2022

A Comparison of Stacking Methods to Estimate Survival Using Residual Lifetime Data from Prevalent Cohort Studies

Zhaoheng Li
Macalester College, zhaohengli02@gmail.com

MACALESTER

A Comparison of Stacking Methods to Estimate Survival Using Residual Lifetime Data from Prevalent Cohort Studies

Zhaoheng Li

Vittorio Addona, Advisor

Bryan Martin, Reader

Kelsey Grinde, Reader

April 2022

Macalester College
Department of Mathematics, Statistics and Computer Science

Abstract

Prevalent cohort studies are widely used for their cost-efficiency and convenience. However, in such studies, only the residual lifetime can be observed. Traditionally, researchers rely on self-reported onset times to infer the underlying survival distribution, which may introduce additional bias that confounds downstream analysis. This study compares two stacking procedures and one mixture model approach that uses only residual lifetime data while leveraging the strengths of different estimators. Our simulation results show that the two stacked estimators outperform the nonparametric maximum likelihood estimator (NPMLE) and the mixture model, allowing robust and accurate estimations for underlying survival distributions.

Acknowledgements

The stacked survival estimator was developed in collaboration with Dr. David B. Wolfson (McGill University), Dr. James H. McVittie (McGill University), and Dr. Vittorio Addona in the summer of 2020, whose work became the basis of this project. I am also sincerely thankful to Dr. Bryan Martin and Dr. Kelsey Grinde for their careful reading of the project and valuable feedback. Finally, I would like to give special thanks to Dr. Vittorio Addona for being a wondering advisor and collaborator, who showed me directions and offered invaluable support that made this project possible.

Contents

Chapter 1

Introduction

Survival data describes the length of time between an initiating event and a terminating event. In epidemiological studies, the terminating event could be disease relapse or death from disease. For example, survival data is often used to describe the time duration from the onset of a disease to death or the time span from treatment to relapse. Frequently, in such studies, the event of interest is not observed to occur within the study's time window for all subjects. In the medical context, this may happen because the study is scheduled to be terminated due to practical constraints, e.g. budgetary limitations, or because some patients are lost to follow-up, that is, we lose contact with them for any number of possible reasons. As a result, survival data is characterized by so-called *right censoring*. In this case, we will only know that the value is larger than a so-called censoring time, but we will not know its precise value. For example, if a study ends before a patient dies of a certain disease, researchers can only be sure that this patient survives beyond the end of the study, i.e, the date at which the patient is right-censored, but the patient's exact survival time remains unknown. In the remainder of this work, we refer to right-censoring due to a study's termination as administrative censoring to contrast it with loss to follow-up right censoring.

Figure 1.1 displays a schematic of what is termed an *incident cohort study*. For example, suppose the research interest is the survival time of a certain disease since the diagnosis of this disease. Patients recruited in an incident cohort study have not experienced diagnosis yet. Therefore, both the onset time and the failure time or the censoring time are observed. However, although incident cohort studies provide data from the under-

lying target population of interest, they are often time-consuming and expensive. By comparison, *prevalent cohort studies* are more economical and practical, which makes them a common practice for data gathering. These studies are characterized by the identification and enrollment of individuals who have, by the beginning of the study, already experienced the initiating event of interest. These subjects are then followed forward in time for a specified period, with some experiencing the terminating event of interest within administrative censoring.

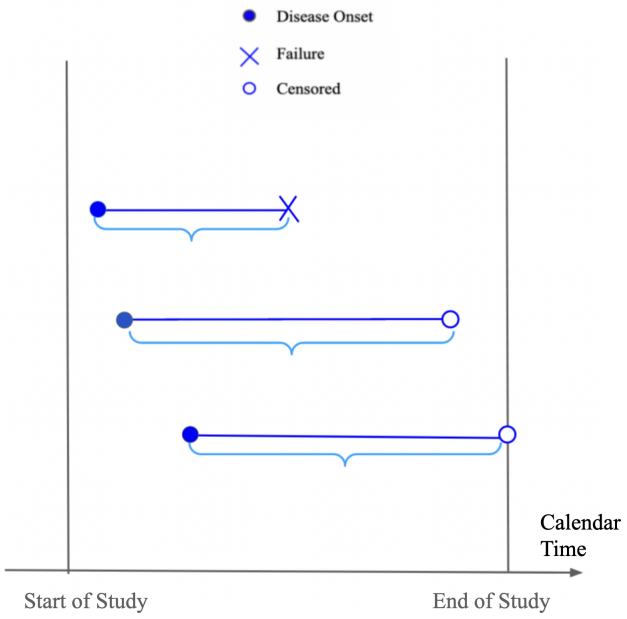


Figure 1.1: Incident Cohort Study, with three hypothetical patients: the terminating event is observed within the study window for the topmost patient, yielding an exact survival time; the other two patients are right-censored, either because they were lost to follow-up (middle) or because the study ended (bottom).

In practice, prevalent cohort studies accrue subjects over a short period of time, but they are often theorized as slicing into a population at a specific time point to identify subjects. Figure 1.2 displays a pictorial representation of a prevalent cohort study. Let the date of screening for our prevalent cohort study be denoted by R, and suppose that an individual i has an onset of the condition of interest at O_i. The full survival time of

interest is T_i, which is composed of the backward recurrence time W_i and the forward recurrence time $T_i - W_i$ [4]. In related literature, some authors refer to the backward recurrence time as the current duration [10] and the forward recurrence time as the residual lifetime [3, 12]. Since the backward recurrence time occurs before the start of the study, it is only possible to observe the forward recurrence time during the study. Let C_i denote the censoring time, that is, the time from the start of the study until loss-to-follow-up or administrative censoring. Then the actually observed time since the beginning of the study, i.e, the prevalent day, until failure or censoring is $X_i = \min(C_i, T_i - W_i)$, which is *left-truncated* and possibly right-censored. The term "left-truncation" refers to the idea that some patients (e.g., see the topmost individual in Figure 1.2) will have experienced the terminating event before even being observable in the study. These subjects can be thought of as being "truncated" from our sample, that is, missed entirely due to our sampling scheme. Moreover, this illustrates an added complication of the prevalent cohort study design: the subjects that we do observe do not constitute a random sample from the target population, but rather, they tend to have longer survival times. This can be understood as follows: a subject is only observable in a prevalent cohort study if $O_i < R$ and $O_i + T_i \geq R$, that is, conditional on surviving long enough to make it into the study. The observed data from a prevalent cohort study can thus be summarized by the triple $\mathcal{O}_i = \{W_i, X_i, \delta_i\}$, where $\delta_i = \mathbb{1}_{C_i > T_i - W_i}$ represents the status indicator for the i^{th} subject.

Since all recruited subjects into prevalent cohort studies have already experienced the onset or initiating event, prevalent cohort studies demand less time and financial resources than incident cohort studies. Ideally, the full prevalent observation could be used (that is, the sum of the *recalled* backward recurrence time and the observed forward recurrence time). These "recalled" backward recurrence times, W_i, can be obtained by interviewing subjects, for example, in order to determine as good as possible when the initiating event occurred. In practice, however, attempts to recall when onset might have occurred can be unreliable. This can pose yet another challenge to the estimation of the underlying survival curve from the full left-truncated times obtained from a prevalent cohort study.

This practical difficulty of left-truncated data obtained from a prevalent cohort study provides the motivation to only use the possibly censored forward recurrence times or residual lifetimes, X_i, in order to estimate the underlying survival distribution. Under the assumption of a stationary in-

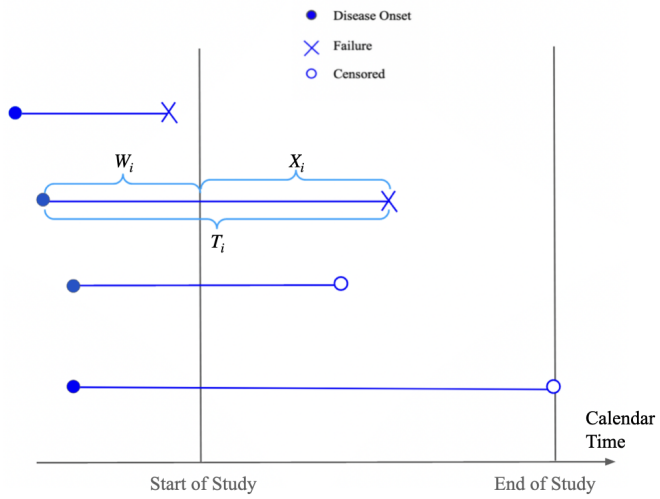


Figure 1.2: Example of a prevalent cohort study. Each line represents an individual, and only the second, third, and fourth individuals are recruited into the study. Only forward recurrence time, X_i, can be observed directly for recruited subjects. Individual 2 yields an exact observation, while the data for individuals 3 and 4 are right-censored. $\delta_i = 1$ for exact observations, i.e, the second individual in the figure; $\delta_i = 0$ for censored observations, i.e, the third and fourth individuals

cidence rate (i.e., of a roughly constant onset incidence), the relationship between the underlying survival curve, S, and the forward recurrence time density function, f_{fwrd}, is well documented in the literature [1]. We can thus exploit this relationship to estimate the survival curve of interest, S, from the target population, using data generated by f_{fwrd}.

Regardless of the data at our disposal, estimates of S can be arrived at by adopting either a *parametric* or a *non-parametric* approach. Each of these broad approaches, however, has advantages and weaknesses. On one hand, any single parametric model could be biased if its distributional assumptions are not met; on the other hand, even an optimal non-parametric estimator (i.e., the non-parametric maximum likelihood estimator, or NPMLE) may yield wide confidence intervals due to its intrinsic robustness stemming from a lack of assumptions about the data-generating process. Moreover, as we will see in Section 2.4, the NPMLE, being solely data driven, is only defined over a limited support corresponding to the study's follow-

up period. Consequently, there is no hope, from the NPMLE, of being able to estimate S at times that lie beyond the length of the study's follow-up period.

An alternative approach that we explore throughout this paper is that of combining parametric and non-parametric estimators to arrive at new estimators which overcome the respective shortcomings (e.g., potential misspecification, and lack of precision and definition) of the individual component estimators. Traditionally, so-called *mixture models* can be used to combine parametric models from the same family. However, the inclusion of a non-parametric estimator, and the fact that mixture models have traditionally only allowed for multiple groups sharing a single parametric family signify that they may not adequately serve our purposes, which in part aims to introduce a process that is more robust to model misspecification.

By contrast, a *stacking* approach [2, 21] may provide more flexibility in terms of allowing for both non-parametric and distinct parametric family components to combine into a single estimator of S. This may allow us to benefit from desirable properties of both the parametric and non-parametric estimation procedures. Broadly speaking, *stacking* first obtains estimates for each candidate model, and then optimizes an objective function to arrive at a set of weights for combining the candidate models into a single estimator. Nevertheless, there is still much room to explore different modeling choices. For example, Smyth and Wolpert propose a method to form a stacked estimator where the stacking is carried out on the density functions of the data, which in their case is assumed to be uncensored [17]. Alternatively, Wey, Connett, and Rudser [20] extend the concept of stacking to a combination of survival functions using potentially right-censored data. This study modified both methods so that they can be applied to forward recurrence time data collected in prevalent cohort studies.

In Section 2, we establish our model assumptions and the relationship between forward recurrence times and underlying survival times, laying the foundation for our procedures. In Section 2.5, we elaborate on the details of how a stacking approach can be used to linearly combine both parametric and non-parametric estimators with the goal of estimating the underlying target survival curve, S. In Section 2.5.2, we introduce an alternative stacking method that combines density estimators instead of survival curves. Section 2.6 presents a third strategy that uses a mixture model approach. We evaluated the three methods through simulation studies,

whose results are reported in Section 3. Discussions for the simulation results can be found in Section 4. We conclude in Section 5 that either the stacked survival estimator or the stacked density estimator can be used to obtain an accurate and robust estimator for survival data collected from prevalent cohort study designs using only the observed forward recurrence time data.

Chapter 2

Methods

2.1 Model Assumptions

Our method is based on the relationship between the forward recurrence time data density function and the underlying (or target) survival distribution, S. This relationship is based on two frequently made assumptions [20]:

i. *Non-informative censoring*: We assume that the process by which subjects are right-censored is non-informative for their survival time. This assumption is very common in the literature, and it can be understood as independence between the censoring time random variable and the failure time random variable. Alternatively, we can think of this assumption as stating that knowledge of the censoring time does not provide any information about the occurrence of the terminating event being imminent, say. For a censored data point, we only know that the survival time is greater than the observed censoring time.

ii. *Stationary incidence process*: We assume that the rate of disease occurrence is constant in the population. This is another common assumption in the literature. If we are not prepared to make this assumption, an alternative is to conduct an analysis conditional on the observed truncation times, or to assume a parametric form for the onset process. If we are not prepared to make any assumption about the disease onset process, then it would become impossible to estimate survival due to an identifiability issue. That is, observed data could provide information about S, or could be due to changes in the frequency of onsets (or initiating events). We note that, as a consequence

of this *stationarity* assumption, the truncation time distribution is uniform, and the forward and backward recurrence time distributions are identical [1].

2.2 Relation between Residual Lifetime Distribution and the Underlying Survival Time Distribution

Suppose that we denote the density function of the forward recurrence times, $(T_i - W_i)$, by f_{fwrd}. Then, under assumptions i and ii, it is well-known that we have the following relationship between f_{fwrd} and the target (underlying) survival curve, S [1, 9, 11, 19]:

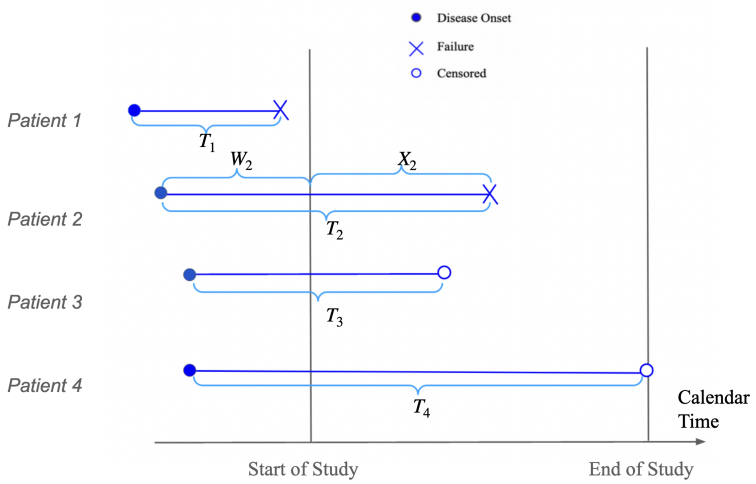
Theorem 1.

$$f_{fwrd}(t) = \frac{S(t)}{\mu} \tag{2.1}$$

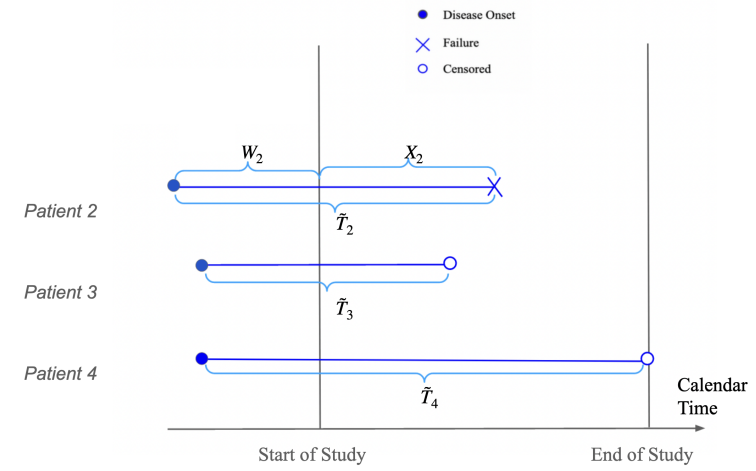
where $\mu = E[T]$

The following proof [1] uses the length-biased distribution of T, $\tilde{T}$. The use of $\tilde{T}$ reflects that the survival data collected in prevalent cohort studies is length biased, i.e, it is more likely for individuals with longer survival time to be recruited into the study. An illustration of the distinction between T and $\tilde{T}$ is presented in Figure 2.1, where $\tilde{X}$ and $\tilde{W}$ are the length-biased distributions of X and W.

Proof. Denote by $\tilde{T} = \tilde{W}_i + \tilde{X}_i$ the length-biased distribution of T. The probability to observe $\tilde{T}_i$ is proportional to T_i. Under stationary incidence rates, W follows a uniform distribution f_W with a corresponding cumulative density function F_W. Given that T has a density function f_T and mean survival time $E[T] = \mu$, write



(a)



(b)

Figure 2.1: $\tilde{T}$ is the length-biased distribution of T. (a): T_1, T_2, T_3, and T_4 are the survival time of four individuals with distribution T; (b): only T_2,T_3, and T_4 are sampled from the length biased distribution $\tilde{T}$ since individual 1 experiences the failure event before the prevalent day when the study begins.

$$f_{\tilde{T},\tilde{W}}(t,w) = f(t,w|T \geq W)$$

$$= \frac{f(t)f_W(w)}{\int_0^\infty f_W(u)S(u)du}\mathbb{1}(t \geq w)$$

$$= \frac{f(t)}{\int_0^\infty S(u)du}\mathbb{1}(t \geq w)$$

Since $\int_0^\infty S(u)du = \mu$, it follows that

$$f_{\tilde{T},\tilde{W}}(t,w) = \frac{f(t)}{\mu}$$

and

$$f_{\tilde{W}}(w) = \int_w^\infty f_{\tilde{T},\tilde{W}}(t,w)dt$$

$$= \frac{S(w)}{\mu}$$

Under the assumption of stationary onsets, we shall show below that the length-biased backward and forward recurrence times are identical in distribution.

Theorem 1 yields the following useful result for $f_{\tilde{T}}$:

$$f_{\tilde{T}}(t) = f(t|T \geq W)$$

$$= \int_0^\infty f(t,w|T \geq W)dw$$

$$= \int_0^t f(t,w|T \geq W)dw$$

$$= \int_0^t \frac{f(t)f_W(w)}{\int_0^\infty f_W(u)S(u)du}dw$$

$$= \frac{f(t)F_W(t)}{\int_0^\infty f_W(u)S(u)du}$$

Under the stationarity assumption, ii, we have $F_W(t) = tf_W(t)$, and the above equation reduces to the well-known length-biased distribution of T, that is,

$$f_{\tilde{T}}(t) = \frac{tf(t)}{\mu}$$

Thus,

$$f_{\tilde{W}}(w|\tilde{T}=t) = \frac{f_{\tilde{T},\tilde{W}}(t,w)}{f_{\tilde{T}}(t)} = \frac{1}{t}.$$

To derive the distribution for the forward recurrence times, X, we write:

$$f_{fwrd}(x|\tilde{T}=t) = f_{\tilde{W}(t-x|\tilde{T}=t)=\frac{1}{t}}$$

Then,

$$\begin{aligned} f_{fwrd}(x) &= \int_x^\infty f_{\tilde{T},X}(t,x)dt \\ &= \int_x^\infty f_{\tilde{X}}(x|\tilde{T}=t)f_{\tilde{T}}(t)dt \\ &= \int_x^\infty \frac{1}{u}\frac{uf(u)}{\mu}du \\ &= \frac{S(x)}{\mu} \end{aligned}$$

□

Finally, this implies

$$f_{fwrd}(0) = \frac{S(0)}{\mu} = \frac{1}{\mu}$$

Thus, from Equation (2.1), $f_{fwrd}(t) = S(t)f_{fwrd}(0)$, which suggests a natural estimator of S to be

$$\hat{S}(t) = \frac{\hat{f}_{fwrd}(t)}{\hat{f}_{fwrd}(0)} \tag{2.2}$$

Equation (2.1) is used to construct likelihood functions used in parametric estimations, while Equation (2.2) is explicitly used in the non-parametric

estimation once the estimated density $\hat{f}_{fwrd}(t)$ is obtained using an algorithm due to Denby and Vardi [5], as explained in Section 2.4.

2.3 Parametric Estimator

We use maximum likelihoods estimators for parametric models. Specifically, let θ be the parameter vector of the assumed parametric distribution. The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is obtained by maximizing the likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_{fwrd}^{\delta_i}(x_i) S_{fwrd}^{1-\delta_i}(x_i) \tag{2.3}$$

Substituting equation (2.1) allows us to re-write the likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{S(X_i;\boldsymbol{\theta})^{\delta_i} \left(\int_{z>x_i} S(z;\boldsymbol{\theta}) \right)^{1-\delta_i}}{\mu(\boldsymbol{\theta})} \tag{2.4}$$

where $\mu(\boldsymbol{\theta}) = \int_0^\infty S(u;\boldsymbol{\theta}) du$.

A natural parametric estimator, $\hat{S}$, for S is thus:

$$\hat{S}(t) = S(t;\hat{\boldsymbol{\theta}}) \tag{2.5}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$.

2.4 Non-parametric Estimator

As parametric models may have misspecified assumptions, a non-parametric estimator can be considered to provide a more robust option. Since we are employing forward recurrence time data, a standard non-parametric approach for right-censored data (i.e., the Kaplan-Meier curve applied to the forward recurrence times) would ignore the special structure suggested by equation (2.1). In particular, (2.1) implies that the forward recurrence time density is a non-increasing function, a fact that can be exploited in the estimation procedure.

Denby and Vardi proposed an algorithm for determining the non-parametric maximum likelihood estimate (NPMLE) of a non-increasing density function when using potentially right-censored data [5]. When the largest observation is censored, the likelihood function in (2.3) only has a supremum but no maximum. Thus, Denby and Vardi proposed the *M-restricted maximum likelyhood estimate (MLE) of the density*, where all the remaining probability mass is placed at an extremely large time, M.

Definition 2.4.1. (M-Restricted Maximum Likelihood Estimate) Denote by D_M the set of all nonincreasing left continuous density functions with support on $(0, M]$. The M-restricted MLE is [5]:

$$\max_{g \in D_M} L(g|\text{data})$$

This problem always has a solution, and can be solved iteratively using a version of the Renewal Theory (RT) algorithm [5, 18].

However, this M-restricted NPMLE has a local bias near 0. Note that this bias also occurs in the absence of censoring, where the asymptotic properties of the NPMLE under the constraint of decreasing density have been established [16]. This leads to the conjecture that this phenomenon might be the result of a different rate of convergence near 0 [5]. As equation (2.2) reveals that our proposed estimator depends on $\hat{f}_{fwrd}(0)$, it is necessary to correct this local bias.

This motivates our use of the *corrected Denby-Vardi estimator*. The details on the *corrected Denby-Vardi estimator* can be found in [5], which aims to flatten the peak near 0. In what proceeds, when we refer to the Denby-Vardi estimator, we tacitly mean this corrected version. In particular, we submit the corrected Denby-Vardi estimator as a candidate model for the stacking procedure that is described in Section 2.5.

2.5 Stacking Procedures

2.5.1 Stacking survival curves

Assuming that there are m candidate models, we first compute the corresponding forward survival functions estimators,

$$\hat{S}_{1,fwrd}(x), \hat{S}_{2,fwrd}(x), \ldots, \hat{S}_{m,fwrd}(x),$$

using the maximum likelihood procedure for the parametric models, and the Denby-Vardi estimator in the non-parametric case. We propose an estimator to balance the pros and cons for different models by combining them linearly, resulting in an estimator of the form

$$\hat{S}_{fwrd}(x) = \sum_{k=1}^{m} a_k \hat{S}_{k,fwrd}(x)$$

The weights for the linear combination are obtained via the algorithm proposed by Wey, Connett, and Rudser [20], in which we minimize the squared errors of forward survival function as measured by inverse-probability-of-censoring weighting Brier Score. We now describe this procedure in more details.

Definition 2.5.1. (Brier Score)

Squared error for survival functions at a given time point t is measured by the **Brier Score**. In the absence of censoring, we have

$$BS(t) = \frac{1}{n} \sum_{i=1}^{n} \left(Z_i(t) - \hat{S}(t) \right)^2$$

where $Z_i(t) = \mathbb{1}(t_i > t)$ and t_i is the event time for the i^{th} observation.

To evaluate the estimator's performance at time t, if the failure event of the i^{th} observation does not occur by t, then it is $1 - \hat{S}(t) = \hat{F}(t)$ that contributes to the Brier Score. In this case, smaller estimated survival probability at t, $\hat{S}(t)$, is penalized. By contrast, if the failure event of the i^{th} observation has occurred by time t, greater $\hat{S}(t)$ is penalized.

Since t_i may not be observed, *inverse-probability-of-censoring-weights (IPCW)* [13, 20] are used to account for the probability of an observation being censored. This adjusts for a potential bias resulting from loss-to-follow-up.

Definition 2.5.2. (Inverse-Probability-of-Censoring-Weighted (IPCW) Brier Score)

Let t_i be the event time for the i^{th} observation in Definition 2.5.1, c_i the censoring time, G the survival function of the censoring distribution, $T_i(t) = \min\{t_i, t\}$, and $\Delta_i(t) = \mathbb{1}_{T_i < c_i}$. We can define [20].

$$IPCW - BS(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i(t)}{G(T_i(t))} \times \left(Z_i(t) - \hat{S}(t)\right)^2$$

We note the following items regarding use of IPCW Brier Score:

(1) The weight for an uncensored observation depends on whether the event occurs by time t

(2) Censored observations with $c_i > t$ also contribute to IPCW-BS(t)

(3) Censored observations with $c_i < t$ only contribute indirectly through the estimation of the censoring distribution.

To optimize for the weights $\hat{a}_1, \ldots, \hat{a}_m$ to linearly combine m models, the IPCW Brier Score is minimized over a set of time points, $t_1, \ldots, t_s$, under the constraints $\hat{a_k} \geq 0$ and $\sum_{k=1}^{m} a_k = 1$. That is:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \sum_{r=1}^{s}\sum_{i=1}^{n}\frac{\Delta_i(t)}{G(T_i(t))} \times \left(Z_i(t) - \sum_{k=1}^{m} a_k \hat{S}_k^{(-i)}(t)\right)^2 \tag{2.6}$$

where $\mathbf{a} = \{a_1, a_2, \ldots, a_m\}$ and $\hat{S}_k^{(-i)}$ is the survival estimate from the k^{th} model leaving the i^{th} observation out, resulting in n-fold cross-validation. In practice, 5-fold cross validation is used instead to facilitate computational efficiency. Furthermore, for computational simplicity, this work chooses to optimize the IPCW Brier Score over nine time points that are equally spaced in the range of observed forward recurrence time data as implemented in [15].

It can be shown that there exists a set of weights such that the stacked model performs at least as good as the best candidate available in terms of squared error. It is true, however, that the estimated weights obtained from any particular data set may differ from the ideal set of weights [20].

Finally, the forward stacked survival estimator obtained in this fashion ensures that the corresponding density function has a non-increasing density, since a density function is the negative derivative of a survival curve,

$$f_{fwrd}(x) = -\frac{\mathrm{d}}{\mathrm{d}x} S_{fwrd}(x) = \sum_{k=1}^{m} a_k f_{k,fwrd}(x),$$

thus yielding a linear combination of non-increasing density functions for forward recurrence time data.

Transferring the weights obtained from the forward recurrence distributions, we can obtain a stacked estimator for the underlying survival function, S, in the following intuitive manner:

$$\hat{S}(t) = \sum_{k=1}^{m} \hat{a}_k \hat{S}_k(t)$$

As discussed in Section 4, we note that this transfer of weights is ad-hoc, but one that is interpretable given the assumption that the underlying survival distribution can be described by a stacking model.

2.5.2 Stacking density functions

In Section 2.5.1, we discuss a method for stacking a set of survival functions. Alternatively, after obtaining estimates for k individual models using residual lifetime data, we can stack their density functions

$$f_{fwrd}(x) = \sum_{k=1}^{m} \pi_k f_{k,fwrd}(x).$$

where the set of linear combination coefficients, $\pi_1, \pi_2, \ldots, \pi_m$, can be obtained by an EM algorithm so that the likelihood of the stacked density model to observe the forward recurrence time data could be maximized [17]. Algorithm 1 describes below the framework to optimize for the set of linear combination coefficients π_k:

Cross-validation is used in the training processing to prevent overfitting. To save computational time, we used a 5-fold cross validation for simulation results presented in Section 3.2 as described in Algorithm 2.

Algorithm 1 Optimizing for the coefficients for the stacked-density-estimator through EM algorithm

1: Set $\pi_j^{(0)} = \frac{1}{m}$ for all j.
2: Denote by

$$r_{ij}^{(t+1)} = \frac{\pi_j^{(t)} L_j(x_i)}{\sum_{k=1}^{m} \pi_k^{(t)} L_k(x_i)}$$

the responsibility that model k takes for data point i at the $(t+1)^{th}$ iteration. Here, $L_k(x_i)$ is defined by Equation 2.3. Alternatively, r_{ij} can be thought as the probability for data point i to come from candidate model j
3: Iteratively update by

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} r_{ij}}{n}$$

Algorithm 2 Cross-Validation for stacked-density estimator

Split the data into 5 folds for cross-validation. Obtain 5 estimators for each of the M candidate models. Step 1 and 2 are the same as those for stacking survival functions. Compute the likelihood for each of the N observations under each of the M models and obtain an $N \times M$ matrix L where L_{ij} is the likelihood to observe data i under model j as obtained from equation 2.4. Although 2.4 is an expression for parametric models, likelihood for the NPMLE is calculated in the same fashion. Here the likelihood calculation differs from Smyth and Wolpert's work because of censoring. Feed L into equations (1) and (2) to maximize cross-validated likelihood. After a set of weights π's are obtained according to forward survival time distributions, transfer the weights in the same fashion as Section 2.5.1 describes.

2.6 Mixture Model

In contrast to stacking procedures, we acknowledge that a mixture model could be adopted to combine models from different families if all candidate models are parametric.

Suppose there are d parametric models with parameters $\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \ldots, \boldsymbol{\theta_d}$, and let $\sigma_1, \sigma_2, \ldots, \sigma_d$ be the mixing probabilities where $\sigma_k > 0$ and $\sum_{k=1}^{d} \sigma_k =$

1. Similar to the setting of stacking, we assume the underlying full survival distribution follows a mixture model. That is

$$S(t;\boldsymbol{\theta},\boldsymbol{\sigma}) = \sum_{k=1}^{d} \sigma_k S_k(t;\theta_k).$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_d\}$ and $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \ldots, \sigma_d\}$

This implies

$$\begin{aligned}
\mu_{\boldsymbol{\theta},\boldsymbol{\sigma}} &= \int_0^\infty \sum_{k=1}^{d} \sigma_k S_k(t;\boldsymbol{\theta}_k)\mathrm{d}t \\
&= \sum_{k=1}^{d} \sigma_k \int_0^\infty S_k(t;\boldsymbol{\theta}_k)\mathrm{d}t \\
&= \sum_{k=1}^{d} \sigma_k \mu_k(\boldsymbol{\theta}_k)
\end{aligned}$$

where $\mu_k(\boldsymbol{\theta}_k)$ represents the expected value of the k^{th} parametric model.

It follows that the likelihood can be written as

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\sigma}) &= \prod_{i=1}^{n} \left(\frac{S(t_i;\boldsymbol{\theta},\boldsymbol{\sigma})}{\mu_{\boldsymbol{\theta},\boldsymbol{\sigma}}}\right)^{\delta_i} \left(\int_{t_i}^\infty \frac{S(t;\boldsymbol{\theta},\boldsymbol{\sigma})}{\mu_{\boldsymbol{\theta},\boldsymbol{\sigma}}}\mathrm{d}t\right)^{1-\delta_i} \\
&= \mu_{\boldsymbol{\theta},\boldsymbol{\sigma}}^{n} \prod_{i=1}^{n} \left(S(t_i;\boldsymbol{\theta},\boldsymbol{\sigma})\right)^{\delta_i} \left(\int_{t_i}^\infty S(t_i;\boldsymbol{\theta},\boldsymbol{\sigma})\mathrm{d}t\right)^{1-\delta_i} \\
&= \left(\sum_{k=1}^{d} \sigma_k \mu_k(\boldsymbol{\theta}_k)\right)^{n} \prod_{i=1}^{n} \left(S(t_i;\boldsymbol{\theta},\boldsymbol{\sigma})\right)^{\delta_i} \left(\int_{t_i}^\infty S(t;\boldsymbol{\theta},\boldsymbol{\sigma})\mathrm{d}t\right)^{1-\delta_i}
\end{aligned}$$

The set of parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$, are estimated simultaneously by maximizing numerically the above likelihoods, and the optimized weights $\boldsymbol{\sigma}$ thus can be applied directly to combine the associated underlying survival functions.

One potential advantage of the mixture model approach is that there is no ad-hoc transferring of mixture weights, as it is needed for the stacking procedures. In Section 3, we present extensive simulation results to compare the two stacking procedures with the mixture model approach.

Chapter 3

Simulation Study

Right-censored forward recurrence time data is simulated using the framework employed by McVittie et al.[15]. Briefly, we first generate the onset time O from a $Uniform(50)$ distribution and then simulate the failure time T assuming different underlying true survival distributions. Cases with $T < 50 - O$ are filtered out to represent individuals unobservable in a prevalent cohort study whose failure event occurs before the study initiation. We keep generating the onset-failure pair (O, T) until a dataset with sample size n is obtained. The residual lifetimes are then calculated as $X_i = T_i - (50 - O_i)$ for $i = 1, 2, \ldots, n$. Both administrative censoring and random censoring are considered. To generate a dataset with administrative censoring rate, q, the censoring time C^* is set to be the $1 - q^{\text{th}}$ quantile of X. For random censoring, the censoring time C_i is generated from an exponential distribution. All codes for simulations can be found in Section A.3.

3.1 Comparison with the Denby-Vardi NPMLE

We simulated Weibull (2,2) data under 10%, 20%, and 30% administrative censoring rate and compared the prediction accuracy of NPMLE and the two stacked estimators. Figure 3.1 shows the discrete integrated squared survival errors (DISSE) [15] for the NPMLE and the stacked estimator including the NPMLE, Weibull, Lognormal, Log-Logistic, and Gamma mod-

els as a measure of mean squared errors (MSE), where

$$\text{DISSE} = \sum_{i=1}^{k} (t_i - t_{i-1})(\hat{S}(t_i) - S(t_i))^2$$

Essentially, the DISSE is a measure to numerically compute the sum of squared errors based on the survival curves. It is calculated during the time span 0-10, which covers nearly all the non-zero density weight for the Weibull(2,2) distribution. Both stacked estimators produce smaller DISSEs compared with the Denby-Vardi estimator, where the stacked survival estimator yields a slightly smaller DISSE, which is reasonable since the stacking weights for stacked survival estimators are optimized by directly minimizing the IPCW Brier Score as a measure of MSE. The NPMLE performs poorly as the administrative censoring increases due to its lack of ability to predict any potentially non-zero survival probability after the end of the study, i.e., after administrative censoring. Table 3.1 and Table 3.2 present the optimized stacking weights in the stacked survival and stacked density estimators, respectively. For both estimators, the stacked estimator's weights on the NPMLE diminish quickly due to increasing administrative censoring rates, which corroborated the observations in [15]. The dominant weights the Weibull model receives in Table 3.1 also demonstrates that under this simulation setting, the stacked survival procedure is able to shift the weights towards the correct parametric family.

Administrative Censoring Rates	Stacking Weights				
	NPMLE	Weibull	Loglogistic	Lognormal	Gamma
10%	0.004	0.981	0.010	0.002	0.003
20%	6.017×10^{-4}	0.974	0.023	1.329×10^{-7}	1.911×10^{-3}
30%	2.983×10^{-9}	0.847	0.111	1.042×10^{-7}	0.042

Table 3.1: Weights for the five candidate models in the stacked survival procedure under varying rates of administrative censoring

Administrative Censoring Rates	Stacking Weights				
	NPMLE	Weibull	Loglogistic	Lognormal	Gamma
10%	0.012	0.393	0.234	0.276	0.086
20%	0.012	0.352	0.210	0.283	0.142
30%	0.071	0.082	0.088	0.598	0.161

Table 3.2: Weights for the five candidate models in the stacked density procedure under varying rates of administrative censoring

The superiority of a stacked estimator over the NPMLE in the presence of a high administrate censoring rate is also demonstrated by an application on the dementia dataset collected by the Canadian Study of Health and Aging (CSHA) through a prevalent cohort study during the year span 1991-1996. However, the nature of dementia makes the recalled disease onset times susceptible to bias, thus motivating the use of an estimator that is based solely on the observed residual lifetime. Figure 3.2 shows the estimated survival curves by the non-parametric corrected Denby-Vardi estimator and a stacked survival estimator that is composed of the NPMLE, Weibull, Log-Normal, and Gamma models. The Denby-Vardi NPMLE drops to 0 after 60 months when the study ended, while by comparison, the stacked estimator can capture the tail behavior by incorporating parametric models.

3.2 Comparison with Stacked Density

A simulation study is also conducted to compare the stacked-survival-function estimator with the stacked-density-function estimator assuming random right-censoring. Since the former aims at minimizing mean squared errors as measured by IPCW Brier Score while the latter aims at maximizing likelihoods, both the DISSE and the Kullback-Leibler divergence (KLD) are used as evaluation metrics.

The inclusion of a non-parametric model would cause the stacked estimator to be non-parametric in nature. Due to the construction of the stacked estimator, it is impractical to obtain an estimated density function for the underlying survival time. Therefore, in order to compute the Kullback-Leibler divergence, the Denbi-Vardi NPMLE is excluded from the candidate models for stacking in the following simulations.

In the first scenario, a single distribution is used to simulate failure times. A Weibull (2,2) distribution and a Weibull (0.75,3) distribution are used to generate survival times exhibiting an increasing hazard and a decreasing hazard, respectively. The censoring time distribution is adjusted such that the censoring rate is kept around 30%. In each scenario, fifty datasets with a sample size of 125 are generated with a random censoring rate of 30%. Two types of stacked models are considered, depending on when the correct model–the Weibull model in our case–is included

or not, where the latter scenario aims at reflecting the fact that in reality the correct model family is usually unknown. That is, we want a simulation scenario that assesses the robustness of our procedures in situations where the underlying truth is not a candidate model for stacking. When the correct model is included, we consider stacking all four parametric models–Weibull, Log-Logistic, Log-Normal, and Gamma–and stacking only Weibull and Gamma models. The stacking weights, the DISSE and KLD of individual models, the stacked-survival-function estimator, and the stacked-density-function estimator are reported in Tables 3.3-3.8. All simulation results are obtained using a sample size of 125.

Overall, the two stacked estimators have similar performances in both DISSE and KLD. We also note that all models perform worse in the decreasing hazard case, possibly due to the asymptotic behavior of the Weibull(0.75,3) hazard function at the time point 0. We repeated the simulation scheme for Gamma models under both increasing and decreasing hazards and obtained similar results as reported in the supplementary materials.

Model	DISSE	KLD
Weibull	0.02153351	3.720×10^{-4}
Loglogistic	0.04721325	16.101×10^{-4}
Lognormal	0.03657954	11.003×10^{-4}
Gamma	0.0363825	6.634×10^{-4}

Table 3.3: DISSE, and KLD for individual parametric estimators under a Weibull (2,2) distribution

Weights				DISSE	KLD
Weibull	Loglogistic	Lognormal	Gamma		
0.811	−−	−−	0.189	0.02378625	4.287×10^{-4}
0.573	0.131	0.152	0.144	0.02796682	6.432×10^{-4}
−−	0.208	0.438	0.354	0.03734887	9.891×10^{-4}

Table 3.4: Stacking weights, DISSE, and KLD for stacked-survival-function estimator under a Weibull(2,2) distribution

Apart from simulating data from a single distribution, we also simulated data from a mixed distribution. Though this is a less likely scenario in real life, this simulation is conducted to examine the performance

Weights				DISSE	KLD
Weibull	Loglogistic	Lognormal	Gamma		
0.736	−−	−−	0.264	0.02634225	4.597×10^{-4}
0.602	0.143	0.114	0.140	0.03187206	7.526×10^{-4}
−−	0.150	0.351	0.498	0.03710031	10.546×10^{-4}

Table 3.5: Stacking weights, DISSE, and KLD for stacked-density-function estimator under a Weibull(2,2) distribution

Model	DISSE	KLD
Weibull	0.1371731	0.8285432
Loglogistic	0.4891242	9.682708
Lognormal	0.1873343	34.62643
Gamma	0.2923004	2.168843

Table 3.6: DISSE, and KLD for individual parametric estimators under a Weibull(0.75, 3) distribution

Weights				DISSE	KLD
Weibull	Loglogistic	Lognormal	Gamma		
0.6855028	−−	−−	0.3144972	0.1486238	1.158258
0.2081505	0.2756152	0.2366610	0.2795733	0.2514906	2.46644
−−	0.2882678	0.3057394	0.4059928	0.2660464	2.724732

Table 3.7: Stacking weights, DISSE, and KLD for stacked-survival-function estimator under a Weibull(0.75, 3) distribution

Weights				DISSE	KLD
Weibull	Loglogistic	Lognormal	Gamma		
0.6109639	−−	−−	0.3890361	0.164746	1.207803
0.2591137	0.1087186	0.2329598	0.3992079	0.2120041	3.378248
−−	0.1119335	0.3379473	0.5501193	0.2227059	3.982036

Table 3.8: Stacking weights, DISSE, and KLD for stacked-density-function estimator under a Weibull(0.75, 3) distribution

of the stacked-survival-function and the stacked-density-function estimators in the absence of any single correct parametric model. Fifty datasets of sample size 125 with 30% random right-censoring rate are generated, where half of the data comes from a Weibull(2,2) distribution and the other

half comes from a Gamma(2,2) distribution. We consider two sets of candidate models for stacking: a Weibull-Gamma stacking model and a stacking model including Weibull, Log-Logistic, Log-Normal, and Gamma models. The stacking weights, DISSE, and KLD for the two estimators are reported in Table 3.9. Both estimators have similar performance as measured by DISSE, while the stacked-density-function estimator has a slightly better KLD. However, in the case of stacking only Weibull and Gamma models, the stacking weights estimated by the stacked-survival-function estimator are much closer to the underlying 50-50 truth.

Estimator	Stacking Weights				DISSE	KLD
	Weibull	Log-Logistic	Log-Normal	Gamma		
stacked-survival-function estimator	0.440	−−	−−	0.560	0.036	0.109
	0.227	0.168	0.217	0.387	0.021	0.065
stacked-density-function estimator	0.333	−−	−−	0.667	0.036	0.104
	0.300	0.109	0.304	0.287	0.019	0.064

Table 3.9: Stacking weights, DISSE, and KLD for the stacked-survival-function estimator and the stacked-density-function estimator under a mixed distribution. Both estimators utilize only the residual lifetime data.

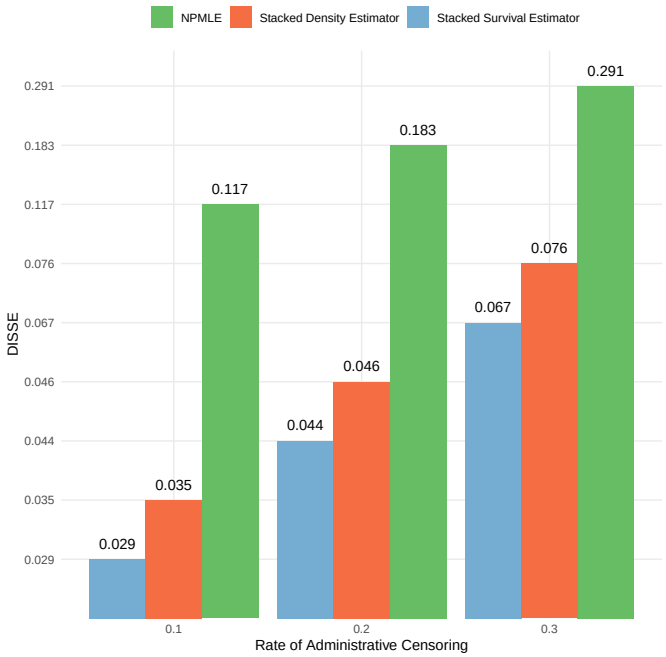
3.3 Evaluation for A Mixture Model Approach

We generate forward recurrence time data from a Weibull (2,2) distribution as described in Section 3 to evaluate the performance of the mixture model approach. 50 datasets were simulated, each of sample size of 125. The parametric nature of this approach excludes the inclusion of the NPMLE. Furthermore, the number of parameters present in the mixture model poses difficulty for estimation due to its high dimensionality. For example, we are not able to obtain optimized parameters when we use all four parametric families–Weibull, Lognormal, Loglogistic, and Gamma–as components for the mixture model, which involves 12 parameters in total. Moreover, the estimated parameters for individual parametric models could be unreasonable, which results in the platform-like segments in Figure 3.3. As suggested by Table 3.10, the model performs poorly except for the Weibull-Gamma mixture case, which still has a large variance and suffers from unrealistic bumps compared with the results in Table 3.4. Additional simulations presented in the supplementary materials suggest this problem persists even if the models contained in the mixture come from the same parametric family. However, as Figure S1 suggests, switching back to using

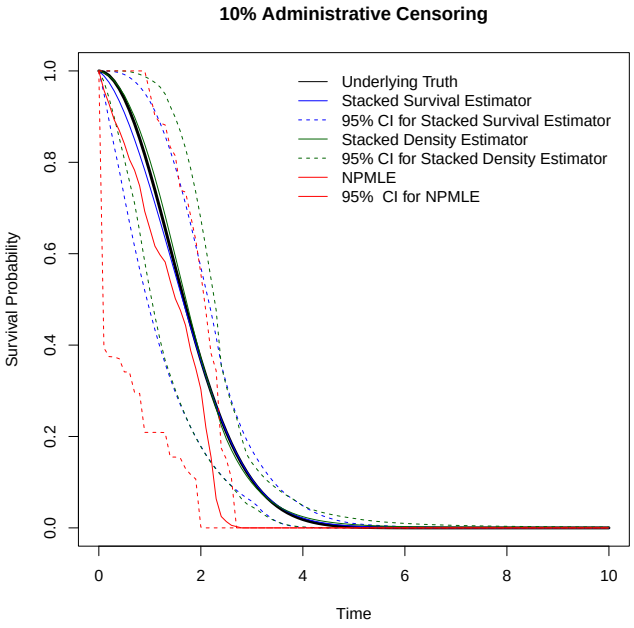
full survival data rather than the residual lifetime data avoids the problem, which is confirmed by previous literature [6, 7, 14]. It could be that the bumpiness when only forward recurrence time data is used is a result of numerical issues during the optimization process or concavity-related issues, as discussed in the supplementary materials.

	Weibull-Gamma	Lognormal-Loglogistic-Gamma
DISSE	0.082	0.281

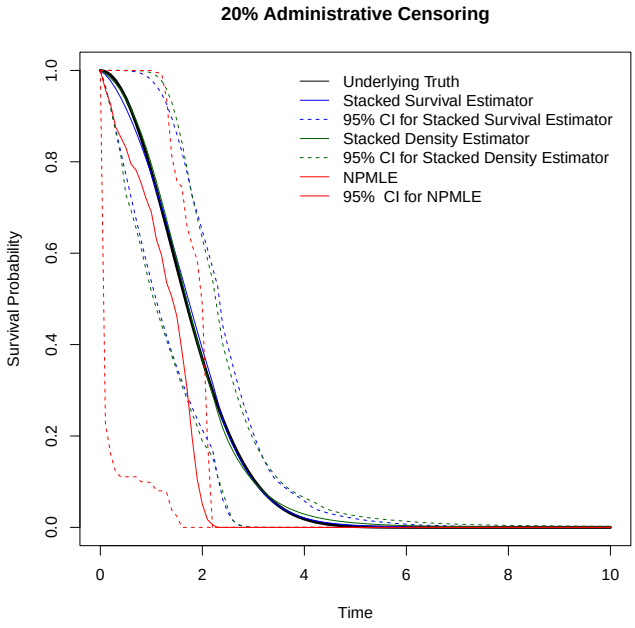
Table 3.10: DISSE of mixture models using residual lifetime only under Weibull (2,2) truth


NPMLE
Stacked Density Estimator
Stacked Survival Estimator
DISSE
0.291
0.183
0.117
0.076
0.067
0.046
0.044
0.035
0.029
0.1
0.2
0.3
Rate of Administrative Censoring

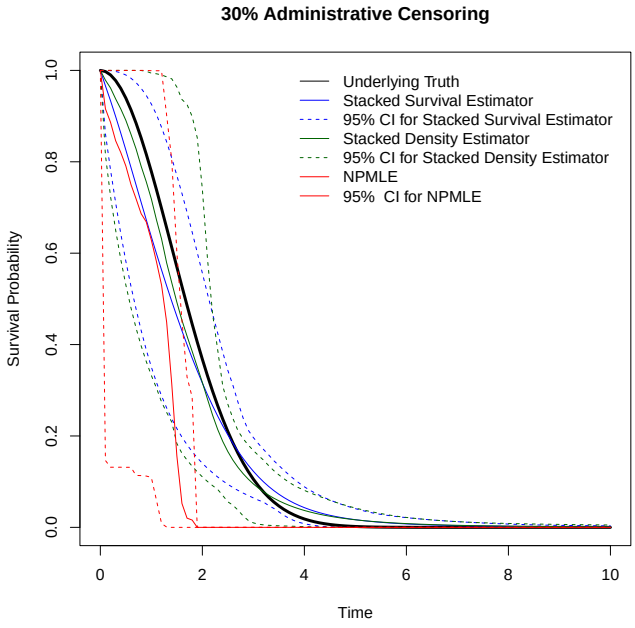
(a)


10% Administrative Censoring
Underlying Truth
Stacked Survival Estimator
95% CI for Stacked Survival Estimator
Stacked Density Estimator
95% CI for Stacked Density Estimator
NPMLE
95% CI for NPMLE
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
Time

(b)



(c)



(d)

Figure 3.1: (a) MSE for NPMLE, stacked survival estimator, and stacked density estimator under increasing rates of administrative censoring; (b)-(d) predicted survival curves for NPMLE and the two stacked estimators under 10% , 20 %, and 30% administrative censoring. Both stacking approaches outperform the NPMLE in the presence of administrative censoring.

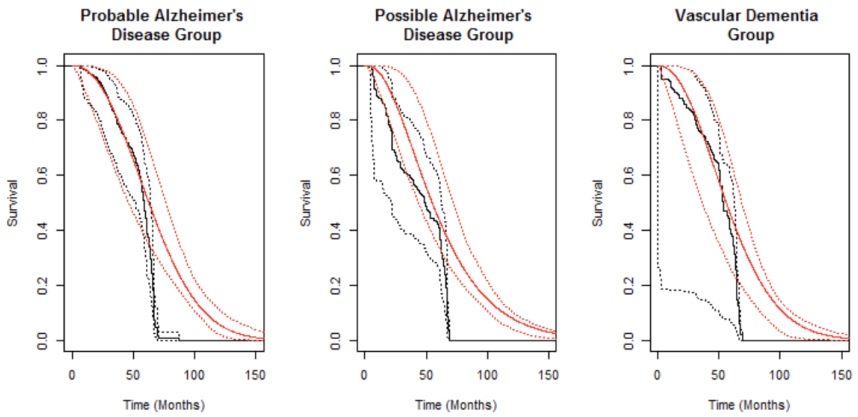
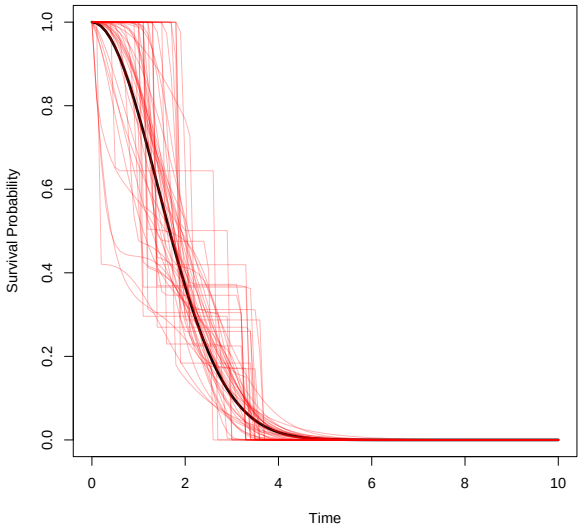
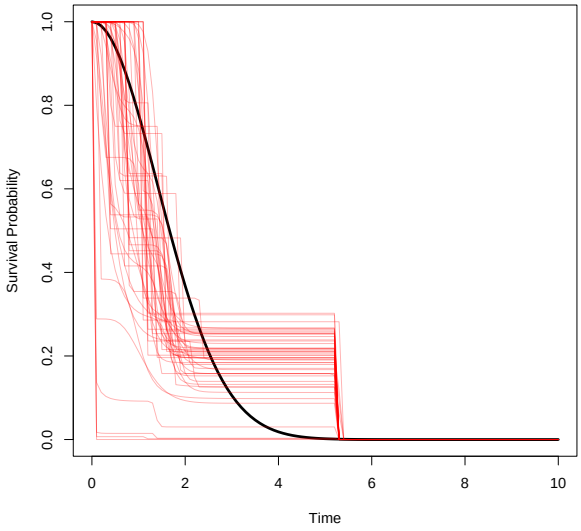


Figure 3.2: NPMLE (in black) and the stacked estimator (in red) with 95% bootstrapped pointwise confidence interval for three dementia subgroups in CSHA [15]



(a) Weibull-Gamma mixture model



(b) Lognormal-Loglogistic-Gamma mixture model

Figure 3.3: Survival curves estimated by mixture models that include (a) Weibull and Gamma models and (b) Lognormal, Loglogistic, and Gamma models, excluding the correct parametric Weibull distribution; data was generated under a Weibull(2,2) distribution

Chapter 4

Discussion

4.1 Alternative Estimation Methods

In comparison to the ad-hoc weight transferring involved in the stacked-survival-function estimator and the stacked-density-function estimator discussed throughout our work, Equation (2.1) and Equation (2.2) allow alternative estimators as described below:

1. **Stacking Forward Survival with Inversion**

   Once a set of weights are optimized via Brier Score to obtain a linear combination of forward survival functions $\hat{S}_{fwrd}(t) = \sum_{k=1}^{m} \alpha_i \hat{S}_{k,fwrd}(t)$,rather than transferring the weights to combine the underlying survival functions, the relationship $S(t) = \frac{f_{fwrd}(t)}{f_{fwrd}(0)}$ can be used again.

   Specifically, since

   $$\hat{f}_{fwrd}(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\hat{S}_{fwrd}(t) = -\sum_{k=1}^{m} \alpha_i \hat{f}_{k,fwrd}(t),$$

   we can define an estimator as follows:

   $$\hat{S}(t) = \frac{\hat{f}_{fwrd}(t)}{\hat{f}_{fwrd}(0)} = \frac{\sum_{k=1}^{m} \alpha_i \hat{f}_{k,fwrd}(t)}{\sum_{k=1}^{m} \alpha_i \hat{f}_{k,fwrd}(0)}$$

2. **Stacking Forward Density with Inversion**

   Corresponding to the stacked-density-function estimator described in Section 2.5.2, after estimating $\hat{f}_{fwrd}(t) = \sum_{k=1}^{m} \alpha_i \hat{f}_{k,fwrd}(t)$ via the EM algorithm, we can use $S(t) = \frac{f_{fwrd}(t)}{f_{fwrd}(0)}$ to estimate $\hat{S}(t)$.

   Though these two approaches avoid the ad-hoc transferring of weights, we lose the clear interpretability of the methods in Sections 2.5.1 and 2.5.2, since the estimated $\hat{S}(t)$ can no longer be written as a linear combination of individual survival functions $\hat{S}_i(t)$. There, these two methods are not recommended for practice.

   Similarly, with respect to the mixture model adopted in Section 3.3 where the underlying distribution is assumed to be a mixture, we can assume instead that the residual lifetime data follows a mixed distribution. In this case, write the residual lifetime density distribution is $f_{fwrd}(t) = \sum_{k=1}^{d} \sigma_i f_{k,fwrd}(t;\theta_k)$. It follows that the likelihood for right-censored residual lifetime data with sample size n is

$$\mathcal{L}(\theta,\sigma) = \prod_{i=1}^{n} \left(\sum_{k=1}^{d} \sigma_k f_{k,fwrd}(t_i;\theta_k) \right)^{\delta_i} \left(\int_{t_i}^{\infty} \sum_{k=1}^{d} \sigma_k f_{k,fwrd}(t;\theta_k) \mathrm{d}t \right)^{1-\delta_i}$$

   After optimizing for $\hat{\theta}$ and $\hat{\sigma}$, we could estimate the underlying survival distribution by:

$$S(t;\hat{\theta},\hat{\sigma}) = \frac{\sum_{k=1}^{d} \hat{\sigma}_k f_{k,fwrd}(t;\hat{\theta}_k)}{\sum_{k=1}^{d} \hat{\sigma}_k f_{k,fwrd}(0;\hat{\theta}_k)}.$$

   This approach is not adopted because its assumption is less natural and causes difficulty in interpretation. That being said, future work could investigate the performance of each of these alternatives.

4.2 Performance of Model-Mixture Approach

We observe that mixture models produce sub-optimal performances when we base the estimation only on forward recurrence times. We hypothesize

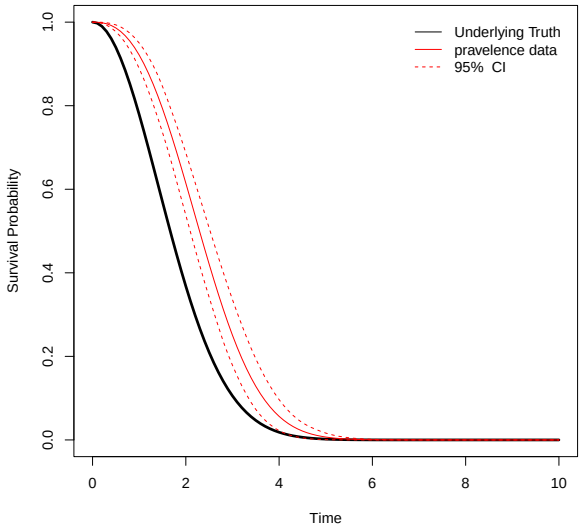
two reasons: the complex forms of likelihood functions and high dimensionality. Equation (2.4) suggests that the likelihood function we use has a complicated form that involves integrating survival functions when the estimation is based solely on the information provided by forward recurrence times. Moreover, as the mixture model optimizes for both parameters for individual model candidates and mixture weights simultaneously, the dimension of the space the optimization occurs over rises drastically as the number of candidate models increases. The high dimensionality may contribute to non-identifiability, where there is no single set of parameters that could give optimal performance.

4.3 Choice between the Two Stacked Approchaes

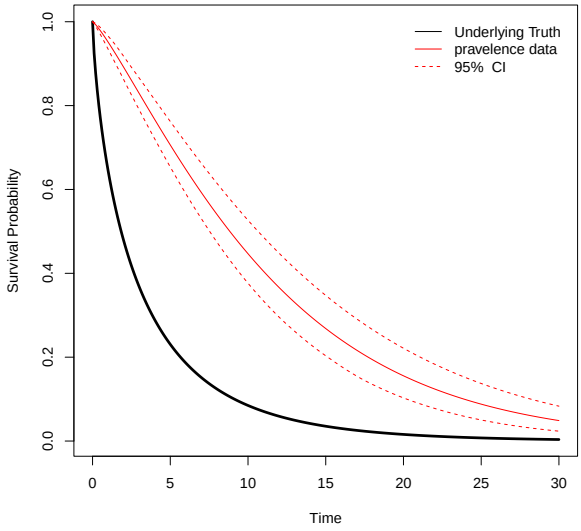
Simulation results in Section 3 suggest that both the stacked-survival estimator and the stacked-density estimator perform relatively well when applied to residual lifetime data. Since stacking weights in the stacked-survival estimator are optimized by minimizing the IPCW Brier Score, while in the stacked-density estimator are obtained via maximizing likelihoods, we adopted both the DISSE and the KLD as metrics to evaluate their performances. Though often the stacked-survival estimator generates a smaller DISSE as expected, we also observed in Table 3.7 and Table 3.8 that the stacked-survival estimator sometimes gives a slightly higher DISSE. We conjecture that this might be due to the fact that our stacked-survival procedure only optimizes the IPCW Brier Score on nine equally spaced points for computational efficiency, resulting in stacked weights that deviate from the true optimal value. That having been said, based on the simulation results, both estimators give comparable performances in different simulation settings. In terms of computational efficiency, both stacking methods require estimating individual candidate models. However, simulation results suggest that optimizing stacking weights for the stacked-survival estimator runs roughly three times faster than the stacked-density estimator. Optimizing fifty datasets each of a sample size of 125 requires approximately 20 seconds for the former while 60 seconds for the latter using CPU. Thus, either method could be used to estimate survival data collected in prevalent cohort studies using only residual lifetime information, while the stacked-survival estimator is slightly superior in terms of running time.

4.4 On the Case of Decreasing Hazard

Simulation results suggest that all estimators, individual or stacked, all perform better in the presence of increasing hazards in comparison to decreasing hazards. Though a clear explanation of the reasons behind this observation is still under investigation, we hypothesize that this might be due to the length-biased survival data collected during prevalent cohort studies. Using the simulation scheme laid out in Section 3.2, we simulated fifty datasets, each containing 125 subjects with a 30% random censoring. Figure 4.1 shows the estimates and 95% confidence intervals using a Weibull model fitted to full prevalent data generated from a Weibull (2,2) distribution with an increasing hazard and a Weibull (0.75, 3) distribution with a decreasing hazard, respectively. We chose not to adjust for length bias in order to gauge the effects of length bias on estimation. It appears estimation suffers from more severe bias under a decreasing hazard than under an increasing hazard, which is reasonable since in the former case, a larger proportion of subjects with earlier failure events would be excluded from sampling. It is possible that the larger MSE or KLD we observed in data simulated under decreasing hazards in Section 3 is a result of remnant length bias that is more severe for survival data with decreasing hazards.



(a)



(b)

Figure 4.1: Weibull estimates unadjusted for length-bias and their 95% confidence intervals using simulated full survival data from prevalent study (red) (a) an increasing hazard using Weibull(2,2) and (b) a decreasing hazard using Weibull(0.75,3) with 30% random censoring. Under a decreasing hazard, the estimates are impacted more severely by length-bias.

Chapter 5

Concluding Remarks

Survival data collected in prevalent cohort studies are partially observed. Though residual lifetimes are directly observed, researchers have relied on self-reported onset times to estimate underlying survival distributions. However, in certain settings, self-reported onset times may not be reliable, and thus may introduce additional bias that produces undesirable estimates. This project exploited the relationship between the density of the forward recurrence time and the underlying survival distribution under the assumption of stationary onset times to avoid using self-reported onset times.

Though this relationship is well-known, improvements could be further achieved by combining non-parametric estimators and parametric estimators from different families together in order to obtain an estimator can enjoy the low variance of parametric models while avoiding model misspecifications. Previously, [15] developed the stacked-survival estimator to address this issue. This study investigated the alternative option to stack density functions rather than survival functions and explored the possibility to apply a mixture model under the setting of residual-lifetime-based survival estimation. Results have shown that the stacked survival estimator outperforms NPMLE alone, especially in the presence of high administrative censoring. Simulation results also suggested that the stacked survival estimator and the stacked density estimator have similar performances in terms of DISSE and KLD. Though previous literature has applied a mixture model for full survival data, our results suggest that mixture models are not suitable for estimations based on residual lifetimes.

Lastly, we acknowledge that the stacking methods discussed in this

work do not incorporate covariates to predict survival times. Conventionally, covariate effects in survival analysis are estimated by parametric models or semi-parametric models such as the Cox model, though [8] proposed a method for covariate-adjusted non-parametric survival curve estimation. It remains to be investigated how to incorporate covariates for the prediction of survival time using forward-recurrence-time-based estimation procedures.

Bibliography

[1] Masoud Asgharian, David B. Wolfson, and Xun Zhang. "Checking Stationarity of the Incidence Rate Using Prevalent Cohort Survival Data". In: *Statistics in Medicine* 25.10 (May 30, 2006), pp. 1751–1767. ISSN: 02776715, 10970258. DOI: 10.1002/sim.2326. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.2326 (visited on 02/23/2022).

[2] Leo Breiman. "Stacked Regressions". In: *Machine Learning* 24.1 (July 1996), pp. 49–64. ISSN: 0885-6125, 1573-0565. (Visited on 03/23/2022).

[3] K. C. G. Chan, Y. Q. Chen, and C.-Z. Di. "Proportional Mean Residual Life Model for Right-Censored Length-Biased Data". In: *Biometrika* 99.4 (Dec. 1, 2012), pp. 995–1000. ISSN: 0006-3444, 1464-3510. (Visited on 03/23/2022).

[4] David Roxbee Cox. *Renewal theory*. Methuen, 1962.

[5] L. Denby and Y. Vardi. "The Survival Curve with Decreasing Density". In: *Technometrics* 28.4 (Nov. 1986), p. 359. ISSN: 00401706. DOI: 10.2307/1268985. JSTOR: 1268985.

[6] Ülkü Erişoğlu, Murat Erişoğlu, and Hamza Erol. "A Mixture Model Of Two Different Distributions Approach To The Analysis Of Heterogeneous Survival Data". In: (June 29, 2011). (Visited on 03/24/2022).

[7] V. T. Farewell. "The Use of Mixture Models for the Analysis of Survival Data with Long-Term Survivors". In: *Biometrics* 38.4 (Dec. 1982), p. 1041. ISSN: 0006341X. JSTOR: 2529885.

[8] Honghua Jiang et al. "Covariate-Adjusted Non-Parametric Survival Curve Estimation". In: *Statistics in Medicine* 30.11 (May 20, 2011), pp. 1243–1253. ISSN: 02776715. DOI: 10.1002/sim.4216. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.4216 (visited on 03/05/2022).

[9] N. Keiding. "Estimating Time to Pregnancy from Current Durations in a Cross-Sectional Sample". In: *Biostatistics* 3.4 (Dec. 1, 2002), pp. 565–578. ISSN: 14654644, 14684357. (Visited on 04/18/2022).

[10] Niels Keiding et al. "The Current Duration Approach to Estimating Time to Pregnancy: Current Duration Approach to TTP". In: *Scandinavian Journal of Statistics* 39.2 (June 2012), pp. 185–204. ISSN: 03036898. (Visited on 03/23/2022).

[11] Niels Keiding et al. "The Current Duration Approach to Estimating Time to Pregnancy: Current Duration Approach to TTP". In: *Scandinavian Journal of Statistics* 39.2 (June 2012), pp. 185–204. ISSN: 03036898. (Visited on 04/18/2022).

[12] Sotirios Losidis and Konstadinos Politis. "Moments of the Forward Recurrence Time in a Renewal Process". In: *Methodology and Computing in Applied Probability* 22.4 (Dec. 2020), pp. 1591–1600. ISSN: 1387-5841, 1573-7713. (Visited on 03/23/2022).

[13] Karen Lostritto, Robert L. Strawderman, and Annette M. Molinaro. "A Partitioning Deletion/Substitution/Addition Algorithm for Creating Survival Risk Groups". In: *Biometrics* 68.4 (Dec. 2012), pp. 1146–1156. ISSN: 0006341X. (Visited on 04/11/2022).

[14] Gj McLachlan and Dc McGiffin. "On the Role of Finite Mixture Models in Survival Analysis". In: *Statistical Methods in Medical Research* 3.3 (Oct. 1994), pp. 211–226. ISSN: 0962-2802, 1477-0334. (Visited on 03/24/2022).

[15] James H. McVittie et al. "Stacked Survival Models for Residual Lifetime Data". In: *BMC Medical Research Methodology* 22.1 (Dec. 2022), p. 10. ISSN: 1471-2288. DOI: 10.1186/s12874-021-01496-3. URL: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-021-01496-3 (visited on 03/05/2022).

[16] B. L. S. Prakasa Rao. "Estimation of a Unimodal Density". In: *Sankhyā: The Indian Journal of Statistics* 31.1 (Mar. 1969), pp. 23–36. DOI: https://www.jstor.org/stable/25049557.

[17] Padhraic Smyth and David Wolpert. "Linearly Combining Density Estimators via Stacking". In: *Machine Learning* 36.1/2 (July 1999), pp. 59–83. ISSN: 08856125. DOI: 10.1023/A:1007511322260. URL: http://link.springer.com/10.1023/A:1007511322260 (visited on 02/23/2022).

[18] Y. Vardi. "Nonparametric Estimation in Renewal Processes". In: *The Annals of Statistics* 10.3 (Sept. 1, 1982). ISSN: 0090-5364. (Visited on 03/23/2022).

[19] Ted Westling and Marco Carone. "A Unified Study of Nonparametric Inference for Monotone Functions". In: *The Annals of Statistics* 48.2 (Apr. 1, 2020). ISSN: 0090-5364. (Visited on 04/18/2022).

[20] Andrew Wey, John Connett, and Kyle Rudser. "Combining Parametric, Semi-Parametric, and Non-Parametric Survival Models with Stacked Survival Models". In: *Biostatistics (Oxford, England)* 16.3 (July 2015), pp. 537–549. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxv001. pmid: 25662068.

[21] David H. Wolpert. "Stacked Generalization". In: *Neural Networks* 5.2 (Jan. 1992), pp. 241–259. ISSN: 08936080. (Visited on 03/23/2022).

Appendix A

Supplementary material

A.1 Additional Simulations for Comparison between Two Stacked Estimators

A.1.1 Data Generated from a Gamma (2,1) Distribution with an Increasing Hazard

Model	MSE	KLD
Weibull	0.02658461	0.1138012
Loglogistic	0.06089398	0.3611299
Lognormal	0.03772015	0.3773242
Gamma	0.03360495	0.1368174

Table S1: DISSE, and KLD for individual parametric estimators under a Gamma (2,1) distribution

Weights				MSE	KLD
Weibull	Loglogistic	Lognormal	Gamma		
0.4335266	−−	−−	0.5664734	0.03179796	0.1226104
0.3182486	0.3070779	0.1729068	0.2017667	0.03655998	0.1875371
0.4839072	0.2978241	0.2182687	−−	0.03527199	0.1657934

Table S2: Stacking weights, DISSE, and KLD for stacked-survival-function estimator under a Gamma(2, 1) distribution

Weights				MSE	KLD
Weibull	Loglogistic	Lognormal	Gamma		
0.5361875	−−	−−	0.4638125	0.03033129	0.08992327
0.4729728	0.1830774	0.2482173	0.0957324	0.03155281	0.1299453
0.5382773	0.1858990	0.2758237	−−	0.03137829	0.1301424

Table S3: Stacking weights, DISSE, and KLD for stacked-density-function estimator under a Gamma(2, 1) distribution

A.1.2 Data Generated from a Gamma (0.8,5) Distribution with a Decreasing Hazard

Model	MSE	KLD
Weibull	2.105723	4.251685
Loglogistic	3.473024	23.08009
Lognormal	2.944581	109.9397
Gamma	2.077842	6.262271

Table S4: DISSE, and KLD for individual parametric estimators under a Gamma (0.8,5) distribution

Weights				MSE	KLD
Weibull	Loglogistic	Lognormal	Gamma		
0.4300011	−−	−−	0.5699989	2.072471	4.706692
0.27769444	0.26727826	0.05555535	0.39947196	2.440104	6.849669
0.63363418	0.27347327	0.09289255	−−	2.549522	7.360706

Table S5: Stacking weights, DISSE, and KLD for stacked-survival-function estimator under a Gamma(0.8, 5) distribution

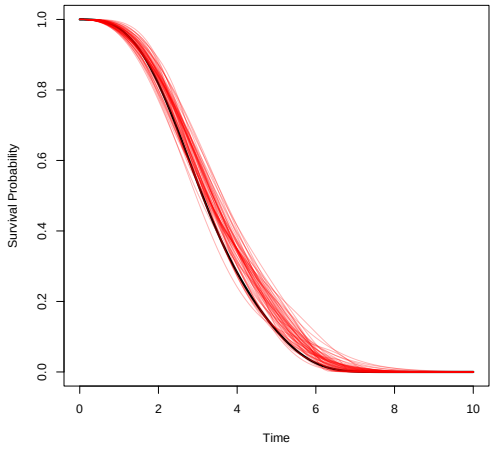
A.2 Abnormal Behavior of Mixture Models When residual lifetime Data is Used

To identify the cause of the bumpiness observed in the mixture models, we simulated fifty datasets from a mixture distribution of Weibull(3,3) and Weibull(5,5) using a sample size of 1000 and a right-censoring rate of 0.15, where 80% data comes from Weibull(3,3) and 20% data comes from Weibull(5,5).

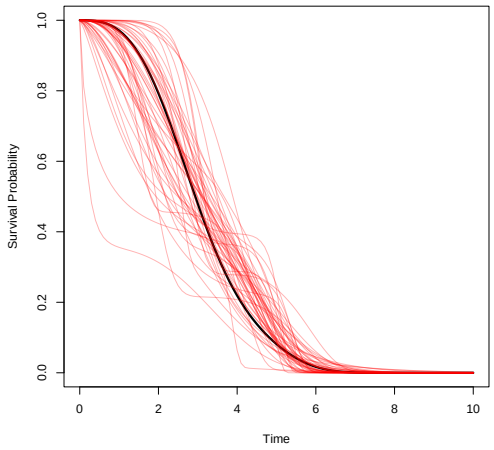
Weights				MSE	KLD
Weibull	Loglogistic	Lognormal	Gamma		
0.4798271	−−	−−	0.5201729	2.086596	4.948595
0.3059291	0.1305337	0.1928793	0.3706579	2.342791	6.958089
0.6551547	0.1446893	0.2001560	−−	2.442306	7.240001

Table S6: Stacking weights, DISSE, and KLD for stacked-density-function estimator under a Gamma(0.8, 5) distribution

The residual lifetime data is obtained using the same simulation scheme as described in section 3. We considered a mixture model with two Weibull components and a Weibull-Log-Logistic mixture model, and compared the estimated survival curves using either the full survival data or the residual lifetime in Figure S1, which shows the estimated survival curves for fifty simulated datasets of sample size 125 under a 15% rate of random censoring. We simulated this set of data under a smaller rate of random censoring to focus on comparing the effects of using forward recurrence data versus fully observed survival data.



(a) Estimated survival curves by a Weibull-Weibull mixture model using full survival data



(b) Estimated survival curves by a Weibull-Weibull mixture model using only residual lifetime data

Figure S1: Switching from using the full survival data to the residual lifetime data causes the bumpiness observed in mixture models despite different selections of mixture components. Data is generated from a mixture of Weibull(3,3) and Weibull(5,5) distribution

A.3 Codes Availability

All codes and resources used in this project can be found at `https://www.dropbox.com/sh/ct1qkqn3m7z10j7/AAA_FHKqhfYLjHhKxOVbNMgCa?dl=0`. Scripts for the stacked survival estimator are based on [15].