On inference in parametric survival data models

Nils Lid Hjort

University of Oslo and Norwegian Computing Centre

-- April 1990 --

ABSTRACT. The usual parametric models for survival data are of the following form. Some parametrically specified hazard rate $\alpha(s,\theta)$ is assumed for possibly censored random life times $X_1^0,\ldots,X_n^0$; one observes only $X_i = \min\{X_i^0, c_i\}$ and $\delta_i = I\{X_i^0 \leq c_i\}$ for certain censoring times c_i that either are given or come from some censoring distribution. We study the following problems: What do the maximum likelihood estimator and other estimators really estimate when the true hazard rate $\alpha(s)$ is different from the parametric hazard rates? What is the limit distribution of an estimator under such outside-the-model circumstances? How can traditional model-based analyses be made model-robust? Does the model-agnostic viewpoint invite alternative estimation approaches? What are the consequences of carrying out model-based and model-robust bootstrapping? How do theoretical and empirical influence functions generalise to situations with censored data? How do methods and results carry over to more complex models for life history data like regression models and Markov chains?

KEY WORDS: *agnostic parameter estimation; censored data; distance measures; hazard regression; incorrect model; influence function; maximum likelihood; parametric and nonparametric bootstrapping*

1. Introduction

This paper is about aspects of maximum likelihood and related estimation methods applied to parametric survival data models. The aspects we shall care about include large-sample behaviour when the parametric model is a nonperfect approximation to the true model; distance measures from true to parametric model; model-based and model-robust estimation of the approximate covariance matrix; measures of influence; natural alternative estimation procedures suggested by the agnostic point of view; model-based and model-robust ways of bootstrapping; and similar questions for hazard rate regression models. Indeed, Section 2 studies limit behaviour of the maximum likelihood estimator when the parametric model is incorrect, Section 3 finds influence functions under censoring, and in Section 4 the general methods are used to assess the behaviour of various bootstrapping schemes. The apparatus developed in Sections 3 and 4 can be used to prove some known results anew, and should be useful also in other survival data models and for other estimators than the maximum likelihood one. Some new estimation methods are discussed in Section 5, and Section 6 treats two regression models for hazard rates. Complementary remarks are offered in the final Section 7.

A recurrent theme underlying our article is the point of view that (i) parametric models are usually incorrect, (ii) that estimation and inference in parametric models nevertheless can be a useful enterprise, (iii) provided the statistician knows what she is doing.

Even statisticians admit (i). Traditional and valid arguments favouring (ii) include matters of sample size versus nonparametrics and the value of simplifying and synthesising to aid understanding of complex phenomena. The following reasoning also supports (ii) and pertains to the present paper. We view a parametric estimation procedure as an attempt to find the best fitting or most appropriate parametric approximant to the more elusive true model. An estimator for the parameter vector θ will typically be consistent for a certain θ_0 that is most appropriate, or least false, in the sense of minimising a suitable distance measure between true model and parametric model. Accordingly estimating the least false parameter is a meaningful statistical operation, even outside model conditions (i.e. even if the minimum distance is positive), provided only that the distance measure itself is reasonable. Regarding (iii) above, as far as the first order large sample consequences of an incorrect parametric model is concerned the single technical complication will be seen to be a different expression for the limiting covariance matrix of the estimators. A consistent estimator for this more general covariance matrix can be constructed explicitly, or approximated by appropriate resampling, or reached as a by-product of empirical influence functions.

Different estimation methods may correspond to different distance measures and thus different least false parameters. It often enhances one's understanding of an estimation procedure to view it in this light, i.e. by exhibiting the accompanying distance measure between truth and approximating model. Of course this agnostic point of view can be the explicit motivation for some estimators in the first place; an empirical counterpart can be constructed for a given distance measure and then be minimised for the given data.

The results of this paper give precise statistical substance to fitting and analysing data with a wrong model, and suggest that it even can be fruitful. This is not to say that one shouldn't assess the adequacy of one's model or compare different natural models; one should indeed, and general methods for doing this can be found in Hjort (1990a). But the agnostic point of view and results under such is meant to free statisticians from the irongrip of that part of traditional methodology which has 'the parametric model is assumed to be absolutely correct' as basic assumption. This should have some pragmatic value as well, since practitioners often try out a variety of models while knowing that neither of them is likely to be quite correct. The theory developed below gives a recipe for bettering this practice by using corrected approximate covariance matrices for the estimators.

One can also usefully define and study situations where the amount of misspecification is moderate. This is done on a general basis in Hjort (1990b). Included there is a result which says that it is actually advantageous, in terms of precision of estimators, to stick to a given model even when it is moderately incorrect, and the precise 'tolerance radius' around the model against various types of model departures is also found.

The points of view expressed above are not entirely new, but relatively few publications have discussed behaviour of model-derived estimates under fixed alternative conditions. The basic and not so difficult result (1.3) below has appeared a couple of times under various guises, and sometimes rather implicitly, see Cox (1962) and Reeds (1978) for early examples and Hjort (1986a, 1986b, 1988) and Linhart and Zucchini (1986) for recent ones in different settings. The remainder of this section is a concise treatment of the simpler non-censored i.i.d.-case. It is included here since the viewpoint and results do not appear to

be well known, and since our results perhaps will be easiest to understand and appreciate when compared to corresponding statements for this simpler classical framework.

Let $X_1, \ldots, X_n$ be independent from some unknown distribution F with density f, and suppose the data are to be fitted to some p-dimensional parametric family of densities $\{f_\theta : \theta \in \Theta\}$. Where notationally convenient we shall write $f(x,\theta)$ instead of $f_\theta(x)$ and so on. Note that we do not assume the true f to belong to the parametric class, unlike what is typically the case in textbook treatments of this problem. The maximum likelihood estimator $\widehat{\theta}$ maximises the observed likelihood $L_n(\theta)$ w.r.t. the parameter. Since the simple average $n^{-1}\log L_n(\theta)$ tends to $E_F \log f_\theta(X) = \int f \log f_\theta \, dx$ in probability $\widehat{\theta}$ intuitively aims at becoming close to the parameter value θ_0 that maximises this expression, or, equivalently, minimises the Kullback–Leibler distance

$$d[f, f_\theta] = \int f(x) \log\{f(x)/f_\theta(x)\} \, dx \tag{1.1}$$

from true model to parametric model. We think of $\theta_0 = \theta_0(F)$, which is indeed uniquely defined in most cases, as the *least false* or *most fitting* parameter value.

We summarise below the behaviour of $\widehat{\theta}$ for large n under the present outside-the-model circumstances. The arguments needed to prove the results can be seen as more careful versions of the 'traditional ones' that are used under model circumstances (see e.g. Lehmann, 1983, Ch. 6). Consider the p-vector U_n of first order derivatives and the $p \times p$-matrix I_n of second order derivatives of $n^{-1}\log L_n(\theta)$. $\widehat{\theta}$ is a solution to the maximum likelihood equations $U_n(\theta) = 0$, so by Taylor expansion $0 = U_n(\widehat{\theta}) = U_n(\theta_0) + I_n(\widetilde{\theta})(\widehat{\theta} - \theta_0)$, which leads to

$$\sqrt{n}(\widehat{\theta} - \theta_0) = \{-I_n(\widetilde{\theta})\}^{-1} \sqrt{n} U_n(\theta_0), \tag{1.2}$$

in which $\widetilde{\theta}$ lies somewhere between θ_0 and $\widehat{\theta}$. Two matrices therefore determine the limit distribution: the limit $J = J(F, \theta_0)$ of $-I_n(\theta_0)$, obtained by the law of large numbers, and the covariance matrix $K = K(F, \theta_0)$ of $\sqrt{n} U_n(\theta_0)$, obtained from the central limit theorem. More precisely,

$$J = -\int \frac{\partial^2 \log f(x,\theta_0)}{\partial\theta\partial\theta} \, dF(x) \quad \text{and} \quad K = \int \Big(\frac{\partial \log f(x,\theta_0)}{\partial\theta}\Big)\Big(\frac{\partial \log f(x,\theta_0)}{\partial\theta}\Big)' \, dF(x).$$

Natural estimators for these $p \times p$ matrices are $\widehat{J} = J(\widehat{F}, \widehat{\theta})$ and $\widehat{K} = K(\widehat{K}, \widehat{\theta})$, that is

$$\widehat{J} = -\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta\partial\theta} \log f(X_i, \widehat{\theta}), \; \widehat{K} = K(\widehat{F}, \widehat{\theta}) = \frac{1}{n}\sum_{i=1}^n \Big(\frac{\partial}{\partial\theta} \log f(X_i, \widehat{\theta})\Big)\Big(\frac{\partial}{\partial\theta} \log f(X_i, \widehat{\theta})\Big)'.$$

$\widehat{F}$ is the empirical distribution which places weight $1/n$ on each data point.

RESULT. *Under traditional regularity conditions $\widehat{\theta}$ is consistent for the least false parameter θ_0. Furthermore,*

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d J^{-1} N_p\{0, K\} = N_p\{0, J(F,\theta_0)^{-1} K(F,\theta_0) J(F,\theta_0)^{-1}\}, \tag{1.3}$$

and $\widehat{J}$ and $\widehat{K}$ are consistent estimators for J and K.

The result (1.3) is the appropriate generalisation of the classical textbook result, in which $f(x) = f(x, \theta_0)$ is assumed, and where it is easy to show that the two matrices are equal,

$$J(F_\theta, \theta) = K(F_\theta, \theta). \tag{1.4}$$

We can now distinguish between model-based and model-robust inference about θ_0. In the first case θ_0 is true, and one uses $\widetilde{J}^{-1}/n$ as the estimate of the covariance matrix for $\widehat{\theta}$, where $\widetilde{J}$ could be either $J(\widehat{F}, \widehat{\theta})$ or $J(F(.,\widehat{\theta}), \widehat{\theta})$. In the second case θ_0 has the wider interpretation of being merely most fitting, and one uses $\widehat{J}^{-1}\widehat{K}\widehat{J}^{-1}/n$ instead.

EXAMPLE 1.1. Suppose nonnegative data are fitted to the exponential distribution with density $f_\theta(x) = \theta \exp(-\theta x)$. Then $d[f, f_\theta] = \int_0^\infty f(x) \log f(x)\, dx - \int_0^\infty (\log \theta - \theta x) f(x)\, dx$ is minimised for the least false parameter $\theta_0 = 1/\mu(F)$, where $\mu(F) = E_F X$. One finds $J = 1/\theta_0^2$ and $K = \mathrm{Var}_F X = \sigma^2(F)$. The model-based asymptotic variance of $\widehat{\theta} = 1/\widehat{\mu}$ is $n^{-1}\theta_0(F)^2$, estimated by $n^{-1}\widehat{\theta}^2$, whereas the model-robust version is $n^{-1}\sigma^2(F)\theta_0(F)^4$, estimated by $n^{-1}\widehat{\sigma}^2\widehat{\theta}^4$. □

Next turn attention to bootstrapping. Model-based bootstrapping consists of drawing samples $X_1^*, \ldots, X_n^*$ from the parametrically estimated $F(.,\widehat{\theta})$, and computing bootstrap estimates $\widehat{\theta}^* = \widehat{\theta}(X_1^*, \ldots, X_n^*)$. Nonparametric or model-robust bootstrapping on the other hand samples X_i^*'s from $\widehat{F}$. The (first-order) large sample behaviour of $\widehat{\theta}^*$ can be analysed and characterised by the methods already used. Think of $\theta_0 = \mathrm{ml}(F)$, the maximiser of $\int \log f_\theta(x)\, dF(x)$, as a functional operating on the space of distributions. Observe that both $\mathrm{ml}(\widehat{F})$ and $\mathrm{ml}(F(.,\widehat{\theta}))$ are equal to $\widehat{\theta}$. By (1.2) and (1.3) we have

$$\sqrt{n}\{\mathrm{ml}(\widehat{F}) - \mathrm{ml}(F)\} \doteq_d J(F, \mathrm{ml}(F))^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \log f(X_i, \mathrm{ml}(F))}{\partial \theta}, \tag{1.5}$$

where $U_n \doteq_d V_n$ means that $U_n - V_n$ tends to zero in probability. More precise information can be gathered using methods presented in Section 4.

Consider first parametric bootstrapping, which uses $\widehat{\theta}^*$ computed from $F(.,\widehat{\theta})^*$, say, the empirical distribution of X_i^*'s from $F(.,\widehat{\theta})$. Then

$$\begin{aligned}\sqrt{n}(\widehat{\theta}^*_{\mathrm{pb}} - \widehat{\theta}) &= \sqrt{n}\{\mathrm{ml}(F(.,\widehat{\theta})^*) - \mathrm{ml}(F(.,\widehat{\theta}))\} \doteq_d J(F(.,\widehat{\theta}),\widehat{\theta})^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \log f(X_i^*, \widehat{\theta})}{\partial \theta} \\ &\doteq_d J(F(.,\widehat{\theta}),\widehat{\theta})^{-1} N_p\{0, K(F(.,\widehat{\theta}),\widehat{\theta})\} = N_p\{0, J(F(.,\widehat{\theta}),\widehat{\theta})^{-1}\}.\end{aligned} \tag{1.6}$$

Correspondingly, for nonparametric bootstrapping one has

$$\begin{aligned}\sqrt{n}(\widehat{\theta}^*_{\mathrm{nb}} - \widehat{\theta}) &= \sqrt{n}\{\mathrm{ml}(\widehat{F}^*) - \mathrm{ml}(\widehat{F})\} \doteq_d J(\widehat{F},\widehat{\theta})^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \log f(X_i^*, \widehat{\theta})}{\partial \theta} \\ &\doteq_d J(\widehat{F},\widehat{\theta})^{-1} N_p\{0, K(\widehat{F},\widehat{\theta})\} = N_p\{0, \widehat{J}^{-1}\widehat{K}\widehat{J}^{-1}\}.\end{aligned} \tag{1.7}$$

Several conclusions can be drawn from this. First, the nonparametric bootstrap always works, in the large sample first order sense, in that the bootstrap distribution always

mimics the true distribution, even when the parametric model is incorrect; the distribution of $\sqrt{n}(\widehat{\theta}^*_{\rm nb}-\widehat{\theta})$ tends with probability one to the same as does $\sqrt{n}(\widehat{\theta}-\theta_0)$, cf. (1.3). Secondly, the parametric bootstrap only works when the model is correct, otherwise it does not reflect the real sampling variability. Thirdly, we should note that the sampling variability of $\widehat{\theta}^*_{\rm nb}$ is typically much larger than that of $\widehat{\theta}^*_{\rm pb}$. This is related to the observation that *if* the model happens to be correct, then both $\widehat{J}^{-1}\widehat{K}\widehat{J}^{-1}$ and $\widehat{J}^{-1}$ estimate the same quantity, namely the asymptotic covariance matrix of $\sqrt{n}(\widehat{\theta}-\theta_0)$, but the first is less stable than the second.

In situations where interest centres on another parameter $\mu=\mu(\theta)$ the discussion here applies to $\widehat{\mu}=\mu(\widehat{\theta})$ and $\widehat{\mu}^*=\mu(\widehat{\theta}^*)$ instead.

EXAMPLE 1.2. Let $\widehat{V}_{\rm nb}$ and $\widehat{V}_{\rm pb}$ be the bootstrap estimates of the variance of $\widehat{\theta}$ in the exponential situation treated above. Then it can be shown that

$$\frac{\mathrm{Var}\{\widehat{V}_{\rm pb}\}}{\mathrm{Var}\{\widehat{V}_{\rm nb}\}} \doteq \frac{\mathrm{Var}\{\widehat{\theta}^2/n\}}{\mathrm{Var}\{\widehat{\theta}^4\widehat{\sigma}^2/n\}} \doteq \frac{4\theta_0^4/n}{8\theta_0^4/n}=\frac{1}{2}$$

if the exponential model prevails. See also further comments, examples, and amendments in Hjort (1988). □

2. Theory for incorrectly specified parametric survival data models

Suppose $X_1^0,\ldots,X_n^0$ are lifetimes for n individuals drawn from a homogeneous population with underlying hazard rate $\alpha(s)=f(s)/F[s,\infty)$ for $s\geq 0$. Suppose that one observes only $X_i=\min\{X_i^0,c_i\}$ and $\delta_i=I\{X_i^0\leq c_i\}$, where the censoring variables c_i are independent of the lifetimes and come from some censoring distribution G. A parametric model is proposed of the type $\alpha(s)\approx\alpha_\theta(s)=\alpha(s,\theta)$. In this section the large-sample properties of the maximum likelihood estimator outside model conditions are derived, parallelling the treatment of the traditional non-censored type problem in Section 1.

The treatment below extends that of Borgan (1984) and Hjort (1986a). The mathematical techniques needed to derive results involve central limit theorems and inequalities for martingales and integrals of previsible functions with respect to martingales. The necessary technicalities resemble those thoroughly presented in Andersen and Gill (1982), Borgan (1984), Andersen and Borgan (1985), and Hjort (1986a). This allows us to skip most of the formal details here. New proofs of some of the older results can also be constructed as a by-product of the general machinery of influence functions and differentiable functionals developed in Sections 3 and 4 below.

We must start by defining the maximum likelihood estimator. Introduce the counting process N, the at-risk process Y, and the associated martingale M by

$$N(t)=\sum_{i=1}^n I\{X_i\leq t,\delta_i=1\},\quad Y(t)=\sum_{i=1}^n I\{X_i\geq t\},\quad M(t)=N(t)-\int_0^t Y(s)\alpha(s)\,ds. \tag{2.1}$$

Notice that M employs the true hazard rate $\alpha(s)$ rather than some $\alpha(s,\theta_0)$. With conditions about the censoring mechanism much weaker than the random censorship assumption

used here the likelihood can be written

$$L_n(\theta) = \exp\Big[\int_0^T \{\log \alpha(s,\theta) dN(s) - Y(s)\alpha(s,\theta)\, ds\}\Big],$$

where $[0,T]$ is the time interval over which the processes are observed. We will assume T finite to get certain martingale arguments below easily through, but extension to the full half-line is possible with appropriate extra conditions. Among the important properties of M is the fact that $W_n = \int_0^t H_n(s)\, dM(s)/\sqrt{n}$ converges in distribution to $W = \int_0^t h(s)\, dV(s)$, provided H_n is previsible (the value of $H_n(s)$ is known already at time $s-$) and converges uniformly, in probability, to a deterministic function h. Here V is a Gaussian zero-mean process with independent increments and $\mathrm{Var}\, dV(s) = y(s)\alpha(s)\, ds$, and $y(s)$ is the limit in probability of $Y(s)/n$, namely

$$y(s) = \Pr\{X_i \geq s\} = \Pr\{X_i^0 \geq s, c_i \geq s\} = F[s,\infty)G[s,\infty). \tag{2.2}$$

Note that W is normal with mean zero and variance $\int_0^t h^2 y\alpha\, ds$.

Consider first

$$\begin{aligned}\frac{1}{n}\log L_n(\theta) &= \frac{1}{n}\int_0^T \{\log \alpha(s,\theta) dN(s) - Y(s)\alpha(s,\theta) ds\} \\ &= \frac{1}{n}\int_0^T \{\log \alpha_\theta (dM + Y\alpha\, ds) - Y\alpha_\theta ds\} \to_p \int_0^T y(\alpha \log \alpha_\theta - \alpha_\theta)\, ds.\end{aligned}$$

Maximising $L_n(\theta)$ should therefore in the end amount to maximising the right hand expression here, which is the same as minimising the distance

$$d[\alpha,\alpha_\theta] = \int_0^T y\{\alpha(\log\alpha - \log\alpha_\theta) - (\alpha - \alpha_\theta)\}\, ds \tag{2.3}$$

from the true model to the approximating parametric model. $d[\alpha,\alpha_{\theta_0}]$ is always non-negative and is zero only if $\alpha(s) = \alpha(s,\theta_0)$ a.e. on $[0,T]$, in which case θ_0 indeed is the "true" parameter. In general we can only reckon with a *least false* parameter value θ_0 which minimises (2.3). Observe that the value of θ_0 may depend upon the censoring distribution through $y(s) = F[s,\infty)G[s,\infty)$. Note also that (2.3) properly generalises the Kullback–Leibler distance (1.1). See Remark 7A and Section 5B.

THEOREM 2.1. *Suppose that there is a unique minimiser θ_0 of (2.3); that $\alpha(s,\theta)$ is three times differentiable in a neighbourhood $N(\theta_0)$ of θ_0; that these functions are bounded over $[0,T]\times N(\theta_0)$; that $\alpha(s)$ and $\alpha(s,\theta_0)$ are bounded away from zero as s runs from 0 to T; and finally that the J matrix appearing below is positive definite. [Somewhat weaker sufficient conditions can be put up in the style of Borgan (1984, Section 4; note the corrigendum p. 275).] Then the maximum likelihood estimator $\widehat{\theta}$ is consistent for the least false parameter θ_0. Consider matrices $J = J(\alpha,y,\theta_0)$ and $K = K(\alpha,y,\theta_0)$ defined as follows:*

$$J = \int_0^T y(s)\Big[\psi(s,\theta_0)\psi(s,\theta_0)'\alpha(s,\theta_0) - D\psi(s,\theta_0)\{\alpha(s) - \alpha(s,\theta_0)\}\Big]\, ds,$$

$$K = \int_0^T y(s)\Big[\psi(s,\theta_0)\psi(s,\theta_0)'\alpha(s) + \{\psi(s,\theta_0)E(s)' + E(s)\psi(s,\theta_0)'\}\alpha(s,\theta_0)\Big]\, ds,$$

in which $\psi(s,\theta) = \partial \log \alpha(s,\theta)/\partial\theta$, $D\psi(s,\theta) = \partial^2 \log \alpha(s,\theta)/\partial\theta\partial\theta$, *and* $E(s) = \int_0^s y(t)\,\psi(t,\theta_0)\{\alpha(t) - \alpha(t,\theta_0)\}\,dt$ *(in particular,* $E(0) = E(T) = 0$*). Then*

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d J^{-1} N_p\{0, K\} = N_p\{0, J(\alpha, y, \theta_0)^{-1} K(\alpha, y, \theta_0) J(\alpha, y, \theta_0)^{-1}\}.$$

PROOF: It was indeed shown already in Hjort (1986a) that $\widehat{\theta}$ is consistent for this most fitting parameter θ_0. There is even almost sure convergence in the present random censorship situation, a fact used in Section 4.

Next turn to the limit distribution of $\widehat{\theta}$. The idea is to consider the vector U_n of first order partial derivatives and the matrix I_n of second order partial derivatives of $n^{-1} \log L_n$ and apply (1.2) again, in this more difficult situation. One finds

$$\begin{aligned} U_n(\theta) &= \frac{1}{n} \int_0^T \psi(s,\theta)\{dN(s) - Y(s)\alpha(s,\theta)\,ds\} \\ &= \frac{1}{n} \int_0^T \psi_\theta\{dM + Y(\alpha - \alpha_\theta)\,ds\} \to_p \int_0^T y\psi_\theta(\alpha - \alpha_\theta)\,ds \end{aligned}$$

and

$$\begin{aligned} I_n(\theta) &= \frac{1}{n} \int_0^T \big[D\psi(s,\theta)\{dN(s) - Y(s)\alpha(s,\theta)\,ds\} - \psi(s,\theta)Y(s)\alpha(s,\theta)\psi(s,\theta)'\,ds\big] \\ &\to_p \int_0^T y\{D\psi(.,\theta)(\alpha - \alpha_\theta)\,ds - \psi_\theta\psi_\theta'\alpha_\theta\,ds\}. \end{aligned}$$

In particular $-I_n(\theta_0)$ tends to the J matrix in probability. Note next that $U_n(\theta)$ tends to zero when $\theta = \theta_0$. Furthermore,

$$\sqrt{n}\,U_n(\theta_0) = \int_0^T \psi(s,\theta_0)\big[dM(s)/\sqrt{n} + \sqrt{n}\{Y(s)/n - y(s)\}\{\alpha(s) - \alpha(s,\theta_0)\}\,ds\big].$$

Here $V_n = M/\sqrt{n}$ has a limit process V described before (2.2) and $Z_n = \sqrt{n}(Y/n - y)$ converges in distribution to a Gaussian zero-mean process $Z(.)$, in the function space $D[0,T]$ of left-continuous functions with right hand limits, by the theory presented for example in Billingsley (1968, Section 13). One has

$$\operatorname{cov}\{Z_n(s), Z_n(t)\} = \operatorname{cov}\big[I\{X_i^0 \geq s, c_i \geq s\}, I\{X_i^0 \geq t, c_i \geq t\}\big] = y(s \vee t) - y(s)y(t),$$

writing $s \vee t = \max\{s,t\}$. Also, if N_i, Y_i, M_i are the counting process, at risk process, and martingale for individual no. i, then

$$\operatorname{cov}\{dV_n(s), Z_n(t)\} = \operatorname{cov}\{dM_i(s), Y_i(t)\} = E\{dN_i(s) - Y_i(s)\alpha(s)\,ds\}Y_i(t)$$

can be seen to equal $-\alpha(s)ds\,y(t)$ for $s < t$ and 0 for $s \geq t$. These are also expressions for $\operatorname{cov}\{Z(s), Z(t)\}$ and $\operatorname{cov}\{dV(s), Z(t)\}$. That indeed $(V_n, Z_n) \to_d (V, Z)$ in $D[0,T] \times D[0,T]$ and

$$\sqrt{n}\,U_n(\theta_0) \to_d \int_0^T \psi(s,\theta_0)\big[dV(s) + Z(s)\{\alpha(s) - \alpha(s,\theta_0)\}\,ds\big] = \text{one} + \text{two} \qquad (2.4)$$

hold, where an expression for $K = \text{VAR}\{\text{one}+\text{two}\}$ for this necessarily Gaussian limit vector is derived below, can be shown combining function space asymptotics from Billingsley (1968) and Andersen and Borgan (1985).

To find K, observe first that

$$\text{VAR}\{\text{one}\} = \int_0^T y(s)\psi(s,\theta_0)\psi(s,\theta_0)'\alpha(s)\,ds.$$

Write $\Delta(s) = \alpha(s) - \alpha(s,\theta_0)$ for the difference between true hazard and most fitting hazard. Then

$$\begin{aligned}\text{VAR}\{\text{two}\} &= \int_0^T \int_0^T \psi(s,\theta_0)\psi(t,\theta_0)'\Delta(s)\Delta(t)\{y(s\vee t) - y(s)y(t)\}\,dsdt \\ &= \int_0^T \int_0^t \{\psi(s,\theta_0)\psi(t,\theta_0)' + \psi(t,\theta_0)\psi(s,\theta_0)'\}\Delta(s)\Delta(t)y(t)\,dsdt,\end{aligned}$$

since $\int_0^T y(s)\psi(s,\theta_0)\Delta(s)\,ds$ is zero. Finally we need

$$\begin{aligned}&E[\{\text{one}\}\{\text{two}\}' + \{\text{two}\}\{\text{one}\}'] \\ &\qquad = -\int_0^T \int_0^t \{\psi(s,\theta_0)\psi(t,\theta_0)' + \psi(t,\theta_0)\psi(s,\theta_0)'\}\alpha(s)\Delta(t)y(t)\,dsdt.\end{aligned}$$

Write $\alpha(s) = \alpha(s,\theta_0) + \Delta(s)$ here, and find that some terms luckily cancel each other out:

$$K = \int_0^T y\psi_{\theta_0}\psi_{\theta_0}'\alpha\,ds - \int_0^T \int_0^t [\psi(s,\theta_0)\psi(t,\theta_0)' + \psi(t,\theta_0)\psi(s,\theta_0)']\alpha(s,\theta_0)\Delta(t)y(t)\,dsdt.$$

The alternative formula given in the theorem follows upon clever integration by parts. □

Suppose for a minute that the model is in fact true, so that $\alpha(s) = \alpha(s,\theta_0)$. Then J and K agree, and there is an identity

$$J(\alpha_\theta, y, \theta) = K(\alpha_\theta, y, \theta) = \int_0^T y(s)\psi(s,\theta)\psi(s,\theta)'\alpha(s,\theta)\,ds \tag{2.5}$$

which generalises (1.4). The model-based statement $\sqrt{n}(\widehat{\theta} - \theta_0) \to_d N_p\{0, J^{-1}\}$ was one of the main results of Borgan (1984), and further discussion, including matters of optimality, can be found in Hjort (1986a).

To carry out valid large-sample inference about the most fitting parameter θ_0, for example setting an approximate confidence interval for one of the parameter components, one needs a consistent estimator for the asymptotic covariance matrix. Estimators for J and K can be constructed in several ways. The most natural estimators come forward when we express them as functions of the true cumulative hazard $A(.) = \int_0^{\cdot} \alpha(s)\,ds$, the limiting at risk proportion $y(.)$, and the parameter θ_0, and then insert consistent estimators $\widehat{A}(.) = \int_0^{\cdot} dN(s)/Y(s)$, $\widehat{y}(.) = Y(.)/n$, and $\widehat{\theta}$ for these. This leads to

$$\widehat{J} = \int_0^T \frac{Y(s)}{n}\psi(s,\widehat{\theta})\psi(s,\widehat{\theta})'\alpha(s,\widehat{\theta})\,ds - \int_0^T \frac{Y(s)}{n}D\psi(s,\widehat{\theta})\Big\{\frac{dN(s)}{Y(s)} - \alpha(s,\widehat{\theta})\,ds\Big\}, \tag{2.6}$$

and three different expressions for $\widehat{K}$:

$$\begin{aligned}\widehat{K} &= \int_0^T \frac{Y(s)}{n}\psi(s,\widehat{\theta})\psi(s,\widehat{\theta})'\frac{dN(s)}{Y(s)} + \int_0^T \{\psi(t,\widehat{\theta})\widehat{E}(t)' + \widehat{E}(t)\psi(t,\widehat{\theta})'\}\alpha(t,\widehat{\theta})\,dt, \\ &= \int_0^T \psi(s,\widehat{\theta})\psi(s,\widehat{\theta})'\frac{dN(s)}{n} \\ &\quad - \int_0^T \int_0^t \{\psi(s,\widehat{\theta})\psi(t,\widehat{\theta})' + \psi(t,\widehat{\theta})\psi(s,\widehat{\theta})'\}\alpha(s,\widehat{\theta})ds\left\{\frac{dN(t)}{n} - \frac{Y(t)}{n}\alpha(t,\widehat{\theta})dt\right\} \\ &= \frac{1}{n}\sum_{i=1}^n \{\psi(x_i,\widehat{\theta})\delta_i - A^d(x_i,\widehat{\theta})\}\{\psi(x_i,\widehat{\theta})\delta_i - A^d(x_i,\widehat{\theta})\}'.\end{aligned} \tag{2.7}$$

Here $\widehat{E}(t) = \int_0^t \{Y(s)/n\}\psi(s,\widehat{\theta})\{dN(s)/Y(s) - \alpha(s,\widehat{\theta})\,ds\}$ and $A^d(t,\theta) = \int_0^t \psi_\theta \alpha_\theta\,ds$ is the derivative w.r.t. θ of $A(t,\theta) = \int_0^t \alpha_\theta\,ds$. It takes some algebraic skill to show that these are equivalent expressions. The third formula is computationally more convenient and also emerges naturally from the discussion of influence functions in the next section. The important statistical consistency property is however most easily proved using the first formula.

This impliues, for an example, that the ellipsoïd

$$\{\theta\colon(\theta-\widehat{\theta})'\widehat{J}\widehat{K}^{-1}\widehat{J}(\theta-\widehat{\theta}) \le \gamma_{p,.90}/n\}$$

defines an asymptotically correct and model-robust 90% confidence region for the most fitting parameter θ_0, when $\gamma_{p,.90}$ is the upper 10% point of the χ^2_p distribution.

EXAMPLE 2.1. Study once more the exponential model where $\alpha(s,\theta) = \theta$. The maximum likelihood estimator is $\widehat{\theta} = N(T)/\int_0^T Y(s)\,ds = \sum_{i=1}^n \delta_i/\sum_{i=1}^n x_i$. It converges to the most appropriate parameter value $\theta_0 = \int_0^T y(s)\alpha(s)\,ds/\int_0^T y(s)\,ds$, i.e. a y-weighted average of the true hazard rate, by an application of the theorem. Furthermore, the second term of the J expression vanishes, and

$$J = \frac{1}{\theta_0^2}\int_0^T y\alpha\,ds, \quad K = \frac{1}{\theta_0^2}\int_0^T y\alpha\,ds + \frac{2}{\theta_0}\int_0^T\int_0^t y(s)\{\alpha(s)-\theta_0\}ds\,dt,$$

with accompanying estimates $\widehat{J} = \{N(T)/n\}/\widehat{\theta}^2$, $\widehat{K} = \frac{1}{n}\sum_{i=1}^n(\delta_i/\widehat{\theta} - x_i)^2$, cf. (2.6) and (2.7). The asymptotic variance of $\sqrt{n}(\widehat{\theta}-\theta_0)$ is estimated by respectively

$$\frac{\widehat{\theta}^2}{N(T)/n} \quad \text{or} \quad \frac{\widehat{\theta}^4}{\{N(T)/n\}^2}\frac{1}{n}\sum_{i=1}^n\Big(\frac{\delta_i}{\widehat{\theta}} - x_i\Big)^2,$$

under and outside model circumstances. Note that these expressions reduce to those of Example 1.1 when there is no censoring. □

3. Influence functions

This section studies influence functions for estimator functionals in the presence of censoring, and some of their uses are indicated.

The influence function of an estimator is an infinite population concept. Consider for concreteness the non-censored situation of Section 1 first, where data come from F. Assume that an estimator $\widehat{\theta}$ can be expressed as $S(\widehat{F})$, where $\widehat{F}$ is the empirical distribution. Its target value is $\theta_0 = S(F)$. The *influence function* $I(F,x)$ for such a functional is the derivative of $S(F_\varepsilon) = S((1-\varepsilon)F + \varepsilon I_x)$ at $\varepsilon = 0$, writing I_x to denote point mass at x. The ordinary maximum likelihood estimator is for example $\widehat{\theta} = \mathrm{ml}(\widehat{F})$, where $\mathrm{ml}(F)$ is the maximiser of $\int \log f_\theta(x)\, dF(x)$. One can demonstrate that

$$I(F,x) = \lim_{\varepsilon\to 0} \varepsilon^{-1}\{\mathrm{ml}(F_\varepsilon) - \mathrm{ml}(F)\} = J(F,\mathrm{ml}(F))^{-1}\frac{\partial \log f(x,\mathrm{ml}(F))}{\partial\theta}, \tag{3.1}$$

cf. (1.2) and (1.3). — Influence functions are useful for several purposes. It can indicate sensitivity against possible outliers; it provides a tool with which to find the limit distribution of estimators; data-based *empirical influence functions* can be constructed and used to assess the influence of individual data points; it can sometimes be used to construct new estimators with specific desiderata; and empirical and theoretical influence functions enter naturally in studies of the bootstrap and other resampling procedures. General references include Efron (1982), Reid (1983), and Hampel, Ronchetti, Rousseeuw, and Stahel (1986).

A natural task is now to explore influence functions for estimators in the random censorship model of Section 2. Reid (1981) and Reid, Crépeau, and Knafl (1985) have also studied influence functions with censored data, but the present situation is not covered by their work. Let us redescribe the problem in a way suiting the task. We will limit discussion to the maximum likelihood method. The model has been described by saying that partially observed (X_i^0, c_i) pairs come from $F \times G$. Let $H = H_{F,G}$ be the inherited distribution for data pairs $(X_i,\delta_i) = (\min\{X_i^0, c_i\}, I\{X_i^0 \le c_i\})$ in $[0,\infty)\times\{0,1\}$. H has subdistribution functions $H^0(t) = \Pr\{X_i \le t, \delta_i = 0\}$ and $H^1(t) = \Pr\{X_i \le t, \delta_i = 1\}$. The data collection can be represented by the N and Y processes of (2.1), or equivalently by the proportion at risk process $\widehat{y}(s) = Y(s)/n$ with limit $y(s) = F[s,\infty)G[s,\infty)$, and the Nelson–Aalen estimator $\widehat{A}(t) = \int_0^t dN(s)/Y(s)$ with limit $A(t) = \int_0^t \alpha(s)\, ds$. The $\widehat{\theta}$ estimator solves $\int_0^T \psi(s,\theta)\widehat{y}(s)\{d\widehat{A}(s) - \alpha(s,\theta)\, ds\} = 0$ and converges to θ_0, the solution of $\int_0^T \psi(s,\theta)y(s)\{dA(s) - \alpha(s,\theta)\, ds\} = 0$. We may view θ_0 as defined by the pair (F,G), or by (A,y), or by $H = (H^0, H^1)$. Observe that A and y can be recovered from H, by

$$\begin{aligned} y(s) &= \Pr\{X_i \ge s\} = (H^0 + H^1)[s,\infty), \\ y(s)dA(s) &= dH^1(s) = \Pr\{X_i \in [s, s+ds], \delta_i = 1\}. \end{aligned} \tag{3.2}$$

In particular, the maximum likelihood method can be viewed as a functional $\mathrm{ml}(H)$ on the space of $H = (H^0, H^1)$ distributions, and $\widehat{\theta} = \mathrm{ml}(\widehat{H})$, where $\widehat{H}$ is the empirical distribution of data pairs $(x_1,\delta_1),\ldots,(x_n,\delta_n)$.

One might consider several influence measures, corresponding to altering different aspects of the model. One can consider variations in $\mathrm{ml}(A,y)$ when F is replaced by $F_\varepsilon = (1-\varepsilon)F + \varepsilon I_x$, when G is replaced by $G_\varepsilon = (1-\varepsilon)G + \varepsilon I_c$, or both, or replacing (F,G) by $(1-\varepsilon)F\times G + \varepsilon I_{(x,c)}$. These would give different generalisations $I(F,G,x,c)$ of (3.1). The way data are captured suggests however that we should consider local variation of H in the direction of a given point (x,δ) in $[0,\infty)\times\{0,1\}$.

THEOREM 3.1. *Let $\theta_0 = \mathrm{ml}(H)$ for some H under consideration. Under the regularity conditions of Theorem 2.1 the maximum likelihood estimator has influence function*

$$\begin{aligned} I\big(H,(x,\delta)\big) &= \lim_{\varepsilon\to 0} \varepsilon^{-1}\{\mathrm{ml}((1-\varepsilon)H + \varepsilon I_{(x,\delta)}) - \mathrm{ml}(H)\} \\ &= J(H,\theta_0)^{-1}\Big[\psi(x,\theta_0)I\{\delta=1\} - \frac{\partial}{\partial\theta}\int_0^x \alpha(s,\theta_0)\,ds\Big] \\ &= J(H,\theta_0)^{-1}\int_0^T \psi(s,\theta_0)\{dN_{x,\delta}(s) - Y_{x,\delta}(s)\alpha(s,\theta_0)\,ds\}. \end{aligned}$$

Here $J(H,\theta_0)$ is $J(\alpha,y,\theta_0)$ from Section 2, and $N_{x,\delta}(t) = I\{x\le t,\delta=1\}$ and $Y_{x,\delta}(s) = I\{x\ge s\}$ are counting process and at risk process for the single pair (x,δ).

PROOF: Write $H_\varepsilon = (1-\varepsilon)H + \varepsilon I_{(x,\delta)}$. This H_ε gives rise to y_ε and A_ε as follows, using (3.2):

$$\begin{aligned} y_\varepsilon(s) &= (1-\varepsilon)y(s) + \varepsilon\, I\{x\ge s\}, \\ y_\varepsilon(s)dA_\varepsilon(s) &= (1-\varepsilon)y(s)\,dA(s) + \varepsilon\, I\{x\in[s,s+ds],\delta=1\}. \end{aligned}$$

We are to find $\theta_\varepsilon = \mathrm{ml}(H_\varepsilon)$, the solution of

$$u_\varepsilon(\theta) = \int_0^T \psi_\theta(s)y_\varepsilon(s)\{dA_\varepsilon(s) - \alpha_\theta(s)\,ds\} = 0.$$

This can be done by carrying out a first order Taylor expansion analysis. The result is $\theta_\varepsilon - \theta_0 \doteq \{-\frac{\partial u}{\partial\theta}\}_0^{-1}\{\frac{\partial u}{\partial\varepsilon}\}_0\,\varepsilon$, where the partial derivatives of $u_\varepsilon(\theta)$ are evaluated at $\varepsilon=0$ and $\theta=\theta_0$. Some analysis demonstrates that $\varepsilon^{-1}(\theta_\varepsilon-\theta_0)$ tends to the limit given in the theorem. When evaluating $\frac{\partial}{\partial\varepsilon}u_\varepsilon(\theta)$ it is crucial to note that A_ε has a point mass of size $\doteq \varepsilon I\{\delta=1\}/y(x)$ at x. See also Section 4 for a refinement. □

The result of the theorem generalises (3.1), since $\log f(x,\theta) = \log\alpha(x,\theta) - A(x,\theta)$ with derivative $\psi(x,\theta) - A^d(x,\theta)$, and $\delta=1$ in the non-censored case.

The result of Theorem 3.1 is also suggested by the proof of Theorem 2.1, where we in effect showed

$$\widehat\theta - \theta_0 = \mathrm{ml}(\widehat H) - \mathrm{ml}(H) \doteq_d J(H,\theta_0)^{-1}\frac{1}{n}\sum_{i=1}^n \int_0^T \psi(s,\theta_0)\{dN_i(s) - Y_i(s)\alpha(s,\theta_0)\,ds\}, \quad (3.3)$$

writing N_i and Y_i for the counting process and at risk process of individual no. i. Theorem 2.1 could alternatively have been derived after Theorem 3.1 using general asymptotic theory of estimators with influence functions, see e.g. Reid (1983), Gill (1989), and the present Section 4.

Measures of influence for the individual data pairs can be proposed. Let $\widehat H_{(i)}$ be the empirical distribution when (x_i,δ_i) is deleted from the data set. Then

$$\widehat\theta = \mathrm{ml}\big((1-\tfrac{1}{n})\widehat H_{(i)} + \tfrac{1}{n}I_{(x_i,\delta_i)}\big) \doteq \mathrm{ml}(\widehat H_{(i)}) + \tfrac{1}{n}I\big(\widehat H_{(i)},(x_i,\delta_i)\big),$$

which invites using a cross validation type influence measure $I(\widehat{H}_{(i)},(x_i,\delta_i)) \doteq n(\widehat{\theta}-\widehat{\theta}_{(i)})$ for the i'th data pair, where $\widehat{\theta}_{(i)}$ is computed leaving this pair out. It is somewhat simpler to use the approximation

$$\widehat{I}_i = I(\widehat{H},(x_i,\delta_i)) = J(\widehat{H},\widehat{\theta})^{-1}\int_0^T \psi(s,\widehat{\theta})\{dN_i(s) - Y_i(s)\alpha(s,\widehat{\theta})\,ds\} = \widehat{J}^{-1}\widehat{L}_i \qquad (3.4)$$

instead. Note that $J(\widehat{H},\widehat{\theta})$ simply is the $\widehat{J}$ of (2.6), that $\sum_{i=1}^n \widehat{I}_i = 0$, and that

$$\frac{1}{n}\sum_{i=1}^n \widehat{I}_i\widehat{I}_i' = \widehat{J}^{-1}\Big\{\frac{1}{n}\sum_{i=1}^n \widehat{L}_i\widehat{L}_i'\Big\}\widehat{J}^{-1} = \widehat{J}^{-1}\widehat{K}\widehat{J}^{-1} = \widehat{\Sigma}, \qquad (3.5)$$

the estimated asymptotic covariance matrix for $\sqrt{n}(\widehat{\theta}-\theta_0)$, cf. some algebraic manipulations summed up in (2.7).

We propose using the $\widehat{I}_i$'s as a data-analytic tool, to screen data for possible outliers and to identify data pairs with possibly unduly influence. A further suggestion is to "sphere" them, computing $\widehat{\Sigma}^{-1/2}\widehat{I}_i = \widehat{J}^{1/2}\widehat{K}^{-1/2}\widehat{J}^{-1/2}\widehat{L}_i$ instead. These have mean zero and covariance matrix the identity, which should make outliers more easily detectable.

REMARK. Note that we end up with the model-robust covariance estimator since Theorem 3.1 was derived under the agnostic point of view. The influence function under model conditions is similar but with a simpler J^{-1} matrix, see Theorem 2.1. As an example, suppose $F_\theta(t) = 1-\exp(-t^\theta)$ is the Weibull distribution (with a single parameter). Then the estimated influence function is

$$\widehat{I}(x,\delta) = \widetilde{J}^{-1}\{(1+\log x^{\widehat{\theta}})\delta - x^{\widehat{\theta}}\log x^{\widehat{\theta}}\}/\widehat{\theta},$$

where $\widetilde{J}$ is $J(H(.,\widehat{\theta}),\widehat{\theta})$ in the model-based case and $J(\widehat{H},\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^n\{\delta_i + x_i^{\widehat{\theta}}(\log x_i^{\widehat{\theta}})^2\}/\widehat{\theta}^2$ in the model-agnostic case. These are different. In the uncensored $[0,\infty)$ case the first number is simply $1.3504^2/\widehat{\theta}^2$ [from $1+\Gamma''(2) = (1-\gamma)^2+\pi^2/6 = 1.3504^2$]. □

4. Model-based and model-robust bootstrapping

This section briefly studies the large sample behaviour of some natural bootstrapping schemes. The aim is to use the available data to come up with simulated versions $\widehat{\theta}^*$ of the maximum likelihood estimator $\widehat{\theta}$ in such a way that important quantities related to the (partially unknown) distribution of $\widehat{\theta}$ can be estimated from the empirical distribution of $\widehat{\theta}^*$. If interest focusses on some real-valued $\mu = \mu(\theta)$, then the discussion below applies to $\widehat{\mu} = \mu(\widehat{\theta})$ and $\widehat{\mu}^* = \mu(\widehat{\theta}^*)$ instead.

4A. Preliminaries: the maximum likelihood functional. Recall from Section 3 that the maximum likelihood procedure can be seen as a functional operating on distributions $H = (H^0,H^1)$ for (X,δ). The estimator aims at ml(H), the maximiser of $\int_0^T y(\log\alpha_\theta dA - \alpha_\theta\,ds)$, or, equivalently, the solution of $\phi(H,\theta)=0$, where

$$\begin{aligned}\phi(H,\theta) &= \int_0^T y(s)\psi(s,\theta)\{dA(s)-\alpha(s,\theta)\,ds\}\\ &= \int_0^T \psi(s,\theta)\big\{dH^1(s) - (H^0+H^1)[s,\infty)\alpha(s,\theta)\,ds\big\},\end{aligned} \qquad (4.1)$$

utilising the (3.2) correspondence between (A, y) and H (and we could think of $\mathrm{ml}(H)$ as $\mathrm{ml}(A, y)$ instead). The nonparametric estimate $\widehat{H}$ for H is the empirical distribution of the data pairs (x_i, δ_i). There is a small class of parametric counterparts $H(., \widehat{\theta})$ that corresponds to using $A(t, \widehat{\theta}) = \int_0^t \alpha(s, \widehat{\theta})\, ds$ for A and any consistent estimate $\widetilde{y}(t)$ for $y(t)$, for example $\widehat{y}(t) = \exp\{-A(t, \widehat{\theta})\}\widehat{G}[t, \infty)$, employing the Kaplan–Meier estimate $\widehat{G}$ for G. Observe that both $\mathrm{ml}(\widehat{H})$ and $\mathrm{ml}(H(., \widehat{\theta}))$ indeed are equal to $\widehat{\theta}$.

We shall establish that the ml functional is sufficiently smooth, in a precise sense, and shall have occasion to use this to rigorously justify that various natural bootstrapping schemes actually work. For a pair of distributions $H = (H^0, H^1)$ and $H_0 = (H_0^0, H_0^1)$ for (X, δ), consider the supremum type norm

$$\begin{aligned} \|H - H_0\|^2 &= \|H_1^1 - H_0^1\|^2 + \|H_1^0 - H_0^0\|^2 \\ &= \sup_{0 \le t \le T} |H_1^1(t) - H_0^1(t)|^2 + \sup_{0 \le t \le T} |H_1^0(t) - H_0^0(t)|^2. \end{aligned}$$

LEMMA. *The* ml *functional is locally Lipschitz differentiable w.r.t. the norm* $\|H - H_0\|$, *under the conditions underlying Theorems 2.1 and 3.1. In other words*

$$\mathrm{ml}(H) - \mathrm{ml}(H_0) = \int_{[0,\infty)\times\{0,1\}} I(H_0, (x, \delta))\, d(H - H_0)(x, \delta) + r(H_0, H),$$

where $r(H_0, H) = O(\|H - H_0\|^2)$ *as this distance tends to zero.*

PROOF: Single out some H_0 and write $\theta_0 = \mathrm{ml}(H_0)$ in what follows. Consider

$$B(H, (x, \delta)) = \int_0^T \psi(s, \mathrm{ml}(H))\big\{dN_{x,\delta}(s) - Y_{x,\delta}(s)\alpha(s, \mathrm{ml}(H))\, ds\big\},$$

so that the influence function $I(H, (x, \delta))$ of Theorem 3.1 can be written $J(H, \mathrm{ml}(H))^{-1}$ $B(H, (x, \delta))$. Note that $B(.,.)$ acts as a functional derivative of $\phi(H, \theta)$ w.r.t. H in that

$$\phi(H, \theta_0) - \phi(H_0, \theta_0) = \int B(H_0, (x, \delta))\, d(H - H_0)(x, \delta)$$

(even without a remainder term). Write for convenience $D\phi(H, \theta)$ for the $p \times p$ matrix of (ordinary) partial derivatives of $\phi(H, \theta)$ w.r.t. θ. Note that $D\phi(H_0, \theta_0)$ is nothing but the $-J(H_0, \theta_0)$ matrix involved in Theorems 2.1 and 3.1.

We have accordingly derivatives of $\phi(H, \theta)$ in both directions, and can try Taylor expansion. Assume that

$$\phi(H, \theta) = \phi(H_0, \theta_0) + D\phi(H_0, \theta_0)(\theta - \theta_0) + \int B(H_0, (x, \delta))\, d(H - H_0)(x, \delta) + r_0(H, \theta) \quad (4.2)$$

for suitable remainder term $r_0(H, \theta)$. Then solving $\phi(H, \theta) = 0$ to find $\mathrm{ml}(H)$ gives

$$\begin{aligned} \mathrm{ml}(H) - \mathrm{ml}(H_0) &= -D\phi(H_0, \theta_0)^{-1}\Big[\int B(H_0, (x, \delta))\, d(H - H_0)(x, \delta) + r_0(H, \mathrm{ml}(H))\Big] \\ &= \int I(H_0, (x, \delta))\, d(H - H_0)(x, \delta) + J(H_0, \theta_0)^{-1} r_0(H, \mathrm{ml}(H)), \end{aligned}$$

and the lemma is proved provided we can show $r_0(H, \mathrm{ml}(H)) = O(\|H - H_0\|^2)$. For this is suffices to prove that $r_0(H,\theta) = O(\|H - H_0\|\,\|\theta - \theta_0\|)$ in (4.2), in conjunction with $\mathrm{ml}(H) - \mathrm{ml}(H_0) = O(\|H - H_0\|)$. But

$$\begin{aligned} r_0(H,\theta) &= \phi(H,\theta) - \phi(H,\theta_0) - D\phi(H_0,\theta_0)(\theta - \theta_0) \\ &= \left[D\phi(H,\theta_0) + O(\|\theta - \theta_0\|) - D\phi(H_0,\theta_0)\right](\theta - \theta_0) \\ &= O(\|H - H_0\|\,\|\theta - \theta_0\|), \end{aligned}$$

using regularity conditions about third order partial derivatives etcetera. □

Suppose $\widetilde{H}$ is some estimate of H, and let $\widetilde{H}^*$ be the empirical distribution of data pairs (x_i^*, δ_i^*) obtained via some scheme or other. Then

$$\begin{aligned} \widetilde{\theta}^* - \widetilde{\theta} &= \mathrm{ml}(\widetilde{H}^*) - \mathrm{ml}(\widetilde{H}) \\ &= \frac{1}{n}\sum_{i=1}^n I(\widetilde{H}, (x_i^*, \delta_i^*)) + r(\widetilde{H}, \widetilde{H}^*) \\ &= J(\widetilde{H},\widetilde{\theta})^{-1}\frac{1}{n}\sum_{i=1}^n \int_0^T \psi(s,\widetilde{\theta})\{dN_i^*(s) - Y_i^*(s)\alpha(s,\widetilde{\theta})\,ds\} + r(\widetilde{H},\widetilde{H}^*), \end{aligned} \tag{4.3}$$

where $N_i^*(t) = I\{x_i^* \le t, \delta_i^* = 1\}$ and $Y_i^*(t) = I\{x_i^* \ge t\}$ are associated with data pair (x_i^*, δ_i^*), cf. Theorem 3.1. To arrive safely at an a.s. limit distribution result for $\sqrt{n}(\widetilde{\theta}^* - \widetilde{\theta})$ a necessity is a.s. convergence to 0 of $\sqrt{n}\,r(\widetilde{H}, \widetilde{H}^*)$. This follows if $\widetilde{H}^*$ is close enough to $\widetilde{H}$ (a statistical question) and ml(.) is smooth enough (a function space calculus question). The latter point is dealt with in the lemma. Regarding the first point, note that if $\widetilde{H}^*$ is the empirical distribution of data from $\widetilde{H}$, then $\|\widetilde{H}^* - \widetilde{H}\| = O(\{n^{-1}\log\log n\}^{1/2})$ with probability 1 by well-known fluctuation estimates in the Glivenko–Cantelli theorem, from which it follows that $\sqrt{n}\|\widetilde{H}^* - \widetilde{H}\|^2 = O(n^{-1/2}\log\log n)$ a.s. This is also true when $\widetilde{H}$ is non-continuous, and when $\widetilde{H} = \widetilde{H}_n$ itself is random and converges to some fixed H, i.e. $\sqrt{n}\|\widetilde{H}_n^* - \widetilde{H}_n\|^2$ is still $O(n^{-1/2}\log\log n)$ a.s. when $\widetilde{H}_n^*$ is the empirical distribution of data from $\widetilde{H}_n$. See Shao (1989) for similar remarks.

4B. Parametric bootstrapping. Simulate pseudo-data $(X_1^*, \delta_1^*), \ldots, (X_n^*, \delta_n^*)$ from the parametrically estimated model. In other words, simulate X_i^{0*} from the distribution with hazard rate $\alpha(.,\widehat{\theta})$ and c_i^* from $\widehat{G}$, independently, and form $X_i^* = \min\{X_i^{0*}, c_i^*\}$, $\delta_i^* = I\{X_i^{0*} \le c_i^*\}$. (This is actually semi-parametric bootstrapping.) Compute $\widehat{\theta}^*$ from this pseudo-data set, i.e. from the empirical distribution $H(.,\widehat{\theta})^*$, say, of the n pseudo-pairs. Then from (4.3), letting $dM_i^*(s) = dN_i^*(s) - Y_i^*(s)\alpha(s,\widehat{\theta})\,ds$,

$$\begin{aligned} \sqrt{n}(\widehat{\theta}_{\mathrm{pb}}^* - \widehat{\theta}) &= \sqrt{n}\{\mathrm{ml}(H(.,\widehat{\theta})^*) - \mathrm{ml}(H(.,\widehat{\theta}))\} \\ &= J(H(.,\widehat{\theta}),\widehat{\theta})^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^T \psi(s,\widehat{\theta})\,dM_i^*(s) + \sqrt{n}\,r(H(.,\widehat{\theta}), H(.,\widehat{\theta})^*). \end{aligned}$$

This can be used to prove

$$\sqrt{n}(\widehat{\theta}_{\mathrm{pb}}^* - \widehat{\theta}) \to_d N_p\{0, J(H(.,\theta_0),\theta_0)^{-1}\} \text{ a.s.} \tag{4.4}$$

The notation emphasises that there is convergence in distribution with probability 1, i.e. the data-conditional distribution converges to the right limit for almost all sequences of outcomes (X_i, δ_i). Note that the J matrix obtained here is of the 'under true model' type, and is simpler than in the general case described in Theorem 2.1; in fact

$$J(H(.,\theta_0),\theta_0) = \int_0^T y(s)\psi(s,\theta_0)\psi(s,\theta_0)'\alpha(s,\theta_0)\,ds.$$

The first technical point to observe when proving (4.4) is that the M_i^*'s become orthogonal martingales in the conditional framework given data, with variance processes $Y_i^*(s)\alpha(s,\widehat{\theta})\,ds$, and that the proof of Theorem 2.1 works in this framework, with $\alpha(s) = \alpha(s,\widehat{\theta})$ as the underlying true model. See Akritas (1988) for somewhat similar arguments carefully spelled out in a somewhat similar situation. The second point is that the remainder term goes a.s. to zero, actually as $O(n^{-1/2}\log\log n)$ by the lemma and the remark ending 4A.

Sometimes c_i's are known, in which case it is natural to just put $c_i^* = c_i$ in the bootstrapping scheme above, or perhaps more information is otherwise available about the distribution G. Suppose c_i^* is drawn from G_i instead of the sometimes coarse Kaplan–Meier estimate $\widehat{G}$. The limit distribution argument above rests crucially on convergence of $n^{-1/2}\sum_{i=1}^n \int_0^T \psi(s,\widehat{\theta})\,dM_i^*(s)$. This is a martingale with variance equal to the mean value of $n^{-1}\sum_{i=1}^n \int_0^T \psi(s,\widehat{\theta})\psi(s,\widehat{\theta})'Y_i^*(s)\alpha(s,\widehat{\theta})\,ds$, which is $\int_0^T \psi(s,\widehat{\theta})\psi(s,\widehat{\theta})'\widetilde{y}(s)\alpha(s,\widehat{\theta})\,ds$, where $\widetilde{y}(s) = \exp\{-A(s,\widehat{\theta})\}\widetilde{G}[s,\infty)$ and $\widetilde{G}[s,\infty) = n^{-1}\sum_{i=1}^n G_i[s,\infty)$. If only $\widetilde{G}(.)$ tends in probability to the true $G(.)$ then martingale limit methods of Helland (1982) can be called upon to show that (4.4) holds again. This takes in particular care of the situation with known c_i's. One has the same (first order) limit distribution as with $\widehat{G}$ but presumably less sampling variability for fixed n.

4C. Nonparametric bootstrapping. This time draw X_i^{0*} from the nonparametric Kaplan–Meier estimate $\widehat{F}$ instead, in tandem with an independent c_i^* from $\widehat{G}$, as above. This happens to be equivalent to drawing (X_i^*, δ_i^*) pairs independently from $\widehat{H}$, as explained in Efron (1981). Somewhat more elaborate arguments are needed in this case. Let $dM_i^*(s) = dN_i^*(s) - Y_i^*(s)\,d\widehat{A}(s)$. The M_i^*'s become orthogonal martingales in the data-conditional framework, with variance process $Y_i^*(s)d\widehat{A}(s)\{1 - d\widehat{A}(s)\}$. From (4.3)

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_{\rm nb} - \widehat{\theta}) &= \sqrt{n}\{{\rm ml}(\widehat{H}^*) - {\rm ml}(\widehat{H})\} \\ &= J(\widehat{H},\widehat{\theta})^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^T \psi(s,\widehat{\theta})\big[dM_i^*(s) + Y_i^*(s)\{d\widehat{A}(s) - \alpha(s,\widehat{\theta})\,ds\}\big] \\ &\quad + \sqrt{n}\,r(\widehat{H},\widehat{H}^*).\end{aligned}$$

The remainder term again goes a.s. to zero by the efforts of 4A, and $J(\widehat{H},\widehat{\theta})$, which is $\widehat{J}$ of (2.6), is strongly consistent for $J = J(H,\theta_0)$ under the present conditions. The middle term can be written

$$\int_0^T \psi(s,\widehat{\theta})\Big[dM^*(s)/\sqrt{n} + \sqrt{n}\{Y^*(s)/n - \widehat{y}(s)\}\,\{d\widehat{A}(s) - \alpha(s,\widehat{\theta})\,ds\}\Big]$$

and resembles an expression used in the proof of Theorem 2.1. This proof can in fact be copied and used in the present problem with suitable delicate alterations, to show that the middle term tends in distribution a.s. to $N_p\{0, K(H,\theta_0)\}$, where the K matrix is as in Theorem 2.1. The details require some modest machinery for discrete time martingales, as in Helland (1982), and can be taken care of by means similar to those in the Appendix of Hjort (1985b). The end result is

$$\sqrt{n}(\widehat{\theta}_{\rm nb} - \widehat{\theta}) \to_d N_p\{0, J^{-1}KJ^{-1}\} \text{ a.s.} \tag{4.5}$$

4D. Discussion. The consequences of (4.4) and (4.5) are more or less as for the classical non-censored case, discussed briefly after (1.7). The nonparametric bootstrap always works correctly, in the first order large sample sense, as a consequence of (4.5) and Theorem 2.1. The parametric bootstrap creates the correct amount of variability only if the model itself is correct. Otherwise either under- or overestimation could result. (4.4) is statistically meaningful even when the model is wrong, in that it tells about the estimation uncertainty in a situation with data from a correct model at the least false θ_0. If the model does happen to be adequate, then both $\widehat{\theta}^*_{\rm nb}$ and $\widehat{\theta}^*_{\rm pb}$ have the same limit distributions, but the nonparametric one will usually have larger sampling variability. This is for example clear when one writes down the necessary expressions in the situation with censored data from an exponential distribution.

There are other bootstrapping schemes. We noted that all sensible ways of drawing c_i^*'s in the parametric case gives the same large sample behaviour for $\widehat{\theta}^*_{\rm pb}$. This is not quite the case for $\widehat{\theta}^*_{\rm nb}$. If one uses the empirical distribution $\widetilde{G}$ in the case of known c_i's, then the nonparametric scheme with X_i^{0*}'s from $\widehat{F}$ is first of all not equivalent to drawing pairs (X_i^*, δ_i^*)'s from $\widehat{H}$ anymore, and secondly the limit distribution of $\sqrt{n}(\widehat{\theta}^*_{\rm nb} - \widehat{\theta})$ exists but is slightly different from that of $\sqrt{n}(\widehat{\theta} - \theta_0)$.

Our justification proof for the bootstrap schemes used local Lipschitz differentiability of the ml functional. Results (4.4) and (4.5) could have been reached in other ways as well. Rather general function space methods in Gill (1989) and Csörgő and Mason (1989) could be used, but would give somewhat weaker results, without the extra bonus of speed of convergence which our Lipschitz method gives. On the other hands the methods used by these authors would give results even without the almost sure convergence details that partly underlie our proof, and this is relevant in more complex counting process models where perhaps only weak consistency can be proved for $\widehat{\theta}$. It is also worth pointing out that the technical matters were helped by the assumed finiteness of the observation interval $[0,T]$. With likelihoods on the full halfline $[0,\infty)$ the ml functional would not be quite Lipschitz differentiable, and there would also have been difficulties with applying the implicit function theorem, when solving for θ in $\phi(H,\theta) = 0$, if one were to use Gill's machinery.

5. Other estimation methods

We have concentrated on the maximum likelihood estimator $\widehat{\theta}$ in previous sections. Hjort (1986a, Section 3) proved that several of the familiar asymptotic optimality properties enjoyed by this method in classical situations carry over to the present censored data

framework. These properties have however as basic assumption that the parametric model is indeed correct. There is therefore still interest in studying other estimation schemes, that perhaps might be somewhat less inefficient than $\widehat{\theta}$ under the ideal model's home turf conditions but that for example could have better robustness properties outside model conditions. This section briefly discusses some possibilities.

5A. Bayes estimators. If $\pi(\theta)\,d\theta$ is a prior density for θ then the Bayes estimator is $\widehat{\theta}_B = E\{\theta|\text{data}\} = \int \theta L_n(\theta)\pi(\theta)\,d\theta / \int L_n(\theta)\pi(\theta)\,d\theta$. But as far as first order asymptotic behaviour is concerned such estimators are equivalent to the maximum likelihood solution, i.e. $\sqrt{n}(\widehat{\theta}_B - \widehat{\theta})$ goes to zero in probability, even outside model conditions, according to Hjort (1986a, Section 2).

5B. M-type estimators. We saw in Example 2.1 that the maximum likelihood solution in the constant hazard rate model tends to $\theta_0 = \int_0^T y\alpha\,ds / \int_0^T y\,ds$, a weighted average of the true hazard rate over the observation interval. As a consequence small s-values are given much more weight than larger s-values. Perhaps more disturbing is the fact that the somewhat problem-irrelevant censoring distribution G is involved in θ_0, through $y(s) = F[s,\infty)G[s,\infty)$. This is a general feature of the maximum likelihood approach, see (2.3). One could argue that the most fitting constant hazard rate should be $\theta_1 = \int_0^T \alpha\,ds / \int_0^T ds$ instead, or at least that it should be freed of its dependence upon G.

This corresponds to a different weighting of the log-likelihood. Consider in general terms the *weighted likelihood*

$$WL_n(\theta) = \exp\Big[\int_0^T W_n(s)\big\{\log\alpha_\theta(s)\,dN(s) - Y(s)\alpha_\theta(s)\,ds\big\}\Big], \tag{5.1}$$

where $W_n(.)$ is a weight function tending in probability to some $w(.)$, and where the notation is as in Section 2. The corresponding maximum weighted likelihood estimator $\widehat{\theta}_w$ maximises this function, and also solves $\int_0^T W_n\psi_\theta\{dN - Y\alpha_\theta\,ds\} = 0$. An alternative term suggested by an analogy to the non-censored i.i.d. situation is *M-estimators*.

A result about the asymptotic behaviour of such estimators (and more general ones) was reached in Hjort (1985a, Section 4), but only under model conditions. It is now possible to go through the arguments of Section 2 and 3 and apply them to M-estimators. Under appropriate and mild regularity conditions, which include $W_n(s) \to_p w(s)$, it holds that $n^{-1}\log WL_n(\theta)$ tends to $\int_0^T wy(\alpha\log\alpha_\theta - \alpha_\theta)\,ds$, that $\widehat{\theta}_w$ is consistent for the (new) least false parameter $\theta_{0,w}$ that minimises the differently weighted distance measure

$$d_w[\alpha,\alpha_\theta] = \int_0^T wy\big\{\alpha(\log\alpha - \log\alpha_\theta) - (\alpha - \alpha_\theta)\big\}\,ds, \tag{5.2}$$

cf. (2.3), in particular each M-estimator is consistent at the model, and that

$$\sqrt{n}(\widehat{\theta}_w - \theta_{0,w}) \to_d J_w^{-1}N_p\{0,K_w\} = N_p\{0,J_w^{-1}K_wJ_w^{-1}\}, \tag{5.3}$$

in which

$$J_w = \int_0^T wy\big[\psi(.,\theta_{0,w})\psi(.,\theta_{0,w})'\alpha(.,\theta_{0,w}) - D\psi(.,\theta_{0,w})\{\alpha - \alpha(.,\theta_{0,w})\}\big]\,ds,$$

$$K_w = \mathrm{VAR} \int_0^T w(s)\psi(s,\theta_{0,w})\big[dV(s) - Z(s)\{\alpha(s) - \alpha(s,\theta_{0,w})\}\,ds\big],$$

cf. (2.4). We point out that the weight function $W_n(s)$ is allowed to be random here, it can for example be previsible (its value at time s is known at time $s-$), or of the form $G_n(s,\widehat{\theta})$, where $G_n(s,\theta_{0,w})$ is previsible and converges to $w(s,\theta_{0,w})$ in probability. (Such a function's value at time s is *not* known at time $s-$, since it employs $\widehat{\theta}$, which requires all the $[0,T]$-data to be computed.)

This apparatus can now be used to construct a *modified maximum likelihood estimator* $\widehat{\theta}_m$ that avoids being dependent upon the censoring distribution G. The point is to use $W_n(s) = \widehat{G}[s,\infty)^{-1}$, where $\widehat{G}[s,\infty) = \prod_{u\le s}\{1 - dN_c(u)/Y(u)\}$ is the Kaplan–Meier estimator based on the observed censoring times. The accompanying distance measure for $\widehat{\theta}_m$ is (5.2) above with $y(s)w(s) = y(s)G[s,\infty)^{-1} = F[s,\infty) = \exp\{-A(s)\}$, and is perhaps an even more appropriate generalisation of Kullback–Leibler's information distance than (2.3), see Remark 7A. The modified $\widehat{\theta}_m$ is consistent for $\theta_{0,m}$, for example, $\theta_{0,m} = \int_0^T e^{-A}\alpha\,ds/\int_0^T e^{-A}\,ds$ in the exponential model. This points out anew that different estimators might converge to different least false values when the model is incorrect; $\widehat{\theta}_m$ aims here at a value more tied to the 'inverse expected time to failure' interpretation of θ than to the 'constant hazard rate' interpretation.

Another interesting choice is $W_n(s) = \widehat{y}(s)^{-1} = \widehat{F}[s,\infty)^{-1}\widehat{G}[s,\infty)^{-1}$. It converges to $y(s)^{-1}$ and has the effect of freeing the estimator from its dependence on $y(.)$, i.e. from favouring portions of $[0,T]$ with large y over portions with small y. In the exponential case this modificator estimates $\theta_{0,w} = \int_0^T \alpha(s)\,ds/T$, the neutrally weighted hazard rate.

Using the modified estimator entails a loss in efficiency at the model, as $J_w^{-1}K_wJ_w^{-1}$ is a larger matrix than J^{-1}. As an example, study the exponential model, suppose that $\alpha(s) = \theta_0$ prevails, and assume that the censoring distribution is $G(t) = 1 - \exp(-g\theta_0)$, which corresponds to an expected frequency $1/(g+1)$ of (x_i,δ_i) pairs where x_i^0 is truly observed. The maximum likelihood estimator $\widehat{\theta}$ and the two modificators $\widehat{\theta}_{m1}$ and $\widehat{\theta}_{m2}$ mentioned above all take the form $\int_0^T W_n dN/\int_0^T W_nY\,ds$, using respectively $W_n(s) = 1$, $W_n(s) = \widehat{G}[s,\infty)^{-1}$, and $W_n(s) = \widehat{y}(s)^{-1}$. All three are consistent for θ_0 (since the model is in command), and their asymptotic variances can be shown to be respectively

$$\frac{1}{n}\frac{\theta_0^2}{1-\varepsilon},\quad \frac{1}{n}\frac{\theta_0^2}{1-g}\frac{1-\varepsilon^{1-g}}{(1-\varepsilon)^2},\quad \frac{1}{n}\frac{\theta_0^2}{1+g}\frac{(1/\varepsilon)^{1+g}-1}{(\log 1/\varepsilon)^2},$$

in which $\Pr\{X^0 \le T\} = 1 - \exp(-\theta_0 T) = 1 - \varepsilon$. The third estimator is too defensive it its avoidance of the model, and is much worse than the two others for most combinations of g and ε. The second estimator does not lose much efficiency for values of g that signal low or moderate amounts of censoring, say $g \le \frac{1}{2}$. The efficiency loss becomes significant in cases with more than a moderate amount of censoring.

The influence function of an M-estimator can also be found, using arguments presented in Section 3. With notation as there it becomes

$$I\big(H,(x,\delta)\big) = J_w^{-1}\int_0^T w(s)\psi(s,\theta_{0,w})\{dN_0(s) - Y_0(s)\alpha(s,\theta_{0,w})\,ds\}, \qquad (5.4)$$

and an estimator for it can easily be constructed, along with empirical influence measures of the type $\widehat{I}_i = I(\widehat{H}, (x_i, \delta_i))$. If the maximum likelihood method looks non-robust, in that the influence function given in Theorem 3.1 is sensitive to large values of x, then a more robust estimator can be constructed by using an appropriate deflating w-function.

5C. Dynamic likelihood and smoothing. A choice of W_n different in spirit from those considered above is $W_n(s) = w(s) = I\{s \in B\}$, for a suitable subinterval B of $[0,T]$. The resulting estimator $\widehat{\theta}_B$ uses only data for individuals who are at risk at the beginning of B and information about what happens to them during B, and aims at a locally most appropriate $\theta_{0,B}$, the parameter value that minimises $d_B[\alpha, \alpha_\theta]$, say, which is as in (2.3) but integrated only over B. Such estimates could be computed for different subintervals and compared, for example for model checking purposes. (Much more general model checking procedures are in Hjort (1990a).)

A similar but more ambitious idea, both statistically and computationally, is to use a local $B(s) = (s - \frac{1}{2}h, s + \frac{1}{2}h]$ around each given s, to compute an estimate $\widehat{\theta}(s) = \widehat{\theta}_{B(s)}$ based only on local data. This corresponds to a dynamic or local likelihood approach, and is somewhat similar in motivation to work by Hastie and Tibshirani (1987). Choosing once again the constant hazard rate model as an example, $\widehat{\theta}(s) = N\{B(s)\}/\int_{B(s)} Y(s)\,ds$ becomes the dynamic hazard rate estimate at s. This is similar to but not the same as kernel smoothing of the Nelson–Aalen estimator $\int_0^{\cdot} dN/Y$, which is Ramlau-Hansen's (1983) way of nonparametrically estimating a hazard rate. We can also pass to general kernel function smoothing, and for each s maximise

$$\int_0^T K(s-u)\{\log \alpha_\theta(u) dN(u) - Y(u)\alpha_\theta(u)\,du\}$$

to obtain the local or dynamic or smoothing $\widehat{\theta}(s)$, where K is symmetric with maximum at zero. We view these methods as semiparametric approaches to the estimation of a parametric model. (Other useful approaches sharing this particular characteristic are discussed in Hjort (1986b).) Observe that our method can be used also to construct a "dynamic semiparametric density estimator" via $f_\theta(t) = \alpha_\theta(t)\exp\{-A_\theta(t)\}$, and of course works in cases without censoring as well. A dynamic estimator of the normal density can for example easily be constructed, of the form $\widehat{f}(t) = N\{\widehat{\mu}(t), \widehat{\sigma}^2(t)\}$, where $\widehat{\mu}(t)$ and $\widehat{\sigma}(t)$ are obtained locally. These matters will be pursued elsewhere.

6. Regression models for hazard rates

So far we have considered lifetimes to have been drawn from a homogeneous population. Statistically more challenging and important problems arise when the individuals under study also have covariate measurements that may influence the lifetime distribution. In this section two regression models for hazard rates are studied, the traditional semiparametric Cox model with unspecified baseline hazard rate and the fully parametric Cox model with parametric baseline hazard rate. Once more the questions to be discussed include behaviour of the maximum likelihood estimators outside the narrow model assumptions, agnostic estimation of the covariance matrix, and influence measures.

The data set is $(x_1, \delta_1, z_1), \ldots, (x_n, \delta_n, z_n)$, where x_i and δ_i are as in previous sections and z_i is a q-dimensional covariate measurement vector for individual no. i. The hazard rate for this individual is $\alpha_i(s) = \alpha(s|z_i)$. The Cox model postulates that

$$\alpha_i(s) = \alpha(s) \exp(\beta' z_i) = \alpha(s) \exp(\beta_1 z_{i,1} + \cdots \beta_q z_{i,q}), \quad i = 1, \ldots, n, \tag{6.1}$$

where $\alpha(.)$ is an unspecified hazard rate and β is a vector of coefficients. These are traditionally estimated by maximum partial likelihood, see Gill (1984) for a good account of the theory. However, the behaviour of the estimates outside the narrow proportional hazards assumption seems not to have been studied in the literature, except for Hjort (1986a), where the point limit β_0 of the estimates is identified outside model conditions. In 6B below also the limit distribution is found, and a consistent estimate is provided for the covariance matrix.

The success of Cox regression analysis has perhaps had the unintended side effect that practitioners too seldomly invest efforts in studying the baseline hazard $\alpha(.)$. A parametric version, say

$$\alpha_i(s) = \alpha(s, \theta) \exp(\beta' z_i), \quad i = 1, \ldots, n, \tag{6.2}$$

for some p-dimensional θ, if found to be adequate, would lead to more precise estimation of survival probabilities and related quantities and concurrently contribute to a better understanding of the survival phenomenon under study. This is the model studied in 6A below. References where such models have been used, with $\alpha(s, \theta)$ corresponding to the exponential, Weibull, log-normal distribution, or to piece-wise constant hazards, can be found in Kalbfleisch and Prentice (1980, Chapter 3) and Borgan (1984). Hjort (1990a) provides goodness of fit tests for models of type (6.2).

REMARK. For ease of exposition we shall assume throughout this section that data are *i.i.d. realisations* of a triple (X, Δ, Z), with appropriate distribution H in $[0, \infty) \times \{0, 1\} \times \mathcal{R}^q$, and that (X, Δ) come from life time X^0 and censoring time C in the manner described in Section 2. The sequence (N_i, Y_i, M_i) of individual analogues to (2.1), that is $dN_i(s) = I\{X_i \in [s, s + ds], \Delta_i = 1\}$, $Y_i(s) = I\{X_i \geq s\}$, and $M_i(t) = N_i(t) - \int_0^t Y_i(s)\alpha_i(s)\, ds$, also become i.i.d. The i.i.d.-assumption is not crucial; most of the reasoning and results below continue to be valid for example in a setting with non-random censoring times and covariates, with suitable modifications. □

6A. Parametric Cox regression. The model postulates (6.2). Assume that the true state of affairs is of the form $\alpha_i(s) = \alpha(s|z_i) = \alpha_0(s)h_0(z_i)$ for some $\alpha_0(.)$ and some $h_0(.)$; this is the hazard rate that would have been seen if a large data set were collected from individuals with the same covariate vector z_i. The maximum likelihood estimators $\widehat{\theta}$, $\widehat{\beta}$ maximise

$$\frac{1}{n} \log L_n = \frac{1}{n} \sum_{i=1}^{n} \int_0^T \Big[\{\log \alpha(s, \theta) + \beta' z_i\}\, dN_i(s) - Y_i(s)\alpha(s, \theta) \exp(\beta' z_i)\, ds \Big].$$

They also solve $U_n(\theta,\beta)=0$, where U_n has components

$$U_n^{(1)}(\theta,\beta)=\frac{1}{n}\sum_{i=1}^n\int_0^T\psi_\theta\{dN_i-Y_i\alpha_\theta\exp(\beta' z_i)\,ds\}=\int_0^T\psi_\theta\{dG_n^{(0)}-Q_n^{(0)}\alpha_\theta\,ds\},$$

$$U_n^{(2)}(\theta,\beta)=\frac{1}{n}\sum_{i=1}^n\int_0^T z_i\{dN_i-Y_i\alpha_\theta\exp(\beta' z_i)\,ds\}=\int_0^T\{dG_n^{(1)}-Q_n^{(1)}\alpha_\theta\,ds\}.$$

A certain amount of exra notation is necessary here. We use

$$R_n^{(0)}(s)=\frac{1}{n}\sum_{i=1}^n Y_i(s)h_0(z_i)\to_p r^{(0)}(s)=EI\{X\geq s\}h_0(Z),$$

$$R_n^{(1)}(s)=\frac{1}{n}\sum_{i=1}^n Y_i(s)z_ih_0(z_i)\to_p r^{(1)}(s)=EI\{X\geq s\}Zh_0(Z),$$

$$dG_n^{(0)}(s)=\frac{1}{n}\sum_{i=1}^n dN_i(s)\to_p dG^{(0)}(s)=EI\{X\in[s,s+ds],\Delta=1\}=r^{(0)}(s)\alpha_0(s)\,ds,$$

$$dG_n^{(1)}(s)=\frac{1}{n}\sum_{i=1}^n z_i dN_i(s)\to_p dG^{(1)}(s)=EZI\{X\in[s,s+ds],\Delta=1\}=r^{(1)}(s)\alpha_0(s)\,ds,$$

$$Q_n^{(0)}(s,\beta)=\frac{1}{n}\sum_{i=1}^n Y_i(s)\exp(\beta' z_i)\to_p q^{(0)}(s,\beta)=EI\{X\geq s\}\exp(\beta' Z),$$

$$Q_n^{(1)}(s,\beta)=\frac{1}{n}\sum_{i=1}^n Y_i(s)z_i\exp(\beta' z_i)\to_p q^{(1)}(s,\beta)=EI\{X\geq s\}Z\exp(\beta' Z),$$

$$Q_n^{(2)}(s,\beta)=\frac{1}{n}\sum_{i=1}^n Y_i(s)z_iz_i'\exp(\beta' z_i)\to_p q^{(2)}(s,\beta)=EI\{X\geq s\}ZZ'\exp(\beta' Z).$$

If the model is perfect, then $h_0(z)=\exp(\beta_0' z)$ for some β_0 and $R_n^{(0)}(s)=Q_n^{(0)}(s,\beta_0)$, $r^{(0)}(s)=q^{(0)}(s,\beta_0)$, etc.

To study the behaviour of the estimators, observe that the components of $U_n(\theta,\beta)$ have limits

$$\begin{aligned}u^{(1)}(\theta,\beta)&=\int_0^T\psi_\theta\{dG^{(0)}-q^{(0)}(.,\beta)\alpha_\theta\,ds\}=\int_0^T\psi_\theta\{r^{(0)}\alpha_0-q^{(0)}(.,\beta)\alpha_\theta\}\,ds,\\ u^{(2)}(\theta,\beta)&=\int_0^T\{dG^{(1)}-q^{(1)}(.,\beta)\alpha_\theta\,ds\}=\int_0^T\{r^{(1)}\alpha_0-q^{(1)}(.,\beta)\alpha_\theta\}\,ds.\end{aligned}\tag{6.3}$$

These functions determine the limit (θ_0,β_0) of $(\widehat{\theta},\widehat{\beta})$, see the theorem below. Taking second partial derivatives of $n^{-1}\log L_n(\theta,\beta)$ gives a matrix $I_n(\theta,\beta)$, and an expression for its limit in probability can be found. Let J be the limit of $-I_n(\theta_0,\beta_0)$. It has blocks

$$J_{11}=\int_0^T q^{(0)}(.,\beta_0)\psi(.,\theta_0)\psi(.,\theta_0)'\alpha(.,\theta_0)\,ds-\int_0^T D\psi(.,\theta_0)\{r^{(0)}\alpha_0-q^{(0)}(.,\beta_0)\alpha(.,\theta_0)\}\,ds,$$

$$J_{12} = \int_0^T \psi(.,\theta_0) q^{(1)}(.,\beta_0)' \alpha(.,\theta_0)\, ds, \quad J_{22} = \int_0^T q^{(2)}(.,\beta_0)\alpha(.,\theta_0)\, ds.$$

Let finally

$$K = \mathrm{VAR}\begin{bmatrix} \int_0^T \psi(s,\theta_0)\{dN_i(s) - Y_i(s)\alpha(s,\theta_0)\exp(\beta_0' Z_i)\, ds\} \\ \int_0^T Z_i\{dN_i(s) - Y_i(s)\alpha(s,\theta_0)\exp(\beta_0' Z_i)\, ds\} \end{bmatrix}.$$

A somewhat complicated explicit expression can be given for K, as in the proof of Theorem 2.1, this time involving α_0 and h_0, but we will be content with this description and the consistent estimator below. The following result generalises a theorem of Borgan (1984) to outside-the-model conditions.

THEOREM 6.1. *Assume that the equations $u^{(1)}(\theta,\beta) = 0$, $u^{(2)}(\theta,\beta) = 0$ have a unique solution (θ_0,β_0). Suppose further that the regularity conditions on $\alpha(s,\theta)$ stated in Theorem 2.1 hold, that J is positive definite, that the covariates z_i are uniformly bounded as n grows, and that $h_0(z)$ is bounded away from zero and infinity in this bounded domain. Then the maximum likelihood estimators $(\widehat\theta,\widehat\beta)$ are consistent for the least false parameter values (θ_0,β_0). These also minimise the distance measure $d[\alpha_0(.)h_0(.), \alpha_\theta(.)\exp(\beta'.)]$ given in (6.6) below. Furthermore,*

$$\begin{bmatrix} \sqrt{n}(\widehat\theta - \theta_0) \\ \sqrt{n}(\widehat\beta - \beta_0) \end{bmatrix} \to_d J^{-1} N_{p+q}\{0,K\} = N_{p+q}\{0, J^{-1}KJ^{-1}\},$$

where J and K are given above. A consistent estimator for J is $\widehat J$, with blocks

$$\widehat J_{11} = \int_0^T Q_n^{(0)}(.,\widehat\beta)\psi(.,\widehat\theta)\psi(.,\widehat\theta)'\alpha(.,\widehat\theta)\, ds - \int_0^T D\psi(.,\widehat\theta)\{dG_n^{(0)}(s) - Q_n^{(0)}(.,\widehat\beta)\alpha(.,\widehat\theta)\, ds\},$$

$$\widehat J_{12} = \int_0^T \psi(.,\widehat\theta) Q_n^{(1)}(.,\widehat\beta)'\alpha(.,\widehat\theta)\, ds, \quad \widehat J_{22} = \int_0^T Q_n^{(2)}(.,\widehat\beta)\alpha(.,\widehat\theta)\, ds.$$

Finally

$$\widehat K = \frac{1}{n}\sum_{i=1}^n \widehat L_i \widehat L_i', \quad \widehat L_i = \begin{bmatrix} \psi(x_i,\widehat\theta)\delta_i - \exp(\widehat\beta' z_i) A^d(x_i,\widehat\theta) \\ z_i\{\delta_i - \exp(\widehat\beta' z_i) A(x_i,\widehat\theta)\} \end{bmatrix}$$

is a consistent estimator for K.

We note that the regularity conditions can be weakened, along the lines of Borgan (1984, Section 6), but that those given here should be satisfied in most practical applications. Note also that uniqueness of the root of $u(\theta,\beta) = 0$, or of the minimiser of the (6.6) distance, follows if the log-likelihood function is concave.

PROOF: The consistency part can essentially be handled using methods of Hjort (1986a, Theorem 2.3). The asymptotic normality part is similar to the proof of Theorem 2.1, again using the Taylor expansion argument (1.2). One has to employ the martingales $M_i(t) = N_i(t) - \int_0^t Y_i(s)\alpha_0(s)h_0(z_i)\, ds$ where Borgan (1984) was allowed by the model to use $N_i(t) - \int_0^t Y_i(s)\alpha(s,\theta_0)\exp(\beta_0' z_i)\, ds$, and take the additional variability into account. The variables whose covariance matrix defines K above split into a martingale term and an

additional more complicated term that comes from incorrectness of the model; only under model circumstances does the second term vanish and K become equal to J. Consistency of $\widehat{J}$ and $\widehat{K}$ can be established using martingale inequalities and uniform convergence in probability arguments that for example can be gleaned from Hjort (1990a, Section 2). Let's leave it at that. □

Measures of influence become even more important in the presence of covariates. Let H be the distribution of (X, Δ, Z), and let (x, δ, z) be fixed. Then $H_\varepsilon = (1-\varepsilon)H + \varepsilon I_{(x,\delta,z)}$ represents a small perturbation of H in direction (x, δ, z), and the least false $(\theta_\varepsilon, \beta_\varepsilon)$ determined by H_ε can be studied. It is by the theorem the solution to $u_\varepsilon^{(1)}(\theta,\beta) = 0$, $u_\varepsilon^{(2)}(\theta,\beta) = 0$, where u_ε is as in (6.3), but with

$$\begin{aligned} dG_\varepsilon^{(j)}(s) &= (1-\varepsilon)dG^{(j)}(s) + \varepsilon z^j\, I\{x \in [s, s+ds], \delta = 1\}, \\ q_\varepsilon^{(j)}(s,\beta) &= (1-\varepsilon)q^{(j)}(s,\beta) + \varepsilon z^j\, I\{x \geq s\} \exp(\beta' z), \end{aligned} \tag{6.4}$$

for j equal to 0 and 1. Note that $G_\varepsilon^{(0)}$ and $G_\varepsilon^{(1)}$ have positive point masses at x if $\delta = 1$. Analysis as in the simpler case covered by Theorem 3.1 gives at the end of the night the influence function

$$\begin{aligned} I(H,(x,\delta,z)) &= \lim_{\varepsilon\to 0} \begin{bmatrix} \{\theta_0(H_\varepsilon) - \theta_0(H)\}/\varepsilon \\ \{\beta_0(H_\varepsilon) - \beta_0(H)\}/\varepsilon \end{bmatrix} \\ &= J^{-1} \begin{bmatrix} \psi(x,\theta_0)\delta - \exp(\beta_0' z) A^d(x,\theta_0) \\ z\{\delta - \exp(\beta_0' z) A(x_i,\theta_0)\} \end{bmatrix} \\ &= J^{-1} \begin{bmatrix} \int_0^T \psi(s,\theta_0)\{dN_0(s) - Y_0(s)\alpha(s,\theta_0)\exp(\beta_0' z)\, ds\} \\ \int_0^T z\{dN_0(s) - Y_0(s)\alpha(s,\theta_0)\exp(\beta_0' z)\, ds\} \end{bmatrix}, \end{aligned}$$

in which N_0 and Y_0 are counting process and at risk process for (x, δ). Natural diagnostic measures for influence are

$$\widehat{I}_i = I(\widehat{H},(x_i,\delta_i,z_i)) = \widehat{J}^{-1}\widehat{L}_i = \widehat{J}^{-1} \begin{bmatrix} \int_0^T \psi(s,\widehat{\theta})\{dN_i(s) - Y_i(s)\alpha(s,\widehat{\theta})\exp(\widehat{\beta}' z_i)\, ds\} \\ \int_0^T z_i\{dN_i(s) - Y_i(s)\alpha(s,\widehat{\theta})\exp(\widehat{\beta}' z_i)\, ds\} \end{bmatrix}, \tag{6.5}$$

where an alternative expression for $\widehat{L}_i$ is given in the theorem, and $\widehat{H}$ is the empirical distribution of the n triples (x_i, δ_i, z_i). It is also an approximation to the crossvalidated $I(\widehat{H}_{(i)},(x_i,\delta_i,z_i))$ and to $(n(\widehat{\theta} - \widehat{\theta}_{(i)}), n(\widehat{\beta} - \widehat{\beta}_{(i)}))$, see Section 3. A further important property of these empirical influence measures is that their empirical covariance matrix becomes $\widehat{\Sigma} = \widehat{J}^{-1}\widehat{K}\widehat{J}^{-1}$, as in (3.5). We propose computing the sphered influence measures $\widehat{\Sigma}^{-1/2}\widehat{I}_i$, which have mean zero and empirical covariance matrix the identity in dimension $p+q$, to screen data for outliers and for individual data triples with particular influence.

Let us end this subsection with exhibiting the distance measure between hazard rates with respect to which (θ_0, β_0) chosen by the maximum likelihood procedure is least false, cf. the first part of the Introduction. We reach slightly more general insight by writing $\alpha(s|z) = \alpha_\theta(s)h_\beta(z)$ for the parametric model, instead of the special case (6.2), and $\alpha_0(s)h_0(z)$ for the true model. Under these circumstances one can show that

$$\frac{1}{n}\log L_n(\theta,\beta) \to_p \int_0^T \{r^{(0)}(s)\log \alpha_\theta(s)\,\alpha_0(s) + r^{(1)}(s,\beta)\alpha_0(s) - q^{(0)}(s,\beta)\alpha_\theta(s)\}\, ds,$$

where the functions entering the integrand are given below, and also expressed as integrals over the covariate space $\mathcal{Z}$ with respect to the covariate distribution $D(dz)$ for Z:

$$r^{(0)}(s) = EI\{X \geq s\}h_0(Z) = \int_{\mathcal{Z}} y(s|z)h_0(z)\,D(dz),$$
$$r^{(1)}(s,\beta) = EI\{X \geq s\}h_0(Z)\log h_\beta(Z) = \int_{\mathcal{Z}} y(s|z)h_0(z)\log h_\beta(z)\,D(dz),$$
$$q^{(0)}(s,\beta) = EI\{X \geq s\}h_\beta(Z) = \int_{\mathcal{Z}} y(s|z)h_\beta(z)\,D(dz).$$

These equations also feature the z-dependent $y(s|z) = \Pr\{X \geq s|z\}$. Consider the z-dependent hazard distance from $\alpha_0(.)h_0(z)$ to $\alpha_\theta(.)h_\beta(z)$, as measured by the already encountered distance measure (2.3), that is

$$d_z[\alpha_0(.)h_0(z), \alpha_\theta(.)h_\beta(z)] = \int_0^T y(s|z)\Big[\alpha_0(s)h_0(z)\log\frac{\alpha_0(s)h_0(z)}{\alpha_\theta(s)h_\beta(z)} - \{\alpha_0(s)h_0(z) - \alpha_\theta(s)h_\beta(z)\}\Big]\,ds.$$

It is now a matter of careful checking to see that maximising the limit of $n^{-1}\log L_n(\theta,\beta)$ is the same as minimising the z-weighted distance function

$$d[\alpha_0\,h_0, \alpha_\theta\,h_\beta] = \int_{\mathcal{Z}} d_z[\alpha_0(.)h_0(z), \alpha_\theta(.)h_\beta(z)]\,D(dz). \tag{6.6}$$

6B. Semiparametric Cox regression. The model postulates (6.1), where $\alpha(.)$ is left unspecified. Let us, conservatively and counterbalancedly, assume only that $\alpha_i(s) = \alpha(s|z_i) = \alpha(s)h_0(z_i)$ for some $\alpha(.)$ and some $h_0(.)$. The Cox estimator maximises the partial log likelihood

$$\log L_n(\beta) = \sum_{i=1}^n \int_0^T \Big[\beta' z_i - \log\Big\{\sum_{j=1}^n Y_j(s)\exp(\beta' z_j)\Big\}\Big]\,dN_i(s),$$

see for example Gill (1984). It is also a root of

$$U_n(\beta) = \frac{1}{n}\sum_{i=1}^n \int_0^T \Big\{z_i - \frac{Q_n^{(1)}(s,\beta)}{Q_n^{(0)}(s,\beta)}\Big\}\,dN_i(s) = \int_0^T \{dG_n^{(1)}(s) - E_n(s,\beta)\,dG_n^{(0)}(s)\},$$

where notation is as in 6A and $E_n = Q_n^{(1)}/Q_n^{(0)}$, with limit $e(s,\beta) = q^{(1)}(s,\beta)/q^{(0)}(s,\beta)$. We have

$$U_n(\beta) \to_p u(\beta) = \int_0^T \{dG^{(1)}(s) - e(s,\beta)\,dG^{(0)}(s)\} = \int_0^T \{r^{(1)}(.) - \frac{q^{(1)}(.,\beta)}{q^{(0)}(.,\beta)}r^{(0)}(.)\}\,dA(s). \tag{6.7}$$

If the model is perfect, then $r^{(0)} = q^{(0)}(.,\beta_0)$ and $r^{(1)} = q^{(1)}(.,\beta_0)$ for some β_0, and in particular $u(\beta_0) = 0$. The consistency part of the theorem below generalises this; once more there is a least false parameter β_0 even when the (6.1) model is incorrect.

We shall also need the second order partial derivatives of $n^{-1}\log L_n$, aiming once more at establishing limit distributions via Taylor expansion and (1.2). One finds

$$-I_n(\beta_0) = \int_0^T \Big\{ \frac{Q_n^{(2)}(s,\beta_0)}{Q_n^{(0)}(s,\beta_0)} - E_n(s,\beta_0)E_n(s,\beta_0)' \Big\}\, dG_n^{(0)}(s)$$
$$\to_p J = \int_0^T \Big\{ \frac{q^{(2)}(s,\beta_0)}{q^{(0)}(s,\beta_0)} - e(s,\beta_0)e(s,\beta_0)' \Big\}\, r^{(0)}(s)\, dA(s).$$

Observe that the formula usually given in the literature for this information matrix has the model-based $q^{(0)}(s,\beta_0)$ in lieu of our agnostic $r^{(0)}(s)$. Let finally

$$K = \mathrm{VAR}\int_0^T \{Z - e(s,\beta_0)\}\Big\{dN_0(s) - Y_0(s)\exp(\beta_0' Z)\frac{r^{(0)}(s)}{q^{(0)}(s,\beta_0)}\alpha(s)\,ds\Big\},$$

where $N_0(t) = I\{X \le t, \Delta = 1\}$, $Y_0(s) = \{X \ge s\}$, and (X,Δ,Z) has distribution H. A long and complicated explicit expression can be obtained for $K = K(H)$, but the description here suffices for our purposes. What is important is knowing the existence of this matrix and how it enters the limit distribution, and having an explicit consistent estimator, which we provide below.

THEOREM 6.2. *Suppose that the regularity conditions of Hjort (1986a, Theorem 4.1) hold. Then the Cox estimator $\widehat\beta$ is consistent for the least false parameter value β_0 that uniquely solves $u(\beta) = 0$. This parameter value also minimises the distance function $d[h_0(.),\exp(\beta'.)]$ given in (6.9) below. Furthermore*

$$\sqrt{n}(\widehat\beta - \beta_0) \to_d J^{-1}N_q\{0,K\} = N_q\{0, J^{-1}KJ^{-1}\},$$

where J and K are given above, and

$$\widehat J = \int_0^T \Big\{ \frac{Q_n^{(2)}(s,\widehat\beta)}{Q_n^{(0)}(s,\widehat\beta)} - E_n(s,\widehat\beta)E_n(s,\widehat\beta)' \Big\}\, dG_n^{(0)}(s),$$

$$\widehat K = \frac{1}{n}\sum_{i=1}^n \widehat L_i \widehat L_i', \quad \widehat L_i = \int_0^T \{z_i - E_n(s,\widehat\beta)\}\Big\{dN_i(s) - Y_i(s)\exp(\widehat\beta' z_i)\frac{dG_n^{(0)}(s)}{Q_n^{(0)}(s,\widehat\beta)}\Big\}$$

are consistent estimators.

PROOF: Consistency and uniqueness of β_0 was established in Hjort (1986a, Section 4). Methods provided there are also sufficient to demonstrate $-I_n(\widetilde\beta) \to_p J$ in the appropriate analogue of (1.2). What remains to be shown, therefore, is that $\sqrt{n}\,U_n(\beta_0) \to_d N_q\{0,K\}$. This can be done along the lines of the proof of Theorem 2.1, though matters become much more involved. One establishes that

$$\sqrt{n}\,U_n(\beta_0) \doteq_d \frac{1}{\sqrt{n}}\sum_{i=1}^n \{z_i - E_n(s,\beta_0)\}\, dM_i(s) + \int_0^T \sqrt{n}\{W_n(s) - w(s)\}\, dA(s),$$

where $M_i(t) = N_i(t) - \int_0^t Y_i(s)\alpha(s)h_0(z_i)\,ds$ is a martingale and where

$$W_n(s) = \frac{1}{n}\sum_{i=1}^{n} Y_i(s)\{z_i - e(s,\beta_0)\}\Big\{h(z_i) - \frac{r^{(0)}(s)}{q^{(0)}(s,\beta_0)}\exp(\beta_0' z_i)\Big\}$$

with expected value $w(s) = r^{(1)}(s) - e(s,\beta_0)r^{(0)}(s)$. Note that our M_i is different from the traditionally employed $N_i(t) - \int_0^t Y_i(s)\alpha(s)\exp(\beta_0' z_i)\,ds$, see Andersen and Gill (1982) and Gill (1984). These authors found themselves in the luxurious possession of a perfect model, in which case the second term above vanishes, since $W_n(s)$ then is zero. The rest of the proof therefore generalises the model-based proof by establishing joint convergence in distribution of the martingale $n^{-1/2}\sum_{i=1}^n \int_0^{\cdot}\{z_i - E_n(s,\beta_0)\}\,dM_i(s)$ and $\sqrt{n}\{W_n(s) - w(s)\}$, in the function space $D[0,T]\times D[0,T]$, and finally computing the covariance matrix, which indeed becomes K. Inserting consistent estimators for unknown parameters and functions in this expression gives a consistent estimator $\widetilde{K}$. A long algebraic exercise reminiscent of the manipulations that led to the simplified third expression in (2.7) shows that $\widetilde{K} = \widehat{K}$. □

Let us also provide the influence function for the semiparametric Cox model. This can be found by analysis similar to that carried out in 6A for the parametric Cox model. Suppose $\beta_0(H_\varepsilon)$ is the least false parameter vector under $H_\varepsilon = (1-\varepsilon)H + \varepsilon I_{(x,\delta,z)}$, i.e. the solution to $u_\varepsilon(\beta) = 0$, where u_ε is as in (6.7) but with appropriate $dG_\varepsilon^{(0)}$ and $dG_\varepsilon^{(1)}$ instead, as in (6.4). One finds that $\{\beta_0(H_\varepsilon) - \beta_0(H)\}/\varepsilon$ tends to $\{-\frac{\partial u}{\partial\beta}\}_0^{-1}\{\frac{\partial u}{\partial\varepsilon}\}_0$, where the partial derivatives are evaluated at $\beta = \beta_0$ and $\varepsilon = 0$. The first matrix $\{-\frac{\partial u}{\partial\beta}\}_0$ is simply J. Taking the point masses at x for both $G_\varepsilon^{(0)}$ and $G_\varepsilon^{(1)}$ into account one reaches

$$\begin{aligned} I(H,(x,\delta,z)) &= J^{-1}\Big[\Big\{z - \frac{q^{(1)}(x,\beta_0)}{q^{(0)}(x,\beta_0)}\Big\}\delta - \int_0^x \frac{\exp(\beta_0' z)}{q^{(0)}(s,\beta_0)}\Big\{z - \frac{q^{(1)}(s,\beta_0)}{q^{(0)}(s,\beta_0)}\Big\}\,dG^{(0)}(s)\Big] \\ &= J^{-1}\int_0^T \Big\{z - \frac{q^{(1)}(s,\beta_0)}{q^{(0)}(s,\beta_0)}\Big\}\Big\{dN_0(s) - Y_0(s)\exp(\beta_0' z)\frac{r^{(0)}(s)}{q^{(0)}(s,\beta_0)}\,\alpha(s)ds\Big\}; \end{aligned}$$

N_0 and Y_0 belong once more to the single triple (x,δ,z).

The discussion ending subsection 6A can now be repeated with small changes. The empirical influence function is $I(\widehat{H},(x,\delta,z))$, and the natural influence measure for data triple (x_i,δ_i,z_i) becomes

$$\widehat{I}_i = I\big(\widehat{H},(x_i,\delta_i,z_i)\big) = \widehat{J}^{-1}\widehat{L}_i, \tag{6.8}$$

where $\widehat{L}_i$ is given the Theorem 6.2. These sum to zero and have empirical covariance matrix equal to the important $\widehat{\Sigma}$, the estimate for the limiting covariance matrix of $\sqrt{n}(\widehat{\beta} - \beta_0)$. The sphered versions $\widehat{\Sigma}^{-1/2}\widehat{I}_i$ have the identity matrix as empirical covariance matrix, and sore thumbs should stick out.

Reid, Crépeau, and Knafl (1985) also gave an influence function for the Cox regression model. They used another method and did not make it clear that their evaluations in fact were valid also outside the model conditions. They reached an influence measure in their formula (2), given in a form very different from ours, but it turns out to be identical to (6.8).

Let us finally provide the distance measure under which the β_0 parameter chosen by the Cox method is least false, in the spirit of the introductory remarks of Section 1. Let us be slightly more general and allow $\alpha(s|z) = \alpha(s)h_\beta(z)$ for the model, instead of (6.1), and suppose the truth is $\alpha(s)h_0(z)$. Then $\frac{1}{n}\log L_n(\beta)$ can be shown to converge in probability to

$$\lambda(h_0, h_\beta) = \int_0^T \{r^{(1)}(s,\beta) - r^{(0)}(s)\log q^{(0)}(s,\beta)\}\alpha(s)\,ds,$$

using the same notation as in (6.6). One can now show that the maximum of $\lambda(h_0, g)$ over all g functions is $\lambda(h_0, h_0)$. [One possibility is to prove it first in the simple case of a finite support $\{z_1, \ldots, z_m\}$ for the design variable distribution $D(dz)$ for Z, where the problem becomes one of maximising a given function with respect to $g(z_1), \ldots, g(z_m)$. Then one can pass to the general case with appropriate limit arguments.] Hence there is a natural distance measure with respect to which Cox's maximum partial likelihood estimator converges to the least false value:

$$\begin{aligned} d[h_0(.), h_\beta(.)] &= \lambda(h_0, h_0) - \lambda(h_0, h_\beta) \\ &= \int_0^T \Big[EI\{X \geq s\}h_0(Z)\log\frac{h_0(Z)}{h_\beta(Z)} - EI\{X \geq s\}h_0(Z)\log\frac{EI\{X \geq s\}h_0(Z)}{EI\{X \geq s\}h_\beta(Z)}\Big]\,dA(s) \\ &= \int_{\mathcal{Z}}\int_0^T y(s|z)\Big[\log\frac{h_0(z)}{h_\beta(z)} - \log\frac{EI\{X \geq s\}h_0(z)}{EI\{X \geq s\}h_\beta(z)}\Big]dA(s)\,h_0(z)D(dz). \end{aligned} \tag{6.9}$$

7. Discussion and concluding remarks

In this final section a couple of complementary remarks are offered, some of which point to further research.

7A. Some identities in the absence of censoring. General formulae were derived under censoring circumstances in Section 2, and these should reduce to the more familar ones of Section 1 when no censoring is present and the observation period is $[0, \infty)$. Without censoring the y of (2.2) is simply $\exp(-A)$, writing A and A_θ for the cumulative hazard rates. The identities below are valid in this $y = \exp(-A)$ case.

The new formula for the limit of $n^{-1}\log L_n(\theta)$ is $\int_0^T y(\alpha\log\alpha_\theta - \alpha_\theta)\,dt$. The densities can be written $f_\theta = \alpha_\theta\exp(-A_\theta)$ and $f = \alpha\exp(-A)$. Integration by parts yields

$$\int_0^T y(\alpha\log\alpha_\theta - \alpha_\theta)\,dt = \int_0^T f\log f_\theta\,dt - e^{-A(t)}A_\theta(T), \tag{7.1}$$

and we have $\int_0^\infty f\log f_\theta\,dt$ when T grows. When this identity for α is applied also to α_θ, we find for the new distance measure (2.3) between hazard rates

$$d[\alpha, \alpha_\theta] = \int_0^T f\log(f/f_\theta)\,dt - e^{-A(T)}\{A(T) - A_\theta(T)\}. \tag{7.2}$$

Accordingly this distance generalises the Kullbak–Leibler information distance.

The new formula for the limit of $n^{-1}\partial \log L_n(\theta)/\partial\theta$ is $\int_0^T y\psi_\theta(\alpha-\alpha_\theta)\,dt$. Taking partial derivatives of the first identity gives

$$\int_0^T y\psi_\theta(\alpha-\alpha_\theta)\,dt = \int_0^T f\frac{\partial \log f_\theta}{\partial\theta}\,dt - e^{-A(T)}\int_0^T \alpha_\theta\psi_\theta\,dt,$$

and we have the appropriate limit when T grows. Taking second order partial derivatives of the same identity yields

$$\int_0^T e^{-A}\left[\psi_\theta\psi_\theta'\alpha_\theta - D\psi(.,\theta)(\alpha-\alpha_\theta)\right]dt$$
$$= -\int_0^T f\frac{\partial^2 \log f_\theta}{\partial\theta\partial\theta}\,dt + e^{-A(T)}\int_0^T \left[\psi_\theta\psi_\theta' + D\psi(s,\theta)\right]\alpha_\theta\,dt.$$

In particular the J matrix of Section 2 becomes $-E_F\partial^2 \log f_\theta/\partial\theta\partial\theta$ when T reaches infinity.

Consider finally the K matrix. One can show that

$$K = \int_0^T e^{-A}\psi_{\theta_0}(\psi_{\theta_0})'\alpha\,dt + \int_0^T \left[E(\psi_{\theta_0})' + \psi_{\theta_0}E'\right]\alpha_{\theta_0}\,dt$$
$$= \int_0^T (\psi_{\theta_0} - A^d_{\theta_0})(\psi_{\theta_0} - A^d_{\theta_0})'e^{-A}\alpha\,dt + e^{-A(T)}A^d_{\theta_0}(T)A^d_{\theta_0}(T)',$$

where $A^d_\theta(t) = \int_0^t \alpha_\theta(s)\psi_\theta(s)\,ds$ is the derivative of $A_\theta(t)$ w.r.t. θ. Note that the usual score function is the derivative of the logarithm of $f_\theta(t) = \alpha_\theta(t)\exp\{-A_\theta(t)\}$, that is, $L_\theta(t) = \psi_\theta(t) - A^d_\theta(t)$. In the limit as T grows we have $K = \int_0^\infty L_\theta(t)L_\theta(t)'\,dF(t)$, as we should.

7B. General counting process models. For ease of exposition our basic framework has been that of the random censorship model. Most of our arguments use martingale theory only, however, and go through with minor modifications for general and multivariate parametric counting process models, see Andersen and Borgan (1985) for a review of relevant methods. One particular detail that does become more difficult is that of almost sure convergence of the maximum likelihood estimator. In the structurally simplest versions of a parametric counting process models, as in Borgan (1984) and Hjort (1986a), only convergence in probability has been established. This does not affect the theory of Sections 2, 3, 5, but some small amendments are called for regarding the equivalent of Section 4 for such general models. The principal difference is that results (4.4) and (4.5) for the bootstrap must be phrased differently; the bootstrap distributions converge in probability only. This will follow by applying the apparatus of Section 4 without Lipschitz diffentiability but with Hadamard differentiability instead, see Gill (1989, Section 4). The methods of Csörgő and Mason could conceivably also be used.

7C. Bootstrapping in regression models for survival data. Section 4 treated only homogeneous models. Consider for concreteness the parametric Cox model (6.2) for data (X_i, δ_i, z_i) with distribution H. More than simply 'model-based' and 'model-robust' bootstrapping schemes can be proposed in such a situation. Scheme 1 could be to generate

z_i^* from some estimated covariance distribution, nonparametric or parametric, and then X_i^{0*} from the distribution with hazard $\alpha(s,\widehat{\theta})\exp(\widehat{\beta}'z_i^*)$ along with c_i^* from some suitable G_i, for example the Kaplan–Meier estimate for the censoring distribution. One might also just keep $z_i^* = z_i$ for individual i. This scheme gives one way of obtaining $(X_i^*, \delta_i^*, z_i^*)$, trying to be as faithful to the postulated model as possible. Scheme 2 could be to resample triplets, i.e. from the empirical distribution $\widehat{H}$. This method ignores all the finer structure of the model. Scheme 3 could be in the semiparametric Cox spirit and simulate X_i^{0*} from the estimated distribution $\widehat{F}_i(t) = 1 - \prod_{[0,t]}\{1 - d\widehat{A}(s)\}^{\exp(\widehat{\beta}'z_i)}$, cf. Hjort (1985b, Section 1). Scheme 4 could use a nonparametric smoother for the relative risk part instead of $\exp(\widehat{\beta}'z_i)$. As indicated each of these schemes will have its sub-schemes.

The first order behaviour of all these schemes can be sorted out with the methods developed in this paper, under and outside model conditions. This also goes for similar schemes for the semiparametric Cox model. This careful cataloguing is left for future work. Let us merely mention one result, which judicious calculations will show: All schemes indicated above are first order asymptotically correct if the (6.2) model is correct, in the sense that $(\sqrt{n}(\widehat{\theta}^* - \widehat{\theta}), \sqrt{n}(\widehat{\beta}^* - \widehat{\beta}))'$ has the same limiting distribution, with probability 1, as $(\sqrt{n}(\widehat{\theta} - \theta_0), \sqrt{n}(\widehat{\beta} - \beta_0))'$. See the first part of Hjort (1985b) for the kind of arguments that would be needed, in addition to Sections 3 and 4 of the present paper. Scheme 1 would however display smaller sampling variability than Scheme 2.

7D. Finer bootstrap analysis. Our study has been a first order large sample one, regarding both behaviour of estimates and of bootstrapped versions of them. One could enter the more difficult world of second order expansions and second order correct confidence intervals as well. At least in the random censorship model it should be possible to show that bootstrapping based on studentised statistics provide second order correct intervals, that is, approximate the distribution of $t = \sqrt{n}\{\mu(\widehat{\theta}) - \mu(\theta_0)\}/\widehat{\tau}$ with that of $t^* = \sqrt{n}\{\mu(\widehat{\theta}^*) - \mu(\widehat{\theta})\}/\widehat{\tau}^*$, where $\widehat{\tau}$ is an estimate of the limiting standard deviation for $\sqrt{n}\{\mu(\widehat{\theta}) - \mu(\theta_0)\}$ and $\widehat{\tau}^*$ its bootstrap sister. Methods of Hall (1988) are relevant here, as would second order methods for martingales, as rudimentarily presented in the Appendix of Hjort (1985b). In the latter paper second order correct intervals of Efron's ABC variety are constructed for the parameters in Cox' regression model.

References

Akritas, M.G. (1986). Bootstrapping the Kaplan–Meier estimator. *J. Amer. Statist. Assoc.* **81**, 1032–1038.

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

Andersen, P.K. and Borgan, Ø. (1985). Counting process models for life history data: A review (with discussion). *Scand. J. Statist.* **12**, 97–158.

Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

Borgan, Ø. (1984). Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data. *Scand. J. Statist.* **11**, 1–16. Corrigendum, *ibid.* p. 275.

Cox, D.R. (1962). Further results on tests on separate families of hypotheses. *J. Royal Statist. Soc. B* **24**, 406–424.

Csörgő, S. and Mason, D.M. (1989). Bootstrapping empirical processes. *Ann. Statist.* **17**, 1447–1471.

Efron, B. (1981). Censored data and the bootstrap. *J. Amer. Statist. Assoc.* **76**, 312–319.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans.* SIAM–NSF, CBMS #38, Philadelphia.

Fernholz, L.T. (1983). *von Mises calculus for statistical functionals.* Lecture Notes in Statistics. Springer, New York.

Gill, R.D. (1984). Understanding Cox's regression model: a martingale approach. *J. Amer. Statist. Assoc.* **79**, 441–447.

Gill, R.D. (1989). Non- and semi-parametric maximum likelihood estimation and the von Mises method (part I, with discussion). *Scand. J. Statist.* **16**, 97–128.

Hall, P. (1988). Theoretical discussion of bootstrap confidence intervals (with discussion). *Ann. Statist.* **16**, 927–985.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley, Singapore.

Hastie, T.J. and Tibshirani, R. (1986). Generalized additive models (with discussion). *Statist. Science* **1**, 297–318.

Helland, I. (1982). Central limit theorems for martingales with discrete or continuous time. *Scand. J. Statist.* **9**, 79–94.

Hjort, N.L. (1985a). Discussion contribution to Andersen and Borgan's review article. *Scand. J. Statist.* **12**, 141–150.

Hjort, N.L. (1985b). Bootstrapping Cox's regression model. Technical Report NSF–241, Department of Statistics, Stanford University.

Hjort, N.L. (1986a). Bayes estimators and asymptotic efficiency in parametric counting process models. *Scand. J. Statist.* **13**, 63–85.

Hjort, N.L. (1986b). Discussion contribution to Diaconis and Freedman's "On the consistency of Bayes estimates". *Ann. Statist.* **14**, 49–55.

Hjort, N.L. (1988). Discussion contribution to Hinkley's lectures on bootstrapping techniques. To appear in *Scand. J. Statist.*

Hjort, N.L. (1990a). Goodness of fit tests in models for life history data based on cumulative hazard rates. *Ann. Statist.* **18**, xxx-yyy.

Hjort, N.L. (1990b). Estimation in moderately misspecified models. Technical report, Department of Mathematics, University of Oslo.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data.* Wiley, Singapore.

Lehmann, E.L. (1983). *Theory of Point Estimation.* Wiley, Singapore.

Linhart, H. and Zucchini, W. (1986). *Model Selection.* Wiley, Singapore.

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

Reeds, J.A. (1976). On the definition of a von Mises functional. Graduate thesis, University of Harvard.

Reeds, J.A. (1978). Jackknifing maximum likelihood estimates. *Ann. Statist.* **6**, 727–739.

Reid, N. (1979). Influence functions for censored data. *Ann. Statist.* **7**, 78–92.

Reid, N. (1983). Influence functions. In *Encyclopedia of Statistical Science* **4**, eds. Kotz, Johnson, Read, 117–119. Wiley, New York.

Reid, N., Crépeau, H., and Knafl, G. (1985). Influence functions for proportional hazards regression. *Biometrika* **72**, 1–9.

Shao, J. (1989). Functional calculus and asymptotic theory for statistical analysis. *Statist. and Probab. Letters* **8**, 397–405.