Bernoulli **13**(1), 2007, 279–299
DOI: 10.3150/07-BEJ6068

Estimating optimal step-function approximations to instantaneous hazard rates

MOULINATH BANERJEE[1] and IAN W. MCKEAGUE[2]

[1]*Department of Statistics, University of Michigan, 1085 South University, Ann Arbor MI 48109, USA*

[2]*Department of Biostatistics, Columbia University, 722 West 168th Street, New York NY 10032, USA. E-mail: im2131@columbia.edu*

We investigate the problem of estimating the best binary decision tree approximation to the baseline hazard function in the Cox proportional hazards model. Our motivation is to find an effective way of condensing key functional information in the baseline hazard into a small number of estimable parameters. The parameters consist of a threshold and two hazard levels, one to the left of the threshold and one to the right, defined in terms of the best L^2 approximation to the nonparametric baseline hazard function. Estimators of these parameters are introduced and shown to converge at cube-root rate to a non-normal limit distribution. Two alternate ways of constructing confidence intervals for the threshold are compared. Results from a simulation study and an example concerning a threshold for the age of onset of schizophrenia in a large cohort study are discussed.

Keywords: binary decision tree; change-point; cube root; misspecified model; proportional hazards; split point

1. Introduction

Many people become informed about studies of disease risk through their mainstream media. For effective communication of public health information of this type, it is crucial to report the key statistical conclusions in ways that are understandable to non-scientists (Brownson and Remington 2002). The Cox proportional hazards model is often suitable in this regard because it provides an estimate of instantaneous relative risk r for an exposed individual compared with an unexposed individual (holding all other risk factors constant) that only involves a single regression parameter, and r does not depend on time. Information about the *baseline* hazard rate $\lambda(t)$, however, is not as easily reported because it depends on time, often in a complicated fashion, and it can be difficult to interpret plots of the Breslow estimator of the cumulative baseline hazard (e.g. Figure 1) and kernel estimates of $\lambda(t)$ itself. Yet all the information about how disease risk evolves temporally is contained in the baseline, so it would be helpful to find an effective way of condensing that information into a form that can be communicated easily. In the present paper we investigate how this can be done in terms of the best-fitting binary decision tree

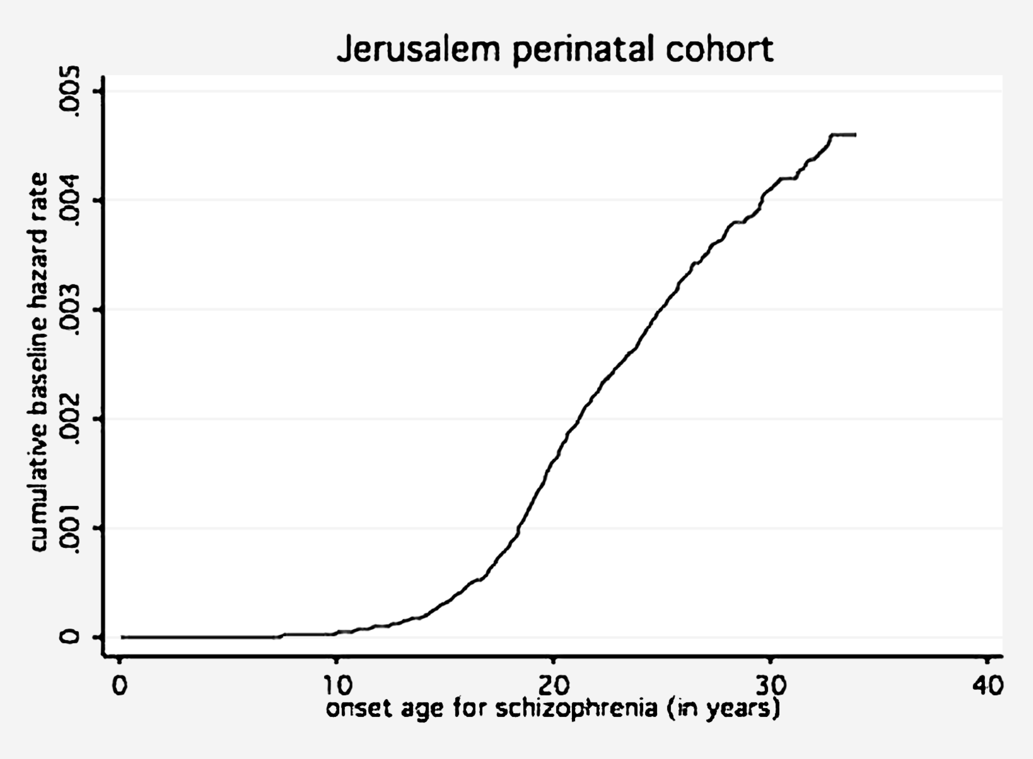

Figure 1. Breslow estimator of the cumulative baseline hazard function for the onset age of schizophrenia in the Jerusalem perinatal cohort.

approximation to $\lambda(t)$. Binary decision trees are step-functions with a single jump and have just three parameters, all immediately interpretable, so they provide an excellent means of condensing the information in the instantaneous risk into a tractable form.

We take the point of view that to effectively extract information about $\lambda(t)$, as t varies over a specified follow-up period, it is important to use a global approximation with only a few parameters. As such, a binary decision tree provides a compelling *working model* for $\lambda(t)$, and a suitable compromise between interpretability and predictive power. We are interested in the parameters of the best-fitting working model, that is, the parameter values that minimize the L^2 distance between the binary decision tree and the true $\lambda(t)$. The binary decision tree is defined in terms of a threshold and two hazard levels, one to the left of the threshold and one to the right. This threshold is similar to the split point used in classification and regression trees (CART). In recent work, Banerjee and McKeague (2006) studied split point estimation in the setting of nonparametric regression, and applied the approach to the estimation of a pollution threshold; see also Bühlmann and Yu (2002). The present paper is the first attempt to develop split point methods in a semiparametric setting for the purpose of condensing information in the nonparametric part of the model.

It is well known that binary decision trees have poor predictive power in comparison with other learning methods, but are more attractive in terms of interpretability (see Hastie *et al.* 2001: 313). In the present context, however, interpretability and condensation of information override the need for *local* predictive power. To appreciate this point, note that a local abrupt change in $\lambda(t)$ would not necessarily coincide closely with the threshold in the best binary decision tree approximation unless that is its main global feature. Moreover, the binary decision tree threshold parameter exists even when no abrupt changes are present in $\lambda(t)$ (see the simulation example in Section 3). Indeed, our approach does not make any assumptions about the presence of abrupt changes in $\lambda(t)$, and is complimentary to change-point analysis in which the aim is to estimate the locations of assumed jump discontinuities in an otherwise smooth curve.

Change-point methods are well developed in the nonparametric regression and survival analysis literature; see Chang *et al.* (1994), Müller and Wang (1994), Gijbels *et al.* (1999), Antoniadis *et al.* (2000), Antoniadis and Gijbels (2002), Dempfle and Stute (2002), and Wu *et al.* (2003). Change-point models of proportional hazards type have been studied by Luo *et al.* (1997), Pons (2002, 2003) and Dupuy (2006). In change-point analysis, however, no distinction is made between the working model that has the jump point and the model that is assumed to generate the data. In contrast, in the present setting we need to develop a model-robust approach that applies under misspecification of the discontinuous working model by a smooth curve. Under a misspecified Cox model, Breslow's estimator and the maximum partial likelihood estimator for the regression parameters are known to converge at $\sqrt{n}$ rate (Lin and Wei 1989); in that case, however, the working model is the Cox model itself and the misspecification is general.

The main result of this paper shows that the estimators of the three parameters in the working binary decision tree model converge at cube-root rate to a non-normal continuous limit distribution (a scaled Chernoff distribution). We examine two alternate ways of constructing confidence intervals for the threshold, one based on the usual Wald approach and the other by inverting a deviance statistic. The cube-root rate is in marked contrast to change-point estimators which converge at rate n under the optimistic assumption that the change-point model is correctly specified.

The paper is organized as follows. The main results are presented in Section 2. In Section 3 we describe results from a simulation study and an application to estimating a threshold for the onset age for schizophrenia based on data from a large cohort study. In Section 4 we discuss the broader implications of our results and prospects for future work on misspecified threshold models in other survival analysis settings. Proofs are contained in Section 5.

2. Threshold estimation under proportional hazards

The Cox proportional hazards model assumes that the conditional hazard function for the failure time T of an individual with a p-vector of covariates Z can be written as

$$\lambda(t|Z) = \lambda(t)\exp\{\beta^{\mathrm{T}} Z\},$$

where $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a p-vector of unknown regression coefficients and $\lambda(t)$ is an unspecified baseline hazard function. Here β is restricted to a compact neighborhood $\mathcal{B}$ of the true value β_0, and the covariates are assumed to be bounded. Let $X = \min\{T, C\}$ be a possibly right-censored failure time, where T and the censoring time C are assumed to be conditionally independent given Z. Let $\delta = 1\{T \leqslant C\}$ denote the indicator that the failure time is observed.

To effectively condense the information in the baseline hazard function we need to consider an approximating family of parametrically specified functions with the parameters having an appealing interpretation. We propose to use a binary tree approximation identified by a 3-vector of parameters $(\lambda_l, \lambda_u, d)$, where d is the threshold (or jump point), λ_l is the value to the left of the jump, and λ_u is the value to the right of the jump. The subscripts in λ_l and λ_u stand for 'lower' and 'upper' approximations, respectively. Best projected values of these parameters in the L^2 sense are defined by

$$(\lambda_l^0, \lambda_u^0, d^0) = \underset{\lambda_l, \lambda_u, d}{\operatorname{argmin}} \int_0^\tau \left[\lambda(t) - \overline{\lambda}(t; \lambda_l, \lambda_u, d)\right]^2 \mathrm{d}t, \tag{2.1}$$

where

$$\overline{\lambda}(t; \lambda_l, \lambda_u, d) = \lambda_l 1(t \leqslant d) + \lambda_u 1(t > d)$$

is the binary tree approximation to $\lambda(t)$ and $\tau > 0$ is a given terminal time. The optimal threshold d^0 is the main parameter of interest; it most accurately splits the time interval into two subintervals with the risk changing abruptly at the boundary.

Suppose we have n independent and identically distributed observations (X_i, δ_i, Z_i) of (X, δ, Z). It is natural to estimate $(\lambda_l^0, \lambda_u^0, d^0)$ by expressing the integral in (2.1) in terms of the cumulative baseline hazard function $\Lambda(t) = \int_0^t \lambda(u)\,\mathrm{d}u$ and plugging in its Breslow estimator

$$\hat{\Lambda}_n(t) = \mathbb{P}_n\left[\frac{\delta 1\{X \leqslant t\}}{S^{(0)}(\hat{\beta}, X)}\right].$$

Here $\mathbb{P}_n$ is the empirical distribution of the observations, $\hat{\beta}$ the maximum partial likelihood estimator of β_0, $S^{(0)}(\beta, t) = \mathbb{P}_n[Y(t)\mathrm{e}^{\beta^{\mathrm{T}} Z}]$, and $Y(t) = 1\{X \geqslant t\}$ is the at-risk indicator. In the case of no covariates, the Breslow estimator reduces to the Nelson–Aalen estimator.

We need to apply various results of Andersen and Gill (1982) on the asymptotic distribution of $\hat{\beta}$ and $\hat{\Lambda}_n(t)$, for which it suffices to assume that the matrix Σ (in the notation of their Theorem 4.1) is positive definite, $\lambda(t)$ is bounded on $[0, \tau]$, and $P(X > \tau) > 0$. Note that, under these conditions, $s^{(0)}(\beta, t) = \mathrm{E}[Y(t)\mathrm{e}^{\beta^{\mathrm{T}} Z}]$ is bounded away from zero as a function of $(\beta, t) \in \mathcal{B} \times [0, \tau]$. In addition, we need the following two conditions to establish our main results:

Conditions

(A1) There is a unique vector $(\lambda_l^0, \lambda_u^0, d^0)$ with $\lambda_l^0 \neq \lambda_u^0$ and $0 < d^0 < \tau$ that minimizes the integral on the right-hand side of (2.1).

(A2) λ is continuously differentiable in a neighbourhood of d^0, and $\lambda'(d^0) \neq 0$.

Under the above conditions we find the following set of normal equations:

$$\lambda_l^0 = \frac{\Lambda(d^0)}{d^0}, \qquad \lambda_u^0 = \frac{\Lambda(\tau) - \Lambda(d^0)}{\tau - d^0}, \qquad \lambda(d^0) = \frac{\lambda_l^0 + \lambda_u^0}{2},$$

obtained by setting the partial derivatives of the integral on the right-hand side of (2.1) with respect to $(\lambda_l, \lambda_u, d)$ to zero. The parameters λ_l^0 and λ_u^0 are seen to have the attractive interpretation as the mean hazard levels in the regions separated by the threshold d^0.

Expanding the integral in (2.1) and discarding the terms not involving $(\lambda_l, \lambda_u, d)$ shows that

$$(\lambda_l^0, \lambda_u^0, d^0) = \operatorname*{argmin}_{\lambda_l, \lambda_u, d} \mathbb{M}(\lambda_l, \lambda_u, d),$$

where the criterion function is

$$\mathbb{M}(\lambda_l, \lambda_u, d) \equiv (\lambda_l^2 - \lambda_u^2)\, d + \lambda_u^2\, \tau + 2\,(\lambda_u - \lambda_l)\, \Lambda(d) - 2\lambda_u\, \Lambda(\tau).$$

Natural estimates of $(\lambda_l^0, \lambda_u^0, d^0)$ are obtained by replacing the unknown $\Lambda(t)$ in the above expression by the Breslow estimator and then minimizing the resulting quantity with respect to $(\lambda_l, \lambda_u, d)$:

$$(\hat{\lambda}_l, \hat{\lambda}_u, \hat{d}_n) = \operatorname*{argmin}_{\lambda_l, \lambda_u, d} \mathbb{M}_n(\lambda_l, \lambda_u, d),$$

where

$$\mathbb{M}_n(\lambda_l, \lambda_u, d) = (\lambda_l^2 - \lambda_u^2) d + \lambda_u^2 \tau + 2(\lambda_u - \lambda_l)\hat{\Lambda}_n(d) - 2\lambda_u \hat{\Lambda}_n(\tau).$$

Here and in the following, whenever we refer to a minimizer, we mean some choice of minimizer rather than the set of all minimizers (similarly for maximizers), and we include the possibility of replacing $\mathbb{M}_n$ by its left-continuous version in d, in order to guarantee the existence of a minimizer. Note that

$$\hat{\lambda}_l = \frac{\hat{\Lambda}_n(\hat{d}_n)}{\hat{d}_n}, \qquad \hat{\lambda}_u = \frac{\hat{\Lambda}_n(\tau) - \hat{\Lambda}(\hat{d}_n)}{\tau - \hat{d}_n}, \tag{2.2}$$

corresponding to the first two normal equations.

Our first result gives the joint asymptotic distribution of these estimators.

Theorem 2.1. *If* (A1) *and* (A2) *hold, then*

$$n^{1/3}\left(\hat{\lambda}_l - \lambda_l^0, \hat{\lambda}_u - \lambda_u^0, \hat{d}_n - d^0\right) \to_d (c_1, c_2, 1) \operatorname*{argmax}_t Q(t),$$

where

$$Q(t) = aW(t) - bt^2,$$

W is a standard two-sided Brownian motion process on the real line, $a^2 = \lambda(d^0)/s^{(0)}(\beta_0, d^0)$,

$$b = b_0 - \frac{1}{8}|\lambda_l^0 - \lambda_u^0|\left(\frac{1}{d^0} + \frac{1}{\tau - d^0}\right) > 0,$$

with $b_0 = |\lambda'(d^0)|/2$, *and*

$$c_1 = \frac{\lambda_u^0 - \lambda_l^0}{2d^0}, \qquad c_2 = \frac{\lambda_u^0 - \lambda_l^0}{2(\tau - d^0)}.$$

2.1. Wald-type confidence intervals

It can be shown using Brownian scaling (see Banerjee and Wellner 2001) that

$$Q(t) =_d a\,(a/b)^{1/3}\, Q_1((b/a)^{2/3}\, t), \tag{2.3}$$

where $Q_1(t) = W(t) - t^2$, so the limit in the above theorem can be expressed more simply as

$$(c_1,\, c_2,\, 1)\,(a/b)^{2/3} \operatorname*{argmax}_t Q_1(t).$$

Let $p_{\alpha/2}$ denote the upper $\alpha/2$-quantile of the distribution of $\operatorname{argmax}_t Q_1(t)$ (this is symmetric about 0), known as Chernoff's distribution. Accurate values of $p_{\alpha/2}$, for selected values of α, are available in Groeneboom and Wellner (2001), where numerical aspects of Chernoff's distribution are studied. Utilizing the above theorem, this allows us to construct approximate $100(1-\alpha)\%$ confidence limits simultaneously for all the parameters $(\lambda_l^0, \lambda_u^0, d^0)$ in the working model:

$$\hat\lambda_l \pm \hat c_1 \hat\delta_n, \quad \hat\lambda_u \pm \hat c_2 \hat\delta_n, \quad \hat d_n \pm \hat\delta_n, \qquad \text{where } \hat\delta_n = n^{-1/3}(\hat a/\hat b)^{2/3} p_{\alpha/2}, \tag{2.4}$$

given consistent estimators $\hat c_1$, $\hat c_2$, $\hat a$, $\hat b$ of the nuisance parameters. Estimates of c_1 and c_2 are obtained by standard plug-in. The derivative of the baseline hazard function at d^0, appearing in b, can be estimated without difficulty using kernel smoothing (see Ramlau-Hansen 1983). In view of the third normal equation, the numerator of a^2 is estimated by the average of $\hat\lambda_l$ and $\hat\lambda_u$, and the denominator by $S^{(0)}(\hat\beta, \hat d_n)$.

2.2. Confidence sets based on deviance

Another strategy is to use a deviance function as an asymptotic pivot, which can be inverted to provide a confidence set for d^0; cf. the use of a residual sum of squares statistic in Banerjee and McKeague (2006). Define the deviance as

$$\mathbb{D}_n(d) = \mathbb{M}_n(\hat\lambda_l^d, \hat\lambda_u^d, d) - \mathbb{M}_n(\hat\lambda_l, \hat\lambda_u, \hat d_n),$$

where $\hat\lambda_l^d$ and $\hat\lambda_u^d$ are defined as in (2.2) but with $\hat d_n$ replaced by d. Our next result provides the asymptotic distribution of this statistic at $d = d^0$.

Theroem 2.2. *If* (A1) *and* (A2) *hold, then*

$$n^{2/3}\mathbb{D}_n(d^0) \to_d 2|\lambda_l^0 - \lambda_u^0| \max_t Q(t),$$

where Q is given in Theorem 2.1.

Using the Brownian scaling (2.3), the above limiting distribution can be expressed more simply as

$$2|\lambda_l^0 - \lambda_u^0| a(a/b)^{1/3} \max_t Q_1(t).$$

This leads to the following approximate $100(1-\alpha)\%$ confidence set for the threshold:

$$\{d : \mathbb{D}_n(d) \leqslant 2n^{-2/3}|\hat{\lambda}_l - \hat{\lambda}_u|\hat{a}(\hat{a}/\hat{b})^{1/3} q_\alpha\}, \tag{2.5}$$

where q_α is the upper α-quantile of $\max_t Q_1(t)$. This set is a finite union of intervals, so in practice (as in the next section) we only report the end-points of the component containing $\hat{d}_n$, at the expense of a slight under-coverage.

3. Numerical examples

In this section we study examples with simulated data and an application to a large cohort study concerning the onset age for schizophrenia.

For our simulation example, we consider a p-dimensional covariate $Z \sim$ Unif $[0, 1]^p$, for $p = 1$ and $p = 5$, and specify the baseline hazard function as $\lambda(t) = t$, the regression parameter $\beta_0 = (1, 1, \ldots, 1)^{\mathrm{T}}/p$, the censoring time C as exponential with mean 3, and the terminal time as $\tau = 1.5$. For both $p = 1$ and $p = 5$, about 65% of the failure times are uncensored and occur before τ, and roughly 10% exceed τ. The threshold is unique, $d^0 = 0.75$, and the conditions of Theorem 2.1 are satisfied. Although the best-fitting binary decision tree here provides a relatively crude approximation to $\lambda(t)$, the aim is to condense the information in $\lambda(t)$ rather than provide good local predictive power. Indeed, we chose this example to show that our approach works well even when there is no abrupt change in $\lambda(t)$. To provide a fair comparison between the Wald and deviance type confidence intervals, we used the true values of the nuisance parameters a and b to calibrate the intervals. Tables 1 and 2 report the coverage and average lengths of nominal 95% confidence intervals, and show that the deviance type confidence interval performs somewhat better than the Wald type one, with average length about 15% less while maintaining close to 95% coverage, except possibly at low sample sizes.

Next we apply our approach to estimate a threshold for the onset age of schizophrenia in the Jerusalem perinatal cohort comprising 92 000 individuals born during 1964–1976 to Israeli women living in Jerusalem and the adjoining rural areas. We restrict attention to 87 642 of these individuals for whom complete covariate information is available, and treat X as the (possibly right-censored) age in years at onset of schizophrenia. Malaspina *et al.* (2001) found a steady increase in schizophrenia risk with advancing paternal age, so we include paternal age at the time of the individual's birth as a covariate, along with two other covariates: indicator of male, and indicator of low socio-economic status. The Breslow estimator of the cumulative baseline hazard function is plotted in Figure 1. The Ramlau-Hansen kernel

Table 1. Coverage and average confidence interval length, $p = 1$

	Wald		Deviance type	
n	Coverage	Length	Coverage	Length
50	98.4	1.77	93.6	1.10
100	99.0	1.42	93.1	1.02
150	97.6	1.16	93.9	0.90
200	98.0	1.05	95.1	0.85
250	98.0	0.95	94.6	0.79
300	97.5	0.89	94.5	0.74
350	95.6	0.81	94.5	0.69
400	94.8	0.76	96.4	0.66
450	94.0	0.73	94.4	0.62
500	93.7	0.70	93.8	0.59

Table 2. Coverage and average confidence interval length, $p = 5$

	Wald		Deviance type	
n	Coverage	Length	Coverage	Length
50	92.4	2.44	83.4	0.98
100	95.4	1.47	89.8	0.96
150	96.0	1.23	92.3	0.89
200	96.8	1.09	92.0	0.82
250	94.5	0.99	93.5	0.78
300	94.6	0.89	94.9	0.73
350	95.3	0.82	93.5	0.68
400	93.9	0.79	92.4	0.65
450	91.9	0.73	94.4	0.61
500	92.4	0.70	92.0	0.59

estimate of $\lambda'(d^0)$, which is a part of $\hat{b}$, uses a bandwidth of one year and the Epanechnikov kernel.

The threshold estimate for the onset age is $\hat{d}_n = 16.69$ years, with 95% confidence intervals 15.79–17.59 and 16.29–17.06, for the Wald and deviance type methods, respectively. Note that the confidence interval based on deviance is considerably tighter. The point estimates of the baseline hazard levels are $\hat{\beta}_l = 2.39 \times 10^{-5}$ and $\hat{\beta}_u = 20.16 \times 10^{-5}$, showing about a tenfold increase in risk across the threshold. We set the terminal time at $\tau = 30$ years, but the results are virtually the same for any τ greater than 25 years.

4. Discussion

In this paper we have studied the effect of misspecification of a binary decision tree for the baseline hazard function in the Cox model. The convergence rate of the threshold estimator is found to be $n^{1/3}$, which is much slower than the rate of n that is obtained under correctly specified change-point models. Furthermore, the limit distribution of the threshold estimator is expressed in terms of the maximizer of a Gaussian process (a scaled Chernoff distribution) with continuous sample paths, as opposed to a limiting jump process of the type that arises in change-point estimation problems. The estimators of the hazard levels on either side of the threshold are also $n^{1/3}$-consistent (with scaled Chernoff distributions appearing once again in the limit), in contrast to the (correctly specified) change-point scenario where the hazard level estimators are $\sqrt{n}$-consistent with normal limits.

While, in this paper, we have considered a threshold in time, it is also of considerable interest to study covariate thresholds, which would be relevant, for example, to the study of paternal age-related effects in schizophrenia risk (Malaspina *et al.* 2001). Pons (2003) studies a covariate change-point Cox model, but the behaviour of the change-point estimator she proposes is not known in the misspecified setting. For this problem, it is not unreasonable to expect similar results to what we have obtained here, but its treatment is well beyond the scope of the present paper. Going beyond the Cox model, one can, for instance, consider threshold estimation in covariate or time for right-censored transformation models; Kosorok and Song (2006) study estimation of a covariate threshold in a correctly specified change-point model of this type. However, inference under misspecification or for time thresholding in this setting remains to be developed.

The extension of our working model to allow general parametric models before and after the threshold (rather than constants) is straightforward. We refer to Banerjee and McKeague (2006) for the way this can be done in the nonparametric regression framework; a similar extension goes through in the present setting.

Inspection of the proofs in the next section shows that they utilize an intricate combination of empirical process and martingale/counting process theory. Initially we attempted to use empirical process techniques exclusively, but this was not possible. In addition, we found that it was not feasible to establish a crucial weak convergence part of the proof (Lemma 5.1) using the martingale central limit theorem; even though $f_{n,t}$ is a martingale in $t > 0$ and has zero mean by the martingale property of M, the martingale property of $f_{n,t}$ fails when $t < 0$.

The determination of a threshold level for the hazard function and mean hazard levels on either side of the threshold are potentially important in a variety of biomedical contexts beyond the examples mentioned earlier. For example, patients suffering from terminal illnesses are often monitored over time in order to determine when a drastic – but potentially life-saving – medical intervention should be used. A similar issue arises on a population level when deciding the time (or age) at which vaccination is advisable from a public health point of view. In such scenarios, it is important to estimate a threshold in time for the hazard rate, after which the intervention may be justified. The methods of this paper can be used to determine a confidence interval not only for the threshold, but also for the relative risk λ_u^0/λ_l^0 across the threshold. Large values of this relative risk would indicate a greater necessity for medical intervention in the time zone given by the confidence interval for the threshold. Procedures of this type could also have a beneficial impact on health care policies pertaining to diagnostic testing.

We have restricted attention to the classical case of right censoring, but there is another form of censoring under which inference becomes considerably more difficult, namely interval censoring; here, the exact time to failure is observed for none of the individuals being studied. Rather, the failure time is only known to lie in a random interval. Interval censoring arises extensively in HIV/AIDS studies. In contrast to right-censored data, the cumulative baseline hazard Λ in the Cox model with interval-censored data can only be estimated at rate $n^{1/3}$ using the maximum likelihood estimator $\hat{\Lambda}_n$ (Huang 1996). This strongly suggests that if we base our criterion function $\mathbb{M}_n$ on $\hat{\Lambda}_n$, the estimate of d^0 obtained by minimizing $\mathbb{M}_n$ will converge at a rate slower than $n^{1/3}$. Preliminary simulations indicate that this is indeed the case. However, we believe that the asymptotics in this situation will be considerably harder than in the right-censored setting; little is currently known about the global behaviour of $\hat{\Lambda}_n$ to derive a minimax rate of convergence for $\hat{d}_n$. One way to bypass this problem might be to use a different estimate of Λ (possibly a smoothed estimate with a faster rate of convergence, under appropriate regularity conditions) in $\mathbb{M}_n$, leading to a possibly faster rate of convergence for (the corresponding) $\hat{d}_n$. In any case, this problem needs detailed investigation and is left as a topic for future research.

5. Proofs

The notation $\lesssim$ means that the left-hand side is bounded by a generic constant times the right-hand side. For a vector v, let $|v|$ be its Euclidean norm.

Proof of Theorem 2.1. The proof uses a standard strategy applicable to M-estimators in which we establish the rate of convergence (this step is delayed until the end of the proof) and the weak convergence of a suitably localized version of the criterion function, and then apply the argmax continuous mapping theorem (cf. van der Vaart and Wellner 1996: 288).

First note that the normalized estimators can be expressed as

$$n^{1/3}\left(\hat{\lambda}_l - \lambda_l^0,\, \hat{\lambda}_u - \lambda_u^0,\, \hat{d}_n - d^0\right) = \operatorname*{argmin}_h \mathbb{Q}_n(h), \tag{5.6}$$

where the localized version of the criterion function is given by

$$\mathbb{Q}_n(h) = n^{2/3}\left\{\mathbb{M}_n(\lambda_l^0 + h_1\, n^{-1/3},\, \lambda_u^0 + h_2\, n^{-1/3},\, d^0 + h_3\, n^{-1/3}) - \mathbb{M}_n(\lambda_l^0,\, \lambda_u^0,\, d^0)\right\},$$

$h = (h_1,\, h_2,\, h_3) \in \mathbb{R}^3$. To apply the argmax continuous mapping theorem we need to show that $\mathbb{Q}_n$ converges in distribution (and that its minimizer is tight; see the last step of the proof). The weak convergence is established in the space $B_{\text{loc}}(\mathbb{R}^3)$ of locally bounded functions on $\mathbb{R}^3$ equipped with the topology of uniform convergence on compacta. This is done by decomposing the process into parts with random and non-random limits, $\mathbb{Q}_n = \mathbb{Q}_{n,1} + \mathbb{Q}_{n,2}$, where

$$\mathbb{Q}_{n,1}(h) = n^{2/3}\Big\{(\mathbb{M}_n - \mathbb{M})(\lambda_l^0 + h_1\,n^{-1/3}, \lambda_u^0 + h_2\,n^{-1/3}, d^0 + h_3\,n^{-1/3}) - (\mathbb{M}_n - \mathbb{M})(\lambda_l^0, \lambda_u^0, d^0)\Big\}$$

and

$$\mathbb{Q}_{n,2}(h) = n^{2/3}\Big\{\mathbb{M}(\lambda_l^0 + h_1\,n^{-1/3}, \lambda_u^0 + h_2\,n^{-1/3}, d^0 + h_3\,n^{-1/3}) - \mathbb{M}(\lambda_l^0, \lambda_u^0, d^0)\Big\}.$$

A second-order Taylor expansion of $\mathbb{Q}_{n,2}(h)$, the gradient of which vanishes at $h = 0$, shows that it converges to $h^{\mathrm{T}}\,V\,h/2$ uniformly on every compact rectangle in $\mathbb{R}^3$, where V is the (positive-definite) Hessian matrix of the function $\mathbb{M}$ at $(\lambda_l^0, \lambda_u^0, d^0)$. Explicitly,

$$V = \begin{pmatrix} 2\,d^0 & 0 & \lambda_l^0 - \lambda_u^0 \\ 0 & 2\,(\tau - d^0) & \lambda_l^0 - \lambda_u^0 \\ \lambda_l^0 - \lambda_u^0 & \lambda_l^0 - \lambda_u^0 & 2\lambda_u^0 - \lambda_l^0\lambda'(d^0) \end{pmatrix}.$$

Now consider the first term in $\mathbb{Q}_n$. Using the fact that

$$(\mathbb{M}_n - \mathbb{M})\,(\lambda_l, \lambda_u, d) = 2\,(\lambda_u - \lambda_l)\,(\hat{\Lambda}_n - \Lambda)(d) - 2\lambda_u\,(\hat{\Lambda}_n - \Lambda)(\tau),$$

after some algebra we can write

$$\begin{aligned} \mathbb{Q}_{n,1}(h) = {} & 2\,(\lambda_u^0 - \lambda_l^0)\,n^{2/3}\,[(\hat{\Lambda}_n(d^0 + h_3\,n^{-1/3}) - \hat{\Lambda}_n(d^0)) - (\Lambda(d^0 + h_3\,n^{-1/3}) - \Lambda(d^0))] \\ & + 2\,(h_2 - h_1)\,n^{1/3}\,(\hat{\Lambda}_n - \Lambda)(d^0 + h_3\,n^{-1/3}) - 2\,h_2\,n^{1/3}\,(\hat{\Lambda}_n - \Lambda)(\tau). \end{aligned}$$

From Theorem 3.4 of Andersen and Gill (1982), $\sup_{0\leqslant d\leqslant\tau}|\hat{\Lambda}_n(d) - \Lambda(d)| = O_p(n^{-1/2})$, and it follows that the last two terms in the above display are $o_p(1)$ uniformly over every compact rectangle in $\mathbb{R}^3$. The first term can be written as $2(\lambda_u^0 - \lambda_l^0)\mathbb{Q}_{n,1}(h_3)$, where

$$\begin{aligned} \mathbb{Q}_{n,1}(t) = {} & n^{2/3}\int_{d^0}^{d^0+tn^{-1/3}} \left\{\frac{1}{S^{(0)}(\hat{\beta}, u)} - \frac{1}{S^{(0)}(\beta_0, u)}\right\}\mathbb{P}_n\mathrm{d}N(u) \\ & + \sqrt{n}\mathbb{P}_n\left[n^{1/6}\int_{d^0}^{d^0+tn^{-1/3}} \frac{\mathrm{d}M(u)}{s^{(0)}(\beta_0, u)}\right] \\ & + n^{2/3}\int_{d^0}^{d^0+tn^{-1/3}} \left\{\frac{1}{S^{(0)}(\beta_0, u)} - \frac{1}{s^{(0)}(\beta_0, u)}\right\}\mathbb{P}_n\mathrm{d}M(u) \\ & - n^{2/3}\int_{d^0}^{d^0+tn^{-1/3}} 1\{\mathbb{P}_n Y(u) = 0\}\lambda(u)\,\mathrm{d}u \\ = {} & A_n(t) + B_n(t) + C_n(t) + D_n(t). \end{aligned} \tag{5.7}$$

Here $N(t) = \delta 1(X \leqslant t)$ is the basic counting process, and $M(t) = N(t) - \int_0^t Y(u)\mathrm{e}^{\beta_0^{\mathrm{T}} Z}\lambda(u)\,\mathrm{d}u$ is the martingale part of $N(t)$, and $1/0 = 0$. In Lemma 5.1 we show that $B_n(t)$ converges to

$aW(t)$ in distribution in the space $B_{\text{loc}}(\mathbb{R})$. Moreover, as we show later in the proof, all other terms in the above display are asymptotically negligible. This allows us to conclude that $\mathbb{Q}_n(h)$ converges in distribution to $L(h) = \tilde{a}W(h_3) + h^{\mathrm{T}} V h/2$, where $\tilde{a} = 2|\lambda_l^0 - \lambda_u^0|a$. The argmax continuous mapping theorem then implies that $\operatorname{argmin}_h \mathbb{Q}_n(h)$ converges in distribution to $\operatorname{argmin}_h L(h)$; cf. the proof of Theorem 2.1 in Banerjee and McKeague (2006). Note that

$$\min_h L(h) = \min_{h_3}\left\{\tilde{a}W(h_3) + \min_{h_1,h_2} h^{\mathrm{T}} V h/2\right\}$$

and that we can find $\operatorname{argmin}_{h_1,h_2} h^{\mathrm{T}} V h/2$ explicitly. After some routine calculus, we conclude that the limiting distribution of (5.6) can be expressed as

$$(c_1,\, c_2,\, 1)\operatorname*{argmin}_t \{aW(t) + bt^2\} =_d (c_1,\, c_2,\, 1)\operatorname*{argmax}_t Q(t),$$

the limit stated in the theorem.

We now examine the three remainder terms in (5.7) and show that they are asymptotically negligible uniformly over t in any compact interval. For the first term, a Taylor expansion gives

$$A_n(t) = n^{-1/6} H(\beta^*,\, t)^{\mathrm{T}}\{\sqrt{n}(\hat{\beta} - \beta_0)\}, \tag{5.8}$$

where β^* is on the line segment between $\hat{\beta}$ and β_0,

$$H(\beta,\, t) = -n^{1/3}\int_{d^0}^{d^0+tn^{-1/3}} \frac{S^{(1)}(\beta,\, u)}{S^{(0)}(\beta,\, u)^2}\, \mathbb{P}_n \mathrm{d}N(u),$$

and $S^{(1)}(\beta,\, u) = \mathbb{P}_n[ZY(u)\mathrm{e}^{\beta^{\mathrm{T}} Z}]$. For any $K > 0$,

$$\sup_{\beta\in\mathcal{B},|t|<K} |H(\beta,\, t)| \leqslant \left[\sup_{\beta\in\mathcal{B},u\in[0,\tau]} \frac{|S^{(1)}(\beta,\, u)|}{S^{(0)}(\beta,\, u)^2}\right]\left[n^{1/3}\mathbb{P}_n 1\{|X - d^0| < Kn^{-1/3}\}\right].$$

The second term on the right of the above display converges in probability to $2Kp_X(d^0)$, where p_X is the pdf of X, and the first term is bounded in probability by elementary Glivenko–Cantelli type arguments. It follows that $|A_n(t)| = n^{-1/6}O_p(1)|\sqrt{n}(\hat{\beta} - \beta_0)| \to_p 0$ uniformly over t belonging to a compact interval. For the third term in (5.7), given $K > 0$,

$$\begin{aligned}\sup_{|t|<K} |C_n(t)| &\leqslant n^{2/3}\int_{d^0-Kn^{-1/3}}^{d^0+Kn^{-1/3}} \left|\frac{1}{S^{(0)}(\beta_0,\, u)} - \frac{1}{s^{(0)}(\beta_0,\, u)}\right| (\mathbb{P}_n\, \mathrm{d}N(u) + \lambda(u) S^{(0)}(\beta_0,\, u)\, \mathrm{d}u)\\ &= C_{n,1} + C_{n,2}\end{aligned}$$

where

$$\begin{aligned}C_{n,1} &\leqslant n^{1/3}\sup_{u\in[0,\tau]}\left|\frac{1}{S^{(0)}(\beta_0,\, u)} - \frac{1}{s^{(0)}(\beta_0,\, u)}\right| n^{1/3}\mathbb{P}_n 1\{|X - \mathrm{d}^0| < Kn^{-1/3}\}\\ &= n^{1/3} O_p(n^{-1/2})[2Kp_X(d^0) + o_p(1)] \to_p 0\end{aligned}$$

and

$$C_{n,2} \leqslant n^{1/3} \sup_{u\in[0,\tau]} \left| \frac{1}{S^{(0)}(\beta_0,\, u)} - \frac{1}{s^{(0)}(\beta_0,\, u)} \right| n^{1/3} \int_{d^0-Kn^{-1/3}}^{d^0+Kn^{-1/3}} \lambda(u) S^{(0)}(\beta_0,\, u)\,\mathrm{d}u$$

$$= n^{1/3} O_p(n^{-1/2})[2K\lambda(d^0)s^{(0)}(\beta_0,\, d^0) + o_p(1)] \rightarrow_p 0,$$

as required. Here we have used the fact that $\{Y(u)\mathrm{e}^{\beta_0^{\mathrm{T}} Z},\, u \in [0,\, \tau]\}$ is Donsker, and the limit of $S^{(0)}(\beta_0,\, \cdot)$, namely $s^{(0)}(\beta_0,\, \cdot)$, is bounded away from zero on $[0,\, \tau]$, to handle the supremum term. The last term, $D_n(t)$, in (5.7) is asymptotically negligible uniformly over t in a compact interval, since $P(\mathbb{P}_n Y(u) = 0) \leqslant P(X \leqslant \tau)^n \rightarrow 0$ for any $u < \tau$.

It remains to derive the convergence rates of the estimates promised earlier. This will be done by applying a slight extension of Theorem 3.2.5 of van der Vaart and Wellner (1996) in which we only need to check the moment condition of that theorem on a set Ω_n with $P^\star(\Omega_n) \rightarrow 1$. Let θ generically denote the vector $(\lambda_l,\, \lambda_u,\, d)$, and $\theta_0 = (\lambda_l^0,\, \lambda_u^0,\, d^0)$. Since $\mathbb{M}$ is uniquely minimized at θ_0 and is twice continuously differentiable, the condition $\mathbb{M}(\theta) - \mathbb{M}(\theta_0) \geqslant d^2(\theta,\, \theta_0)$ holds with d being the l_∞ norm in $\mathbb{R}^3$ (this is equivalent to the usual Euclidean metric, but somewhat simpler to work with). The rate of convergence will be derived in terms of the expected continuity modulus of $\sqrt{n}(\mathbb{M}_n - \mathbb{M})$ at θ_0:

$$\sqrt{n}\,\mathrm{E}^\star \left[\sup_{d(\theta,\theta_0)<\epsilon} |(\mathbb{M}_n - \mathbb{M})(\theta) - (\mathbb{M}_n - \mathbb{M})(\theta_0)| 1_{\Omega_n} \right] \tag{5.9}$$

for $\epsilon > 0$. The set Ω_n will be defined below in such a way that it is tractable to compute a bound on the above expectation. In what follows, we simplify the notation by using E and P in place of $\mathrm{E}^\star$ and $P^\star$.

Straightforward algebra shows that

$$(\mathbb{M}_n - \mathbb{M})(\theta) - (\mathbb{M}_n - \mathbb{M})(\theta_0) = 2[(\lambda_u - \lambda_l) - (\lambda_u^0 - \lambda_l^0)](\hat{\Lambda}_n - \Lambda)(d) - 2(\lambda_u - \lambda_u^0)(\hat{\Lambda}_n - \Lambda)(\tau)$$
$$+ 2\,(\lambda_u^0 - \lambda_l^0)\,[(\hat{\Lambda}_n - \Lambda)(d) - (\hat{\Lambda}_n - \Lambda)(d^0))].$$

This implies that (5.9) is bounded above by

$$6\epsilon R_n + 2|\lambda_l^0 - \lambda_u^0| S_n(\epsilon), \tag{5.10}$$

where $R_n = \sqrt{n}\mathrm{E}[\sup d \in [0,\, \tau]|\hat{\Lambda}_n(d) - \Lambda(d)| 1_{\Omega_n}]$ and

$$S_n(\epsilon) = \sqrt{n}\mathrm{E}\left[\sup_{|d-d^0|\leqslant\epsilon} |(\hat{\Lambda}_n - \Lambda)(d) - (\hat{\Lambda}_n - \Lambda)(d^0)| 1_{\Omega_n} \right].$$

First consider $S_n(\epsilon)$, and a similar decomposition to (5.7):

$$
\begin{aligned}
(\hat{\Lambda}_n - \Lambda)(d) - (\hat{\Lambda}_n - \Lambda)(d^0) &= \int_{d^0}^{d} \left\{ \frac{1}{S^{(0)}(\hat{\beta}, u)} - \frac{1}{S^{(0)}(\beta_0, u)} \right\} \mathbb{P}_n \,\mathrm{d}N(u) \\
&\quad + \mathbb{P}_n \left[\int_{d^0}^{d} \frac{\mathrm{d}M(u)}{s^{(0)}(\beta_0, u)} \right] \\
&\quad + \int_{d^0}^{d} \left\{ \frac{1}{S^{(0)}(\beta_0, u)} - \frac{1}{s^{(0)}(\beta_0, u)} \right\} \mathbb{P}_n \,\mathrm{d}M(u) \\
&\quad - \int_{d^0}^{d} 1\{\mathbb{P}_n Y(u) = 0\} \lambda(u) \,\mathrm{d}u \\
&= \bar{A}_n(d) + \bar{B}_n(d) + \bar{C}_n(d) + \bar{D}_n(d). \qquad (5.11)
\end{aligned}
$$

For the first term in this decomposition,

$$
\sqrt{n}\mathrm{E}\left[\sup_{|d-d^0| \leqslant \epsilon} |\bar{A}_n(d)| 1_{\Omega_n} \right] \leqslant \left(\mathrm{E}|\sqrt{n}(\hat{\beta} - \beta_0)|^{3/2} 1_{\Omega_n} \right)^{2/3} \left(\mathrm{E} \sup_{|d-d^0| \leqslant \epsilon} |\bar{H}(\beta^*, d)|^3 1_{\Omega_n} \right)^{1/3},
$$

where we have used Hölder's inequality (with $p = 3/2$, $q = 3$), and a Taylor expansion (similar to (5.8)) with β^* on the line segment between $\hat{\beta}$ and β_0, and

$$
\bar{H}(\beta, d) = -\int_{d^0}^{d} \frac{S^{(1)}(\beta, u)}{S^{(0)}(\beta, u)^2} \mathbb{P}_n \,\mathrm{d}N(u).
$$

From (2.6) in Andersen and Gill (1982),

$$
\sqrt{n}(\hat{\beta} - \beta_0) 1_{\Omega_n} = \left(n^{-1} \mathcal{I}(\beta^*, \tau) \right)^{-1} \sqrt{n} U(\tau),
$$

where Ω_n is the event on which the matrix $\mathcal{I}(\beta^*, \tau)$ is non-singular, $\mathcal{I}$ is defined by Andersen and Gill (see the display after their (2.5)), the matrix inverse of a singular matrix is defined to be zero, and

$$
U(t) = \mathbb{P}_n \int_0^t \left[Z - \frac{S^{(1)}(\beta_0, u)}{S^{(0)}(\beta_0, u)} \right] \mathrm{d}M(u). \qquad (5.12)
$$

Here note that $P(\Omega_n) \to 1$ by the proof of Andersen and Gill's Theorem 3.2. From that proof, $\|n^{-1}\mathcal{I}(\beta^*, \tau) - \Sigma\| \to_p 0$, where $\|\cdot\|$ is the operator norm, and Σ is an invertible matrix, so $P(\Omega_n) \to 1$. Also, by a matrix inequality (cf. Lemma 7 of Gandy and Jensen 2005), it follows that for some constant C, $\|n\mathcal{I}(\beta^*, \tau)^{-1} - \Sigma^{-1}\| \leqslant C\|n^{-1}\mathcal{I}(\beta^*, \tau) - \Sigma\|$ on a (new) set Ω_n (which is a subset of the previous Ω_n) such that $P(\Omega_n) \to 1$, so on this set

$$
|\sqrt{n}(\hat{\beta} - \beta_0)| \leqslant \|\Sigma^{-1}\| |n^{1/2} U(\tau)| + C\|n^{-1}\mathcal{I}(\beta^*, \tau) - \Sigma\| |n^{-1/2} U(\tau)|,
$$

and applying Hölder's inequality (with $p = 4$, $q = 4/3$),

$$\mathrm{E}|\sqrt{n}(\hat{\beta}-\beta_0)1_{\Omega_n}|^{3/2} \leqslant \mathrm{E}|n^{1/2}U(\tau)|^{3/2} + \left(\mathrm{E}\|n^{-1}\mathcal{I}(\beta^*, \tau)-\Sigma\|^6 1_{\Omega_n}\right)^{1/4}\left(\mathrm{E}|n^{1/2}U(\tau)|^2\right)^{3/4}.$$

Since $U(t)$ is a stochastic integral with respect to a counting process martingale, we can compute its second moment explicitly in terms of the intensity of the counting process:

$$\mathrm{E}|n^{1/2}U(\tau)|^2 = \mathrm{E}\left\{\mathbb{P}_n\int_0^\tau\left[Z-\frac{S^{(1)}(\beta_0, u)}{S^{(0)}(\beta_0, u)}\right]^2 Y(u)\mathrm{e}^{\beta_0^{\mathrm{T}}Z}\lambda(u)\,\mathrm{d}u\right\}$$

$$\leqslant O(1)+\mathrm{E}\left[\frac{1\{\mathbb{P}_nY(\tau)>0\}}{\mathbb{P}_nY(\tau)}\right]^2,$$

which is uniformly bounded in n (cf. Lemma 1 of McKeague and Utikal 1990), where we have used the boundedness of the covariates and the baseline hazard function. Next, consider

$$\mathrm{E}\|n^{-1}\mathcal{I}(\beta^*, \tau)-\Sigma\|^6 1_{\Omega_n} \leqslant \mathrm{E}\|n^{-1}\mathcal{I}(\beta^*, \tau)\|^6 1_{\Omega_n} + \|\Sigma\|^6.$$

We show that the first term is uniformly bounded, using a new Ω_n by intersecting the previous one with $\{\inf_{u\in[0,\tau],\beta\in\mathcal{B}} S^{(0)}(\beta, u)>c\}$, where $c=\inf_{u\in[0,\tau],\beta\in\mathcal{B}} s^{(0)}(\beta, u)/2>0$. By the uniform convergence in probability of $S^{(0)}$ to $s^{(0)}$, we still have $P^\star(\Omega_n)\to 1$. Inspecting the expression for $\mathcal{I}$ in Andersen and Gill (1982), note that

$$\sup_{t\in[0,\tau],\beta\in\mathcal{B}}\left\|\frac{1}{n}\mathcal{I}(\beta, t)\right\|1_{\Omega_n}<C,$$

where C is a constant that does not depend on n. It follows that $\mathrm{E}|\sqrt{n}(\hat{\beta}-\beta_0)|^{3/2}1_{\Omega_n}$ is uniformly bounded. To complete our work with the first term in (5.11), we now turn to $\mathrm{E}\sup_{|d-d0|\leqslant\epsilon}|\overline{H}(\beta^*, d)|^3 1_{\Omega_n}$. Let Ω_n with $P(\Omega_n)\to 1$ now be chosen (as a subset of the earlier one) so that

$$\sup_{u\in[0,\tau],\beta\in\mathcal{B}}\frac{|S^{(1)}(\beta, u)|}{S^{(0)}(\beta, u)^2}1_{\Omega_n}$$

is bounded by a constant. Writing $A_i=1\{|X_i-d^0|\leqslant\epsilon\}$ and using the independence of the A_i, we have

$$\mathrm{E}\sup_{|d-d^0|\leqslant\epsilon}|\overline{H}(\beta^*, d)|^3 1_{\Omega_n} \leqslant \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n A_i\right]^3$$

$$=\frac{1}{n^3}\left[\sum_{i=1}^n \mathrm{E}(A_i^3)+\sum_{i\neq j\neq k}\mathrm{E}(A_iA_jA_k)+\sum_{i\neq j}\mathrm{E}(A_i^2A_j)\right]$$

$$\leqslant\frac{1}{n^3}\left[n\epsilon+n^3\epsilon^3+n^2\epsilon^2\right],$$

where in the last step we use the uniform boundedness of the density of X over $[0, \tau]$. Thus, the first term in the decomposition (5.11) satisfies

$$\sqrt{n}\mathrm{E}\left[\sup_{|d-d^0|\leqslant\epsilon} |\bar{A}_n(d)| 1_{\Omega_n}\right] \lesssim \frac{\epsilon^{1/3}}{n^{2/3}} + \epsilon + \frac{\epsilon^{2/3}}{n^{1/3}}. \tag{5.13}$$

Now consider the second term in (5.11). Noting that the martingale integral has zero mean, we may write $\bar{B}_n$ in terms of the empirical process $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$. Using an inequality of van der Vaart and Wellner (1996: 291), we then have

$$\sqrt{n}\mathrm{E}\left[\sup_{|d-d^0|\leqslant\epsilon} |\bar{B}_n(d)|\right] = \mathrm{E}\left[\sup_{|d-d^0|\leqslant\epsilon} |\mathbb{G}_n f_d|\right] \lesssim J(1, \mathcal{M}_\epsilon)(P M_\epsilon^2)^{1/2}, \tag{5.14}$$

where $J(1, \mathcal{M}_\epsilon)$ is an entropy integral and M_ϵ is an envelope (specified later) for the class of functions $\mathcal{M}_\epsilon = \{f_d : |d - d^0| \leqslant \epsilon\}$, with

$$\begin{aligned} f_d &= \frac{\delta[1(X \leqslant d) - 1(X \leqslant d^0)]}{s^{(0)}(\beta_0, X)} - \mathrm{e}^{\beta_0^{\mathrm{T}} Z}\left[\int_{d^0}^{d} \frac{1(X \geqslant u)\lambda(u)}{s^{(0)}(\beta_0, u)}\,\mathrm{d}u\right] \\ &= f_{1,d} - f_{2,d}. \end{aligned}$$

We now set about finding an upper bound on the entropy integral. Set $\mathcal{M}_{1,\epsilon} = \{f_{1,d} : |d - d^0| \leqslant \epsilon\}$. It can be seen that $\mathcal{M}_{1,\epsilon}$ is obtained by multiplying a fixed function by members of a Vapnik–Chervonenkis class of functions, and is therefore itself Vapnik–Chervonenkis, so

$$\sup_Q N(\eta \|M_{1,\epsilon}\|_{Q,2}, \mathcal{M}_{1,\epsilon}, L_2(Q)) \lesssim \eta^{-\mathrm{V}_1},$$

for all $\eta > 0$ and some constant $V_1 > 0$, where $M_{1,\epsilon} = K_1 1\{|X - d^0| \leqslant \epsilon\}$ is an envelope for $\mathcal{M}_{1,\epsilon}$. Set

$$\begin{aligned} \mathcal{M}_{2,\epsilon} &= \{f_{2,d} : |d - d^0| \leqslant \epsilon\} = \{f_{2,d} : d^0 \leqslant d \leqslant d^0 + \epsilon\} \cup \{f_{2,d} : d^0 - \epsilon \leqslant d \leqslant d^0\} \\ &= \mathcal{M}_{2,\epsilon}^+ \cup \mathcal{M}_{2,\epsilon}^-. \end{aligned}$$

Note that $\mathcal{M}_{2,\epsilon}^+$ is obtained by multiplying $\mathrm{e}^{\beta_0^{\mathrm{T}} Z}$ by members of a class of functions, say $\mathcal{G}_\epsilon$, that is contained in the class of monotone increasing functions taking values in $[-K\epsilon, K\epsilon]$, for some constant $K > 0$. Taking $G_\epsilon = K\epsilon$ as an envelope for this class, we have

$$\sup_Q \log N(\eta \|G_\epsilon\|_{Q,2}, \mathcal{G}_\epsilon, L_2(Q)) \lesssim \eta^{-1},$$

for all $\eta > 0$, where we have used a simple extension of Theorem 2.7.5 of van der Vaart and Wellner (1996). Now setting $M_{2,\epsilon} = BG_\epsilon$ as an envelope for $\mathcal{M}_{2,\epsilon}^+$, where B is a bound on $\mathrm{e}^{\beta_0^{\mathrm{T}} Z}$, we obtain

$$\sup_Q \log N(\eta \|M_{2,\epsilon}\|_{Q,2}, \mathcal{M}_{2,\epsilon}^+, \mathrm{L}_2(\mathrm{Q})) \lesssim \eta^{-1}.$$

By choosing K large enough, we can ensure that $M_{2,\epsilon}$ is an envelope for $\mathcal{M}_{2,\epsilon}^-$ as well, so the same argument as before shows that the above display also holds with $\mathcal{M}^{+}_{2,\epsilon}$ replaced by $\mathcal{M}^{-}_{2,\epsilon}$. Thus

$$\sup_{Q} \log N(\eta \| M_{2,\epsilon}\|_{Q,2}, \mathcal{M}_{2,\epsilon}, L_2(Q)) \lesssim \eta^{-1}.$$

Noting that $\mathcal{M}_\epsilon \subset \mathcal{M}_{1,\epsilon} - \mathcal{M}_{2,\epsilon}$, and that $M_\epsilon = M_{1,\epsilon} + M_{2,\epsilon}$ is an envelope for this class, it is easily deduced that

$$\sup_{Q} \log N(\eta \| M_{\epsilon}\|_{Q,2}, \mathcal{M}_{\epsilon}, L_2(Q)) \lesssim \eta^{-1} - \mathrm{V}_1 \log \eta.$$

Thus,

$$\begin{aligned} J(1, M_\epsilon) &\equiv \sup_{Q} \int_0^1 \sqrt{1 + \log N(\eta \, \| M_\epsilon\|_{Q,2}, \mathcal{M}_\epsilon, L_2(Q))} \, \mathrm{d}\eta \\ &\lesssim \int_0^1 \sqrt{\eta^{-1} - V_1 \log \eta} \, \mathrm{d}\eta < \infty. \end{aligned}$$

Finally, noting that $P M_\epsilon^2 \lesssim \epsilon$, we conclude from (5.14) that

$$\sqrt{n}\mathrm{E}\left[\sup_{|d-d^0| \leqslant \epsilon} |\bar{B}_n(d)| \right] \lesssim \sqrt{\epsilon}. \tag{5.15}$$

Now consider the third term in (5.11). Note that $S^{(0)}(\beta_0, u)$ is bounded away from zero on Ω_n, and the total variation of the (random) signed measure $\mathbb{P}_n \, \mathrm{d}M$ over the interval $[d^0 - \epsilon, d^0 + \epsilon]$ is bounded by $U_n = \mathbb{P}_n\{|X - d^0| \leqslant \epsilon\} + O(\epsilon)$, so we have

$$\begin{aligned} \sqrt{n}\mathrm{E}\left[\sup_{|d-d^0| \leqslant \epsilon} |\bar{C}_n(d)| 1_{\Omega_n} \right] &\lesssim \mathrm{E}\left[\sup_{|u-d^0| \leqslant \epsilon} \sqrt{n} |S^{(0)}(\beta_0, u) - s^{(0)}(\beta_0, u)| U_n \right] \\ &\lesssim \left(E \sup_{|u-d^0| \leqslant \epsilon} |\mathbb{G}_n 1\{X \geqslant u\} \mathrm{e}^{\beta_0^{\mathrm{T}} Z}|^2 \right)^{1/2} \left(E U_n^2 \right)^{1/2} \\ &\lesssim \left(\frac{\epsilon}{n} + \epsilon^2 \right)^{1/2} \end{aligned} \tag{5.16}$$

using the Cauchy–Schwarz inequality. The above second moment term involving $\mathbb{G}_n$ is bounded by Theorem 1.3 of Talagrand (1994). Here Talagrand's result is applied to the uniformly bounded Vapnik–Chervonenkis class of functions $\mathcal{F} = \{(x, z) \mapsto 1\{x \geqslant u\}\mathrm{e}^{\beta_0^{\mathrm{T}} z}, u \in [0, \tau]\}$, where z is confined to a bounded set in $\mathbb{R}^p$. It follows from Theorem 2.6.7 of van der Vaart and Wellner (1996) that the covering number of $\mathcal{F}$ satisfies the assumption of Talagrand's theorem. A similar argument shows that R_n in (5.10) is bounded.

The fourth term in (5.11) vanishes on Ω_n, so it makes no contribution to (5.9). Combining our results for the various other terms, (5.13), (5.15) and (5.16), along with (5.10), we find that the expected continuity modulus (5.9) is of order

$$\phi_n(\epsilon) = \epsilon + \frac{\epsilon^{1/3}}{n^{2/3}} + \epsilon + \frac{\epsilon^{2/3}}{n^{1/3}} + \sqrt{\epsilon} + \left(\frac{\epsilon}{n} + \epsilon^2\right)^{1/2}$$

for $\epsilon > 0$. The leading term in $\phi_n(\epsilon)$ is $\sqrt{\epsilon}$. Solving $r_n^2 \phi_n(1/r_n) \leqslant \sqrt{n}$ yields the rate of convergence $r_n \leqslant n^{1/3}$, and we conclude from Theorem 3.2.5 of van der Vaart and Wellner (1996) that

$$n^{1/3}\left(\hat{\lambda}_l - \lambda_l^0, \hat{\lambda}_u - \lambda_u^0, \hat{d}_n - d^0\right) = O_p(1).$$

This completes the proof. □

Lemma 5.1. *The process $B_n(t)$ in the proof of Theorem 2.1 converges in distribution in the space $B_{\text{loc}}(\mathbb{R})$ to the Gaussian process $Q(t) = aW(t)$ defined in the statement of the theorem.*

Proof. We appeal to Theorem 2.11.22 of van der Vaart and Wellner (1996), which gives a central limit theorem for processes indexed by classes of functions changing with n, under various assumptions including an entropy integral condition. In the notation of that theorem, we can write $B_n(t) = \mathbb{G}_n f_{n,t}$, where $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ is the empirical process and

$$f_{n,t} = n^{1/6} \int_{d^0}^{d^0 + tn^{-1/3}} \frac{\mathrm{d}M(u)}{s^{(0)}(\beta_0, u)},$$

with t restricted to $[-K, K]$ for $K > 0$. Note that the martingale integral above has zero mean. We now verify that the limiting covariance function

$$R(s, t) = \lim_{n\to\infty} \mathrm{E}(f_{n,s} f_{n,t})$$

of $B_n(t)$ coincides with that of $aW(t)$. By the martingale property, $\mathrm{E}(f_{n,s} f_{n,t}) = 0$ if s and t are of opposite sign, so $R(s, t) = 0$ whenever s and t have opposite sign. The predictable quadratic variation process of M is the integrated intensity of the counting process N, so for $s, t > 0$,

$$\begin{aligned}
\mathrm{E}(f_{n,s} f_{n,t}) &= n^{1/3}\mathrm{E}\int_{d^0}^{d^0 + (s\wedge t)n^{-1/3}} \frac{Y(u)\lambda(u)\mathrm{e}^{\beta_0^{\mathrm{T}} Z}}{s^{(0)}(\beta_0, u)^2}\,\mathrm{d}u \\
&= n^{1/3}\int_{d^0}^{d^0 + (s\wedge t)n^{-1/3}} \frac{\lambda(u)}{s^{(0)}(\beta_0, u)}\,\mathrm{d}u \\
&\to \frac{\lambda(d^0)}{s^{(0)}(\beta_0, d^0)}(s \wedge t).
\end{aligned}$$

Thus, $R(s, t) = a^2(s \wedge t)$ for $s, t > 0$. It can be checked similarly that for $s, t < 0$, $R(s, t) = a^2(-s \wedge -t)$. Hence the limiting covariance function of $B_n(t)$ is indeed that of $aW(t)$. The entropy integral condition can be checked in a very similar way to the steps involved with $\bar{B}_n(t)$ in the proof of Theorem 2.1, using the envelope function

$$F_n(X, \delta, Z) = n^{1/6} K_1 1\{|X - d^0| \leqslant Kn^{-1/3}\} + K_2 n^{-1/6}\tau,$$

where the first term bounds the counting process part of $f_{n,t}$ and the second term bounds the compensator part, for (sufficiently large) constants K_1 and K_2. We omit the details of the remainder of the proof as they are similar to arguments used in Banerjee and McKeague (2006). □

Proof of Theorem 2.2. Writing $\hat{\lambda}_l^{d^0}$ and $\hat{\lambda}_u^{d^0}$ as $\hat{\lambda}_l^0$ and $\hat{\lambda}_u^0$, respectively, we have

$$\begin{aligned} n^{2/3}\,\mathbb{D}_n(d^0) &= n^{2/3}\,[\mathbb{M}_n(\hat{\lambda}_l^0, \hat{\lambda}_u^0, d^0) - \mathbb{M}_n(\hat{\lambda}_l, \hat{\lambda}_u, \hat{d}_n)] \\ &= n^{2/3}\,[\mathbb{M}_n(\lambda_l^0, \lambda_u^0, d^0) - \mathbb{M}_n(\hat{\lambda}_l, \hat{\lambda}_u, \hat{d}_n)] \\ &\quad - n^{2/3}\,[\mathbb{M}_n(\lambda_l^0, \lambda_u^0, d^0) - \mathbb{M}_n(\hat{\lambda}_l^0, \hat{\lambda}_u^0, \hat{d}_n)] \equiv I_n + J_n. \end{aligned}$$

Now, $I_n = -\min_h \mathbb{Q}_n(h)$ by (5.6) and converges in distribution to $-\min_h L(h)$, using Theorem 5.1 of Banerjee and McKeague (2006), and this simplifies to the limit stated in the theorem. It only remains to show that $J_n = o_p(1)$. Straightforward algebra allows J_n to be written as

$$\begin{aligned} &n^{2/3}((\hat{\lambda}_l^0)^2 - (\lambda_l^0)^2)d^0 - 2\,n^{2/3}(\hat{\lambda}_l^0 - \lambda_l^0)\hat{\Lambda}_n(d^0) \\ &\quad + n^{2/3}(\tau - d^0)((\hat{\lambda}_u^0)^2 - (\lambda_u^0)^2) - 2\,n^{2/3}(\hat{\lambda}_u^0 - \lambda_u^0)(\hat{\Lambda}_n(\tau) - \hat{\Lambda}_n(d^0)), \end{aligned}$$

which is simply

$$n^{2/3} d^0 \left[\hat{\lambda}_l^0 - \lambda_l^0\right]\left[\hat{\lambda}_l^0 + \lambda_l^0 - 2\frac{\hat{\Lambda}_n(d^0)}{d^0}\right] + n^{2/3}\left[\tau - d^0\right]\left[\hat{\lambda}_u^0 - \lambda_u^0\right]\left[\hat{\lambda}_u^0 + \lambda_u^0 - 2\frac{\hat{\Lambda}_n(\tau) - \hat{\Lambda}_n(d^0)}{\tau - d^0}\right].$$

Using the fact that $\hat{\lambda}_l^0 = \hat{\Lambda}_n(d^0)/d^0$ and $\hat{\lambda}_u^0 = (\hat{\Lambda}_n(\tau) - \hat{\Lambda}_n(d^0))/(\tau - \mathrm{d}^0)$, this simplies to

$$-n^{2/3}(\hat{\lambda}_l^0 - \lambda_l^0)^2 - n^{2/3}\,(\tau - d^0)(\hat{\lambda}_u^0 - \lambda_u^0)^2$$

which is $O_p(n^{-1/6})$ (using the $\sqrt{n}$-consistency of $\hat{\lambda}_u^0$ and $\hat{\lambda}_l^0$ for λ_u^0 and λ_l^0 respectively) and hence $o_p(1)$. This completes the proof.

Acknowledgements

Moulinath Banerjee was partially supported by a University of Michigan Rackham grant, and Ian McKeague by National Science Foundation grant DMS-0505201. The authors are grateful to Dolores Malaspina and Susan Harlap for helpful discussions and permission to use data from the Jerusalem Perinatal Cohort Schizophrenia Study.

References

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

Antoniadis, A. and Gijbels, I. (2002) Detecting abrupt changes by wavelet methods. *J. Nonparametr. Statist.*, **14**, 7–29.

Antoniadis, A., Gijbels, I. and MacGibbon, B. (2000) Nonparametric estimation for the location of a change-point in an otherwise smooth hazard function under random censoring. *Scand. J. Statist.*, **27**, 501–519.

Banerjee, M. and McKeague, I. W. (2006) Confidence sets for split points in decision trees. *Ann. Statist.* To appear.

Banerjee, M. and Wellner, J.A. (2001) Likelihood ratio tests for monotone functions. *Ann. Statist.*, **29**, 1699–1731.

Brownson, R. and Remington, P.L. (2002) *Communicating Public Health Information Effectively: A Guide for Practitioners*. Washington DC: American Public Health Association.

Bühlmann, P. and Yu, B. (2002) Analyzing bagging. *Ann. Statist.*, **30**, 927–961.

Chang, I.S., Chen, C.H. and Hsiung, C.A. (1994) Estimation in change-point hazard rate models with random censorship. In E. Carlstein, H.G. Mueller and D. Siegmund (eds), *Change-Point Problems*, IMS Lecture Notes-Monogr. Ser. 23, pp. 78–92. Hayward, CA: Institute of Mathematical Statistics.

Dempfle, A. and Stute, W. (2002) Nonparametric estimation of a discontinuity in regression. *Statist. Neerlandica*, **56**, 233–242.

Dupuy J.-F. (2006) Estimation in a change-point hazard regression model. *Statist. Probab. Lett.*, **76**, 182–190.

Gandy, A. and Jensen, U. (2005) On goodness-of-fit tests for Aalen's additive risk model. *Scand. J. Statist.*, **32**, 425–445.

Gijbels, I., Hall, P. and Kneip, A. (1999) On the estimation of jump points in smooth curves. *Ann. Inst. Statist. Math.*, **51**, 231–251.

Groeneboom, P. and Wellner, J.A. (2001) Computing Chernoff's distribution. *J. Comput. Graph. Statist.*, **10**, 388–400.

Hastie, T., Tibshirani, R. and Friedman, J.H. (2001) *The Elements of Statistical Learning*. New York: Springer-Verlag.

Huang, J. (1996) Efficient estimation for the Cox model with interval censoring. *Ann. Statist.*, **24**, 540–568.

Kosorok, M.R. and Song, R. (2006) Further details on inference under right censoring for transformation models with a change-point based on a covariate threshold. BMI Technical Report 194, University of Wisconsin, Madison.

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

Luo, X., Turnbull, B.W. and Clark, L.C. (1997) Likelihood ratio tests for a changepoint with survival data. *Biometrika*, **84**, 555–565.

Malaspina, D., Harlap, S., Fennig, S., Heiman, D., Nahon, D., Feldman, D. and Susser, E.S. (2001) Advancing paternal age and the risk of schizophrenia. *Arch. Gen. Psychiatry*, **58**(4), 361–367.

McKeague, I.W. and Utikal, K. (1990) Inference for a nonlinear counting process regression model. *Ann. Statist.*, **18**, 1172–1187.

Müller, H.G. and Wang, J.L. (1994) Change-point models for hazard functions. In E. Carlstein, H.G. Mueller and D. Siegmund (eds), *Change-Point Problems*, IMS Lecture Notes Monogr. Ser, 23, pp. 224–241. Hayward, CA: Institute of Mathematical Statistics.

Pons, O. (2002) Estimation in a Cox regression model with a change-point at an unknown time. *Statistics*, **36**, 101–124.

Pons, O. (2003) Estimation in a Cox regression model with a change-point according to a threshold in a covariate. *Ann. Statist.*, **31**, 442–463.

Ramlau-Hansen, H. (1983) Smoothing counting process intensities by means of kernel functions. *Ann. Statist.*, **11**, 453–466.

Talagrand, M. (1994) Sharper bounds for Gaussian and empirical processes. *Ann. Probab.*, **22**, 28–76.

van der Vaart, A. and Wellner, J.A. (1996) *Weak Convergence and Empirical Processes.* New York: Springer-Verlag.

Wu, C.Q., Zhao, L.C. and Wu, Y.H. (2003) Estimation in change-point hazard function models. *Statist. Probab. Lett.*, **63**, 41–48.

Received April 2006 and revised September 2006