# Maximum likelihood estimation in the additive hazards model

Lu, C.Y.; Goeman, J.; Putter, H.

Received: 11 July 2021 | Accepted: 30 August 2022

DOI: 10.1111/biom.13764

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

# Maximum likelihood estimation in the additive hazards model

Chengyuan Lu | Jelle Goeman | Hein Putter

Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

**Correspondence**
Chengyuan Lu, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands.
Email: c.lu@lumc.nl

**Abstract**

The additive hazards model specifies the effect of covariates on the hazard in an additive way, in contrast to the popular Cox model, in which it is multiplicative. As the non-parametric model, additive hazards offer a very flexible way of modeling time-varying covariate effects. It is most commonly estimated by ordinary least squares. In this paper, we consider the case where covariates are bounded, and derive the maximum likelihood estimator under the constraint that the hazard is non-negative for all covariate values in their domain. We show that the maximum likelihood estimator may be obtained by separately maximizing the log-likelihood contribution of each event time point, and we show that the maximizing problem is equivalent to fitting a series of Poisson regression models with an identity link under non-negativity constraints. We derive an analytic solution to the maximum likelihood estimator. We contrast the maximum likelihood estimator with the ordinary least-squares estimator in a simulation study and show that the maximum likelihood estimator has smaller mean squared error than the ordinary least-squares estimator. An illustration with data on patients with carcinoma of the oropharynx is provided.

KEYWORDS

additive hazards, constrained optimization, maximum likelihood

## 1 | INTRODUCTION

The fundamental concept in survival analysis is the hazard rate. There are several regression models describing the hazard rate, such as multiplicative risk models and additive risk models. The dominant hazard model in survival analysis is the multiplicative proportional hazards model (Cox, 1972). In many applications, the proportional hazards assumption is not met. In such cases, developing models that adequately describe the non-proportional effect of the covariate(s) is not straightforward (Gore et al., 1984; Perperoglou et al., 2006; Schemper, 1992; van Houwelingen & Putter, 2012) and different models should be considered.

In this paper, we focus on the non-parametric additive hazards model proposed by Aalen (1980, 1989), extensively studied in Martinussen and Scheike (2006). The additive hazards model defines the hazard rate as a linear form of the vector of covariates. Unconventionally, it does not naturally force the hazard rate to be positive. However, it has several useful properties, as set out by Aalen et al. (2008). First, the additive hazards model is internally consistent: it retains its additive structure if covariates are measured with uncertainty or are dropped from the linear expression, provided the left out covariates are independent of the other covariates or if the covariates are jointly normally distributed (Aalen et al., 2015; Martinussen et al., 2020). This contrasts with the proportional hazards model, which loses its proportional hazards property when covariates are omitted (Bretagnolle & Huber-Carol, 1988; Struthers & Kalbfleisch, 1986; Schumacher et al., 1987). This drawback of proportional hazards has triggered a recent debate about the danger of using the hazard ratio as a causal effect measure in survival analysis (Aalen et al., 2015; Hernán, 2010). A second advantage of additive hazards is that it allows implementation of dynamic structures (Martinussen et al., 2000), such as self-exciting processes, which are impossible to incorporate in most other nonlinear regression models.

The usual way to estimate the parameters in the additive hazards model is by Aalen's method, which uses ordinary least squares (OLS) to estimate the cumulative effect of the covariates. Aalen's OLS method is straightforward to implement, but it has the disadvantage that it does not guarantee a positive hazard for all time points. One method that generally yields efficient estimators is maximum likelihood (ML). To the best of our knowledge, the ML method has not been considered in any detail in the context of the additive hazards model, since no analytical expression for the maximum likelihood estimator (MLE) was available. The objective of this paper is to derive the MLE for the additive hazards model and to determine its advantages and limitations and in comparison to Aalen's OLS method.

We show that the MLE of the cumulative baseline hazard and cumulative covariate effects is a step function changing values only at the event time points, and that the maximum of the log-likelihood can be found by separately maximizing the log-likelihood contributions corresponding to each of the event time points, as with Aalen's OLS solution. We define the constraint domain given by the positivity of hazard, and show that the maximization problem is equivalent to fitting a series of identity-link Poisson regression models under non-negativity constraints. This problem has been studied before by Marschner (2010) and Marschner et al. (2012), who proposed an EM algorithm to obtain MLEs. In contrast, our solution is analytical, with a computation time at each time point that is linear in the sample size and in the number of covariates.

In Section 2, we introduce notation and in Section 3 we discuss Aalen's OLS method. Section 4 contains the theoretical results on the MLE. In Section 5, we report on the results of a simulation study comparing ML with OLS. Section 6 illustrates our methods on real data from a randomized clinical trial on patients with carcinoma of the oropharynx. The paper ends with a discussion.

## 2 | NOTATION AND MODEL DEFINITION

We use bold letters to indicate vectors and matrices. Define $T^*$ to be the time to event, $C$ to be the time to censoring, and let $T = \min(T^*, C)$, and $\Delta = I(T^* \leq C)$ the event indicator. We observe $(t_i, \delta_i, \mathbf{x}_i^*)$, $i = 1, \ldots, n$, with $\mathbf{x}_i^* = (x_{i1}, \ldots, x_{ip})^\top$ a $p$-vector of covariates. We assume that the covariates $x_{ij}$ are restricted to the interval $[0, 1]$. This looks like a severe restriction, but it is not, because for any covariate with a minimum of $a$ and a maximum of $b$ in the observed data, we can rescale the covariate to be $(x - a)/(b - a)$, which takes values in $[0, 1]$. Extend $\mathbf{x}_i^*$ with the constant $x_{i0} = 1$, obtaining $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$. Define $\mathcal{E}$ to be the set of event time points $t_i$, with corresponding $\delta_i = 1$ and let $|\mathcal{E}| = K$. We assume there are no ties, and define the ordered sequence of event time points $t_1^* < \cdots < t_K^*$. Finally, define $Y_i(t) = I(t_i \geq t)$ as the at risk indicator of subject $i$, taking the value 1 if subject $i$ is at risk for the event of interest at time $t$ and the value 0 otherwise.

The additive hazards model (Aalen, 1980, 1989) assumes the hazard rate given $\mathbf{x}_i^*$ to be of the form

$$h(t \mid \mathbf{x}_i^*) = \mathbf{x}_i^\top \boldsymbol{\beta}(t) = \beta_0(t) + \beta_1(t) x_{i1} + \cdots + \beta_p(t) x_{ip}.$$

The parameters $\beta_j(t)$ allow effects of the covariates to change over time, thus, the additive model is fully nonparametric.

## 3 | AALEN's METHOD

To estimate the parameters $\boldsymbol{\beta}(t)$, the most commonly used approach is Aalen's OLS method. Let us define the counting process $N_i(t)$ as the number of events experienced by subject $i$ before or at time $t$. Then, the intensity $\lambda_i(t)$ of the counting process has the form

$$\lambda_i(t) = Y_i(t)(\beta_0(t) + \beta_1(t) x_{i1} + \cdots + \beta_p(t) x_{ip}) = Y_i(t) h(t \mid \mathbf{x}_i^*).$$

The formula $dN_i(t) = \lambda_i(t)dt + dM_i(t)$ gives equations

$$d\mathbf{N}(t) = \mathbf{X}(t) d\mathbf{B}(t) + d\mathbf{M}(t),$$

where $\mathbf{N}(t) = (N_1(t), N_2(t), \ldots, N_n(t))^\top$ is the vector of counting processes, $\mathbf{X}(t)$ is the matrix of covariates multiplied with $Y_i(t)$ at the $i$th row, $\mathbf{B}(t) = (B_0(t), B_1(t), \ldots, B_p(t))^\top$ is the cumulative beta defined as $B_j(t) = \int_0^t \beta_j(s) ds$ and $\mathbf{M}(t) = (M_1(t), \ldots, M_n(t))^\top$, where $M_i(t)$ is the $i$th martingale error term. OLS regression thus gives Aalen's estimator of the $\boldsymbol{\beta}(t)$ if $X(t)$ is of full rank:

$$d\widehat{\mathbf{B}}(t) = (\mathbf{X}(t)^\top \mathbf{X}(t))^{-1}\mathbf{X}(t)^\top d\mathbf{N}(t).$$

When $X(t)$ is not of full rank, we set $d\widehat{\mathbf{B}}(t) = 0$ (Aalen et al., 2008).

## 4 | MAXIMUM LIKELIHOOD ESTIMATION

In this section, we will derive an analytic solution for the MLE under the natural constraint that the hazard is nonnegative at each time point for all covariate values within the domain. The proofs will be given in Web Appendix A.

### 4.1 | The shape of the likelihood

To begin with, let us consider the likelihood function for the additive hazards model. The log-likelihood contribution of subject $i$ with an event or censored at time $t_i$ is given by

$$\ell_i = \delta_i \log\{h(t_i \mid \mathbf{x}_i^*)\} + \log\{S(t_i \mid \mathbf{x}_i^*)\},$$

where $S(t \mid \mathbf{x}_i^*) = P(T^* > t \mid \mathbf{x}_i^*) = \exp(-H(t \mid \mathbf{x}_i^*))$ is the survival probability and $H(t \mid \mathbf{x}_i^*)$ the cumulative hazard of subject $i$. The log-likelihood is therefore given by (Aalen et al., 2008)

$$\ell = \sum_{i=1}^{n} \{\delta_i \log x_i^\top \boldsymbol{\beta}(t_i) - \mathbf{x}_i^\top \mathbf{B}(t_i)\}. \qquad (1)$$

Our first result is

**Proposition 1.** *The likelihood function $\ell$ is unbounded as a function of $\boldsymbol{\beta}(t)$.*

We therefore introduce the natural constraint that the hazard for each possible covariate value of our data is nonnegative at each time point, that is, we require that for all $t \geq 0$,

$$h(t \mid \mathbf{x}^*) \geq 0, \text{ for } \mathbf{x}^* \in \{0,1\}^p, \qquad (2)$$

which implies positivity of the hazard for $\mathbf{x}^* \in [0,1]^p$. Our objective is to maximize the total log-likelihood given in (1), subject to the constraint (2).

Let us assume $\hat{\mathbf{B}}(t)$ is the function which maximizes the likelihood function (1), subject to the constraint (2). Since $\hat{\mathbf{B}}(t)$ is an estimator of $\mathbf{B}(t)$, which is generally the negative logarithm of the survival function, we may assume that it is right continuous with left limits (cadlag). Thus, we may decompose $\hat{\mathbf{B}}(t)$ as the summation of a continuous function $\hat{\mathbf{B}}_c(t)$ and a step function $\hat{\mathbf{B}}_s(t)$.

**Lemma 1.** *Under the constraint $h(t \mid \mathbf{x}^*) \geq 0$ for $\mathbf{x}^* \in \{0,1\}^p$, the log-likelihood can only achieve a maximum if $\hat{\mathbf{B}}_c(t) \equiv 0$ and $\hat{\mathbf{B}}_s(t)$ is a step function with jumps only at the event time points in $\mathcal{E}$.*

Lemma 1 implies that the maximization problem is the same as maximizing the total log-likelihood with respect to the jumps of $\mathbf{B}(t)$ at the event time points. Denoting the jump of $B_j(t)$ at the $k$th event time point $t_k^*$ as $\beta_{kj}$, $k = 1, \ldots, K$, $j = 0, \ldots, p$, this leads to $B_j(t) = \sum_{k:t_k^* \leq t} \beta_{kj}$.

We assume that there are no ties. The total log-likelihood $\ell$ from (1) can now be rewritten in terms of the $\beta_{kj}$ as

$$\ell = \sum_{i=1}^{n} \left\{ \delta_i \log\left( \sum_{j=0}^{p} x_{ij}\beta_{k(i),j} \right) - \sum_{j=0}^{p} x_{ij} \sum_{k:t_k^* \leq t_i} \beta_{kj} \right\},$$

where $k(i)$ is the index of the event time points $t_k^*$ corresponding to $t_i$, that is, $t_{k(i)}^* = t_i$ (in case $\delta_i = 0$, $k(i)$ can be chosen as an arbitrary index in $\{1, \ldots, n\}$). We reorder this sum over subjects as a sum over the distinct event time points $t_k^*$, obtaining

$$\ell = \sum_{k=1}^{K} \left\{ \log\left( \sum_{j=0}^{p} x_{i(k),j}\beta_{kj} \right) - \sum_{j=0}^{p} \beta_{kj} \sum_{l:t_l \geq t_k^*} x_{lj} \right\} = \sum_{k=1}^{K} \ell_k^*,$$

with

$$\ell_k^* = \log\left( \sum_{j=0}^{p} x_{i(k),j}\beta_{kj} \right) - \sum_{j=0}^{p} \beta_{kj} s_{kj}, \qquad (3)$$

$s_{kj} = \sum_{l:t_l \geq t_k^*} x_{lj}$, and where $i(k)$ is the subject index corresponding to the $k$th event time point, that is, $t_{i(k)} = t_k^*$.

It is easy to see that each term $\ell_k^*$ is a function only of the variables $\beta_{kj}$, $j = 0, \ldots, p$. Since the terms $\ell_k^*$ do not have parameters in common, we may reduce the problem by separately maximizing each $\ell_k^*$ as a function of $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kp})$.

For the remainder of this section, we will fix (any) one of the time points $t_k^*$. To simplify notation, we are going to suppress dependence on $k$, and consider maximization of

$$\ell^* = \log\left( \sum_{j=0}^{p} x_j\beta_j \right) - \sum_{j=0}^{p} s_j\beta_j = \log\left( \mathbf{x}^\top \boldsymbol{\beta} \right) - \mathbf{s}^\top \boldsymbol{\beta} \qquad (4)$$

with respect to $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$, where $\mathbf{s} = (s_0, s_1, \ldots, s_p)^\top$, $x_j$ stands for the $j$th element of the covariate vector of the subject that failed at time $t_k^*$, and $x_0 = 1$. Appendix C of the Supporting Information details the connection with Poisson regression with identity link, in parallel to the connection of Cox's proportional hazards model with Poisson regression with log link (Holford, 1980; Laird & Olivier, 1981; McCullagh & Nelder, 1989).

## 4.2 | Maximum likelihood with the full constraint matrix

We will now give an alternative description of the constraint (2) using matrix form. Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ be the vector of parameters of the function $\ell^*$. We can restate the constraint (2) by the matrix $M_\mathbf{D}$ through the inequalities $M_\mathbf{D}\boldsymbol{\beta} \geq \mathbf{0}$. We construct the matrix $M_\mathbf{D}$ such that the rows consist of all the $2^p$ possible $(p+1)$-tuples $(m_0, m_1, \ldots, m_p)$ in $\{0,1\}^{p+1}$ with $m_0 = 1$. So, $M_\mathbf{D}$ is a $2^p \times (p+1)$ matrix.

The domain $\mathbf{D} \subset \mathbb{R}^{p+1}$ defined by constraint (2) is a polytope enclosed by the flat boundaries defined by $\gamma_i = M_i^\top \boldsymbol{\beta} = 0$, $i = 1, \ldots, 2^p$, where $M_i$ is the $i$th row of $M_\mathbf{D}$. We can now show that the maximum point lies on the boundary of this polytope.

**Lemma 2.** *If the maximum point of $\ell^*$ in the domain $\mathbf{D}$ exists, then the maximum point lies on the intersection of $p$ flat boundaries given by $\gamma_{i_j} = M_{i_j}^\top \boldsymbol{\beta} = 0$, $j = 1 \ldots, p$.*

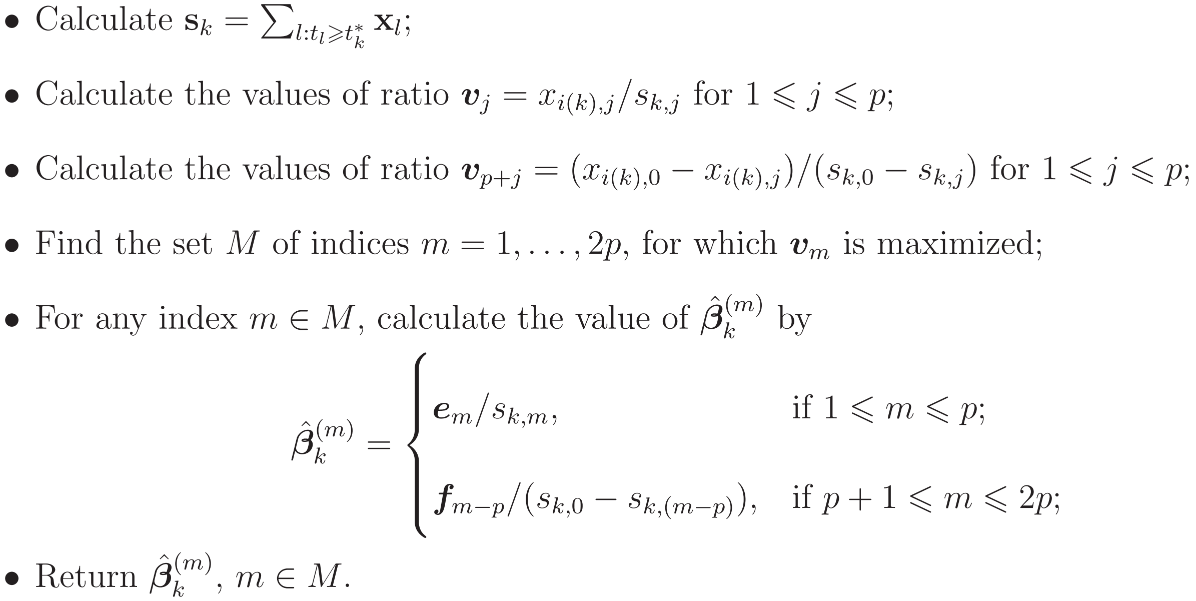
In Lemma 2, we have given the geometric property of the maximum point of the $\ell^*$. The intersection of $p$ flat boundaries is a one-dimensional line. So, the maximum point always lies on a one-dimensional line edge on the boundary of the domain. The following lemma defines all the one-dimensional line edges on the boundary of the domain. There are $2p$ such line edges.

**Lemma 3.** *The one-dimensional line edges on the boundary of the domain $M_\mathbf{D}\boldsymbol{\beta} \geq 0$ are given by $(0, \ldots, 0, l, 0, \ldots, 0)$, $l > 0$, with the position of $l$ ranging from 2 to $p+1$, and $(l, \ldots, 0, -l, 0, \ldots, 0)$, with the position of $-l$ ranging from 2 to $p+1$.*

Let us denote by $\boldsymbol{e}_j$ the $(p+1)$-vector $(0, 0, \ldots, 0, 1, 0, \ldots, 0)^\top$, with the 1 at position $j+1$, and by $\boldsymbol{f}_j$ the $(p+1)$-vector $(1, 0, \ldots, 0, -1, 0, \ldots, 0)^\top$, with the $-1$ at position $j+1$, and define the sets $\mathcal{A}^+ = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p\}$ and $\mathcal{A}^- = \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_p\}$. We call the set $\mathcal{A} = \mathcal{A}^+ \cup \mathcal{A}^-$, a finite set of cardinality $2p$, the set of admissible directions. Lemma's 2 and 3 imply that the maximum of $\ell^*$ lies on one of the edges $\boldsymbol{v} \cdot l$, with $\boldsymbol{v} \in \mathcal{A}$. We now come to our main theorem which describes the structure of the MLE of the additive hazards model under the constraint of non-negativity of the hazards. Let $i(k)$ denote the index of the subject that fails at $t_k^*$, the MLE $\hat{\boldsymbol{\beta}}_k$ at the event time point $t_k^*$ can be determined as follows:

**Theorem 1.** *For an additive hazards model $h(t \mid \mathbf{x}_i^*) = \boldsymbol{\beta}(t)\mathbf{x}_i$ with constraints $h(t \mid \mathbf{x}^*) \geq 0$, with $\mathbf{x}^* \in [0,1]^p$, assuming no ties, the values returned by Algorithm 1are maximum likelihood estimators.*

**Algorithm 1** Algorithm to find the maximum point within the domain **D**.



- Calculate $\mathbf{s}_k = \sum_{l:t_l \geqslant t_k^*} \mathbf{x}_l$;
- Calculate the values of ratio $\boldsymbol{v}_j = x_{i(k),j}/s_{k,j}$ for $1 \leqslant j \leqslant p$;
- Calculate the values of ratio $\boldsymbol{v}_{p+j} = (x_{i(k),0} - x_{i(k),j})/(s_{k,0} - s_{k,j})$ for $1 \leqslant j \leqslant p$;
- Find the set $M$ of indices $m = 1, \ldots, 2p$, for which $\boldsymbol{v}_m$ is maximized;
- For any index $m \in M$, calculate the value of $\hat{\boldsymbol{\beta}}_k^{(m)}$ by

$$\hat{\boldsymbol{\beta}}_k^{(m)} = \begin{cases} \boldsymbol{e}_m/s_{k,m}, & \text{if } 1 \leqslant m \leqslant p; \\ \boldsymbol{f}_{m-p}/(s_{k,0} - s_{k,(m-p)}), & \text{if } p+1 \leqslant m \leqslant 2p; \end{cases}$$

- Return $\hat{\boldsymbol{\beta}}_k^{(m)}$, $m \in M$.

Theorem 1 offers some insight into the MLE. There are only two possible patterns of the jump $\hat{\boldsymbol{\beta}}(t_k^*)$. Either there is an increment of the coefficient $\hat{\beta}_j(t_k^*)$ for one covariate or there is a decrease of $\hat{\beta}_j(t_k^*)$ for one covariate which is compensated by an increment of the intercept. So, this estimator is sparse in the change of $\boldsymbol{\beta}(t)$.

Theorem 1 also implies a condition for the existence and uniqueness of the likelihood estimator. The likelihood estimator exists if and only if $s_{k,m} \neq 0$ and $s_{k,0} - s_{k,(m-p)} \neq 0$. Equality corresponds to the case where all $x_i$'s are zeros or all $x_i$'s are ones, $i \geq i(k)$. Uniqueness is not guaranteed by Theorem 1. In fact, all choices of the ratios $x_{i(k),j}/s_{k,j}$ and/or $(x_{i(k),0} - x_{i(k),j})/(s_{k,0} - s_{k,j})$ found by Algorithm 1 result in the same maximum. In that case any $\boldsymbol{v} \in \mathcal{A}$ corresponding to such a maximum is allowed, as well as any convex combination, as stated in the following theorem.

**Theorem 2.** *If $\hat{\boldsymbol{\beta}}_k^{(m_1)}$ and $\hat{\boldsymbol{\beta}}_k^{(m_2)}$ are solutions given by Algorithm 1 which maximize the likelihood $\ell$, then any convex combination $\lambda\hat{\boldsymbol{\beta}}_k^{(m_1)} + (1-\lambda)\hat{\boldsymbol{\beta}}_k^{(m_2)}$, with $0 \leq \lambda \leq 1$, attains the same maximized likelihood.*

Theorem 2 shows that the points of maxima form a flat top and are closed under convex combinations. In practice, we suggest an average of all solutions given by Theorem 1. By Theorem 2, this is also a solution which maximizes the likelihood.

With our result on the MLE, we also have the following result which reveals the connection between the MLE and Aalen's estimator.

**Proposition 2.** *For an additive hazards model with one binary covariate, the maximum likelihood estimator and the Aalen estimator coincide.*

This can also be explained intuitively by Theorem 1 as follows. For one binary covariate, Aalen's OLS method can be viewed as selecting one of the two competing trends at each event time. Theorem 1 shows that the nature of MLE is similar; the jump of the coefficient takes place for only one covariate, that is, only one trend is selected at each event time.

## 4.3 | Example

We illustrate Theorem 1 using a simple example with two covariates. Let us assume that at a certain event time point $t^*$ a subject fails with covariate values $x_1 = 0$ and $x_2 = 1$, so we have $\mathbf{x}^* = (0, 1)$, and $\mathbf{x} = (1, 0, 1)$. Suppose that $\mathbf{s} = (8, 5, 6)^\top$. Then, the log-likelihood is given by

$$\ell^* = \log(\beta_0 + \beta_2) - 8\beta_0 - 5\beta_1 - 6\beta_2. \tag{5}$$

In this case, the domain is defined by the matrix $\mathbf{M_D}$ through the boundary conditions $M_{\mathbf{D}}\boldsymbol{\beta} \geq 0$, where

$$M_{\mathbf{D}} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

Theorem 1 directs us to compare the values of $x_1/s_1$, $x_2/s_2$, $(x_0 - x_1)/(s_0 - s_1)$, and $(x_0 - x_2)/(s_0 - s_2)$, namely 0, $\frac{1}{6}$, $\frac{1}{3}$, 0, corresponding to the one-dimensional line edges $\boldsymbol{v} \cdot l$, with $\boldsymbol{v} = (0, 1, 0)^\top, (0, 0, 1)^\top, (1, -1, 0)^\top$ and $(1, 0, -1)^\top$, respectively, in the set of admissible directions $\mathcal{A}$. The third of these numbers is the largest, so the one-dimensional line $\boldsymbol{v} \cdot l$, with $\boldsymbol{v} = (1, -1, 0)^\top$ is the edge where the maximum point lies. Maximizing over $l$ gives $l = \frac{1}{\mathbf{s}^\top \boldsymbol{v}} = \frac{1}{3}$, $\hat{\boldsymbol{\beta}}(t) = (\frac{1}{3}, -\frac{1}{3}, 0)$ and the maximized log-likelihood $\ell^* = \log(\frac{1}{3}) - 1$.

## 4.4 | Maximum likelihood with a partial constraint matrix

In the previous subsection, we showed that the maximum point of the likelihood function lies on one of the $2p$ one-dimensional line edges of the boundary. This result was derived for the case where the domain is given by the matrix $M_{\mathbf{D}}$, whose rows are all the combinations of 1 and 0 in the last $p$ columns. Sometimes we need more flexible domains, for example, when dummy covariates are introduced or when it is known that some coefficients are positive. In this more general case Theorem 1 does not apply, and numerical solutions are needed.

We consider a general domain defined by $M_D\boldsymbol{\beta} \geq 0$ for some matrix $M_D$, where the matrix $M_D$ can be an arbitrary matrix of $p+1$ columns. We assume that the rank of $M_D$ is $p+1$. We assert that Lemma 2 still applies for this constraint matrix $M_D$. The maximum point lies on a one-dimensional edge on the boundary, defined as the intersection of $p$ flat boundaries. So, any local maximum point of a one-dimensional edge is a potential solution of the ML of the domain $D$. Given any one-dimensional edge, we have $p$ equations $\gamma_i = M_i\boldsymbol{\beta} = 0$. Combined with the first-order condition which is also a linear equation, we have $p + 1$ linear equations whose solution is the maximum point on this edge. By testing its derivatives and constraint conditions, we can check if it is a maximum point of the domain $D$. However, it is difficult to find the right edge on which the maximum point lies. We introduce three methods to find the right edge: the naïve method, the ascending method, and the descending method.

The *naïve method* loops through all the possible combinations of $p$ rows of $M_D$. Every combination of $p$ rows defines a one-dimensional edge and gives a possible solution of the maximum point $\boldsymbol{\beta}^*$. We can test its derivatives and constraint conditions to check if it is the maximum point we want to find. This naïve method is easy to understand, but not efficient, since the complexity grows approximately with the $p$th power of number of the rows of $M_D$.

The *ascending method* searches for a sequence of $\boldsymbol{\beta}^*$ with increasing likelihood. More specifically, the algorithm starts from a $p$-combination of rows, whose maximum point is $\boldsymbol{\beta}_0^*$. We assume this edge is inside the domain. If $\boldsymbol{\beta}_0^*$ is not the maximum point of the domain $D$, its gradient points to the inside of the domain. Following this direction, we can find an adjacent one-dimensional edge whose maximum point $\boldsymbol{\beta}_1^*$ has a larger likelihood. Since this algorithm starts from an edge already on the domain, as a result, it is quite efficient when a good starting edge of the domain is known. To find a new edge, we use the gradient of the log-likelihood, constrained to the adjacent hyperplanes. This method works well if the geometry of the domain is not too complicated.

The *descending method* is the reverse of the ascending method. Assuming that we find a maximum point $\boldsymbol{\beta}^*$ on a one-dimensional edge, it is not necessary that this $\boldsymbol{\beta}^*$ satisfies all the constraint conditions $M_D\boldsymbol{\beta}^* \geq 0$. However, it may satisfy some of the constraint conditions, that is, we have $\gamma_i = M_i\boldsymbol{\beta}^* \geq 0$ for some row vector $M_i$. So, for a submatrix $M'_D$ consisting of these rows, we have $M'_D\boldsymbol{\beta} \geq 0$. In many cases, this means we find a maximum point of the domain $D'$ defined by $M'_D\boldsymbol{\beta} \geq 0$. This new domain $D'$ is larger than the domain $D$ and we have $D' \supset D$. So, the ML of domain $D'$ is larger than the ML of domain $D$. Starting from this domain $D'$, we can add the constraints step by step, which gives smaller and smaller domains and finally reaches the domain $D$ that we want. Correspondingly, we have a series of decreasing ML. This gives the name of the descending method. The descending method also requires a starting edge whose maximum point is the maximum point of some larger domain $D' \supset D$. We can apply the naïve method to find such an edge and this is the time consuming part of the algorithm.

Details and pseudo-code of all three methods are given in Web Appendix B.

**TABLE 1** Comparison of the complexity of the maximum likelihood (ML) method and Aalen's ordinary least-squares method; computation time in seconds

| Number of covariates | 2 | 4 | 8 | 12 | 16 |
|---|---|---|---|---|---|
| ML method (ahMLE) | 0.21 | 0.32 | 0.46 | 0.59 | 0.98 |
| ML method (addreg) | 34.5 | 3.2 min | > 1 h | > 1 h | > 1 h |
| Aalen's method (ahMLE) | 0.20 | 0.35 | 0.47 | 0.77 | 1.04 |
| Aalen's method (timereg) | 1.32 | 2.37 | 4.29 | 6.59 | 10.07 |

## 5 | SIMULATION

We have shown the theoretical background and algorithms in Section 4. The aim of the following simulation study is to show the feasibility of our MLE and compare its performance with the OLS estimator.

We assume $\boldsymbol{\beta}(t) = \boldsymbol{\beta}t$, a linear function of $t$ with $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$ constant. Hence, the true hazard has $p$ covariates and it is given by $h(t \mid \mathbf{x}) = \sum_j \beta_j x_j t$. This corresponds to a Weibull distribution with shape $b = 2$ and rate $a = \sum \beta_j x_j / 2$, using the parameterization $h(t; a, b) = abt^{b-1}$ for the Weibull hazard. We randomly generate covariate data $\mathbf{x}_i$, for $i = 1, \dots, n$, with independent uniform distributions on $[0, 1]$. By the given parameters and randomly generated covariates, we generate the survival times as simulated data. We use both the ML method and Aalen's OLS method to fit the additive hazards model. We use the MLE with full constraint matrix introduced in Section 4, as implemented in the R package ahMLE (2022). For assessing computing time, we also used the R package addreg (2017), which is not specifically designed for additive hazards but more generally to Poisson generalized linear models with identity link. The addreg package results always converge to a solution that is close to the package ahMLE results but not exactly. We verified that the addreg solution never gave a higher likelihood than our new algorithm. For Aalen's method, we use the R packages ahMLE and also timereg (2022) to assess computing time. These two packages also gave identical results.

We first assessed the computation time of the new ML method on the simulated data. Theorem 1 implies that the complexity of the algorithm to obtain the MLE is linear in the number of covariates $p$. Hence, the method should be faster than Aalen's method whose complexity is quadratic in $p$. We generated simulated data as above with sample size $n = 500$. Table 1 shows the computation time of the two methods with a varying number $p$ of covariates. It is seen that as $p$ increases, the ML method runs faster than Aalen's OLS method, although timings are comparable overall. Implementation of the MLE is faster in ahMLE, compared to addreg, and implementation of the OLS

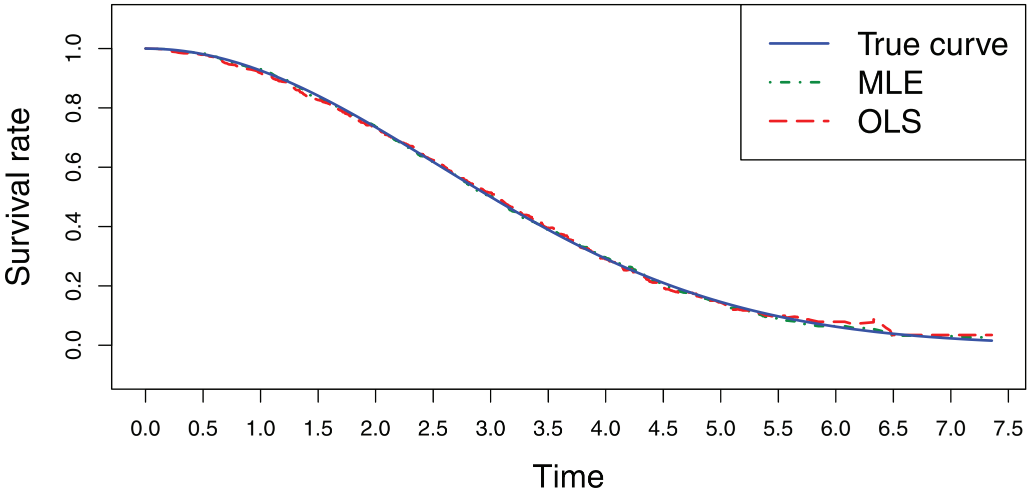


FIGURE 1 Estimated survival curves by ordinary least squares (OLS) and by maximum likelihood (MLE). This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

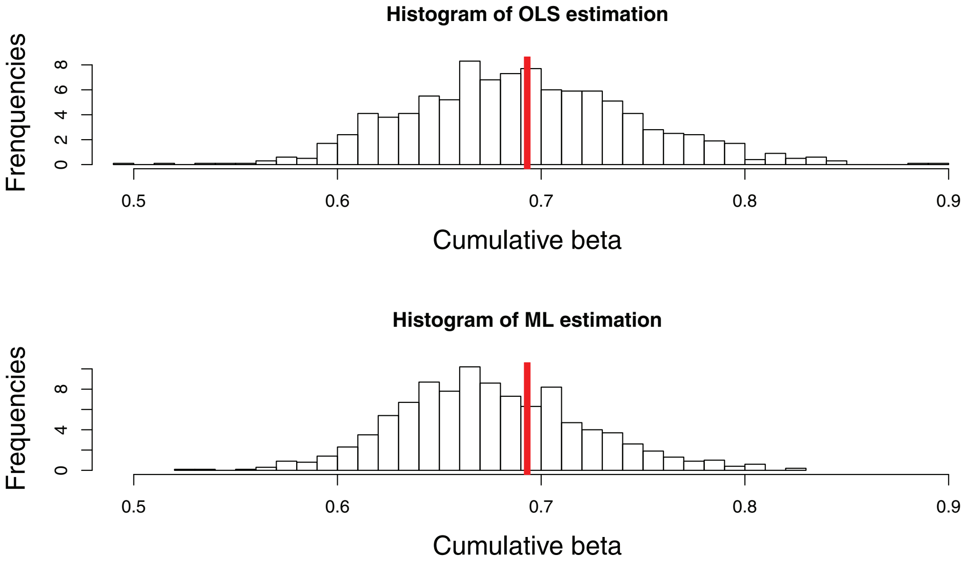


FIGURE 2 Histograms of estimated cumulative hazards at time $t = 3$ (median) by ordinary least squares (OLS) and maximum likelihood (ML). This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

estimator is faster in ahMLE, compared to timereg, although it should be acknowledged that timereg provides computes additional output.

Our second aim is to compare the performance of the two estimators. We take $p = 4$ and $\boldsymbol{\beta}(t) = (0.05, 0.02, 0.04, 0.06, 0.08)t$, with sample size $n = 250, 500, 1000$. For each subject an independent censoring time $C_i$ is generated with uniform distribution over (2.5,7.5), giving a censoring rate of about 22% in the simulation data.

We choose a subject with data of values $x = (0.4, 0.6, 0.4, 0.6)$. Figure 1 shows the true survival curve and a single realization of the estimated survival curves by the two methods. The survival curves given by both estimators are similar and approximate the true survival curve well.

We repeat the same simulation set-up for 1000 replications with the same subject. We check the estimated cumulative hazards at the three quartiles and the 90% percentile of the true time to event distribution for that subject, at $t = 1.93$, 3.00, 4.24 and 5.47, respectively. Figure 2 shows histograms of the estimated cumulative hazards for the two methods at the median. The true value of the cumulative hazard is 0.693, shown by the red vertical line. The

**TABLE 2** Bias (mean estimate minus true value), empirical SE (the standard deviation of the estimates around their mean), and RMSE (root mean squared error, square root of the mean-squared difference between estimate and truth) of cumulative hazards estimators by OLS, OLSR, and ML

| True survival probability | Time | True cumulative hazard | Method | Estimated cumulative hazard | Bias | Empirical SE | RMSE |
|---|---|---|---|---|---|---|---|
| $n = 250$ | | | | | | | |
| 0.75 | 1.93 | 0.288 | OLS | 0.290 | 0.002 | 0.045 | 0.045 |
| | | | OLSR | 0.416 | 0.128 | 0.05 | 0.14 |
| | | | MLE | 0.285 | −0.003 | 0.038 | 0.038 |
| 0.5 | 3.00 | 0.693 | OLS | 0.698 | 0.004 | 0.079 | 0.079 |
| | | | OLSR | 1.00 | 0.31 | 0.09 | 0.327 |
| | | | MLE | 0.684 | −0.009 | 0.068 | 0.068 |
| 0.25 | 4.24 | 1.386 | OLS | 1.386 | −0.001 | 0.150 | 0.150 |
| | | | OLSR | 1.99 | 0.61 | 0.163 | 0.635 |
| | | | MLE | 1.363 | −0.023 | 0.123 | 0.125 |
| 0.10 | 5.47 | 2.304 | OLS | 2.300 | −0.003 | 0.385 | 0.385 |
| | | | OLSR | 3.23 | 0.93 | 0.41 | 1.02 |
| | | | MLE | 2.285 | −0.018 | 0.285 | 0.286 |
| $n = 1000$ | | | | | | | |
| 0.75 | 1.93 | 0.288 | OLS | 0.286 | −0.0001 | 0.022 | 0.022 |
| | | | OLSR | 0.417 | 0.129 | 0.028 | 0.133 |
| | | | MLE | 0.282 | −0.006 | 0.018 | 0.019 |
| 0.5 | 3.00 | 0.693 | OLS | 0.694 | 0.0007 | 0.038 | 0.038 |
| | | | OLSR | 0.995 | 0.302 | 0.046 | 0.307 |
| | | | MLE | 0.677 | −0.016 | 0.032 | 0.036 |
| 0.25 | 4.24 | 1.386 | OLS | 1.382 | −0.005 | 0.076 | 0.076 |
| | | | OLSR | 1.97 | 0.588 | 0.093 | 0.598 |
| | | | MLE | 1.347 | −0.039 | 0.064 | 0.075 |
| 0.10 | 5.47 | 2.304 | OLS | 2.295 | −0.008 | 0.160 | 0.160 |
| | | | OLSR | 3.263 | 0.961 | 0.179 | 0.98 |
| | | | MLE | 2.251 | −0.052 | 0.133 | 0.143 |
| $n = 5000$ | | | | | | | |
| 0.75 | 1.93 | 0.288 | OLS | 0.288 | −0.00016 | 0.01 | 0.01 |
| | | | OLSR | 0.414 | 0.126 | 0.014 | 0.128 |
| | | | MLE | 0.282 | −0.006 | 0.009 | 0.011 |
| 0.5 | 3.00 | 0.693 | OLS | 0.694 | 0.0005 | 0.018 | 0.018 |
| | | | OLSR | 0.99 | 0.30 | 0.023 | 0.30 |
| | | | MLE | 0.675 | −0.018 | 0.016 | 0.025 |
| 0.25 | 4.24 | 1.386 | OLS | 1.387 | −0.0002 | 0.033 | 0.033 |
| | | | OLSR | 1.972 | 0.586 | 0.044 | 0.59 |
| | | | MLE | 1.347 | −0.039 | 0.031 | 0.050 |
| 0.10 | 5.47 | 2.303 | OLS | 2.300 | −0.003 | 0.070 | 0.070 |
| | | | OLSR | 3.265 | 0.962 | 0.090 | 0.972 |
| | | | MLE | 2.240 | −0.06 | 0.059 | 0.086 |

Abbreviations: MLE, maximum likelihood estimator; OLS, ordinary least squares; OLSR.

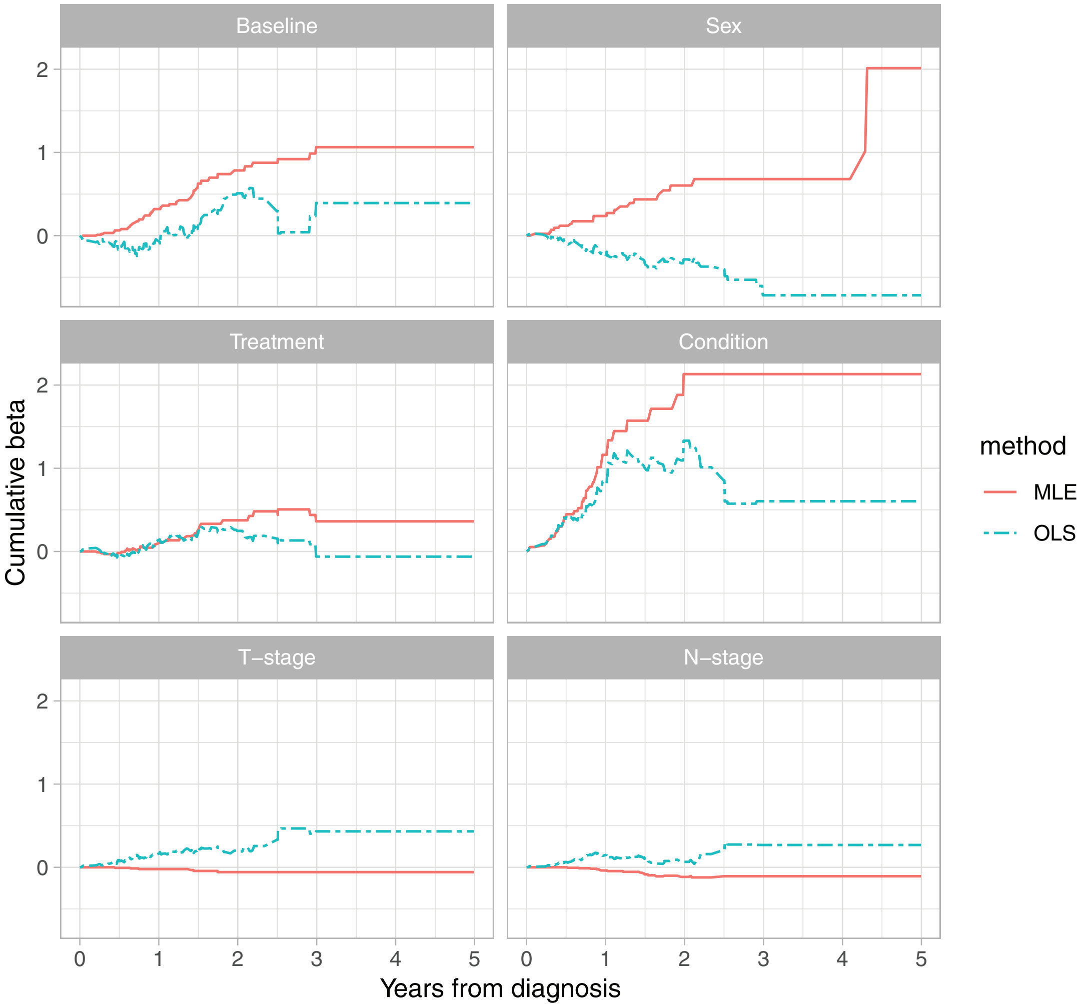


FIGURE 3 Estimated cumulative beta's of factors by maximum likelihood (MLE) and ordinary least squares (OLS). This figure appears in color in the electronic version of this paper, and any mention of color refers to that version.

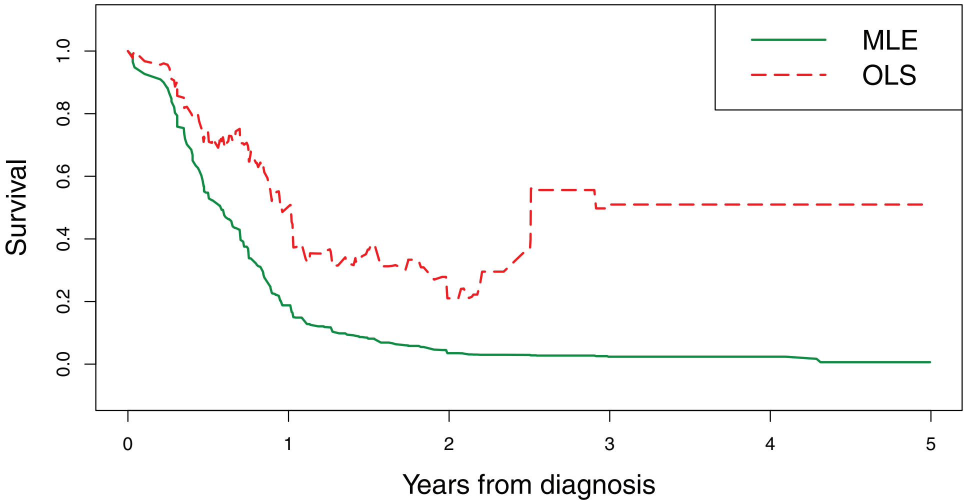


FIGURE 4 Estimated survival curves by the two methods maximum likelihood (MLE) and ordinary least squares (OLS). This figure appears in color in the electronic version of this paper, and any mention of color refers to that version

OLS estimate is quite symmetric around the true value, while the MLE is slightly skewed to the left. However, the variability of the ML estimation is smaller.

This behavior is also seen in Table 2, which shows that for all four time points and all sample sizes, the MLE has more bias but smaller variance. The bias in the ML estimate does not seem to vanish with increasing sample size, eventually dominating the mean-squared error. We believe that this persistence of the bias arises as a consequence of the constraints. To check this, we have added a third method to the simulation, which imposes the constraints of the ML method to Aalen's estimation procedure. We use the R package quadprog (2019) to implement this method, which we call Ordinary Least Square methods Restricted (OLSR). The results from OLSR show much more persistent bias than the MLE.

## 6 | APPLICATION

We apply our methods to data from a clinical trial with 195 patients with carcinoma of the oropharynx by the Radiation Therapy Oncology Group in the United States. The data are introduced by Kalbfleisch and Prentice (2002, Section 1.1.2, Web Appendix A) and used for illustration of additive hazards models in Aalen et al. (2008). Patients were randomized into two treatment groups ("standard" and "experimental" treatment), and survival times were measured in days from diagnosis. Seven covariates were included in the data: sex, treatment, grade, age, condition, T-stage, and N-stage. Following  Aalen et al. (2008), we take all the covariates as continuous. We rescale them to [0,1].

The cumulative beta's of the baseline and covariates sex, treatment group, condition, T-stage and N-stage are shown in Figure 3. The OLS baseline hazard estimate decreases over the first 9 months and between 2 and 2.5 years, which would correspond to a negative hazard for subjects with all covariates 0, and an increasing survival curve. The MLE of the baseline hazard is monotone increasing over time, as desired. For treatment effect, condition, T-stage and N-stage factors the two estimated curves of the cumulative beta are very close to each other in these cases. For sex, the two methods show opposite trends. The differences in estimates between the two methods may partly be explained by the large number of covariates in the model, in relation to the modest sample size, leading to highly variable estimates for both methods.

To illustrate the difference in the behavior of the survival curves between ML and OLS, we choose a subject who is female of age 55 years in the treatment group, of cancer grade moderate differentiated, of condition restricted worked, of second T-stage and first N-stage. The covariates of the subject are given by $\mathbf{x} = (1, 1, 0, 1, -0.5, 1, 1, 1)$. The estimated survival curves according to the OLS and ML models are shown in Figure 4. The two estimators give similar results in the first half year when the risk set is large. After that, the OLS survival curve starts to fluctuate wildly, while the ML estimate shows a steady decrease.

## 7 | DISCUSSION

In this paper, we have presented an ML approach to the additive hazards model as an alternative to the OLS approach. We constructed an MLE under the natural constraint that the hazard is non-negative at each time point for each possible combination of covariate values. The MLE is characterized by increments of the cumulative hazards at the observed event time points, as the OLS. We derived an explicit expression of the MLE. Its computational complexity is linear in the number of covariates for each time point, while that of the OLS is quadratic. A simulation study showed that the MLE sacrifices a small amount of bias in favor of smaller variability, leading to a smaller mean-squared error. This bias seems to persists for large sample size. Still, we found that the bias was much smaller than that of a version of Aalen's estimator constrained to have positive increments. We showed in the Supporting Information that the additive hazards model is a special case of the Poisson identity link regression model, parallel to the well-established connection between the Cox proportional hazards model and the log link Poisson model.

We have not developed any theoretical results on the asymptotic properties of the MLE, although these are important to derive. Obtaining such asymptotic results is challenging, because maximum likelihood estimation under constraints represents a non-standard setting for which asymptotic theory (typically derived under strict regularity conditions) is not readily available. The fact that the number of constraints grows with $n$ adds to the complexity of this problem. It would be valuable to develop asymptotic theory and investigate efficiency of the MLE. Other issues of interest to explore are cross-validation and penalized likelihood.

## ACKNOWLEDGMENTS

We would like to acknowledge support by the Leiden University Fund/Hakkenberg $\alpha\,\beta\,\gamma$-integratie Fonds, www.luf.nl.

## OPEN RESEARCH BADGES

This article has earned Open Data and Open Materials badges. Data and materials are available at: https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Fbiom.13764&file=biom13764-sup-0002-SuppMat.zip.

## ORCID

*Chengyuan Lu* https://orcid.org/0000-0002-8486-3609
*Hein Putter* https://orcid.org/0000-0001-5395-1422

## REFERENCES

Aalen, O. (1980) A model for nonparametric regression analysis of counting processes. In: Klonecki W., Kozek A., & Rosiński J. (Eds.). *Mathematical statistics and probability theory. Lecture Notes in Statistics*, vol. 2 (pp. 1–25). Springer, New York.

Aalen, O. (1989) A linear regression model for the analysis of life times. *Statistics in Medicine*, 8, 907–925.

Aalen, O., Borgan, O. & Gjessing, H. (2008) *Survival and event history analysis: a process point of view*. Springer, New York.

Aalen, O.O., Cook, R.J. & Røysland, K. (2015) Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Analysis*, 21(4), 579–593.

addreg. (2017) Methods for fitting identity-link GLMS and GAMS to discrete data, using em-type algorithms with more stable convergence properties than standard methods. Available from: https://cran.r-project.org/web/packages/addreg/ [Accessed 30th July 2022].

ahMLE. (2022) Methods for the additive hazard model. Available from: https://cran.r-project.org/web/packages/ahMLE/ [Accessed 30th July 2022].

Bretagnolle, J. & Huber-Carol, C. (1988) Effects of omitting covariates in Cox's model for survival data. *Scandinavian Journal of Statistics*, 15, 125–138.

Cox, D.R. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187–202.

Gore, S.M., Pocock, S.J. & Kerr, G.R. (1984) Regression models and non-proportional hazards in the analysis of breast cancer survival. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 33(2), 176–195.

Hernán, M.A. (2010) The hazards of hazard ratios. *Epidemiology*, 21(1), 13–15.

Holford, T.R. (1980) The analysis of rates and of survivorship using log-linear models. *Biometrics*, 36, 299–305.

invGauss. (2022) Threshold regression that fits the (randomized drift) inverse Gaussian distribution to survival data. Available from: https://cran.r-project.org/web/packages/invGauss/ [Accessed 30th July 2022].

Kalbfleisch, J.D. & Prentice, R.L. (2002) *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Laird, N. & Olivier, D. (1981) Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76, 231–240.

Marschner, I.C. (2010) Stable computation of maximum likelihood estimates in identity link Poisson regression. *Journal of Computational and Graphical Statistics*, 19(3), 666–683.

Marschner, I.C., Gillett, A.C. & O'Connell, R.L. (2012) Stratified additive Poisson models: Computational methods and applications in clinical epidemiology. *Computational Statistics & Data Analysis*, 56(5), 1115–1130.

Martinussen, T. & Scheike, T.H. (2006) *Dynamic regression models for survival data*. Springer, New York.

Martinussen, T., Scheike, T.H., et al. (2000) A nonparametric dynamic additive regression model for longitudinal data. *The Annals of Statistics*, 28(4), 1000–1025.

Martinussen, T., Vansteelandt, S. & Andersen, P.K. (2020) Subtleties in the interpretation of hazard contrasts. *Lifetime Data Analysis*, 26(4), 833–855.

McCullagh, P. & Nelder, J. (1989) Generalized linear models, Second Edition. *Monographs on statistics and applied probability series*. Chapman & Hall.

Perperoglou, A., Le Cessie, S. & van Houwelingen, H.C. (2006) Reduced-rank hazard regression for modelling non-proportional hazards. *Statistics in Medicine*, 25(16), 2831–2845.

quadprog (2019) Functions to solve quadratic programming problems. Available from: https://cran.r-project.org/web/packages/quadprog/ [Accessed 30th July 2022].

Schemper, M. (1992) Cox analysis of survival data with non-proportional hazard functions. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 41(4), 455–465.

Schumacher, M., Olschewski, M. & Schmoor, C. (1987) The impact of heterogeneity on the comparison of survival times. *Statistics in Medicine*, 6, 773–784.

Struthers, C.A. & Kalbfleisch, J.D. (1986) Misspecified proportional hazards models. *Biometrika*, 73, 363–369.

timereg (2022) Flexible regression models for survival data. Available from: https://cran.r-project.org/web/packages/timereg/. [Accessed 30th July 2022].

van Houwelingen, H. & Putter, H. (2012) *Dynamic prediction in clinical survival analysis*. CRC Press, Boca Raton.

## SUPPORTING INFORMATION

Web Appendix A referenced in Section 4, Web Appendix B referenced in Section 4.4, and Web Appendix C, referenced in Section 4.1 are available with this paper at the Biometrics Website on Wiley Online Library. The R codes for the results of figures and tables are available at the Biometrics Website on Wiley Online Library.

Data S1

**How to cite this article:** Lu, C., Goeman, J., & Putter, H. (2022) Maximum likelihood estimation in the additive hazards model. *Biometrics*, 1–11. https://doi.org/10.1111/biom.13764