 Vol 32: 425–445, 2005

On Goodness-of-Fit Tests for Aalen's Additive Risk Model

AXEL GANDY

Department of Stochastics, University of Ulm

UWE JENSEN

Institute for Applied Mathematics and Statistics, University of Hohenheim

ABSTRACT. In this paper we propose goodness-of-fit tests for Aalen's additive risk model. They are based on test statistics the asymptotic distributions of which are determined under both the null and alternative hypotheses. The results are derived using martingale techniques for counting processes. An important feature of these tests is that they can be adjusted to particular alternatives. One of the alternatives we consider is Cox's multiplicative risk model. It is perhaps remarkable that such a test needs no estimate of the baseline hazard in the Cox model. We present simulation studies which give an impression of the performance of the proposed tests. In addition, the tests are applied to real data sets.

Key words: Aalen model, additive risk model, counting process, Cox regression, goodness-of-fit, martingale residuals, survival analysis

1. Introduction

In Aalen's additive risk model which is used in survival analysis and recurrent event analysis an n-variate counting process $\boldsymbol{N}(t)=(N_1(t),\dots,N_n(t))^{\mathrm{T}}$ is observed together with certain covariates $Y_{ij}(t), j=1,\dots,k$ for each component $N_i(t)$. The covariates and the counting process are linked together by the assumption that the intensity $\boldsymbol{\lambda}(t)=(\lambda_1(t),\dots,\lambda_n(t))^{\mathrm{T}}$ of $\boldsymbol{N}(t)$ has the form

$$\lambda_i(t)=\sum_{j=1}^{k} Y_{ij}(t)\alpha_j(t), \tag{1}$$

where $\alpha_j(t)$ are the unknown deterministic quantities that need to be estimated. The model was originally introduced by Aalen (1980). Further discussion of it including asymptotic results for estimators of $\alpha_j(t)$ can be found in Aalen (1989, 1993), Andersen *et al*. (1993), Huffer & McKeague (1991) and McKeague (1988a).

Goodness-of-fit of the model has been discussed in several papers. In McKeague & Utikal (1991) the fit of Aalen's model was compared to the fit in a larger class of models. In order to achieve an asymptotic χ^2-distribution it is necessary to partition the observation interval and the space the covariates take their values in. In particular, if the covariates are multidimensional (i.e. $k>1$) this partitioning might prove difficult. The simulation study in McKeague & Utikal (1991) shows that large sample sizes are required to ensure that the observed level is close to the nominal level.

Kim *et al*. (1998), Yuen & Burke (1997) and Song *et al*. (1996) considered goodness-of-fit in a smaller semiparametric model in which

$$\lambda_i(t)=(\lambda_0(t)+\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i)R_i(t), \tag{2}$$

where $\boldsymbol{Z}_i \in \mathbb{R}^p$ are time-independent covariates, $R_i(t) \in \{0, 1\}$ are at risk-indicators, $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown regression parameter, and $\lambda_0(t)$ is an unobservable baseline intensity.

Aalen (1993) suggested the use of so-called martingale residuals. He used graphical methods to assess goodness-of-fit and discussed some suggestions for formal tests based on grouped martingale residuals.

In Grønnesby & Borgan (1996) martingale residual processes grouped after a certain 'risk score' were considered. Time-independent covariates $\boldsymbol{Z}_i$ are assumed and the 'risk score' $\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{Z}_i$ is obtained by fitting the smaller model (2). The sum of the martingale residuals in each group is used to construct an asymptotically χ^2-distributed test statistic.

In this paper, we propose tests based on an idea similar to the above-mentioned martingale residuals. Assume that $\boldsymbol{c}(t) = (c_1(t), \ldots, c_n(t))^{\mathrm{T}}$ is a vector of predictable (and observable) stochastic processes that is perpendicular to the columns of $\boldsymbol{Y}(t) = (Y_{ij}(t))_{i=1,\ldots,n,\, j=1,\ldots,k}$, in the sense that $\boldsymbol{Y}(t)^{\mathrm{T}}\boldsymbol{c}(t) = \boldsymbol{0}$. Then under Aalen's model and some regularity conditions,

$$\hat{T}(t) := \frac{1}{\sqrt{n}} \int_0^t \boldsymbol{c}(s)^{\mathrm{T}}\, \mathrm{d}\boldsymbol{N}(s)$$

is a local martingale with mean zero since $\int_0^t \boldsymbol{c}(s)^{\mathrm{T}}\boldsymbol{\lambda}(s)\,\mathrm{d}s = \int_0^t \boldsymbol{c}(s)^{\mathrm{T}}\boldsymbol{Y}(s)\boldsymbol{\alpha}(s)\,\mathrm{d}s = 0$. So $\hat{T}(t)$ should fluctuate around 0. We will consider $\boldsymbol{c}(t)$ defined via a projection of some suitably chosen vector $\boldsymbol{d}(t)$ onto the space orthogonal to the columns of $\boldsymbol{Y}(t)$. Our test statistic $\hat{T}(t)$ can be interpreted as the sum of weighted martingale residuals with time-dependent weights $d_i(t)$. We will suggest choices of $\boldsymbol{d}(t)$ to detect certain alternatives. One of the choices of $\boldsymbol{d}(t)$ we consider is designed to detect the multiplicative risk model introduced by Cox (1972). In some cases, as for example in the Cox model, the appropriate choice of $\boldsymbol{d}(t)$ may depend on a finite-dimensional parameter $\boldsymbol{\beta} \in \mathbb{R}^p$, $p \in \mathbb{N}$. Consequently, the statistic $\hat{T}(t)$ depends on $\boldsymbol{\beta}$. Replacing this parameter by an estimate $\hat{\boldsymbol{\beta}}$ results in a new statistic $\hat{T}^*(t)$. We show that if $\hat{\boldsymbol{\beta}}$ is $n^{1/2}$-consistent then under mild conditions the statistics $\hat{T}(t)$ and $\hat{T}^*(t)$ are asymptotically equivalent, i.e. $\hat{T}(t) - \hat{T}^*(t) \xrightarrow{P} 0$. This allows to use $\hat{T}^*(t)$ instead of $\hat{T}(t)$.

Here is an outline of the following sections. In section 2, we show asymptotic results concerning $\hat{T}$ and $\hat{T}^*$ which we use in section 3 to construct tests. Some of these tests are of Kolmogorov–Smirnov type. In section 4, we consider the behaviour of the tests under alternatives and suggest how $\boldsymbol{d}(t)$ can be chosen. We discuss conditions under which our asymptotic results hold for these choices of $\boldsymbol{d}(t)$. In section 5, we consider the case where the individuals are i.i.d. and show consistency of our tests. In section 6, we present simulation studies similar to those of McKeague & Utikal (1991). Our tests perform satisfactorily if (1) holds, even for small sample sizes. Under alternatives, our tests achieve a far better power than the test proposed in McKeague & Utikal (1991) (which is not surprising since our tests are adjusted to detect certain alternatives). In section 7, we apply our tests to a data set from software reliability (Gandy & Jensen, 2004), the Stanford heart transplant data (Miller & Halpern, 1982) and the primary biliary cirrhosis (PBC) data (Fleming & Harrington, 1991). Finally, we present some conclusions in section 8.

We always write matrices and vectors in bold face ($\boldsymbol{M}, \boldsymbol{x}$) and if we refer to elements of a matrix or a vector we denote the elements by M_{ij} or x_i. Furthermore, $\boldsymbol{M_i}$ denotes the ith row of $\boldsymbol{M}$, i.e. if $\boldsymbol{M}$ has k columns, $\boldsymbol{M_i} = (M_{i1}, \ldots, M_{ik})$. Stochastic convergence is denoted by $\xrightarrow{P}$ and convergence in distribution in the sense of Billingsley (1999) is denoted by $\xrightarrow{\mathrm{d}}$. Convergence in this paper will always be as $n \to \infty$ unless indicated otherwise. If a matrix $\boldsymbol{A} \in \mathbb{R}^{k \times k}$ is not invertible then $\boldsymbol{A}^{-1}$ is defined to be the $k \times k$ matrix with all elements equal to 0. For vectors, $\|\cdot\|$ denotes the Euclidean norm and for matrices, $\|\cdot\|$ denotes the corresponding matrix norm, i.e. $\|\boldsymbol{A}\| = \sup \|\boldsymbol{A}\boldsymbol{x}\|$ where the sup is over all vectors $\boldsymbol{x}$ such that $\|\boldsymbol{x}\| = 1$. For sequences (X_n), $n \in \mathbb{N}$ of random variables we say $X_n = O_P(1)$ if for each $\epsilon > 0$ there exists $K > 0$ such that $\sup_{n \in \mathbb{N}} P(|X_n| > K) < \epsilon$. If (a_n), $n \in \mathbb{N}$ is another sequence we write $X_n = O_P(a_n)$ if $X_n / a_n = O_P(1)$. If $(\boldsymbol{X}_n)$, $n \in \mathbb{N}$ is a sequence of random vectors, $O_P(\cdot)$ is defined in the same way with $|\cdot|$ replaced by $\|\cdot\|$.

2. Asymptotic results

We want to be able to study the behaviour of the test statistics that will be defined in section 3 not only under the assumption that Aalen's model is the true underlying model, i.e. (1) holds, but also under alternative hypotheses (see section 4). Therefore, we derive asymptotic results that do not require that (1) holds. We only assume that our counting process $\boldsymbol{N}(t)$ has an intensity $\boldsymbol{\lambda}(t)$.

More precisely for some τ, $0 < \tau < \infty$, let $\mathfrak{T} = [0, \tau]$ be the interval and $(\Omega, \mathcal{F}, P)$ the probability space on which all stochastic processes in this paper are defined. $D(\mathfrak{T})$ denotes the space of càdlàg functions from $\mathfrak{T}$ to $\mathbb{R}$, equipped with the Skorokhod topology and its Borel-σ-algebra. Let $k \in \mathbb{N}$ be the number of covariates per component of $\boldsymbol{N}(t)$. For each $n \in \mathbb{N}$, $n \geq k$, let the following stochastic elements be given on a filtered probability space. To ease notation we will not make the dependence on n explicit. Let $\boldsymbol{N}(t) = (N_1(t), \ldots, N_n(t))^{\mathrm{T}}$ be an adapted multivariate counting process whose elements have no common jumps. We assume that $\boldsymbol{N}(t)$ admits an intensity $\boldsymbol{\lambda}(t) = (\lambda_1(t), \ldots, \lambda_n(t))^{\mathrm{T}}$, i.e. $\boldsymbol{\lambda}(t)$ is a locally bounded predictable process and $\boldsymbol{M}(t) := (M_1(t), \ldots, M_n(t))^{\mathrm{T}} := \boldsymbol{N}(t) - \int_0^t \boldsymbol{\lambda}(s)\,\mathrm{d}s$ is a local martingale. Furthermore, let the covariates $\boldsymbol{Y}(t) = (Y_{ij}(t))$ be an $n \times k$ matrix of locally bounded predictable processes. Let $\boldsymbol{c}(t) = (c_1(t), \ldots, c_n(t))^{\mathrm{T}}$ be a vector of locally bounded predictable processes that satisfies $\boldsymbol{Y}(t)^{\mathrm{T}} \boldsymbol{c}(t) = \boldsymbol{0}$ for all $t \in \mathfrak{T}$. As we aim at a convergence result under Aalen's model as well as under alternatives the asymptotic result will be for

$$T(t) := \frac{1}{\sqrt{n}} \int_0^t \boldsymbol{c}(s)^{\mathrm{T}}\, \mathrm{d}\boldsymbol{M}(s).$$

As estimator for the variance of $T(t)$ we will use the variation process

$$\hat{G}(t) := [T](t) = \frac{1}{n} \sum_{i=1}^{n} \int_0^t c_i^2(s)\, \mathrm{d}N_i(s).$$

Of course, if (1) holds true, then for all $t \in \mathfrak{T}$:

$$\hat{T}(t) - T(t) = \frac{1}{\sqrt{n}} \int_0^t \boldsymbol{c}(s)^{\mathrm{T}} \boldsymbol{\lambda}(s)\, \mathrm{d}s = \frac{1}{\sqrt{n}} \int_0^t \boldsymbol{c}(s)^{\mathrm{T}} \boldsymbol{Y}(s) \boldsymbol{\alpha}(s)\, \mathrm{d}s = 0.$$

We introduce some additional notation to be used throughout the paper which will ease the presentation considerably. For matrices $\boldsymbol{A} \in \mathbb{R}^{n \times a}$, $\boldsymbol{B} \in \mathbb{R}^{n \times b}$ and vectors $\boldsymbol{x} \in \mathbb{R}^n$ with a, b, $n \in \mathbb{N}$ we define

$$\overline{\boldsymbol{A}} = \frac{1}{n} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{1}, \quad \overline{\boldsymbol{AB}} = \frac{1}{n} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{B}, \quad \overline{\boldsymbol{AxB}} = \frac{1}{n} \boldsymbol{A}^{\mathrm{T}} \mathrm{diag}(\boldsymbol{x}) \boldsymbol{B},$$

where $\boldsymbol{1} = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^n$ and $\mathrm{diag}(\boldsymbol{x})$ denotes the diagonal matrix with the elements of $\boldsymbol{x}$ on its diagonal. More generally, if $k \in \mathbb{N}$, $a_1, \ldots, a_k \in \mathbb{N}$ and for all $j = 1, \ldots, k$, $\boldsymbol{A}^j \in \mathbb{R}^{n \times a_j}$ we define for all $b_j \in \{1, \ldots, a_j\}$, $j = 1, \ldots, k$,

$$\overline{\boldsymbol{A}^1 \cdots \boldsymbol{A}^k}_{b_1, \ldots, b_k} := \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{k} A^j_{ib_j} = \frac{1}{n} \sum_{i=1}^{n} A^1_{ib_1} \cdots A^k_{ib_k}.$$

We call $\overline{\boldsymbol{A}^1 \cdots \boldsymbol{A}^k}$ *product mean*. The $\boldsymbol{A}^j$ are allowed to be random and may also depend on further parameters which will be indicated in parentheses, e.g. if $\boldsymbol{A}^1$ and $\boldsymbol{A}^2$ depend on $t \in \mathfrak{T}$

then $\overline{A^1A^2}(t)=(1/n)A^1(t)^{\mathrm{T}}A^2(t)$. If $A^1,\ldots,A^k$ depend on $\theta\in\Theta$, where Θ is some measurable space, we say that $\overline{A^1\cdots A^k}$ converges *uip* (uniformly in probability) on Θ if there exists a deterministic, measurable, bounded function $\overrightarrow{A^1\cdots A^k}:\Theta\to\mathbb{R}^{a_1\times\cdots\times a_k}$ such that for all $b_j\in\{1,\ldots,a_j\}$, $j=1,\ldots,k$,

$$\sup_{\theta\in\Theta}\left|\overline{A^1\cdots A^k}_{b_1,\ldots,b_k}(\theta)-\overrightarrow{A^1\cdots A^k}_{b_1,\ldots,b_k}(\theta)\right|\xrightarrow{P}0.$$

We will use $\Theta=\mathfrak{T}$ and $\Theta=\mathcal{C}\times\mathfrak{T}$ for some $\mathcal{C}\subset\mathbb{R}^p$.

Using the new notation with the above $\boldsymbol{c}$ and $\boldsymbol{\lambda}$ the following proposition is an immediate consequence of Rebolledo's theorem (Andersen *et al.*, (1993), p. 78).

Proposition 1

If $\overline{c\lambda c}$ *converges uip on* $\mathfrak{T}$ *and* $n^{-\frac{1}{2}}\sup_{i=1,\ldots,n;\, t\in\mathfrak{T}}|c_i(t)|\xrightarrow{P}0$ *then*

$$T\xrightarrow{\mathrm{d}}m \text{ in } D(\mathfrak{T}),$$

where m *is a continuous, mean-zero Gaussian martingale with covariance* $\mathrm{Cov}(m(s),m(t))=\int_0^{s\wedge t}\overrightarrow{c\lambda c}(u)\,\mathrm{d}u$. *Furthermore,* $\hat{G}(t)\xrightarrow{P}\mathrm{Var}(m(t))$ *uniformly in* $t\in\mathfrak{T}$.

A natural way to define $\boldsymbol{c}(t)$ is to take a locally bounded predictable stochastic process $\boldsymbol{d}(t)$ and project it onto the space orthogonal to the columns of $\boldsymbol{Y}(t)$. Let $\boldsymbol{Q}(t)$ be the projection matrix onto this space. Being a projection matrix, $\boldsymbol{Q}(t)$ satisfies $\boldsymbol{Q}(t)=\boldsymbol{Q}(t)^{\mathrm{T}}=\boldsymbol{Q}^2(t)$. If $\boldsymbol{Y}(t)$ has full rank then $\boldsymbol{Q}(t)=\boldsymbol{I}-\boldsymbol{Y}(t)(\boldsymbol{Y}(t)^{\mathrm{T}}\boldsymbol{Y}(t))^{-1}\boldsymbol{Y}(t)^{\mathrm{T}}$ (if $\boldsymbol{Y}(t)$ does not have full rank then $(\boldsymbol{Y}(t)^{\mathrm{T}}\boldsymbol{Y}(t))^{-1}$ can be replaced by a generalized inverse). So the setup we are dealing with is $\boldsymbol{c}(t)=\boldsymbol{Q}(t)\boldsymbol{d}(t)$. If we choose $d_i(t)=1$ for all i and t then $\hat{T}(t)=n^{-1/2}\mathbf{1}^{\mathrm{T}}\int_0^t\boldsymbol{Q}(s)\,\mathrm{d}\boldsymbol{N}(s)$ is the scaled sum of the so-called martingale residuals (Aalen 1989, 1993).

The following conditions will be needed to show weak convergence of $T(t)$. They are similar to conditions used in McKeague (1988a). In fact, our first condition (D1) contains the conditions (A2), (A3) and parts of (A1) of McKeague (1988a).

(D1) $\overline{YY}$ converges uip on $\mathfrak{T}$, $\overrightarrow{YY}$ is continuous, $\overrightarrow{YY}(t)$ is invertible for all $t\in\mathfrak{T}$ and $n^{-1/2}\sup_{i=1,\ldots,n;\, t\in\mathfrak{T}}|Y_{ij}(t)|\xrightarrow{P}0$ for $j=1,\ldots,k$.

(D2) $\overline{Y\lambda Y}$, $\overline{Yd}$, $\overline{d\lambda d}$, $\overline{Y\lambda d}$ converge uip on $\mathfrak{T}$ and $n^{-1/2}\sup_{i=1,\ldots,n;\, t\in\mathfrak{T}}|d_i(t)|\xrightarrow{P}0$.

Theorem 1

Assume the conditions (D1) and (D2) hold and that $\boldsymbol{c}(t)=\boldsymbol{Q}(t)\boldsymbol{d}(t)$. *Then*

$$T\xrightarrow{d}m \text{ in } D(\mathfrak{T}),$$

where m *is a continuous, mean-zero Gaussian martingale with covariance* $\mathrm{Cov}(m(s),m(t))=G(s\wedge t)$, $G(t)=\int_0^t\overrightarrow{(Qd)\lambda(Qd)}(u)\,\mathrm{d}u$. *Furthermore,* $\hat{G}(t)\xrightarrow{P}G(t)$ *uniformly in* $t\in\mathfrak{T}$.

Proof. Lemma 9 in appendix B ensures $n^{-1/2}\sup_{i=1,\ldots,n;\, t\in\mathfrak{T}}|(\boldsymbol{Q}(t)\boldsymbol{d}(t))_i|\xrightarrow{P}0$. By lemma 8 in appendix B, $\overline{c\lambda c}=\overline{(Qd)\lambda(Qd)}$ converges uip on $\mathfrak{T}$. Hence we can apply proposition 1.

We can actually extend the above to allow $\boldsymbol{d}$ to depend on a finite-dimensional parameter. We will make use of this extension in section 4.3 to detect Cox's model as alternative. Suppose $\mathcal{B}\subset\mathbb{R}^p$ is an open set and $\boldsymbol{d}:\mathcal{B}\times\mathfrak{T}\to\mathbb{R}^n$ is a random function. We require that for fixed

$\boldsymbol{\beta} \in \mathcal{B}$ the process $\boldsymbol{d}(\boldsymbol{\beta}, \cdot) : \mathfrak{T} \to \mathbb{R}^n$ is a locally bounded predictable stochastic process. In our test statistic we would like to use $\boldsymbol{d}$ with a certain $\boldsymbol{\beta}^{\mathbf{0}} \in \mathcal{B}$ which in general is not observable. We assume that we have an estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}^{\mathbf{0}}$. Note that we do not require $\boldsymbol{d}(\hat{\boldsymbol{\beta}}, \cdot)$ to be adapted. This will allow $\hat{\boldsymbol{\beta}}$ to use the entire information up to time τ. The modification of the test statistic $\hat{T}$ we get is

$$\hat{T}^*(t) := n^{-1/2} \int_0^t \boldsymbol{d}(\hat{\boldsymbol{\beta}}, s)^{\mathrm{T}} \boldsymbol{Q}(s)\, \mathrm{d}\boldsymbol{N}(s).$$

As before the convergence results will not be for $\hat{T}^*$ itself but for

$$T^*(t) := n^{-1/2} \int_0^t \boldsymbol{d}(\hat{\boldsymbol{\beta}}, s)^{\mathrm{T}} \boldsymbol{Q}(s)\, \mathrm{d}\boldsymbol{M}(s).$$

Under Aalen's model we have $\hat{T}^*(t) - T^*(t) = 0$. As an estimator for the variance of T^* we will use

$$\hat{G}^*(t) := \frac{1}{n} \int_0^t \boldsymbol{d}(\hat{\boldsymbol{\beta}}, s)^{\mathrm{T}} \boldsymbol{Q}(s) \text{ diag } (\mathrm{d}\boldsymbol{N}(s)) \boldsymbol{Q}(s) \boldsymbol{d}(\hat{\boldsymbol{\beta}}, s).$$

Using $\boldsymbol{d}(t) := \boldsymbol{d}(\boldsymbol{\beta}^{\mathbf{0}}, t)$ in the definition of T and $\hat{G}$ we will show that $T^* - T$ and $\hat{G}^* - \hat{G}$ both converge to zero. For this we need some conditions.

(D3) $\boldsymbol{d}(\cdot, s, \omega) : \mathcal{B} \to \mathbb{R}^n$ is twice continuously differentiable for fixed $s \in \mathfrak{T}$, $\omega \in \Omega$.

We use $\nabla_\nu d_i(\boldsymbol{\beta}, s) := \frac{\partial}{\partial \beta_\nu} d_i(\boldsymbol{\beta}, s)$, $\nabla_\nu \boldsymbol{d}(\boldsymbol{\beta}, s) := \frac{\partial}{\partial \beta_\nu} \boldsymbol{d}(\boldsymbol{\beta}, s)$, $\nabla_\nu \nabla_\eta \boldsymbol{d}(\boldsymbol{\beta}, s) := \frac{\partial}{\partial \beta_\nu} \frac{\partial}{\partial \beta_\eta} \boldsymbol{d}(\boldsymbol{\beta}, s)$ and $\nabla \boldsymbol{d}(\boldsymbol{\beta}, s) := (\nabla_1 \boldsymbol{d}(\boldsymbol{\beta}, s), \ldots, \nabla_p \boldsymbol{d}(\boldsymbol{\beta}, s))$.

(D4) $\overline{\boldsymbol{\lambda}}$, $\overline{\boldsymbol{Y\lambda Y}}$, $\overline{(\nabla \boldsymbol{d})\boldsymbol{\lambda Y}}(\boldsymbol{\beta}^{\mathbf{0}}, \cdot)$, $\overline{(\nabla \boldsymbol{d})\boldsymbol{\lambda}(\nabla \boldsymbol{d})}(\boldsymbol{\beta}^{\mathbf{0}}, \cdot)$ converge uip on $\mathfrak{T}$. There exists an open set $\mathcal{C} \subset \mathcal{B}$ such that $\boldsymbol{\beta}^{\mathbf{0}} \in \mathcal{C}$, $\overline{\boldsymbol{dY}}$, $\overline{(\nabla \boldsymbol{d})\boldsymbol{Y}}$, $\overline{(\nabla_\nu \nabla_\eta \boldsymbol{d})\boldsymbol{Y}}$, $\nu, \eta = 1, \ldots, p$ converge uip on $\mathcal{C} \times \mathfrak{T}$ and $\overrightarrow{\boldsymbol{dY}}(\cdot, t)$, $\overrightarrow{(\nabla \boldsymbol{d})\boldsymbol{Y}}(\cdot, t)$, $\overrightarrow{(\nabla_\nu \nabla_\eta \boldsymbol{d})\boldsymbol{Y}}(\cdot, t)$, $\nu, \eta = 1, \ldots, p$ are continuous in $\boldsymbol{\beta}^{\mathbf{0}}$ uniformly in $t \in \mathfrak{T}$. For $\nu, \eta = 1, \ldots, p, j = 1, \ldots, k$,

$$n^{-1/4} \sup_{\substack{i=1,\ldots,n \\ t \in \mathfrak{T}}} |Y_{ij}(t)| \xrightarrow{P} 0, \qquad n^{-1/4} \sup_{\substack{i=1,\ldots,n \\ t \in \mathfrak{T},\, \boldsymbol{\beta} \in \mathcal{C}}} |d_i(\boldsymbol{\beta}, t)| \xrightarrow{P} 0,$$

$$n^{-1/4} \sup_{\substack{i=1,\ldots,n \\ t \in \mathfrak{T},\, \boldsymbol{\beta} \in \mathcal{C}}} |(\nabla_\nu d_i)(\boldsymbol{\beta}, t)| \xrightarrow{P} 0 \qquad \text{and } n^{-1/2} \sup_{\substack{i=1,\ldots,n \\ t \in \mathfrak{T},\, \boldsymbol{\beta} \in \mathcal{C}}} |(\nabla_\nu \nabla_\eta d_i)(\boldsymbol{\beta}, t)| \xrightarrow{P} 0.$$

Theorem 2

Suppose that (D1), (D3) *and* (D4) *hold.*

If $\boldsymbol{c}(t) = \boldsymbol{Q}(t)\boldsymbol{d}(\boldsymbol{\beta}^{\mathbf{0}}, t)$ *and* $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}} = O_P(n^{-1/2})$ *then*

$$T^* - T \xrightarrow{P} 0 \text{ and } \hat{G}^* - \hat{G} \xrightarrow{P} 0 \text{ uniformly in } t \in \mathfrak{T}.$$

The proof is relegated to appendix C.

3. Construction of test statistics

We construct tests for the hypothesis that Aalen's model is true, i.e. for

$$H_0 : \boldsymbol{\lambda}(t) = \boldsymbol{Y}(t)\boldsymbol{\alpha}(t) \text{ for some deterministic, bounded, measurable } \boldsymbol{\alpha} : \mathfrak{T} \to \mathbb{R}^k.$$

We only describe the construction of tests from $\hat{T}$ and $\hat{G}$. Construction of tests from $\hat{T}^*$ and $\hat{G}^*$ is done in a similar way.

We assume that the conditions of theorem 1 hold. Under H_0, we are in the situation that $\hat{T} \xrightarrow{d} m$ where m is a mean zero Gaussian martingale whose variance $G(t)$ can be estimated consistently by $\hat{G}(t)$ uniformly in t. There are various ways to construct asymptotic tests in such a situation. We present three test statistics the asymptotic distributions of which are explicitly known. We always require $G(\tau) > 0$.

$$T^{(1)} := \hat{G}(\tau)^{-1/2}\hat{T}(\tau) \xrightarrow{d} N(0,1).$$

$$T^{(2)} := \sup_{t\in\mathfrak{T}} \left| \frac{\hat{G}(\tau)^{1/2}}{\hat{G}(\tau)+\hat{G}(t)} \hat{T}(t) \right| \xrightarrow{d} \sup_{t\in[0,\frac{1}{2}]} |W^0(t)|,$$

where $W^0(t)$ is a Brownian bridge. This transformation can for example be found in Hall & Wellner (1980).

$$T^{(3)} := \sup_{t\in\mathfrak{T}} \left| \hat{G}(\tau)^{-\frac{1}{2}} \hat{T}(t) \right| \xrightarrow{d} \sup_{t\in[0,\,1]} |W(t)|,$$

where $W(t)$ is a Brownian motion. The convergence is based on the fact that $m(t) \stackrel{d}{=} W(G(t))$, where G and m are as in theorem 1.

An explicit formula for the asymptotic distribution of $T^{(2)}$ can be found in Hall & Wellner (1980). Formulas for the asymptotic distribution of $T^{(3)}$ can be derived from Borodin & Salminen (2002). For the test statistics $T^{(2)}$ and $T^{(3)}$ we always reject at the upper tail. For the test statistic $T^{(1)}$ we will indicate whether we use a two-sided test or a one-sided test (rejecting at the upper tail).

4. Behaviour under alternatives

In this section we will consider choices of $\boldsymbol{d}(t)$ in theorem 1 and $\boldsymbol{d}(\boldsymbol{\beta}, t)$ in theorem 2 to test against three particular alternatives. We start with propositions concerning the asymptotic behaviour of $\hat{T}$ and $\hat{T}^*$ under alternatives. We need the following condition:

(D5) $\overline{\boldsymbol{d\lambda}}$ and $\overline{\boldsymbol{Y\lambda}}$ converge uip on $\mathfrak{T}$.

Proposition 2

If (D1), (D2) and (D5) hold and $\boldsymbol{c}(t) = \boldsymbol{Q}(t)\boldsymbol{d}(t)$ *then*

$$n^{-1/2}\hat{T}(t) \xrightarrow{P} H(t) := \int_0^t \overrightarrow{(\boldsymbol{Qd})\boldsymbol{\lambda}}(s)\,\mathrm{d}s$$

uniformly in $t \in \mathfrak{T}$.

Proof. Since $n^{-1/2}\hat{T}(t) = n^{-1/2}(\hat{T}(t) - T(t)) + n^{-1/2}T(t)$ and $n^{-1/2}T(t) \xrightarrow{P} 0$ uniformly in $t \in \mathfrak{T}$ by theorem 1, it is enough to consider $n^{-1/2}(\hat{T}(t) - T(t)) = \int_0^t \overline{(\boldsymbol{Qd})\boldsymbol{\lambda}}(s)\,\mathrm{d}s$. By lemma 8 $\overline{(\boldsymbol{Qd})\boldsymbol{\lambda}}$ converges uip on $\mathfrak{T}$.

In the next proposition we consider the case where $\boldsymbol{d}$ may depend on a finite-dimensional parameter. In condition (D2) we set $\boldsymbol{d}(t) := \boldsymbol{d}(\boldsymbol{\beta^0}, t)$.

Proposition 3

Suppose (D1), (D2), (D3) and (D4) hold, $\overline{\boldsymbol{Y\lambda}}$ *converges uip on* $\mathfrak{T}$, *and* $\overline{\boldsymbol{d\lambda}}$ *converges uip on* $\mathcal{C} \times \mathfrak{T}$ *(where* $\mathcal{C}$ *is as in (D4)). If* $\boldsymbol{c}(t) = \boldsymbol{Q}(t)\boldsymbol{d}(\boldsymbol{\beta^0}, t)$ *and* $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^0} = O_P(n^{-1/2})$ *then*

$$n^{-1/2}\hat{T}^*(t) \xrightarrow{P} H(t)$$

uniformly in $t \in \mathfrak{T}$, *where* $H(t) = \int_0^t \overrightarrow{(\boldsymbol{Qd})\boldsymbol{\lambda}}(\boldsymbol{\beta^0}, s)\,\mathrm{d}s$.

Proof. By lemma 10, $\overline{dY}(\hat{\beta}, \cdot) \to \overrightarrow{dY}(\beta^0, \cdot)$ and $\overline{d\lambda}(\hat{\beta}, \cdot) \to \overrightarrow{d\lambda}(\beta^0, \cdot)$ converge uip on $\mathfrak{T}$. Lemma 8 implies convergence of $\overline{(Qd)\lambda}(\hat{\beta}, \cdot) \to \overrightarrow{(Qd)\lambda}(\beta^0, \cdot)$ uip on $\mathfrak{T}$. Hence, $n^{-1/2}(\hat{T}^*(t) - T^*(t)) = \int_0^t \overline{(Qd)\lambda}(\hat{\beta}, s)\,ds$ converges uip on $\mathfrak{T}$ to $H(t)$. Considering that theorem 1 together with theorem 2 implies $n^{-1/2}T^* \to 0$ uip on $\mathfrak{T}$ finishes the proof.

Under H_0 we know that $\hat{T} = T$ (or $\hat{T}^* = T^*$) and thus $H(t) = 0$ in this case. For alternatives in which $H(t) \neq 0$, the proposed tests from section 3 can be seen to be consistent (note that theorem 1 yields the stochastic convergence of $\hat{G}(t)$). More about $H(t) \neq 0$ will be said in section 5.

For the three alternatives we consider we will suggest choices of $\boldsymbol{d}(t)$ (or $\boldsymbol{d}(\boldsymbol{\beta}, t)$). We also consider the conditions required for the theorems and propositions presented so far in these special cases.

4.1. Using an estimator of the intensity

If the intensity of the alternative (say $\gamma(s)$) is completely known then we suggest to choose $\boldsymbol{d}(s) = \gamma(s)$. Then under the alternative $\overrightarrow{(Qd)\lambda}(t) = (1/n)(Q(t)\gamma(t))(Q(t)\gamma(t)) \geq 0$, and proposition 2 shows that $H(t) \geq 0$, enabling us to use one-sided tests.

However, the intensity of the alternative is usually not known precisely. If we have a uniformly consistent estimator of the intensity under the alternative, using this estimator as choice for $\boldsymbol{d}$ allows us to use a one-sided test as the following lemma shows. Recall that $\boldsymbol{d}$ has to be predictable.

Lemma 1

Suppose that (D1), (D2) and (D5) hold, that $\boldsymbol{c}(t) = \boldsymbol{Q}(t)\boldsymbol{d}(t)$, $\overline{\boldsymbol{\lambda\lambda}}$ *converges uip on* $\mathfrak{T}$, $\sup_{i=1,\dots,n;\, t\in\mathfrak{T}} |d_i(t) - \lambda_i(t)| \xrightarrow{P} 0$ *and* $\sup_{t\in\mathfrak{T}} |\overline{\lambda}(t)| = O_P(1)$, $\sup_{t\in\mathfrak{T}} \overline{|\boldsymbol{Y}|}_j(t) = O_P(1)$, $j = 1, \dots, k$ (*where* $\overline{|\boldsymbol{Y}|}_j(t) := \frac{1}{n}\sum_{i=1}^n |Y_{ij}(t)|$). *Then*

$$n^{-1/2}\hat{T}(t) \xrightarrow{P} H(t) = \int_0^t \overrightarrow{(\boldsymbol{Q\lambda})\boldsymbol{\lambda}}(s)\,ds \geq 0.$$

Proof. By proposition 2, $H(t) = \int_0^t \overrightarrow{(\boldsymbol{Qd})\boldsymbol{\lambda}}(s)\,ds$. Consider the decomposition

$$\overline{(\boldsymbol{Qd})\boldsymbol{\lambda}} = \overline{(\boldsymbol{Q}(\boldsymbol{d}-\boldsymbol{\lambda}))\boldsymbol{\lambda}} + \overline{(\boldsymbol{Q\lambda})\boldsymbol{\lambda}}. \tag{3}$$

The second term on the right-hand side of (3) converges uip on $\mathfrak{T}$ by lemma 8. We will show that the first term on the right-hand side of (3) is asymptotically negligible. Since

$$|\overline{(\boldsymbol{d}-\boldsymbol{\lambda})\boldsymbol{\lambda}}(t)| \leq \frac{1}{n}\sum_{i=1}^n |d_i(t) - \lambda_i(t)|\lambda_i(t) \leq \sup_{\substack{i=1,\dots,n\\ s\in\mathfrak{T}}} |d_i(s) - \lambda_i(s)|\overline{\lambda}(t) \xrightarrow{P} 0$$

uniformly in $t \in \mathfrak{T}$ and similarly $|\overline{(\boldsymbol{d}-\boldsymbol{\lambda})\boldsymbol{Y}}_j(t)| \xrightarrow{P} 0$ uniformly in $t \in \mathfrak{T}$, we can apply lemma 8 and get convergence of $\overline{(\boldsymbol{Q}(\boldsymbol{d}-\boldsymbol{\lambda}))\boldsymbol{\lambda}}$ uip on $\mathfrak{T}$ and $\overrightarrow{(\boldsymbol{Q}(\boldsymbol{d}-\boldsymbol{\lambda}))\boldsymbol{\lambda}}(t) = 0$.

Similarly, under the conditions of proposition 3, it can be shown that, $n^{-1/2}\hat{T}^*$ converges to a non-negative process. A condition needed for this is $\sup_{i=1,\dots,n;\, t\in\mathfrak{T}} |d_i(\hat{\boldsymbol{\beta}}, t) - \lambda_i(t)| \xrightarrow{P} 0$ for which the following is sufficient: $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}^0$, $d_i(\boldsymbol{\beta}^0, t) = \lambda_i(t)$ and a type of 'equicontinuity' of $d_i(\boldsymbol{\beta}, t)$ at $\boldsymbol{\beta}^0$: for each $\epsilon > 0$ there exists a neighbourhood $\mathcal{C}$ of $\boldsymbol{\beta}^0$ such that $P(\sup_{i=1,\dots,n;\, t\in\mathfrak{T},\, \boldsymbol{\beta}\in\mathcal{C}} |d_i(\boldsymbol{\beta}, t) - d_i(\boldsymbol{\beta}^0, t)| > \epsilon) \to 0$.

4.2. Additional covariate

As a second alternative, we consider a model with additional covariates:

$$H_d : \boldsymbol{\lambda}(t) = \boldsymbol{X}(t)\boldsymbol{\beta}(t), \text{ for some deterministic, bounded, measurable } \boldsymbol{\beta} : \mathfrak{T} \to \mathbb{R}^{k+k'},$$

where $\boldsymbol{X}$ is an $n \times (k + k')$ matrix of predictable, observable processes whose submatrix consisting of the first k columns equals $\boldsymbol{Y}$. The parameter $\boldsymbol{\beta}$ is unknown.

We choose $\boldsymbol{d} := (X_{1,k+1}, \ldots, X_{n,k+1})^{\mathrm{T}}$, i.e. $\boldsymbol{d}$ is column $k + 1$ of $\boldsymbol{X}$. Using the boundedness of $\boldsymbol{\beta}$, the following lemma is immediate.

Lemma 2

Suppose $\overline{\boldsymbol{XX}}$, $\overline{\boldsymbol{XXX}}$ converge uip on $\mathfrak{T}$, $\overrightarrow{\boldsymbol{YY}}$ is continuous, $\overrightarrow{\boldsymbol{YY}}(t)$ is invertible for all $t \in \mathfrak{T}$ and $n^{-1/2} \sup_{i=1,\ldots,n,\, t\in\mathfrak{T}} |X_{ij}(t)| \xrightarrow{P} 0$ for $j = 1, \ldots, k$.

If H_d is true then (D1), (D2) and (D5) hold.

Since H_0 is included in H_d, we need not consider H_0 separately.

4.3. Cox's model as alternative

The third alternative is the model introduced by Cox (1972). Let the covariates in Cox's model be the entries of an $n \times p$ matrix $\boldsymbol{Z}$ of locally bounded predictable processes. Of course, we could have $\boldsymbol{Z}(t) = \boldsymbol{Y}(t)$ or $\boldsymbol{Z}(t)$ could consist of certain columns of $\boldsymbol{Y}(t)$, but we do not require this to be the case. Let R_i, $i = 1, \ldots, n$ be predictable processes taking values in $\{0, 1\}$ and let $\mathcal{B} \subset \mathbb{R}^p$ be open and convex. Cox's model assumes that

$$H_c : \lambda_i(t) = \lambda_0(t) R_i(t) \exp(\boldsymbol{Z}_i(t)\boldsymbol{\beta}^{\boldsymbol{0}}), \quad i = 1, \ldots, n,$$

for some $\boldsymbol{\beta}^{\boldsymbol{0}} \in \mathcal{B}$ and some deterministic, bounded, measurable $\lambda_0 : \mathfrak{T} \to [0, \infty)$, where $\boldsymbol{Z}$ and R_i are observable, $\boldsymbol{\beta}^{\boldsymbol{0}}$ and λ_0 are not. We will base our test on $\hat{T}^*$ with

$$d_i(\boldsymbol{\beta}, s) := R_i(s) \exp(\boldsymbol{Z}_i(s)\boldsymbol{\beta}), \quad i = 1, \ldots, n.$$

The parameter $\boldsymbol{\beta}^{\boldsymbol{0}}$ is usually estimated by a partial maximum likelihood approach (Cox, 1972; Andersen & Gill, 1982), i.e. $\hat{\boldsymbol{\beta}}$ is the maximizer of $C(\boldsymbol{\beta}, \tau)$ where

$$C(\boldsymbol{\beta}, t) := \int_0^t \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{Z}(s)^{\mathrm{T}}\, \mathrm{d}\boldsymbol{N}(s) - \int_0^t \log(n\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s)) n \,\mathrm{d}\overline{\boldsymbol{N}}(s).$$

Note that we do not need to estimate $\lambda_0(t)$. Since under H_c we have $\overline{(\boldsymbol{Qd})\boldsymbol{\lambda}}(\boldsymbol{\beta}^{\boldsymbol{0}}, t) = (1/n)(\boldsymbol{Q}(t)\boldsymbol{d}(\boldsymbol{\beta}^{\boldsymbol{0}}, t))^{\mathrm{T}}(\boldsymbol{Q}(t)\boldsymbol{d}(\boldsymbol{\beta}^{\boldsymbol{0}}, t))\lambda_0(t) \geq 0$, proposition 3 can be used to see $H(t) \geq 0$ and hence one-sided tests can be applied to detect Cox's model.

Most conditions required for theorems 1, 2 and proposition 3 can be easily transformed to the special situation we are considering. However, the condition $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\boldsymbol{0}} = O_P(n^{-1/2})$ needs special attention. If Cox's model (H_c) holds, asymptotic normality of $\hat{\boldsymbol{\beta}}$ is shown in Andersen & Gill (1982 p. 1105). This implies $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\boldsymbol{0}} = O_P(n^{-1/2})$. For our purposes we need to know the asymptotic behaviour of $\hat{\boldsymbol{\beta}}$ under H_0 as well. Asymptotic normality of $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^a)$ for some $\boldsymbol{\beta}^a$ under misspecified models has been considered previously by Lin & Wei (1989), Sasieni (1993) and Fine (2002) under an i.i.d. setup allowing only one event per individual. We relax these requirements but confine ourselves to showing $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^a = O_P(n^{-1/2})$ under H_0, where

$$\boldsymbol{\beta}^a := \arg\max_{\boldsymbol{\beta}\in\mathcal{B}} a(\boldsymbol{\beta}, \tau)$$

and $a(\boldsymbol{\beta}, t) := \int_0^t \left(\boldsymbol{\beta}^{\mathrm{T}} \overrightarrow{\boldsymbol{ZY}}(s) - \log(\overrightarrow{\boldsymbol{d}}(\boldsymbol{\beta}, s)) \overrightarrow{\boldsymbol{Y}}(s)^{\mathrm{T}}\right) \boldsymbol{\alpha}(s)\, \mathrm{d}s$. The main idea for the proof of this is similar to the proof under H_c given in Andersen & Gill (1982). First, uniform stochastic convergence of $X(\boldsymbol{\beta}, t) := (1/n)C(\boldsymbol{\beta}, t) + (\log n)\overline{N}(t)$ to $a(\boldsymbol{\beta}, t)$ is shown. Since $X(\boldsymbol{\beta}, \tau)$ is concave in $\boldsymbol{\beta}$, convex analysis can be used to transfer the convergence of $X(\boldsymbol{\beta}, \tau)$ to its maximizer $\hat{\boldsymbol{\beta}}$. A suitable Taylor expansion of $X(\boldsymbol{\beta}, \tau)$ around $\boldsymbol{\beta}^a$ allows the statement about $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^a)$.

Lemma 3
Suppose that

(i) for each $\boldsymbol{\beta} \in \mathcal{B}$, $\overline{\boldsymbol{d}}(\boldsymbol{\beta}, \cdot)$ *converges uip on* $\mathfrak{T}$ *and the mapping* $\overrightarrow{\boldsymbol{d}}(\boldsymbol{\beta}, \cdot)$ *is bounded away from* 0,
(ii) $\overline{\boldsymbol{Y}}$, $\overline{\boldsymbol{YZ}}$ *and* $\overline{\boldsymbol{ZZY}}$ *converge uip on* $\mathfrak{T}$,
(iii) $\boldsymbol{\beta}^a$ *exists and is unique.*

If H_0 *holds then* $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}^a$. *If furthermore there exists an open* $\mathcal{C} \subset \mathcal{B}$ *with* $\boldsymbol{\beta}^a \in \mathcal{C}$ *such that*

(iv) $\overline{\boldsymbol{d}}$, $\overline{\boldsymbol{Zd}}$, $\overline{\boldsymbol{ZdZ}}$ *converge uip on* $\mathcal{C} \times \mathfrak{T}$,
(v) $\int_0^\tau (\overrightarrow{\boldsymbol{d}}(\boldsymbol{\beta}^a, s)^{-2}\overrightarrow{\boldsymbol{Zd}}^{\otimes 2}(\boldsymbol{\beta}^a, s) - \overrightarrow{\boldsymbol{d}}(\boldsymbol{\beta}^a, s)^{-1}\overrightarrow{\boldsymbol{ZdZ}}(\boldsymbol{\beta}^a, s))\overrightarrow{\boldsymbol{Y}}(s)^{\mathrm{T}}\boldsymbol{\alpha}(s)\,\mathrm{d}s$ *is invertible and*
(vi) $\overline{\boldsymbol{d}}(\boldsymbol{\beta}^a, t) - \overrightarrow{\boldsymbol{d}}(\boldsymbol{\beta}^a, t)$, $\overline{\boldsymbol{Zd}}(\boldsymbol{\beta}^a, t) - \overrightarrow{\boldsymbol{Zd}}(\boldsymbol{\beta}^a, t), \overline{\boldsymbol{Y}}(t) - \overrightarrow{\boldsymbol{Y}}(t), \overline{\boldsymbol{ZY}}(t) - \overrightarrow{\boldsymbol{ZY}}(t) = O_P(n^{-1/2})$ *uniformly in* $t \in \mathfrak{T}$

then $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^a = O_P(n^{-\frac{1}{2}})$

Above, we used the notation $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}$ for column vectors $\boldsymbol{a}$. The proof is relegated to appendix C.

5. The i.i.d. case

In this section, we want to show consistency of our tests in the case of i.i.d. observations. The main tool we use is the following. Let $\mathcal{A}$ be the set of processes x whose paths are left-continuous with right-hand limits and satisfy $E\sup_{t\in\mathfrak{T}} |x(t)|^2 < \infty$. Suppose (a_i, b_i), $i \in \mathbb{N}$ are i.i.d. and $a_i, b_i \in \mathcal{A}$ for all $i \in \mathbb{N}$. Then

$$E\sup_{t\in\mathfrak{T}} |a_1(t)b_1(t)| \le (E\sup_{t\in\mathfrak{T}} |a_1(t)|^2 E\sup_{t\in\mathfrak{T}} |b_1(t)|^2)^{0.5} < \infty$$

and hence by the strong law of large numbers of Rao (1963),

$$\sup_{t\in\mathfrak{T}} \left| \frac{1}{n}\sum_{i=1}^{n} a_i(t)b_i(t) - E(a_1(t)b_1(t)) \right| \to 0 \quad \text{almost surely.}$$

Hence, $\overline{\boldsymbol{ab}}$ converges uip on $\mathfrak{T}$ and $\overrightarrow{\boldsymbol{ab}}(t) = E[a_1(t)b_1(t)]$. For fixed $t \in \mathfrak{T}$, we will interpret $a_1(t)$ and $b_1(t)$ as elements of the Hilbert space $L_2 := L_2(P)$ of square integrable random variables with the usual scalar product $<\cdot,\cdot>_2$ and thus we may write

$$\overrightarrow{\boldsymbol{ab}}(t) = <a_1(t), b_1(t)>_2.$$

The following lemma shows how the orthogonal projection $\boldsymbol{Q}(t) : \mathbb{R}^n \to \mathbb{R}^n$ carries over to the orthogonal projection $Q_t : L_2 \to L_2$ onto $\operatorname{span}(Y_{11}(t), \ldots, Y_{1k}(t))^{\perp}$.

Lemma 4
Suppose $(a_i, b_i, Y_{i1}, \ldots, Y_{ik})$, $i \in \mathbb{N}$ *are i.i.d. and* $a_i, b_i, Y_{i1}, \ldots, Y_{ik} \in \mathcal{A}$.

If $(<Y_{1v}(t), Y_{1\mu}(t)>_2)_{v,\mu=1,\ldots,k}$ *is invertible for all* $t \in \mathfrak{T}$ *and continuous in* $t \in \mathfrak{T}$, *then* $\overline{\boldsymbol{a}(\boldsymbol{Qb})}$ *converges uip on* $\mathfrak{T}$ *and*

$$\overrightarrow{\boldsymbol{a}(\boldsymbol{Qb})}(t) = <a_1(t), Q_t b_1(t)>_2.$$

Proof. As already noted, $\overline{\boldsymbol{ab}}$, $\overline{\boldsymbol{aY}}$, $\overline{\boldsymbol{YY}}$ and $\overline{\boldsymbol{Yb}}$ converge uip on $\mathfrak{T}$ and

$$\overrightarrow{\boldsymbol{ab}}(t) = < a_1(t), b_1(t) >_2, \qquad \overrightarrow{\boldsymbol{aY}}(t) = (< a_1(t),\ Y_{1j}(t) >_2)^{\mathrm{T}}_{j=1,\ldots,k},$$

$$\overrightarrow{\boldsymbol{YY}}(t) = (< Y_{1\nu}(t),\ Y_{1\mu}(t) >_2)_{\nu,\,\mu=1,\ldots,k}, \qquad \overrightarrow{\boldsymbol{Yb}}(t) = (< Y_{1j}(t), b_1(t) >_2)_{j=1,\ldots,k}.$$

By lemma 8, $\overline{\boldsymbol{a(Qb)}}$ converges uip on $\mathfrak{T}$ and

$$\overrightarrow{\boldsymbol{a(Qb)}}(t) = < a_1(t), b_1(t) >_2 - (< a_1(t),\ Y_{1j}(t) >_2)^{\mathrm{T}}_j (\overrightarrow{\boldsymbol{YY}}(t))^{-1} (< Y_{1j}(t), b_1(t) >_2)_j.$$

It remains to use the linearity of $< \cdot, \cdot >_2$ and lemma 12 to accomplish the proof.

We want to use the above to show consistency of our tests for the cases of section 4. Generally speaking, we are interested in showing $H(t) = 0$ for all $t \in \mathfrak{T}$ iff Aalen's model (H_0) holds true.

Consider the setup of section 4.1 where $d_i(t)$ is a uniformly consistent estimator of $\lambda_i(t)$. Suppose $(d_i,\ \lambda_i,\ Y_{i1}, \ldots, Y_{ik})$ are i.i.d. and $d_i,\ \lambda_i,\ Y_{i1}, \ldots, Y_{ik} \in \mathcal{A}$. Then under the conditions of lemma 1,

$$H(t) = \int_0^t < \lambda_1(s), Q_s \lambda_1(s) >_2 \,\mathrm{d}s = \int_0^t \| Q_s \lambda_1(s) \|_2^2 \,\mathrm{d}s,$$

where $\|x\|_2^2 = < x,\ x >_2$. Hence, $H(t) = 0$ iff $\lambda_1(t) \in \mathrm{span}(Y_{11}(t), \ldots,\ Y_{1k}(t))$ for almost all $t \in \mathfrak{T}$, i.e. Aalen's model holds true.

Now consider the setup of section 4.2 for detecting an additional covariate. Assume $(X_{i1}, \ldots, X_{i,\,k+k'})$ are i.i.d. and $X_{i1}, \ldots, X_{i,\,k+k'} \in \mathcal{A}$. If $< X_{1,\,k+1},\ X_{k+v} >_2 = 0$ for $2 \le v \le k'$ then under the conditions of lemma 2 by proposition 2

$$\begin{aligned} H(t) &= \int_0^t < X_{1,\,k+1}, Q_s \lambda_1(s) >_2 \,\mathrm{d}s = \int_0^t < X_{1,\,k+1}(s), Q_s X_{1,\,k+1}(s) >_2 \beta_{k+1}(s)\,\mathrm{d}s \\ &= \int_0^t \| Q_s X_{1,\,k+1}(s) \|_2^2 \beta_{k+1}(s)\,\mathrm{d}s. \end{aligned}$$

Note that in this case since β_{k+1} may change signs on $\mathfrak{T}$ the test statistic $T^{(1)}$ based on $\hat{T}(\tau)$ does not ensure a consistent test whereas the sup-based tests $T^{(2)}$ and $T^{(3)}$ ensure consistency.

If in the setup of section 4.3, where we want to detect Cox's model, we assume that $(d_i(\boldsymbol{\beta}^{\mathbf{0}}, \cdot),\ Y_{i1}, \ldots,\ Y_{ik})$ are i.i.d. and $d_i(\boldsymbol{\beta}^{\mathbf{0}}, \cdot),\ Y_{i1}, \ldots,\ Y_{ik} \in \mathcal{A}$ then we may use proposition 3 to get $H(t) = \int_0^t \| Q_s d_1(\boldsymbol{\beta}^{\mathbf{0}},\ s) \|_2^2 \lambda_0(s)\,\mathrm{d}s$ and hence

$$H(t) = 0\ \forall t \in \mathfrak{T} \text{ iff } (\lambda_0(t) = 0 \text{ or } d_1(\boldsymbol{\beta}^{\mathbf{0}}, t) \in \mathrm{span}(Y_{11}(t), \ldots,\ Y_{1k}(t)) \text{ for almost all } t \in \mathfrak{T}).$$

6. Simulation results

Our simulation study uses true models which were also considered in McKeague & Utikal (1991). As covariates, we take independent random variables x_i, $i = 1, \ldots, n$ that are uniformly distributed on [0, 1]. The simulation is for classical survival analysis, i.e. we have $\lambda_i(t) = 0$ if $N_i(t-) = 1$. We assume independent right censoring with i.i.d. random variables C_i, $i = 1, \ldots, n$, following an exponential distribution with parameter chosen such that 27% of the observations before τ are censored. Let $R_i(t) := I\{C_i > t,\ N_i(t-) = 0\}$. In our simulations we have $\boldsymbol{Y}_i(t) = (1,\ x_i) R_i(t)$.

First, we consider $\boldsymbol{d}$ chosen as in section 4.3 to make our tests powerful against Cox's model. As covariates $\boldsymbol{Z}$ for Cox's model we use $\boldsymbol{Z}_i(t) = (x_i)$. Table 1 gives levels and powers at the asymptotic 5 per cent level for the tests of section 3. We also display some results from the

Table 1. *Observed levels and powers for tests against Cox's model as alternative (section 4.3) with asymptotic level 5%. The number of samples was 10,000. We used* $\tau = 2$*. The test based on* $T^{(1)}$ *is one-sided. Results from McKeague & Utikal (1991, Table 1b) are also displayed (sample size 1000)*

	Observed level $[\lambda_i(t) = (1 + x_i)R_i(t)]$				Observed power $[\lambda_i(t) = 0.5\exp(2x_i)R_i(t)]$			
n	$T^{(1)}$	$T^{(2)}$	$T^{(3)}$	McKU	$T^{(1)}$	$T^{(2)}$	$T^{(3)}$	McKU
75	0.0350	0.0272	0.0341		0.1693	0.0262	0.0538	
150	0.0430	0.0399	0.0431		0.2996	0.0838	0.1354	
300	0.0452	0.0420	0.0442	0.212	0.5198	0.2286	0.3084	0.243
600	0.0490	0.0472	0.0459		0.8060	0.5320	0.6276	
1200	0.0468	0.0485	0.0486	0.106	0.9734	0.8847	0.9256	0.579

Table 2. *Observed levels and powers using* $d_i = I\{x_i \notin [0.25, 0.75]\}$ *and* $\tau = 10$ *with asymptotic level 5%. Tests are as derived in section 3. The number of samples was 10,000. The test based on* $T^{(1)}$ *is two-sided. Results from McKeague & Utikal (1991, Table 1b) are also displayed (sample size 1000)*

	Observed level $[\lambda_i(t) = (1 + x_i)R_i(t)]$			Observed power $[\lambda_i(t) = \min(x_i, 1 - x_i)R_i(t)]$			
n	$T^{(1)}$	$T^{(2)}$	$T^{(3)}$	$T^{(1)}$	$T^{(2)}$	$T^{(3)}$	McKU
50	0.0417	0.0187	0.0269	0.8746	0.7170	0.8194	
100	0.0455	0.0287	0.0346	0.9959	0.9828	0.9919	
180	0.0455	0.0348	0.0397	1.0000	1.0000	1.0000	0.912

goodness-of-fit test suggested by McKeague & Utikal (1991, Table 1b). In the simulation where the true model is an Aalen model most observed levels are close to or below the nominal level of 5 per cent. The tests are conservative for small n. If $\lambda_i(t) = 0.5\exp(2x_i)R_i(t)$ then the power increases as n increases with best results for the one-sided test based on $T^{(1)}$.

We also consider the alternative $\lambda_i(t) = \min(x_i, 1 - x_i)R_i(t)$. To detect this alternative we choose $d_i = I\{x_i \notin [0.25, 0.75]\}$. Results of the simulations are in Table 2. The test is sensitive against this alternative for small sample sizes. In the simulations where H_0 holds, the observed level is close to or below 5 per cent again.

Comparing these results to those of McKeague & Utikal (1991) we see that our tests are much better at attaining the prescribed level in the simulations in which H_0 holds. Furthermore, we get a greater power against the stated alternatives. Of course, the greater power is not surprising since the test of McKeague & Utikal (1991) is an omnibus test and our test was designed to detect these specific alternatives.

7. Application to real data sets

We applied our tests to three real data sets. The first from software reliability is chosen to illustrate the applicability of the methods outside classical survival analysis where we may have multiple events per point process. In the second example we can show that the fit of an additive model suggested by Grønnesby & Borgan (1996) to the PBC data (Fleming & Harrington, 1991) is not ideal. In the last example we see that for the well-known Stanford heart transplant data Miller & Halpern (1982) an Aalen model cannot be rejected when tested against the standard Cox model.

Table 3. *p-values for the software reliability data sets from Gandy & Jensen (2004) using $d_i(t) = 1$. The test based on $T^{(1)}$ is two-sided*

Test statistic	$T^{(1)}$	$T^{(2)}$	$T^{(3)}$
One covariate	0.0100	0.0022	0.0155
Three covariates	0.7805	0.3655	0.5914

7.1. A data set from software reliability

We start by considering a data set from software reliability used in Gandy & Jensen (2004). It contains bug reports of 73 open source software projects. Two different sets of covariates were considered in the aforementioned paper. The first includes only the current size of the source code of the projects as covariate. The second includes the size of the source code of the projects a fixed time ago, changes in the size of source code since then and the number of recent bug reports. Note that in both cases no baseline in the form of a covariate identically equal to 1 was included. Checking this using $\boldsymbol{d}(t) = \mathbf{1}$ as suggested in section 4.2 we get the p-values of Table 3 for our tests applied to the two different sets of covariates. In the data set with one covariate all three tests suggest a bad fit of the model that could possibly be improved by including a baseline. In the case of three covariates the hypothesis that the Aalen model is the correct one is supported. This agrees with the conclusion of Gandy & Jensen (2004).

7.2. PBC data

Our next example considers the PBC data presented in Fleming & Harrington (1991), where it is analysed at length using Cox's model. It contains data about the survival of 312 patients with PBC. We use the corrections of the data set given by Fleming & Harrington (1991 p. 188). The final Cox model of Fleming & Harrington (1991) uses the covariates age, oedema, log(albumin), log(bilirubin) and log(prothrombin time).

In Grønnesby & Borgan (1996) the data set is analysed with Aalen's model using the covariates baseline, bilirubin, oedema dichotomized (0.5 pooled together with 0), albumin (zero for the highest half then linear), prothrombin time (zero for the lowest half then linear), age, interaction of age and prothrombin time. Grønnesby & Borgan (1996) investigate the fit of the linear model and the final Cox model of Fleming & Harrington (1991) and conclude that 'the fit of both models is acceptable'. Their formal goodness-of-fit test for the linear model yields a p-value of 0.075 for Aalen's model and 0.197 for the Cox model. Furthermore, they mention that the linear model 'suffers from [...] negative estimated intensities'.

Using the just mentioned coding of covariates for both models our one-sided test of Aalen's model against Cox's model based on $T^{(1)}$ leads to a p-value of 0.034 rejecting Aalen's model at the 5 per cent level.

7.3. Stanford heart transplant data

The last example is concerned with the Stanford heart transplant data given by Miller & Halpern (1982). We consider those $n = 157$ patients receiving heart transplants with complete records. Two covariates are of interest: age at time of transplant (which we denote a_i) and a donor–recipient mismatch score (which we denote b_i). Among the models fitted by Miller & Halpern (1982) are several Cox models. The mismatch score b_i is not significant in a Cox model with covariates $\mathbf{Z}_i = (a_i, b_i)$. Based on a graphical method they state that the fit of the model with covariates $\mathbf{Z}_i = (a_i)$ is 'not ideal'. To improve the fit they consider a model based on the covariates $\mathbf{Z}_i = (a_i, a_i^2)$ for which they find no lack of fit. These findings are supported using formal tests by Lin *et al.* (1993) and Marzec & Marzec (1997).

In the sequel we fit and test Aalen's additive model. Let $R_i(t)$ denote the at-risk indicator of the ith patient. If we use the model $\boldsymbol{Y}_i(t) = (1, a_i)R_i(t)$ our two-sided test based on $T^{(1)}$ using $d_i(t) = b_i$ is not significant (p-value 0.401) whereas the test using $d_i(t) = a_i^2$ is significant (p-value 0.004). Testing against Cox's model with the one-sided test based on $T^{(1)}$ and $\mathbf{Z}_i = (a_i)$ is significant as well (p-value 0.0167). If we include a_i^2 as suggested by our test, our test against Cox's model with the one-sided test based on $T^{(1)}$ and $\mathbf{Z}_i = (a_i, a_i^2)$ is not significant (p-value 0.153). It is not clear whether the Cox model or the additive model has a better fit.

8. Conclusions and outlook

As mentioned in the introduction goodness-of-fit for Aalen's additive risk model has been considered before only by a few authors who concentrated on smaller classes of models, on graphical methods or presented a so-called omnibus test, which is not directed against specific alternatives. In this paper, we propose goodness-of-fit tests based on martingale residuals which can be adjusted to detect particular alternatives. They are characterized by the following properties:

- The test statistics are asymptotically distribution free.
- The asymptotic distribution of the test statistics can be determined under both the null and alternative hypotheses.
- In the important i.i.d. case consistency can be proven. Note that in this case most of the required conditions are satisfied if the covariates are bounded.
- The tests can be tailored to detect specific alternatives, in particular Cox's model, by an appropriate choice of $\boldsymbol{d}(t)$.
- No estimate of the baseline $\lambda_0(t)$ is needed in the test against Cox's model.

The next to last property gives some freedom to choose $\boldsymbol{d}(t)$ and the natural question arises how two different choices could be compared and what could be an optimal choice then. This question needs further investigation.

For some tests, as for example for the test against Cox's model, we have to insert estimates for $\boldsymbol{\beta}^{\mathbf{0}}$, which destroy the property of predictability of $\boldsymbol{d}(t)$. But fortunately, it can be shown (see section 2) that the resulting test statistics are asymptotically equivalent to those with the true parameter. To assess how much power is lost due to this estimation in the Cox model we conducted some simulation studies which indicate that the loss is small. For example in the setup of section 6 with $\lambda_i(t) = 0.5\exp(2x_i)R_i(t)$ and a sample size of $n = 300$ we used $d_i(t) = \exp(2x_i)$ and got a p-value of 0.5267 for the one-sided test based on $T^{(1)}$, the simulation from section 6 with the estimated parameter resulted in a p-value of 0.5198.

The simulation studies in section 6 showed that even for moderate sample sizes the prescribed level is met. Of course, no general suggestions for the sample size to attain a certain power can be made, because this depends on the alternative to be detected.

In McKeague & Sasieni (1994), a restriction of Aalen's model is discussed, where some of the covariates are required to have time-independent influence. Formally, the model is given by

$$H_S: \quad \boldsymbol{\lambda}(t) = \boldsymbol{Y}^c(t)\boldsymbol{\alpha}^c + \boldsymbol{Y}^v(t)\boldsymbol{\alpha}^v(t)$$

for some $\boldsymbol{\alpha}^c \in \mathbb{R}^{k_c}$ and some deterministic, bounded, measurable $\boldsymbol{\alpha}^v : [0, \tau] \to \mathbb{R}^{k_v}$, where $\boldsymbol{Y}^v(t) = (Y_{ij}^v(t))$ is an $n \times k_v$ matrix of locally bounded predictable processes and $\boldsymbol{Y}^c(t) = (Y_{ij}^c(t))$ is an $n \times k_c$ matrix of locally bounded predictable processes. Of course, the test proposed earlier can be used for this submodel as well since H_S is a restriction of Aalen's additive model. However, the test statistic can be modified to make use of the information that $\boldsymbol{\alpha}^c$ is time-independent. As a new test statistic we propose

$$\hat{T} = \frac{1}{\sqrt{n}} \int_0^\tau \boldsymbol{d}(s)^{\mathrm{T}} \boldsymbol{Q}_v(s) \left(\mathrm{d}\boldsymbol{N}(s) - \boldsymbol{Y}^c(s)\hat{\boldsymbol{\alpha}}^c \, \mathrm{d}s \right)$$

where

$$\hat{\boldsymbol{\alpha}}^c = \left(\int_0^\tau \boldsymbol{Y}^c(s)^{\mathrm{T}} \boldsymbol{Q}_v(s) \boldsymbol{Y}^c(s) \, \mathrm{d}s \right)^{-1} \int_0^\tau \boldsymbol{Y}^c(s)^{\mathrm{T}} \boldsymbol{Q}_v(s) \, \mathrm{d}\boldsymbol{N}(s)$$

and $\boldsymbol{Q}_v(s) := \boldsymbol{I} - \boldsymbol{Y}^v(s)(\boldsymbol{Y}^v(s)^{\mathrm{T}}\boldsymbol{Y}^v(s))^{-1}\boldsymbol{Y}^v(s)^{\mathrm{T}}$. It can be shown that $\hat{T}$ contains an orthogonal projection in a suitable space. A detailed discussion of the properties of this test statistic would go beyond the scope of this article and will appear elsewhere.

Of course, there are many other regression models (e.g. Martinussen & Scheike, 2002; Scheike & Zhang, 2002), to which our tests could also be adjusted.

Acknowledgements

We thank the editor Thomas Scheike, an Associate Editor and a referee for their constructive comments which helped improve the content of the paper. Financial support of the Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged.

References

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

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

Aalen, O. O. (1993). Further results on the non-parametric linear regression model in survival analysis. *Statist. Med.* **12**, 1569–1588.

Andersen, P. K. & Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100–1120.

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993). *Statistical models based on counting processes*. Springer-Verlag, New York.

Billingsley, P. (1999). *Convergence of probability measures*, 2nd edn. Wiley, New York.

Borodin, A. N. & Salminen, P. (2002). *Handbook of Brownian motion – facts and formulae*, 2nd edn. Birkhäuser, Basel.

Cox, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34**, 187–220.

Fine, J. P. (2002). Comparing nonnested Cox models. *Biometrika* **89**, 635–647.

Fleming, T. R. & Harrington, D. P. (1991). *Counting processes and survival analysis*. Wiley, New York.

Gandy, A. & Jensen, U. (2004). A nonparametric approach to software reliability. *Appl. Stoch. Models Bus. Ind.* **20**, 3–15.

Grønnesby, J. K. & Borgan, Ø. (1996). A method for checking regression models in survival analysis based on the risk score. *Lifetime Data Anal.* **2**, 315–328.

Hall, W. & Wellner, J. (1980). Confidence bands for a survival curve from censored data. *Biometrika* **67**, 133–144.

Huffer, F. W. & McKeague, I. W. (1991). Weighted least squares estimation of Aalen's additive risk model. *J. Am. Statist. Assoc.* **86**, 114–129.

Kim, J., Song, M. S. & Lee, S. (1998). Goodness-of-fit tests for the additive risk model with ($p > 2$)-dimensional time-invariant covariates. *Lifetime Data Anal.* **4**, 405–416.

Lin, D. Y. & Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *J. Amer. Statist. Assoc.* **84**, 1074–1078.

Lin, D. Y., Wei, L. J. & Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–572.

Martinussen, T. & Scheike, T. H. (2002). A flexible additive multiplicative hazard model. *Biometrika* **89**, 283–298.

Marzec, L. & Marzec, P. (1997). Generalized martingale-residual processes for goodness-of-fit inference in Cox's type regression models. *Ann. Statist.* **25**, 683–714.

McKeague, I. W. (1988a). Asymptotic theory for weighted least squares estimators in Aalen's additive risk model. *Contemp. Math.* **80**, 139–152.

McKeague, I. W. (1988b). A counting process approach to the regression analysis of grouped survival data. *Stochastic Process. Appl.* **28**, 221–239.

McKeague, I. W. & Sasieni, P. D. (1994). A partly parametric additive risk model. *Biometrika* **81**, 501–514.

McKeague, I. W. & Utikal, K. J. (1991). Goodness-of-fit tests for additive hazards and proportional hazards models. *Scand. J. Statist.* **18**, 177–195.

Miller, R. & Halpern, J. (1982). Regression with censored data. *Biometrika* **69**, 521–531.

Rao, R. R. (1963). A law of large numbers for D[0,1]-valued random variables. *Theor. Probab. Appl.* **8**, 70–74.

Rudin, W. (1974). *Real and complex analysis*, 2nd edn. McGraw-Hill, New York.

Sasieni, P. (1993). Some new estimators for Cox regression. *Ann. Statist.* **21**, 1721–1759.

Scheike, T. H. & Zhang, M.-J. (2002). An additive multiplicative Cox–Aalen regression model. *Scand. J. Statist.* **29**, 75–88.

Song, M.-U., Jeong, D.-M. & Song, J.-K. (1996). Checking the additive risk model with martingale residuals. *J. Korean Statist. Soc.* **25**, 433–444.

Yuen, K. C. & Burke, M. D. (1997). A test of fit for a semiparametric additive risk model. *Biometrika* **84**, 631–639.

Received January 2004, in final form February 2005

Uwe Jensen, Institut für Angewandte Mathematik und Statistik (110), Universität Hohenheim, 70593 Stuttgart, Germany.
E-mail: jensen@uni-hohenheim.de

Appendix

Appendix A. Convergence of inverted matrices

This section shows how the uniform stochastic convergence of time-dependent random matrices can be carried over to their inverses. We use this to get uniform stochastic convergence of $(\overline{\boldsymbol{YY}}(t))^{-1}$ from that of $\overline{\boldsymbol{YY}}(t)$. A similar lemma was already stated in McKeague (1988a) without detailed proof. We correct a slight error and give a full proof.

Lemma 5

Suppose $\boldsymbol{A}^{(n)}(t)$, $t \in \mathfrak{T}$, $n \in \mathbb{N}$ are $k \times k$ matrices of random processes. If there exists a continuous function $\boldsymbol{a} : \mathfrak{T} \to \mathbb{R}^{k\times k}$ such that $\boldsymbol{a}(t)$ is invertible for all $t \in \mathfrak{T}$ and $\sup_{t\in\mathfrak{T}} \|\boldsymbol{A}^{(n)}(t) - \boldsymbol{a}(t)\| \xrightarrow{P} 0$ then

(i) $P(\boldsymbol{A}^{(n)}(t)$ *is invertible* $\forall t \in \mathfrak{T}) \xrightarrow{P} 1$,

(ii) $\exists K > 0$ *s.t.* $P(\| (\boldsymbol{A}^{(n)}(t))^{-1} \| < K \forall t \in \mathfrak{T}) \xrightarrow{P} 1$ *and*

(iii) $\sup_{t\in\mathfrak{T}} \|(\boldsymbol{A}^{(n)}(t))^{-1} - \boldsymbol{a}^{-1}(t)\| \xrightarrow{P} 0$.

To prove lemma 5 we need the following two lemmas.

Lemma 6

Let $0 < \tau < \infty$, $\mathfrak{T} = [0, \tau]$ and $k \in \mathbb{N}$. If $\boldsymbol{a} : \mathfrak{T} \to \mathbb{R}^{k\times k}$ is a continuous mapping such that $\boldsymbol{a}(s)$ is invertible for all $s \in \mathfrak{T}$ then there exists an $\epsilon > 0$ such that for all $\boldsymbol{B} : \mathfrak{T} \to \mathbb{R}^{k\times k}$, $\sup_{s\in\mathfrak{T}} \|\boldsymbol{a}(s) - \boldsymbol{B}(s)\| < \epsilon$ implies $\boldsymbol{B}(s)$ invertible $\forall s \in \mathfrak{T}$.

Proof. Since $\boldsymbol{a}(\mathfrak{T})$ is compact, the set of invertible matrices $\mathrm{GL} \subset \mathbb{R}^{k\times k}$ is open and $\boldsymbol{a}(\mathfrak{T}) \subset \mathrm{GL}$ we can find $s_1, \dots, s_v \in \mathfrak{T}$ and $\delta_1, \dots, \delta_v > 0$ such that $\boldsymbol{a}(\mathfrak{T}) \subset \bigcup_{i=1}^{v} U(\boldsymbol{a}(s_i), \delta_i)$ and $U(\boldsymbol{a}(s_i), 2\delta_i) \subset \mathrm{GL}$, where $U(\boldsymbol{C}, \xi) := \{\boldsymbol{D} \in \mathbb{R}^{k\times k} : \|\boldsymbol{C} - \boldsymbol{D}\| < \xi\}$. It can be verified that $\epsilon := \min\{\delta_1, \dots, \delta_v\}$ satisfies the claim.

Lemma 7

Let $p \in \mathbb{N}$. If $\boldsymbol{C}, \boldsymbol{D} \in \mathbb{R}^{p\times p}$ are invertible and $\|\boldsymbol{D}^{-1}\|\|\boldsymbol{D} - \boldsymbol{C}\| < 1$ then

$$\|\boldsymbol{C}^{-1} - \boldsymbol{D}^{-1}\| \leq \frac{\|\boldsymbol{D}^{-1}\|^2\|\boldsymbol{C} - \boldsymbol{D}\|}{1 - \|\boldsymbol{D}^{-1}\|\|\boldsymbol{C} - \boldsymbol{D}\|}. \tag{4}$$

Proof. Let $\boldsymbol{I}$ denote the unit matrix in $\mathbb{R}^{p\times p}$.

$$\begin{aligned}\|\boldsymbol{C}^{-1} - \boldsymbol{D}^{-1}\| &= \|\boldsymbol{C}^{-1}(\boldsymbol{D} - \boldsymbol{C})\boldsymbol{D}^{-1}\| \leq \|\boldsymbol{C}^{-1}\|\|\boldsymbol{C} - \boldsymbol{D}\|\|\boldsymbol{D}^{-1}\| \\ &= \|\boldsymbol{C}^{-1}\boldsymbol{D}\boldsymbol{D}^{-1}\|\|\boldsymbol{C} - \boldsymbol{D}\|\|\boldsymbol{D}^{-1}\| \leq \|\boldsymbol{C}^{-1}\boldsymbol{D}\|\|\boldsymbol{D}^{-1}\|^2\|\boldsymbol{C} - \boldsymbol{D}\|.\end{aligned}$$

By the assumption, $\|\boldsymbol{D}^{-1}(\boldsymbol{D} - \boldsymbol{C})\| \leq \|\boldsymbol{D}^{-1}\|\|\boldsymbol{D} - \boldsymbol{C}\| < 1$. Hence,

$$\begin{aligned}\|\boldsymbol{C}^{-1}\boldsymbol{D}\| &= \|(\boldsymbol{D}^{-1}\boldsymbol{C})^{-1}\| = \|(\boldsymbol{I} - (\boldsymbol{D}^{-1}(\boldsymbol{D} - \boldsymbol{C})))^{-1}\| = \|\sum_{n=0}^{\infty}(\boldsymbol{D}^{-1}(\boldsymbol{D} - \boldsymbol{C}))^n\| \\ &\leq \sum_{n=0}^{\infty}\|\boldsymbol{D}^{-1}(\boldsymbol{D} - \boldsymbol{C})\|^n = \left(1 - \|\boldsymbol{D}^{-1}(\boldsymbol{D} - \boldsymbol{C})\|\right)^{-1} \leq \left(1 - \|\boldsymbol{D}^{-1}\|\|\boldsymbol{C} - \boldsymbol{D}\|\right)^{-1}.\end{aligned}$$

Proof of lemma 5. Choose $\epsilon > 0$ as in lemma 6. Then

$$P(\exists t \in \mathfrak{T} \text{ s.t. } \boldsymbol{A}^{(n)}(t) \text{ is singular}) \leq P(\|\boldsymbol{a} - \boldsymbol{A}^{(n)}\| \geq \epsilon) \xrightarrow{P} 0.$$

Since $\boldsymbol{a}$, taking inverses and $\|\cdot\|$ are continuous mappings, the compactness of $\mathfrak{T}$ implies that $\{\|\boldsymbol{a}^{-1}(s)\| : s \in \mathfrak{T}\}$ is compact. Hence, there exists a constant $L > 0$ such that $\sup_{t\in\mathfrak{T}} \|\boldsymbol{a}^{-1}(t)\| \leq L$.

On the event $D_n := \{\|\boldsymbol{a}^{-1}(t)\|\|\boldsymbol{a}(t) - \boldsymbol{A}^{(n)}(t)\| < \frac{1}{2}, \boldsymbol{A}^{(n)}(t) \text{ invertible } \forall t \in \mathfrak{T}\}$ we can use lemma 7 to see that $\forall s \in \mathfrak{T}$, $\|\boldsymbol{a}^{-1}(s) - (\boldsymbol{A}^{(n)}(s))^{-1}\| \leq 2L^2\|\boldsymbol{a}(s) - \boldsymbol{A}^{(n)}(s)\|$. Since $D_n \supset \{\|\boldsymbol{a}(t) - \boldsymbol{A}^{(n)}(t)\| < (2L)^{-1}, \boldsymbol{A}^{(n)}(t) \text{ invertible } \forall t \in \mathfrak{T}\}$, we have $P(D_n) \to 1$. Hence, $\sup_{t\in\mathfrak{T}} \|(\boldsymbol{A}^{(n)}(t))^{-1} - \boldsymbol{a}^{-1}(t)\| \xrightarrow{P} 0$.

Let $K := L + 1$. Since $\|(\boldsymbol{A}^{(n)}(t))^{-1}\| \leq L + \|(\boldsymbol{A}^{(n)}(t))^{-1} - \boldsymbol{a}^{-1}(t)\|$, we get (iii).

Remark 1. In McKeague (1988b p. 231), it is stated that $\|\boldsymbol{C} - \boldsymbol{D}\| < \|\boldsymbol{D}\|$ (instead of $\|\boldsymbol{D}^{-1}\| \|\boldsymbol{D} - \boldsymbol{C}\| < 1$) implies (4). This is not true as the following example shows:

Let $\boldsymbol{D} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{pmatrix}$, $\boldsymbol{C} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{3} \end{pmatrix}$. Then $\boldsymbol{C}$ and $\boldsymbol{D}$ are invertible and $\|\boldsymbol{C} - \boldsymbol{D}\| = \frac{1}{2} < 1 = \|\boldsymbol{D}\|$. But $\|\boldsymbol{D}^{-1}\| = \|\begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}\| = 3$ and hence, $1 - \|\boldsymbol{D}^{-1}\|\|\boldsymbol{C} - \boldsymbol{D}\| = 1 - \frac{3}{2} < 0$, which shows that the right-hand side of (4) is negative.

Appendix B. Some technical lemmas

In this section we present some technical lemmas mainly concerned with convergence and projections. We assume that we are in the setup introduced at the beginning of section 2.

Lemma 8
Let $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\lambda} : \mathfrak{T} \to \mathbb{R}^n$ be random functions.

1. *If (D1) holds and $\overline{\boldsymbol{ab}}$, $\overline{\boldsymbol{aY}}$ and $\overline{\boldsymbol{bY}}$ converge uip on $\mathfrak{T}$ then $\overline{\boldsymbol{a}(\boldsymbol{Qb})} = \overline{(\boldsymbol{Qa})\boldsymbol{b}}$ converges uip on $\mathfrak{T}$ and $\overrightarrow{\boldsymbol{a}(\boldsymbol{Qb})}(t) = \overrightarrow{(\boldsymbol{Qa})\boldsymbol{b}}(t) = \overrightarrow{\boldsymbol{ab}}(t) - \overrightarrow{\boldsymbol{aY}}(t)(\overrightarrow{\boldsymbol{YY}}(t))^{-1}\overrightarrow{\boldsymbol{Yb}}(t)$ for all $t \in \mathfrak{T}$.*
2. *If (D1) holds and $\overline{\boldsymbol{a\lambda b}}$, $\overline{\boldsymbol{aY}}$, $\overline{\boldsymbol{bY}}$, $\overline{\boldsymbol{a\lambda Y}}$, $\overline{\boldsymbol{b\lambda Y}}$ and $\overline{\boldsymbol{Y\lambda Y}}$ converge uip on $\mathfrak{T}$ then $\overline{(\boldsymbol{Qa})\boldsymbol{\lambda}(\boldsymbol{Qb})}$ converges uip on $\mathfrak{T}$.*

Proof. By (D1) we can use lemma 5 (see appendix A) to find $K > 0$ such that for the event $A_n := \{\overline{\boldsymbol{YY}}(t)$ is invertible, $\|(\overline{\boldsymbol{YY}}(t))^{-1}\| < K \forall t \in \mathfrak{T}\}$ we have $P(A_n) \to 1$. On A_n the projection $\boldsymbol{Q}(t)$ can be written as $\boldsymbol{Q}(t) = \boldsymbol{I} - \boldsymbol{Y}(t)(\overline{\boldsymbol{YY}}(t))^{-1}\frac{1}{n}\boldsymbol{Y}(t)^{\mathrm{T}}$ for all $t \in \mathfrak{T}$. Dropping the dependence on $t \in \mathfrak{T}$, on A_n this implies $\overline{\boldsymbol{a}(\boldsymbol{Qb})} = \overline{\boldsymbol{ab}} - \overline{\boldsymbol{aY}}(\overline{\boldsymbol{YY}})^{-1}\overline{\boldsymbol{Yb}}$ and $\overline{(\boldsymbol{Qa})\boldsymbol{\lambda}(\boldsymbol{Qb})} = \overline{\boldsymbol{a\lambda b}} - \overline{\boldsymbol{aY}}(\overline{\boldsymbol{YY}})^{-1}\overline{\boldsymbol{Y\lambda b}} - \overline{\boldsymbol{a\lambda Y}}(\overline{\boldsymbol{YY}})^{-1}\overline{\boldsymbol{Yb}} - \overline{\boldsymbol{aY}}(\overline{\boldsymbol{YY}})^{-1}\overline{\boldsymbol{Y\lambda Y}}(\overline{\boldsymbol{YY}})^{-1}\overline{\boldsymbol{Yb}}$. Lemma 5 together with the assumptions (which include the boundedness of the limits) finish the proof.

Lemma 9
Suppose (D1) holds, $\overline{\boldsymbol{aY}}$ converges uip on $\mathfrak{T}$ and for some $\alpha > 0$ we have $n^{-\alpha}\sup_{i=1,\dots,n;\, t\in\mathfrak{T}} |a_i(t)| \xrightarrow{P} 0$, $n^{-\alpha}\sup_{i=1,\dots,n;\, t\in\mathfrak{T}} |Y_{ij}(t)| \xrightarrow{P} 0$ for $j = 1, \dots, k$. Then

$$n^{-\alpha} \sup_{\substack{i=1,\dots,n \\ t\in\mathfrak{T}}} |(\boldsymbol{Q}(t)\boldsymbol{a}(t))_i| \xrightarrow{P} 0.$$

Proof. Let A_n be defined as in the proof of lemma 8. On A_n we have for all $t \in \mathfrak{T}$ and $i = 1, \dots, n$,

$$\begin{aligned} |(\boldsymbol{Q}(t)\boldsymbol{a}(t))_i| &= \left|\sum_{j=1}^n Q_{ij}(t)a_j(t)\right| = \left|a_i(t) - \boldsymbol{Y}_i(t)(\overline{\boldsymbol{YY}}(t))^{-1}\overline{\boldsymbol{Ya}}(t)\right| \\ &\leq |a_i(t)| + \|\boldsymbol{Y}_i(t)\| \left\|(\overline{\boldsymbol{YY}}(t))^{-1}\right\| \|\overline{\boldsymbol{Ya}}(t)\| \\ &\leq |a_i(t)| + \sqrt{k} \max_{j=1,\dots,k} |Y_{ij}(t)| K \|\overline{\boldsymbol{Ya}}(t)\|. \end{aligned}$$

Lemma 10
Let $\mathcal{B} \subset \mathbb{R}^p$ be an open set, $a_n, a : \mathcal{B} \times \mathfrak{T} \to \mathbb{R}$ be random functions. Suppose $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}^{\mathbf{0}} \in \mathcal{B}$, and suppose there exist an open neighbourhood $\mathcal{C} \subset \mathcal{B}$ of $\boldsymbol{\beta}^{\mathbf{0}}$ such that $a_n \xrightarrow{P} a$ uniformly on $\mathcal{C} \times \mathfrak{T}$ and $a(\cdot, t) : \mathcal{B} \to \mathbb{R}$ is continuous at $\boldsymbol{\beta}^{\mathbf{0}}$ uniformly in $t \in \mathfrak{T}$. Then $a_n(\hat{\boldsymbol{\beta}}, t) \xrightarrow{P} a(\boldsymbol{\beta}^{\mathbf{0}}, t)$ uniformly in $t \in \mathfrak{T}$.

Proof. For all $t \in \mathfrak{T}$, $|a_n(\hat{\boldsymbol{\beta}}, t) - a(\boldsymbol{\beta}^{\mathbf{0}}, t)| \leq |a_n(\hat{\boldsymbol{\beta}}, t) - a(\hat{\boldsymbol{\beta}}, t)| + |a(\hat{\boldsymbol{\beta}}, t) - a(\boldsymbol{\beta}^{\mathbf{0}}, t)|$. The continuity of a shows that the second term converges to 0 uniformly on $\mathfrak{T}$. Since $\mathcal{C}$ is an open neighbourhood of $\boldsymbol{\beta}^{\mathbf{0}}$, $P(\hat{\boldsymbol{\beta}} \in \mathcal{C}) \to 1$ and hence the convergence of a_n implies $|a_n(\hat{\boldsymbol{\beta}}, t) - a(\boldsymbol{\beta}^{\mathbf{0}}, t)| \xrightarrow{P} 0$ uniformly in $t \in \mathfrak{T}$.

Lemma 11
If $e_i : \mathfrak{T} \to \mathbb{R}$, $i \in \mathbb{N}$ are random functions, $\sup_{i=1,\dots,n;\, t\in\mathfrak{T}} |e_i(t)| \xrightarrow{P} 0$ and $\overline{\boldsymbol{\lambda}}$ converges uip on $\mathfrak{T}$ then (assuming that the integrals exist)

$$\frac{1}{n}\int_0^t \boldsymbol{e}(s)^{\mathrm{T}}\,\mathrm{d}\boldsymbol{N}(s) \xrightarrow{P} 0 \quad \text{and} \quad \frac{1}{n}\int_0^t \boldsymbol{e}(s)^{\mathrm{T}}\,\mathrm{d}\boldsymbol{M}(s) \xrightarrow{P} 0 \quad \textit{uniformly in } t \in \mathfrak{T}.$$

Proof. Since $\langle \overline{\boldsymbol{M}}\rangle(\tau) = \langle \frac{1}{n}\sum_{i=1}^{n}(N_i(\cdot) - \int_0^{\cdot} \lambda_i(s)\,\mathrm{d}s)\rangle(\tau) = \frac{1}{n^2}\sum_{i=1}^{n}\int_0^{\tau}\lambda_i(s)\,\mathrm{d}s = \frac{1}{n}\int_0^{\tau}\overline{\boldsymbol{\lambda}}(s)\,\mathrm{d}s \xrightarrow{P} 0$, Lenglart's inequality implies $\overline{\boldsymbol{M}} \to 0$ uip on $\mathfrak{T}$. Therefore,

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^t |\mathrm{d}N_i(s)| = \overline{\boldsymbol{N}}(t) = \overline{\boldsymbol{M}}(t) + \int_0^t \overline{\boldsymbol{\lambda}}(s)\,\mathrm{d}s \xrightarrow{P} \int_0^t \overrightarrow{\boldsymbol{\lambda}}(s)\,\mathrm{d}s$$

uniformly in $t \in \mathfrak{T}$. Hence,

$$\left|\frac{1}{n}\int_0^t \boldsymbol{e}(s)^{\mathrm{T}}\,\mathrm{d}\boldsymbol{N}(s)\right| \le \frac{1}{n}\sum_{i=1}^{n}\left|\int_0^t e_i(s)\,\mathrm{d}N_i(s)\right| \le \frac{1}{n}\sum_{i=1}^{n}\int_0^t |\mathrm{d}N_i(s)| \sup_{\substack{i=1,\ldots,n\\ s\in\mathfrak{T}}} |e_i(s)| \xrightarrow{P} 0$$

uniformly in $t \in \mathfrak{T}$. The second statement can be shown similarly.

Lemma 12

Suppose H is a vector space over $\mathbb{R}$ with scalar product $<\cdot,\cdot>$. If $y_1,\ldots,y_k \in H$ are such that the matrix $\boldsymbol{A} := (<y_\eta, y_\xi>)_{\eta,\,\xi=1,\ldots,k}$ is invertible then

$$Q: H \to H, x \mapsto x - (y_1,\ldots,y_k)\boldsymbol{A}^{-1}\left(<y_1,x>,\ldots,<y_k,x>\right)^{\mathrm{T}}$$

is the orthogonal projection onto the space orthogonal to G, where G is the space spanned by $y_1,\ldots,y_k$. Furthermore, $<Qx, y> = <x, Qy> = <Qx, Qy>$ for all $x, y \in H$.

Proof. For $x \in H$ and $i \in \{1,\ldots,k\}$,

$$\begin{aligned}<y_i, Qx> &= <y_i, x> - \left(<y_i,y_1>,\ldots,<y_i,y_k>\right)\boldsymbol{A}^{-1}\left(<y_1,x>,\ldots,<y_k,x>\right)^{\mathrm{T}}\\ &= <y_i,x> - <y_i,x> = 0.\end{aligned}$$

Hence Q maps into $G^{\perp}$. Clearly $P: H \to H$, $x \mapsto x - Qx$ maps into G. Since P and Q are linear, Q is the orthogonal projection onto $G^{\perp}$ (see e.g. Rudin, 1974, p. 84). The remainder of the lemma are properties of orthogonal projections.

Appendix C. Proofs

Proof of theorem 2. First, we prove the convergence of $T^* - T$. By Taylor's theorem,

$$\begin{aligned}T^*(t) - T(t) &= n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}})^{\mathrm{T}}\frac{1}{n}\int_0^t \left(\boldsymbol{Q}(s)\nabla\boldsymbol{d}(\boldsymbol{\beta}^{\mathbf{0}},s)\right)^{\mathrm{T}}\,\mathrm{d}\boldsymbol{M}(s)\\ &\quad + n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}})^{\mathrm{T}}\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\int_0^t \left(n^{-\frac{1}{2}}\boldsymbol{Q}(s)\nabla_{\upsilon}\nabla_{\eta}\boldsymbol{d}(\tilde{\boldsymbol{\beta}},s)\right)_i \mathrm{d}M_i(s)\right)_{\upsilon\eta} n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}}),\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}$ is between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^{\mathbf{0}}$. Note that $\tilde{\boldsymbol{\beta}}$ depends on s, ω and i.

We will show that both terms on the right-hand side converge to 0 uip on $\mathfrak{T}$.

$$\begin{aligned}n\left\langle \frac{1}{n}\int_0^t \left(\boldsymbol{Q}(s)\nabla\boldsymbol{d}(\boldsymbol{\beta}^{\mathbf{0}},s)\right)^{\mathrm{T}}\,\mathrm{d}\boldsymbol{M}(s)\right\rangle &= \frac{1}{n}\int_0^t \left(\boldsymbol{Q}(s)\nabla\boldsymbol{d}(\boldsymbol{\beta}^{\mathbf{0}},s)\right)^{\mathrm{T}}\ \mathrm{diag}(\boldsymbol{\lambda}(s))\boldsymbol{Q}(s)\nabla\boldsymbol{d}(\boldsymbol{\beta}^{\mathbf{0}},s)\,\mathrm{d}s\\ &= \int_0^t \overline{(\boldsymbol{Q}\nabla\boldsymbol{d})\boldsymbol{\lambda}(\boldsymbol{Q}\nabla\boldsymbol{d})}(\boldsymbol{\beta}^{\mathbf{0}},s)\,\mathrm{d}s\end{aligned}$$

which converges uip on $\mathfrak{T}$ by lemma 8. By Lenglart's inequality we can conclude that

$$\frac{1}{n}\int_0^{\cdot} \left(\boldsymbol{Q}(s)\nabla \boldsymbol{d}(\boldsymbol{\beta}^{\mathbf{0}}, s)\right)^{\mathrm{T}} \,\mathrm{d}\boldsymbol{M}(s) \to 0 \quad \text{uip on } \mathfrak{T}.$$

For $v, \eta = 1, \dots, p$ we have that $\overline{(\nabla_v \nabla_\eta \boldsymbol{d})\boldsymbol{Y}}(\tilde{\boldsymbol{\beta}}, \cdot)$ converges uip on $\mathfrak{T}$ by lemma 10 and thus lemma 9 implies $n^{-1/2} \sup_{i=1,\dots,n, \, t \in \mathfrak{T}} |(\boldsymbol{Q}(t)\nabla_v \nabla_\eta \boldsymbol{d}(\tilde{\boldsymbol{\beta}}, t))_i| \xrightarrow{P} 0$. Hence, we can use the assumptions together with lemma 11 to get

$$\frac{1}{n}\sum_{i=1}^{n} \int_0^{\cdot} \left(n^{-1/2}\boldsymbol{Q}(s)\nabla_v \nabla_\eta \boldsymbol{d}(\tilde{\boldsymbol{\beta}}, s)\right)_i \,\mathrm{d}M_i(s) \to 0 \text{ uip on } \mathfrak{T}.$$

Next, we show the convergence of $\hat{G}^*(t) - \hat{G}(t)$. Since

$$\frac{\partial}{\partial \beta_v}\left(\frac{1}{n}\int_0^t \boldsymbol{d}(\boldsymbol{\beta}, s)^{\mathrm{T}}\boldsymbol{Q}(s) \,\mathrm{diag}(\mathrm{d}\boldsymbol{N}(s))\boldsymbol{Q}(s)\boldsymbol{d}(\boldsymbol{\beta}, s)\right)$$
$$= \frac{1}{n}\int_0^t 2(\nabla_v \boldsymbol{d})(\boldsymbol{\beta}, s)^{\mathrm{T}}\boldsymbol{Q}(s) \,\mathrm{diag}(\mathrm{d}\boldsymbol{N}(s))\boldsymbol{Q}(s)\boldsymbol{d}(\boldsymbol{\beta}, s),$$

a Taylor approximation yields

$$\hat{G}^*(t) - \hat{G}(t) = n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}})^{\mathrm{T}} \frac{1}{n}\int_0^t 2n^{-1/2}(\nabla \boldsymbol{d})(\tilde{\boldsymbol{\beta}}, s)^{\mathrm{T}}\boldsymbol{Q}(s) \,\mathrm{diag}(\mathrm{d}\boldsymbol{N}(s))\boldsymbol{Q}(s)\boldsymbol{d}(\tilde{\boldsymbol{\beta}}, s)$$
$$= n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}})^{\mathrm{T}} \frac{1}{n}\sum_{i=1}^{n}\int_0^t 2\left(n^{-1/4}\boldsymbol{Q}(s)(\nabla \boldsymbol{d})(\tilde{\boldsymbol{\beta}}, s)\right)_i \left(n^{-1/4}\boldsymbol{Q}(s)\boldsymbol{d}(\tilde{\boldsymbol{\beta}}, s)\right)_i \mathrm{d}N_i(s)$$

with $\tilde{\boldsymbol{\beta}}$ between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^{\mathbf{0}}$. Lemma 9 together with lemma 10 shows that we can use lemma 11 to get that $\hat{G}^* - \hat{G}$ converges to zero uip on $\mathfrak{T}$.

Proof of lemma 3. Let $\boldsymbol{\beta} \in \mathcal{B}$ and $X(\boldsymbol{\beta}, t) := (1/n)C(\boldsymbol{\beta}, t) + (\log n)\overline{N}(t)$. It can be shown that $X(\boldsymbol{\beta}, t)$ is concave in $\boldsymbol{\beta} \in \mathcal{B}$ for each $t \in \mathfrak{T}$. Furthermore,

$$X(\boldsymbol{\beta}, t) = \frac{1}{n}\int_0^t \left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}(s)^{\mathrm{T}} - \log(\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s))\mathbf{1}^{\mathrm{T}}\right) \mathrm{d}\boldsymbol{N}(s)$$
$$= \frac{1}{n}\int_0^t \left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}(s)^{\mathrm{T}} - I\{\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s) > 0\}\log(\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s))\mathbf{1}^{\mathrm{T}}\right) \mathrm{d}\boldsymbol{N}(s)$$

and hence for each $\boldsymbol{\beta} \in \mathcal{B}$, $X(\boldsymbol{\beta}, t) - A(\boldsymbol{\beta}, t)$ is a local square integrable martingale, where

$$A(\boldsymbol{\beta}, t) := \frac{1}{n}\int_0^t \left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}(s)^{\mathrm{T}} - I\{\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s) > 0\}\log(\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s))\mathbf{1}^{\mathrm{T}}\right) \boldsymbol{\lambda}(s)\,\mathrm{d}s.$$

For $\boldsymbol{\beta} \in \mathcal{B}$,

$$nB(\boldsymbol{\beta}, t) := n\langle X(\boldsymbol{\beta}, \cdot) - A(\boldsymbol{\beta}, \cdot)\rangle(t)$$
$$= \frac{1}{n}\int_0^t \left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}(s)^{\mathrm{T}} - I\{\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s) > 0\}\log(\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s))\mathbf{1}^{\mathrm{T}}\right)\mathrm{diag}(\boldsymbol{\lambda}(s))$$
$$\times \left(\boldsymbol{Z}(s)\boldsymbol{\beta} - I\{\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s) > 0\}\log(\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s))\mathbf{1}\right) \mathrm{d}s$$
$$= \int_0^t \left(\boldsymbol{\beta}^{\mathrm{T}}\overline{\boldsymbol{Z\lambda Z}}(s)\boldsymbol{\beta} - 2\log(\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s))I\{\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s) > 0\}\boldsymbol{\alpha}(s)^{\mathrm{T}}\overline{\boldsymbol{YZ}}(s)\boldsymbol{\beta}\right.$$
$$\left. + I\{\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s) > 0\}\log^2(\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s))\overline{\boldsymbol{Y}}(s)^{\mathrm{T}}\boldsymbol{\alpha}(s)\right)\mathrm{d}s$$
$$\xrightarrow{P} \int_0^t \left(\boldsymbol{\beta}^{\mathrm{T}}\overrightarrow{\boldsymbol{Z\lambda Z}}(s)\boldsymbol{\beta} - 2\log(\overrightarrow{\boldsymbol{d}}(\boldsymbol{\beta}, s))\boldsymbol{\alpha}(s)^{\mathrm{T}}\overrightarrow{\boldsymbol{YZ}}(s)\boldsymbol{\beta}\right.$$
$$\left. + \log^2(\overrightarrow{\boldsymbol{d}}(\boldsymbol{\beta}, s))\overrightarrow{\boldsymbol{Y}}(s)^{\mathrm{T}}\boldsymbol{\alpha}(s)\right)\mathrm{d}s$$

uniformly in $t \in \mathfrak{T}$. The convergence of $\overline{\boldsymbol{Z\lambda Z}}$ is implied by that of $\overline{\boldsymbol{ZZY}}$. Since $\overrightarrow{\boldsymbol{d}}$ is bounded away from 0, the convergence of $\overline{\boldsymbol{d}}$ implies the convergence of $I\{\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s)>0\}$ and $\log(\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s))$.

By the above and Lenglart's inequality (see e.g. Andersen & Gill, 1982, appendix I), $\sup_{t\in\mathfrak{T}} |X(\boldsymbol{\beta}, t) - A(\boldsymbol{\beta}, t)| \xrightarrow{P} 0$ for all $\boldsymbol{\beta} \in \mathcal{B}$. Furthermore, since

$$A(\boldsymbol{\beta}, t) = \int_0^t \left(\boldsymbol{\beta}^{\mathrm{T}}\overline{\boldsymbol{YZ}}^{\mathrm{T}}(s)\boldsymbol{\alpha}(s) - I\{\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s)>0\}\log(\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s))\overline{\boldsymbol{Y}}(s)^{\mathrm{T}}\boldsymbol{\alpha}(s)\right) \mathrm{d}s$$

we can use assumptions (i) and (ii) to get $\sup_{t\in\mathfrak{T}} |A(\boldsymbol{\beta}, t) - a(\boldsymbol{\beta}, t)| \xrightarrow{P} 0$. Hence, $\sup_{t\in\mathfrak{T}} |X(\boldsymbol{\beta}, t) - a(\boldsymbol{\beta}, t)| \xrightarrow{P} 0$. By Andersen & Gill (1982 Cor.II.2) this implies $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^a \xrightarrow{P} 0$.

To show $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^a = O_P(n^{-1/2})$ we proceed as follows. Let

$$U(\boldsymbol{\beta}, t) := \frac{\partial}{\partial \boldsymbol{\beta}} X(\boldsymbol{\beta}, t) = \frac{1}{n}\int_0^t \boldsymbol{Z}(s)\,\mathrm{d}\boldsymbol{N}(s) - \int_0^t \overline{\boldsymbol{d}}(\boldsymbol{\beta}, s)^{-1}\overline{\boldsymbol{Zd}}(\boldsymbol{\beta}, s)\,\mathrm{d}\overline{\boldsymbol{N}}(s),$$

$$J(\boldsymbol{\beta}, t) := \left(\frac{\partial}{\partial \boldsymbol{\beta}}\right)^2 X(\boldsymbol{\beta}, t) = \int_0^t \left(\overline{\boldsymbol{d}}(\boldsymbol{\beta}, s)^{-2}\overline{\boldsymbol{Zd}}(\boldsymbol{\beta}, s)^{\otimes 2} - \overline{\boldsymbol{d}}(\boldsymbol{\beta}, s)^{-1}\overline{\boldsymbol{ZdZ}}(\boldsymbol{\beta}, s)\right) \mathrm{d}\overline{\boldsymbol{N}}(s).$$

A Taylor expansion of U around $\boldsymbol{\beta}^a$ yields

$$U(\boldsymbol{\beta}, t) - U(\boldsymbol{\beta}^a, t) = J(\tilde{\boldsymbol{\beta}}, t)(\boldsymbol{\beta} - \boldsymbol{\beta}^a)$$

for some $\tilde{\boldsymbol{\beta}}$ between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^a$. By definition of $\hat{\boldsymbol{\beta}}$ we have $U(\hat{\boldsymbol{\beta}}, \tau) = 0$ and hence

$$-n^{1/2}U(\boldsymbol{\beta}^a, \tau) = J(\tilde{\boldsymbol{\beta}}, \tau)n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^a).$$

We will show $U(\boldsymbol{\beta}^a, \tau) = O_P(n^{-1/2})$ and $J(\tilde{\boldsymbol{\beta}}, \tau) \xrightarrow{P} \overrightarrow{J}(\boldsymbol{\beta}^a, \tau)$ where

$$\overrightarrow{J}(\boldsymbol{\beta}, t) := \int_0^t \left(\overrightarrow{\boldsymbol{d}}(\boldsymbol{\beta}, s)^{-2}\overrightarrow{\boldsymbol{Zd}}(\boldsymbol{\beta}, s)^{\otimes 2} - \overrightarrow{\boldsymbol{d}}(\boldsymbol{\beta}, s)^{-1}\overrightarrow{\boldsymbol{ZdZ}}(\boldsymbol{\beta}, s)\right)\overrightarrow{\boldsymbol{Y}}(s)^{\mathrm{T}}\boldsymbol{\alpha}(s)\,\mathrm{d}s.$$

Since we assumed $\overrightarrow{J}(\boldsymbol{\beta}^a, \tau)$ to be invertible we will get $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^a = O_P(n^{-1/2})$.

The convergence of $J(\hat{\boldsymbol{\beta}}, t)$ is immediate from $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}^a$, lemma 10 and the assumptions.

The boundedness of $U(\hat{\boldsymbol{\beta}}, \tau)$ needs some more work. Let

$$\begin{aligned} V(t) &:= \frac{1}{n}\int_0^t \boldsymbol{Z}(s)\boldsymbol{\lambda}(s)\,\mathrm{d}s - \int_0^t \overline{\boldsymbol{d}}(\boldsymbol{\beta}^a, s)^{-1}\overline{\boldsymbol{Zd}}(\boldsymbol{\beta}^a, s)\overline{\boldsymbol{\lambda}}(s)\,\mathrm{d}s \\ &= \int_0^t \left(\overline{\boldsymbol{Z\lambda}}(s) - \overline{\boldsymbol{d}}(\boldsymbol{\beta}^a, s)^{-1}\overline{\boldsymbol{Zd}}(\boldsymbol{\beta}^a, s)\overline{\boldsymbol{Y}}(s)\right)\boldsymbol{\alpha}(s)\,\mathrm{d}s \end{aligned}$$

and

$$\overrightarrow{V}(t) := \int_0^t \left(\overrightarrow{\boldsymbol{Z\lambda}}(s) - \overrightarrow{\boldsymbol{d}}(\boldsymbol{\beta}^a, s)^{-1}\overrightarrow{\boldsymbol{Zd}}(\boldsymbol{\beta}^a, s)\overrightarrow{\boldsymbol{Y}}(s)\right)\boldsymbol{\alpha}(s)\,\mathrm{d}s.$$

By definition of $\boldsymbol{\beta}^a$ we have $\overrightarrow{V}(\tau) = 0$. Hence,

$$n^{1/2}U(\boldsymbol{\beta}^a, \tau) = n^{1/2}(U(\boldsymbol{\beta}^a, \tau) - V(\tau)) + n^{\frac{1}{2}}(V(\tau) - \overrightarrow{V}(\tau)). \tag{5}$$

We show that both terms on the right-hand side are stochastically bounded. Since

$$\langle n^{1/2}\left(U(\boldsymbol{\beta}^a, \cdot) - V(\cdot)\right)\rangle(\tau) = \int_0^\tau \left(\overline{\boldsymbol{Z\lambda Z}}(s) - \overline{\boldsymbol{d}}(\boldsymbol{\beta}^a, s)^{-2}\overline{\boldsymbol{Zd}}(\boldsymbol{\beta}^a, s)^{\otimes 2}\overline{\boldsymbol{\lambda}}(s)\right) \mathrm{d}s$$

converges in probability, $\langle n^{1/2}\left(U(\boldsymbol{\beta}^a, \cdot) - V(\cdot)\right)\rangle(\tau)$ is stochastically bounded. Hence, using Lenglart's inequality, we can conclude that $n^{1/2}\left(U(\boldsymbol{\beta}^a, t) - V(t)\right)$ is stochastically bounded uniformly in $t \in \mathfrak{T}$. The second term on the right-hand side of (5) can be dealt with as follows.

Since

$$V(\tau) - \overrightarrow{V}(\tau) = \int_0^\tau \Big[\Big(\overline{\boldsymbol{ZY}}(s) - \overrightarrow{\boldsymbol{ZY}}(s) \Big) + \Big(\overline{\boldsymbol{d}}(\boldsymbol{\beta}^a, s)^{-1} \overline{\boldsymbol{Zd}}(\boldsymbol{\beta}^a, s) \overline{\boldsymbol{Y}}(s) - \overrightarrow{\boldsymbol{d}}(\boldsymbol{\beta}^a, s)^{-1} \overrightarrow{\boldsymbol{Zd}}(\boldsymbol{\beta}^a, s) \overrightarrow{\boldsymbol{Y}}(s) \Big) \Big] \boldsymbol{\alpha}(s)\, \mathrm{d}s,$$

the assumptions imply $V(\tau) - \overrightarrow{V}(\tau) = O_P(n^{-\frac{1}{2}})$.