Semiparametric inference for the recurrent event process by means of a single-index model
Olivier Bouaziz, Ségolen Geffray, Olivier Lopez

To cite this version:
Olivier Bouaziz, Ségolen Geffray, Olivier Lopez. Semiparametric inference for the recurrent event process by means of a single-index model. 2010. hal-00446528v1

HAL Id: hal-00446528
https://hal.archives-ouvertes.fr/hal-00446528v1
Submitted on 19 May 2010 (v1), last revised 16 Sep 2014 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Semiparametric inference for the recurrent event process by means of a single-index model

Olivier Bouaziz¹, Ségolen Geffray², Olivier Lopez³

Abstract

In this paper, we introduce new parametric and semiparametric regression techniques for a recurrent event process subject to random right censoring. We develop models for the cumulative mean function and provide asymptotically normal estimators. Our semiparametric model which relies on a single-index assumption can be seen as a reduction dimension technique that, contrary to a fully nonparametric approach, is not stroke by the curse of dimensionality when the number of covariates is high. We discuss data-driven techniques to choose the parameters involved in the estimation procedures and provide a simulation study to support our theoretical results.

Key words: asymptotic normality, dimension reduction, empirical processes, recurrent events, right-censoring, single-index model.

¹ Laboratoire de Modélisation Aléatoire, Université Paris X, 200 avenue de la République, 92000 Nanterre, France, E-mail: obouaziz@u-paris10.fr.
² IRMA, UMR 7501, 7 rue René-Descartes, 67084 Strasbourg Cedex, France, E-mail: geffray@math.unistra.fr
³ Laboratoire de Statistique Théorique et Appliquée, Université Paris VI, 175 rue du Chevaleret, 75013 Paris, France, E-mail: olivier.lopez0@upmc.fr.
This work is supported by French Agence Nationale de la Recherche (ANR) ANR Grant “Prognostic” ANR-09-JCJC-0101-01.
1 Introduction

The modeling of recurrent events has become a crucial issue in various application fields of statistical inference such as clinical and epidemiological studies, insurance or actuarial science. Among many examples, one can mention the modeling of asthma, of epileptic seizures or of repeated warranty claims. In these settings, regression models are a valuable tool for predicting or identifying the causes which influence the number of such events occurring during a given time period. A natural way to measure the impact of covariates on the recurrent event process consists of estimating the conditional cumulative mean function. In this paper, our aim consists of developing both parametric and semiparametric inference for this conditional cumulative mean function. To that aim, we introduce new estimators and study their asymptotic behavior. We also discuss the data-driven way of calibrating the parameters involved in the estimation procedures.

In the literature, various authors have studied Cox regression models adapted to the recurrent event context. For example, in the absence of dependent death, Prentice et al. (1981) considered Cox-type regression models which allow the intensity of the recurrent event process to depend on the individual’s prior failure history through stratification. Allowing for independent censoring and time-dependent covariates, Andersen and Gill (1982) carried out Cox-type regression analysis for the intensity of the recurrent process which is assumed to be a time-transformed Poisson process. Andersen et al. (1993) also adopted modeling techniques based on the intensity process in the presence of censoring under a non-homogeneous Markov assumption. Lin et al. (2000) provided asymptotic distribution theory for the fitting of Cox-type marginal models without the Poisson assumption. Lawless and Nadeau (1995) proposed a semiparametric regression model where the conditional cumulative mean function is proportional to an unknown baseline function through a coefficient that depends parametrically on the covariates. More recently, Ghosh and Lin (2003) performed semiparametric regression with a scale-change model that formulates the marginal distributions of the recurrent event process and death as two joint accelerated failure time models while leaving the dependence structure unspecified.

The main advantage of these kinds of models stands in the simplicity of the regression function. But they unfortunately face the disadvantage (with respect to a purely nonparametric approach) to rely on strong modeling assumptions that may not hold in practice.

In this work, we first study a general parametric regression model for the recurrent event process. We then study a semiparametric generalization which relies on a single-
index assumption. We propose a new procedure to estimate both the index and the conditional cumulative mean regression function and provide a detailed asymptotic study of the proposed estimators. This single-index model can be seen as a compromise between a parametric approach and a nonparametric one. In particular, while allowing full flexibility, the nonparametric approach is known to fail when the number of covariates is high (greater than 3 in practice) which is the so-called “curse-of-dimensionality”. It turns out that single-index models rely on a dimension reduction assumption which allows to achieve better convergence rates and still ensures enough flexibility to be adapted to a large number of practical cases. This model can also be seen as a generalization of Cox regression model. Compared to uncensored single-index models adapted to mean-regression, see e.g. Ichimura (1993), in the specific setting of recurrent events, the presence of censoring usually deteriorates the quality of estimation in the tail of the distribution. Therefore, in our approach, we introduce a weight function designed to compensate the lack of information induced by censoring. The main novelty of our procedure stands in the fact that this weight function may be chosen using data-driven techniques.

The paper is organized as follows. In Section 2, we define the parametric and semi-parametric models and explain the general methodology. Asymptotic results are presented in Section 3. Simulation studies are carried out in Section 4 to investigate the performance of our methods for finite sample size. Technical results are postponed to the Appendix in Section 6.

2 Model assumptions and methodology

In this section, we present the general setting. Specifically, Section 2.1 introduces the different regression models. Section 2.2 presents the estimation procedures. They are based on a least-squares type criterion and on a rescaled process defined in Section 2.2.1 which permits to correct the impact of censoring.

2.1 Regression models for recurrent events

Consider the recurrent event process $N^*(t)$ which denotes the number of recurrent events occurring in the time interval $[0, t]$. This process can be seen as a piecewise constant function with jump only on $[0, D]$ where $D$ can be random. In clinical applications, this time $D$ may stand for the death time of a patient. For insurance applications, $D$ can represent the warranty length (which can be random if the client has the possibility of
breaking the contract) or the lifetime of the insured good. In this paper, we aim to infer on the cumulative conditional mean function given for $t \geq 0$ by

$$
\mu(t|z) = E[N^*(t)|Z = z],
$$

where $Z$ is a $d$–dimensional vector of covariates.

We now present the two different models for $\mu$ that are studied throughout this paper.

**Model 1 : parametric case.**

$$
\mu(t|z) = \mu_0(t, z; \theta_0),
$$

where $\theta_0 \in \Theta \subset \mathbb{R}^d$ is unknown and $\mu_0$ is a known function.

**Model 2 : semiparametric case.**

$$
\mu(t|z) = \mu_{\theta_0}(t, \theta'_0 z),
$$

where $\theta_0 \in \Theta \subset \mathbb{R}^d$, $\mu_{\theta}(t, u) = E[N^*(t)|\theta'Z = u]$ and the family of functions $\mathcal{F} = \{\mu_{\theta} : \theta \in \Theta\}$ is unknown. We impose that the first component of $\theta_0$ is 1 to identify this parameter. Another equivalent condition could consist of imposing that $\theta_0$ is of norm 1 for some given norm on $\mathbb{R}^d$.

The appealing feature of the first model stands in the simplicity of the regression function. However, like every parametric procedure, it relies on strong assumptions which have few chances to hold in practice. On the opposite, a fully nonparametric procedure requires fewer assumptions but suffers from the so-called “curse of dimensionality” when the number of covariates is high. Therefore, the second model appears as a good compromise between the parametric approach and the nonparametric one. Indeed it is more flexible than a fully parametric one but is not stroke by the curse of dimensionality since it relies on a dimension reduction assumption. Moreover, model 2 can be seen as a generalization of widely studied models. For example, the models $\mu(t|z) = \mu_0(t) \exp(\theta'_0 z)$ and $\mu(t|z) = \mu_0(t \exp(\theta'_0 z))$ where $\mu_0$ is an unknown baseline function correspond respectively to the popular Cox regression model and to the accelerated failure time model and are covered by model 2 as special cases.

One does not generally observe $N^*$ on the whole time interval $[0, D]$ because the random variable $D$ is subject to right-censoring. Let $C$ be a positive random variable standing for the censoring time. The observation time $T$ is then given by $T = D \wedge C$. Hence, instead of observing $N^*(t)$ for $t \in [0, D]$, one only observes $N(t) = N^*(t \wedge T)$ for $t \in [0, D]$. Letting $\delta = I(D \leq C)$, the observations consist of $n$ i.i.d. replications.
(\(T_i, \delta_i, Z_i, N_i(\cdot)\)) of \((T, \delta, Z, N(\cdot))\). Let us introduce the distribution functions of the observed variables in the censored data model:

\[
\begin{align*}
H(t) &= P(T \leq t), \\
F(t) &= P(D \leq t), \\
G(t) &= P(C \leq t).
\end{align*}
\]

We also define \(\tau_H = \inf(t : H(t) = 1)\) the right endpoint of the support of the random variable \(T\). In the sequel, we use some assumptions needed to identify these distribution functions.

**Assumption 1.** Assume that

\[
\begin{align*}
P(dN^*(C) \neq 0) &= 0, \\
P(D = C) &= 0.
\end{align*}
\]

This is a common assumption in the context of recurrent events which prevents us from ties between the occurrence times of death, censoring and recurrent events.

**Assumption 2.** Assume that

\[
\begin{align*}
C \perp \perp (N^*, D), \\
P(C \leq t | N^*, Z, D) &= P(C \leq t | N^*, D) \text{ for } t \in [0, \tau_H].
\end{align*}
\]

Assumption 2 holds in the particular case where \(C\) is independent of \((N^*, D, Z)\) but is more general since it does not require the independence between \(C\) and \(Z\). Similar kinds of assumptions are often considered in the literature on the Kaplan-Meier estimator for the survival distribution function, see e.g. Stute (1993).

### 2.2 Estimation procedure

#### 2.2.1 The rescaled process

One of the difficulties we face when estimating the conditional expectation of \(N^*\) is that the process \(N^*\) is not directly observed. Hence, the most natural criteria we may like to use can not be computed since they rely on \(N^*\). Therefore, we introduce a rescaled process \(Y\) designed to compensate the censoring effects. We define

\[
Y(t) = \int_0^t \frac{dN(s)}{1 - G(s-)}
\]

(2.3)
In the definition (2.3), the denominator is decreasing when \( s \) grows to infinity. This means that we allow more weight to the events we observe when \( s \) is large. This compensates the lack of observations due to censoring for \( s \) large. Under Assumptions 1 and 2, we have

\[
E[dN(s)|Z] = E[dN^*(s \land C)|Z] \\
= E[dN^*(s)I(s \leq C)|Z] \\
= E[dN^*(s)|Z](1 - G(s-))
\]

so that

\[
E[Y(t)|Z] = E[N^*(t)|Z].
\]

However, the rescaled process \( Y \) cannot be computed in practice since it relies on the distribution function \( G \) which is usually unknown. But the process \( Y(t) \) can be estimated for \( t \geq 0 \) by

\[
\hat{Y}(t) = \int_0^t \frac{dN(s)}{1 - \hat{G}(s-)},
\]

where \( \hat{G} \) denotes the Kaplan-Meier estimator of \( G \) given for \( t \geq 0 \) by

\[
\hat{G}(t) = 1 - \prod_{i: T_i \leq t} \left(1 - \frac{1}{\sum_{j=1}^n I(T_j \geq T_i)} \right)^{1-\delta_i}.
\]

### 2.2.2 The parametric case

Going back to the definition of the conditional expectation, it is quite natural to perform estimation of \( \theta_0 \) in the parametric model using minimization of a least-squares-type criterion. Once again, since \( N^* \) is unavailable, we consider a criterion based on the estimated rescaled process \( Y \).

Let \( w \) denote a measure such that \( w([0, \infty)) < \infty \) and define

\[
M_w(\theta, \mu_0) = \int_0^\tau H E[\mu_0(t, Z; \theta)^2] dw(t) - 2 \int_0^\tau H E[Y(t)\mu_0(t, Z; \theta)] dw(t).
\]

By definition of the conditional expectation, the true parameter value \( \theta_0 \) satisfies

\[
\theta_0 = \arg \min_{\theta \in \Theta} M_w(\theta, \mu_0).
\]

To estimate \( \theta_0 \), it is natural to replace the function \( M_w \) by an empirical version, that is

\[
M_{n,w}(\theta, \mu_0) = \frac{1}{n} \sum_{i=1}^n \int_0^{T_i(n)} \mu_0(t, Z_i; \theta)^2 dw(t) - \frac{2}{n} \sum_{i=1}^n \int_0^{T_i(n)} \hat{Y}_i(t) \mu_0(t, Z; \theta) dw(t),
\]
where $T_{(n)}$ is the greatest order statistics associated to the sample $T_1,\ldots,T_n$. Then we define an estimator of $\theta_0$ as

$$
\hat{\theta}(w) = \arg \min_{\theta \in \Theta} M_{n,w}(\theta, \mu_0).
$$

(2.6)

In the above definition, we emphasize the fact that this estimator depends on the choice of the measure $w$. This measure $w$ is an important feature of our procedure. First, in some situations, the statistician may wish to give more weight to some time intervals which are of higher importance. Moreover, the measure $w$ is also useful to control the rescaled process. Indeed, in equation (2.4), the denominator goes to zero when $s$ grows large and $w$ can be precisely designed to avoid the practical problems caused by these too small denominators. Therefore, the finite sample behavior of our estimation procedure strongly relies on a wise choice of the measure $w$.

In Section 3.3, we obtain an asymptotic representation of $\hat{\theta}(w)$ as a process indexed by $w$ which holds uniformly in $w \in \mathcal{W}$ where $\mathcal{W}$ is a set of measures in which the statistician plans to choose $w$. We then discuss in Section 3.5 the adaptive choice of $w$.

2.2.3 The semiparametric case

In the semiparametric case, the family of functions $\mu_\theta$ is unknown. However, the criterion used for the parametric case can be slightly modified to estimate $\theta_0$. We can write

$$
\hat{\theta}_0 = \arg \min_{\theta \in \Theta} M_w(\theta, \mu_\theta),
$$

where

$$
M_w(\theta, \mu_\theta) = \int_0^{T_H} E[\mu_\theta(t, \theta'Z)^2] \, dw(t) - 2 \int_0^{T_H} E[Y(t)\mu_\theta(t, \theta'Z)] \, dw(t).
$$

Using a family of nonparametric estimators $\hat{\mu}_\theta$ of $\mu_\theta$, we define the estimator of $\theta_0$ as

$$
\hat{\theta}(w) = \arg \min_{\theta \in \Theta} M_{n,w}(\theta, \hat{\mu}_\theta),
$$

(2.7)

where

$$
M_{n,w}(\theta, \hat{\mu}_\theta) = n^{-1} \sum_{i=1}^n \int_0^{T_{(n)}} \hat{\mu}_\theta(t, \theta'Z_i)^2 \, dw(t) - 2n^{-1} \sum_{i=1}^n \int_0^{T_{(n)}} \hat{Y}_i(t)\hat{\mu}_\theta(t, \theta'Z_i) \, dw(t).
$$

In Section 3.4, we derive an asymptotic representation of $\hat{\theta}(w)$ (see Theorem 3.3) regardless of the type of nonparametric estimators $\hat{\mu}_\theta$ used in the computation and provided these
nonparametric estimators satisfy a list of uniform convergence conditions. Nevertheless, let us give a precise example of $\hat{\mu}_\theta$ using kernel estimators. The convergence properties of this type of estimator is derived in Section 6.2.

Recall that (see Ghosh and Lin (2000) for instance)

$$\mu_\theta(t, u) = \int_{0}^{t} [1 - F_\theta(s - |u|)] dR_\theta(s|u),$$

where $dR_\theta(t|u) = E[dN^*(t)|D \geq t, \theta'Z = u]$ and where $F_\theta(s|u) = P(D \leq s|\theta'Z = u)$.

Using the identifiability Assumptions 1 and 2, this can be rewritten as

$$\mu_\theta(t, u) = \int_{0}^{t} \frac{E[dN(s)|\theta'Z = u]}{1 - G(s-)}.$$

(2.8)

We estimate the numerator in (2.8) using a kernel estimator and the denominator by the Kaplan-Meier estimator $\hat{G}$, leading to

$$\hat{\mu}_{\theta,h}(t, u) = \int_{0}^{t} \frac{\sum_{i=1}^{n} K \left( \frac{\theta'Z_i - u}{h} \right) dN_i(s)}{\sum_{j=1}^{n} K \left( \frac{\theta'Z_j - u}{h} \right) [1 - \hat{G}(s-)]},$$

(2.9)

where $K$ is a kernel function and $h$ a bandwidth sequence tending to zero. In Section 3.2, we list some conditions on $K$ and $h$. How to choose the bandwidth from the data in practice is considered at the end of Section 3.7.

3 Asymptotic results

In this section, we provide asymptotic properties for our estimators. In Section 3.1, we first expose and briefly discuss a list of technical assumptions on the model and on the different elements needed for the estimation procedures. In Section 3.2, we expose our main lemma, which is the keystone of our theoretical results. In the next two sections we give asymptotic representations of $\hat{\theta}(w)$ for the parametric and semiparametric models. We then discuss the adaptive choice of the measure $w$ in order to improve the performance of our procedure in Section 3.3. The variance of the limiting process is estimated in Section 3.6 and the choice of the bandwidth $h$ in (2.9) is highlighted in Section 3.7.

3.1 Exposition and discussion of assumptions

In order to obtain our asymptotics results, we first need to impose some conditions on different classes of functions.
Let us introduce some notations about the covering number. Let $\mathcal{F}$ be a class of functions with envelope $F$. Define, for a probability measure $Q$, the norm $\| \cdot \|_{p,Q}$ as the norm of $L^p(Q)$. The covering number of the class $\mathcal{F}$ for the measure $Q$ denoted by $N(\varepsilon, \mathcal{F}, \| \cdot \|_{p,Q})$ is the smaller number of $L^p(Q)$-balls of radius $\varepsilon$ needed to cover the set $\mathcal{F}$. The uniform covering number is defined as $N(\varepsilon, \mathcal{F}, \| \cdot \|_p) = \sup Q N(\varepsilon \|F\|_{p,Q}, \mathcal{F}, \| \cdot \|_{p,Q})$ where the supremum is taken over all probability measures. In what follows, we say that a class of functions $\mathcal{F}$ is a $\| \cdot \|_p$–VC–class of functions if there exists $\alpha$ and $C$ such that $N(\varepsilon, \mathcal{F}, \| \cdot \|_{p}) \leq C \varepsilon^{-\alpha}$.

A class of functions $\mathcal{F}$ is said to satisfy one of the following property if the corresponding condition holds.

**Property 1.** For a class of functions $\mathcal{F} = \{ f : (t, z) \in [0, \tau_H] \times \mathcal{Z} \mapsto f(t, z) \}$ and for any $\tau < \tau_H$, define
$$\mathcal{F}_\tau = \{ f(t, \cdot), t \in [0, \tau] \},$$
which is a set of functions defined on $\mathcal{Z}$. Then, for any $\tau < \tau_H$, $\mathcal{F}_\tau$ is a VC-class of functions.

**Property 2.** For a class of functions $\mathcal{F} = \{ f : (t, z) \in [0, \tau_H] \times \mathcal{Z} \mapsto f(t, z) \}$, the family of functions defined by $\{ (z, y) \mapsto \int_0^{\tau_H} y(t) f(t, z) dw(t), f \in \mathcal{F}, w \in \mathcal{W} \}$ is Glivenko Cantelli.

In Section 6.3.3 in the appendix, we give a general type of sufficient conditions to fulfill this property. It is easy to check that these technical assumptions are verified when the following conditions hold altogether:

- $\mathcal{F}$ is a class of polynomial functions $f(t, z)$ (with bounded coefficients),
- $dE[Y(t)] = g(t)dt$ for some polynomial function $g(t)$,
- the class of measures is of the form $\mathcal{W} = \{ w : dw(t) = W_0(t) d\tilde{w}(t) \}$ where $W_0(t)$ is a decreasing function (of order $t^{-k}$ for $k$ sufficiently high or exponential) and where $\tilde{w}$ belongs to a class of monotone positive uniformly bounded functions sufficiently small (for example, piecewise constant bounded functions with a finite number of jumps).

**Property 3.** Let $\mathcal{F} = \{ f_\theta : (t, z) \in [0, \tau_H] \times \mathcal{Z} \mapsto f_\theta(t, z), \theta \in \Theta \}$ be a family of functions indexed by $\theta$. For any $f_{\theta_1}, f_{\theta_2} \in \mathcal{F}$ and $z \in \mathcal{Z}$, we have
$$\sup_{w \in \mathcal{W}} \int_0^{\tau_H} \| f_{\theta_1}(t, z) - f_{\theta_2}(t, z) \| dw(t) \leq C \| \theta_1 - \theta_2 \|,$$
where $C$ is a positive constant.

We now introduce the assumptions needed to derive the asymptotic normality of $\hat{\theta}$ in the parametric and semiparametric models.

**Assumptions for the parametric model.**

In the estimation procedures, we consider integrated versions of the rescaled process with respect to a measure $w$ belonging to a class of measures $W$. Detailed comments on this family and its role in the statistical procedure are discussed in Section 3.5. We need the following assumption for this class of measures.

**Assumption 3.** Assume there exists some probability measure $w_0$ and a positive constant $C_0$ such that, for any $w \in W$,

$$\int_{\tau_H} dw(s) \leq C_0 W_0(t),$$

where $W_0(t) = \int_{\tau_H} dw_0(s)$ can be written as

$$W_0(t) = W_1(t)W_2(t)$$

where $W_1$ and $W_2$ are two positive and non-increasing functions satisfying

1. $\int_{0}^{\tau_H} W_1^2(t)[1 - F(t-)]^{-1}[1 - G(t-)]^{-2}dG(t) < \infty$,
2. $\int_{0}^{\tau_H} W_2(t)E[dN^*(t)] < \infty$,
3. $\lim_{t \to \tau_H} W_2(t) = 0$.

In particular, Assumption 3 holds when all the measures $w$ have their support included in a common compact subspace strictly included in $[0, \tau_H]$. On the other hand, since the function $W_1$ controls $1 - \hat{G}(s-)$ in $\hat{Y}(s)$ for $s$ in the vicinity of the tail of the distribution, Assumption 3 also allows to consider measures $w$ which are supported in the whole interval $[0, \tau_H]$. Taking $W_1(t) = (1 - H(t))^{1/2}(1 - G(t))^{\varepsilon}$ for some $\varepsilon > 0$ would be sufficient to obtain (1). Moreover, in the case where $\tau_H = \infty$, if we suppose that, for $\alpha > 0$, we have $E[N^*(t)] \sim \alpha t$ when $t \to \infty$, we could take for example $W_2(t) = t^{-\beta}$ for $\beta > 1$ to fulfill (2) and (3).

We also need the following Hölder condition on the process $N$. This is a technical assumption used in the proof of our main lemma.
Assumption 4. Suppose that

\[ E \left( \sup_{t \leq \tau, t' \leq \tau} \frac{|N(t) - N(t')|}{|t - t'|^{\alpha}} \right) < \infty. \]

Let \( \nabla_{\theta} \mu_0(s, z; \theta_1) \) (resp. \( \nabla^2_{\theta} \mu_0(s, z; \theta_1) \)) denote the vector of partial derivatives (resp. the Hessian matrix) of \( \mu_0(s, z; \theta) \) with respect to all the components of \( \theta \) evaluated at \( \theta_1 \). The following assumption can be understood as a regularity assumption on the regression model.

Assumption 5. Assume that, for all \( w \in W \), the matrix

\[ \Sigma_{w, \cdot} = \int_0^T E[\nabla_{\theta} \mu_0(t, Z, \theta_0) \nabla_{\theta} \mu_0(t, Z, \theta_0)'] dw(t) \]

is invertible. Moreover, assume that the classes of functions \( \{ \mu_0(\cdot, \cdot; \theta), \theta \in \Theta \} \), \( \{ \nabla_{\theta} \mu_0(\cdot, \cdot; \theta), \theta \in \Theta \} \) and \( \{ \nabla^2_{\theta} \mu_0(\cdot, \cdot; \theta), \theta \in \Theta \} \) satisfy Properties 1, 2 and 3.

Additional assumptions for the semiparametric model.

The following assumption is similar to Assumption 3. Here, \( \nabla_{\theta} \mu_{\theta_1}(s, z) \) (resp. \( \nabla^2_{\theta} \mu_{\theta_1}(s, z) \)) denotes the vector of partial derivatives (resp. the Hessian matrix) of \( \mu_\theta(s, \theta'z) \) with respect to all the components of \( \theta \) evaluated at \( \theta_1 \). Note that the gradient vector \( \nabla_{\theta} \mu_{\theta_1}(s, z) \) does not only depend on \( \theta'z \) but also depends on the whole vector \( z \). We give an explicit expression of this gradient in Lemma 6.5.

Assumption 6. Assume that, for all \( w \in W \), the matrix

\[ \Sigma_{w, sp} = \int_0^T E[\nabla_{\theta} \mu_0(t, Z) \nabla_{\theta} \mu_0(t, Z)'] dw(t) \]

is invertible. Moreover, assume that the classes of functions \( \{ \mu_\theta(\cdot, \cdot; \theta), \theta \in \Theta \} \), \( \{ \nabla_{\theta} \mu_\theta(\cdot, \cdot; \theta), \theta \in \Theta \} \) and \( \{ \nabla^2_{\theta} \mu_\theta(\cdot, \cdot; \theta), \theta \in \Theta \} \) satisfy Properties 1, 2 and 3.

As announced, we need uniform convergence properties for the nonparametric estimators \( \hat{\mu}_\theta \).

Assumption 7. Define \( \bar{\mu}_\theta(t, u) = \sup(\mu_\theta(t, u), 1) \).

(1) Assume that

\[
\begin{align*}
\sup_{t \leq T(n), \theta \in \Theta, z \in Z} \left| \frac{\hat{\mu}_\theta(t, \theta'z) - \mu_\theta(t, \theta'z)}{\mu_{\theta_0}(t, \theta'_0z)^{\lambda_1 + \lambda_2}} \right| &= o_p(1), \\
\sup_{t \leq T(n), \theta \in \Theta, z \in Z} \left| \frac{\nabla_{\theta} \hat{\mu}_\theta(t, z) - \nabla_{\theta} \mu_\theta(t, z)}{\mu_{\theta_0}(t, \theta'_0z)^{\lambda_1 + \lambda_2}} \right| &= o_p(1), \\
\sup_{t \leq T(n), \theta \in \Theta, z \in Z} \left| \frac{\nabla^2_{\theta} \hat{\mu}_\theta(t, z) - \nabla^2_{\theta} \mu_\theta(t, z)}{\mu_{\theta_0}(t, \theta'_0z)^{\lambda_1 + \lambda_2}} \right| &= o_p(1),
\end{align*}
\]

where \( \lambda_1, \lambda_2 \) are such that \( \lambda_1 + \lambda_2 \geq 1 \).
(2) Assume also that
\[ \sup_{t \leq T(n), z \in Z} |\hat{\mu}_{\theta_0}(t, \theta'_0 z) - \mu_{\theta_0}(t, \theta'_0 z)| = O_P(\varepsilon_n), \]
\[ \sup_{t \leq T(n), z \in Z} \|\nabla_{\theta_0} \hat{\mu}_{\theta_0}(t, z) - \nabla_{\theta_0} \mu_{\theta_0}(t, z)\| = O_P(\varepsilon'_n), \]
where \( \varepsilon_n, \varepsilon'_n = o_P(n^{-1/2}) \).

**Assumption 8.** Assume that
\[ \sup_{z \in Z} \int_{0}^{T} (E[N^*(t)|Z = z])^{2(\lambda_1 + \lambda_2)} dw(t) < \infty \]
where \( \lambda_1, \lambda_2 \) were defined in Assumption 7.

The following assumption is essential to the empirical process theory used in our proofs. We assume that the nonparametric estimators and \( \mu_{\theta_0} \) belong to some Donsker classes of functions.

**Assumption 9.** Assume that there exists some Donsker classes of functions \( \mathcal{G} \) and \( \mathcal{H} \) such that for all \( w \in \mathcal{W} \),
\[ (z, y) \mapsto \int_{0}^{T} (\mu_{\theta_0}(t, \theta'_0 z) - y(t))\nabla_{\theta_0} \mu_{\theta_0}(t, z)dw(t) \in \mathcal{G}, \]
\[ z \mapsto \int_{0}^{T} \mu_{\theta_0}(t, \theta'_0 z)\nabla_{\theta_0} \mu_{\theta_0}(t, z)dw(t) \in \mathcal{H}. \]

Moreover, assume that, almost surely for \( n \) large enough,
\[ (z, y) \mapsto \int_{0}^{T} (\mu_{\theta_0}(t, \theta'_0 z) - y(t))\nabla_{\theta_0} \hat{\mu}_{\theta_0}(t, z)dw(t) \in \mathcal{G}, \]
\[ z \mapsto \int_{0}^{T} \hat{\mu}_{\theta_0}(t, \theta'_0 z)\nabla_{\theta_0} \mu_{\theta_0}(t, z)dw(t) \in \mathcal{H}. \]

To give examples of such kind of classes, consider \( \mathcal{F} \) and \( \mathcal{W} \) as defined in the discussion following Property 1 and suppose, in addition, that the functions \( (t, u) \to W_0(t)f(t, u) \) for \( f \in \mathcal{F} \) (\( f \) is defined on \( \mathbb{R}^2 \) since \( \theta'_0 z \in \mathbb{R} \)) are twice continuously differentiable with bounded derivatives up to order 2. It follows from the results of Section 6.3.3 and from the decomposition of the gradient vector \( \nabla_{\theta_0} \mu_{\theta_0}(t, z) \) obtained in Lemma 5.3 that we can consider \( \mathcal{H} = \mathcal{G} = \mathcal{F}' + z\mathcal{F}' \) where \( \mathcal{F}' = \{(u, y) \to \int_{0}^{T} (f_1(t, u) - y(t))f_2(t, u)dw(t), w \in \mathcal{W}, f_1, f_2 \in \mathcal{F}\} \).
3.2 The main lemma

From a theoretical viewpoint, the main issue stands in studying the difference between $Y$ and its estimated version. The following lemma provides an asymptotic representation for a class of empirical sums in which the process $\hat{Y}$ is involved.

**Lemma 3.1.** Let $\mathcal{F}$ be a class of functions with bounded envelope $\Phi$ satisfying Property 4 and assume that Assumptions 3 and 4 hold. Define, for any function $f \in \mathcal{F}$,

$$ S_n(f, w) = \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{T_i} Y_i(t) f(t, Z_i) dw(t) $$

and

$$ \hat{S}_n(f, w) = \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{T_i(n)} \hat{Y}_i(t) f(t, Z_i) dw(t). $$

(1) Assume that $\sup_{w \in W} E[S_n(\Phi, w)] < \infty$. Then, for all $f \in \mathcal{F}$,

$$ \hat{S}_n(f, w) - S_n(f, w) = \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{T_i} \eta_{k-}(T_i, \delta_i) E[f(t, Z) d\mu(s|Z)] dw(t) + R_n(f, w), $$

where

$$ \eta_k(T, \delta) = \frac{(1 - \delta) I(T \leq t) - \int_{0}^{t} I(T \geq s) dG(s)}{1 - H(T^-)} - \int_{0}^{t} \frac{I(T \geq s) dG(s)}{[1 - H(s^-)][1 - G(s^-)]} $$

and where

$$ \sup_{w \in W, f \in \mathcal{F}} |R_n(f, w)| = o_P(n^{-1/2}). $$

Moreover, if the measures $w$ are all supported in $[0, \tau]$ for some $\tau < \tau_H$, then

$$ \sup_{w \in W, f \in \mathcal{F}} |R_n(f, w)| = O_P(n^{-1} \log n). $$

(2) If $\hat{f}$ denotes a family of nonparametric estimators of functions $f \in \mathcal{F}$ satisfying

$$ \sup_{f \in \mathcal{F}} \|\hat{f} - f\|_\infty = o_P(1), $$

then

$$ \sup_{w \in W} |\hat{S}_n(\hat{f}, w) - S_n(\hat{f}, w)| = o_P(n^{-1/2}). $$

Moreover, if the measures $w$ are all supported in $[0, \tau]$ for some $\tau < \tau_H$, then

$$ \sup_{w \in W} |\hat{S}_n(\hat{f}, w) - S_n(\hat{f}, w)| = O_P(n^{-1} \log n). $$

The proof is postponed to Section 5.3. With the estimated rescaled process $\hat{Y}$ at hand, we can now propose $M$–estimation procedures to estimate the regression function in both the parametric and semiparametric cases.
3.3 Asymptotic normality of \( \hat{\theta} \) in the parametric case

Let \( \Rightarrow \) denote the weak convergence.

**Theorem 3.2.** Assume that (2.7) holds. Under Assumptions 1 to 5, the estimator in (2.7) admits the following asymptotic representation

\[
\hat{\theta}(w) - \theta_0 = \Sigma_{w,p}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( \int_{0}^{T_n} [Y_i(t) - \mu_0(t, Z_i; \theta_0)] \nabla_\theta \mu_0(t, Z_i; \theta_0) dw(t) \right) 
+ \int_{0}^{T_n} \int_{0}^{t} \nabla_{\theta} \mu_0(t, Z_i; \theta_0) d\mu_0(s, Z_i; \theta_0) dw(t) \right\} + R_n(w),
\]

where \( \sup_{w \in W} |R_n(w)| = o_P(n^{-1/2}) \). As a consequence, for any \( w \in W \),

\[
\sqrt{n} (\hat{\theta}(w) - \theta_0) \Rightarrow \mathcal{N}(0, V_{w,p}),
\]

where \( V_{w,p} = \Sigma_{w,p} \Delta_{w,p} \Sigma_{w,p}^{-1} \) and \( \Delta_{w,p} \) is the covariance matrix of each term of the i.i.d. sum in the asymptotic expansion.

**Proof.** Write

\[
M_{n,w}(\theta, \mu_0) = -2 \hat{S}_n(\mu_0(\cdot, \cdot; \theta), w) + n^{-1} \sum_{i=1}^{n} \int_{0}^{T_n} \mu_0(t, Z_i; \theta)^2 dw(t).
\]

Then, use the asymptotic representation of Lemma 3.1. Uniform consistency of \( \hat{\theta}(w) \) follows from the uniform convergence of \( M_{n,w}(\theta, \mu_0) \) which is obtained from Properties 2 and 3 for the classes of functions \( \{\mu_0(\cdot, \cdot; \theta), \theta \in \Theta\} \) and \( \{\nabla_\theta \mu_0(\cdot, \cdot; \theta), \theta \in \Theta\} \) (see Assumption 3).

To obtain the uniform CLT property for \( \hat{\theta}(w) \), use a Taylor expansion of \( \nabla_\theta M_{n,w}(\theta, \mu_0) \) around \( \theta_0 \):

\[
\nabla_\theta M_{n,w}(\hat{\theta}, \mu_0) = \nabla_\theta M_{n,w}(\theta_0, \mu_0) + \nabla^2_\theta M_{n,w}(\hat{\theta}, \mu_0)(\hat{\theta} - \theta_0),
\]

for some \( \tilde{\theta} \) between \( \hat{\theta} \) and \( \theta_0 \). The left-hand side of (3.1) is zero by definition of \( \hat{\theta} \). Moreover, the matrix \( \nabla^2_\theta M_{n,w}(\tilde{\theta}, \mu_0) \) is almost surely invertible for \( n \) large enough under Assumption 4 since \( \hat{\theta} \) (and consequently \( \tilde{\theta} \)) tends to \( \theta_0 \) almost surely. This leads to

\[
\hat{\theta} - \theta_0 = -\nabla^2_\theta M_{n,w}(\tilde{\theta}, \mu_0) \nabla_\theta M_{n,w}(\theta_0, \mu_0).
\]

Write

\[
\nabla^2_\theta M_{n,w}(\hat{\theta}, \mu_0) = -2 \left[ \hat{S}_n(\nabla^2_\theta \mu_0(\cdot, \cdot; \hat{\theta}), w) - \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{T_n} \left( \nabla_\theta \mu_0(t, Z_i; \hat{\theta}) \nabla_\theta \mu_0(t, Z_i; \hat{\theta})' + \mu_0(t, Z_i; \hat{\theta}) \nabla^2_\theta \mu_0(t, Z_i; \hat{\theta}) \right) dw(t) \right] + R_n(\theta, w)
\]

14
where \( R_n(\theta, w) \) comes from the change in the integration bounds of \([0, T(n)]\) by \([0, \tau_H]\) and can be seen to tend uniformly to zero from Lebesgue’s dominated convergence since the term inside the integral is bounded. From Lemma 3.1, the almost sure convergence of \( \tilde{\theta} \) and the fact that \( \{\nabla^2 \mu_0(\cdot, \cdot, \theta) \in \Theta\} \) satisfies Property 3 (see Assumption 3), we get that \( \tilde{S}_n(\nabla^2 \mu_0(\cdot, \cdot, \tilde{\theta}), w) \) converges to \( \int_0^{\tau_H} E[Y(t)\nabla^2 \mu_0(t, Z; \theta_0)]dw(t) \) uniformly in \( w \). The second part converges uniformly to its expectation over \( \Theta \) as a consequence of the Glivenko-Cantelli property of classes of functions satisfying Property 3. This shows that

\[
\sup_w |\nabla^2 M_{n,w}^{-1}(\tilde{\theta}, \mu_0) - \nabla^2 M_{w}^{-1}(\theta_0, \mu_0)| = o_P(1).
\]

On the other hand, we write

\[
\nabla_{\theta} M_{n,w}(\theta_0, \mu_0) = -2 \left[ \tilde{S}_n(\nabla_{\theta} \mu_0(\cdot, \cdot; \theta_0), w) - \frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau_H} \mu_0(t, Z; \theta_0) \nabla_{\theta} \mu_0(t, Z; \theta_0) dw(t) \right]
\]

\[
+ \frac{2}{n} \sum_{i=1}^{n} \int_{T(n)}^{\tau_H} \mu_0(t, Z; \theta_0) \nabla_{\theta} \mu_0(t, Z; \theta_0) dw(t).
\]

Using Lebesgue’s dominated convergence theorem, the last term tends uniformly to zero at a \( n^{-1/2} \) rate. Finally, the asymptotic representation follows from Lemma 3.1.

\[\Box\]

### 3.4 Asymptotic normality of \( \hat{\theta} \) in the semiparametric case

**Theorem 3.3.** Assume that (2.4) holds. Under Assumptions 4 to 4 and 6 to 9, the estimator in (2.7) admits the following asymptotic representation

\[\hat{\theta}(w) - \theta_0 = \Sigma_{w,sp}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( \int_0^{\tau_H} [Y_i(t) - \mu_{\theta_0}(t, \theta'_0 Z_i)] \nabla_{\theta} \mu_{\theta_0}(t, Z_i) dw(t) \right) + \int_0^{\tau_H} \int_0^t \eta_n(s - (T_i, \delta_i)) \mu_{\theta_0}(s, \theta'_0 Z) dw(t) \right\} + R_n(w),\]

where \( \sup_{w \in W} |R_n(w)| = o_P(n^{-1/2}) \). As a consequence, for any \( w \in W \),

\[\sqrt{n}(\hat{\theta}(w) - \theta_0) \Rightarrow N(0, V_{w,sp}),\]

where \( V_{w,sp} = \Sigma_{w,sp}^{-1} \Delta_{w,sp} \Sigma_{w,sp}^{-1} \) and \( \Delta_{w,sp} \) is the covariance matrix of each term of the i.i.d. sum in the asymptotic expansion.

**Proof.** The consistency of the preliminary estimator can be proved in the same way as in the proof of Theorem 3.2, using now the second part of Lemma 3.1 and the uniform consistency of \( \hat{\mu}_\theta \) (Assumption 4). Asymptotic normality comes from the fact that

\[\hat{\theta} - \theta_0 = - \nabla^2 M_{n,w}^{-1}(\hat{\theta}, \hat{\mu}_\theta) \nabla_{\theta} M_{n,w}(\theta_0, \hat{\mu}_\theta).\]
The fact that
\[
\sup_w \| \nabla_\theta^2 M_{n,w}^{-1}(\bar{\theta}, \hat{\mu}_0) - \nabla_\theta^2 M_w^{-1}(\theta_0, \mu_{\theta_0}) \| = o_P(1)
\]
can be shown in the same way as in the proof of Theorem 3.2 using now the second part of Lemma 3.1. The big issue consists of proving the asymptotic representation of \(\nabla_\theta M_{n,w}(\theta_0, \mu_{\theta_0})\). Write
\[
\nabla_\theta M_{n,w}(\theta_0, \hat{\mu}_{\theta_0}) = -2 \left[ \hat{S}_n(\nabla_\theta\hat{\mu}_{\theta_0}(\cdot, \theta_0'), w) - \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{T(n)} \hat{\mu}_{\theta_0}(t, \theta_0' Z_i) \nabla_\theta \hat{\mu}_{\theta_0}(t, Z_i) dw(t) \right].
\]
Using the second part of Lemma 3.1, this can be rewritten as
\[
\nabla_\theta M_{n,w}(\theta_0, \hat{\mu}_{\theta_0}) = \nabla_\theta M_{n,w}(\theta_0, \mu_{\theta_0})
\]
\[
- \frac{2}{n} \sum_{i=1}^{n} \int_{0}^{\tau_H} \hat{\mu}_{\theta_0}(t, \theta_0' Z_i) \nabla_\theta \hat{\mu}_{\theta_0}(t, \theta_0' Z_i) Y_i(t) \frac{\nabla_\theta \mu_{\theta_0}(t, Z_i) - \nabla_\theta \hat{\mu}_{\theta_0}(t, Z_i)}{\hat{\mu}_{\theta_0}(t, \theta_0' Z_i)} dw(t) + \sum_{i=1}^{n} \int_{0}^{\tau_H} \frac{\hat{\mu}_{\theta_0}(t, \theta_0' Z_i) - \mu_{\theta_0}(t, \theta_0' Z_i)}{\hat{\mu}_{\theta_0}(t, \theta_0' Z_i)} \nabla_\theta \hat{\mu}_{\theta_0}(t, Z_i) dw(t) + \sum_{i=1}^{n} \int_{0}^{\tau_H} \frac{\hat{\mu}_{\theta_0}(t, \theta_0' Z_i) - \mu_{\theta_0}(t, \theta_0' Z_i)}{\hat{\mu}_{\theta_0}(t, \theta_0' Z_i)} (\nabla_\theta \mu_{\theta_0}(t, Z_i) - \nabla_\theta \hat{\mu}_{\theta_0}(t, Z_i)) \frac{\nabla_\theta \mu_{\theta_0}(t, Z_i) - \nabla_\theta \hat{\mu}_{\theta_0}(t, Z_i)}{\hat{\mu}_{\theta_0}(t, \theta_0' Z_i)} dw(t)
\]
\[
+ R_{4n}(w)
\]
where \(R_{4n}(w)\) comes from Lemma 3.1 and the change in the integration bound of \([0, T(n)]\) by \([0, \tau_H]\). Using the same arguments as in the proof of Theorem 3.2, we deduce that
\[
\sup_w \| R_{4n}(w) \| = o_P(n^{-1/2}).
\]
Using the uniform convergence rates of \(\hat{\mu}_{\theta_0}\) and \(\nabla_\theta \hat{\mu}_{\theta_0}\), we get straightforwardly that \(\sup_w \| R_{3n}(w) \| = o_P(n^{-1/2})\). Using the uniform convergence of \(\nabla_\theta \hat{\mu}_{\theta_0}\), we see that the term \(R_{1n}\) can be decomposed into
\[
R_{1n}(w) = n^{-1} \sum_{i=1}^{n} [f_w(Z_i, Y_i) - f_{n,w}(Z_i, Y_i)]
\]
where \(f_w\) and \(f_{n,w}\) both belong (almost surely for \(n\) large enough) to the class \(\mathcal{G}\) defined in Assumption 3.2 and with \(\sup_w \| f_w - f_{n,w} \|_\infty \to 0\) a.s. Therefore, using the asymptotic equicontinuity of the Donsker class \(\mathcal{G}\) (see e.g. Section 2.1.2 in Van der Vaart and Wellner (1996)), this shows that
\[
\sup_w \| R_{1n}(w) - \int [f_w(z, y) - f_{n,w}(z, y)] dP_{Z,Y}(z, y) \| = o_P(n^{-1/2}).
\]
Moreover, it is clear that \( \int [f_w(z, y) - f_{n,w}(z, y)] dP_{Z,Y}(z, y) = 0 \) using the fact that 
\[ \nabla_{\theta} \mu_{\theta_0}(t, z) - \nabla_{\theta} \hat{\mu}_{\theta_0}(t, z) \] 
is a function of \( z \) only and that \( E[\mu_{\theta_0}(t, \theta'_0 Z_i) - Y_i(t)|Z_i] = 0 \).

The term \( R_{2n}(w) \) can be handled in the same way using now the Donsker class \( \mathcal{H} \) in Assumption 9, observing that \( (\hat{\mu}_{\theta_0}(t, \theta'_0 z) - \mu_{\theta_0}(t, \theta'_0 z)) \) is a function of \( \theta'_0 z \) only and getting from Lemma 6.5 that \( E[\nabla_{\theta} \mu_{\theta_0}(t, Z)|\theta'_0 Z] = 0. \)

### 3.5 Adaptive choice of \( w \)

The representations of Theorems 3.2 and 3.3 hold uniformly in \( w \in \mathcal{W} \). Therefore, the asymptotic normality of our estimators of the parameter remains valid if we replace \( w \) by a data-driven measure \( \hat{w} \) that converges to a specific optimal measure \( w_0 \). We give some indications on a method to obtain such kind of data-driven measure adapted to our estimation problem.

The empirical measure \( \hat{w} \) will be defined as the minimizer of some criterion. Since it is generally impossible to perform minimization on the functional space \( \mathcal{W} \), we minimize over a growing subset \( \mathcal{W}_n \). The adaptive procedure we propose consists of first estimating the asymptotic covariance matrix \( V_{w,sp} \) (or \( V_{w,p} \) in the parametric case) for any \( w \in \mathcal{W}_n \). From the asymptotic variance estimators, we derive the estimation of the mean squared error \( E[\|\hat{\theta}(w) - \theta_0\|^2] \). We then take \( \hat{w} \) as the element of \( \mathcal{W}_n \) such that the estimated mean squared error is minimal over \( \mathcal{W}_n \). Then, our final estimator is

\[ \hat{\theta} = \hat{\theta}(\hat{w}). \]

The uniform convergence of the remainder term in the representations of Theorems 3.2 and 3.3 provides the asymptotic normality of \( \hat{\theta} \) in the case where \( \Sigma_{\hat{w}} \to \Sigma_{w_0} \) a.s. for some \( w_0 \in \mathcal{W} \).

### 3.6 Estimation of the variance

We show how to estimate the variance in the representation of Theorem 3.3 and we propose an estimator of the mean squared error of \( \theta_0 \). Denote by \( W_{n,w} \) the term between brackets in the representation of Theorem 3.3 so that

\[ \hat{\theta}(w) - \theta_0 = \Sigma_{w,sp}^{-1} W_{n,w} + R_n(w), \]

where \( \sup_{w \in \mathcal{W}} |R_n(w)| = o_P(n^{-1/2}) \). The quantity \( W_{n,w} \) can be estimated in the following way

\[ \hat{W}_{n,w} = \frac{1}{n} \sum_{i=1}^{n} \hat{w}(\delta_i, Z_i, T_i, Y_i; w), \]
where
\[
\hat{\psi}(\delta, Z, T, Y; w) = \int_0^{T(n)} (Y(t) - \hat{\mu}(t, \hat{\theta}'Z)) \nabla_\theta \hat{\mu}(t, Z)dw(t)
\]
\[
+ \int_0^{T(n)} \int_0^t \hat{\eta}_s(T, \delta) n^{-1} \sum_{i=1}^n \left( \nabla_\theta \hat{\mu}(t, \hat{\theta}'Z_i) \right) dw(t),
\]
and \(\hat{\eta}(T, \delta) = \frac{(1 - \delta)I(T \leq t)}{1 - \hat{H}(T -)} - \int_0^t \frac{I(T \geq s) d\hat{G}(s)}{(1 - \hat{H}(s -))(1 - G(s -))}\)
and \(\hat{H}\) is the empirical estimator of \(H\). Therefore, the quantity \(\Delta_{w,sp}\) can be estimated by
\[
\hat{\Delta}_{w,sp} = \frac{1}{n} \sum_{i=1}^n \left( \hat{\psi}(\delta, Z, T, Y; w) - \frac{1}{n} \sum_{i=1}^n \hat{\psi}(\delta, Z, T, Y; w) \right)^2,
\]
where \(\otimes 2\) denotes the product of the matrix with its transpose. To consistently estimate \(\Sigma_{w,sp}\), we use
\[
\hat{\Sigma}_{w,sp} = \frac{1}{n} \sum_{i=1}^n \int_0^{T(n)} \nabla_\theta \hat{\mu}_\theta(t, Z_i) \nabla_\theta \hat{\mu}_\theta(t, Z_i)'.
\]
A consistent estimator of \(V_{w,sp}\) can then be computed from \(\hat{V}_{w,sp} = \hat{\Sigma}_{w,sp}^{-1} \hat{\Delta}_{w,sp} \hat{\Sigma}_{w,sp}^{-1}\).
Finally, we take \(\hat{E}_T^2 = \hat{W}_{n,w} \hat{\Sigma}_{w,sp}^{-1} \hat{\Sigma}_{w,sp}^{-1} \hat{W}_{n,w}\) as mean squared error estimate.

### 3.7 Estimation of the nonparametric part

In the semiparametric model, estimation of the finite dimensional parameter \(\theta_0\) is only the first step of the method. With our estimator \(\hat{\theta}\) at hand, we wish to estimate the conditional mean function \(\mu(t|z)\). Different strategies can be proposed to perform this estimation. For this final estimator, there is no theoretical need to use the same kind of nonparametric estimator as in the computation of \(\hat{\theta}\). Proposition 3.4 below states that, under some convergence assumptions for the nonparametric estimator used in this second step, the asymptotic behavior of the final semiparametric estimator of \(\mu\) is identical to the asymptotic behavior of a purely nonparametric estimator in the case where \(\theta_0\) is exactly known.

**Proposition 3.4.** Let \(\Theta^*\) be some neighborhood of \(\theta_0\), and let \(T\) be a set on which
\[
\sup_{\theta \in \Theta^*, t \in T, z \in Z} \| \nabla_\theta \mu_\theta(t, z) \| < \infty.
\]
Let \(\hat{\mu}_\theta\) be a family of nonparametric estimators of \(\mu_\theta\) satisfying the assumption
\[
\sup_{\theta \in \Theta^*, t \in T, z \in Z} \| \nabla_\theta \hat{\mu}_\theta(t, z) - \nabla_\theta \mu_\theta(t, z) \| = O_P(1).
\]
Then, we have

\[
\sup_{t \in T, z \in Z} |\hat{\mu}(t, \hat{\theta} z) - \hat{\mu}_{\theta_0}(t, \theta_0 z)| = O_P(n^{-1/2}).
\]

**Proof.** This is a direct consequence of a Taylor expansion of \(\hat{\mu}\) around \(\theta_0\). From Theorem 3.3 we have \(\hat{\theta} - \theta_0 = O_P(n^{-1/2})\). Then, the boundedness of \(\nabla_\theta \mu_{\theta_0}(t, z)\) and the uniform convergence in assumption (3.2) give the result. \(\square\)

In the kernel estimator example of equation (2.9), a crucial issue stands in the choice of the bandwidth which strongly influences the performance of the nonparametric estimation. A first method to define our final estimator of \(\mu_{\theta_0}\) consists of using an arbitrary sequence of bandwidth \(h\) to compute \(\hat{\theta}\), then of using cross-validation techniques to select a bandwidth \(\hat{h}\). The final estimator is finally set as \(\hat{\mu}_{\hat{\theta}, \hat{h}}(t, \theta_0 z)\). However, it seems more appealing to us to define a procedure which can be seen as an extension of the adaptive choice of bandwidth proposed by H"ardle et al. (1993) and Delecroix et al. (2006). An interesting feature of this technique is that it selects an adaptive bandwidth \(\hat{h}\) and a direction \(\hat{\theta}\) at the same time. Indeed, define

\[
(\hat{\theta}, \hat{h}) = \arg \min_{\theta \in \Theta, h \in \mathcal{H}} M_{n, w}(\theta, \hat{\mu}_{\theta, h}). \tag{3.3}
\]

The uniform in bandwidth consistency of the kernel estimators we use (see Section 3.2) ensures us that \(\hat{\theta}\) has the same asymptotic properties as in Theorem 3.3. On the other hand, Proposition 3.5 below shows that the adaptive bandwidth \(\hat{h}\) is asymptotically equivalent to the bandwidth we could obtain using a classical cross-validation technique in the case where the parameter \(\theta_0\) is exactly known.

**Proposition 3.5.** For some positive constants \(a, c\) and \(C\), let \(\mathcal{H} = [cn^{-a}, Cn^{-a}]\) be a set of bandwidth satisfying Assumption 7 and let

\[
h_0 = \arg \min_{h \in \mathcal{H}} M_{n, w}(\theta_0, \hat{\mu}_{\theta_0, h}).
\]

Under the assumptions of Theorem 3.3 and provided that \(\sup_{h \in \mathcal{H}, t \in \mathbb{R}^+, z \in \mathbb{Z}} |\hat{\mu}_{\theta, h}(t, \theta' z) - \mu_{\theta, h}(t, \theta' z)| = o_P(1)\), we have

\[
\hat{h}/h_0 \to 1 \ \text{a.s.}
\]

**Proof.** Define \(\phi(h/h_0) = M_{n, w}(\theta_0, \hat{\mu}_{\theta_0, h})\) and \(\phi_n(h/h_0) = \arg \min_{\theta \in \Theta} M_{n, w}(\theta, \hat{\mu}_{\theta, h})\). By definition of \(h_0\) and \(\hat{h}\) we have \(\arg \min_{s \in [c, C]} \phi(s) = 1\) and \(\hat{h}/h_0 = \arg \min_{s \in [c, C]} \phi_n(s)\).
Now write
\[
\phi_n(s) = \phi(s) - \frac{2}{n} \sum_{i=1}^{n} \int_{0}^{T_H} Y_i(t)(\hat{\mu}_{\theta,sha}(t, \theta'Z_i) - \hat{\mu}_{\theta,h}(t, \theta'Z_i))dw(t)
\]
\[+ \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{T_H} (\hat{\mu}_{\theta,sha}(t, \theta'Z_i)^2 - \hat{\mu}_{\theta,h}(t, \theta'Z_i)^2)dw(t) + M_{n,w}(\theta, \hat{\mu}_{\theta,h}).\]

Using Lemma 5.1 and the uniform in bandwidth consistency of \(\hat{\mu}_{\theta,h}\), the second and third terms in the decomposition tend to zero uniformly in \(s\). On the other hand, the last term does not depend on \(s\). This shows that \(\hat{h}/h_0 \to 1\) a.s. \(\square\)

4 Simulations

We present here some empirical evidence of the good behavior of our semiparametric estimation procedure for finite sample sizes.

In our simulation study, we consider the case where, conditionally on \(Z_i\), the process \(N^\ast\) is an homogeneous Poisson process with intensity \(\theta_0'Z_i + \alpha\), that is
\[
E[N^\ast(t)|Z_i] = (\theta_0'Z_i + \alpha)t, \quad i = 1, \ldots, n.
\]

We take \(\alpha = 5\) and \(\theta_0 = (1, 1.6, 1.25, 0.7)'\). We consider 4-dimensional covariates \(Z_i \sim \otimes^4\mathcal{U}[1, 2]\) for \(i = 1, \ldots, n\). The variables \(Y_i\) for \(i = 1, \ldots, n\) are generated according to a Weibull distribution with parameters \((10, 1.09)\). The censoring distribution is selected to be Weibull with parameters \((4, \lambda)\). Taking \(\lambda = 1.38\) or \(\lambda = 1\) leads to respectively 30\% or 50\% of censoring and an average of 20 or 18 recurrents events per sample. In our results, we emphasize the impact of the two parameters involved in our semiparametric procedure, namely the bandwidth of the nonparametric kernel estimators and the measure \(w\).

First, we consider the case of a fixed bandwidth and show how the adaptive choice of \(\hat{w}\) can improve the estimation performance of the parameter \(\theta_0\). The nonparametric estimators are kernel estimators computed using a Gaussian kernel and a bandwidth \(h_0 = 0.2\). We consider a set of discrete measures supported on \(\mathcal{I} = \{0.1, 0.2, \ldots, 1.2\}\). Hence, for any function \(f\), the integral with respect to \(w\) reduces to a finite sum. Indeed, we have
\[
\int f(t)dw(t) = \sum_{k \in \mathcal{I}} f(k)w(\{k\}).
\]
Moreover, we consider only a finite number of choices for the weights \( w(\{k\}) \), that is
\[
\begin{cases}
  w(\{k\}) = 1 & \text{for } k = 0.1, \ldots, 0.8 \\
  w(\{k\}) \in \{0.25, 0.5, 0.75, 1\} & \text{for } k = 0.9, 1, 1.1, 1.2.
\end{cases}
\]

The intuition is that our procedure should allocate smaller weights to large values of \( T_i \) since the behavior of the Kaplan-Meier estimator is known to be less effective in this part of the distribution (and contributes significantly to the variance). Our estimator \( \hat{\theta} = \hat{\theta}(\hat{w}, h_0) \) is then compared to the estimator \( \tilde{\theta} \) obtained for the measure \( w_0 \) which puts mass 1 at every point of \( \mathcal{I} \).

In the table below, we report our results over 100 simulations of samples of size 100 for two different rates of censoring (\( p = 30\% \) and \( p = 50\% \)). Recalling that the first component of \( \theta_0 \) is imposed to be one, we only have to estimate the three other components. For each estimator, the Mean Squared Error (MSE) \( E(\|\hat{\theta} - \theta_0\|^2) \) is decomposed into bias and variance.

### Table 1

<table>
<thead>
<tr>
<th>( p = 30% )</th>
<th>Bias</th>
<th>Variance</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\theta} )</td>
<td>((-0.322))</td>
<td>(\begin{bmatrix} 0.452 &amp; 0.111 &amp; 0.041 \ 0.111 &amp; 0.42 &amp; 0.009 \ 0.041 &amp; 0.009 &amp; 0.249 \end{bmatrix})</td>
<td>1.2645</td>
</tr>
<tr>
<td>( \tilde{\theta} )</td>
<td>((-0.129))</td>
<td>(\begin{bmatrix} 0.2 &amp; 0.062 &amp; 0.047 \ 0.062 &amp; 0.272 &amp; -0.004 \ 0.047 &amp; -0.004 &amp; 0.168 \end{bmatrix})</td>
<td>0.6855</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( p = 50% )</th>
<th>Bias</th>
<th>Variance</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\theta} )</td>
<td>((-0.428))</td>
<td>(\begin{bmatrix} 0.478 &amp; 0.129 &amp; 0.156 \ 0.129 &amp; 0.386 &amp; 0.034 \ 0.156 &amp; 0.034 &amp; 0.335 \end{bmatrix})</td>
<td>1.4903</td>
</tr>
<tr>
<td>( \tilde{\theta} )</td>
<td>((-0.276))</td>
<td>(\begin{bmatrix} 0.242 &amp; 0.035 &amp; 0.033 \ 0.035 &amp; 0.234 &amp; 0.023 \ 0.033 &amp; 0.023 &amp; 0.199 \end{bmatrix})</td>
<td>0.8433</td>
</tr>
</tbody>
</table>

We also compute the average weights of \( \hat{w} \) for the last four points of \( \mathcal{I} \). For 30\% of censoring, we have: \( E[\hat{w}(\{0.9\})] = 0.7775, E[\hat{w}(\{1\})] = 0.6525, E[\hat{w}(\{1.1\})] = 0.6075 \)
and $E[\hat{w}([1.2])] = 0.535$ and for 50% of censoring, $E[\hat{w}([0.9])] = 0.7825$, $E[\hat{w}([1])] = 0.6825$, $E[\hat{w}([1.1])] = 0.575$ and $E[\hat{w}([1.2])] = 0.4875$. Clearly, choosing the measure from the data improves both the bias and the variance of our estimator. Moreover the weights of $\hat{w}$ get smaller for large values of $k$, especially when the proportion of censored data is high. Consequently, the adaptive measure seems to have a significant impact on the quality of the estimation of $\theta_0$.

Next, we show how the choice of the parameter $h$ influences the quality of estimation. We consider the fixed measure $w_0$ which puts the same weights 1 at each point. The bandwidth $\hat{h}$ is chosen adaptively in a regular grid of length 0.05 in the set $[0.05, 0.3]$. The performance of the resulting estimator presented below is compared with the estimator $\tilde{\theta}$ of the previous table and shows significant improvement of its MSE.

<table>
<thead>
<tr>
<th></th>
<th>Bias</th>
<th>Variance</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\theta}_{w_0,\hat{h}}, p = 30%$</td>
<td>$-0.19$</td>
<td>$0.216$ $0.08$ $-0.08$</td>
<td>0.967</td>
</tr>
<tr>
<td></td>
<td>$-0.155$</td>
<td>$0.08$ $0.351$ $-0.009$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.084</td>
<td>$-0.08$ $-0.009$ $0.174$</td>
<td></td>
</tr>
<tr>
<td>$\hat{\theta}_{w_0,\hat{h}}, p = 50%$</td>
<td>$-0.281$</td>
<td>$0.244$ $-0.056$ $-0.081$</td>
<td>1.126</td>
</tr>
<tr>
<td></td>
<td>$-0.309$</td>
<td>$-0.056$ $0.256$ $0.027$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$-0.114$</td>
<td>$-0.081$ $0.027$ $0.17$</td>
<td></td>
</tr>
</tbody>
</table>

### 5 Conclusion

We proposed a new procedure to estimate the conditional cumulative mean function of the recurrent event process. We considered both parametric and semiparametric models for the conditional cumulative mean function. Our semiparametric single-index model can be seen as a generalization of both the Cox model and the accelerated failure time model. Moreover, a new feature of our procedure stands in the measure $w$ involved in our estimators which is designed to prevent us from problems in the tail of the distribution due to the presence of censoring. Then, we proposed a data-driven method to choose this measure adaptively. Our criterion is based on the minimization of the mean squared error for the estimation of $\theta_0$ but our procedure is flexible enough to allow the use of any other criteria more adapted to the context. For example, we could consider a criterion directly based on the error of the estimation of $\mu$.  

22
In this work, we mainly focused on kernel estimators for estimating the nonparametric part of our model, providing methods to choose the smoothing parameters from the data. Nevertheless, all our results are still valid for a general class of nonparametric estimators and do only rely on convergence properties. Hence, other kinds of estimators may be used provided they satisfy these conditions.

6 Appendix

6.1 Proof of Lemma 3.1

Let

\[ S_n^{T(n)}(f, w) = \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{T(n)} Y_i(t) f(Z_i, t) dw(t). \]

Write

\[ \tilde{S}_n(f, w) = S_n^{T(n)}(f, w) + \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{T(n)} f(Z_i, t) \int_{0}^{t} \frac{\hat{G}(s) - G(s)}{[1 - G(s)][1 - G(s)-]} dw(t) \]

\[ = S_n^{T(n)}(f, w) + R_n(f, w). \]

Decompose \( f \) into its positive and negative parts denoted respectively by \( f^+ \) and \( f^- \). The expectations of the two resulting sums \( S_n^{T(n)}(f^+, w) \) and \( S_n^{T(n)}(f^-, w) \) go to zero faster than \( n^{-1/2} \) using Lebesgue’s dominated convergence. This entails that

\[ \sup_{f \in \mathcal{F}, w \in \mathcal{W}} |S_n^{T(n)}(f, w) - S_n(f, w)| = o_P(n^{-1/2}). \]

Let \( \tau < \tau_H \) and define \( w_\tau(t) = w(t)I(t \leq \tau) \). On \([0, \tau]\), we use the asymptotic i.i.d. expansion of the Kaplan-Meier estimator \( \hat{G} \) proposed by Gijbels and Veraverbeke (1991) which can also be deduced from Stute (1995):

\[ \frac{\hat{G}(t) - G(t)}{1 - G(t)} = \frac{1}{n} \sum_{j=1}^{n} \eta_k(T_j, \delta_j) + \tilde{R}_n(t), \]

where \( \sup_{t \leq \tau} |\tilde{R}_n(t)| = O_{a.s.}(n^{-1} \log n) \) and

\[ \eta_k(T, \delta) = \frac{(1 - \delta)I(T \leq t)}{1 - H(T^-)} - \int_{0}^{t} \frac{I(T \geq s)dG(s)}{[1 - H(s-)][1 - G(s-)]}. \]

Moreover, we recall that \( \sup_{t \leq \tau} |\hat{G}(t) - G(t)| = O_P(n^{-1/2}) \) (see Gill (1983), Theorem 2.1) and that \( \sup_{t \leq \tau} (1 - G(t))(1 - \hat{G}(t))^{-1} = O_P(1) \) (see Gill (1983), Lemma 2.6). Then, we
decompose

\[ R_n(f, w_r) = \frac{1}{n^2} \sum_{i,j} \int_0^{T(n)} f(Z_i, t) \int_0^t \eta_{k-}(T_j, \delta_j) dN_i(s) \frac{1}{1 - G(s)} \, dw_r(t) + R'_n(f, w_r). \]

Using the fact that \( \mathcal{F} \) is an uniformly bounded class, that \( \int dw_r \leq C_0 \) from Assumption 3 and that \( E[N_i(\tau)] < \infty \) for all \( \tau \), we deduce that \( \sup f, w \left| R'_n(f, w_r) \right| = O_P(n^{-1}) \). The first term in \( R_n(f, w_r) \) can be rewritten as

\[
\frac{1}{n} \sum_{j=1}^n \int_0^{T(n)} \int_0^t \eta_{k-}(T_j, \delta_j) E[f(Z, t) \mu(s | Z)] \, dw_r(t) + \int \left( \frac{1}{n^2} \sum_{i,j} \psi_{f,t}(Z_i, N_i, T_j, \delta_j) \right) \, dw_r(t),
\]

where

\[
\psi_{f,t}(Z_i, N_i, T_j, \delta_j) = \int_0^t \eta_{k-}(T_j, \delta_j) \left\{ \frac{f(Z_i, t) dN_i(s)}{1 - G(s)} - E[f(Z, t) \mu(s | Z)] \right\}.
\]

Observe that, with probability tending to one, the upper bound \( T(n) \) in the integrals can be replaced by \( \tau < \tau_H \). Let \( f, f' \in \mathcal{F} \) and \( t, t' \in [0, \tau] \). We have

\[
|\psi_{f,t}(Z_i, N_i, T_j, \delta_j) - \psi_{f',t'}(Z_i, N_i, T_j, \delta_j)| \leq C_\tau \left( \|f - f'\|_\infty N_i(\tau) + |t - t'|^\alpha \sup_{t,t' \leq \tau} \frac{N_i(t) - N_i(t')}{|t - t'|^\alpha} \right),
\]

where \( C_\tau < \infty \) and \( \alpha > 0 \). Let \( \mathcal{H}_\tau \) denote the set of all functions \( \psi_{f,t} \) when \( f \) ranges \( \mathcal{F} \) and \( t \) ranges \([0, \tau]\). It follows from (6.1) and Assumption 3 that \( \mathcal{H}_\tau \) is a \( \| \cdot \|_2 \)-VC-class of functions. From this, using the Glivenko-Cantelli property of \( \mathcal{H}_\tau \),

\[
\sup_{f,t \leq \tau} \left| \frac{1}{n^2} \sum_{i=1}^n \psi_{f,t}(Z_i, N_i, T_i, \delta_i) \right| = O_P(n^{-1})
\]

and

\[
\sup_{f,t \leq \tau} \left| \frac{1}{n^2} \sum_{i \neq j} \psi_{f,t}(Z_i, N_i, T_j, \delta_j) \right| = O_P(n^{-1}),
\]

since this can be seen as the supremum of a second order degenerate \( U \)-process indexed by \( \mathcal{H}_\tau \) (see Sherman (1994)). This leads to the i.i.d. representation for \( \hat{S}_n(f, w_r) \) for any \( \tau < \tau_H \).

Similarly, write

\[
\hat{S}_n(\hat{f}, w_r) = S_n^{T(n)}(\hat{f}, w_r) + R_n(\hat{f}, f, w_r) + R_n(f, w_r)
\]
and using the fact that \( \sup_{f \in \mathcal{F}} \| f - \hat{f} \|_\infty = o_P(1) \) and that \( \sup_{t \leq \tau} |\hat{G}(t) - G(t)| = O_P(n^{-1/2}) \), we deduce that \( \sup_{f,w} |R_n(\hat{f} - f, w_r)| = o_P(n^{-1/2}) \). The representation for \( \hat{S}_n(\hat{f}, w_r) \) follows.

Now, we make \( \tau \) tend to \( \tau_0 \). Let \( \hat{P}_n^\tau(f, w) = \hat{S}_n(f, w) - S_n^{T(\tau)}(f, w) \) and \( P_n^\tau(f, w) = \hat{S}_n(f, w_r) - S_n^{T(\tau)}(f, w_r) \). Since the class \( F \) is uniformly bounded, we get

\[
|\hat{P}_n^\tau(f, w) - P_n^\tau(f, w)| \leq \frac{M}{n} \sum_{i=1}^{n} \int_{0}^{T(\tau)} \int_{0}^{t} \frac{|\hat{G}(s-) - G(s-)|}{[1 - G(s-)][1 - \hat{G}(s-)]} dN_i(s) dw(t) \leq \frac{M'}{n} \sum_{i=1}^{n} \int_{0}^{T(\tau)} \frac{W_0(s \vee \tau)|\hat{G}(s-) - G(s-)|dN_i(s)}{[1 - G(s-)][1 - \hat{G}(s-)]},
\]

where the last inequality is obtained from Fubini’s theorem and Assumption 3. From Theorem 1.2 in Gill (1983), Assumption 3 and the fact that \( \sup_{t \leq T(\tau)} [1 - G(t-)][1 - \hat{G}(t-)]^{-1} = O_P(1) \) (see again Gill, 1983), we get that

\[
|\hat{P}_n^\tau(f, w) - P_n^\tau(f, w)| \leq \frac{A_n}{n} \sum_{i=1}^{n} \int_{0}^{T(\tau)} \frac{W_2(s \vee \tau)dN_i(s)}{1 - G(s-)},
\]

where \( A_n = O_P(n^{-1/2}) \). The result follows from Lemma 6.6.

### 6.2 Uniform convergence of the nonparametric estimators

In this section, we show that the kernel estimator \( \hat{\mu}_{\theta,h} \) defined by (2.9) satisfies the convergence rates required by Assumption 10. Introduce the quantity

\[
\hat{\mu}_{\theta,h}(t, u) = \sum_{i=1}^{n} \int_{0}^{t} K \left( \frac{\theta'Z_i - u}{h} \right) dN_i(s) \frac{1}{\sum_{j=1}^{n} K \left( \frac{\theta'Z_j - u}{h} \right) [1 - G(s-)]},
\]

We first study the convergence rate of the difference between \( \hat{\mu}_{\theta,h} \) and \( \mu_{\theta} \) and their derivatives. Since no Kaplan-Meier functions are involved in this expression, we can use classical results on uniform convergence of kernel estimators, mainly from Einmahl and Mason (2005).

**Assumption 10.** Assume that

1. \( K \) has a compact support, say \([-1, 1] \), \( \int_R K(s) ds = 1 \) and \( \sup_{x} |K(x)| < \infty \),
2. \( K \) is a twice differentiable and two order kernel with derivatives of order 0, 1 and 2 of bounded variation,
(3) \( K := \{ K((x - \cdot)/h) : h > 0, x \in \mathbb{R}^d \} \) is a pointwise measurable class of functions,

(4) \( h \in \mathcal{H}_n \subset [an^{-\alpha}, bn^{-\alpha}] \) with \( a, b > 0 \) and \( \alpha \in (1/8, 1/5) \).

We also introduce a trimming function in order to prevent from denominator close to zero in the definition of \( \hat{\mu}_{\theta,h} \). Indeed, to ensure uniform consistency of our estimator, we need to bound this denominator away from zero. We use the same methodology as in Delecroix et al. (2006). Let \( \theta_0 Z \) denote the density of \( \theta_0 Z \) and define the “ideal” trimming function \( J_{\theta_0}(\theta_0 Z, c) = I(\theta_0 Z \in B_0) \) where \( B_0 = \{ u : \theta_0 Z(u) \geq c \} \) for some constant \( c > 0 \). As in Delecroix et al. (2006) (see also Lopez (2009)), we first assume that we know some set \( B \) on which \( \inf \{ f_{\theta Z}(\theta'z) : z \in B, \theta \in \Theta \} > c \) where \( c \) is a strictly positive constant. In a preliminary step, we can use this set \( B \) to compute the preliminary trimming \( J_B(z) = I(z \in B) \). Using this trimming function and a deterministic sequence of bandwidth \( h_0 \) satisfying (4) in Assumption 10, we define a preliminary estimator \( \theta_n \) of \( \theta_0 \) as

\[
\theta_n(w) = \arg \min_{\theta \in \Theta} M_{n,w}(\theta, \hat{\mu}_{\theta}) J_B(z).
\]

Given this preliminary consistent estimator of \( \theta_0 \), we use the following trimming \( J_n(\theta_0 Z, c) = I(\hat{\theta}_{\theta Z}(\theta_0'Z) \geq c) \) which appears to be asymptotically equivalent to \( J_{\theta_0}(\theta_0'Z, c) \) (see e.g. Lopez (2009)). Then, our final estimator consists of

\[
\hat{\theta}(w) = \arg \min_{\theta \in \Theta_n} M_{n,w}(\theta, \hat{\mu}_{\theta}) J_n(\theta_n'Z, c),
\]

where \( \Theta_n \) is a shrinking neighborhood of \( \theta_0 \) accordingly to our preliminary estimator \( \theta_n \).

As announced, the next proposition gives the rates of convergence of \( \hat{\mu}_{\theta,h} \) and its derivatives. Since we need a convergence over \( \theta \in \Theta \), the trimming we need to use is \( J_\theta(\theta'Z, c) := I(\hat{\theta}_{\theta Z}(\theta'Z) \geq c) \). But notice that \( J_{\theta_0}(\theta_0'Z, c) \) can be replaced by \( J_\theta(\theta'Z, c/2) \) on shrinking neighborhoods of \( \theta_0 \).

**Proposition 6.1.** Under Assumption 10, for \( z \) such that \( J_\theta(\theta'z, c) > 0 \), we have

\[
\sup_{t \leq T_{(n),\theta,z,h}} \sqrt{\frac{nh}{\log n}} \left| \frac{\mu_0(t, \theta'z) - \mu_0(t, \theta'z)}{\hat{\mu}_{\theta_0}(t, \theta_0'z)^{\lambda_1 + \lambda_2}} \right| = O_P(1), \quad (6.2)
\]

\[
\sup_{t \leq T_{(n),\theta,z,h}} \sqrt{\frac{nh^3}{\log n}} \left| \nabla_\theta \mu_0(t, z) - \nabla_\theta \mu_0(t, z) \right| = O_P(1), \quad (6.3)
\]

\[
\sup_{t \leq T_{(n),\theta,z,h}} \sqrt{\frac{nh^5}{\log n}} \left| \nabla_\theta^2 \mu_0(t, z) - \nabla_\theta^2 \mu_0(t, z) \right| = O_P(1). \quad (6.4)
\]
Proof. The proofs of (6.2)-(6.4) are all similar. The most delicate term to handle, coming from (6.4), is

$$
\hat{A}_{n,h}^{n,h}(t,z) := \frac{1}{n} \sum_{i=1}^{n} \frac{(Z_i - z)^2}{n^3} \left( \hat{\mu}_{\theta_0}(t, \theta_0' z)^{\lambda_1 + \lambda_2} K'' \left( \frac{\theta' Z_i - \theta' z}{h} \right) \right) \int_0^t \frac{dN_i(s)}{1 - G(s)}.
$$

Consider the class of functions $K$ introduced in Assumption 10. From Nolan and Pollard (1987), it can easily be seen that, using a kernel $K$ satisfying Assumption 10, for some $C > 0$ and $\nu > 0$, we have $N(\epsilon, K, \| \cdot \|_{\infty}) \leq C \epsilon^{-\nu}$, $0 < \epsilon < 1$.

Then, concerning the uniformity with respect to $\theta$, Lemma 22 (ii) of Nolan and Pollard (1987) shows that the family of functions \{(Z, N) \mapsto \hat{A}_{n,h}^{n,h}(t,z)\} satisfies the assumptions of Proposition 1 in Einmahl and Mason (2005).

Define

$$
\tilde{A}_{h}(t,z) := \frac{1}{h^3} E \left[ \frac{(Z - z)^2}{\hat{\mu}_{\theta_0}(t, \theta_0' z)^{\lambda_1 + \lambda_2} K'' \left( \frac{\theta' Z - \theta' z}{h} \right)} \right] \int_0^t \frac{dN(s)}{1 - G(s)},
$$

$$
A_{h}(t,z) := \frac{\partial^2}{\partial u^2} \left( E \left[ \frac{(Z - z)^2}{\hat{\mu}_{\theta_0}(t, \theta_0' z)^{\lambda_1 + \lambda_2}} \int_0^t \frac{dN(s)}{1 - G(s)} \right] \right)_{u=\theta' z}
$$

and apply Talagrand’s inequality (see Talagrand (1994), see also Einmahl and Mason (2005)) to obtain that

$$
\sup_{t \leq T(n) \theta, z,h} |\hat{A}_{n,h}(t,z) - \tilde{A}_{n,h}(t,z)| = O_P\left(n^{-1/2} h^{-5/2} (\log n)^{1/2}\right).
$$

For the bias term, classical kernel arguments (see for instance Bosq and Lecoutre (1997)) show that

$$
\sup_{t \leq T(n) \theta, z,h} |\hat{A}_{n,h}(t,z) - A_{n,h}(t,z)| = O(h^2).
$$

It remains to study $\hat{\mu}_{\theta,h} - \tilde{\mu}_{\theta,h}$. The following lemma gives some precision on the difference between the Kaplan Meier weights of $\hat{\mu}_{\theta,h}$ and the “ideal” weights involving the true function $G$ in $\tilde{\mu}_{\theta}$.

Lemma 6.2. Let $\hat{W}(s) = (1 - \hat{G}(s))^{-1}$, $\tilde{W}(s) = (1 - G(s))^{-1}$ and

$$
C_G(t) = \int_0^t \frac{dG(s)}{(1 - G(s))(1 - H(s))}.
$$

(1) We have

$$
\sup_{t \leq T(n)} \frac{1 - G(t)}{1 - \hat{G}(t)} = O_P(1).
$$

27
(2) For all $0 \leq \alpha \leq 1$ and $\varepsilon > 0$, we have

$$|\hat{W}(s) - \bar{W}(s)| \leq R_n(s)\bar{W}(s)C_G(s)^{\alpha(1/2+\varepsilon)},$$

where $\sup_{s \leq T(n)} R_n(s) = O_P(n^{-\alpha/2})$.

Proof. (1) This result is a consequence of Lemma 2.6 in Gill (1983).

(2) For $0 \leq \alpha \leq 1$ and $\varepsilon > 0$, write

$$\hat{W}(s) - \bar{W}(s) = \bar{W}(s)C(s)^{\alpha(1/2+\varepsilon)} \left( Z_G(s)C(s)^{-1/2-\varepsilon} \right)^\alpha \left( Z_G(s) \right)^{1-\alpha} \frac{1 - G(s-)}{1 - \bar{G}(s-)},$$

where $Z_G(s) = \left( \bar{G}(s-) - G(s-) \right) \left( 1 - G(s-) \right)^{-1}$. Since $\int_0^T C_G(s)^{-1-2\varepsilon} dC_G(s) < \infty$, apply Theorem 1 in Gill (1983) and use the first part of the current lemma to conclude the proof.

From the definition of our estimator, problems arise when studying $\hat{\mu}_{\theta,\varepsilon}$ for $t$ in the tail of the distribution. This is a common problem when studying Kaplan-Meier estimators but it can be circumvented by some moment conditions on the response and censoring distribution. For instance, in the classical censored framework, Stute (1995) used the function $C_G$ to compensate the bad behavior of Kaplan Meier estimator in the tail of the distribution. The following assumption gives a similar moment condition but adapted to our recurrent event context.

**Assumption 11.** Assume that, for some $\eta > 0$,

$$\sup_{t,z} \frac{C_G(t)^{7/20+\eta}}{\mu_{\theta_0}(t, \theta_0^* z)^{\lambda_1}} < \infty$$

and

$$\sup_{t,z} \frac{\int_0^T (1 - G(s-))E[N^*(s)dN^*(s)]}{(1 - G(t-))\mu_{\theta_0}(t, \theta_0^* z)^{2\lambda_2}} < \infty,$$

where $\lambda_1$ and $\lambda_2$ are defined in Assumption 3.

**Proposition 6.3.** Under Assumptions 11 and 14, for $z$ such that $J_0(\theta^* z, \varepsilon) > 0$, we have

$$\sup_{t \leq T(n), \theta, z, h} \left| \frac{\bar{\mu}_{\theta}(t, \theta^* z) - \bar{\mu}_{\theta}(t, \theta^* z)}{\mu_{\theta_0}(t, \theta_0^* z)^{\lambda_1+\lambda_2}} \right| = O_P \left( n^{-7/20} \right),$$

and

$$\sup_{t \leq T(n), \theta, z, h} \left| \frac{\bar{\mu}_{\theta}(t, \theta^* z) - \bar{\mu}_{\theta}(t, \theta^* z)}{\mu_{\theta_0}(t, \theta_0^* z)^{\lambda_1+\lambda_2}} \right| = O_P \left( n^{-7/20} \right),$$

and

$$\sup_{t \leq T(n), \theta, z, h} \left| \frac{\bar{\mu}_{\theta}(t, \theta^* z) - \bar{\mu}_{\theta}(t, \theta^* z)}{\mu_{\theta_0}(t, \theta_0^* z)^{\lambda_1+\lambda_2}} \right| = O_P \left( n^{-7/20} \right).$$
Proof. We only prove (6.7) since (6.5) and (6.6) can be handled similarly. Let us consider the following term involving the second derivative of $K$

$$\frac{1}{n h^3} \sum_{i=1}^{n} (Z_i - z)^2 K'' \left( \frac{\theta' Z_i - \theta' z}{h} \right) \left( \hat{\mu}_{\theta_0}(t, \theta' Z_i) \right)^{-1} \int_0^t \left( \hat{W}(s) - \bar{W}(s) \right) dN_i(s).$$

From Lemma 6.2, this term can be bounded by

$$O_P(n^{-\alpha/2} h^{-2}) \left| \frac{1}{nh} \sum_{i=1}^{n} K'' \left( \frac{\theta' Z_i - \theta' z}{h} \right) \hat{\mu}_{\theta_0}(t, \theta' Z_i)^{-(\lambda_1 + \lambda_2)} \int_0^t \hat{W}(s) C_G(s)^{\alpha(1/2 + \varepsilon)} dN_i(s) \right|$$

(6.8)

where the $O_P-$ rate does not depend on $t, \theta, z$ nor $h$. Now, consider the family of functions indexed by $t, \theta, z$ and $h$,

$$\left\{ (Z, N) \mapsto K'' \left( \frac{\theta' Z - \theta' z}{h} \right) \hat{\mu}_{\theta_0}(t, \theta' Z)^{-(\lambda_1 + \lambda_2)} \int_0^t \hat{W}(s) C_G(s)^{\alpha(1/2 + \varepsilon)} dN(s) \right\}.$$

This family is Euclidian (see Nolan and Pollard (1987)) for an envelope

$$\sup_{t, z} \frac{\hat{W}(t) C_G^{(1/2 + \varepsilon)}(t) \bar{N}(t)}{\mu_{\theta_0}(t, \theta' Z)^{\lambda_1 + \lambda_2}}$$

which is, for $\alpha = 7/10$, square integrable from Assumption 11. Then, using the results of Sherman (1994), the second part of (6.8) is $O_P(1)$ uniformly in $t, \theta, z$ and $h$.

Finally, combination of Propositions 6.1 and 6.3 leads to the following result.

Corollary 6.4. Under Assumptions 10 and 11, for $z$ such that $J_\theta(\theta' z, \alpha) > 0$,

$$\sup_{t \leq T(n), \theta, z, h} \left| \hat{\mu}_{\theta}(t, \theta' z) - \mu_{\theta}(t, \theta' z) \right| \left| \bar{N}(t) \right| = o_P(n^{-1/2}).$$

6.3 Technical lemmas

6.3.1 Gradient vector in the single-index model

Lemma 6.5. If the function $\theta \mapsto \mu_{\theta}(t|\theta' z)$ is differentiable, we have

$$\nabla_{\theta} \mu_{\theta_0}(t|Z) = \mu'_{\theta_0}(t|\theta' Z)[Z - E(Z|\theta' Z)],$$

where $\mu'_{\theta_0}(t|u) = \frac{d}{du} \mu_{\theta_0}(t|u)$. As a consequence

$$E \left[ \nabla_{\theta} \mu_{\theta_0}(t|Z) | \theta' Z \right] = 0.$$
Proof. Observe that $\mu_\theta(t|\theta' Z) = E[\mu_{\theta_0}(t|\theta'_0 Z)|\theta' Z]$ and let $\alpha(Z, \theta) = \theta'_0 Z - \theta' Z$ for $\theta \in \Theta$. We have
\[
\mu_\theta(t|\theta' Z) = E\left[\mu_{\theta_0}(t|\alpha(Z, \theta) + \theta'_2 Z)|\theta'_2 Z\right].
\]
Defining $\Gamma(\theta_1, \theta_2) = E\left[\mu_{\theta_0}(t|\alpha(Z, \theta_1) + \theta'_2 Z)|\theta'_2 Z\right]$, we have $\Gamma(\theta, \theta) = \mu_\theta(t|\theta' Z)$, which leads to
\[
\nabla_{\theta_1} \Gamma(\theta_1, \theta_0)|_{\theta_1 = \theta_0} = -\mu_{\theta_0}(t|\theta'_0 Z)E[Z|\theta'_0 Z],
\]
\[
\nabla_{\theta_2} \Gamma(\theta_0, \theta_2)|_{\theta_2 = \theta_0} = Z\mu'_{\theta_0}(t|\theta'_0 Z).
\]

\[\square\]

6.3.2 Auxiliary lemma for tightness conditions

Lemma 6.6. Let $F$ be a class of functions. Let $P_n(t, f)$ be a process on $[0; \tau_H] \times F$. Define, for any $\tau \in [0; \tau_H]$, $R_n(\tau, f) = P_n(\tau_H, f) - P_n(\tau, f)$. Assume that for any $\tau$ such that $\tau < \tau_H$
\[
P_n(t, f) \Rightarrow W(V_f(t)) \in D([0; \tau]), f \in F,
\]
where $W(V_f(t))$ is a Gaussian process with covariance function $V_f$ and $D$ denotes the set of càdlàg functions.

Assume that, for a sequence of random variables $(Z_n)$ and two functions $\Gamma$ and $\Gamma_n$, the following conditions hold,

(1) $\lim_{\tau \to \tau_H} V_f(\tau) = V_f(\tau_H)$ with $\sup_{f \in F} |V_f(\tau_H)| < \infty$,

(2) $|R_n(\tau', f)| \leq Z_n \times \Gamma_n(\tau)$ for all $\tau < \tau' < \tau_H$,

(3) $Z_n = O_P(1)$,

(4) $\Gamma_n(\tau) \to \Gamma(\tau)$ in probability,

(5) $\lim_{\tau \to \tau_H} \Gamma(\tau) = 0$.

Then $P_n(\tau_H, f) \Rightarrow N(0, V_f(\tau_H))$.

Proof. From Theorem 13.5 in Billingsley (1999) and condition (1), it suffices to show that, for all $\varepsilon > 0$
\[
\lim_{\tau \to \tau_H} \limsup_{n \to \infty} P\left(\sup_{\tau \leq t \leq \tau_H, f \in F} |R_n(t, f)| > \varepsilon\right) = 0.
\]
Using condition (2), the probability in equation (6.10) is bounded, for all $M > 0$, by

$$P(\|\Gamma_n(\tau) - \Gamma(\tau)\| > \varepsilon/M - \Gamma(\tau)) + P(Z_n > M). \quad (6.11)$$

Moreover, from condition (4), we can state that

$$\limsup_{n \to \infty} P(\|\Gamma_n(\tau) - \Gamma(\tau)\| > \varepsilon/M - \Gamma(\tau)) = I(\varepsilon/M - \Gamma(\tau) \geq 0).$$

Since $\Gamma(\tau) \to 0$ (condition (5)), we can deduce that

$$\lim_{\tau \to \tau_H} \limsup_{n \to \infty} P(\|\Gamma_n(\tau) - \Gamma(\tau)\| > \varepsilon/M - \Gamma(\tau)) = 0.$$ 

As a consequence,

$$\lim_{\tau \to \tau_H} \limsup_{n \to \infty} P \left( \sup_{\tau \leq t \leq \tau_H, f \in F} |R_n(t, f)| > \varepsilon \right) \leq \lim_{M \to \infty} \limsup_{n \to \infty} P(Z_n > M) = 0,$$

using the fact that $Z_n = O_P(1)$ (condition (3)).

### 6.3.3 Covering number results

In this section, we determine covering numbers of some particular classes of functions. From these computations, we can easily deduce sufficient conditions to check Property 2 and Assumption 4.

**Proposition 6.7.** Let $F$ be a class of functions $f(t, z)$ with envelope $F$ defined on $\mathbb{R} \times \mathbb{R}^d$ with continuous derivative with respect to the first component. Let $\tilde{F}$ be the envelope of the class of functions $\partial f(s, z)/\partial s$. Let $W_0(t)$ be a positive bounded decreasing function and set $W = \{w : dw(t) = W_0(t)\tilde{w}(t), \tilde{w} \in \tilde{W}\}$ where $\tilde{W}$ is a class of monotone positive functions uniformly bounded by a same bound $M \geq 0$. Let $W$ be an envelope function for $W$.

Assume that $E[(\int_0^{\tau_H} F(t, z)W_0(t)dY(t))^2] < \infty$, $E[(\int_0^{\tau_H} F(t, z)Y(t)dW_0(t))^2] < \infty$ and $E[(\int_0^{\tau_H} \tilde{F}(t, z)W_0(t)Y(t)dt)^2] < \infty$.

Then, the class of functions $H = \{(z, y) \to \int_0^{\tau_H} f(t, z)g(t)dw(t), f \in F, w \in W\}$ has a uniform covering number satisfying, for some constant $C$,

$$N(\varepsilon, H, \| \cdot \|_2) \leq C N(\varepsilon, W_0F, \| \cdot \|_2) N(\varepsilon, \tilde{W}, \| \cdot \|_2).$$

**Proof.** Let $Q$ be a probability measure and introduce $N_1 = N_Q(\varepsilon\|W_0F\|_Q, W_0F, \| \cdot \|_{2,Q})$ and $N_2 = N_Q(\varepsilon\|W\|_Q, \tilde{W}, \| \cdot \|_{2,Q})$. Let $\{f^{W_i}_j\}_{1 \leq i \leq N_1}$ (respectively $\{\tilde{w}_j\}_{1 \leq j \leq N_2}$) be the center
of the $\varepsilon - \| \cdot \|_{2,Q}$ balls needed to cover $W_0\mathcal{F}$ (respectively $\hat{W}$). Writing $dw = W_0 d\tilde{w}$, we have for any $1 \leq i \leq N_1$ and $1 \leq j \leq N_2$

$$\left| \int_0^{\tau_H} Y(t)f(t,z)W_0(t)dw(t) - \int_0^{\tau_H} Y(t)f_i^W(t,z)d\tilde{w}_j(t) \right| \leq \left| \int_0^{\tau_H} Y(t)(f(t,z)W_0(t) - f_i^W(t,z))d\tilde{w}_j(t) \right| + \left| \int_0^{\tau_H} Y(t)f(t,z)W_0(t)(d\tilde{w} - d\tilde{w}_j)(t) \right|.$$ 

For any $f \in \mathcal{F}$, there exists some $i$ such that the first term is seen to be less than $C_1\varepsilon$ in $L^2(Q)$–norm. For the second term, there also exists some $j$ such that this is smaller than $C_2\varepsilon$, which can be seen using integration by parts. The result follows. 

\[ \square \]

\section*{References}


