Lifetime Data Analysis, 1, 131–144 (1995)

Gaussian Models for Degradation Processes-Part I: Methods for the Analysis of Biomarker Data

KJELL A. DOKSUM
University of California, Berkeley

SHARON-LISE T. NORMAND sharon@orfeo.med.harvard.edu
Department of Health Care Policy, Harvard Medical School, Boston, MA 02115

Received February, 1995; accepted March 22, 1995

Abstract. We present two stochastic models that describe the relationship between biomarker process values at random time points, event times, and a vector of covariates. In both models the biomarker processes are degradation processes that represent the decay of systems over time. In the first model the biomarker process is a Wiener process whose drift is a function of the covariate vector. In the second model the biomarker process is taken to be the difference between a stationary Gaussian process and a time drift whose drift parameter is a function of the covariates. For both models we present statistical methods for estimation of the regression coefficients. The first model is useful for predicting the residual time from study entry to the time a critical boundary is reached while the second model is useful for predicting the latency time from the infection until the time the presence of the infection is detected. We present our methods principally in the context of conducting inference in a population of HIV infected individuals.

Keywords: Degradation processes, Gaussian process models, Inverse Gaussian distribution, regression models, CD4, HIV, residual time, latency time.

1. Introduction

Biomarker series are important health indicators that represent the immunological progression of a disease. For example, in HIV infected individuals, typical biomarkers will be immune system markers such as CD4 counts or CD8 counts measured repeatedly over time. A decrease in CD4 counts over time is an indicator of the immunological progression of HIV infection, and consequently, of declining health. In this paper we consider two models for biomarker series based on Gaussian processes. Both models we present incorporate covariates, such as age and sex, that may influence failure experience, but are not health status measurements that can be influenced by treatments or interventions.

It is important to note that biomarker processes fall outside the usual regression paradigm because they can not be used as covariates if the response is life time and the main interest is in treatment efficacy. Thus, once a treatment has been assigned, it is not advisable to condition on biomarker process values because such conditioning can mask important health status differences. Consider, for example, the degradation of the immune system of HIV infected individuals. Data on such processes are available in numerous studies, including the San Francisco Mens Health Study (Winkelstein et al, 1987), the New York Blood Center Study (Cleary et al, 1988), the Multicenter AIDS Cohort Study (Kaslow et al, 1987), the Multicenter Canadian AZT Trial (Montaner et al, 1993), and the Toronto Sexual

Contact Study (Calzavara et al, 1993). CD4 counts will tend to decline once the individual is infected with the HIV virus. Two treatment groups may exhibit the same survival experience when we condition on the CD4 values even though one group has average survival twice that of the other group. This could easily happen because an effective treatment would yield higher CD4 counts as well as longer survival than an unsuccessful one. However, comparing subjects with the same CD4 count from the two groups could hide improved survival experience. Thus it is important to use models that give joint distributions for the biomarker process values and event times. The Gaussian process approach provides such models.

The question of how to model the joint distribution of event times and biomarker process values has recently been discussed by Lefkopoulou and Zelen (1995), Jewell and Kalbfleisch (1992), as well as by several of the authors in the volume AIDS Epidemiology edited by Jewell et al (1992). Our approach, where we model the biomarker process as a Gaussian process $X(t)$ which is associated with the event time T, is similar to the approach presented in Berman (1990; 1994) and in Doksum and Hóyland (1992).

In this article we develop two stochastic models that connect biomoarker processes, event time, and covariates of interest. In Section 2 we ask: if somebody walks into a clinic today and discovers for the first time that he/she is HIV positive, then given todays biomarker value and other covariate values, what is the distribution of the time until that individual's biomarker reaches a critical value. We model the biomarker series of infected subjects as a Weiner process with a drift parameter which depends on time and covariates. We specify the likelihood function for the biomarker increments and derive the maximum likelihood estimators for the regression parameters as well as their standard errors. The maximum likelihood estimator of the expected time until the critical value is first reached is derived and its corresponding standard error is provided.

In Section 3 we ask: if somebody walks into a clinic today and discovers for the first time that she/he is HIV positive, then given todays biomarker value and other covariate values, what is the distribution of the time from HIV infection. In Section 3, we develop a model for the biomarker values of infected subjects, which incorporates biomarker information from a group of uninfected individuals as well as covariate information. Unlike in Section 2, our model is based on a stationary Gaussian process and is not conditional on the initial biomarker value. We define latency time as the time from HIV infection until the time infection is detected. By making use of Bayes' theorem, we derive the conditional distribution of the length of the latency time given the individual's biomarker and covariate information in Section 3.5.

Finally, in Section 4 we summarize our results and indicate how the models presented in Sections 2 and 3 are related.

In a subsequent article (Part II), we apply our methods to the San Francisco Mens Health Study cohort (Winkelstein et al, 1987). This data set contains 381 HIV seroprevalent men, 549 HIV seronegative mean, and 44 men who seroconverted during the study period which spanned the interval between March 1, 1984 and March, 1991. We model the joint distribution of CD4 counts, covariates and event times, and we provide estimates of regression coefficients as well as their corresponding standard errors. Moreover, we employ model diagnostic techniques to check the appropriateness of the models.

2. The Wiener Process Approach. Predicting Residual Time

2.1. *The Likelihood. Estimation of Parameters*

Let $X_0(t)$ denote the value of the biomarker process at time $t \geq t_0$, where t_0 denotes the time the infection was detected. In many medical applications $X_0(t)$ is the level of a biomarker process such as a CD4 blood cell count and t_0 is the time HIV is first detected. Using these conventions, we may assume $X_0(t_0) > 0$. Also, because $X_0(t_0)$ is the initial value, its level can not be affected by treatments and consequently $X_0(t_0)$ is an ancillary covariate. Without loss of generality we set $t_0 = 0$. Our analysis will be conditional on the level $X_0(0)$ and will be based on the process

$$X(t) = \log[X_0(t)/X_0(0)], \qquad t \geq 0.$$

We assume that for a given individual in a sample of n subjects we observe $T_1, \ldots, T_k$, $X_0(0), X(T_1), \ldots, X(T_k), \mathbf{Z}$, where $T_1, \ldots, T_k$ are observation times, $X(T_1), \ldots, X(T_k)$ are the biomarker process values at these times, and $\mathbf{Z}$ is a $(d \times 1)$ covariate vector consisting of covariates such as age, sex, etc.

We write the joint density of $\{(T_j, X(T_j)); j = 1, \ldots, k\}$ given $X_0(0) = x_0$ and the covariates $\mathbf{Z} = \mathbf{z}$ as

$$f_{\boldsymbol{\theta}}(t_1, \ldots, t_k, x_1, \ldots, x_k \mid x_0, \mathbf{z}) = f_{\boldsymbol{\theta}}(x_1, \ldots, x_k \mid x_0, \mathbf{z}, t_1, \ldots, t_k) \times f(t_1, \ldots, t_k \mid x_0, \mathbf{z})$$

where $\boldsymbol{\theta}$ is a parameter vector that determines the model. This notation implies that $T_1, \ldots, T_k$ are uninformative; that is, their joint distribution does not depend on $\boldsymbol{\theta}$. In fact, in many applications, $T_1, \ldots, T_k$ are appointment times assigned by clinics. Our likelihood is now

$$L(\boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(x_1, \ldots, x_k \mid x_0, \mathbf{z}, t_1, \ldots, t_k).$$

We assume the model where, given $X_0(0) = x_0 > 0$, $X(t)$ is a Wiener process with drift $-\eta t$ and diffusion constant δ^2. That is, $X(t), t \geq 0$, is an independent increment process with $X(0) = 0$, mean $E(X(t)) = -\eta t$, and each increment $X(t) - X(s), 0 < s < t$, has variance $(t - s)\delta^2$. Normand and Doksum (1994) show that a linear drift model is reasonable for calibrated log CD4 counts. We call the negative of the slope of the mean of the biomarker process, $\frac{-d}{dt}E(X(t)) = \eta$, the *degradation rate*.

To obtain a simple expression for the likelihood, we introduce the biomarker increments

$$Y_j = X(t_j) - X(t_{j-1}) \overset{\text{indep.}}{\sim} \mathrm{N}\left((t_j - t_{j-1})\eta, (t_j - t_{j-1})\delta^2\right)$$

for $j = 1, \ldots, k$ and $t_0 = 0$. Note that $(Y_1, \ldots, Y_k)$ is a one-to-one function of $X(t_1), \ldots, X(t_k)$.

For the i^{th} individual of a sample of n subjects, we use the notation $\mathbf{t}_i = (t_{i1} - t_{i0}, t_{i2} - t_{i1}, \ldots, t_{ik_i} - t_{ik_i-1})'$ for the vector of observation time increments, $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ik_i})'$ for the vector of biomarker increments, and $\mathbf{Z}_i$ for the $d \times 1$ vector of patient covariates. We

assume that η_i depends linearly on the covariates and write $\eta_i = \mathbf{Z}_i' \boldsymbol{\beta}$ where $\boldsymbol{\beta}$ is a $(d \times 1)$ vector of degradation regression coefficients. Then, the log-likelihood for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \delta^2)$ is proportional to

$$N \log \delta^2 + \sum_{i=1}^{n} \frac{(\mathbf{Y}_i - Z_{ti}\,\boldsymbol{\beta})' D_i^{-1} (\mathbf{Y}_i - Z_{ti}\,\boldsymbol{\beta})}{\delta^2} \tag{1}$$

where $N = \sum_{i=1}^{n} k_i$, Z_{ti} is the $k_i \times d$ matrix formed by the product $\mathbf{t}_i \times \mathbf{Z}_i'$, and D_i is a $k_i \times k_i$ diagonal matrix with the j^{th} diagonal entry $t_{ij} - t_{ij-1}$. Note that (1) can be maximized explicitly and the maximum likelihood estimates are

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} Z_{ti}' D_i^{-1} Z_{ti} \right)^{-1} \sum_{i=1}^{n} Z_{ti}' D_i^{-1} \mathbf{Y}_i \tag{2}$$

$$\hat{\delta}^2 = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{Y}_i - Z_{ti}\hat{\boldsymbol{\beta}})' D_i^{-1} (\mathbf{Y}_i - Z_{ti}\hat{\boldsymbol{\beta}}). \tag{3}$$

$\hat{\boldsymbol{\beta}}$ and $\hat{\delta}^2$ can be used to estimate and test effects of covariates using

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = \hat{\delta}^2 \left(\sum_{i=1}^{n} Z_{ti}' D_i^{-1} Z_{ti} \right)^{-1}$$

as an estimate of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$. Of particular interest would be the case where one of the covariates corresponds to a treatment variable, such as AZT. In this case our results provide an estimate (and corresponding standard error) of the treatment regression coefficient.

2.2. *Predicting Residual Time*

Suppose that interest centers on predicting the time from study entry to the time a boundary is crossed. Let c_r denote the (critical) boundary for the biomarker process. For example, if interest centers on estimating the time at which an HIV infected individual has AIDS, an important critical boundary for the CD4 process would be $c_r = 200$ (see CDC (1993)). For the transformed process, $X(t) = \log[X_0(t)/X_0(0)]$, where $X_0(t)$ is the biomarker count at time t, we let $c = \log[c_r/X_0(0)]$. Let T denote the (residual) time from zero until the process $X(t)$ crosses c. In our model where, given $X_0(0) = x_0 > 0$, $X(t)$ is a Wiener process with drift $-\eta t$, the conditional distribution of T given $X_0 = x_0$ is inverse Gaussian $IG(t \mid \mu, \lambda)$ with parameters $\mu = c/\eta$ and $\lambda = c^2/\delta^2$. The density is

$$f(t \mid x_0) = \sqrt{\frac{\lambda}{2\pi t^3}} \exp\left\{ -\frac{\lambda}{2\mu^2} \frac{(t-\mu)^2}{t} \right\}, \quad t > 0,\ \mu > 0,\ \lambda > 0.$$

This distribution has many nice properties; see Chhikara and Folks (1989). In particular, we have

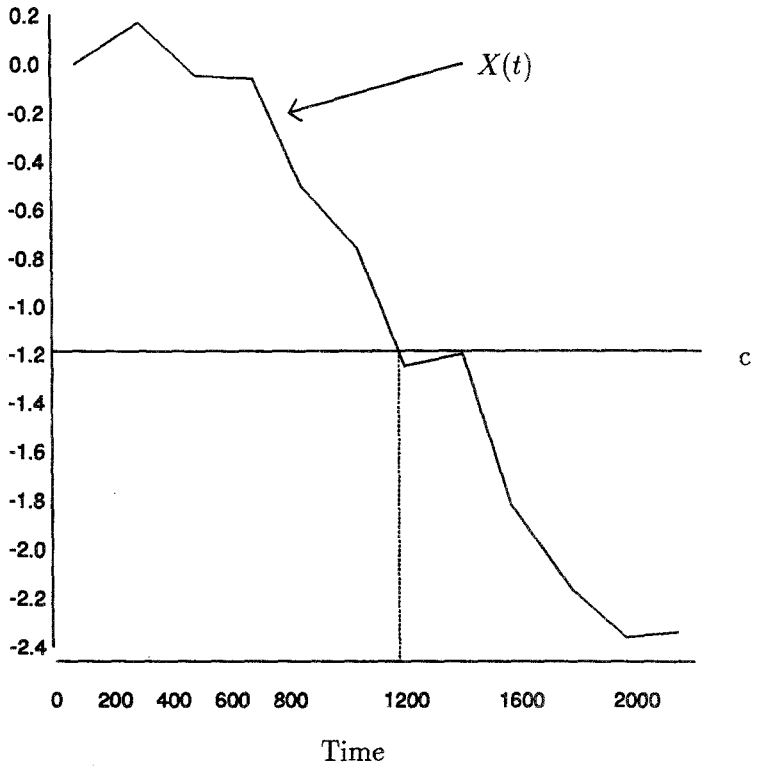

Figure 1. Predicting Residual Time T. The y-axis represents $X(t) = \log[X_0(t)/X_0(0)]$ for an HIV positive individual, where the x-axis represents visit times (in days), $X_0(t) = \text{CD4}(t)$. The solid horizontal line represents the critical boundary defined as $c = \log[200/X_0(0)]$ and the intersection of the dotted vertical line with the x-axis indicates the estimated expected time, $E(T \mid X_0(0))$, until the CD4 count first crosses the boundary.

PROPOSITION 1 *For a subject with covariate vector* $\mathbf{Z}_i$ *and initial marker value* $X_{0i}(0) = x_{0i}$, *the conditional distribution of the residual time* T_i *until the biomarker process crosses some boundary* c_r *is* $IG(t|\mu_i, \lambda_i)$, *where* $\mu_i = c_i/\eta_i$, $\eta_i = \mathbf{Z}_i'\boldsymbol{\beta}$, $\lambda_i = c_i^2/\delta^2$, *and* $c_i = \log[c_r/x_{0i}]$.

Using the results of Section 2.1, the maximum likelihood estimator of the expected time until the biomarker process first crosses the boundary c_r is

$$\hat{\mu}_i = \frac{c_i}{\mathbf{Z}_i'\hat{\boldsymbol{\beta}}} \text{ with } \widehat{\text{Var}}(\hat{\mu}_i) = \frac{c_i^2 \mathbf{Z}_i' \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) \mathbf{Z}_i}{\left(\mathbf{Z}_i'\hat{\boldsymbol{\beta}}\right)^4} \tag{4}$$

where $\widehat{\text{Var}}(\hat{\mu}_i)$ is derived using the delta-method.

3. The Stationary Gaussian Process Approach. Modeling Latency Time

3.1. *The Likelihood. Estimation of Parameters*

In this section we will, in addition to the biomarker process $X_0(t)$ for infected subjects, consider a process $\tilde{X}_0(t)$ of individuals not (yet) infected. We assume a model where $W(t) = \log \tilde{X}_0(t)$, $t \geq 0$, is a stationary Gaussian process with

$$W(t) \sim N(\mu, \sigma^2) \text{ and } \mathrm{Cov}(W(t'), W(t)) = r(t - t')\sigma^2, \qquad t' < t$$

for some unknown correlation function $r(\cdot)$. Because $W(t)$ represents the biomarker process *before* infection, we assume this process is a random function subject to normal fluctuations around a constant level. This is in contrast to the biomarker process $X(t)$ for the infected individuals (Section 2), in which we postulated that $X(t)$ is the sample function of a Weiner process with a linear drift, $-\eta t$. In this section we denote $\log X_0(t)$ by $V(t)$ and use a model in which $V(t)$ is distributed as $W(t)$ minus a drift. In Section 4 we relate the Wiener process model in Section 2 to the model of this section.

Let s_0 be the time of infection and let S denote the latency time, that is, the time from s_0 until the time s_0+S when the infection is detected. S is assumed to be independent of $W(t)$. Using the stationarity assumption we can show that, without loss of generality, we can assume $s_0 = 0$. Our model for the biomarker process $X_0(t)$ assumes that $V(t) = \log X_0(t)$, $t \geq 0$, has the same distribution as $W(t) - \Delta t$, $t \geq 0$. In this model we refer to $\frac{-d}{dt}E(V(t)) = \Delta$ as the degradation rate.

In the remainder of this Section we focus on the HIV example where $W(t)$ and $V(t)$ represent the logarithm of the CD4 counts of HIV negative and positive individuals respectively (recall from Section 2 that a linear drift model is reasonable for calibrated CD4 counts).

For each of n_P HIV positive individuals we observe $\{V(s_j + S); j = 0, 1, \ldots, k\}$ and $\mathbf{Z}_P$, where $\{s_j + S\}$ are the observation times, $\{V(s_j + S)\}$ are the CD4 values at these times, with $s_0 = 0$, and $\mathbf{Z}_P$ is a $(d_P \times 1)$ covariate vector. In addition, for each member of an independent sample of n_N HIV negative subjects from the same population we observe $\{W(u_j); j = 1, \ldots, k'\}$, and $\mathbf{Z}_N$, where u_j and $W(u_j)$ are, respectively, the observation time and the (nondegraded) CD4 value at the jth time, and $\mathbf{Z}_N$ is a $(d_N \times 1)$ covariate vector.

The likelihood of the data $(V, W) = \{(V(S + s_j), W(u_{j'})\}$ is

$$\prod_{n_N}\prod_{n_P} f_W(w|\mathbf{z}_N) \int_0^\infty f_V(v|\mathbf{z}_P, s) f_S(s)\, ds, \tag{5}$$

where the product is taken over all n_N HIV negative individuals and n_P HIV positive individuals. In (5), $f_W(w|\mathbf{z}_N)$ is the multivariate Normal density with mean μ and covariance matrix $r(|u_j - u_l|)\sigma^2$, $\{j, l = 0, 1, \ldots k\}$; $f_V(v|\mathbf{z}_P, s)$ is a multivariate Normal density with mean $\mu - (s_j + s)\Delta$, and covariance matrix $r(|s_j - s_l|)\sigma^2$, $\{j, l = 0, 1, \ldots k\}$; and $f_S(s)$ is the latency time probability density function.

In order to model the dependence of the biomarker series on the covariates, we introduce the parameterization $\Delta = \mathbf{Z}'_P\,\beta$ and $\mu = \mathbf{Z}'_N\,\alpha$ where β and α are the regression parameters for the HIV positive and HIV negative individuals respectively. One approach to obtaining

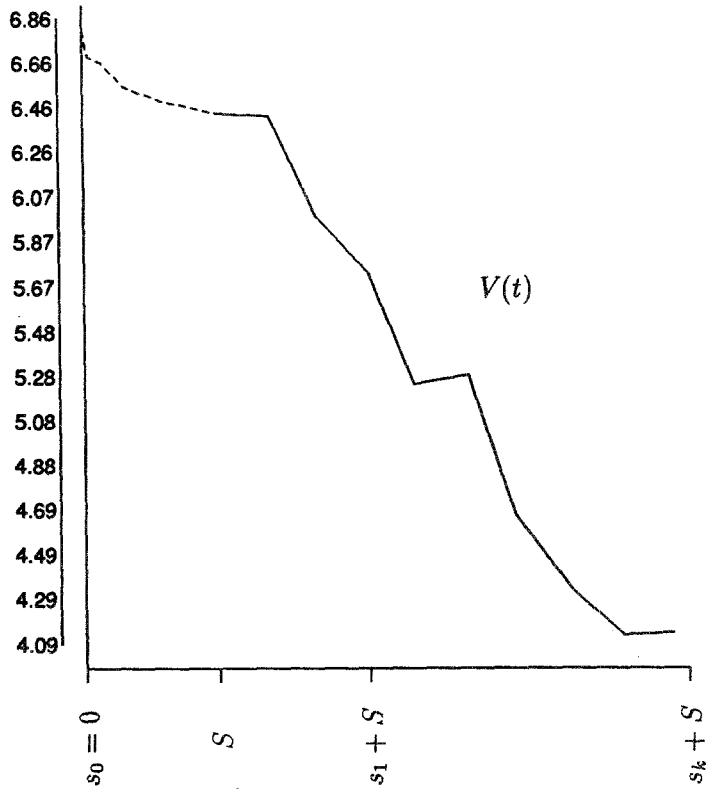

Figure 2. Modeling Latency Time. The y-axis represents $V(t) = \log X_0(t)$ and the x-axis represents time. The solid line in the figure indicates the logarithm of the observed CD4 counts from the first visit to the kth visit for an infected subject. The dashed line represents the unobserved portion of the series from time of infection ($s_0 = 0$) to the time when infection is detected ($s_0 + S = S$).

estimators of Δ and μ is to parameterize $f_S(s)$ and $r(t)$, and then maximize the likelihood in (5). Possible parameterization choices for the latency density, $f_S(s)$, and for the correlation function, $r(t)$, are $f_S(s) = \xi^{-1} \exp\{-\xi^{-1}s\}$ where ξ is the mean of the latency time S (see Berman, 1990), and $r(t)$ as a first order auto-regressive correlation function, $\exp\{-\tau|t|\}$ where $\tau > 0$.

3.2. Ad Hoc Estimates for a Semiparametric Model

Alternatively, to maximize the likelihood, consistent "ad hoc" estimates of the regression coefficients are available which can be used on their own without specifying a parametric form of the correlation function $r(t)$, or as a first step in an iteration procedure to find the maximum likelihood estimate. Because of independence across subjects, if we consider

only the first observation of the HIV negative subjects (j = 1), then the maximum likelihood estimator of α based on $(\mathbf{W}_1, Z_N)$ is the least squares estimator

$$\hat{\alpha}_1 = \left(Z_N' Z_N\right)^{-1} Z_N' \mathbf{W}_1 \tag{6}$$

where Z_N is the $n_N \times d_N$ matrix of covariates for the HIV negative group and $\mathbf{W}_1$ is the $n_N \times 1$ vector of log CD4 counts at time $j = 1$.

To estimate β, we use the following result:

PROPOSITION 2 *The distribution of* $Y_j = V(s_j + S) - V(s_{j-1} + S)$ *is Normal with mean* $(s_j - s_{j-1})\Delta$ *and variance* $2\sigma^2[1 - r(s_j - s_{j-1})]$.

Proof. Condition on $S = s$, and note that because S is independent of W, the conditional distribution of

$$Y_j = W(s_j + s) - W(s_{j-1} + s) - (s_j - s_{j-1})\Delta$$

is the stated Normal distribution. Because the conditional distribution is the same for all s, this is also the unconditional distribution by the iterated expectation theorem.

Using Proposition 2, if we set $\check{Y}_j = \frac{Y_j}{(s_j - s_{j-1})}$, then $\mathrm{E}(\check{Y}_j) = \mathbf{Z}_P' \beta$. Because the $\check{Y}_j$'s are independent across individuals, then if we consider only the first increment, a natural estimator of β is

$$\hat{\beta}_1 = \left(Z_P' Z_P\right)^{-1} Z_P' \check{\mathbf{Y}}_1 \tag{7}$$

where Z_P is the $n_P \times d_P$ matrix of covariates for the HIV positive group and $\check{\mathbf{Y}}_1$ is the $n_P \times 1$ vector of log CD4 process increment slopes, $\frac{Y_1}{(s_1 - s_0)}$, between visit times 0 and 1.

Remark 1. The estimator for α specified in equation (6) based on the log CD4 values at the first visit is $\sqrt{n_N}$ consistent. We can get another estimator for α using the log CD4 values at the second visit, $\hat{\alpha}_2$, and generally, for the jth visit, $\hat{\alpha}_j$. A natural weighted estimator for α is then $\sum_{j=1}^{m} n_{N_j} \hat{\alpha}_j / \sum_{j=1}^{m} n_{N_j}$ where n_{N_j} is the number of HIV negative subjects contributing an observation at visit j and m is the number of visits at which we can compute $\hat{\alpha}$.

The estimator for β is $\sqrt{n_P}$ consistent provided the correlation function $r(t)$ is bounded away from one for t bounded away from zero, and provided the between-visit times, $s_j - s_{j-1}$, are bounded away from zero. The estimator $\hat{\beta}_1$ specified in equation (7) is obtained from the increment in CD4 values between the first and second visit times. We can also use the increments in CD4 values between the $(j-1)$st and jth visit times, thereby obtaining $\hat{\beta}_j$. As with α, a natural weighted estimator for β is then $\sum_j n_{P_j} \hat{\beta}_j / \sum_j n_{P_j}$, where n_{P_j} is the number of HIV positive subjects contributing observations at visit times $j - 1$ and j.

Remark 2. An ad hoc estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n_N - 2} \sum_{i=1}^{n_N} \left(\mathbf{W}_{i1} - Z_{Ni}' \hat{\beta}\right)^2$$

where $\mathbf{W}_{i1}$ is the log CD4 count at time $j = 1$ for the i^{th} subject and $\mathbf{Z}_{Ni}$ is the $d_N \times 1$ vector of covariates for the i^{th} subject.

Remark 3. The advantage of this "ad hoc" approach, over maximizing the likelihood in (5), is that it does not depend on specifying a parametric form for the covariance function r. Thus this approach is semiparametric.

3.3. *Estimating Mean Latency Time*

Our approach to modeling latency time is a modification and extension to include covariates of the method proposed by Berman (1990). In the previous subsection, we saw how to estimate the parameters α and β in the stationary process model using independent samples of HIV negative and positive subjects. Here we focus on the distribution of the latency time S for HIV positive individuals. We will consider a model in which $\mu = \mathbf{Z}'_N \alpha$ represents the mean log CD4 process level of an uninfected subject with covariate vector $\mathbf{Z}_N$, while $\Delta = \mathbf{Z}'_P \beta$ represents the rate of decline in the mean log CD4 process of an infected individual with covariate vector $\mathbf{Z}_P$. In what follows, we assume $\mathbf{Z}_P$ contains $\mathbf{Z}_N$ in addition to one or more treatment variables.

Our unconditional (before conditioning on $(\mathbf{Z}_N, \mathbf{Z}_P)$) model is

$$V(s_j + S) = W_0(s_j + S) + \mathbf{Z}'_N \alpha - (s_j + S)\mathbf{Z}'_P \beta, \qquad j = 0, \ldots, k \tag{8}$$

where $W_0(t) = W(t) - \mu$ is a stationary Gaussian process with mean zero and covariance function $r(|t - s|)\sigma^2$. Moreover, we assume that $W_0(t)$, $t \geq 0$, and $(S, \mathbf{Z}_N, \mathbf{Z}_P)$ are independent. In the model specified in (8), S denotes the unconditional latency time for a subject drawn from the population.

Rearranging (8) to form an equation for the latency time S and introducing the subscript i for the i^{th} HIV positive subject with covariate vectors $\mathbf{Z}_{Ni}$, $\mathbf{Z}_{Pi}$, we can write our conditional model for S_i given $\mathbf{Z}_{Ni}, \mathbf{Z}_{Pi}$ as

$$S_i = \Delta_i^{-1}[W_i(s_{ij} + S_i) - V_i(s_{ij} + S_i) - \Delta_i s_{ij}].$$

Because $E(W_i(s_{ij} + S_i)) = \mu_i = \mathbf{Z}'_{Ni} \alpha$, it follows that

$$E(S_i) = \Delta_i^{-1}\left[\mu_i - E(V_i(s_{ij} + S_i)) - \Delta_i s_{ij}\right] \tag{9}$$

$$= \Delta_i^{-1}\left[E[V_i(s_{ij})] - E[V_i(s_{ij} + S_i)]\right], \qquad j = 0, \ldots, k_i. \tag{10}$$

Note that (10) represents the expected latency time as the expected drop in the log CD4 level from time s_{ij} to time $s_{ij} + S_i$, divided by the degradation rate, Δ_i. This gives a very appealing geometric interpretation (see Figure 3). Using equation (9), a natural estimator of the mean latency time $\mu_{S_i} = E(S_i)$ is

$$\hat{\mu}_{S_i} = \hat{\Delta}_i^{-1}\left[\hat{\mu}_i - \frac{1}{k_i}\sum_{j=0}^{k_i}\left([V_i(s_{ij} + S_i)] + \hat{\Delta}_i s_{ij}\right)\right] \tag{11}$$

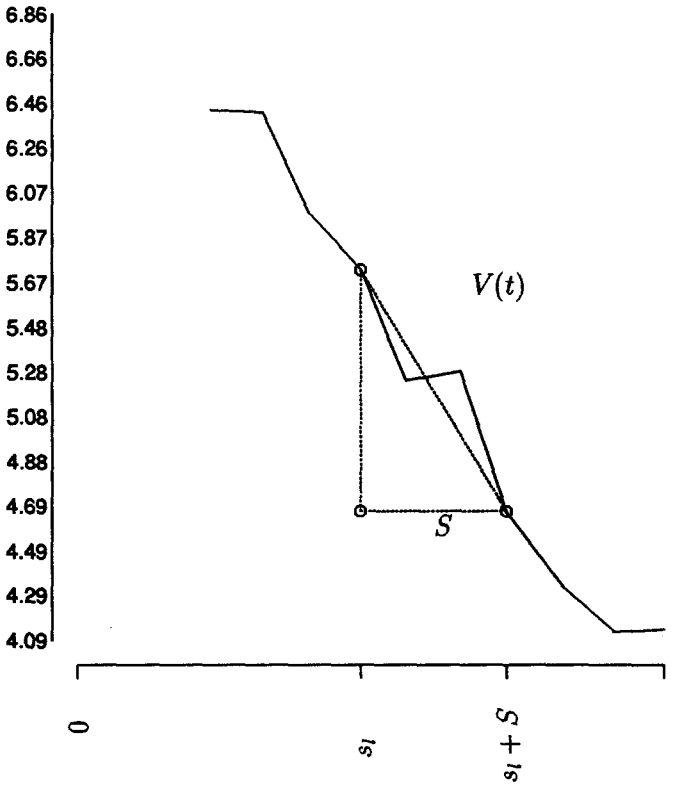

Figure 3. Estimating Mean Latency Time S. The y-axis represents $V(t) = \log X_0(t)$ for an infected subject and the x-axis represents time, t. S is predicted as the drop in $V(t)$ from time s_i to time $s_i + S$, divided by the rate of decline, Δ.

where $s_{ij} = \sum_{l=1}^{j}(t_{i,l+1} - t_{i,l})$, the time from the first to $(j+1)$st visit, and $s_{i0} = 0$. This is an estimator for the conditional mean latency time for the i^{th} individual given his/her covariate vectors. An estimator of the population mean latency time, $\mu_S = E(S)$, where S is the unconditional latency time for an individual selected at random from the population of infected HIV individuals, is

$$\hat{\mu}_S = \frac{1}{n_P}\sum_{i=1}^{n_P}\hat{\mu}_{S_i}.$$

3.4. *A Consistent Estimator of Mean Latency Time*

$\hat{\mu}_{S_i}$ is not a very efficient estimator of $E(S_i \mid \mathbf{Z}_{Pi}, \mathbf{Z}_{Ni})$ because it is only based on $k_i + 1$ observations. It is possible to develop a more efficient and consistent estimator of μ_{S_i} by

noting that under the model assumptions specified in equation (8), μ_{S_i} is a function of the covariate vector $\mathbf{Z}_{Pi}$ through $\Delta_i = \mathbf{Z}'_{Pi}\boldsymbol{\beta}$ and $\mu_i = \mathbf{Z}'_{Ni}\boldsymbol{\alpha}$ only. Thus we can write

$$\mu_{S_i} = g(\Delta_i, \mu_i), \qquad i = 1, 2, \ldots, n_P$$

for some function g. Note that the distribution of the latency time depends on the level μ_i that the individual would have had if the individual was not infected. This function g, and thus μ_{S_i}, can be estimated by computing a nonparametric estimator from $\{(\hat{\Delta}_i, \hat{\mu}_i, \hat{\mu}_{S_i}); i = 1, 2, \ldots, n_p\}$. For example, let K be the Epanechnikov kernel

$$K(u) = I(|u| \leq 1)(1 - u^2)0.75$$

where $I(\cdot)$ is the indicator function. Then, an estimator of μ_{S_i} is

$$\tilde{\mu}_{S_i} = \frac{\sum_{l'=1}^{n_N}\sum_{i'=1}^{n_P} K\left(\frac{\hat{\Delta}_{i'}-\hat{\Delta}_i}{h_P}\right) K\left(\frac{\hat{\Delta}_{l'}-\hat{\Delta}_i}{h_N}\right)\hat{\mu}_{S_{i'}}}{\sum_{l'=1}^{n_N}\sum_{i'=1}^{n_P} K\left(\frac{\hat{\Delta}_{i'}-\hat{\Delta}_i}{h_P}\right) K\left(\frac{\hat{\Delta}_{l'}-\hat{\Delta}_i}{h_N}\right)} \tag{12}$$

where h_P and h_N are the bandwidths which determine the index set over which the local weighted average of the $\hat{\mu}_{S_{i'}}$ estimates is taken. Here h_P and h_N tend to zero as n_P and n_N tend to infinity. Alternatively, more efficient locally linear nonparametric regression estimators could be employed (for example, see Cleveland and Devlin (1988), Fan (1993), and Ruppert and Wand (1994)). Note that from Figure 3 it is clear that the consistency of $\tilde{\mu}_{S_i}$ is not dependent on the normality assumption of $V(t)$; however, the consistency is dependent on the linearity assumption, $E(V(t) \mid \mathbf{Z}_{Pi}, \mathbf{Z}_{Ni}) = \mu_i - \Delta_i t$ in t.

3.5. *The Distribution of Latency Time Given the Initial CD4 Value*

Let S_i denote the latency time for an individual with covariate vectors $\{\mathbf{Z}_{Pi}, \mathbf{Z}_{Ni}\}$ and assume that conditional on $\{\mathbf{Z}_{Pi}, \mathbf{Z}_{Ni}\}$, model (8) holds. To obtain the distribution of S_i given the marker value $V_i(S_i)$ at the first visit time, it is convenient to rescale by setting

$$V'_i(t) = \sigma^{-1}[V_i(t) - \mu_i] \text{ and } S'_i = \sigma^{-1}\Delta_i S_i.$$

Then $V'_i(t)$ has a $N(-\sigma^{-1}\Delta_i t, 1)$ distribution and the distribution of $V'_i(S_i)$ given $S'_i = s'$ is $N(-s', 1)$. Let $q_i(s')$ denote the marginal (prior) distribution of S'_i, then, by Bayes Theorem, the (posterior) density of S'_i given $V'_i(S_i) = v'$ is

$$q_i(s'|v') = \frac{\phi(v' + s')q_i(s')}{\int_0^\infty \phi(v' + s')q_i(s')\,ds'}; \qquad s' \geq 0$$

where ϕ is the standard Normal density. In the special case where q_i is the Exponential density, $q_i(t) = \xi_i^{-1}e^{-\xi_i^{-1}t}$, $t \geq 0$ (see Berman, 1990),

$$q_i(s'|v') = \frac{\phi(v' + s' + \xi_i^{-1})}{\int_0^\infty \phi(v' + s' + \xi_i^{-1})\,ds'}, \qquad s' \geq 0.$$

This is a truncated Normal density. The mean ξ_i of the exponential distribution can be estimated as $\hat{\xi}_i = \hat{\sigma}^{-1}\hat{\Delta}_i\mu^*_{S_i}$ where $\mu^*_{S_i}$ can be either $\hat{\mu}_{S_i}$ of (11) or $\tilde{\mu}_{S_i}$ of (12), depending on the model chosen. Thus we have estimates of all the parameters appearing in the posterior distribution of S_i given $V_i(S_i)$.

Remark 4. These results could be extended from conditioning only on the initial CD4 process values to conditioning on all the CD4 process values. This is accomplished in the case of no covariates by Berman (1994, Theorem 4.1 with $\nu = 0$, $\tau^2 = 0$). Conditioning on the covariates only changes the mean of S_i; the shape of the posterior distribution of S_i given the CD4 process values $V(s_{ij} + S_i)$, $j = 1, 2, \ldots, k_i$ does not change and will be as in Berman (1994). In this case, an estimate of the correlation function, $r(t)$, will be needed. Estimation of the correlation function could be computed using nonparametric methods or by introducing a parametric form for the correlation function, $r(t)$, such as the first order autoregressive function, $r(t) = \exp(-\tau|t|)$. τ can be estimated as indicated in Section 3.1.

4. Discussion

In this paper we described two stochastic models which give joint distributions of biomarker processes, event times, and covariates of particular relevance in monitoring HIV infected individuals. We introduced a Weiner process model to predict residual time in Section 2 and a stationary Gaussian process model to model latency time in Section 3. There is a simple connection between the Weiner process model and the more general Gaussian process model which we have proposed. To establish this connection, first transform the time scale to the unit interval. This could be accomplished by dividing time the by the length of the study or by using a nonlinear transformation, such as $t \to 1 - \exp(-t) = u$. The distributional results of Section 2 are invariant to a time scale transformation (Doksum and Hóyland (1992)).

Note that the process $X(t)$ of Section 2 can be written using the notation in Section 3 as

$$X(t) = V(t) - V(0), \qquad 0 \le t \le 1.$$

Thus, $E(X(t)) = -\Delta t$, and we can identify Δ with η. The covariance function of $V(t) - V(0)$ is

$$\sigma^2\,[r(t-s) - r(t) - r(s) + 1], \qquad 0 \le s \le t \le 1.$$

This reduces to the covariance function, $\mathrm{Cov}(X(t), X(s)) = \delta^2 s$ of Section 2 by setting $\sigma^2 = \delta^2/2$ and $r(t) = 1 - t$, $0 \le t \le 1$. An interesting question that needs to be addressed empirically is when the simple correlation structure of the Weiner process model is appropriate. Finally note that in Section 3 notation, $t_{ij} = s_{ij} + S_i$. Thus, the model process described in Section 2 is a special case of the process presented in Section 3. Because the event time in Section 2 was "residual time" and was "latency time" in Section 3, it was convenient to use different notation in the two sections.

In addition to providing methods for the analysis of events time, our methods make it possible to study the effects of covariates, such as sociodemographic factors, risk exposure,

and treatment, by providing estimators and standard errors of regression coefficients. We presented our methods in the context of CD4 counts for HIV infected individuals in which interest was centered on predicting residual time or estimating latency time of the disease. However, we note that our methods are applicable in a broad range of problems in which interest centers on examining event times when biomarker information is available.

Berman (1990) conducted a data analysis of CD4 counts in HIV infected individuals but did not incorporate covariate information. In Part II, we analyze CD4 counts using data from the San Francisco Mens Health Study (Winkelstein et al, 1987) when covariate information is available. We employ model diagnostics to check the assumptions of normality and linearity, and we examine the correlation structure of the CD4 process. Of the assumptions listed above, the most crucial assumption is that of linearity. We demonstrate, that after calibrating for changes in immunological techniques for measuring CD4 counts, the linearity assumption is appropriate.

Acknowledgments

Professor Doksum's work was partially supported by Grant CA-56713, awarded by the National Cancer Institute, Department of Health and Human Services and by Grant DMS-93-07403, awarded by the National Science Foundation. Professor Normand's work was partially supported by Grant CA-61141, awarded by the National Cancer Institute.

References

S. M. Berman, "A Stochastic Model for the Distribution of HIV Latency Time Based on T4 Counts," *Biometrika* vol. 77 pp. 733–741, 1990.

S. M. Berman, "Perturbation of Normal Random Vectors by Nonnormal Translations, and an Application to HIV Latency Time Distributions," *Annals of Applied Probability*, vol. 4 pp. 968–980, 1994.

L. M. Calzavara, R. A. Coates, J. M. Raboud, et al, "Ongoing high-risk sexual behaviors in relation recreational drug use in sexual encounters: Analysis of 5 years of data from the Toronto Sexual Contact Study," *Ann. Epidemiol.* vol. 3 pp. 272–280, 1993.

CDC, Revision of the HIV classification system and the AIDS surveillance definition. Center for Disease Control and Prevention (Atlanta), 1993.

R. S. Chhikara and L. Folks, *The Inverse Gaussian Distribution. Theory, Methodology and Applications*, Marcel Dekker: New York, 1989.

P. D. Cleary, E. Singer, et al, "Sociodemographic and behavioural characteristics of HIV antibody-positive blood donors," *American Journal of Public Health* vol. 78 pp. 953–957, 1988.

W. S. Cleveland and S. J. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting," *Journal of the American Statistical Association* vol. 83 pp. 596–610, 1988.

K. A. Doksum and A. Hóyland, "Models for Variable-Stress Accelerated Life Testing Experiments Based on Wiener Processes and the Inverse Gaussian Distribution," *Technometrics* vol. 34 pp. 74–82, 1992.

J. Fan, "Locally linear regression smoothers and their minimax efficiencies," *Annals of Statistics* vol. 21 pp. 196–216, 1993.

N. P. Jewell and J. D. Kalbfleisch, "Marker models in survival analysis and applications to issues associated with AIDS," In *AIDS Epidemiology: Methodological Issues* (N. P. Jewell, K. Dietz, and V. T. Farewell, eds.), Birkhäuser: Boston, 1992, pp. 211–230.

J. P. Jewell, K. Dietz, and V. T. Farwell, *AIDS Epidemiology: Methodological Issues*, Birkhäuser: Boston, 1992.

R. A. Kaslow, D. G. Ostrow, R. Detels, et al, "The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants", *American Journal of Epidemiology* vol. 126 pp. 310–318, 1987.

M. Lefkopoulou and M. Zelen , "Intermediate clinical events, surrogate markers and survival," *Proceedings of the 1994 Conference on Lifetime Data Models in Reliability and Survival Analysis*, in press, 1995.

J. S. Montaner, J. Singer, M. T. Schechter, et al, "Clinical correlates of in vitro HIV-1 resistance of zidovudine. Results of the Multicenter Canadian AZT Trial," *AIDS* vol. 7 pp. 189–196, 1993.

S. L. Normand and K. A. Doksum, "Nonparametric calibration methods for longitudinal data," Technical Report #HCP-1994-3, Department of Health Care Policy, Harvard Medical School, Boston, MA, 1994.

D. Ruppert and M. P. Wand, "Multivariate locally weighted least squares regression," *Annals of Statistics* vol. 22, to appear, 1994.

W. Winkelstein Jr, D. M. Lyman, N. Padian, R. Grant, M. Samuel, R. E. Anderson, W. Lang, J. Riggs, and J. A. Levy, "Sexual practices and risk of infection by the human immunodeficiency virus," *Journal of the American Medical Association* vol. 257 pp. 321–325, 1987.