Journal of Statistical Research *ISSN 0256 - 422 X*
2011, Vol. 45, No. 2, pp. 77-94
Bangladesh

SIMULTANEOUS ESTIMATION OF BASELINE HAZARD RATE AND REGRESSION PARAMETERS IN COX PROPORTIONAL HAZARDS MODEL WITH MEASUREMENT ERROR

ALEXANDER KUKUSH
National Taras Shevchenko University of Kyiv
01601 Kyiv, Volodymyrska st. 64, Ukraine
Email: alexander_kukush@univ.kiev.ua

SÁNDOR BARAN, ISTVÁN FAZEKAS
Faculty of Informatics, University of Debrecen
H-4026 Debrecen, Kassai str. 26, Hungary
Email: baran.sandor@inf.unideb.hu, fazekas.istvan@inf.unideb.hu

ELENA USOLTSEVA
National Taras Shevchenko University of Kyiv
01601 Kyiv, Volodymyrska st. 64, Ukraine
Email: elena_usolceva@ukr.net

SUMMARY

We consider Cox proportional hazards model under covariate measurement error and investigate a simultaneous estimation method for the baseline hazard and covariate parameter. We show the strong consistency of the estimators and we also estimate their rate of convergence. Simulation results are also presented to illustrate the theoretical ones.

Keywords and phrases: Cox proportional hazards model, measurement error, strong consistency, Kullback-Leibler distance.

AMS Classification: Primary 62N01; Secondary 62N02.

1 Introduction

Cox semiparametric proportional hazards model is very popular in biometrics and medical statistics. Often the variables of interest cannot be observed directly and actually the so called surrogate data are observed instead. This is modeled by the presence of measurement errors. Ignoring the difference between the variables of interest and surrogate data leads to "naive" estimators of regression parameters that are usually severely biased. Recently, the discussion on measurement error in the Cox model has become vivid, see e.g. the corresponding paper of Augustin [3] and references therein. However, still there are only few results on the consistency of reasonable estimators in the Cox model under measurement errors.

In the present paper we deal with censored observations under measurement error and propose to estimate the baseline hazard function and regression parameters simultaneously. The baseline hazard function is not parameterized and belongs to a compact set of continuous positive functions. We use the partial log-likelihood function and correct it for censoring and measurement error following the ideas of Augustin [3]. Our simultaneous estimator maximizes the corrected objective function on the compact parameter set. Under mild assumptions we prove the strong consistency of the estimators and give the rate of convergence in terms of Kullback-Leibler distance between the true and estimated density of (Y, Δ, X), where Y is the censored lifetime, Δ is the censorship indicator, and X is the variable of interest. A computer simulation is presented to illustrate theoretical findings.

The paper is organized as follows. Section 2 gives the model of observations and constructs the estimators. Our main consistency result is presented in Section 3. The rate of convergence is derived in Section 4, simulation results are reported in Section 5, and Section 6 concludes.

2 Model and Estimator

Consider the Cox proportional hazards model [5], where the intensity of failure for the survival time T at time point t of an individual given a covariate vector X is specified as

$$\Lambda(t \,|\, X; \lambda, \beta) := \lambda(t) \exp(\beta^\top X), \tag{2.1}$$

where β is a k-dimensional parameter belonging to a parameter set $\Theta_\beta \subset \mathbb{R}^k$, while $\lambda(t) \in \Theta_\lambda \subset C[0, \tau]$, $\tau > 0$, is the baseline hazard function, i.e. the hazard function for $X = 0$, and Θ_λ consists of positive functions. This means that the conditional pdf of T given X equals

$$f_T(t \,|\, X; \lambda, \beta) = \Lambda(t \,|\, X; \lambda, \beta) \exp\left(-\int_0^t \Lambda(s \,|\, X; \lambda, \beta)\, \mathrm{d}s \right) \tag{2.2}$$

and

$$\int_0^\infty \Lambda(t \,|\, X; \lambda, \beta)\, \mathrm{d}t = \infty.$$

Hence,

$$\Lambda(t \,|\, X; \lambda, \beta) = -\frac{\mathrm{d}}{\mathrm{d}t} \log G_T(t \,|\, X; \lambda, \beta) = \frac{f_T(t \,|\, X; \lambda, \beta)}{G_T(t \,|\, X; \lambda, \beta)}, \tag{2.3}$$

where $G_T(t \,|\, X; \lambda, \beta) := 1 - F_T(t \,|\, X; \lambda, \beta)$ is the conditional survival function of T given X. However, instead of the lifetimes T one can usually observe a censored lifetime $Y := \min\{T, C\}$, where C is the censor distributed on $[0, \tau]$, together with the censorship indicator $\Delta := \mathbf{1}_{\{T \le C\}}$. Obviously, Y is also distributed on the interval $[0, \tau]$. Further, the censor C is independent of X and T.

The statistical problem is to estimate parameter β and baseline hazard λ on the basis of triples (Y_i, Δ_i, X_i), $i = 1, 2, \dots n$, of observations of censored lifetimes, corresponding censorship indicators and covariates, respectively. We assume that observed lifetimes $T_1, T_2, \dots, T_n$, censors $C_1, C_2, \dots, C_n$ and covariates $X_1, X_2, \dots, X_n$ are independent copies of T, C and X, respectively.

In the classical case, that is if the covariates can directly be observed, $\lambda(t)$ and β can be estimated by maximization of the partial (or Breslow's) log-likelihood function [4, 7]

$$Q_n(\lambda, \beta) := \frac{1}{n} \sum_{i=1}^{n} q(Y_i, \Delta_i, X_i; \lambda, \beta), \tag{2.4}$$

where

$$q(Y, \Delta, X; \lambda, \beta) := \Delta\big(\log \lambda(Y) + \beta^\top X\big) - \mathrm{e}^{\beta^\top X} \int_0^Y \lambda(u) \mathrm{d}u.$$

In the present paper we assume the existence of measurement errors in the covariates, that is instead of X_i we observe

$$W_i = X_i + U_i, \qquad i = 1, 2, \dots, n, \tag{2.5}$$

where the errors $\{U_i\}$ are independent copies of a k-dimensional random vector U with known moment generating function $M_U(\beta) := \mathsf{E}\mathrm{e}^{\beta^\top U}$, and independent of $\{X_i, T_i, C_i\}$. In this case, according to the ideas of Augustin [3], objective function $Q_n(t)$ has to be corrected for measurement errors with the help of deconvolution method [10]. The corrected objective function is defined as

$$Q_n^{cor}(\lambda, \beta) := \frac{1}{n} \sum_{i=1}^{n} q^{cor}(Y_i, \Delta_i, W_i; \lambda, \beta), \tag{2.6}$$

where

$$q^{cor}(Y, \Delta, W; \lambda, \beta) := \Delta\big(\log \lambda(Y) + \beta^\top W\big) - \frac{\mathrm{e}^{\beta^\top W}}{M_U(\beta)} \int_0^Y \lambda(u) \mathrm{d}u.$$

Naturally, we have

$$\mathsf{E}\big(q^{cor}(Y, \Delta, W; \lambda, \beta) \,\big|\, Y, \Delta, X\big) = q(Y, \Delta, X; \lambda, \beta),$$

almost surely, implying

$$\mathsf{E}q^{cor}(Y, \Delta, W; \lambda, \beta) = \mathsf{E}q(Y, \Delta, X; \lambda, \beta) =: q_\infty(\lambda, \beta). \tag{2.7}$$

The corrected estimators $(\widehat{\lambda}_n, \widehat{\beta}_n)$ of (λ, β) are defined as

$$(\widehat{\lambda}_n, \widehat{\beta}_n) := \arg \max_{(\lambda, \beta) \in \Theta} Q_n^{cor}(\lambda, \beta), \tag{2.8}$$

where $\Theta := \Theta_\lambda \times \Theta_\beta$. If the parameter sets are compact, then Θ will also be a compact set in $C[0, \tau] \times \mathbb{R}^k$. Since Q_n^{cor} is continuous, the maximum in (2.8) will obviously be attained.

3 Strong Consistency

We prove the strong consistency of the estimators defined by (2.6) under the assumptions below

(i) $\Theta_\lambda \subset C[0,\tau]$ is the following compact convex set of positive functions

$$\Theta_\lambda := \big\{ f : [0,\tau] \to \mathbb{R} \,\big|\, f(t) \ge a, \forall t \in [0,\tau] \text{ and } |f(t) - f(s)| \le L|t-s|, \forall t,s \in [0,\tau] \big\},$$

where $a > 0$ and $L > 0$ are fixed constants.

(ii) $\Theta_\beta \subset \mathbb{R}^k$ is compact and convex.

(iii) Measurement error U has zero mean and for a fixed $\varepsilon > 0$,

$$\mathsf{E}\mathrm{e}^{D\|U\|} < \infty \qquad \text{where} \qquad D := \max_{\beta \in \Theta_\beta} \|\beta\| + \varepsilon.$$

(iv) $\mathsf{E}\mathrm{e}^{D\|X\|} < \infty$, where $D > 0$ is the constant defined in (iii).

(v) τ is the right endpoint of the distribution of C, i.e. $\mathsf{P}(C > \tau) = 0$ and for all $\varepsilon > 0$ we have $\mathsf{P}(C > \tau - \varepsilon) > 0$.

(vi) The covariance matrix S_X of the random vector X is positive definite.

Theorem 1. *Consider the Cox proportional hazards model with measurement error defined by* (2.1) *and* (2.5) *with true parameters* $\lambda_0(t)$ *and* β_0, *and assume that conditions (i)–(vi) are satisfied. Then* $(\widehat{\lambda}_n, \widehat{\beta}_n)$ *are strongly consistent estimators of the true parameters* (λ_0, β_0), *that is*

$$\sup_{t \in [0,\tau]} |\widehat{\lambda}_n(t) - \lambda_0(t)| \to 0 \qquad \textit{and} \qquad \widehat{\beta}_n \to \beta_0$$

almost surely as $n \to \infty$.

Proof. In order to prove the strong consistency of the estimators $(\widehat{\lambda}_n, \widehat{\beta}_n)$ one has to prove

(a) $\sup_{(\lambda,\beta)\in\Theta} \left| Q_n^{cor}(\lambda,\beta) - q_\infty(\lambda,\beta) \right| \to 0$ almost surely as $n \to \infty$;

(b) $q_\infty(\lambda,\beta) \le q_\infty(\lambda_0,\beta_0)$, and equality holds if and only if $\lambda \equiv \lambda_0$ and $\beta = \beta_0$.

Let $\frac{\partial q^{cor}}{\partial \lambda}$ denote the Fréchet derivative of q^{cor} with respect to the function λ which is a linear functional on $C[0,\tau]$. Hence, $\left\| \frac{\partial q^{cor}}{\partial \lambda} \right\|$ is the norm of this linear functional corresponding to the supremum norm on $C[0,\tau]$, while for $h \in C[0,\tau]$ the expression $\left\langle \frac{\partial q^{cor}}{\partial \lambda}, h \right\rangle$ means the effect of the functional $\frac{\partial q^{cor}}{\partial \lambda}$ on h. E.g. if

$$q^{cor}(\lambda) = \int_0^\tau \lambda^2(t)\mathrm{d}t, \quad \text{then} \quad \left\langle \frac{\partial q^{cor}}{\partial \lambda}, h \right\rangle = 2\int_0^\tau \lambda(t)h(t)\mathrm{d}t.$$

According to results of [8] to verify (a) it suffices to show (a1) $Q_n^{cor}(\lambda,\beta)\to q_\infty(\lambda,\beta)$ almost surely as $n\to\infty$ for all $(\lambda,\beta)\in\Theta$;

(a2) $\mathsf{E}\sup\limits_{(\lambda,\beta)\in\Theta}\left\|\dfrac{\partial q^{cor}}{\partial\lambda}(Y,\Delta,W;\lambda,\beta)\right\|<\infty,\quad \mathsf{E}\sup\limits_{(\lambda,\beta)\in\Theta}\left\|\dfrac{\partial q^{cor}}{\partial\beta}(Y,\Delta,W;\lambda,\beta)\right\|<\infty;$

(a3) $q_\infty(\lambda,\beta)$ is continuous in (λ,β).

By conditions (i)–(iv) for all fixed $(\lambda,\beta)\in\Theta$ we have $\mathsf{E}\left|q^{cor}(Y,\Delta,W;\lambda,\beta)\right|<\infty$. As our observations (Y_i,Δ_i,W_i), $i=1,2,\dots n$, are i.i.d., the strong law of large numbers applies, so $Q_n^{cor}(\lambda,\beta)\to\mathsf{E}q^{cor}(Y,\Delta,W;\lambda,\beta)$ almost surely as $n\to\infty$, that together with (2.7) implies (a1).

Next, for $h\in C[0,\tau]$

$$\left\langle\frac{\partial q^{cor}}{\partial\lambda}(Y,\Delta,W;\lambda,\beta),h\right\rangle=\frac{\Delta h(Y)}{\lambda(Y)}-\frac{\mathrm{e}^{\beta^\top W}}{M_U(\beta)}\int_0^Y h(u)\mathrm{d}u,$$

yielding

$$\left\|\frac{\partial q^{cor}}{\partial\lambda}(Y,\Delta,W;\lambda,\beta)\right\|\le\sup_{\lambda\in\Theta_\lambda}\left\|\frac{1}{\lambda(Y)}\right\|+\frac{\tau\mathrm{e}^{D(\|X\|+\|U\|)}}{\min_{\beta\in\Theta_\beta}M_U(\beta)},$$

so by (i)–(iv) the first condition of (a2) holds.

Further,

$$\frac{\partial q^{cor}}{\partial\beta}(Y,\Delta,W;\lambda,\beta)=\Delta W-\left(M_U(\beta)W-\mathsf{E}\big(U\mathrm{e}^{\beta^\top U}\big)\right)\mathrm{e}^{\beta^\top W}M_U^{-2}(\beta)\int_0^Y\lambda(u)\mathrm{d}u,$$

which directly implies the second condition of (a2). We remark that the extra term $\varepsilon>0$ in the definition of the constant D used in conditions (iii) and (iv) is needed to ensure

$$\mathsf{E}\sup_{\beta\in\Theta_\beta}\left\|U\mathrm{e}^{\beta^\top U}\right\|<\infty\qquad\text{and}\qquad\mathsf{E}\sup_{\beta\in\Theta_\beta}\left\|X\mathrm{e}^{\beta^\top X}\right\|<\infty,$$

respectively.

Finally, by definition (2.7)

$$q_\infty(\lambda,\beta)=\mathsf{E}q(Y,\Delta,X;\lambda,\beta)=\mathsf{E}\left(\Delta\big(\log\lambda(Y)+\beta^\top X\big)-\mathrm{e}^{\beta^\top X}\int_0^Y\lambda(u)\mathrm{d}u\right).$$

Obviously, $q(Y,\Delta,X;\lambda,\beta)$ is continuous in (β,λ), and by conditions (i) and (ii)

$$\left|q(Y,\Delta,X;\lambda,\beta)\right|\le\left|\Delta\big(\log\lambda(Y)+\beta^\top X\big)\right|+\left|\mathrm{e}^{\beta^\top X}\int_0^\tau\lambda(u)\mathrm{d}u\right|\le\mathcal{C}\left(1+\|X\|+\mathrm{e}^{D\|X\|}\right),\quad(3.1)$$

where $\mathcal{C}$ is a positive constant. Conditions (iii) and (iv) imply that the right hand side of (3.1) has a finite mean, that together with the dominated convergence theorem implies (a3).

To verify (b) we are going to use the following general result that is quite well known in information theory (see e.g. [1, Lemma 8.3.1]).

Lemma 3.1. *Let* ϱ *and* ϱ_0 *be arbitrary densities with respect to a* σ*-finite measure* μ *on the* σ*-field* $\mathcal{B}(\mathbb{R}^k)$. *If*

$$\int_{\mathbb{R}^k} \varrho_0(x)\log\varrho(x)\mathrm{d}\mu(x) \quad \textit{is finite, then} \quad \int_{\mathbb{R}^k} \varrho_0(x)\log\varrho_0(x)\mathrm{d}\mu(x)$$

exists and

$$\int_{\mathbb{R}^k} \varrho_0(x)\log\varrho_0(x)\mathrm{d}\mu(x) \geq \int_{\mathbb{R}^k} \varrho_0(x)\log\varrho(x)\mathrm{d}\mu(x). \tag{3.2}$$

If

$$\int_{\mathbb{R}^k} \varrho_0(x)\log\varrho_0(x)\mathrm{d}\mu(x) \quad \textit{is finite, then} \quad \int_{\mathbb{R}^k} \varrho_0(x)\log\varrho(x)\mathrm{d}\mu(x)$$

exists and (3.2) *holds.*

Equality in (3.2) *is attained if and only if* $\varrho(x) = \varrho_0(x)$ *for almost all* x *with respect to the measure* μ.

Consider the couple (Y, Δ) that is distributed on $\mathcal{X} := \mathbb{R}^+ \times \{0,1\}$ with $\mathbb{R}^+ := (0,\infty)$, and consider on $\mathcal{X}$ measure $\mu = \lambda_1 \times \lambda_c$, where λ_1 and λ_c denote the Lebesgue and the counting measure, respectively.

To simplify notation, for a moment let us fix the covariate vector X. First we show that the joint pdf of (Y,Δ) with respect to μ equals

$$f(y,\delta|X;\lambda_0,\beta_0) := f_T^{\delta}(y|X;\lambda_0,\beta_0)G_T^{1-\delta}(y|X;\lambda_0,\beta_0)f_C^{1-\delta}(y)G_C^{\delta}(y), \qquad (y,\delta)\in\mathcal{X}, \tag{3.3}$$

where f_T and f_C are the densities, while G_T and G_C are the survival functions of T and C, respectively. Now, to verify (3.3) it suffices to prove that for all $A \in \mathcal{B}(\mathbb{R}^+)$

$$\int_{A\times\{0\}} f(y,\delta|X;\lambda_0,\beta_0)\,\mathrm{d}\mu(y,\delta) = \mathsf{P}(Y\in A, \Delta = 0), \tag{3.4}$$

$$\int_{A\times\{1\}} f(y,\delta|X;\lambda_0,\beta_0)\,\mathrm{d}\mu(y,\delta) = \mathsf{P}(Y\in A, \Delta = 1) \tag{3.5}$$

hold. For the left hand side of (3.4) we have

$$\int_{A\times\{0\}} f(y,\delta|X;\lambda_0,\beta_0)\,\mathrm{d}\mu(y,\delta) = \int_A f_C(y)G_T(y)\mathrm{d}y = \mathsf{E}\mathbf{1}_A(C)G_T(C),$$

while for the right hand side

$$\begin{aligned}\mathsf{P}(Y\in A,\Delta=0) &= \mathsf{P}(C\in A, T\geq C) = \mathsf{E}(\mathbf{1}_A(C)\mathbf{1}_{\{T\geq C\}}) \\ &= \mathsf{E}\big(\mathbf{1}_A(C)\mathsf{E}(\mathbf{1}_{\{T\geq C\}}\,|\,C)\big) = \mathsf{E}\mathbf{1}_A(C)G_T(C),\end{aligned}$$

which clearly proves (3.4). Equality (3.5) can be proved in the same way.

Further, as the distribution of C is concentrated on the interval $[0,\tau]$, for $y > \tau$ we have $f_C(y) = 0$ and $G_C(y) = 0$, implying $f(y,\delta|X;\lambda,\beta) = 0$ for $y > \tau$, $\delta \in \{0,1\}$. Hence, $f(y,\delta|X;\lambda,\beta)$ and the corresponding log-likelihood function

$$\ell(y,\delta|X;\lambda,\beta) := \log f(y,\delta|X;\lambda,\beta) = \ell_T(y,\delta|X;\lambda,\beta) + \ell_C(y,\delta), \tag{3.6}$$

where

$$\begin{aligned}\ell_T(y,\delta|X;\lambda,\beta) &:= \delta \log f_T(y|X;\lambda,\beta) + (1-\delta)\log G_T(y|X;\lambda,\beta),\\ \ell_C(y,\delta) &:= (1-\delta)\log f_C(y) + \delta \log G_C(y),\end{aligned}$$

depends on the values of $f_T(y|X;\lambda,\beta)$ only on the interval $[0,\tau]$. Applying Lemma 3.1 for $f(y,\delta|X;\lambda_0,\beta_0)$ and $f(y,\delta|X;\lambda,\beta)$ we obtain

$$\mathsf{E}\ell(Y,\Delta|X;\lambda,\beta) \le \mathsf{E}\ell(Y,\Delta|X;\lambda_0,\beta_0), \tag{3.7}$$

which is equivalent to

$$\mathsf{E}\ell_T(Y,\Delta|X;\lambda,\beta) \le \mathsf{E}\ell_T(Y,\Delta|X;\lambda_0,\beta_0), \tag{3.8}$$

and equalities in (3.7) and (3.8) hold if and only if

$$\begin{aligned}&f_T^{\delta}(y|X;\lambda_0,\beta_0)\, G_T^{1-\delta}(y|X;\lambda_0,\beta_0) f_C^{1-\delta}(y) G_C^{\delta}(y)\\ &\qquad = f_T^{\delta}(y|X;\lambda,\beta) G_T^{1-\delta}(y|X;\lambda,\beta) f_C^{1-\delta}(y) G_C^{\delta}(y)\end{aligned} \tag{3.9}$$

almost everywhere on $[0,\tau]$ with respect to measure μ. For $\delta = 1$ condition (3.9) reduces to

$$f_T(y|X;\lambda_0,\beta_0) G_C(y) = f_T(y|X;\lambda,\beta) G_C(y) \tag{3.10}$$

almost everywhere with respect to the Lebesgue measure λ_1, and by condition (v) — since for $y < \tau$ we have $G_C(y) > 0$ — (3.10) is equivalent to

$$f_T(y|X;\lambda_0,\beta_0) = f_T(y|X;\lambda,\beta) \qquad \text{almost everywhere with respect to } \lambda_1. \tag{3.11}$$

If (3.11) is valid, then (3.9) is true for $\delta = 0$. Hence, we have equality in (3.8) if and only if (3.11) holds.

Now, using (2.3) one can easily see that $q_\infty(\lambda,\beta) = \mathsf{E}\ell_T(Y,\Delta|X;\lambda,\beta)$, so from inequality (3.8) we obtain

$$q_\infty(\lambda,\beta) = \mathsf{E}\ell_T(Y,\Delta|X;\lambda,\beta) \le \mathsf{E}\ell_T(Y,\Delta|X;\lambda_0,\beta_0) = q_\infty(\lambda_0,\beta_0), \tag{3.12}$$

where equality is attained if and only if

$$f_T(t|X;\lambda_0,\beta_0) = f_T(t|X;\lambda,\beta) \qquad \text{almost surely for almost all } t \in [0,\tau]. \tag{3.13}$$

Suppose now that (3.13) holds. Representation (2.3) implies that

$$\Lambda(t|X;\lambda_0,\beta_0) = \Lambda(t|X;\lambda,\beta) \qquad \text{almost surely for almost all } t \in [0,\tau], \tag{3.14}$$

so $\exp\big((\beta-\beta_0)^\top X\big)$ is constant, and in this way $(\beta-\beta_0)^\top X$ is also constant with probability one. Hence, $\mathsf{Var}\big((\beta-\beta_0)^\top X\big)=0$ that together with condition (vi) yields $\beta=\beta_0$. Obviously, using again (3.14) one can see that $\beta=\beta_0$ implies $\lambda_0(t)=\lambda(t),\ t\in[0,\tau]$, which together with (3.12) and (3.13) completes the proof of condition (b). □

Remark 1. It is not necessary for the censor C to have a pdf. It can have any (e.g. discrete) distribution μ_C on $[0,\tau]$ provided the survival function $G_C(y)$ is positive for all $y<\tau$. In this case the reasoning concerning the joint density (3.3) of (Y,Δ) has to be corrected. Couple (Y,Δ) is distributed on $\mathcal{X}=(0,\tau]\times\{0,1\}=\big((0,\tau]\times\{0\}\big)\cup\big((0,\tau]\times\{1\}\big)=:\mathcal{X}_0\cup\mathcal{X}_1$, and measure μ on $\mathcal{X}$ has to be defined separately on $\mathcal{X}_0$ and $\mathcal{X}_1$ as $\mu(A\times\{1\}):=\lambda_1(A)$ and $\mu(A\times\{1\}):=\mu_C(A)$, respectively, where $A\in\mathcal{B}\big((0,\tau]\big)$. Using symbolic notation $\mu=\lambda_1\times\delta_1+\mu_C\times\delta_0$, where δ_1 and δ_0 are Dirac measures concentrated at 1 and 0, respectively. In this way the density of C with respect to μ_C is $f_C(y)\equiv 1$, and (3.3) takes the form

$$f(y,\delta|X;\lambda_0,\beta_0):=f_T^{\delta}(y|X;\lambda_0,\beta_0)G_T^{1-\delta}(y|X;\lambda_0,\beta_0)G_C^{\delta}(y),\qquad (y,\delta)\in\mathcal{X},\tag{3.15}$$

specifying a density on $\mathcal{X}$ with respect to measure μ. Now, the validity of (3.4) and (3.5) for the density defined by (3.15) can be checked in the same way as before.

4 Kullback-Leibler Distance of the True and Estimated Density Functions

To estimate the rate of convergence of the estimators $(\widehat{\lambda}_n,\widehat{\beta}_n)$ defined by (2.8) to the true parameter values (λ_0,β_0), besides conditions (i) – (vi) of the strong consistency we need an additional one, namely

(vii) $S_n(\lambda,\beta)/\sqrt{n}$ converges in distribution in $C(\Theta)$ to a Gaussian measure, where

$$S_n(\lambda,\beta):=n\big(Q_n^{cor}(\lambda,\beta)-q_\infty(\lambda,\beta)\big)=\sum_{i=1}^{n}\big(q^{cor}(Y_i,\Delta_i,W_i;\lambda,\beta)-\mathsf{E}q^{cor}(Y,\Delta,W;\lambda,\beta)\big).\tag{4.1}$$

However, according to the statement of Lemma 4.1, which is an application of Theorem 2 of [11], the following assumption is sufficient to check the validity of condition (vii).

(vii') $\mathsf{E}e^{2D_\beta\|X\|}<\infty$ and $\mathsf{E}e^{2D_\beta\|U\|}<\infty$, where $D_\beta:=\max_{\beta\in\Theta_\beta}\|\beta\|>0$.

Lemma 4.1. *Consider the Cox proportional hazards model with measurement error defined by* (2.1) *and* (2.5). *Under assumptions (i) – (iv) condition (vii') is a sufficient condition for (vii).*

Proof. Let us consider $q^{cor}(Y,\Delta,W;\lambda,\beta)$ as a random element on $\Theta=\Theta_\lambda\times\Theta_\beta\subset C[0,\tau]\times\mathbb{R}^k$ and let

$$\rho\big((\lambda_1,\beta_1),(\lambda_1,\beta_1)\big):=\sup_{t\in[0.\tau]}|\lambda_1(t)-\lambda_2(t)|+\|\beta_1-\beta_2\|.$$

By conditions (i) and (ii), (Θ, ϱ) is a compact metric space, so according to Theorem 2 of [11] applied to centered random elements $q^{cor}(Y_i, \Delta_i, W_i; \lambda, \beta) - \mathsf{E}q^{cor}(Y, \Delta, W; \lambda, \beta)$ to prove asymptotic normality of $S_n(\lambda, \beta)$ it suffices to show

(c1) $\mathsf{P}\big(q^{cor} \in \mathsf{Lip}(\varrho)\big) = 1$, where $\mathsf{Lip}(\varrho) \subset C(\Theta)$ is the set of Lipschitz functions on Θ with respect to metric ϱ;

(c2) $\int\limits_0^1 H_\varrho^{1/2}(\Theta, v)\mathrm{d}v < \infty$,

where for a compact metric space (S, ϱ), function $H_\varrho(S, \varepsilon)$ (or simply $H(S, \varepsilon)$) is the ε-entropy of S, that is $H_\varrho(S, \varepsilon) := \log N_\varrho(S, \varepsilon)$, and $N_\varrho(S, \varepsilon)$ (or simply $N(S, \varepsilon)$) is the minimum number of balls with diameter not greater than 2ε which cover S (see e.g. [6]);

(c3) $\mathsf{E}\big\|q^{cor}(Y, \Delta, W; \lambda, \beta)\big\|_\varrho^2 < \infty$, where $\|\cdot\|_\varrho$ is the norm induced by the metric ϱ, that is for $g \in \mathsf{Lip}(\Theta)$ we have $\|g\|_\varrho := d(g) + |g(\lambda^*, \beta^*)|$, where

$$d(g) := \sup_{(\lambda_1,\beta_1)\neq(\lambda_2,\beta_2)} \frac{|g(\lambda_1, \beta_1) - g(\lambda_2, \beta_2)|}{\varrho\big((\lambda_1, \beta_1), (\lambda_2, \beta_2)\big)},$$

and (λ^*, β^*) is some fixed element in Θ.

Now, as a consequence of condition (a2) of the proof of Theorem 1 we have

$$\sup_{(\lambda,\beta)\in\Theta} \left\| \frac{\partial q^{cor}}{\partial \lambda}(Y, \Delta, W; \lambda, \beta) \right\| < \infty \quad \text{and} \quad \sup_{(\lambda,\beta)\in\Theta} \left\| \frac{\partial q^{cor}}{\partial \beta}(Y, \Delta, W; \lambda, \beta) \right\| < \infty$$

almost surely, that directly implies (c1).

Next, consider the compact metric spaces Θ_λ and Θ_β with the supremum norm and with the Euclidean norm, respectively, and let $0 < \varepsilon$ be an arbitrary constant. Obviously,

$$N_\varrho(\Theta, 2\varepsilon) \le N(\Theta_\lambda, \varepsilon) N(\Theta_\beta, \varepsilon),$$

hence

$$H_\varrho^{1/2}(\Theta, 2\varepsilon) \le \sqrt{2}\big(H^{1/2}(\Theta_\lambda, \varepsilon) + H^{1/2}(\Theta_\beta, \varepsilon)\big).$$

Now, as for $\varepsilon \le 1$ we have $N(\Theta_\beta, \varepsilon) \le \mathcal{C}\varepsilon^{-k}$ with some positive constant $\mathcal{C}$, implying $H(\Theta_\beta, \varepsilon) \le \log \mathcal{C} - k \log \varepsilon$, using $\int_0^1 (-\log u)^{1/2}\mathrm{d}u < \infty$, we obtain

$$\int\limits_0^1 H^{1/2}(\Theta_\beta, u)\mathrm{d}u < \infty. \tag{4.2}$$

Further, according to the results of Potapov [9], if Θ_λ is of "uniform type", that is there exist $b > 1$, $\mathcal{C} > 0$ and $\nu_0 > 0$, such that for all $0 < \nu < \nu_0$, inequality

$$H(\Theta_\lambda, b\nu) + \mathcal{C} \le H(\Theta_\lambda, \nu) \tag{4.3}$$

holds, then for $\varepsilon \le 1$ we have $H(\Theta_\lambda, \varepsilon) \le \mathcal{C}\varepsilon^{-1}$ implying

$$\int_0^1 H^{1/2}(\Theta_\lambda, u)\mathrm{d}u < \infty. \tag{4.4}$$

However, as Θ_λ is compact and since convex, it is also connected, by Lemma 1 of [9] there exists $\nu_0 > 0$ such that for all $0 < \nu < \nu_0/4$

$$H(\Theta_\lambda, 4\nu) + 1 \le H(\Theta_\lambda, \nu)$$

is satisfied, which proves (4.3).

In this way, since

$$\int_0^1 H_\varrho^{1/2}(\Theta, v)\mathrm{d}v = 2\int_0^{1/2} H_\varrho^{1/2}(\Theta, 2u)\mathrm{d}u \le 2^{3/2}\left(\int_0^{1/2} H^{1/2}(\Theta_\lambda, u)\mathrm{d}u + \int_0^{1/2} H^{1/2}(\Theta_\beta, u)\mathrm{d}u\right),$$

(c2) follows from (4.2) and (4.4).

Finally, using conditions (i) and (ii) after short calculations one can see that there exists a positive constant $\mathcal{C}$ such that

$$\left\|q^{cor}(Y, \Delta, W; \lambda, \beta)\right\|_\varrho^2 \le \mathcal{C}\Big(1 + \|W\| + \mathrm{e}^{D_\beta\|W\|}\Big),$$

thus (c3) is a direct consequence of (vii'). □

Now, we can formulate our result about the rate of convergence of the estimators.

Theorem 2. *Consider the Cox proportional hazards model with measurement error defined by* (2.1) *and* (2.5) *with true parameters* $\lambda_0(t)$ *and* β_0, *and assume that conditions (i), (ii), (iv) – (vi) and (vii') hold. Then*

$$\mathcal{D}\big(f(Y, \Delta, X; \lambda_0, \beta_0), f(Y, \Delta, X; \widehat{\lambda}_n, \widehat{\beta}_n)\big) = \frac{O_p(1)}{\sqrt{n}},$$

where for densities f_1 *and* f_2 *with respect to a measure* μ,

$$\mathcal{D}\big(f_1, f_2\big) := \int f_1(x)\log\frac{f_1(x)}{f_2(x)}\mathrm{d}\mu(x)$$

denotes the Kullback-Leibler distance of f_1 *and* f_2.

Proof. By the definition of the estimators $(\widehat{\lambda}_n, \widehat{\beta}_n)$ we have $Q_n^{cor}(\widehat{\lambda}_n, \widehat{\beta}_n) \ge Q_n^{cor}(\lambda_0, \beta_0)$ implying

$$0 \le q_\infty(\lambda_0, \beta_0) - q_\infty(\widehat{\lambda}_n, \widehat{\beta}_n) \le 2\sup_{(\lambda,\beta)\in\Theta}\left|Q_n^{cor}(\lambda, \beta) - q_\infty(\lambda, \beta)\right|. \tag{4.5}$$

According to Lemma 4.1, condition (vii') implies the asymptotic normality of $S_n(\lambda, \beta)/\sqrt{n}$ where $S_n(\lambda, \beta)$ is the sum defined by (4.1). Hence, the right hand side of (4.5) is $O_p(1)/\sqrt{n}$.

Let $(\lambda,\beta)\in\Theta$. Using the same ideas as in the proof (3.12) with the help of (3.3) and (3.6) we obtain

$$
\begin{aligned}
0 \le q_\infty(\lambda_0,\beta_0) - q_\infty(\lambda,\beta) &= \mathsf{E}\Big(\mathsf{E}\big(\ell_T(Y,\Delta|X;\lambda_0,\beta_0)\big|X\big)\Big) - \mathsf{E}\Big(\mathsf{E}\big(\ell_T(Y,\Delta|X;\lambda,\beta)\big|X\big)\Big) \\
&= \mathsf{E}\Big(\mathsf{E}\big(\ell(Y,\Delta|X;\lambda_0,\beta_0) - \ell(Y,\Delta|X;\lambda,\beta)\big|X\big)\Big) = \mathsf{E}\left(\mathsf{E}\left(\log\frac{f(Y,\Delta|X;\lambda_0,\beta_0)}{f(Y,\Delta|X;\lambda,\beta)}\middle|X\right)\right) \\
&= \mathsf{E}\left(\log\frac{f(Y,\Delta|X;\lambda_0,\beta_0)}{f(Y,\Delta|X;\lambda,\beta)}\right) = \mathsf{E}\left(\log\frac{f_T^{\Delta}(Y|X;\lambda_0,\beta_0)G_T^{1-\Delta}(Y|X;\lambda_0,\beta_0)}{f_T^{\Delta}(Y|X;\lambda,\beta)G_T^{1-\Delta}(Y|X;\lambda,\beta)}\right) \qquad (4.6) \\
&= \mathsf{E}\left(\log\frac{f_T(Y|X;\lambda_0,\beta_0)}{f_T(Y|X;\lambda,\beta)}\mathsf{P}(\Delta=1\,|X)\right) + \mathsf{E}\left(\log\frac{G_T(Y|X;\lambda_0,\beta_0)}{G_T(Y|X;\lambda,\beta)}\mathsf{P}(\Delta=0\,|X)\right) \\
&= \mathsf{E}\left(\log\frac{f_T(T|X;\lambda_0,\beta_0)}{f_T(T|X;\lambda,\beta)}\mathsf{P}(T\le C\,|X)\right) + \mathsf{E}\left(\log\frac{G_T(C|X;\lambda_0,\beta_0)}{G_T(C|X;\lambda,\beta)}\mathsf{P}(T> C\,|X)\right) \\
&= \mathsf{E}\left(\int_0^C f_T(t|X;\lambda_0,\beta_0)\log\frac{f_T(t|X;\lambda_0,\beta_0)}{f_T(t|X;\lambda,\beta)}\mathrm{d}t\right) + \mathsf{E}\left(\int_C^\tau f_T(t|X;\lambda_0,\beta_0)\log\frac{G_T(C|X;\lambda_0,\beta_0)}{G_T(C|X;\lambda,\beta)}\mathrm{d}t\right) \\
&= \mathsf{E}\left(\int_0^C f_T(t|X;\lambda_0,\beta_0)\log\frac{f_T(t|X;\lambda_0,\beta_0)}{f_T(t|X;\lambda,\beta)}\mathrm{d}t\right) + \mathsf{E}\left(G_T(C|X;\lambda_0,\beta_0)\log\frac{G_T(C|X;\lambda_0,\beta_0)}{G_T(C|X;\lambda,\beta)}\right).
\end{aligned}
$$

Further, let $f(y,\delta,x|\lambda,\beta)$ be the joint density of the triple of censored lifetime, censorship indicator and covariate X on $[0,\tau]\times\{0,1\}\times\mathbb{R}^k$, which by (3.3) equals

$$
\begin{aligned}
f(y,\delta,x|\lambda,\beta) &= f(y,\delta|X;\lambda_0,\beta_0)f_X(x) \\
&= f_T^{\delta}(y|X;\lambda_0,\beta_0)G_T^{1-\delta}(y|X;\lambda_0,\beta_0)f_C^{1-\delta}(y)G_C^{\delta}(y)f_X(x),
\end{aligned}
$$

where $f_X(x)$ is the density of X. However, we do not assume that X has a pdf with respect to the Lebesgue measure. Let μ_X be the distribution of X on $\mathbb{R}^k$ and one can consider $f(y,\delta,x|\lambda,\beta)$ as a density with respect to the product measure $\lambda_1\times\lambda_c\times\mu_X$. Hence, from (4.6) we obtain

$$
q_\infty(\lambda_0,\beta_0) - q_\infty(\lambda,\beta) = \mathsf{E}\left(\log\frac{f(Y,\Delta,X;\lambda_0,\beta_0)}{f(Y,\Delta,X;\lambda,\beta)}\right) = \mathcal{D}\big(f(Y,\Delta,X;\lambda_0,\beta_0), f(Y,\Delta,X;\lambda,\beta)\big),
$$

that completes the proof. □

5 Simulation Results

To illustrate the behavior of the proposed estimator we performed computer simulations using Matlab (version 2008a). For optimization Matlab function `fmincon` was used. Naturally, objective function $Q_n^{cor}(\lambda,\beta)$ can not be maximized numerically with respect to a function $\lambda(t)\in C[0,\tau]$, so we applied two approximation methods of λ based on m points where m is an increasing function of the sample size. The first method is spline interpolation on m equidistant points, the second is Chebyshev interpolation, that is polynomial interpolation, where the nodes are roots of the

mth Chebyshev polynomial of the first kind. In both cases the objective function is maximized with respect to β and to the values of $\lambda(t)$ in the node points.

In our example 100 independent samples were simulated and using subsamples of increasing size the estimates of the parameters were calculated. For parameter β the mean, the standard deviation and the mean absolute error (MAE) was calculated for each sample size considered. For the function $\lambda(t)$ first the means of the estimated values at node points were calculated and the approximations were based on these mean values. To check the fit of the approximation we estimated the deviation in supremum norm from the true function. We remark that lifetimes T were generated with the help of inversion method.

Example 5.1. Consider the Cox proportional hazards model with measurement error where β is one-dimensional with true value $\beta_0 = 1$, while $\lambda_0(t) = 10 + t$. Covariate X and measurement error U are both normal with means -4 and 0 and standard deviations 0.4 and 0.1, respectively.

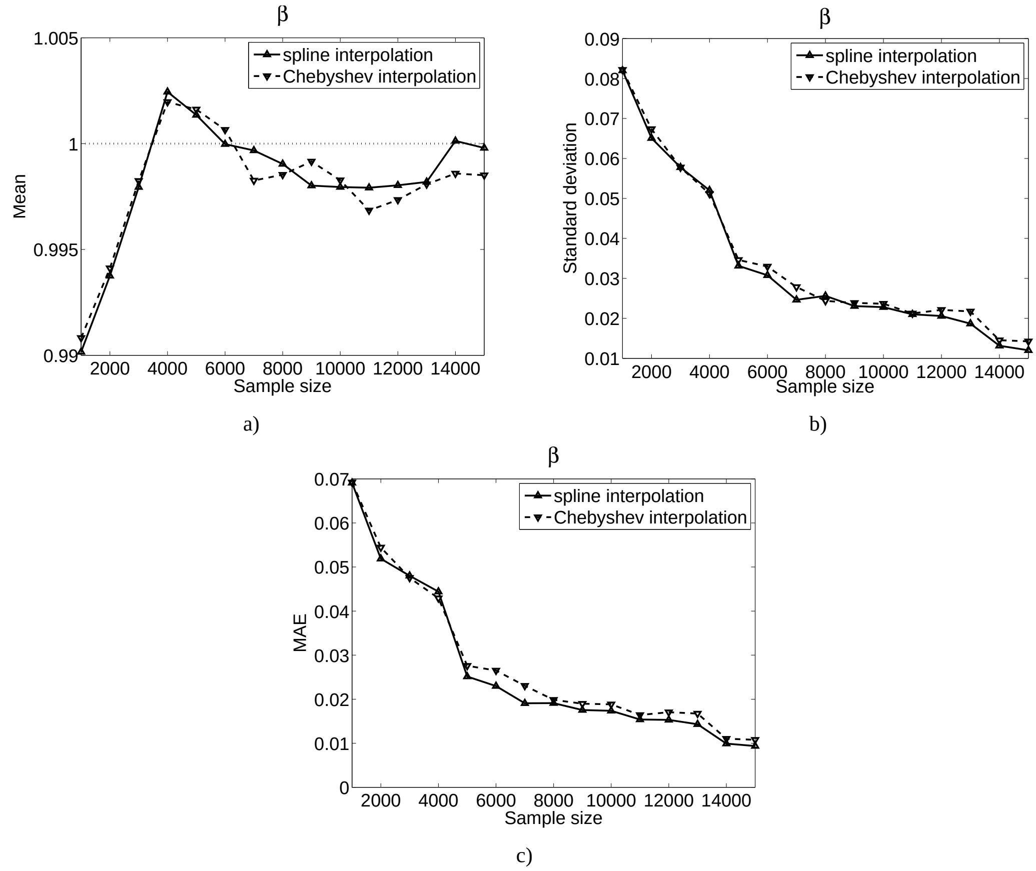


Figure 1: Means, standard deviations and MAEs of the estimates of β.

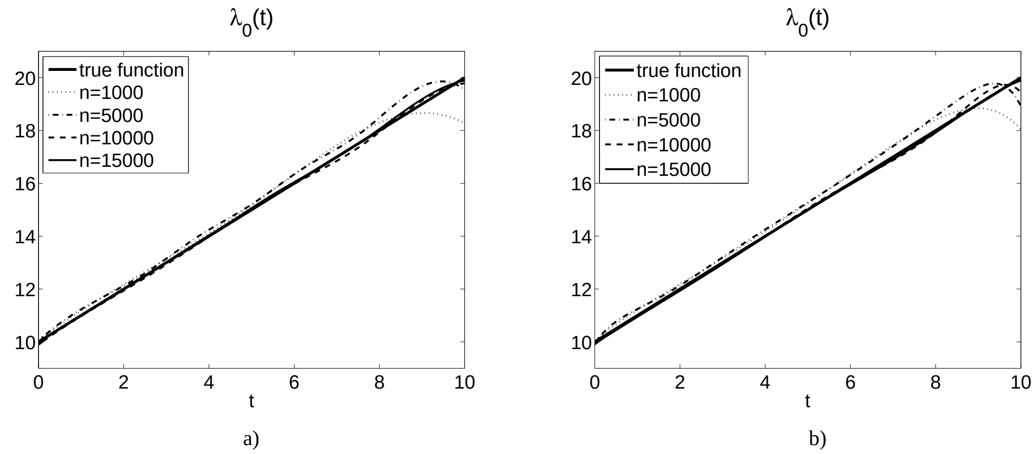


Figure 2: Estimates of $\lambda_0(t)$ based on a) spline approximation; b) Chebyshev approximation.

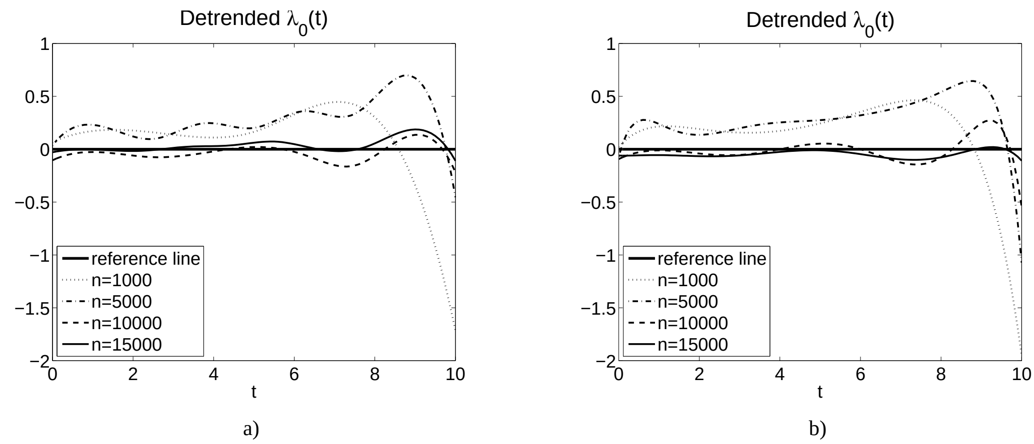


Figure 3: Detrended estimates of $\lambda_0(t)$ based on a) spline approximation; b) Chebyshev approximation.

Censor C is defined on the interval $[0, 10]$, and has a pdf of triangular shape, i.e.

$$f_C(x) := \begin{cases} x/50, & \text{if } x \in [0, 10]; \\ 0, & \text{otherwise.} \end{cases}$$

Using this settings approximately 25% of the lifetimes are censored. Sample size varies between 1000 and 15000 with steps of 1000 and the number of nodes $m := \lfloor \log(n) + 0.5 \rfloor$, that is the logarithm of the sample size rounded to the nearest integer.

Figures 1a, 1b and 1c show the means, standard deviations and MAEs of the estimated values of the parameter β for both approximation methods, plotted versus the sample size. These figures clearly show the convergence of the estimator to the true parameter value and also show a slight advantage of the spline interpolation.

Concerning the other parameter, Figures 2a and 2b show the estimates of $\lambda_0(t)$ based on spline and Chebyshev approximation, respectively, for four different sample sizes, while on Figures 3a and 3b the detrended estimates, i.e. the deviations from the true parameter function, are given. Observe, that Chebyshev approximation gives slightly better result which is more clearly observable on Figure 4, where the deviations in supremum norm of the estimates from the true $\lambda_0(t)$ are plotted versus the sample size.

Example 5.2. Consider the settings of Example 5.1 but assume that $m := \lfloor n^{1/3} + 0.5 \rfloor$, that is the cube root of the sample size rounded to the nearest integer. In this case number m of nodes increases with the increase of the sample size more drastically than in Example 5.1, so e.g. for sample size $n = 15000$ the order of the approximating polynomial in Chebyshev approximation is 24. However, the high order (in practice orders higher than 20 should be avoided) induces some extra fluctuation in the Chebyshev approximation and the results became worse than the results for smaller sample sizes. In this way large deviations of $\lambda_0(t)$ from its estimator might be consequences of the error of the approximation.

Naturally, for spline approximation the higher the number of nodes, the better results we obtain. For this reason in the present example we consider only the results of the spline approximation.

Figure 5a shows the means, while 5b the standard deviations and the MAEs of estimates of β. On Figures 6a and 6b the estimates of λ_0 and their detrended versions are plotted for four different sample sizes, while Figure 7 shows deviations in supremum norm of the estimates from the true baseline hazard. Comparing Figures 6a, 6b and 7 with Figures 2a, 3a and 4, respectively, one can clearly see the advantage of the increase of the nodes of spline approximation.

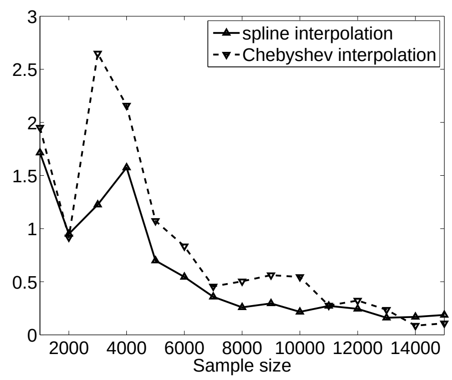
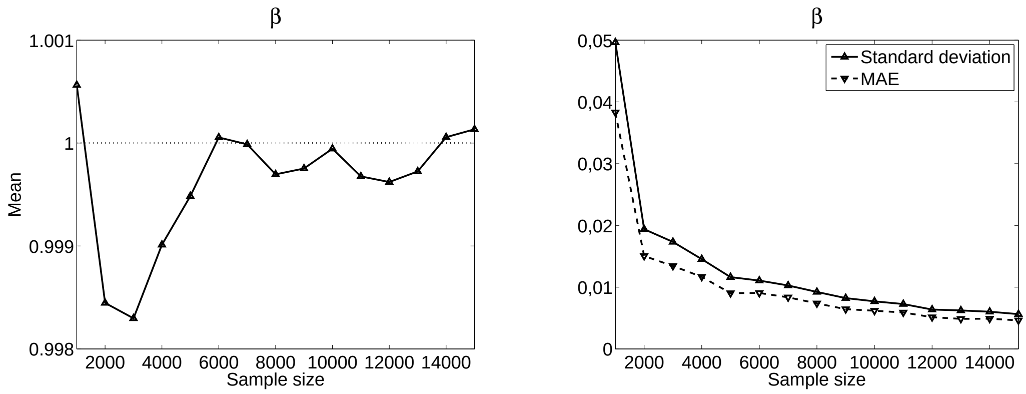


Figure 4: Deviations in supremum norm of the estimates of $\lambda_0(t)$ from the true function.

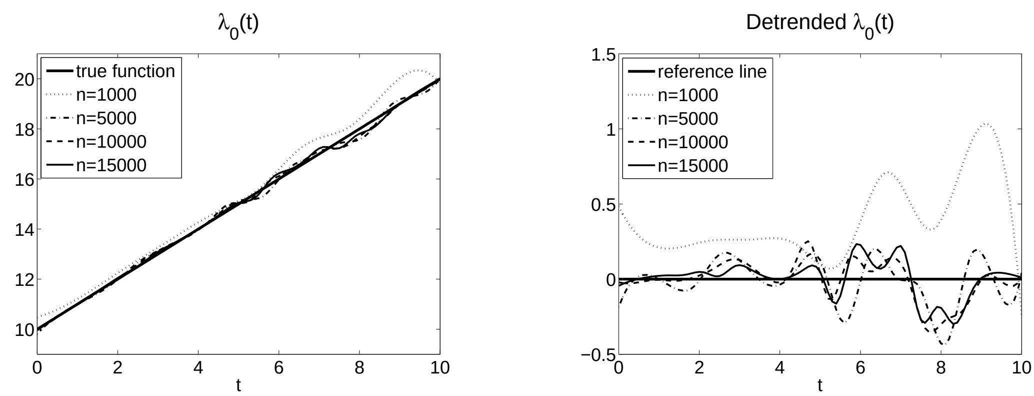


a) b)

Figure 5: Means, standard deviations and MAEs of the estimates of β.

Example 5.3. Let the covariates X, measurement errors U and censors C be the same as in Example 5.1, $\beta_0 = 1$ but now we have a Weibull hazard function, that is $\lambda_0(t) = 3/2t^2$. In this case similarly to the previous examples approximately 25% of the lifetimes are censored. Sample size varies between 1000 and 10000 with steps of 500 and the number of nodes $m := \lfloor n^{1/3} + 0.5 \rfloor$. We remark that compared to case of the linear baseline hazard, the optimization algorithm used considerably fewer steps to find the optimal points of the objective function.

Again, on Figure 8a the means, while on Figure 8b the standard deviations and the MAEs of estimates of β are plotted against the sample size. Here one can clearly observe the convergence



a) b)

Figure 6: Estimates and detrended estimates of $\lambda_0(t)$

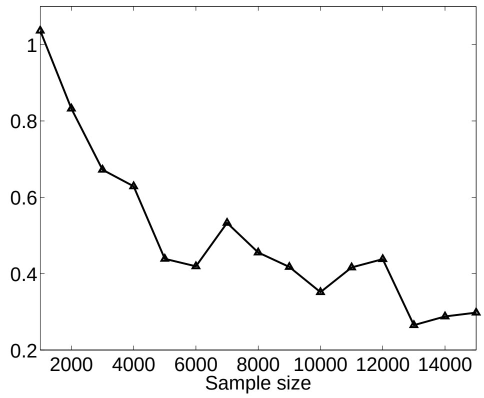


Figure 7: Deviations in supremum norm of the estimates of $\lambda_0(t)$ from the true function.

of the estimator $\widehat{\beta}_n$. Further, Figures 9a and 9b show the estimates of $\lambda_0(t)$ and their detrended versions for four different sample sizes, while on Figure 10 one can see the deviations in supremum norm of the estimates from the true baseline hazard.

6 Conclusion

We dealt with Cox proportional hazards model under censoring and measurement error and proved the consistency of simultaneous estimators of the baseline hazard function and regression parameters. The estimators are constructed via maximization over the infinite-dimensional compact set.

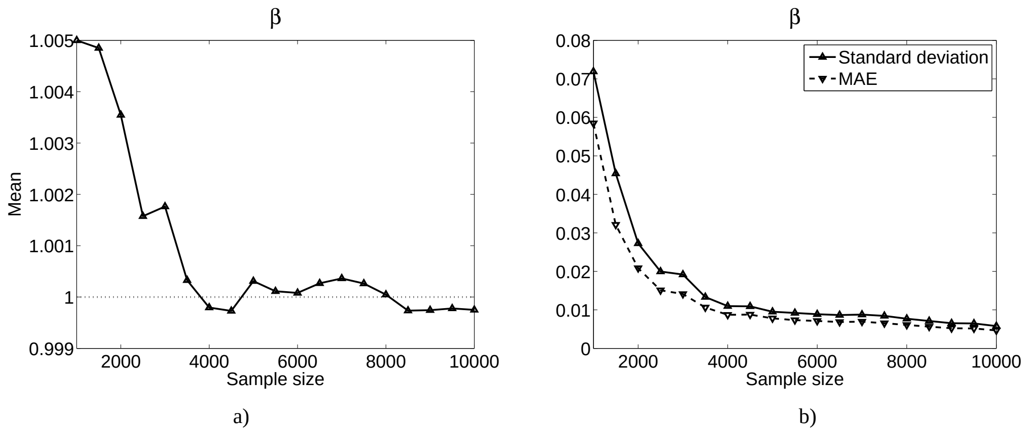
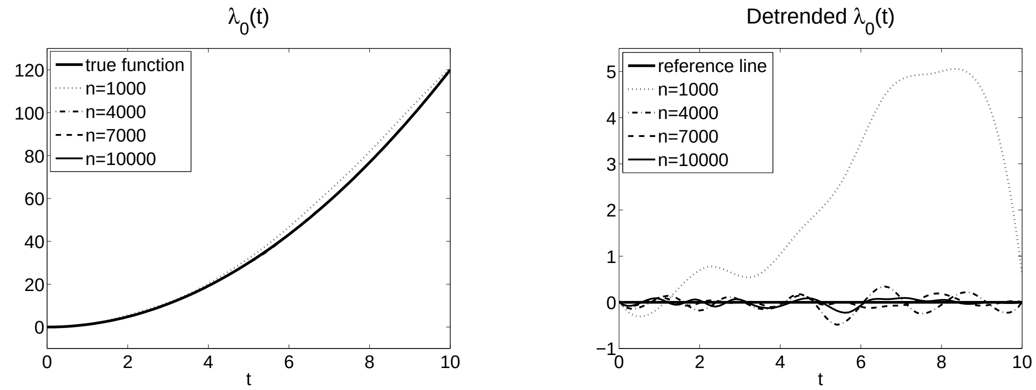


Figure 8: Means, standard deviations and MAEs of the estimates of β.



a) b)

Figure 9: Estimates and detrended estimates of $\lambda_0(t)$

The recurrent algorithms for searching the maximum points are contained in [2] and can be applied in this case.

In future research we intend to elaborate specific numerical methods to compute the estimates. Also we intend to give the rate of convergence in terms of deviation of the estimators from the true values and show the asymptotic normality of some linear functionals of the estimators.

Besides this based on our simulation results we are also going to check the consistency for the regression spline or polynomial interpolation estimators of λ.

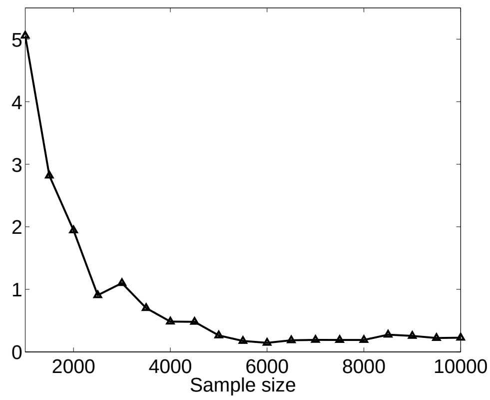


Figure 10: Deviations in supremum norm of the estimates of $\lambda_0(t)$ from the true function.

Acknowledgements

Research has been supported by the Hungarian Scientific Research Fund under Grant No. OTKA T079128/2009 and partially supported by TÁMOP 4.2.1./B-09/1/KONV-2010-0007/IK/IT project. The project is implemented through the New Hungary Development Plan co-financed by the European Social Fund, and the European Regional Development Fund.

References

[1] Ash, R. B. (1965). *Information Theory.* Interscience Publishers, New York.

[2] Aubin, J.-P. and Ekeland, I. (1984). *Applied Nonlinear Analysis.* Wiley, New York.

[3] Augustin, T. (2004). An exact corrected log-likelihood function for Cox's proportional hazards model under measurement error and some extensions. *Scandinavian Journal of Statistics* **31**, *43–50.*

[4] Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, *89–99.*

[5] Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society. Series B* **34**, *187–220.*

[6] Dudley, R. M. and Strassen, V. (1969). The central theorem and ε-entropy. *Lecture Notes in Mathematics* **89**, *224-231.*

[7] Kay, R. (1977). Proportional hazard regression models and the analysis of censored survival data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **26**, *227–237.*

[8] Kukush, A., Schneeweiss, H. and Wolf, R. (2005). Relative efficiency of three estimators in a polynomial regression with measurement errors. *Journal of Statistical Planning and Inference* **27**, *179–203*

[9] Potapov, V. N. (1997). ε-Entropy of compact sets in C and tabulation of continuous functions. *Siberian Mathematical Journal* **38**, *759–771.*

[10] Stefanski, L. A. (1989). Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Communications in Statistics – Theory and Methods* **18** *4335–4358.*

[11] Zinn, J. (1977). A note on the central limit theorem in Banach spaces. *The Annals of Probability* **5**, *283–286.*