NONPARAMETRIC ESTIMATION WITH INCOMPLETE DATA: EFFICIENT NONPARAMETRIC HAZARD RATE ESTIMATION WITH LEFT TRUNCATED AND RIGHT CENSORED DATA

by Sam Efromovich

Endowed Professor

Head of Actuarial Program, Department of Mathematical Sciences, UTDallas

This work was partially supported by NSF Grant DMS–0906790 and NSA Grant H982301310212.

Talk134r in pap134

EXAMPLES OF INCOMPLETE DATA

We are interested in a random variable X^* which cannot be observed directly.

Typical cases are TRUNCATION and CENSORING.

Examples from Klugman, Panjer and Willmot "LOSS MODELS":

Left Truncation: An ordinary deductible of d is applied (a loss below d may not be known).

Right Censoring: A policy limit u is applied. If loss exceeds u its value is not recorded.

Left Truncation and Right Censoring: In constructing a mortality table, it is common to follow a group of people of various ages over only a few years. When a person joins a study, he or she is alive at that time. This person's age at death must be at least as great as the age at entry to the study and thus has been *left truncated*. If the person is alive at the end of the study or leaves the study prior to its end, then *right censoring* occurs because the person's age at death is unknown.

Two more examples:

WHEL STUDY: Contains data on the time from a breast cancer surgery to the cancer relapse. The information is available only for women who at the time of baseline had no relapsed cancer, and this creates *left truncation*. At the same time, some participants either dropped the study, or died during the study, or had no relapse during the study, and then these events created *right censoring*.

CHANNING HOUSE is a retirement center located in Palo Alto, California, and its distinctive feature is that all residents of the community are covered by a health care program with a zero deductible. The resident's age at entry to the community as well as the age at leaving the community or death were recorded. Note that when a person joined the community, he or she was alive at that time, and the person's age at death was at least as great as the age at entry to the community and thus was *left truncated*. If the resident was alive when the study ended or the resident left the community before the study ended, *right censoring* had occurred.

IMPORTANT REMARKS:

(i) The fundamental difference between censoring and truncation is that the former informs us about missing of an underlying observation of interest while the latter does not.

(ii) In what follows, I will not use classical ideas and methods of product-limit estimation proposed and explored by Kaplan and Meier, Peterson, Nelson and Aalen, and many other great statisticians. Instead, the classical method of moment estimation, namely that

the sample mean is a good estimate

of the population mean

is used.

Consequences of ignoring left truncation and right censoring via a simple example

Suppose that two independent random variables U^* and V^* have Gamma distribution with (shape,scale) parameters $(2, 2)$ and $(5, 1)$, respectively, and then the left truncation occurs at a fixed point $a = 5$ and this generates truncated random variables U and V, respectively. Then $\mathbb{E}\{U^* - V^*\} = -1$ while $\mathbb{E}\{U - V\} = .5$. Let us also note that there is no chance to estimate distributions of U^* and V^* for values smaller than the point of truncation a. Furthermore, if then right censoring at a fixed point 10 is applied to U and V, and we denote corresponding censored random variables as U' and V', then $\mathbb{E}\{U' - V'\} = 0.11$. As we see, ignoring truncation and/or censoring creates a bias.

Modeling Left Truncated and Right Censored Data

Data consist of n observations. Observations are collected via a hidden sequential sampling from a triplet of mutually independent and nonnegative random variables (T^*, X^*, Z^*) where:

T^* is the truncating random variable,

X^* is the random variable of interest,

Z^* is the censoring random variable.

Suppose that (T_k^*, X_k^*, Z_k^*) is the kth realization of (T^*, X^*, Z^*) and the available sample of truncated and censored statistics is of size $l-1$, $l-1 \leq \min(k-1, n-1)$. If $T_k^* > \min(X_k^*, Z_k^*)$ then left truncation of the kth realization occurs meaning that:

(i) kth triplet is not observed;

(ii) the fact that the kth observation has occurred is unknown;

(iii) next realization of the triplet should be waited for.

On the other hand, if $T_k^* \leq \min(X_k^*, Z_k^*)$ then an observation $(T_l, Y_l, R_l) := (T_k^*, \min(X_k^*, Z_k^*), I(X_k^* \leq Z_k^*))$ is available. Sequential sampling from the triplet (T^*, X^*, Z^*) stops as soon as $l = n$.

Note that the hidden mechanism of collecting data can be described via a negative binomial experiment such that:

(i) The experiment stops as soon as nth "success" occurs;

(ii) Data are collected only when a "success" occurs;

(iii) There is no information on how many "failures" occurred between "successes";

(iii) The probability of "success" is

$$p := \mathbb{P}(T^* \leq \min(X^*, Z^*)) = \int_0^\infty f^{T^*}(t) G^{X^*}(t) G^{Z^*}(t) dt. \tag{0.1}$$

While we do not know the total number of hidden "failures", the negative binomial distribution sheds some light on the hidden number N of "failures", and in particular the mean and variance of N are

$$\mathbb{E}(N) = n(1-p)p^{-1}, \quad \mathrm{Var}(N) = n(1-p)p^{-2}.$$

HAZARD RATE

By definition, the hazard rate function of nonnegative random variable X^* is

$$h^{X^*}(x) := \lim_{v \to 0} \frac{\mathbb{P}(x \leq X^* \leq x + v | X^* \geq x)}{v} = \frac{f^{X^*}(x)}{G^{X^*}(x)} = -\frac{dG^{X^*}(x)/dx}{G(x)},$$

where $f^{X^*}(x)$ is the probability density of X^*,

$$G^{X^*}(x) := \int_x^\infty f^{X^*}(u)du = 1 - F^{X^*}(x), \quad G^{X^*}(x) > 0, \ \ x \geq 0,$$

is the survivor function, and $F^{X^*}(x)$ is the cumulative distribution function of X^*. If one thinks about X^* as a time to an event-of-interest, then $h^{X^*}(x)dx$ represents the instantaneous likelihood that the event occurs within the interval $(x, x + dx)$ given that the event has not occurred at time x. The hazard rate quantifies the trajectory of imminent risk, and it may be referred to by other names in different sciences, for instance as the failure rate in reliability theory and the force of mortality in sociology.

CLASSICAL PROPERTIES OF THE HAZARD RATE

(i) The hazard rate, similarly to the probability density or the survivor function, characterizes the random variable X^*. Namely, if the hazard rate is known then the corresponding probability density is

$$f^{X^*}(x) = h^{X^*}(x)e^{-\int_0^x h^{X^*}(v)dv},$$

and the survivor function is

$$G^{X^*}(x) = e^{-\int_0^x h^{X^*}(v)dv}.$$

The preceding identity follows from integrating both sides of

$$h^{X^*}(x) = -\frac{dG^{X^*}(x)/dx}{G^{X^*}(x)}$$

and then using $G^{X^*}(0) = 1$.

(ii) The hazard rate is nonnegative and has the same smoothness as the corresponding density.

(iii) The hazard rate is not integrable on its support because a hazard rate must satisfy $\int_0^\infty h(x)dx = \infty$. Hence it is natural to study it over a finite interval, for instance $[a, a+1]$, $a \geq 0$.

(iv) The hazard rate of the minimum of two independent random variables is the sum of the hazard rates of the two random variables.

EXAMPLES:

(i) A familiar example is the constant hazard rate of an exponential random variable (the rate is equal to the reciprocal of the mean), and inverse is also valid - a constant hazard rate implies exponential distribution. A constant hazard rate has coined the name memoryless for exponential distribution.

(ii) Another interesting example is the Weibull distribution

$$f^{X^*}(x, k, \lambda) = (k/\lambda)(x/\lambda)^{k-1} e^{-(x/\lambda)^k} I(x > 0),$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter.

If $k < 1$ then the hazard rate is decreasing ("infant mortality");

If $k > 1$ then the hazard rate is increasing ("aging" process).

ASYMPTOTIC THEORY

The studied risk is the Mean Integrated Squared Error (MISE) over a unit interval $[a, a+1]$. In the case of observing data without truncation and censoring, for α-fold differentiable functions the familiar minimax rate $n^{-2\alpha/(2\alpha+1)}$ of the MISE convergence is well known (Ibragimov and Khasminskii 1981 and Efromovich 1999).

Under a mild assumption, the minimax rate is preserved for observations with left truncation and right censoring.

Hence our aim is to find a corresponding sharp minimax constant of the MISE convergence.

To present solution of the problem, it is convenient to begin with explanation of the used minimax approach via the game theory.

MINIMAX GAME

There are three players: the dealer, the statistician and nature.

• The dealer chooses all parameters of an underlying function class as well as nuisance functions defining left truncation and right censoring mechanisms, and then presents them to nature.

• Nature chooses most difficult, for estimation, hazard rate from the dealer's class and generates a corresponding sample of observations from the hazard rate.

• The statistician and the dealer try to find best estimate of the hazard rate.

• If an observer is present, who may know everything about the game, including the underlying hazard rate, then the observer is traditionally called oracle and the oracle also can suggest an estimate.

NOTATION

In what follows $\varphi_0(x) = 1$ and $\varphi_j(x) = 2^{1/2}\cos(\pi j(x-a))$, $j > 0$ are elements of the classical cosine basis on the interval $[a, a+1]$), α is a positive integer number, $\theta_j = \int_a^{a+1} \varphi_j(x)h(x)dx$ are Fourier coefficients of $h^{X^*}(x)$ on $[a, a+1]$. $\mathbb{E}_{h^{X^*}}\{\cdot\}$ denotes the expectation given a hazard rate function h^{X^*} (remember that the hazard rate characterizes a random variable), and $o_s(1)$ are generic sequences that are vanishing as $s \to \infty$.

LOCAL SOBOLEV CLASS OF α-DIFFERENTIABLE FUNCTIONS

$$\mathcal{S}(\alpha, Q, h_0^{X^*}, \beta) := \Big\{ h : \quad h(x) = h_0^{X^*}(x) + \sum_{j=1}^{\infty} \theta_j \varphi_j(x) I(x \in [a, a+1]);$$

$$\sum_{j=1}^{\infty} (\pi j)^{2\alpha} \theta_j^2 \le Q < \infty, \ \alpha \in \{1, 2, \ldots\}, \ \sup_{x \in [a, a+1]} |\sum_{j=1}^{\infty} \theta_j \varphi_j(x)| < \ln^{-1}(n+3);$$

$$\int_0^{a+1} h_0^{X^*}(v) dv < \infty, \quad \inf_{x \in [a, a+1]} h_0^{X^*}(x) \ge 0,$$

$$\sum_{j=0}^{\infty} (1 + j^{2\alpha+\beta}) [\int_a^{a+1} h_0^{X^*}(x) \varphi_j(x) dx]^2 < \infty, \ \beta > 0 \Big\}.$$

The underlying idea of this class is that all considered functions are not farther than $\ln^{-1}(n+3)$ in $L_\infty([a, a+1])$-norm from the pivot. This is why the class is called local. The last line indicates that the pivot should be smoother than a "regular" function from the class.

Golubev (1991) shows that a pivot does not affect the sharp constant in the probability density, regression and spectral density estimation problems.

Next theorem shows that the pivot does affect the constant in the hazard rate case.

THEOREM 1 (Lower Bound for Dealer-Estimators). Let us assume that $G^{Z^*}(x)F^{T^*}(x)G^{X^*}(x) > 0$ for $x \in [a, a+1]$. Then the following lower bound for the local minimax MISE holds,

$$\inf_{\check{h}^*} \sup_{h^{X^*} \in \mathcal{S}(\alpha, Q, h_0^{X^*}, \beta)} \mathbb{E}_{h^{X^*}} \Big\{ \int_a^{a+1} (\check{h}^*(x) - h^{X^*}(x))^2 dx \Big\}$$

$$\geq P(\alpha, Q)([\int_a^{a+1} h_0^{X^*}(v) g^{-1}(v) dv] n^{-1})^{2\alpha/(2\alpha+1)}(1 + o_n(1)),$$

where the infimum is taken over all possible dealer-estimators $\check{h}^*$ based on a left truncated and right censored sample $(Y_1, T_1, R_1), \ldots, (Y_n, T_n, R_n)$, distributions F^{T^*}, F^{Z^*} and parameters $(\alpha, Q, h_0^{X^*}, \beta)$ of the underlying Sobolev class,

$$P(\alpha, Q) := Q^{1/(2\alpha+1)}(2\alpha+1)^{1/(2\alpha+1)} \Big[\frac{\alpha}{\pi(\alpha+1)}\Big]^{2\alpha/(2\alpha+1)},$$

$$g(v) := \mathbb{P}(T \leq v \leq Y) = \mathbb{P}(T^* \leq v \leq Y^* | T^* \leq Y^*) = [p^{-1} G^{Z^*}(v) F^{T^*}(v)] G^{X^*}(v),$$

and $p = \mathbb{P}(T^* \leq \min(X^*, Z^*))$.

ADAPTIVE DATA-DRIVEN ESTIMATOR

The estimator is defined as

$$\hat{h}(x) := \sum_{k=1}^{K_n} \Big[1 - \frac{\hat{d}n^{-1}}{L_k^{-1}\sum_{j\in B_k}\hat{\theta}_j^2}\Big] I\Big(L_k^{-1}\sum_{j\in B_k}\hat{\theta}_j^2 > (\hat{d}+1/\ln(n))n^{-1}\Big)\sum_{j\in B_k}\hat{\theta}_j\varphi_j(x),$$

where

$$\hat{\theta}_j := \sum_{l=1}^{n} R_l\varphi_j(Y_l)\eta_l^{-1}I(Y_l\in[a,a+1]),$$

$$\eta_l := \sum_{s=1}^{n} I(T_s \le Y_l \le Y_s),$$

$$\hat{d} := n\sum_{l=1}^{n}\eta_l^{-2}R_l I(X_l\in[a,a+1]).$$

THEOREM 2. **Let** $G^{Z^*}(x)F^{T^*}(x)G^{X^*}(x) > 0$ **for** $x\in[a,a+1]$**. Then the adaptive estimator** $\hat{h}(x)$ **is sharp-minimax, that is**

$$\sup_{h^{X^*}\in\mathcal{S}(\alpha,Q,h_0^{X^*},\beta)} \mathbb{E}_{h^{X^*}}\{\int_a^{a+1}(\hat{h}(x)-h^{X^*}(x))^2dx\}$$

$$\le P(\alpha,Q)([\int_a^{a+1} h_0^{X^*}(x)g^{-1}(x)dx]n^{-1})^{2\alpha/(2\alpha+1)}(1+o_n(1)).$$

NUMERICAL STUDY

The proposed estimator is compared with an oracle-estimator motivated by a kernel estimator of Uzunogullari and Wang (1992),

$$\check{h}(x, b^*(x)) = n^{-1} \sum_{l=1}^{n} K\Big(\frac{x - Y_l}{b^*(x)}\Big) \frac{R_l}{b^*(x) g(Y_l)} I(Y_l \in [a - tb^*(a), a + 1 + tb^*(a+1)]),$$

where $K(x)$ is the Gaussian kernel and $b^*(x)$ is the golden-rule oracle's bandwidth

$$b^*(x) = n^{-1/5} \frac{[h(x) \int K^2(t)dt]^{1/5}}{[h''(x) \int t^2 K(t)dt]^{2/5} (g(x)^{1/5}}.$$

The oracle uses an underlying hazard rate to calculate the bandwidth and information about p^{-1}, F^{T^*}, G^{X^*} and G^{Z^*} to calculate the underlying function $g(x)$. Furthermore, the oracle uses observations Y_l from an increased interval $[a - tb^*(a), a + 1 + tb^*(a + 1)]$ to take into account boundary issues. To deal with boundary issues, the cosine series estimator is enriched by two polynomial functions x and x^2 via using Gram-Schmidt orthonormalization.

EXPERIMENT

The underlying distributions of X^* are either Weibull $W(\gamma, \beta)$, where γ and β are shape and scale parameters, respectively, or it is a Bath-tub (BT) distribution generated by $X^* := \min(V_1, V_2)$ with V_1 and V_2 being $W(0.3, 1)$ and $W(15, 1)$, respectively. The underlying distributions of T^* and Z^* are either exponential or uniform. Several different intervals $[a, b]$ and $n \in \{100, 200, 300, 400, 500, 1000\}$ are considered. Then for each experiment, that is the underlying distributions, interval $[a, b]$ and sample size, 5000 simulations are conducted. For each simulation the empirical integrated squared error of the oracle (ISEO) and the empirical integrated squared error of the proposed data-driven series estimator (ISEE) are calculated. Then the median ratio (over 5000 simulations) of ISEO/ISEE is shown in Table 1 as well as the average number of observations fallen within a studied interval $[a, b]$. Note that each entry in Table 1 is written as A/B where A is the median ratio of ISEO/ISEE and B is the average number of observations fallen within $[a, a + 1]$.

Table 1. Results of Monte Carlo simulations. Distributions are denoted as $W(\gamma, \beta)$, BT, $U(c_1, c_2)$ and $E(\lambda)$ for Weibull with shape parameter γ and scale parameter β, Bathtub corresponding to the minimum of random variables with distributions $W(0.3, 1)$ and $W(15, 1)$, uniform on the interval $[c_1, c_2]$, and exponential with $1/\lambda$ being the mean, respectively. For each experiment, which is defined by the three distributions, sample size n, and the interval of estimation $[a, b]$, 5000 samples are generated and then for each sample the oracle estimate and the proposed data-driven series estimate are calculated and then the corresponding empirical integrated squared errors over the interval of interest $[a, b]$ are calculated and denoted as ISEO and ISEE, respectively. A corresponding entry in the Table is written as A/B where A is the median of 5000 ratios ISEO/ISEE and B is the average number of observations fallen within the considered interval $[a, b]$. The fifth column shows coefficients of difficulty d^*.

					n					
X^*	T^*	Z^*	$[a, b]$	d^*	100	200	300	400	500	1000
$W(3, 4)$	$U(0, 3)$	$U(3, 10)$	$[0.5, 4]$	1.84	0.80/59	0.92/118	1.07/177	1.20/237	1.15/295	1.41/590
$W(3, 4)$	$U(0, 3)$	$U(3, 10)$	$[1, 4]$	1.79	0.71/58	0.87/116	0.88/174	1.10/232	1.01/290	1.32/580
$W(3, 4)$	$U(0, 3)$	$U(3, 10)$	$[1, 5]$	6.88	0.71/80	0.89/161	1.02/241	1.03/320	1.26/400	1.41/800
$W(1.2, 5)$	$E(1)$	$E(0.05)$	$[1, 5]$	1.59	0.92/48	1.03/95	1.27/143	1.33/200	1.45/251	1.43/480
$W(0.5, 2)$	$E(2)$	$E(0.05)$	$[0.5, 8]$	4.70	0.72/46	0.80/91	0.86/138	0.97/182	1.07/230	1.44/462
$W(0.5, 2)$	$E(5)$	$E(0.05)$	$[0.1, 3]$	1.89	0.69/47	0.83/95	0.90/142	0.94/190	1.01/237	1.24/474
$W(0.3, 1)$	$E(2)$	$E(0.05)$	$[0.2, 6]$	2.45	0.73/32	0.92/64	1.00/96	1.08/128	1.19/160	1.58/320
$W(3, 2)$	$E(1)$	$E(0.15)$	$[1, 2]$	1.90	0.91/38	0.99/74	1.13/111	1.11/151	1.24/188	1.56/371
$W(3, 2)$	$E(1)$	$E(0.15)$	$[0.5, 2]$	2.08	0.84/45	0.91/91	1.09/137	1.04/182	1.15/229	1.42/456
$W(3, 2)$	$E(1.5)$	$E(0.1)$	$[0.5, 2.5]$	6.70	0.76/72	0.94/144	1.03/215	1.12/287	1.10/360	1.47/720
BT	$E(80)$	$E(0.5)$	$[0.05, 0.9]$	1.74	0.78/45	0.90/90	1.08/135	1.16/180	1.38/225	1.40/450

CONCLUSION FROM TABLE 1

Only for the smallest sample sizes $n \leq 200$ the oracle dominates the estimator. The main reason of this domination is that sizes of subsamples, used by the estimator and shown in denominators of corresponding cells, are small for nonparametric estimation. Furthermore, remember that the oracle uses the unknown function $g(x)$ and the underlying hazard rate function $h^{X^*}(x)$ to calculate the golden-rule bandwidth, and the latter is helpful for the smallest samples. Coefficients of difficulty d^* shed light on relative complexity of a particular experiment.

ANALYSIS OF REAL DATA

CHANNING HOUSE: It is a retirement center located in Palo Alto, California, and its distinctive feature is that all residents of the community are covered by a health care program with a zero deductible, that is, no additional financial burden to the residents. The data were collected between the opening of the house in January 1964 and July 1975. In that period of time 97 men and 365 women passed through the center, and among them 130 women and 46 men died at Channing House. The resident's age at entry to the community as well as the age at leaving the community or death were recorded. Note that when a person joined the community, he or she was alive at that time, and the person's age at death was at least as great as the age at entry to the community and thus was left truncated. If the resident was alive when the study ended or the resident left the community before the study ended, right censoring had occurred. Differences between the survival hazard rates for male and female residents was the primary aim of the study.

WHEL STUDY: **The project, supported by NIH, was designed to address the question of whether high intake of vegetables and fruits could reduce breast cancer recurrence. The project included 3088 women previously treated for breast cancer and who were cancer-free at the baseline. Participants were randomly assigned to either an intensive diet intervention or to a comparison group, this was done from year 1995 to year 2000, and then participants were followed upon through 2006. Plasma carotenoids concentrations, including alpha-carotene, beta-carotene, lutein, lycopene and cryptoxanthin, were measured at baseline using blood samples (only 3044 participants provided blood samples and therefore are considered in the example). One of the main aims of the study was to examine the relationship between a plasma carotenoids concentration (as an indicator of dietary intake) and recurrence-free survival. The variable of interest, X^*, is the recurrence-free survival time which is defined as the time from the date of initial breast cancer diagnosis (for all our purposes that date can be considered as the date of surgery) to the date of diagnosis of the cancer relapse.**

In our analysis of the data, for each type of plasma carotenoid, we divided 3044 patients, who provided blood samples at baseline, into two groups using the median concentration as the threshold. For example, if alpha-carotene is the variable of interest, then participants whose plasma alpha-carotene was above the median value were classified as group 1 and participants with plasma alpha-carotine below the median were classified as group 2. Then for each group the proposed adaptive hazard rate estimator was used to estimate the hazard rate for the time from surgery to relapse. Similarly paired estimates were calculated for other types of carotenoids.

T-TEST

In all our practical examples we compared hazard rates for two populations, and in all cases nonparametric estimates have indicated that average hazard rates are different. Hence, it is reasonable to evaluate an integrated hazard rate over the interval of interest $\mu = \int_a^b h^{X^*}(x)dx$ **for each population and then compare them. Let** μ_1 **and** μ_2 **denote integrated over** $[a, b]$ **hazard rates for first and second populations considered in the above-discussed examples. Then it is natural to test the null hypothesis** $\mu_1 = \mu_2$ **versus the alternative hypothesis** $\mu_1 > \mu_2$**. To shed a different light on the hypotheses, remember that a survival function** $G^{X^*}(x) = e^{-\int_0^x h^{X^*}(v)dv}$**, and then for any two populations with hazard rates** $h^{X_1^*}$ **and** $h^{X_2^*}$ **we have**

$$\ln\left(\left[\frac{G^{X_1^*}(b)}{G^{X_1^*}(a)}\right]/\left[\frac{G^{X_2^*}(b)}{G^{X_2^*}(a)}\right]\right) = \int_a^b [h^{X_2^*}(x) - h^{X_1^*}(x)]dx = \mu_2 - \mu_1.$$

In particular, if $a = 0$ **then we get** $G^{X_1^*}(b)/G^{X_2^*}(b) = e^{\mu_2 - \mu_1}$**.**

Because $\mu = \mathbb{E}_{h^{X^*}}\{Rg^{-1}(Y)I(Y \in [a, a+b])\}$ **, the natural estimate of** μ **is** $\hat{\mu} := \sum_{l=1}^n R_l \eta_l^{-1} I(Y_l \in [a, a+b])$**. In its turn, this estimate allows us to use a standard t-test for comparison between means of two populations.**

Table 2. Results for hypothesis testing.

Example	$\hat{\mu}_1$	$\hat{\mu}_2$	$\text{Var}(\hat{\mu}_1)$	$\text{Var}(\hat{\mu}_2)$	p-value
Channing House	1.08	0.85	0.0263	0.0065	0.1003
Alpha-carotene	0.16	0.20	0.00013	0.00017	0.0047
Beta-carotene	0.16	0.21	0.00013	0.00017	0.0017
Cryptoxanthin	0.17	0.19	0.00014	0.00016	0.1195
Lycopene	0.16	0.20	0.00014	0.00017	0.0115

CONFIDENCE BANDS

For Channing House example and WHEL data are shown in Figure 5. Here, following Wasserman (2005), we show confidence bands for $E\{\hat{h}(x)\}$. The top diagram in Figure 5 exhibites 80 percent confidence bands for male and female hazard rates. As we see, there is a relatively small interval (about 20 months) around 930 months where the two confidence bands do not intersect. Furthermore, there is a larger interval in time (about 200 months) where each estimate is not covered by confidence band for another edstimate. The left-bottom diagram shows 90 percent confidence bands for two groups of participants in the WHEL study with different levels of alpha-carotene. As we see, there is a relatively large period in time when the two confidence bands do not intersect. To see something similar for beta-carotine, it is required to use 80 percent confidence bands (see the right-bottom diagram). Note that it is typical for nonparametric confidence bands to be much wider near boundaries (see Wasserman 2005). For the Channing House example this phenomenon is more pronounce for the right boundary, and for the WHEL study both boundaries are affected.

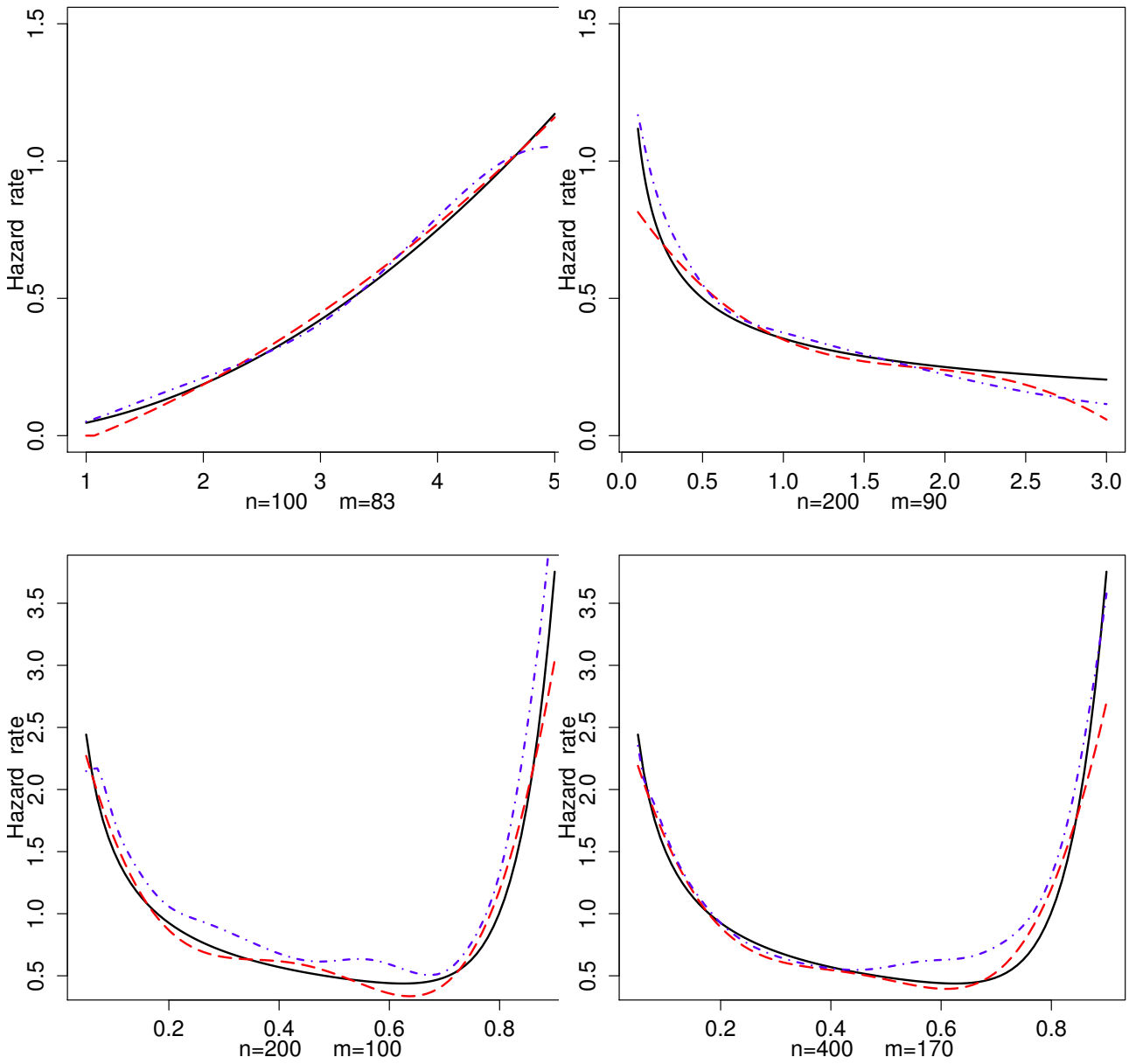

Figure 1: Estimation of a hazard rate function (solid line) by oracle kernel estimator (dot-dash line) and proposed estimator (longdash line). The top left, right top and two bottom diagrams correspond to the third, sixth and eleventh experiments described in Table 1. Subtitles show the total sample size n and the number m of Ys observed within an interval of estimation.

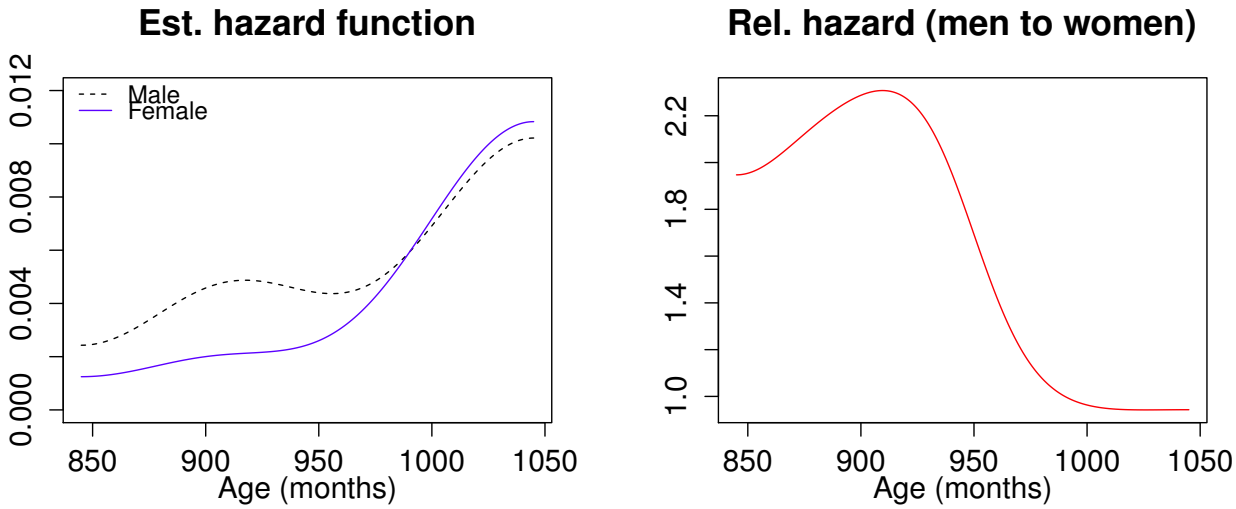

Figure 2: Analysis of the Channing House data. The left diagram shows estimated hazard rates for male (the dotted line) and female (the solid line) residents. The right diagram shows the relative hazard rate which is the ratio of the male's estimated hazard rate to the female's estimated hazard rate.

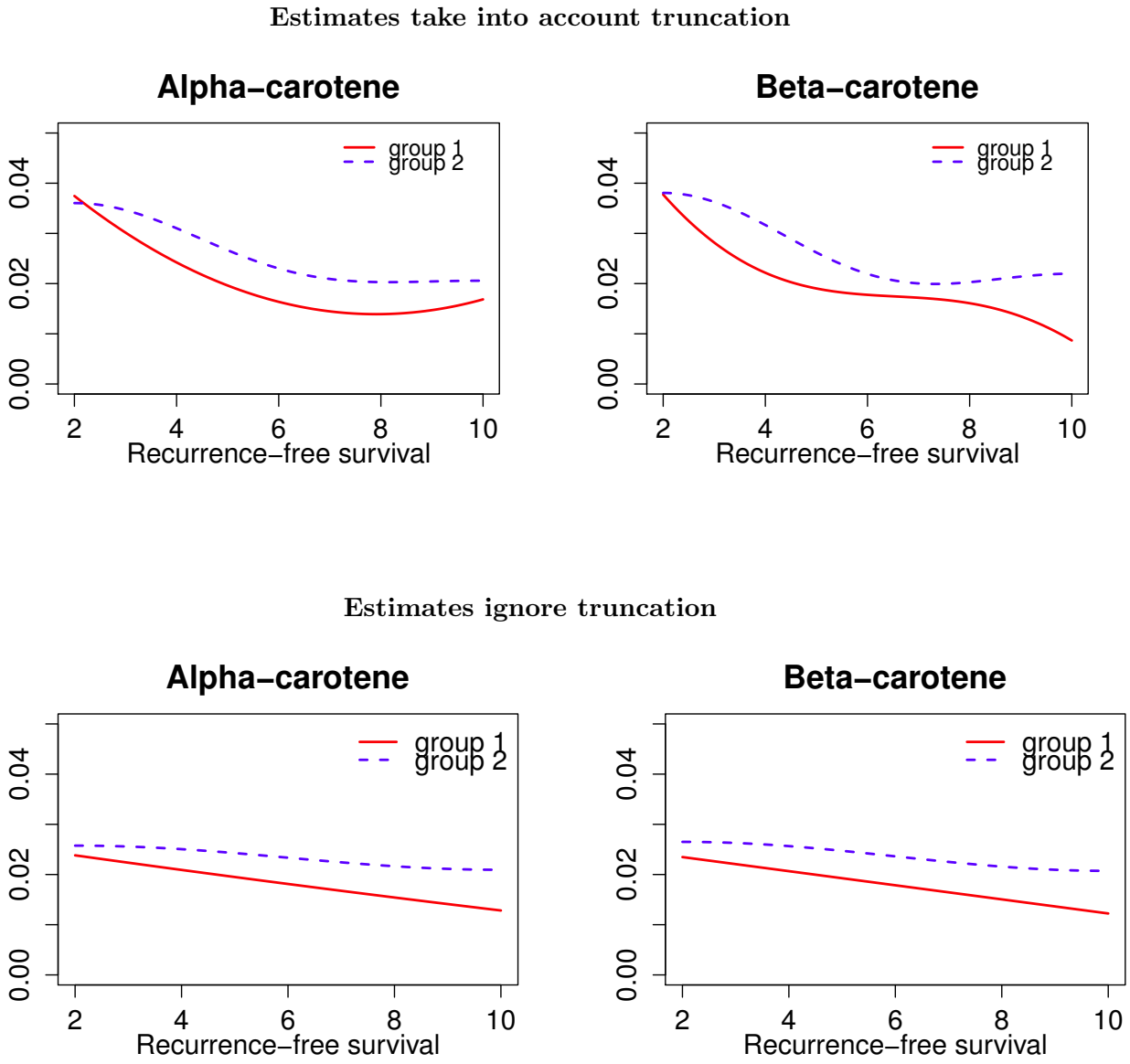

Figure 3: Estimated hazard rates for two groups of women in WHEL study. The two top diagrams show the proposed estimates, the two bottom diagrams show estimates that ignore left truncation. Women with larger than the median alpha-carotene level (the left column of diagrams) or larger than the median beta-carotene level (the right column of diagrams) belong to the second group, and correspondingly others to the first group.

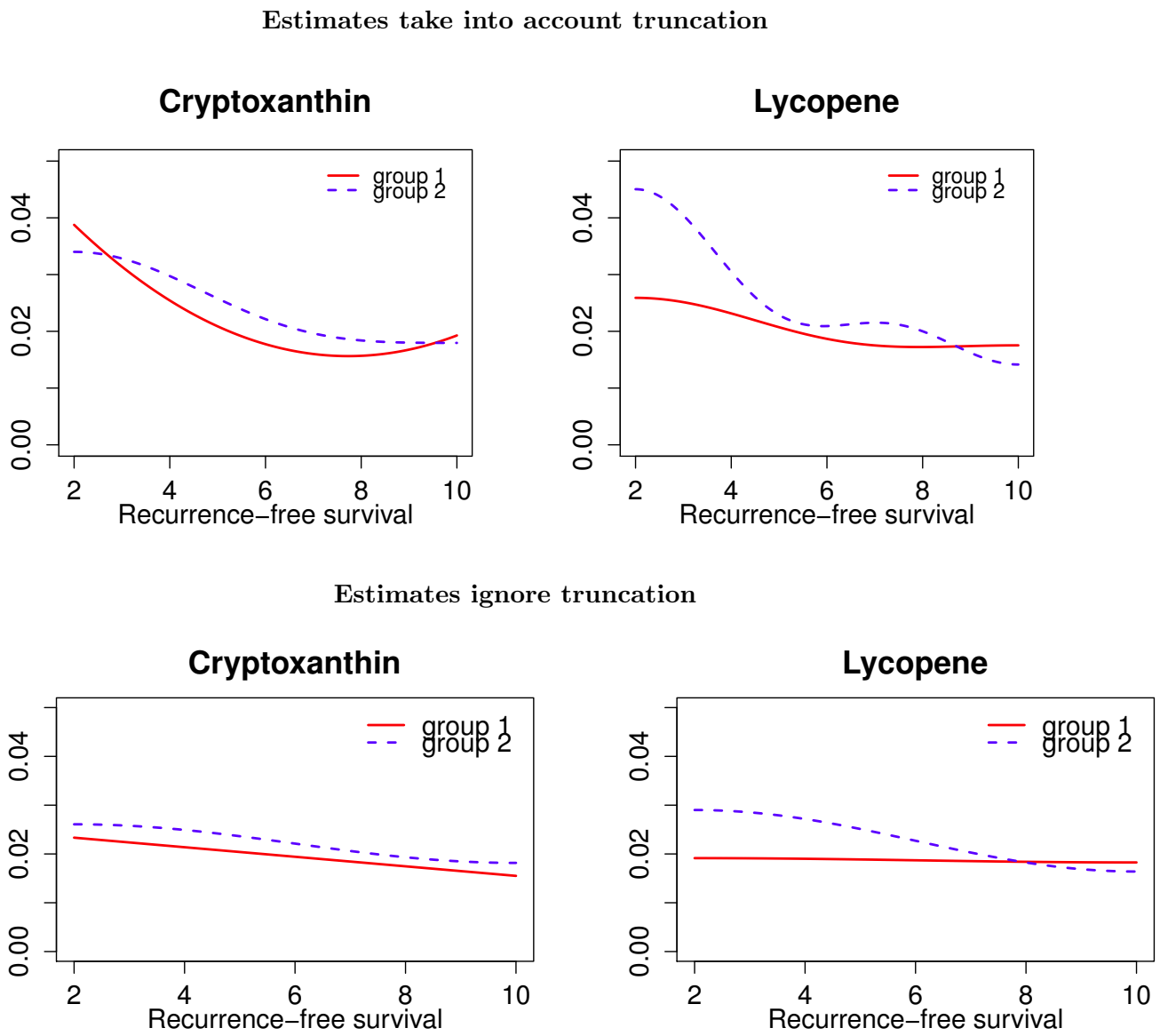

Figure 4: Estimated hazard rates for two groups of women in WHEL study. Structure of Figure 3 is identical to Figure 2 only here the effects of larger concentrations of cryptoxanthin and lycopene are studied.

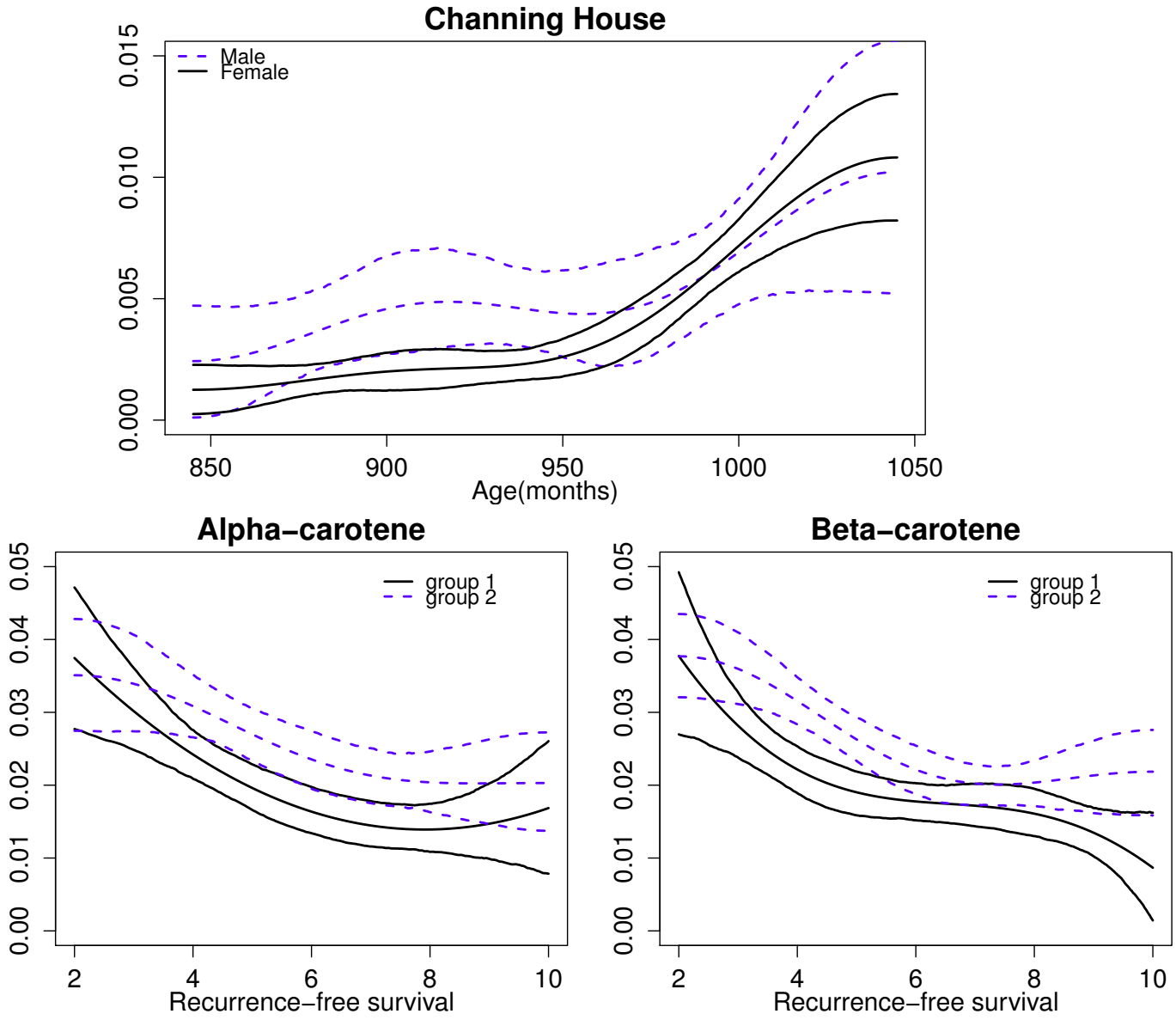

Figure 5: Confidence bands. The top diagram shows 80 perecent confidence bands for the Channing House data. Two bottom diagrams devoted to WHEL study and show 90 and 80 percent confidence bands for alpha-carotine and beta-carotine, respectively.