SEMIPARAMETRIC MODELING OF LABELED CELL KINETICS, WITH APPLICATION TO ISOTOPE LABELING OF ERYTHROCYTES

Hans-Georg Müller[1], Chun-Lung Su[1], Stephen R. Dueker[2], Yumei Lin[2], Andrew Clifford[2], Bruce A. Buchholz[3] and John S. Vogel[3]

Second Revision June 2002

Abstract

We propose a stochastic model for the kinetics of cells which have been tagged with a chemical label. The proposed model consists of two components: A parametrically specified distribution for the time to incorporation of the label into the cells, and a non-parametric survival function reflecting the survival time of the label-cell combination. The target quantity of this modeling approach is the fraction of labeled cells among all cells, viewed as a function of time. Longitudinal measurements of this labeled cell fraction are available from a recent experiment with folate labeled red blood cells. The proposed semiparametric model is fitted to these data and some of the implications are explored. The proposed method also includes bootstrap-based inference.

Keywords and phrases: Cell Survival, Labeled Cell Fraction, Smoothing

[1]Department of Statistics, University of California, One Shields Ave., Davis, CA 95616
[2]Department of Nutrition, University of California, One Shields Ave., Davis, CA 95616
[3]Center for Accelerator Mass Spectrometry, Lawrence Livermore National Laboratory, Livermore, CA 94551

1. Introduction

The labeling of cells with a chemical marker is a classical technique for in vivo studies of cell behavior and survival. Survival characteristics are typically inferred from the behavior of the associated label. The method of cell labeling dates back to Shemin and Rittenberg (1946) who introduced this technique with the purpose of determining the life span of erythrocytes. Shemin and Rittenberg labeled erythrocytes with ^{15}N-glycine. Using a parametric mathematical modeling approach, they determined that the average lifetime of erythrocytes is approximately 120 days, and this value is accepted to the present day. Recent work along these lines is reviewed in Macallan et al. (1998) and Hellerstein (1999).

In this paper, we propose an alternative approach based on semiparametric modeling. The proposed model implies a stochastic interpretation for the survival of labeled red blood cells (erythrocytes), extending the work of Bergner (1962). We demonstrate the usefulness of the proposed model with an application to the analysis of data that were generated in a recent labeling experiment, where instead of using ^{15}N-glycine, the labeling of erythrocytes was accomplished with an $80nmol$ (corresponding to $100nCi$) dose of ^{14}C-folic acid. A detailed description of this experiment and the data that were generated can be found in Clifford et al. (1998).

The proposed class of models more generally applies to experiments in which a cohort of cells is exposed to a label *in vivo*, and where the incorporation of the label to an exposed cell may be viewed as a chance event. The observations consist of measurements of the fraction of labeled cells, longitudinally measured over time. The typical time dynamic behavior of the labeled cell fraction exhibits an initial delay. In the case of erythrocytes, this delay corresponds to a phase of label incorporation during erythropoiesis which lasts for a few days. One then observes a sharp initial rise of the labeled cell fraction, as labeled erythrocytes enter the circulating blood stream. This is followed by a protracted decline. The decline reflects a loss of labeled cells, that can

be caused by two distinct mechanisms: (a) Release of the label from a labeled cell, which then is transformed into an unlabeled cell; or (b) Removal of a labeled cell from the circulating compartment being studied, usually due to cell death.

In the following, survival refers to the survival of the label-cell unit. Either one of the two above mentioned events constitutes a terminal event for the label-cell unit. If one intends to draw more specific conclusions on actual cell survival and lifespan, one needs to base such inference on the assumption that the loss of label-cell complexes is solely due to cell death.

The purpose of this paper is to propose a stochastic model for the time course of the label-cell units. We demonstrate how distinct components of the labeled cell fraction function can be identified by appropriate estimation methods, based on observed trends in labeled cell fractions. One component reflects the process of cell labeling, during which label-cell units are formed, and a second component corresponds to the survival of these label-cell units. The proposed semiparametric model is quite general, the application to ^{14}C-folate labeling of erythrocytes being a special case of the general technique that we develop here.

The underlying idea of the proposed modeling approach is to postulate a parametric model for the time to incorporation of the labels into the cells, and a smooth nonparametric function for the survival time of the label-cell unit. The combination of these two components then leads to a flexible semiparametric model. Of particular interest is the estimation of the survival component that corresponds to the declining phase of the label-cell units. This requires separate identification of the distinct rising and declining phases of the circulating label-cell units. We note that nonparametric modeling entails very few assumptions about the unknown shape of the survival function of the label-cell units and therefore is a preferred method. So far only parametric models have been described in the literature. As we demonstrate in the following, parametric approaches are often limited in scope and are not easily applicable to the data that motivate our study.

2. A Stochastic Model For Labeled Cell Kinetics

The experiment was carried out with a human subject. Oral ingestion of the label occurred at time $t = 0$, the origin of the time axis. An event of interest is the time of labeling of a circulating erythrocyte. The labeling agent, which in our example was ^{14}C-folic acid, becomes available in the bone marrow where the cells are born and reside for only a very short time before they enter the circulating blood stream. Only cells residing in the marrow and responsive to labeling at the time the label enters the bloodstream are subject to being labeled. The incorporation of the label into the cell is a random event that may or may not happen during a cell's stay in the marrow. We are only able to observe the time course of the labeled cells which constitute the cohort of cells of interest. The labeled cells leave the marrow and enter the blood stream at a random time, denoted by T_1. The statistical distribution of the random variable T_1, which is the time of entry of the labeled cell into the blood stream, is assumed to be of known shape such as a normal distribution with unknown parameters. The second important event is the time where the label-cell unit disappears, due to cell death or dissociation of the label-cell unit. The timing of this event corresponds to a second random time T_2. If $T_2 \leq T_1$, a cell remains unlabeled and is not in the labeled cohort.

We assume that the distribution functions of the two random variables T_1 and T_2 are F_1 and F_2, respectively. Additional model assumptions are as follows:

(A1) The random times T_1 and T_2 are independent.

(A2) The form of the distribution function F_1 of T_1 is known, subject to unknown location and scale parameters. Formally, $T_1 \sim f_1$ with probability density function (pdf) $f_1 = f\left(\frac{\cdot - \mu}{\sigma}\right)$, where f is a known pdf and μ and σ are unknown location and scale parameters of this distribution.

(A3) The distribution of T_2 is smooth and otherwise unknown. More precisely, this distribution possesses a probability density function (pdf) f_2, which is smooth, i.e., a continuous second derivative exists. It is not assumed that f_2 belongs to a parametric class of pdfs.

Assumption (A1) reflects that time of entry into and disappearance from the blood stream are considered to be unrelated to each other. Assumption (A2) means that time of emergence is distributed according to a distribution with known shape and unknown parameters. Often it will be reasonable to assume that this distribution is symmetric around its mean. If the time to labeling is assumed to be generated by a sum of many small independent random influences, F_1 may be chosen as the normal distribution function Φ.

The measurements obtained in a labeled cell experiment target the fraction of labeled cells in dependence on time t. The fraction of labeled cells as a function of t is denoted as *the labeled cell fraction function*

$$\begin{aligned} g(t) &= P(T_1 < t < T_2) \\ &= P(T_1 < t)P(t < T_2) \\ &= F_1(t)(1 - F_2(t)) \\ &= F_1(t)\overline{F}_2(t). \end{aligned} \tag{1}$$

This is an immediate consequence of (A1). Here $\overline{F}_2(t) = 1 - F_2(t)$ denotes the survival function associated with label-cell unit survival time T_2. With (A2), we arrive at the model

$$g(t) = F_1\left(\frac{t-\mu}{\sigma}\right)\overline{F}_2(t). \tag{2}$$

This model is semiparametric, with the parametric part F_1 and the nonparametric part F_2.

We note that since analysis of the survival of the label-cell unit is of primary interest, a fully nonparametric approach, where both F_1 and F_2 are nonparametric, is not feasible, since F_1 and F_2 would then not be identifiable as separate functions. In this case, the distribution of the time to emergence could not be decoupled from the survival distribution. While the function $g(\cdot)$ could be estimated nonparametrically from the data by using any of a variety of smoothing methods, we would not be able to separate out F_2, which is crucial for a determination of the survival of the label-cell unit.

An alternative option is a fully parametric model, where both F_1 and F_2 are assumed to fall into a parametric class of models. One might choose for instance $F_1 \equiv \Phi$ and $F_2(t) = e^{-\lambda t}$, assuming an exponential distribution for survival time, leading to

$$g(t) = \Phi\left(\frac{t-\mu}{\sigma}\right) e^{-\lambda t}. \tag{3}$$

This corresponds to a nonlinear regression model with regression function g and unknown parameters μ, σ and λ. The mean survival time in this fully parametric model is given by $\theta = ET_2 = 1/\lambda$; we can view μ and σ as nuisance parameters in this model. Other parametric models could be considered as well, such as models where the Gaussian distribution for T_1 is replaced by a Gamma distribution or where the assumption of an exponential distribution for T_2 is replaced by assuming Gamma or Weibull distributions.

3. Modeling the Data

In the experiment on erythrocyte kinetics that serves as our motivating example, the fraction of labeled cells is measured repeatedly at predetermined times. These measurements correspond to

$$Y_i = \frac{\text{Number of label} - \text{cell units at time } t_i}{\text{Total number of cells at time } t_i} \ , \ i = 1, \ldots, n \ .$$

The measurements are obtained by using highly sensitive accelerator mass spectrometry (Vogel, Turteltaub and Nelson 1995), allowing measurements in the attomol range.

Our basic model assumption is

$$E(Y_i) = P(T_1 < t_i < T_2) = g(t_i). \tag{4}$$

More precisely, the measurements Y_i are related to the function g by a fixed design regression model,

(A4) $Y_i = g(t_i) + \epsilon_i,\ i = 1, \ldots, n.$

Here, the measurement errors ϵ_i are assumed to be independent and identically distributed random variables with $E\epsilon_i = 0$ and $E\epsilon_i^2 = \sigma^2 < \infty$. The measurement errors cause fluctuations in the observed data Y_i around $g(t_i)$. The regression function g is the labeled cell fraction function, corresponding to the expected value of the measurements Y_i. This function is assumed to follow the semiparametric model (1).

Furthermore, we assume that the times of measurements t_i are obtained via a smooth design density h with positive and compact support, i.e., that the times of measurement t_i are determined by

(A5) $\int_0^{t_i} h(x)dx = \frac{i-1}{n-1}$.

Under (A1)-(A5), choosing F_1 as a normal distribution and F_2 as an exponential distribution as in (3), the parameters of this fully parameterized model can be estimated consistently (as $n \to \infty$) by nonlinear least squares (see Lai, Robbins and Wei, 1979; Wu, 1981). The least squares estimates are given by

$$(\hat{\lambda}, \hat{\mu}, \hat{\sigma}) = \underset{\lambda,\mu,\sigma>0}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \Phi\left(\frac{t_i - \mu}{\sigma}\right) e^{-\lambda t_i} \right)^2. \tag{5}$$

Once $(\hat{\lambda}, \hat{\mu}, \hat{\sigma})$ have been determined numerically, we obtain the consistent estimate $\hat{\theta} = \widehat{ET_2} = \hat{\lambda}^{-1}$ for the mean label survival time $\theta = ET_2$ and $\hat{\overline{F}}_2(t) = e^{-\hat{\lambda}t}$ for the

survival function of the label-cell unit. The pdf of the time to emergence T_1 is then estimated as $\varphi\left(\frac{\cdot-\hat{\mu}}{\hat{\sigma}}\right)$. For the labeled cell fraction function g, the resulting parametric estimate is

$$\hat{g}_P(t) = \Phi\left(\frac{t-\hat{\mu}}{\hat{\sigma}}\right) e^{-\hat{\lambda}t}. \tag{6}$$

We remark here that the density $\varphi(\frac{\cdot-\hat{\mu}}{\hat{\sigma}})$ may not be compatible with $T_1 \geq 0$ and we therefore fitted also truncated versions of this distribution. However, the resulting fits for the function g were worse and therefore our fits allow that, technically, $T_1 < 0$ may occur, which might be interpreted as an immediate labeling of such a cell; in practice, this event has a low probability of less than .05 for our fitting procedures.

4. Smoothing and Monotone Smoothing

In the proposed semiparametric approach, we relax the assumption that the survival time distribution F_2 in model (1) is known. Instead, we only require F_2 to be smooth (twice continuously differentiable), according to assumption (A3), while F_1 (the distribution of time to labeling), is assumed to fall into a parametric class. Accordingly, we use smoothing methods for the estimation of the nonparametric component. Denote a smoothing algorithm applied to the scatterplot $(X_i, Y_i)_{i=1,\ldots,n}$, using bandwidth b and evaluated at x, by $S(x, (X_i, Y_i)_{i=1,\ldots,n}, b)$. One can choose from a variety of smoothing methods such as kernel estimates or smoothing splines (Simonoff, 1996).

A commonly used smoothing method is local fitting of lines by means of weighted least squares (Fan and Gijbels, 1995). This corresponds to

$$S_L(x, (t_i, Y_i)_{i=1,\ldots,n}, b) = \hat{a}_0$$

where $(\hat{a}_0, \hat{a}_1)$ are the minimizers of the weighted least squares expression

$$\sum_{i=1}^{n} \left[Y_i - (a_0 + a_1(t_i - t))\right]^2 K\left(\frac{t-t_i}{b}\right), \ t \in [0, \tau]. \tag{7}$$

Here, $K \geq 0$ is a weight function, which is usually assumed to be symmetric with support [-1,1]. A suitable class of weight functions is $K(u) = (1-u^2)^\mu$ for $\mu \geq 1$ an integer, $K(u) = 0$ for $|u| > 1$; in our application we chose $\mu = 3$. Furthermore, $b > 0$ is a bandwidth (see Fan and Gijbels, 1996 or Müller, 1988). Then (7) corresponds to fitting a line to the data in the window $[t-b, t+b]$ and evaluating it at t to obtain the smoothed estimate. We note that case weights can be added when dealing with heteroscedastic situations.

Since the design of the t_i is non-equidistant in typical cell labeling experiments, in many cases it will be desirable to implement design-adaptive local bandwidth choice (see Müller and Stadtmüller, 1987a). An alternative is bandwidth choice by cross-validation (see Eubank, 1999, or Simonoff, 1996). A basic requirement for the local weighted least squares method is that each window, for which a weighted least squares line is to be fitted, contains at least three data points. To indicate local bandwidth choice, we replace the bandwidth b appearing in the smoother S_L (7) by b_t. A reasonable option is to choose b_t as an increasing function of t for the typical case of a design which becomes increasingly sparse towards the right. In our application below we use a linearly increasing local bandwidth which is simple to implement and adjusts to the declining design density as time advances. We then obtain nonparametric estimates

$$\hat{g}_N(t) = S_L(t, (t_i, Y_i)_{i=1,\ldots,n}, b_t) = \hat{a}_0(t) \tag{8}$$

for $g(t)$.

If we choose local bandwidths b_t, the simplest method to ensure that the required minimum number of points falls into the smoothing window is to require the following for a given integer $k \geq 3$: The minimum bandwidth for smoothing at t is given by b_t =min$\{b : [t-b, t+b]$ contains at least k elements of the sequence $\{t_1, \ldots, t_n\}\}$. Such *design adaptive local bandwidth choices* are easy to implement once k has been specified and these choices automatically adapt to non-equidistant designs. As long as assumption (A5) is satisfied, there exists a constant C such that $b_t \sim C\frac{k}{n}$, and the

usual asymptotic consistency results for local linear fitting then remain valid as long as $k = k_n \to \infty, k_n/n \to 0$. This translates into $b_t \to 0, nb_t \to \infty$ for the local bandwidths b_t. Consistency follows and rates of convergence can be obtained as well. For a twice continuously differentiable function g, the local mean squared error has the usual rate $MSE(\hat{g}_N(t)) \sim Cn^{-4/5}$.

In order to model the disappearance of the label-cell units in our application, one of the components of the model will be a survival function, and we aim to estimate this survival function via nonparametric regression. As a survival function is constrained to be monotone decreasing, this requires the nonparametric estimation of monotone regression functions. Starting with a bandwidth choice b or b_t, an implementation of the locally weighted least squares nearest neighbor smoothers $S_L(t_j)$ at the design points is straightforward. If the true function is strictly monotone and smooth, well-known theorems on uniform convergence (e.g., Müller and Stadtmüller, 1987b) imply that the function estimates also will be monotone for sufficiently large samples. We therefore include a version in our applications where we just apply the smoother S_L.

To ensure monotonicity in the finite sample situation, we consider a second version where we apply an explicit monotonization step. This can be implemented via the *pool adjacent violators algorithm* (PAVA) described by Barlow et al. (1972). The combination of PAVA with smoothing has been proposed in Friedman and Tibshirani (1984). Denoting the resulting monotonized estimator, defined on the design points, by $S_P(t_j)$, a linear interpolation step is added to define final monotonized regression estimators at all points $t \in [0, \tau]$. For given t, let t_j, t_{j+1} with $t_j \leq t_{j+1}$ be the two elements of the sequence (t_i) which are closest to t. Then the final estimator is

$$S_M(t, (t_i, Y_i)_{i=1,\ldots,n}, k) = S_P(t_j) + \frac{S_P(t_{j+1}) - S_P(t_j)}{t_{j+1} - t_j}(t - t_j) \tag{9}$$

Data-adaptive choices of the bandwidths b or b_t can be based on cross-validation, plug-in methods (see Wand and Jones, 1995), or visual inspection.

5. Semiparametric Modeling

5.1 The Model

We consider now the semiparametric version of model (1), choosing $F_1 \equiv \Phi$ and F_2 as a smooth nonparametric function. Then the labeled cell fraction function becomes

$$g(t) = \Phi\left(\frac{t-\mu}{\sigma}\right)\overline{F}_2(t). \tag{10}$$

This model has the parametric components μ and σ and the nonparametric component $\overline{F}_2(t)$. The two components are intertwined and we can obtain reasonable estimates for each of them when given the other component. This motivates an iterative algorithm based on successive updating of each component, given a current value for the other component. Initialization and updating steps for both nonparametric and parametric components are as follows.

5.2 Initialization

For the *initialization step*, we need to find starting values $\mu^{(0)}, \sigma^{(0)}$ for μ and σ. These can be obtained from the fully parametric fit, by solving (5) for μ and σ. This requires a nonlinear least squares minimization step.

5.3 Nonparametric Updating Step

For the *nonparametric updating step* in the j-th iteration, we proceed as follows: Given current parameter estimates $\mu^{(j)}, \sigma^{(j)}$, define transformed data

$$Y_i^{(j+1)} = Y_i / \Phi\left(\frac{t_i - \mu^{(j)}}{\sigma^{(j)}}\right),\ i = 1, \ldots, n. \tag{11}$$

The transformed data $Y_i^{(j+1)}$ follow the model

$$Y_i^{(j+1)} = \overline{F}_2(t_i) + e_i^{(j+1)}$$

with errors $e_i^{(j+1)}$, which will be slightly correlated and slightly off center; the deviation from being centered is tied to the degree to the distance between $\mu^{(j)}, \sigma^{(j)}$ and μ, σ.

We then obtain an estimate of the survival function $\overline{F}_2$ via a monotonized nonparametric regression fit to the data $(t_i, Y_i^{(j+1)})$. The nonparametric estimate $\hat{\overline{F}}_2^{(j+1)}$ has to satisfy the constraints:

$$\hat{\overline{F}}_2^{(j+1)}(0) = 1 \text{ and } \hat{\overline{F}}_2^{(j+1)}(t_1) \geq \hat{\overline{F}}_2^{(j+1)}(t_2) \text{ for } t_1 \leq t_2. \tag{12}$$

For this restricted curve estimation problem, we employ the monotonized estimates (9) and introduce an additional truncation to ensure that $\hat{\overline{F}}_2^{(j+1)}(t) \leq 1$, for all $t \geq 0$, and $\hat{\overline{F}}_2^{(j+1)}(0) = 1$:

$$\hat{\overline{F}}_2^{(j+1)}(t) = \begin{cases} 1 & t = 0 \\ \min\{1, S_M(t, (t_i, Y_i^{(j+1)})_{i=1,\ldots,n}, b_t)\} & t > 0 \end{cases} \tag{13}$$

In case that $\hat{\overline{F}}_2^{(j+1)}(0) = 1$, but $\hat{\overline{F}}_2^{(j+1)}(t) \leq 1 - \delta$ for some $\delta > 0$ for $t > 0$, i.e., if $\hat{\overline{F}}_2^{(1)}(t)$ has a discontinuity at $t = 0$, we linearly interpolate $\hat{\overline{F}}_2^{(j+1)}(0)$ and $\hat{\overline{F}}_2^{(j+1)}(\rho)$, for a small $\rho > 0$, and thus ensure that the estimated survival function $\hat{\overline{F}}_2^{(j+1)}(t)$ is continuous. In an alternative non-monotonized version we replaced S_M in (13) by the smoother S_L (7).

5.4 Parametric Updating Step

In the *parametric updating step*, we next update $\mu,\ \sigma$ by obtaining nonlinear least squares regression estimates for $\mu,\ \sigma$ in the model

$$Z_i^{(j+1)} = \frac{Y_i}{\hat{\overline{F}}_2^{(j+1)}(t_i)} = \Phi\left(\frac{t_i - \mu}{\sigma}\right) + \tilde{e}_i^{(j+1)}$$

with new errors $\tilde{e}_i^{(j+1)},\ i = 1, \ldots, n$. The least squares estimates corresponding to the parametric updates are then given by

$$(\mu^{(j+1)}, \sigma^{(j+1)}) = \underset{\mu,\sigma>0}{\arg\min} \sum_{i=1}^{n} \left(Z_i^{(j+1)} - \Phi\left(\frac{t_i - \mu}{\sigma}\right)\right)^2. \tag{14}$$

5.5 Model Fits and Inference

The updated parameter estimates $\mu^{(j+1)}, \sigma^{(j+1)}$ obtained from (14) are then entered into (11), and $\hat{\overline{F}}_2^{(j+1)}$ is updated to $\hat{\overline{F}}_2^{(j+2)}$ via (13). Obtaining the nonlinear least squares estimators for model (14) with $\hat{\overline{F}}_2^{(j+1)}$ replaced by $\hat{\overline{F}}_2^{(j+2)}$, we obtain new updates $\mu^{(j+2)}$, $\sigma^{(j+2)}$ of $\mu,\ \sigma$, etc. The final estimates $\hat{\mu},\ \hat{\sigma}$ for $\mu,\ \sigma$ and the final nonparametric estimate of the cell survival function $\hat{\overline{F}}_2(t)$ are obtained at convergence of these iteration steps.

The resulting fit for the semiparametric model (10) becomes

$$\hat{g}_{SP}(t) = \Phi\left(\frac{t - \hat{\mu}}{\hat{\sigma}}\right) \hat{\overline{F}}_2(t). \tag{15}$$

An estimate of the mean survival time $\theta = E(T_2)$ is given by

$$\hat{\theta} = -\int_0^\tau t d\hat{\overline{F}}_2(t) = \int_0^\tau \hat{\overline{F}}_2(t) dt. \tag{16}$$

We propose a bootstrap method to obtain inference for these estimates. Several versions of the regression bootstrap were investigated. The following implementation which is based on resampling from the residuals (see, e.g., Davison, 1997) worked reasonably well. Given the model fit $\hat{g}_{SP}$ (15), we collect the residuals

$$e_i \ = \ Y_i \ - \ \hat{g}_{SP}(t_i), \qquad i = 1, \ldots, n,$$

and then resample with equal probability and replacement from the residuals $\{e_1, \ldots, e_n\}$ to obtain the resampled residuals $\{e_1^*, \ldots, e_n^*\}$ and the resampled data

$$Y_i^* \;=\; \hat{g}_{SP}(t_i) \;+\; e_i^*, \qquad i = 1, \ldots, n.$$

For each of these bootstrap samples, we run the iteration (11)-(14) to arrive at the sample of bootstrap estimates $(\hat{\mu}_k^*, \hat{\sigma}_k^*, \hat{\theta}_k^*, \hat{\overline{F}}_{2,k}^*(\cdot))$, $k = 1, \ldots, K$, where K is large; in our application, we choose $K = 1000$. The $\hat{\theta}_k^*$ are obtained via (16) from $\hat{\overline{F}}_{2,k}^*(\cdot)$. We then determine the empirical γ-quantiles q_γ^* from the empirical bootstrap distributions of $\hat{\mu}^*$, $\hat{\sigma}^*$ and $\hat{\theta}^*$ to obtain level $(1-\alpha)$-confidence intervals $[q_{\alpha/2}^*(\hat{\mu}^*), q_{1-\alpha/2}^*(\hat{\mu}^*)]$ for μ, $[q_{\alpha/2}^*(\hat{\sigma}^*), q_{1-\alpha/2}^*(\hat{\sigma}^*)]$ for σ, and $[q_{\alpha/2}^*(\hat{\theta}^*), q_{1-\alpha/2}^*(\hat{\theta}^*)]$ for the mean survival time θ.

For functions $\hat{\overline{F}}_2$ and g we obtain bootstrap estimates for $k = 1, \ldots, K$ through $\hat{\overline{F}}_{2,k}^*(\cdot)$ as defined above and

$$\hat{g}_{SP,k}^*(t) = \Phi\left(\frac{t - \hat{\mu}_k^*}{\hat{\sigma}_k^*}\right) \hat{\overline{F}}_{2,k}^*(t),$$

see (15). These bootstrap curve estimates can then be used to construct pointwise confidence intervals for the functions $g(t), \overline{F}_2(t)$ at any given argument t. For level $(1-\alpha)$-confidence intervals, this is achieved by finding the empirical γ-quantiles q_γ^* from the empirical bootstrap distributions of $\hat{\overline{F}}_2^*(t)$ and of $\hat{g}_{SP}^*(t)$ to obtain the $(1-\alpha)$-confidence intervals $[q_{\alpha/2}^*(\hat{\overline{F}}_2^*(t)), q_{1-\alpha/2}^*(\hat{\overline{F}}_2^*(t))]$ for $\overline{F}_2(t)$ and $[q_{\alpha/2}^*(\hat{g}_{SP}^*(t)), q_{1-\alpha/2}^*(\hat{g}_{SP}^*(t))]$ for $g(t)$. These confidence intervals are illustrated for $\alpha = 0.05$ in Figure 2, lower panel, and in Figure 4 (see section 6 for further details).

5.6 Remarks on Asymptotics

We conclude this section with some remarks on the asymptotics as $n \to \infty$. We consider the implementation of the model with the non-monotonized smoother S_L, focussing on the first iteration step (11) with $j = 0$. For large enough sample sizes this smoother will have the required monotonicity property. Let γ, ζ be constants with $0 < \gamma, \zeta < 1$. A linear smoother S_L as employed in $\hat{g}_N$ (8) under some basic regularization assumptions has the asymptotic consistency property $\hat{g}_N(t) - g(t) = O_p(n^{-\gamma})$, typically with $\gamma = 2/5$ for twice continuously differentiable target functions.

Assume that for the initial values for (μ, σ) as obtained in the initialization step of the iterative algorithm it holds that

$$(\mu_0, \sigma_0) \;=\; (\mu, \sigma) \;+\; O_p(n^{-\zeta}).$$

Since $E|Y_i| < \infty$, this implies

$$Y_i^{(1)} - Y_i / \Phi\left(\frac{t-\mu}{\sigma}\right) = O_p(n^{-\zeta}).$$

The linearity of the smoother and a simple calculation then imply that

$$S_L(t, (t_i, Y_i^{(1)})_{i=1,\ldots,n}, b_t) \;-\; \overline{F}_2(t) \;=\; O_p(n^{-\zeta}) \;+\; O_p(n^{-\gamma}).$$

As long as the initial estimators μ_0, σ_0 converge with a rate faster than the nonparametric rate, then the nonparametric components will be estimated with the usual rate and will also have the same asymptotic distribution. It is not necessary that the initial estimates have the $\sqrt{n}$ convergence rate. In practice, this means that the initial parameter estimates should not be too far from the true targets in order to make the method work. We can expect this to be the case in our application to red blood cell kinetics to be discussed in the next section. There the fully parametric model fit (6), on which the initial parameter estimates are based, is reasonably close to the final fit.

As for the convergence of the parametric components in the course of the iteration algorithm, we first note that the parametric updating step for the relevant nonlinear least squares estimates (14) can be expressed via iterative linear least squares, implementing a Newton-Raphson iteration within each updating step. Note that

$$Z_i^{(j+1)} \;=\; \frac{Y_i}{\overline{F}_2(t_i)} \left(1 + O_p(n^{-\zeta})\right)$$

where the remainder terms on the r.h.s. are uniformly bounded in probability. Combining this with the linearity of the updating steps shows that the usual $\sqrt{n}$-convergence rate will be achieved for $(\hat{\mu}, \hat{\sigma})$. Similarly, asymptotic normality will also be obtained under suitable regularity conditions.

6. Application to Folate Labeled Erythrocytes

In this experiment, $n = 59$ observations of the fraction of labeled erythrocytes among all erythrocytes were obtained from a healthy adult male volunteer. Details regarding the experiment can be found in Clifford et al. (1998). The subject ingested a bolus dose of ^{14}C-folic acid that equilibrated with the body folate pool during the first day. During this time a cohort of developing erythrocytes was labeled. The measurements of the labeled cell fraction, obtained with an advanced atomic mass spectroscopy technique were obtained over a time range of 200 days. Further details about the unique and ultra-sensitive measurement techniques that were employed can be found in Buchholz et al. (1999). A definining feature of this experiment is the use of very small doses (true tracer) and radioactivity levels, enabled by new measurement techniques and the innovative use of ^{14}C-folic acid as the labeling agent.

As can be seen from Figure 1, most measurements Y_i were clustered around the first five days, with little change in observed labeled cell fractions which remain close to 0 during this time period. At the later measurement times, measurements became increasingly sparse, with increasing gaps between them. Accordingly, a design-adaptive bandwidth was selected for the smoothing steps, with $b = b_t$, where $b_t = 10(1 + t/50),\ 0 \le t \le 25,\ b_t = 10(1 + t/50 + (t - 25)/25),\ 25 < t \le 50,\ b_t = 30,\ t > 50$. Figure 1 also displays the fully nonparametric estimate, obtained with local linear fitting (7).

The shape of the fitted labeled cell fraction function g turns out to be quite interesting: The first part consists of an expected delayed initial rapid increase of labeled cells over the first 15 days, reflecting the labeling process. This phase tops out at the peak, which is observed at about 20 days. It is followed by a period of decline in the labeled cell fraction, indicating that during that time period some label-cell units disappear from the circulation. While these features are as expected, this declining phase is followed by what seems to be a plateau phase, between days 70 and 125, during which

label-cell units cease to disappear. This then gives way to a final phase characterized by a marked and terminal decline in the number of label-cell units.

A comparison of parametric and semiparametric function estimates is shown in Figure 2, upper panel. The solid curve is the fit obtained from the parametric model (3), fitting a normal model for the labeling part and an exponential model for the survival part. It appears there is lack of fit in the declining phase of the labeled cell fraction function $g(t)$, especially the plateau phase is not reflected in the parametric fit. More general survival models such as Gamma or Weibull distributions for the survival part did not improve upon this situation. On the other hand, the initial rising phase appears to be sufficiently well reflected by the function Φ.

The dashed and dash-dotted function fits displayed in the upper panel of Figure 2 correspond to semiparametric fits (15) of model (10). The dashed fit is obtained by using the PAVA-based monotonized smoother (9), while the dash-dotted fit uses the straightforward (non-monotonized) version of the kernel smoother in the local weighted least squares form (8). Both semiparametric implementations are seen to provide good overall fits, with only minor differences. The semiparametric fit is seen to be clearly better than the fit provided by the parametric model. The differences between the two semiparametric fits are almost negligible. We conclude that in this application the monotonization step does not make a big difference, while the adequacy of the parametric fit is in doubt. This conclusion is supported by the 90%- and 95%-confidence intervals that are shown

For the smoothing step in the fitting of the semiparametric model, the truncation to 1 near the origin as described in (13) comes into force, as for very small values of t, the uncorrected estimate of $\hat{\overline{F}}_2$ came out with $\hat{\overline{F}}_2(t) > 1$. We note that (16) provides an estimate for the mean survival time for the label-cell units through the semiparametric model. Numerical integration for the monotonized version yielded an estimated mean survival time of $\hat{\theta} = 82.3$ days, while the semiparametric fit without monotonization led to an estimated mean survival time of $\hat{\theta} = 82.5$ days. The parameter estimates for

μ and σ in the parametric part of (10) were found to be $\hat{\mu} = 14.8$ days and $\hat{\sigma} = 8.5$ days for the monotonized version, with similar values for the non-monotonized version.

We generated 1000 bootstrap samples as described in section 5.5, for each of which the fitting procedure was repeated. The 95%−confidence intervals for the monotonized version were found to be: For μ, $[13.55, 17.46]$, for σ, $[7.49, 10.43]$, and for θ, $[79.47, 85.45]$, with similar results for the non-monotonized version.

For the parametric model the estimated survival time is well known to be $1/\hat{\lambda}$, which came out to be 93 days, quite similar to the nonparametric fit, notwithstanding the differences between the two approaches and the fact that the parametric model clearly did not provide a satisfactory fit to the data. The estimates for μ and σ in (3) were obtained as $\hat{\mu} = 16.3$ days and $\hat{\sigma} = 9.8$ days. This conforms with the slightly delayed rise in the labeled cell fraction function near the origin in Figure 2, upper panel, for the parametric model as compared to the semiparametric fits.

The bootstrap method as outlined in subsection 5.5 was also applied to obtain pointwise confidence intervals for the labeled cell fraction function $g(t)$, based on the semiparametric model implemented with the monotonized smoother. This fit corresponds to the dashed curve in the upper panel of Figure 2 and to the solid curve in the lower panel. Along with the semiparametric function estimate, the lower panel of Figure 2 also depicts pointwise 95% bootstrap confidence intervals. These are seen to be of reasonably small width, substantiating some of the features displayed by the semiparametric fit, such as the plateau.

In addition, the upper panel of Figure 3 depicts the data $(t_i, Y_i^{(\infty)})$, illustrating the nonparametric updating step (11)-(13) at convergence, along with the final estimate of the survival function $\hat{\overline{F}}_2 = \hat{\overline{F}}_2^{(\infty)}$ (13). We find that the nonparametric estimate provides a good fit to these raw data.

The lower panel of Figure 3 analogously depicts the data $(t_i, Z_i^{(\infty)})$, illustrating the parametric updating step (14) at convergence. The scatterplot is overlaid with the final estimate of the parametric part of the cell labeling function, $\Phi\left(\frac{t_i - \hat{\mu}}{\hat{\sigma}}\right)$, where

$\hat{\mu} = \hat{\mu}^{(\infty)}, \hat{\sigma} = \hat{\sigma}^{(\infty)}$ are the estimates of the parametric components at convergence.

Finally, the estimated survival function $\hat{\overline{F}}_2$ and bootstrap based pointwise 95% confidence intervals for the true survival function $\overline{F}_2$ can be viewed in Figure 4. Again, the bootstrap confidence intervals, while not simultaneous, give additional evidence for the presence of the observed features in the true survival function $\overline{F}_2$.

7. Concluding Remarks

If there exists indeed a plateau in the labeled cell fraction curve $g(t)$, as suggested by the semiparametric fit and the confidence intervals, this would indicate that there is a "stable" phase in which the label-cell units stay intact, and virtually no cell death nor removal of label occurs. This phase is followed by a fairly sudden onset of cell-label removal in the final phase. It thus appears that the survival of the label-cell complexes can be segmented into three phases:

An initial rapid decline in the survival function, followed by a constant period, which then gives way to a second rapid decline. In terms of the hazard function or force of mortality, $h(t) = \frac{dF(t)/dt}{\overline{F}(t)}$, one is reminded of the "bathtub" shape known to apply to human mortality. This suggestive structure will have to be confirmed through the analysis of data from additional individuals.

The sizeable literature on determination of the lifespan of erythrocytes focuses exclusively on the fitting of parametric survival models. Among such models, the exponential distribution with survival function $\overline{F}(t) = exp(-\lambda t)$ and mean lifetime $E(T) = 1/\lambda$ has acquired a prominent place (see Clifford et al., 1998), but other parametric models have been considered as well (Barosi et al., 1983; Haurie et al., 2000).

The resulting parametric model fits are considerably less flexible than the nonparametric approach outlined here. A parametric method will only find features in the data which are already incorporated *a priori* in the model. Such methods are thus ill suited

if a time course is not very well defined or does not fall into a preconceived class of functions. This is clearly the case for the data at hand.

For example, a plateau in the survival of the label-cell units is not easily modelled parametrically. We conclude that the fitting of non- and semiparametric models with their data-driven flexible features is well suited for the exploratory analysis of cell kinetic data, especially when the detection of new features and the generation of hypotheses is of interest.

Acknowledgements

We wish to thank two referees and an associate editor for most helpful remarks. This research was supported by NIH Grant R01 DK45939 and by NSF Grants DMS 96-25984 and DMS 99-71602.

References

Barlow, R., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions.* New York: Wiley.

Barosi, G., Baraldi, A., Bonomi, F., Cazzola, M., Dacco, M., Spriano, P., and Stefanelli, M. (1983). Competing models for the analysis of red-cell survival obtained with the Cr-51 labeling technique. *Scandinavian Journal of Hematology* **31**, 381–388.

Bergner, P.E. (1962). On the stochastic interpretation of cell survival curves. *Journal of Theoretical Biology* **2**, 279–295.

Buchholz, B.A, Arjomand, A., Dueker, S.R., Schneider, P.D., Clifford, A. (1999). Intrinsic erythrocyte labeling and attomole pharmacokinetic tracing of C-14-labeled

folic acid with accelerator mass spectrometry. *Analytical Biochemistry* **269**, 348–352.

Clifford, A.J., Arjomand, A., Dueker, S.R., Schneider, P.D., Buchholz, B.A, Vogel, J.S. (1998). The dynamics of folate acid metabolism in an adult given a small tracer dose of ^{14}C-folic acid. *Advances in Experimental Medicine and Biology* **445**, 239–251.

Davison, A.C. (1997). *Bootstrap Methods and Their Application.* Cambridge University Press.

Eubank, R.L. (1999). *Spline Smoothing and Nonparametric Regression.* New York: Dekker (Second Edition).

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications.* London: Chapman and Hall.

Friedman, J. and Tibshirani, R. (1984). The monotone smoothing of scatterplots. *Technometrics* **26**, 243–250.

Haurie, C., Dale, D.C., Rudnicki, R., Mackey, M.C. (2000). Modeling complex neutrophil dynamics in the grey collie. *Journal of Theoretical Biology* **204**, 505–519.

Hellerstein, M.C. (1999). Measurement of T-cell kinetics: recent methodological advances. *Immunology Today* **20**, 438–441.

Lai, T. L., Robbins, H., Wei, C. Z. (1979). Strong consistency of least squares estimates in multiple regression II. *Journal of Multivariate Analysis* **9**, 343–361.

Macallan, D. C., Fullerton, C. A., Neese, R. A., Haddock, K., Park, S. S. and Hellerstein, M. K. (1998). Measurement of cell proliferation by labeling of DNA with stable isotope-labeled glucose: Studied in vitro, in animals and in humans. *Proceedings of the National Academy of Sciences USA* **95**, 708–713.

Müller, H. G. (1988). *Nonparametric Regression Analysis of Longitudinal Data.* New York: Springer.

Müller, H. G. and Stadtmüller, U. (1987a). Variable bandwidth kernel estimators of regression curves. *Annals of Statistics* **15**, 182–201.

Müller, H. G. and Stadtmüller, U. (1987b). Estimation of heteroscedasticity in regression analysis. *Annals of Statistics* **15**, 610–625.

Shemin D. and Rittenberg D. (1946). The life span of the human red blood cell *Journal of Biological Chemistry* **166**, 627–636.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics.* New York: Springer.

Vogel J.S., Turteltaub K.W., and Nelson, D.E. (1995). Accelerator mass spectrometry isotope quantification at attomol sensitivity. *Analytical Chemistry* **67**, 353A–359A.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* London: Chapman and Hall.

Wu, C. F (1981). Asymptotic theory of nonlinear least squares estimation. *Annals of Statistics* **9**, 501–513.

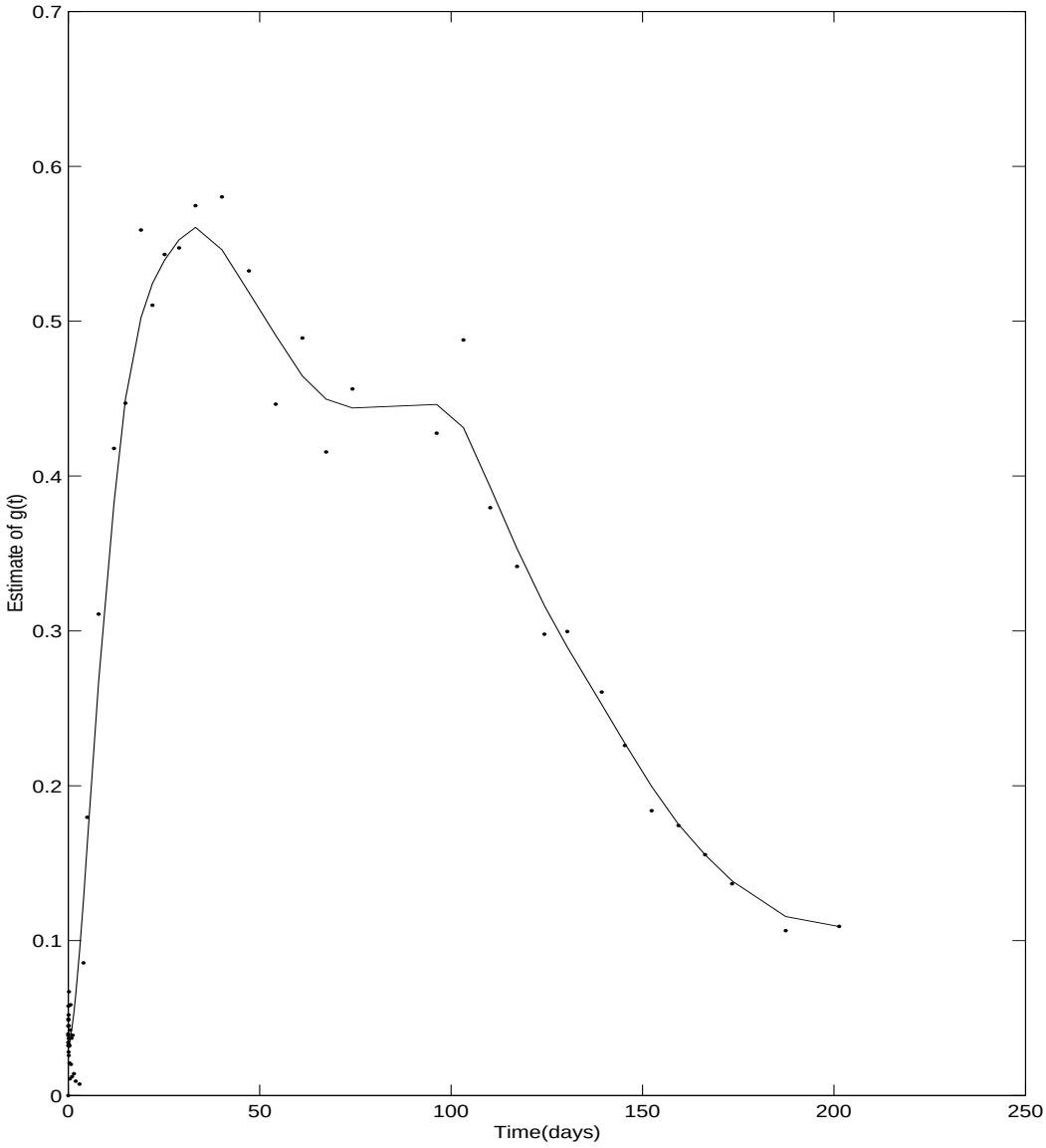

Figure 1. Original measurements (dots) and fully nonparametric local least squares fit (7) for the labeled cell fraction function $g(t)$ in the ^{14}C-folic acid erythrocyte labeling experiment.

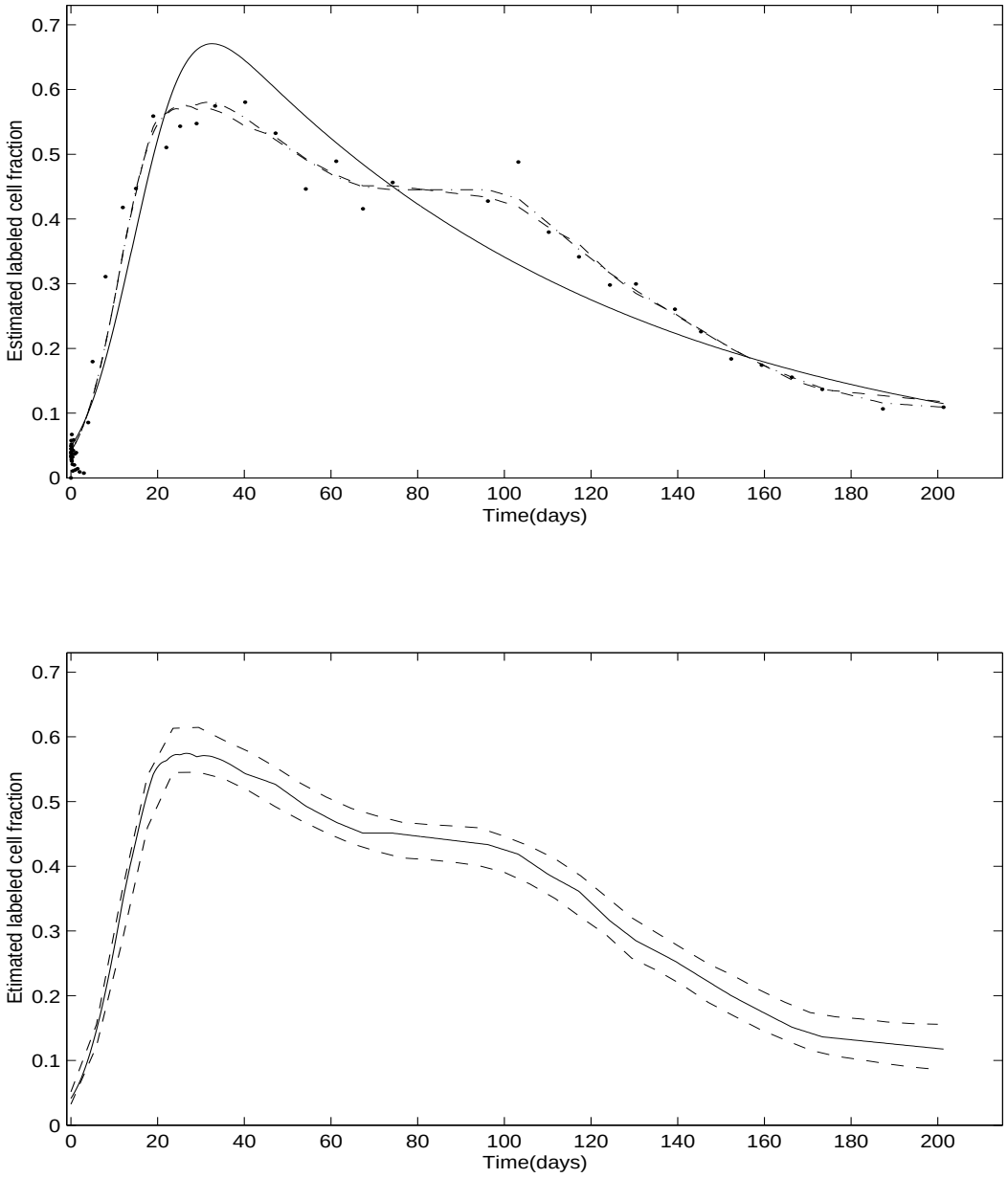

Figure 2. *Upper Panel:* Original measurements (dots) and parametric model fit (3) (solid), plus two semiparametric model fits (10). All three model fits target the labeled cell fraction function. The dashed fit has been obtained with the monotonized version using the pool adjacent violators algorithm (PAVA), while the dash-dot fit is based on a non-monotonized nonparametric fit. *Lower Panel:* Monotonized fit (dashed function in upper panel) here depicted as solid curve, with pointwise 95% bootstrap confidence bands (dashed).

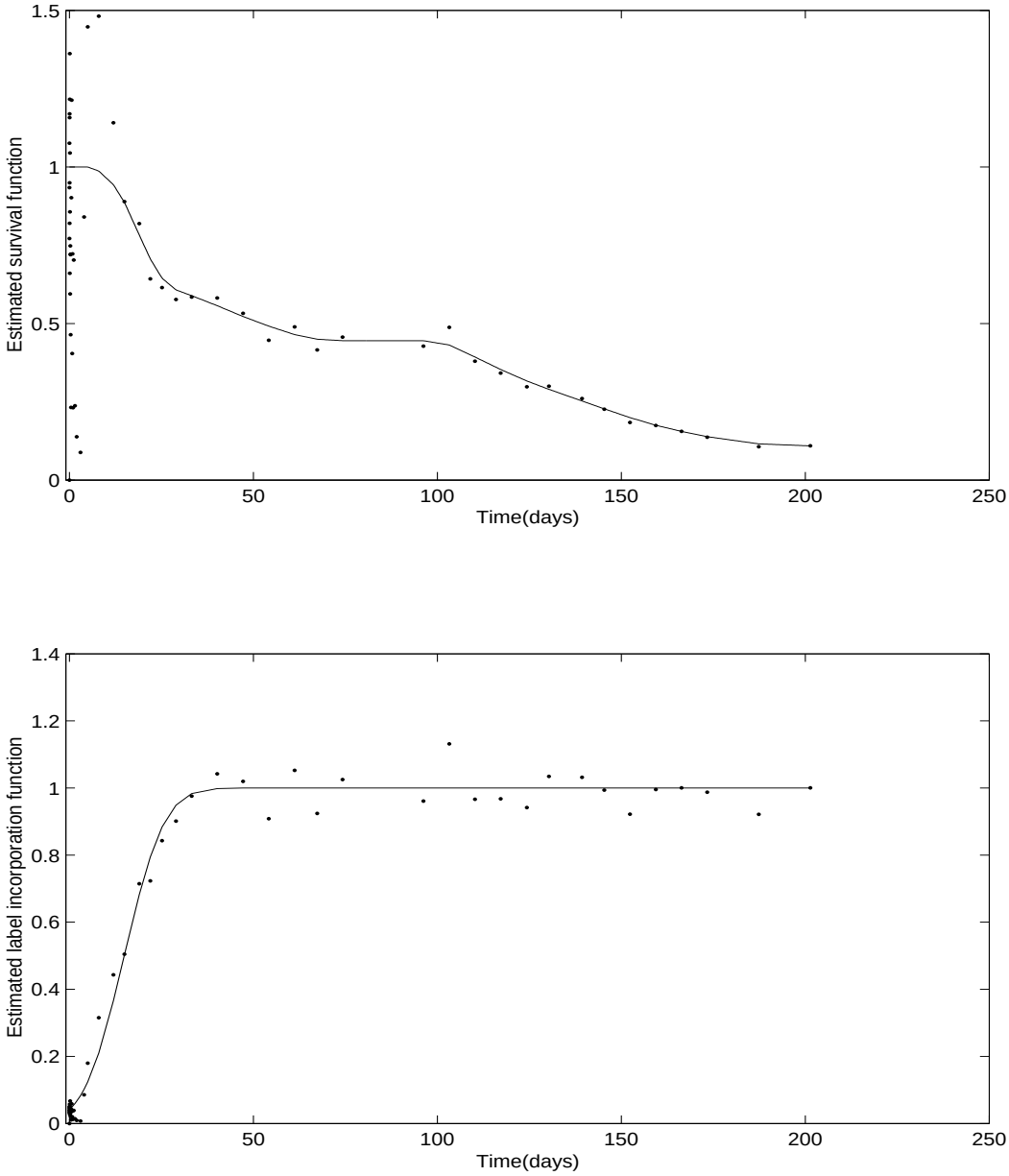

Figure 3. Estimation of the survival function $\overline{F}_2$ and the parametric label incorporation function F_1 (2) at the last iteration step. *Upper Panel:* Estimated monotonized survival function $\hat{\overline{F}}_2$ and corresponding raw data (11), illustrating the nonparametric updating step at the last iteration. *Lower Panel:* Estimated label incorporation function with raw data (14), illustrating the parametric updating step at the last iteration.

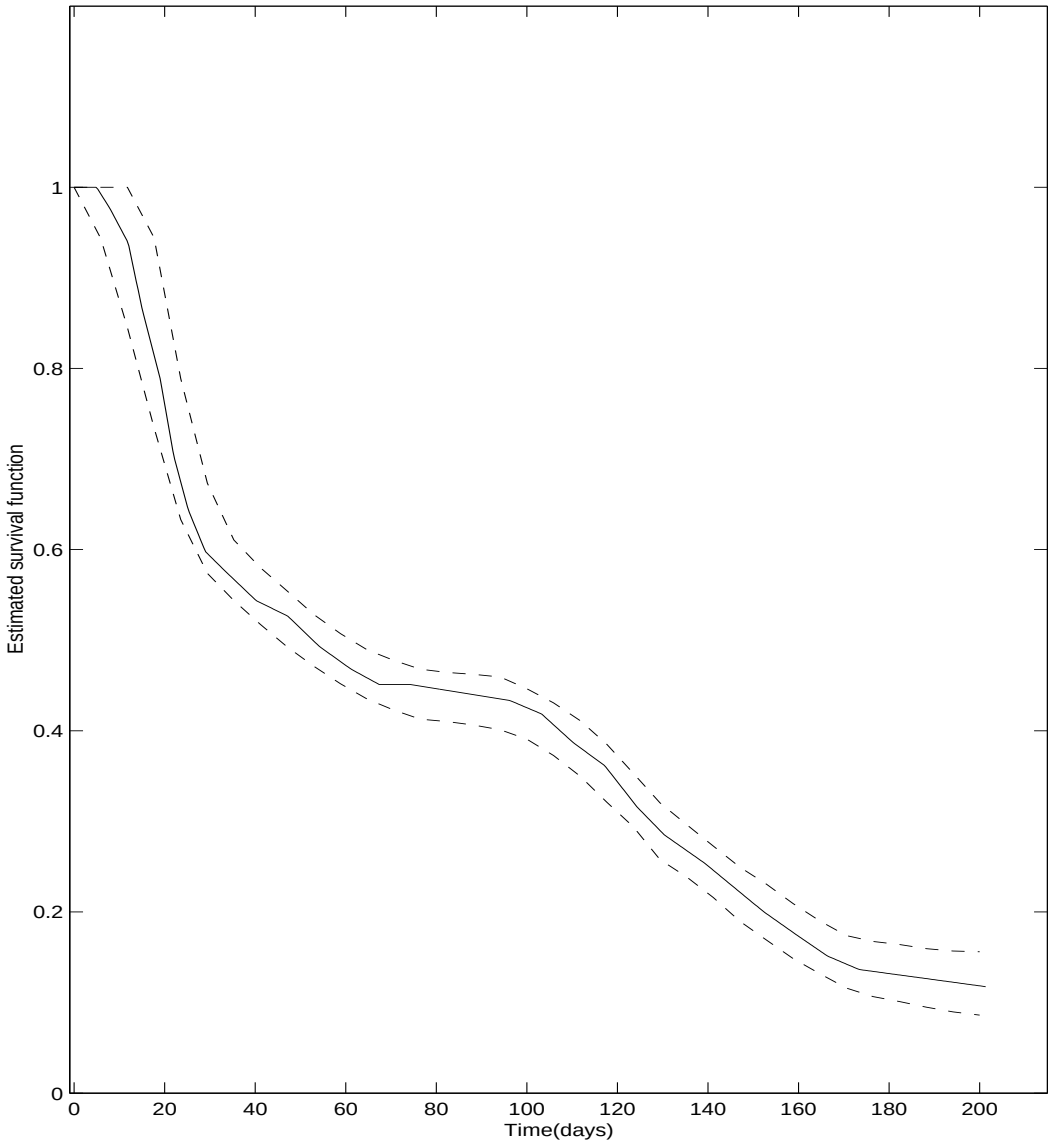

Figure 4. The estimated survival function estimate $\hat{\overline{F}}_2$, same as in upper panel of Figure 3, with pointwise 95% bootstrap confidence bands (dashed).