Semiparametric likelihood inference for heterogeneous survival data under double truncation based on a Poisson birth process

Achim Dörre¹

Received: 31 December 2020 / Revised: 1 May 2021 / Accepted: 17 May 2021 / Published online: 6 July 2021
© The Author(s) 2021

Abstract

We study a selective sampling scheme in which survival data are observed during a data collection period if and only if a specific failure event is experienced. Individual units belong to one of a finite number of subpopulations, which may exhibit different survival behaviour, and thus cause heterogeneity. Based on a Poisson process model for individual emergence of population units, we derive a semiparametric likelihood model, in which the birth distribution is modeled nonparametrically and the lifetime distributions parametrically, and define maximum likelihood estimators. We propose a Newton–Raphson-type optimization method to address numerical challenges caused by the high-dimensional parameter space. The finite-sample properties and computational performance of the proposed algorithms are assessed in a simulation study. Personal insolvencies are studied as a special case of double truncation and we fit the semiparametric model to a medium-sized dataset to estimate the mean age at insolvency and the birth distribution of the underlying population.

Keywords Double truncation · Survival analysis · Semiparametric models · Maximum likelihood

1 Introduction

In survival analysis, missing data are a central notion, as full observability of all units is rarely achievable. Consequently, statistical methods and approaches are designed to consider the various types of missing information to obtain valid inference. The

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s42081-021-00128-w>.

✉ Achim Dörre
achim.doerre@gmail.com

¹ Department of Economics, University of Rostock, Rostock, Germany

Fig. 1 Illustration of the truncation mechanism. Only units that fail within the data collection period are sampled (solid circles). Units that fail before or after the data collection period (empty circles) are not sampled

situation we address in the following is at the intersection of survival analysis and retrospective studies, where truncated data result from sampling only failed units within a certain data collection period.

Example 1 Consider personal insolvencies in Germany (see also Sect. 5). When a person becomes insolvent in Germany, individual information on the person is made available due to legal processes. Since the information is accessible for a limited amount of time through a public website, only units with their insolvency event occurring within the data collecting period can be observed. We have collected data from this source from March 2018 to December 2018 and sampled data on insolvent persons born between 1920 and 2000 in all 16 federal states of Germany. Figure 1 illustrates this truncation mechanism.

This sampling design is a special case of random double truncation, as the lifetimes are truncated if they are either too short or too long. Random double truncation may occur as an unavoidable characteristic of data collection procedures or may be chosen deliberately as a cost-effective alternative to other sampling designs. In the example above, the stochastic process of births over time poses a specific component of the truncation setup, which is addressed here.

Random double truncation is relevant to many areas including medicine (see Lagakos et al. 1988; Scheike and Keiding 2006; Moreira et al. 2014), astronomy (see Efron and Petrosian 1999), economics and engineering (see Lawless and Kalbfleisch 1992; Ye and Tang 2016). In recent years, attention to this area of research has been growing (see Emura and Konno 2012; Zhang 2015; Emura et al. 2017; Dörre and Emura 2019; de Uña-Álvarez 2020). This is partly due to the fact that administrative and economically collected data are often observational and frequently subject to selective criteria or constraints which can be posed in a truncation framework.

Among the various existing models and methods for doubly truncated data that have been investigated, the nonparametric maximum likelihood estimator under random double truncation (NPMLE, see Shen 2010a; Moreira and Uña-Álvarez 2010; Emura et al. 2015) has perhaps been studied most extensively. The NPMLE is generally applicable under double truncation, but can be computationally demanding and

requires caution regarding identifiability aspects (see Shen 2010a) and the existence of the estimator (see Xiao and Hudgens 2019).

Parametric models have been developed partly to address the drawbacks of the nonparametric approach (see Emura et al. 2017; Dörre and Emura 2019). In Dörre (2020), fully parametric models have been studied in a setting similar to the one studied here. Finally, semi-parametric models (see Wang 1989; Shen 2010b) have been proposed in the context of double truncation.

Most existing models for doubly truncated data are primarily designed for homogeneous populations and can not readily be extended to heterogeneous populations in an efficient manner. However, heterogeneity may be an immanent feature of survival data. For instance, in the example described above, it can not be assumed that the underlying lifetime distributions of people in different federal states are identical. While existing approaches such as the NPMLE could be applied to each subpopulation separately, this procedure would not be efficient if the subpopulations have common structural properties.

By incorporating available information on the structure of the study population, we derive a semiparametric method for survival data under double truncation. The survival data are modeled as heterogeneous in that the population may consist of separate subpopulations with potentially different lifetime distributions.

In Sect. 2, we specify the sampling setup and parameter space, and derive a corresponding likelihood model in order to define proper maximum likelihood estimates. By reconsidering the model structure, we address asymptotic properties and discuss notions of point processes and inverse probability weighting in Sect. 3. In Sect. 4, finite-sample properties of the proposed maximum likelihood estimates and numerical optimization methods are examined by means of simulation studies. We apply the methods to study personal insolvencies in Germany in Sect. 5.

2 Model and methods

2.1 General framework

We suppose that the units under investigation experience two subsequent events, generically called 'birth' and 'failure', throughout time. The meaning of these events may differ with respect to the particular context, and constitute a survival analysis setup, where the 'lifetime' is defined as the duration between birth and failure. Furthermore, we suppose that the population of interest is composed of a finite number m of distinct subpopulations, which may exhibit different lifetime distributions.

Units emerge throughout time according to stochastic processes, yielding random subpopulation sizes N_j , $j = 1, \dots, m$. Each unit is born at a random time b and has a random lifetime y . The corresponding failure event therefore occurs at $b + y$. The stochastic process describing the birth of units implicitly corresponds to a distribution $G(\cdot)$ of births in the population.

Due to the sampling setup, only units that fail within the data collection interval $[\tau^L, \tau^R]$ are observed. Therefore, any unit fulfilling the selection criterion

$$\tau^L \leq b + y \leq \tau^R \quad (1)$$

is being sampled, and for sampled units we assume that full information is available, i.e. birth b , lifetime y and the variable $x \in \{1, \dots, m\}$ indicating the subpopulation to which the unit belongs (see Fig. 2). Units not fulfilling this selection criterion are truncated in the sense that no information about those units is available.

The sample consists of $n \leq N$ units, and is subdivided into random subsamples of sizes n_j , $j = 1, \dots, m$, according to the considered subpopulations. Note that the subpopulation sizes N_j are unobservable, whereas the subsample sizes n_j are observed. We denote the data as $D_n = \{(b_i, x_i, y_i) : i = 1, \dots, n\}$.

Throughout, we make the following assumptions:

- (A) For each $j = 1, \dots, m$, the units in the j th subpopulation emerge independently of all other subpopulations and stochastically within a common fixed time interval $[s, t]$ according to a Poisson process with intensity function $b \mapsto \lambda_j(b) > 0$, $b \in [s, t]$.
- (B) The intensity functions are proportional, i.e. there is a function $b \mapsto \lambda_0(b)$ and $a_j > 0$, $j = 1, \dots, m$, such that $\lambda_j(b) = a_j \lambda_0(b)$.
- (C) The lifetime distributions are continuous and parametric, i.e. for each $j = 1, \dots, m$, there is a true parameter θ_j such that $f_j(\cdot | \theta_j)$ is the density function of y conditional on $x = j$.
- (D) The truncation mechanism is such that a unit (b, x, y) from the population is sampled if and only if it fails within $[\tau^L, \tau^R]$, i.e. if $\tau^L \leq b + y \leq \tau^R$.
- (E) The lifetimes y and births b are independent.

The notation introduced in Assumption (C) reflects that different subpopulations may possess different distribution classes. Furthermore, we write $\theta = (\theta_1, \dots, \theta_m)$, where needed. Regarding Assumption (B), we generally set $a_1 = 1$, such that $\lambda_0(\cdot) \equiv \lambda_1(\cdot)$, to avoid non-identifiability of the proportionality factors a_j , $j = 2, \dots, m$.

Assumption (E) is necessary in some derivations below. In particular, this assumption requires that the lifetime distribution in any subpopulation does not change over the course of time. In specific applications, this assumption may be questionable. For testing the (quasi-)independence assumption, several approaches have been studied; see Tsai (1990) and Martin and Betensky (2005). Recent papers which address the aspect of the independence assumption within the truncation context are Chiou et al. (2018, 2019) and Moreira et al. (2021).

2.2 Proportional birth intensity model

The distribution function of a randomly chosen birth b generated by a Poisson process with intensity function $\lambda(\cdot)$ is equal to

$$G(b | \lambda(\cdot)) = \frac{\int_s^b \lambda(c) \, dc}{\int_s^t \lambda(c) \, dc}, \quad b \in [s, t]. \quad (2)$$

Fig. 2 The general framework of the sampling setup. Units emerge in subpopulations according to a Poisson process with proportional intensity functions. Depending on the individual time of birth b and lifetime y , a unit is being sampled if it fulfills the selection criterion $\tau^L \leq b + y \leq \tau^R$

By Assumption (B) and Eq. (2), it follows that there is a common birth distribution for all considered subpopulations, such that the birth distribution functions of the subpopulations fulfill $G_j \equiv G$ for all $j = 1, \dots, m$.

By Assumptions (A) and (D), the observed units of any subpopulation j form a random point process with regard to the observed births b_i where $x_i = j$. It can be shown that the observed point process of each subpopulation j is also a Poisson process. Furthermore, the sum of all observed birth processes is a Poisson process, too, by the independence stated in Assumption (A). For any subpopulation $j = 1, \dots, m$, we denote the cumulative intensity of the latent birth process as $\Lambda_j := \int_s^t \lambda_j(b) db$ and the cumulative intensity of the observed birth process as $\Lambda_j^{\text{obs}} := \int_s^t \int_{\mathbb{R}} \lambda_j(b) f_j(y|\theta_j) \mathbf{I}(\tau^L \leq b + y \leq \tau^R) dy db$.

The probability of observing one unit, conditional on it belonging to subpopulation j with given birth time b , denoted as $\pi_j(\theta_j|b)$, depends on the lifetime distribution of the corresponding subpopulation, and is equal to

$$\pi_j(\theta_j|b) = \int_{\tau^L - b}^{\tau^R - b} f_j(y|\theta_j) dy.$$

The intensity function of the observed point process of subpopulation j is equal to

$$\lambda_j^{\text{obs}}(b) = \pi_j(\theta_j|b) \lambda_j(b), b \in [s, t]. \quad (3)$$

Similarly, $\Lambda_j^{\text{obs}} = \Lambda_j \bar{\pi}_j(\theta_j|G)$, where

$$\bar{\pi}_j(\theta_j|G) = \int_s^t \pi_j(\theta_j|b) dG(b) \quad (4)$$

is the probability of observing a random unit from subpopulation j , given birth distribution function G . To see this, note that

$$\begin{aligned}\Lambda_j^{\text{obs}} &= \int_s^t \int_{\mathbb{R}} \lambda_j(b) f_j(y|\theta_j) \mathbf{I}(\tau^L \leq b + y \leq \tau^R) \, dy \, db \\ &= \int_s^t \lambda_j(b) \int_{\mathbb{R}} f_j(y|\theta_j) \mathbf{I}(\tau^L \leq b + y \leq \tau^R) \, dy \, db \\ &= \int_s^t \lambda_j(b) \int_{\tau^L - b}^{\tau^R - b} f_j(y|\theta_j) \, dy \, db \\ &= \int_s^t \lambda_j(b) \pi_j(\theta_j|b) \, db\end{aligned}$$

Using Eq. (2), it follows that $\lambda_j(b)db = \Lambda_j dG(b)$, and therefore

$$\Lambda_j^{\text{obs}} = \int_s^t \pi_j(\theta_j|b) \Lambda_j \, dG(b) = \Lambda_j \bar{\pi}_j(\theta_j|G).$$

By $h(\cdot)$, we denote the density function of x , i.e. the probability of a unit from the population belonging to a specific subpopulation. It is a consequence of Assumption (B) that this density function only depends on $\mathbf{a} = (a_1, \dots, a_m)$, representing a multinomial experiment with

$$h(j|\mathbf{a}) = \frac{a_j}{\sum_{k=1}^m a_k}, \quad j = 1, \dots, m. \quad (5)$$

Furthermore, the sum of all latent birth processes itself forms a Poisson process, and has cumulative intensity $\Lambda_1 \sum_{j=1}^m a_j$. Similarly, it holds that the sum of all observed birth processes is a Poisson process, too, for which the cumulative intensity is equal to $\Lambda_1 \sum_{j=1}^m a_j \bar{\pi}_j(\theta_j|G)$. The probability of observing a single random unit from the population is equal to

$$\pi(G, \mathbf{a}, \theta) = \sum_{j=1}^m h(j|\mathbf{a}) \bar{\pi}_j(\theta_j|G) = \frac{\sum_{j=1}^m a_j \bar{\pi}_j(\theta_j|G)}{\sum_{j=1}^m a_j}, \quad (6)$$

where $\bar{\pi}_j(\theta_j|G)$ is defined as in Eq. (4). Finally, we state the following result, which is fundamental for constructing the likelihood function.

Lemma 1 *Under Assumptions (A)–(E), the point process $\eta(b, x, y) := \sum_{i=1}^n \mathbf{I}(b_i \leq b, x_i \leq x, y_i \leq y)$, which counts the observed units in the three-dimensional space $[s, t] \times \{1, \dots, m\} \times \mathbb{R}_{\geq 0}$, is a Poisson process with intensity function*

$$\mu(b, x, y) = \pi(G, \mathbf{a}, \theta) a_x \lambda_1(b) f_x(y|\theta_x) \mathbf{I}(\tau^L \leq b + y \leq \tau^R).$$

The cumulative intensity is equal to $\pi(G, \mathbf{a}, \theta) \Lambda_1 \sum_{j=1}^m a_j$.

Proof It can be shown by elementary calculations that $\eta(\cdot)$ is Poisson distributed with the corresponding intensity. Assumption (A) ensures independent increments of $\eta(\cdot)$. A detailed proof is omitted. $\square$

2.3 Likelihood function

Under Assumptions (A)–(E), the observed Poisson process has the density function (see Snyder and Miller 1991 and Theorem 3.1.1 in Reiss 1993):

$$f(D_n|G, \mathbf{a}, \boldsymbol{\theta}) = \left(\prod_{i=1}^n a_{x_i} \lambda_1(b_i) f_{x_i}(y_i|\boldsymbol{\theta}_{x_i}) \right) \exp \left\{ -\pi(G, \mathbf{a}, \boldsymbol{\theta}) \Lambda_1 \sum_{j=1}^m a_j \right\} \quad (7)$$

Note that Eq. (7) does not consist of iid factors and is therefore not ideal for likelihood inference. By conditioning on the random sample size n , however, we construct a likelihood function with proper iid terms. For more technical details in a similar model, see Dörre (2020).

To estimate the birth distribution function G nonparametrically, we place positive probability weights precisely on all observed births $b_1, \dots, b_n$. Specifically, we set $\Delta G_i = P(b = b_i)$ for each $i = 1, \dots, n$, and treat ΔG_i as parameters of the unknown birth distribution. The corresponding birth distribution results as $G(b) = \sum_{i=1}^n \Delta G_i \cdot I(b_i \leq b)$ for $b \in [s, t]$. By conditioning Eq. (7) on the random sample size n , we obtain the corresponding likelihood function

$$L(G, \mathbf{a}, \boldsymbol{\theta}|D_n) = \prod_{i=1}^n \frac{\Delta G_i \cdot h(x_i|\mathbf{a}) f_{x_i}(y_i|\boldsymbol{\theta}_{x_i})}{\pi(G, \mathbf{a}, \boldsymbol{\theta})}, \quad (8)$$

where $\Delta G = (\Delta G_1, \dots, \Delta G_n) \in [0, 1]^n$ with $\sum_{i=1}^n \Delta G_i = 1$, $\mathbf{a} = (a_1, \dots, a_m) \in \mathbb{R}_{>0}^m$ with $a_1 = 1$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_m) \in \Theta_1 \times \dots \times \Theta_m$. We define the maximum likelihood estimator $(\hat{G}^{(n)}, \hat{\mathbf{a}}, \hat{\boldsymbol{\theta}})$ as the maximizer of Eq. (8), i.e.

$$L(\hat{G}^{(n)}, \hat{\mathbf{a}}, \hat{\boldsymbol{\theta}}|D_n) = \max L(G, \mathbf{a}, \boldsymbol{\theta}|D_n).$$

Altogether, this model entails $n + m + \sum_{j=1}^m \dim(\Theta_j) - 2$ parameters, not counting a_1 and ΔG_n due to being subject to constraints. The model is semiparametric in the sense that the number of parameters increases linearly with growing sample size. The likelihood function Eq. (8) can be rearranged as

$$L(G, \mathbf{a}, \boldsymbol{\theta}|D_n) = \pi(G, \mathbf{a}, \boldsymbol{\theta})^{-n} \prod_{i=1}^n \Delta G_i \prod_{j=1}^m \left\{ \left(\frac{a_j}{\sum_{j=1}^m a_j} \right)^{n_j} \cdot \prod_{i:x_i=j} f_j(y_i|\boldsymbol{\theta}_j) \right\},$$

so that the log-likelihood function can be written as

$$\begin{aligned} \log L(G, \mathbf{a}, \boldsymbol{\theta} | D_n) = & -n \log \pi(G, \mathbf{a}, \boldsymbol{\theta}) + \sum_{i=1}^n \log \Delta G_i - n \log \left(\sum_{j=1}^m a_j \right) \\ & + \sum_{j=1}^m \left\{ n_j \log a_j + \sum_{i: x_i=j} \log f_j(y_i | \boldsymbol{\theta}_j) \right\}, \end{aligned}$$

where $\pi(G, \mathbf{a}, \boldsymbol{\theta})$ is obtained as in Eq. (6) and, following Eq. (4),

$$\bar{\pi}_j(\boldsymbol{\theta}_j | G) = \sum_{i=1}^n \pi_j(\boldsymbol{\theta}_j | b_i) \Delta G_i$$

for any distribution G with discrete jumps ΔG_i at $b_1, \dots, b_n$.

Note that the likelihood function Eq. (8) represents an inverse probability weighting (IPW) approach, where the probability weights $\pi(G, \mathbf{a}, \boldsymbol{\theta})$ are determined endogenously and equally for every observed unit. By conditioning on certain parts of the data, other inverse probability weighting likelihood functions are conceivable under the considered truncation scheme. For instance, for any $j = 1, \dots, m$, an estimator $\hat{\boldsymbol{\theta}}_j^c$ of $\boldsymbol{\theta}_j$ may be defined as the maximizer of

$$\tilde{L}_j(\boldsymbol{\theta}_j | D_n) := \prod_{i: x_i=j} \frac{f_j(y_i | \boldsymbol{\theta}_j)}{\pi_j(\boldsymbol{\theta}_j | b_i)}, \quad (9)$$

and $\tilde{L}_j$ is valid in the sense that $\hat{\boldsymbol{\theta}}_j^c \rightarrow \boldsymbol{\theta}_j$ as $n_j \rightarrow \infty$, which is not proved here. Using Eq. (8) has the advantage of leading to generally more precise estimates due to the stronger assumptions with regard to the birth distribution, while Eq. (9) has the advantage of not being vulnerable to misspecification with regard to the lifetime distribution in other subpopulations than j or the birth distribution G .

2.4 Birth intensity estimation

The proposed likelihood model Eq. (8) can be regarded as an inverse probability weighted approach, in which every sampled unit is weighted by the inverse probability of observing a single random unit from the population. Roughly speaking, every sampled unit corresponds to $\pi(G, \mathbf{a}, \boldsymbol{\theta})^{-1}$ hypothetical units in the population, out of which one single unit fulfills the selection criterion in Eq. (1) and has thus been sampled. Similarly, a sampled unit from the j th subpopulation with birth b_i corresponds to $\pi_j(\boldsymbol{\theta}_j | b_i)^{-1}$ hypothetical units before the selection. Based on this perspective and following Eq. (3), we define

$$\hat{\Lambda}_j := \sum_{i: x_i=j} \pi_j(\boldsymbol{\theta}_j | b_i)^{-1} \quad (10)$$

as an estimate of the cumulative birth intensity of the j th subpopulation.

The IPW perspective also offers an alternative approach to estimate the proportionality factors $a_2, \dots, a_m$. Due to the proportional intensity assumption, it holds that $\Lambda_j = a_j \Lambda_1$ for $j = 2, \dots, m$. One simple way to estimate the coefficients $a_2, \dots, a_m$ is therefore to use Eq. (10), by which

$$\hat{a}_j^{\text{simple}} = \frac{\hat{\Lambda}_j}{\hat{\Lambda}_1} = \frac{\sum_{i:x_i=j} \pi_j(\hat{\theta}_j|b_i)^{-1}}{\sum_{i:x_i=1} \pi(\hat{\theta}_1|b_i)^{-1}} \quad (11)$$

is obtained. Note also that this approach is feasible whenever a consistent estimate of θ is available, and no estimate of G is required. We show in Sect. 4, that this method leads to valid, yet inefficient estimates.

3 Large-sample properties

Studying the asymptotic behaviour of the proposed estimators is based on the sub-sample sizes growing indefinitely. We note, however, that $n_j \rightarrow \infty$ if and only if $\Lambda_j \rightarrow \infty$, and thus asymptotics can not be studied without changing the birth intensity function. As a consequence of this perspective, it is not possible to estimate Λ_j consistently in the sense that $|\hat{\Lambda}_j - \Lambda_j|$ converges to 0 for $n \rightarrow \infty$. Nevertheless, Eq. (2) shows that the birth distribution G , which we estimate, is invariant with respect to scaling the birth intensity function λ , as long as its shape is not changed.

Due to Assumptions (A)–(E), n_j is Poisson distributed with mean $\Lambda_j^{\text{obs}} = \Lambda_j \bar{\pi}_j(\theta_j|G)$. Therefore, when $\Lambda_j \rightarrow \infty$, it holds that $n_j \rightarrow \infty$, almost surely, and the expression

$$n_j^{-1} \sum_{i:x_i=j} \pi_j(\theta_j|b_i)^{-1}$$

converges to $E(\pi_j(\theta_j|b)^{-1}|x = j)$ by the law of large numbers, where $E(\cdot)$ is the conditional expectation with respect to the law of observed units in subpopulation j . It holds that

$$E(\pi_j(\theta_j|b)^{-1}|x = j) = \int_s^t \pi_j(\theta_j|b)^{-1} dG_j^{\text{obs}}(b),$$

where

$$dG_j^{\text{obs}}(b) = \frac{\pi_j(\theta_j|b)\lambda_j(b)}{\int_s^t \pi_j(\theta_j|c)\lambda_j(c) dc} = \pi_j(\theta_j|b)\lambda_j(b)/\Lambda_j^{\text{obs}}$$

is the density function of observed random births b in subpopulation j . Therefore,

$$E(\pi_j(\theta_j|b)^{-1}|x=j) = \frac{1}{\Lambda_j^{\text{obs}}} \int_s^t \lambda_j(b) db = \frac{\Lambda_j}{\Lambda_j^{\text{obs}}},$$

and thus

$$\lim_{\Lambda_j \rightarrow \infty} \frac{\hat{\Lambda}_j}{\Lambda_j} = \lim_{\Lambda_j \rightarrow \infty} \frac{n_j}{\Lambda_j^{\text{obs}}} = 1,$$

since n_j is Poisson distributed with parameter Λ_j^{obs} . This derivation shows that $\hat{\Lambda}_j$ as defined in Eq. (10) is a reasonable estimate of Λ_j . Consequently, $\hat{a}_j^{\text{simple}}$ as defined in Eq. (11) is expected to be a reasonable estimate of a_j .

To study the large-sample properties of the proposed estimators based on Eq. (8), we consider the asymptotic behaviour of the log-likelihood function. Define

$$Q_n(G, \mathbf{a}, \boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^n \{ \log f_{x_i}(y_i | \boldsymbol{\theta}_{x_i}) + \log h(x_i | \mathbf{a}) + \log \Delta G_i - \log \pi(G, \mathbf{a}, \boldsymbol{\theta}) \}$$

as the stochastic criterion function, which is maximized by the maximum likelihood estimator $(\hat{G}^{(n)}, \hat{\mathbf{a}}, \hat{\boldsymbol{\theta}})$. Suppose first that the true birth distribution G is known. A conditional criterion function can be defined as

$$Q_n(\mathbf{a}, \boldsymbol{\theta} | G) := \frac{1}{n} \sum_{i=1}^n \{ \log f_{x_i}(y_i | \boldsymbol{\theta}_{x_i}) + \log h(x_i | \mathbf{a}) - \log \pi(G, \mathbf{a}, \boldsymbol{\theta}) \}$$

and it holds due to the law of large numbers that

$$Q_\infty(\mathbf{a}, \boldsymbol{\theta} | G) := \lim_{n \rightarrow \infty} Q_n(\mathbf{a}, \boldsymbol{\theta} | G) = E_{\mathbf{a}, \boldsymbol{\theta}} \left[\log \left(\frac{f_x(y | \boldsymbol{\theta}_x) h(x | \mathbf{a})}{\pi(G, \mathbf{a}, \boldsymbol{\theta})} \right) \right], \quad (12)$$

almost surely, where $E_{\mathbf{a}, \boldsymbol{\theta}}$ denotes the expectation with respect to the distribution conditional on $\tau^L \leq b + y \leq \tau^R$. For known G , the density function of an observed random pair (x, y) under truncation is precisely $f_x(y | \boldsymbol{\theta}_x) h(x | \mathbf{a}) / \pi(G, \mathbf{a}, \boldsymbol{\theta})$, such that Eq. (12) in fact represents the Kullback–Leibler divergence with respect to $(\mathbf{a}, \boldsymbol{\theta})$ up to a constant additional term. It is well-known that this expression is uniquely maximized at the true parameter values $(\mathbf{a}, \boldsymbol{\theta})$ under appropriate regularity assumptions (see Lemma 5.35 in van der Vaart 1998).

Now, more generally, assume that some consistent estimator of G is available, i.e. an estimator $\hat{G}^{(n)}$ based on the data D_n such that

$$\|\hat{G}^{(n)} - G\|_\infty = \sup_{b \in [s, t]} |\hat{G}^{(n)}(b) - G(b)| \rightarrow 0,$$

almost surely, for $n \rightarrow \infty$. Reconsidering Eq. (4) and Eq. (6) and $0 < \bar{\pi}_j(\theta_j|G) \leq 1$ for every $j = 1, \dots, m$, it follows that

$$\left| \pi(\widehat{G}^{(n)}, \mathbf{a}, \boldsymbol{\theta}) - \pi(G, \mathbf{a}, \boldsymbol{\theta}) \right| \rightarrow 0, \quad n \rightarrow \infty.$$

Therefore, $|Q_n(\mathbf{a}, \boldsymbol{\theta}|\widehat{G}^{(n)}) - Q_n(\mathbf{a}, \boldsymbol{\theta}|G)| \rightarrow 0$ for any $\mathbf{a}$ and $\boldsymbol{\theta}$ and thus

$$\lim_{n \rightarrow \infty} Q_n(\mathbf{a}, \boldsymbol{\theta}|\widehat{G}^{(n)}) = E_{\mathbf{a}, \boldsymbol{\theta}} \left[\log \left(\frac{f_x(y|\boldsymbol{\theta}_x)h(x|\mathbf{a})}{\pi(G, \mathbf{a}, \boldsymbol{\theta})} \right) \right], \quad (13)$$

as in Eq. (12). It follows that $\widehat{\mathbf{a}}$ and $\widehat{\boldsymbol{\theta}}$ are consistent under appropriate regularity assumptions, if $\widehat{G}^{(n)}$ is a consistent estimator of G .

Assuming asymptotic normality of $\widehat{\boldsymbol{\gamma}} = (\widehat{\mathbf{a}}, \widehat{\boldsymbol{\theta}})$ with the variance given by the Fisher information, it is possible to derive asymptotic confidence intervals in a conventional way. To do so, we abbreviate $\boldsymbol{\gamma} = (\mathbf{a}, \boldsymbol{\theta})$ and define the matrix A_n of all negative partial derivatives of the log-likelihood function

$$A_{n,jk} = - \frac{\partial^2 \log L(\widehat{G}^{(n)}, \widehat{\mathbf{a}}, \widehat{\boldsymbol{\theta}}|D_n)}{\partial \gamma_j \partial \gamma_k}.$$

Given that A_n is a regular matrix, the asymptotic variance matrix of $\widehat{\boldsymbol{\gamma}}$ can be estimated by the inverse $\widehat{V}_n := A_n^{-1}$. The level $(1 - \alpha)$ confidence interval for the j th component of $\boldsymbol{\gamma}$ is then obtained as

$$\widehat{\gamma}_j \pm z_{1-\alpha/2} \sqrt{\widehat{V}_{n,jj}},$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution.

Similarly, confidence intervals for aspects of the latent distribution, say the mean lifetime $E(y|\boldsymbol{\theta}_j)$ in the j th subpopulation, can be defined. To do so, we first generate a large number M of iid draws

$$\boldsymbol{\gamma}^1, \dots, \boldsymbol{\gamma}^M \sim N(\widehat{\boldsymbol{\gamma}}, \widehat{V}_n).$$

An estimated level $(1 - \alpha)$ confidence interval of the mean lifetime is then obtained using the $\alpha/2$ and $1 - \alpha/2$ quantiles of $\{E(y|\boldsymbol{\theta}_j^1), \dots, E(y|\boldsymbol{\theta}_j^M)\}$.

4 Numerical implementation and simulation

4.1 Numerical implementation

To find the maximum likelihood estimate for given data D_n , we propose a Newton–Raphson-type algorithm. In this procedure, the stopping criterion is defined with

respect to a distance function which indicates the average change in an iteration with respect to the different parameter types. We define the distance function

$$d((\tilde{G}, \tilde{\mathbf{a}}, \tilde{\boldsymbol{\theta}}), (G, \mathbf{a}, \boldsymbol{\theta})) = \frac{1}{n} \sum_{i=1}^n |\Delta \tilde{G}_i - \Delta G_i| + \frac{1}{m} \sum_{j=1}^m |\tilde{a}_j - a_j| + \frac{1}{m} \sum_{j=1}^m \|\tilde{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|,$$

in which the number n of support points is incorporated into measuring the distance of two distribution functions $\tilde{G}$ and G . The following algorithm is proposed for determining the maximum likelihood estimate.

Algorithm MNS (Model-driven Newton-Raphson Search)

Step 1 Set $\epsilon > 0$ and initial values $a_j^0 = n_j/n_1$ and $\boldsymbol{\theta}_j^0$ for $j = 1, \dots, m$.

Step 2 For every $i = 1, \dots, n$, determine

$$N_i^0 = \pi_{x_i}(\boldsymbol{\theta}_{x_i}^0 | b_i)^{-1}$$

and set $\Delta G_i^0 = N_i^0 / \sum_{i=1}^n N_i^0$ for $i = 1, \dots, n$.

Step 3 Perform univariate Newton-Raphson steps for maximizing $\log L(G^0, \mathbf{a}, \boldsymbol{\theta} | D_n)$ for fixed G^0 with regard to all components of $\mathbf{a}$ and $\boldsymbol{\theta}$ until convergence. Denote the result as $(\mathbf{a}^1, \boldsymbol{\theta}^1)$.

Step 4 For every $i = 1, \dots, n$, determine

$$N_i^1 = \pi_{x_i}(\boldsymbol{\theta}_{x_i}^1 | b_i)^{-1}$$

and set $\Delta G_i^1 = N_i^1 / \sum_{i=1}^n N_i^1$ for $i = 1, \dots, n$. Set $v = 1$ and abbreviate $\boldsymbol{\gamma}^v := (\Delta G^v, \mathbf{a}^v, \boldsymbol{\theta}^v)$.

Step 5 Set $\tilde{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^v$. For each component k of $\tilde{\boldsymbol{\gamma}}$ except ΔG_n and a_1 , perform a damped univariate Newton-Raphson step as

$$\begin{aligned} \tilde{\boldsymbol{\gamma}}' &= \tilde{\boldsymbol{\gamma}} \\ \tilde{\gamma}'_k &= \tilde{\gamma}_k - \omega \frac{\partial \log L(\tilde{\boldsymbol{\gamma}} | D_n)}{\partial \gamma_k} \bigg/ \frac{\partial^2 \log L(\tilde{\boldsymbol{\gamma}} | D_n)}{\partial \gamma_k \partial \gamma_k} \\ \tilde{\boldsymbol{\gamma}} &= \tilde{\boldsymbol{\gamma}}' \end{aligned}$$

Step 6 Set $\boldsymbol{\gamma}^{v+1} = \tilde{\boldsymbol{\gamma}}$ and increase v by 1.

Step 7 If $d(\boldsymbol{\gamma}^v, \boldsymbol{\gamma}^{v-1}) \geq \epsilon$, return to Step 5; else stop and set $(\hat{G}^{(n)}, \hat{\mathbf{a}}, \hat{\boldsymbol{\theta}}) = \boldsymbol{\gamma}^v$.

For the initial values $\boldsymbol{\theta}_j^0$, we use the maximum likelihood estimates of $\boldsymbol{\theta}_j$ in the model without truncation. The key step in this algorithm is that the initial value G^0 is determined utilizing the initial values $\boldsymbol{\theta}_j^0$ and using Eq. (10). Intuitively, for every i , N_i^0 is the hypothetical number of population units among which the single unit i has been observed. By standardizing these hypothetical population numbers as done

in Step 4, the corresponding distribution function is derived. It turns out (see Sect. 4.2), that choosing ΔG^0 in this way remarkably reduces the computational runtime in comparison to alternative optimization strategies.

The damping factor $0 < \omega \leq 1$ is used to avoid obtaining invalid parameter values, such as $\Delta G_i \leq 0$ or $a_j \leq 0$ for any $j \in \{1, \dots, m\}$. Depending on the sample size, ω may need to be chosen close to zero initially. When $\mathbf{y}^v$ is sufficiently close to the maximum likelihood estimator, the damping factor can be dropped, i.e. $\omega = 1$. Similarly, the Newton–Raphson method in Step 3 of the algorithm MNS may potentially yield inadmissible values a_j due to numerical reasons. To avoid this issue, one may need to adjust the optimization to ensure robustness (e.g., using a damping factor).

4.2 Simulation studies

Simulation study on validity

We formulate the following simulation setup to study the finite-sample properties of the proposed method (see Table 1). Population units are born in the time interval $[s, t] = [0, 8]$. In all considered setups, we consider $m = 2$ subpopulations and set $a_2 = 1.5$. The birth intensity function of subpopulation 1 is defined as a piecewise constant function $\lambda_1(b) = \lambda_{11}I(b \leq (t + s)/2) + \lambda_{12}I(b > (t + s)/2)$. Lifetimes are exponentially distributed according to $y \mapsto f_j(y|\theta_j) = \theta_j \exp(-\theta_j y)$ with parameters θ_1 and θ_2 in the respective subpopulations. For each simulation setup, we generate 1000 random datasets and determine the estimates using the algorithm MNS. The damping factor is set as $\omega = 1$ and $\epsilon = 10^{-4}$.

Following Table 1, comparisons between the Setups 1–2 ($\lambda_{11} > \lambda_{12}$) and Setups 3–4 ($\lambda_{11} = \lambda_{12}$) are of primary interest, as these setups respectively concern the same birth distributions. The Setups 5 and 6 differ from Setup 4 only in that τ^R is increased. The probability of observing a random unit ranges from 0.138 to 0.261 in the considered setups. Note also that although the birth distribution is in fact parametric in these simulation setups, the birth distribution is estimated nonparametrically.

The simulation results (see Table 2) indicate that the proposed method yields valid estimates of the lifetime parameters θ . While the bias of the estimates is notable due to the relatively small sample sizes, the MSEs are adequately small and decrease with growing sample size (see Setup 1 vs. 2 and Setup 3 vs. 4). Comparing the Setups

Table 1 Simulation setups

Setup	τ^L	τ^R	λ_{11}	λ_{12}	a_2	θ_1	θ_2	$\pi(G, \mathbf{a}, \theta)$
1	8	10	100	50	1.50	0.40	0.20	0.138
2	8	10	200	100	1.50	0.40	0.20	0.138
3	8	10	50	50	1.50	0.40	0.20	0.165
4	8	10	100	100	1.50	0.40	0.20	0.165
5	8	11	100	100	1.50	0.40	0.20	0.219
6	8	12	100	100	1.50	0.40	0.20	0.261

Table 2 Simulation results for the estimated lifetime parameters $(\hat{\theta}_1, \hat{\theta}_2)$

Setup	$E(n)$	Bias		MSE	
		θ_1	θ_2	θ_1	θ_2
1	206	−0.02192	−0.02440	0.00295	0.00123
2	412	−0.02661	−0.02622	0.00176	0.00098
3	165	0.01216	0.01018	0.00370	0.00115
4	329	0.00793	0.00968	0.00172	0.00057
5	437	−0.01354	−0.01272	0.00140	0.00053
6	521	−0.02570	−0.02558	0.00167	0.00096

$E(n)$ denotes the mean sample size

Table 3 Simulation results for the estimation of a_2

Setup	Bias		MSE	
	$\hat{a}_2$	$\hat{a}_2^{\text{simple}}$	$\hat{a}_2$	$\hat{a}_2^{\text{simple}}$
1	0.12396	0.13181	0.06697	0.08177
2	0.12750	0.12912	0.03980	0.04283
3	−0.00335	0.01394	0.05382	0.07316
4	−0.01435	−0.00262	0.02855	0.03738
5	0.06664	0.07167	0.02745	0.03705
6	0.12337	0.11916	0.03787	0.04545

5 and 6 with Setup 4, the MSE is not substantially decreased despite the increased sample size (and increased selection probability). This shows that the structure of the truncation mechanism, including the definition of the data collection period, may influence the precision of estimates in a nontrivial manner, and larger samples not necessarily correspond to higher precision.

Table 3 summarises the performance of $\hat{a}_2$ and the comparison to the simple estimate $\hat{a}_2^{\text{simple}}$. It turns out that the biases of both estimates are considerable, yet of a similar magnitude. While $\hat{a}_2^{\text{simple}}$ is a valid estimator of a_2 , $\hat{a}_2$ is more precise in all considered setups.

Finally, we study the performance of the algorithm MNS and the estimated birth distribution. For the latter purpose, we define the distances

$$\|G^0 - \hat{G}^{(n)}\|_{\infty}, \|G^0 - G\|_{\infty} \quad \text{and} \quad \|\hat{G}^{(n)} - G\|_{\infty}.$$

The results displayed in Table 4 demonstrate that the initial values as defined in Step 2 in the algorithm MNS are suitable for optimization. On average, the first Newton–Raphson-type search in Step 3 in the algorithm MNS required about 6 to 8 iterations before reaching the stopping criterion. Considering that the number of parameters in the studied simulation setups is $n + 3$, requiring on average less than 40 iterations in the final Newton–Raphson-type search in Step 5 in the algorithm MNS seems quite satisfactory.

Table 4 Performance of the algorithm MNS ($E(\kappa_3)$ and $E(\kappa_5)$ denote the respective mean number of Newton–Raphson iterations in Step 3 and Step 5 of the algorithm MNS)

Setup	$E(\kappa_3)$	$E(\kappa_5)$	$E(\ G^0 - \widehat{G}^{(n)}\ _\infty)$	$E(\ G^0 - G\ _\infty)$	$E(\ \widehat{G}^{(n)} - G\ _\infty)$
1	7.5	35.3	0.00368	0.07482	0.07413
2	7.3	20.5	0.00117	0.05541	0.05505
3	6.4	29.7	0.00412	0.08080	0.08150
4	6.2	14.1	0.00107	0.05813	0.05828
5	6.9	17.8	0.00114	0.05106	0.05087
6	7.4	25.1	0.00158	0.05801	0.05711

Assessing the estimation of the birth distribution G , the mean distance of $\widehat{G}^{(n)}$ to G , denoted as $E(\|G^0 - G\|_\infty)$ is remarkably small in all considered setups and generally decreases with growing sample size, ceteris paribus. Note also that the distance of G^0 to $\widehat{G}^{(n)}$, denoted as $E(\|G^0 - \widehat{G}^{(n)}\|_\infty)$, is relatively small, again demonstrating that the suggested choice of initial values is appropriate. Nevertheless, $\|G^0 - G\|_\infty \leq \|G^0 - \widehat{G}^{(n)}\|_\infty + \|\widehat{G}^{(n)} - G\|_\infty$, and the final Newton–Raphson method in Step 5 in the algorithm MNS with regard to all parameters is necessary to find the maximum likelihood estimate.

Simulation study on computational performance

We further evaluate the proposed initial parameter choice in the Newton–Raphson algorithm by comparing its performance to reasonable alternatives. A local random search (LRS) is a conceptually simple approach to finding local maxima of the likelihood function (see Zhigljavsky 1991). Given numerically permissible initial values, the algorithm successively finds parameters with higher likelihood value, until a defined termination criterion is met. Reasonable termination criteria are, e.g., that no improvements are found for a defined number of steps or that a preset number of iterations has been performed. The algorithm LRS has the advantages of using computationally inexpensive iterations and of being relatively robust in that the local step sizes can be chosen sufficiently conservative to avoid numerically non-permissible parameters. The algorithm LRS does not utilize the local shape of the log-likelihood function, but just requires the function's value at candidate points. See the Appendix for an implementation of the algorithm LRS.

In addition, we consider a Newton–Raphson-type algorithm, which we call 'Naive Newton–Raphson Search' (NNS, see Appendix). This algorithm addresses a drawback of the algorithm LRS using the local shape of the log-likelihood function for updating parameter values in each iteration. The algorithm NNS differs from the algorithm MNS in that the initial birth distribution is chosen naively, i.e. without utilizing the IPW relation expressed in Eq. (10).

We apply the three algorithms on 100 simulated datasets each with the following three setups. Births are randomly drawn in the interval $[s, t] = [0, 8]$ from $m = 2$ subpopulations with intensity $\lambda_1(\cdot) \equiv \lambda_1 \in \{100, 200, 500\}$ and proportion factor

$a_2 = 1$. Lifetimes are exponentially distributed according to $f_j(y|\theta_j) = \theta_j \exp(-\theta_j y)$ with parameters $\theta_1 = 0.4$ and $\theta_2 = 0.15$. If a unit fails within $[\tau^L, \tau^R] = [8, 10]$, it is sampled.

For each algorithm and iteration, the distance of the current log-likelihood value to the maximal value is calculated. For the algorithms NNS and the MNS, 10 iterations have been performed, and took roughly the same runtime. Since an iteration in the algorithm LRS is computationally much cheaper than for the algorithms NNS and MNS, the number of iterations of the algorithm LRS has been adjusted to account for this difference. Therefore, the runtime of all three algorithms has been approximately the same in each setup. The initial values are chosen as $\theta_j^0 = n_j / \sum_{i:x_i=j} y_i$ for each $j \in \{1, 2\}$, which would be the subpopulation-specific maximum likelihood estimate of θ_j if the data were not subject to truncation, and $a_2^0 = 1$.

Figure 3 shows that the three considered algorithms seem to form a clear order with regard to performance. Due to having the same initial values, the algorithms LRS and NNS start with the same log-likelihood value. The algorithm LRS quickly improves over the algorithm NNS, and then progresses rather slowly towards the maximum. The algorithm MNS clearly has the best performance of the three considered algorithms in the considered setups. Interestingly, the initial value of the algorithm MNS has higher log-likelihood value than the algorithms NNS or LRS achieve within the considered runtime. Moreover, after few steps, the algorithm MNS essentially reaches the maximum of the log-likelihood. This simulation demonstrates that a model-driven initial choice of the parameters ΔG_i can remarkably reduce the runtime of the numerical optimization.

5 Application

Personal insolvency occurs when a person is unable to pay their debts to other entities. In Germany, personal insolvencies are reported publicly due to legal regulations. Reports on the individual insolvencies are available through a public website for 14 days, before public access is removed again. The sample, which results from continually tracking such public reports, includes information on all persons who become insolvent during the timespan of data collection, and is thus subject to the described double truncation framework. From March 2018 to December 2018, $n = 73,370$ units that were born between 1920 and 2000 have been sampled (see Table 5 for basic summary statistics).

Individual information includes the date of birth b_i , the age at insolvency y_i (adjusted to biological age 18, when legal age is reached in Germany) and the federal state x_i in which the insolvency is reported. There are $m = 16$ federal states, and they constitute the subpopulations in this analysis. For the lifetimes in any subpopulation $j \in \{1, \dots, 16\}$, we assume a Gamma distribution $\text{Ga}(\alpha_j, \beta_j)$, defined as

$$f_j(y|\alpha_j, \beta_j) = \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} y^{\alpha_j-1} \exp(-\beta_j y), \quad y > 0.$$

Fig. 3 Distance to the maximum log-likelihood value for the algorithms NNS, LRS and MNS. (top panel: $\lambda_1 = 100$, middle panel: $\lambda_1 = 200$, bottom panel: $\lambda_1 = 500$; dashed lines indicate the 80% pointwise confidence band; *The number of iterations of the algorithm LRS have been adjusted to yield comparable runtime to the algorithms NNS and MNS)

In this notation, $\theta_j = (\alpha_j, \beta_j)$ and $\theta = ((\alpha_1, \beta_1), \dots, (\alpha_{16}, \beta_{16}))$. The mean lifetime in the j th subpopulation is equal to α_j/β_j . As stated in the Assumptions (A)–(E), we implicitly assume that births are formed independently across the $m = 16$ federal states, and that the corresponding intensity functions are proportional. We have considered this assumption by checking the observed empirical birth distributions for each

Table 5 Descriptive statistics of personal insolvencies for 16 federal states in Germany between March 2018 and December 2018

	Size	Mean age at insolvency	Lifetime variance
Brandenburg	2554	42.45	164.01
Berlin	3304	45.39	155.10
Baden-Wuerttemberg	6766	45.09	168.35
Bavaria	8172	45.05	183.27
Bremen	917	43.33	176.28
Hesse	4682	44.95	159.22
Hamburg	2076	44.49	172.46
Mecklenburg-Vorpommern	1800	42.31	155.65
Lower Saxony	9812	42.55	176.09
North Rhine-Westphalia	18119	44.40	160.80
Rhineland-Palatinate	3232	44.85	173.04
Schleswig-Holstein	3167	43.90	162.70
Saarland	1207	43.00	177.20
Saxony	3412	43.37	178.34
Saxony-Anhalt	2428	41.81	157.43
Thuringia	1722	43.15	147.81
Total	73370	44.05	168.55

state in the dataset, and while the impact of the lifetime distributions is neglected via this inspection, we found that there is no apparent reason to discard the proportionality assumption. Note that Assumption (D) is fulfilled by the procedure the data have been collected. Since there is perhaps a substantial percentage of people who never experience insolvency (i.e. 'immortal' units), the population of interest is implicitly defined as the population of all people in Germany who eventually experience insolvency. Therefore, the birth distribution estimated under the truncation setup does not necessarily represent the actual birth distribution known for Germany. The model altogether contains 73,417 parameters. We apply the algorithm MNS to determine the maximum likelihood estimates.

As initial values in Step 1 of the algorithm MNS, we choose $a_j^0 = n_j/n_1$. Reasonable initial values (α_j^0, β_j^0) were obtained by first fitting the model to a random subsample having size 2000. We choose $\omega = 0.4$ as the damping factor. The resulting estimates are displayed in Table 6.

Figure 4 summarizes the estimates of the proportion factors and the mean age at insolvency for each subpopulation. The results indicate distinct differences between the federal states, with the estimated proportion factors generally reflecting the proportions of the federal state sizes known for Germany (numerical values not displayed).

Similarly, the estimated mean age at insolvency varies across federal states, ranging from 56.4 in Saxony-Anhalt to 70.6 in Bavaria. In comparison to the observed mean ages at insolvency (see Table 5) in the sample, the estimates are considerably larger. Without accounting for the truncation mechanism, the mean ages are systematically

Table 6 Maximum likelihood estimates

	$\hat{a}_j$	$(\hat{\alpha}_j, \hat{\beta}_j)$
Brandenburg	(1.000)	(2.589, 0.058)
Berlin	1.350	(3.365, 0.075)
Baden-Wuerttemberg	2.988	(2.958, 0.060)
Bavaria	3.880	(2.748, 0.052)
Bremen	0.410	(2.345, 0.045)
Hesse	1.962	(3.075, 0.066)
Hamburg	0.910	(2.757, 0.056)
Mecklenburg-Vorpommern	0.610	(2.800, 0.071)
Lower Saxony	3.996	(2.314, 0.047)
North Rhine-Westphalia	7.320	(2.977, 0.065)
Rhineland-Palatinate	1.480	(2.765, 0.054)
Schleswig-Holstein	1.293	(2.753, 0.059)
Saarland	0.527	(2.297, 0.044)
Saxony	1.305	(2.760, 0.062)
Saxony-Anhalt	0.793	(2.754, 0.072)
Thuringia	0.580	(3.221, 0.084)

Fig. 4 Estimates of the coefficients a_j and estimated mean biological ages at insolvency for $m = 16$ federal states in Germany. Squares left to the federal state names indicate the estimated proportion factors $\hat{a}_j$, where darker squares indicate higher birth intensity. Numbers in parentheses indicate the mean ages at insolvency. Vertical bars display the 90% confidence interval of the mean lifetime for each subpopulation

Fig. 5 Estimated birth distribution for Germany for (biological) births between 1920 and 2000 (solid line: maximum likelihood estimate, dashed line: empirical distribution function)

underestimated. We note that larger subsample sizes correlate with smaller confidence intervals for the mean biological age at insolvency in the respective subpopulations, which is not unexpected.

Figure 5 indicates that the truncation mechanism substantially distorts the birth distribution, as the empirical distribution function differs significantly from the estimate. In particular, more recent births between 1985 and 2000 seem to be overrepresented in the sample in comparison to the population of interest, while people born between 1940 and 1980 are underrepresented.

6 Discussion

We have focused on a joint model for the birth distribution and the lifetime distribution for heterogeneous survival data under double truncation. In doing so, we have derived maximum likelihood estimates for all involved parameters.

The situation is cast as a high-dimensional problem, where the number of parameters increases linearly with sample size. The main advantage is that the model can be treated with well-known concepts for parametric models, and implemented in a standard fashion. One substantial drawback of the large number of parameters is that the optimization method for finding the maximum likelihood estimates needs to be carefully adjusted.

Concerning the implementation of the model, we have demonstrated that caution in the choice of initial values in a Newton–Raphson-type algorithm is important. Using a generic choice is not ideal and can lead to impractical runtime. Instead, we propose model-driven initial values, which greatly reduce computational runtime. In our calculations, we have found that with this approach, the model can be fitted to even medium-sized datasets such as considered in the application ($n > 50,000$, $m > 10$) within reasonable runtime. However, we note that further improvements on the optimization method are certainly possible and worth studying. This also concerns the choice of the damping factor ω , which has been introduced to avoid obtaining invalid

parameter values within the optimization algorithm. We have not included a general guideline on how to properly choose ω , because it may depend on many variables such as the sample size, the setting of the birth interval $[s, t]$, the data collection period $[\tau^L, \tau^R]$ and also on the parametric families used for modeling the lifetimes.

Similarly, alternative optimization methods such as the algorithm LRS discussed in Sect. 4, though being theoretically basic, can be valuable in certain situations, particularly when initial values of the Newton–Raphson methods are too far from the optimum. Owing to the relatively robust search strategy, the algorithm LRS can be used as a preparation for more sophisticated optimization methods such as the algorithm MNS.

All three considered algorithms require initial values θ_j^0 , $j = 1, \dots, m$, for which we suggest using the conventional maximum likelihood estimate in the absence of truncation. Though it has not been further examined in the simulation study, this choice has generally lead to surprisingly satisfactory results. Alternative initial values might yield better numerical performance under certain circumstances.

We have sketched how the asymptotic behaviour of the estimates $\hat{a}$ and $\hat{\theta}$ can be studied for given consistent estimator $\hat{G}^{(n)}$. It remains to prove that the proposed nonparametric estimator $\hat{G}^{(n)}$ is in fact consistent.

Acknowledgements The author is grateful to Takeshi Emura for the invitation to contribute to the special issue and for numerous inspiring discussions on the addressed topic. For their many constructive and effective comments, the author wishes to express his gratitude to the anonymous reviewers. He also thanks Rafael Weißbach and Jacobo de Uña-Álvarez for many stimulating conversations.

Funding Open Access funding enabled and organized by Projekt DEAL. No particular funding has been received for this work.

Availability of data and material (data transparency) The data were collected from a public website.

Declarations

Conflict of interest There are no competing interests.

Code availability (software application or custom code) Implementation has been performed as custom code and can be made publicly available.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix

Algorithm LRS (Local Random Search)

Step 1 Set initial values $a_j^0 = n_j/n_1$ and θ_j^0 for $j = 1, \dots, m$, and $\Delta G_i^0 = n^{-1}$ for $i = 1, \dots, n$. Set search parameters $q_G, q_a, q_\theta \in (0, 1)$.

Set $(G^{\max}, \mathbf{a}^{\max}, \boldsymbol{\theta}^{\max}) = (G^0, \mathbf{a}^0, \boldsymbol{\theta}^0)$ and $LL^{\max} = \log L(G^{\max}, \mathbf{a}^{\max}, \boldsymbol{\theta}^{\max} | D_n)$.

Step 2 For each parameter component, draw a random candidate value as follows:

2.1 For each $i = 1, \dots, n$, draw $z_i \sim U((1 - q_G)\Delta G_i^{\max}, (1 + q_G)\Delta G_i^{\max})$.

Set $\Delta G_i = z_i / \sum_{i=1}^n z_i$ for $i = 1, \dots, n$.

2.2 For each $j = 2, \dots, m$, draw $a_j \sim U((1 - q_a)a_j^{\max}, (1 + q_a)a_j^{\max})$.

2.3 For each $j = 1, \dots, m$, draw $\theta_j \sim U((1 - q_\theta)\theta_j^{\max}, (1 + q_\theta)\theta_j^{\max})$.

Step 3 If $\log L(G, \mathbf{a}, \boldsymbol{\theta} | D_n) > \log L(G^{\max}, \mathbf{a}^{\max}, \boldsymbol{\theta}^{\max} | D_n)$, set $(G^{\max}, \mathbf{a}^{\max}, \boldsymbol{\theta}^{\max}) = (G, \mathbf{a}, \boldsymbol{\theta})$ and $LL^{\max} = \log L(G^{\max}, \mathbf{a}^{\max}, \boldsymbol{\theta}^{\max} | D_n)$.

Step 4 If the termination criterion is fulfilled, set $(\hat{G}^{(n)}, \hat{\mathbf{a}}, \hat{\boldsymbol{\theta}}) = (G^{\max}, \mathbf{a}^{\max}, \boldsymbol{\theta}^{\max})$; else return to Step 2.

As in the algorithm MNS, we use the maximum likelihood estimates of $\boldsymbol{\theta}_j$ in the model without truncation for the initial values $\boldsymbol{\theta}_j^0$. In Step 2, the parameters $q_G, q_a, q_\theta \in (0, 1)$ define the maximum relative step size for each parameter and iteration. An efficient choice of these values ensures a sufficiently large probability of finding parameter values with increased log-likelihood value, and a sufficiently large average step size to enable a feasible runtime by avoiding unnecessarily many iterations. An optimal choice generally depends on the setup, in particular the sample size and the considered lifetime distributions. In the studied simulation setup, we have found that choosing $q_G = n^{-1}$, $q_a = 0.05$ and $q_\theta = 0.025$ seems to be a suitable practical choice for small to moderate sample sizes.

Algorithm NNS (Naive Newton-Raphson Search)

Step 1 Set $\epsilon > 0$, $v = 0$ and initial values $a_j^0 = n_j/n_1$ and $\boldsymbol{\theta}_j^0$ for $j = 1, \dots, m$, and $\Delta G_i^0 = n^{-1}$ for $i = 1, \dots, n$. Abbreviate $\boldsymbol{\gamma}^v := (G^v, \mathbf{a}^v, \boldsymbol{\theta}^v)$.

Step 2 Set $\tilde{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^v$. For each component k of $\tilde{\boldsymbol{\gamma}}$, perform a damped univariate Newton-Raphson step as

$$\begin{aligned}\tilde{\boldsymbol{\gamma}}' &= \tilde{\boldsymbol{\gamma}} \\ \tilde{\gamma}'_k &= \tilde{\gamma}_k - \omega \frac{\partial \log L(\tilde{\boldsymbol{\gamma}} | D_n)}{\partial \gamma_k} \bigg/ \frac{\partial^2 \log L(\tilde{\boldsymbol{\gamma}} | D_n)}{\partial \gamma_k \partial \gamma_k} \\ \tilde{\boldsymbol{\gamma}} &= \tilde{\boldsymbol{\gamma}}'\end{aligned}$$

Step 3 Set $\boldsymbol{\gamma}^{v+1} = \tilde{\boldsymbol{\gamma}}$.

Step 4 If $d(\boldsymbol{\gamma}^v, \boldsymbol{\gamma}^{v+1}) \geq \epsilon$, return to Step 2; else stop and set $(\hat{G}^{(n)}, \hat{\mathbf{a}}, \hat{\boldsymbol{\theta}}) = \boldsymbol{\gamma}^v$.

The damping factor ω is incorporated as in the algorithm MNS. For sample sizes $n \approx 1500$, a damping factor of $\omega = 0.25$ seemed suitable for both avoiding invalid parameter values and still achieving satisfactory runtime of the algorithm. Note that

in Step 2, the current parameter $\tilde{\gamma}$ is adjusted in each univariate Newton–Raphson step. As in the algorithm MNS, we use the maximum likelihood estimates of θ_j in the model without truncation for the initial values θ_j^0 .

References

- Chiou, S. H., Austin, M. D., Qian, J., & Betensky, R. A. (2019). Transformation model estimation of survival under dependent truncation and independent censoring. *Statistical Methods in Medical Research*, 28(12), 3785–3798. <https://doi.org/10.1177/0962280218817573>.
- Chiou, S. H., Qian, J., Mormino, E., & Betensky, R. A. (2018). Permutation tests for general dependent truncation. *Computational Statistics and Data Analysis*, 128, 308–324. <https://doi.org/10.1016/j.csda.2018.07.012>.
- de Uña-Álvarez, J. (2020). Nonparametric estimation of the cumulative incidences of competing risks under double truncation. *Biometrical Journal*, 62(3), 852–867. <https://doi.org/10.1002/bimj.201800323>.
- Dörre, A. (2020). Bayesian estimation of a lifetime distribution under double truncation caused by time-restricted data collection. *Statistical Papers*, 61(3), 941–965. <https://doi.org/10.1007/s00362-017-0968-7>.
- Dörre, A., & Emura, T. (2019). *Analysis of doubly truncated data: An introduction*. Berlin: Springer.
- Efron, B., & Petrosian, V. (1999). Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association*, 94(447), 824–834.
- Emura, T., Hu, Y., & Konno, Y. (2017). Asymptotic inference for maximum likelihood estimators under the special exponential family with double-truncation. *Statistical Papers*, 58, 877–909. <https://doi.org/10.1007/s00362-015-0730-y>.
- Emura, T., & Konno, Y. (2012). Multivariate normal distribution approaches for dependently truncated data. *Statistical Papers*, 53(1), 133–149.
- Emura, T., Konno, Y., & Michimae, H. (2015). Statistical inference based on the nonparametric maximum likelihood estimator under double-truncation. *Lifetime Data Analysis*, 21(3), 397–418.
- Lagakos, S. W., Barraj, L. M., & de Gruttola, V. (1988). Nonparametric analysis of truncated survival data with application to AIDS. *Biometrika*, 75(3), 515–523.
- Lawless, J. F., & Kalbfleisch, J. D. (1992). Some issues in the collection and analysis of field reliability data. In J. P. Klein & P. K. Goel (Eds.), *Survival analysis: State of the Art. Nato Science (series E: applied sciences)* (Vol. 211, pp. 141–152). Dordrecht: Springer.
- Martin, E. C., & Betensky, R. A. (2005). Testing quasi-independence of failure and truncation times via conditional Kendall's tau. *Journal of the American Statistical Association*, 100(470), 484–492.
- Moreira, C., & de Uña-Álvarez, J. (2010). Bootstrapping the NPMLE for doubly truncated data. *Journal of Nonparametric Statistics*, 22, 567–583.
- Moreira, C., de Uña-Álvarez, J., & Braekers, R. (2021). Nonparametric estimation of a distribution function from doubly truncated data under dependence. *Computational Statistics*, <https://doi.org/10.1007/s00180-021-01085-4>.
- Moreira, C., de Uña-Álvarez, J., & van Keilegom, I. (2014). Goodness-of-fit tests for a semiparametric model under random double truncation. *Computational Statistics*, 29, 1365–1379. <https://doi.org/10.1007/s00180-014-0496-z>.
- Reiss, R.-D. (1993). *A course on point processes*. New York: Springer.
- Scheike, T. H., & Keiding, N. (2006). Design and analysis of time-to-pregnancy. *Statistical Methods in Medical Research*, 15, 127–140.
- Shen, P. (2010a). Nonparametric analysis of doubly truncated data. *Annals of the Institute of Statistical Mathematics*, 62(5), 835–853.
- Shen, P. (2010b). Semiparametric analysis of doubly truncated data. *Communications in Statistics Theory and Methods*, 39(17), 3178–3190. <https://doi.org/10.1080/03610920903219272>.
- Snyder, D. L., & Miller, M. I. (1991). *Random point processes in time and space* (2nd ed.). Berlin: Springer.
- Tsai, W. Y. (1990). Testing the assumption of independence of truncation time and failure time. *Biometrika*, 77(1), 169–177.
- van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511802256>.

- Wang, M.-C. (1989). A semiparametric model for randomly truncated data. *Journal of the American Statistical Association*, 84(407), 742–748.
- Xiao, J., & Hudgens, M. G. (2019). On nonparametric maximum likelihood estimation with double truncation. *Biometrika*, 106(4), 989–996.
- Ye, Z., & Tang, L. (2016). Augmenting the unreturned for field data with information on returned failures only. *Technometrics*, 58(4), 513–523.
- Zhang, X. (2015). Nonparametric inference for an inverse-probability-weighted estimator with doubly truncated data. *Communications in Statistics Simulation and Computation*, 44(2), 489–504.
- Zhigljavsky, A. A. (1991). *Theory of global random search*. Berlin: Springer Science + Business.