Received: 13 December 2023 | Revised: 16 August 2024 | Accepted: 29 September 2024

DOI: 10.1111/sjos.12758

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

# Data-driven estimation for multithreshold accelerated failure time model

**Chuang Wan**[1] | **Hao Zeng**[2] | **Wenyang Zhang**[3] | **Wei Zhong**[4] | **Changliang Zou**[5]

[1]Department of Statistics, School of Economics, Jinan University, Guangzhou, China

[2]Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, China

[3]Faculty of Business Adminstration, University of Macau, Macau, China

[4]MOE Key Lab of Econometrics, WISE, Department of Statistics & Data Science, School of Economics, Xiamen University, Xiamen, China

[5]NITFID, School of Statistics and Data Science, LPMC and KLMDASR and LEBPS, Nankai University, Tianjin, China

**Correspondence**
Hao Zeng, Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen 518055, China.
Email: zenghao.acmail@gmail.com

**Funding information**
Universidade de Macau, Grant/Award Numbers: CPG2023-00041-FBA, CPG2024-00004-FBA, SRG2023-00027-FBA; National Statistical Science Research Grants of China, Grant/Award Number: 2022LD08; National Natural Science Foundation of China, Grant/Award Numbers: 11925106, 11931001, 11931014, 12226007, 12231011, 12301344, 12326325, 12471265, 71988101; National Key Research and Development Program of China, Grant/Award Numbers: 2022YFA1003703, 2022YFA1003800

## Abstract

This article develops a novel estimation framework for the multithreshold accelerated failure time model, which has distinct linear forms within different subdomains. One major challenge is to determine the number of threshold effects. We first show the selection consistency of a modified Bayesian information criterion under mild conditions. It is useful sometimes but heavily depends on the penalization magnitude, which usually varies from the model configuration and data distribution. To address this issue, we leverage a cross-validation criterion alongside an order-preserved sample-splitting scheme to yield a consistent estimation. The new criterion is completely data driven without additional parameters and thus robust to model setting and data distributions. The asymptotic properties for the parameter estimates are also carefully established. Additionally, we propose an efficient score-type test to examine the existence of threshold effects. The new statistic is free of estimating any potential threshold effects and is thus suitable for multithreshold scenarios. Numerical experiments validate the reliable finite-sample performance of our methodologies, which corroborates the theoretical results.

**KEYWORDS**

cross-validation, information criterion, multithreshold accelerated failure time model, sample splitting, score test

All authors equally contributed to the study and are listed in the alphabetical order.

# 1 | INTRODUCTION

## 1.1 | Literature review

Threshold regression has emerged as a widely applied and powerful statistical tool for examining the dynamic relationship between the response and the predictors. It enables us to capture many nonlinear phenomena, providing interpretable and meaningful results, thus leading to its great popularity. For example, in biomedical research and precision medicine, dividing the observed sample into distinct subgroups according to some quantitative covariates is capable of identifying target subgroups from the population and facilitating personalized treatment recommendations (He et al., 2018; He et al., 2023; Wei et al., 2022; Yan et al., 2021). The earlier developed methodologies for threshold models primarily focused on single threshold setting, for example, Hansen (2000), Kosorok and Song (2007), Lee and Seo (2008), Lee et al. (2020, 2011), Li and Zhang (2011), Xu et al. (2014). However, in many practical applications, the single threshold setting is undoubtedly insufficient to capture the necessary nonlinearity and also mitigates model misspecification risks. The problem of multiple thresholds is arguably more challenging since the number of thresholds is usually unknown to us.

The existing literature on determination for the number of thresholds can be roughly classified into three categories based on different perspectives. First, determining the number of thresholds or change-points, say $K$, is essentially a model selection problem, since $K$ derives model dimension. The information criterion- (IC) based approaches are routinely employed by balancing a term quantifying model fit with a term accounting for model complexity. Many works have been done to study and improve the performance of IC in multiple change-point context. For instance, Gonzalo and Pitarakis (2002) developed a sequential model selection procedure using Bayesian Information Criterion (BIC) to assess the model complexity. In a Gaussian setting, Ninomiya (2015) proposed an adaptation of the Akaike information criterion (AIC). In the context of multiple structural break models[1], Hall et al. (2015) examined the performance of various information criteria and observed that the BIC and the Hannan–Quinn Information Criterion (HQIC) outperformed the AIC, while the AIC is generally inconsistent. Several ad hoc criteria for the change-point problem were also proposed, for instance, the improved BIC (Hannart & Naveau, 2012) and the minimum description length criterion (Aue et al., 2014). While the IC is well grounded for regular statistical models, different penalization magnitudes are required to accommodate different contexts in the change-point problem. We stress that the penalization magnitude is crucial as it directly relates model complexity, and more importantly, choosing the optimal IC is not realistic in practice since the optimal penalization magnitude typically varies from model setting and error distribution (Zhang & Siegmund, 2007; Zou et al., 2020).

The second category is based on the sequential testing strategy pioneered by Bai and Perron (1998). The main concept involves starting with testing $\mathbb{H}_0 : K = 0$ against $\mathbb{H}_1 : K > 0$. If $\mathbb{H}_0$ is rejected, then test $\mathbb{H}_0 : K = 1$ versus $\mathbb{H}_1 : K > 1$. This process continues by incrementing the number of change-points until the null hypothesis cannot be rejected. Along this line, Lee and Lam (2020) proposed a likelihood ratio test for Cox regression to decide whether and where to split the sample, with the segmentation positions serving as the change-point estimators. Contributions in this domain also include Hall et al. (2012), He et al. (2013), Dette and Gösmann (2020), Gierz and Park (2022), among others. Despite the ease of implementation of the sequential testing procedure, there are still some limitations. First, the limiting behavior of the test statistic is generally not a tractable distribution, which may involve some nuisance parameters such as thresholds

(Wang & Zou, 2023). Second, the accumulation of type-I errors at each step of sequential testing poses challenges in controlling the false discovery rate. Moreover, constructing the sequential testing statistics requires splitting the sample into multiple subperiods, which may result in a loss of statistical power as the candidate number of change-points increases (Qu, 2008).

As for the third category, researchers found that the change-point or threshold detection problem can be reformulated into a high-dimensional regression with sparsity in the synthetic parameter vector. Thus, modern feature selection techniques (Fan & Li, 2001; Tibshirani, 1996) can be employed to estimate the number of change-points and their locations, simultaneously. A pivotal contribution in this area can be traced back to Harchaoui and Lévy-Leduc (2010), who used the LASSO algorithm to estimate the locations of multiple change-points in one-dimensional piece-wise constant signals. Subsequently, Li and Jin (2018) proposed a two-stage procedure based on group fused penalty functions to identify multiple thresholds in the segmented accelerated failure time model. The following work of Li and Jin (2018) made extensions and adaptation to other aspects, such as the model-based subgroup analysis (Wang et al., 2019), the multiple change-plane model (Li et al., 2021), the multithreshold proportional hazards model (Wang et al., 2022) and the multithreshold structural equation model (Wang & Li, 2023). However, the performance of shrinkage-based threshold detection algorithms is directly linked to the tuning parameters in the penalty term. Although these methods employed the BIC for selecting an appropriate tuning parameter, their theoretical justification remains an open problem.

## 1.2 | Model setting

The main objective of this paper is to estimate the multithreshold accelerated failure time model proposed by Li and Jin (2018) in a data-driven fashion. Let $T_i$ be an uncensored response of interest, such as the survival time or the logarithm of failure time, and $\mathbf{X}_i$ be an observable $p$-dimensional regressors. Due to some reasons, such as patient withdrawal from the study or deaths, not all $T_i$'s are available. Hence we only observe

$$Y_i = \min(T_i, C_i) \quad \text{and} \quad \delta_i = 1(T_i \le C_i),$$

where $C_i$ denotes the independent and identically distributed (i.i.d.) censoring variable, and $\delta_i$ is the censoring indicator. We consider the following multiple threshold regression

$$T_i = \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k^\star + \varepsilon_i, \ \tau_k^\star < Z_i \le \tau_{k+1}^\star, \quad k = 0,\ldots,K^\star, \quad i = 1,\ldots,n, \tag{1}$$

where $K^\star$ is the unknown number of change-points/thresholds, $\boldsymbol{\beta}_0^\star,\ldots,\boldsymbol{\beta}_{K^\star}^\star$ are $p \times 1$ unknown parameter vectors for the $K^\star + 1$ subgroups separated by the change-points $\mathcal{T}_{K^\star}^\star = (\tau_1^\star,\ldots,\tau_{K^\star}^\star)$, $Z_i$ is the univariate threshold variable and $\varepsilon_i$ is a scalar random error. Without loss of generality, we assume that the threshold observations $\{Z_i, i = 1,\ldots,n\}$ have been arranged in ascending order, that is, $Z_1 < Z_2 < \cdots < Z_n$.

## 1.3 | Main contributions

In this paper, we first propose a modified information criterion for model (1) and demonstrate its consistency in selecting the number of thresholds under mild conditions. It is also interesting to

see that the conventional BIC is not consistent. While the information criterion is commonly used in practice, there is a lack of universal guidance on selecting penalty factors, leaving a gap between theory and practice. To bridge this gap, we further develop a new order-preserved cross-validation (OPCV) criterion based on sample splitting. Specifically, we begin by dividing all observations into two sets: one with even-indices and one with odd indices, based on the order of the sorted threshold variable. We then alternatively use one set as the training set and the other as the testing set. We fit candidate threshold models on the training set and assess prediction error on the testing set. This process is repeated with the roles of training and testing sets reversed. The optimal number of thresholds is determined by minimizing the total prediction error, resulting in a special twofold cross-validation criterion. One desirable feature of OPCV criterion is its completely data-driven principles without any tuning parameters. Thus, it can be robust to model settings and data distributions. We provide theoretical evidence to support the consistency of the estimated number of thresholds.

Before estimating $K$, it is essential to conduct a testing procedure to determine if there are statistically significant threshold effects. This test problem differs from the hypothesis test for a given number of change-points like in Bai and Perron (1998). Here we mainly focus on the presence of threshold effects, that is, $K = 0$ against $K > 1$. Therefore, we do not need to specify the exact values of change-points. In such a scenario, the Wald and the likelihood ratio tests are troublesome, as they rely on the alternative models that require estimating under the presence of threshold effects. Additionally, their limiting distributions are also hard to derive. Inspired by works of Qu (2008) and Zhong et al. (2022), we introduce a score-type testing procedure to examine the existence of threshold effect. The proposed test is akin to a weighted CUSUM-type statistic, which only requires estimating under the null model, that is, a completely linear model, making it suitable for multithreshold scenarios. We demonstrate that the asymptotic null distribution of the test statistic is not pivotal and propose a convenient simulation procedure to compute the critical values. Our method is implemented in the R package `MTAFT`, available at the author's website https://github.com/zenghao-stat/MTAFT.

The rest of the article is organized as follows. In Section 2, we present the procedures for estimating the number of thresholds using different criteria and show the selection consistency. Theoretical properties for the parameter estimates and the new score-type test statistic are constructed in Section 3. Monte Carlo simulation studies are conducted in Section 4. Section 5 contains an empirical application on the German breast cancer dataset. We draw some concluding remarks in Section 6. Mathematical proofs of the theorems, the threshold detection algorithms and additional simulation results are collected in Appendix S1.

## 2 | ESTIMATING THE NUMBER OF THRESHOLDS

As we mentioned earlier, the consistency of BIC in Li and Jin (2018) remains under-explored. Thus in this section, we first propose a modified information criterion that can correctly determine the number of thresholds for model (1). An important finding is that the detected number of thresholds by the BIC of Li and Jin (2018) may be over-estimated. However the information criteria could be useful sometimes, but are still ad hoc. Toward this end, we formalize a cross-validation framework inspired by Zou et al. (2020) in the second part.

## 2.1 | The modified Bayesian information criterion

Due to censoring, we only observe $\mathbb{D} = \{\mathbf{D}_i, i = 1,\ldots,n\}$, where $\mathbf{D}_i = (Y_i, \delta_i, \mathbf{X}_i, Z_i)$. To estimate the unknown coefficients $\boldsymbol{\beta}_k$'s, we apply the Stute's method to fit the accelerated failure time model (Stute, 1996. Denote $\mathbb{D}(\tau_k, \tau_{k+1}) = \{\mathbf{D}_i : \tau_{k-1} < Z_i \le \tau_k\}$ as the collection of observations for the $k$th subgroup and $b_k$ is the cardinality. Following the convention, we assume $\tau_0$ and $\tau_{K+1}$ are known lower and upper bounds for the support of $Z_i$. Let $Y_{[1,k]} \le \ldots \le Y_{[b_k,k]}$ be the ordered failure times and $\delta_{[\ell,k]}$ be the concomitant of the $\ell$th-ordered $Y_{[\ell,k]}$, $\ell = 1,\ldots,b_k$. Then, the Kaplan–Meier weights are defined as

$$w_{[1,k]} = \frac{\delta_{[1,k]}}{b_k}, \quad w_{[\ell,k]} = \frac{\delta_{[\ell,k]}}{(b_k - \ell + 1)} \prod_{j=1}^{\ell-1} \left( \frac{b_k - j}{b_k - j + 1} \right)^{\delta_{[j,k]}}, \ \ell = 2,\ldots,b_k. \tag{2}$$

Let $r_{Y_i,k} \in \{1,\ldots,b_k\}$ be the rank of $Y_i$ in $\mathbb{D}(\tau_k, \tau_{k+1})$, that is, $Y_i = Y_{[r_{Y_i,k},k]}$. For simplicity, we let $Y_{[r_{Y_i},k]} = Y_{[r_{Y_i,k},k]} = Y_i$ and $w_{[r_{Y_i},k]} = w_{[r_{Y_i,k},k]}$. Generally, the number of thresholds is estimated by minimizing a prespecified information criterion, that is, $\widehat{K}_{\mathrm{mBIC}} = \arg\min_{K \in \mathcal{K}} \mathrm{mBIC}(K)$ where $\mathcal{K} = \{K : K \in \mathbb{N},\ K \ge 1\}$ is a candidate set for $K$. For model (1), we propose a modified Bayesian information criterion (mBIC),

$$\mathrm{mBIC}(K) = \log \left[ \sum_{k=0}^{K} \frac{b_k}{n} \left\{ \sum_{\mathbf{D}_i \in \mathbb{D}(\tau_k, \tau_{k+1})} w_{[r_{Y_i},k]} \left( Y_i - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k \right)^2 \right\} \right] + p(K+1) \frac{c_0 (\log n)^{\delta_0}}{n}, \tag{3}$$

for some constants $c_0, \delta_0 > 0$. We note that Lebarbier (2005), Zhang and Siegmund (2007) and Wang et al. (2019) also proposed modified forms of penalized information criterion functions for estimating multiple change-points, which share some similarities with the proposed mBIC in (3). Lebarbier (2005) proposed the least-squares estimation with a penalty factor for Gaussian time series that obeyed independence, while Lavielle and Moulines (2000) obtained consistency results in the least-squares setting with time dependence. To achieve selection consistency, they also specified two positive constants that need to be determined within the penalty factor. Zhang and Siegmund (2007) adapted the classic BIC from a Bayesian prospective to identify the number of breaks in the mean of an independent Gaussian process. By assuming a uniform prior distribution over the break dates and regime means, they derived an asymptotic approximation for the posterior probability to adjust to the penalty magnitude, involving both the number of breaks and the intervals between regimes. When the dimensions of the covariates in the threshold models are relatively large, Wang et al. (2019) apply the extended BIC proposed by Wang et al. (2009) to enhance the penalty magnitude in the BIC, which plays the same role with the proposed mBIC.

When $c_0 = 1$ and $\delta_0 = 1$, the information criterion in (3) becomes the conventional BIC applied in Li and Jin (2018). However, Li and Jin (2018) did not provide any theoretical justification for the BIC, and we aim to fill this gap here. Denote the distributions of $T_i$ and $Y_i$ as $F_k$ and $H_k$, respectively, for $k = 0,\ldots,K^\star$. The distribution of log censoring times $\{C_1,\ldots,C_n\}$ is denoted as $G$. We define $U_{H_k} = \inf\{x : H_k(x) = 1\}$ to be the least upper bound for the support of $H_k$. Similarly, we define $U_{F_k}$ and $U_G$ in a similar way to $U_{H_k}$ by replacing $H_k$ with $F_k$ and $G$. We formally introduce the following technical assumptions.

(C1) $T_i$ and $C_i$ are independent, and the censoring mechanism $\delta_i = \min(T_i, C_i)$ is independent of the covariates $\mathbf{X}_i$, for all $i = 1,\ldots,n$. Besides, $U_{F_k} < U_G$ or $U_{F_k} = U_G = \infty$.

(C2) (i) $E(\|\mathbf{X}\|^4) < \infty$ and $E(T_i^2) < \infty$. (ii) $0 < E(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}) = \mathbf{\Sigma}_0$ is finite and nonsingular.

(C3) $E(\varepsilon_i|\mathbf{X}_i) = 0$ and $\varepsilon_i = \sigma(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k^\star)\varepsilon_i^\star$, where $\varepsilon_i^\star, i = 1,\ldots,n$, are i.i.d. from distribution function $F_{\varepsilon^\star}$. Moreover, suppose that $U_{F_k} > U_{F_0}$ and there exists some positive constant $U_{F_k}^\star < U_{F_k} - F_{\varepsilon^\star}$ such that $\sigma(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k^\star) = 0$ if $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k^\star > U_{F_k}^\star$, and 1 otherwise.

(C4) There exist positive definite matrices $E[\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}1\{Z_i \in (\tau_k^\star - \varsigma, \tau_k^\star)\}]$ and $E[\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}1\{Z_i \in (\tau_k^\star, \tau_k^\star + \varsigma)\}]$ in a small $\varsigma$-neighborhood of each of the true threshold parameters $\tau_1^\star,\ldots,\tau_{K^\star}^\star$, where $\varsigma \in (0, \min_{1\le k\le K^\star}(\tau_{k+1}^\star - \tau_k^\star)/4]$.

(C5) There exist $0 < a_1 < \cdots < a_{K^\star} < 1$ such that $\Pr(Z_i \le \tau_j^\star) = a_j$ for $j = 1,\cdots,K^\star$.

Condition (C1) implies that $\delta_i$ is independent of $\mathbf{X}_i$ conditional on $C_i$, a common assumption in the independent censoring framework, see for example, Stute (1993, 1996). An equivalent condition is the Markov assumption of Stute and Wang (1993), where $\Pr(T_i \le C_i|\mathbf{X}_i, T_i) = \Pr(T_i \le C_i|T_i)$. From (C1), $U_{H_k} = \min(U_{F_k}, U_G)$. Condition (C2) (i) involves some unconditional moment restrictions on $\mathbf{X}$ and $T$; while (ii) is a standard invertibility condition because threshold points are estimated through a global weighted least square search. Condition (C3) accommodates heteroscedastic error, as demonstrated in Li and Jin (2018). (C3) also implies $T_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k^\star$ if $U_{F_k}^\star + U_{F_0} < Y_i \le U_{F_k}$. Condition (C4) imposes restrictions on the regressors in the vicinity of the threshold points to prevent local collinearity issue and ensure the identification of threshold effects. Notice that we allow the threshold variable $Z_i$ to be a component of $\mathbf{X}_i$, in which case Condition (C4) is automatically satisfied if the threshold variable $Z_i$ has a positive density within a small neighborhood of each threshold. Condition (C5) dictates that the true number of thresholds is fixed with asymptotically distinct multithreshold effects.

Theorem 1 below demonstrates the selection consistency of the mBIC in (3) for identifying the number of thresholds when $\delta_0 > 2$ for a sufficiently large $n$.

**Theorem 1.** *Under the Conditions (C1)–(C5) and $\delta_0 > 2$, $\widehat{K}_{\mathrm{mBIC}}$ is consistent in the sense that $Pr(\widehat{K}_{\mathrm{mBIC}} = K^\star) \to 1$, as $n \to \infty$.*

We briefly outline the process for establishing the consistency of the mBIC. Intuitively, for the underfitting case with $K < K^\star$, it can be shown that $\Pr\{\min_K \mathrm{mBIC}(K) > \mathrm{mBIC}(K^\star)\} \to 1$ as $n \to \infty$ by noting that the discrepancy in the first term of Equation (3) becomes predominate. On the other hand, for the overfitting case with $K > K^\star$, while the first term does not decrease too much as $K$ increases, the difference in the second term, i.e., the penalty factor could be relatively large. According to the proof of Theorem 1, we choose the penalty factor as $c_0(\log n)^{\delta_0}/n$ for $\delta_0 > 2$, which is in accordance with Liu et al. (1997) under no-censoring context. This choice can be explained by the fact that the decrement of the first term for the overfitting case has an upper bound of the order $(\log n)^2/n$ (see, Lemma 2 in Appendix S1), which is smaller than the penalty factor with order $(\log n)^{\delta_0}/n$ for $\delta_0 > 2$ as $n \to \infty$.

We stress that the mBIC and BIC mainly differ in the severity of the penalty factor for overfitting, with the BIC being a relatively conservative criterion. Nevertheless, there is no universally optimal pair of $c_0$ and $\delta_0$ that always works well for all models. From the perspective of model specification, a relatively larger penalty term would be preferable for easily identified models as it can markedly diminish the risk of overestimation. However, if the change-point signal is hard to identify, for example, some of $\|\boldsymbol{\beta}_{k+1}^\star - \boldsymbol{\beta}_k^\star\|$'s are small, then the penalty term cannot be too large to avoid probable underestimation. Another factor influencing the choice is the error distribution. It is well-known that a heavy-tailed distribution is more likely to produce extreme values, which resemble change-points, and a severe penalty is therefore essential. Hence, it is difficult to select

an appropriate penalty factor that suits diverse model settings and error distributions, creating a disparity between theoretical principles and practical application.

## 2.2 | The OPCV criterion

We suggest a new data-driven approach to estimate the number of thresholds through an OPCV scheme. The key idea involves dividing the dataset into a training set and a validation set, in which we use the training set to construct a candidate multithreshold model and evaluate the prediction error on the validation set. However, since the threshold model has an intrinsic order structure, traditional random data splitting is not a good choice. Inspired by Zou et al. (2020), we split the sample based on the parity of the threshold variable locations, being either odd or even, that is,

$$\mathbb{D}_O = \{(Y_{2i-1}, \delta_{2i-1}, \mathbf{X}_{2i-1}, Z_{2i-1}), i = 1,\ldots,N\} \quad \text{and} \quad \mathbb{D}_E = \mathbb{D} \setminus \mathbb{D}_O,$$

where we assume $n = 2N$ for simplicity. This partitioning strategy makes the difference between the training and validation sets minimal, which mostly preserves the original threshold structure within each subset. Compared to the traditional randomly drawing method, the variability caused by the extra randomness can be greatly reduced.

Next, to avoid numerically estimating $\boldsymbol{\beta}_k$'s, we apply the score function to measure the goodness-of-fit. Specially, for a given $\mathcal{T}_K = (\tau_1,\ldots,\tau_K)$, the score function corresponding to the weighted least squares loss of the Stute estimator on $\mathbb{D}(\tau_k, \tau_{k+1})$ is defined as

$$\mathbf{S}_i(\mathbf{D}_i; \boldsymbol{\gamma}) = b_k w_{[r_{Y_i},k]} \mathbf{X}_i (Y_i - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\gamma}),$$

where $\boldsymbol{\gamma}$ is a given $p$-vector. Ideally, $E\left(\overline{\mathbf{S}}_k\right) \neq E\left(\overline{\mathbf{S}}_{k+1}\right)$ implies a change on $\boldsymbol{\beta}$ at $Z_i = \tau_k$, where

$$\overline{\mathbf{S}}_k = b_k^{-1} \sum_{\mathbf{D}_i \in \mathbb{D}(\tau_k, \tau_{k+1})} \mathbf{S}_i(\mathbf{D}_i; \boldsymbol{\gamma}). \tag{4}$$

Hence, given a candidate model $\mathcal{M}_K$, which is specified by a set of thresholds $\mathcal{T}_K = (\tau_1,\ldots,\tau_K)$ and the corresponding parameters $\boldsymbol{\mu}_k^\star \equiv E\left(\overline{\mathbf{S}}_k\right)$'s, the goodness-of-fit measure for $\mathcal{M}_K$ is

$$C(\mathcal{M}_K; \mathbb{D}) = \sum_{k=0}^{K} \sum_{\mathbf{D}_i \in \mathbb{D}(\tau_k, \tau_{k+1})} \left\{\mathbf{S}_i(\mathbf{D}_i; \boldsymbol{\gamma}) - \boldsymbol{\mu}_k^\star\right\}^{\mathrm{T}} \boldsymbol{\Omega}_n \left\{\mathbf{S}_i(\mathbf{D}_i; \boldsymbol{\gamma}) - \boldsymbol{\mu}_k^\star\right\}, \tag{5}$$

where $\boldsymbol{\Omega}_n$ is a prespecified weight matrix. We remark that it can be verified that (5) is invariant to $\boldsymbol{\gamma}$ and $\boldsymbol{\Omega}_n$ and their choices are not crucial. In practice, one can simply set $\boldsymbol{\gamma}$ as the Stute estimator of $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}_n$ as the identity matrix.

To conduct the cross-validation method, let us use $\mathbb{D}_O$ as the training set for change-point detection and $\mathbb{D}_E$ for validation. Given $K$, obtain the set of change-points $\widehat{\mathcal{T}}_K^O = (\widehat{\tau}_1^O,\ldots,\widehat{\tau}_K^O)$ based on $\mathbb{D}_O$ using a change-point detection algorithm, say $\mathcal{A}(K; \mathbb{D}_O)$. Denote the trained change-point model as $\widehat{\mathcal{M}}_K^O$. The goodness-of-fit criterion on the validation set $\mathbb{D}_E$ is therefore

$$C(\widehat{\mathcal{M}}_K^O; \mathbb{D}_E) = \sum_{k=0}^{K} \sum_{\mathbf{D}_i \in \mathbb{D}_E(\widehat{\tau}_k^O, \widehat{\tau}_{k+1}^O)} \left\{\mathbf{S}_i(\mathbf{D}_i; \boldsymbol{\gamma}) - \overline{\mathbf{S}}_k^O\right\}^{\mathrm{T}} \boldsymbol{\Omega}_n \left\{\mathbf{S}_i(\mathbf{D}_i; \boldsymbol{\gamma}) - \overline{\mathbf{S}}_k^O\right\}, \tag{6}$$

**Algorithm 1.** The OPCV procedure to detect the threshold effects

**Step 1.** For each $K \in \mathcal{K}$, use the detection algorithm $\mathcal{A}(K; \mathbb{D}_O)$ to obtain the threshold model $\widehat{\mathcal{M}}_K^O$.
**Step 2.** Compute $\bar{\mathbf{S}}_k^O$ for $k = 0, \ldots, K$ based on $\mathbb{D}_O$, and calculate the prediction measure $C(\widehat{\mathcal{M}}_K^O; \mathbb{D}_E)$ based on $\mathbb{D}_E$.
**Step 3.** Repeat Steps 1 and 2 to obtain $C(\widehat{\mathcal{M}}_K^E; \mathbb{D}_O)$ by switching the positions of $\mathbb{D}_O$ and $\mathbb{D}_E$.
**Step 4.** The estimated number of threshold effects is given by (7). The threshold parameters can be estimated by using the whole sample.

where $\bar{\mathbf{S}}_k^O$ is defined in a similar way to (4) by replacing $\mathbb{D}(\tau_k, \tau_{k+1})$ with $\mathbb{D}_O(\hat{\tau}_k^O, \hat{\tau}_{k+1}^O)$. Repeat above process by interchanging $\mathbb{D}_O$ and $\mathbb{D}_E$, which yields $C(\widehat{\mathcal{M}}_K^E; \mathbb{D}_O)$. The estimated number of thresholds is

$$\hat{K}_{\mathrm{OPCV}} = \arg\min_{K \in \mathcal{K}} \Big\{ C(\widehat{\mathcal{M}}_K^O; \mathbb{D}_E) + C(\widehat{\mathcal{M}}_K^E; \mathbb{D}_O) \Big\}. \tag{7}$$

After obtaining $\hat{K}_{\mathrm{OPCV}}$, the final change-point estimators can be further refined by using the entire dataset. Since the candidate model $\mathcal{M}_K$ needs to be estimated based on the available sample, the OPCV procedure facilitates the determination of the number of thresholds in practice. We summarize the procedures in Algorithm 1.

It is well-known that CV is a widespread strategy in performing model selection due to its simplicity and universality. Arlot and Celisse (2010) provided guidelines for choosing the best CV procedure based on the problem at hand, including tasks like change-point detection. Another related work to this paper is Arlot and Celisse (2011) who proposed to use a CV-based empirical risk to detect univariate mean change with heterogeneity. By an extensive simulation study, they demonstrated the robustness of the CV procedure against heteroscedasticity. Despite the advantages, the reduced sample size in each divided fold may potentially impact the estimation accuracy of change-points. To mitigate this limitation, we recommend two practical suggestions. First, one can implement the change-point detection methods that maintain decent detection accuracy when dealing with smaller sample sizes or weaker signals. Second, a multiple-splitting scheme could be employed to improve the stability of the selection procedure. Please refer to Remark 1 for more details.

We now establish the selection consistency of the OPCV criterion in Theorem 2. To facilitate the proof, we introduce the following condition.

(C6) Denote the true jump signal strength as $\boldsymbol{\mu}_{k+1}^\star - \boldsymbol{\mu}_k^\star = E(\mathbf{X}_i \mathbf{X}_i^{\mathrm{T}})(\boldsymbol{\beta}_{k+1}^\star - \boldsymbol{\beta}_k^\star)$. There exist some positive $\phi$ such that

$$0 < \inf_{|z - \tau_k^\star| < \phi} E\big\{ \mathbf{X}_i \mathbf{X}_i^{\mathrm{T}} (\boldsymbol{\beta}_{k+1}^\star - \boldsymbol{\beta}_k^\star) | Z_i = z \big\} < \infty.$$

Condition (C6) specifies the fixed-threshold-effect setting (Chan, 1993). It asserts that $\boldsymbol{\delta}_k^\star = \boldsymbol{\beta}_{k+1}^\star - \boldsymbol{\beta}_k^\star \neq 0$ for all $n$, thus excluding the diminishing setup where $\boldsymbol{\delta}_k^\star \to 0$ as $n \to \infty$. Condition (C6) also rules out the continuous threshold regression studied by, for example, Zhong et al. (2022) and Wan et al. (2023). Theorem 2 presents the selection consistency of the OPCV estimator in (7).

**Theorem 2.** *Under the Conditions (C1)–(C6), $\widehat{K}_{\mathrm{OPCV}}$ estimated by (7) is consistent, that is $Pr(\widehat{K}_{\mathrm{OPCV}} = K^\star) \to 1$ as $n \to \infty$.*

We now illustrate the validity of the OPCV criterion. Note that $\mathcal{C}(\mathcal{M}_K; \mathbb{D})$ can be further decomposed into

$$\begin{aligned}
&\mathcal{C}(\mathcal{M}_K; \mathbb{D}) \\
&= \sum_{k=0}^{K} \sum_{\mathbf{D}_i \in \mathbb{D}(\tau_k, \tau_{k+1})} \left\{ \mathbf{S}_i(\mathbf{D}_i; \boldsymbol{\gamma}) - \overline{\mathbf{S}}_k \right\}^{\mathrm{T}} \boldsymbol{\Omega}_n \left\{ \mathbf{S}_i(\mathbf{D}_i; \boldsymbol{\gamma}) - \overline{\mathbf{S}}_k \right\} + \sum_{k=0}^{K} b_k \left( \overline{\mathbf{S}}_k - \boldsymbol{\mu}_k^\star \right)^{\mathrm{T}} \boldsymbol{\Omega}_n \left( \overline{\mathbf{S}}_k - \boldsymbol{\mu}_k^\star \right) \\
&\equiv \mathcal{S}_{\mathbf{S}}^2(\mathcal{T}_K; \boldsymbol{\Omega}_n) + \mathcal{D}(\mathcal{M}_K; \mathbb{D}).
\end{aligned}$$

First, in the underfitting case (i.e., $K < K^\star$), $\mathcal{S}_{\mathbf{S}}^2(\mathcal{T}_K; \boldsymbol{\Omega}_n) - \mathcal{S}_{\mathbf{S}}^2(\mathcal{T}_{K^\star}; \boldsymbol{\Omega}_n)$ could be quite large, compared to the second term, which helps prevent underfitting. In the overfitting case (i.e., $K > K^\star$), $\mathcal{S}_{\mathbf{S}}^2(\mathcal{T}_K; \boldsymbol{\Omega}_n) - \mathcal{S}_{\mathbf{S}}^2(\mathcal{T}_{K^\star}; \boldsymbol{\Omega}_n)$ would be quite small, but the term $\mathcal{D}(\mathcal{M}_K; \mathbb{D})$ would dominate $\mathcal{D}(\mathcal{M}_{K^\star}; \mathbb{D})$ under certain conditions. Consequently, $\mathcal{D}(\mathcal{M}_K; \mathbb{D})$ plays the role of the penalty term in the mBIC or BIC but in a data-driven manner. This actually avoids setting a use-specified penalty factor in the information criterion.

*Remark* 1. Our OPCV procedure is actually a specially designed two-fold CV variant. It is well known that a multiple-splitting scheme can improve the selection stability. A $S$-fold CV where $S$ ($S \geq 2$) is a given positive integer can be conducted as follows. For each split $s = 1, \ldots, S$, we randomly draw the indices $i_s \in \{2i - 1, 2i\}$ for $i = 1, \ldots, N$, and denote the drawn sample as $\mathbb{D}_{O_s} = \{(Y_{i_s}, \delta_{i_s}, \mathbf{X}_{i_s}, Z_{i_s}), i = 1, \ldots, N\}$. Run the change-point detection algorithm $\mathcal{A}$ on the data split $\mathbb{D}_{O_s}$ to obtain $\widehat{\mathcal{M}}_K^{O_s}$ and calculate the goodness-of-fit criterion on $\mathbb{D} \setminus \mathbb{D}_{O_s}$. The number of change-points estimator by the $S$-fold CV is $\widehat{K} = \arg\min_{K \in \mathcal{K}} \sum_{s=1}^{S} \mathcal{C}(\widehat{\mathcal{M}}_K^{O_s}; \mathbb{D} \setminus \mathbb{D}_{O_s})$.

*Remark* 2. For each $K \in \mathcal{K}$, we need to find $K$ estimated threshold points by applying the algorithm $\mathcal{A}$. The `MTAFT` package offers two algorithm choices for $\mathcal{A}$. One is the dynamic programming (DP) algorithm (Auger & Lawrence, 1989) that searches for globally optimal solution by minimizing the cost for all possible segmented sequences. The DP implementation can be performed in the order of $O(K_{\max} n^2)$, where $K_{\max}$ is an upper bound of thresholds. The other is the wild binary segmentation (WBS) algorithm (Fryzlewicz, 2014) that finds locally optimal solutions. It should be noted that the presented WBS algorithm is slightly different from the classical version in Fryzlewicz (2014), which needs to predefine a cutoff value, and stops until no detected change magnitude exceeds the threshold. In the revised version, we nest the change-point estimates for further model selection, which is free of specifying any nuisance parameters. This algorithm has the complexity of order $O(n \log n)$, which is more efficient than the DP algorithm, particularly for large sample sizes. However, it may miss some true threshold effects for minor and weak changes. We summarize the detailed procedures for these algorithms in Appendix S1. In our numerical studies, we mainly apply the DP algorithm for more accurate estimation. We also conduct comparisons with the penalized two-stage algorithm proposed by Li and Jin (2018) through simulation studies detailed in Appendix S1.

# 3 | ASYMPTOTICS AND TESTING

## 3.1 | Asymptotic properties of the parameter estimates

Having established the selection consistency regarding the number of thresholds, we next study the asymptotic properties of the estimators for both the thresholds and the regression coefficients given that the number of thresholds $K^\star$ is known in this section. Given the estimated thresholds $\widehat{\tau}_k, k = 1,\ldots, K^\star$, we estimate the slope parameters $\boldsymbol{\beta}_k$ for each subsample $\mathbb{D}(\widehat{\tau}_k, \widehat{\tau}_{k+1})$ separately by using the Stute's estimator, denoted by $\widehat{\boldsymbol{\beta}}_k$.

(C7) The threshold variable $Z_i$ has a positive and bounded probability density function, which is continuous within some small neighborhoods of the true thresholds.

Condition (C7) is standard in threshold regression, which can be relaxed along the line of Yu and Zhao (2013). Theorem 3 below justifies the consistency of the threshold estimators and the asymptotic normality of the subgroup-specific slope estimators.

**Theorem 3.** *For model (1), under Conditions (C.1)–(C.7), we have*

*(i)* $\widehat{\tau}_k - \tau_k^\star = O_p(n^{-1})$ *for some* $k = 1,\ldots, K^\star$.
*(ii) the Stute's estimators* $\widehat{\boldsymbol{\beta}}_k$ *based on the estimated* $\widehat{K}$ *and* $\widehat{\tau}_k$*'s is asymptotically normal estimates of* $\boldsymbol{\beta}_k^\star, k = 0,\ldots, K^\star$.

We provide some comments on Theorem 3. First, the asymptotic performance for threshold parameters is built under the fixed-threshold-effect framework. Another prevalent assumption is the diminishing-threshold-effect framework of Hansen (2000), which assumes $\boldsymbol{\delta}_k^\star \to 0$ as $n \to \infty$. In this case, the convergence rate of $\widehat{\tau}_k$ would be slowed down and also depend on the diminishing rate, but a tractable asymptotic distribution could be obtained. However, this case has poor finite sample performance especially when the true threshold effect is indeed fixed, a limitation underscored by Yu (2015). Second, it can be show that $n(\widehat{\tau}_k - \tau_k^\star)$ converges to the argmin of a compound Poisson process (Yu & Phillips, 2018). However, this limiting distribution is too complicated to be used for statistical inference and thus is omitted. Third, the slope estimates are asymptotically normal, which lends support to the inference, such as conducting hypothesis tests or constructing confidence intervals.

## 3.2 | Testing for the existence of threshold effects

Before conducting detection for the multithreshold model (1), it is necessary to first guarantee that there indeed exists at least one threshold effect. Otherwise, the detection procedure could be meaningless. Thus, we are interested in the following null and alternative hypotheses:

$$\mathbb{H}_0 : K = 0, \quad \text{versus} \quad \mathbb{H}_1 : K \geq 1. \tag{8}$$

To test (8), one may employ the classical Wald-type test statistics based on the asymptotic properties of the Stute estimators under $\mathbb{H}_1$, or the likelihood-ratio-type test. However, both statistics require the estimation of $K$ in the presence of threshold effects, which is intractable. Besides,

deriving the asymptotic properties for the alternative estimators is also theoretically challenging. Note that when $\mathbb{H}_0$ holds, the original model is completely linear, that is,

$$Y_i = \min(T_i, C_i) \quad \text{and} \quad T_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_o + \varepsilon_{oi}, \tag{9}$$

where $\boldsymbol{\beta}_o$ and $\varepsilon_{oi}$ are the regression coefficients and the error term under the null. Then the estimator for $\boldsymbol{\beta}_o$ is

$$\widehat{\boldsymbol{\beta}}_o = \left(\sum_{i=1}^{n} w_{[r_{Y_i}]}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-1}\left(\sum_{i=1}^{n} w_{[r_{Y_i}]}\mathbf{X}_i Y_i\right),$$

where the weights $w_{[r_{Y_i}]}$, $i = 1,\ldots,n$ are defined in a similar way in (2) by letting $\mathbb{D}(\tau_k, \tau_{k+1}) = \mathbb{D}$ and $b_k = n$.

This motivates us to consider the score-type statistic. The building block of our proposed test statistic is

$$\mathbf{R}_n(\tau) = n^{1/2}\sum_{i=1}^{n} w_{[r_{Y_i}]}\left\{\mathbf{X}_i 1(Z_i \le \tau) - \widehat{\mathbf{Q}}_1(\tau)\widehat{\mathbf{Q}}^{-1}\mathbf{X}_i\right\}\widehat{\varepsilon}_{oi}, \quad \tau \in \Gamma,$$

where $\widehat{\mathbf{Q}}_1(\tau) = \sum_{i=1}^{n} w_{[r_{Y_i}]}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}1(Z_i \le \tau)$, $\widehat{\mathbf{Q}} = \sum_{i=1}^{n} w_{[r_{Y_i}]}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}$, $\tau$ is a random scalar denoting any potential change-point location and $\Gamma$ is the range set of all $\tau$'s. It is standard to assume $\Gamma = [Z_{(\varrho)}, Z_{(1-\varrho)}]$ where $Z_{(\varrho)}$ and $Z_{(1-\varrho)}$ are the $\varrho$ and $(1-\varrho)$ quantiles of $Z_i$, and $\varrho$ is usually 0.1 or 0.15. Note that $\tau$ is not identifiable under the null hypothesis. As in the usual threshold effect testing such as Lee et al. (2011), we adopt the union–intersection principle (Roy, 1953) and take a supremum of the building blocks:

$$L_n = \sup_{\tau\in\Gamma}\|\mathbf{R}_n(\tau)\|, \tag{10}$$

where $\|\cdot\|$ is the Euclidean norm. The main idea of our score-type statistic is quite intuitive. Under $\mathbb{H}_0$, it is easy to show that $\widehat{\boldsymbol{\beta}}_o$ is consistent and $\widehat{\mathbf{Q}}_1(\tau)\widehat{\mathbf{Q}}^{-1}n^{1/2}\sum_{i=1}^{n} w_{[r_{Y_i}]}\mathbf{X}_i\widehat{\varepsilon}_{oi} = o_p(1)$. The score function $n^{1/2}\sum_{i=1}^{n} w_{[r_{Y_i}]}\mathbf{X}_i\widehat{\varepsilon}_{oi}$ would show a random pattern around 0, which makes $L_n$ relatively small. On the other hand, if there exists any threshold effect, $\widehat{\boldsymbol{\beta}}_o$ would significantly differ from the true value and $\widehat{\varepsilon}_{oi}$ would be far from zero in a systematic fashion concerning $T_i$. Thus a large $L_n(\tau)$ is the strong evidence against $\mathbb{H}_0$. Here we recenter $w_{[r_{Y_i}]}\mathbf{X}_i 1(Z_i \le \tau)$ by $w_{[r_{Y_i}]}\widehat{\mathbf{Q}}_1(\tau)\widehat{\mathbf{Q}}^{-1}\mathbf{X}_i$ as the effect of $\widehat{\boldsymbol{\beta}}_o$ will not disappear asymptotically under $\mathbb{H}_1$. Recentering is to offset such effect.

Next, we consider the following local alternative model,

$$Y_i = \min(T_i, C_i) \quad \text{and} \quad T_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_o + \mathbf{X}_i^{\mathrm{T}}n^{-1/2}\boldsymbol{\delta}1(Z_i \le \tau) + \varepsilon_i, \tag{11}$$

where $\boldsymbol{\delta}$ is a $p$-dimensional vector that may depend on $Z_i$. We then formally illustrate the limiting behavior of $L_n$.

**Theorem 4.** *Suppose the Conditions (C1)–(C3) hold. Under the local alternative model in (11), we have*

**Algorithm 2.** Simulating critical values for the statistic $L_n$

**Step 1.** Generate i.i.d. $N(0,1)$ random variables $\{\upsilon_i\}_{i=1}^n$.

**Step 2.** Calculate the bootstrap test statistic $L_n^{(b)} = \sup_{\tau\in\Gamma} \|\mathbf{R}_n^*(\tau)\|$, where

$$\mathbf{R}_n^*(\tau) = n^{1/2} \sum_{i=1}^{n} w_{[r_{Y_i}]} \left\{ \mathbf{X}_i 1(Z_i \le \tau) - \widehat{\mathbf{Q}}_1(\tau)\widehat{\mathbf{Q}}^{-1}\mathbf{X}_i \right\} \widehat{\varepsilon}_{oi}\upsilon_i.$$

**Step 3.** Repeat Steps 1 and 2 $B$ times to generate $\{L_n^{(b)}\}_{b=1}^B$.

**Step 4.** If $\widehat{p}_n = B^{-1}\sum_{b=1}^B 1\{L_n^{(b)} \ge L_n\} \le \alpha$, we reject $\mathbb{H}_0$ at $\alpha$; otherwise, accept $\mathbb{H}_0$.

$$L_n \xrightarrow{d} \sup_{\tau\in\Gamma} \|\mathbf{R}(\tau) + \left\{\mathbf{Q}_1(\tau\wedge\tau^\star) - \mathbf{Q}_1(\tau)\mathbf{Q}^{-1}\mathbf{Q}_1(\tau^\star)\right\}\boldsymbol{\delta}\|,$$

*where* $\mathbf{R}(\tau)$ *is mean zero Gaussian process with covariance kernel*

$$E\left[\left\{\mathbf{X}_i 1(Z_i \le \tau) - \mathbf{Q}_1(\tau)\mathbf{Q}^{-1}\mathbf{X}_i\right\}\left\{\mathbf{X}_i 1(Z_i \le \tau') - \mathbf{Q}_1(\tau')\mathbf{Q}^{-1}\mathbf{X}_i\right\}^{\mathrm{T}}\varepsilon_i^2\right],$$

*with* $\mathbf{Q}_1(\tau\wedge\tau^\star) = E\left\{\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} 1(Z_i \le \tau\wedge\tau^\star)\right\}$, $\mathbf{Q}_1(\tau) = \mathbf{Q}_1(\tau\wedge\tau)$ *and* $\mathbf{Q} = E\left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)$.

Theorem 4 implies that the asymptotic distribution under $\mathbb{H}_0$ (i.e., $\boldsymbol{\delta} = 0$) is not pivotal, and the critical values of $L_n$ cannot be tabulated directly. But we can characterize the limiting behavior of the test statistic by simulating the Gaussian process $\mathbf{R}_n(\tau)$. The detailed procedure is shown in Algorithm 2.

The following theorem verifies the validity of the bootstrap resampling scheme in Algorithm 2.

**Theorem 5.** *Under the conditions in Theorem 4 and* $\mathbb{H}_0$, $\mathbf{R}_n^*(\tau)$ *defined in Algorithm 2 converges to* $\mathbf{R}(\tau)$ *as* $n \to \infty$.

## 4 | NUMERICAL STUDIES

We conduct simulation studies to investigate the finite sample performance of our proposed methodologies. All simulation results reported are obtained across 500 replicated samples.

### 4.1 | Power analysis

We first evaluate the performance of the proposed score-type test statistics for testing the existence of threshold effects. The datasets were generated from the following two cases:

**Example 1.** $T_i = 1 + X_i 1(Z_i < 0.5) + (1+\alpha)X_i 1(Z_i \ge 0.5) + \varepsilon_i$;

**Example 2.** $T_i = 1 + X_i 1(Z_i < 0.3) + (1+\alpha)X_i 1(0.3 \le Z_i < 0.6) + X_i 1(Z_i \ge 0.6) + \varepsilon_i$.

For both cases, $Z_i \overset{i.i.d}{\sim} U[0,1]$, $X_i \overset{i.i.d}{\sim} N(0,1)$, $\varepsilon \overset{i.i.d}{\sim} N(0,0.5)$ and $Y_i = \min(T_i, C_i)$ is observed, where $C_i \overset{i.i.d}{\sim} N(2,16)$ for the censoring rate of 40%, and $N(4.5,16)$ for the censoring rate of 20%. Here $\alpha$ controls the degree of departure from the null hypothesis, and $\alpha = 0$ implies no threshold

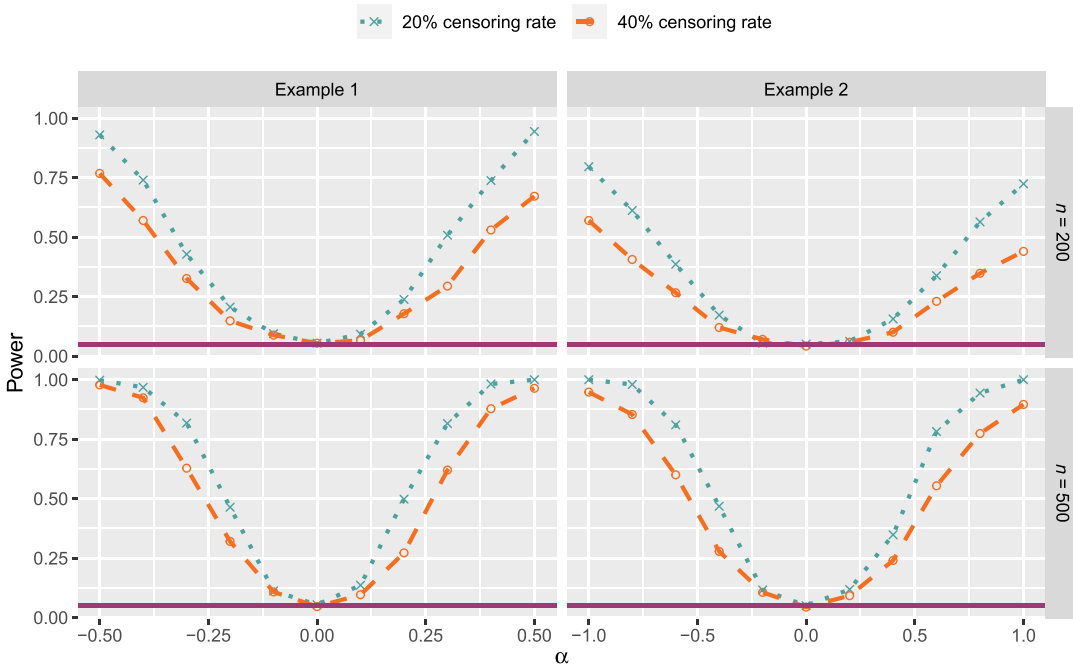


**FIGURE 1** The local power at censoring rates 20% and 40% obtained by 500 repetitions with sample sizes $n = 200$ and 500 across varying $\alpha$'s.

effect in the model. In terms of our working model, the true parameters in Example 1 are $K = 1$, $\tau = 0.5$ and $(\boldsymbol{\beta}_0^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}}) = (1, 1, 1, 1 + \alpha)$. For Example 2, the true parameters are $K = 2$, $(\tau_1, \tau_2) = (0.3, 0.6)$ and $(\boldsymbol{\beta}_0^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}) = (1, 1, 1, 1 + \alpha, 1, 1)$. We consider two sample sizes $n = 200$ and 500, respectively.

We have drawn the power curves across different values of $\alpha$'s in Figure 1, and the exact $p$-values can be found in Appendix S1. The nominal significance level is set as 5%. We observe the following findings. First and most importantly, our test exhibits acceptable type I errors, which matches the nominal level, and the power is also decent. Second, as $\alpha$ moves away from zero, the power for Example 2 also gradually approaches one. This illustrates that our score-type test statistic can detect the presence of multithreshold effects effectively. Third, as expected, our test statistic demonstrates higher detecting powers as the sample size increases or the censoring rate decreases.

## 4.2 | Selection consistency

We next evaluate the performance of proposed methods in estimating the number of true thresholds $K^\star$. We generate random samples from model (1) in the following:

> **Example 3.** $K^\star = 2$, $\mathbf{X}_i = (x_{1,i}, \ldots, x_{4,i})^{\mathrm{T}}$ with $x_{1,i} = 1$ and $x_{j,i} \overset{i.i.d}{\sim} N(0, 1)$, $j = 2, 3, 4$. $(\boldsymbol{\beta}_0^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}) = (2, 1, 1, 1, 1, -1, -1, 0.5, 0, 1, 0.5, -1)$, $C_i \overset{i.i.d.}{\sim} N(2, 16)$ and $Z_i = x_{2,i}$. The true thresholds are $\tau_1^\star = -0.5244$ and $\tau_2^\star = 0.2533$, corresponding to the 30% and 60% lower percentiles of the standard normal distribution;

**Example 4.** The same setting as Example 3 except $C_i \overset{i.i.d.}{\sim} N\left(\sum_{j=1}^{4} x_{j,i}, 16\right)$;

**Example 5.** The model is set as $T_i = \beta_{0i} + \beta_{1i}x_i + \sigma\varepsilon_i$, in which $K^\star = 3$ and the threshold parameters are $\tau_k^\star = k/(K^\star + 1) + U(-1, 1) \times n^{-3/5}$ The coefficients $\beta_{0i} = \beta_{1i}$ is generated from the signal vector $\boldsymbol{\gamma}$ =(0,1.64, −1.66, 1.32, −1.32, 0.98, −1.39, 1.29, −1.53, 1.68, −1.37). Specifically, let $\gamma_{k-1}$ denote the $k$th element of $\boldsymbol{\gamma}$ and $\beta_{1i} = \gamma_{\mathrm{mod}(J_0+k,11)}$ for $\tau_k^\star < Z_i \le \tau_{k+1}^\star$, $k = 0, \ldots, K^\star$, where $J_0$ is an integer randomly sampling from $\{1, \ldots, 11\}$ and $\mathrm{mod}(a, b)$ represents the modulo operator. The covariate is distributed from $\sqrt{3}/2\sigma_\beta U(-1, 1)$ with $\sigma_\beta = \mathrm{sd}(\beta_{1i}'\mathrm{s})$. We select $\sigma$ to make the signal-to-noise $\sigma_\beta/\sigma$ equal to 3.

Examples 3 and 4 share similarities with those presented in Li and Jin (2018) with $p = 4$, while Example 5 is motivated by Zou et al. (2020). Two scenarios of the error distribution are considered: (i) $\varepsilon_i \overset{i.i.d.}{\sim} N(0, 0.5)$, and (ii) $\varepsilon_i \overset{i.i.d}{\sim} t(3)/\sqrt{6}$. The sample size $n$ is 500, and the censoring rates are all around 40%. We compare the following approaches: the two-stage penalized estimation algorithm proposed by Li and Jin (2018), with the optimal tuning parameter selected using the BIC and mBIC criteria (Pen+BIC and Pen+mBIC); and the DP algorithm in conjunction with the BIC, mBIC and CV criteria (DP+BIC, DP+mBIC and DP+CV). The CV criterion requires estimation for each $K \in \mathcal{K}$, while the detection method in Li and Jin (2018) cannot yield such results. Hence, we exclude the Pen+CV approach from the simulation study. To get a broader picture of the performance of the mBIC, we explore two different settings $c_0 = 0.299, \delta_0 = 2.01$ and $c_0 = 0.299, \delta_0 = 2.10$, denoted as mBIC1 and mBIC2, respectively. We label the DP algorithm with mBIC1 as DP+mBIC1, with similar conventions applied to other methodologies. To assess the finite-sample performance of threshold detection, we analyze the frequency of $\widehat{K}$ over 500 repetitions. To compare the accuracy of the estimated thresholds, we also calculate the Hausdorff distances defined as $\max\{\max_i \min_j |\tau_i^\star - \widehat{\tau}_j|, \max_j \min_i |\tau_i^\star - \widehat{\tau}_j|\}$.

The Hausdorff distance and the frequency of $\widehat{K}$ for various methods are shown in Figures 2 and 3, respectively. From the results, we yield several key findings. First, from the distribution of $\widehat{K}$, the conventional BIC always overestimates $K$, which implies that the penalization magnitude is inadequate. In terms of the Hausdorff distance, the biases of the superfluous threshold estimates generated by the BIC are nonnegligible, accompanied by relatively large SDs. These facts align with Theorem 4 that the BIC used in Li and Jin (2018) is not a consistent measure. Second, by imposing a heavier penalization magnitude, mBIC2 is outperformed by mBIC1 for both the Pen and DP algorithms. While mBIC1 and mBIC2 exhibit higher correct identification probabilities under normal noise, they slightly overestimate the number of thresholds under $t(3)/\sqrt{6}$ error. This highlights the significant impact of penalization magnitude and error distribution on the performance of information criterion methods, as reported by Zou et al. (2020). Third, the adoption of our OPCV procedure mitigates overfitting tendencies and substantially enhances correct identification probabilities, which shows its stability. More importantly, the OPCV criterion is completely data driven without setting any tuning parameters. Fourth, when the tuning parameter in the shrinkage estimation from Li and Jin (2018) is selected by the mBIC, the performance improves significantly. We see that Pen+mBIC1 and Pen+mBIC2 have greater Hausdorff distances and higher detection accuracy rates than Pen+BIC. The penalized algorithm developed by Li and Jin (2018) proves advantageous in detecting the threshold effects compared to the DP algorithm. More comparative studies on computation efficiency and estimation accuracy across various detection algorithms are conducted in Appendix S1.

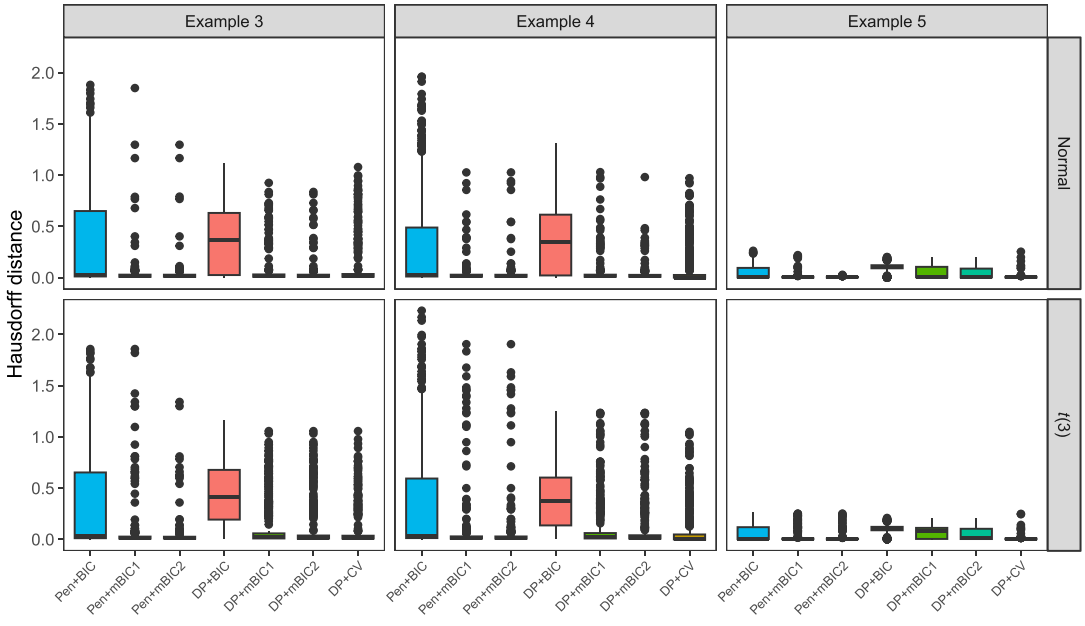


FIGURE 2 The Hausdorff distance under Examples 3–5 for various detection methods.

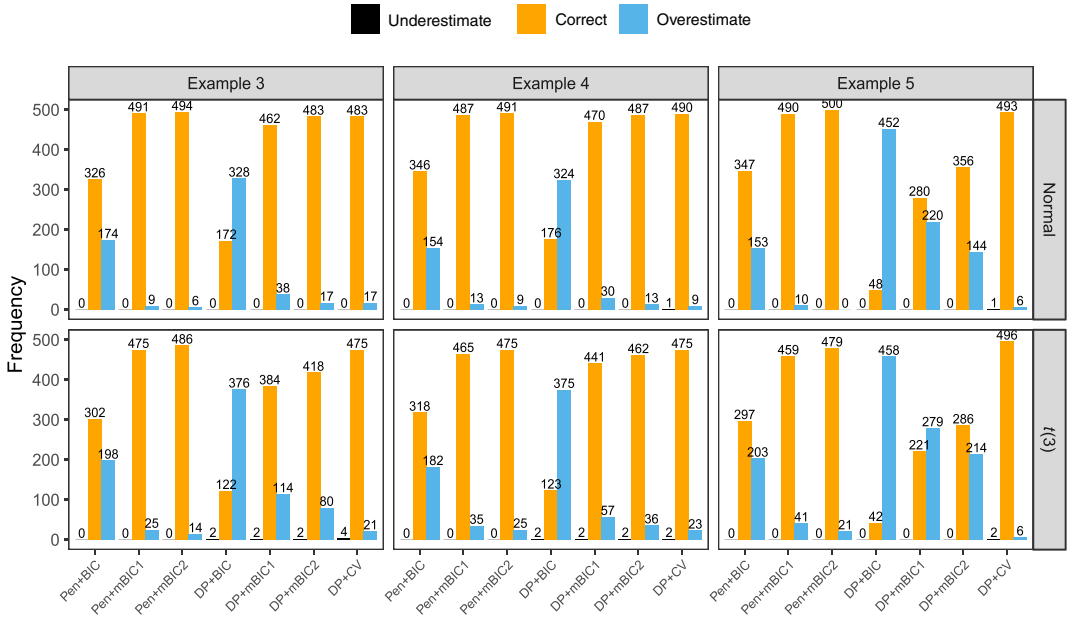


FIGURE 3 The frequency of estimated $K$ under Examples 3–5 for various detection methods.

**TABLE 1** Performance of the estimated coefficients under Examples 3 and 4 using the order-preserved cross-validation (OPCV) criterion in conjunction with the dynamic programming algorithm.

| | Subgroup 1 | | | | Subgroup 2 | | | | Subgroup 3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\beta_{01}$ | $\beta_{02}$ | $\beta_{03}$ | $\beta_{04}$ | $\beta_{11}$ | $\beta_{12}$ | $\beta_{13}$ | $\beta_{14}$ | $\beta_{21}$ | $\beta_{22}$ | $\beta_{23}$ | $\beta_{24}$ |
| Example 3 with $\varepsilon_i \sim N(0, 0.5)$ | | | | | | | | | | | | |
| Bias | 0.004 | 0.001 | −0.003 | −0.003 | −0.009 | 0.012 | 0.005 | −0.001 | −0.011 | 0.000 | 0.006 | 0.003 |
| SD | 0.202 | 0.155 | 0.081 | 0.078 | 0.092 | 0.382 | 0.080 | 0.084 | 0.140 | 0.130 | 0.075 | 0.069 |
| SE | 0.186 | 0.142 | 0.078 | 0.076 | 0.087 | 0.341 | 0.076 | 0.077 | 0.132 | 0.124 | 0.066 | 0.067 |
| CR | 0.932 | 0.932 | 0.921 | 0.937 | 0.946 | 0.917 | 0.930 | 0.919 | 0.932 | 0.908 | 0.905 | 0.939 |
| Example 3 with $\varepsilon_i \sim t(3)/\sqrt{6}$ | | | | | | | | | | | | |
| Bias | −0.022 | −0.007 | −0.005 | −0.002 | −0.010 | −0.003 | 0.002 | −0.002 | −0.013 | −0.002 | −0.004 | 0.007 |
| SD | 0.181 | 0.137 | 0.074 | 0.078 | 0.087 | 0.354 | 0.080 | 0.076 | 0.130 | 0.126 | 0.069 | 0.071 |
| SE | 0.167 | 0.127 | 0.068 | 0.069 | 0.080 | 0.309 | 0.069 | 0.068 | 0.122 | 0.114 | 0.062 | 0.062 |
| CR | 0.924 | 0.945 | 0.940 | 0.933 | 0.931 | 0.924 | 0.929 | 0.927 | 0.940 | 0.952 | 0.945 | 0.945 |
| Example 4 with $\varepsilon_i \sim N(0, 0.5)$ | | | | | | | | | | | | |
| Bias | 0.004 | 0.001 | −0.003 | −0.003 | −0.009 | 0.012 | 0.005 | −0.001 | −0.011 | 0.000 | 0.006 | 0.003 |
| SD | 0.202 | 0.155 | 0.081 | 0.078 | 0.092 | 0.382 | 0.080 | 0.084 | 0.140 | 0.130 | 0.075 | 0.069 |
| SE | 0.186 | 0.142 | 0.078 | 0.076 | 0.087 | 0.341 | 0.076 | 0.077 | 0.132 | 0.124 | 0.066 | 0.067 |
| CR | 0.932 | 0.932 | 0.921 | 0.937 | 0.946 | 0.917 | 0.930 | 0.919 | 0.932 | 0.908 | 0.905 | 0.939 |
| Example 4 with $\varepsilon_i \sim t(3)/\sqrt{6}$ | | | | | | | | | | | | |
| Bias | −0.022 | −0.007 | −0.005 | −0.002 | −0.010 | −0.003 | 0.002 | −0.002 | −0.013 | −0.002 | −0.004 | 0.007 |
| SD | 0.181 | 0.137 | 0.074 | 0.078 | 0.087 | 0.354 | 0.080 | 0.076 | 0.130 | 0.126 | 0.069 | 0.071 |
| SE | 0.167 | 0.127 | 0.068 | 0.069 | 0.080 | 0.309 | 0.069 | 0.068 | 0.122 | 0.114 | 0.062 | 0.062 |
| CR | 0.924 | 0.945 | 0.940 | 0.933 | 0.931 | 0.924 | 0.929 | 0.927 | 0.940 | 0.952 | 0.945 | 0.945 |

Abbreviation: CR, confidence interval.

## 4.3 | Regression coefficient estimation

Finally, we evaluate the estimation performance of the regression coefficients $\widehat{\boldsymbol{\beta}}_k$, when the number of thresholds is correctly estimated. Table 1 summarizes the simulation results for Examples 3 and 4 using the proposed OPCV criterion. Here are the key observations from the simulation results. The biases of estimators (Bias) are minimal, indicating that the proposed coefficient estimators are consistent. The SD are quite close to the estimated SEs, and the coverage rates of the Wald-type confidence intervals (CR) are also close to the nominal level 95%. The simulation results under $N(0, 0.5)$ noise show slightly superior estimation accuracy compared to those under $t(3)/\sqrt{6}$ noise. In summary, the proposed method demonstrates favorable finite sample performance and maintains effectiveness even when the assumption of independence censoring is violated in Example 4.

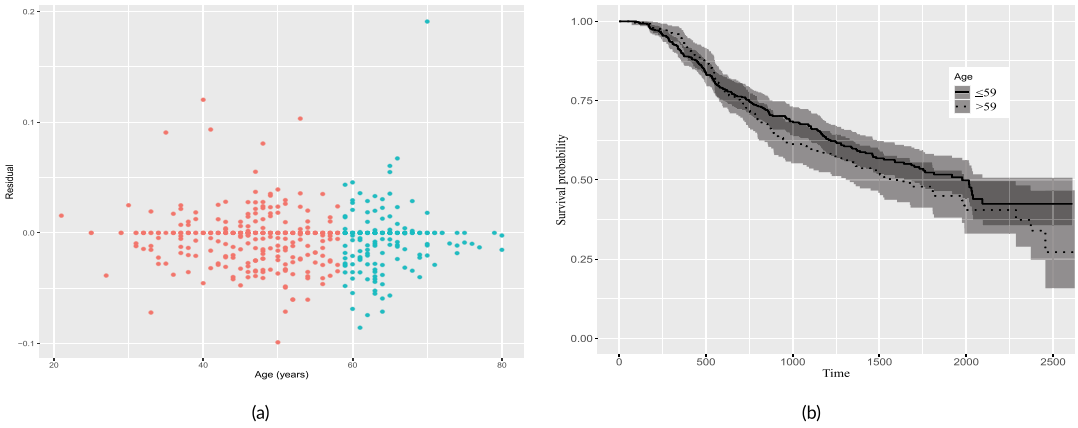


**FIGURE 4** Panel (a) displays the residual from the linear AFT model versus age. Panel (b) displays the Kaplan–Meier curves and 95% confidence bands for German breast cancer data.

## 5 | GERMAN BREAST CANCER DATA ANALYSIS

To demonstrate our methods in practical analysis, we re-visit the German breast cancer dataset (Schmoor et al., 1996), which was collected from July 1984 to December 1989. This dataset comprises 686 patients with primary node-positive breast cancer, 387 of whom experienced right-censored. For breast cancer, it is widely acknowledged that age is a crucial factor that may influence the survival risk. Age at diagnosis sparks debates such as "How young is very young?" and "How old is too old?" Younger patients face distinct concerns such as fertility preservation, pregnancy, lactation, sexual function, and cosmetic outcomes compared to older women. Consequently, a young age at breast cancer diagnosis is considered an independent adverse prognostic factor associated with a higher risk of relapse and mortality. Conversely, increasing age is linked to higher breast cancer-specific mortality in postmenopausal women. Our study aimed to determine optimal cutoff values for categorizing breast cancer patients as "young" or "old." In this analysis, we select age at diagnosis as the threshold variable $Z_i$, ranging from 21 to 80 years, encompassing both younger and older cohorts. The baseline covariates of interest $\mathbf{X}_i$ include age, tumor grade, and a hormonal treatment indicator. Similar to previous studies, we dichotomize the tumor grade variable into grade I and grade II/III, respectively.

We first ignore the potential heterogeneity and fit a linear AFT regression model by using the failure time as the response and the three baseline covariates as predictors. The estimation results are summarized in Table 3, and the residual versus age plot is shown in Figure 4a. By visual inspection, the residuals appear to cluster into two groups with the cutting value being 59. It is reasonable to suspect that there is a threshold effect around $Z_i = 59$, that is, age 59. Given the absence of prior information regarding the number of threshold effects, we employ the score test procedure with 1000 bootstrap replications. The resulting $p$-value is 0.094, indicating statistical significance at a 0.1 significance level for the observed jump threshold effect pattern. Consequently, it is valid to proceed with the estimation of multithreshold accelerated failure time model. The estimated threshold parameters are presented in Table 2, using the method of Li and Jin (2018) with the smoothly clipped absolute deviation (SCAD) and minimax concave (MCP) penalties, along with various criteria in conjunction with the dynamic programming algorithm. We use the dynamic programming algorithm due to its superior estimation accuracy, particularly

**TABLE 2** Threshold detection results of various methods.

| | SCAD | MCP | BIC | mBIC1 | mBIC2 | CV |
|---|---|---|---|---|---|---|
| $\widehat{K}$ | 0 | 0 | 4 | 2 | 1 | 1 |
| $\widehat{\tau}_k$'s | — | — | (46, 51, 57, 61) | (57, 61) | 59 | 59 |

Abbreviations: BIC, Bayesian information criterion; CV, cross-validation.

**TABLE 3** Analysis results including the estimates of regression coefficients (Coef.), the SE and the $p$-values by Wald test for two subgroups divided at $\widehat{\tau} = 59$.

| | Subgroup 1 | | | Subgroup 2 | | | Full sample | | |
|---|---|---|---|---|---|---|---|---|---|
| | Coef. | SE | $p$-Values | Coef. | SE | $p$-Values | Coef. | SE | $p$-Values |
| Intercept | 7.5424 | 0.1820 | 0.0000*** | 8.3127 | 0.5568 | 0.0000*** | 7.4391 | 0.1274 | 0.0000*** |
| Age | −0.0067 | 0.0031 | 0.0340* | −0.0086 | 0.0086 | 0.3180 | 0.0046 | 0.0019 | 0.0177* |
| Treatment | 0.3096 | 0.0560 | 0.0000*** | −0.0206 | 0.0726 | 0.7770 | 0.1185 | 0.0435 | 0.0066** |
| Tumor grade | −0.0696 | 0.1051 | 0.5080 | −0.4250 | 0.0932 | 0.0000*** | −0.4459 | 0.0564 | 0.0000*** |

*Note*: *, ** and *** represent significant at the level 0.05, 0.01, and 0.001, respectively.

for moderate sample sizes. From Table 2, it is evident that the shrinkage method proposed by Li and Jin (2018) does not identify any threshold effect. This is understandable because variable selection techniques such as the SCAD penalty typically requires a stronger signal to select important features. Hence, the Pen+SCAD and Pen+MCP inherit this limitation and fail to identify the underlying threshold information. Interestingly, the BIC yields four estimated threshold parameters, while mBIC1 and mBIC2 result in 2 and 1 threshold effects, respectively. Intuitively, as the magnitude of penalization increases, the number of detected thresholds decreases. Similar to mBIC2, the cross-validation approach detects only one change-point with the same threshold estimate $\widehat{\tau} = 59$. Figure 4b graphically depicts the Kaplan–Meier curves for individuals in the two estimated subgroups ($Z_i \leq 59$ and $Z_i > 59$). We see that younger patients generally have higher survival probabilities than older patients.

To further ascertain the number of threshold effects, we apply the proposed score test for the subgroup 1 (age ≤59) and the subgroup 2 (age >59), respectively. The resulting $p$-values are 0.387 and 0.165, indicating that neither subgroup exhibits a threshold effect. As demonstrated by the simulation, the performance of information criterion is sensitive to the magnitude of penalization and data distribution, making it not inherently a data-driven selector. We substantiate this argument through empirical analysis. We also apply an AFT mixture model (Shokoohi et al., 2019) to estimate the heterogeneous subgroups, which automatically determines the number of components without any configuration adjustments. This approach reveals the presence of two distinct cluster within our dataset, aligning with our earlier observations. Further exploration of the age effect within these subgroups showed a notable gap between clusters among individuals aged 50–60, as depicted in Figure S1 of Appendix S1. The exact age at which this separation occurs is indeterminate due to the soft classification nature of mixture models.

Table 3 summarizes the regression coefficient estimation results. As shown by the table, the age effect is −0.0067 for individuals below the age of 59, and it notably jumps to −0.0086 for those aged 59 and older. Importantly, the former parameter estimate is statistically significant at the 0.05 significance level. The hormonal treatment shows a substantial positive effect on survival time, but this positive relationship becomes statistically insignificant after the estimated

change-point. Tumor grade, on the other hand, shows no significant association within subgroup 1 but exhibits a strong negative effect on the status of patients when the age exceeds the threshold of 59 years. However, if we ignore the threshold effect, the heterogeneous effects of covariates will be missed. From Table 3, the age and hormonal treatment show no significant impact on survival time in subgroup 2, but show a positive relationship when considering the entire sample. Notably, the tumor grade exerts a pronounced negative effect on survival time in the full sample, whereas in subgroup 1, this negative effect is less prominent. This emphasizes that disregarding the threshold effect in regression analysis may overlook heterogeneous information and potentially lead to misleading conclusions.

## 6 | CONCLUDING REMARKS

In this paper, we first show the selection consistency of the modified Bayesian information criterion to select the number of thresholds within the framework of the multithreshold accelerated failure time model. However, its empirical performance generally depends on the choice of its penalization magnitude, which usually varies with model setting and error distribution. To address this variability, we propose a new OPCV criterion based on data splitting inspired by Zou et al. (2020). The OPCV can not only yield a consistent estimation for the number of thresholds but also make the estimation procedure data driven and more convenient for practical use. The work done here stems from Li and Jin (2018) but has made sufficient progress beyond their work.

There are still several issues that warrant further exploration in future research. First, it is beneficial to consider the interval censored failure time as it becomes increasingly common in medical studies. The likelihood-based objective function in Lu (2008) can be adapted to facilitate the relevant inference. Moreover, along this line, our framework can be directly extended to deal with many other survival models with threshold structure, such as the Cox proportional hazards model, cure rate model, and others. But the specific cost function and estimation routine may incur technical difficulties. Second, the DP algorithm may be time consuming or even infeasible from computational considerations for the dataset with a large sample size. In such scenarios, the WBS algorithm and its variants are promising to find an approximate solution at the cost of certain loss of estimation precision. Additionally, it is also of interest to exploit new algorithms to find the change-points exactly in a fast way. Third, the threshold models depict the nonlinear pattern by introducing a piece-wise linear regression form. As an alternative, the varying coefficient model is another important tool to model for nonlinearity by assuming the regression coefficients to be a smooth function, see Chen et al. (2012), Lin, Fei, and Li (2016), and Lin, Tan, and Li (2016). We acknowledge that these models may perform well sometimes, especially in capturing the local effect change, but its performance heavily depends on the sample sizes due to its nonparametric nature. Conducting more comparative studies is necessary to illustrate the usefulness of our threshold models. These directions warrant future research.

### ACKNOWLEDGMENTS

We thank the Editor, the Associate Editor and referees for their encouragements and insightful comments which have substantially improved the paper. Zhong also thanks the supports of Fujian Key Lab of Statistics, Fujian Key lab of Digital Finance.

### ENDNOTE

[1]The structural break models can be regarded as a specialized form of threshold regression, whose threshold variable is the time index.

## ORCID

*Hao Zeng* https://orcid.org/0000-0002-0084-5367
*Wei Zhong* https://orcid.org/0000-0002-8868-4069


## REFERENCES


Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79.

Arlot, S., & Celisse, A. (2011). Segmentation of the mean of heteroscedastic data via cross-validation. *Statistics and Computing*, *21*, 613–632.

Aue, A., Cheung, R. C., Lee, T. C., & Zhong, M. (2014). Segmented model selection in quantile regression using the minimum description length principle. *Journal of the American Statistical Association*, *109*, 1241–1256.

Auger, I. E., & Lawrence, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, *51*, 39–54.

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

Chan, K.-S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *The Annals of Statistics*, *21*, 520–533.

Chen, K., Lin, H., & Zhou, Y. (2012). Efficient estimation for the cox model with varying coefficients. *Biometrika*, *99*, 379–392.

Dette, H., & Gösmann, J. (2020). A likelihood ratio approach to sequential change point detection for a general class of parameters. *Journal of the American Statistical Association*, *115*, 1361–1377.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

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

Gierz, K., & Park, K. (2022). Detection of multiple change points in a weibull accelerated failure time model using sequential testing. *Biometrical Journal*, *64*, 617–634.

Gonzalo, J., & Pitarakis, J.-Y. (2002). Estimation and model selection based inference in single and multiple threshold models. *Journal of Econometrics*, *110*, 319–352.

Hall, A. R., Han, S., & Boldea, O. (2012). Inference regarding multiple structural changes in linear models with endogenous regressors. *Journal of Econometrics*, *170*, 281–302.

Hall, A. R., Osborn, D. R., & Sakkas, N. (2015). Structural break inference using information criteria in models estimated by two-stage least squares. *Journal of Time Series Analysis*, *36*, 741–762.

Hannart, A., & Naveau, P. (2012). An improved bayesian information criterion for multiple change-point models. *Technometrics*, *54*, 256–268.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica*, *68*, 575–603.

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

He, P., Fang, L., & Su, Z. (2013). A sequential testing approach to detecting multiple change points in the proportional hazards model. *Statistics in Medicine*, *32*, 1239–1245.

He, Y., Lin, H., & Tu, D. (2018). A single-index threshold cox proportional hazard model for identifying a treatment-sensitive subset based on multiple biomarkers. *Statistics in Medicine*, *37*, 3267–3279.

He, Y., Zhou, L., Xia, Y., & Lin, H. (2023). Center-augmented $\ell 2$-type regularization for subgroup learning. *Biometrics*, *79*, 2157–2170.

Kosorok, M. R., & Song, R. (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *The Annals of Statistics*, *35*, 957–989.

Lavielle, M., & Moulines, É. (2000). Least-squares estimation of an unknown number of shifts in a time series. *Journal of Time Series Analysis*, *21*, 33–59.

Lebarbier, É. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, *85*, 717–736.

Lee, C. Y., Chen, X., & Lam, K. F. (2020). Testing for change-point in the covariate effects based on the cox regression model. *Statistics in Medicine*, *39*, 1473–1488.

Lee, C. Y., & Lam, K. (2020). Survival analysis with change-points in covariate effects. *Statistical Methods in Medical Research*, *29*, 3235–3248.

Lee, S., & Seo, M. H. (2008). Semiparametric estimation of a binary response model with a change-point due to a covariate threshold. *Journal of Econometrics*, *144*, 492–499.

Lee, S., Seo, M. H., & Shin, Y. (2011). Testing for threshold effects in regression models. *Journal of the American Statistical Association*, *106*, 220–231.

Li, J., & Jin, B. (2018). Multi-threshold accelerated failure time model. *The Annals of Statistics*, *46*, 2657–2682.

Li, J., Li, Y., Jin, B., & Kosorok, M. R. (2021). Multithreshold change plane model: Estimation theory and applications in subgroup identification. *Statistics in Medicine*, *40*, 3440–3459.

Li, J., & Zhang, W. (2011). A semiparametric threshold model for censored longitudinal data analysis. *Journal of the American Statistical Association*, *106*, 685–696.

Lin, H., Fei, Z., & Li, Y. (2016). A semiparametrically efficient estimator of the time-varying effects for survival data with time-dependent treatment. *Scandinavian Journal of Statistics*, *43*, 649–663.

Lin, H., Tan, M. T., & Li, Y. (2016). A semiparametrically efficient estimator of single-index varying coefficient cox proportional hazards models. *Statistica Sinica*, *26*(2), 779–807.

Liu, J., Wu, S., & Zidek, J. V. (1997). On segmented multivariate regression. *Statistica Sinica*, *7*, 497–525.

Lu, W. (2008). Maximum likelihood estimation in the proportional hazards cure model. *Annals of the Institute of Statistical Mathematics*, *60*, 545–574.

Ninomiya, Y. (2015). Change-point model selection via aic. *Annals of the Institute of Statistical Mathematics*, *67*, 943–961.

Qu, Z. (2008). Testing for structural change in regression quantiles. *Journal of Econometrics*, *146*, 170–184.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, *24*, 220–238.

Schmoor, C., Olschewski, M., & Schumacher, M. (1996). Randomized and non-randomized patients in clinical trials: Experiences with comprehensive cohort studies. *Statistics in Medicine*, *15*, 263–271.

Shokoohi, F., Khalili, A., Asgharian, M., & Lin, S. (2019). Capturing heterogeneity of covariate effects in hidden subpopulations in the presence of censoring and large number of covariates. *The Annals of Applied Statistics*, *13*, 444–465.

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis*, *45*, 89–103.

Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics*, *23*, 461–471.

Stute, W., & Wang, J.-L. (1993). The strong law under random censorship. *The Annals of Statistics*, *21*, 1591–1607.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*, 267–288.

Wan, C., Zhong, W., Zhang, W., & Zou, C. (2023). Multikink quantile regression for longitudinal data with application to progesterone data analysis. *Biometrics*, *79*(2), 747–760.

Wang, B., Li, J., & Wang, X. (2022). Multi-threshold proportional hazards model and subgroup identification. *Statistics in Medicine*, *41*, 5715–5737.

Wang, G., & Zou, C. (2023). Cpss: An r package for change-point detection by sample-splitting methods. *Journal of Quality Technology*, *55*, 61–74.

Wang, H., Li, B., & Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *71*, 671–683.

Wang, J., & Li, J. (2023). Multi-threshold structural equation model. *Journal of Business & Economic Statistics*, *41*, 377–387.

Wang, J., Li, J., Li, Y., & Wong, W. K. (2019). A model-based multithreshold method for subgroup identification. *Statistics in Medicine*, *38*, 2605–2631.

Wei, K., Zhu, H., Qin, G., Zhu, Z., & Tu, D. (2022). Multiply robust subgroup analysis based on a single-index threshold linear marginal model for longitudinal data with dropouts. *Statistics in Medicine*, *41*, 2822–2839.

Xu, G., Sen, B., & Ying, Z. (2014). Bootstrapping a change-point cox model for survival data. *Electronic Journal of Statistics*, *8*, 1345–1379.

Yan, X., Yin, G., & Zhao, X. (2021). Subgroup analysis in censored linear regression. *Statistica Sinica*, *31*, 1027–1054.

Yu, P. (2015). Adaptive estimation of the threshold point in threshold regression. *Journal of Econometrics*, *189*, 83–100.

Yu, P., & Phillips, P. C. (2018). Threshold regression asymptotics: From the compound poisson process to two-sided Brownian motion. *Economics Letters*, *172*, 123–126.

Yu, P., & Zhao, Y. (2013). Asymptotics for threshold regression under general conditions. *The Econometrics Journal*, *16*, 430–462.

Zhang, N. R., & Siegmund, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, *63*, 22–32.

Zhong, W., Wan, C., & Zhang, W. (2022). Estimation and inference for multi-kink quantile regression. *Journal of Business & Economic Statistics*, *40*, 1123–1139.

Zou, C., Wang, G., & Li, R. (2020). Consistent selection of the number of change-points via sample-splitting. *The Annals of Statistics*, *48*, 413–439.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Wan, C., Zeng, H., Zhang, W., Zhong, W., & Zou, C. (2025). Data-driven estimation for multithreshold accelerated failure time model. *Scandinavian Journal of Statistics*, *52*(1), 447–468. https://doi.org/10.1111/sjos.12758