ON AN EXTENSION OF THE PROMOTION TIME CURE MODEL

BY JAD BEYHUM† ANOUAR EL GHOUCH † FRANÇOIS PORTIER *
AND INGRID VAN KEILEGOM † †

Télécom Paris *, KU Leuven † and UCLouvain † †

We consider the problem of estimating the distribution of time-to-event data that are subject to censoring and for which the event of interest might never occur, i.e., some subjects are cured. To model this kind of data in the presence of covariates, one of the leading semiparametric models is the promotion time cure model [43], which adapts the Cox model to the presence of cured subjects. Estimating the conditional distribution results in a complicated constrained optimization problem, and inference is difficult as no closed-formula for the variance is available.

We propose a new model, inspired by the Cox model, that leads to a simple estimation procedure and that presents a closed formula for the variance. In this paper we show (i) that the new model contains as a special case the promotion time cure model with an exponential link, and hence we have a simpler way to estimate the latter model than what is done so far in the literature; (ii) that in the latter special case, both estimators are equal to the partial likelihood estimator under the usual Cox model; (iii) that the estimators under the new model have certain asymptotic properties when the model is correct and when it is misspecified; (iv) that the error of LASSO type estimators is of order \( \sqrt{\log(nd)/n} \) in the case of high-dimensional covariates with dimension \( d \) and sample size \( n \). We also study the practical behaviour of our estimation procedure by means of simulations, and we apply our model and estimation method to a breast cancer data set.

1. Introduction. Survival analysis has been the subject of many statistical studies in the past decades (see e.g. [15, 32]) and is commonly used in clinical trials (see e.g. [9]), where the traditional main goal is to explain the death of patients having a certain disease. When analysing the effect of some covariates \( X \in \mathbb{R}^d \) on a survival time \( T \in \mathbb{R}_{\geq 0} \), a common approach in the literature is based on semiparametric estimation. The seminal paper by [10] introduces the so-called semiparametric proportional hazards model, often referred to as the Cox model, which is given by the following set of conditional survival functions defined on \( \mathbb{R}_{\geq 0} \times \mathbb{R}^d : \)

\[ P_1 = \left\{ (t, x) \mapsto S(t|x) = \exp(-\exp(\gamma^T x)\Lambda(t)) : \gamma \in \mathbb{R}^d, \Lambda \in \mathcal{G} \right\}, \]

where \( \mathcal{G} \) is the space of absolutely continuous cumulative hazard functions defined on \( \mathbb{R}_{\geq 0} \). In this standard semiparametric model the elements are characterized by the Euclidean parameter \( \gamma \), called the regression vector, and the infinite dimensional parameter \( \Lambda \), called the cumulative hazard. These parameters are estimated by maximizing the profile likelihood for \( \gamma \) [10] and by computing the Breslow estimator for \( \Lambda \) [6]. Both estimators are nonparametric maximum likelihood estimators (NPMLE), as defined in [22]. Known asymptotic results include the asymptotic normality [1], the semiparametric efficiency of the regression parameters [28] as well as the cumulative hazard [4, 17], and the validity of general bootstrap schemes [41].

In many data sets, especially the ones arising from clinical trials, a certain proportion of the individuals will never experience the event of interest. These individuals are referred to as the cured subjects. As the survival function \( t \mapsto S(t) \) does not tend to 0 as \( t \to +\infty \) in that case (but rather
tends to the proportion of cured subjects), specific models need to be considered to account for the 
*improperness* of the distribution of $T$. The *promotion time cure model* is an extension of the Cox model specially designed to handle the presence of cured subjects in the data. It is defined as the set of conditional survival functions defined on $\mathbb{R}_{\geq 0} \times \mathbb{R}^d$, given by

$$
P_2 = \left\{ (t, x) \mapsto S(t|x) = \exp \left( -\eta(\beta_1 + \beta_2^T x)G(t) \right) \right. : \left. \beta = (\beta_1, \beta_2) \in \mathbb{R}^{d+1}, \ G \in \mathcal{F} \right\},
$$

where $\mathcal{F}$ denotes the space of absolutely continuous cumulative distribution functions on $\mathbb{R}_{\geq 0}$ and $\eta : \mathbb{R} \to \mathbb{R}_{>0}$ is a given function. This model was introduced by [43] and seems appropriate to treat cure data as, for every $x \in \mathbb{R}^d$, $\lim_{t \to +\infty} S(t|x) > 0$, so that each subject has a positive chance of being cured. In model $P_2$, the parameter vector $\beta$ has an intercept whereas $\gamma$ in model $P_1$ does not. This is because $\lim_{t \to +\infty} \Lambda(t) = +\infty$ and hence an intercept in model $P_1$ would not be identified, whereas in model $P_2$ the function $\Lambda(t)$ is replaced by $G(t)$, which tends to 1 as $t \to +\infty$. Estimation of $P_2$ has been studied by [34, 35, 36, 7, 14, 37, 45, 26], among many others. Certain parallels might be drawn between the statistical properties related to the estimators of the classical Cox model and the ones related to the promotion time cure model. The classical estimators of $\beta$ and $G$ are the NPMLE’s [45]. In [45], the authors show that the resulting NPMLE is asymptotically normal and moreover that the estimated vector of regression parameters is semiparametrically efficient. In [26] it is shown that the whole model is estimated efficiently and the validity of a general weighted bootstrap is proved.

There is still an important difference between the NPMLE’s associated to models $P_1$ and $P_2$. The NPMLE of the Cox model has a much simpler expression than the NPMLE of the promotion time cure model. Within model $P_1$, the estimated regression parameter maximizes a known (explicit) objective function and the estimated cumulative hazard is expressed through a closed formula [1]. Within Model $P_2$, the estimated regression parameter is also the maximizer of a certain objective function, but this time the objective function is implicitly defined [26]. Moreover, the same is true for the estimated cumulative hazard in $P_2$, which is only known up to some quantity implicitly defined. The previous features involve important complications that intervene at two different stages. First, estimators from $P_2$ are more difficult to describe, theoretically, than estimators from $P_1$. This eventually deteriorates the accuracy of the confidence intervals or of the testing procedures. Second, the computation of the estimators in $P_2$ has some numerical difficulties, e.g., long computation time, problems with local minima, etc. Given this, the question is to know, whether or not, it is legitimate to rely on a complicated estimation procedure for $P_2$? In other words, does the presence of cured subjects in the data prevents us from having an estimation procedure as simple as in the Cox model?

The aim of this paper is to provide a new model dedicated to cure data analysis and for which the NPMLE overpasses the previous difficulties associated with $P_2$.

The undesirable complications when estimating $P_2$ come from the particular nature of the parameter space $\mathcal{F}$. This space is formed by cumulative distribution functions $G$ that satisfy the constraint $\lim_{t \to +\infty} G(t) = 1$. Such a constraint is taken into account with the help of a Lagrange procedure involving an additional parameter being implicitly defined, the *Lagrange multiplier*. It turns out that this constraint can be alleviated by including an additional parameter in the model, replacing $G$ by $\theta F$, with $\theta > 0$ and $F$ a cumulative distribution function. We define the set of conditional survival functions $\mathbb{R}_{\geq 0} \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, given by

$$
P_3 = \left\{ (t, x) \mapsto S(t|x) = \exp(-g(\gamma, x)\theta F(t)) : (\gamma, \theta) \in B \times \mathbb{R}_{>0}, \ F \in \mathcal{F} \right\},
$$

where $B \subseteq \mathbb{R}^q$, $g : \mathbb{R}^q \times \mathbb{R}^d \to \mathbb{R}_{>0}$ is a given function and $q \in \mathbb{N}$. Note that in the present form, $P_3$ handles biological models as developed in [7] to analyse time to relapse of cancer through
the distribution of the carcinogenic cells. It includes also a cure version of the Cox model when \( g(\gamma, x) = \exp(\gamma^T x) \). In this case, it coincides with \( P_2 \) for which \( \eta = \exp \). Otherwise \( P_2 \) and \( P_3 \) are different. In \( P_3 \), the role of \( \theta \) is interpreted as a simple multiplicative effect on the cumulative distribution, whereas the effect of \( \beta_1 \) in \( P_2 \) must be analysed depending on the shape of the function \( \eta \).

The main contributions of the paper are listed below.

(i) As the NPMLE of \( P_3 \) is much simpler to evaluate than the one associated to \( P_2 \), the proposed methodology provides a significant improvement in terms of computational ease. In particular, we show that the NPMLE’s associated with \( P_2 \) and \( P_3 \) coincide when \( \eta = \exp \) and \( g(\gamma, x) = \exp(\gamma^T x) \). Hence our approach provides a new way to compute the NPMLE of \( P_2 \) when \( \eta = \exp \) (most commonly used) which is simpler than the existing procedure [45, 26].

(ii) We show that when the link function equals \( \eta = \exp \) and when \( g(\gamma, x) = \exp(\gamma^T x) \), the NPMLE under model \( P_3 \), or equivalently the NPMLE under model \( P_2 \), coincides with the partial likelihood estimator under the usual Cox model, which ignores the presence of a cure fraction. Hence, the NPMLE under the promotion time cure model and its asymptotic variance can be computed using software for the Cox model. This leads to substantially faster and more accurate estimation of the NPMLE, its variance and corresponding confidence intervals.

(iii) We derive the asymptotics of the NPMLE associated with \( P_3 \), both when the model is correct and when it is misspecified. As in the case of the Cox model, we have closed-formulas for the variance of the limiting Gaussian distributions. This allows us to develop some tests and to build confidence intervals on some quantities of interest as for instance the proportion of cure given the value of a covariate \( x \). The finite sample size accuracy of the confidence intervals is investigated with the help of simulations.

(iv) We consider the high-dimensional setup in which the number of covariates can be larger than the sample size, and define a LASSO type estimator of \( \gamma_0 \) in the case where \( g(\gamma, x) = \exp(\gamma^T x) \). We establish an upper bound on the estimation error associated to the parameter \( \gamma_0 \) under this setup and derive the rate of convergence of an estimator of the cure rate.

In Section 2 we present the framework of the paper and derive the NPMLE of model \( P_3 \). Section 3 considers the links with the NPMLE of \( P_2 \) and with the partial likelihood estimator under the Cox model. In Section 4, the asymptotic behaviour of the NPMLE of \( P_3 \) is studied when the model is correctly specified, while Section 5 focuses on the asymptotics when the model is misspecified. Section 6 focuses on the behavior of a LASSO type estimator of the cure rate in the special case where we have high-dimensional covariates. In Sections 7 and 8, we provide simulations and a real data analysis to give some insights in the finite sample performance of our approach. Finally, some conclusions and discussion is given in Section 9, whereas the proofs are collected in the supplementary material ([3]).

2. The data, the model, the estimator.

2.1. Framework. We focus on the standard right censoring context: the lifetime \( T \) of interest is right censored by some random variable \( C \) so that we only observe \( Y = \min(T, C), \delta = 1\{T \leq C\} \) and the vector of covariates \( X \). This means that we know whether the variable of interest \( T \) has been observed or censored. The covariates \( X \) are in contrast always observed, and we further denote by \( S \subseteq \mathbb{R}^d \) their support. We suppose conditional independence between \( T \) and \( C \), given \( X \). In practice, as is the case for instance in clinical trials, \( C \) might be bounded. This prevents us from observing any cured subjects, defined by \( T = +\infty \). A way around this problem is to assume the
existence of a threshold \( \tau \in \mathbb{R} \) such that

\[ \{ T > \tau \} \Rightarrow \{ T = +\infty \}. \]

Therefore whenever \( Y \) will be observed to be greater than \( \tau \), the individual will be known to be cured. Let \( P \) denote the probability measure associated to \((Y, C, X)\).

2.2. Nonparametric maximum likelihood. Let \((T_i, C_i, X_i)_{i \in \mathbb{N}}\) denote a sequence of independent and identically distributed random variables with law \( P \), as described in the previous subsection. The underlying probability measure is denoted by \( \mathbb{P} \). The estimator we consider shall be based on the observed variables: \( Y_i = \min(T_i, C_i) \), \( \delta_i = 1\{T_i \leq C_i\} \), \( X_i \), \( i = 1, \ldots, n \). Let \( R_i(y) = \Delta_i 1\{Y_i \geq y\} + (1 - \Delta_i) \), with \( \Delta_i = 1\{Y_i \leq \tau\} \), be the at risk process for individual \( i \).

Under the current data generating process, assuming that \( F \) is absolutely continuous, and assuming non-informative censoring [29], the likelihood of one single observation \((y, \delta, x)\) in model \( \mathcal{P}_3 \) is given by

\[
\text{Lik}(y, \delta, x) = \{g(\gamma, x)\theta f(y)\}^\delta \exp \left[ -g(\gamma, x)\theta \Delta F(y) + (1 - \Delta) \right], \tag{2.1}
\]

where \( f \) stands for the derivative of \( F \). Model \( \mathcal{P}_3 \) can be re-written as the set of all survival functions of the form \( \exp(-g(\gamma, x)\Lambda(t)) \) where \( \gamma \in \mathbb{R}^d \) and \( \Lambda \) belongs to \( \mathcal{G} \), the space of absolutely continuous cumulative hazards \( \Lambda \) such that \( \Lambda(\tau) = \lim_{y \to +\infty} \Lambda(y) = \theta \). Note that there is a one-to-one relationship between the two sets of parameters \((\theta, F)\) and \( \Lambda \), i.e., \( \Lambda = \theta F \) and \( \theta = \lim_{t \to +\infty} \Lambda(t) \).

As a consequence the likelihood in (2.1) can be expressed in terms of \((\gamma, \theta, F)\) or equivalently, in terms of \((\gamma, \Lambda)\). Switching from one parametrization to another is straightforward. For the sake of simplicity, we derive the NPMLE with respect to \((\gamma, \Lambda)\) in the next few lines. By following [22], the NPMLE is defined as

\[
(\hat{\gamma}, \hat{\Lambda}) = \arg\max_{\gamma \in B, \Lambda} \sum_{i=1}^{n} \left[ \delta_i \log(g(\gamma, X_i)\Lambda(Y_i)) - g(\gamma, X_i)\Delta_i \Lambda(Y_i) + (1 - \Delta_i) \Lambda(+\infty) \right], \tag{2.2}
\]

the maximum is taken over \( \Lambda \) lying in the space of cumulative hazard functions possibly discrete, and \( \Lambda(Y) = \Lambda(y) - \lim_{t \to y^-} \Lambda(t) \) is the size of the jump of \( \Lambda \) at \( y \). As is common practice for computing the NPMLE in semiparametric models, the above NPMLE might be profiled over the nuisance parameter \( \Lambda \) [23, 17]. Maximizing along submodels \( d\Lambda_s = (1 + sh)d\Lambda \), \( s \in \mathbb{R} \), with \( h \) a bounded real function, the value of \( \Lambda \) which maximizes (2.2), for each \( \gamma \in \mathbb{R}^d \), is a solution of

\[
n^{-1} \sum_{i=1}^{n} \delta_i h(Y_i) - \int \hat{Q}_\gamma(u)h(u)d\Lambda(u) = 0,
\]

with \( \hat{Q}_\gamma(u) = n^{-1} \sum_{i=1}^{n} g(\gamma, X_i)R_i(u) \). The solution of the previous equation is given by

\[
\hat{\Lambda}_\gamma(y) = n^{-1} \sum_{i=1}^{n} \frac{\delta_i 1\{Y_i \leq y\}}{\hat{Q}_\gamma(Y_i)}, \quad y \in \mathbb{R}_{\geq 0}.
\]

This is then plugged into (2.2) to get that

\[
\left\{ \begin{array}{l}
\hat{\gamma} \in \arg\max_{\gamma \in B} \prod_{i=1}^{n} \left\{ g(\gamma, X_i)/\hat{Q}_\gamma(Y_i) \right\}^{\delta_i} \\
\hat{\Lambda}(y) = n^{-1} \sum_{i=1}^{n} \hat{Q}_\gamma(Y_i)^{-1} \delta_i 1\{Y_i \leq y\}, \quad y \in \mathbb{R}_{\geq 0}.
\end{array} \right. \tag{2.3}
\]
Back to the parameters \((\theta, F)\) of Model \(\mathcal{P}_3\), the NPMLE is given by

\[
\begin{align*}
\hat{\theta} &= n^{-1} \sum_{i=1}^{n} \hat{Q}_\gamma(Y_i)^{-1} \delta_i \\
\hat{F}(y) &= (\hat{\theta} n)^{-1} \sum_{i=1}^{n} \hat{Q}_\gamma(Y_i)^{-1} \delta_i 1_{\{Y_i \leq y\}}, \quad y \in \mathbb{R}_{\geq 0}.
\end{align*}
\]  

(2.4)

At fixed sample size \(n\), the quantities involved in the previous equations are well defined as soon as, for instance, there exists \(i\) such that \(\delta_i = 1\) and the function \(\gamma \mapsto g(\gamma, x)\) is continuous on \(B\), for every \(x \in S\).

Note also that the estimation of the parameters depends only on the observed variables \((Y_i, \delta_i, X_i)\) such that \(Y_i \leq \tau\), and \((\Delta_i, X_i)\) such that \(Y_i > \tau\), \(i = 1, \ldots, n\). It results that moving the threshold over \([Y_{(n, \delta)}, +\infty)\), with \(Y_{(n, \delta)} = \max_{i=1, \ldots, n} Y_i \delta_i\) has no effect on the NPMLE. In practice the threshold could then be fixed at \(Y_{(n, \delta)}\).

An important point in many situations is to evaluate the proportion of cured subjects in the population under study for a given covariate vector \(x \in S\), i.e., \(p_0(x) = \exp(-g(\gamma_0, x) \theta_0)\). The estimator of \(p(x)\), within our framework, naturally follows from the plug-in rule:

\[
\hat{p}(x) = \exp(-g(\hat{\gamma}, x) \hat{\theta}).
\]  

(2.5)

3. Links between models \(\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3\), and their corresponding estimators. The aim of this section is twofold. We will first study the link between models \(\mathcal{P}_2\) and \(\mathcal{P}_3\) and between their corresponding estimators, and we will next focus on the link between the estimators under model \(\mathcal{P}_3\) and under the classical Cox model, which ignores the presence of a cure fraction.

3.1. Promotion time cure estimator. The NPMLE for \(\mathcal{P}_2\) is given by [26]:

\[
\begin{align*}
\hat{\beta} &\in \arg\max_{(\beta_1, \beta_2) \in \mathbb{R}^d} \prod_{i=1}^{n} \left\{ \left( \frac{\eta(\beta_1 + \beta_2^T X_i)}{Q_{2, \beta}(Y_i) - \hat{\lambda}_\beta} \right)^{\delta_i} \exp(-\hat{\lambda}_\beta) \right\} \\
\hat{G}(y) &= n^{-1} \sum_{i=1}^{n} \frac{\delta_i 1_{\{Y_i \leq y\}}}{Q_{2, \hat{\lambda}_\beta}(Y_i) - \hat{\lambda}_\beta},
\end{align*}
\]  

(3.1)

where \(Q_{2, \beta}(u) = n^{-1} \sum_{i=1}^{n} \eta(\beta_1 + \beta_2^T X_i) R_i(u)\) and for every \(\beta \in \mathbb{R}^d\), \(\hat{\lambda}_\beta\) is the smallest number verifying \(\sum_{i=1}^{n} \delta_i / (Q_{2, \beta}(Y_i) - \hat{\lambda}_\beta) = n\). Because the function \(\beta \mapsto \hat{\lambda}_\beta\) is implicitly defined, it is more difficult to compute the NPMLE of \(\mathcal{P}_2\) through (3.1), than the one of \(\mathcal{P}_3\) through (2.3) and (2.4).

In particular, solving (3.1) requires to run an optimization procedure over \(\beta\) for which, at each iteration, we shall evaluate \(\hat{\lambda}_\beta\), by an additional procedure. When \(\eta = \exp\), it is actually useless to solve (3.1), since it gives the same results as (2.3) and (2.4). This is the statement of the following proposition, the proof of which is given in Appendix A.

**Proposition 3.1.** Suppose that \(B = \mathbb{R}^d\), \(\eta = \exp\) and \(g(\gamma, x) = \exp(\gamma^T x)\) for every \(x \in \mathbb{R}^d\). If there exists \(i\) such that \(\delta_i = 1\), then \(\hat{\beta}^T = (\log(\hat{\theta}), \hat{\gamma}^T)\) and \(\hat{G} = \hat{F}\).

Hence our approach provides a new way to compute the NPMLE of \(\mathcal{P}_2\) in the common case where \(\eta = \exp\), which is simpler than the existing procedure [45, 26].
3.2. Cox and Breslow estimators. Model $\mathcal{P}_3$ is aimed to handle the presence of cured subjects in the data whereas the traditional Cox model, $\mathcal{P}_1$, is not. However, when $g(\gamma, x) = \exp(\gamma^T x)$, the estimator $\widehat{\gamma}$ in (2.3) is equal to the partial likelihood estimator of the Cox model:

**Proposition 3.2.** Suppose that $g(\gamma, x) = \exp(\gamma^T x)$ for every $x \in \mathbb{R}^d$. Then, the estimators $\widehat{\gamma}$ and $\widehat{\lambda}$ defined in (2.3) are respectively equal to the partial likelihood estimator of $\gamma$ and the Breslow estimator of $\lambda$ under the Cox model.

**Remark 1.** Note that by combining Propositions 3.1 and 3.2, we can conclude that the NPMLE under the promotion time cure model with $\eta = \exp$, can be computed by using the classical Cox estimator, which is computationally much easier and for which existing software packages can be used.

A consequence of Proposition 3.2 is that the derivation of the asymptotics for $(\widehat{\gamma}, \widehat{F}, \widehat{\theta})$ is somewhat similar as in the case of the Cox and Breslow estimator, provided for instance in [1]. An interesting difference with the Cox and Breslow estimator comes from the fact that

$$\min_{i=1, \ldots, n} \widehat{Q}_\gamma(Y_i) \geq n^{-1} \sum_{i=1}^{n} g(\gamma, X_i)(1 - \Delta_i).$$

From the framework described in the previous section, we deduce that $E[1 - \Delta_i \mid X] > 0$ and $E[g(\gamma_i, X_i) (1 - \Delta_i)] > 0$, for every $\gamma \in \mathbb{R}^d$. Consequently, the decreasing function $u \mapsto E[g(\gamma, X) R(u)]$ is bounded from below. In Lemma ??, see the Appendix, this property is shown to hold for $\widehat{Q}_{\gamma}$, uniformly in $\gamma$, with probability going to 1. This raises a significant difference with respect to classical Cox estimators in which the quantity corresponding to $\widehat{Q}_{\gamma}$ would go to 0 at infinity. This in turn implies that the weak convergence of the rescaled $\widehat{\delta}$ will still hold over $\mathbb{R}_{\geq 0}$. This is in contrast with the case of the Cox model for which such a convergence holds on bounded intervals. We refer to [1] for a discussion on the study of the Breslow estimator over $[0, +\infty)$.

4. Asymptotics. The asymptotic analysis of the NPMLE associated to model $\mathcal{P}_3$ is inspired by the approach developed for the Cox model in [1]. The main differences rely on: (i) the identification in model $\mathcal{P}_3$ is more difficult than in the Cox model due to the presence of $g$ which (in general) breaks the convexity properties that are useful in [1]; and (ii) the presence of cure which will affect the theoretical analysis.

We may first derive the asymptotic behaviour of the $Z$-estimator $\widehat{\gamma}$, and then rely on functional Delta-method type arguments, to describe $\widehat{\lambda}$. The monographs of [39] and [17] will be of good help at each of these steps to rely on suitable empirical process techniques. The preliminary study of $\widehat{\gamma}$ and $\widehat{\lambda}$ (given in Sections 4.1, 4.2 and 4.3) will provide the basis to describe the behaviour of $\widehat{p}(x)$ defined in (2.5).

As it is common for $M$-estimators, the asymptotic study of $\widehat{\gamma}$ starts with the establishment of its consistency. In contrast, for $\widehat{\lambda}$, we will rely on the explicit formula (2.3) to directly show the weak convergence of $\widehat{\lambda}$.

The model is said to be well-specified if the conditional survival function of $T$ given $X = x$ is given by $S_0(t \mid x) = \exp(-g(\gamma_0, x) \theta_0 F_0(t))$, for some $\gamma_0 \in B$, $\theta_0 \in \mathbb{R}_{\geq 0}$, and $F_0$ an absolutely continuous cumulative distribution function. This is assumed throughout the section and we put $A_0 = \theta_0 F_0$.

For each $i = 1, \ldots, n$, define the counting process $N_i(y) = \delta_1\{y_i \leq y\}$, recall that $R_i(y) =$
\[ M_i(y) = N_i(y) + g(\gamma_0, X_i) \int_0^y R(u)\theta_0 dF_0(u). \]

As in [12, Theorem 1.3.1], we are based on the following martingale property: for each \( i \), \( y \mapsto M_i(y) \) is a martingale with respect to the \( \sigma \)-field \( \mathcal{F}_y \) generated by \( \{N_i(u), 1_{\{\gamma_i \leq \delta \leq 0\}}, X_i : 0 \leq u \leq y\} \). That is \( y \mapsto g(\gamma_0, X_i) \int_0^y R(u)\theta_0 dF_0(u) \) is the compensator of the counting process \( N_i \). In particular, we have the formula [12, Theorem 1.5.1]

\[ E[\delta h(Y, X)] = \int E[h(u, X)g(\gamma_0, X)R(u)]\theta_0 dF_0(u) \]

for any bounded measurable function \( h \). Finally, the following identity shall be useful: for any bounded measurable functions \( h \) and \( \tilde{h} \), we have [12, Theorem 2.4.2]

\[ E \left[ \int h(u)dM(u) \int \tilde{h}(u)dM(u) \right] = \int h(u)\tilde{h}(u)E[g(\gamma_0, X)R(u)]d\Lambda_0(u). \]

4.1. Consistency of \( \hat{\gamma} \). The following hypothesis helps to control the complexity of the underlying class of functions as well as to guarantee the continuity of the function to maximize. Let \( | \cdot |_k \) denote the \( \ell_k \)-norm.

(H1) The set \( B \subseteq \mathbb{R}^q \) is compact. There exist functions \( m_1 : \mathcal{S} \to \mathbb{R}_{\geq 0} \) and \( M_1 : \mathcal{S} \to \mathbb{R}_{\geq 0} \) such that for every \( x \in \mathcal{S} \) and every \( \gamma \in B \), we have \( 0 < m_1(x) \leq g(\gamma, x) \leq M_1(x) \) and \( E[|\log(m_1(X))|], E[|\log(M_1(X))|] \) and \( E[M_1^2(X)] \) are finite. There exists a function \( c_1 : \mathcal{S} \to \mathbb{R}_{\geq 0} \) such that for every \( x \in \mathcal{S} \) and every \( (\gamma, \tilde{\gamma}) \in B^2 \),

\[ |g(\gamma, x) - g(\tilde{\gamma}, x)| \leq |\gamma - \tilde{\gamma}| c_1(x), \]

with \( 0 < E[c_1^2(X)] < +\infty \).

The two following assumptions are of good help to obtain the uniqueness of the maximizer of the likelihood function. Note that (H2) implies that every individual has a chance to be cured.

(H2) The variables \( T \) and \( C \) are independent given \( X \). Moreover, \( P(C > \tau | X) > 0 \) a.s., \( P(T = +\infty | X) > 0 \) a.s., and \( P(T \in (\tau, +\infty)) = 0 \).

(H3) For any \( \gamma \in \mathbb{R}^d \), \( \text{var}(g(\gamma_0, X)/g(\gamma, X)) = 0 \) implies that \( \gamma = \gamma_0 \).

**Proposition 4.1.** Under (H1)–(H3), we have that \( \hat{\gamma} \overset{P}{\to} \gamma_0 \).

We now discuss assumption (H3) by considering some examples.

**Example 1 (Cox with cure).** When \( g(\gamma, x) = \exp(\gamma^T x) \), (H3) is equivalent to the statement that \( \text{var}(X) \) has full rank.

Without specifying \( g(\gamma, x) = \exp(\gamma^T x) \), identifiability might not hold. Indeed, consider the case where \( g(\gamma, x) = |\gamma^T x| \), then of course different pairs \( (\theta, \gamma) \) could lead to the same function \( x \mapsto \theta|\gamma^T x| \). A possibility when facing such difficulties is to restrict \( \gamma \) to the unit sphere in \( \mathbb{R}^d \). Then identifiability might be recovered. We refer to this model as a directional model.
EXAMPLE 2 (directional model). Suppose that $g(\gamma, x) = \eta(\gamma^T x)$ with $|\gamma|_2 = 1$, $\gamma_1 > 0$. One can typically think of functions of the form $g(\gamma, x) = |\gamma^T x|^k$, for some $k \geq 1$. Such models allow for a geometric interpretation in the same vein as the single-index models [11]. The information available from the covariates $X$ to predict $Y$ is contained in the linear transformation $P_\gamma X$, where $P_\gamma$ stands for the orthogonal projector on span($\gamma$). For more details about identifiability of single-index models, we refer to Theorem 1 in [18] as well as Theorem 1 in [25] where $X$ is required to possess a density.

If $|\gamma|_2 = 1$ does not hold, then identifiability could fail unless more specific forms are considered for $\eta$. An example where identifiability is still satisfied is given below.

EXAMPLE 3 (Modified Cox). An interesting choice is when $g(\gamma, x) = \exp(\rho_k(\gamma^T x))$, where $\rho_k(t) = \text{sign}(t)|t|^k$, for $k > 1$. In the following lines, we obtain (H3) under the assumption that $X$ has a continuous density and $B(0, r)$ is included in the support of $X$. Suppose that $\rho_k(\gamma_0^T x) - \rho_k(\gamma^T x)$ is constant for almost every $x \in B(0, r)$. Suppose that $\gamma$ and $\gamma_0$ are linearly independent. Then, take $\alpha \in B(0, r)$ such that $\alpha^T \gamma = 0 \neq \alpha^T \gamma_0$. Let $g(x) = \rho_k(\gamma_0^T x) - \rho_k(\gamma^T x)$ and $K$ be a probability density function. For any $s \in (0, 1)$ we have, using approximation theory, that $(g * K_h)(s\alpha) \rightarrow s^k \rho_k(\gamma_0^T \alpha)$ as $h \rightarrow 0$, where $K_h(\cdot) = K(\cdot/h)/h^d$. Hence for any $s \in (0, 1)$, $s^k \rho_k(\gamma_0^T \alpha)$ is constant which is impossible. Supposing that $\gamma$ and $\gamma_0$ are linearly dependent, we directly obtain that $\gamma = \gamma_0$.

4.2. Asymptotic normality of $\hat{\gamma}$. We now introduce some notations that will be useful to express the asymptotic normality results. For every $y \in \mathbb{R}_{\geq 0}$, $\gamma \in \mathbb{R}^d$, $Q_\gamma(y) = E[g(\gamma, X)R(y)]$, $d_\gamma(x) = \nabla_\gamma g(\gamma, x)/g(\gamma, x)$, $h_\gamma(y) = \nabla_\gamma Q_\gamma(y)/Q_\gamma(y)$. We define

\begin{equation}
I_0 = \int E\left[\left\{d_0(X) - h_0(u)\right\}\left\{d_0(X) - h_0(u)\right\}^T g(\gamma_0, X)R(u)\right] d\Lambda_0(u),
\end{equation}

where $d_0 = d_{\gamma_0}$ and $h_0 = h_{\gamma_0}$. We require the following assumptions to obtain an asymptotic decomposition for $\hat{\gamma}$.

(H4) The matrix $I_0$ has full rank and $\gamma_0$ is in the interior of $B$.

(H5) For every $x \in S$, $\gamma \mapsto g(\gamma, x)$ is differentiable and there exists a function $c_2 : S \rightarrow \mathbb{R}_{\geq 0}$ such that for every $x \in S$ and every $(\gamma, \tilde{\gamma}) \in B^2$,

\begin{equation}
|\nabla_\gamma g(\gamma, x) - \nabla_\gamma g(\tilde{\gamma}, x)|_1 \leq |\gamma - \tilde{\gamma}|_1 c_2(x),
\end{equation}

with $0 < E[c_2^2(X)] < +\infty$. Moreover there exists a function $M_2 : S \rightarrow \mathbb{R}_{\geq 0}$ such that, for every $x \in S$, $|\nabla_\gamma g(\gamma, x)|_1 < M_2(x)$ where $E[M_2(X)]$, $E[M_2^2(X)/m_1(X)]$, $E[(c_2(X) + M_2(X))^2M_1(X)/m_1^2(X)]$, and $E[M_2^2(X)(c_1(X) + M_1(X))^2M_1(X)/m_1^4(X)]$ are finite.

**Proposition 4.2.** Under (H1)-(H5), we have that

\begin{equation}
n^{1/2}(\hat{\gamma} - \gamma_0) = n^{-1/2}I_0^{-1} \sum_{i=1}^n \left(\int (d_0(X_i) - h_0(u))dM_i(u) + o_p(1),
\end{equation}

and in particular, using Lemma ??, see the Appendix, combined with (4.2), it holds that $n^{1/2}(\hat{\gamma} - \gamma_0) \xrightarrow{d} N(0, I_0^{-1})$.  

8
4.3. Weak convergence of $\hat{\Lambda}$. Based on the decomposition obtained for $\hat{\gamma}$, we can now obtain a uniform representation of the process $\{n^{1/2}(\hat{\Lambda}(y) - \Lambda_0(y)) : y \in \mathbb{R}_{\geq 0}\}$. This is the statement of the next Proposition.

**Proposition 4.3.** Under (H1)-(H5), we have that

\begin{equation}
\sup_{y \in \mathbb{R}_{\geq 0}} \left| n^{1/2}(\hat{\Lambda}(y) - \Lambda_0(y)) - \left\{ n^{-1/2} \sum_{i=1}^{n} \int_{0}^{y} \frac{dM_{i}(u)}{Q_{0}(u)} - \int_{0}^{y} h_{0}(u)^{T}d\Lambda_{0}(u)n^{1/2}(\hat{\gamma} - \gamma_{0}) \right\} \right| = o_{\mathbb{P}}(1).
\end{equation}

In particular, using Lemma 2.6, the two terms involved in the decomposition are asymptotically independent and $n^{1/2}(\hat{\Lambda} - \Lambda_0)$ converges weakly to a tight centered Gaussian process in $\mathcal{C}(\mathbb{R}_{\geq 0})$ with covariance process given by

$$(y, y') \mapsto \int_{0}^{\min(y, y')} \frac{d\Lambda_{0}(u)}{Q_{0}(u)} + \left( \int_{0}^{y} h_{0}(u)^{T}d\Lambda_{0}(u) \right) I_{0}^{-1} \left( \int_{0}^{y'} h_{0}(u)d\Lambda_{0}(u) \right).$$

The two previous propositions, Proposition 4.2 and 4.3, form the basis of the next analysis, which ultimately describes the estimator $\hat{\rho}(x)$ of the cure proportion $p(x)$. The following results will be obtained as (almost direct) consequences of Propositions 4.2 and 4.3 and so are referred to as corollaries.

4.4. Asymptotic normality of $\hat{\theta}$. Since $\hat{\theta} = \lim_{y \to +\infty} \hat{\Lambda}(y) = \hat{\Lambda}(\tau)$ and $\theta_0 = \Lambda_0(\tau)$, the weak convergence of $n^{1/2}(\hat{\theta} - \theta_0)$ is deduced from the weak convergence of $n^{1/2}(\hat{\Lambda} - \Lambda_0)$ as the finite dimensional laws converge in distribution. The expression for the asymptotic variance is deduced from the one given in Proposition 4.3.

**Corollary 4.4.** Under (H1)-(H5), $n^{1/2}(\hat{\theta} - \theta_0)$ converges in distribution to a centered Gaussian distribution with variance

\begin{equation}
\nu_0 = \int \frac{d\Lambda_0(u)}{Q_0(u)} + \left( \int h_0(u)^{T}d\Lambda_0(u) \right) I_0^{-1} \left( \int h_0(u)d\Lambda_0(u) \right).
\end{equation}

As $\hat{F} = \hat{\Lambda}/\hat{\theta}$, invoking some Delta-metho arguments, the weak convergence of the process $n^{1/2}(\hat{F} - F_0)$ can be established. This however is not needed in the following.

4.5. Cure rate estimation. Recall that the cure proportion associated to $x \in S$ is given by $p_0(x) = \exp(-g(\gamma_0, x)\theta_0)$ and that the estimator is $\hat{p}(x) = \exp(-g(\hat{\gamma}, x)\hat{\theta})$. A Taylor development gives that

$n^{1/2}(\hat{p}(x) - p_0(x)) = -p_0(x)g(\gamma_0, x) \left\{ n^{1/2}(\hat{\theta} - \theta_0) + \theta_0 d_0(x)^{T}n^{1/2}(\hat{\gamma} - \gamma_0) \right\} + o_{\mathbb{P}}(1).$

Injecting (4.6) and (4.7) in the previous display leads to the following statement.

**Corollary 4.5.** Under (H1)-(H5), for a given $x \in S$, we have that

\begin{equation}
n^{1/2}(\hat{p}(x) - p_0(x)) = -p_0(x)g(\gamma_0, x)n^{-1/2} \sum_{i=1}^{n} \left\{ \int \frac{dM_{i}(u)}{Q_{0}(u)} + u_{0}(x)^{T}I_{0}^{-1} \int (d_{0}(X_{i}) - h_{0}(u))dM_{i}(u) \right\} + o_{\mathbb{P}}(1),
\end{equation}

9
where $u_0(x) = \theta_0d_0(x) - \int h_0(u)d\Lambda_0(u)$. Consequently, $n^{1/2}(\hat{p}(x) - p_0(x))$ converges in distribution to a centered Gaussian distribution with variance

$$v_p(x) = p_0(x)^2g(\gamma_0, x)^2 \left( \int \frac{d\Lambda_0(u)}{Q_0(u)} + u_0(x)^T I_0^{-1} u_0(x) \right).$$

Note that a similar result can be obtained concerning the estimator $\exp(-g(\hat{\gamma}, x)\Lambda(y))$ of the survival function $S_0(y|x)$ but we prefer to omit this for the sake of brevity.

5. Misspecification. Misspecification occurs when the probability distribution of $T$ given $X$ does not belong to the model of interest $\mathcal{P}_3$. This common case in practice is difficult to handle in theory because the estimator is likely to lie at the boundary of the domain making unavailable the vanishing gradient property or the useful equation (4.1), which involves the true values of the parameters $(\gamma_0, \Lambda_0)$. The approach taken here bears resemblance with the seminal approach proposed in [42] in that the consistency is established when the parameter of interest $\gamma^*$ might lie at the boundary of $B$ whereas the asymptotic normality is obtained when $\gamma^*$ is an interior point of $B$. Here the parameter of interest $\gamma^*$ is defined, not as the parameter value for which the model is true, but simply as the value that maximizes the expected partial likelihood $\gamma \mapsto E[\delta \log(g(\gamma, X)/Q_\gamma(Y))]$. The estimator is defined as before as a maximum of

$$\gamma \mapsto \sum_{i=1}^n \delta_i \log(g(\gamma, X_i)/Q_{\gamma}(Y_i)).$$

In contrast with [42], we face a semiparametric model which implies the presence of a functional estimator in the likelihood $\hat{Q}_\gamma$. One way to deal with this type of maximum likelihood estimators is to follow the approach proposed in [8]. However another approach is investigated here, making full use of the properties already established to deal with the correctly specified case.

To obtain the consistency, assumption (H3) needs to be replaced by the following one. Note that the assumption below was also needed for the consistency of the Cox estimator under misspecification [31].

(H6) The function $\gamma \mapsto E[\delta \log(g(\gamma, X)/Q_\gamma(Y))]$ defined on $B$ achieves its maximum at a unique point $\gamma^* \in B$.

**Proposition 5.1.** Under (H1), (H2) and (H6), we have that $\hat{\gamma} \xrightarrow{\mathbb{P}} \gamma^*$.

For the asymptotic normality, the following condition will replace (H4). Interestingly, the approach taken in the correctly specified case avoids the twice differentiability of $g$ whereas here we will require the existence of $A^*(X, Y)$ defined as $\nabla^2 \log(g(\gamma, X)/Q_\gamma(Y))$ evaluated at $\gamma^*$.

(H7) The point $\gamma^*$ is an interior point of $B$. For every $x \in \mathcal{S}$, the function $\gamma \mapsto g(x, \gamma)$ is twice differentiable at $\gamma^*$ and the matrix $V^* = E[\delta A^*(X, Y)]$ is invertible.

To state our final result we need to introduce the influence function, $\psi_\gamma$, defined by, for all $\gamma \in B$,

$$\psi_\gamma(y, \delta, x) = \{ \delta(d_\gamma(x) - h_\gamma(y)) - (L_{1,\gamma}(y, x) - E_{(Y,\delta)}L_{1,\gamma}(y, x)) \},$$

$$L_{1,\gamma}(Y, X_i) = \nabla_\gamma g(\gamma, X_i)E_{(Y,\delta)} \left[ \delta \frac{R_\gamma(Y)}{Q_\gamma(Y)} - g(\gamma, X_i)E_{(Y,\delta)} \left[ \frac{R_\gamma(Y)h_\gamma(Y)}{Q_\gamma(Y)} \right] \right].$$

where $E_{(Y,\delta)}$ stands for the expectation with respect to $(Y, \delta)$ keeping the other random variables fixed.
Proposition 5.2. Under (H1), (H2), (H5), (H6), (H7), we have that

\[ n^{1/2}(\hat{\gamma} - \gamma^*) = V^* - 1 n^{-1/2} \sum_{i=1}^{n} \psi_{\gamma^*}(Y_i, \delta_i, X_i) + o_P(1). \]

Moreover \( n^{1/2}(\hat{\gamma} - \gamma^*) \xrightarrow{d} N(0, V^* - 1 \text{var}(\psi_{\gamma^*}(Y, \delta, X))V^* - 1). \)

The asymptotic variance of \( \hat{\gamma} \) does not embrace the same form as the one obtained in the case of the correctly specified model. Of course starting from the variance formula in the misspecified case, given in the previous proposition, and supposing that the model is correctly specified, we recover the initial variance formula. This is done by checking that

\[ n^{-1} \sum_{i=1}^{n} L_{1, \gamma_0}(Y_i, X_i) = n^{-1} \sum_{i=1}^{n} \int (d_0(X_i) - h_0(u))g(\gamma_0, X_i)R_i(u)d\Lambda_0(u), \]

and using the classical formula that expresses the Fisher information matrix \( I_0 \) as the Hessian matrix \( V^* \) (see for instance equation (3.1) in [42]). Because in general the formulas are different, the estimator employed for the variance should depend on whether the model is correctly specified or not. Hence one might want to test if the model is correctly specified before estimating the variance. Finally these differences, that arise between the correctly specified and the misspecified case, might be used to conduct a goodness-of-fit test. This type of approach was initially proposed in [42] where the author recommends to consider the difference between the two variance estimates (in the misspecified and the correctly specified case) as the test statistic. To the best of our knowledge, this has not been exploited yet under the proportional hazards model and, consequently, this might open new avenues for further research on testing for misspecification in the proportional hazards model.

Note in passing that, in the misspecified case, convergence results concerning the other estimators, \( \hat{\Lambda}, \hat{\theta} \) and \( \hat{p} \), can be obtained in a similar way as in the correctly specified case, that is, as a byproduct of the analysis of \( \hat{\gamma} \).

6. High dimensional estimation.

6.1. High dimensional set-up. In this section, we study the high-dimensional setup in which the number of covariates can be larger than the sample size. We consider a sequence of data generating processes indexed by \( n \). The distribution of \( X \) and the true parameter \( \gamma_0 \) are now allowed to depend on \( n \). For the sake of clarity, we assume that the parameters \( \theta_0 \) and \( F_0 \) (and so \( \Lambda_0 \)) do not vary with \( n \). This restriction could be alleviated with the help of additional technicalities.

We focus on the asymptotic regime where \( n \) goes to infinity and \( d = d_n \geq 1 \) is a function of \( n \) which may go to infinity with \( n \). The asymptotic regime satisfies

\[ \frac{\log(nd)}{s_0} n \to 0 \quad \text{as} \quad n \to \infty. \]

where \( s_0 = |\gamma_0|_0 \geq 1 \) and \( |\cdot|_0 \) stands for the number of non-zero coefficients. As a consequence, we allow for scenarios where the number of covariates \( d \) is larger than the sample size \( n \).

As in the previous section, we are interested in the estimation of the parameter and the cure rate under this setup, as these are the quantities of interest in many applications, but we can also
consider estimators of $\Lambda_0$, $\theta_0$ and $S_0$ along the same lines. The results we obtain in this framework are valid for functions $g$ of the form

$$g(\gamma, x) = g(\gamma^T x), \quad x \in \mathbb{R}^d,$$

with from now on $g : \mathbb{R} \to \mathbb{R}_{\geq 0}$. This includes, as an example, the Cox model with cure fraction, i.e., when $g(\gamma, X) = \exp(\gamma^T X)$. Note that the Cox model has the advantage of having a concave likelihood function which is suitable when dealing with high dimensional data [5, 13, 16, 44]. In what follows, we cannot use the results of these papers as we consider more general functions $g$ for which the likelihood might not be concave. Remark finally that the fact that $g(\gamma, x) = g(\gamma^T x)$ is natural as it makes the function $g$ independent of the dimension.

To handle the potentially large number of covariates, we rely on the following LASSO estimator [33]:

$$\hat{\gamma}_L \in \arg\min_{\gamma \in B} \frac{1}{n} \sum_{i=1}^{n} \delta_i \left( \log \left( \hat{Q}_\gamma (Y_i) \right) - \log (g(\gamma^T X_i)) \right) + \lambda_n |\gamma|_1,$$

where $B \subset \mathbb{R}^d$, $\hat{Q}_\gamma (u) = n^{-1} \sum_{i=1}^{n} g(X_i^T \gamma)R_i(u)$ for all $u \geq 0$ and $(\lambda_n)_{n \geq 1}$ is a positive sequence of regularization parameters such that

$$\lambda_n = \sqrt{\frac{\log(nd)}{n}}.$$

This choice ensures that the empirical likelihood is correctly penalized.

Such a penalized approach takes advantage of sparse situations, i.e., where the number of informative covariates is small. This is formally stated in the theoretical results that follows.

6.2. Main results. We cannot rely on results from the few other papers ([30, 16, 19, 21]) that have developed an high-dimensional estimation theory for nonconvex penalized loss because their frameworks are not adapted to the problem of the present paper. Indeed, in all these works, the empirical loss is a sum of i.i.d. variables which prevents us to use directly their results due to the presence of $\hat{Q}_\gamma (u)$. The paper [30] derives an oracle inequality for the likelihood at the true parameter but does not give a convergence result on the estimator of the parameter. [16, 19] make restrictive global (restricted) convexity assumptions, which are hard to justify in our framework for a general function $g$. Finally, by relying on the uniform convergence of the gradient of the empirical likelihood, we bear resemblance with [21] although they do not provide a high-level result on the rate of convergence of the estimator of the parameter.

We develop our own set of results which are specific to the model of the present paper and rely on the assumptions below. The following regularity condition is imposed on the function $g$.

(H8) $g : \mathbb{R} \mapsto \mathbb{R}_{\geq 0}$ is two times differentiable with continuous second derivative.

Let $S$ denote the support of $X$. As in [30], we require that $\sup_{\gamma \in B, x \in S} |\gamma^T x|$ is bounded. To do this, we assume that the set $B$ is included in the centered $\ell_1$-ball with radius $r$, that is

$$B \subset B(r) = \{ \gamma \in \mathbb{R}^d : |\gamma|_1 \leq r \}$$

and that the covariates are bounded:

(H9) There exists a constant $M_X$ such that for all $n \geq 1$, $|x|_\infty \leq M_X$ for all $x \in S$. 

12
The radius \( r \) does not depend on the sample size. The bound \( \sup_{\gamma \in B, x \in \mathcal{S}} |\gamma^T x| \leq r M_X \) will be useful to control the behaviour of the empirical log-likelihood and its gradient.

For \( \gamma \in B \), let \( L(\gamma) = E[\delta \{ \log(Q_{\gamma}(Y)) - \log(g(\gamma^T X)) \}] \) be minus the expected log-likelihood. The function \( L \) depends on \( n \). In order to prove consistency of the estimator, we also need an identification assumption.

(H10) For all \( \eta > 0 \), there exists \( \epsilon > 0 \) such that for all \( n \geq 1 \),

\[
\inf_{\gamma \in B : |\gamma - \gamma_0|_2 \geq \eta} L(\gamma) - L(\gamma_0) \geq \epsilon.
\]

This is a shape restriction on the expected log-likelihood function. When \( n \) is fixed, this condition holds if the expected log-likelihood is continuous and \( \gamma_0 \) is its unique maximizer (see for instance, [24]). We just require this condition to be satisfied with an \( \epsilon \) which does not depend on \( n \). It plays the same role as the margin condition in [30] but is different because it relates the likelihood to the parameter rather than to the scalar product of the parameter and the regressors. Notice that we have to impose a global condition because the problem is possibly nonconvex. If the likelihood were convex, then only a local condition would suffice (e.g. that the eigenvalues of the Hessian of minus the expected likelihood are bounded from below by a positive constant in a neighborhood of the true parameter). As a last remark, it would be possible to relax this assumption by letting \( \epsilon \) to go to 0 too quickly.

We also assume that the smallest eigenvalue of the Hessian of \( L \) is bounded from below in an \( \ell_2 \)-ball around \( \gamma_0 \).

(H11) There exist constants \( \rho_*, \eta_* > 0 \) such that for all \( n \geq 1 \)

\[
\inf_{\gamma \in B, |\gamma - \gamma_0|_2 \leq \eta_*} \rho_{\min}(\nabla^2 L(\gamma)) \geq \rho_*,
\]

where \( \rho_{\min}(\cdot) \) is the minimal eigenvalue.

This is a local strict convexity assumption classical in the \( M \)-estimation literature (see e.g. [38] in which a similar restriction is made but, again, with respect to the scalar product between the parameters and the regressors). Here, the condition is required to hold uniformly in \( n \).

In order to illustrate our conditions, notice that, when \( g = \exp \), hypotheses (H10) and (H11) hold when the minimal eigenvalue of the Hessian of \( L \) at \( \gamma_0 \) (which depends on \( n \)) is bounded from below uniformly in \( n \) as shown in the following lemma.

**Lemma 6.1.** Let \( g = \exp \) and suppose that Assumption (H9) is fulfilled. If there exists \( \kappa_* > 0 \) such that for all \( n \geq 1 \), \( \rho_{\min}(\nabla^2 L(\gamma_0)) \geq \kappa_* \), then Assumptions (H10) and (H11) are satisfied.

We have the following result on the estimation of the true parameter.

**Proposition 6.2.** Suppose that Assumptions (H8), (H9), (H10), and (H11) are fulfilled. Then, under (6.1), it holds with probability going to 1,

\[
|\hat{\gamma}_L - \gamma_0|_1 \leq \frac{12}{\rho_* s_0} \sqrt{\frac{\log(nd)}{n}}.
\]

Once the parameter \( \hat{\gamma}_L \) has been estimated based on (6.2), the estimate of the cure rate at \( x \in \mathcal{S}_n \) is given by \( \hat{p}_L(x) = \exp(-g(\hat{\gamma}_L^T x) \hat{\theta}_L) \), where \( \hat{\theta}_L = n^{-1} \sum_{i=1}^n \hat{Q}_{\hat{\gamma}_L}(Y_i)^{-1} \delta_i \). The following result provides an upper bound on the estimation error associated to the cure rate.
Proposition 6.3. Suppose that Assumptions (H8), (H9), (H10), and (H11) are fulfilled and that (6.1) holds. Let \( \{x_n\}_n \) be a sequence of points such that \( x_n \in S \). Then, we have

\[
\hat{p}_L(x_n) - p_0(x_n) = O_p \left( s_0 \sqrt{\frac{\log(nd)}{n}} \right).
\]

7. Simulation study. We performed some extensive Monte Carlo simulations in order to assess the performance of our suggested estimators. Our implementation is done in the statistical computing language R [27], and the code is available on github (https://github.com/anouel/ExtPromotionTime).

The simulations were performed under a variety of conditions on the censoring rate, sample size and cure rate. The data were generated according to the following model:

\[
S(t|x_1, x_2) = \exp \left[ - \exp \left\{ \Gamma(\gamma_0_1 x_1 + \gamma_0_2 x_2) \right\} \theta_0 F_0(t) \right].
\]

In the above model, we chose the link function \( \Gamma(\cdot) \) to be either the identity, the cubic or the sine function. For clarity, in the first part of this simulation study we will focus on the case of the identity function. With few exceptions, all our comments and findings also apply to the case where \( \Gamma(\cdot) = (\cdot)^3 \) and \( \Gamma(\cdot) = \sin(\cdot) \). In all our simulations, \( \log(\theta_0) = 0.1, \gamma_0_1 = -2, \gamma_0_2 = 1, F_0 \) is the cumulative distribution function of a uniform variable on \([0, 1] \), \( X_1 \) is a uniformly distributed random variable on \([\alpha, \alpha + 1] \), \( X_2 \) is a normal random variable with mean \( \alpha \) and standard deviation \( 1/12 \), and \( X_1 \) and \( X_2 \) are independent. The censoring variable is exponential with parameter \( \lambda \) and is independent of \( (X_1, X_2) \). By varying the latter we mainly control the censoring rate, while by varying \( \alpha \) we control the cure rate.

Suppose we have a sample \( (Y_i, \delta_i, X_i), i = 1, \ldots, n \) from the distribution described above, with \( X_i = (X_{i1}, X_{i2})^T \). We obtain \( \hat{\gamma} = (\hat{\gamma}_1, \hat{\gamma}_2)^T \), the estimator of \( \gamma_0 = (\gamma_0_1, \gamma_0_2)^T \), by maximizing the partial likelihood function given by (2.3), using the Newton-Raphson algorithm. We get \( \hat{\theta} \), the estimator of \( \theta_0 \), by applying (2.4). The cure probability estimator is then obtained by

\[
\hat{p}(x_1, x_2) = \exp \left[ - \exp \left\{ \Gamma(\hat{\gamma}_1 x_1 + \hat{\gamma}_2 x_2) \right\} \hat{\theta} \right].
\]

Using the plug-in principle together with (4.4), we obtain an estimator for the asymptotic variance-covariance matrix of \( \hat{\gamma} \) which is given by \( \hat{I}^{-1}/n \), where

\[
\hat{I} = n^{-1} \sum_{i=1}^{n} \delta_i \left\{ (d_{\gamma}(X_i) - \hat{h}_{\gamma}(Y_i))(d_{\gamma}(X_i) - \hat{h}_{\gamma}(Y_i))^T \right\},
\]

with, for every \( \gamma \in B \), \( \hat{h}_{\gamma}(y) = \nabla_{\gamma} \hat{Q}_{\gamma}(y)/\hat{Q}_{\gamma}(y) \). Similarly, using (4.8), we obtain an estimator of the asymptotic variance of \( \hat{\theta} \) given by \( \hat{v}_\theta/n \), where

\[
\hat{v}_\theta = n^{-1} \sum_{i=1}^{n} \frac{\delta_i}{Q_{\gamma}(Y_i)^2} + \left( n^{-1} \sum_{i=1}^{n} \frac{\delta_i \hat{h}_{\gamma}(Y_i)}{Q_{\gamma}(Y_i)} \right)^T \hat{I}^{-1} \left( n^{-1} \sum_{i=1}^{n} \frac{\delta_i \hat{h}_{\gamma}(Y_i)}{Q_{\gamma}(Y_i)} \right).
\]

And using the expression for the variance of \( \hat{p} \) given in Corollary 4.5, we obtain an estimator of the asymptotic variance of \( \hat{p}(x) \) given by \( \hat{v}_p/n \), where

\[
\hat{v}_p = \hat{p}(x)^2 g(\hat{\gamma}, x)^2 \left( n^{-1} \sum_{i=1}^{n} \frac{\delta_i}{Q_{\gamma}(Y_i)^2} + \hat{u}(x)^T \hat{I}^{-1} \hat{u}(x) \right),
\]
with \( \hat{u}(x) = \hat{\theta} d\gamma(x) - n^{-1} \sum_{i=1}^n \delta_i h\gamma(Y_i) / \hat{Q}_\gamma(Y_i) \) and \( x = (x_1, x_2)^T \).

We perform \( N = 2000 \) replications for four sample sizes \((n = 100, n = 200, n = 400, n = 600)\), three levels of censoring \((20\%, 40\% \text{ and } 60\%)\) and three levels of cure \((20\%, 40\% \text{ and } 60\%)\). For every scenario and every replication, we calculate the estimators \( \hat{\gamma}_1, \hat{\gamma}_2, \hat{\theta} \) and \( \hat{p}(x_1, x_2) \) together with their estimated asymptotic variance \((\hat{AV}ar)\) and the corresponding asymptotic 95\% confidence intervals based on the asymptotic normality. Based on the 2000 replications, we also calculate the empirical bias, the empirical variance \((VAR)\), the empirical mean squared error \((MSE)\) of every estimator together with the empirical coverage probability \((COV)\) for the confidence intervals. In the case of the cure probability \( p(x_1, x_2) \) we did the calculations for \( x_2 = 0 \) and every quantile of \( X_1 \) corresponding to the probability levels \(0.01, 0.02, \ldots, 0.99\). We summarize the results by taking the average of the resulting 99 empirical COV’s. Due to space limitations, we provide below only some selected but representative scenarios.

Figure 1 provides the boxplots for \( \hat{\gamma}_1, \hat{\gamma}_2 \) and \( \hat{\gamma}_0 = \log(\hat{\theta}) \). By comparing the upper and lower part \((n = 100 \text{ vs } n = 600)\) of this figure, we clearly see that the performance of the estimators improves with increasing sample size both in terms of bias and variance. This confirms the consistency of these estimators. This figure also shows the effect of the cure rate and the censoring rate. As expected, increasing the latter rates results in a larger bias and, especially, in a larger variance of the estimators. This effect can also be seen in Figure 2 which provides the boxplots for the asymptotic estimated variances. Compared to the censoring rate, the cure rate seems to have no, or very limited, effect on \( \hat{\gamma}_1 \) and \( \hat{\gamma}_2 \), but it does affect the bias and the variance of \( \hat{\gamma}_0 \). In fact, when the percentage of cure increases, the bias and the variance of \( \hat{\gamma}_0 \) decrease (and so does the MSE). Globally, it seems that the estimation of \( \hat{\gamma}_0 \) is more difficult than the estimation of \( \hat{\gamma}_1 \) and \( \hat{\gamma}_2 \). This is especially the case when the censoring percentage is large and the cure probability is small. If moreover the sample size is small, then the bias can be quite large.

As we said before, Figure 2 provides the boxplots for the asymptotic estimated variances. The plots suggest that the proposed estimators are consistent (note that the y-axis in the upper and the lower plots do not have the same scale). Basically the remarks we made above on the effect of the proportion of cure and censoring remain valid for the proposed estimators of the variances. Again, it can be seen that estimating the variance of \( \hat{\gamma}_0 \) is more difficult and can lead to, relatively, large variances especially when the censoring and cure rates are large and the sample size is small.

Figure 3 which provides some Q-Q plots for the estimated parameters confirms the validity of the normal approximation of the sampling distributions of \( \hat{\gamma}_1 \) and \( \hat{\gamma}_2 \). However, this approximation seems to be less accurate for \( \hat{\theta} \) even when \( n = 600 \) (figure not shown here). In fact the sampling distribution of the latter tends to be positively skewed especially when the censoring rate is large. Applying the logarithmic transformation, seems to solve the problem as it makes the distribution more symmetric (see the Q-Q plot for \( \hat{\gamma}_0 \) in Figure 3).
Fig 1: Boxplots of $\hat{\gamma}_1$, $\hat{\gamma}_2$ and $\hat{\gamma}_0$ for $n = 100$ and $n = 600$ and for $\Gamma(\cdot) = \cdot$. The empirical mean of the estimates is indicated by a +. The true values are indicated by a horizontal line.
Fig 2: Boxplots of $\hat{AVar}(\gamma_1), \hat{AVar}(\gamma_2)$ and $\hat{AVar}(\gamma_0)$ for $n = 100$ and $n = 600$ and for $\Gamma(\cdot) = \cdot$. The empirical mean of $\hat{AVar}$ is indicated by a $+$, the empirical variance of the estimates $(\hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_0)$ is indicated by a $\times$. 
In Table 1 we give the MSE and the variance for some of the studied scenarios and for \( \Gamma(\cdot) = \cdot \), \( \Gamma(\cdot) = (\cdot)^3 \) and \( \Gamma(\cdot) = \sin(\cdot) \). It is clear from these results that the variance is the dominant component in the mean squared errors. It can also be observed that the obtained results with the link \( \Gamma(\cdot) = \cdot \) and \( \Gamma(\cdot) = (\cdot)^3 \) are globally better than the corresponding results obtained with \( \Gamma(\cdot) = \sin(\cdot) \). Table 1 also shows the coverage probabilities (COV) of the 95\% asymptotic confidence intervals for the parameters \( \gamma_1, \gamma_2 \) and \( \gamma_0 = \log(\theta) \). The confidence intervals for the latter are based on the asymptotic normality of \( \hat{\theta} \) and the Delta method. Globally, the obtained COV’s are close to the nominal level. With \( n = 100 \), the confidence intervals tend to be liberal when \( \Gamma(\cdot) = \sin(\cdot) \) especially for \( \gamma_0 \). This also happens for \( \gamma_1 \) when \( \Gamma(\cdot) = (\cdot)^3 \).
In the case of the identity link function, we also perform a simulation study to compare our method with the classical approach based on the parametrization $P_2$. For the latter, the calculation was done using the algorithm described in [20], implemented in the R package “miCoPTCM” (see [2]), which, to the best of our knowledge, represents the most up-to-date algorithm currently available in the literature. First, in terms of bias, MSE and VAR, the two approaches give almost exactly the same results for all settings (the results are not shown here). This is not surprising since, under this setting, i.e. with the exponential link function, the two methods are equivalent. But this is not the case for COV; see Table 2. In fact, it can be clearly seen from this table, that $P_2$ always leads to liberal confidence intervals as the observed coverage probabilities are smaller than the desired 95%, especially for a small sample size ($n = 100$). In contrast, our method gives coverage probabilities that are typically closer to (or higher than) the nominal level. This confirms our remark regarding the advantage of having a closed formula to calculate the confidence intervals. Second, in terms of
computation time, after calculating the estimators and their asymptotic variances using the two approaches, we observed the following. In case of small sample sizes ($n \leq 200$), the computational effort, measured by CPU time, is more or less the same for both methods but as the sample size increases the difference becomes larger and larger in favor of our method. More precisely, for a sample of 400 observations, the computational effort required by the classical algorithm is roughly 2 times larger than that required by the algorithm of the present paper. For 600 observations, the effort is nearly 5 times larger.

<table>
<thead>
<tr>
<th>$n$</th>
<th>%cure</th>
<th>%cens</th>
<th>COV for $P_2$</th>
<th>COV for $P_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>10</td>
<td>20</td>
<td>0.928 0.808 0.722</td>
<td>0.962 0.902 0.879</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>20</td>
<td>0.948 0.820 0.777</td>
<td>0.970 0.918 0.905</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>40</td>
<td>0.938 0.814 0.763</td>
<td>0.962 0.913 0.896</td>
</tr>
<tr>
<td>100</td>
<td>40</td>
<td>40</td>
<td>0.959 0.852 0.845</td>
<td>0.973 0.932 0.923</td>
</tr>
<tr>
<td>100</td>
<td>40</td>
<td>60</td>
<td>0.957 0.858 0.848</td>
<td>0.980 0.927 0.924</td>
</tr>
<tr>
<td>600</td>
<td>10</td>
<td>20</td>
<td>0.957 0.920 0.906</td>
<td>0.968 0.948 0.941</td>
</tr>
<tr>
<td>600</td>
<td>20</td>
<td>20</td>
<td>0.961 0.938 0.932</td>
<td>0.968 0.956 0.952</td>
</tr>
<tr>
<td>600</td>
<td>20</td>
<td>40</td>
<td>0.957 0.927 0.919</td>
<td>0.972 0.954 0.942</td>
</tr>
<tr>
<td>600</td>
<td>40</td>
<td>40</td>
<td>0.959 0.948 0.946</td>
<td>0.968 0.962 0.956</td>
</tr>
<tr>
<td>600</td>
<td>40</td>
<td>60</td>
<td>0.964 0.948 0.945</td>
<td>0.969 0.964 0.956</td>
</tr>
</tbody>
</table>

**Table 2**

Empirical coverage probability (COV) for nominal 95% confidence intervals for $\gamma_1$, $\gamma_2$ and $\gamma_0$ using parametrizations $P_2$ and $P_3$.

Figure 4 shows the empirical coverage probabilities (COV) of the confidence intervals for $p(x_1, x_2)$. We can see that these COV’s can be quite unsatisfactory especially in the left tail of the support of $X_1$ even when the sample size is relatively large. To correct for this, we apply the logit transformation and the Delta method to construct confidence intervals for $\log(p/(1-p))$ and transform back (taking the logistic transformation) to get confidence intervals for the cure probabilities. This leads to very satisfactory results with coverage probabilities close to the nominal level both in the middle and in the tails especially when the sample size is large.
Fig 4: Empirical coverage probabilities of nominal 95% confidence intervals for the cure probability as a function of $x_1$ for $x_2 = 0$. The coverage probabilities obtained without transformation are indicated by a +, those obtained after a logit transformation are indicated by a ×. The proportion of censoring and the cure rate both equal 0.40.

Regarding the case of high-dimensional data, we performed a simulation study to check the validity of the proposed LASSO-type estimator of the cure rate and compared its performance with the standard estimator, i.e., the one based on the unconstrained likelihood. Data were generated as described above using the identity link function and various sets of inactive covariates of size varying from 0 to 800. All these covariates were generated independently from a normal distribution with mean $\alpha$ and standard deviation 1. The results of this study are summarized in Table 3, which clearly shows that the LASSO estimator performs much better than the standard estimator. As expected, the latter becomes very unstable as the number of covariates becomes larger and larger and breaks down completely when this number exceeds the sample size.
Table 3

<table>
<thead>
<tr>
<th>Number of inactive covariates</th>
<th>Standard Bias</th>
<th>Standard MSE</th>
<th>LASSO Bias</th>
<th>LASSO MSE</th>
<th>Standard Bias</th>
<th>Standard MSE</th>
<th>LASSO Bias</th>
<th>LASSO MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>16.05</td>
<td>8.00</td>
<td>15.29</td>
<td>7.76</td>
<td>5.45</td>
<td>5.52</td>
<td>4.36</td>
<td>5.41</td>
</tr>
<tr>
<td>5</td>
<td>-12.98</td>
<td>8.20</td>
<td>1.38</td>
<td>7.57</td>
<td>-9.56</td>
<td>5.66</td>
<td>-2.73</td>
<td>5.30</td>
</tr>
<tr>
<td>10</td>
<td>-44.84</td>
<td>8.70</td>
<td>-7.91</td>
<td>7.59</td>
<td>-28.36</td>
<td>5.54</td>
<td>-6.02</td>
<td>4.97</td>
</tr>
<tr>
<td>200</td>
<td>-1492.66</td>
<td>228.12</td>
<td>-49.83</td>
<td>7.87</td>
<td>-1011.87</td>
<td>109.82</td>
<td>-27.23</td>
<td>5.21</td>
</tr>
<tr>
<td>600</td>
<td>NA</td>
<td>NA</td>
<td>-47.07</td>
<td>7.93</td>
<td>NA</td>
<td>NA</td>
<td>-31.41</td>
<td>5.53</td>
</tr>
<tr>
<td>800</td>
<td>NA</td>
<td>NA</td>
<td>-48.67</td>
<td>8.32</td>
<td>NA</td>
<td>NA</td>
<td>-41.53</td>
<td>5.42</td>
</tr>
</tbody>
</table>

10000×Bias and 10000×MSE for the estimated cure rate using the standard method and the LASSO method.

8. Real data application. To illustrate the application of our model, the proposed methodology is applied on a real data set from a breast cancer study. The dataset consists of 286 patients that experienced a lymph-node-negative breast cancer between 1980 to 1995 [40]. The event time of interest is the time to distant metastasis (DM). Among the 286 patients, 107 experienced a relapse from breast cancer. As can be seen from Figure 5, the Kaplan-Meier estimator of the survival function shows a large plateau at about 0.60. Furthermore, 88% of the censored observations are in the plateau. A cure model seems therefore appropriate for these data.

![Fig 5: Kaplan-Meier estimator of the survival curve for time to distant metastasis for breast cancer survival data (censored observations are indicated by +).](image)

We consider two covariates: the age of the patient (ranges from 26 to 83 with a median of 52 years) and the estrogen receptor (ER) status, which is a binary variable equaling 0 (ER−) in the case of less than 10 fmol per mg protein (77 patients in total) and equaling 1 (ER+) when 10 fmol per mg protein or more (209 patients in total). We analyse the data using the semiparametric model given in (7.1) and we choose the link function \( \Gamma(x) \) to be either \( x^k \) or \( \sin(x^k) \) with \( k = 1, \ldots, 8 \). In Table 4 we report the values of the obtained profile log-likelihood (PLL) as given by (2.3), and the obtained full log-likelihood (FLL) as given by (2.2). Based on this, and given that the number of parameters is fixed, one may conclude that, in terms of the likelihood, the model that fits best these data is the model with the sine function and \( k = 4 \).
\[ \Gamma(x) = x^k \]

<table>
<thead>
<tr>
<th>( k )</th>
<th>PLL</th>
<th>FLL</th>
<th>( k )</th>
<th>PLL</th>
<th>FLL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25.0</td>
<td>-687.1</td>
<td>2</td>
<td>25.9</td>
<td>-686.2</td>
</tr>
<tr>
<td>3</td>
<td>25.2</td>
<td>-686.9</td>
<td>4</td>
<td>25.9</td>
<td>-686.2</td>
</tr>
<tr>
<td>5</td>
<td>25.4</td>
<td>-686.7</td>
<td>6</td>
<td>26.1</td>
<td>-686.0</td>
</tr>
<tr>
<td>7</td>
<td>25.6</td>
<td>-686.5</td>
<td>8</td>
<td>25.3</td>
<td>-686.8</td>
</tr>
</tbody>
</table>

\[ \Gamma(x) = \sin(x^k) \]

<table>
<thead>
<tr>
<th>( k )</th>
<th>PLL</th>
<th>FLL</th>
<th>( k )</th>
<th>PLL</th>
<th>FLL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25.3</td>
<td>-686.8</td>
<td>2</td>
<td>28.1</td>
<td>-684.0</td>
</tr>
<tr>
<td>3</td>
<td>25.4</td>
<td>-686.7</td>
<td>4</td>
<td>24.1</td>
<td>-688.0</td>
</tr>
<tr>
<td>5</td>
<td>25.2</td>
<td>-686.9</td>
<td>6</td>
<td>25.5</td>
<td>-686.6</td>
</tr>
<tr>
<td>7</td>
<td>25.6</td>
<td>-686.5</td>
<td>8</td>
<td>25.8</td>
<td>-686.3</td>
</tr>
</tbody>
</table>

Table 4

The profile log-likelihood (PLL) and the full log-likelihood (FLL) for different link functions.

So, with \( X_1 \) being the age and \( X_2 \) being the ER status, the selected model is given by

\[ S(t|x_1, x_2) = \exp \left[ -\exp \left\{ \sin(\gamma_1 x_1 + \gamma_2 x_2)^4 \right\} \theta F(t) \right]. \]

The estimated parameters together with their standard deviations (STD) and 95% confidence intervals (CI) are given in Table 5 below.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>STD</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \gamma_1 )</td>
<td>0.018</td>
<td>0.001</td>
<td>[0.016, 0.020]</td>
</tr>
<tr>
<td>( \gamma_2 )</td>
<td>-1.625</td>
<td>0.118</td>
<td>[-1.857, -1.393]</td>
</tr>
<tr>
<td>( \theta )</td>
<td>0.340</td>
<td>0.047</td>
<td>[0.247, 0.433]</td>
</tr>
</tbody>
</table>

Table 5

The parameters of the selected model.

9. Discussion. The promotion time cure model is commonly used for survival data with a cure fraction. Unlike the well-known Cox model, for which the nonparametric maximum likelihood estimators can be easily obtained, the classical parametrization of this model leads to some computational and theoretical complications caused by the fact that the estimated regression parameters can only be obtained as maximizers of a certain implicitly defined objective function. To overcome this drawback a new parametrization was proposed which leads to a simple estimation procedure. This makes the theoretical developments more transparent, the proofs easier to follow and helps to get closed formulas for the asymptotic variances. It also leads to a significant improvement in terms of computation time. For example, with a sample of size 600, the computational effort, measured by CPU time, required by the classical algorithm is roughly 5 times larger than that required by the algorithm of the present paper. The derived asymptotic results together with the closed formulas for the variance allow us to build better confidence intervals for some quantities of interest as for instance the proportion of cure for a given set of covariates, which plays an important role in cure models. The simulation study confirms the theoretical results and reveals some other interesting findings such as, for example, the fact that the Delta method based on the logit transformation produces better confidence intervals compared to the classical Wald method. It may be interesting to consider other transformations like the arcsine and the probit function, but this was not investigated here. A more interesting question that deserves an in-depth investigation is the choice of the link function \( g \). For that one needs to find an "optimal" balance between model complexity and model fitting of the data.

Supplementary material. The supplement gathers the proofs of all the results along with some useful technical lemmas.
Acknowledgements. J. Beyhum and I. Van Keilegom were supported by the European Research Council (2016-2021, Horizon 2020/ ERC grant agreement No. 694409). F. Portier was supported by Fonds de la Recherche Scientifique (FNRS) A4/5 FC 2779/2014-2017 No. 22342320, and A. El Ghouch was supported by the PDR (convention PDR.T.0080.16), a funding instrument of the FNRS.

References.


