Simple nonparametric estimation of a conditional survival function when the response is interval-censored

Thierry Duchesne
(Thierry.Duchesne@mat.ulaval.ca)
Mohammad Hossein Dehghan

Département de mathématiques et de statistique, Université Laval
Unité de recherche en santé des populations, CHA

Biostatistics Seminar Series
Epidemiology, Biostatistics & Occupational Health
McGill University
February 2, 2010

1Research funded by NSERC.
Outline

1. Introduction
   - Definition of the problem
   - Notation & background

2. Generalized Turnbull
   - Weighted self-consistency equations
   - Computations and weighted EM-algorithm

3. Variance and bandwidth
   - Variance estimation
   - Bandwidth selection

4. Example
   - Follow-up of HIV patients in Kinshasa

5. Further work
   - Several covariates
   - Time-varying covariates

6. Conclusion
Definition of the problem

Nonparametric inference problem

Let \((T_1, Z_1), \ldots, (T_n, Z_n)\) be \(n\) independent observations such that 
\[ \text{Pr}[T_i > t | Z_i = z] = S(t | z). \] (E.g., \(T_i = \text{time to seroconversion}, \ Z_i = \text{age at entry in study}\)
Nonparametric inference problem

- Let \((T_1, Z_1), \ldots, (T_n, Z_n)\) be \(n\) independent observations such that \(\text{Pr}[T_i > t | Z_i = z] = S(t | z)\). (E.g., \(T_i = \text{time to seroconversion}, Z_i = \text{age at entry in study}\))

- We do observe \(Z_1, \ldots, Z_n\) exactly, but we do not necessarily precisely know the value of each of \(T_1, \ldots, T_n\).

- Instead, for (some of the) \(i\)'s, we observe \((L_i, R_i, Z_i), i = 1, \ldots, n\), with \(L_i \leq T_i \leq R_i\). (E.g., \(L_i = \text{time of last negative test}, R_i = \text{time of first positive test}\))
Nonparametric inference problem

- Let \((T_1, Z_1), \ldots, (T_n, Z_n)\) be \(n\) independent observations such that \(\Pr[T_i > t | Z_i = z] = S(t \mid z)\). (E.g., \(T_i = \text{time to seroconversion}, Z_i = \text{age at entry in study}\))

- We do observe \(Z_1, \ldots, Z_n\) exactly, but we do not necessarily precisely know the value of each of \(T_1, \ldots, T_n\).

- Instead, for (some of the) \(i\)'s, we observe \((L_i, R_i, Z_i)\), \(i = 1, \ldots, n\), with \(L_i \leq T_i \leq R_i\). (E.g., \(L_i = \text{time of last negative test}, R_i = \text{time of first positive test}\))

- For simplicity, assume that censoring interval boundaries \(L_i\) and \(R_i\) are jump times of a point process independent of \(T_i\) that bracket \(T_i\) (hence \(L_i < T_i < R_i\)). (E.g., jump times = times of visits at clinic)
Our objective

Goal: Quick and simple nonparametric (model free) estimation of $S(t|z_0)$, $t \in (0, \Upsilon)$ for a given $z_0$. 

Why would we want to do this? 

Descriptive analysis (E.g., plots of $S(t|z)$ vs $t$ for “key” values of $z$) 

Help with model specification (E.g., Sun (2006) suggests graphical checks based on ad hoc estimates of $S(t|z)$ to determine whether a proportional hazards model is reasonable)
Definition of the problem

Our objective

**Goal:** Quick and simple nonparametric (model free) estimation of $S(t|z_0)$, $t \in (0, \Upsilon)$ for a given $z_0$.

Why would we want to do this??

- Descriptive analysis *(E.g., plots of $S(t|z)$ vs $t$ for “key” values of $z$)*
Definition of the problem

Our objective

Goal: Quick and simple nonparametric (model free) estimation of $S(t|z_0), t \in (0, \Upsilon)$ for a given $z_0$.

Why would we want to do this??

- Descriptive analysis (*E.g., plots of* $S(t|z)$ *vs* $t$ *for “key” values of* $z$*)

- Help with model specification (*E.g., Sun (2006) suggests graphical checks based on ad hoc estimates of* $S(t|z)$ *to determine whether a proportional hazards model is reasonable*)
Strategy and notation

We start with the simplest problem and slowly build up to the problem of interest:

- Estimators for simple problems are well known and understood.
Strategy and notation

We start with the simplest problem and slowly build up to the problem of interest:

- Estimators for simple problems are well known and understood.
- Allows us to proceed “by analogy” ...
We start with the simplest problem and slowly build up to the problem of interest:

- Estimators for simple problems are well known and understood.
- Allows us to proceed “by analogy” ...

Notation:
Let \(1\{A\}\) be the indicator of event \(A\).
Let \(\delta_i = 1\) if \(R_i = T_i\) and \(\delta_i = 0\) if \(R_i = \infty\).
Consider the case with no covariate, or equivalently, \( Z_i = z_0 \) w.p. 1 (i.e., \( S(t|z_0) \equiv S(t) \)).
Consider the case with no covariate, or equivalently, $Z_i = z_0$ w.p. 1 (i.e., $S(t|z_0) \equiv S(t)$).

- When $L_i = R_i = T_i \forall i$ (complete data), then nonparametric MLE of $S(t)$ corresponds the complement of the empirical CDF, i.e.,

$$\hat{S}_{ECDF}(t) = \sum_{i=1}^{n} \frac{1}{n} \mathbb{1}\{T_i > t\}.$$
Consider the case with no covariate, or equivalently, $Z_i = z_0$ w.p. 1 (i.e., $S(t|z_0) \equiv S(t)$).

- When $L_i = R_i = T_i \forall i$ (complete data), then nonparametric MLE of $S(t)$ corresponds the complement of the empirical CDF, i.e.,

$$
\hat{S}_{ECDF}(t) = \sum_{i=1}^{n} \frac{1}{n} 1\{T_i > t\}.
$$

- When $L_i = T_i$ and $R_i = T_i$ or $R_i = \infty$ (right-censored data), the nonparametric MLE of $S(t)$ is the Kaplan-Meier estimator, i.e.,

$$
\hat{S}_{KM}(t) = \prod_{i: T_i \leq t} \left(1 - \frac{\delta_i}{\sum_{j=1}^{n} 1\{T_j \geq T_i\}}\right).
$$
No covariate, interval-censored data

More generally, the nonparametric MLE of $S(t)$ is Turnbull’s self-consistent estimator. Let $0 = \tau_0 < \tau_1 < \cdots < \tau_g$ be the ordered distinct values of $\{L_i, R_i, i = 1, \ldots, n\}$. Let $B_j = (\tau_{j-1}, \tau_j)$. Let $\alpha_{ij} = 1\{(L_i, R_i) \supseteq B_j\}.$
Turnbull’s estimator gives probability $p_j$ to interval $B_j$ by solving the “self-consistency” equations

$$p_j = \sum_{i=1}^{n} \frac{1}{n} \frac{\alpha_{ij} p_j}{\sum_{k=1}^{g} \alpha_{ik} p_k}, \quad j = 1, \ldots, g.$$
Turnbull’s estimator gives probability $p_j$ to interval $B_j$ by solving the “self-consistency” equations

$$p_j = \sum_{i=1}^{n} \frac{1}{n} \sum_{k=1}^{g} \frac{\alpha_{ij} p_j}{\alpha_{ik} p_k}, \quad j = 1, \ldots, g.$$ 

The estimator of $S(t)$ is then

$$\hat{S}_T(t) = \sum_{j: \tau_j > t} p_j.$$
Turnbull’s estimator

Turnbull’s estimator gives probability \( p_j \) to interval \( B_j \) by solving the “self-consistency” equations

\[
p_j = \sum_{i=1}^{n} \frac{1}{n} \frac{\alpha_{ij} p_j}{\sum_{k=1}^{g} \alpha_{ik} p_k}, \quad j = 1, \ldots, g.
\]

The estimator of \( S(t) \) is then

\[
\hat{S}_T(t) = \sum_{j: \tau_j > t} p_j.
\]

\( \hat{S}_T(t) \) reduces to \( \hat{S}_{KM} \) when data are right-censored and to \( \hat{S}_{ECDF} \) when data are complete.
Now back to the case with $Z_i$ not necessarily always equal to $z_0$. Let $k(u)$ be a density function symmetric around 0 and let $k_h(u) = (1/h)k(u/h)$. Set $\omega_i^h(z_0) = k_h(z_i - z_0)/\sum_{j=1}^n k_h(z_j - z_0)$. 
Now back to the case with $Z_i$ not necessarily always equal to $z_0$. Let $k(u)$ be a density function symmetric around 0 and let $k_h(u) = (1/h)k(u/h)$. Set $\omega_i^h(z_0) = k_h(z_i - z_0) / \sum_{j=1}^{n} k_h(z_j - z_0)$.

When $L_i = R_i = T_i \ \forall i$ (complete data), then the Nadayara-Watson estimator of $S(t|z_0) = E[1\{T_i > t\}|Z_i = z_0]$ is given by

$$\hat{S}_{NW}(t|z_0) = \sum_{i=1}^{n} \omega_i^h(z_0) 1\{T_i > t\}.$$
Now back to the case with \( Z_i \) not necessarily always equal to \( z_0 \). Let \( k(u) \) be a density function symmetric around 0 and let \( k_h(u) = (1/h)k(u/h) \). Set \( \omega_i^h(z_0) = k_h(z_i - z_0) / \sum_{j=1}^{n} k_h(z_j - z_0) \).

When \( L_i = R_i = T_i \ \forall i \) (complete data), then the Nadayara-Watson estimator of \( S(t|z_0) = E[\mathbf{1}\{T_i > t\}|Z_i = z_0] \) is given by

\[
\hat{S}_{NW}(t|z_0) = \sum_{i=1}^{n} \omega_i^h(z_0) \mathbf{1}\{T_i > t\}.
\]

When all weights are equal, then \( \hat{S}_{NW}(t|z_0) \) reduces to \( \hat{S}_{ECDF}(t) \).
With covariate, right-censored data

Under right-censored data, Beran (1981) proposed the Generalized Kaplan-Meier estimator of $S(t|z_0)$. It can be written as

$$\hat{S}_{GKM}(t|z_0) = \prod_{i: T_i \leq t} \left( 1 - \frac{\delta_i \omega_i^h(z_0)}{\sum_{j=1}^{n} 1\{T_j \geq T_i\} \omega_j^h(z_0)} \right).$$
Under right-censored data, Beran (1981) proposed the Generalized Kaplan-Meier estimator of \( S(t|z_0) \). It can be written as

\[
\hat{S}_{GKM}(t|z_0) = \prod_{i:T_i \leq t} \left( 1 - \frac{\delta_i \omega_i^h(z_0)}{\sum_{j=1}^{n} 1\{T_j \geq T_i\} \omega_j^h(z_0)} \right).
\]

⇒ Beran replaced the indicators \( \delta_i \) and \( 1\{T_j \geq T_i\} \) in the Kaplan-Meier by weighted versions.
Under right-censored data, Beran (1981) proposed the Generalized Kaplan-Meier estimator of $S(t|z_0)$. It can be written as

$$
\hat{S}_{GKM}(t|z_0) = \prod_{i: T_i \leq t} \left(1 - \frac{\delta_i \omega_i^h(z_0)}{\sum_{j=1}^{n} 1\{T_j \geq T_i\} \omega_j^h(z_0)} \right).
$$

⇒ Beran replaced the indicators $\delta_i$ and $1\{T_j \geq T_i\}$ in the Kaplan-Meier by weighted versions.

⇒ When there is no censoring, $\hat{S}_{GKM}(t|z_0)$ reduces to $\hat{S}_{NW}(t|z_0)$. 
Under right-censored data, Beran (1981) proposed the Generalized Kaplan-Meier estimator of $S(t|z_0)$. It can be written as

$$\hat{S}_{GKM}(t|z_0) = \prod_{i:T_i \leq t} \left( 1 - \frac{\delta_i \omega^h_i(z_0)}{\sum_{j=1}^n 1\{T_j \geq T_i\} \omega^h_j(z_0)} \right).$$

⇒ Beran replaced the indicators $\delta_i$ and $1\{T_j \geq T_i\}$ in the Kaplan-Meier by weighted versions.

⇒ When there is no censoring, $\hat{S}_{GKM}(t|z_0)$ reduces to $\hat{S}_{NW}(t|z_0)$.

⇒ When all weights are equal, $\hat{S}_{GKM}(t|z_0)$ reduces to $\hat{S}_{KM}(t)$. 
Weighted self-consistency equations

First try: replace indicators with weighted indicators

Turnbull’s self-consistency equations are

\[ p_j = \sum_{i=1}^{n} \frac{1}{n} \sum_{k=1}^{g} \alpha_{ij} p_j \sum_{k=1}^{g} \alpha_{ik} p_k, \quad j = 1, \ldots, g. \]

**Question:** Can we do as Beran did to go from \( \hat{S}_{KM} \) to \( \hat{S}_{GKM} \) and replace \( \alpha_{ij} \) by a weighted version?
**Weighted self-consistency equations**

**First try: replace indicators with weighted indicators**

Turnbull’s self-consistency equations are

\[ p_j = \sum_{i=1}^{n} \frac{1}{n} \frac{\alpha_{ij}p_j}{\sum_{k=1}^{g} \alpha_{ik}p_k}, \quad j = 1, \ldots, g. \]

**Question:** Can we do as Beran did to go from \( \hat{S}_{KM} \) to \( \hat{S}_{GKM} \) and replace \( \alpha_{ij} \) by a weighted version?

No! Weight must depend on covariate \( Z_i \) \( \Rightarrow \omega_i \) and not \( \omega_j \) \( \Rightarrow \) cancels out in self-consistency equations.
Approach via self-consistency equations

There is hope though ...

- Efron showed that $\hat{S}_{KM}$ satisfies self-consistency equations.
Approach via self-consistency equations

There is hope though ...

- Efron showed that $\hat{S}_{KM}$ satisfies self-consistency equations.
- Turnbull’s self-consistency equations reduce to those of Efron’s under right-censoring.
There is hope though ...

- Efron showed that $\hat{S}_{KM}$ satisfies self-consistency equations.
- Turnbull’s self-consistency equations reduce to those of Efron’s under right-censoring.
- $\Rightarrow$ What self-consistency equations does $\hat{S}_{GKM}$ satisfy?
Weighted self-consistency equations

Second try: replace $1/n$ by $\omega_i^h(z_0)$

If we replace $1/n$ by $\omega_i^h(z_0)$ in Turnbull’s equations, then we get

$$p_j^h(z_0) = \sum_{i=1}^{n} \omega_i^h(z_0) \frac{\alpha_{ij} p_j^h(z_0)}{\sum_{k=1}^{g} \alpha_{ik} p_k^h(z_0)}, \quad j = 1, \ldots, g.$$ 

This works well:
Weighted self-consistency equations

Second try: replace $1/n$ by $\omega_i^h(z_0)$

If we replace $1/n$ by $\omega_i^h(z_0)$ in Turnbull’s equations, then we get

$$p_j^h(z_0) = \sum_{i=1}^{n} \omega_i^h(z_0) \frac{\alpha_{ij} p_j^h(z_0)}{\sum_{k=1}^{g} \alpha_{ik} p_k^h(z_0)}, \quad j = 1, \ldots, g.$$ 

This works well:

- $\hat{S}_{GKM}$ does solve these equations;
If we replace $1/n$ by $\omega^h_i(z_0)$ in Turnbull’s equations, then we get

$$p^h_j(z_0) = \sum_{i=1}^{n} \omega^h_i(z_0) \frac{\alpha_{ij} p^h_j(z_0)}{\sum_{k=1}^{g} \alpha_{ik} p^h_k(z_0)}, \quad j = 1, \ldots, g.$$ 

This works well:

- $\hat{S}_{GKM}$ does solve these equations;
- as easy to implement numerically as Turnbull’s estimator;
Second try: replace $1/n$ by $\omega_i^h(z_0)$

If we replace $1/n$ by $\omega_i^h(z_0)$ in Turnbull’s equations, then we get

$$p_j^h(z_0) = \sum_{i=1}^{n} \omega_i^h(z_0) \frac{\alpha_{ij} p_j^h(z_0)}{\sum_{k=1}^{g} \alpha_{ik} p_k^h(z_0)}, \quad j = 1, \ldots, g.$$ 

This works well:

- $\hat{S}_{GKM}$ does solve these equations;
- as easy to implement numerically as Turnbull’s estimator;
- can be justified as a local likelihood estimator;
Second try: replace $1/n$ by $\omega_i^h(z_0)$

If we replace $1/n$ by $\omega_i^h(z_0)$ in Turnbull’s equations, then we get

$$p_j^h(z_0) = \sum_{i=1}^{n} \omega_i^h(z_0) \frac{\alpha_{ij} p_j^h(z_0)}{\sum_{k=1}^{g} \alpha_{ik} p_k^h(z_0)}, \quad j = 1, \ldots, g.$$  

This works well:

- $\hat{S}_{GKM}$ does solve these equations;
- as easy to implement numerically as Turnbull’s estimator;
- can be justified as a local likelihood estimator;
- reduces to $\hat{S}_T$ when weights are equal, reduces to $\hat{S}_{GKM}$ when only right-censored data, reduces to $\hat{S}_{NW}$ under complete data.
Iterative solution algorithm

The self consistency equations suggest the following scheme for numerical evaluation:

**Algorithm 1**

**Step 1.** Set initial values $p_j^{h:(0)}(z_0), j = 1, \ldots, g$ (e.g., $1/g \forall j$, or $1/M$ for each of the $M$ innermost intervals). Set $r = 1$. 


The self consistency equations suggest the following scheme for numerical evaluation:

**Algorithm 1**

**Step 1.** Set initial values $p_j^{h:(0)}(z_0), j = 1, \ldots, g$ (e.g., $1/g \forall j$, or $1/M$ for each of the $M$ innermost intervals). Set $r = 1$.

**Step 2.** Compute

$$p_j^{h:(r)}(z_0) = \sum_{i=1}^{n} \omega_i^{h}(z_0) \frac{\alpha_{ij} p_j^{h:(r-1)}(z_0)}{\sum_{k=1}^{g} \alpha_{ik} p_k^{h:(r-1)}(z_0)}, \quad j = 1, \ldots, g.$$
Iterative solution algorithm

The self consistency equations suggest the following scheme for numerical evaluation:

**Algorithm 1**

**Step 1.** Set initial values $p_j^{h:(0)}(z_0)$, $j = 1, \ldots, g$ (e.g., $1/g \forall j$, or $1/M$ for each of the $M$ innermost intervals). Set $r = 1$.

**Step 2.** Compute

$$p_j^{h:(r)}(z_0) = \sum_{i=1}^{n} \omega_i^h(z_0) \frac{\alpha_{ij} p_j^{h:(r-1)}(z_0)}{\sum_{k=1}^{g} \alpha_{ik} p_k^{h:(r-1)}(z_0)}, \quad j = 1, \ldots, g.$$  

**Step 3.** If $d(p^{(r)}, p^{(r-1)}) < \epsilon$, stop, otherwise put $r = r + 1$ and return to Step 2.
Algorithm 1 as a “weighted” EM-algorithm

Consider the “weighted marginal log-likelihood”

\[
\ell_\omega = \sum_{i=1}^{n} \omega_i^h(z_0) \ln \{S(L_i) - S(R_i)\}.
\]

Let \((L_i, R_i)\) be the observed data and \((T_i, L_i, R_i)\) be the complete data. Then the complete data log-likelihood is

\[
\ell_{\text{Comp}} = \sum_{j=1}^{g} \left( \sum_{i=1}^{n} \omega_i^h(z_0) \mathbf{1}\{T_i \in B_j\} \right) \ln p_j.
\]
Algorithm 1 as a “weighted” EM-algorithm

Consider the “weighted marginal log-likelihood”

$$\ell_\omega = \sum_{i=1}^{n} \omega_i^h(z_0) \ln \{ S(L_i) - S(R_i) \}.$$  

Let $(L_i, R_i)$ be the observed data and $(T_i, L_i, R_i)$ be the complete data. Then the complete data log-likelihood is

$$\ell_{\text{Comp}} = \sum_{j=1}^{g} \left( \sum_{i=1}^{n} \omega_i^h(z_0) 1\{T_i \in B_j\} \right) \ln p_j.$$  

It is easy to show that the E-step yields

$$Q(p|p^{(r)}) = \sum_{j=1}^{g} d^*(p^{(r)}) p_j^{(r)} \ln p_j,$$

where $d^*_j(p^{(r)}) = \sum_{i=1}^{n} \omega_i^h(z_0) \alpha_{i,j} / \sum_{k=1}^{g} \alpha_{i,k} p_k^{(r)}$. 

Algorithm 1 as a “weighted” EM-algorithm

Consider the “weighted marginal log-likelihood”

\[ \ell_\omega = \sum_{i=1}^{n} \omega_i^h(z_0) \ln \{ S(L_i) - S(R_i) \}. \]

Let \((L_i, R_i)\) be the observed data and \((T_i, L_i, R_i)\) be the complete data. Then the complete data log-likelihood is

\[ \ell_{Comp} = \sum_{j=1}^{g} \left( \sum_{i=1}^{n} \omega_i^h(z_0) \mathbf{1}\{T_i \in B_j\} \right) \ln p_j. \]

It is easy to show that the E-step yields

\[ Q(p|p^{(r)}) = \sum_{j=1}^{g} d^*(p^{(r)}) p_j^{(r)} \ln p_j, \]

where

\[ d^*_j(p^{(r)}) = \sum_{i=1}^{n} \omega_i^h(z_0) \alpha_{ij} / \sum_{k=1}^{g} \alpha_{ik} p_k^{(r)}. \]

It is as easy to show that \( p_j^{h:(r+1)}(z_0) \) defined in Algorithm 1 maximizes \( Q(p|p^{(r)}) \) w.r.t. \( p \) (M-step).
So Algorithm 1 is an EM-algorithm applied to a **weighted** likelihood ... 
⇒ Are the numerical properties of the EM-algorithm preserved under a weighted likelihood??
Numerical convergence of weighted EM-algorithms

So Algorithm 1 is an EM-algorithm applied to a weighted likelihood...

⇒ Are the numerical properties of the EM-algorithm preserved under a weighted likelihood??

**Lemma** If $\ell_{obs}(\theta) = \sum_{i=1}^{n} \omega_i \ln f_{obs}(x_i; \theta)$ with $0 < \omega_i < 1$ and \{$(\theta^{(r)})$, $r = 0, 1, 2, \ldots$\} is the sequence of parameter estimates obtained with the EM-algorithm, then

$$\ell_{obs}(\theta^{(r+1)}) \geq \ell_{obs}(\theta^{(r)}) \forall r = 0, 1, \ldots$$
Numerical convergence of weighted EM-algorithms

So Algorithm 1 is an EM-algorithm applied to a weighted likelihood ...

⇒ Are the numerical properties of the EM-algorithm preserved under a weighted likelihood??

**Lemma** If $\ell_{obs}(\theta) = \sum_{i=1}^{n} \omega_i \ln f_{obs}(x_i; \theta)$ with $0 < \omega_i < 1$ and \{ $\theta^{(r)}$, $r = 0, 1, 2, \ldots$ \} is the sequence of parameter estimates obtained with the EM-algorithm, then

$$\ell_{obs}(\theta^{(r+1)}) \geq \ell_{obs}(\theta^{(r)}) \forall r = 0, 1, \ldots$$

⇒ Under general conditions, the EM algorithm preserves its monotone convergence property with weighted likelihoods.
Sun (2001) has proposed bootstrap algorithms to get pointwise variance estimates of $\hat{S}_T(t)$.

Our simulations show that the following two versions of his algorithm provide good pointwise variance estimators for $\hat{S}_{GT}(t|z_0)$. 
Bootstrap variance estimation

**Bootstrap algorithm**

1. Fix $t$ and $z_0$. Compute bandwidth $h(n)$. 

Variance estimation
Bootstrap variance estimation

Bootstrap algorithm

1. Fix $t$ and $z_0$. Compute bandwidth $h(n)$.
2. For $k$ in 1 to $K$:
   Step (i) Sample $n$ observations with replacement from $(L_i, R_i, Z_i)$,
   $i = 1, \ldots, n \Rightarrow (L_i^{(k)}, R_i^{(k)}, Z_i^{(k)}), i = 1, \ldots, n$. 
**Bootstrap algorithm**

1. Fix $t$ and $z_0$. Compute bandwidth $h(n)$.

2. For $k$ in 1 to $K$:

   Step (i) Sample $n$ observations with replacement from $(L_i, R_i, Z_i)$, $i = 1, \ldots, n \Rightarrow (L_i^{(k)}, R_i^{(k)}, Z_i^{(k)}), i = 1, \ldots, n$.

   Step (ii) Version A Compute $\hat{S}_{GT}^{h(n)}(t|z_0)$ using the bootstrap sample from Step (i).

   Step (ii) Version B Compute a bandwidth $h^{(k)}(n)$ and then $\hat{S}_{GT}^{h^{(k)}(n)}(t|z_0)$ using the bootstrap sample from Step (i).
Bootstrap variance estimation

Bootstrap algorithm

1. Fix $t$ and $z_0$. Compute bandwidth $h(n)$.
2. For $k$ in 1 to $K$:
   - Step (i) Sample $n$ observations with replacement from $(L_i, R_i, Z_i)$, $i = 1, \ldots, n \Rightarrow (L_i^{(k)}, R_i^{(k)}, Z_i^{(k)}), i = 1, \ldots, n$.
   - Step (ii) Version A Compute $\hat{S}_{GT}^{h(n)}(t|z_0)$ using the bootstrap sample from Step (i).
   - Step (ii) Version B Compute a bandwidth $h^{(k)}(n)$ and then $\hat{S}_{GT}^{h^{(k)}(n)}(t|z_0)$ using the bootstrap sample from Step (i).

Use the sample variance of the $K$ values $\hat{S}_{GT}^{(k)}(t|z_0)$ as the variance estimate.
Variance estimation

What our simulations suggest

\( T \sim \text{Weibull}(\text{shape} = 3, \text{scale} = 1.5Z_0), \ n = 50, \ r = 4 \) (small censoring intervals) and \( z_0 = 15 \)

| \( t \) | \( S_{True}(t|z_0) \) | \( \sigma^2_{True} \) | \( \hat{\sigma}^2_{\text{Boot.}A} \) | \( \hat{\sigma}^2_{\text{Boot.}B} \) | \( \text{MSE} \) | \( (\text{BIAS})^2 \) |
|---|---|---|---|---|---|---|
| 10.3 | 0.908 | 0.00307 | 0.00313 | 0.00377 | 0.00386 | 0.000794 |
| 14.5 | 0.763 | 0.00642 | 0.00639 | 0.00769 | 0.00837 | 0.001960 |
| 20.0 | 0.495 | 0.00860 | 0.00847 | 0.00950 | 0.00915 | 0.000561 |
| 25.5 | 0.235 | 0.00628 | 0.00620 | 0.00693 | 0.00688 | 0.000606 |
| 29.7 | 0.100 | 0.00379 | 0.00370 | 0.00416 | 0.00575 | 0.001960 |

Version A slightly underestimates, Version B slightly overestimates

Computational burden of Version B not worth it if goal of analysis is descriptive...
Variance estimation

What our simulations suggest

\[ T \sim \text{Weibull}(\text{shape} = 3, \text{scale} = 1.5Z_0), \ n = 50, \ r = 4 \text{ (small censoring intervals)} \] and \[ z_0 = 15 \]

| \( t \) | \( S_{\text{True}}(t|z_0) \) | \( \sigma^2_{\text{True}} \) | \( \hat{\sigma}^2_{\text{Boot},A} \) | \( \hat{\sigma}^2_{\text{Boot},B} \) | MSE | \( (\text{BIAS})^2 \) |
|----|----------------|-----------------|----------------|----------------|------|-----------------|
| 10.3 | 0.908 | 0.00307 | 0.00313 | 0.00377 | 0.00386 | 0.000794 |
| 14.5 | 0.763 | 0.00642 | 0.00639 | 0.00769 | 0.00837 | 0.001960 |
| 20.0 | 0.495 | 0.00860 | 0.00847 | 0.00950 | 0.00915 | 0.000561 |
| 25.5 | 0.235 | 0.00628 | 0.00620 | 0.00693 | 0.00688 | 0.000606 |
| 29.7 | 0.100 | 0.00379 | 0.00370 | 0.00416 | 0.00575 | 0.001960 |

- Version A slightly underestimates, Version B slightly overestimates
Variance estimation

What our simulations suggest

\[ T \sim \text{Weibull}(\text{shape } = 3, \text{scale } = 1.5Z_0), \ n = 50, \ r = 4 \text{ (small censoring intervals)} \text{ and } z_0 = 15 \]

| \( t \) | \( S_{\text{True}}(t|z_0) \) | \( \sigma_{\text{True}}^2 \) | \( \hat{\sigma}_{\text{Boot}.A}^2 \) | \( \hat{\sigma}_{\text{Boot}.B}^2 \) | \( \text{MSE} \) | \( (\text{BIAS})^2 \) |
|---|---|---|---|---|---|---|
| 10.3 | 0.908 | 0.00307 | 0.00313 | 0.00377 | 0.00386 | 0.000794 |
| 14.5 | 0.763 | 0.00642 | 0.00639 | 0.00769 | 0.00837 | 0.001960 |
| 20.0 | 0.495 | 0.00860 | 0.00847 | 0.00950 | 0.00915 | 0.000561 |
| 25.5 | 0.235 | 0.00628 | 0.00620 | 0.00693 | 0.00688 | 0.000606 |
| 29.7 | 0.100 | 0.00379 | 0.00370 | 0.00416 | 0.00575 | 0.001960 |

- Version A slightly underestimates, Version B slightly overestimates
- Computational burden of Version B not worth it if goal of analysis is descriptive...
Because the number of innermost intervals \((B_j’s\ with \ p_j > 0)\) changes when \(h\) changes if \(k(u)\) is of finite support, bandwidth selection is much easier with \(k(u) = (2\pi)^{-1/2} \exp(-u^2/2), -\infty < u < \infty\). Under this gaussian (normal) kernel, two methods investigated:
Bandwidth selection rules

Because the number of innermost intervals \((B_j\text{'s with } p_j > 0)\) changes when \(h\) changes if \(k(u)\) is of finite support, bandwidth selection is much easier with \(k(u) = (2\pi)^{-1/2} \exp(-u^2/2)\), \(-\infty < u < \infty\). Under this gaussian (normal) kernel, two methods investigated:

- Silverman’s rule of thumb (easy and “cheap” to implement; not tailor made for the problem at hand)

\[
h_{ROT} = 1.06\hat{\sigma}_z n^{-1/5},
\]

with \(\hat{\sigma}_z^2\) the sample variance of \(Z_1, \ldots, Z_n\).
Bandwidth selection rules

- $D$-fold likelihood cross-validation (Pan, 2000) (costly implementation; proven its worth when smoothing interval censored data in $t$)
Bandwidth selection rules

- $D$-fold likelihood cross-validation (Pan, 2000) (costly implementation; proven its worth when smoothing interval censored data in $t$)

$$
\ell_{CV}(h) = \sum_{j=1}^{D} \sum_{i \in G_j} \ln \left\{ \sum_{k=1}^{m_{h}^{-j}(z_0)} p_{k,h}^{-j}(z_0) \left( K_*^t \left[ \min \{a, (R - v_k^{(-j)})/\alpha \} \right] \right. \right. \\
\left. \left. - K_*^t \left[ \max \{-a, (L - v_k^{(-j)})/\alpha \} \right] \right) \right\},
$$

with $K_*^t(s) = \int_{-a}^{s} K_t(u) \, du$, $K_t$ a density on $(-a,a)$, and $\alpha$ another bandwidth.

With $K_t$ normal, $\alpha(h) = \tilde{T} \left\{ 20 \left[ (\bar{Z} - z_0)^2 / (\gamma^2) \right] + 10 \right\} h / \gamma$, with

$$\tilde{T}_i = (L_i + R_i) / 2$$

and $\gamma = \text{midpoint between } \bar{Z} \text{ and } z_0$. 
Optimal \( (h_{opt}) \), average cross-validation \( (h_{CV}) \) and average rule-of-thumb \( (h_{ROT}) \) bandwidths, with 
\( T \sim \text{Weibull}(\text{shape} = 3, \text{scale} = 1.5Z_0) \) and \( Z \sim U(5, 25) \) for various values of \( n, r \) and \( z_0 \) under the gaussian kernel.

<table>
<thead>
<tr>
<th>( z_0 )</th>
<th>8</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r = 0.25 )</td>
<td>( h_{opt} )</td>
<td>2.2</td>
<td>2.8</td>
<td>3.8</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>( h_{CV} )</td>
<td>2.1</td>
<td>3.0</td>
<td>3.6</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>( h_{ROT} )</td>
<td>2.8</td>
<td>2.8</td>
<td>2.8</td>
<td>2.8</td>
</tr>
<tr>
<td>( r = 4 )</td>
<td>( h_{opt} )</td>
<td>2.0</td>
<td>2.5</td>
<td>3.4</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td>( h_{CV} )</td>
<td>2.1</td>
<td>2.9</td>
<td>3.1</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td>( h_{ROT} )</td>
<td>2.8</td>
<td>2.8</td>
<td>2.8</td>
<td>2.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( z_0 )</th>
<th>8</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r = 0.25 )</td>
<td>( h_{opt} )</td>
<td>2.0</td>
<td>2.6</td>
<td>3.1</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>( h_{CV} )</td>
<td>1.9</td>
<td>2.7</td>
<td>3.1</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>( h_{ROT} )</td>
<td>2.4</td>
<td>2.4</td>
<td>2.4</td>
<td>2.4</td>
</tr>
<tr>
<td>( r = 4 )</td>
<td>( h_{opt} )</td>
<td>1.8</td>
<td>2.0</td>
<td>2.9</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>( h_{CV} )</td>
<td>2.0</td>
<td>2.8</td>
<td>3.0</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td>( h_{ROT} )</td>
<td>2.4</td>
<td>2.4</td>
<td>2.4</td>
<td>2.4</td>
</tr>
</tbody>
</table>
Recap

To get a reasonable $\hat{S}_{GT}(t|z_0)$ for descriptive or model selection purposes, use gaussian kernel and $h_{ROT}$. 
Recap

- To get a reasonable $\hat{S}_{GT}(t|z_0)$ for descriptive or model selection purposes, use gaussian kernel and $h_{ROT}$.

- Bootstrap (Version A) variance estimates are quite reasonable.
Follow-up of HIV patients in Kinshasa

Kinshasa (Democratic Republic of the Congo) HIV study

746 female sex workers from Kinshasa were periodically tested for HIV. All were HIV negative at the beginning of the study and 70 eventually tested positive. Time of sero-conversion ($T_i$) is interval-censored by time of last negative test ($L_i$) and time of first positive test ($R_i$).

Is the Cox model a reasonable approach to link $Z_i$, age of women at entry in the study to $T_i$, time until sero-conversion? If so, plots of $\ln \{-\ln \hat{S}(t|z)\}$ vs $\ln t$ for a few values of $z$ should show curves that are vertical shifts of each other.
Nonparametric estimates of $S(t|z)$ for various values of $z_0$ and with $h_{ROT} = 2.6$. 

![Graph showing nonparametric estimates of $S(t|z)$ for different values of $z$ and bandwidth $h_{ROT}$.](image)
Nonparametric estimates of $S(t|z)$ for various values of $z_0$ and with $h_{CV} = 3$, $h_{ROT} = 2.6$ and $h = 4$. 

![Graph of Nonparametric estimates of $S(t|z)$ for various values of $z_0$ and with $h_{CV} = 3$, $h_{ROT} = 2.6$ and $h = 4$.](image)
Is the Cox model reasonable??

![Graph showing nonparametric estimates](image-url)
Several covariates

Multi-dimensional covariate $\mathbf{Z}$

- We can define $\omega_i^h(z_0) = k_h(Z_i - z_0) / \sum_{j=1}^{n} k_h(Z_j - z_0)$, with $k_h$ a multivariate density with mode at zero and scale parameter $h$ and the estimator proposed still works.
Several covariates

**Multi-dimensional covariate** $\mathbf{Z}$

- We can define $\omega_i^h(z_0) = k_h(\mathbf{Z}_i - z_0) / \sum_{j=1}^{n} k_h(\mathbf{Z}_j - z_0)$, with $k_h$ a multivariate density with mode at zero and scale parameter $h$ and the estimator proposed still works.

- Choice of $h$ extremely difficult and variance quickly explodes due to curse of dimensionality
Several covariates

Multi-dimensional covariate $Z$

- We can define $\omega_i^h(z_0) = \frac{k_h(Z_i - z_0)}{\sum_{j=1}^n k_h(Z_j - z_0)}$, with $k_h$ a multivariate density with mode at zero and scale parameter $h$, and the estimator proposed still works.

- Choice of $h$ extremely difficult and variance quickly explodes due to curse of dimensionality.

- If some elements of $Z$ are categorical, then things get trickier...
Several covariates

Multi-dimensional covariate $Z$

- We can define $\omega_i^h(z_0) = k_h(Z_i - z_0)/\sum_{j=1}^n k_h(Z_j - z_0)$, with $k_h$ a multivariate density with mode at zero and scale parameter $h$ and the estimator proposed still works.

- Choice of $h$ extremely difficult and variance quickly explodes due to curse of dimensionality

- If some elements of $Z$ are categorical, then things get trickier ...

$\Rightarrow$ Method proposed here better suited for univariate data exploration.
What if $Z_i$ is replaced by $\{Z_i(s), s \in \mathcal{I}_i\}$?

In some situations, we may observe the values of a covariate that varies in time, say $\{Z_i(s), s \in \mathcal{I}_i\}$ for some set of observation times $\mathcal{I}_i$:

- Covariate defined as a function of time, e.g., interaction between a covariate $\tilde{Z}_i$ and a function of time, such as $\{Z_i(s) = \tilde{Z}_i \ln(s), s > 0\}$;
In some situations, we may observe the values of a covariate that varies in time, say \( \{Z_i(s), s \in \mathcal{I}_i\} \) for some set of observation times \( \mathcal{I}_i \):

- Covariate defined as a function of time, e.g., interaction between a covariate \( \tilde{Z}_i \) and a function of time, such as \( \{Z_i(s) = \tilde{Z}_i \ln(s), s > 0\} \);

- Covariate measured at the jump times of the bracketing point process, e.g., at visit times \( \{Z_i(s), s \in \{s_{i1}, \ldots, s_{in_i}\}\} \).
What if $Z_i$ is replaced by $\{Z_i(s), s \in \mathcal{I}_i\}$?

In some situations, we may observe the values of a covariate that varies in time, say $\{Z_i(s), s \in \mathcal{I}_i\}$ for some set of observation times $\mathcal{I}_i$:

- Covariate defined as a function of time, e.g., interaction between a covariate $\tilde{Z}_i$ and a function of time, such as $\{Z_i(s) = \tilde{Z}_i \ln(s), s > 0\}$;
- Covariate measured at the jump times of the bracketing point process, e.g., at visit times $\{Z_i(s), s \in \{s_{i1}, \ldots, s_{in_i}\}\}$;

$\Rightarrow$ **Q1**: Can we use a (weighted) EM-algorithm to extend methods of joint modelling of longitudinal and survival data?
What if $Z_i$ is replaced by $\{Z_i(s), s \in \mathcal{S}_i\}$?

In some situations, we may observe the values of a covariate that varies in time, say $\{Z_i(s), s \in \mathcal{S}_i\}$ for some set of observation times $\mathcal{S}_i$:

- Covariate defined as a function of time, e.g., interaction between a covariate $\tilde{Z}_i$ and a function of time, such as $\{Z_i(s) = \tilde{Z}_i \ln(s), s > 0\}$;
- Covariate measured at the jump times of the bracketing point process, e.g., at visit times $\{Z_i(s), s \in \{s_{i1}, \ldots, s_{in_i}\}\}$;

$\Rightarrow$ **Q1:** Can we use a (weighted) EM-algorithm to extend methods of joint modelling of longitudinal and survival data?

$\Rightarrow$ **Q2:** Can the nonparametric estimator proposed here be extended to handle time-varying covariates?
Q1: Many papers that propose methods for joint longitudinal and survival data analysis use the EM-algorithm to condition on unobserved random effects. We add the $T_i$ to the missing data and replace the event times by $(L_i, R_i)$ in the observed data.
Q1: Many papers that propose methods for joint longitudinal and survival data analysis use the EM-algorithm to condition on unobserved random effects. We add the $T_i$ to the missing data and replace the event times by $(L_i, R_i)$ in the observed data.

Q2: In the case where $Z_i(s)$ is a function of $\tilde{Z}_i$ and time, then the covariate path is entirely determined by $\tilde{Z}_i$ and thus $Pr[T_i > t | \{Z_i(s), s \in S_i\}] = Pr[T_i > t | \tilde{Z}_i]$. Hence the method presented here applies with $\omega_i$ defined as a function of $\tilde{Z}_i$. The case where $\{Z_i(s), s \in S_i\}$ appears more difficult without “parameterizing” the covariate paths or some assumptions (e.g., collapsibility, semi-parametric) on $S(t | \{Z_i(s), s \in S_i\})$. 
Final thoughts

- Estimator $\hat{S}_{GT}$ with $h_{ROT}$ as easy to implement as $S_T$ and gives pretty good results.
- Helpful for descriptive analysis or regression model specification when response is interval-censored and covariate is continuous. Less helpful for formal inference, such as significance tests of covariate effect.
- Inferential properties can presumably be improved, but can simplicity be retained?