Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Computational Statistics and Data Analysis 169 (2022) 107405

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

Missing link survival analysis with applications to available pandemic data

María Luz Gámiz [a], Enno Mammen [b], María Dolores Martínez-Miranda [a,*], Jens Perch Nielsen [c]

[a] Department of Statistics and Operations Research, University of Granada, Spain
[b] Institute of Applied Mathematics, Heidelberg University, Germany
[c] Bayes Business School, City, University of London, UK

ARTICLE INFO

Article history:
Received 30 March 2021
Received in revised form 25 November 2021
Accepted 27 November 2021
Available online 13 December 2021

Keywords:
Double one-sided cross-validation
Hazard
Local linear estimation
Missing data

ABSTRACT

It is shown how to overcome a new missing data problem in survival analysis. Iterative nonparametric techniques are utilized and the missing data information is both estimated and used for further estimation in each iterative step. Theory is developed and a good finite sample performance is illustrated by simulations. The main motivation is an application to French data on the temporal development of the number of hospitalized Covid-19 patients.

1. Introduction

This paper is introducing a new missing link data problem for survival analysis. The missing link is referring to the missing link between the time of origin of a failure and the failure time itself. Information is available about origins and information is available about failures, but the link between these pieces of information is missing. During the recent Covid-19 pandemic, this kind of missing link survival data was omnipresent. This paper is a methodological paper proposing a new tool in survival analysis including both underlying theory and finite sample studies. While we have seen important advances in nonparametric and semiparametric survival analysis and reliability theory including highly multidimensional problems and time-dependent covariates (see for example Andersen et al. (1993), Martinussen and Scheike (2006) and Gámiz et al. (2011)), survival models have not - to our knowledge - been developed taking account of missing information on the origin of durations as investigated in this paper. The missing data survival model described and analysed in this paper is similar but different from the backcalculation method of Brookmeyer and Gail (1988) and Brookmeyer (1996), see also Jewell (1990) for a review paper on missing data statistical modelling used in the time of the AIDS epidemic. We believe that our approach would have been useful had it already been developed at the outbreak of the AIDS epidemic. One could for example had been interested in the connection between the total number of tested HIV-positive individuals and the number of onsets of AIDS. The approach advocated for in this paper is based on easy-to-collect data and can be used as a benchmark method, for example when catching up in the confusing beginning of a pandemic, where there is no time for complicated data discussions across boundaries and districts. In recent years, there has been an increasing

* Correspondence to: Department of Statistics and Operations Research, Campus Fuentenueva, 18071 Granada, Spain.
E-mail address: mmiranda@ugr.es (M.D. Martínez-Miranda).

https://doi.org/10.1016/j.csda.2021.107405
0167-9473/

interest in landmarking future forecasts based on marker information for example, see Ferrer et al. (2019), Proust-Lima et al. (2016), Blanche et al. (2015) and Proust-Lima et al. (2014), however such methodology seems to be too complicated to serve as a benchmark methodology in the beginning of a pandemic. Our methodology is simpler than the EM-algorithm where complicated conditional means have to be considered, see for example Allassonniere and Chevallier (2021) and Zhao et al. (2020) for recent contributions in this direction. It is also different from missing data work as introduced in Heckman (1979), because in our model it is the starting point of a duration that is missing. The reason our algorithm is working anyway is that the starting point of the duration can be recovered via the changes in new cases over time, the pattern of changes of new cases translates into a pattern of durations. The preliminary theory provided in this paper provide evidence for this claim.

Our method is illustrated on recent data from the Covid-19 pandemic that is available for most countries or even regions within countries. The empirical illustration of this paper focuses on duration effects on mortality and recovery of hospitalized Covid-19 patients via aggregated data. Since the beginning of the Covid-19 pandemic there has been daily information on the number of hospitalized patients with the virus. Often data are aggregated and contain only information about daily numbers of patients staying in hospitals or leaving the hospital at this day. Thus key statistical information has been lost including information on current duration in hospital of each Covid-19 patient. In this paper we show that one can analyse this new sampling scheme - with important duration information missing - almost as well as if one indeed had full information from the beginning. This is only one important building block while understanding the development of a pandemic. Other building blocks like the spread of the virus provide similar missing data issues when cross-country data is applied. We believe that the insights of this paper provide an important first step towards making mathematical statistical sense of the kind of data that is actually available during a pandemic.

Forecasting the development of a pandemic is complicated and the mathematical analysis of the missing data application in this paper is non-trivial providing a new theoretical problem of mathematical statistics. However, the data input needed to apply the forecasting methodology suggested by this paper is easy to understand and monitor. Also the forecast itself, the output, is easy to understand and apply to monitor the development of the pandemic. The ambition of this paper is to provide the first indication of a new methodology that can be communicated to epidemiologists or other practitioners in the field. When the input and the output is easy to understand, one could imagine or hope for that our method could enter basic textbooks in the field, see for example Jewell (2004). Such basic communication is possible when both input and output are easy to understand even when the theoretical mathematical statistical steps are highly sophisticated.

While missing data analyses have a long tradition in mathematical statistics, our problem is different from the problem of Rubin (1996). In our aggregated data, one cannot isolate and impute the missing data via the modelling provided in Rubin (1996) or the many other papers working with multiple imputation or missing covariates, see Liu and Hu (2020) for a recent example. Also, aggregated data analysis as in Farebrother (1979) or King (1997) do not match the aggregated data problem we face with our Covid-19 data, because there is no linear transformation available defining the missing data problem. The related approach involving an original underlying continuous stochastic process model, as we indeed also have, in Lawless and McLeish (1984) does also not match to our type of data, because we do not have observed information of the exact timing between aggregated data points. It is exactly this timing that is our missing data. In other words: we have identified a new data missing problem in survival analysis that is applicable to the kind of data all of us have witnessed during the year 2020. Our practical and theoretical solutions provided below show that one can overcome this missing data problem with almost as good results as had we known the missing information on duration.

2. Counting process formulation of the simplest version of the missing link survival analysis problem

In this section we consider the simplest possible version of the missing link survival analysis problem with only two observed counting processes that we call $\bar{N}_1$ and $\bar{N}_2$. The first of these two counting processes counts arrivals and the second counts departures. The link between these arrivals and departures is missing, the two observed counting processes do not contain this information. In our applied work on hospitalized Covid-19 patients later in the paper, the situation is slightly more complicated with two different possible reasons for leaving the hospital, namely death or recovery. However, for clarity we present the fundamental problem in the simplest possible setting. In Section 2.1 below, we introduce the problem in the situation where data are observed continuously and in Section 2.2 we make the necessary amendments to the situation where data are only observed within discrete intervals. In our practical study later in this paper, data are only observed on a daily basis and therefore follow the notation and set-up of Section 2.2 below. In Appendix A we provide a brief glossary with this notation.

2.1. The fundamental missing link survival analysis problem with continuous data

Let us assume that subjects arrive to a system at random times modelled by a counting process $\bar{N}_1$. Specifically $\bar{N}_1(t)$ counts the number of subjects that enter the system during the interval $(0, t]$. Each subject entering the system gives rise to a new counting process $N_{2,i}$ that can only take values in zero or one. This counting process starts at the time of arrival jumping to one when the event under study is happening (if this event is happening at all). We assume that the intensity function of the process $N_{2,i}$ fits the multiplicative Aalen model and can be written $\lambda_{2,i}(s) = \alpha(s)Y_{2,i}(s)$, where $Y_{2,i}$ is a predictable process taking value 1 when the subject i is in the system and 0 otherwise and α is an unknown (deterministic)

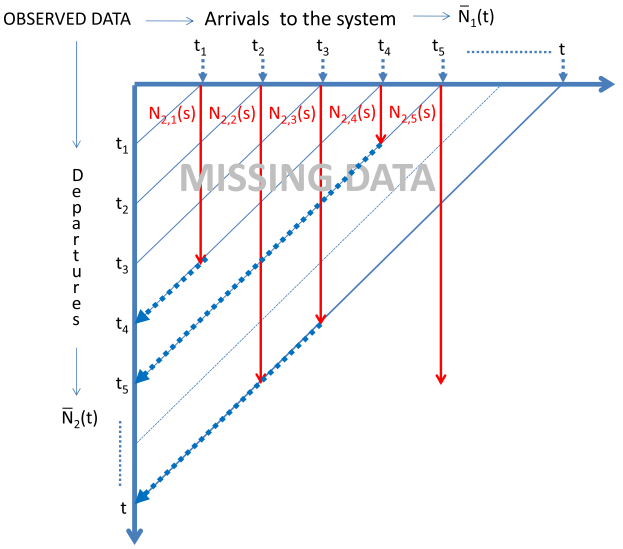

Fig. 1. Missing data problem. A simple case.

hazard function. When full information is available of the stochastic processes $N_{2,i}$ and $Y_{2,i}$, one could estimate the hazard function α, by usual kernel smoothing

$$\widehat{\alpha}_b(s) = \frac{\sum_{i=1}^n \int_0^{+\infty} \bar{K}_{s,b}(s-u)dN_{2,i}(u)}{\sum_{i=1}^n \int_0^{+\infty} \bar{K}_{b,s}(s-u)Y_{2,i}(u)du}$$

where $\bar{K}_{\cdot,b}(\cdot)$ is the local-linear kernel, and b is the bandwidth, see Nielsen and Tanggaard (2001). However we do not observe the stochastic processes $N_{2,i}$ and $Y_{2,i}$ directly. Instead, we observe the counting process $\bar{N}_1$ above counting arrivals and the following counting process

$$\bar{N}_2(t) = \sum_{\{i:t_i \le t\}} N_{2,i}(t - t_i)$$

Fig. 1 is a graphical simplification of the real situation. The available data are the two counting processes $\bar{N}_1$ and $\bar{N}_2$. The arrivals are represented on the horizontal axis of Fig. 1 by the counting process $\bar{N}_1(t)$; and, on the other hand, we have the counts of departures from the system by calendar time, represented on the vertical axis of the figure by the counting process $\bar{N}_2(t)$. For example, the plot is illustrating the particular case that $\bar{N}_1(t_5) = 5$ and $\bar{N}_2(t_5) = 2$. The time spent in the system for these 2 subjects that leave the system at or before t_5 are respectively $s_1 = t_4 - t_1$ and $s_2 = t_5 - t_4$. The subject entering at time t_5 still remains in the system at the current time, that is t. Exact information about durations represented in the plot is not provided.

In the next subsection we will see that the set-up of this subsection immediately generalizes to the further missing data case where observations are only observed in discrete intervals. While this generalization is immediate, the discrete set-up leads to a quite different and perhaps more tedious notation than the elegant continuous counting process formulation above. Our choice of nonparametric smoothing method above, Gámiz et al. (2016), is exactly chosen because it can be immediately transferred to the discrete data set-up of the next subsection. We do not know of other hazard smoothing methods, where this transfer to discrete data is this simple and immediate.

2.2. The fundamental missing link survival analysis problem with discretely observed data

Often occurrences and exposures are not observed continuously and the only data available is discretely collected during pre-specified time intervals. In the following we introduce notation for discretized observed versions of the above continuous counting processes: $\bar{N}_1(t)$, for the arrivals of subjects to the system; and, $\bar{N}_2(t)$, for the number of departures, when observed, not continuously, but on a discrete grid of time points. This grid not necessarily has to be equidistant, but for simplicity in the notation and without any loss of generality, we consider the set $\{1, 2, \ldots, M\}$.

Let us define, for $x = 1, 2, \ldots, M$,

$$E_{x,1} = \int_{x-1}^{x} d\bar{N}_1(u);$$

the total number of subjects entering the system in the interval $(x-1, x]$; and,

$$O_x = \int_{x-1}^{x} d\bar{N}_2(u);$$

are the total number of subjects that leave the system in the interval $(x-1, x]$.
As mentioned above, each subject that enters the system originates a counting process associated with its survival time inside the system, $N_{2,i}$ for the ith subject arriving at time t_i. Let $N_{2,x}$ be the aggregated counting process for all subjects arriving in the interval $(x-1, x]$, that is

$$N_{2,x}(s) = \sum_{\{i:t_i \in (x-1,x]\}} N_{2,i}(s)$$

for $s > 0$. When the processes $N_{2,i}$ are observed (full information), we can also define the following counts

$$O_{x,d} = \int_{d-1}^{d} dN_{2,x-d+1}(s)$$

for $1 \le d \le x \le M$, which are the total number of subjects that enter the system at time $x-d+1$ and leave at time x.
It must be satisfied that $O_x = \sum_{d=1}^{x} O_{x,d}$, for all $1 \le x \le M$, that is, we need to check that

$$O_x = \sum_{d=1}^{x} \int_{d-1}^{d} dN_{2,x-d+1}(s).$$

To get this, first we define $v = x-d+s$ and do the corresponding change of variable in the integral, then exchange the sum with the integral

$$\sum_{d=1}^{x} O_{x,d} = \sum_{d=1}^{x} \int_{x-1}^{x} dN_{2,x-d+1}(v-(x-d)) = \int_{x-1}^{x} \sum_{d=1}^{x} dN_{2,x-d+1}(v-(x-d))$$

then we do the following change of variable in the sum $z = x-d$

$$\sum_{d=1}^{x} O_{x,d} = \int_{x-1}^{x} \sum_{z=0}^{x-1} dN_{2,z+1}(v-z) = \int_{x-1}^{x} d\bar{N}_2(v) = O_x$$

When full information is available, these counts are of course directly observable. However, our missing link survival data problem implies that full data information is not available, and we are left to do our survival analysis with occurrence counts data like O_x above together with discrete exposure data like E_x below

$$E_x = \sum_{r=1}^{x} E_{r,1} - \sum_{r=1}^{x-1} O_r$$

for $x = 1, 2, \ldots, M$, and depending on duration, we have the number of subjects that remain in the system on the day x with duration exactly equal to d,

$$E_{x,d} = E_{x-d+1,1} - \sum_{s=1}^{d-1} O_{x-s,d-s}$$

for $1 \le d \le x \le M$. It can be checked that $E_x = \sum_{d=1}^{x} E_{x,d}$.

In Section 3 and onwards we will discuss our discrete missing data problem in a slightly more complicated situation than the one above, namely with two possible reasons for leaving the system instead just one as above. These two reasons (death or recovery in our applications) give rise to slightly more complicated notation.

3. The intuition behind our new missing data methodology

In this section we illustrate the situation in Section 2.2 via our concrete Covid-19 data on hospitalizations and with two possible reasons for leaving the hospital: death or recovery. Our missing data problem is of a typical survival analysis nature: while data collectors take great care with occurrences, they tend to be less careful with information about exposure. In our data we have daily accounts of people leaving the hospital and whether they left alive. That is good information on occurrence. We do know how many people are at risk of dying at hospital every single day, but the data collectors have not kept record of the duration distribution of this exposure. That is the main reason that standard survival models cannot be used on this type of pandemic data. Our methodology is an iterative procedure overcoming this missing data survival problem. First we assume that we know the answer to our problem: the duration distribution of people coming into hospital. Given this distribution we can forecast future exposures of every daily entry to hospital and aggregate this information to give us the relative daily distribution of exposure on duration. So, we do not use forecast exposure directly, only the information it provides on the relative daily distribution of exposures. This information is of course biased by the first a priori assumption on knowing the duration distribution in the first place, but it does give us a place to start our analyses. In the second step we use the relative exposure duration information collected in the first step to estimate the duration distribution using standard non-parametric survival smoothing techniques. This distribution could be our answer, but it is even better to iterate the above two steps until convergence. Our simulation study below shows that the method almost estimates the duration distribution as well as had the data collector indeed collected the detailed duration information on the daily exposure.

4. The formal model and discussion

We have decided to formulate our model via two-dimensional Poisson point processes to be able to derive some asymptotic theory indicating that our approach indeed works as our simulation study indicates it does. We could also have chosen to formulate our model via the standard continuous counting process Aalen survival model that is so common in survival analysis, see Section 2 above and Gámiz et al. (2016), but that would leave us with two later translations. First a translation describing the relationship between the continuous model and the discrete data actually observed, and secondly a translation from the continuous counting process martingale theory to the two-dimensional Poisson process considered for our first theoretical insight on this extremely difficult queuing problem. So, eventually we decided to be direct in our model formulation and work directly with the two-dimensional Poisson process. The formal model formulation includes the crucial time points for the i-th individual and the available information. We assume that we have a data set that contains only partial information on i.i.d. time points e_i and t_i $(i = 1, \ldots, n)$. Here e_i is the moment the i-th individual is hospitalized, and t_i is the moment he/she abandons the hospital due to death or recovery. We are interested in the distribution of $d_i = t_i - e_i$ i.e. in the length of time that patient i has spent in the hospital until he/she recovers or dies.

We assume that for all time points t in an interval $[0, M]$ our data set contains the values of the number of individuals who entered the hospital before t, the number of patients who left the hospital before t after recovery and the number of patients who left the hospital before t because of death.

We model the tuples (e_i, t_i) as generated by a two-dimensional Poisson point process ξ. For an interval $[a_s, b_s] \times [a_w, b_w]$ we assume that $\xi([a_s, b_s] \times [a_w, b_w])$ is a Poisson random variable with mean $\int_{a_s}^{b_s} \gamma(s)\mathrm{d}s\{S(a_w) - S(b_w)\}$. Here $\gamma(s)$ is the intensity of the number of patients admitted in a hospital. Furthermore, for the random duration W at the hospital $S(w) = \exp\left\{-\int_0^w \alpha(v)\mathrm{d}v\right\}$ is the survival function and $\alpha(w)$ is the corresponding hazard function. We make the central assumption that the distribution of W does not depend on the date when the patient enters a hospital. And by the model assumption of a Poisson point process the Poisson random variables $\xi(B_1)$ and $\xi(B_2)$ are independent for two disjoint subsets B_1, B_2 of $\mathbb{R}^2$. These two properties are important for our approach. The assumption of Poisson distributions is made mainly for having a simplified mathematical discussion but it is not essential for our approach.

This model is related to the literature on statistical inference for M/G/∞ queues studied in queuing theory. Such models have been discussed under conditions where no information is available on which incoming person is identical with a served leaving person. This has been done for discrete and continuous time, see e.g. Pickands and Stine (1997), Bingham and Pitts (1999) and Goldenshluger (2016). In all these papers it has been assumed that the intensity $\gamma(s)$ is constant. Then the statistical analysis is simplified by the stationarity of the process. Our data do not show stationarity and for this reason it requires other approaches. The paper Goldenshluger and Koops (2018) is an example where non-constant intensities of incoming claims are considered as in our paper, but this is done under smoothness conditions on γ that our approach does not need.

In the next section we consider a discretized version of the discussed model that better fits to our data set where only aggregated daily observations are observed.

5. Specified model formulation via aggregated Covid-19 data

We formulate our data via an aggregated data formulation to avoid the perhaps more challenging continuous time counting process formulation used in Section 2.1 above. The approach of this section is therefore an adaptation of Section 2.2 to the Covid-19 hospitalization data. Patients data are almost always observed on a monthly basis, a weekly basis, a daily

basis or an hourly basis for example. The aggregated data formulation below is therefore an attempt to reach a wider audience in the statistical community without sacrificing the generality of the applicability of the method. The Covid-19 data studied in this paper are aggregated on a daily basis.

We consider a model where one observes counts of daily occurrences and exposures for a population of individuals with different age and thus different hazard rates. Formally the model can most easily be described by defining independent Poisson random variables $O_{x,d}$ for (x,d) in a set that will be specified below. The variables $O_{x,d}$ denote the number of occurrences at day x for individuals of age d. In our data application $O_{x,d}$ is the number of individuals who leave hospital at day x and have been in the hospital for d days. In the notation of the last section we have $O_{x,d} = \xi((x-d-1, x-d] \times (x-1, x])$. Again, we will distinguish two types of occurrences: death and recovery. We denote by $O^D_{x,d}$ the number of deaths occurring at day x for individuals having stayed for d days at hospital. Furthermore, $O^R_{x,d}$ denotes the number of individuals who have stayed at hospital for d days and leave the hospital at day x. We denote the total number $O^D_{x,d} + O^R_{x,d}$ by $O_{x,d}$. We assume that there is an upper bound for the days staying at the hospital which we denote by $\mathcal{D}+1$. Then the number of patients that enter the hospital at a day x is given by $O_{x,1} + O_{x+1,2} + \cdots + O_{x+\mathcal{D},\mathcal{D}+1}$, which we denote by $E_{x,1}$. Furthermore, we observe $O^v_x = O^v_{x,1} + \cdots + O^v_{x,\mathcal{D}+1}$, with v equal to R, D or to a blank. This is the number of all occurrences at day x counting recoveries, deaths or the sum of both, respectively.

We assume that $O^v_{x,d}$ with v equal to R or D have a Poisson distribution with parameter $\gamma_{x-d+1}[(1-\alpha_1)\cdots(1-\alpha_{d-1})\alpha^v_d]$ for some $\gamma_x, \alpha^v_d > 0$ with $\alpha_d = \alpha^D_d + \alpha^R_d$. Furthermore, we assume that the variables $O^R_{x,d}$ and $O^D_{x,d}$ are independent. Then also the variables $O_{x,d}$ are independent Poisson random variables and $E_{x,1}$ has a Poisson distribution with parameter $\gamma_x = \sum_{d=1}^{\mathcal{D}} \gamma_x(1-\alpha_1)\cdots(1-\alpha_{d-1})\alpha_d + \gamma_x(1-\alpha_1)\cdots(1-\alpha_{\mathcal{D}})$. Thus γ_x is the number of expected incoming patients at day x. Note that in the notation of the last section we have $\gamma_x = \int_{x-1}^{x} \gamma(v)\mathrm{d}v \approx \gamma(x)$ and $\alpha_d = \{S(d-1) - S(d)\}/S(d-1) = 1 - \exp\{-\int_{d-1}^{d} \alpha(t)\mathrm{d}t\}$, which is approximately equal to $\alpha(d)$ if α is continuous in the interval $[d-1, d]$.

We assume that no patient has entered the hospital before day 1. This means that $\gamma_x = 0$ for $x \le 0$. Furthermore, we define $O^v_{x,d}$ for all patients that arrived at the hospital between day 1 and day M. Thus we define $O^v_{x,d}$ for all values (x,d) that fulfil both, $x = 1, \ldots, M+\mathcal{D}$ and $d = \max(1, x-M+1), \ldots, \mathcal{D}+1$. Furthermore, α^v_d denotes the probability that a patient who is in the hospital at a day x and has been in the hospital for d days leaves the hospital at that day for reason v. These are the parameters we are interested in. The values of γ_x are nuisance parameters.

We consider the case that all our information consists on only observing the sums $E_{x,1}$ and O^v_x, for $v \in \{D, R\}$, over a period $x = 1, \ldots, M$, for some $M \ge \mathcal{D}+1$. In particular the values of $O^D_{x,d}$ and $O^R_{x,d}$ are unobserved. This is equivalent to assuming that one observes the values of E_x and of O^v_x for $v \in \{D, R\}$ and $x = 1, \ldots, M$, where E_x is the observed total number of people in hospital on the day x. To see this statement note that $E_{1,1} = E_1$ and that for $x > 1$ the values of $E_{x,1}$ are given by

$$E_{x,1} = E_x - (E_{x-1} - O_{x-1}). \tag{1}$$

We assume for simplicity that the marginal distributions of $E_{x,1}$ and of O^v_x for $v \in \{D, R\}$ are Poisson. This is done mainly for having a more transparent description of the model and simpler arguments for its mathematical study but we argue that it is not essential for the performance of our approach.

In the next section we will describe an estimator for the hazards α^v_d $(d = 1, \ldots, \mathcal{D})$ for $v \in \{D, R\}$.

6. Hazard estimation for augmented data

We now describe procedures for the estimation of the hazards α^v_d $(d = 1, \ldots, \mathcal{D})$ for $v \in \{D, R\}$. For this purpose in a first step we estimate α_d $(d = 1, \ldots, \mathcal{D})$ by an iterative procedure. This estimate will be used in a second step to get estimates of α^D_d and α^R_d.

In the iteration cycles of the first step the values of the fits of $O_{x,d}$ and of $E_{x,d}$ are updated. Here $E_{x,d}$ is the unobserved total number of people in a hospital on the day x with duration d. Notice that $E_x = \sum_{d=1}^{\mathcal{D}+1} E_{x,d}$.

We consider two alternative procedures. In the first procedure the unknown values of $O_{x,d}$ and of $E_{x,d}$ are fitted by random quantities (random version). In the second procedure the algorithm is deterministic (deterministic version). These values will then be used to get estimates of the total sums $O_{+,d}$ and $E_{+,d}$, where $O_{+,d} = O_{1,d} + \cdots + O_{M,d}$ and $E_{+,d} = E_{1,d} + \cdots + E_{M,d}$. The number $O_{+,d}$ is the total number of recoveries and deaths for people who have been in hospital exactly d days and leave the hospital in the period $x = 1, \ldots, M$. The exposure, $E_{+,d}$, is the total number of people who, at a day $x = 1, \ldots, M$, have been in hospital exactly d days. At the end of each cycle the estimate of α_d is updated by the smoothed ratio of the fitted values of $O_{+,d}$ and of $E_{+,d}$.

The fitted values after the r-th iteration are denoted by $\hat{O}^{(r)}_{x,d}$, $\hat{E}^{(r)}_{x,d}$, $\hat{O}^{(r)}_{+,d}$, $\hat{E}^{(r)}_{+,d}$, and $\hat{\alpha}^{(r)}_d$ respectively.

We now describe the iteration cycles of the first step of the algorithm. At the start $r = 0$ we generate initial values for $\hat{O}^{(0)}_{x,d}$ and $\hat{E}^{(0)}_{x,d}$ and we make an initial choice for $\hat{\alpha}^{(0)}_d$ $(d = 1, \ldots, \mathcal{D})$, e.g. an Exponential distribution.

r-th iteration cycle of the first step of the algorithm

(i) For all $x = 1, \ldots, M$ proceed as follows. Put $\hat{E}^{(r)}_{x,1} = E_{x,1}$. In the random version of the algorithm generate a random variable $\hat{O}^{(r)}_{x,1}$ with Binomial distribution with parameters $\hat{E}^{(r)}_{x,1}$ and $\hat{\alpha}^{(r)}_1$. In the deterministic version of the algorithm choose $\hat{O}^{(r)}_{x,1}$ as the mean $\hat{\alpha}^{(r)}_1 \hat{E}^{(r)}_{x,1}$ of this Binomial distribution. At the start $r = 0$ we also tried a Binomial distribution with probability parameter $1/O_x$.

(ii) Then, for $d = 2$ define $\hat{E}^{(r)}_{x+d-1,d} = \hat{E}^{(r)}_{x+d-2,d-1} - \hat{O}^{(r)}_{x+d-2,d-1}$. Now, in the random version of the algorithm, $\hat{O}^{(r)}_{x+d-1,d}$ is generated as a Binomial random variable with parameters $\hat{E}^{(r)}_{x+d-1,d}$ and $\hat{\alpha}^{(r)}_d$. In the deterministic version of the algorithm, $\hat{O}^{(r)}_{x+d-1,d}$ is again the mean $\hat{\alpha}^{(r)}_d \hat{E}^{(r)}_{x+d-1,d}$ of the Binomial distribution. At the start $r = 0$ we also used the mean of a Binomial distribution with probability parameter $1/O_{x+d-1}$.

(iii) These calculations are repeated for $d = 3, \ldots, \min(M - x + 1, \mathcal{D} + 1)$. For all $x = 1, \ldots, M$ this gives the values of the following occurrences and exposures: $\{(\hat{O}^{(r)}_{x,1}, \hat{E}^{(r)}_{x,1}), \ldots, (\hat{O}^{(r)}_{M^*,\mathcal{D}^*}, \hat{E}^{(r)}_{M^*,\mathcal{D}^*}\}$, being $M^* = \min(M, x + \mathcal{D})$ and $\mathcal{D}^* = \min(M - x + 1, \mathcal{D} + 1)$. Thus, after having repeated the procedure for all $x = 1, \ldots, M$, we have the values of $\hat{O}^{(r)}_{x,d}$ and $\hat{E}^{(r)}_{x,1}$ for all (x, d), with $x = 1, \ldots, M$ and $d = 1, \ldots, \min(x, \mathcal{D})$. Furthermore to simplify notation, we put $\hat{O}^{(r)}_{x,d} = \hat{E}^{(r)}_{x,d} = 0$ for $x, d = 1, \ldots, \mathcal{D} + 1$ $(x < d)$.
In the resulting matrix, $\hat{O}^{(r)}_{x,d}$, with $x = 1, \ldots, M$, $d = 1, \ldots, \mathcal{D} + 1$, the sum of elements in the x-th row is the marginal distribution corresponding to occurrences according to notification time x, whereas the sum of the elements in the d-th column is the marginal distribution of duration. In the same way, in the resulting matrix $\hat{E}^{(r)}_{x,d}$, with $x = 1, \ldots, M$ and $d = 1, \ldots, \mathcal{D}+1$, the sum of elements in the x-th row is the marginal distribution of exposure according to notification time x, whereas the sum of the elements in the d-th column is the marginal distribution of duration.

(iv) Normalize the two matrices to obtain the distribution of occurrences and exposures for duration d, given the notification day x. So, define

$$\hat{q}^{(r)}_{x,d} = \frac{\hat{O}^{(r)}_{x,d}}{\sum_{d'=1}^{\mathcal{D}+1} \hat{O}^{(r)}_{x,d'}} \quad \text{and} \quad \hat{h}^{(r)}_{x,d} = \frac{\hat{E}^{(r)}_{x,d}}{\sum_{d'=1}^{\mathcal{D}+1} \hat{E}^{(r)}_{x,d'}},$$

for occurrences and exposures, respectively, for $x = 1, \ldots, M$ and $d = 1, \ldots, \mathcal{D}$.

(v) Finally the simulated occurrences and exposures for duration d are obtained as

$$\hat{O}^{(r)}_{+,d} = \sum_{x=1}^{M} \hat{q}^{(r)}_{x,d} O_x \quad \text{and} \quad \hat{E}^{(r)}_{+,d} = \sum_{x=1}^{M} \hat{h}^{(r)}_{x,d} E_x,$$

for $d = 1, \ldots, \mathcal{D} + 1$, where O_x and E_x are the observed occurrences and exposures at notification day x, for $x = 1, \ldots, M$.

(vi) For $y = 1, \ldots, \mathcal{D}$, the estimated hazard function is updated by local-linear smoothing (Nielsen and Tanggaard, 2001):

$$\hat{\alpha}^{(r+1)}_y = \frac{\sum_{d=1}^{\mathcal{D}+1} \bar{K}_{b,y}(y - d) \hat{O}^{(r)}_{+,d}}{\sum_{d=1}^{\mathcal{D}+1} \bar{K}_{b,y}(y - d) \hat{E}^{(r)}_{+,d}} \tag{2}$$

with a suitable bandwidth choice such as the double one-sided cross-validation method of Gámiz et al. (2016) (see also Mammen et al. (2011)). Here $\bar{K}_{b,s}$ is a local linear kernel

$$\bar{K}_{b,s}(s - u) = \frac{a_2(s) - a_1(s)(s - u)}{a_0(s)a_2(s) - \{a_1(s)\}^2} K_b(s - u),$$

where $a_j(s) = \int (s - u)^j K(u) Y(u) \mathrm{d}u$, for $j = 0, 1, 2$, and $K_b(\cdot) = b^{-1} K(\cdot/b)$, with a bandwidth parameter $b > 0$ and a kernel K being a symmetric probability density function. For $d = \mathcal{D} + 1$ we set $\hat{\alpha}^{(r+1)}_d = 1$.

(vii) These steps are iterated until $\max_{1 \le d \le \mathcal{D}} \{|\hat{\alpha}^{(r+1)}_d - \hat{\alpha}^{(r)}_d| / \hat{\alpha}^{(r)}_d\} < \epsilon$, with $\epsilon > 0$ small enough.

In the second step of the algorithm the final hazard estimator for duration, $\hat{\alpha}_y$, is split into hazards for duration due to death, $\hat{\alpha}^D_y$ and hazards due to recovery, $\hat{\alpha}^R_y$, as follows:

$$\hat{\alpha}^D_y = \frac{\sum_{d=1}^{\mathcal{D}+1} \bar{K}_{b,y}(y - d) \hat{O}^{D}_{+,d}}{\sum_{d=1}^{\mathcal{D}+1} \bar{K}_{b,y}(y - d) \hat{E}_{+,d}} \quad \text{and} \quad \hat{\alpha}^R_y = \frac{\sum_{d=1}^{\mathcal{D}+1} \bar{K}_{b,y}(y - d) \hat{O}^{R}_{+,d}}{\sum_{d=1}^{\mathcal{D}+1} \bar{K}_{b,y}(y - d) \hat{E}_{+,d}},$$

with $\hat{O}^{D}_{+,d} = \sum_{x=1}^{M} \hat{q}_{x,d}\ O^D_x$, where O^D_x is the total number of deaths registered on the day x, $\hat{q}_{x,d} = \hat{q}^{(r)}_{x,d}$ and $\hat{E}_{+,d} = \hat{E}^{(r)}_{+,d}$ for the last value of r; and $\hat{O}^{R}_{+,d} = \sum_{x=1}^{M} \hat{q}_{x,d}\ O^R_x$, where O^R_x is the total number of recoveries observed on the day x $(x = 1, \ldots, M)$.

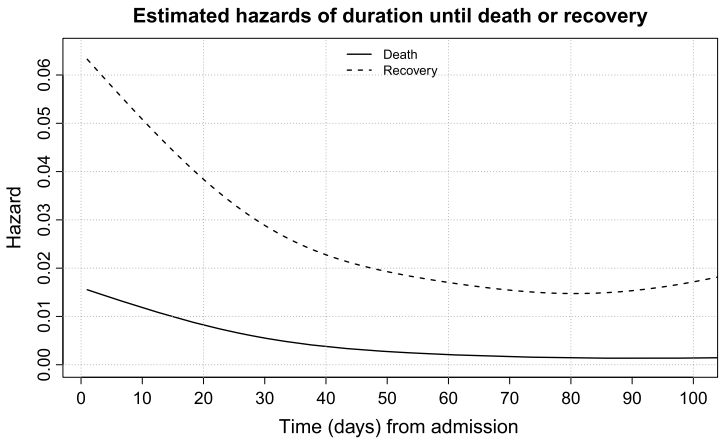

Fig. 2. Estimated hazards of the elapsed time since admission until death (solid) or recovery (small dashes).

In Section 8 we will study the performance of the estimators by simulations. There we will also compare them with oracle estimators that make use of the unobserved quantities $O^{D}_{+,d}$. These quantities allow to calculate the values of $\hat{O}^{D}_{+,d}$, $\hat{O}^{R}_{+,d}$ and $\hat{E}_{+,d}$ so we can calculate the following infeasible estimators, which we call oracle estimators:

$$\hat{\alpha}_y^{\text{oracle},v} = \frac{\sum_{d=1}^{\mathcal{D}+1} \bar{K}_{b,y}(y-d) O^{v}_{+,d}}{\sum_{d=1}^{\mathcal{D}+1} \bar{K}_{b,y}(y-d) E_{+,d}},$$

with v equal to D, R or a blank.

In Section 9 we will discuss the asymptotic performance of our estimator. For simplicity we will only do that for a modified version where no smoothing is applied in the updating of the hazard $\hat{\alpha}_y^{(r+1)}$. This means that we replace (2) by the update

$$\hat{\alpha}_y^{(r+1)} = \frac{O^{(r)}_{+,y}}{E^{(r)}_{+,y}} \tag{3}$$

for $y = 1, \ldots, \mathcal{D}$ and put $\hat{\alpha}_y^{v} = \hat{O}^{v}_{+,d}/\hat{E}_{+,d}$, with v equal to D, R or a blank. There we will also compare these estimators with the oracle estimators that are now given as $\hat{\alpha}_y^{\text{oracle},v} = O^{v}_{+,d}/E_{+,d}$, again with v equal to D, R or a blank. We will argue that our estimators converge with the same rate of convergence as the oracle estimator only if the values of γ_x show an irregular pattern. Otherwise the rate may be slower.

7. Empirical analysis of Covid-19 duration in France

We consider publicly available data from France which consist of aggregated counts of daily accumulated occurrences (deaths and recoveries) and counts of daily hospitalized people. The data record extends from 24 January 2020, however there is no reliable information on the number of hospitalized persons until 18 March 2020. So, in our analysis we have taken data only from 18 March 2020 to 1 November 2020 (following the notation of the previous sections we have that $M = 229$), which amounts to about $n = 166987$ observations.

Our goal is to estimate the hazard function of the elapsed time since admission in hospital until death or recovery. Through the deterministic algorithm presented above we estimate these two hazards separately. Results were very similar for the random version as expected for the large sample size we have (see Section 8). On the one hand we consider the length of time in hospital until death, and, on the other hand, the length of time in hospital until the individual is discharged. In the second step of the algorithm we need to choose the bandwidth b for the hazard estimators. In this data analysis we have used the double one-sided cross-validation method of Gámiz et al. (2016) that is implemented in the R package DOvalidation (Gámiz et al., 2017). The final hazard estimates are shown in Fig. 2. The hazard for the time to leave the hospital alive since admission indicates that there is around a 6% chance of recovery every day in the beginning of hospitalization. This percentage is decreasing till around 2% for longer durations. The hazard for the time-to-death in hospital indicates that the probability of dying is around 1.5% per day in the beginning of hospitalization and it decreases below 0.5% for longer hospitalizations.

Age and gender seem to be risk factors for Covid-19 (ISARIC (2020), Horwitz et al. (2021)). Using available information about gender and age we have rerun the algorithm for subgroups. Fig. 3 shows the hazard estimates considering four age groups: less than 40 years, between 40 and 60, between 60 and 80, and 80 years and above. We can see that age is a

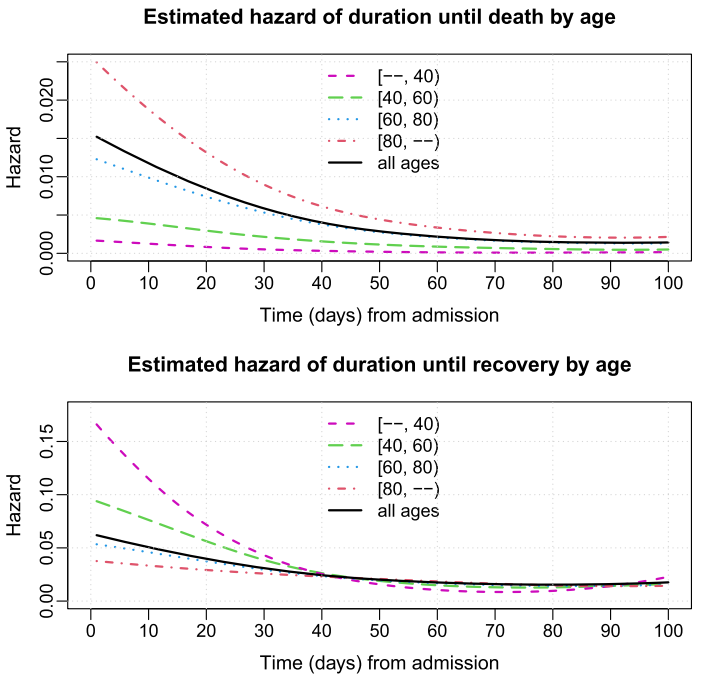

Fig. 3. Estimated hazards for deaths (top) or recoveries (bottom) for all individuals (solid) and by age groups: less than 40 years (pink small dashes), between 40 and 60 (green dashes), between 60 and 80 (blue dots), and more than 80 (red dot-dash). (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

significant marker in both the chances of recovering (graph on the top) and dying (graph on the bottom). Older people (above 80 years) have about 2.5% risk of dying at the beginning of the hospitalization, roughly double than people between 60 and 80 years, and almost five times bigger than those in the 40-60 age group. The risk of dying is decreasing with the time of hospitalization as well as differences among age groups. For the recovery rates, at the beginning of hospitalization the chances of recovering are above 10% for the younger people, around 5% for ages between 60 and 80 years and about 4% for the older ones (above 80). Again differences among groups are attenuated as the time in hospital increases.

The hazard estimates for women and men are shown in Fig. 4. Gender seems to have less influence in both the chances of dying and recovering. Death hazards of men and women only differ slightly at the beginning of the hospitalization. The hazards are around 1.5% at the beginning of the hospitalization time and move below 0.5% for longer times. Women and men recovery rates almost overlap with slightly better chance of recovery for women.

From the hazard derived by our algorithm we have estimated the expected remaining time (or residual time) in hospital as a function of the duration. This is the additional time that a subject is expected to stay in the hospital conditioned to the number of days he/she has already been hospitalized. Formally, let W_d be the remaining time in hospital for a patient who arrived before day d. The expected remaining time in hospital is $E(W_d) = \{S(d)\}^{-1} \int_d^\infty S(w)\mathrm{d}w$. We estimate this expression by plugging-in an estimator of the survival function $S(\cdot)$ derived from the algorithm. The results considering age groups are shown in the top panel of Fig. 5 and compared with results considering the full sample (all ages). For the full sample, the estimated remaining time agrees with the first increasing and after decreasing hazard estimate shown on the bottom graph. For subjects just admitted in hospital the mean hospitalization time is about 20 days however, as time passes, this value increases reaching to 40 days after the first month in hospital. The expected remaining time in hospital decreases gradually after two months of stay. We can see that the estimated residual time in hospital changes with the age of the subject (see also Fig. 5). At the beginning of the hospitalization people between 40 and 60 years are expected to stay for around two weeks (a bit less, about 10 days for ages below 40 years), while older people are expected to stay in hospital for a bit more than three weeks. As the time passes, younger people (40-60 years) who have been in hospital for one week are expected to stay around three more weeks hospitalized, and again this time increases above four weeks for subjects above 60 years.

The residual time has also been estimated separately for men and women and shown in Fig. 6. We can see that residual time in hospital for men and women is very similar, with slight differences for longer stays where women are expected to stay about 2 days less than men.

Also we have estimated the cause-specific (death or recovery) probabilities of leaving the hospital in terms of length of time since admission. Defining random variables for each subject: W^D (time from admission until death) and W^R (time from admission until recovery). The probability of getting out alive is

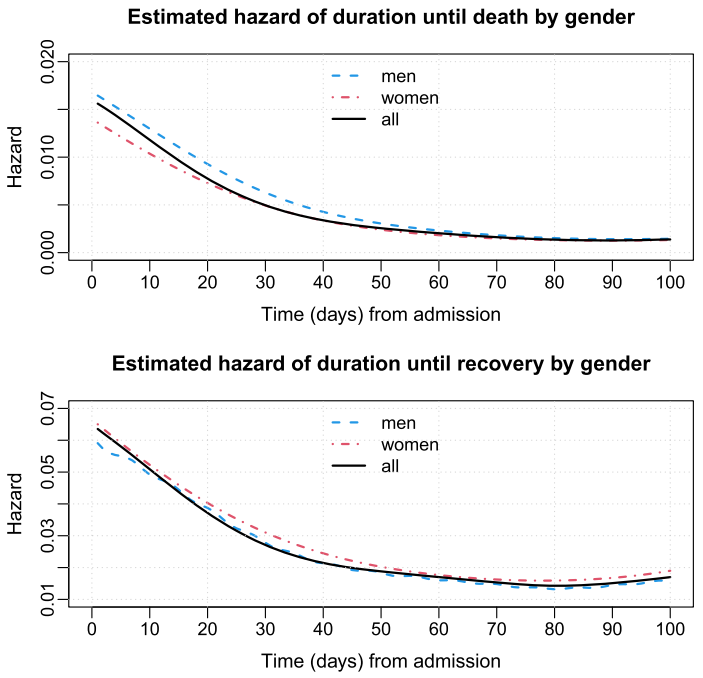

Fig. 4. Estimated hazards for deaths (top) or recoveries (bottom) for all individuals (solid), for men (blue small dashes) and for women (red dot-dash).

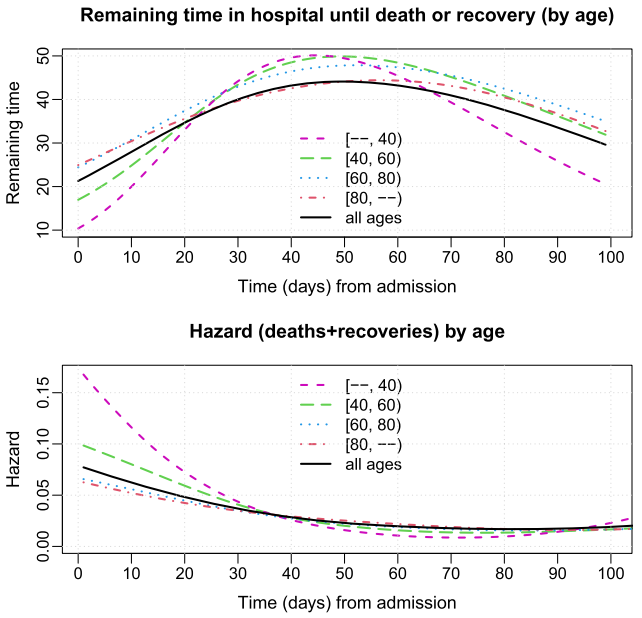

Fig. 5. Estimated remaining time in hospital for all individuals (solid) and by age groups: less than 40 years (pink small dashes), between 40 and 60 (green dashes), between 60 and 80 (blue dots), and more than 80 (red dot-dash).

$$\mathrm{pr}\{W^D > W^R \mid W > d\} = \{S(d)\}^{-1} \int_d^\infty S^D(w)\alpha^R(w)\mathrm{d}w,$$

and the probability of dying in hospital is

$$\mathrm{pr}\{W^R > W^D \mid W > d\} = \{S(d)\}^{-1} \int_d^\infty S^R(w)\alpha^D(w)\mathrm{d}w.$$

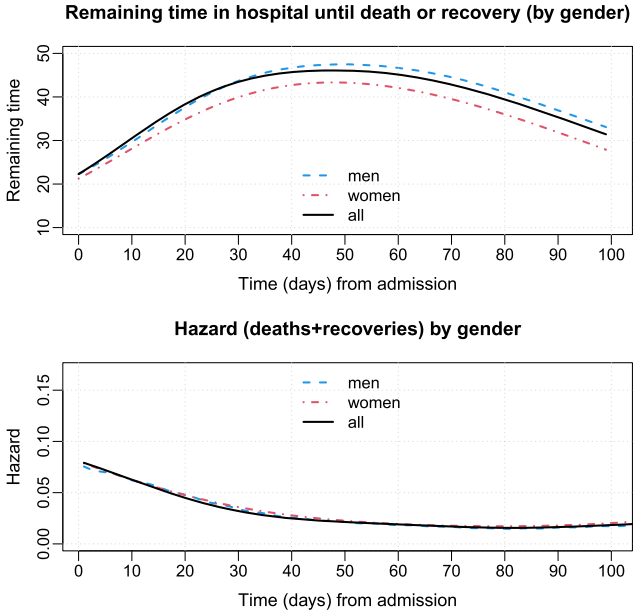

Fig. 6. Estimated remaining time in hospital for all individuals (solid), for men (blue small dashes) and for women (red dot-dash).

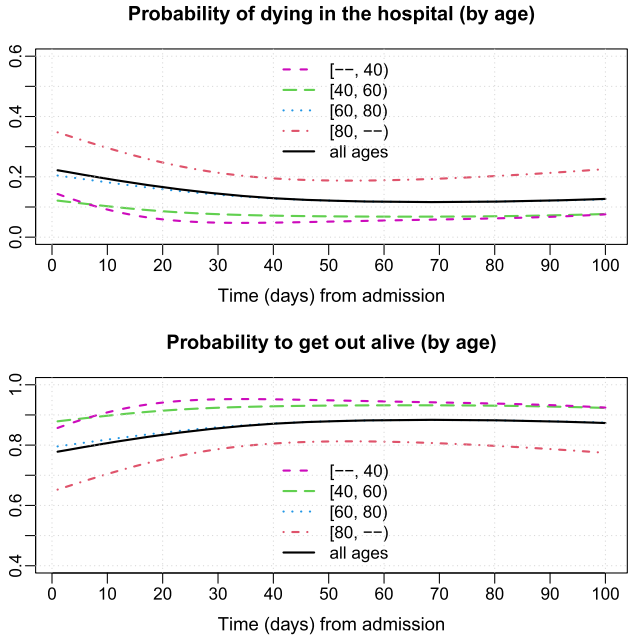

Fig. 7. Cause-specific (death or recovery) probabilities of leaving the hospital in terms of length of time since admission. Analysis for all individuals (solid) and by age groups: less than 40 years (pink small dashes), between 40 and 60 (green dashes), between 60 and 80 (blue dots), and more than 80 (red dot-dash).

The results are shown in Fig. 7 and Fig. 8, considering age and gender classification, respectively, and compared to the results considering the full sample. From Fig. 7 the probability of getting out of the hospital alive (bottom graph) slightly increases with the time that the subject has already been hospitalized. This increase is more substantial for the older people (above 80 years) in the first month in hospital. Similarly the probability of dying in hospital decreases with hospitalization time. For the older hospitalized people (above 80 years) the probability of dying in the first day in hospital is about 0.35, this reduces to about 0.2 for people between 60 and 80 years, and about 0.1 for people below 60 years. After one month is hospital people above 80 years will reduce their probability of dying up to 20%. Considering a gender classification (see Fig. 8) we can see that the probability of getting out of the hospital alive or dead varies with gender in about 5%. This

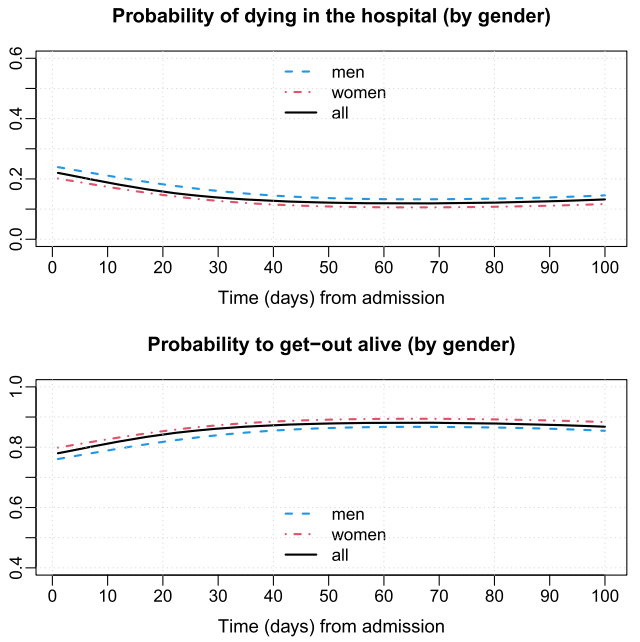

Fig. 8. Cause-specific (death or recovery) probabilities of leaving the hospital in terms of length of time since admission. Analysis for all individuals (solid), for men (blue small dashes) and for women (red dot-dash).

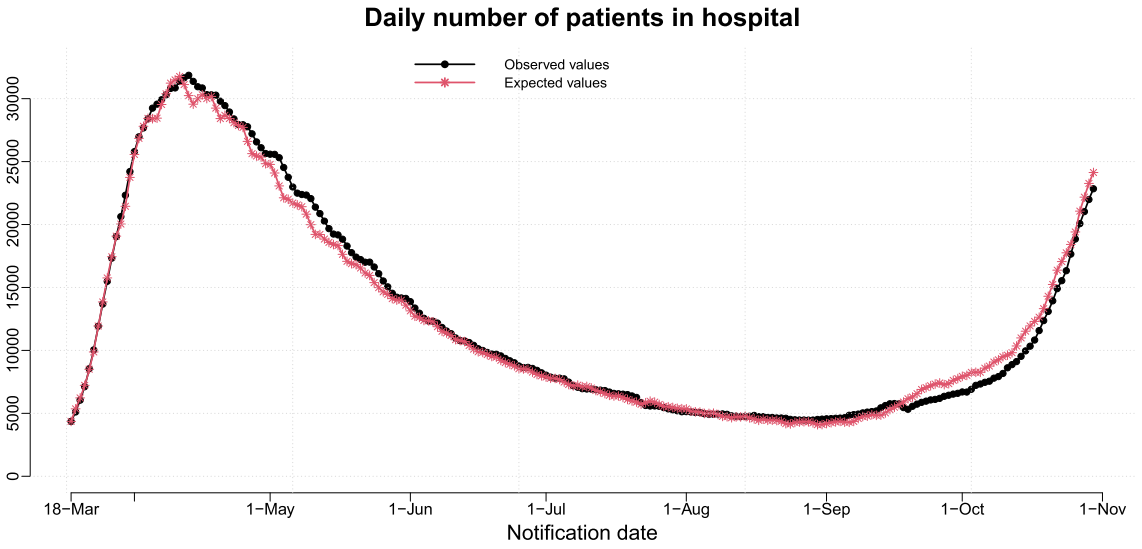

Fig. 9. Time series of daily hospitalized. The black curve represents the number of patients in hospital reported every day since March 18th until November 1st (E_x). The red curve shows the expected number of patients in hospital every day in the same period of time ($\widehat{E}_x$).

difference decreases until around 2% as time in hospital passes. Women in hospital will survive the disease with about 2% more chances than men.

Finally, to confirm whether the estimated hazard function works properly in the real data application, we have produced Fig. 9 where the observed time series of patients in hospitals reported every day is compared with the expected daily number of patients in hospitals estimated every day using our methodology.

8. Simulations

In this section we evaluate the performance of the deterministic algorithm and the random version considering simulated data. The simulation settings have been chosen to mimic the design of the Covid-19 data analysed above. We have defined two theoretical models consisting of a model for the new arrivals and hazard specifications for the time spent in hospital until death, α^D, and time until recovery, α^R. In both cases we assume a non-homogeneous Poisson process (NHPP)

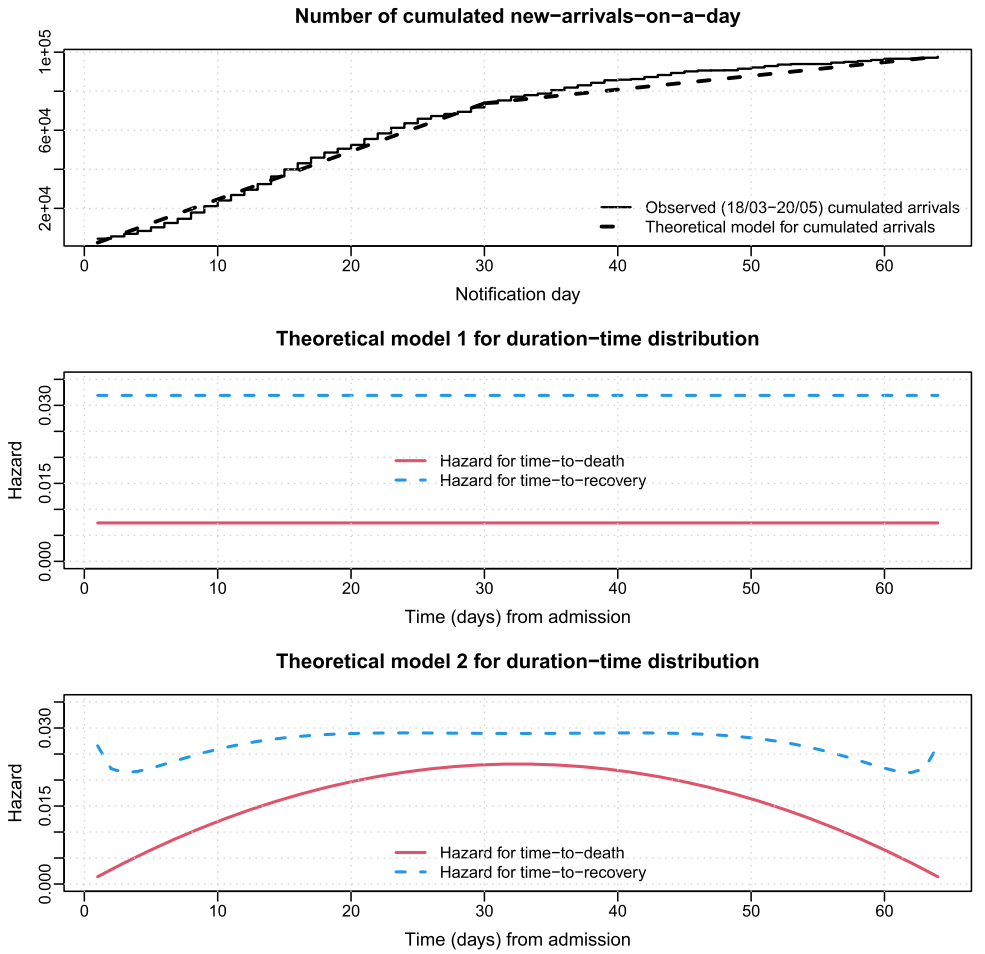

Fig. 10. True simulated models: cumulated new hospitalized by day (top); two hazard models (middle and bottom) for the time spent in hospital until death (red solid) and for time spent in hospital until recovery (blue small-dashes).

for the new arrivals with piecewise-constant intensity function. The intensity has been estimated from the Covid-19 data (restricted to the first $M = 64$ days, this is, until 20th May) considering one change-point. This model is shown in the first plot of Fig. 10 and represents well the observed cumulative new arrivals in the data, also shown in the same plot. The hazard models are plotted in the second and third panels of Fig. 10. Model 1 assumes a constant hazard for time to death and time to recovery, this is, exponential distributions with rate 0.0074 for deaths and rate 0.0315 for recoveries (both values have been estimated again from the data). Model 2 assumes non-constant hazards described by the following Beta models: $\alpha^D(x) = B(x/65; 2, 2)/65$ and $\alpha^R(x) = (0.6/65)\{B(x/65; 0.5, 0.5) + B(x/65; 2, 4) + B(x/65; 4, 2)\}$, for $x \in (0, 65)$, where $B(t; a, b)$ denotes the density at t of a Beta distribution with parameters (a, b).

From each model we have simulated data with sample sizes $n = 10^4, 10^5$ and 10^6, following the steps described in Appendix B. For each case we simulate 500 samples and construct two types of hazard estimators. On the one hand the oracle estimators, $\hat{\alpha}^{\text{oracle},D}$ and $\hat{\alpha}^{\text{oracle},R}$, using full information. And on the other hand the estimators $\hat{\alpha}^D$ and $\hat{\alpha}^R$ derived from the two versions of the algorithm described above, using only partial information. The performance of all methods is evaluated through the integrated squared error (ISE) criterion. Giving an estimator $\hat{\alpha}$ of a hazard α, we define $\text{ISE}(\hat{\alpha}) = \int \left\{\hat{\alpha}(w) - \alpha(w)\right\}^2 \mathrm{d}w$. All hazard estimators have been computed using infeasible optimal bandwidths derived from the same ISE criterion.

The simulation results are presented in Tables 1 and 2. For each model and sample size we report the average (along the 500 samples) of the ISE values (MISE), as well as the median of these values (MedISE). Additionally, the MISE has been decomposed into bias (ISB) and variance (MIV) terms, computed as $\text{ISB} = \int \{\overline{\alpha}(w) - \alpha(w\}^2 \mathrm{d}w$ and $\text{MIV} = \sum_{s=1}^{500} \int \left\{\hat{\alpha}^{(s)}(w) - \overline{\alpha}^{(s)}(w)\right\}^2 \mathrm{d}w/500$, respectively, where $\overline{\alpha}(w) = \sum_{s=1}^{500} \hat{\alpha}^{(s)}(w)/500$, with $\hat{\alpha}^{(s)}$ being the estimator computed from the s-th sample.

From the numbers reported in the tables we can see that the deterministic and random version perform very similarly, with slightly better performance of the deterministic algorithm for Model 1, and the random version for Model 2. These

Table 1
Simulation results based on 500 samples from Model 1. Numbers have been multiplied by 10^9.

		Full information		Partial information			
		Oracle		Deterministic		Random	
n	Criteria	Death	Recovery	Death	Recovery	Death	Recovery
10^3	MedISE	42.4690	227.9892	42.4895	520.497	43.0862	372.1278
	MISE	87.8287	414.9680	99.8704	1295.563	134.8471	1982.1454
	ISB	0.0424	1.4865	8.1855	203.066	13.3475	274.0193
	MIV	87.7863	413.4815	91.6850	1092.497	121.4996	1708.1261
10^4	MedISE	0.4868	2.1474	0.8224	11.7252	4.0191	73.5455
	MISE	0.9047	4.3089	1.6167	24.0084	5.1414	88.8570
	ISB	0.0001	0.0136	0.6567	12.0568	3.5391	66.8636
	MIV	0.9046	4.2953	0.9601	11.9516	1.6023	21.9934
10^5	MedISE	0.0054	0.0201	0.0638	1.2736	0.0444	0.7981
	MISE	0.0101	0.0406	0.0759	1.3955	0.0632	1.1078
	ISB	0.0001	0.0001	0.0629	1.2403	0.0430	0.8035
	MIV	0.0100	0.0406	0.0130	0.1552	0.0202	0.3043
10^6	MedISE	0.0000	0.0002	0.0076	0.1435	0.0061	0.1184
	MISE	0.0001	0.0004	0.0077	0.1451	0.0064	0.1200
	ISB	0.0000	0.0000	0.0076	0.1439	0.0058	0.1085
	MIV	0.0001	0.0004	0.0001	0.0012	0.0007	0.0116

Table 2
Simulation results based on 500 samples from Model 2. Numbers have been multiplied by 10^9.

		Full information		Partial information			
		Oracle		Deterministic		Random	
n	Criteria	Death	Recovery	Death	Recovery	Death	Recovery
10^3	MedISE	363.0705	412.9618	1304.1770	479.9957	2392.4805	1696.5205
	MISE	747.8199	801.7007	1787.1386	1381.4157	3577.6675	4253.8616
	ISB	191.2974	222.4912	1181.2439	466.8975	2121.0992	1946.5344
	MIV	556.5226	579.2095	605.8947	914.5182	1456.5683	2307.3272
10^4	MedISE	8.2421	12.6770	148.1062	84.8517	65.2069	37.6386
	MISE	11.0830	19.0822	185.0273	153.7580	88.3409	50.3968
	ISB	3.2460	8.9466	156.5900	99.3605	67.3017	34.8546
	MIV	7.8369	10.1356	28.4372	54.3976	21.0392	15.5422
10^5	MedISE	0.1733	0.6863	15.7356	3.6894	9.5892	3.3615
	MISE	0.2012	0.7013	15.5417	3.8643	10.7606	4.3471
	ISB	0.0378	0.3130	15.3606	3.5399	9.8182	3.2527
	MIV	0.1633	0.3883	0.1811	0.3244	0.9424	1.0945
10^6	MedISE	0.0026	0.0174	1.5606	0.3777	1.2725	0.4097
	MISE	0.0030	0.0207	1.5612	0.3757	1.2221	0.4098
	ISB	0.0006	0.0062	1.5600	0.3735	1.2047	0.3821
	MIV	0.0025	0.0144	0.0012	0.0022	0.0174	0.0277

improvements seem to be more significant for the smaller sample sizes. Notice that in our data application the sample size was about 170000 so it is expected from these results that both algorithms work similarly. On the other hand, estimating with partial information, as the two versions of the algorithm do, leads to a loss in terms of the MISE (also MedISE) as expected. Looking at the bias/variance decomposition we can see that the major contribution in the MISE corresponds to the bias, for the two versions of the algorithm. Meanwhile the oracle estimator exhibits quite small bias terms in all cases.

9. Asymptotics

In this section we will discuss asymptotics for the iterative estimation procedure introduced in Section 6. We concentrate on asymptotics for the estimator $\hat{\alpha}$ and provide some heuristics implying that our approach estimates $\hat{\alpha}$ at the optimal asymptotic rate of convergence. There are several ways for an asymptotic framework for our model. There is the length M of the observation period and the number of incoming patients $E_{x,1}$, where we could assume that their expectation γ_x is of order I, uniformly in x for some $I \to \infty$. Furthermore, there is the bandwidth b used in the smoothing step and there is the support of α where we could assume that $\alpha_d = a(d/(\mathcal{D}+1))$ for a fixed smooth function a and some support length parameter $\mathcal{D}$. In a very general asymptotic approach one would let M, I, and $\mathcal{D}$ converge to infinity and b converge to zero. For simplicity below we will discuss the case where $\mathcal{D}$ and b are fixed and where only I and M converge to infinity. For simplicity we assume that $b = 0$, i.e. that there is no smoothing. That means that $\hat{\alpha}_y^{(r+1)}$ is updated by (3).

In this setting we now discuss the first step of our algorithm where $\alpha_1, \ldots, \alpha_D$ are estimated. We make the distributional assumptions stated in Section 5. Thus we assume that $O_{x,d}$ are independent Poisson random variables with parameter $\gamma_{x-d+1}(1-\alpha_1)\cdots(1-\alpha_{d-1})\alpha_d$. For our asymptotic discussion we assume that the parameters $\alpha_1, \ldots, \alpha_D$ are fixed and that the nuisance parameters γ_x, i.e. the expected number of incoming patients at day x, converge to infinity:

(A1) For some constants $0 < c < C < \infty$ and $I \to \infty$ it holds that $cI \le \gamma_x \le CI$, for $x = 1, \ldots, M$.

For $Q = (Q_0, \ldots, Q_{\mathcal{D}+1})$ with $1 = Q_0 \ge \cdots \ge Q_{\mathcal{D}+1} = 0$ we define for $1 \le d \le \mathcal{D}+1$

$$E^*_{x,d}(Q) = Q_{d-1}E_{x-d+1,1},$$
$$O^*_{x,d}(Q) = (Q_{d-1} - Q_d)E_{x-d+1,1}.$$

For an estimator $\hat\alpha$ of α and for the underlying α we define $\hat Q$ and Q^* by $\hat Q_0 = Q^*_0 = 1$, $\hat Q_{\mathcal{D}+1} = Q^*_{\mathcal{D}+1} = 0$, and $\hat Q_d = (1-\hat\alpha_d)\cdots(1-\hat\alpha_1)$, $Q^*_d = (1-\alpha_d)\cdots(1-\alpha_1)$ for $1 \le d \le \mathcal{D}$. Note that $E^*_{x,d}(\hat Q) = E^*_{x,d}(Q) = 0$ for $x, d = 1, \ldots, \mathcal{D}+1$ $(x < d)$, because of $E_{x,1} = 0$ for $x < 0$. For the description of our iterative procedure define

$$q^*_{x,d}(\hat Q) = \frac{O^*_{x,d}(\hat Q)}{\sum_{l=1}^{\mathcal{D}+1} O^*_{x,l}(\hat Q)}, \quad h^*_{x,d}(\hat Q) = \frac{E^*_{x,d}(\hat Q)}{\sum_{l=1}^{\mathcal{D}+1} E^*_{x,l}(\hat Q)},$$
$$O^*_{+,d}(\hat Q) = \sum_{x=1}^{M} q^*_{x,d}(\hat Q)O_x, \quad E^*_{+,d}(\hat Q) = \sum_{x=1}^{M} h^*_{x,d}(\hat Q)E_x.$$

In the r-th iteration cycle of the first step of our algorithm we replace $\hat\alpha^{(r-1)}$ by

$$\hat\alpha^{(r)}(y) = \frac{\sum_{d=1}^{\mathcal{D}+1} \bar K_{b,y}(y-d)O^*_{+,d}(\hat Q^{(r-1)})}{\sum_{d=1}^{\mathcal{D}+1} \bar K_{b,y}(y-d)E^*_{+,d}(\hat Q^{(r-1)})},$$

with $\hat Q^{(r-1)}_d = (1-\hat\alpha^{(r-1)}_d)\cdots(1-\hat\alpha^{(r-1)}_1)$. As said, we will discuss the estimator only for the case $b = 0$ of no smoothing. Then we have the update $\hat\alpha^{(r)}_d = O^*_{+,d}(\hat Q^{(r-1)})/E^*_{+,d}(\hat Q^{(r-1)})$, for $d = 1, \ldots, \mathcal{D}$. With

$$E^*_x(Q) = \sum_{d=1}^{\mathcal{D}+1} Q_{d-1}E_{x-d+1,1},$$
$$O^*_x(Q) = \sum_{d=1}^{\mathcal{D}+1} (Q_{d-1} - Q_d)E_{x-d+1,1},$$

we have $O^*_x(\hat Q) = \sum_{d=1}^{\mathcal{D}+1} \hat\alpha_d \hat Q_{d-1}E_{x-d+1,1}$ and we get

$$\hat\alpha^{(r)}_d = \frac{\sum_{x=1}^{M} \frac{O^*_{x,d}(\hat Q^{(r-1)})O_x}{O^*_x(\hat Q^{(r-1)})}}{\sum_{x=1}^{M} \frac{E^*_{x,d}(\hat Q^{(r-1)})E_x}{E^*_x(\hat Q^{(r-1)})}} = \frac{\sum_{x=1}^{M} \frac{\hat\alpha^{(r-1)}_d \hat Q^{(r-1)}_{d-1}E_{x-d+1,1}O_x}{O^*_x(\hat Q^{(r-1)})}}{\sum_{x=1}^{M} \frac{\hat Q^{(r-1)}_{d-1}E_{x-d+1,1}E_x}{E^*_x(\hat Q^{(r-1)})}} = \hat\alpha^{(r-1)}_d \frac{\sum_{x=1}^{M} \frac{E_{x-d+1,1}O_x}{O^*_x(\hat Q^{(r-1)})}}{\sum_{x=1}^{M} \frac{E_{x-d+1,1}E_x}{E^*_x(\hat Q^{(r-1)})}},$$

which is equivalent to

$$\hat\alpha^{(r)}_d = \hat\alpha^{(r-1)}_d \left[1 + \frac{\sum_{x=1}^{M} E_{x-d+1,1}\left\{\frac{O_x}{O^*_x(\hat Q^{(r-1)})} - \frac{E_x}{E^*_x(\hat Q^{(r-1)})}\right\}}{\sum_{x=1}^{M} \frac{E_{x-d+1,1}E_x}{E^*_x(\hat Q^{(r-1)})}}\right].$$

We now assume that this iterative algorithm converges to a fix point $\hat\alpha$, $\hat Q$ of the equation. Then we get the following equation for the fix point:

$$\hat\alpha_d = \hat\alpha_d \left[1 + \frac{\sum_{x=1}^{M} E_{x-d+1,1}\left\{\frac{O_x}{O^*_x(\hat Q)} - \frac{E_x}{E^*_x(\hat Q)}\right\}}{\sum_{x=1}^{M} \frac{E_{x-d+1,1}E_x}{E^*_x(\hat Q)}}\right].$$

This is equivalent to

$$F(\hat Q) = 0, \tag{4}$$

where

$$F_d(Q) = \frac{1}{IM}\sum_{x=1}^{M} E_{x-d+1,1}\left\{\frac{O_x}{O_x^*(Q)} - \frac{E_x}{E_x^*(Q)}\right\},$$

for $d = 1, \ldots, \mathcal{D}$. For the derivative F' of F we have that

$$F'_{d,d'}(Q) = \frac{1}{IM}\sum_{x=1}^{M} E_{x-d+1,1}\left\{\frac{O_x}{O_x^*(Q)^2}(E_{x-d'+1,1} - E_{x-d',1}) - \frac{E_x}{E_x^*(Q)^2}E_{x-d'+1,1}\right\}.$$

Using that $I^{-1}O_{x,d} - I^{-1}(Q^*_{d-1} - Q^*_d) = o_P(1)$ uniformly for $1 \le x \le M$, $1 \le d \le \mathcal{D}$, we get that

$$F'(Q^*) = \Gamma_{M,I} + o_p(1),$$

where

$$\Gamma_{M,I,d,d'} = \frac{1}{M}\sum_{x=1}^{M}\gamma_{x-d+1}\left\{\frac{\gamma_{x-d'+1} - \gamma_{x-d'}}{\sum_{s=1}^{\mathcal{D}+1}\gamma_{x-s+1}(Q^*_{s-1} - Q^*_s)} + \frac{\gamma_{x-d'+1}}{\sum_{s=1}^{\mathcal{D}+1}\gamma_{x-s+1}Q^*_{s-1}}\right\}.$$

In the following lemma we will state that $\hat{Q}$ achieves an optimal rate of convergence if $\Gamma_{M,I}$ is uniformly invertible:

(A2) $\Gamma_{M,I}$ is invertible with bounded operator norm of the inverse.

Lemma 1. *Make Assumptions (A1) and (A2). Then it holds that the equation $F(\hat{Q}) = 0$ in (4) has a solution $\hat{Q}$ with $\hat{Q}_d - Q^*_d = O_P((IM)^{-1/2})$.*

For a proof of the lemma we make use of the Newton-Kantorovich theorem, see Deimling (1985), Section 15, for example. According to this theorem it suffices to show that $\|F'(Q^*)^{-1}F(Q^*)\| = O_P((IM)^{-1/2})$, $\|F'(Q^*)^{-1}\| = O_P(1)$, and $\|F'(Q^1) - F'(Q^2)\| = O_P(1)$, uniformly for vectors Q^1 and Q^2 in a neighbourhood of Q^*. These claims can be easily shown. E.g. one shows $\|F(Q^*)\| = O_P((IM)^{-1/2})$ by using that given $E_{x,1}$ $(x = 1 \ldots, M)$, the summands $V_x = E_{x-d+1,1}[O_x/O_x^*(Q^*) - E_x/E_x^*(Q^*)] = E_{x-d+1,1}[\{O_x - O_x^*(Q^*)\}/O_x^*(Q^*) - \{E_x - E_x^*(Q^*)\}/E_x^*(Q^*)]$ have conditional mean zero and are of order $O_P(I^{-1/2})$, and that V_x and V_z are conditional independent for $|x - z| > \mathcal{D} + 1$. For a check of the other claims note also that the denominators $O_x^*(Q)$ and $E_x^*(Q)$ can be easily bounded from below.

Assumption (A2) is a strong condition that requires that γ_x depends irregularly on x. For a constant γ_x or for a γ_x that smoothly depends on x condition (A2) may be violated. But also in the case that (A2) does not hold we conjecture that one can show consistency but only with a slower rate of convergence. One can show that $\sqrt{I}(F'(Q^*) - \Gamma_{M,I})$ has a normal limit. Using that multivariate normal random variables take values in a lower dimensional manifold with probability 0 we conjecture that one can show that with high probability the smallest eigenvalue of the random matrix $\sqrt{I}F'(Q^*)$ is bounded from below. Then, by applying again the Newton-Kantorovich theorem, we get a rate of convergence of the order $O_P\{I^{-1/2}(I/M)^{1/2}\} = O_P(M^{-1/2})$ which is slower by a factor $I^{1/2}$. We guess that depending on the irregularity of γ_x all rates between $O_P(M^{-1/2})$ and $O_P\{(IM)^{-1/2}\}$ are possible.

The rates of $\hat{Q} - Q$ carry over to the same rates for $\hat{\alpha}_d - \alpha_d$ $(d = 1, \ldots, \mathcal{D})$ if one assumes that $\alpha_1, \ldots, \alpha_D$ are bounded away from 0 and 1. One can compare the rates with the rates that hold for the oracle estimator, $\hat{\alpha}_d^{\text{oracle}} = O_{+,d}/E_{+,d}$. Because $E_{+,d}$ is of order $O_P(IM)$ one gets that the oracle estimator converges with rate $O_P\{(IM)^{-1/2}\}$, which is identical with the fastest possible rate for $\hat{\alpha}_d - \alpha_d$.

10. Conclusion

Our paper develops a survival analysis approach for aggregated data where only information on the number of individuals entering or leaving a status is available but no information can be collected how long the individuals have been in the status. We show that also in this set-up valid statistical inference can be made on the distribution of durations of individuals in the status. Our main motivation for studying this model came from applications to study the temporal development of durations of hospital stays of Covid-19 patients. A French data set on the number of hospitalized Covid-19 patients served also as a main illustration for the power of our approach. We show that much detailed information can be achieved and that the received statistical information is comparable to the case that the data are fully observed for all individuals. This conclusion is also supported by simulations.

Acknowledgement

The authors thank two anonymous reviewers and the Associate Editor for many valuable comments and suggestions, which have helped to improve the quality of the article. This work has been partially supported by ERDF / Spanish Ministry of Science and Innovation - State Research Agency through grant numbers PID2020-120217RB-I00 and PID2020-116587GB-I00; the Spanish Junta de Andalucía through grant number B-FQM-284-UGR20; and the IMAG-Maria de Maeztu grant CEX2020-001105-/AEI/10.13039/501100011033. Data for the Covid-19 application have been downloaded from http://www.data.gouv.fr/es/datasets/.

Appendix A. Glossary

Table A.3 gives a brief description of the counting processes and other relevant information that are defined and used across the paper.

Table A.3
Glossary of terms.

Continuous time		**Discrete time**	
Time variables		*Time variables*	
t	Notification time $t \in (0, +\infty)$	x	Notification time, $x \in \{1, 2, \ldots\}$
s	Duration time $s > 0$	d	Duration time, $d \in \{1, 2, \ldots, \mathcal{D}+1\}$
Available information		*Available information*	
$\tilde{N}_1(t)$	Number of arrivals in $(0, t]$	$E_{x,1}$	New arrivals at time x
$\tilde{N}_2(t)$	Number of departures (occurrences) in $(0, t]$	O_x	Occurrences notified at time x
		E_x	Number of subjects at risk at x
Non-available information		*Non-available information*	
$N_{2,i}(s)$	1, if subject i arriving at t_i leaves in $(t_i, t_i + s]$; and 0, otherwise	$O_{x,d}$	Occurrences notified at time x with duration d
$N_{2,t}(s)$	Number of departures in $(t, t+s]$ among subjects arriving at time t	$E_{x,d}$	Number of subjects at risk at x with duration d

Appendix B. Simulation scheme

To generate one sample from each model in the simulations we follow the following steps:

1. Fix the sample size n.
2. Generate $E_{1,1}, E_{2,1}, \ldots, E_{M,1}$, with $E_{x,1}$ the number of new arrivals in the interval $(x-1, x]$, given that $n = \sum_{x=1}^{M} E_{x,1}$. Considering the specifications above, we fix some restrictions for the new arrivals, that is $\sum_{x=1}^{M_0} E_{x,1} = n_0$ and $\sum_{x=M_0+1}^{M} E_{x,1} = n - n_0$, with $n_0 = 0.75\ n$.
3. Construct matrices $O_{x,d}$ and $E_{x,d}$, for $1 \le d \le x \le M$, based on the true hazard α. For a fixed $x \in \{1, 2, \ldots, M\}$, put $d = 1$ and randomly simulate $O_{x,d}$ from a Binomial distribution with size $E_{x,d}$ and probability $\alpha(d)$. Then, define $E_{x+1,d+1} = E_{x,d} - O_{x,d}$, and continue simulating $O_{x+1,d+1}$ from a Binomial distribution with size $E_{x+1,d+1}$ and probability $\alpha(d+1)$, continue for $d = 1, 2, \ldots, M - x + 1$, then we construct matrices for the corresponding sequences of occurrences and exposure. To obtain the occurrences attending to a specific cause, death or recovery, we define $O^D_{x,d} = O_{x,d} \frac{\alpha^D(d)}{\alpha(d)}$, for deaths, and $O^R_{x,d} = O_{x,d} - O^D_{x,d}$, for recoveries, for all $1 \le d \le x \le M$.

From a sample we construct the oracle estimators using matrices $O^D_{x,d}$, $O^R_{x,d}$ and $E_{x,d}$, for all $1 \le d \le x \le M$. Full information in terms of duration is then accessible through the sequences: $O^D_d = \sum_{x=1}^{M} O^D_{x,d}$, $O^R_d = \sum_{x=1}^{M} O^R_{x,d}$ and $E_d = \sum_{x=1}^{M} E_{x,d}$, for $1 \le d \le M$. On the other hand, for the two versions of the algorithm in Section 6 we consider that only partial information is available as in our data application. More specifically, the sample information is limited to the following sequences $O^D_x = \sum_{d=1}^{M} O^D_{x,d}$; $O^R_x = \sum_{d=1}^{M} O^R_{x,d}$; and $E_x = \sum_{d=1}^{M} E_{x,d}$, for $1 \le x \le M$. The algorithm is then applied to $\{(O^D_x, O^R_x, E_x), x = 1, \ldots, M\}$ to estimate data related to time-duration in hospital.

References

Allassonniere, S., Chevallier, J., 2021. A new class of stochastic EM algorithms. Escaping local maxima and handling intractable sampling. Comput. Stat. Data Anal. 159, 107159.

Andersen, P., Borgan, O., Gill, R., Keiding, N., 1993. Statistical Models Based on Counting Processes. Springer, New York.

Bingham, N.H., Pitts, S.M., 1999. Nonparametric estimation for the M/G/∞ queue. Ann. Inst. Stat. Math. 51, 71–97.

Blanche, P., Proust-Lima, C., Loubere, L., Berr, C., Dartigues, J.F., Jacqmin-Gadda, H., 2015. Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. Biometrics 71, 102–113.
Brookmeyer, R., 1996. AIDS, epidemics, and statistics. Biometrics 52, 781–796.
Brookmeyer, R., Gail, M.H., 1988. A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. J. Am. Stat. Assoc. 83, 301–308.
Deimling, K., 1985. Nonlinear Functional Analysis. Springer, New York.
Farebrother, R.W., 1979. Estimation with aggregated data. J. Econom. 10, 43–55.
Ferrer, L., Putter, H., Proust-Lima, C., 2019. Individual dynamic predictions using landmarking and joint modelling: validation of estimators and robustness assessment. Stat. Methods Med. Res. 28, 3649–3666.
Gámiz, M.L., Kulasekera, K.B., Limnios, N., Lindqvist, B.H., 2011. Applied Nonparametric Statistics in Reliability. Springer Series in Reliability Engineering. Springer-Verlag, London.
Gámiz, M.L., Mammen, E., Martínez-Miranda, M.D., Nielsen, J.P., 2016. Double one-sided cross-validation of local linear hazards. J. R. Stat. Soc. B 78, 755–779.
Gámiz, M.L., Mammen, E., Martínez-Miranda, M.D., Nielsen, J.P., 2017. DOvalidation: Kernel Hazard Estimation with Best One-Sided and Double One-Sided Cross-Validation. R package version 1.1.0.
Goldenshluger, A., 2016. Nonparametric estimation of service time distribution in the M/G/∞ queue and related estimation problems. Adv. Appl. Probab. 48, 1117–1138.
Goldenshluger, A., Koops, D.T., 2018. Nonparametric estimation of service time characteristics in infinite-server queues with stationary Poisson input. Preprint. arXiv:1805.10353v1.
Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–161.
Horwitz, L.I., Jones, S.A., Cerfolio, R.J., Francois, F., Greco, J., Rudy, B., Petrilli, C.A., 2021. Trends in Covid-19 risk-adjusted mortality rates in a single health system. J. Hosp. Med. 16 (2), 90–92. https://doi.org/10.12788/jhm.3552.
ISARIC, 2020. International Severe Acute Respiratory and Emerging Infections Consortium, COVID-19 Report: 04 October 2020.
Jewell, N.P., 1990. Some statistical issues in studies of the epidemiology of aids. Stat. Med. 9, 1387–1416.
Jewell, N.P., 2004. Statistics for Epidemiology. Chapman&Hall/CRC.
King, G., 1997. A Solution to the Ecological Inference Problem. Reconstructing Individual Behavior from Aggregate Data. Princeton University Press.
Lawless, J.F., McLeish, D.L., 1984. The information in aggregate data from Markov chains. Biometrika 71, 419–430.
Liu, Y., Hu, F., 2020. Balancing unobserved covariates with covariate-adaptive randomized experiments. J. Am. Stat. Assoc. https://doi.org/10.1080/01621459.2020.1825450.
Mammen, E., Martínez-Miranda, M.D., Nielsen, J.P., Sperlich, S., 2011. Do-validation for kernel density estimation. J. Am. Stat. Assoc. 106, 651–660.
Martinussen, T., Scheike, T.H., 2006. Dynamic Regression Models for Survival Data. Statistics for Biology and Health. Springer, New York.
Nielsen, J.P., Tanggaard, C., 2001. Boundary and bias correction in kernel hazard estimation. Scand. J. Stat. 28, 675–698.
Pickands, J., Stine, R.A., 1997. Estimation for an M/G/∞ queue with incomplete information. Biometrika 84, 295–308.
Proust-Lima, C., Dartigues, J.F., Jacqmin-Gadda, H., 2016. Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: a latent process and latent class approach. Stat. Med. 35, 382–398.
Proust-Lima, C., Sene, M., Taylor, M.G., Jacqmin-Gadda, H., 2014. Joint latent class models for longitudinal and time-to-event data: a review. Stat. Methods Med. Res. 23, 70–90.
Rubin, D.B., 1996. Multiple imputation after 18+ years. J. Am. Stat. Assoc. 91, 473–489.
Zhao, J., Kim, H-J., Kim, H-M., 2020. New EM-type algorithms for the Heckman selection model. Comput. Stat. Data Anal. 146, 106930.