}$. The metric projection operator $m_{\mathcal{A}}$ is determined by

$$m_{\mathcal{A}}(\boldsymbol{x}) = \{x^* \in \mathcal{A} : ||\boldsymbol{x} - x^*|| = \text{dist}(\boldsymbol{x}, \mathcal{A})\}.$$

Then, the metric projection of any vector $\boldsymbol{\beta} \in \mathbb{R}^p$ into $\text{St}(1, p)$ is uniquely determiened as $m_{\text{St}(1,p)}(\boldsymbol{\beta}) = \boldsymbol{\beta}/||\boldsymbol{\beta}||$ (Absil and Malick, 2012, Proposition 7). Consequently, the projected posterior distribution of $\boldsymbol{\beta}^*$ is always proper by proposition 3 in Sen et al. (2022) since the posterior of $\boldsymbol{\beta}$ in (1.11) is proper and absolutely continuous. Note that one only samples the posterior of unconstrained $\boldsymbol{\beta}$ and obtains the posterior of $\boldsymbol{\beta}^*$ by projection. Then the point estimate of $\boldsymbol{\beta}^*$ is given by mean or median of the projected posterior. Numerical studies reveal that our estimator of $\boldsymbol{\beta}^*$ enjoys excellent frequentist performance in the sense of low bias and credible intervals that reach the nominal rate, reconciling the frequentist and Bayesian measures of uncertainty quantification.

In summary, our whole posterior inference procedure takes the following steps,

1. **Initialization**. Initialize the MCMC procedure with initial values of $\boldsymbol{\alpha}, \boldsymbol{p}, \boldsymbol{\psi}$ and $\boldsymbol{\nu}$ sampled from their priors. Randomly generate an initial for $\boldsymbol{\beta}$ so that $||\boldsymbol{\beta}|| > 0$.

2. **MCMC**. Draw M posterior samples of $\Theta = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{p}, \boldsymbol{\psi}, \boldsymbol{\nu})$ from the posterior (1.11) by NUTS.

3. **Prediction**. Compute posterior predictive survival functions given $\boldsymbol{z}_0$ following (3.9).

4. **Estimation of $\boldsymbol{\beta}^*$**. Generate the ith posterior sample of parameter $\boldsymbol{\beta}^*$ as $\boldsymbol{\beta}^{(i)}/||\boldsymbol{\beta}^{(i)}||$, where $\boldsymbol{\beta}^{(i)}$ is the ith posterior sample of $\boldsymbol{\beta}$ drawn in Step 2, for $i = 1, \ldots, M$.

3.5 Simulations

Extensive simulations are conducted to evaluate the robustness of prediction of survival outcomes by the proposed `BuLTM` method and performance of the parametric estimation under the nonparametric transformation model setting. We compare `BuLTM` with contemporary Bayesian and frequentist methods. The Bayesian competitor is the `R` package `spBayesSurv` by Zhou and Hanson (2018), which provides a unified two-step Bayesian route for fitting and selecting mainstream transformation models of PH, PO, and AFT. For the frequentist method, we compare with the contemporary `R` package `TransMOdel` (Zhou et al., 2022b) for semiparametric transformation models with pre-specified model error, as an implementation of Chen et al. (2002). Details about reproducibility and simulation results in low censoring cases are put into *Supplements S.6.1* and *S.6.2*.

Simulated survival times are generated following model (3.1). Under each setting, we generate 300 Monte Carlo replicates, each with sample size $n = 200$. The vector of regression coefficients is $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^{\mathrm{T}} = (\sqrt{3}/3, \sqrt{3}/3, \sqrt{3}/3)^{\mathrm{T}}$ such that $||\boldsymbol{\beta}|| = 1$. Therefore, the identified $\boldsymbol{\beta}^*$ estimated by `BuLTM` is coincided with the true $\boldsymbol{\beta}$ in data generation, leading to the interpretation. For covariates $\mathbf{Z} = (z_1, z_2, z_3)$, we set $z_1 \sim \text{Bin}(0.5)$ indicating a discrete/categorical variable, $z_2, z_3 \sim N(0,1)$ as continuous variables with correlation coefficient 0.2, and z_1 is independent of (z_2, z_3).

We assess the performance of `BuLTM` under four true model cases including PH, PO, AFT models, and a case where none of them is the true model.

Case 1. Non- PH/PO/AFT : $\epsilon \sim 0.5N(-0.5, 0.5^2) + 0.5N(1.5, 1^2)$,

$h(t) = \log[(0.8t + t^{1/2} + 0.825)(0.5\Phi_{1,0.3}(t) + 0.5\Phi_{3,0.3}(t) - c_1)], C \sim \mathrm{U}(1.5, 3)$;

Case 2. PH model : $\epsilon \sim \mathrm{EV}(0, 1)$,

$h(t) = \log[(0.8t + t^{1/2} + 0.825)(0.5\Phi_{0.5,0.2}(t) + 0.5\Phi_{2.5,0.3}(t) - c_2)], C \sim \min(\exp(1), 2.5)$;

Case 3: PO model : $\epsilon \sim \mathrm{Logistic}(0, 1)$,

$H(t) = \log[(0.8t + t^{1/2} + 0.825)(0.5\Phi_{0.5,0.2}(t) + 0.5\Phi_{2.5,0.3}(t) - c_3)], C \sim \min(\exp(3/4), 3.5)$;

Case 4: AFT model : $\epsilon \sim N(0, 1^2), h(t) = \log(t), C \sim \min(\exp(3/4), 5)$.

Here $\Phi_{\mu,\sigma}$ denotes the CDF of $N(\mu,\sigma^2)$, EV(a,b) denotes the extreme value distribution such that its exponential follows Weibull$\{\exp(a),1/b\}$, and c_k is the constant such that $\exp\{h(0)\} = H(0) = 0$, for $k = 1,2,3$. The censoring variable C is generated independent of $\mathbf{Z}$, leading to approximately 57%, 58%, 59%, and 61% censoring rates, respectively.

Case 1 can neither be expressed by any of PH, PO, and AFT models nor be incorporated by the class of logarithmic transformations in Chen et al. (2002). In Case 2,
$S_{T|\mathbf{Z}}(t) = \exp\{-\exp[h(t)]\exp(-\boldsymbol{\beta}^T\mathbf{Z})\}$. Therefore, the conditional hazard function is

$$\lambda_{T|\mathbf{Z}}(t) = \exp[h(t)]h'(t)\exp(-\boldsymbol{\beta}^T\mathbf{Z}),$$

which is exactly a PH model, corresponding to $r = 0$ in `TransModel`. In Case 3, $S_{T|\mathbf{Z}}(t) = \{1+\exp[H(t)]\exp(-\boldsymbol{\beta}^T\mathbf{Z})\}^{-1}$. Then, the conditional odds function is

$$\frac{1-S_{T|\mathbf{Z}}(t)}{S_{T|\mathbf{Z}}(t)} = \exp[h(t)]\exp(-\boldsymbol{\beta}^T\mathbf{Z}),$$

which is exactly a PO model, corresponding to $r = 1$ in `TransModel`.

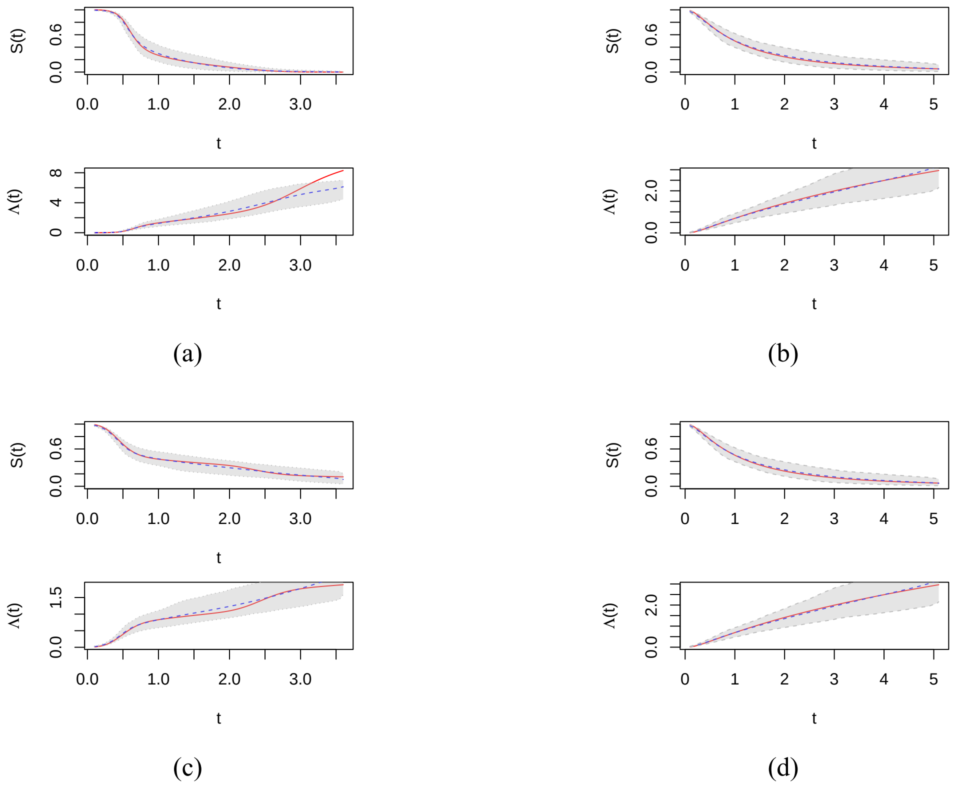
3.5.1 Prediction of conditional survival probability

We assess the accuracy of the prediction of survival outcomes and visualize predictive survival probability and cumulative hazard functions. Following (3.9), `BuLTM` computes the PPD by posterior samples of triplet $(H, S_\xi, \boldsymbol{\beta})$. The accuracy of prediction is assessed by the L_2 distance between real conditional survival curves and the PPD. Numerically, the L_2 distance is approximated by root integrated square error (RISE) on the observed time interval. The smaller RISE, the better the prediction. For each prediction scenario, we compare PPDs of three future observations with different sets of covariates: $\mathbf{Z}_1 = (0,0,0)^{\mathrm{T}}, \mathbf{Z}_2 = (1,1,1)^{\mathrm{T}}$ and $\mathbf{Z}_3 = (0,1,1)^{\mathrm{T}}$, respectively.

Table 3.1 shows that, under these three sets of new observations, `BuLTM` overwhelmingly outruns `spBayesSurv` in the performance of predicting conditional survival probability under non-PH/PO/AFT, PH, and PO models, and is comparable with `spBayesSurv` under the AFT model. It is reasonable that `BuLTM` is superior to the other two approaches in Case 1 since the true model setting is beyond the application scope of `spBayesSurv` and `TransModel`; `BuLTM` still outperforms the other two under in Case 2 since the PH model is a special case of the Weibull mixture model employed by `BuLTM`. For the PO model, the three approaches are comparable; for the AFT model, `spBayesSurv` outperforms since it has already correctly specified the transformation function, resulting in a much simpler problem of density estimation. We find that `TransModel` does not perform well in this case even though we use the standard logistic distribution ($r = 1$) to approximate the Gaussian model error.

Table 3.1: The RISE between true conditional survival functions and functions predicted by `BuLTM` and `spBayesSurv` under Cases 1 to 4.

	Case 1: Non- PH/PO/AFT							Case 2: PH			Case 3: PO			Case 4: AFT		
Z	BuLTM	PH	PO	AFT	$r=0$	$r=0.5$	$r=1$	BuLTM	PH	$r=0$	BuLTM	PO	$r=1$	BuLTM	AFT	$r=1$
$\mathbf{Z}_1$	**0.068**	0.122	0.130	0.117	0.104	0.109	0.104	**0.074**	0.080	0.091	0.010	**0.098**	0.103	0.968	**0.079**	0.102
$\mathbf{Z}_2$	**0.060**	0.128	0.083	0.220	0.099	0.109	0.099	**0.077**	0.084	0.092	**0.125**	0.126	0.127	0.139	**0.125**	0.161
$\mathbf{Z}_3$	**0.079**	0.112	0.100	0.132	0.120	0.123	0.120	**0.100**	0.110	0.113	0.139	0.135	**0.127**	0.157	**0.132**	0.161



(a) (b) (c) (d)

Figure 3.2: The predicted conditional survival probability curve ($S(t)$) and the conditional cumulative hazard function ($\Lambda(t)$) for $Z = (0, 0, 0)^{\mathrm{T}}$; (a), Case 1; (b), Case 2; (c), Case 3; (d) Case 4; real line: true curve; dash line: estimated curve; shadow: 95% empirical point-wise confidence band.

Particularly, `BuLTM` is excellent to predict the survival probability of future observations with the zero covariate vector $\mathbf{Z}$ (also called baseline survival probability in Zhou and Hanson (2018)). We present the average predicted baseline survival probability curves and baseline cumulative hazard curves throughout the simulations in Figure 3.2. The average predicted curves fit the true curves quite well and are covered by the 95% point-wise confidence band, demonstrating the prediction capability of `BuLTM`.

3.5.2 Parametric estimation

We evaluate the performance of `BuLTM` in estimating the identified parameter $\boldsymbol{\beta}^*$, which has the same interpretation as the true unit vector $\boldsymbol{\beta}$ in all simulation settings. We consider the following frequentist operating characteristics for evaluation, the average bias of estimates (BIAS), the square root of the mean squared error of the estimator (RMSE), the average posterior standard error (PSD), the standard error of the estimated values (SDE), and the coverage probability of the 95% credible or confidence interval (CP), as usual. The pointwise bias of `BuLTM` should be computed in a different way from `spBayesSurv`. Among all simulations, we re-scale the mean vector of estimated $\hat{\boldsymbol{\beta}}^*$ into a unit vector and then compute the pointwise bias. Otherwise, the result is surely biased no matter what kind of unit-norm estimator is used. The reason is that `BuLTM` provides an estimate of a unit vector in each replication of simulations, while the element-wise mean of a series of unit vectors is not a unit vector anymore since for unit vectors $v_1, \dots, v_n \in \mathrm{St}(1, p)$, $||n^{-1} \sum_{i=1}^n v_i|| \le 1$ by triangle inequality.

Results of parametric estimation are summarized in Table 3.2 for Cases 1-3. Results under the AFT model are put into *Supplement S.6.3*. It is worth noting that the interpretation of the true $\boldsymbol{\beta}$ in Case 1 is different from that of any semiparametric models fitted by `spBayesSurv` and transformation models fitted by `TransModel`. Therefore, none of them provide reasonable parametric estimation in Case 1, and we leave the place of their assessment results blank. In contrast, the parametric estimation given by `BuLTM` has little bias, the PSD is quite close to the SDE, and the CP is close to the nominal level in this case. In Cases 2 and 3, where the true model is one of PH and PO models, `BuLTM` has a lower bias for most parameters and has lower RMSE for all parameters than the other two methods. Since the "true" regression vector is set to be unit-norm, the three approaches share the same interpretation for $\boldsymbol{\beta}$ in these two cases.

These simulation results demonstrate that `BuLTM` estimates the unit-norm restricted identified parameter quite well.

Table 3.2: Results of estimation of β by `BuLTM`, `spBayeSurv`, and `TransModel` in Cases 1 to 3.

Case 1: Non-PH/PO/AFT		BuLTM					spBayesSurv/TransModel				
	Parameter	BIAS	RMSE	PSD	SDE	CP					
	β_1	-0.003	0.098	0.092	0.097	94.0					
	β_2	-0.006	0.072	0.067	0.071	92.0					
	β_3	0.009	0.072	0.067	0.068	94.0					
		HCase2: PH					HCase3: PO				
Method	Parameters	BIAS	RMSE	PSD	SDE	CP	BIAS	RMSE	PSD	SDE	CP
BuLTM	β_1	0.005	0.159	0.152	0.158	93.7	0.011	0.218	0.211	0.214	92.7
	β_2	**-0.002**	0.122	0.107	0.118	93.3	**-0.000**	0.148	0.146	0.138	95.3
	β_3	**-0.003**	0.109	0.108	0.105	93.3	**-0.011**	0.149	0.146	0.135	95.3
spBayesSurv	β_1	0.018	0.240	0.227	0.240	92.0	**0.000**	0.335	0.315	0.335	94.7
	β_2	0.025	0.137	0.122	0.135	92.7	0.021	0.172	0.167	0.171	94.7
	β_3.	0.023	0.128	0.122	0.126	93.7	0.014	0.164	0.168	0.164	95.0
TransModel	β_1	**0.001**	0.267	0.244	0.267	93.0	0.057	0.369	0.339	0.365	92.0
	β_2	0.008	0.140	0.132	0.140	96.0	0.021	0.196	0.179	0.195	93.0
	β_3	0.017	0.133	0.132	0.132	96.3	0.021	0.190	0.179	0.188	92.3

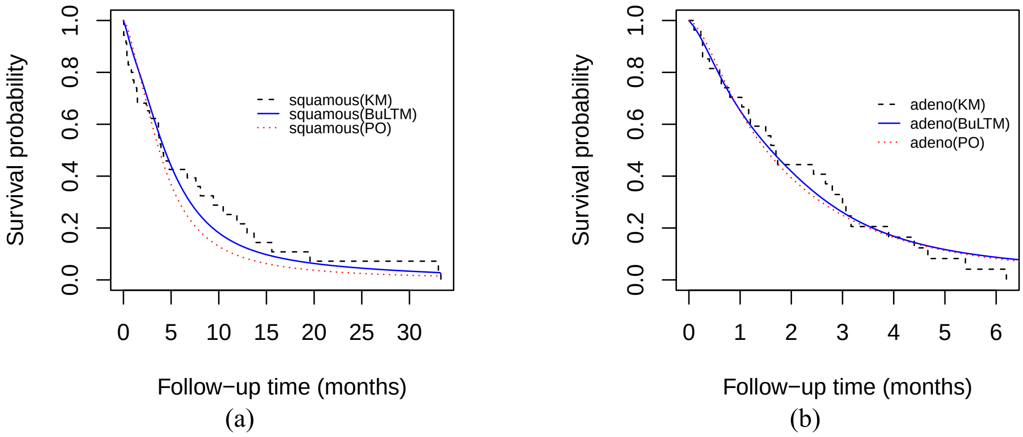
3.6 Applications

3.6.1 PO case: veterans lung cancer data

The first example is the veterans lung cancer dataset from `R` package `survival` (Therneau, 2022). It contains 137 patients from a randomized trial receiving either a standard or a test form of chemotherapy. In the study, the survival time is one of the primary endpoints for the trial and 128 patients were followed to death. We include six covariates, the first five of which are $Z_1 = \text{karno}/10$ (karnofsky score), $Z_2 = \text{prior}/10$ (prior treatment, with 0 for no therapy and 10 otherwise), $Z_3 = \text{age}/100$ (years), $Z_4 = \text{diagtime}/100$ (time in months from diagnosis to randomization), and $Z_5 = I(\text{treatment} = \text{test form of chemotherapy})$. The remaining is the covariate of the cell type which has four categories, adeno, squamous, small cell, and large cell. Thus we include indicator variables to associate with time-to-death, that is, $Z_6 = I(\text{cell type} = \text{squamous})$, $Z_7 = I(\text{celltype} = \text{small})$, and $Z_8 = I(\text{celltype} = \text{large})$.

We fit the nonparametric transformation model for the veterans data by `BuLTM` and compare the results with `spBayesSurv` and `TransModel`. `spBayesSurv` selects the PO model in this case and thus we use $r = 1$ in `TransModel`.

Prediction We compare the curves of estimated survival probability given by `BuLTM` with that of `spBayesSurv` first. We divide the dataset into four strata based on their cell types. For each stratum, the survival curves given by `BuLTM` and `spBayesSurv` are estimated through the predicted survival probability conditional on the mean values of covariates of all individuals within the stratum. For comparison, we use the Kaplan-Meier (K-M) estimator of that stratum as the baseline result. Figure 3.3(a) and 3.3(b) display the results of estimated survival curves. For the squamous stratum, the survival curve given by `BuLTM` is significantly closer to the K-M estimator than that of `spBayesSurv`; for the adeno stratum, the survival curve given by `BuLTM` is slightly closer to the K-M estimator in the middle range of the following-up period. Since `BuLTM` and `spBayesSurv` perform similarly to each other on the remaining two strata, we simply omit their results here. The comparison with the K-M estimator supports the estimation of survival functions given by `BuLTM`.

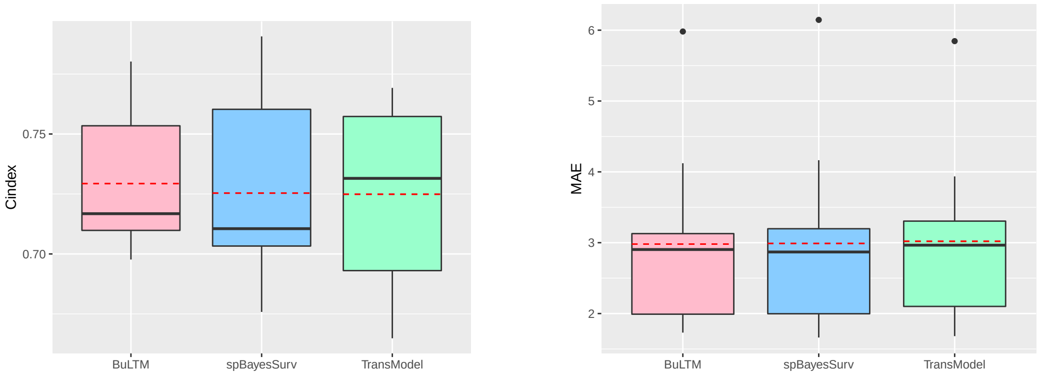


(a)
(b)

Figure 3.3: Estimated curves of survival probability given by BuLTM, spBayesSurv, and the K-M estimator under strata categorized by celltypes; (a) the stratum of squamous; (b) the stratum of large cell.

To further compare their predictive capability, we randomly split the full dataset into the training and testing sets with proportions 90% and 10%, respectively. We repeat this procedure 10 times. We fit survival models based on the training data first and then predict survival outcomes on the testing set. The prediction capability is assessed by the commonly used Concordance index (C index, Harrell et al., 1982), which is an extension of the area under the curve (AUC) as a measure of concordance between a predictive biomarker and the right-censored survival time. A higher C index implies better prediction capability of a model. In this chapter, the C index is computed by R package SurvMetrics (Zhou et al., 2022a) following the procedure in Ishwaran et al. (2008). Details about metrics for prediction evaluation of survival models in this chapter are deferred to *Supplement S.10*. Since most observations in the example are uncensored, a natural prediction of the survival time of a future observation is the median computed from its PPD. And then we use this predicted survival time as the diagnostic marker to compute C index. We also compute the mean of absolute error (MAE) between the predicted survival times and the true survival times of uncensored observations.

Figure 3.4(a) presents the boxplot of C index assessed among the 10 testing sets as well as the mean C index. From the figure we find that BuLTM provides the highest mean C index (0.729) and the most concentrated C indices across 10 testing sets. Although TransModel provides the highest median C index, it has large variation and encounters the worst prediction result. We further compare their MAE of predicted survival times on testing sets in Figure 3.4(b). We find that BuLTM shares almost the same 25% quantile and median of MAE as that of spBayesSurv, while it enjoys a lower mean MAE. Meanwhile, BuLTM outperforms TransModel in both mean and median MAEs. These two results demonstrate that BuLTM is competitive in out-sample prediction on the veterans lung cancer dataset.



(a) (b)

Figure 3.4: (a) The box plot of the C index computed on 10 testing sets; (b) the box plot of MAE between predicted and true survival times of uncensored observations on 10 testing sets.

Estimation of relative risks In terms of estimation of relative risks, we add the smoothed partial rank (SPR) estimator (Song et al., 2007) into our comparison. Although quantitative interpretations of $\boldsymbol{\beta}$ ($\boldsymbol{\beta}^*$ in BuLTM) under different models are different, their qualitative interpretations such as the relative importance of the predictors such as the relative importance of treatment effects are relatively stable (Solomon, 1984). Our analysis demonstrates this point of view since the results of parametric estimation given by different methods are consistent, referred to *Supplement S.8.1*.

According to the model selection result by spBayesSurv, the underlying survival model of this dataset is more likely to be the PO model. Under the PO model, the odds given covariates $\mathbf{Z}$ are proportional to the relative risk $\exp(-\boldsymbol{\beta}^T\mathbf{Z})$ at any time t. Hence, it is important to evaluate the estimated relative risk $\exp(-\hat{\boldsymbol{\beta}}^T\mathbf{Z})$ given by the above three methods ($\exp(-\hat{\boldsymbol{\beta}}^{*T}\mathbf{Z})$ by BuLTM). Naturally, we assess the estimated relative risk through the area under the time-dependent ROC(t) curve (AUC) for censored survival time by treating the survival status as a binary response. Figure 3.5 displays the dynamic AUCs using the estimated relative risks given by BuLTM, spBayesSurv, and TransModel as diagnostics. We find BuLTM and TransModel share almost the same survival AUC curves which are higher than that of spBayesSurv.

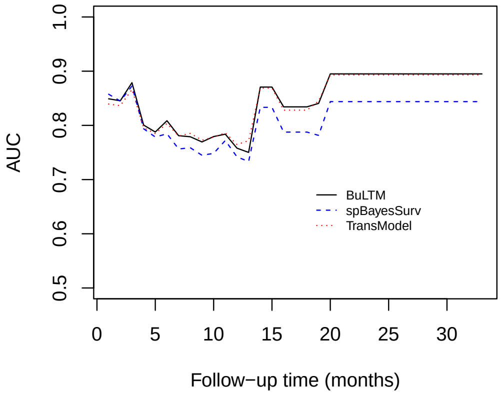


Figure 3.5: Time dependent survival AUC(t) computed by estimated relative risks.

3.6.2 PH case: heart failure clinical records data

We apply BuLTM to analyze the heart failure clinical records data first published by Chicco and Jurman (2020). The dataset records 299 heart failure patients collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad, from April to December 2015 (Ahmad et al., 2017). The dataset consists of 105 women and 194 men, with a range of ages between 40 and 95 years old. In the dataset, 96 observations are recorded as death and the remaining 203 are censored, leading to a censoring rate of 67.9%, which is relatively high. The dataset contains 11 covariates reflecting one's clinical, body, and lifestyle information. Among the 11 covariates, 5 of them are binary variables: anaemia, high blood pressure, diabetes, sex, and smoking. The dataset considers a patient having anaemia if hematocrit levels were lower than 36%, while the criterion for high blood pressure is unclear in the study. Other continuous covariates are age (year), creatinine phosphokinas (level of the creatinine phosphokinas enzyme in the blood, mcg/L), ejection fraction (percentage of blood leaving the heart at each contraction), platelets (platelets in blood, kiloplatelets/mL), serum creatinine (level of creatinine in blood, mg/dL), and serum sodium (level of sodium in blood, mEq/L). The survival times are recorded in days. In our data pre-processing, we transfer the survival time to months by days/30. We report the results of prediction here compared with `spBayesSurv`. Parametric estimation results and estimation of relative risks given by the two methods are similar and deferred to *Supplment S.8.2*.

Prediction Likewise, we compare the curves of estimated survival probability given by `BuLTM` with that of `spBatesSurv` first. In this case, `spBayesSurv` selects the PH model. We consider two strata of observations: the high-risk (HR) stratum where observations have both anaemia and high blood pressure, and the low-risk (LR) stratum where observations have neither anaemia nor high blood pressure. For each stratum, the survival curves given by `BuLTM` and `spBayesSurv` are estimated through the predicted survival probability conditional on the mean values of covariates of all individuals within the stratum. We also use the K-M estimator as the baseline result for comparison.

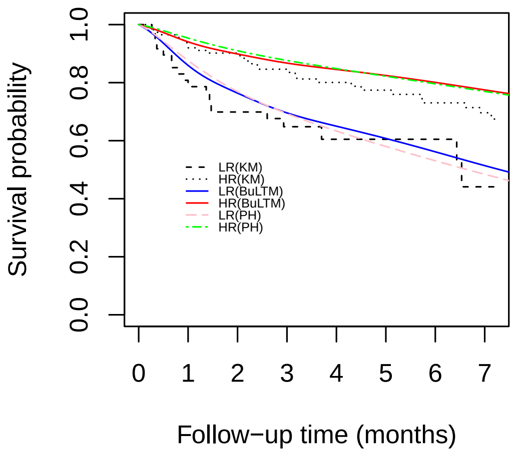
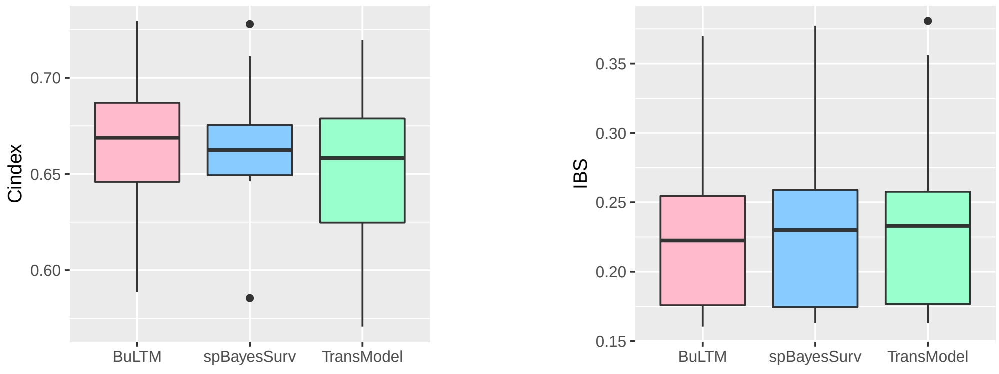


Figure 3.6: Estimated curves of survival probability given by BuLTM, spBayesSurv, and the K-M estimator under high-risk and low-risk strata.

As shown by Figure 3.6, for the LR stratum, the survival curve estimated by BuLTM is closer to the K-M estimator than that of spBayesSurv both at the beginning follow-up time period and months from 5 to 6, and spBayesSurv is closer to the K-M estimator at other times. For the HR stratum, BuLTM performs slightly better at the beginning and provides almost the same result as spBayesSurv at the tail. It is reasonable that BuLTM performs better at the beginning time period on this highly-censored dataset since most quantiles of survival times are distributed at the beginning period and the quantile-knots I-splines prior generates more knots at the beginning. For comparison of their predictive capability on this dataset, we still randomly split the full dataset into the training and testing sets with proportions 90% and 10%, respectively, and repeat this procedure 10 times. Again, we evaluate the predictive capability by the C index. According to the censoring rate (68.9%), we select the 70% quantiles of PPDs to compute the C index. Besides, we consider the Brier score (BS, Graf et al., 1999) to assess the prediction curve error i.e. expected value of the square of the difference between the true survival state of a sample and its predicted survival probability at some specific time points. To evaluate the BS on all follow-up time intervals, we consider the integral of BS functions (IBS) on a given interval as another assessment. As a kind of square error, the lower the IBS, the better the prediction. We don't consider the MAE as an assessment in this case since most observations are censored and hence, the MAE loss is meaningless.



(a) (b)

Figure 3.7: Prediction comparison between BuLTM, spBayesSurv, and TransModel; (a), C index; (b), Integrated Brier score.

As shown by Figure 3.7(a), among the 10 testing sets, BuLTM enjoys a higher median and a higher 75% quantile of C indices. Meanwhile, the average C index of BuLTM (0.669) is again

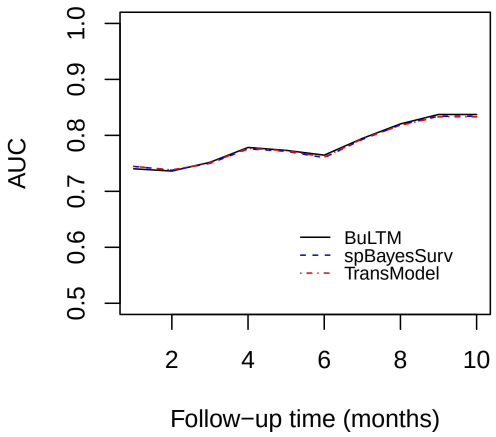


Figure 3.8: Time dependent survival AUC(t) computed by estimated relative risks. (a), method “K-M”; (b), method “NNE”.

slightly higher than that of the PH model (0.664). In terms of the IBS, as shown by Figure 3.7(b), `BuLTM` enjoys a lower median, 75% quantile, and the maximum value than the PH model among the 10 testing sets. The average IBS of `BuLTM` (0.233) is lower than the average value of the PH model (0.238) too. These results support that `BuLTM` has competitive out-sample predictive capability on this dataset.

3.7 Discussion

In this chapter, instead of imposing strong restrictions to make the NTM identified, we assign two weakly informative priors for the nonparametric components, the quantile-knots I-splines prior to the transformation function and a Weibull kernel DPM model to the error distribution, and employ a noninformative prior to the parametric component, to achieve prediction through computing PPDs under NTM (3.1). We are not the daredevils to do so since existing literature has had a few explorations in other environments, where weakly informative priors were modeled to avoid burdensome computation caused by constraints for model identification (McCulloch and Rossi, 1994; Branscum et al., 2008; Burgette et al., 2021; Berchuck et al., 2022; among others).

We explored the use of constrained priors for H while the posterior on the constrained support is too difficult to sample; see *Supplement S.4.1* for details. For posterior inference in `BuLTM`, although we admit that a few inner points of the posterior surface (percentage less than 0.5%) may exceed the maximum tree depth of NUTS (Hoffman et al., 2014) in MCMC sampling, our method enjoys fast convergence and well-mixing of MCMC chains with high effective sample size (ESS) in MCMC diagnosis; see *Supplement S.6.4*; the posterior is neither sensitive to subjective choices of hyperparameters in the weakly informative priors nor similar to priors, referred to *Supplements S.7* and *S.9*, respectively.

Our success on the NTM provides a possible route to address the prediction under a wide range of nonparametric models with unidentified infinite dimensional parameters, where identifying the infinite-dimensional parameters by imposing complicated constraints may encounter computational infeasibility. In this case, one may construct weakly informative nonparametric priors for the infinite-dimensional parameters (with specified center and finite variation, similar to our quantile-knots I-splines prior for h and DPM model for F_ϵ) so as to facilitate MCMC sampling and compute the PDDs for future observations. In the meantime, inference of identified parameters can be obtained by posterior projection to a constrained space that makes the parameter identified. Our results in parametric estimation supply numerical justification for the post-processing, from both aspects of point estimation and uncertainty measurement.

It is intuitive to extend the proposed `BuLTM` method to other types of censorship given that the censoring is noninformative. The analysis of informative censoring or competing risk yields a different likelihood and thus different posterior inference, requiring completely new research. A natural next step work may use the spirit of solving the estimation of the NTM to estimate single-index models from the Bayesian perspective; another natural extension is to study random effects models where the nonparametric transformation acts as the functional random effect.

3.8 Supplement

3.8.1 Deriving $H(0) = 0$ from assumption (A3)

Proof. Suppose $H(0) = a$, where a is a positive constant. It is natural that $Pr\{T > 0\} = 1$. Then we have

$$Pr\{T > 0\} = \int_D Pr\{T > 0|\mathbf{Z} = \boldsymbol{z}\} f_{\mathbf{Z}}(\boldsymbol{z}) d\boldsymbol{z} = 1,$$

where D denotes the support of covariate $\mathbf{Z}$ and $f_{\mathbf{Z}}$ denotes the density of $\mathbf{Z}$. According to the transformation model, $Pr\{T > 0|\mathbf{Z} = \boldsymbol{z}\} = Pr\{H(T) > a|\mathbf{Z} = \boldsymbol{z}\} = Pr\{\xi \exp(\boldsymbol{\beta}^T \boldsymbol{z}) > a\} = Pr\{\xi > a \exp(-\boldsymbol{\beta}^T \boldsymbol{z})\}$. As a counterexample, we suppose the covariate $\mathbf{Z} = Z \sim N(0,1)$ is univariate, the model error $\xi \sim \exp(1)$, and $\boldsymbol{\beta} = \beta_1 = -1$. Since ξ and Z are independent, we have $Pr\{T > 0\} = \int_{\mathbb{R}} \int_{a\exp(z)}^{+\infty} \exp(-t)\phi(z;0,1)dtdz < 1$, where $\phi(\cdot;0,1)$ denotes the density of $N(0,1)$. This contradicts the fact that $Pr\{T > 0\} = 1$. Therefore, $H(0) = 0$. □

3.8.2 The DPM model for S_ξ

A regular Dirichlet process mixture (DPM) model (Lo, 1984) is assigned for S_ξ, the survival probability function of the positive random variable ξ. The DPM is a kernel convolution to the Dirichlet process (DP). We use the stick breaking representation for $G \sim \text{DP}(c, G_0)$ (Sethuraman, 1994)

$$G(\cdot) = \sum_{l=1}^{\infty} p_l \delta_{\theta_l}(\cdot), \theta_l \sim G_0, p_l \sim \text{SB}(1,c)$$

where $\delta(\cdot)$ is the point mass function, and SB is the stick-breaking representation. We call G_0 as the base measure and c as the total mass parameter, acting as the center and precision of the DP, respectively.

Following the above stick-breaking representation, we construct the truncated DPM priors for S_ξ and f_ξ with the Weibull kernel such that

$$S_\xi = 1 - \sum_{l=1}^{L} p_l F_w(\psi_l, \nu_l), f_\xi = \sum_{l=1}^{L} p_l f_w(\psi_l, \nu_l), p_l \sim \text{SB}(1,c), (\psi_l, \nu_l) \sim G_0,$$

where L is the truncation number, and F_w and f_w denote CDF and density of Weibull distribution, respectively. We fix the truncation number L rather than sampling it to simplify computation as a common strategy (Rodriguez et al., 2008). Let $S_\xi^{(\infty)}$ denote the limit of the DPM model, and $S_\xi^{(L)}$ denote the truncated form. The truncation number L is generally selected such that the L_1 error between the limit form and the truncated form, denoted as $\int_0^{+\infty} |S_\xi^{(\infty)}(s) - S_\xi^{(L)}(s)|ds$, is as small as possible. As shown by Ishwaran and James (2002), this L_1 error is bounded by $4n\exp\{-(L-1)/c\}$, where n denotes the sample size. In practice, an error bound of 0.01 is considered to be sufficiently small (Ohlssen et al., 2007). Since we fix the total mass parameter $c = 1$ as a common practice, for sample size $n < 600$, $L = 12$ is a suitable choice of truncation number. In our numerical studies, we find that an L in the range of $10 - 15$ is appropriate to approximate the DPM model well. Users of `BuLTM` are free to adjust the truncation number according to the data size.

Let G_0 be the base measure for (ψ_ι, v_ι). We recommend choosing $G_0 = \text{Gamma}(1, 1) \times \text{Gamma}(1, 1)$ as the specified base measure without any hyperprior for it. The setting of G_0 in our approach implies that $E\{F_\xi(t)\} = 1 - \exp(-t)$ i.e the nonparametric transformation model is centering around the PH model. Such elicitation of the DPM model is a weakly informative prior for S_ξ since the variance of the DP is finite (Nieto-Barajas et al., 2012). Note that it is nontrivial to select the hyperprior for G_0. For the base measure in the DPM with Weibull kernel, Kottas (2006) proposed a Uniform-Pareto (Upar) prior, and Shi et al. (2019) proposed a low information omnibus (LIO) prior, while neither of them is applicable to our method. The Upar prior is not applicable to our unidentified models since the Upar prior is noninformative to (ψ, ν); otherwise, the MCMC algorithm can hardly converge. The LIO prior is a kind of hierarchical specification, which is too complicated to be incorporated into our method with a heavy computation burden.

3.8.3 Relationship between the quantile-knots I-splines prior and the NII process

We summarize the relationship between the quantile-knots I-splines prior and the nonnegative independent increment process here. Let $s_0 = 0 < s_1 < s_2 < \cdots < s_J = \tau$ and we get J disjoint partitions $[0, s_1], (s_1, s_2], \cdots, (s_{J-1}, s_J]$ of D. Note that each I-spline function starts at 0 in an initial flat region, increases in the mid region, and then reaches 1 a the end (Wang and Dunson, 2011b). Therefore, the range of all I-spline functions is $[0, 1]$. Then we determine the I-spline basis functions with knots $s_0 = 0 < s_1 < s_2 < \cdots < s_J = \tau$ and smoothness order $r > 1$ as $\{B_j(t)\}_{j=1}^{K=J+r}$. We call two I-spline functions $B_{j_1}(t)$ and $B_{j_2}(t)$ are "*joint*" on a certain interval D_i for $i = 1, \cdots, J$, if $\exists t' \in D_i$ such that $B_{j_1}(t'), B_{j_2}(t') \in (0, 1)$. Otherwise, they are "*disjoint*" on D_i. We also call an I-spline function $B_j(t)$ "*crosses*" an interval D_i if

$\exists t' \in D_i$ such that $0 < B_j(t_0) < 1$.

We divide all K I-spline basis functions into r groups. Among the r groups, for $\iota = 1, \ldots, r$, the ιth group consists of $B_\iota, B_{\iota+r}, B_{\iota+2r}, \ldots$ such that all I-spline functions in this group are disjoint. That is, for any D_i, only one of the I-spline functions within the ιth group crosses the interval D_i. We define the combination of I-spline functions within the ιth group as

$$H_\iota(t) = \sum_{k\geq 1} \alpha_{\iota+kr} B_{\iota+kr}(t).$$

Then $H_\iota(t)$ has independent increments among all knots $s_0 = 0 < s_1 < s_2 < \cdots < s_J = \tau$, if the coefficients $\{\alpha_{\iota+kr}\}_{k\geq 1}$ are independent positive variables. Therefore, $H\iota$, the combination of I-splines functions within the ιth group is an NII process with independent increment on fixed locations (Phadia, 2015, pp.129). Then we rewrite the equation (6) in the manuscript, the I-splines model into the sum of H_ι

$$H(t) = \sum_{j=1}^{K=J+r} \alpha_j B_j(t) = \sum_{\iota=1}^{r} H_\iota(t).$$

This equation clearly shows that the quantile-knots I-splines prior is a combination of r groups of NII processes. Specifically, when $r = 1$, all I-spline functions are disjoint and therefore, the combination of them reduces to the piecewise exponential model if $\alpha_j \sim \exp(\eta)$ independently. Actually, the first step of determining the initial knots in the quantile-knots I-splines prior is similar to the construction of the piecewise exponential prior in survival models, where partitions of time axis are often taken on empirical quantiles of uncensored survival times (de Castro et al., 2014).

3.8.4 Alternative I-splines priors for H

One may consider other alternative choices of parametric and nonparametric priors for the triplet $(\boldsymbol{\beta}, H, S_\xi)$. Here we introduce some alternative choices of priors. It includes how to construct constrained priors to make the MTM identified. Another construction of I-splines prior with shrinkage prior for H is also given here.

3.8.4.1 Fully identified priors

In this subsection, we discuss the construction of identified priors. Our spirit is from Horowitz's normalization conditions. Like the manuscript, we use the unit scale condition that $||\boldsymbol{\beta}|| = 1$ as an equivalent condition of Horowitz's scale normalization. Rather than applying posterior projection, we assign the uniform distribution on the p-dim unit hypersphere as the prior for the fully identified $\boldsymbol{\beta}$. It is conducted by the following transformation

$$\boldsymbol{\beta}_* \sim N(0, I), \boldsymbol{\beta} = \boldsymbol{\beta}_*/||\boldsymbol{\beta}_*||^{1/2}.$$

Still, we need the location normalization, which assumes that the $H(t_0) = 1$ or $h(t_0) = 0$ for some finite t_0 (Horowitz, 1996). We adopt the I-spline priors as our initial. We formulate H by

$$H(t) = \sum_{j=1}^{K} \alpha_j B_j(t),$$

where $K = J + r$ is the number of I-spline functions; see *Section 3.8.3*. By the characteristic of I-spline functions on interval $D = (0, \tau]$, if $\sum_{j=1}^{K} \alpha_j = 1$, H will surely pass the point $(\tau, 1)$. That is, for $h = \log H$, we have $h(\tau) = 0$. Therefore, the location normalization condition is transferred to a sum-to-one restriction, that is, $(\alpha_1, \ldots, \alpha_K)$ is a K-dim simplex. We consider two choices of priors for the p-dim simplex. The first one is the Dirichlet prior

$$(\alpha_1, \ldots, \alpha_K) \sim \text{Dir}(a_1, \ldots, a_K),$$

where $\{a_j\}_{j=1}^{K}$ are hyperparameters of Dirichlet distribution. Alternatively, we may consider a kind of transformed prior. For $j = 1, \ldots, K$,

$$\alpha_j^* \sim \text{exp}(\eta),\ \alpha_j = \alpha_j^* \Big/ \sum_{j=1}^{K} \alpha_j^*.$$

Both these two priors normalize the location of H and therefore, fully identify the transformation function.

The above priors make the transformation model fully identified. However, with these priors, we find that the MCMC procedure by NUTS converges very slowly and suffers from poor mixing. What's worse, the prediction accuracy is poor. These two drawbacks force us not to work on a fully identified model.

3.8.4.2 The shrinkage prior and comparison

We here introduce the commonly used shrinkage priors for I-spline functions as an alternative to the proposed quantile-knots I-splines prior for H. All I-splines variant priors for H and H' have the same shell

$$H(t) = \sum_{j=1}^{K} \alpha_j B_j(t), H'(t) = \sum_{j=1}^{K} \alpha_j B_j'(t).$$

However, unlike the proposed prior which selects knots from empirical quantiles of observed survival times, the traditional I-splines prior selects sufficiently many (usually from 10 to 30) equally spaced knots from the observed time interval (Cai and Dunson, 2007; Wang and Dunson, 2011a; among others). Then, to avoid overfitting due to using too many knots, one has to incorporate a shrinkage prior for the coefficients α_j to select appropriate I-spline functions. We here consider the truncated generalized double Pareto prior :

$$\alpha_j \sim N^{+}(0, \sigma_j^2), \sigma_j \sim \text{exp}(\eta_j), \eta_j \sim \text{Ga}(\theta, \zeta),$$

where N^{+} denotes the truncated Gaussian distribution such that $\alpha_j > 0$. This is a truncated form of the widely used generalized double Pareto prior as shrinkage prior for coefficients of basis functions (Gelman et al., 2013). In general, $\theta = \zeta = 1$ are typical default hyperparameters. In `BuLTM`, we further simplify this prior as $\sigma_j \sim \text{exp}(1)$. The use of shrinkage prior for I-splines functions may be sensitive to the number of knots (Perperoglou et al., 2019). In our experience, as the number of knots increases, the computation burden of the shrinkage prior becomes heavier while it may not improve the accuracy of final model results. Therefore, the use of shrinkage priors may be accompanied by a time-consuming tuning procedure to determine the best number of equally spaced knots. We compare the shrinkage prior using 15 equally spaced knots and the proposed quantile-knots I-splines prior under model setting Case 1 in the manuscript. Table 3.3 shows the parametric estimation and root integrated square error (RISE) of estimated baseline survival probability functions using these two nonparametric priors in 100 Monte Carlo replications. We find that both priors provide similar estimation results whereas the proposed quantile-knots I-splines prior perform slightly better.

Table 3.3: Parametric estimation results employing two nonparametric priors for H (standard deviation in bracket) and RISE of estimated baseline survival probability functions.

	Quantile-knots	Shrinkage
$\beta_1 = 0.577$	0.579(0.070)	0.581(0.069)
$\beta_2 = 0.577$	0.578(0.050)	0.576(0.049)
$\beta_3 = 0.577$	0.575(0.050)	0.574(0.050)
RISE	0.063	0.064

3.8.5 Proof of Theorem 3.1

Proof. Let $\Theta = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{p}, \boldsymbol{\psi}, \boldsymbol{\nu})$ and $p(\Theta)$ be the product of priors of elements in Θ. To show the posterior $\pi(\Theta)$ is proper is equivalent to show that $\int_{D_\Theta} \pi(\Theta) d\Theta < \infty$, where D_Θ is the domain of Θ.

Let B_j be the I-splines functions, for $j = 1, \ldots, K$. Let $f_w\{\cdot; \psi_l, \nu_l\}$ be the Weibull PDFs with parameters ψ_l and ν_l, for $l = 1, \ldots, L$. By condition (v), let n_1 be the number of uncensored observations and n_0 be the number of censored observations such that $n = n_1 + n_0$, and then we have

$$\begin{aligned}\mathcal{L}(\Theta) < \mathcal{L}^*(\Theta) &\equiv \prod_{i=1}^{n_1} f_\xi\{H(T_i)\exp(-\boldsymbol{\beta}^T\mathbf{Z}_i)\}H'(T_i)\exp(-\boldsymbol{\beta}^T\mathbf{Z}_i) \\ &= \prod_{i=1}^{n_1}\sum_{j=1}^{K}\alpha_j B_j'(T_i)\exp(-\boldsymbol{\beta}^T\mathbf{Z}_i)\sum_{l=1}^{L} p_l f_w\{\exp(-\boldsymbol{\beta}^T\mathbf{Z}_i)\sum_{j=1}^{K}\alpha_j B_j(T_i); \psi_l, \nu_l\}.\end{aligned}$$

By condition (ii), we first integrate out all p_l and it remains to show that

$$\mathcal{A}_l = \int_{D_{\Theta|-p_l}} \left\{ \prod_{i=1}^{n_1} [\exp(-\boldsymbol{\beta}^T \mathbf{Z}_i) f_w\{\exp(-\boldsymbol{\beta}^T \mathbf{Z}_i) \sum_{j=1}^{K} \alpha_j B_j(T_i); \psi_l, \nu_l\} \sum_{j=1}^{K} \alpha_j B_j'(T_i)] \right.$$
$$\left. \times\, p(\Theta| - p_l) d(\Theta| - p_l) \right\} < \infty,$$

for all l, where $\Theta| - p_l$ denotes all parameters except p_ls and $D_{\Theta|-p_l}$ denotes corresponding domains.

Let $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)^T, \boldsymbol{\phi}_i = (B_1'(T_i), \cdots, B_K'(T_i))^T$ and $\mathbf{\Phi}_i = (B_1(T_i), \cdots, B_K(T_i))^T$. For any $0 < T_i < \infty$, by the definition of I-splines function, we have $0 < \boldsymbol{\alpha^T} \phi_i < \infty$ and $0 < \boldsymbol{\alpha^T} \Phi_i < \infty$. Therefore, we have $0 < \boldsymbol{\alpha^T} \phi_i / \boldsymbol{\alpha^T} \Phi_i < \infty$. Let $M_0 = \max(\boldsymbol{\alpha^T} \phi_1 / \boldsymbol{\alpha^T} \Phi_1, \ldots, \boldsymbol{\alpha^T} \phi_{n_1} / \boldsymbol{\alpha^T} \Phi_{n_1})$. Then by condition (iv),

$$\exp(x) f_w\{\exp(x) \boldsymbol{\alpha}^T \boldsymbol{\phi}_i; \psi_l, \nu_l\} \boldsymbol{\alpha}^T \mathbf{\Phi}_i \leq M_0 \{\exp(x) \boldsymbol{\alpha}^T \boldsymbol{\phi}_i\} f_w\{\exp(x) \boldsymbol{\alpha}^T \boldsymbol{\phi}_i; \psi_l, \nu_l\} < \infty$$

for all $x \in \mathbb{R}$.

By condition (v), we can find p uncensored observations such that the $p \times p$ matrix of their covariates, with each row being the vector of covariates of one observation, is full rank. Let Z^* denote that full rank p matrix and let $\boldsymbol{\gamma} = -Z^* \boldsymbol{\beta} = (\gamma_1, \cdots, \gamma_p)^T$. Thus, any $-\boldsymbol{\beta}^T \mathbf{Z}_i$ can be expressed as a linear combination of $(\gamma_1, \ldots, \gamma_p)$ i.e $-\boldsymbol{\beta}^T \mathbf{Z}_i = \sum_{h=1}^{p} c_{ih} \gamma_h$. That is, for $i = 1, \ldots, n_1$

$$f(\gamma_1, \ldots, \gamma_p) = \exp(\sum_{h=1}^{p} c_{ih} \gamma_h) f_w\{\exp(\sum_{h=1}^{p} c_{ih} \gamma_h) \boldsymbol{\alpha}^T \boldsymbol{\phi}_i; \psi_l, \nu_l\} \boldsymbol{\alpha}^T \mathbf{\Phi}_i < \infty.$$

Meanwhile, since Z^* is a one-on-one linear operation of $\boldsymbol{\beta}$, the integrand $\boldsymbol{\beta}$ can be transferred to $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$. Let $T^* = (T_1^*, \ldots, T_p^*)$ denote the survival outcomes of the p subjects with covariates $(Z_1^*, \ldots, Z_p^*)^T = Z^*$. By simple algebra, we have

$$\begin{aligned}
\mathcal{A}_l &\leq M_1 \int_{D_{(\Theta|-p_l,\beta^*)}} p(\Theta|-p_l)[\int_{\mathbb{R}^p} \prod_{h=1}^{p} \exp(\gamma_h) f_w\{\exp(\gamma_h)\boldsymbol{\alpha}^T\boldsymbol{\Phi}_h; \psi_l, \nu_l\} \\
&\quad \times \boldsymbol{\alpha}^T\boldsymbol{\phi}_h d\gamma_1 \cdots d\gamma_p] d(\Theta|-p_l) \\
&\leq M_1 \int_{D_{(\Theta|-p_l,\beta^*)}} p(\Theta|-p_l) d(\Theta|-p_l) \prod_{h=1}^{p} \int_{-\infty}^{+\infty} \exp(\gamma_h) f_w\{\exp(\gamma_h) \sum_{j=1}^{K} \boldsymbol{\alpha}^T\boldsymbol{\Phi}_h; \psi_l, \nu_l\} \\
&\quad \times \boldsymbol{\alpha}^T\boldsymbol{\phi}_h d\gamma_h \equiv \mathcal{B}_l,
\end{aligned}$$

where M_1 is a constant. The first inequality can be derived directly from previous results and the second inequality is the Cauchy–Schwarz inequality.

Finally, we have

$$\begin{aligned}
\mathcal{B}_l &\leq M_1 M_0^p \int_{D_{(\Theta|-p_l)}} p(\Theta|-p_l) d(\Theta|-p_l) \prod_{h=1}^{p} \int_{\infty}^{+\infty} \exp(\gamma_h) f_w\{\exp(\gamma_i); \psi_l, \nu_l\} d\gamma_h \\
&= M_1 M_0^p \int_{D_{(\Theta|-p_l)}} p(\Theta|-p_l) d(\Theta|-p_l) \prod_{h=1}^{p} \int_{0}^{+\infty} f_w\{\exp(\gamma_h); \psi_l, \nu_l\} d\{\exp(\gamma_h)\} \\
&= M_1 M_0^p < \infty.
\end{aligned}$$

The first equation includes product of (p+1) integrals of PDFs $p(\Theta|-p_l)$ and $f_w\{\exp(\gamma_h); \psi_l, \nu_l\}$, $l = 1, \ldots, p$. Therefore, the posterior is proper. □

3.8.6 Additional simulation results

We report additional simulations here. We first introduce the reproducibility of all simulations, and report the results of simulations in highly-censored cases, results of parametric estimation under AFT models, and results of effective sample size (ESS) given by BuLTM in simulations.

3.8.6.1 Reproducibility of simulations

This subsection is about details for the reproducibility of our simulation results. In all simulations, we run four independent parallel chains in BuLTM as the default setting in Stan. The length of each chain is 2500 with the first 500 iterations burn-in and we aggregate four chains to obtain total 8000 posterior samples without any thinning. The MCMC procedure in spBayesSurv draws the same number of samples as ours. In all simulations, we set $L = 12$ for the truncation number of DPM $v = 1$ for the total mass parameter, and $r = 3$ for the order of smoothness of I-spline functions. In case the censoring rate is higher than 50%, we use 5 initial knots; when the censoring rate is less than 50% we use 6 initial knots in constructing the quantile-knots I-splines prior. The coefficients $\{\alpha_j\}_{j=1}^{K}$ are assigned exponential prior with parameter 1. The credible interval of estimates given by BuLTM is the default central posterior interval in Stan; the credible interval of estimates given by spBayesSurv is the highest posterior density interval computed by R package HDInterval; the confidence intervals of TransModel are 95% Wald-type intervals. All numerical studies are realized in R version 4.1.0 with rstan version 2.26.4.

3.8.6.2 Low censoring cases

We assess BuLTM under four cases with high censoring rates. These model settings are similar to the model settings used in the manuscript while the censoring rates are all less than 50%.

Simulated data are generated under the following settings.

LCase 1. Non-PH/PO/AFT : $\epsilon \sim 0.5N(0.5, 0.5^2) + 0.5N\{1, 1\}$,

$h(t) = \log[(0.6t + 0.78t^{1/2} + 0.745)\{0.5\Phi_{0.5,1}(t) + 0.5\Phi_{4,0.5}(t) - c_1\}], C \sim \mathrm{U}(3.5, 5)$;

LCase 2. PH model : $\epsilon \sim \mathrm{EV}(0, 1)$,

$h(t) = \log[(t + 1.213t^{1/2} + 1.5)\{0.5\Phi_{0.5,1}(t) + 0.5\Phi_{3.5,0.3}(t) - c_2\}], C \sim \mathrm{U}(1, 5)$;

LCase 3: PO model : $\epsilon \sim \mathrm{Logistic}(0, 1)$,

$h(t) = \log[(t + 1.213t^{1/2} + 1.5)\{0.5\Phi_{1,0.5}(t) + 0.5\Phi_{4.5,0.3}(t) - c_3\}], C \sim \mathrm{U}(3.5, 5)$;

LCase 4: AFT model : $\epsilon \sim N(0, 1^2), h(t) = \log(t), C \sim \mathrm{U}(2.5, 5)$.

The censoring variable C is generated independent of $\mathbf{Z}$, leading to approximately 39%, 29%, 24%, and 25% censoring rates respectively. For each prediction scenario, we compare the PPDs of three new observations with sets of covariates: $\mathbf{Z}_1 = (0, 0, 0)^T, \mathbf{Z}_2 = (1, 1, 1)^T$ and $\mathbf{Z}_3 = (0, 1, 1)^T$, respectively.

Table 3.4: The RISEs between the conditional survival curves and true curves predicted by `BuLTM`, `spBayesSurv`, and `TransModel` under LCases 1 to 4. Data size $n = 200$.

	Case 1: Non- PH/PO/AFT							Case 2: PH			Case 3: PO			Case 4: AFT		
Z	BuLTM	PH	PO	AFT	$r = 0$	$r = 0.5$	$r = 1$	BuLTM	PH	$r = 0$	BuLTM	PO	$r = 1$	BuLTM	AFT	$r = 1$
$\mathbf{Z}_1$	**0.100**	0.147	0.132	0.156	0.159	0.141	0.127	**0.065**	0.073	0.067	**0.078**	0.083	0.083	0.074	**0.060**	0.094
$\mathbf{Z}_2$	**0.076**	0.163	0.087	0.107	0.152	0.114	0.094	0.138	0.229	**0.117**	0.118	0.122	**0.104**	0.104	**0.090**	0.114
$\mathbf{Z}_3$	**0.103**	0.207	0.139	0.179	0.189	0.156	0.134	0.133	0.220	**0.112**	0.120	0.128	**0.104**	0.113	**0.095**	0.114

Table 3.4 shows that `BuLTM` still works well when the censoring rate goes high. We find that when the censoring rate is lower than 50%, `BuLTM` outperforms `spBayesSurv` under Non-PH/PO/AFT and PH models, is comparable under the PO and the AFT models. This result is in line with the results we report in the manuscript. Results of parametric estimation are summarized in Table 3.5 for LCases 1-3, which are also consistent with the results given by low-censoring cases.

Table 3.5: The performance of parametric estimation of `BuLTM` and `spBayesSurv` under LCases 1-3.

Case 1: Non-PH/PO/AFT		BuLTM					spBayesSurv/TransModel				
	Parameter	BIAS	RMSE	PSD	SDE	CP					
	β_1	0.016	0.090	0.082	0.090	92.0					
	β_2	-0.015	0.070	0.063	0.066	90.3					
	β_3	-0.001	0.069	0.062	0.068	91.7					
		Case 2: PH					Case 3: PO				
Method	Parameter	BIAS	RMSE	PSD	SDE	CP	BIAS	RMSE	PSD	SDE	CP
BuLTM	β_1	-0.013	0.123	0.123	0.121	94.7	0.012	0.169	0.174	0.167	94.0
	β_2	**0.006**	0.083	0.087	0.081	95.0	**-0.008**	0.130	0.123	0.123	92.3
	β_3	0.006	0.086	0.088	0.085	95.0	-0.005	0.130	0.122	0.124	94.0
spBayesSurv	β_1	-0.032	0.172	0.175	0.170	95.0	**0.002**	0.258	0.256	0.259	94.7
	β_2	-0.026	0.088	0.095	0.084	95.3	0.010	0.142	0.136	0.142	94.7
	β_3.	-0.027	0.102	0.095	0.098	93.0	0.013	0.135	0.136	0.135	95.0
TransModel	β_1	**-0.004**	0.172	0.173	0.172	94.3	-0.011	0.283	0.275	0.283	94.3
	β_2	0.007	0.094	0.095	0.094	95.7	**0.008**	0.146	0.145	0.146	95.7
	β_3	**0.005**	0.010	0.010	0.010	96.7	**0.003**	0.148	0.144	0.148	92.3

3.8.6.3 Parametric estimation under AFT models

Results of parametric estimation are given by Table 3.6, where we find `BuLTM` has lower RMSE than `spBayesSurv` for all parameters. In terms of BIAS, `BuLTM` outperforms `spBayesSurv` in the highly-censored case and is comparable in the case with the lower censoring rate. This result as well as results of prediction demonstrate that `BuLTM` performs robustly under the AFT model. We omit the results of `TransModel` here since they cannot correctly specify the model and hence the results of parametric estimation are meaningless.

Table 3.6: Results of estimation of $\boldsymbol{\beta}$ under AFT models.

		LCase4: AFT1, 25% Censored					Case4: AFT2, 61% Censored				
Method	Parameter	BIAS	RMSE	PSD	SDE	CP	BIAS	RMSE	PSD	SDE	CP
BuLTM	β_1	0.017	0.107	0.102	0.107	92.3	**0.011**	0.138	0.130	0.138	94.0
	β_2	-0.009	0.079	0.076	0.076	92.3	**-0.004**	0.101	0.095	0.098	93.0
	β_3	-0.008	0.079	0.077	0.081	92.7	**-0.007**	0.101	0.094	0.093	95.0
spBayesSurv	β_1	**0.000**	0.159	0.150	0.159	90.3	0.016	0.207	0.194	0.206	92.0
	β_2	**0.002**	0.078	0.079	0.078	92.3	0.016	0.105	0.103	0.104	91.0
	β_3	**0.003**	0.084	0.079	0.084	92.0	0.014	0.101	0.103	0.100	93.7

3.8.6.4 Effective sample size of $\boldsymbol{\beta}$

The effective sample size (ESS) is useful as a first-level check when analyzing the reliability of inference. It measures how many independent draws contain the same amount of information as the dependent posterior samples obtained by the MCMC procedure. ESS is usually accompanied by $\hat{R}$, the diagnostics of convergence of MCMC. In an MCMC procedure, especially the case where multiple chains are used, very low ESS may be caused by divergent chains or poor mixing and hence, large $\hat{R}$. If one obtains sufficient ESS (ESS that is greater than 400 is considered to be sufficient by Vehtari et al. (2021)) after sampling, it is highly possible that all chains are converged and well mixed. Therefore, we report ESS of $\boldsymbol{\beta}$ in our simulation studies here as the diagnosis of MCMC.

Results of the average estimated ESS of $\boldsymbol{\beta}$ in all the simulation studies in the manuscript are given by Table 3.7, from which we find in each simulation the ESS of $\boldsymbol{\beta}$ is sufficiently large. This is owed to the NUTS used by `Stan`, which is more possible to sample nearly independent draws (Hoffman et al., 2014). In terms of other parameters, only a few parameters suffer from low ESS in sporadic Monte Carlo replications as a drawback of the analysis of unidentified models. Even so, the MCMC algorithm is still well converged and mixed examined by $\hat{R}$ in `Stan` and thus the final model results are reasonable. Therefore, when using `BuLTM`, one can simply increase the length of MCMC chains to obtain sufficient ESS for all parameters in all situations regardless of the lack of identifiability. Particularly, if one's interest falls on estimating $\boldsymbol{\beta}$, the vector of regression parameters, the length of chains needed is quite small, and the required computation burden is mild.

Table 3.7: The average estimated ESS of $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T$ in simulation studies.

	Case 1	HCase 1	Case 2	HCase 2	Case 3	HCase 3	Case 4	HCase 4
β_1	5935.29	6201.22	6697.80	6187.15	6026.63	5744.58	7014.90	6431.38
β_2	5573.79	6243.93	7305.17	6800.33	6697.69	6302.88	7497.02	6900.31
β_3	5591.05	6193.69	7307.38	6757.07	6689.56	6263.12	7487.56	7053.08

3.8.7 Sensitivity analysis

We analyze the sensitivity of the proposed quantile-knots I-splines prior for H in this section. There are two pre-specified hyperparameters in the prior, the hyperparameter η for the exponential prior, and the number of initial knots. Here we show that the final prediction results are not sensitive to either the initial number of initial knots or the hyperparameter η.

3.8.7.1 Sensitivity of number of initial knots

Sensitivity analysis of the choice of the initial number of basic knots (N_I) in the quantile-knots I-splines prior is conducted by 100 Monte Carlo studies under Case 1 setting in the manuscript. Candidates for the number of initial knots are taken from the range 5 to 11, where we display results of using $5, 6$, and 11 initial knots here for comparison. Results of parametric estimation and the RISE of estimated baseline survival probability curves among different numbers of initial knots are shown in Table 3.8, where we find with different choices of N_I, both results of parametric estimation and RISE of estimated survival probability curves have very mild variation. Figure 3.9 displays plots of average estimated baseline survival probability curves under three choices of the number of initial knots, where we find they are close to each other. This sensitivity analysis numerically demonstrates that the quantile-knots I-splines prior is not sensitive to its choice of the number of knots. And therefore, it is generally tuning-free and computationally expedient.

Table 3.8: Parametric estimation results (standard deviation in bracket) and RISE of estimated baseline survival probability functions under different choices of η .

	$N_I = 5$	$N_I = 6$	$N_I = 11$
$\beta_1 = 0.577$	0.578(0.070)	0.580(0.070)	0.586(0.069)
$\beta_2 = 0.577$	0.575(0.051)	0.575(0.051)	0.572(0.052)
$\beta_3 = 0.577$	0.561(0.058)	0.560(0.058)	0.557(0.058)
RISE	0.063	0.063	0.066

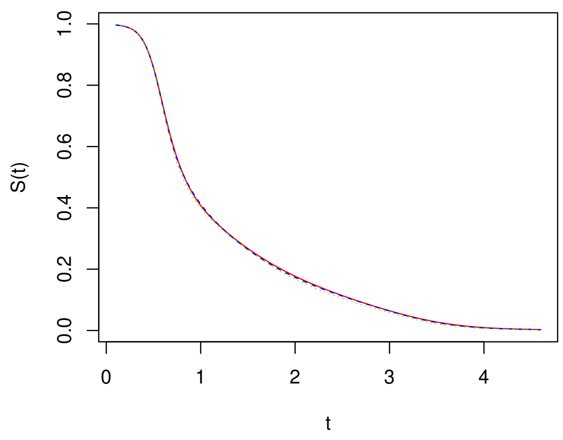


Figure 3.9: Pointwise mean estimated baseline survival probability curves under 100 replications. Real line, $N_I = 5$; dash line, $N_I = 6$; dotted line, $N_I = 11$.

3.8.7.2 Sensitivity of η

Let η be the hyperparameter of exponential prior for coefficients of the quantile-knots I-splines prior in equation (8) in the manuscript. Sensitivity analysis of η is conducted under the setting Case1 in the manuscript. For the sensitivity of η, among 100 Monte Carlo replications, we choose η from three candidates of $\eta = 1, 5,$ and 0.2, corresponding to three levels of informative priors. Notice that we should avoid using too small η since it implies too large prior variance, then the prior is not sufficiently informative anymore. Similarly, too large η induces too small variance, which is too informative to provide sufficient uncertainty.

Results of parametric estimation and RISE of estimated baseline survival curves are given in Table 3.9. From the table, we find that estimation of the parametric component varies quite little among all choices of η and the RISE of estimated baseline survival curves is almost the same with different values of η. For visualization, plots of estimated survival curves given different values of η are shown in Fig 3.10, where we find all estimated curves are close to each other. This sensitivity analysis numerically demonstrates that the quantile-knots I-splines prior are not sensitive to the choice of η within the range 0.2 to 5. Therefore, it is safe to fix η rather than to assign a hyperprior for it.

Table 3.9: Parametric estimation results (standard deviation in bracket) and RISE of estimated baseline survival probability functions under different choices of η.

	$\eta = 1$	$\eta = 5$	$\eta = 0.2$
$\beta_1 = 0.577$	0.580(0.070)	0.571(0.071)	0.592(0.068)
$\beta_2 = 0.577$	0.575(0.052)	0.578(0.051)	0.569(0.052)
$\beta_3 = 0.577$	0.560(0.058)	0.564(0.057)	0.553(0.059)
RISE	0.063	0.064	0.065

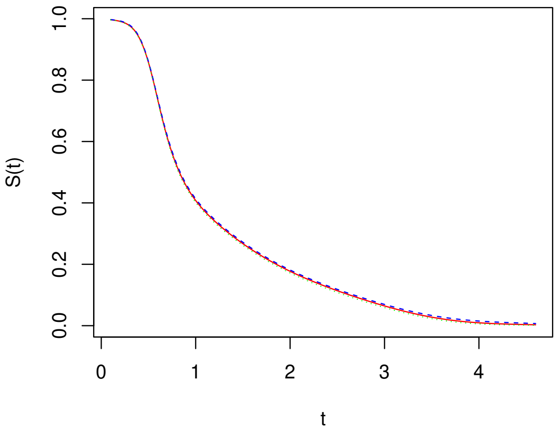


Figure 3.10: Pointwise mean estimated baseline survival probability curves in 100 replications. Real line, $\eta = 1$; dash line, $\eta = 5$; dotted line, $\eta = 0.2$.

3.8.8 Results of parametric estimation on real datasets

3.8.8.1 Veterans lung cancer data

Results of parametric estimation for the veterans lung cancer data given by `BuLTM`, `TransModel` and `spBayesSurv` are displayed in Table 3.10. The three methods provide similar significance levels for all coefficients. Although some signs of estimated coefficients are different, say β_3 and β_7, they are not significant since their credible/confidence intervals cover zero. That implies qualitative interpretations of the estimates of the regression parameter under the three models are stable.

Table 3.10: Results of estimated β for veterans administration lung cancer data. Credible intervals are given on 95% credibility for BuLTM and spBayesSurv. The confidence interval of TransModel is a 95% Wald-type confidence level.

	BuLTM		spBayesSurv		TransModel	
Covariate	Estimate	95%CI	Estimate	95%CI	Estimate	95%CI
Z_1	0.119	(0.045, 0.246)	0.617	(0.449, 0.800)	0.553	(0.368, 0.737)
Z_2	-0.302	(-0.951, 0.897)	-1.391	(-8.597, 6.028)	-0.388	(-8.546, 7.768)
Z_3	-0.006	(-0.700, 0.671)	1.426	(-1.643, 4.477)	0.945	(-2.441, 4.331)
Z_4	0.081	(-0.693, 0.730)	0.033	(-3.533, 3.469)	0.010	(-3.475, 3.496)
Z_5	-0.044	(-0.227, 0.117)	-0.147	(-0.739, 0.487)	-0.278	(-0.963, 0.405)
Z_6	0.350	(0.093, 0.694)	1.387	(0.396, 2.334)	1.995	(0.063, 3.027)
Z_7	-0.005	(-0.242, 0.205)	0.058	(-0.739, 0.916)	0.413	(-0.514, 1.342)
Z_8	0.274	(0.053, 0.571)	1.367	(0.444, 2.308)	1.364	(0.343, 2.385)

3.8.8.2 Heart failure clinical records data

Results of parametric estimation for the heart failure data given by `BuLTM`, `TransModel`, and `spBayesSurv`are displayed in Table 3.11. We find that `BuLTM` is consistent with `spBayesSurv` in the detection of significance, while `TransModel` fails to detect the significance of the covariate Z_9, serum sodium. Existing medical research has evidenced that lower serum sodium was associated with higher in-hospital and 60-day mortality for heart failure patients (Klein et al., 2005). Hence, the results of `BuLTM` and `spBayesSurv` are more reasonable. That explains why the two Bayesian approaches outperform in prediction.

3.8.9 Posterior checking

We assign weakly informative priors for nonparametric components H and S_ξ, which are not fully objective priors. One may worry whether these priors are so informative that the prior-to-posterior updating is not driven by data. We conduct posterior checking on simulation studies and application examples to check the difference between priors and marginal posterior and obtain similar results. Here we take our application to veterans lung cancer data set as an ex-

Table 3.11: Results of estimated $\boldsymbol{\beta}$ in the analysis to heart failure clinical records data. Credible intervals are given on 95% credibility for BuLTM and spBayesSurv. The confidence interval of TransModel is a 95% Wald-type confidence level.

	BuLTM		spBayesSurv		TransModel	
Covariate	Estimate	95%CI	Estimate	95%CI	Estimate	95%CI
$Z_1 =$ age	-0.163	(-0.433, 0.063)	-4.670	(-6.182, -3.135)	-4.631	(-6.474, -2.788)
$Z_2 =$ anemia	-0.013	(-0.036, -0.001)	-0.412	(-0.764, -0.066)	-0.408	(-0.827, 0.012)
$Z_3 =$ creatinine phosphokinase	-0.002	(-0.010, 0.004)	-0.074	(-0.262, 0.113)	-0.075	(-0.293, 0.143)
$Z_4 =$ diabetes	-0.004	(-0.020, 0.008)	-0.117	(-0.476, 0.256)	-0.125	(-0.560, 0.310)
$Z_5 =$ ejection fraction	0.022	(0.008, 0.060)	0.586	(0.386, 0.785)	4.810	(2.773, 6.847)
$Z_6 =$ high blood pressure	-0.015	(-0.042, -0.001)	-0.460	(-0.807, -0.099)	-0.455	(-0.879, -0.031)
$Z_7 =$ platelets	0.076	(-0.033, 0.389)	1.303	(-2.836, 5.327)	1.384	(-3.392, 6.160)
$Z_8 =$ serum creatinine	-0.012	(-0.033, -0.004)	-0.306	(-0.421, -0.183)	-0.313	(-0.453, -0.173)
$Z_9 =$ serum sodium	0.939	(0.787, 0.997)	41.347	(3.248, 74. 256)	43.077	(-2.777, 88.931)
$Z_{10} =$ sex	0.009	(-0.005, 0.033)	0.222	(-0.185, 0.625)	0.224	(-0.269, 0.716)
$Z_{11} =$ smoking	-0.005	(-0.024, 0.010)	-0.133	(-0.542, 0.282)	-0.148	(-0.641, 0.345)

ample. We take $\alpha_j \sim \exp(1)$ for $j = 1, \dots, K$ as weakly informative priors and $p(\boldsymbol{\beta}) \propto 1$ as flat priors. Figure 3.11 compares the priors and marginal posterior of the first eight coefficients of I-spline functions. For all $\{\alpha_j\}_{j=1}^{8}$, their variance is controlled by the weakly informative prior, demonstrating the fact that the impact of priors remedies the flat likelihood. In addition, most of the coefficients in the I-splines prior vary significantly from the prior, evidencing that data drive the prior-to-posterior updating.

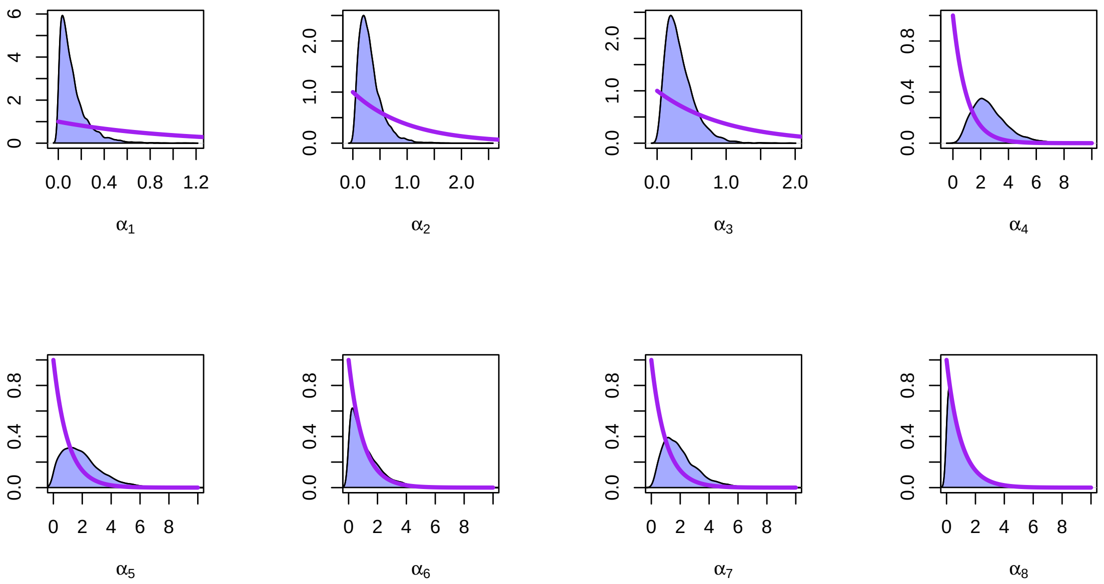


Figure 3.11: Comparison between the the marginal posterior density and priors of $\alpha_1, \dots, \alpha_8$. Shaded region, marginal posterior density; Wide line, prior density of $\exp(1)$.

Note that comparing the prior and posterior of the fully identified parameter $\boldsymbol{\beta}^*$ is meaningless since the projected posterior of $\boldsymbol{\beta}$ is certainly different from its prior. Therefore, in terms of the parametric component, we compare the priors with the marginal posterior of $\boldsymbol{\beta}$, the unconstrained parameter sampled from MCMC. Fig 3.12 shows an apparent difference between flat priors and marginal posterior of $\boldsymbol{\beta}$, demonstrating that the posterior updating is driven by data. An interesting finding is that, even though $\boldsymbol{\beta}$ is unidentified, some of the parameters such as β_1 and β_5 have low posterior variance and posterior intervals that are short enough. This supports the fact that MCMC sampling is workable under unidentified models with weakly informative priors. Meanwhile, we are aware of the necessity of posterior modification by checking the marginal posterior of $\boldsymbol{\beta}$, since the posterior of β_2 and β_4 have heavy-tailed posterior intervals.

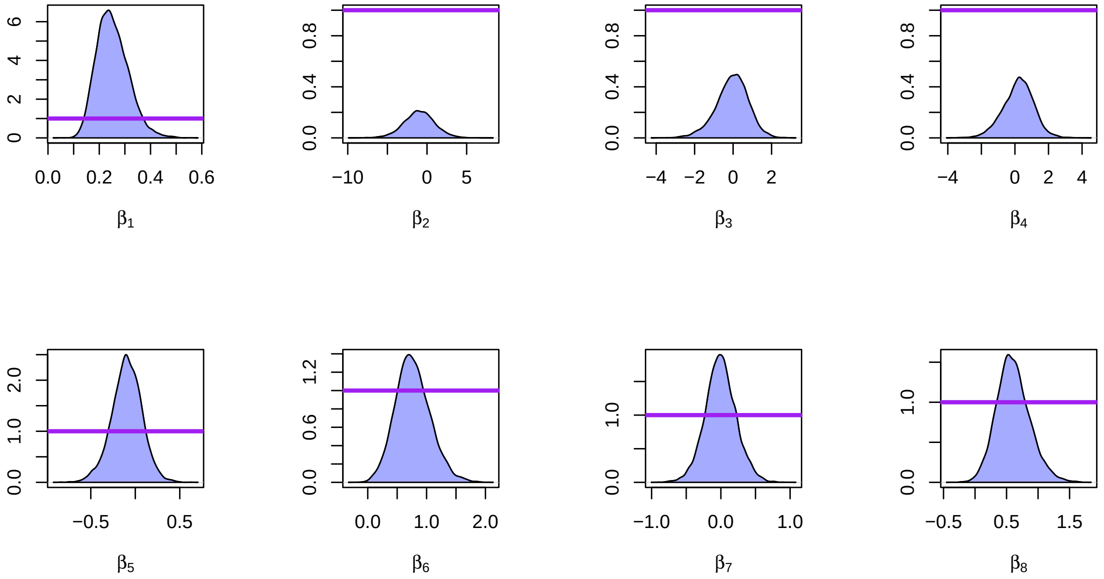


Figure 3.12: Comparison between the the marginal posterior density of β without posterior projection and corresponding priors. The shaded region, posterior density; wide line, flat prior.

3.8.10 Predictive evaluation metrics

In this chapter, we consider two classical metrics to evaluate the predictive capabilities of different survival models, the C index and the integrated Brier score (IBS).

C index

To compute the C-index, we follow the procedure in Ishwaran et al. (2008). We summarize the procedure as follows:

1. Form all possible pairs of survival times over the data.

2. Omit those pairs whose shorter survival time is censored. Omit those tied pairs unless at least one of them is death. Let Permissible denote the total number of permissible pairs.

3. For each untied permissible pair, count 1 if the predicted result is the same as the truth; count 0.5 if the predicted outcomes are tied. For each permissible pair where both are deaths with the same survival time, count 1 if the predicted outcomes are tied; otherwise, count 0.5. For each permissible pair where only one is death and the survival time are tied, count 1 if the death has a worse predicted outcome; otherwise, count 0.5. Let Concordance denote the sum over all permissible pairs.

4. The C index, C, is defined by C = Concordance/Permissible.

IBS

The Brier score (BS) is proposed by Graf et al. (1999) to evaluate prediction at a certain time point t. The BS at time t is formulated as

$$BS(t) = \frac{1}{N}\sum_{i=1}^{n}\left\{\frac{S_{T|\mathbf{Z}}(t|\mathbf{Z}_i)]^2}{\hat{G}(T_i)}I(T_i < t, \delta_i = 1) + \frac{[1 - S_{T|\mathbf{Z}}(t|\mathbf{Z}_i)]^2}{\hat{G}(T_i)}I(T_i \geq t)\right\},$$

where $\hat{G}(T_i)$ denotes estimated survival probability given by the K-M estimator. Then, the IBS is defined as the integral of BS on the interval $(-\infty, \tau)$ for some time $\tau > 0$

$$IBS = \int_{-\infty}^{\tau} BS(t)dt.$$

Chapter 4

Dependent Dirichlet Processes for Analysis of a Generalized Shared Frailty Model

4.1 Introduction

The shared frailty model, coined by Vaupel et al. (1979), has been widely used in the analysis of multivariate survival outcomes that might be associated within subgroups or clusters. Enormous work has been devoted to the development of shared frailty model in both Bayesian and frequency paradigms, and the reviews can be found in Ibrahim et al. (2001); Duchateau and Janssen (2007); Balan and Putter (2020). As an extension of the well-known Cox's proportional hazard model, conditional on the frailty effect, the shared frailty model assumes the hazard ratio between two subjects is proportional to their difference in relative risk scores over time. In addition to the proportional hazard assumption, the shared frailty model fixes the baseline hazard function among all clusters.

Traditional shared frailty models provide a good framework for expediently mathematical tackling the heterogeneity among the multivariate observations, whereas in practice it needs modification and adaption to tolerate complex structure so as to incorporate cross information owing to the intra- and inter-subject variability (Hanson et al. (2012); de Castro et al. (2015)). Take the renowned data on recurrences of bladder cancer for instance (Therneau (2022)). There are three treatment arms, placebo, thiotepa, and pyridoxine. Patients had multiple recurrences

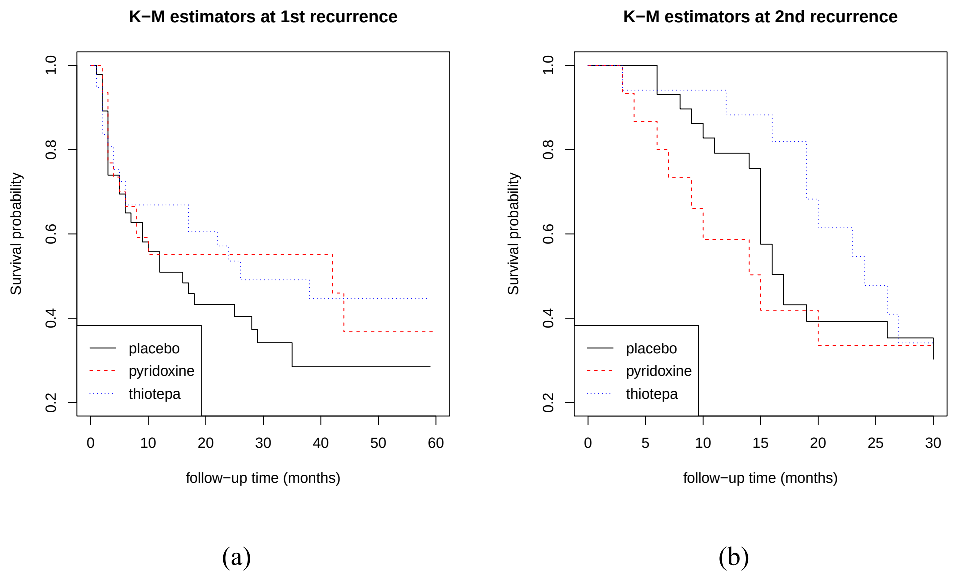


Figure 4.1: The Kaplan-Meier estimator of survival functions for first recurrence time (a) and second recurrence (b) in the bladder cancer data.

of tumors which were sparse beyond the fourth recurrence. Figure 4.1 shows the Kaplan-Meier estimators of the survival function for the times of the first and the second recurrences under three treatments. One observes that, the estimated survival curves at the first recurrence are crossed indicating a crossed hazard and that the proportional hazard assumption is suspected (Zeng and Lin (2007b)); the survival curve of pyridoxine falls below that of placebo at the second recurrence compared to the first recurrence, indicating the functional form of the survival curves varies between recurrences. Neglecting such characteristics of non-proportionality and stratification of recurrences may yield inefficiency by encumbering borrowing strength from potentially related information sources, and consequently may jeopardize the prediction of the global survival times. Moreover, dependency might be existing among the treatment strata and the stratification of recurrences (De Iorio et al. (2004); Hanson et al. (2012)).

Consequently, more complex modeling is needy to characterize the dependence among the baseline hazard functions and treatment strata due to the temporal effects of recurrences. Frequentist inference and computing are pretty challenging and even infeasible. Existing Bayesian literature considered modifications of shared frailty model based on some kind of partial aberrant phenomena (de Castro et al. (2014); Paulon et al. (2020); among others) but rare work has taken bi-level stratification into account (Conlon et al. (2014)), not to mention that dependence among treatment strata (Hanson et al. (2012)).

We propose a generalized shared frailty model (GSFM) for multiple events time data that allows the baseline hazard function to change along with the types of events and treatment strata, strengthening the ability to borrow information from many sources. The proposed model postulates multiplier frailty including both parametric and nonparametric ones, where the parametric frailty random effect accounts for the within subject association by treating each subject as a cluster; and a nonparametric frailty effect represents dependency among treatment strata and temporal recurrences. For the proposed model GSFM, we suggest a Bayesian solution to estimate the regression coefficient vector and the variance parameter of the frailty term, and baseline survival functions stratified by treatments and recurrences. In a Bayesian workflow, the posterior distribution is determined by the combination of observational data in the form of likelihood function and the prior distribution represented based on the background knowledge. From a Bayesian perspective, we model the dependent nonparametric prior through transferring the data context aforementioned into the ANOVA dependent Dirichlet process (ANOVA DDP), which will be further reviewed in Section 2. The construction of No-U-Turn sampler for Markov chain Monte Carlo (MCMC) sampling is automated by Stan (Stan Development Team (2018)) with its R interface (Stan Development Team (2020)). The posterior inference is conducted by Stan as well.

The rest of this chapter is organized as follows. In Section 2, under typical data scenarios of dependence structure, we summarize several modification versions of the dependent Dirichlet process (DDP) initiated from MacEachern's regression spirit that nested dependent predictors into the traditional Dirichlet Process (DP). In Section 3, we postulate the GSFM and transform the dependent dual-stratified multiple events to the survival-function based version of the ANOVA DDP. We have a short comparison between Stan and Nimble, two contemporary Bayesian computing tools based on our user experience. In section 4, we demonstrate the validity of the GSFM and Bayesian inference and analysis of the data on recurrences of bladder cancer. A brief conclusion is contained in Section 5.

4.2 Review of MacEachern's DDP

The DP is the most popular Bayesian nonparametric prior since the seminal work of Ferguson (1974). The belief in data background that there exists some kind of dependence structure stimulates construction and selection of proper dependent prior. Some dependent DPs are constructed for unsupervised purposes such as clustering (Teh et al. (2006); Rodriguez et al. (2008)). The DDP prior adopted in our proposed model is supervised and predictor-dependent, originated from MacEachern (2016); Quintana et al. (2020), named as MacEachern's DDP in two recent review papers, which are interpretive and comprehensive (MacEachern (2016); Quintana et al. (2020)). The key idea behind the MacEachern's DDP is that the distributions of the random measures are marginally DP distributed, validated by in our subsections 3.2 and 3.3. Therefore we here confine how the MacEachern's DDP (henceforth we use the DDP to denote the MacEachern's DDP if the context is clear) came into being expanded from the DP, and compare various modification versions of the DDP under various dependent data structures.

DP vs. DDP

The DP is a distribution on distributions whereas the DDP aims to construct prior for a collection of distributions $\mathcal{F} = \{F_x | x \in \mathcal{X}\}$ indexed by covariate x. In general, there are several representations of the DP such as Polya Urn, Levy measure, and stick-breaking representations (Phadia (2015)). Here we use Sethuraman's stick-breaking construction to connect the DP with the DDP. The stick-breaking construction is a kind of infinite sum representation that divides the DP into two countable series, the weights, and the atoms. Generally, a DP is expressed as a process with two components, the mass parameter determining the weights and the base measure to generate atoms. Through the stick-breaking construction, the DDP can be easily extended from the DP. We list their comparison in Table 4.1, where we can find that the dependency among the covariates set $\mathcal{X}$ is realized by indexing the mass parameter and base measure with the covariate $x \in \mathcal{X}$. More specifically, the dependency can be characterized through the dependency among the weights and atoms in the DDP.

The DDP can be widely applied to scenarios of various dependence data structures. We review modification versions of the DDP from three categories depending on which part it modifies in the stick-breaking representation, weights, atoms, or both. The first is to impose the dependency on the atoms but keep common weights, leading to two typical representatives, ANOVA and Spatial (De Iorio et al. (2004); Gelfand et al. (2005); De Iorio et al. (2009)). The ANOVA type DDP encoded the covariate dependence in the form of regression for the atom processes. The Spatial DDP models for nonstationary spatial random fields with heterogeneous variance. The second category is to modify the weights to be dependent but keep the common atoms. The early and typical work is the time series DDP (Nieto-Barajas et al. (2012)). They introduced a Markov Beta process on the weights to account for the temporal dependency. The third category is to impose dependency on both weights and atoms (DeYoreo and Kottas (2018)). They constructed vector autoregressive and autoregressive models for atoms and weights, respectively. We summarize the aforementioned types of typical modifications in Figure 4.2.

Table 4.1: Comparison of DP & DDP

	DP	DDP
RPM	$F \sim \text{DP}(M, F_0)$	$\mathcal{F} = \{F_x \mid x \in \mathcal{X}, M_x, F_{0x}\}$
Sethuraman's construction	$F(\cdot) = \sum_{h=1}^{\infty} p_h \delta_{\theta_h}(\cdot)$ $p_h \sim \text{SBW}(1, M)$ $\theta_h \sim F_0$	$F_x(\cdot) = \sum_{h=1}^{\infty} p_{xh} \delta_{\theta_{xh}}(\cdot)$ $p_{xh} \sim \text{SBW}(1, M_x)$ $\theta_{xh} \sim F_{0x}$
Convolution	$H(y) = \int k(y \mid \theta) dF(\theta)$	$H_x(y) = \int k(y \mid \theta) dF_x(\theta)$

4.3 Model and Bayesian inference

Consider a clinical trial with multiple event types, for example, the time of the kth recurrence of a certain disease. In the trial, n subjects are divided into G strata of treatment. Our goal is to describe the relationship between the time to the kth recurrence of a subject, and its treatment stratum as well as its vector covariates Z. For a certain subject, the times of recurrences may be dependent since they occur on the same individual and thus we assume an unobservable independent shared-frailty random effect W to account for this dependence. On the other hand,

	DDP			
	Dependent atoms		Dependent weights and atoms	Dependent weights
	ANOVA DDP	Spatial DDP	Discrete-Time-Dependent DPM	tsDDP
DDP	$F_x(\cdot) = \sum_{h=1}^{\infty} p_h \delta_{\theta_{xh}}(\cdot)$	$G_D(\cdot) = \sum_{h=1}^{\infty} p_h \delta_{\theta_{h,D}}(\cdot)$	$F_t(\cdot) = \sum_{h=1}^{\infty} p_{th} \delta_{\theta_{th}}(\cdot)$	$F_d(\cdot) = \sum_{h=1}^{\infty} p_{dh} \theta_h(\cdot)$
Atoms	$\theta_{xh} = \alpha_h^\top d_x$ $\alpha_h \sim F_0$	$\theta_{h,D} \sim G_0$ $G_0^{(n)} \sim N_n(0, \Sigma)$	$\theta_{th} = (\mu_{th}, \Sigma_h)$ $\mu_{th} \sim N(\mu_0, V_o)$ $\mu_{t+1,h} \vert \mu_{th} \sim N(m + \Theta\mu_{th}, V)$	$\theta_h \sim F_0$
Weights	$p_h \sim \text{SBW}(1, M)$	$p_h \sim \text{SBW}(1, M)$	$p_{th} = (1 - \beta_{th}) \Pi_{j=1}^{h=1} \beta_{tj}$ $\beta_{th} = \exp(-(2\alpha)^{-1}(\xi_h + \eta_{th})^2)$ $\xi_h, \eta_{1h} \sim N(0, 1)$ $\eta_{t+1,h} \vert \eta_{th} \sim N(\phi\eta_{th}, 1 - \phi^2)$	$p_{dh} = v_{dh} \Pi_{j=1}^{h-1} (1 - v_{dj})$ $v_{1h} \sim \text{Beta}(1, M)$ $z_{dh} \vert v_{dh} \sim B(m_{dh}, v_{dh})$ $v_{d+1,h} \vert v_{dh} \sim \text{Beta}(1 + z_{dh}, M - z_{dh} + m_{dh})$
Convolution	$H_x = \int N(\alpha^\top d_x, s) dF(\alpha)$ $F \sim \text{DP}(M, F_0)$	$H_D = \int N(\theta_D, \sigma^2) dG(\theta_D)$ $G \sim \text{DP}(M, G_0)$	$H_t = \int N(\mu, \Sigma) dF_t(\mu, \Sigma)$	No convolution
Authors	De Iorio, Müller, Rosner and MacEachern (2004, JASA)	Gelfand, Kottas and MacEachern (2005, JASA)	DeYoreo and Kottas (2018, JASA)	Nieto□Barajas, Müller, Ji, Lu and Mills (2012, Biometrics)

Figure 4.2: Workflows of representative expansions of DDP

we may allow the conditional hazard affiliated with the script pair kj implying distinct survival distributions along with the temporal order of the recurrences of the disease and for specific treatment. For the ith subject in the jth treatment stratum, at the kth recurrence, given the value of frailty variable w_i and its covariate vector z_{kji}, we propose the following frailty model,

$$\lambda_{kj}(t|w_i, z_{kji}) = w_i \lambda_{0kj}(t) \exp(\beta^T z_{kji}), k = 1, \cdots, K, j = 1, \cdots, G, i = 1, \cdots, n_j. \quad (4.1)$$

Model (4.1) is called the *generalized* shared frailty model in the sense that non-proportionality among k-varying recurrences is allowed by the fact that the right-hand baseline hazard has footnotes k and j. We allow dependency among treatment strata in model (4.1). Therefore, the baseline hazard function λ_{0kj} acts as a nonparametric frailty random measure accounting for the dependency owing to the recurrences and treatment scheme.

Model (4.1) is an extension of the classical shared frailty model (4.1.1) on page 101 of Ibrahim et al. (2001) since the baseline hazard function there does not vary from the recurrences and the treatment strata. Model (4.1) has the analog spirit to the frailty model (1) in de Castro et al. (2014), whereas their treatment strata are independent.

4.3.1 Likelihood

The corresponding survival function of model (4.1) is given by:

$$S_{kj}(t|w_i, z_{kji}) = \{S_{0kj}(t)\}^{\exp(\beta^T z_{kji} + v_i)},$$

where S_{0kj} denotes the baseline survival function of the kth recurrence for subjects in the jth treatment stratum, $v_i = \log(w_i)$ denotes logarithm transformation of the frailty effect. Let f_{0kj} be the corresponding baseline density function.

Given the data sample $(Y_{kji}, \delta_{kji}, z_{ji})$, where $Y_{kji} = \min(C_{kji}, T_{kji})$, $\delta_{kji} = I(T_{kji} \leq C_{kji})$, with T_{kji} being the gap time between the $(k-1)$th and kth recurrence of the ith subject in the jth stratum and C_{kij} being the corresponding censoring variable that is independent of T_{kji} given the covariate vector z_{kji}, for $k = 1, \cdots, K, j = 1, \cdots, G, i = 1, \cdots, n_j$, and $\sum_{j=1}^{G} n_j = n$. In the jth stratum, suppose that there are n_{kj} $(n_{kj} \leq n_j)$ subjects suffering from the kth recurrence.

Then the likelihood is written as:

$$\prod_{k=1}^{K}\prod_{j=1}^{G}\prod_{i=1}^{n_{kj}}[\exp(\beta^T z_{kji}+v_i)f_{0kj}(y_{kji})\{S_{0kj}(y_{kji})\}^{(\exp(\beta^T z_{kji}+v_i)-1)}]^{\delta_{kji}}$$
$$\times\{S_{0kj}(y_{kji})\}^{(1-\delta_{kji})\exp(\beta^T z_{kji}+v_i)}.$$

4.3.2 Survival-function based version of the ANOVA DDP

In the Bayesian workflow for the estimation, prior distributions are first determined. We here specify appropriate nonparametric priors for S_{0kj} and f_{0kj}. Since they can be easily derived from one to the other, we here only introduce the priors for S_{0kj}.

We divide S_{0kj} into K groups, and the kth group has G baseline survival functions of different treatment strata at the kth time of recurrence. That is, for a fixed k, $\mathcal{S}_k = \{S_{0kj}, j = 1, \cdots, G\}$ is a collection of baseline survival functions with length G indexed by the categorical covariate j denoting the treatment stratum. The next procedures come from the spirit of De Iorio et al. (2004). As a general example, suppose two dugs A and B will be taken in treatment, with V and U levels of doses, respectively. In this case, $G = VU$ denotes the number of treatment strata and let the level of the jth stratum be (v, w). We write the stick-breaking form of S_{0kj} such that $S_{0kj}(t) = 1 - \sum_{h=1}^{\infty} p_h I(t > \theta_{kjh})$, where $\{p_h\}$ are the strick-breaking weights of the DP. We impose an ANOVA structure on θ_{kjh} :

$$\theta_{kjh} = m_{kh} + A_{kvh} + B_{kwh}, \tag{4.2}$$

where m_{kh} denotes the ANOVA effect shared by all the strata at the kth recurrence, and the rest terms are the ANOVA effects of the jth stratum at the kth recurrence. Let the three components be independently generated from three distributions, and marginally on j, the baseline survival function S_{0kj} follows a DP. The aforementioned procedure implies that $\mathcal{S}_k$ is a survival-function version of the ANOVA DDP.

Since any function in the stick-breaking form is discrete almost surely, we place a convolution through the Dirichlet process mixture (DPM) model (Lo (1984)). Particularly, since the baseline survival functions are defined on the positive half real line, the convolution kernel in DPM should be positive such as log-normal, Gamma, and Weibull. In this chapter, a log-normal kernel is considered. For different recurrences, we treat the relationship among $\mathcal{S}_k$s to be independent.

4.3.3 One-way ANOVA DDP

Considering the data of our interest, where only one drug and one level of dose is used in each treatment stratum, we introduce the modeling of the survival-function version of one-way ANOVA DDP. In this case the prior for the $\mathcal{S}_k$ reduces to a one-way ANOVA form since the dependency among the G treatment strata is explained by only one ANOVA effect. Furthermore, if we set $m_{kh} = 0$, $\alpha_{kh} = (\theta_{k1h}, \cdots, \theta_{kGh})^T$ reduces to a G-variate variable denoting the locations of all G baseline distributions and thus $\theta_{kjh} = \alpha_{kh}^T d_j$, where d_j is the design vector of the jth stratum to select the appropriate ANOVA effects corresponding to j.

With the above notations, we summarize the procedure to construct the survival-function version of one-way ANOVA DDP prior in model (4.1) as follows:

1. Stick-breaking form. For $k = 1, \cdots, K$, let $\mathcal{H}_k$ be the collection of G distribution functions s.t $\mathcal{H}_k = \{H_{kj}, j = 1, \cdots, G\}$. $H_{kj}(\cdot) = \sum_{h=1}^{\infty} p_{kh}\delta_{\theta_{kjh}}(\cdot)$.

2. Convolution step. Let $\alpha_{kh} = (\theta_{k1h}, \cdots, \theta_{kGh})^T$, and d_j be the jth design vector of length G with the jth element being 1 and others being 0. Let $H_{0k} = (H_{0k1}, \cdots, H_{0kG})$ be the collection of base measures, $S_{0kj}(t) = \int S_{\text{LN}}(t|\alpha_k^T d_j, \sigma^2) d\mathcal{H}_k(\alpha, \sigma)$, where S_{LN} denotes the survival function of the log-normal distribution, and $\mathcal{H}_k \sim \text{DP}(M_k, H_{0k})$.

3. Determine the mass parameter and the base measure. For simplicity, we set $M_k = 1$ for all k, which is a commonly used default value of the mass parameter (Gelman et al. (2013)) , $H_{0k}(\theta, \sigma) = N(0, I_G) \times \text{Cauchy}(0, 5)^+$, where Cauchy$^+$ denotes the half_Cauchy distribution.

Step 1 is a standard stick-breaking representation for DP. Step 2 is kernel mixture of DP whereas the kernel is a survival function rather than a cumulative distribution function. The realization of Step 2 is quite straightforward in Stan as it provides the function `lognormal_lccdf` to be used as the kernel of the survival function of the log-normal family.

In Step 3, we specify the base measure as the prior for the location and shape parameters of the log-normal kernel directly rather than adding another hyper prior distribution like De Iorio et al. (2009) did. The main reason is to simplify the computation in Stan. Particularly, inspired by Gelman (2006) and Gelman et al. (2008), we use the half-Cauchy distribution as the non-informative prior for the variance parameter instead of the inverse Gamma prior. In our practice, the choice of half-Cauchy prior significantly improves the speed of convergence and mixture performance of the MCMC chains in our real data analysis and simulation. Another interesting point we met in numerical studies is that the informativeness of the base measure for θ. Here we don't assign the non-informative distribution but a weakly informative one is considered since we find such a weakly informative prior provides better MCMC performance than that of non-informative one with higher effective sample size and better mixture performance. In our other research experience, the weakly informative prior for the variance parameter in the mixing component of the DPM seems to be more preferable.

4.3.4 Other priors and MCMC

In terms of the prior for the parametric prior w_i, we choose log normal prior that $v_i = \log(w_i)$ and $v_i \sim N(0, \tau^2)$, where $\tau > 0$ is an unknown parameter. We further assign a half Cauchy prior for τ s.t $\tau \sim \text{Cauchy}^+(0, 5)$ as a non-informative prior. The prior for the vector of regression coefficients is $\beta \sim N(0, 1000I)$ as a non-informative prior.

We use the truncated Dirichlet process to replace the infinite summand in the DP. The selection of the truncation point is often ad-hoc. Since in Stan the NUTS cannot sampler discrete parameters, we have to fit the truncation number and the mass parameter before the MCMC procedure. In general, the truncation number is set to be large enough s.t the truncated part is negligible. Ohlssen et al. (2007) suggests to use a truncation number L that is greater than $5M + 5$. In our computation, we set $L = 12$.

The MCMC sampling for the posterior distribution is realized in Stan. Stan and its **R** version are widely used in statistical modeling and high-performance statistical computing, especially in Bayesian. Stan realizes the MCMC sampling through the No-U-Turn sampler (NUTS). Stan automates the deriving of the full conditional posterior distribution and NUTS is able to obtain high effective sample size ((Hoffman et al., 2014)).

4.3.5 Stan and NIMBLE: programming styles

The MCMC sampling procedure is implemented in Stan and we also tried to implement the model in NIMBLE, another contemporary Bayesian computing tool in **R**. Stan and NIMBLE are two contemporary Bayesian computing tools that have drawn arising interest for Bayesian analysis but still remain under active development (Kerioui et al. (2020); Ma et al. (2021)). The main advantage of Stan and NIMBLE is that they provide clear automatic posterior sampling procedures based on their specific sampling algorithms without particular justification. Therefore, users can be released from complicated probabilistic deriving and implementation. There has been buzz group discussion about the comparison between Stan and NIMBLE in environments like Stan Development Team (2018) and de Valpine et al. (2021). One comparison on their built-in samplers is demonstrated through implementing weakly informative and informative estimation within the trimmed mean regression model setting (Zhang (2021)). Here we contribute a naive comparison on their programming styles based on the first two authors' experience in coding this project and using Stan and NIMBLE, respectively.

A Bayesian paradigm is made up of three main steps, the prior, likelihood, and the posterior. MCMC generates samples to approximate the posterior distribution. Therefore, what one needs to set in a Bayesian computing tool is the prior and likelihood, let alone Stan or NIMBLE. Nevertheless, Stan and NIMBLE take different programming styles in writing likelihood. In Stan, the default way to present the log likelihood is the syntax `target` and users can add log contribution to it freely, which is similar to the natural language and straightforward to users whatever level of mathematical background. In NIMBLE, the default way is to transfer the likelihood into some standard distributions given by NIMBLE, which may not be friendly for users who have a relatively less mathematical background.

We take fitting the finite mixture of Gaussian model as an example. For a fixed positive integer L, the distribution of Y is given by $F_Y(s) = \sum_{l=1}^{L} p_l N(s|\mu_l, \sigma_l^2)$ and the log-likelihood is log $L(p, \mu, \sigma|Y) = \sum_{i=1}^{n} \sum_{l=1}^{L} \{\log(p_l) + \log \phi(y_i|\mu_1, \sigma_l)\}$, where ϕ denotes the density function of normal distribution. The code for Stan and NIMBLE to implement this model is listed in Listing 1.1 and 1.2, respectively. In Listing 1.1 we clearly find that the contribution to the syntax `target` is just the sum of $\log(p_l)$ and the logarithm of the density of normal distribution denoted by `normal_lpdf`. The rest is to assign a Dirichlet prior for the weights p_l and other parameters. However, in NIMBLE code shown in Listing 1.2, we have to transfer the likelihood into some sampling procedures by IMAGING that there are L clusters of random numbers, the random numbers are i.i.d Gaussian within each cluster, and the probability a random number is drawn from the lth cluster is p_l. Thereafter, the Dirichlet prior is assigned to p_ls. Such imagine matches the Bayesian philosophy but when the likelihood function becomes to be quite complicated, to understand this sampling procedure may not be easy anymore, especially for practitioners not coming from a mathematics or statistics background.

Listing 4.1: Stan code for modeling mixture of Gaussian distribution

```
data{
int<lower=1> N;
vector[N] y;
int<lower=1> L;
}
parameters{
simplex[L] p;
vector[L] mu;
vector<lower=0>[L] sigma;
}
model{
p ~ dirichlet(rep_vector(1, L));
mu ~ normal(0, 100);
sigma ~ cauchy(0, 2.5);
for(i in 1:N){
vector[L] lp_i;
for(l in 1:L){
lp_i[l] = log(p[l]) + normal_lpdf(y[i]|mu[l], sigma[l]);
}
target += log_sum_exp(lp_i);
}
}
```

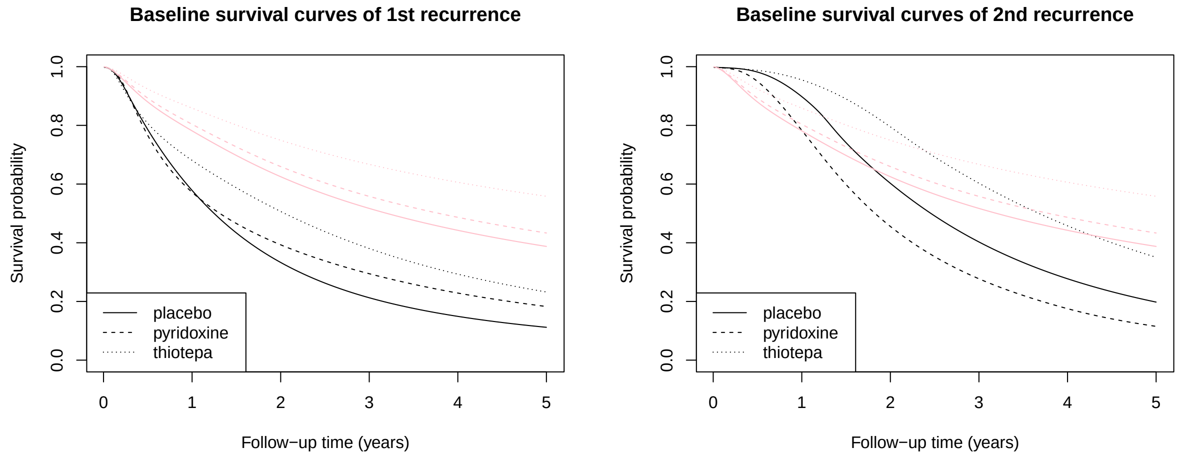
Listing 4.2: NIMBLE code for modeling mixture of Gaussian distribution

```
NimbleCode <- nimbleCode({
  for (i in 1:N) {
    y[i] ~ dnorm(mu_y[z[i]], tau = tau_y[z[i]])
    z[i] ~ dcat(p[1:L])
  }
  for (j in 1:L) {
    mu_y[j] ~ dnorm(0, 0.01)
    tau_y[j] ~ dgamma(0.01, 0.01)
  }
  p[1:L] ~ ddirch(alpha0[1:L])
})
NimbleData <- list(y = y)
NimbleConsts <- list(L = L, N = length(NimbleData$y), alpha0 = rep(1, L))
NimbleInits <- list(mu_y = rnorm(NimbleConsts$L), tau_y = rgamma(
    NimbleConsts$L),p = rep(1/NimbleConsts$L, NimbleConsts$L))
```

4.4 Application: bladder cancer recurrences

We apply the GSFM to analyze the Bladder cancer recurrences data set contained in R package `survival`. Totally 118 subjects in the clinical trial are divided into 3 treatment strata including placebo, pyridoxine (vitamin B6), and thiotepa. Each subject may experience k (from 1 to 9) times of recurrences and may die from or not from the recurrence of bladder cancer. We don't discriminate the death from cancer and the recurrence, and the death from other causes is treated as censoring status. Our interest is the gap time between the $(k-1)$th and the kth recurrences. Besides the treatment schemes, two clinical covariates are considered: the number of tumors at the beginning (x_1) and the size of the largest tumor (x_2) within a subject. The values of these two covariates are evaluated at the beginning of each recurrence interval. This data set was once analyzed for the time between the first to the second recurrence as a univariate time-to-event outcome (Zeng and Lin (2006)). In this chapter, we consider both the first and the second recurrences and thus $K = 2$ here. The two covariates are scaled by divided by 100. To simplify the computation, the follow-up time is transferred from months to years to get lower scalars.

4.4.1 Model checking for baseline survival functions



(a) (b)

Figure 4.3: The estimated baseline survival curves for the first (a) and second (b) recurrence; the black curves are estimated under the proposed generalized shared frailty model, and the pink curves are estimated under the traditional shared frailty model; the real lines, placebo; the dash lines, pyridoxine; the dotted lines, thiotepa.

Before further inference, we need to check whether the proposed model is appropriate. As an alternative, a shared frailty model is fit by R package `spBayeSurv`. In the shared frailty model, the treatment strata are considered as indicator covariates in the parametric term. We run 4 independent MCMC chains for 5000 times with the first 2000 times burn-in and aggregate the rest chains together as the posterior samples under the GSFM. All chains are well mixed and convergent under the GSFM. For the shared frailty model, we run the MCMC 16000 times with the first 6000 times burn-in through R function `survregbayes` using the “IID” Gaussian frailty under “PH” model name. Other settings are default.

The plots of the estimated baseline survival functions under different models stratified by treatment strata can be viewed in figure 4.3. From that, we find the baseline survival functions estimated under the GSFM shows similar trends as that of the K-M estimator in each recurrence, and reflects the crossing survival curves at the first recurrence like the K-M estimator. However, the curves estimated by the shared frailty model are not crossed and cannot change along with recurrences. Therefore, the proposed GSFM is appropriate for the data.

4.4.2 Parametric estimation I: real data

We use the mean of posterior samples (median for τ) as the estimator of parameter and we list the estimation of vector of regression coefficients β and standard deviation parameter τ in Table 4.2.

Table 4.2: The parametric estimation and the MCMC performance for the bladder cancer recurrences data. Est, point estimation; SD, posterior standard deviation; ESS, effective sample size; PACE, the MCMC Pace.

	Est	SD	ESS	PACE
No. tumours	13.849	11.051	1495	0.145
Tumour size	-14.196	12.341	1114	0.194
τ	1.793	0.383	456	0.474

From table 4.2 we find that as the number of tumors at the start point increases, the hazard for recurrences increases as well whereas the larger size of the largest tumor will decrease the hazard. The signs of the effects of the number of tumors and the tumor size are similar to that in Zeng and Lin (2006) who analyzed the first recurrence as a univariate time-to-event data by a transformation model. A slight difference is that under the GSFM, the effect of the number of tumors is not significant (the 95% credible interval covers zero) while the effect reported by Zeng and Lin (2006) is significant. The reason might be that they do not distinguish the two drugs thiotep and pyridoxine but treat them as the same group of treatment. In contrast, in this chapter, we distinguish them and consider their effects as nonparametric components (in baseline survival functions). We conjecture effect of treatment may dominate the performance of therapy, and thus, the clinical effect of the number of tumors becomes less important or significant. In the next subsection, our simulations demonstrate that our MCMC sampling can correctly recognize the significance of the regression parameter.

Besides the parametric estimation result, we also report two metrics about the MCMC performance here. The first one is the effective sample size (ESS), an approximation to the number of "independent" draws in MCMC sampling. It shows that the ESS of all parameters is greater than 400, which is considered to be adequate by Vehtari et al. (2021). The ESS of τ is significantly lower than that of β, a possible reason is that the frailty random effect might be time-dependent $w_i(t)$ rather than a time-fixed effect. Another metric of interest is the average time needed to generate each effective sample, called MCMC Pace. Stan development team emphasized the importance of MCMC Pace, and the definition is given by the team of NIMBLE in de Valpine et al. (2021) as the time-consuming of generating one effective sample. The MCMC Pace to generate τ is much higher than that of β, and we conjecture the possible reason is that the posterior distribution has a long upper tail leading to outliers in posterior samples, which slows down the speed to generate effective samples.

4.4.3 Parametric estimation II: simulation

Another simulation study is considered to evaluate the performance of parametric estimation of the MCMC procedure. Our simulation aims to simulate the occurrences of multiple events on the same individual. We take $K = 2$ and $G = 3$ denoting the number of types of events and the number of treatment strata, respectively. The simulation includes two independent covariates, $x_i \sim \text{Bin}(1, 0.5)$ and $x_2 \sim N(0, 1)$ to incorporate indicator variable and continuous variable as well. For $k = 1, 2, j = 1, 2, 3$, the baseline survival functions S_{0kj} are set as:

- $S_{011} = 1 - 0.5(LN(-0.25, 1) + LN(0.25, 1))$;
- $S_{012} = 1 - 0.5(LN(-0.5, 1) + LN(0.65, 1))$;
- $S_{013} = 1 - 0.5(LN(-0.65, 1) + LN(1.25, 1))$;
- $S_{021} = 1 - LN(0, 1)$;
- $S_{022} = 1 - LN(-0.5, 1)$;
- $S_{023} = 1 - LN(0.5, 1)$

When $k = 1$, the three baseline survival functions are crossed whereas when $k = 2$, the three curves are not. The vector of regression coefficients is $\beta = (1, 1)^T$ and the log frailty random effect $v_i \sim N(0, 1)$ independently. The survival time is generated following model (4.1). The censoring variable of each event is generated from Unif$(4, 6)$ independently, leading to a censoring rate of about 28%. We set the number of subjects to be 90 and they are equally divided into three treatment strata. We repeat the simulation for 150 times.

Table 4.3 summarizes the results for regression parameters β and the standard deviation of frailty effect τ, including the averaged bias (BIAS), root of mean square error (RMSE), posterior estimated standard deviation (ESD) of each point estimate (posterior mean for β and median for τ), the standard deviation (across 150 replicated simulations) of the point estimate (SDE), and the coverage probability (CP) of the 95% credible interval (given by a Wald-type credible interval). The results show that the point estimates of β and τ have quite little bias with low RMSE, ESD values are close to the corresponding SDEs, and the CP values are close to the nominal level 95%.

Table 4.3: Simulation results for the parametric terms. BIAS, averaged bias among the 150 simulations; RMSE, root of mean square error of the estimation; ESD, averaged posterior estimated standard deviation; SDE, the standard deviation of point estimate; CP, the coverage probability of 95% credible interval.

Parameter	BIAS	RMSE	ESD	SDE	CP
$\beta_1 = 1$	-0.062	0.042	0.222	0.196	96.7
$\beta_2 = 1$	-0.025	0.023	0.148	0.152	92.7
$\tau = 1$	-0.078	0.056	0.213	0.224	96.7

4.5 Discussion

In this chapter, we show the power of Bayesian computing illustrated by successfully applying the ANOVA DDP model as the nonparametric prior for a relatively complicated shared frailty model. Our survival-function-based version of the ANOVA DDP, modified based on the ANOVA DDP directly in subsection 3.3, is constructed for the shared frailty model, but can reduce to modeling the univariate dependent survival functions by involving the continuous covariates into the predictor space of the ANOVA DDP. Hence, our work is an extension of De Iorio et al. (2009) to some extent. However, the proposed GSFM is different from the Linear DDP models, the generalization of the accelerated failure time model (Hanson and Jara (2013); Riva-Palacio et al. (2021)). Furthermore, although we point out that there exists potential dual dependence for dual stratification of treatment strata and recurrences, we just simply allow dependence in treatment strata and assume that the recurrences are independent in our methodology demonstration. The dependence across recurrences per subject is dealt with only by the parametric frailty random effect in the proposed shared frailty model. It is more reasonable to incorporate into the baseline survival functions so that the interaction effects between recurrence and treatment may be accounted for. Under the one-level stratification, Hanson et al. (2012) modeled such serial correlation among baseline hazard functions by constructing the so-called dependent tail free process as the prior. It is non-trivial to accommodate dual temporal and stratified dependency as a future research plan.

Chapter 5

Future work: Bayesian tensor factor analysis

Spike-and-slab prior in Bayesian factor analysis

In Part I we propose a Gamma-IBP model for model selection in a high-dimensional regression-like setting. We find that similar IBP-weighted spike-and-slab priors have been applied to Bayesian vector factor analysis (Knowles and Ghahramani (2011); Ročková and George (2016); Ohn and Kim (2022); among others). A common paradigm of Bayesian factor analysis is to assume the potential of infinite latent factors and assign shrinkage priors such as spike-and-slab priors for model selection. This is very different from the frequentist paradigm which needs to pre-specify the number of factors. Hence, motivated by the emergence of multidimensional arrays (tensors) in the current forefront of data science, we may extend our investigations to sparse Bayesian factor analysis or tensor decomposition for tensor data objects. Specifically, for *Bayesian tensor decomposition*, we notice that there have been some Bayesian successes in applications (Xu et al. (2012); Ju et al. (2016); Billio et al. (2023); among others), while explorations to their inferential theories seem to be rare. Although Zhou et al. (2015) studies the posterior contraction rate of their Bayesian tensor factor model, their interest focuses on the probability tensors but not all the real tensors.

Gap between identifiability issues and computational costs

It is well-known that factor analysis encounters identifiability issues. Most of the aforementioned approaches employed the MCMC sampling to combat the model unidentifiability. This is consistent with our findings in Part II that lack of identifiability does not hamper Bayesian prediction if people use MCMC sampling to draw the posterior. Nonetheless, the computation of MCMC sampling is prohibitively expensive when dealing with large-scale and ultrahigh-dimensional data. Hence, considering the tremendous dimensionality of higher-order tensor data, variational Bayes (VB) techniques seem preferable to MCMC in Bayesian factor analysis. Unfortunately, the identifiability issue seems to be a real issue for VB. Take the VB-EM algorithm by Ročková and George (2016) for the Bayesian vector factor model for example. Even though they mitigate identifiability issues by the soft constraint of IBP weights, the posterior distribution of the loading matrix encounters a "magnitude inflation" problem when the dimensionality is much greater than the data size, incurring inconsistent posterior (Ma and Liu, 2022). A similar problem also occurs in variational Bayes matrix factorization (Nakajima and Sugiyama, 2011).

Future work

The above discussions highlight the challenges to address, including tackling the identifiability issue under some tensor decompositions with nice prior elicitation, developing the appropriate VB algorithms, and establishing the posterior inferential theories. Note that in general, the identifiability conditions may be bothersome to mean-filed VB approaches. For example, in matrix/tensor PCA approaches (Hoff (2016); Jiang et al. (2020)), the identifiability requires the loading matrices to be orthornormal, indicating that the posterior on each columns are not independent anymore. Hence, VB methods that relax the mean-filed assumption are needed (Kingma et al. (2016); Saha et al. (2020); among others). Meanwhile, the techniques for the theories about the posterior contraction under the non-mean-filed VB approximation are also expected. We leave them as the future work.

References

Absil, P.-A. and Malick, J. (2012). Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158.

Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., and Raza, M. A. (2017). Survival analysis of heart failure patients: A case study. *PloS one*, 12(7):e0181001.

Anastasiou, A., Chen, Y., Cho, H., and Fryzlewicz, P. (2022). *breakfast: Methods for Fast Multiple Change-Point Detection and Estimation*. R package version 2.3.

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

Bai, R., Rockova, V., and George, E. I. (2020). Spike-and-slab meets LASSO: A review of the spike-and-slab LASSO. *arXiv preprint arXiv:2010.06451*.

Balan, T. A. and Putter, H. (2020). A tutorial on frailty models. *Statistical methods in medical research*, 29(11):3424–3454.

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

Baranowski, R., Chen, Y., and Fryzlewicz, P. (2022). *not: Narrowest-Over-Threshold Change-Point Detection*. R package version 1.3.

Baranowski, R. and Fryzlewicz, P. (2019). *wbs: Wild Binary Segmentation for Multiple Change-Point Detection*. R package version 1.4.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, pages 870–897.

Bardwell, L. and Fearnhead, P. (2017). Bayesian detection of abnormal segments in multiple time series. *Bayesian Analysis*, 12(1):193–218.

Barry, D. and Hartigan, J. A. (1993). A bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319.

Berchuck, S. I., Janko, M., Medeiros, F. A., Pan, W., and Mukherjee, S. (2022). Bayesian nonparametric factor analysis for longitudinal spatial surfaces. *Bayesian Analysis*, 17(2):435–464.

Billio, M., Casarin, R., Iacopini, M., and Kaufmann, S. (2023). Bayesian dynamic tensor regression. *Journal of Business & Economic Statistics*, 41(2):429–439.

Birte, E. and Claudia, K. (2018). A mosum procedure for the estimation of multiple random change points. *Bernoulli*, 24(1):526–564.

Boysen, L., Kempe, A., Liebscher, V., Munk, A., and Wittich, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *The Annals of Statistics*, 37(1):157–183.

Branscum, A. J., Johnson, W. O., Hanson, T. E., and Gardner, I. A. (2008). Bayesian semiparametric ROC curve estimation and disease diagnosis. *Statistics in Medicine*, 27(13):2474–2496.

Burgette, L. F., Puelz, D., and Hahn, P. R. (2021). A symmetric prior for multinomial probit models. *Bayesian Analysis*, 16(3):1–18.

Cai, B. and Dunson, D. B. (2007). Bayesian multivariate isotonic regression splines: applications to carcinogenicity studies. *Journal of the American Statistical Association*, 102(480):1158–1171.

Cappello, L., Madrid Padilla, O. H., and Palacios, J. A. (2023). Bayesian change point detection with spike and slab priors. *Journal of Computational and Graphical Statistics*, (just-accepted):1–24.

Carlstein, E. et al. (1988). Nonparametric change-point estimation. *The Annals of Statistics*, 16(1):188–197.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2017). Stan: a probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32.

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018.

Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069–2101.

Chen, K., Jin, Z., and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, 89(3):659–668.

Chen, S. (2002). Rank estimation of transformation models. *Econometrica*, 70(4):1683–1697.

Cheng, S., Wei, L., and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, 82(4):835–845.

Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics*, 35(3):999–1018.

Chernozhukov, V., Hansen, C., and Liao, Y. (2017). A LAVA attack on the recovery of sums of dense and sparse signals. *The Annals of Statistics*, 45(1):39–76.

Chiappori, P.-A., Komunjer, I., and Kristensen, D. (2015). Nonparametric identification and estimation of transformation models. *Journal of Econometrics*, 188(1):22–39.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2):221–241.

Chicco, D. and Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, 20(1):1–16.

Cho, H. and Fryzlewicz, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):475–507.

Colling, B. and Van Keilegom, I. (2019). Estimation of fully nonparametric transformation models. *Bernoulli*, 25(4B):3762–3795.

Conlon, A., Taylor, J., and Sargent, D. J. (2014). Multi-state models for colon cancer recurrence and death with a cured fraction. *Statistics in medicine*, 33(10):1750–1766.

Cuzick, J. (1988). Rank regression. *The Annals of Statistics*, pages 1369–1389.

de Castro, M., Chen, M.-H., Ibrahim, J. G., and Klein, J. P. (2014). Bayesian transformation models for multivariate survival data. *Scandinavian Journal of Statistics*, 41(1):187–199.

de Castro, M., Chen, M.-H., and Zhang, Y. (2015). Bayesian path specific frailty models for multi-state survival data with applications. *Biometrics*, 71(3):760–771.

De Iorio, M., Johnson, W. O., Müller, P., and Rosner, G. L. (2009). Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics*, 65(3):762–771.

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). An anova model for dependent random measures. *Journal of the American Statistical Association*, 99(465):205–215.

de Valpine, P., Paciorek, C., Turek, D., Michaud, N., Anderson-Bergman, C., Obermeyer, F., Wehrhahn Cortes, C., Rodrìguez, A., Temple Lang, D., and Paganin, S. (2021). *NIMBLE User Manual*. R package manual version 0.11.1.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26(2):403–413.

Depaoli, S., Winter, S. D., and Visser, M. (2020). The importance of prior sensitivity analysis in bayesian statistics: demonstrations using an interactive shiny app. *Frontiers in Psychology*, 11.

DeYoreo, M. and Kottas, A. (2018). Modeling for dynamic ordinal regression relationships: An application to estimating maturity of rockfish in california. *Journal of the American Statistical Association*, 113(521):68–80.

Ding, Y. and Nan, B. (2011). A sieve m-theorem for bundled parameters in semiparametric models, with application to the efficient estimation in a linear model for censored data. *The Annals of Statistics*, 39(6):3032–3061.

Donoho, D. L., Johnstone, I. M., Hoch, J. C., and Stern, A. S. (1992). Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):41–67.

Du, C., Kao, C.-L. M., and Kou, S. (2016). Stepwise signal extraction via marginal likelihood. *Journal of the American Statistical Association*, 111(513):314–330.

Duchateau, L. and Janssen, P. (2007). *The frailty model*. Springer Science & Business Media.

Egleston, B. L., Uzzo, R. G., and Wong, Y.-N. (2017). Latent class survival models linked by principal stratification to investigate heterogenous survival subgroups among individuals with early-stage kidney cancer. *Journal of the American Statistical Association*, 112(518):534–546.

Fang, X., Li, J., and Siegmund, D. (2020). Segmentation and estimation of change-point models: false positive control and confidence regions. *The Annals of Statistics*, 48(3):1615–1647.

Fearnhead, P. (2006). Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16(2):203–213.

Fearnhead, P. and Rigaill, G. (2019). Changepoint detection in the presence of outliers. *Journal of the American Statistical Association*, 114(525):169–183.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2(4):615–629.

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

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

Fryzlewicz, P. (2018). Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *The Annals of Statistics*, 46(6B):3390–3421.

Fryzlewicz, P. (2021). *nsp: Inference for Multiple Change-Points in Linear Models*. R package version 1.0.0.

Fryzlewicz, P. (2023). Narrowest significance pursuit: inference for multiple change-points in linear models. *Journal of the American Statistical Association*, pages 1–25.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383.

Gørgens, T. and Horowitz, J. L. (1999). Semiparametric estimation of a censored regression model with an unknown transformation of the dependent variable. *Journal of Econometrics*, 90(2):155–191.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545.

Hahn, P. R. and Carvalho, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448.

Hanson, T. E. and Jara, A. (2013). Surviving fully bayesian nonparametric regression models. *Bayesian Theory and Applications*, pages 593–615.

Hanson, T. E., Jara, A., Zhao, L., et al. (2012). A bayesian semiparametric temporally-stratified proportional hazards model with spatial frailties. *Bayesian Analysis*, 7(1):147–188.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American statistical Association*, 84(408):986–995.

Härdle, W. K., Müller, M., Sperlich, S., and Werwatz, A. (2004). *Nonparametric and semiparametric models*. Springer Science & Business Media.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247(18):2543–2546.

Haynes, K., Fearnhead, P., and Eckley, I. A. (2017). A computationally efficient nonparametric approach for changepoint detection. *Statistics and Computing*, 27(5):1293–1305.

Hjort, N. L. (1990). Nonparametric bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, 18:1259–1294.

Hoff, P. D. (2016). Equivariant and scale-free tucker decomposition models. *Bayesian Analysis*, 11(3):627–648.

Hoffman, M. D., Gelman, A., et al. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Horowitz, J. L. (1996). Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica*, 64(1):103–137.

Hušková, M. and Slabỳ, A. (2001). Permutation tests for multiple changes. *Kybernetika*, 37(5):605–622.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer Science & Business Media.

Ishwaran, H. and James, L. F. (2002). Approximate dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics*, 11(3):508–532.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860.

James, L. F. (2017). Bayesian Poisson calculus for latent feature modeling via generalized Indian buffet process priors. *The Annals of Statistics*, 45(5):2016–2045.

James, N. A., Matteson, D. S., et al. (2015). ecp: An R Package for nonparametric multiple change point analysis of multivariate data. *Journal of Statistical Software*, 62(i07).

Jeong, S. and Ghosal, S. (2021). Unified bayesian theory of sparse linear regression with nuisance parameters. *Electronic Journal of Statistics*, 15(1):3040–3111.

Jiang, B., Petkova, E., Tarpey, T., and Ogden, R. T. (2020). A bayesian approach to joint modeling of matrix-valued imaging data and treatment outcome with applications to depression studies. *Biometrics*, 76(1):87–97.

Jin, Z., Lin, D., Wei, L., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika*, 90(2):341–353.

Ju, F., Sun, Y., Gao, J., Hu, Y., and Yin, B. (2016). Nonparametric tensor dictionary learning with beta process priors. *Neurocomputing*, 218:120–130.

Jula Vanegas, L., Behr, M., and Munk, A. (2021). Multiscale quantile segmentation. *Journal of the American Statistical Association*, pages 1–14.

Kerioui, M., Mercier, F., Bertrand, J., Tardivon, C., Bruno, R., Guedj, J., and Desmée, S. (2020). Bayesian inference using hamiltonian monte-carlo algorithm for nonlinear joint modeling in the context of cancer immunotherapy. *Statistics in Medicine*, 39(30):4853–4868.

Killick, R. and Eckley, I. (2014). changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3):1–19.

Killick, R., Eckley, I. A., Ewans, K., and Jonathan, P. (2010). Detection of changes in variance of oceanographic time-series using changepoint analysis. *Ocean Engineering*, 37(13):1120–1126.

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

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29.

Kingman, J. (1967). Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78.

Klein, L., O'Connor, C. M., Leimberger, J. D., Gattis-Stough, W., Piña, I. L., Felker, G. M., Adams Jr, K. F., Califf, R. M., and Gheorghiade, M. (2005). Lower serum sodium is associated with increased short-term mortality in hospitalized patients with worsening heart failure: results from the outcomes of a prospective trial of intravenous milrinone for exacerbations of chronic heart failure (optime-chf) study. *Circulation*, 111(19):2454–2460.

Knowles, D. and Ghahramani, Z. (2011). Nonparametric Bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics*, 5(2B):1534–1552.

Ko, S. I., Chong, T. T., Ghosh, P., et al. (2015). Dirichlet process hidden markov multiple change-point model. *Bayesian Analysis*, 10(2):275–296.

Korkas, K. and Fryzlewicz, P. (2020). *wbsts: Multiple Change-Point Detection for Nonstationary Time Series*. R package version 2.1.

Korkas, K. K. and Pryzlewiczv, P. (2017). Multiple change-point detection for non-stationary time series using wild binary segmentation. *Statistica Sinica*, pages 287–311.

Kottas, A. (2006). Nonparametric Bayesian survival analysis using mixtures of Weibull distributions. *Journal of Statistical Planning and Inference*, 136(3):578–596.

Lee, S., Seo, M. H., and Shin, Y. (2016). The lasso for high dimensional regression with a possible change point. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 193–210.

Lenk, P. J. and Choi, T. (2017). Bayesian analysis of shape-restricted functions using Gaussian process priors. *Statistica Sinica*, 27(1):43–69.

Li, H., Munk, A., and Sieling, H. (2016). FDR-control in multiscale change-point segmentation. *Electronic Journal of Statistics*, 10(1):918–959.

Lin, L. and Dunson, D. B. (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika*, 101(2):303–317.

Lin, Y., Luo, Y., Xie, S., and Chen, K. (2017). Robust rank estimation for transformation models with random effects. *Biometrika*, 104(4):971–986.

Linton, O., Sperlich, S., Van Keilegom, I., et al. (2008). Estimation of a semiparametric transformation model. *The Annals of Statistics*, 36(2):686–718.

Liu, C., Martin, R., and Shen, W. (2017). Empirical priors and posterior concentration in a piecewise polynomial sequence model. *arXiv preprint arXiv:1712.03848*.

Lo, A. Y. (1984). On a class of bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*, pages 351–357.

Lu, W. and Ying, Z. (2004). On semiparametric transformation cure models. *Biometrika*, 91(2):331–343.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337.

Ma, Y. and Liu, J. S. (2022). On posterior consistency of Bayesian factor models in high dimensions. *Bayesian Analysis*, 17(3):901–929.

Ma, Z., Hu, G., and Chen, M.-H. (2021). Bayesian hierarchical spatial regression models for spatial data in the presence of missing covariates with applications. *Applied Stochastic Models in Business and Industry*, 37(2):342–359.

MacEachern, S. N. (2016). Nonparametric Bayesian methods: a gentle introduction and overview. *Communications for Statistical Applications and Methods*, 23(6):445–466.

Mallick, B. K. and Walker, S. (2003). A Bayesian semiparametric transformation model incorporating frailties. *Journal of Statistical Planning and Inference*, 112(1-2):159–174.

Martin, A. D., Quinn, K. M., and Park, J. H. (2011). MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software*, 42(9):22.

Martin, R., Mess, R., and Walker, S. G. (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli*, 23(3):1822–1847.

Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345.

McCulloch, R. and Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64(1-2):207–240.

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC.

Meier, A., Kirch, C., and Cho, H. (2021). mosum: A package for moving sums in change-point analysis. *Journal of Statistical Software*, 97:1–42.

Monteiro, J. V., Assunçao, R. M., and Loschi, R. H. (2011). Product partition models with correlated parameters. *Bayesian Analysis*, 6(4):691–726.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Nakajima, S. and Sugiyama, M. (2011). Theoretical analysis of bayesian matrix factorization. *The Journal of Machine Learning Research*, 12:2583–2648.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors1. *The Annals of Statistics*, 42(2):789–817.

Natarajan, A., De Iorio, M., Heinecke, A., Mayer, E., and Glenn, S. (2023). Cohesion and repulsion in bayesian distance clustering. *Journal of the American Statistical Association*, (just-accepted):1–18.

Neelon, B. and Dunson, D. B. (2004). Bayesian isotonic regression and trend analysis. *Biometrics*, 60(2):398–406.

Nieto-Barajas, L. E., Müller, P., Ji, Y., Lu, Y., and Mills, G. B. (2012). A time-series ddp for functional proteomics profiles. *Biometrics*, 68(3):859–868.

Ohlssen, D. I., Sharples, L. D., and Spiegelhalter, D. J. (2007). Flexible random-effects models using bayesian semi-parametric models: applications to institutional comparisons. *Statistics in Medicine*, 26(9):2088–2112.

Ohn, I. and Kim, Y. (2022). Posterior consistency of factor dimensionality in high-dimensional sparse factor models. *Bayesian Analysis*, 17(2):491–514.

Paisley, J. and Jordan, M. I. (2016). A constructive definition of the beta process. *arXiv preprint arXiv:1604.00685*.

Paisley, J. W., Zaas, A. K., Woods, C. W., Ginsburg, G. S., and Carin, L. (2010). A stick-breaking construction of the beta process. In *ICML*.

Park, C. G., Vannucci, M., and Hart, J. D. (2005). Bayesian methods for wavelet series in single-index models. *Journal of Computational and Graphical Statistics*, 14(4):770–794.

Pati, D., Bhattacharya, A., Pillai, N. S., and Dunson, D. (2014). Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *The Annals of Statistics*, pages 1102–1130.

Paulon, G., De Iorio, M., Guglielmi, A., and Ieva, F. (2020). Joint modeling of recurrent events and survival: a bayesian non-parametric approach. *Biostatistics*, 21(1):1–14.

Pein, F., Hotz, T., Sieling, H., and Aspelmeier, T. (2022). *stepR: Multiscale change-point inference*. R package version 2.1-3.

Pein, F., Sieling, H., and Munk, A. (2017). Heterogeneous change point inference. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(4):1207–1227.

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., and Schmid, M. (2019). A review of spline function procedures in R. *BMC Medical Research Methodology*, 19(1):1–16.

Phadia, E. G. (2015). *Prior processes and their applications*. Springer.

Pronzato, L. and Pázman, A. (2013). *Design of experiments in nonlinear models*. Springer.

Pukelsheim, F. (1994). The three sigma rule. *The American Statistician*, 48(2):88–91.

Quinlan, J. J., Page, G. L., and Castro, L. M. (2022). Joint random partition models for multivariate change point analysis. *Bayesian Analysis*, 1(1):1–28.

Quintana, F. A., Mueller, P., Jara, A., and MacEachern, S. N. (2020). The dependent Dirichlet process and related models. *arXiv preprint arXiv:2007.06129*.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, 3(4):425–441.

Ray, K. and Szabó, B. (2022). Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, 117(539):1270–1281.

Riihimäki, J. and Vehtari, A. (2010). Gaussian processes with monotonicity information. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 645–652. JMLR Workshop and Conference Proceedings.

Riva-Palacio, A., Leisen, F., and Griffin, J. (2021). Survival regression models with dependent bayesian nonparametric priors. *Journal of the American Statistical Association*, pages 1–10.

Ročková, V. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *The Annals of Statistics*, 46(1):401–437.

Ročková, V. and George, E. I. (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111(516):1608–1622.

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American statistical Association*, 103(483):1131–1154.

Romano, G., Rigaill, G., Runge, V., and Fearnhead, P. (2022). Detecting abrupt changes in the presence of local fluctuations and autocorrelated noise. *Journal of the American Statistical Association*, 117(540):2147–2162.

Saha, A., Bharath, K., and Kurtek, S. (2020). A geometric variational approach to bayesian inference. *Journal of the American Statistical Association*, 115(530):822–835.

Scheike, T. H. (2006). A flexible semiparametric transformation model for survival data. *Lifetime Data Analysis*, 12(4):461–480.

Scheipl, F., Fahrmeir, L., and Kneib, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, 107(500):1518–1532.

Sen, D., Patra, S., and Dunson, D. (2022). Constrained inference through posterior projections. *arXiv*.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650.

Shi, Y., Martens, M., Banerjee, A., Laud, P., et al. (2019). Low information omnibus (LIO) priors for Dirichlet process mixture models. *Bayesian Analysis*, 14(3):677–702.

Shin, M. and Liu, J. S. (2021). Neuronized priors for Bayesian sparse linear regression. *Journal of the American Statistical Association*, pages 1–16.

Shively, T. S., Sager, T. W., and Walker, S. G. (2009). A Bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):159–175.

Solomon, P. J. (1984). Effect of misspecification of regression models in the analysis of survival data. *Biometrika*, 71(2):291–298.

Song, X., Ma, S., Huang, J., and Zhou, X.-H. (2007). A semiparametric approach for the non-parametric transformation survival model with multiple covariates. *Biostatistics*, 8(2):197–211.

Stan Development Team (2018). The Stan Core Library. Version 2.27.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2.

Stransky, N., Vallot, C., Reyal, F., Bernard-Pierrot, I., De Medina, S. G. D., Segraves, R., De Rycke, Y., Elvin, P., Cassidy, A., Spraggon, C., et al. (2006). Regional copy number–independent deregulation of transcription in cancer. *Nature genetics*, 38(12):1386–1396.

Teh, Y. W., Grür, D., and Ghahramani, Z. (2007). Stick-breaking construction for the Indian buffet process. In *Artificial intelligence and statistics*, pages 556–563. PMLR.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Therneau, T. M. (2022). *A Package for Survival Analysis in R*. R package version 3.4-0.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., et al. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1):1–26.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC. *Bayesian Analysis*, 1(1):1–28.

Vostrikova, L. Y. (1981). Detecting “disorder” in multidimensional random processes. 259(2):270–274.

Wang, L. and Dunson, D. B. (2011a). Bayesian isotonic density regression. *Biometrika*, 98(3):537–551.

Wang, L. and Dunson, D. B. (2011b). Semiparametric Bayes’ proportional odds models for current status data with underreporting. *Biometrics*, 67(3):1111–1118.

Wang, W., He, X., and Zhu, Z. (2020). Statistical inference for multiple change-point models. *Scandinavian Journal of Statistics*, 47(4):1149–1170.

Wang, X. and Berger, J. O. (2016). Estimating shape constrained functions using Gaussian processes. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1–25.

Williamson, S., Wang, C., Heller, K. A., and Blei, D. M. (2010). The IBP compound Dirichlet process and its application to focused topic modeling. In *ICML*.

Wyse, J., Friel, N., Rue, H., et al. (2011). Approximate simulation-free bayesian inference for multiple changepoint models with dependence within segments. *Bayesian Analysis*, 6(4):501–528.

Xu, Z., Yan, F., and Qi, Y. (2012). Infinite tucker decomposition: nonparametric bayesian models for multiway data analysis. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1675–1682.

Yang, Y., Wainwright, M. J., and Jordan, M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics*, pages 2497–2532.

Yao, Y.-C. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *The Annals of Statistics*, pages 1434–1447.

Ye, J. and Duan, N. (1997). Nonparametric $n^{-1/2}$-consistent estimation for the general transformation models. *The Annals of Statistics*, 25(6):2682–2717.

Yen, T.-J. (2011). A majorization–minimization approach to variable selection using spike and slab priors. *The Annals of Statistics*, 39(3):1748–1775.

Zeileis, A., Leisch, F., Hornik, K., and Kleiber, C. (2002). strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7:1–38.

Zeng, D. and Lin, D. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika*, 93(3):627–640.

Zeng, D. and Lin, D. (2007a). Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association*, 102(480):1387–1396.

Zeng, D. and Lin, D. (2007b). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):507–564.

Zhang, L. (2021). A Bayesian Comparison in Stan and NIMBLE by Trimmed mean regression. [M.Phil's dissertation], Hong Kong, China: The Hong Kong Polytechnic University.

Zhao, L., Hanson, T. E., and Carlin, B. P. (2009). Mixtures of polya trees for flexible spatial frailty survival modelling. *Biometrika*, 96(2):263–276.

Zhao, Z., Jiang, F., and Shao, X. (2022). Segmenting time series via self-normalisation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1699–1725.

Zhou, H., Cheng, X., Wang, S., Zou, Y., and Wang, H. (2022a). *SurvMetrics: Predictive Evaluation Metrics in Survival Analysis*. R package version 0.5.0.

Zhou, H. and Hanson, T. (2018). A unified framework for fitting Bayesian semiparametric models to arbitrarily censored survival data, including spatially referenced data. *Journal of the American Statistical Association*, 113(522):571–581.

Zhou, H., Hanson, T., and Zhang, J. (2020). spbayessurv: Fitting bayesian spatial survival models using r. *Journal of Statistical Software*, 92:1–33.

Zhou, J., Bhattacharya, A., Herring, A. H., and Dunson, D. B. (2015). Bayesian factorizations of big sparse tensors. *Journal of the American Statistical Association*, 110(512):1562–1576.

Zhou, J., Zhang, J., and Lu, W. (2022b). TransModel: An R package for linear transformation model with censored data. *Journal of Statistical Software*, 101:1–12.