ELSEVIER Stochastic Processes and their Applications 54 (1994) 257–274

stochastic processes and their applications

Modifications of the EM algorithm for survival influenced by an unobserved stochastic process

Anatoli I. Yashin[1], Kenneth G. Manton*

Duke University, Center for Demographic Studies, Box 90408, 2117 Campus Drive, Durham, NC 27708-0408, USA

Received March 1993; revised February 1994

Abstract

Let $Y = (Y_t)_{t \geq 0}$ be an unobserved random process which influences the distribution of a random variable T which can be interpreted as the time to failure. When a conditional hazard rate corresponding to T is a quadratic function of covariates, Y, the marginal survival function may be represented by the first two moments of the conditional distribution of Y among survivors. Such a representation may not have an explicit parametric form. This makes it difficult to use standard maximum likelihood procedures to estimate parameters – especially for censored survival data. In this paper a generalization of the EM algorithm for survival problems with unobserved, stochastically changing covariates is suggested. It is shown that, for a general model of the stochastic failure model, the smoothing estimates of the first two moments of Y are of a specific form which facilitates the EM type calculations. Properties of the algorithm are discussed.

Keywords: Randomly changing covariates; Missing information principle; Survival analysis; Unobserved stochastic frailty; Random hazard; EM algorithm; Incomplete information; Smoothing estimates

1. Introduction

The EM algorithm (Dempster et al., 1977) is based on the "missing data principle" (Orchard and Woodbury, 1971). It is useful when the marginal distribution is not easily written. In that case it may be possible to maximize a likelihood by sequentially calculating first, the conditional expectation of the log likelihood for complete data (E-step) and then improving the model parameters (M-step).

Kiefer and Wolfowitz (1956) discussed consistency properties necessary for models with nuisance parameters. Laird (1978) showed, for applications without covariates, how the EM algorithm might be used to generate nonparametric maximum likelihood estimators (NPMLE) for discrete mixtures. Lindsay (1983) further generalized

* Corresponding author. Supported by NIA Grant No. AG01159.
[1] Supported by NIH/NIA grant PO1 AG08791-01.

SSDI 0304-4149(94)00012-I

the NPMLE. Wu (1983) showed a modified EM algorithm is guaranteed to converge to a stationary point. Heckman and Singer (1984a, b) showed that the NPMLE could be used to analyze failure times with fixed *covariates* if the hazard factored into independent, covariate and time dependent, functions.

Below, we extend the EM algorithm to model failure time distributions where unobserved influences are not due to a random variable, but a specific type of stochastic process. This uses results (Yashin, 1985) on moment estimation for Gaussian processes where diffusion is generated by a Wiener process. By using ordinary differential equations to describe the change of the first two moments (mean and variance) of the Gaussian process as ancillary information in the likelihood, EM can be used to estimate parameters for survival models influenced by unobserved stochastic processes. Using these ancillary relations in a generalized EM is computationally easier than estimating the marginal likelihood using partial differential equations for the conditional distribution function when no "closed form" expression exists for that distribution requiring numerical integration on each iteration.

Below, we present notation for a stochastic process influencing survival. Then we present the EM algorithm for random variables subject to unobserved "fixed" heterogeneity. For certain distributions, the EM algorithm also maximizes the marginal likelihood. Then we generalize the EM algorithm to estimate parameters for survival functions influenced by unobserved stochastic processes. For this purpose we present a generalization of the EM algorithm for stochastic processes and a proof of its convergence. The proof makes use of the properties of the Radon–Nikodym derivative of probability measures in functional space. The application of the generalized algorithm to survival models with unobserved randomly changing covariates shows that, to estimate parameters, smoothing estimates of unobserved covariates need to be calculated. The equations to calculate smoothing estimates for a general class of multivariate diffusion processes are provided. The properties of the smoothing estimates were examined in Yashin (1991). We discuss types of analyses for which the generalized EM may be useful.

2. Specification of a stochastic survival processes statement of the problem

Let $\lambda \in \Theta \subset \mathbb{R}^k$ be an unknown parameter and, for each $\lambda \in \Theta$, let (Ω, F, P_λ) be a complete probability space on which the nonnegative random variable T and random process $Y = (Y_t)_{t \geq 0}$ are defined. T is time to failure (or death) and Y is an unobserved, random time-varying covariate. T and Y are statistically dependent. The hazard function relating survival, from 0 to t, to the process Y, is assumed quadratic, i.e.,

$$P_\lambda(T > t \mid Y_0^t) = \exp\left\{ - \int_0^t [\lambda_0(u) + Y_u^* Q(\lambda, u) Y_u \, \mathrm{d}u] \right\}, \tag{1}$$

where $*$ represents transposition; $Y_0^t = \{Y_s, 0 \leq s \leq t\}$. For each $u \geq 0$, $\lambda \in \Theta$; $Q(\lambda, u)$ is a nonnegative definite matrix of bounded time dependent coefficients and $\lambda_0(t)$

a hazard rate with λ, an unknown parameter. The quadratic preserves the Gaussian nature of the unobserved process as individuals die. The quadratic has useful mathematical properties both as a hazard function (e.g., quadratic functions for multiple independent failure processes can be summed to yield a quadratic for the total hazard) and for computing parameters in the generalized EM algorithm.

Changes in Y are assumed to satisfy the stochastic differential equation (Yashin, 1985)

$$\mathrm{d}Y_t = [a_0(\lambda, t) + a(\lambda, t)Y_t]\,\mathrm{d}t + b(t)\,\mathrm{d}W_t, \quad Y_0, \tag{2}$$

where Y_0, the initial condition, is a normally distributed random variable; a_0 and a are coefficients (possibly time varying; and dependent on λ) describing deterministic elements of the process; and W_t, $t \geq 0$, is a Wiener process with time dependent scale parameter $b(t)$, e.g., if diffusion increases with age, $b(t)$ increases the size of the average random deviation. Thus, (1) and (2) describe the time dependence of the forces of control and diffusion in the process generating Y.

For estimation, we need to define an observation plan. Let $(\tau, \Delta) = (t_1, \delta_1, t_2, \delta_2, \ldots, t_n, \delta_n)$ be data where t_i, $i = 1, 2, \ldots, n$, are observed independent failure times and δ_i, $i = 1, 2, \ldots, n$, are cases censored before failure, i.e., $\delta_i = 0$ and $\delta_i = 1$ refer to censored and uncensored observations, respectively. The problem is to estimate λ from censored data (τ, Δ). Marginal likelihood procedures, using differential equations to describe the evolution of the process in (2), are discussed in Marchuk et al. (1989) and used in Asachenkov et al. (1988).

2.1. EM algorithm for unobserved random variables

Let $F_\lambda(t, y)$, $g_\lambda(y \mid T = t)$ and $\varphi_\lambda(t)$ be the joint, conditional and marginal density functions for T and Y – in this section a time invariant random variable. The maximum likelihood procedure for estimating λ, given one observation $T = t$, uses the marginal density to define the likelihood, i.e., $\lambda^* = \operatorname{argmax} \varphi_\lambda(t)$. The likelihood for $f_\lambda(t, y)$ may be easier to maximize than $\varphi_\lambda(t)$. In the EM algorithm (Dempster et al., 1977) the conditional expectation of the log likelihood of $f_\lambda(t, y)$ must be maximized in each iteration. In nonrestricted cases this also maximizes the marginal likelihood.

To show parameter estimates improve at each iteration, one can, from the definition of the density function, and Bayes' rule, write

$$\log \varphi_\lambda(t) = \log f_\lambda(t, y) - \log g_\lambda(y \mid T = t). \tag{3}$$

Eq. (3), after averaging, conditional on $T = t$, with respect to measure $P_{\lambda'}$, is

$$\log \varphi_\lambda(t) = E_{\lambda'}[\log f_\lambda(t, Y) \mid T = t] - E_{\lambda'}[\log g_\lambda(Y \mid T = t) \mid T = t]. \tag{4}$$

In the M-step of the EM algorithm, $\lambda^* = \operatorname{argmax} E_{\lambda'}[\log f_\lambda(t, Y) \mid T = t]$ is calculated. To guarantee convergence it is often sufficient to find a λ^* which satisfies the weaker condition, $E_{\lambda'}[\log f_{\lambda^*}(t, Y) \mid T = t] \geq E_{\lambda'}[\log f_{\lambda'}(t, Y) \mid T = t]$). An algorithm with this condition implemented in the M-step can be called a generalized EM algorithm (GEM). For both EM and GEM, the condition

$$\log \varphi_{\lambda^*}(t) - \log \varphi_{\lambda'}(t) \geq -E\left(\log \frac{g_{\lambda^*}(Y \mid T=t)}{g_{\lambda'}(Y \mid T=t)} \mid T=t\right) \geq 0 \tag{5}$$

holds from Jensen's inequality. Thus, in EM and generalized EM (GEM), $\varphi_{\lambda^*}(t) \geq \varphi_{\lambda'}(t)$ so parameter estimates are guaranteed to improve at each iteration (with $\lambda' = \hat{\lambda}_p$ and $\lambda^* = \hat{\lambda}_{p+1}$, for the pth iteration). This is illustrated for distributions of Y often used in survival analysis.

2.1.1. Gamma distribution

If Y is a gamma distributed random variable,

$$g(y) = \frac{\eta^k y^{k-1} e^{-\eta y}}{\Gamma(k)}, \tag{6}$$

then $f_\lambda(t \mid y)$ depends on λ as

$$f_\lambda(t \mid Y=y) = y\lambda\mu(t)e^{-y\lambda H(t)}, \tag{7}$$

where the integrated hazard is $H(t) = \int_0^t \mu(u)\,du$. The marginal distribution, $\varphi_\lambda(t)$, is

$$\varphi_\lambda(t) = \frac{\eta^k k\lambda\mu(t)}{(\eta + \lambda H(t))^{k+1}}. \tag{8}$$

If $T = t$ is observed, the log likelihood is

$$\log \varphi_\lambda(t) \sim \log\lambda - (k+1)\log(\eta + \lambda H(t)) + C, \tag{9}$$

where C does not depend on λ. Maximizing (9) produces

$$\lambda^* = \frac{\eta}{kH(t)}. \tag{10}$$

To apply the EM algorithm to the joint distribution, $f_\lambda(t, y)$, write

$$\log f_\lambda(t, y) = \log g(y) + \log[\lambda\mu(t)y] - \lambda y H(t). \tag{11}$$

In the E-step, the expectation of both parts of (11) conditional on $T = t$ are taken with respect to $P_{\lambda'}$:

$$\begin{aligned} &E_{\lambda'}(\log f_\lambda(t, Y) \mid T=t) \\ &\quad = E_{\lambda'}(\log g(Y) \mid T=t) + \log\lambda + E_{\lambda'}[\log(Y\mu(t)) \mid T=t] \\ &\qquad - \lambda H(t) E_{\lambda'}(Y \mid T=t). \end{aligned} \tag{12}$$

Observe that $E_{\lambda'}(Y \mid T=t)$ is equal to

$$\frac{k+1}{\eta + \lambda' H(t)}. \tag{13}$$

The M-step requires maximizing (12) with respect to λ, setting $\lambda' = \hat{\lambda}_p$ and using (13), to produce

$$\frac{1}{\hat{\lambda}_{p+1}} = \frac{H(t)(k+1)}{\eta + \hat{\lambda}_p H(t)}, \tag{14}$$

and, by inversion,

$$\hat{\lambda}_{p+1} = \frac{\eta + \hat{\lambda}_p H(t)}{(k+1)H(t)}. \tag{15}$$

When $p \uparrow \infty$, $\hat{\lambda}_p \to \lambda^*$, where

$$\lambda^* = \frac{\eta}{kH(t)}. \tag{16}$$

Thus, for gamma distributed Y's, (16) is the same as (10) for the marginal likelihood. With r censored observations, t_i', $i = 1, 2, \ldots, r$, and $n - r$ uncensored observations t_j, $j = 1, 2, \ldots, n - r$, the estimate of $\hat{\lambda}_{p+1}$ in (15) is

$$\begin{aligned} \hat{\lambda}_{p+1} &= \frac{n-r}{\sum_{i=1}^{r} H(t_i') E_{\hat{\lambda}_p}(Y \mid T > t_i') + \sum_{j=1}^{n-r} H(t_j) E_{\hat{\lambda}_p}(Y \mid T = t_j)} \\ &= \frac{n-r}{\sum_{i=1}^{r} [kH(t_i')/(\eta + \hat{\lambda}_p H(t_i'))] + \sum_{j=1}^{n-r} [(k+1)H(t_j)/(\eta + \hat{\lambda}_p H(t_j))]}. \end{aligned} \tag{17}$$

This is an alternative way to calculate maximum likelihood estimators.

2.1.2. Normal distribution

If Y is a Gaussian random variable with mean m and variance γ, then

$$g(y) = \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{(y-m)^2}{2\gamma} \right), \tag{18}$$

with the density for failure time t conditional on y,

$$f(t \mid y) = \lambda \mu(t) y^2 e^{-\lambda H(t) y^2}, \tag{19}$$

where $H(t)$ is the integrated hazard with dependence on y being quadratic to preserve the normal distribution as individuals systematically die off.

If λ is the only unknown parameter, then the log likelihood with k (instead of r) censored and $n - k$ uncensored observations of t is

$$L(t, y) = (n-k)\log\lambda - \lambda y^2 \sum_{i=1}^{k} H(t_i') - \lambda y^2 \sum_{j=1}^{n-k} H(t_j) + c, \tag{20}$$

where c does not depend on λ. Implementation of the E- and M-steps yields

$$\hat{\lambda}_{p+1} = \frac{n-k}{\sum_{i=1}^{k} H(t_i') E_{\hat{\lambda}_p}(Y^2 \mid T > t_i') + \sum_{j=1}^{n-k} H(t_j) E_{\hat{\lambda}_p}(Y^2 \mid T = t_j)}. \tag{21}$$

Note that $E_\lambda(Y^2 \mid T > t_i')$ is the sum of the mean, $m(t)$, and variance, $\gamma(t)$, of Y at t,

$$E_\lambda(Y^2 \mid T > t_i') = m^2(t_i') + \gamma(t_i'), \tag{22}$$

where the mean at t_i' is (Yashin, 1985)

$$m(t_i') = \frac{m}{1 + 2\lambda H(t_i')\gamma}, \tag{23}$$

and the variance is

$$\gamma(t_i') = \frac{\gamma}{1 + 2\lambda H(t_i')\gamma}. \tag{24}$$

To estimate $E_\lambda(Y^2 \mid T = t_j)$ write the density for Y at t_j, using Bayes' rule,

$$g(y \mid T = t_j) = \frac{\partial}{\partial y} P(Y \le y \mid T = t_j), \tag{25}$$

which is equivalent to

$$g(y \mid T = t_j) = \frac{g(y) f(t_j \mid Y = y)}{f(t_j)} = C_0(t_j) y^2 \exp\left(- \frac{(y - m)^2}{2\gamma} \right) \exp(- \gamma H(t_j) y^2), \tag{26}$$

where $C_0(t_j)$ does not depend on y. After simple transformations $g(y \mid T = t_j)$ is

$$g(y \mid T = t_j) = C(t_j) \frac{1}{\sqrt{2\pi G(t_j)}} y^2 \exp\left(- \frac{[y - M(t_j)]^2}{2G(t_j)} \right), \tag{27}$$

with mean

$$M(t_j) = \frac{m}{1 + 2\lambda H(t_j)\gamma}, \tag{28}$$

and variance

$$G(t_j) = \frac{\gamma}{1 + 2\lambda H(t_j)\gamma}. \tag{29}$$

The density normalization constant is

$$C(t_j) = \frac{1}{M^2(t_j) + G(t_j)}. \tag{30}$$

Substituting these expressions in (22) we have

$$E_\lambda(Y^2 \mid T = t_j) = \frac{3G^2(t_j) + 6M^2(t_j)G(t_j) + M^4(t_j)}{G(t_j) + M^2(t_j)}, \tag{31}$$

with $M(t_j)$ and $G(t_j)$ defined in (28) and (29).

3. Generalizing the EM algorithm for stochastic processes: theory

When the probability measure $\{P_\lambda\}$, $\lambda \in \Theta$, corresponds to a random process one must use likelihood ratios instead of likelihood functions. This requires making assumptions about the absolute continuity of the family of probability measures for the parameter set. Assume that for any $\lambda, \lambda' \in \Theta$, measures P_λ and $P_{\lambda'}$ are equivalent on the probability space (Ω, F). Let $\Omega = \Omega_x \times \Omega_y$ and $F = F^x \otimes F^y$, where (Ω_x, F^x) is the probability space where observations (τ, Δ) are defined, and (Ω_y, F^y) is the probability space where an unobserved process, Y, is defined. Denote by $\bar{P}_\lambda$ the restriction of P_λ to F^x. Measures $\bar{P}_\lambda$ and $\bar{P}_{\lambda'}$, are equivalent for all $\lambda, \lambda' \in \Theta$, since P_λ and $P_{\lambda'}$ are equivalent. For $\alpha \in \Theta$, define the maximum likelihood estimate, $\hat{\lambda}$, as

$$\hat{\lambda} = \operatorname{argmax} \frac{\mathrm{d}\bar{P}_\lambda}{\mathrm{d}\bar{P}_\alpha}, \tag{32}$$

where

$$\frac{\mathrm{d}\bar{P}_\lambda}{\mathrm{d}\bar{P}_\alpha} = E_\alpha\left(\frac{\mathrm{d}P_\lambda}{\mathrm{d}P_\alpha}\,\middle|\, F^x\right). \tag{33}$$

Let us define

$$L(\lambda, \alpha) = \log \frac{\mathrm{d}\bar{P}_\lambda}{\mathrm{d}\bar{P}_\alpha} \tag{34}$$

and

$$\mathscr{L}(\lambda, \lambda') = \log \frac{\mathrm{d}P_\lambda}{\mathrm{d}P_{\lambda'}}; \tag{35}$$

then

$$L(\lambda, \lambda') = L(\lambda, \alpha) - L(\lambda', \alpha) = \log E_{\lambda'}\left(\frac{\mathrm{d}P_\lambda}{\mathrm{d}P_{\lambda'}}\,\middle|\, F^x\right). \tag{36}$$

From Jensen's inequality,

$$L(\lambda, \lambda') \geq E_{\lambda'}\left(\log \frac{\mathrm{d}P_\lambda}{\mathrm{d}P_{\lambda'}}\,\middle|\, F^x\right) = E_{\lambda'}(\mathscr{L}(\lambda, \lambda') \,|\, F^x), \tag{37}$$

for which we will use the notation $\Lambda(\lambda, \lambda')$ for the expectation of the likelihood. Assume that for $\lambda \neq \lambda'$, $\Lambda(\lambda, \lambda') > 0$. Hence, $L(\lambda, \lambda') \neq 0$ and $L(\lambda, \lambda') = 0$ if and only if $\lambda = \lambda'$. Let $\hat{P}$ be an arbitrary probability measure on (Ω_x, F^x) which is equivalent to $\bar{P}_\lambda$ for any $\lambda \in \Theta$. Define a new measure $\hat{P}_\lambda$ on (Ω, F) by

$$\hat{P}_\lambda(\mathrm{d}\omega_x, \mathrm{d}\omega_y) = P_\lambda(\mathrm{d}\omega_y \,|\, F^x)\hat{P}(\mathrm{d}\omega_x). \tag{38}$$

The measures $\hat{P}_\lambda$ have the following property.

Theorem 1. *For each $\lambda, \lambda' \in \Theta$, measures $\hat{P}_\lambda$ and $\hat{P}_{\lambda'}$ are equivalent. Their likelihood ratio satisfies*

$$\frac{d\hat{P}_\lambda}{d\hat{P}_{\lambda'}} = \frac{dP_\lambda}{dP_{\lambda'}}\left(\frac{d\bar{P}_\lambda}{d\bar{P}_{\lambda'}}\right)^{-1}. \tag{39}$$

The proof is in the appendix.

The restriction $\hat{P}_\lambda$ on (Ω_x, F^x) is $\hat{P}$ and does not depend on λ; hence,

$$\hat{E}\left(\frac{d\hat{P}_\lambda}{d\hat{P}_{\lambda'}}\bigg| F^x\right) = 1 \tag{40}$$

for any $\lambda, \lambda' \in \Theta$. A corollary of Theorem 1 is

$$\log\frac{d\bar{P}_\lambda}{d\bar{P}_{\lambda'}} = \log\frac{dP_\lambda}{dP_{\lambda'}} - \log\frac{d\hat{P}_\lambda}{d\hat{P}_{\lambda'}}, \tag{41}$$

so that, after conditional averaging,

$$L(\lambda, \lambda') = \Lambda(\lambda, \lambda') - E_{\lambda'}\left(\log\frac{d\hat{P}_\lambda}{d\hat{P}_{\lambda'}}\bigg| F^x\right). \tag{42}$$

Since $\Lambda(\lambda', \lambda') = 0$, $\max_\lambda \Lambda(\lambda, \lambda') \geq 0$. From Jensen's inequality for each $\lambda^* \neq \lambda'$,

$$-E_{\lambda'}\left(\log\frac{d\hat{P}_{\lambda^*}}{d\hat{P}_{\lambda'}}\bigg| F^x\right) \geq -\log E_{\lambda'}\left(\frac{d\hat{P}_{\lambda^*}}{d\hat{P}_{\lambda'}}\bigg| F^x\right) = 0. \tag{43}$$

Hence, if $\lambda^* = \operatorname{argmax} \Lambda(\lambda, \lambda')$ then from (42), $L(\lambda^*, \lambda') \geq 0$ or $L(\lambda^*, \alpha) \geq L(\lambda', \alpha)$, i.e., λ^* improves the value of the likelihood ratio. The EM algorithm modified to estimate the parameters of stochastic processes can now be described.

E-step: Calculate $\Lambda(\lambda, \hat{\lambda}_p)$ on the pth iteration (when $\hat{\lambda}_p$ is known).

M-step: Calculate $\hat{\lambda}_{p+1}$ from the equality

$$\hat{\lambda}_{p+1} = \operatorname{argmax} \Lambda(\lambda, \lambda'_p), \tag{44}$$

or the inequality

$$\Lambda(\hat{\lambda}_{p+1}, \hat{\lambda}_p) \geq 0.$$

According to (37), $L(\hat{\lambda}_{p+1}, \alpha) \geq L(\hat{\lambda}_p, \alpha)$, i.e., the likelihood never decreases. To show the algorithm converges (see Dembo and Zeitouni, 1986), let us note the following result.

Theorem 2. *Assume*:

(a) $\Theta_{\hat{\lambda}_0} - \{\lambda \in \theta \colon L(\lambda, \alpha) \geq L(\hat{\lambda}_0, \alpha)\}$ *is compact for any* $\hat{\lambda}_0 \in \Theta$;

(b) *for any* $\alpha \notin \Theta$, $L(\lambda, \alpha)$ *is continuous in* Θ *with respect to* λ *and differentiable in the interior of* Θ;

(c) $\Lambda(\lambda, \lambda')$ *is continuous with respect to* λ *and* λ';

(d) *all* $\{\hat{\lambda}_p\}$ *are in the interior of* Θ.

Then all limit points of any $\{\hat{\lambda}_p\}$ are stationary points of $L(\cdot, \alpha)$, having the same value L^*. Furthermore $\{L(\hat{\lambda}_p, \alpha)\}$ converges monotonically to L^*.

4. EM algorithm for survival with stochastic covariates

In this section we consider applying the results of Section 3 to survival models with unobserved stochastically changing covariates. We show, for the model represented by Eqs. (1) and (2) (i.e., a diffusion type process with quadratic mortality), that the implementation of the EM algorithm requires calculating "smoothing" estimates of the first two moments of the process – relations for which we can specify an explicit form greatly simplifying computations. The use of smoothing estimates to evaluate the first two moments of the conditional density of Y is a less complex computation than evaluating a conditional distribution of Y of unspecified form. The calculations differ for censored, and uncensored, data.

4.1. General case

Consider the Radon–Nikodym derivative of P_λ, corresponding to independent observations $(Y_0^{t_i}, t_i, \delta_i)$, $i = 1, 2, \ldots, n$, given parameter λ with respect to measure $P_{\lambda'}$, which corresponds to the same observations given λ'. Here $Y_0^{t_i} = \{Y_s, 0 \le s \le t_i\}$, Y_s is given by (2), and t_i, δ_i, $i = 1, 2, \ldots, n$, are n independent observations of failure time T (censored if $\delta_i = 0$ and uncensored if $\delta_i = 1$). T and Y are related by Eq. (1). It can be found directly (or from Kabanov et al. (1978)) that the logarithm of the Radon–Nikodym derivative for these observations is

$$\log \frac{\mathrm{d}P_\lambda}{\mathrm{d}P_{\lambda'}}(Y_0^\tau, \tau, \Delta) = \sum_{i=1}^{n} L_i(\lambda, \lambda'), \tag{45}$$

where $Y_0^\tau = (Y_0^{t_i}, i = 1, 2, \ldots, n)$; $\tau = (t_1, t_2, \ldots, t_n)$; $\Delta = (\delta_1, \delta_2, \ldots, \delta_n)$, and for $L_i(\lambda, \lambda')$, $i = 1, 2, \ldots, n$, we have

$$\begin{aligned} L_i(\lambda, \lambda') = & \int_0^{t_i} Y_u^* [a^*(\lambda', u) - a^*(\lambda, u)] [b(u)b^*(u)]^{-1} \, \mathrm{d}Y_u \\ & - \tfrac{1}{2} \int_0^{t_i} Y_u^* [a^*(\lambda', u) - a^*(\lambda, u)] [b(u)b^*(u)]^{-1} [a(\lambda' u) + a(\lambda, u)] \, Y_u \, \mathrm{d}u \\ & + \delta_i \log \frac{\lambda_0(t_i) + Y_{t_i}^* Q(\lambda, t_i) Y_{t_i}}{\lambda_0(t_i) + Y_{t_i}^* Q(\lambda', t_i) Y_{t_i}} \\ & - \int_0^{t_i} [Y_u^* Q(\lambda, u) Y_u - Y_u^* Q(\lambda', u) Y_u] \, \mathrm{d}u. \end{aligned} \tag{46}$$

For $\Lambda_i(\lambda, \lambda') = E(L_i(\lambda, \lambda') \mid F^x)$ we have

$$\Lambda_i(\lambda, \lambda') = E_{\lambda'}\left(\int_0^{t_i} Y_u^* [a^*(\lambda', u) - a^*(\lambda, u)] [b(u)b^*(u)]^{-1} \, \mathrm{d}Y_u \mid T > t_i \right)$$

$$-\tfrac{1}{2}E_{\lambda'}\left(\int_0^{t_i} Y_u^*[a^*(\lambda',u)-a^*(\lambda,u)][b(u)b^*(u)]^{-1}[a(\lambda',u)+a(\lambda,u)]\,Y_u\,\mathrm{d}u \mid T>t_i\right)$$

$$-\int_0^{t_i}\operatorname{tr}\{[Q(\lambda,u)-Q(\lambda',u)]\gamma(u,t_i)\}\,\mathrm{d}u$$

$$-\int_0^{t_i}[m^*(u,t_i)Q(\lambda,u)m(u,t_i)-m^*(u,t_i)Q(\lambda',u)m(u,t_i)]\,\mathrm{d}u, \tag{47}$$

with $\gamma(u,t)=E_{\lambda'}\{[Y_u-m(u,t_i)][Y_u-m(u,t_i)]^*\mid T>t_i\}$ and $m(u,t)=E_{\lambda'}(Y_u\mid T>t_i)$ if t_i is censored. For the uncensored case,

$$A_i(\lambda,\lambda')=E_{\lambda'}\left(\int_0^{t_i} Y_u^*[a^*(\lambda',u)-a^*(\lambda,u)][b(u)b^*(u)]^{-1}\,\mathrm{d}Y_u \mid T=t_i\right)$$

$$-\tfrac{1}{2}E_{\lambda'}\left(\int_0^{t} Y_u^*[a^*(\lambda',u)-a^*(\lambda,u)][b(u)b^*(u)]^{-1}[a(\lambda',u)+a(\lambda,u)]\,Y_u\,\mathrm{d}u \mid T=t_i\right)$$

$$+E_{\lambda'}\left(\log\frac{\lambda_0(t_i)+Y_{t_i}^*Q(\lambda,t_i)Y_{t_i}}{\lambda_0(t_i)+Y_{t_i}^*Q(\lambda',t_i)Y_{t_i}}\,\middle|\, T=t_i\right)$$

$$-\int_0^{t_i}\operatorname{tr}\{[Q(\lambda,u)-Q(\lambda',u)]\bar{\gamma}(u,t_i)\}\,\mathrm{d}u$$

$$-\int_0^{t_i}[\bar{m}^*(u,t_i)Q(\lambda,u)\bar{m}(u,t_i)-\bar{m}^*(u,t_i)Q(\lambda',u)\bar{m}(u,t_i)]\,\mathrm{d}u, \tag{48}$$

where

$$\bar{m}(u,t_i)=E_{\lambda'}(Y_u\mid T=t_i),\quad \bar{\gamma}(u,t_i)=E_{\lambda'}\{[Y_u-\bar{m}(u,t_i)][Y_u-\bar{m}(u,t_i)]^*\mid T=t_i\}. \tag{49}$$

Since (47) and (48) are cumbersome, we restrict the example to one dimension.

4.2. A one-dimensional unobserved influential process

Let Y be a one-dimensional random process, as in (1) and (2), where $a_0(t)=0$, $\lambda_0(t)=0$, $a(\lambda,t)=a(t)$ and $Q(\lambda,u)=\lambda$, where λ is a scalar. If k failure times are censored, and $n-k$ failure times are uncensored,

$$A(\lambda,\lambda')=(n-k)\log\frac{\lambda}{\lambda'}-(\lambda-\lambda')\sum_{i=0}^{n-k}\int_0^{t_j}E_{\lambda'}(Y_u^2\mid T=t_j)\,\mathrm{d}u$$

$$-(\lambda-\lambda')\sum_{j=0}^{k}\int_0^{t_i'}[m^2(u,t_i')+\gamma(u,t_i')]\,\mathrm{d}u, \tag{50}$$

where $\gamma(u,t)=E_{\lambda'}\{[Y_u-m(u,t_i)]^2\mid T>t_i')$ and $m(u,t')=E_{\lambda'}(Y_u\mid T>t_i')$ are the "smoothing" equations giving the variance and mean of the one-dimensional stochastic process. These are easier to compute than the last integral terms in (47) and (48).

If $\lambda' = \hat{\lambda}_p$, maximizing $\Lambda(\lambda, \hat{\lambda}_p)$ with respect to λ produces

$$\hat{\lambda}_{p+1} = \frac{n-k}{\sum_{i=1}^{k}\int_0^{t_i'}[\gamma(u, t_i') + m^2(u, t_i')]\,\mathrm{d}u + \sum_{j=1}^{n-k}\int_0^{t_j} E_{\lambda'}(Y_u^2 \mid T = t_j)\,\mathrm{d}u}. \tag{51}$$

The smoothing equations for $m(u, t_i')$ and $\gamma(u, t_i')$ are (Yashin, 1991)

$$m(s, t_i') = m(s) - 2\lambda' \int_s^{t_i'} \gamma_{21}(s, u) m(u)\,\mathrm{d}u \tag{52}$$

and

$$\gamma(s, t_i') = \gamma(s) - 2\lambda' \int_s^{t_i'} \gamma_{21}(s, u)\gamma(u)\,\mathrm{d}u, \tag{53}$$

where $\gamma_{21}(s, t_i')$ satisfies

$$\gamma_{21}(s, t_1') = \gamma(s) + \int_s^{t_i'} \gamma_{21}(u, t) a(u)\,\mathrm{d}u - 2\lambda' \int_s^{t_i'} \gamma_{21}(u, t)\gamma(u)\,\mathrm{d}u, \tag{54}$$

and $m(s)$, $\gamma(s)$ satisfy the "filter" equations (Yashin, 1985)

$$m(s) = m + \int_0^s [a(u)m(u) - 2\lambda' m(u)\gamma(u)]\,\mathrm{d}u \tag{55}$$

and

$$\gamma(s) = \gamma + \int_0^s [b^2(u) - 2\lambda'\gamma^2(u)]\,\mathrm{d}u, \tag{56}$$

with $\lambda' = \hat{\lambda}_p$. The estimate $E_{\lambda'}(Y_s^2 \mid T = t_i)$, $s \le t_i$, is (see the appendix)

$$E_{\lambda'}(Y_s^2 \mid T = t_i) = \frac{B^2(s, t_i)[3G^2(s, t_i) + 6M^2(s, t_i)G(s, t_i) + M^4(s, t_i)]}{B^2(s, t_i)[G(s, t_i) + M^2(s, t_i)] + \tilde{\gamma}(s, t_i)} + \frac{\tilde{\gamma}(s, t_i)[M^2(s, t_i) + G(s, t_i)]}{B^2(s, t_i)[G(s, t_i) + M^2(s, t_i)] + \tilde{\gamma}(s, t_i)}, \tag{57}$$

where

$$M(s, t_i) = \frac{m(s)}{1 + 2D(s, t_i)\gamma(s)}, \tag{58}$$

$$G(s, t_i) = \frac{\gamma(s)}{1 + 2D(s, t_i)\gamma(s)}, \tag{59}$$

$$D(s, t_i) = \lambda' \int_s^{t_i} \mu(u) B^2(s, u)\,\mathrm{d}u, \tag{60}$$

$$B(s, t_i) = \exp\left(\int_s^{t_i} [a(u) - 2\lambda\tilde{\gamma}(s, u)]\,\mathrm{d}u\right). \tag{61}$$

$\breve{\gamma}(s, u)$ satisfies

$$\frac{d\breve{\gamma}(s, u)}{du} = 2a(u)\breve{\gamma}(s, u) + b^2(u) - 2\lambda'\breve{\gamma}^2(s, u), \tag{62}$$

with initial condition $\breve{\gamma}(s, s) = 0$, and $m(s)$ and $\gamma(s)$ are (55) and (56).

If $a = b = 0$ (i.e., Y is a random variable with no dynamics, $a = 0$, or $\breve{\gamma}$ diffusion, $b = 0$), then (57) reduces to (31). Observe that when $a = b = 0$, $\breve{\gamma}(s, u) = 0$, then $B(s, t_i) = 1$, $D(s, t_i) = \lambda'(H(t_i) - H(s))$, so

$$m(s) = \frac{m}{1 + 2\lambda H(s)\gamma} \tag{63}$$

and

$$\gamma(s) = \frac{\gamma}{1 + 2\lambda H(s)\gamma}. \tag{64}$$

Substituting $m(s)$, $\gamma(s)$ and $D(s, t_i)$ in (58) and (59) produces

$$M(s, t_i) = \frac{m}{1 + 2\lambda' H(t_i)\gamma} \tag{65}$$

and

$$G(s, t_i) = \frac{\gamma}{1 + 2\lambda' H(t_i)\gamma}, \tag{66}$$

and (57) simplifies to (31). To estimate parameters:

(1) Set $\lambda' = \hat{\lambda}_0$.

(2) Calculate $m(s)$ and $\gamma(s)$ using (55) and (56).

(3) Using (52) and (53) calculate

$$\sum_{i=1}^{k} \int_0^{t_i} [\gamma(u, t_i) + m^2(u, t_i)]\, du, \tag{67}$$

where $t_1, t_2, \ldots, t_k$ are censored failure times.

(4) Calculate $B(s, t_i)$ and $\gamma(s, t_i)$ from (61) and (62).

(5) Calculate $M(s, t_i)$ and $G(s, t_i)$ and $D(s, t_i)$ from (58)–(60).

(6) Calculate $E_{\lambda'}(Y_s^2 \mid T = t_i)$ from (57).

(7) Calculate $\hat{\lambda}_1$ ($\hat{\lambda}_{p+1}$ on the pth cycle) from (51).

(8) Take $\lambda' = \hat{\lambda}_1$ ($\hat{\lambda}_{p+1}$ on the pth cycle) and go to step 2 until convergence to the desired precision is achieved.

Remark. In some applications it may be that the effects of unobserved processes on prior states of the system are important. In this case we can generate the "backward" (in time) smoothing equations (Yashin, 1991) which can be used in place of the "forward" smoothing equations described above. In the "backward" smoothing equation, $m(s, t_i)$, $\gamma(s, t_i)$ and $\gamma_{21}(s, t_i)$ may be written as

$$m(s, t_i) = m(t_i) - \int_s^{t_i} \{a(u)m(u, t_i) + b^2(u)\gamma^{-1}(u)[m(u, t_i) - m(u)]\}\,du, \tag{68}$$

$$\gamma(s, t_i) = \gamma(t) - 2\int_s^{t_i} [a(u)\gamma(u, t_i) - b^2(u)]\,du$$
$$- 2\int_s^{t_i} b^2(u)\gamma^{-1}(u)\gamma(u, t_i)\,du, \tag{69}$$

and

$$\gamma_{21}(s, t_i) = \gamma(t_i) - \int_s^{t_i} a(u)\gamma_{21}(u, t_i)\,du$$
$$- \int_s^{t_i} b^2(u)\gamma^{-1}(u)\gamma_{21}(u, t_i)\,du. \tag{70}$$

Thus, the procedure can be used both to produce estimates conditional on Y and looking at survival in the future; or, conditional on Y, looking at how covariates among survivors evolved to the current time.

5. Discussion

In the case of censored survival data nonparametric methods may be used to estimate a survival function, e.g., Kaplan–Meier (1958); or cumulative hazard (Nelson, 1972; Aalen, 1976) estimators. The hazard rate may be estimated using procedures in Ramlau–Hansen (1983). The asymptotic properties of these procedures make them applicable to a number of problems. The nonparametric approach, however, cannot be used when ancillary data play a significant part in specifying the hazard rate. For example, prior studies may provide important information on age related changes in some variables for individuals. Clinical trials may help identify the appropriate shape of the risk function. To take advantage of this knowledge one needs a different approach. When it is known that a covariate is randomly changing, a stochastic process model can be used to calculate the parameters of the marginal distribution. This distribution, however, may not have an explicit parametric representation. For example, to specify the survival function in a stochastic process model it is often necessary to solve a system of nonlinear differential equations. This means that standard maximum likelihood requires a constrained optimization procedure where constraints are represented by differential equations (Marchuk et al., 1989). Instead of directly constrained maximization, one can use an EM type algorithm. The advantage of EM is that the M-step can be easier to perform. In the examples given, the M-step for the specified process involves only analytically specifiable forms. The more difficult task is performing the E-step. For the process presented this requires numerical integration of nonlinear differential equations for both the filter and smoothing estimates of the first two moments of the unobserved covariates.

6. Conclusion

In this paper we suggest a modification of the EM algorithm to work with unobserved stochastically changing covariates. The likelihood ratio (instead of likelihood function) is used for performing the E- and M-steps of this algorithm. To justify this procedure we prove that the likelihood ratio improves for each iteration. Performing the E-step requires calculating smoothing estimates. In the case of a quadratic hazard, and a Gaussian process for the unobserved covariate, the equations for these estimates may be written in an analytic form. Applying the EM algorithm to censored survival data with unobserved time dependent covariates requires calculating smoothing equations for $E_{\lambda'}(Y_s^2 \mid T > t_i')$ for k censored observations, and $E_{\lambda'}(Y_s^2 \mid T = t_j)$ for $n-k$ uncensored observations. Dembo and Zeitoni (1986) have used the EM algorithm for diffusion type stochastic equations as observed and unobserved processes. Campilo and Le Gland (1989) compared the properties of the EM algorithm with direct likelihood. Their results suggest that the EM algorithm in such applications is analytically simpler than the constrained optimization algorithm but may require more computational effort.

Appendix

Proof of Theorem 1. Let η and ξ be arbitrary F^x-measurable and F-measurable bounded random variables. Consider two forms of $E_\lambda(\eta\zeta)$:

$$E_\lambda(\eta\zeta) = E_{\lambda'}\left(\eta\zeta\frac{\mathrm{d}P_\lambda}{\mathrm{d}P_{\lambda'}}\right) = E_{\lambda'}\left[\eta\, E_{\lambda'}\left(\zeta\frac{\mathrm{d}P_\lambda}{\mathrm{d}P_{\lambda'}}\middle| F^x\right)\right] \tag{A.1}$$

and

$$E_\lambda(\eta\zeta) = E_\lambda[\eta\,[E_\lambda(\zeta \mid F^x)]] = E_{\lambda'}\left[\eta\frac{\mathrm{d}\bar{P}_\lambda}{\mathrm{d}\bar{P}_{\lambda'}}E_\lambda(\zeta \mid F^x)\right]. \tag{A.2}$$

Note that $P_\lambda(\mathrm{d}\bar{P}_\lambda/\mathrm{d}\bar{P}_{\lambda'} > 0) = 1$. To prove this let $A = (\mathrm{d}\bar{P}_\lambda/\mathrm{d}\bar{P}_{\lambda'} = 0)$. Then

$$P_\lambda(A) = \int_A \frac{\mathrm{d}\bar{P}_\lambda}{\mathrm{d}\bar{P}_{\lambda'}}\mathrm{d}P_\lambda = 0$$

by definition of A. Hence $P_\lambda(\bar{A}) = P_\lambda(\mathrm{d}\bar{P}_\lambda/\mathrm{d}\bar{P}_{\lambda'} > 0) = 1$.

Comparing (A.1) and (A.2) we get

$$E_\lambda(\zeta \mid F^x) = \left(\frac{\mathrm{d}\bar{P}_\lambda}{\mathrm{d}\bar{P}_{\lambda'}}\right)^{-1} E_{\lambda'}\left(\zeta\frac{\mathrm{d}P_\lambda}{\mathrm{d}P_{\lambda'}}\middle| F^x\right) \tag{A.3}$$

P_λ-a.s. Note further that $\hat{E}_\lambda(\eta\zeta)$ may be represented as

$$\hat{E}_\lambda(\eta\zeta) = \hat{E}_\lambda[\eta\,\hat{E}_\lambda(\zeta \mid F^x)] = \hat{E}[\eta\, E_\lambda(\zeta \mid F^x)], \tag{A.4}$$

where $\hat{E}_\lambda$ and $\hat{E}$ denote expectation with respect to $\hat{P}_\lambda$ and the restriction of $\hat{P}_\lambda$ on F^x which is $\hat{P}$. Substituting (A.3) in (A.4) produces

$$\hat{E}_\lambda(\eta\zeta) = \hat{E}\left[\eta\left(\frac{d\bar{P}_\lambda}{d\bar{P}_{\lambda'}}\right)^{-1} E_{\lambda'}\left(\zeta\frac{dP_\lambda}{dP_{\lambda'}}\bigg| F^x\right)\right] = \hat{E}_{\lambda'}\left[\eta\zeta\left(\frac{d\bar{P}_\lambda}{d\bar{P}_{\lambda'}}\right)^{-1}\frac{dP_\lambda}{dP_{\lambda'}}\right] \tag{A.5}$$

and

$$\hat{E}_\lambda(\eta\zeta) = \hat{E}_{\lambda'}\left(\eta\zeta\frac{d\hat{P}_\lambda}{d\hat{P}_{\lambda'}}\right). \tag{A.5'}$$

The proof of Theorem 1 follows by comparing (A.5) and (A.5′), arbitrariness of η and ζ, and the "monotone classes" theorem (Dellacherie and Meyer, 1975). □

Calculation of $E_\lambda(Y_s^2 \mid T = t)$: Since further calculations deal with λ' we omit the subscript in $P_{\lambda'}$. Starting with $P(Y_s \le y \mid T = t)$, define the density functions:

$$f_s(y \mid t) = \frac{d}{dy}P(Y_s \le y \mid T = t), \quad h_s(y) = \frac{\partial}{\partial y}P(Y_s \le y)$$

and

$$\varphi_s(t \mid y) = \frac{\partial}{\partial t}P(T \le t \mid Y_s = y), \quad \phi(t) = \frac{d}{dt}P(T \le t).$$

According to Bayes' rule,

$$f_s(y \mid T = t) = \frac{h_s(y)\varphi_s(t \mid Y_s = y)}{\phi(t)}. \tag{A.6}$$

Lemma 1. *Let Y and T be as above, and* $t > s$. *Then* $f_s(y \mid T = t)$,

$$f_s(y \mid T = t) = C(s,t)[B^2(s,t)y^2 + \tilde{\gamma}(s,t)]\frac{1}{\sqrt{2\pi G(s,t)}}\exp\left(-\frac{(y - M(s,t))^2}{2G(s,t)}\right),$$

where

$$C(s,t) = \frac{1}{B^2(s,t)[M^2(s,t) + G(s,t) + \tilde{\gamma}(s,t)]},$$

$m(s)$ *and* $\gamma(s)$ *satisfy* (55) *and* (56).

Proof. Find the conditional density $\varphi_s(t \mid y)$, $t \ge s$. To do this, calculate the conditional survival function $P(T > t \mid Y_s = y)$. If $s < t$ then $\{T > t\} \subseteq \{T > s\}$ and from Bayes' rule,

$$P(T > t \mid Y_s) = P(T > s \mid Y_s)P(T > t \mid T > s, Y_s). \tag{A.7}$$

For the second term on the RHS of (A.7) note that survival conditional on Y is

$$P(T > t \mid T > s,\ Y_s^t) = \exp\left(-\lambda' \int_s^t Y_u^2 \,\mathrm{d}u \right).$$

From Yashin (1985), the conditional survival function is

$$P(T > t \mid T > s,\ Y_s = y) = \exp\left(-\lambda' \int_s^t (\check{m}^2(s,u) + \check{\gamma}(s,u)] \,\mathrm{d}u \right). \tag{A.8}$$

where $\check{m}(s,u)$ and $\check{\gamma}(s,u)$ solutions of

$$\frac{\mathrm{d}\check{m}(s,u)}{\mathrm{d}u} = a(u)m(s,u) - 2\lambda'\check{\gamma}(s,u)m(s,u) \tag{A.9}$$

and

$$\frac{\mathrm{d}\check{\gamma}(s,u)}{\mathrm{d}u} = 2a(u)\check{\gamma}(s,u) + b^2(u) - 2\lambda'\check{\gamma}^2(s,u), \tag{A.10}$$

with initial conditions

$$\check{m}(s,s) = y, \quad \check{\gamma}(s,s) = 0.$$

Under these conditions, $m(s,t)$ is a linear function of y,

$$\check{m}(s,t) = B(s,t)y, \tag{A.11}$$

where

$$B(s,t) = \exp\left(\int_s^t [a(u) - 2\lambda'\check{\gamma}(s,u)] \,\mathrm{d}u \right),$$

and $\check{\gamma}(s,t)$ satisfy (A.10).

Using (A.7), (A.8), and (A.11), the conditional survival function is

$$P(T > t \mid Y_s = y) = P(T > s \mid Y_s = y)\exp\left(-\lambda' \int_s^t [B^2(s,u)y^2 + \check{\gamma}(s,u)] \,\mathrm{d}u \right), \tag{A.12}$$

where $m(s,u)$ and $\gamma(s,u)$ are given by (A.9) and (A.10).

Differentiating (A.12) with respect to t,

$$\varphi_s(t \mid y) = P(T > s \mid Y_s = y)\lambda'[B^2(s,t)y^2 + \check{\gamma}(s,t)] \times \exp\left(-\lambda' \int_s^t [B^2(s,u)y^2 + \check{\gamma}(s,u)] \,\mathrm{d}u \right). \tag{A.13}$$

Note that

$$P(T > s \mid Y_s = y) = P(T > s)\frac{p_s(y \mid T > s)}{h_s(y)}, \tag{A.14}$$

where $P_s(y \mid T > s) = (\partial/\partial y)P(Y_s \le y \mid T > s)$ and $H_s(y) = (\partial/\partial y)P(Y_s \le y)$. From Yashin (1985) the density is

$$p_s(y \mid T > s) = \frac{1}{\sqrt{2\pi\gamma(s)}}\exp\left(-\frac{(y - m(s))^2}{2\gamma(s)}\right),$$

and from (A.6)

$$f_s(y \mid T = t) = C_1(s,t)[B^2(s,t)y^2 + \tilde{\gamma}(s,t)]\exp\left(-\frac{(y - m(s))^2}{2\gamma(s)}\right)\exp(-y^2 D(s,t)), \tag{A.15}$$

where $D(s,t) = \lambda' \int_s^t B^2(s,u)\,du$, and $C_1(s,t)$ does not depend on y.

Lemma 1 follows after simple transformations of (A.15). Directly calculating $E_{\lambda'}(Y_s^2 \mid T = t)$ gives (57). □

References

O. Aalen, Nonparametric inference in connection with multiple decrement models, Scand. J. Statist. 3 (1976) 15–27.

A.L. Asachenkov, B.H. Sobolev and A.I. Yashin, A stochastic process model of oncological disease, Preprint No. 101, Department of Mathematics, USSR Academy of Sciences (1988) (In Russian).

F. Campilo and F. Le Gland, MLE for partially observed diffusions: direct minimization vs. the EM algorithm, Stochastic Process. Appl. 33 (1989) 245–274.

C. Dellacherie and P. Meyer, Probabilities et Potential (Hermann, Paris, 1975) p. 291.

A. Dembo and O. Zeitouni, Parameter estimation of partially observed continuous time stochastic processes via the EM algorithm, Stochastic Process. Appl. 23 (1986) 91–113.

A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), J. Roy. Statist. Soc. 39 Ser. B 1 (1977) 1–38.

Early Breast Cancer Trialists' Collaborative Group, Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy – Part I and II, Lancet 339 (1992) 1–15; 71–85.

J. Heckman and B. Singer, The identifiability of the proportional hazards model, Rev. Econom. Stud. 51 (1984a) 231–241.

J. Heckman and B. Singer, A method for minimizing the impact of distributional assumptions in econometric models for duration data, Econometrics 52 (1984b) 271–320.

Y.M. Kabanov, R.S. Liptser and A.N. Shiryayev, Absolute continuity and singularity of locally absolute continuous probabilistic distributions, Mat. Sb. 3 (1978) 364–415.

E.L. Kaplan and P. Meier, Nonparametric estimation from incomplete observations, J. Amer. Statist. Assoc. 53 (1958) 457–481.

J. Kiefer and J. Wolfowitz, Consistency of the maximum likelihood estimator in the presence of infinitely many parameters, Ann. Math. Statist. 27 (1956) 887–906.

N. Laird, Nonparametric maximum likelihood estimation of a mixing distribution, J. Amer. Statist. Assoc. 73 (1978) 805–811.

B. Lindsay, The geometry of mixture likelihoods. Parts I and II, Ann. Statist. 11 (1983) 86–94; 783–792.

G.I. Marchuk, A.L. Asachenkov, Y.S. Smolyaninov and B.G. Sobolev, On the problem of processing clinical and laboratory data on oncological patients, Soviet J. Numer. Anal. Math. Modelling 4 (5) (1989) 381–396.

W. Nelson, Theory and application of hazard plotting for censored failure data, Technometrics 14 (1972) 945–966.

R. Orchard and M.A. Woodbury, A missing information principle, theory and application, in: L.M. LeCam, J. Neyman and E.L. Scott, eds., Proc. Sixth Berkeley Symp. on Mathematical Statistics and Probability (Univ. of California Press, Berkeley, CA, 1972) pp. 697–715.

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

C.F.J. Wu, On the convergence properties of the EM algorithm, Ann. Statist. 11 (1983) 95–103.

A.I. Yashin, Dynamics in survival analysis: conditional Gaussian property versus Cameron–Martin formula, in: N.V. Krylov, R.S. Liptser and A.A. Novikov, eds., Statistic and Control of Stochastic Processes (Springer, NY, 1985).

A.I. Yashin, Unobserved covariates in survival models: smoothing estimates, Research Report 91-04-01, Center for Population Analysis and Policy, University of Minnesota (1991).