DSSR

Discussion Paper No. 151

A Finite-Horizon Mixture Cure Model with Application to Online Flea Market Data

Yuji Komiyama, Yasumasa Matsuda
and Masakazu Ishihara

May, 2026

Data Science and Service Research
Discussion Paper

Center for Data Science and Service Research
Graduate School of Economic and Management
Tohoku University
27-1 Kawauchi, Aobaku
Sendai 980-8576, JAPAN

A Finite-Horizon Mixture Cure Model with Application to Online Flea Market Data

Yuji Komiyama[1,*] Yasumasa Matsuda[2] Masakazu Ishihara[2,1]

[1]Graduate School of Economics and Management, Tohoku University

[2]Leonard N. Stern School of Business, New York University

[*]Corresponding author: komiyama.yuji.r2@dc.tohoku.ac.jp

Abstract

This study proposes a mixture cure model that latently divides a population based on event occurrence within a finite time horizon. Conventional models rely on event occurrence over an infinite horizon, introducing untestable assumptions that often lead to issues with identifiability and interpretability. By shifting the estimand to a specific period of interest, the proposed approach reduces reliance on these infinite-tail assumptions and aligns interpretations more closely with finite-horizon decision-making objectives. Through simulation studies, we first evaluate the statistical properties of the proposed estimator, including estimation bias and variance. We further show that relying on conventional infinite-horizon models for finite-horizon decision-making can lead to erroneous judgments. Finally, we apply the model to transaction data from Mercari, a Japanese online flea market platform. The empirical results reveal that the proposed model identifies different significant variables compared to the conventional model, offering interpretations that better reflect seasonal variation in user behavior.

Keywords: survival analysis; mixture cure model; decision-making; online flea market.

1 Introduction

In survival analysis, it is usually assumed that if a sufficiently long observation period is ensured for all study subjects, each individual will eventually experience the event of interest. Let T denote the random variable representing the survival time and $F(t)$ its distribution function. The survival function, defined as the probability that the event has not occurred by time t, is given by

$$S(t) = 1 - F(t). \tag{1}$$

Standard survival models assume $\lim_{t\to\infty} S(t) = 0$. In practice, however, not all individuals necessarily experience the event. For example, in medical research where survival time is defined as the time from surgery to recurrence, some patients may be completely cured and never experience recurrence. Similarly, in economics, when survival time is defined as the time from job separation to reemployment, not everyone necessarily finds a new job. To address such situations in which not all individuals experience the event, cure models were introduced, allowing $\lim_{t\to\infty} S(t) > 0$. Individuals who never experience the event are regarded as *cured*.

Cure models can be broadly classified into mixture cure models and promotion time cure models. A mixture cure model assumes that the population consists of two latent subgroups: *a cured group* that will never experience the event, and *a susceptible group* that will eventually experience it. A latent variable representing group membership is introduced. Let $S_\infty(t) = \mathbb{P}(T > t \mid T < \infty)$ denote the survival function of the susceptible group and $1 - \pi_\infty = \mathbb{P}(T = \infty)$ the cure rate. The population survival function $S_{\text{pop}}(t)$ is then

$$S_{\text{pop}}(t) = (1 - \pi_\infty) + \pi_\infty S_\infty(t), \tag{2}$$

where $\lim_{t\to\infty} S_\infty(t) = 0$ is assumed. In contrast, the promotion time cure model proposed by Yakovlev et al. (1996) is based on biological mechanisms and has a structure in which the same parameters determine both the cure rate and survival, making it impossible to separate these effects. For interpretability, the present study focuses on the mixture cure model. By allowing these quantities to depend on covariates x, the model separates two components of covariate effects: the *incidence* component, which models whether an individual belongs to the susceptible group through $\pi_\infty(x) = \mathbb{P}(T < \infty \mid x)$, and the *latency* component, which models the event-time distribution conditional on being susceptible through $S_\infty(t \mid x) = \mathbb{P}(T > t \mid x, T < \infty)$. Both components are defined in terms of whether the event ultimately occurs over an infinite time horizon.

In many applied settings, however, the question of whether the event *ultimately* occurs is not the relevant question; decisions are instead governed by finite time horizons. A retailer deciding whether to discount or remove an unsold product needs to know whether it will sell within the next few months, not whether it would eventually sell given infinite time. A clinician evaluating a treatment is concerned with whether relapse occurs within a specific follow-up window. In each case, the finite-horizon answer is what drives the decision, and the infinite-horizon answer may be neither available nor necessary. For a given finite horizon $c > 0$, this question—whether the event occurs by time c, i.e., $\mathbb{P}(T < c \mid x)$—is a fundamentally different estimand from the infinite-horizon counterpart $\mathbb{P}(T < \infty \mid x)$, and the two need not agree. A covariate that reduces the ultimate event probability $\mathbb{P}(T < \infty \mid x)$ may simultaneously increase the short-term event probability $\mathbb{P}(T < c \mid x)$ for a given horizon c. Consequently, adopting an infinite-horizon model to inform finite-horizon decisions can lead to erroneous conclusions, yet this mismatch has received little attention in the literature.

A concrete instance of this mismatch arises in consumer-to-consumer (C2C) online marketplaces, where the seller is an individual consumer rather than a retailer and must physically store the listed item at home until the transaction is completed. The seller therefore bears holding costs—storage-space occupation, ongoing management effort, and the opportunity cost of alternative disposal channels such as secondhand shops, donation, or discarding—that accumulate with elapsed time, so the seller's welfare depends not on whether the item ever sells but on whether it sells within a period that the seller can tolerate holding it. The substantive estimand is then $\mathbb{P}(T < c \mid x)$ for a finite c rather than $\mathbb{P}(T < \infty \mid x)$, making this setting a canonical case in which the finite-horizon framework developed in this study is required rather than merely convenient. We analyze this case in Section 4 using transaction data from Mercari (Mercari, Inc., 2023), one of the largest C2C online flea market platforms in Japan.

To address this gap, we shift the mixture cure framework itself—latent binary classification of the population with separate incidence and latency components—to a prespecified finite time horizon $[0, c)$, where c is chosen by the analyst. The contributions of this study are threefold. First,

we develop a finite-horizon mixture cure framework in which the incidence component estimates $\mathbb{P}(T < c \mid x)$ and the latency component models the conditional event-time distribution on $[0, c)$. This framework enables joint estimation and interpretation of the probability of event occurrence by time c and the conditional timing of the event within $[0, c)$, thereby providing a model suited to finite-horizon decision-making. Second, by restricting the analysis to $[0, c)$, the model avoids reliance on assumptions about tail behavior beyond c: the event-time distribution on (c, ∞) is left unspecified, and the possibility that the event occurs after c is not ruled out. Incidence and latency interpretations therefore concern only behavior on $[0, c)$ and are explicitly contingent on the analyst's choice of c. Third, through simulation studies we demonstrate that conventional infinite-horizon models can yield sign-reversed covariate effects relative to the finite-horizon truth, producing misleading guidance for finite-horizon decisions.

There exist numerous prior studies on mixture cure models. The mixture cure model was first proposed by Boag (1949); Berkson and Gage (1952), who considered models without covariates. Later, Farewell (1977) modeled the probability of not being cured using logistic regression. Furthermore, Kuk and Chen (1992) extended the approach to allow the probability of not being cured to depend on covariates via logistic regression, and modeled the survival function for susceptible individuals with the Cox proportional hazards model (Cox, 1972). Subsequent studies further developed estimation methods (Peng and Dear, 2000; Sy and Taylor, 2000). There are also extensions incorporating accelerated failure time (AFT) models (Li and Taylor, 2002) and nonparametric approaches (López-Cheda et al., 2017). This model is also extended to the economic field (Dirick et al., 2019) and the marketing field (Kumar et al., 2018), among others.

To address the lack of information at infinite time, Taylor (1995) introduced an assumption called the zero-tail constraint. Under this assumption, all censored observations with survival times longer than the maximum observed event time are treated as cured individuals. In other words, it is assumed that the follow-up period is sufficiently long. This assumption is reasonable when the survival function becomes flat at a non-zero value beyond a certain point, but in other cases it may lead to overestimation of the cure rate. To address this problem, Peng (2003) proposed a model that, while expressing the survival function for susceptible individuals with the Cox proportional hazards model, introduces a decaying baseline survival function such as the Weibull or exponential distribution beyond the maximum observed event time. This approach aims to mitigate the influence of the zero-tail constraint. However, this method still relies on the assumption that survival times follow a certain decaying distribution, and does not fundamentally resolve identifiability issues. In recent years, various attempts have been made to relax this assumption. For example, to test whether truly cured individuals exist in the population when follow-up is limited, Escobar-Bach and Keilegom (2019) proposed a new estimator for the cure rate using extrapolation techniques from extreme value theory. Tests for the sufficiency of follow-up duration (Ping Xie et al., 2024), and nonparametric estimation methods based on extreme value theory (Beirlant et al., 2026), have also been developed. Additionally, Safari et al. (2023) showed that when some cured individuals are known in advance (e.g., in settings where a medical test definitively verifies cure), including this information in estimation asymptotically reduces the variance of estimators. Compared to these previous works, the present study differs in that it shifts the mixture cure framework itself to the finite interval $[0, c)$. Rather than attempting to recover infinite-horizon quantities under limited follow-up, our model directly targets finite-horizon estimands—the probability of event occurrence

by time c and the conditional event-time distribution on $[0, c)$—ensuring identifiability through the boundary condition without reliance on assumptions about tail behavior beyond c. Accordingly, estimation of the cure rate is not the main objective in our model.

Literature that critically examines the practical interpretation of mixture cure models remains scarce. While Amico and Keilegom (2018) notes that "Indeed, the incidence models the long-term effect of covariates on the cure status, which is something permanent, whereas the latency focuses on the short-term, time-dependent effect that only concerns uncured observations," such discussions are largely confined to the infinite-horizon framework. If we analyze survival time for which longer duration is preferable (such as the time to recurrent cancer), the conventional interpretation may be useful. However, if we analyze the survival time for which shorter duration is preferable (such as the time to sale of a product), the conventional interpretation may not be useful. To the best of our knowledge, no existing studies have problematized the inconsistency between these conventional interpretations and the practical needs of finite-horizon decision-making. By addressing this gap, the proposed model offers a more practical interpretation than conventional models, particularly in scenarios where cured individuals exist but the primary analytical focus is on whether the event occurs within a finite time horizon.

The remainder of this paper is organized as follows. Section 2 introduces the proposed mixture cure model, providing mathematical definitions and justifications for the modeling framework. In Section 3, we evaluate the desirable statistical properties of the proposed model, such as estimation bias and the standard deviation of estimates, using simulated data. Furthermore, we demonstrate that relying on conventional models—which divide the population based on infinite-horizon event occurrence—can lead to erroneous judgments in finite-horizon decision-making. Section 4 presents an application to transaction data from Mercari to analyze user behavior. The empirical analysis highlights that the proposed model offers more practical and intuitive interpretations, especially regarding the seasonality of products, and reveals differences in findings compared to conventional mixture cure models. Finally, Section 5 summarizes the main findings and discusses their implications.

2 Methodology

In this section, we first define the mathematical notation required for the discussion of survival analysis. Then, we outline the proposed model, followed by a discussion on the estimation method.

2.1 Preliminaries

For each individual $i = 1, \ldots, N$, let T_i be the true survival time, C_i be the right-censoring time, $t_i = \min(T_i, C_i)$ be the observed time, and $\delta_i = \mathbf{1}(T_i \leq C_i)$ be the censoring indicator. Here, $\mathbf{1}(\cdot)$ is the indicator function. Let T be the random variable representing the survival time, with its marginal distribution function $F(t)$ and survival function $S(t) = 1 - F(t)$. Let $x_i = (1, \tilde{x}_i) \in \mathcal{X} \subset \mathbb{R}^{p+1}$ be the covariate vector corresponding to each individual. Here, 1 is the constant term, $\tilde{x}_i \in \mathbb{R}^p$ is the covariate vector excluding the constant term, and $\mathcal{X}$ represents the covariate vector space. Let $F(t \mid x)$ be the distribution function of the survival time conditional on x, and $S(t \mid x) = 1 - F(t \mid x)$ be the conditional survival function. Furthermore, we assume non-informative censoring, $T_i \perp C_i \mid x_i$. Thus, the observed data can be expressed as $\mathcal{D} = \{(x_i, \tilde{x}_i, t_i, \delta_i)\}_{i=1}^N$.

In the conventional mixture cure model (Sy and Taylor, 2000), it is assumed that the population is latently divided into two latent subgroups: a cured group in which the event will never occur in the future, and a susceptible group in which the event will occur at some point up to infinity. The population survival function S_{pop} is expressed as:

$$S_{\text{pop}}(t \mid x) = (1 - \pi_\infty(x)) + \pi_\infty(x) S_\infty(t \mid x), \quad t \in [0, \infty) \tag{3}$$

Here, $\pi_\infty(x) \in (0, 1)$, and it is assumed that $\lim_{t\to\infty} S_\infty(t \mid x) = 0$. In this case, the cure rate is defined as $\lim_{t\to\infty} S_{\text{pop}}(t \mid x) = 1 - \pi_\infty(x) = \mathbb{P}(T = \infty \mid X = x)$. The cure rate represents the probability that an individual with covariate x will never experience the event in the future. $S_\infty(t \mid x) = \mathbb{P}(T > t \mid X = x, T < \infty)$ is the survival function of the susceptible group. In conventional models (Sy and Taylor, 2000; Peng and Dear, 2000), it is common to assume a logistic regression model for $\pi_\infty(x)$ and a Cox proportional hazards model for $S_\infty(t \mid x)$.

2.2 Proposed Model

In this study, we classify the population based on the occurrence of the event at a finite time point $c \in (0, \infty)$, rather than whether the event occurs in infinite time. Here, c is a prespecified value chosen by the analyst according to the decision-relevant time horizon, and the interpretation of the model differs depending on the chosen value of c. Specifically, the subpopulation in which the event does not occur by $t = c$ is regarded as *the event-free group up to time c*, and the subpopulation in which the event occurs by $t = c$ is regarded as *the event group up to time c*. The population survival function S_{pop} is expressed as:

$$S_{\text{pop}}(t \mid x) = (1 - \pi_c(x)) + \pi_c(x) S_c(t \mid x), \quad t \in [0, c) \tag{4}$$

Here, $\pi_c(x) \in (0, 1)$. The survival function $S_c(t \mid x) = \mathbb{P}(T > t \mid X = x, T < c)$ of the event group up to time c is defined on $t \in [0, c)$, and we impose $\lim_{t\to c-} S_c(t \mid x) = 0$ so that $\lim_{t\to c-} S_{\text{pop}}(t \mid x) = 1 - \pi_c(x) = \mathbb{P}(T \geq c \mid X = x)$ holds. $1 - \pi_c(x)$ represents the probability that an individual with covariate x will not experience the event by $t = c$. It should be noted that the domain of the model changes between Equation (3) and Equation (4). Since the proposed model restricts the domain to $t \in [0, c)$, no assumptions are made for $t \geq c$. Therefore, the possibility of the event occurring after c is allowed, and individuals who actually experience the event after c are classified into the event-free group up to time c. By restricting the analysis target to an observable and meaningful range in this way, clear estimation within the finite time horizon $[0, c)$ becomes possible without relying on the assumption of infinite time. As a result, practical interpretations, such as measuring the probability of event occurrence within period c and the effect on the time to the event, become easier.

Let f_c be the probability density function of the event group up to time c, and f_{0c} be the baseline probability density function when $\tilde{x} = 0$. Following Sy and Taylor (2000), we propose a model that assumes proportional hazards for the survival function of the event group up to time c. That is, we assume that the survival function $S_c(t \mid \tilde{x})$ of the event group up to time c is expressed in a Cox proportional hazards form:

$$h_c(t \mid \tilde{x}; \beta) = h_{0c}(t) \exp(\tilde{x}^\top \beta) \tag{5}$$

Here, $h_c(t \mid \tilde{x}; \beta) = f_c(t \mid \tilde{x}; \beta)/S_c(t \mid \tilde{x}; \beta)$ is the hazard function, $h_{0c}(t) = f_{0c}(t)/S_{0c}(t)$ is the baseline hazard function, and $\exp(\tilde{x}^\top \beta)$ describes the effect of the covariate $\tilde{x}$. Note that to ensure the identifiability of the baseline survival function, the covariate vector $\tilde{x}$ does not include a constant term. Since $S_c(t \mid \tilde{x}; \beta) = \exp\left(-\int_0^t h_c(u \mid \tilde{x}; \beta)\, du\right)$, we can rewrite Equation (5) as:

$$S_c(t \mid \tilde{x}; \beta) = S_{0c}(t)^{\exp(\tilde{x}^\top \beta)}$$

where $S_{0c}(t) = \exp\left(-\int_0^t h_{0c}(u)\, du\right)$ is the baseline survival function.

Because the conditional event-time distribution of the event group up to time c is supported on the finite interval $[0, c)$, the boundary condition $\lim_{t \to c-} S_{0c}(t) = 0$ must hold, which in turn forces $\lim_{t \to c-} h_{0c}(t) = \infty$. Directly modeling such a divergent baseline hazard is difficult. We therefore take the baseline density f_{0c} as the primary object of nonparametric modeling on $[0, c)$, and induce the baseline survival function and the baseline hazard from it via

$$S_{0c}(t) = 1 - \int_0^t f_{0c}(u)\, du, \qquad h_{0c}(t) = \frac{f_{0c}(t)}{S_{0c}(t)}.$$

Since f_{0c} is a probability density on $[0, c)$ by construction, the boundary condition $\lim_{t \to c-} S_{0c}(t) = 0$ holds by definition.

A key advantage of this density-based construction is that the standard Cox interpretation of the regression coefficient β is preserved on $[0, c)$. Although the induced $h_{0c}(t)$, and hence $h_c(t \mid \tilde{x}; \beta)$, diverges as $t \to c-$, the hazard ratio between any two covariate values $\tilde{x}_1$ and $\tilde{x}_2$ is

$$\frac{h_c(t \mid \tilde{x}_1; \beta)}{h_c(t \mid \tilde{x}_2; \beta)} = \exp\left((\tilde{x}_1 - \tilde{x}_2)^\top \beta\right),$$

which is finite and time-independent because the divergent baseline cancels in the ratio. Hence β retains exactly the same interpretation as in the standard Cox proportional hazards regression (Cox, 1972), and the proposed model reconciles the boundary divergence required by the finite-horizon constraint with the familiar covariate-effect interpretation of the Cox model.

To realize this design, we represent the baseline density f_{0c} as a mixture of normalized cubic B-spline basis functions $\{\tilde{B}_{i,3}(t)\}_{i=1}^K$. Following de Boor (1978), let $B_{i,3}(t)$ denote the standard cubic B-spline basis function defined by a knot sequence. We define the normalized basis as

$$\tilde{B}_{i,3}(t) = \frac{B_{i,3}(t)}{\int_0^c B_{i,3}(u)\, du}$$

so that $\int_0^c \tilde{B}_{i,3}(t)\, dt = 1$ holds. Cubic B-splines have the property of being able to uniformly approximate smooth curves if the knots are appropriately set. Moreover, since $B_{i,3}(t)$ is a piecewise polynomial, the integral $\int_0^t \tilde{B}_{i,3}(u)\, du$ in (6) can be written in closed form, which facilitates efficient computation.

Concretely, we define the baseline density on $[0, c)$ as the linear combination

$$f_{0c}(t; \alpha) = \sum_{i=1}^K \gamma_i(\alpha) \tilde{B}_{i,3}(t),$$
$$\gamma_i(\alpha) = \frac{\exp(\alpha_i)}{\sum_{j=1}^K \exp(\alpha_j)}, \quad \alpha_K = 0$$

Here, $\alpha = (\alpha_1, \ldots, \alpha_K)$ (where $\alpha_K = 0$ is fixed to avoid unidentifiability arising from the shift invariance of the softmax function), $\gamma_i(\alpha) > 0$ for all i (since the softmax function is strictly positive), and $\sum_{i=1}^{K} \gamma_i(\alpha) = 1$. Thus $f_{0c}(\cdot; \alpha)$ is a valid probability density on $[0, c)$ with $\lim_{t \to c-} \int_0^t f_{0c}(u; \alpha)\, du = 1$. The baseline survival function induced from f_{0c} is

$$S_{0c}(t; \alpha) = 1 - \sum_{i=1}^{K} \gamma_i(\alpha) \int_0^t \tilde{B}_{i,3}(u)\, du \tag{6}$$

for $t \in [0, c)$. Then $S_{0c}(t; \alpha) > 0$ for all $t \in [0, c)$ and $\lim_{t \to c-} S_{0c}(t; \alpha) = 0$. This ensures that the Cox-type density

$$\begin{aligned} f_c(t \mid \tilde{x}; \beta, \alpha) &= -\frac{d}{dt} S_c(t \mid \tilde{x}; \beta, \alpha) \\ &= \exp(\tilde{x}^\top \beta) f_{0c}(t; \alpha) S_{0c}(t; \alpha)^{\exp(\tilde{x}^\top \beta) - 1} \end{aligned}$$

is well-defined on $[0, c)$; in particular, when $\exp(\tilde{x}^\top \beta) - 1 < 0$, the factor $S_{0c}(t; \alpha)^{\exp(\tilde{x}^\top \beta) - 1}$ remains finite because $S_{0c}(t; \alpha) > 0$ on the defined interval. The knot sequence fixes the endpoints of $[0, c]$, extends with equally spaced knots outside the interval, and places internal knots at the empirical quantiles of the observed event times. The resulting latency model can be viewed as a sieve-based proportional hazards model (Cox, 1972) on the finite interval $[0, c)$: the Cox proportional hazards structure governs covariate effects, while the baseline density is approximated nonparametrically as a B-spline mixture, from which the baseline survival and hazard functions are induced.

We assume a logistic regression model for $\pi_c(x)$:

$$\pi_c(x; b) = \frac{1}{1 + \exp(-x^\top b)} \tag{7}$$

Here, $b = (b_1, \ldots, b_{p+1})$ is the parameter vector of the logistic regression model. Therefore, the parameters to be estimated are (b, β, α), and the total number is $2p + K (= (p + 1) + p + (K - 1))$.

2.3 Estimation Method

Following Sy and Taylor (2000), we introduce a latent variable z_i representing membership in the event-free group up to time c or the event group up to time c, and perform parameter estimation using the EM algorithm. Here, if $z_i = 0$, the i-th individual belongs to the event-free group up to time c, and if $z_i = 1$, they belong to the event group up to time c. Therefore, each data point is represented by $\{(\tilde{x}_i, x_i, t_i, \delta_i, z_i)\}_{i=1}^{N}$. When $t_i < c$, if $\delta_i = 1$, the individual belongs to the event group up to time c, and if $\delta_i = 0$, they belong to either the event-free group up to time c or the event group up to time c. On the other hand, when $t_i \geq c$, all are considered to belong to the event-free group up to time c. Based on the above, the observed log-likelihood based on the observed data is

expressed as:

$$\begin{aligned}\ell(b,\beta,\alpha) = &\sum_{i:t_i<c} \delta_i \log\left(\pi_c(x_i;b) f_c(t_i \mid \tilde{x}_i;\beta,\alpha)\right) \\ &+ \sum_{i:t_i<c} (1-\delta_i) \log\left((1-\pi_c(x_i;b)) + \pi_c(x_i;b) S_c(t_i \mid \tilde{x}_i;\beta,\alpha)\right) \\ &+ \sum_{i:t_i\geq c} \log(1-\pi_c(x_i;b))\end{aligned}$$

In the softmax function used in the proposed model, there is near-unidentifiability with respect to $\{\alpha_i\}_{i=1}^{K-1}$. In other words, when a certain α_i takes a large value, slightly changing α_j $(j \neq i)$ may hardly change the value of $\ell(b,\beta,\alpha)$. As a result, the negative Hessian matrix of $\ell(b,\beta,\alpha)$ may not be a positive definite matrix. Therefore, estimation by Newton's method can become unstable. To prevent this, we specify a prior distribution for $\{\alpha_i\}_{i=1}^{K-1}$. Namely, we assume a normal distribution $\alpha \sim \mathcal{N}(0, \lambda^{-1} I_{K-1})$. Here, λ is a regularization parameter (hyperparameter). This corresponds to L2 regularization. We consider MAP estimation.

$$\begin{aligned}\hat{\theta}_{\mathrm{MAP}} &= \operatorname*{argmax}_{\theta} p(\theta \mid \mathcal{D}) \\ &= \operatorname*{argmax}_{\theta} \left(\log p(\mathcal{D} \mid \theta) + \log p(\theta|\lambda)\right) \\ &= \operatorname*{argmax}_{\theta} \left(\ell(b,\beta,\alpha) - \frac{\lambda}{2}||\alpha||^2\right)\end{aligned}$$

The complete-data log-likelihood is

$$\begin{aligned}\ell_C(b,\beta,\alpha;z) &= \ell_1(b;z) + \ell_2(\beta,\alpha;z), \qquad (8)\\ \ell_1(b;z) &= \sum_{i=1}^{N} \left\{z_i \log\left(\pi_c(x_i;b)\right) + (1-z_i)\log\left(1-\pi_c(x_i;b)\right)\right\}, \\ \ell_2(\beta,\alpha;z) &= \sum_{i=1}^{N} \left\{z_i \delta_i \log f_c(t_i \mid \tilde{x}_i;\beta,\alpha) + z_i(1-\delta_i)\log S_c(t_i \mid \tilde{x}_i;\beta,\alpha)\right\} - \frac{\lambda}{2}\|\alpha\|^2\end{aligned}$$

Let $\theta^{(m)} = (b^{(m)}, \beta^{(m)}, \alpha^{(m)})$ be the parameters at the m-th iteration. In the E-step, $w_i^{(m)}$ is calculated as follows:

$$w_i^{(m)} = \mathbb{E}[Z_i|\theta^{(m)},\mathcal{D}] = \begin{cases} 1, & \text{if } \delta_i = 1 \text{ and } t_i < c, \\ \left.\dfrac{\pi_c(x_i;b) S_c(t_i \mid \tilde{x}_i;\beta,\alpha)}{1-\pi_c(x_i;b)+\pi_c(x_i;b) S_c(t_i \mid \tilde{x}_i;\beta,\alpha)}\right|_{\theta=\theta^{(m)}}, & \text{if } \delta_i = 0 \text{ and } t_i < c, \\ 0, & \text{if } t_i \geq c. \end{cases}$$

In the conventional model (Sy and Taylor, 2000), the E-step conditions only on δ_i: if $\delta_i = 1$, the individual is known to be susceptible; if $\delta_i = 0$, the membership remains ambiguous. In contrast, the proposed model introduces the condition $t_i \geq c$, under which $w_i^{(m)} = 0$ is assigned deterministically. This yields a key advantage: individuals with follow-up times beyond c are definitively classified as not having experienced the event by time c, thereby extracting more information from the data than the conventional formulation. By substituting this into Equation (8), we obtain the expected log-likelihood $\tilde{\ell}_C(b,\beta,\alpha;w^{(m)}) = \tilde{\ell}_1(b;w^{(m)}) + \tilde{\ell}_2(\beta,\alpha;w^{(m)})$, where $w^{(m)} = (w_i^{(m)}; i = 1,\ldots,N)^\top$. In

the M-step, we maximize $\tilde{\ell}_C$ with respect to b, β, and α. Since $\tilde{\ell}_C$ is expressed as the sum of $\tilde{\ell}_1$ and $\tilde{\ell}_2$, and the parameters are partitioned between the two terms, maximizing each of them will maximize $\tilde{\ell}_C$. They are each estimated using a quasi-Newton method. By the Laplace approximation, the posterior distribution $p(\theta|\mathcal{D})$ can be approximated by a Gaussian distribution with mean $\hat{\theta}_{\mathrm{MAP}}$ and precision matrix A around $\theta = \hat{\theta}_{\mathrm{MAP}}$, where

$$A = -\nabla_\theta^2 \left(\ell(b, \beta, \alpha) - \frac{\lambda}{2} \|\alpha\|^2 \right) \Big|_{\theta = \hat{\theta}_{\mathrm{MAP}}}.$$

Using this, we calculate the Bayesian credible intervals. The hyperparameter λ is estimated by the empirical Bayes method (MacKay, 1992). The estimation procedure is as follows (see Appendix A for details). Let $m = 1, 2, \ldots$ index the outer iterations of empirical Bayes updates, let $\lambda^{(0)} > 0$ be an initial value for λ, and set $\hat{\lambda}^{(0)} = \lambda^{(0)}$.

1. Given $\hat{\lambda}^{(m-1)}$, perform MAP estimation with hyperparameter $\hat{\lambda}^{(m-1)}$ to obtain $\hat{\theta}_{\mathrm{MAP}}$.

2. Using $\hat{\theta}_{\mathrm{MAP}}$, find all roots $\lambda_j^{(m)}$ $(j = 1, \ldots, J_m)$ of $g(\lambda) = 0$ such that

$$g(\lambda) = \sum_{i=1}^{M_\alpha} \frac{\mu_i}{\mu_i + \lambda} - \lambda \|\hat{\alpha}\|^2 = 0, \quad g'(\lambda_j^{(m)}) < 0$$

   where $M_\alpha = K - 1$ is the dimension of α, $H_{xy} = -\nabla_{xy}^2 \log p(\mathcal{D}|\theta)|_{\theta=\hat{\theta}}$ for $x, y \in \{\tilde{\theta}, \alpha\}$ with $\tilde{\theta} = (b, \beta)$, and μ_i $(i = 1, \ldots, M_\alpha)$ are the eigenvalues of $S := H_{\alpha\alpha} - H_{\alpha\tilde{\theta}} H_{\tilde{\theta}\tilde{\theta}}^{-1} H_{\tilde{\theta}\alpha}$. Set $\hat{\lambda}^{(m)} = \mathrm{argmax}_{\lambda \in \{\lambda_1^{(m)}, \ldots, \lambda_{J_m}^{(m)}\}} \log p(\mathcal{D}|\lambda)$.

3. Repeat the above steps. If

$$\left| \log p(\mathcal{D}|\hat{\lambda}^{(m)}) - \log p(\mathcal{D}|\hat{\lambda}^{(m-1)}) \right| < \epsilon$$

   for a certain threshold $\epsilon > 0$, the procedure is regarded as converged.

4. Using the converged $\hat{\lambda}^{(m)}$, perform MAP estimation again and take the resulting $\hat{\theta}_{\mathrm{MAP}}$ as the final estimator.

3 Simulation Studies

In this section, we evaluate the proposed model using simulation studies and demonstrate the potential pitfalls of conventional models in finite-horizon settings.

3.1 Scenario A

In Scenario A, we evaluate the performance of the proposed model using synthetic data. The proposed model serves as the true data-generating mechanism. Although the model is estimated via Bayesian MAP inference, we evaluate its performance using frequentist criteria. We generate synthetic data given sample size N, number of covariates p, and c, and apply the proposed EM algorithm and variance estimation based on the Laplace approximation. We evaluate the performance using empirical bias, coverage probability (CP), and credible interval width for the regression coefficients. The covariate vector $\tilde{x}_i$ combines p_{cont} continuous covariates drawn from $N(0, 1)$ and

categorical covariates (one-hot encoded). We set $p_{\text{cont}} = 1$ and three categorical variables with 4, 3, and 2 levels, yielding $p = 7$, to mirror the predominantly categorical structure of the real data in Section 4. The vector $x_i = (1, \tilde{x}_i^\top)^\top$ is fixed across all replications. In each replication, the latent variable z_i is generated as $z_i \sim \text{Bernoulli}(\pi_c(x_i; b))$ based on the logistic regression model (7). The coefficient vector is set to $b = (b_1, \ldots, b_{p+1})^\top$, where the intercept b_1 controls the overall proportion of the event group up to time c. The true value of the coefficients except for b_1 is $(b_2, \ldots, b_{p+1}) = (-0.3,\ 0.5,\ 0.4,\ 0.2,\ 0.0,\ -0.2,\ -0.5)^\top$.

For individuals with $z_i = 1$, the event time T_i is generated from the survival function $S_c(t \mid \tilde{x}_i; \beta)$ using the inverse transform sampling. For individuals with $z_i = 0$, the event time is set to $T_i = c + U_i$, where $U_i \sim \text{Exponential}(\lambda_{>c})$ with $\lambda_{>c} = 0.05$. This setting allows us to evaluate model performance when events for the "cured" group technically occur after c, but are unobserved within the finite window. For the baseline survival distribution of the event group up to time c, we adopt a distribution on the finite interval $[0, c)$ with shape parameter $\eta = 1.5$:

$$S_{0c}(t) = 1 - (t/c)^\eta, \quad 0 \le t < c.$$

Using this baseline in the Cox proportional hazards form yields the conditional survival function for data generation. The estimated model approximates this baseline via the B-spline representation in Equation (6). The true value of the coefficients for the latency (survival) model is $\beta = (0.3,\ -0.4,\ -0.2,\ 0.0,\ 0.2,\ 0.4,\ 0.5)^\top$.

Independent censoring times C_i are generated from an exponential distribution with rate λ_{cens}. We consider two sub-scenarios (A-1, A-2) by varying the event group up to time c proportion and censoring rate.

1. Scenario A-1 represents a high event group up to time c proportion (70%) with a standard censoring rate (30% among $z_i = 1$).

2. Scenario A-2 represents a low event group up to time c proportion (30%) with a standard censoring rate (30% among $z_i = 1$).

The parameters b_1 and λ_{cens} were tuned using a monotone line search on a pilot dataset of $N = 100,000$ to achieve these target proportions. The final values are: Scenario A-1 ($b_1 = 0.928, \lambda_{\text{cens}} = 0.06$); Scenario A-2 ($b_1 = -0.838, \lambda_{\text{cens}} = 0.06$).

We perform simulations with sample sizes $N = 500, 1000$, setting the number of replications $M = 500$, $K = 7$, and $c = 10$. We evaluate the estimation accuracy of β and b using bias, coverage probability, and credible interval width. Additionally, the baseline survival function is evaluated using the Root Mean Integrated Squared Error (RMISE):

$$\text{RMISE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{c}\int_0^c \left(\hat{S}_{0c,m}(t) - S_{0c}(t)\right)^2 dt\right]}.$$

Here, $\hat{S}_{0c,m}(t)$ denotes the estimated baseline survival function in the m-th replication, and $S_{0c}(t)$ is the true baseline survival function. The integral is evaluated numerically using the trapezoidal rule. Specifically, the interval $[0, c]$ is divided into J equally spaced grid points $t_1, \ldots, t_J$ with $t_{j+1} - t_j = c/J$; in this simulation, we set $J = 1000$. The simulation results are summarized in

Tables 1 and 2. As shown in Table 1, both scenarios A-1 and A-2 demonstrate a tendency for the bias and standard deviation to decrease as the sample size increases. The coverage probabilities are mostly close to 95%, which are near the nominal level. Furthermore, transitioning from scenario A-1 to A-2, the bias and standard deviation of the parameters in the incidence part decrease slightly, while those in the latency part increase. This behavior is consistent with the following mechanism: a decrease in the event rate up to time c leads to an increased proportion of the event-free group up to time c. Consequently, the number of individuals with an observation period greater than c increases, which in turn increases the number of individuals with $w_i^{(m)} = 0$, thereby slightly stabilizing the estimation for the incidence part. At the same time, this decrease in the event rate reduces the effective sample size available for estimating the latency part. Additionally, Table 2 shows that for both A-1 and A-2, the RMISE decreases as the sample size increases, whereas transitioning from A-1 to A-2, the RMISE increases as the event rate decreases. The former observation is consistent with an increase in the effective sample size for estimating the baseline function as the overall sample size grows. The latter is consistent with a reduction in the effective sample size for baseline function estimation due to the decreased event rate. Based on these findings, it is suggested that the proposed model performs the estimations correctly. Additionally, we conducted a sensitivity analysis by varying the number of basis functions $K \in \{5, 7, 10, 15\}$. The estimated regression coefficients remained largely unchanged across these choices, indicating that the proposed model is robust with respect to the specification of K.

3.2 Scenario B

In this subsection, we show that when decisions are made over a finite horizon, a model that assumes a different structure—such as the conventional infinite-horizon cure model—can yield misleading conclusions and lead to incorrect decisions. Using simulation data, we verify that the proposed finite-horizon model correctly answers the two questions of interest: "What is the probability that the event occurs within the c-period?" and "What is the time until the event occurs among those who experience it within the c-period?"

In the simulation for this scenario, we consider a binary covariate x (e.g., treatment vs. control, or exposed vs. unexposed). The covariate is generated as $x \sim \text{Bernoulli}(0.5)$. Conditional on x, the ultimate event indicator $Y \in \{0, 1\}$ (whether the event ever occurs) and the event time T are generated from a cure-type exponential model. The incidence part is specified so that

$$\pi(x) = P(Y = 1 \mid x) = \begin{cases} 0.5, & \text{if } x = 1, \\ 0.8, & \text{if } x = 0, \end{cases}$$

and, given $Y = 1$, the latency part follows an exponential distribution

$$T \mid (Y = 1, x) \sim \text{Exp}(\lambda(x)), \qquad \lambda(x) = \begin{cases} 7.0, & \text{if } x = 1, \\ 0.4, & \text{if } x = 0. \end{cases}$$

Thus $x = 1$ reduces the ultimate probability of the event but shortens the time to the event among susceptible individuals. Figure 1a shows the Kaplan–Meier estimator for the susceptible-only population. The susceptible-only curve for $x = 1$ approaches zero within the first few time units. Figure 1b shows the Kaplan–Meier estimator for the overall population. The $x = 1$ group exhibits

Table 1: Simulation results for proposed model. Comparison of empirical bias (Bias), empirical standard deviation (SD), 95% coverage probability (CP), and mean 95% credible interval width (Width) across all scenarios ($N = 500$ vs $N = 1000$, $M = 500$ replications).

		$N = 500$				$N = 1000$			
Param	**True**	**Bias**	**SD**	**CP**	**Width**	**Bias**	**SD**	**CP**	**Width**
Scenario A-1: High Event Rate (70%)									
Incidence Model Parameters (Logistic)									
b_1	0.928	0.024	0.343	0.952	1.315	0.027	0.236	0.960	0.932
b_2	-0.300	-0.014	0.134	0.966	0.540	-0.010	0.092	0.950	0.366
b_3	0.500	-0.000	0.380	0.942	1.451	0.011	0.265	0.952	0.999
b_4	0.400	0.005	0.366	0.952	1.458	0.003	0.266	0.938	1.012
b_5	0.200	0.012	0.360	0.942	1.381	0.002	0.243	0.944	0.948
b_6	0.000	0.001	0.328	0.940	1.253	-0.018	0.234	0.936	0.881
b_7	-0.200	-0.005	0.333	0.952	1.249	-0.035	0.224	0.940	0.865
b_8	-0.500	-0.020	0.262	0.958	1.027	-0.010	0.192	0.944	0.714
Latency Model Parameters (Survival)									
β_1	0.300	0.002	0.078	0.932	0.290	0.003	0.050	0.946	0.196
β_2	-0.400	0.034	0.179	0.930	0.678	0.019	0.121	0.954	0.489
β_3	-0.200	0.037	0.181	0.946	0.722	0.024	0.132	0.940	0.512
β_4	0.000	0.037	0.182	0.954	0.735	0.028	0.125	0.940	0.501
β_5	0.200	0.026	0.157	0.958	0.626	0.024	0.105	0.968	0.445
β_6	0.400	0.029	0.165	0.952	0.664	0.020	0.112	0.958	0.461
β_7	0.500	0.019	0.140	0.952	0.553	0.010	0.099	0.940	0.384
Scenario A-2: Low Event Rate (30%)									
Incidence Model Parameters (Logistic)									
b_1	-0.838	-0.020	0.311	0.968	1.286	-0.001	0.230	0.960	0.907
b_2	-0.300	-0.013	0.138	0.938	0.516	-0.008	0.090	0.942	0.352
b_3	0.500	0.015	0.357	0.954	1.377	-0.013	0.253	0.950	0.961
b_4	0.400	0.010	0.350	0.958	1.418	-0.007	0.270	0.938	0.988
b_5	0.200	-0.017	0.361	0.962	1.425	-0.010	0.243	0.948	0.972
b_6	0.000	-0.014	0.300	0.942	1.169	-0.002	0.211	0.962	0.829
b_7	-0.200	-0.005	0.310	0.952	1.204	-0.003	0.203	0.958	0.838
b_8	-0.500	-0.001	0.250	0.948	0.983	0.000	0.178	0.948	0.683
Latency Model Parameters (Survival)									
β_1	0.300	0.003	0.112	0.962	0.462	0.003	0.077	0.942	0.308
β_2	-0.400	0.070	0.262	0.966	1.035	0.062	0.186	0.952	0.748
β_3	-0.200	0.094	0.294	0.936	1.123	0.056	0.203	0.942	0.790
β_4	0.000	0.089	0.299	0.942	1.190	0.069	0.201	0.956	0.788
β_5	0.200	0.066	0.229	0.962	0.976	0.019	0.176	0.942	0.686
β_6	0.400	0.050	0.271	0.954	1.061	0.035	0.185	0.944	0.722
β_7	0.500	0.029	0.243	0.936	0.891	0.022	0.147	0.962	0.607

Table 2: RMISE summary by scenario and sample size.

Scenario	**RMISE**
Scenario A-1 ($N = 500$):	0.040987
Scenario A-1 ($N = 1000$):	0.029934
Scenario A-2 ($N = 500$):	0.058478
Scenario A-2 ($N = 1000$):	0.044473

an early concentration of events, with the survival curve dropping sharply and then flattening around 0.5, whereas the $x = 0$ group shows a more gradual decline. Independent censoring times are sampled as $C \sim \text{Uniform}(0, 8)$, and we observe $(t_i, \delta_i) = (\min\{T_i, C_i\}, \mathbf{1}\{T_i \leq C_i\})$ for $n = 1{,}000$ individuals.

When we fit the conventional mixture cure model (Sy and Taylor, 2000) that assumes an infinite time horizon, using x as the covariate in both the incidence and latency components, the estimated incidence coefficient for x is negative, whereas the latency coefficient is positive (Figures 1c and 1d). Thus the conventional model suggests that $x = 1$ reduces the eventual probability of the event, which is consistent with the data-generating mechanism ($\pi(1) < \pi(0)$). In contrast, when we fit the proposed finite-horizon mixture model with various cutoffs $c \in \{0.1, 0.5, \ldots, 6.1\}$, the estimated coefficients for x in the incidence change sign as c increases. This captures that $x = 1$ increases the probability of the event within the c-period and shortens the time to the event. For larger c, the incidence coefficient decreases and eventually turns negative, reflecting the transition from the short-horizon question (event probability within c) to the long-horizon question (ultimate event probability). This means the effect of x on $\mathbb{P}(T < c \mid x = 1)$ changes as c changes. If we want to know the effect of x on $\mathbb{P}(T < c \mid x = 1)$ for a specific c, we need to fit the model with the cutoff c. This sign reversal illustrates how the choice of time horizon c fundamentally affects the interpretation of the incidence effect. In cases like the present data, where this tendency is clearly observed, it may be possible to detect it. However, in real-world data, many variables are involved, so this tendency may not be identifiable.

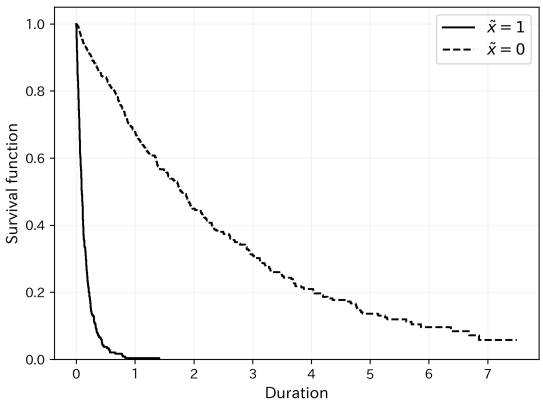
4 Real Data Analysis

In this section, we apply the proposed model to transaction data from Mercari and analyze user behavior in the online flea market.

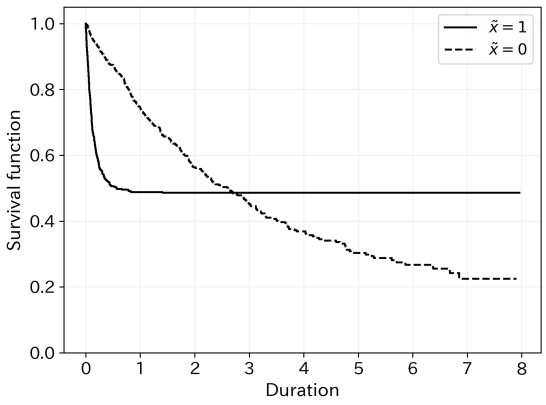
4.1 Data Description

Mercari is an online flea market platform operated by Mercari, Inc., where users can list items for sale and other users can purchase those items. The Mercari dataset contains structured information such as item prices, categories, and shipping methods, as well as unstructured information including item titles, descriptions, and thumbnail images. Seller-specific information, such as transaction history or reputation measures, is not available. We analyze the time from product listing to transaction completion as survival time. A transaction completion involves a sequence of steps, including payment by the buyer, shipment by the seller, receipt of the item, and submission of a review. In this study, a transaction completion is treated as the event of interest ($\delta_i = 1$). Items that have not completed a transaction by the data extraction date are treated as right-censored ($\delta_i = 0$). The target population consists of items listed in 2020, and the dataset is a snapshot collected in June 2023. As a result, temporal changes in product attributes after listing (e.g., price updates, description edits, or image changes) are not observed. Items that remain unsold as of June 2023 are therefore censored. Under this setup, it is possible to observe whether an item was sold for at least two and a half years after listing.

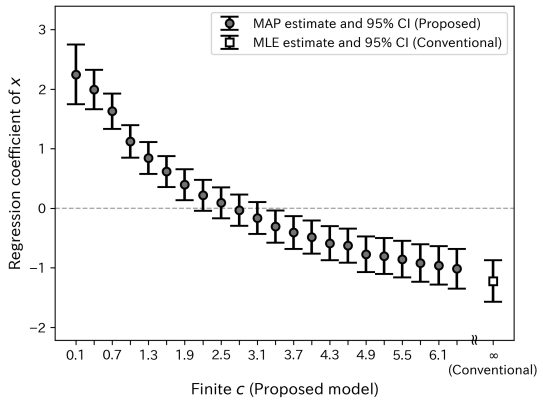
Since the dataset spans from 2020 to June 2023, every item listed in 2020 has at least two and a half years of follow-up, and whether a transaction occurred within any reasonable horizon



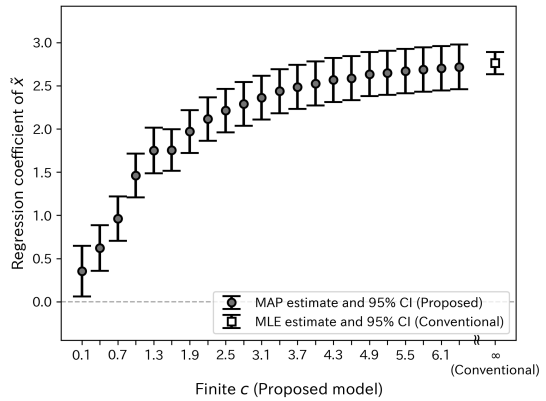
(a) Kaplan–Meier Estimator (Susceptible Only)



(b) Kaplan–Meier Estimator



(c) Incidence Coefficient Estimates



(d) Latency Coefficient Estimates

Figure 1: Summary of results for Scenario B.

c is directly observed. To demonstrate the proposed model in the more realistic setting where some individuals are right-censored before c, we introduce an artificial administrative censoring date of January 1, 2021, assuming that product attributes observed in June 2023 are identical to those at that date. Items unsold by January 1, 2021 are treated as censored at that date. In this censoring mechanism, the survival and censoring times can be considered independent conditional on covariates. Mercari offers a wide range of product categories, including men's, women's, tickets, and furniture. In this analysis, we focus on items belonging to the most granular category, T-shirts/Cut and Sewn (Short Sleeve/Sleeveless), while retaining higher-level categories (e.g., men's or women's) as covariates. Among highly frequent fine-grained categories—such as "Others," "Idol," "Manga," and T-shirts/Cut and Sewn (Short Sleeve/Sleeveless)—we selected this category under the assumption that attribute variability within the category is relatively small. From this category, we randomly sampled 100,000 items, restricting attention to items priced below 15,000 JPY and with a survival time of at least one day. This restriction is motivated by the fact that transaction completion includes shipping after purchase, which we assume takes at least one day; survival times shorter than one day are therefore treated as outliers and excluded. After this filtering, the sample size is $N = 98{,}258$. Descriptive statistics are summarized in Table 3. As shown in the table, the dataset contains many categorical variables, with price being the only continuous variable. The number of listings increases from March, peaks in May, and then decreases monotonically toward December, indicating seasonality in listing activity. The standard deviation of prices is also relatively large. Figure 2 presents the Kaplan–Meier estimate of the survival function for the sampled data. The survival curve drops sharply immediately after listing and then levels off around 30%, suggesting that most items are sold shortly after listing, while a substantial fraction remains unsold. It is unlikely that all unsold items would eventually complete a transaction given a sufficiently long listing period, which motivates the use of cure-type models. Beyond the generic case for cure-type modeling, the C2C holding-cost structure discussed in Section 1 applies directly to Mercari: sellers are individual consumers whose welfare depends on selling within a tolerable period, so the substantive estimand is $\mathbb{P}(T < c \mid x)$ for a finite c rather than $\mathbb{P}(T < \infty \mid x)$. In line with this rationale, we set the cutoff at $c = 1$ week and 3 months in the proposed model, representing horizons within which the seller can plausibly tolerate holding the item. Under this choice of c, the proposed model can be interpreted as evaluating (i) the effect of covariates on the probability of transaction completion within one week or three months, respectively, and (ii) the effect of covariates on the speed of transaction completion among items that complete a transaction within this period.

4.2 Data Preprocessing and Model Fitting

For the application of the analytical models, we conducted data preprocessing. Specifically, the dataset was randomly split into training and test sets with a 4:1 ratio. We then prepared covariates by creating dummy variables for whether the listing date was on a holiday or weekend and for each month, as well as for categorical features such as brand, size, item condition, shipping lead time, and shipping charge burden. For brands, we extracted the top five most frequent brands in both the men's and women's training data, and created dummy variables for their union (excluding 'No brand'). For each categorical predictor (including listing month), the reference category was set to the most frequent level in the training data. Variables that may be modified after listing (item name, and item description) were excluded from the analysis to focus on intrinsic product attributes and

Table 3: Descriptive statistics of the Mercari data (T-shirts/Cut and Sewn (Short Sleeve/Sleeveless))

Sample Composition	
Sample size N	98,258
Number of observed events	66,298
Number of censored	31,960
Censoring rate	32.5%
Observed survival time (hours)	min 25 / Q1 169 / median 675 / Q3 2,937 / max 8,783
Main Attributes	
Top-level category	Men's 55.2%, Women's 44.8%
Item condition	No noticeable damage or stains 43.4%, Brand new/unused 30.2%, Almost unused 13.3%, Slightly damaged/stained 10.7%, Damaged/stained 2.1%, Overall poor condition 0.3%
Shipping lead time	1–2 days 51.4%, 2–3 days 34.5%, 4–7 days 14.1%
Shipping charge burden	Shipping included (seller) 99.0%, Cash on delivery (buyer) 0.97%
Listing month	Jan 3.0%, Feb 3.5%, Mar 6.3%, Apr 10.0%, May 19.2%, Jun 14.4%, Jul 13.6%, Aug 12.5%, Sep 7.2%, Oct 4.4%, Nov 3.1%, Dec 2.6%
Anonymous shipping	Anonymous 86.5%, Non-anonymous 13.5%
Price	min 300 JPY, Q1 800 JPY, median 1,500 JPY, Q3 2,800 JPY, max 15,000 JPY, mean 2,266.3 JPY (SD 2,342.3 JPY)
Size	M 36.7%, L 22.3%, S 14.9%, FREE SIZE 12.4%, XL(LL) 8.8%, XS(SS) 2.1%, 2XL(3L) 1.8%, 3XL(4L) 0.6%, 4XL(5L) or larger 0.3%, XXS or smaller 0.2%, missing 0.1%

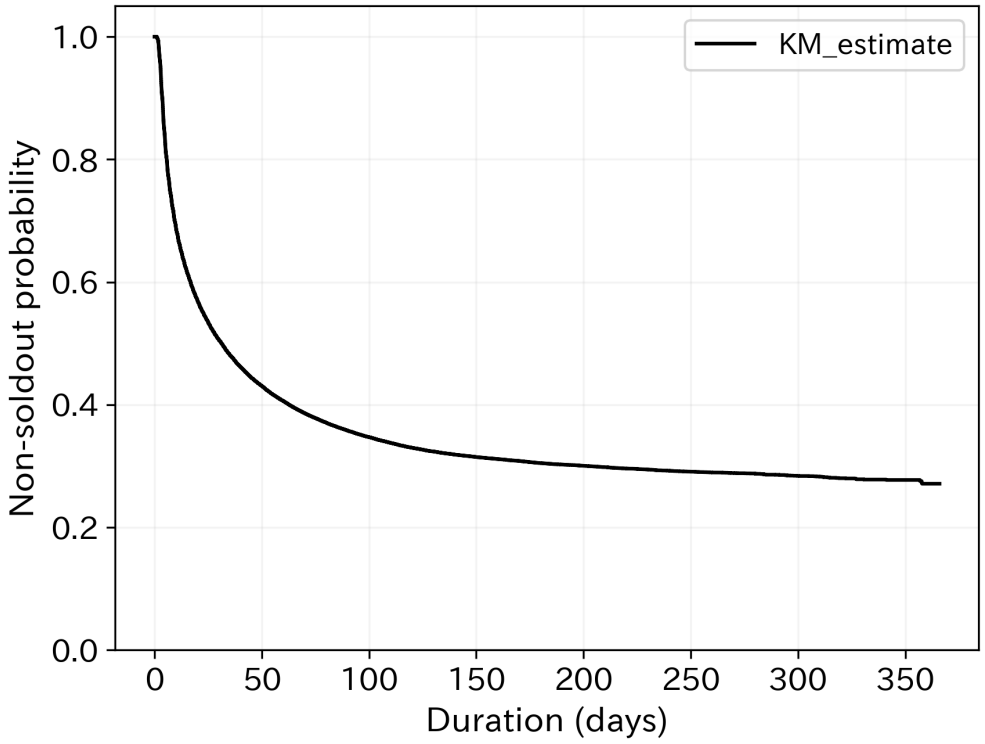


Figure 2: Kaplan–Meier Estimator

listing conditions that are fixed at the time of listing. Finally, an intercept was added and covariates for each item x_i were obtained, resulting in $p = 41$. In addition to the proposed model, we used the conventional model (Sy and Taylor, 2000) as a benchmark for analysis. Training was performed on the training set, and evaluation was carried out on the test set.

The main hyperparameter for the proposed model, K, was chosen by comparing the training-data posterior $p(\theta \mid \mathcal{D})$ across $K \in \{10, 20, 30, 40\}$; gains from increasing K beyond 20 were negligible, so we fixed $K = 20$. A sensitivity analysis with respect to K indicated that the substantive results were largely unchanged. For cross-model comparison, we measured $\overline{\mathrm{AUC}}(25, c)$ (Pölsterl, 2020; Uno et al., 2007) on the test data at $c = 1$ week and $c = 3$ months (Table 4), with all times in hours so that the lower endpoint 25 coincides with the minimum observed duration after excluding listings with time-to-sale of at most one day. Because both models use the same covariate structure and differ only in the target of inference (finite vs. infinite horizon), differences in $\overline{\mathrm{AUC}}(25, c)$ are expected to be small; the present study does not deliberately alter model flexibility.

As for model interpretation, we compare the effects of gender, size, and listing month between the proposed model ($c = 1$ week and $c = 3$ months) and the conventional model; among the covariates examined, these three showed the largest discrepancies across the two approaches.

Figure 3 compares the Women's effect with Men's as the reference category. The conventional mixture cure model decomposes covariate effects into an incidence component for the odds of eventual susceptibility ($T < \infty$) and a latency component for the hazard of sale timing conditional on $T < \infty$, whereas the proposed model targets the odds of sale within horizon c and the hazard of sale timing conditional on $T < c$ among listings with $T < c$. Under $c = 1$ week, the odds ratio exceeds one, so Women's listings are estimated to have a higher probability than Men's of selling within one week. By contrast, for the proposed model with $c = 3$ months and for the conventional

model, the odds ratios are below one and the hazard ratios are above one, so the directions of the estimated odds and hazard effects agree between these two specifications. Although the estimated coefficients share the same signs across these two specifications, the underlying model structures differ, so these coefficients do not admit the same interpretation.

Figure 4 compares size effects with M as the reference category (Figure 4a for odds ratios; Figure 4b for hazard ratios). For odds, the proposed model with $c = 1$ week indicates that XXS or smaller has a significantly positive effect on the probability of completing a transaction within one week. By contrast, for the conventional model—which targets the odds of eventually completing a transaction—and for the proposed model with $c = 3$ months—which targets the odds of completing a transaction within three months—XXS or smaller is not significant. At the same time, 4XL(5L) or larger is significantly positive for odds in all three specifications. For hazard ratios, only the proposed model with $c = 3$ months assigns a significantly positive effect to 4XL(5L) or larger. Taken together, these estimates suggest that on Mercari, extremely small and extremely large sizes—sizes that are rarely offered in mainstream business-to-consumer apparel retail—materially affect both transaction completion and how readily a completion occurs within the relevant horizon. For odds ratios and for hazard ratios separately, the covariates that emerge as statistically significant can differ depending on whether the conventional model or the proposed model with a given horizon c is used; care is therefore required when translating significance patterns into decisions. In particular, if the operational objective is to complete a transaction within one week but the conventional model is used, the significant covariates can differ from those under the proposed model with $c = 1$ week, which may lead to erroneous decisions.

Figure 5 summarizes listing-month effects with May as the reference month: Figure 5a reports odds ratios and Figure 5b reports hazard ratios. The interpretive gap between the conventional and proposed specifications is largest for the odds ratios. In the proposed model with $c = 1$ week, the coefficient for January contrasts the odds that an item listed in January completes a transaction within one week of listing relative to items listed in May. In the proposed model with $c = 3$ months, the coefficient for January contrasts the odds of completing a transaction within three months of listing—for items listed in January, this corresponds to sale by approximately April relative to the May benchmark. The conventional model shows relatively large odds ratios in January and February, followed by a monotone decline toward December. By contrast, the proposed model with $c = 1$ week yields the strongest odds ratios from May through August relative to the other months, whereas the proposed model with $c = 3$ months shows an increase from January to May and a decline from July onward. Because the data analyzed here are for T-shirts/Cut and Sewn (Short Sleeve/Sleeveless) items, a pattern reflecting stronger demand in months closer to the summer season in Japan is substantively plausible. This seasonal pattern is not mirrored by the conventional odds profile, whereas the proposed odds profiles are more consistent with such domain knowledge. For hazard ratios, the conventional model and the proposed model with $c = 3$ months both exhibit seasonality in listing month, whereas the proposed model with $c = 1$ week yields estimates that are comparatively flat across months, suggesting little seasonality in the hazard among items that sell within one week. A plausible interpretation is that listings that clear within one week are broadly popular items for which seasonal variation is secondary. Aligning the estimand—and hence the model and horizon c—with the decision criterion therefore facilitates interpretations that match the operational objective.

Table 4: $\overline{\mathrm{AUC}}(25, c)$ on the test data by horizon c.

Model	$c = 1$ week	$c = 3$ months
Conventional model	0.5841	0.5979
Proposed model	0.5960	0.5995

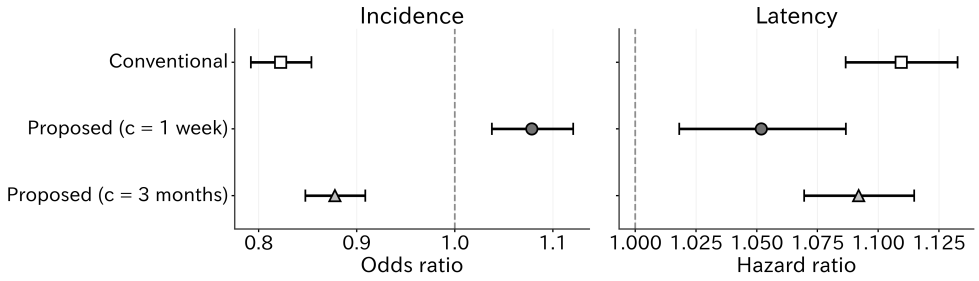


Figure 3: Comparison of the estimated Women's effect with Men's as the reference category.

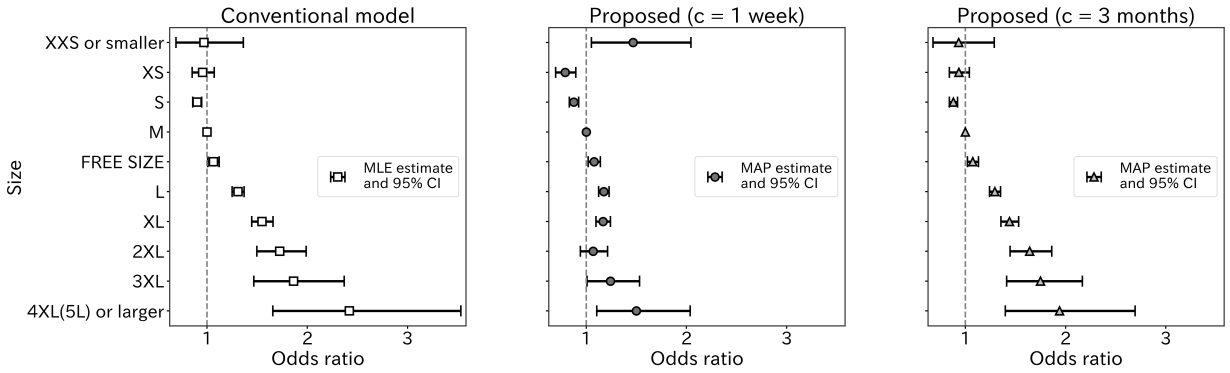
Finally, regarding the nearly monotone decline of the conventional model's odds ratios for listing month in the odds-ratio panel (Figure 5a), we considered whether the observation period might affect the estimation of the coefficients. For example, items listed in January can be observed for at least 330 days, while those listed in December can be observed for at most 30 days. To examine this issue, we also conducted the same conventional model estimation for categories other than T-shirts/Cut and Sewn (Short Sleeve/Sleeveless). As a result, the estimated coefficients did not show a monotonic decrease, and thus we concluded that the observation period in the conventional model does not have a major impact on the estimation of the coefficients.

5 Conclusion

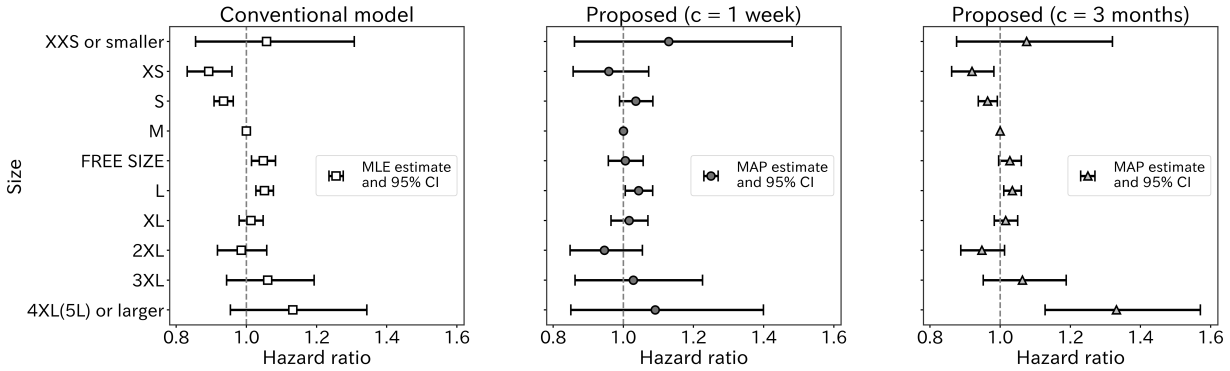
This study proposed a mixture cure model based on a prespecified finite time c, in which the population is latently divided according to whether the event occurs within $[0, c)$ rather than at infinite time. By imposing the boundary constraint $\lim_{t \to c-} S_c(t \mid x) = 0$, the model achieves identifiability without relying on untestable assumptions about tail behavior, and classifies censored observations with $t_i \geq c$ deterministically as belonging to the event-free group, thereby extracting additional information relative to the conventional formulation.

Simulation studies verified the statistical properties of the proposed estimator. In Scenario A, empirical bias and standard deviation decreased as the sample size increased, and coverage probabilities remained close to the nominal 95% level across a range of event rates and covariate structures, indicating that the EM-based estimator behaves as expected. Scenario B revealed a qualitatively important pitfall of conventional infinite-horizon models: when the covariate simultaneously increases the short-term event probability and reduces the long-term probability, the incidence coefficient estimated by the conventional model can reverse sign relative to the finite-horizon truth. This sign reversal demonstrates that using an infinite-horizon model for finite-horizon decision-making can lead to erroneous conclusions, and it underscores the practical value of explicitly specifying a decision-relevant horizon c.

In the application to the Mercari dataset, the proposed model—fit with a three-month hori-

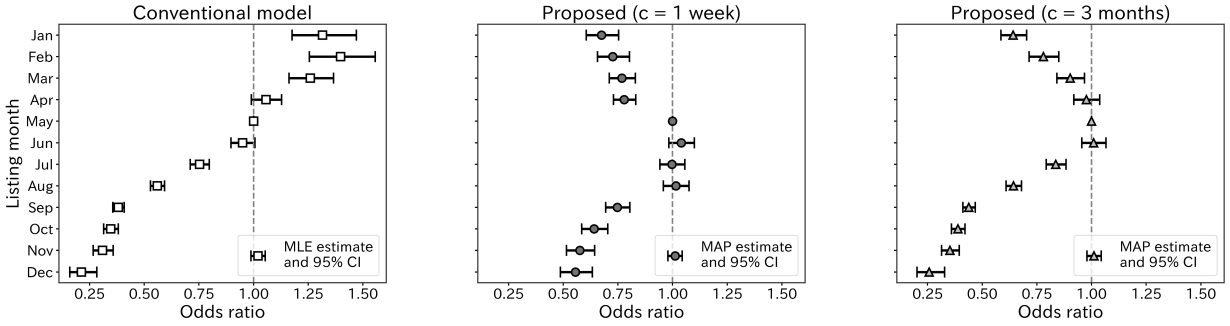


(a) Odds ratio.

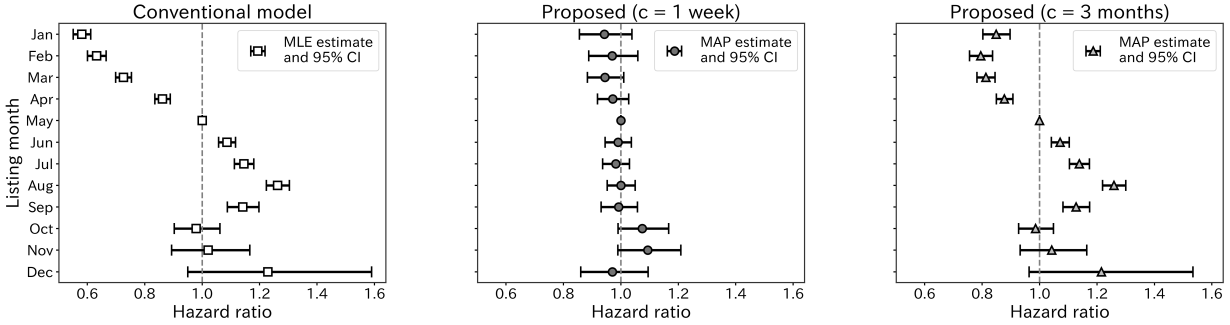


(b) Hazard ratio.

Figure 4: Comparison of size effect with M as the reference category.



(a) Odds ratio.



(b) Hazard ratio.

Figure 5: Comparison of listing month effect with May as the reference category.

zon reflecting the operational decision window—identified qualitatively different patterns from the conventional model. The estimated listing month effect captured a summer-season demand peak consistent with prior knowledge about short-sleeve apparel, a pattern that was not apparent in the conventional model. The significance of size effects also differed between the two models, illustrating that the choice of estimand (finite vs. infinite horizon) can alter substantive conclusions in practice.

A key characteristic of the proposed framework is that all model interpretations are explicitly scoped to the analyst-specified horizon c, which is chosen a priori to reflect the decision-relevant time window rather than inferred from data. Accordingly, the model makes no claim about event probabilities beyond c, and conclusions may differ across different choices of c, as demonstrated in Scenario B. When the appropriate horizon is unclear, sensitivity analysis over multiple values of c is advisable. Future extensions include the incorporation of time-dependent covariates, left truncation, interval censoring, and the derivation of asymptotic distributional theory for the proposed estimators.

Acknowledgements

This study was supported by the Sunaga Shigemitsu Economics Support Fund of the Tohoku University Fund. We also used the Mercari Dataset provided by Mercari, Inc. through the IDR Dataset Provision Service of the National Institute of Informatics. We hereby express our gratitude to all those involved.

A Empirical Bayes Method

We want to find the hyperparameter λ that maximizes:

$$p(\mathcal{D}|\lambda) = \int p(\mathcal{D} \mid \theta)p(\theta|\lambda)d\theta \tag{9}$$

Here, let the dimensions of θ, b, β, α be $M, M_b, M_\beta, M_\alpha$, respectively. We consider this following MacKay (1992). In many cases, $p(\mathcal{D}|\lambda)$ cannot be obtained analytically. Therefore, letting

$$f(\theta) = p(\mathcal{D} \mid \theta)p(\theta|\lambda)$$

we have

$$\log f(\theta) = \log p(\mathcal{D} \mid \theta) + \log p(\theta|\lambda)$$

Considering MAP estimation,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, f(\theta)$$

Since $\nabla_\theta f(\theta)|_{\theta=\hat{\theta}} = 0$, using the Laplace approximation yields:

$$f(\theta) \approx f(\hat{\theta}) \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^\top A(\theta - \hat{\theta})\right)$$

where $A = -\nabla_\theta^2 \log f(\theta)|_{\theta=\hat{\theta}}$. Integrating this with respect to θ,

$$p(\mathcal{D}|\lambda) \approx f(\hat{\theta})(2\pi)^{\frac{M}{2}}|A|^{-\frac{1}{2}}$$

Taking the logarithm,

$$\log p(\mathcal{D}|\lambda) \approx \log f(\hat{\theta}) + \frac{M}{2}\log 2\pi - \frac{1}{2}\log|A|$$

From $f(\hat{\theta}) = p(\mathcal{D}|\hat{\theta})p(\hat{\theta}|\lambda)$, we obtain:

$$\log p(\mathcal{D}|\lambda) \approx \log p(\mathcal{D}|\hat{\theta}) + \log p(\hat{\theta}|\lambda) + \frac{M}{2}\log 2\pi - \frac{1}{2}\log|A| \tag{10}$$

Here, if we set $p(\theta \mid \lambda) = p(b \mid \lambda)p(\beta \mid \lambda)p(\alpha \mid \lambda) \propto p(\alpha \mid \lambda)$, then $p(\alpha \mid \lambda) = \mathcal{N}(\alpha|0, \lambda^{-1}I_{M_\alpha}) = \left(\frac{\lambda}{2\pi}\right)^{\frac{M_\alpha}{2}} \exp\left(-\frac{\lambda}{2}\alpha^\top\alpha\right)$. Taking the logarithm gives:

$$\begin{aligned}\log p(\alpha|\lambda) &= \frac{M_\alpha}{2}\log\lambda - \frac{M_\alpha}{2}\log 2\pi - \frac{\lambda}{2}\alpha^\top\alpha \\ &= \frac{M_\alpha}{2}\log\lambda - \frac{\lambda}{2}\alpha^\top\alpha + \text{const}\end{aligned} \tag{11}$$

Substituting Equation (11) into Equation (10), we get:

$$\log p(\mathcal{D}|\lambda) \approx \log p(\mathcal{D}|\hat{\theta}) + \frac{M_\alpha}{2}\log\lambda - \frac{\lambda}{2}\hat{\alpha}^\top\hat{\alpha} - \frac{1}{2}\log|A| + \text{const} \tag{12}$$

Here,

$$\begin{aligned}A &= -\nabla_\theta^2 \log f(\theta)|_{\theta=\hat{\theta}} \\ &= -\nabla_\theta^2 \log p(\mathcal{D}|\theta)|_{\theta=\hat{\theta}} - \nabla_\theta^2 \log p(\theta|\lambda)|_{\theta=\hat{\theta}} \\ &= H + K\end{aligned}$$

where $H = -\nabla_\theta^2 \log p(\mathcal{D}|\theta)|_{\theta=\hat{\theta}}$ and $K = -\nabla_\theta^2 \log p(\theta|\lambda)|_{\theta=\hat{\theta}}$. Furthermore, letting $\tilde{\theta} = (b, \beta)$, with $\theta = (\tilde{\theta}, \alpha)$, we define:

$$H = \begin{pmatrix} H_{\tilde{\theta}\tilde{\theta}} & H_{\tilde{\theta}\alpha} \\ H_{\alpha\tilde{\theta}} & H_{\alpha\alpha} \end{pmatrix}, \quad H_{xy} = -\nabla_{xy}^2 \log p(\mathcal{D} \mid \theta)\big|_{\theta=\hat{\theta}}.$$

Then A can be written as:

$$A = \begin{pmatrix} H_{\tilde{\theta}\tilde{\theta}} & H_{\tilde{\theta}\alpha} \\ H_{\alpha\tilde{\theta}} & H_{\alpha\alpha} + \lambda I_{M_\alpha} \end{pmatrix}$$

By decomposition using the Schur complement,

$$\begin{aligned}|A| &= |H_{\tilde{\theta}\tilde{\theta}}| \cdot \left|H_{\alpha\alpha} + \lambda I_{M_\alpha} - H_{\alpha\tilde{\theta}}H_{\tilde{\theta}\tilde{\theta}}^{-1}H_{\tilde{\theta}\alpha}\right| \\ &= |H_{\tilde{\theta}\tilde{\theta}}| \cdot |S + \lambda I_{M_\alpha}| \quad \left(S := H_{\alpha\alpha} - H_{\alpha\tilde{\theta}}H_{\tilde{\theta}\tilde{\theta}}^{-1}H_{\tilde{\theta}\alpha}\right)\end{aligned}$$

Here, since S is a real symmetric matrix, it is diagonalizable by an orthogonal matrix. Let its eigenvalues be $\mu_1, \ldots, \mu_{M_\alpha}$. Letting v_i be the eigenvector corresponding to μ_i,

$$Sv_i = \mu_i v_i$$

$$\begin{aligned}(S + \lambda I_{M_\alpha})v_i &= Sv_i + \lambda I_{M_\alpha} v_i \\ &= \mu_i v_i + \lambda v_i \\ &= (\mu_i + \lambda)v_i\end{aligned}$$

Thus, the eigenvalues of $S + \lambda I_{M_\alpha}$ are $\mu_1 + \lambda, \ldots, \mu_{M_\alpha} + \lambda$, so

$$|S + \lambda I_{M_\alpha}| = \prod_{i=1}^{M_\alpha} (\mu_i + \lambda)$$

Therefore,

$$\begin{aligned}\log |A| &= \log |H_{\tilde{\theta}\tilde{\theta}}| + \log \prod_{i=1}^{M_\alpha} (\mu_i + \lambda) \\ &= \log |H_{\tilde{\theta}\tilde{\theta}}| + \sum_{i=1}^{M_\alpha} \log (\mu_i + \lambda)\end{aligned}$$

Differentiating the Laplace-approximated marginal likelihood in Equation (12) with respect to λ, with $\hat{\alpha}$ and μ_i evaluated at the current MAP estimate, yields the following empirical Bayes updating equation.

$$\begin{aligned}\frac{\partial \log p(\mathcal{D}|\lambda)}{\partial \lambda} &\approx \frac{M_\alpha}{2\lambda} - \frac{1}{2}\hat{\alpha}^\top \hat{\alpha} - \frac{1}{2}\sum_{i=1}^{M_\alpha} \frac{1}{\mu_i + \lambda} = 0 \\ M_\alpha - \lambda ||\hat{\alpha}||^2 - \left(M_\alpha - \sum_{i=1}^{M_\alpha} \frac{\mu_i}{\mu_i + \lambda} \right) &= 0 \\ \sum_{i=1}^{M_\alpha} \frac{\mu_i}{\mu_i + \lambda} - \lambda ||\hat{\alpha}||^2 &= 0\end{aligned}$$

Let

$$g(\lambda) = \sum_{i=1}^{M_\alpha} \frac{\mu_i}{\mu_i + \lambda} - \lambda \|\hat{\alpha}\|^2$$

For the precision matrix A to be positive definite at the local maximum of the posterior distribution, it is necessary that $\mu_i + \lambda > 0$ for all i. Therefore, letting $\mu_{\min}$ be the minimum value of μ_i, the domain of λ is $\lambda > \max(0, -\mu_{\min})$. When $\mu_{\min} < 0$, as $\lambda \to (-\mu_{\min})^{+0}$ at the left end of the domain, $g(\lambda) \to -\infty$, and also as $\lambda \to \infty$, $g(\lambda) \to -\infty$. Furthermore, the derivative of $g(\lambda)$,

$$g'(\lambda) = -\sum_{i=1}^{M_\alpha} \frac{\mu_i}{(\mu_i + \lambda)^2} - \|\hat{\alpha}\|^2$$

can change sign multiple times due to the mixture of signs of μ_i. As a result, there can be multiple (0, 2, 4, ...) solutions (stationary points) satisfying $g(\lambda) = 0$ within the interval.

References

Amico, M. and Keilegom, I. V. (2018). Cure Models in Survival Analysis. *Annual Review of Statistics and Its Application*, 5:311–342.

Beirlant, J., Bladt, M., and Van Keilegom, I. (2026). Nonparametric cure models through extreme-value tail estimation. *Scandinavian Journal of Statistics*, pages 1–17.

Berkson, J. and Gage, R. P. (1952). Survival Curve for Cancer Patients Following Treatment. *Journal of the American Statistical Association*, 47(259):501–515.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):15–53.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 34(2):187–202.

de Boor, C. (1978). *A Practical Guide to Splines*. Springer, New York.

Dirick, L., Bellotti, T., Claeskens, G., and Baesens, B. (2019). Macro-Economic Factors in Credit Risk Calculations: Including Time-Varying Covariates in Mixture Cure Models. *Journal of Business & Economic Statistics*, 37(1):40–53.

Escobar-Bach, M. and Keilegom, I. V. (2019). Non-Parametric Cure Rate Estimation Under Insufficient Follow-Up by Using Extremes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(5):861–880.

Farewell, V. T. (1977). A model for a binary variable with time-censored observations. *Biometrika*, 64(1):43–46.

Kuk, A. Y. and Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79(3):531–541.

Kumar, V., Leszkiewicz, A., and Herbst, A. (2018). Are you Back for Good or Still Shopping Around? Investigating Customers' Repeat Churn Behavior. *Journal of Marketing Research*, 55(2):208–225.

Li, C.-S. and Taylor, J. M. G. (2002). A semi-parametric accelerated failure time cure model. *Statistics in Medicine*, 21(21):3235–3247.

López-Cheda, A., Cao, R., Jácome, M. A., and Van Keilegom, I. (2017). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Computational Statistics & Data Analysis*, 105:144–165.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation*, 4(3):415–447.

Mercari, Inc. (2023). Mercari Dataset. Informatics Research Data Repository, National Institute of Informatics. (dataset). https://doi.org/10.32130/idr.17.1.

Peng, Y. (2003). Estimating baseline distribution in proportional hazards cure models. *Computational Statistics & Data Analysis*, 42(1):187–201.

Peng, Y. and Dear, K. B. G. (2000). A Nonparametric Mixture Model for Cure Rate Estimation. *Biometrics*, 56(1):237–243.

Ping Xie, Escobar-Bach, M., and Van Keilegom, I. (2024). Testing for Sufficient Follow-Up in Censored Survival Data by Using Extremes. *Biometrical Journal*, 66(7):e202400033.

Pölsterl, S. (2020). scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6.

Safari, W. C., López-de Ullibarri, I., and Jácome, M. A. (2023). Latency function estimation under the mixture cure model when the cure status is available. *Lifetime Data Analysis*, 29(3):608–627.

Sy, J. P. and Taylor, J. M. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, 56(1):227–236.

Taylor, J. M. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics*, pages 899–907.

Uno, H., Cai, T., Tian, L., and Wei, L. J. (2007). Evaluating prediction rules for t-year survivors

with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537.

Yakovlev, A. Y., Tsodikov, A. D., and Asselain, B. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*, volume 1. World Scientific.