Harvard University
Harvard University Biostatistics Working Paper Series

Year 2003 *Paper* 11

Empirical and Kernel Estimation of Covariate Distribution Conditional on Survival Time

Xiaochun Li* Ronghui Xu†

*Harvard School of Public Health and Dana Farber Cancer Institute, xiaochun@jimmy.harvard.edu

†Harvard School of Public Health and Dana Farber Cancer Institute, rxu@hsph.harvard.edu

http://biostats.bepress.com/harvardbiostat/paper11

Empirical and Kernel Estimation of Covariate Distribution Conditional on Survival Time

Xiaochun Li and Ronghui Xu
Department of Biostatistics
Harvard School of Public Health and Dana-Farber Cancer Institute
Boston, MA 02115, USA.
xiaochun@jimmy.harvard.edu rxu@jimmy.harvard.edu

October 28, 2003

Abstract

In biomedical research there is often interest in describing covariate distributions given different survival groups. This is not immediately available due to censoring. In this paper we develop an empirical estimate of the conditional covariate distribution under the proportional hazards regression model. We show that it converges weakly to a Gaussian process and provide its variance estimate. We then apply kernel smoothing to obtain an estimate of the corresponding density function. The density estimate is consistent and has the same rate of convergence as the classical kernel density estimator. We have developed an R package to implement our methodology, which is demonstrated through the Mayo Clinic primary biliary cirrhosis data.

Note: The figures in this paper are also available in color.

1 Introduction

Recently with the advancement of biomedicine, there has been an increasing number of long-term survivors in some of the disease areas such as cancer. In the organizations we work in, clinicians are often interested in describing patient characteristics according to different survival groups. For example, in many types of cancer five years is considered long-term with regard to survival, we are then interested describing the covariate distribution among long-term survivors versus short-term survivors. Our work in this paper was motivated by collaborative projects such as predicting long-term survival in chemonaive patients with advanced non-small cell lung cancer treated by standard chemotherapy, using the Eastern Cooperative Oncology Group database (not yet published). Such knowledge can potentially help clinical decision making if we can predict a patient's survival based on the characteristics at study entry. Although regression models are almost always used to directly identify important prognostic factors, graphics is often more conducive for clinicians to examine and to understand how the covariate distributions vary for different groups of survivors. This type of covariate analysis is commonly done in studies with categorical outcomes. But in studies where the categorization of the outcome is based on time to event, the conditional covariate distributions are not immediately available when there is censoring. In this paper we study both the empirical and smooth estimates of the conditional covariate distribution given survival for this purpose.

Xu and O'Quigley (2000) developed an estimate of survival probabilities given any range of the covariates under the proportional hazards regression model (Cox, 1972). As an intermediate step they derived an empirical estimate of the covariate distribution given any fixed survival time. Here we will further develop the estimator, both to show that the estimated distribution function converges weakly to a Gaussian process and to give its estimated

variance. In addition, we apply kernel smoothing to obtain the density function when the covariate is continuous. We show that the estimated density is consistent and has the same convergence rate as for the classical kernel density estimate.

The rest of the paper is organized as following. In the next section we present both the empirical estimate of the conditional covariate distribution and the smooth estimate of the corresponding density. Asymptotic results are given for both estimators. We illustrate the application of our estimators in Section 3, using the Mayo Clinic primary biliary cirrhosis data set. Section 4 contains some further discussion. Proofs of the asymptotic results as well as some details about the software are given in the appendix.

2 Estimating the conditional distribution of Z given T

In a survival study with n subjects, let $T_1, T_2, ..., T_n$ be the potential failure times, and $C_1, C_2, ..., C_n$ be the potential censoring times for the individuals $i = 1, 2, ..., n$. Let $X_i = \min(T_i, C_i)$, and $\delta_i = I(T_i \leq C_i)$. Define also $Y_i(t) = I(X_i \geq t)$. It is assumed that the failure time of each subject is related to its covariates, or explanatory variables, Z_i, $i = 1, 2, ..., n$. We assume (T_i, C_i, Z_i), $i = 1, 2, ..., n$, to be a random sample from the joint distribution of (T, C, Z). The proportional hazards model (Cox, 1972) postulates a simplified form for the relationship between the hazard function $\lambda(\cdot)$ for a subject at time t and the observed value of the covariate Z as

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta' Z\}, \tag{1}$$

where $\lambda_0(t)$ is a fixed 'baseline' hazard function, and β is the unknown log relative risk parameter. Generally in model (1) the covariate Z is a $p \times 1$ vector. Statistical inference on β is traditionally carried out through maximizing the partial likelihood, and we denote $\hat{\beta}$ the maximum partial likelihood estimate of β.

2.1 Empirical estimate

Although the Cox model specifies the distribution of T given the covariates, we have at any fixed time t two different conditional distributions of Z on T that are relevant. One is conditioning on $T \geq t$, which can be interpreted as the covariate distribution among all the subjects that are alive at time t, and can be readily estimated by the empirical distribution of Z in the risk set at time t if the censoring is independent of T and Z. The independent censoring assumption is often satisfied in randomized controlled clinical trials, such as in the applications that we are interested in, and is often satisfied in prognostic factors studies. Extensions to conditional independent censoring is described in the last section. Another conditional distribution of Z is that given $T = t$. Under the assumption that T has a continuous distribution we usually observe only one failure at a time and it is difficult to estimate this latter conditional distribution based on a single observation, or a few in the case of ties. We can, however, obtain a consistent estimate by leaning on the model.

Denote

$$\pi_i(\beta, t) = \frac{Y_i(t)\exp(\beta' Z_i)}{\sum_{j=1}^n Y_j(t)\exp(\beta' Z_j)}. \tag{2}$$

Then the product of the $\pi_i(\beta, X_i)$'s over the observed failure times gives the partial likelihood. Xu and O'Quigley (2000) showed that for any fixed time t, $\{\pi_i(\hat{\beta}, t)\}_{i=1}^n$ provides a consistent estimate of the conditional distribution of Z given $T = t$ under model (1). More precisely, we have

$$\widehat{F}(z|t) = \hat{P}(Z \leq z | T = t) = \sum_{\{j: Z_j \leq z\}} \pi_j(\hat{\beta}, t). \tag{3}$$

When Z is multi-dimensional, '$Z \leq z$' in the above means component-wise less than or equal to. In practice, we might be more interested in estimating the covariate distribution given

$a < T \leq b$, where $0 \leq a < b \leq \infty$. In this case we have

$$\hat{F}(z|a < T \leq b) = \frac{\int_a^b \hat{F}(z|t)d\hat{F}(t)}{\hat{F}(b) - \hat{F}(a)}, \tag{4}$$

where $\hat{F}(\cdot)$ is a consistent estimate of the marginal distribution function of T, such as the Kaplan-Meier (1958) estimate under the independent censoring assumption.

Theorem 1 *$\sqrt{n}\{\hat{F}(\cdot|t) - F(\cdot|t)\}$, where $\hat{F}(\cdot|t)$ is defined in (3), converges weakly to a zero-mean Gaussian process, whose variance and covariance can be estimated as given in Appendix A.*

Theorem 2 *$\sqrt{n}\{\hat{F}(\cdot|a < T \leq b) - F(\cdot|a < T \leq b)\}$, where $\hat{F}(\cdot|a < T \leq b)$ is defined in (4), converges weakly to a zero-mean Gaussian process, whose variance and covariance can be estimated as given in Appendix A.*

When Z is discrete, suppose that z is one of the mass points. Then

$$\hat{p}(z|t) = \hat{P}(Z = z|T = t) = \frac{\sum_{j=1}^n Y_j(t) \exp(\hat{\beta}' Z_j) I(Z_j = z)}{\sum_{j=1}^n Y_j(t) \exp(\hat{\beta}' Z_j)}. \tag{5}$$

Similar to (4) we also have

$$\hat{P}(Z = z|a < T \leq b) = \frac{\int_a^b \hat{p}(z|t)d\hat{F}(t)}{\hat{F}(b) - \hat{F}(a)}, \tag{6}$$

Results similar to the above theorems hold for (5) and (6); see appendix.

2.2 Kernel smooth estimate

When Z is continuous, formula (3) gives $\widehat{F}(z|t)$, an estimate of $F(z|t)$, which can be rather jagged as it has jumps at distinct Z_j's. In general $\widehat{F}(\cdot|t)$ gets rougher as t gets larger because the risk set becomes smaller; see Figure 1 (a).

It is hard to perceive the data structure via the plot of $\widehat{F}(\cdot|t)$. Even if the covariate data has stochastic ordering given the ordered survival times, the conditional distributions are

not intuitive and hard to illustrate to medical researchers. In contrast, densities are usually much more informative visually.

The density function of $P(Z \leq z|T = t)$ is

$$f_t(z) \equiv f(z|t) = \frac{\exp(\beta' z)h_t(z)}{\int \exp(\beta' u)h_t(u)\,du},$$

where $h_t(z)$ is the conditional density of Z given $T \geq t$ (Xu and O'Quigley, 2000). Hence to estimate $f_t(z)$, we need only to estimate $h_t(z)$. The conditional cumulative distribution function corresponding to $h_t(z)$ can be consistently estimated by $\sum_{Z_j \leq z} Y_j(t) / \sum_1^n Y_j(t)$, the empirical distribution of the covariates in the risk set at time t. Therefore one way to obtain a smoothed estimate of $f_t(z)$ is to use a kernel density estimator for $h_t(z)$.

In classic one-dimensional kernel density estimation (cf. Fan and Gijbels, 1996) we have n observations, $Z_1, Z_2, \ldots, Z_n$. A kernel density estimate is defined as:

$$\widehat{f}(z) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{z - Z_i}{b}\right), \tag{7}$$

where K is a unimodal density function and b is a positive number. In the smoothing literature, K is known as the kernel function and b is a smoothing parameter called bandwidth. Bandwidth b controls the amount of smoothing by weighting the Z_i's for the estimation of the density at z. If K has a compact support, b also controls how many Z_i's around z to be included in the estimation of $f(z)$. Although K does not need to be a function with compact support, certain regularity conditions need to be imposed on the tails of K so that remote points from z have a nearly diminished effect on $\widehat{f}(z)$.

Since we want to smooth $h_t(z)$, the covariate density function conditional on $T \geq t$, the sample size is $\sum_1^n Y_i(t)$ instead of n in (7). We shall define the kernel estimate analogously,

$$\widehat{h}_t(z) = \frac{1}{b \sum_1^n Y_i(t)} \sum_{i=1}^{n} Y_i(t) K\left(\frac{z - Z_i}{b}\right). \tag{8}$$

Therefore, the conditional density $f_t(z)$ is estimated by

$$\widehat{f}_{t,\hat{\beta}}(z) \equiv \frac{e^{\hat{\beta}z}\widehat{h}_t(z)}{\int e^{\hat{\beta}u}\widehat{h}_t(u)\,du}\ , \tag{9}$$

where $\hat{\beta}$ is the maximum partial likelihood estimate. If K is chosen to be the normal kernel, i.e. density of $N(0,1)$, (9) becomes

$$\widehat{f}_{t,\hat{\beta}}(z) = \frac{\exp(\hat{\beta}z)\sum_1^n K\{(z-Z_i)/b\}\,Y_i(t)}{b\exp\{(b\hat{\beta})^2/2\}\sum_1^n \exp(\hat{\beta}Z_i)Y_i(t)}. \tag{10}$$

Let $R(K) = \int K^2(s)\,ds$ and $\sigma_K^2 = \int K(s)s^2\,ds$, where $K(\cdot)$ is the kernel used to estimate $f(z|t)$. The following theorem summarizes the asymptotic property of $\widehat{f}_{t,\hat{\beta}}(z)$ defined in (9).

Theorem 3

$$\widehat{f}_{t,\hat{\beta}}(z) - f_t(z) = c_1 b^2 + \frac{c_2}{\sqrt{nb}} + o(b^2) + o_p(\frac{1}{\sqrt{nb}}), \tag{11}$$

where

$$c_1 = \frac{e^{\beta z}\sigma_K^2(h_t^{(2)}(z) - \beta^2 h_t(z))}{2\int e^{\beta u}h_t(u)\,du} \tag{12}$$

$$c_2 = \frac{e^{\beta z}}{\int e^{\beta u}h_t(u)\,du}\left(\frac{h_t(z)R(K)}{C_p}\right)^{1/2} \tag{13}$$

and $C_p = P(C > t)P(T > t)$, *provided that the following hold:*

1. *kernel* K *is a symmetric density function with*

$$R(K) < \infty \qquad \textit{and} \qquad 0 < \sigma_K^2 < \infty; \tag{14}$$

2. $h_t^{(2)} < \infty$;

3. $b \to 0$ *and* $nb \to \infty$;

4. *censoring* C *is independent of* T *and* Z;

5. *regularity conditions such that* $\sqrt{n}(\hat{\beta} - \beta)$ *converges in distribution to normal with mean 0 and finite variance;*

6. *covariate* Z *is bounded.*

Remark : By Theorem 3, we know that $\widehat{f}_{t,\hat{\beta}}(z)$ is consistent and that the best attainable rate for convergence is still $n^{-4/5}$ in mean squared errors, as for the classic kernel density estimate. Note that the asymptotic variance of $\widehat{f}_{t,\hat{\beta}}(z)$ for a given t has an inverse relationship with the constant $C_p = P(C > t)P(T > t)$, which tells us that the variance of the estimate is large when t is large, namely, when the risk set is small.

Definition (8) generalizes in a straightforward way to multiple dimensions in the case of continuous covariates:

$$\widehat{h}_t(z) = \frac{1}{b_1 \dots b_p \sum_1^n Y_i(t)} \sum_{i=1}^{n} Y_i(t) \prod_{j=1}^{p} K\left(\frac{z_j - Z_{ij}}{b_j}\right), \tag{15}$$

where $Z_i = (Z_{i1}, \dots, Z_{ip})'$, $z = (z_1, \dots, z_p)'$ and b_j is the bandwidth for the j-th dimension. Note that the same univariate kernel is used in each dimension. We also note that in practice the application of high dimensional estimator will be very limited, as discussed later. For $p = 2$, if K is the density of $N(0, 1)$, (9) becomes

$$\widehat{f}_{t,\hat{\beta}}(z) = \frac{\exp(\hat{\beta}'z) \sum_1^n K\{(z_1 - Z_{i1})/b_1\} K\{(z_2 - Z_{i2})/b_2\} Y_i(t)}{b_1 b_2 \exp[\{(b_1\hat{\beta}_1)^2 + (b_2\hat{\beta}_2)^2\}/2] \sum_1^n \exp(\hat{\beta}' Z_i) Y_i(t)}, \tag{16}$$

where $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)'$. Formulae can also be easily extended to covariates of mixed types; we shall not pursue this further here.

The smoothed conditional density of Z given $a < T \leq b$ is

$$\frac{\int_a^b \widehat{f}_{t,\hat{\beta}}(z) d\hat{F}(t)}{\hat{F}(b) - \hat{F}(a)} = \frac{\sum_{t_k \in (a,b]} \widehat{f}_{t_k,\hat{\beta}}(z) \Delta\hat{F}(t_k)}{\hat{F}(b) - \hat{F}(a)}, \tag{17}$$

where $\Delta\hat{F}(t_k)$'s are jumps at t_k's if $\hat{F}$ is the Kaplan-Meier estimator.

3 Application and Software

We shall demonstrate our methodology using the data set of a trial in primary biliary cirrhosis (*PBC*) of the liver from Mayo Clinic. The *PBC* data set is described and analyzed in Fleming and Harrington (1991). This is a double-blinded randomized trial comparing D-penicillamine with placebo. There are a total of 418 patients in the data set, with 161 deaths. It turned out that the treatment had a negligible effect on prognosis, so that the two arms may be combined for study of the natural history of the disease. Detailed analysis in Fleming and Harrington (1991) identified several important prognostic factors. Here for illustration purposes we consider four of them: *age, log(albumin), edema* and *log(bilirubin).* Normal kernels are used below for the estimation of conditional densities.

Under a proportional hazards model, the relative risk of death for age (in years) is 1.04 (p-value < 0.0001). That is, an additional year in age is associated with 4% increase in the relative risk. Figure 1 (b) shows that as years of survival increase, the distribution of age shifts towards younger. Figure 1 (c) shows the age distributions for patients who survived less than 5 years versus those who survived greater than 5 years. Figure 1 (d) shows that the age distributions for those who survived less than 2 years and those who survived between 2 and 5 years are, however, similar. Figure 1 (b) also shows that age distributions are quite different between the longest survival group and the rest. These seem to indicate that age might have a bigger impact in distinguishing long versus short term survivors, than in different groups of short term survivors.

In a similar way Figure 2 shows that longer survivors tend to have higher albumin levels (in gm/dl, on the log-scale), with relative risk 0.0082 (p-value < 0.0001). This is consistent with the fact that lower than normal levels of albumin in blood indicates dysfunction of liver. Notice that the distributions of *log(albumin)* appear nearly equally distinct for different

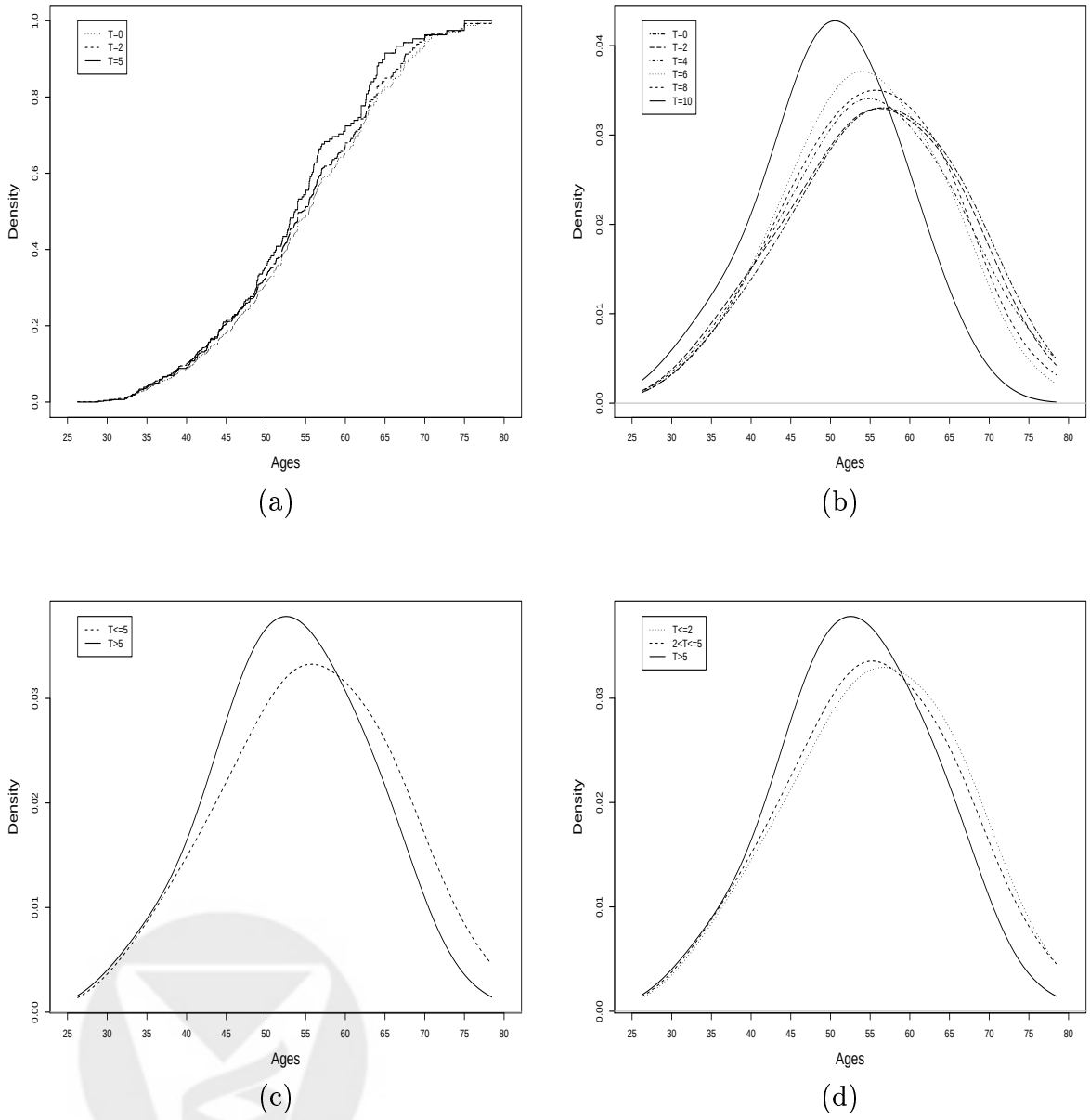
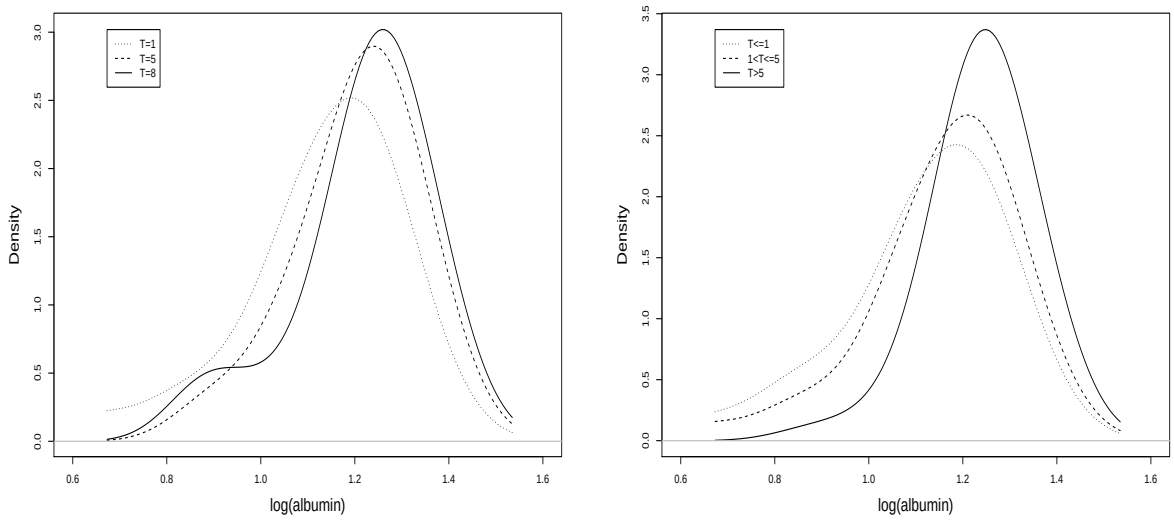


Figure 1: *Age distribution conditional on survival time, bandwidth* $b = 5$ *for (b)-(d).*

groups of survivors as shown in the figures.



(a) (b)

Figure 2: *Log(albumin) distribution conditional on survival time, bandwidth* $b = 0.08$.

The variable *edema* has three levels: 1) no edema and no diuretic therapy for edema, 2) edema present for which no diuretic therapy was given, or edema resolved with diuretic therapy, and 3) edema despite diuretic therapy. The relative risks are 10.3 between levels 3 and 1 and 2.54 between levels 2 and 1 (p-value < 0.0001). Figure 3 shows shifts in proportions of the three categories as survival increases, indicating the adverse association of severity of edema and survival.

Figure 4 shows the clear inverse relationship between bilirubin levels (in mg/dl, on the log-scale) and survival. The relative risk here is 2.69 (p-value < 0.0001).

Finally Figure 5 shows the contours of the joint distributions of *log(albumin)* and *log(bilirubin)* at survival times 1 year (a) and 5 years (b). The mode of the joint density at 5 years has moved towards the lower right corner as compared to 1 year, indicating higher albumin as well as lower bilirubin among longer survivors. Notice that the same trend was observed in

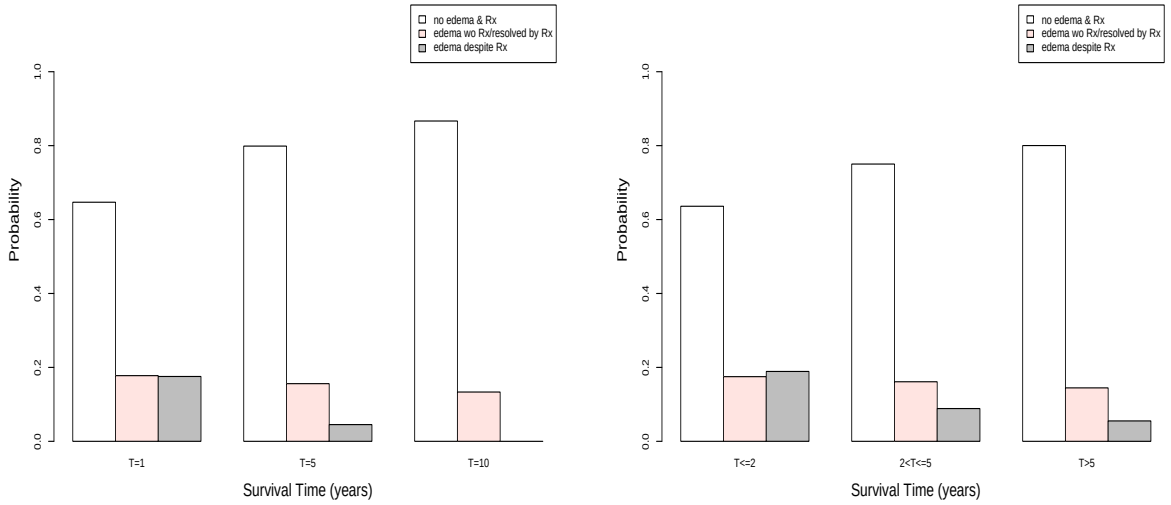


Figure 3: *Edema distribution conditional on survival time.*

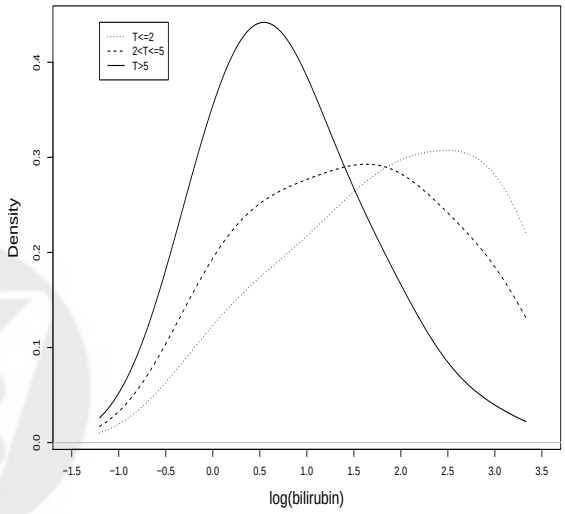
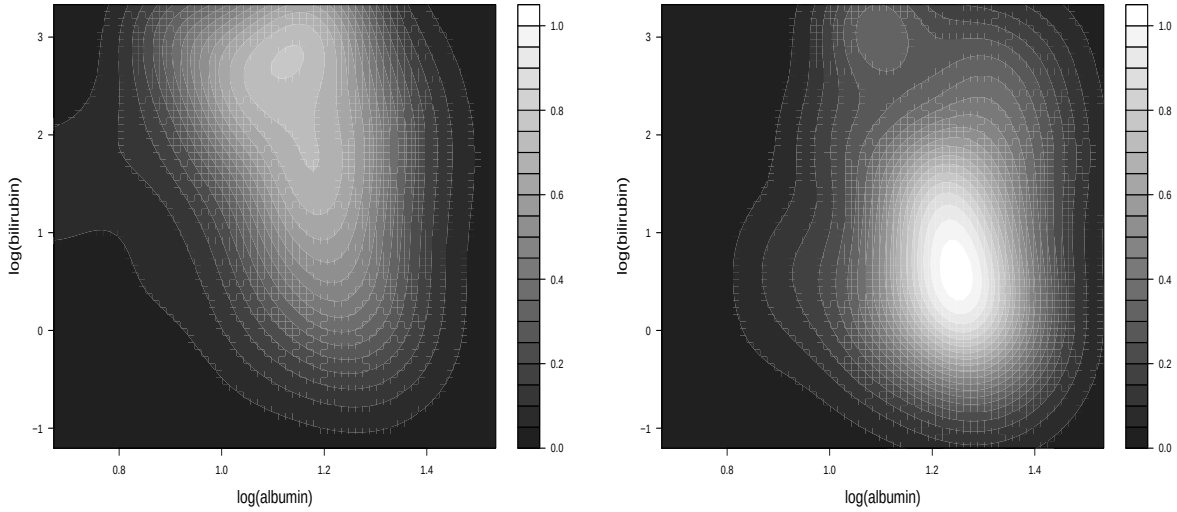


Figure 4: *Log(bilirubin) distribution conditional on survival time, bandwidth* $b = 0.5$.

the univariate estimates. Here we also see that the conditional distribution of covariates at a later time has smaller spread than that at an earlier time point, which appears consistent with the known fact that the population at risk becomes more homogeneous over time. The same was also observed in the univariate estimates.



(a) T=1 (year) (b) T=5 (years)

Figure 5: *Joint distribution of Log(albumin) and Log(bilirubin) conditional on survival time, bandwidth* $b = (0.08, 0.5)$.

The above analysis was done using an *R* package *SurvCov* that we have developed, available from *http://biowww.dfci.harvard.edu/~xiaochun*. The *PBC* data set is given as an example contained in the package. Bandwidths for the estimation of the conditional densities are chosen through 'trial and error'; see Appendix C for more details.

4 Discussion

We have considered estimating covariate distributions given survival, which was motivated by our applied work. We studied the asymptotic properties of the estimators and have

developed software to implement the estimates. In the illustrations we did not plot the pointwise confidence intervals of the estimated curves, since they are often of less interest in exploratory analyses; but they certainly are available following the asymptotic results. The model we have considered is the most commonly used Cox proportional hazards model. However, as shown in Xu and O'Quigley (2000) their *Theorem 1* also applies to other types of multiplicative hazards models, such as the time-varying regression effects $\beta(t)$ model (Xu and Adak, 2002). The methods of this paper can then be applied to these other types of models if needed.

In the above we have assumed that C is independent of T and Z. While this is often satisfied in randomized controlled clinical trials and prognostic factors studies where the methods developed here are likely to be used, under the Cox model we sometimes relax the assumption to conditional independence of C and T given Z. In this latter case following Xu and O'Quigley (2000) if we can discretize Z into a finite number of categories, the conditional distribution of Z given $T = t$ is $f(z|t) = \exp(\beta' z)S(t|z)g(z)/\sum_s \exp(\beta' s)S(t|s)g(s)$, where the summation is over the categories of Z, $S(t|z)$ is the conditional survival probability given z and $g(z)$ is the probability of category z. We can estimate $S(t|z)$ by either the subgroup Kaplan-Meier estimate within the category of value z or the predicted survival probabilities under the Cox model, and estimate the marginal failure time distribution $F(\cdot)$ in (4) and (6) by the weighted Kaplan-Meier estimate of Murray and Tsiatis (1996). For density estimate in this case we then smooth the marginal distribution $g(\cdot)$ of Z.

Finally although there is no apparent mathematical difficulties in applying our methodology to multiple dimensions, the use of the estimates and their graphical displays are limited beyond two dimensions for the purposes of this paper. Aside from the curse of dimensionality, finding a graphical display so that the details of the densities are not obscured by the dimensionality is in itself a challenge. On the other hand, the trend as well as other

characteristics observable in the univariate conditional distributions appear quite adequate for practical applications in our opinion.

Appendix

A Asymptotics of the empirical estimates

1. Estimator $\hat{F}(z|t)$

Under the usual regularity and continuity conditions for the proportional hazards model Andersen and Gill (1982), it is straightforward to show that $\hat{F}(z|t)$ is a consistent estimate of $F(z|t)$. We now show that $\hat{F}(z|t)$ is asymptotically equivalent to a sum of i.i.d. random variables. Note that $\hat{\beta}-\beta = I^{-1}(\check{\beta})U(\beta)$, where $\check{\beta}$ is on the line segment between $\hat{\beta}$ and the 'true' β under model (1), and $U(\cdot)$ and $I(\cdot)$ are the score function and the negative second derivative of the log partial likelihood, respectively. In the following $\dot{\beta}$ is also on the line segment between $\hat{\beta}$ and β. We have

$$\begin{aligned}
\sqrt{n}\hat{F}(z|t) &= \sqrt{n}\hat{F}_{\beta}(z|t) + \sqrt{n}(\hat{\beta}-\beta)'\frac{\partial \hat{F}(z|t)}{\partial \beta}\Big|_{\dot{\beta}} \\
&= \frac{n^{-1/2}\sum Y_j(t)\exp(\beta' Z_j)I(Z_j \le z)}{n^{-1}\sum Y_j(t)\exp(\beta' Z_j)} + \sqrt{n}(\hat{\beta}-\beta)'\left[\frac{\sum Y_j(t)Z_j\exp(\dot{\beta}' Z_j)I(Z_j\le z)}{\sum Y_j(t)\exp(\dot{\beta}' Z_j)}\right. \\
&\quad \left. - \frac{\sum Y_j(t)\exp(\dot{\beta}' Z_j)I(Z_j\le z)\cdot \sum Y_j(t)Z_j\exp(\dot{\beta}' Z_j)}{\{\sum Y_j(t)\exp(\dot{\beta}' Z_j)\}^2}\right] \\
&= n^{-1/2}\sum_{j=1}^{n}\left\{A(t)\cdot Y_j(t)\exp(\beta' Z_j)I(Z_j\le z) - \frac{B(t)'}{I_0}U_j(\beta)\right\} + o_p(1),
\end{aligned}$$

where $A(t) = \{s_0(\beta,t)\}^{-1}$, $B(t) = E\{Y(t)Z\exp(\beta' Z)I(Z\le z)\}/s_0(\beta,t) - E\{Y(t)\exp(\beta' Z)I(Z\le z)\}s_1(\beta,t)/s_0(\beta,t)^2$, $s_0(\beta,t) = E\{Y(t)\exp(\beta' Z)\}$, $s_1(\beta,t) = E\{Y(t)Z\exp(\beta' Z)\}$, I_0 is the expected information for β under the Cox model, and $U_j(\beta) = \int\{Z_j - s_1(\beta,t)/s_0(\beta,t)\}dM_j(t)$ with $M_j(T) = I(T_j\le t, T_j\le C_j) - Y_j(t)\exp(\beta' Z_j)\Lambda_0(t)$ and $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ (Andersen

and Gill, 1982).

Therefore $\sqrt{n}\{\hat{F}(\cdot|t) - F(\cdot|t)\}$ converges weakly to a Gaussian process, whose variance and covariance can be estimated by the sample variance and covariance of the above i.i.d random variables, with β, A, B, U_j and I_0 replaced by their sample-based estimates $\hat{\beta}$, $\hat{A}$, $\hat{B}$, $\widehat{U_j}$ and $\hat{I}_0$, respectively.

Similar result holds for $\hat{p}(z|t)$ with '$\leq z$' in the above replaced by '=z'.

2. Estimator $\hat{F}(z|a < T \leq b)$

The asymptotic calculation for $\hat{F}(z|a < T \leq b)$ is a bit more complex. First we have $\sqrt{n}\hat{F}(z|a < T \leq b) = \sqrt{n}\hat{F}_\beta(z|a < T \leq b) + \sqrt{n}(\hat{\beta} - \beta)'\partial\hat{F}(z|a < T \leq b)/\partial\beta|_{\ddot{\beta}}$, where $\ddot{\beta}$ is on the line segment between $\hat{\beta}$ and β. We can show that the second term is asymptotically equivalent to a sum of i.i.d. random variables like in the above. For the first term, although using the central limit theorem of Stute (1995) for Kaplan-Meier integrals we can still write it as sum of i.i.d. random variables, variance estimation following that is not straightforward since it involves functions of the unknown censoring distribution. Here we will use the empirical influence function instead. This approach was used in Reid and Crépeau (1985) and G. Knafl in an unpublished dissertation at the Northwestern University, as well as in Xu and Harrington (2001). The idea is to express the estimate as a functional of the empirical distribution of the data, and the true parameter as the same functional of the true distribution, and then find the Gâteaux derivative of this functional. Let $H_n(x, \delta, z)$ be the empirical distribution function of the triples (X_i, δ_i, Z_i), $i = 1, ..., n$, i.e. putting mass n^{-1} on each triple. Let $H(x, \delta, z)$ be the corresponding joint distribution function of (X, δ, Z). We also use $H(x, z)$ for the joint marginal distribution function of (X, Z), $H(x)$ for the marginal distribution function of X, and $H_n(x, z)$ and $H_n(x)$ for the corresponding empirical distribution functions.

The numerator in $\hat{F}_\beta(z|a < T \leq b)$ can be written $\sum_{i=1}^n \delta_i W(X_i)\hat{F}(z|X_i)I(a < X_i \leq b)$,

where $W(X_i)$ is the jump of the Kaplan-Meier curve at a failure time X_i. When there are no ties $W(X_i) = \hat{S}(X_i)/\{1 - H_n(X_i)\}$, where $\hat{S}(\cdot)$ is the Kaplan-Meier estimate of the marginal survival function of T. In practice ties may be split randomly. Then (4) as a functional of H_n (and S) can be written

$$g(H_n) = \int \delta \cdot \frac{\int_{\tilde{x} \geq x} e^{\beta' \tilde{z}} I(\tilde{z} \leq z) dH_n(\tilde{x}, \tilde{z})}{\int_{\tilde{x} \geq x} e^{\beta' \tilde{z}} dH_n(\tilde{x}, \tilde{z})} \cdot \frac{\hat{S}(x)}{1 - H_n(x)} I(a < x \leq b)\, dH_n(x, \delta, z) / \{\hat{S}(a) - \hat{S}(b)\}. \tag{18}$$

The population parameter for the above is

$$\begin{aligned} g(H) &= \int \delta \cdot \frac{\int_{\tilde{x} \geq x} e^{\beta' \tilde{z}} I(\tilde{z} \leq z) dH(\tilde{x}, \tilde{z})}{\int_{\tilde{x} \geq x} e^{\beta' \tilde{z}} dH(\tilde{x}, \tilde{z})} \cdot \frac{S(x)}{1 - H(x)} I(a < x \leq b)\, dH(x, \delta, z) / \{S(a) - S(b)\} \\ &\equiv g_{num}(H) / \{S(a) - S(b)\}. \end{aligned} \tag{19}$$

Following Reid (1981) $S(x) = \exp\{-\int_0^x dF^u(\tilde{x})/(1 - H(\tilde{x})\}$, where $F^u(x) = P(X \leq x, \delta = 1)$ is the sub-distribution function for uncensored data. Denote $1\{x_0, \delta_0, z_0\}$ the distribution function that puts unit mass at (x_0, δ_0, z_0). Taking the limit of $\epsilon^{-1}\{g[(1 - \epsilon)H + \epsilon 1\{x_0, \delta_0, z_0\}] - g(H)\}$ as $\epsilon \to 0$ we obtain

$$\begin{aligned} IC(x_0, \delta_0, z_0) - \mu &= \frac{\partial}{\partial \epsilon} g[(1 - \epsilon)H + \epsilon 1\{x_0, \delta_0, z_0\}]\Big|_{\epsilon = 0} \\ &= \frac{g'_{num}}{S(a) - S(b)} - g(H) \times \frac{S'(a) - S'(b)}{S(a) - S(b)}, \end{aligned} \tag{20}$$

where $IC(x, \delta, z)$ is the influence function, $\mu = \int IC(x, \delta, z) dH(x, \delta, z)$, and

$$\begin{aligned} g'_{num} &= -g_{num}(H) + \delta_0 \cdot \frac{\int_{\tilde{x} \geq x_0} e^{\beta' \tilde{z}} I(\tilde{z} \leq z_0) dH(\tilde{x}, \tilde{z})}{\int_{\tilde{x} \geq x_0} e^{\beta' \tilde{z}} dH(\tilde{x}, \tilde{z})} \cdot \frac{S(x_0)}{1 - H(x_0)} I(a < x_0 \leq b) \\ &\quad + \int \delta \left[\frac{I(x_0 \geq x, z_0 \leq z) e^{\beta' z_0}}{\int_{\tilde{x} \geq x} e^{\beta' \tilde{z}} dH(\tilde{x}, \tilde{z})} - \frac{\int_{\tilde{x} \geq x} e^{\beta' \tilde{z}} I(\tilde{z} \leq z) dH(\tilde{x}, \tilde{z})}{\int_{\tilde{x} \geq x} e^{\beta' \tilde{z}} dH(\tilde{x}, \tilde{z})} \times \left\{ \frac{I(x_0 \geq x) e^{\beta' z_0}}{\int_{\tilde{x} \geq x} e^{\beta' \tilde{z}} dH(\tilde{x}, \tilde{z})} \right. \right. \\ &\quad + \left. \left. \frac{\delta_0 I(x_0 \leq x)}{1 - H(x_0)} - \int_0^{x \wedge x_0} \frac{dF^u(\tilde{x})}{\{1 - H(\tilde{x})\}^2} - \frac{H(x) - I(x_0 \leq x)}{1 - H(x)} \right\} \right] \frac{S(x)}{1 - H(x)} I(a < x \leq b)\, dH(x, \delta, z), \\ S'(a) &= S(a) \left[-\frac{\delta_0 I(x_0 \leq a)}{1 - H(x_0)} + \int_0^{a \wedge x_0} \frac{dF^u(\tilde{x})}{\{1 - H(\tilde{x})\}^2} \right], \\ S'(b) &= S(b) \left[-\frac{\delta_0 I(x_0 \leq b)}{1 - H(x_0)} + \int_0^{b \wedge x_0} \frac{dF^u(\tilde{x})}{\{1 - H(\tilde{x})\}^2} \right]. \end{aligned}$$

It follows then that $\sqrt{n}\{\hat{F}(z|a<T\le b)-F(z|a<T\le b)\}$ is asymptotically normal with mean zero and variance that can be consistently estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\left[\{EIC(x_i,\delta_i,z_i)-\mu\}+\frac{\widehat{U_i}(\hat{\beta})}{\hat{I}_0}\frac{\int_a^b \hat{B}(t)d\hat{F}(t)}{\hat{S}(a)-\hat{S}(b)}\right]^2, \tag{21}$$

where $\hat{B}$, $\widehat{U_i}$ and $\hat{I}_0$ are the same as for $\hat{F}(z|t)$, $EIC()$ is the empirical influence function, and

$$EIC(x_i,\delta_i,z_i)-\mu=\frac{\widehat{g'_{num}}}{\hat{S}(a)-\hat{S}(b)}-\hat{F}(z|a<T\le b)\times\frac{\widehat{S'}(a)-\widehat{S'}(b)}{\hat{S}(a)-\hat{S}(b)}, \tag{22}$$

with

$$\begin{aligned}
\widehat{g'_{num}} &= -\int_a^b \hat{F}(z|t)d\hat{F}(t)+\delta_i\hat{F}(Z_i|X_i)W(X_i)I(a<X_i\le b)\\
&+ \sum_{j=1}^{n}\delta_j W(X_j)I(a<X_j\le b)\left[\frac{nI(X_i\ge X_j,Z_i\le Z_j)\exp(\hat{\beta}'Z_i)}{\sum_l Y_l(X_j)\exp(\hat{\beta}'Z_l)}-\hat{F}(Z_j|X_j)\right.\\
&\times\left\{\frac{nI(X_i\ge X_j)\exp(\hat{\beta}'Z_i)}{\sum_l Y_l(X_j)\exp(\hat{\beta}'Z_l)}+\frac{n\delta_i I(X_i\le X_j)}{\sum_l Y_l(X_i)}-\sum_{X_l\le(X_i\wedge X_j)}\frac{n\delta_l}{\{\sum_k Y_k(X_l)\}^2}\right.\\
&\left.\left.-\frac{n-\sum_l Y_l(X_j)-nI(X_i\le X_j)}{\sum_l Y_l(X_j)}\right\}\right],\\
\widehat{S'}(a) &= \hat{S}(a)\left[-\frac{n\delta_i I(X_i\le a)}{\sum_l Y_l(X_i)}+\sum_{X_l\le(X_i\wedge a)}\frac{n\delta_l}{\{\sum_k Y_k(X_l)\}^2}\right],\\
\widehat{S'}(b) &= \hat{S}(b)\left[-\frac{n\delta_i I(X_i\le b)}{\sum_l Y_l(X_i)}+\sum_{X_l\le(X_i\wedge b)}\frac{n\delta_l}{\{\sum_k Y_k(X_l)\}^2}\right].
\end{aligned}$$

Similarly we have the variance estimate of $\sqrt{n}\{\hat{p}(z|a<T\le b)-p(z|a<T\le b)\}$ as in the above with $\hat{F}$ replaced by $\hat{p}$, '$\le z$' replaced by '$=z$', and '$Z_i\le Z_j$' replaced by '$Z_i=Z_j$'. The covariances of the limiting Gaussian processes in z can also be estimated by the sample covariances of these i.i.d. random variables.

B Asymptotics of the kernel estimate

Lemma 1 *Let $C_p = P(C > t)P(T > t)$. Under the conditions in Theorem 3, the following holds*

$$A \equiv \frac{1}{nb}\sum_1^n K\left(\frac{z - Z_j}{b}\right) Y_j(t) = Q_0 + Q_1 b^2 + \frac{Q_2}{\sqrt{nb}} + o(b^2) + o_p(\frac{1}{\sqrt{nb}}), \tag{23}$$

where

$$\begin{aligned} Q_0 &= C_p\, h_t(z), & (24)\\ Q_1 &= \frac{C_p}{2}\, h_t^{(2)}(z)\sigma_K^2, & (25)\\ Q_2^2 &= C_p\, h_t(z) R(K). & (26) \end{aligned}$$

The proof is similar to that in classic kernel density estimation; for example see Scott (1992). By Chebyshev's inequality, we have $X = E(X) + O_p(\sqrt{Var(X)})$. The factor C_p comes from the fact that if we let $A_j = K\{(z - Z_j)/b\}Y_j(t)$,

$$\begin{aligned} E(\frac{1}{nb}\sum_1^n A_j) &= \frac{1}{b}P(C > t)\int K\left(\frac{z-u}{b}\right) P(Z = u, T > t)\, du \\ &= P(C > t)P(T > t)\int K(s) h_t(z - bs)\, ds. \end{aligned}$$

We can then apply a Taylor expansion to $h_t(z - bs)$ and finish the proof in the same way as in the classic setting. Similarly, we have

Lemma 2 *Under the conditions in Theorem 3, the following holds*

$$\begin{aligned} B \equiv \int e^{\beta z} A\, dz &= \frac{1}{nb}\sum_1^n Y_j(t)\int e^{\beta z} K\left(\frac{z - Z_j}{b}\right) dz \\ &= U_0 + U_1 b^2 + + o(b^2) + O_p(n^{-1/2}), \end{aligned} \tag{27}$$

where

$$U_0 = C_p \int e^{\beta u} h_t(u)\, du, \tag{28}$$

$$U_1 = \frac{C_p \beta^2 \sigma_K^2}{2} \int e^{\beta u} h_t(u)\, du. \tag{29}$$

Proof: Let $u = (z - Z_j)/b$, $B = n^{-1} \sum_1^n Y_j(t) e^{\beta Z_j} \int e^{b\beta u} K(u)\, du$. By the conditions on K and boundedness of Z,

$$\begin{aligned}
E(B) &= C_p \int e^{\beta u} h_t(u)\, du \int e^{b\beta u} K(u)\, du \\
&= U_0 + U_1 b^2 + o(b^2), \\
Var(B) &\leq n^{-1} E(Y_j(t) e^{2\beta Z_j}) \left(\int e^{b\beta u} K(u)\, du \right)^2 \\
&= n^{-1} C_p \int e^{2\beta u} h_t(u)\, du \left(\int e^{b\beta u} K(u)\, du \right)^2 \\
&= O(n^{-1}).
\end{aligned}$$

The results in Lemmas 1 and 2 yield the following:

Lemma 3 *Under the conditions in Theorem 3,*

$$A/B = \frac{Q_0}{U_0} + \left(\frac{Q_1}{U_0} - \frac{Q_0 U_1}{U_0^2} \right) b^2 + \left(\frac{Q_2}{U_0} \right) \frac{1}{\sqrt{nb}} + o(b^2) + o_p((nb)^{-1/2}). \tag{30}$$

From Lemma 2 we have $Var(B) = O(n^{-1})$. It can be easily shown that $Cov(A, B) = O(n^{-1})$. Use delta-method together with the results so far obtained and Lemma 3 follows.

Since $\widehat{f}_{t,\beta}(z) = e^{\beta z} A/B$, from Lemma 3 we have the following,

$$\widehat{f}_{t,\beta}(z) = f_t(z) + c_1 b^2 + c_2/\sqrt{nb} + o(b^2) + o_p((nb)^{-1/2}),$$

where

$$c_1 = \frac{e^{\beta z}\sigma_K^2(h_t^{(2)}(z) - \beta^2 h_t(z))}{2\int e^{\beta u} h_t(u)\, du} \tag{31}$$

$$c_2 = \frac{e^{\beta z}}{\int e^{\beta u} h_t(u)\, du}\left(\frac{h_t(z)R(K)}{C_p}\right)^{1/2}. \tag{32}$$

Lemma 4 *Under the conditions in Theorem 3, and* $\dot{\beta} \in (\hat{\beta}, \beta)$*, we have*

$$\frac{d\widehat{f}_{t,\beta}(z)}{d\beta}\,|_{\dot{\beta}} \xrightarrow{p} \frac{df_t(z)}{d\beta}, \tag{33}$$

$$\hat{\beta} - \beta = O_p(n^{-1/2}). \tag{34}$$

Proof: Omitted.

Since

$$\widehat{f}_{t,\hat{\beta}}(z) - f_t(z) = \widehat{f}_{t,\beta}(z) - f_t(z) + \frac{d\widehat{f}_{t,\beta}(z)}{d\beta}\,|_{\dot{\beta}}\,(\hat{\beta} - \beta),$$

using the results in Lemmas in this section, we have proven Theorem 3.

C R package *SurvCov*

R (Ihaka and Gentleman, 1996) is a language and environment for statistical computing and graphics, and available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form from *http://www.r-project.org*. *R* can be considered as a dialect of *S*, much code written for *S* runs unaltered under *R*.

After installing the package *SurvCov* and invoking *R*, for a demo type

```
R>  library(SurvCov)
R>  demo(PBC)
```

All examples will run.

In our package *SurvCov* we did not implement an automatic selection of the optimal bandwidth, the reason being that this package is a graphical tool to visually examine the stochastic trend, if any, in the distributions of a covariate conditional on survival times or intervals of survival time. The optimal bandwidths at different survival times are of the same order $n^{-1/5}$, with corresponding scaling constants. If we choose one bandwidth of the optimal order at one time point (say, by using the results in Theorem 3), the conditional density estimates using this bandwidth at other survival times are still asymptotically unbiased and consistent.

References

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

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

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications.* Chapman & Hall Ltd.

Finkelstein, D., Ettinger, D., and Ruckdeschel, J. (1986). Long-term survivors in metastatic non-small-cell lung cancer: an eastern cooperative oncology group study. *Journal of Clinical Oncology*, 4:702–709.

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

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314.

Murray, S. and Tsiatis, A. (1996). Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics*, 52:137–151.

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

Scott, D. W. (1992). *Multivariate density estimation: theory, practice, and visualization.* John Wiley & Sons.

Stute, W. (1995). The central limit theorem under random censorship. *Annals of Statistics*, 23:422–439.

Xu, R. and Adak, S. (2002). Survival analysis with time-varying regression effects using a tree-based approach. *Biometrics*, 58:305–315.

Xu, R. and Harrington, D. (2001). An semiparametric estimate of treatment effects with censored data. *Biometrics*, 57:875–885.

Xu, R. and O'Quigley, J. (2000). Proportional hazards estimate of the conditional survival function. *Journal of the Royal Statistical Society, Series B, Methodological*, 62:667–680.