The International Journal of Biostatistics

Volume 2, Issue 1

2006

Article 1

A Regression Model for Dependent Gap Times

Robert L. Strawderman, *Cornell University*

Recommended Citation:

Strawderman, Robert L. (2006) "A Regression Model for Dependent Gap Times," *The International Journal of Biostatistics*: Vol. 2: Iss. 1, Article 1.

DOI: 10.2202/1557-4679.1005

A Regression Model for Dependent Gap Times

Robert L. Strawderman

Abstract

A natural choice of time scale for analyzing recurrent event data is the "gap" (or sojourn) time between successive events. In many situations it is reasonable to assume correlation exists between the successive events experienced by a given subject. This paper looks at the problem of extending the accelerated failure time (AFT) model to the case of dependent recurrent event data via intensity modeling. Specifically, the accelerated gap times model of Strawderman (2005), a semiparametric intensity model for independent gap time data, is extended to the case of multiplicative gamma frailty. As argued in Aalen & Husebye (1991), incorporating frailty captures the heterogeneity between subjects and the "hazard" portion of the intensity model captures gap time variation within a subject. Estimators are motivated using semiparametric efficiency theory and lead to useful generalizations of the rank statistics considered in Strawderman (2005). Several interesting distinctions arise in comparison to the Cox-Andersen-Gill frailty model (e.g., Nielsen et al, 1992; Klein, 1992). The proposed methodology is illustrated by simulation and data analysis.

KEYWORDS: accelerated failure time model, counting process, frailty model, intensity model, rank regression, recurrent events

Author Notes: I would like to thank the Editor and two referees for useful suggestions that improved the paper. The financial support of NIH grant GM056182 is gratefully acknowledged.

1 INTRODUCTION

Recurrent events are prevalent in longitudinal studies in biomedical and public health settings, particularly cancer, AIDS, cardiovascular and renal disease, and mental health. These data have mostly been analyzed using marginal models, assuming the potential for correlation exists between events within a subject (Prentice *et al.* 1981; Wei *et al.* 1989; Lin *et al.* 1998; Lin *et al.* 2000; Huang 2002; and, Chang 2004). Marginal models focus on fixed covariate effects, and the correlation between event times is not modeled. The resulting estimators are robust to the correlation structure between events. These inefficient estimates are also sensitive to censoring assumptions that, unfortunately, are often violated for recurrent event outcomes (Kalbfleisch and Prentice, 2002, Ch 9). Importantly, marginal models are only useful for studying population-level differences in event rates; they are not particularly useful for prediction or for studying longitudinal process dynamics precisely because the information required for doing so is not used.

Intensity models for the analysis of recurrent event data are being investigated with increasing frequency (Nielsen *et al.* 1992; Peña *et al.* 2001; Miloslavsky *et al.* 2004; Aalen *et al.* 2004; Strawderman, 2005). Intensity (or event history) models fully specify the correlation structure between events. The resulting estimators are generally more efficient than those obtained under a comparable marginal model and also more robust to censoring assumptions. Intensity models are especially useful in prediction. However, because the full correlation structure is specified, these models are also prone to misspecification. For example, the independence assumptions associated with a Poisson or renewal process, though potentially realistic in reliability applications, are often criticized as being too strong for biological applications.

Frailty models have been used in survival analysis as a way to capture dependence between subjects; see Hougaard (2000) for a review. Most of these efforts have focused on the Cox regression model (Cox, 1972a) and related variants; see, for example, Klein (1992) and Parner (1998). Aalen and Husebye (1991), Nielsen *et al.* (1992), Peña *et al.* (2001), and Duchateau *et al.* (2003) consider intensity models with frailty for modeling recurrent event data. Klein (1992) specifically considers the Cox regression model, and Nielsen *et al.* (1991) provide the necessary development for incorporating frailty into the Andersen-Gill model (Andersen and Gill, 1982). Hereafter, we refer to these combined developments as the Cox-Andersen-Gill (CAG) frailty model. Duchateau *et al.* (2003) consider both this model and the incorporation of frailty into the modulated renewal process model of Cox (1972b); see also Oakes (1992) and Oakes and Cui (1994).

Recently, the problem of incorporating frailty into the AFT model has also been considered. Compared to the Cox regression model, the AFT model has a “quite direct physical interpretation” (DR Cox, as quoted in Reid 1994, p. 450). Pan (2001) incorporates gamma frailty into this model as a way to accommodate clustered survival data. Lambert *et al.* (2004) consider parametric models in a similar context. These authors implicitly assume that cluster sizes are fixed and noninformative and that event times are independently censored. Neither of these assumptions holds in the case of recurrent event data. Specifically, identifying subjects as clusters and gap times within subjects as the events of interest, the cluster size (i.e., number of events) is evidently random and informative. Moreover, for a subject observed to experience k events, the effective censoring time for the $(k+1)^{st}$ gap time depends on the previous k gap times, violating the independent censorship assumption. Further comments on the distinction between modeling clustered time-to-event and recurrent event data can be found at the end of Section 2.

Strawderman (2005) develops the accelerated gap times (AGT) model, an accelerated failure time model applicable to recurrent event data. In the AGT model, covariates serve to accelerate or decelerate the gap times between events. Given a set of baseline covariates, the AGT model assumes that the observation on a given subject is the realization of a censored renewal process. The resulting renewal structure stems from assuming that the gap times of the corresponding uncensored recurrent event process are mutually independent. Analogously to Nielsen *et al.* (1992) and Duchateau *et al.* (2003), this paper incorporates gamma frailty into the AGT model in order to relax the gap time independence assumption. Hereafter, the resulting model is referred to as the “dependent accelerated gap times” (DAGT) model.

The remainder of this paper will proceed as follows. In Section 2, the DAGT model is developed for the case of baseline covariates. In Section 3, semi-parametric efficiency considerations are used to motivate an appropriate class of weighted logrank-type estimating functions for the regression parameter, the marginal baseline intensity function, and the variance of the gamma frailty distribution. Computational methods are discussed in detail. The weighted bootstrap, recently considered in Kosorok, Lee, and Fine (2004) for the case of the Cox regression model with frailty, is proposed as a way to obtain standard error estimates. Sections 4 and 5 respectively apply the proposed methods to simulated data and the bladder tumor data of Byar (1980). Section 6 concludes the paper with a discussion.

2 THE DAGT MODEL

Consider first a single subject with a p -dimensional bounded covariate vector Z experiencing repeated events at calendar times $0 < U_1 < U_2 < \dots$. Define j^{th} gap time as $T_j = U_j - U_{j-1}$, where $j \geq 1$ and $U_0 = 0$; that is, T_j is the time elapsed between the $(j-1)^{st}$ and j^{th} event. Let W denote a $\text{Gamma}(\eta_0^{-1}, \eta_0^{-1})$ random variable, where $\eta_0 \in [0, \infty)$. The parameterization of the gamma distribution being used here is the same as in Nielsen *et al.* (1992); hence, $E[W] = 1$ and $\text{Var}(W) = \eta_0$. We interpret $\eta_0 = 0$ as being equivalent to $W = 1$ with probability one. Conditionally on W , suppose $V_1, V_2, \dots$ are independent and identically absolutely continuous random variables with hazard function $W\lambda_0(\cdot)$, where $\lambda_0(u) > 0$ for $u > 0$ and $\int_0^t \lambda_0(u)du < \infty$ for every $t < \infty$. Conditionally upon Z and W , suppose $T_1, T_2, \dots$ are independent random variables with $T_j \stackrel{d}{=} V_j \exp(-\theta'_0 Z)$ for each $j \geq 1$. Given Z only, the model assumes exchangeability of $V_j, j \geq 1$; Z then directly accelerates or decelerates these 'baseline' gap times that correspond to $Z = 0$.

To accommodate censoring it is convenient to reformulate this model using intensities. Given Z and W , the hazard of T_j is easily shown to be $W\lambda_0(te^{\theta'_0 Z})e^{\theta'_0 Z}$. Define $U_m = \sum_{j=1}^m T_j$, and let $N^v(t) = \max(m : U_m \leq t)$ for $t \geq 0$. Let $N^v(t) = \max(m : U_m \leq t)$ for $t \geq 0$, where $U_m = \sum_{j=1}^m T_j$. Then, given Z and W , $N^v(t)$ is a renewal process with cumulative intensity

$$\Lambda(t|Z) = W \int_0^t \lambda_0\{e^{\theta'_0 Z} R^v(s)\} e^{\theta'_0 Z} ds, \quad (1)$$

where $R^v(s) = s - U_{N^v(s-)}$ (cf. Daley and Vere-Jones, 1988, p. 523). Since $\text{Var}(W) = \eta_0$, the degree of gap time dependence decreases as η_0 decreases. The AGT intensity model of Strawderman (2005) assumes independence and is recovered as $\eta_0 \rightarrow 0$. Peña *et al.* (2001) consider estimation (respectively, without and with frailty) in the model obtained by setting $\theta_0 = 0$.

REMARK: By (1), $E[\Lambda(t|Z)|Z] = \int_0^t E[W\lambda_0\{e^{\theta'_0 Z} R^v(s)\}|Z] e^{\theta'_0 Z} ds$. However, $E[W\lambda_0\{e^{\theta'_0 Z} R^v(s)\}|Z] \neq \lambda_0\{e^{\theta'_0 Z} R^v(s)\}$; consequently, the marginal and conditional interpretations of the regression coefficients do not coincide with regard to the impact of covariates on the average number of events. However, since $E[W\lambda_0(t)] = \lambda_0(t)$, the proposed model is essentially equivalent to a linear regression model for the logarithm of the gap times with an exchangeable error structure (i.e., without censoring), suggesting that these two interpretations do coincide for the effect of the regression coefficients on gap times.

Suppose there are n subjects, the i^{th} being observed over $[0, \tau_i]$, $\tau_i > 0$, and let the observed data be $\mathbf{O} = \{N_i^\dagger(u), Y_i^\dagger(u), Z_i, u \geq 0, i = 1 \dots n\}$, where $N_i^\dagger(u) = N_i^v(u \wedge \tau_i)$, $Y_i^\dagger(u) = I(\tau_i \geq u)$, and $a \wedge b = \min(a, b)$. We suppose $E[N_i^\dagger(\tau_i)] < \infty$ for $i \geq 1$. The major modeling assumption of this paper is then the following: given the observed covariates $Z_1 \dots Z_n$ and unobserved frailties $W_1 \dots W_n$, the processes $(N_i^\dagger(\cdot), i = 1 \dots n)$ form a multivariate counting process, the i^{th} component having intensity

$$\lambda_i(u) = W_i \lambda_0\{e^{\theta'_0 Z_i} R_i(u)\} e^{\theta'_0 Z_i} Y_i^\dagger(u) \quad (2)$$

for $R_i(u) = u - U_{iN_i^\dagger(u-)}$. As a result, for $i = 1 \dots n$ and $t \geq 0$,

$$M_i^\dagger(t) = N_i^\dagger(t) - W_i \int_0^t \lambda_0\{e^{\theta'_0 Z_i} R_i(u)\} e^{\theta'_0 Z_i} Y_i^\dagger(u) du \quad (3)$$

are orthogonal local square integrable mean zero martingales with respect to the history jointly generated by the “full” data $\mathbf{F} = \mathbf{O} \cup \{W_1 \dots W_n\}$. Implied here is the assumption that censoring is noninformative (i.e., on the calendar time scale) for the model parameters in (1), conditionally on the frailty and covariates. The corresponding conditional likelihood function constructed from $\mathbf{O} | (W_1 \dots W_n)' = w$ is then a product of two terms, one involving the intensity model (2) and the other consisting of terms involving the censoring process. It is assumed that this latter term does not involve w ; in Nielsen *et al.* (1992), this is referred to as “Conditional on $(W_1 \dots W_n)' = w$, censoring is noninformative of w .” Under these two assumptions, the partial likelihood under $\mathbf{F}$ for the model parameters represents a full, rather than partial, likelihood for the frailty parameter η_0 (Nielsen *et al.* 1992); similarly, the “partial marginal” likelihood obtained by integrating out w yields the full likelihood for η_0 assuming that only $\mathbf{O}$ is observed.

REMARK: The partial marginal likelihood for η_0 depends on the intensity function for $N_i^\dagger(\cdot)$ with respect to the observed data filtration. Similarly to Nielsen *et al.* (1992), this is easily shown to be

$$\frac{1 + \eta_0 N_i^\dagger(u-)}{1 + \eta_0 A_i^\dagger(u, \theta_0, \Lambda_0)} \lambda_0\{e^{\theta'_0 Z_i} R_i(u)\} e^{\theta'_0 Z_i} Y_i^\dagger(u),$$

where

$$A_i^\dagger(t, \theta_0, \Lambda_0) = \int_0^t \lambda_0\{e^{\theta_0' Z_i} R_i(u)\} e^{\theta_0' Z_i} Y_i^\dagger(u) du. \quad (4)$$

This remains a valid intensity model if η_0 exceeds each of $-1/N_i^\dagger(\tau_i)$ and $-1/A_i^\dagger(\tau_i, \theta_0, \Lambda_0)$ for $i = 1 \dots n$. In particular, the model permits negative η_0 , allowing for negative association despite the lack of a frailty interpretation. See Nielsen *et al.* (1992, pp. 27-31) or Andersen *et al.* (1993, pp. 667-669) for further discussion on the implications of allowing $\eta_0 < 0$.

REMARK: Pan (2001) proposes to model clustered survival data by introducing a frailty W into the AFT model. Structurally, Pan's model appears similar to that considered in this paper. Two important distinctions between Pan's model and the DAGT model were mentioned in the Introduction. A third important distinction is now possible by considering the intensity-based formulation of Pan's model. Considering a generic cluster, let $V_j(s) = I\{\nu_j \leq s, \delta_j = 1\}$, $s \geq 0$ denote the event counting process associated with member j of the cluster, where (ν_j, δ_j) , $j = 1 \dots n$ are respectively the observed survival times and corresponding censoring indicators of all cluster members. The cluster size n is regarded as fixed; more generally, one may view the analysis as being conditional upon the realization of a random cluster size assuming it is independent of all underlying failure, censoring, frailty and covariate information. In either case, Pan's model may be reformulated as follows: given a frailty W , $\{V_j(s), s \geq 0, j = 1 \dots n\}$ forms a multivariate counting process, the j^{th} component having cumulative intensity function $\mathcal{V}_j(s) = W \int_0^{s \wedge \nu_j} \lambda_0(u e^{\theta_0' Z}) e^{\theta_0' Z} du$. In comparison, (4) shows that the DAGT model involves a single frailty variable applied to a single intensity function. Moreover, under Pan's assumptions, each of the n processes $\{V_j(s) - \mathcal{V}_j(s), s \geq 0, j = 1 \dots n\}$ are local mean zero martingales with respect to the filtration jointly generated by the data (i.e., conditionally upon W and n). Hence, for example, $E[V_j(s) - \mathcal{V}_j(s) | W, Z, n] = 0$ for any $j = 1 \dots n$. With recurrent event data and letting T_j denote the j^{th} gap time for a given subject, one might formally proceed in Pan's setting by conditioning on $n = N^\dagger(\tau) + 1$ and defining $(\nu_j, \delta_j) = (T_j, 1)$ for $j \leq n - 1$ and $(\nu_j, \delta_j) = (R^v(\tau), 0)$ for $j = n$. In this case, the centered processes $\{V_j(s) - \mathcal{V}_j(s), s \geq 0, j = 1 \dots n\}$ no longer possess martingale structure with respect to an appropriate filtration; see, for example, Strawderman (2005). In fact, while $E[\sum_{j=1}^n \{V_j(s) - \mathcal{V}_j(s)\} | W, Z] = 0$ for any $s \geq 0$, it is not true in general that $E[V_j(s) - \mathcal{V}_j(s) | W, Z, n] = 0$ for a given $s \geq 0$ or $j = 1 \dots n$. For example, it is easy to prove that equality gen-

erally fails for $j = 1$ with $\theta_0 = 0$ and $\lambda_0(u) = \lambda_0$ (i.e., $N^\dagger(\cdot)$ is a homogeneous Poisson process, given W).

3 SEMIPARAMETRIC ESTIMATION

3.1 INTRODUCTION

If $\Lambda_0(\cdot) = \int_0^\cdot \lambda_0(u)du$ is parametrically specified, standard likelihood methods can be used for parameter estimation. However, with $\Lambda_0(\cdot)$ left unspecified, the model is semiparametric and parameter estimation becomes more challenging. Methods for semiparametric estimation of $\psi_0 = (\theta_0, \eta_0, \Lambda_0)$ are developed in the next several subsections.

Strawderman (2005) considers semiparametric estimation and asymptotics for θ_0 and $\Lambda_0(\cdot)$ in the absence of frailty, deriving a class of weighted logrank-type estimating functions for θ_0 and also an Aalen-type estimator for $\Lambda_0(\cdot)$. The class of estimating functions for θ_0 directly generalizes those studied in Tsiatis (1990) and Ritov (1990) to the case where each subject can experience more than one event. These results, incorporated as part of an EM-type estimation procedure, can be used to devise estimation procedures appropriate for the DAGT model.

To set ideas, a brief review of estimation for the CAG frailty model is helpful. The parameters of this model will also be denoted by $\psi_0 = (\theta_0, \Lambda_0, \eta_0)$, it being understood that the intensity model parameterization and resulting estimation procedures differ from those of the DAGT model. For now, it is assumed that frailty variance η_0 is known. Nielsen *et al.* (1992) and Klein (1992) point out that estimation of θ_0 and Λ_0 are straightforward using the Expectation-Maximization (EM) algorithm. The E-step consists of computing $E[W_i|\mathbf{O}_i]$, $i = 1 \dots n$, quantities that exist in closed form for the gamma frailty model. The M-step consists of computing the MLEs of θ_0 and Λ_0 based on the full data loglikelihood with W_i replaced by a suitable estimate of $E[W_i|\mathbf{O}_i]$, $i = 1 \dots n$. The resulting MLEs for θ_0 and Λ_0 are minor modifications of the standard partial likelihood and Breslow estimators. Iterating between these E and M steps eventually yields the observed data MLEs of θ_0 and Λ_0 (i.e., for η_0 known).

Given the structural similarity of the CAG frailty and DAGT models, a basic road map for estimation in the latter class of models now presents itself. Specifically, under the DAGT model, $W_i|\mathbf{O}_i$ is distributed as $\text{Gamma}(\xi_0 + N_i, \xi_0 + A_i^\dagger(\theta_0, \Lambda_0))$, where $N_i = N_i^\dagger(\tau_i -)$ is the number of events experienced by subject i , $A_i^\dagger(\theta_0, \Lambda_0)$ equals (4) evaluated at $t = \infty$, and $\xi_0 = \eta_0^{-1}$. The resulting E-step is therefore functionally identical to that under the CAG frailty

model, where $W_i|\mathbf{O}_i$ is distributed as $\text{Gamma}(\xi_0 + N_i, \xi_0 + A_i^{(cag)}(\theta_0, \Lambda_0))$ for $A_i^{(cag)}(\theta_0, \Lambda_0) = \int_0^\infty \lambda_0(u) e^{\theta'_0 Z_i} Y_i^\dagger(u) du$.

Let $\widehat{W}_i$ denote suitable estimates of $E[W_i|\mathbf{O}_i]$ for $i = 1 \dots n$. Define $\widehat{\mathbf{F}}$ to be the observed data $\mathbf{O}$ augmented with these estimates regarded as fixed realizations of $W_1 \dots W_n$. Adapting the results in Strawderman (2005, §3), the efficient scores for (θ_0, Λ_0) under $\widehat{\mathbf{F}}$ are easily computed and (in principle) yield estimates of (θ_0, Λ_0) . The process of iterating between the aforescribed E-step and updating the estimates of (θ_0, Λ_0) via these estimating equations clearly mimics the EM algorithm used for the CAG frailty model. However, there are also fundamental differences. For example, it will be seen in Section 3.2 that the efficient (full data) score for θ_0 depends on

$$Q(u) = 1 + u \frac{\lambda'_0(u)}{\lambda_0(u)},$$

a quantity not easily estimated. Similarly to Strawderman (2005), $Q(\cdot)$ is therefore replaced by a weight function that is easily computed. Since the efficient full data scores are not used, efficiency of the estimators obtained from this iterative procedure cannot be expected. In fact, the estimating equations do not correspond to the derivatives of a meaningful likelihood function. As a result, the ascent property typically associated with the EM algorithm is lost, for there is no longer an obvious (or unique) objective function being maximized. The use of general estimating equations in place of true likelihood scores in missing data problems is considered in Elashoff and Ryan (2004), the result being dubbed an “Expectation Substitution” (ES) algorithm. Section 3.2 develops the required E- and S-steps and Section 3.3 describes the algorithm used to estimate ψ_0 . Further computational details are provided in Section 3.4; a proposal for estimating standard errors is given in Section 3.5.

3.2 KEY ELEMENTS OF AN ES ALGORITHM

Throughout this section, it is implicitly assumed that η_0 is known and equal to η ; estimation of the frailty variance is considered later. We begin by computing the conditional expectations required for the E-step. Then, we define estimators for use in the S-step. For this we utilize the notation and results of Strawderman (2005). For any θ , define

$$Y_i(t|\theta) = \sum_{j=1}^{N_i} I(T_{ij} e^{\theta' Z_i} \geq t) + I[e^{\theta' Z_i} \{\tau_i - U_{iN_i}\} \geq t] \quad (5)$$

and

$$N_i(t|\theta) = \int_0^\infty I\{R_i(v)e^{\theta'Z_i} \leq t\} N_i^\dagger(dv) \equiv \sum_{j=1}^{N_i} I(T_{ij}e^{\theta'Z_i} \leq t), \quad (6)$$

where T_{ij} is the j^{th} gap time for the i^{th} subject and U_{iN_i} is the calendar time of the last event observed for subject i . Notice that $N_i(t|\theta_0)$ and $Y_i(t|\theta_0)$ respectively keep track of events and “at risk” status on the baseline gap time scale. For any θ and cumulative hazard function $L(\cdot)$, define

$$M_i(t|\theta, L) = N_i(t|\theta) - W_i \int_0^t Y_i(u|\theta) L(du). \quad (7)$$

For example, with $\theta = \theta_0$ and $L(u) = \Lambda_0(u)$, we obtain the centered gap time process $M_i(t|\theta_0, \Lambda_0) = N_i(t|\theta_0) - W_i \int_0^t Y_i(u|\theta_0) \Lambda_0(du)$. Using (6) and Lemma 2 of Strawderman (2005), we have

$$M_i(t|\theta_0, \Lambda_0) = \int_0^\infty I\{R_i(v)e^{\theta_0'Z_i} \leq t\} M_i^\dagger(dv),$$

where $M_i^\dagger(\cdot)$ is defined given in (3), demonstrating that $M_i(t|\theta_0, \Lambda_0)$ has mean zero. Finally, for any θ and $u \geq 0$, define

$$\mathcal{E}(u|\theta) = \frac{E\{W_1 Z_1 Y_1(u|\theta)\}}{E\{W_1 Y_1(u|\theta)\}}.$$

Expectations needed for E-step:

The primary function of the E-step is to replace all “missing” data (e.g., the W_i 's) with their best unbiased predictors based on $\mathbf{O}$. Suppose $\theta_0 = \theta$ and $\Lambda_0 = \Lambda$; then, our assumptions thus far imply $W_i|\mathbf{O}_i$ is distributed as $\text{Gamma}(\eta^{-1} + N_i, \eta^{-1} + A_i^\dagger(\theta, \Lambda))$, where N_i and $A_i^\dagger(\theta, \Lambda)$ are defined earlier. Using Lemma 2 of Strawderman (2005),

$$A_i^\dagger(\theta, \Lambda) \equiv \sum_{j=1}^{N_i+1} \Lambda(X_{ij}(\theta)), \quad (8)$$

where $X_{ij}(\theta) = T_{ij}e^{\theta'Z_i}$ for $j \leq N_i$ and $X_{ij}(\theta) = (\tau_i - U_{iN_i})e^{\theta'Z_i}$ for $j = N_i + 1$.

It follows that

$$E_{\theta, \eta, \Lambda}[W_i | \mathbf{O}_i] = \frac{1 + \eta N_i}{1 + \eta \sum_{j=1}^{N_i+1} \Lambda(X_{ij}(\theta))}.$$

With η fixed and substituting estimates $\hat{\theta}$ and $\hat{\Lambda}$ for θ and Λ , the E-step reduces to computing

$$\widehat{W}_i = \frac{1 + \eta N_i}{1 + \eta \sum_{j=1}^{N_i+1} \widehat{\Lambda}(X_{ij}(\hat{\theta}))}, \quad i = 1 \dots n. \quad (9)$$

As expected, $\widehat{W}_i \rightarrow 1$ as $\eta \rightarrow 0$.

Estimators needed for s-step

Let the “data” $\widehat{\mathbf{F}}$ consist of $\mathbf{O}$ and (9), regarding the latter as realizations of $W_1 \dots W_n$. Then, the efficient scores for θ_0 and Λ_0 under $\widehat{\mathbf{F}}$ are easily found by adapting the arguments used to prove Proposition 1 of Strawderman (2005). Specifically, it can be shown that these scores take the form

$$S_{\text{eff}}(\theta_0) = \sum_{i=1}^n \int_0^\infty Q(u) \{Z_i - \mathcal{E}(u|\theta_0)\} \widehat{M}_i(du|\theta_0, \Lambda_0) \quad (10)$$

and

$$B_{\theta_0, \Lambda_0}(\gamma) = \sum_{i=1}^n \int_0^\infty \gamma(u) \widehat{M}_i(du|\theta_0, \Lambda_0) \quad (11)$$

where $\gamma(\cdot)$ is some bounded deterministic function and $\widehat{M}_i(du|\theta_0, \Lambda_0)$ is defined as in (7) with $\widehat{W}_i$ in place of W_i . Replacing θ_0 and Λ_0 by generic parameters θ and Λ yields estimating equations for both parameters. For example, solving $B_{\theta, \Lambda}(\gamma) = 0$ for $\Lambda(\cdot)$ is equivalent to solving $\sum_{i=1}^n N_i(du|\theta) - \Lambda(du) \{ \sum_{i=1}^n \widehat{W}_i Y_i(u|\theta) \} = 0$. Assuming $\gamma(\cdot)$ is not identically zero, the solution is

$$\widehat{\Lambda}(t|\theta) = \int_0^t \frac{\sum_{i=1}^n N_i(du|\theta)}{\sum_{i=1}^n \widehat{W}_i Y_i(u|\theta)}.$$

A computationally equivalent form of this estimator is

$$\widehat{\Lambda}(t|\theta) = \sum_{i=1}^n \sum_{j=1}^{N_i} \frac{I\{X_{ij}(\theta) \leq t\}}{\sum_{r=1}^n \sum_{s=1}^{N_r+1} \widehat{W}_r I\{X_{rs}(\theta) \geq X_{ij}(\theta)\}}, \quad (12)$$

where $X_{ij}(\theta), j = 1 \dots N_i + 1, i = 1 \dots n$ and $\widehat{W}_1 \dots \widehat{W}_n$ are defined as in the E-step. Given the data $\widehat{\mathbf{F}}$ and an estimate $\widehat{\theta}$ of θ_0 , $\Lambda_0(t)$ can be estimated via $\widehat{\Lambda}(t) = \widehat{\Lambda}(t|\widehat{\theta})$.

An estimating equation for θ_0 based on $\widehat{\mathbf{F}}$ is obtained from (10) by substituting θ in for θ_0 and then estimating both $\mathcal{E}(\cdot|\theta)$ and $Q(\cdot)$. Estimating $\mathcal{E}(\cdot|\theta)$ is straightforward. First, write $E\{Z_i^{\otimes k} W_i Y_i(u|\theta)\} = E\{Z_i^{\otimes k} Y_i(u|\theta) E[W_i|\mathbf{O}_i]\}$ for $k = 0, 1, 2$ and $i = 1 \dots n$. Then, replacing $E[W_i|\mathbf{O}_i]$ with $\widehat{W}_i$ in (9), we estimate $\mathcal{E}(\cdot|\theta)$ using $\mathbb{E}(t|\theta) = \mathbb{S}^{(1)}(t|\theta)/\mathbb{S}^{(0)}(t|\theta)$, where

$$\mathbb{S}^{(k)}(t|\theta) = \frac{1}{n} \sum_{j=1}^n Z_j^{\otimes k} \widehat{W}_j Y_j(t|\theta). \quad (13)$$

Reliable nonparametric estimation of $Q(\cdot)$ is quite difficult. In the case of the AGT model, Strawderman (2005) proposes to replace $Q(\cdot)$ with a data-dependent weight, a strategy commonly used for semiparametric estimation under the AFT model (e.g., Ritov, 1990; Tsiatis, 1990). Without significant loss of generality, we write the required data-dependent weight as $H(u|\theta)$, the notation emphasizing the fact that choices involving η or Λ generally involve estimates of these quantities that depend only on θ .

Replacing θ_0 by θ in (10) and substituting $H(\cdot|\theta)$ for $Q(\cdot)$, we obtain

$$\sum_{i=1}^n \int_0^\infty H(u|\theta) \{Z_i - \mathbb{E}(u|\theta)\} \widehat{M}_i(du|\theta, \Lambda), \quad (14)$$

a class of weighted estimating functions indexed by $H(\cdot|\theta)$. Regardless of $\Lambda(\cdot)$, observe that

$$\sum_{i=1}^n \int_0^\infty H(u|\theta) \{Z_i - \mathbb{E}(u|\theta)\} \widehat{W}_i Y_i(u|\theta) \Lambda(du) = 0.$$

Consequently, using (14) is equivalent to using

$$S_H(\theta) = \frac{1}{n} \sum_{i=1}^n \int_0^\infty H(u|\theta) \{Z_i - \mathbb{E}(u|\theta)\} N_i(du|\theta),$$

which, similarly to (12), may be written

$$S_H(\theta) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{N_i} H(X_{ij}(\theta)|\theta) \{Z_i - \widehat{\mathbb{E}}(X_{ij}(\theta)|\theta)\}, \quad (15)$$

where

$$\widehat{\mathbb{E}}(u|\theta) = \frac{\sum_{k=1}^n \sum_{r=1}^{N_k+1} \widehat{W}_k Z_k I\{X_{kr}(\theta) \geq u\}}{\sum_{k=1}^n \sum_{r=1}^{N_k+1} \widehat{W}_k I\{X_{kr}(\theta) \geq u\}}.$$

The estimating function (15) is discontinuous in θ and a solution $\widehat{\theta}$ to $S_H(\theta) = 0$ does not exist in general. Hereafter, $\widehat{\theta}$ is defined as any zero-crossing of $S_H(\theta)$ or minimizer of $\|S_H(\theta)\|$, where $\|v\| = (v'v)^{1/2}$ for a vector v .

Two interesting choices of $H(u|\theta)$ include the “logrank” weight $H(u|\theta) \equiv 1$ and the “Gehan” weight $H(u|\theta) \equiv \mathbb{S}^{(0)}(u|\theta)$. Upon simplification, the Gehan weight is seen to yield

$$S_G(\theta) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^{N_i} \sum_{k=1}^n \sum_{r=1}^{N_k+1} (Z_i - Z_k) \widehat{W}_k I\{X_{kr}(\theta) \geq X_{ij}(\theta)\},$$

which is the gradient of the weighted convex objective function (cf. Fyngenson and Ritov, 1994; Strawderman 2005)

$$L_G(\theta) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^{N_i} \sum_{k=1}^n \sum_{r=1}^{N_k+1} \widehat{W}_k \{\log X_{kr}(\theta) - \log X_{ij}(\theta)\}^+$$

for $a^+ = \max\{a, 0\} \equiv (|a| + a)/2$.

The calculations of this section assume $\eta_0 = \eta$ is known. In addition, these calculations are clearly only meaningful when considered as part of a larger EM-type iteration scheme; for example, $\widehat{W}_i$, $i = 1 \dots n$ are considered fixed throughout but in fact depend on η and on the observed data through the estimates of $\widehat{\theta}$ and $\widehat{\Lambda}$. Joint estimation of ψ_0 is considered in Section 3.3.

3.3 ESTIMATING ψ_0

Nielsen *et al.* (1992) propose two possible EM-type algorithms for estimating ψ . The first is a standard EM algorithm, in which all parameters are jointly estimated based on iteratively maximizing the expected full data loglikelihood function. The second proposal involves using the EM algorithm to estimate θ_0 and Λ_0 , assuming $\eta_0 = \eta$ is fixed; this algorithm is then run for a grid of

η values, with $\hat{\eta}$ computed from the resulting approximation to the observed data profile likelihood function.

A third possibility here is to use the ECME algorithm of Liu and Rubin (1994). For simplicity of exposition, suppose the model parameters have been partitioned into 2 groups, say Ξ_1 and Ξ_2 . The basic ECME algorithm then replaces the M-step by two steps. The first step determines an estimate of Ξ_1 by maximizing the expected full data loglikelihood, fixing Ξ_2 at its current estimate. The second step estimates Ξ_2 by maximizing the observed data loglikelihood, fixing Ξ_1 at its current estimate. For example, in the case of the CAG frailty model, one might estimate θ_0 and Λ_0 using the expected full data loglikelihood with η fixed at a current estimate; then, η can be estimated from the corresponding observed data loglikelihood, with θ_0 and Λ_0 fixed at current estimates. Liu and Rubin (1994) show that the ECME algorithm converges under conditions similar to the EM algorithm while typically improving the rate of convergence.

An algorithm for estimating ψ_0 under the DAGT model can be developed along similar lines. Specifically, let $\text{ES}(\eta)$ denote the E- and S-steps described in Section 3.2. Fixing $\eta = \hat{\eta}_{[0]}$, we run $\text{ES}(\hat{\eta}_{[0]})$ to obtain estimates $\hat{\theta}_{[1]}$ and $\hat{\Lambda}_{[1]}$. Then, $\hat{\eta}_{[1]}$ is determined by maximizing the observed data loglikelihood for η in (16) below, computed assuming that $\theta = \hat{\theta}_{[1]}$ and $\Lambda = \hat{\Lambda}_{[1]}$. The observed data loglikelihood for η when $\theta_0 = \theta$ and $\Lambda_0 = \Lambda$ equals

$$\ell_P(\eta) = \sum_{i=1}^n \left(\sum_{j=1}^{N_i} \log [1 + \eta(j-1)] - (1 + \eta N_i) \left\{ \eta^{-1} \log \left[1 + \eta A_i^\dagger(\theta, \Lambda) \right] \right\} \right) \quad (16)$$

up to a constant independent of η ; this follows easily from (8) and the general representation for the counting process likelihood under a multiplicative gamma frailty model given in Nielsen *et al.* (1992, §3). Iterating this process to convergence yields an estimator $\hat{\psi}$ of ψ_0 . Specifically, the following algorithm is used for estimating ψ_0 :

Algorithm 1:

1. Initialize $\hat{\eta}_{[0]}$ and set $k = 1$.
2. run $\text{ES}(\hat{\eta}_{[k-1]})$ to obtain $\hat{\theta}_{[k]}$ and $\hat{\Lambda}_{[k]}(\cdot) = \hat{\Lambda}(\cdot | \hat{\theta}_{[k]})$.
3. Compute $\hat{\eta}_{[k]}$ by maximizing $\ell_P(\eta)$ with $\theta = \hat{\theta}_{[k]}$ and $\Lambda(\cdot) = \hat{\Lambda}_{[k]}(\cdot)$.
4. If convergence achieved, set $\hat{\psi} = (\hat{\theta}_{[k]}, \hat{\Lambda}_{[k]}(\cdot), \hat{\eta}_{[k]})$; else, return to Step 2.

REMARK: The loglikelihood (16) is in fact the proper marginal loglikelihood for η assuming $\theta_0 = \theta$ and $\Lambda_0(\cdot) = \Lambda(\cdot)$. However, for $\eta_0 = \eta$, the estimators for θ_0 and $\Lambda_0(\cdot)$ obtained by iterating between the E- and S-steps in Section 3.2 are not MLEs and their substitution into (16) does not yield a proper profile loglikelihood. In particular, efficiency of $\hat{\eta}$ cannot be expected.

For reasons explained in Section 3.4, Step 1 differs from that described in Nielsen *et al.* (1992), who initialize estimation under $\eta = 0$ (i.e., independence). The stopping criteria used for assessing convergence in Step 4 is $\|S_H(\hat{\theta}_{[k]})\|_\infty \leq \|S_H(\hat{\theta}_{[0]})\|_\infty$ and $\max\{|D_{[k],1}|, |D_{[k],2}|, |D_{[k],3}|\} < 10^{-4}$, where

$$D_{[k],1} = \frac{1}{n} \sum_{i=1}^n (\widehat{W}_{i,[k]} - \widehat{W}_{i,[k-1]}), \quad D_{[k],2} = \frac{\ell_P(\hat{\eta}_{[k]}) - \ell_P(\hat{\eta}_{[k-1]})}{10|\ell_P(\hat{\eta}_{[k-1]})|},$$

and $D_{[k],3} = \|S_H(\hat{\theta}_{[k]})\| - \|S_H(\hat{\theta}_{[k-1]})\|$. Since $n^{-1} \sum_{i=1}^n \widehat{W}_{i,[k]}$ should be close to 1.0 at convergence, the use of the mean difference across successive iterations is both an absolute and relative error criterion. The value of $\ell_P(\hat{\eta}_{[k]})$ used in the above criterion is that computed in Step 3. The use of a slightly less stringent relative error criterion for changes in $\ell_P(\eta)$ helps to prevent this particular criterion from requiring too many steps in cases where (16) is relatively flat. Monitoring the magnitude of $\|S_H(\hat{\theta}_{[k]})\|$ is not a good strategy for declaring convergence due to the fact that this norm is not guaranteed to be equal to zero for any θ . Instead, the change in norm is monitored and augmented with the condition $\|S_H(\hat{\theta}_{[k]})\|_\infty \leq \|S_H(\hat{\theta}_{[0]})\|_\infty$, ensuring that the final solution is no worse than the initial solution. Finally, if convergence is not achieved according to the above criteria by 100 iterations, iteration is stopped without declaring convergence. This stopping rule, though effective and conservative in the author's experience, has no theoretical justification and other stopping rules could certainly be employed.

3.4 CONVERGENCE, STARTING VALUES, AND OTHER DETAILS

Algorithm 1 involves the use of distinct estimating functions rather than the optimization of a single objective function. Hence, it is most appropriately classified as an example of a nonlinear block Gauss-Seidel algorithm; see Ortega and Rheinboldt (2000, §7.4). Convergence of Gauss-Seidel-type algorithms typically require smoothness conditions that are violated by (15) (cf. Ortega and Rheinboldt, 2000, Thm. 10.3.5). As a result, Algorithm 1 of Section 3.3 has no guarantee of local or global convergence. Importantly, this is a consequence

of the nature of the semiparametric estimation problem under study, not the specific implementation considered in this paper. The problem considered in Pan (2001) is similar in this regard and the lack of guaranteed convergence also extends to his algorithm. However, the conditions needed for convergence are typically satisfied in a limiting sense, that is, implied by smoothness conditions needed for establishing consistency and weak convergence of the estimators; see, for example, Strawderman (2005). Hence performance is expected to improve with increasing sample size.

The regularity conditions required for achieving local convergence suggest that it is important to use a good set of starting values. For θ_0 known, $N_i(\cdot|\theta_0)$ and $Y_i(\cdot|\theta_0)$ for $i = 1 \dots n$ correspond to the processes considered in Peña *et al.* (2001). Given a suitable estimate $\tilde{\theta}$, the EM algorithm developed there may thus be applied to the estimated counting and “at risk” processes $N_i(\cdot|\tilde{\theta})$ and $Y_i(\cdot|\tilde{\theta})$, $i = 1 \dots n$ in order to obtain useful starting values for η_0 and $\Lambda_0(\cdot)$. A sensible approach to determining a starting value $\tilde{\theta}$ is to make use of a marginal model. Lin *et al.* (1998) proposed an estimator of θ_0 under a marginal model that subsumes the DAGT model. This estimator is particularly attractive because it can be computed without estimating $\Lambda_0(\cdot)$ and η_0 . In the notation of this paper, their estimator minimizes $\|S_{LWY}(\theta)\|$ for

$$S_{LWY}(\theta) = \sum_{i=1}^n \sum_{j=1}^{N_i} \tilde{H}(U_{ij}e^{\theta'Z_i}|\theta) \left\{ Z_i - \frac{\sum_{k=1}^n Z_k I(U_{ij}e^{\theta'Z_i} \leq \tau_k e^{\theta'Z_k})}{\sum_{k=1}^n I(U_{ij}e^{\theta'Z_i} \leq \tau_k e^{\theta'Z_k})} \right\},$$

where $U_{ij} = \sum_{r=1}^j T_{ir}$ for $j \geq 1$ and $\tilde{H}(\cdot|\theta)$ is some weight function. The resulting estimator is consistent provided censoring is completely independent of the underlying event process. The “Gehan” weight $\tilde{H}(u|\theta) = \sum_{k=1}^n I(\tau_k e^{\theta'Z_k} \geq u)$ is used in computing a solution $\tilde{\theta}$ because the corresponding estimating function reduces to the gradient of a certain convex objective function that can be minimized very efficiently using linear programming methods. The algorithm used here is described in Nielsen (1998). This very fast, stable algorithm is an improved version of the “continuation” algorithms developed in a series of papers by Madsen and Nielsen (1990, 1992, 1993) for L_1 -estimation. We set $\hat{\psi}_{[0]}$ to be the solution obtained from this initial sequence of steps.

All programs used for the simulations and data analysis were written in MATLAB. The computational work required for obtaining starting values in Step 1 of Algorithm 1 is less onerous than it seems, typically requiring much less than half the time required to obtain the estimates in Steps 2 & 3. Computing $\hat{\theta}_{[k]}$ in Step 2 is the most difficult problem and we defer further

discussion until the end of this section. Given $\hat{\theta}_{[k]}$ and $\hat{\eta}_{[k-1]}$ from Step 1, the computation of $\hat{\Lambda}_{[k]}(\cdot)$ is immediate. The computation of $\hat{\eta}_{[k]}$ in Step 3 requires maximization of $\ell_P(\eta)$ in (16). The MATLAB routine `fminbnd`, which uses a combination of golden section search and parabolic interpolation for optimization on bounded intervals, is used to minimize $-\ell_P(\eta)$. The lower bound used in the maximization at the k^{th} iteration is

$$\max \left\{ -\frac{1}{N_i}, \quad i = 1 \dots n; -\frac{1}{\hat{A}_i^*(\tau_i, \hat{\theta}_{[k]}, \hat{\Lambda}_{[k]})}, \quad i = 1 \dots n \right\} + \epsilon,$$

where $\epsilon > 0$ is chosen to be at least two orders of magnitude smaller than the first term. Because this lower bound is negative, it is important that $-\ell_P(\eta)$ remain well defined at $\eta = 0$. Since $\log(1 + \eta x)/\eta = x - (\eta^2/2)x^2 + O((\eta x)^3)$ as $\eta \rightarrow 0$ for any $x \geq 0$, this is indeed the case. For the simulations and data analysis of this paper, the upper bound has been arbitrarily set equal to 50, covering all but the most extreme degrees of gap time dependence.

To compute $\hat{\theta}_{[k]}$ in Step 2, we must minimize $\|S_H(\theta)\|$ in (15); here, $\widehat{W}_i$ in (9) is computed using $\hat{\Lambda}_{[k-1]}(\cdot)$ and $\hat{\eta}_{[k-1]}$. Direct minimization of $\|S_H(\theta)\|$ using a generic optimization routine is hampered by both the lack of smoothness and the possible existence of multiple local minima that may or may not correspond to each component of $S_H(\theta)$ being “small” in magnitude. One way to circumvent the latter problem is to view the minimization of $\|S_H(\theta)\|$ as solving a nonlinear least squares problem. Specifically, notice that $F(\theta) := \|S_H(\theta)\|^2 = \sum_{j=1}^p S_{H,j}^2(\theta)$, where $S_{H,j}(\theta)$ is the j^{th} element of the vector $S_H(\theta)$. Since $S_H(\theta)$ is presumably close to zero for θ near $\hat{\theta}$, each component can be viewed as a residual term that is a nonlinear function of θ . Computational algorithms for nonlinear least squares problems are designed to minimize $F(\theta)$ such that each component of $S_H(\theta)$ is close to zero; see, for example, Kelley (1999). Importantly, it is not required that $S_H(\hat{\theta})$ equals zero, a fact that makes this algorithm potentially useful in this class of problems.

A popular and effective algorithm for nonlinear least squares problems is the Levenberg-Marquardt algorithm, a damped version of the Gauss-Newton algorithm having good global convergence properties (Kelley, 1999, §3.3.5). Most implementations of this algorithm assume $S_H(\theta)$ is smooth and require the user to supply either an exact or approximate Jacobian of the residual vector; see, for example, Kelley (1999, §3.2.3) or Madsen *et al.* (2004). Suppose momentarily that $S_H(\theta)$ were indeed differentiable with Jacobian $J_H(\theta)$ and that we wanted to use the Levenberg-Marquardt algorithm in Step 2 of Algorithm 1 to obtain $\hat{\theta}_{[k]}$ given $\hat{\eta}_{[k-1]}$. Then, setting $\theta_0^* = \hat{\theta}_{[k-1]}$,

the basic Levenberg-Marquardt algorithm would proceed by updating θ_j^* via $\theta_{j+1}^* = \theta_j^* + h$, where h solves

$$(J_H(\theta_j^*)^T J_H(\theta_j^*) + \mu I_p) h = -J_H(\theta_j^*)^T S_H(\theta_j^*).$$

Here, $\mu > 0$ is the so-called “damping parameter” and I_p denotes the $p \times p$ identity matrix. Notice that the matrix pre-multiplying h on the left hand side is always positive definite, ensuring that the next update continues to decrease the objective function $F(\theta)$. When μ is large, the update step is approximately $-\mu^{-1} J_H(\theta_j^*)^T S_H(\theta_j^*)$, a short step in the steepest descent direction; this tends to occur when θ_j^* is far from the solution. When μ is small, the step is approximately $-(J_H(\theta_j^*)^T J_H(\theta_j^*))^{-1} J_H(\theta_j^*)^T S_H(\theta_j^*)$; this tends to occur closer to the solution, where the algorithm behaves more like the standard Gauss-Newton algorithm. The damping parameter μ therefore impacts both the size and direction of the step h . Replacing $J_H(\theta_j^*)$ by an approximation does not fundamentally change the performance of the Levenberg-Marquardt algorithm, with one important exception. Specifically, if the Jacobian approximation used is quite poor, there is no guarantee that the step taken will correspond to a descent direction. This is a particular danger of so-called “secant” implementations of this algorithm (e.g., see Madsen *et al.*, 2004, §3.5), where inaccuracies may result from taking steps h that are either too large or too small.

Of course, $S_H(\theta)$ in (15) is not differentiable, so the Jacobian $J_H(\theta)$ does not exist. Strawderman (2005, §4.3) proposed a novel Monte Carlo procedure for estimating the (limiting) Jacobian of $S_H(\theta)$ in the absence of frailty, that is, with $\widehat{W}_i = 1$ for $i = 1 \dots n$. This procedure is trivially adapted to the problem of approximating the Jacobian of (15), regarding $\widehat{W}_i$, $i = 1 \dots n$ as fixed; we propose using the resulting approximation $\widehat{J}_H(\theta)$ in place of $J_H(\theta)$. Moreover, since computing $\widehat{J}_H(\theta)$ can be rather expensive, we utilize the Levenberg-Marquardt algorithm above with $J_H(\theta_j^*)$ replaced by $\widehat{J}_H(\theta_0^*)$. That is, in determining $\widehat{\theta}_{[k]}$ in Step 2 of Algorithm 1, the initial Jacobian approximation $\widehat{J}_H(\theta_0^*) = \widehat{J}_H(\widehat{\theta}_{[k-1]})$ is not updated during the course of this modified Levenberg-Marquardt algorithm. The rationale for this decision is that the starting value $\widehat{\theta}_{[0]}$ computed as part of Step 1 is already expected to be close to $\widehat{\theta}$, and hence few Jacobian updates should be required. Empirically, implementations that adjusted $\widehat{J}_H(\theta)$ more frequently demonstrated no improvement while substantially increasing computational effort. In fact, simply using $\widehat{J}_H(\theta_0^*) = \widehat{J}_H(\widehat{\theta}_{[0]})$ (i.e., using the same Jacobian approximation each time the Levenberg-Marquardt algorithm is called during the course of

Algorithm 1) proved to be very effective, resulting in further computational savings. The remaining algorithmic details (i.e., adjustment of the damping parameter μ and stopping criteria) follow Madsen *et al.* (2004, §3.2).

REMARK: Strawderman (2005) proposed using a one-step estimator for estimating θ_0 under the AGT model and outlines several of its advantages. Use of the Levenberg-Marquardt algorithm is a generalization of this idea, its first step being more or less equivalent to a one-step estimator when μ is initially chosen to be small. Simulations (not shown) indicate that use of the Levenberg-Marquardt algorithm (i.e., taking multiple steps) in Step 2 of Algorithm 1 reduces estimator bias in the case of the DAGT model.

REMARK: Other optimization methods were considered prior to settling on the adapted Levenberg-Marquardt algorithm described above, specifically two stochastic search methods (simulated annealing, genetic algorithm) and the Nelder-Mead simplex (NMS) algorithm. Convergence toward a solution typically occurs slowly for each of these search-based methods, particularly so with stochastic search. Each of these algorithms is therefore a comparatively impractical choice when combined with the use of the randomly weighted bootstrap for computing standard errors. See Givens and Hoeting (2005, Ch. 3) for a discussion of simulated annealing and genetic algorithms oriented towards statistical optimization problems. Notably, for the simulations of Table 1, the NMS algorithm was also observed to result in a substantially higher frequency of cases in which the program exited without satisfying convergence criteria (i.e., compared to the Levenberg-Marquardt algorithm). The reasons for this convergence failure were not investigated in detail. However, this suboptimal performance may be a consequence of the fact that the NMS algorithm has no guarantee of convergence even for convex, differentiable objective functions (e.g., McKinnon, 1998).

3.5 STANDARD ERRORS & THE WEIGHTED BOOTSTRAP

Kosorok *et al.* (2004) propose the randomly weighted bootstrap as a way for computing standard errors in the CAG frailty model (i.e., with a single event per subject). Let $\omega_1 \dots \omega_n$ denote iid exponential random variables with a unit mean. Define $\varpi_i = \omega_i / \bar{\omega}$, so that $\sum_{i=1}^n \varpi_i = n$; then, $\varpi_1 \dots \varpi_n$ have a Dirichlet distribution. Kosorok *et al.* (2004) first express the marginal log-likelihood function under the indicated model as an expectation with respect to $\mathbb{P}_n$ (i.e., the empirical measure assigning weight n^{-1} to each observation). The measure $\mathbb{P}_n$ is replaced with a weighted measure $\mathbb{P}_n^*$ that places weight ϖ_i/n on observation i , $i = 1 \dots n$. Bootstrap replications are then obtained

by maximizing the resulting reweighted marginal loglikelihood. Kosorok *et al.* (2004) comment that their experience is that Dirichlet weights “work well for semiparametric inference.”

Equivalently, one may proceed by first writing the corresponding score functions in terms of $\mathbb{P}_n$, replacing $\mathbb{P}_n$ by $\mathbb{P}_n^*$, and then solving the reweighted system of estimating equations to obtain bootstrap replicates; see, for example, Ma and Kosorok (2005). Taking the latter as our starting point, the incorporation of weights into the estimation procedure described in the previous sections becomes straightforward. For simplicity, we only consider $H(\cdot|\theta) = 1$ (i.e., the logrank weight).

As argued in the Appendix, Algorithm 1 essentially involves finding the solution to a fixed point system involving three equations. Writing each in terms of $\mathbb{P}_n$, weighted bootstrap versions of these estimating equations are obtained upon replacing $\mathbb{P}_n$ by $\mathbb{P}_n^*$. Each bootstrap replicate of $\hat{\psi}$, say $\hat{\psi}^*$, is obtained by generating a set of bootstrap weights and rerunning Algorithm 1 using $\hat{\psi}$ as the relevant starting value and with $\hat{\Lambda}(t|\theta)$, $S_H(\theta)$ and $\ell_P(\eta)$ respectively replaced by their bootstrap analogs. More specifically, with $\hat{\psi}_{[s-1]}^*$ computed, we determine $\hat{\psi}_{[s]}^*$ as follows:

1. Minimize $\|S_H^*(\theta)\|$, where

$$S_H^*(\theta) = n^{-1} \sum_{i=1}^n \varpi_i \sum_{j=1}^{N_i} \left\{ Z_i - \hat{\mathbb{E}}_{s-1}^*(X_{ij}(\theta)|\theta) \right\}, \quad (17)$$

$$\hat{\mathbb{E}}_{s-1}^*(u|\theta) = \frac{\sum_{k=1}^n \sum_{r=1}^{N_k+1} \varpi_k \hat{W}_{k,[s-1]}^* Z_k I\{X_{kr}(\theta) \geq u\}}{\sum_{k=1}^n \sum_{r=1}^{N_k+1} \varpi_k \hat{W}_{k,[s-1]}^* I\{X_{kr}(\theta) \geq u\}},$$

and

$$\hat{W}_{k,[s-1]}^* = \frac{1 + \hat{\eta}_{[s-1]}^* N_k}{1 + \hat{\eta}_{[s-1]}^* \sum_{m=1}^{N_k+1} \hat{\Lambda}_{[s-1]}^*(X_{km}(\hat{\theta}_{[s-1]}^*))}.$$

2. Compute

$$\hat{\Lambda}_{[s]}^*(t) = \sum_{i=1}^n \varpi_i \sum_{j=1}^{N_i} \frac{I\{X_{ij}(\hat{\theta}_{[s]}) \leq t\}}{\sum_{r=1}^n \varpi_r \sum_{s=1}^{N_r+1} \hat{W}_{r,[s-1]}^* I\{X_{rs}(\hat{\theta}_{[s]}) \geq X_{ij}(\hat{\theta}_{[s]})\}}. \quad (18)$$

3. Maximize

$$\ell_P^*(\eta) = \sum_{i=1}^n \varpi_i \left(\sum_{j=1}^{N_i} \log [1 + \eta(j-1)] - \frac{(1 + \eta N_i)}{\eta} \log [1 + \eta \hat{A}_{i,s}^*] \right) \quad (19)$$

$$\text{for } \hat{A}_{i,s}^* = \sum_{j=1}^{N_i+1} \hat{\Lambda}_{[s]}^*(X_{ij}(\hat{\theta}_{[s]}^*)).$$

Once the desired number of bootstrap replicates is obtained, standard errors and related quantities (e.g., percentile-based confidence intervals) are obtained in the usual fashion.

4 SIMULATION RESULTS

Two sets of simulations were done to evaluate the proposed estimators in the presence of gap time dependence for the case of the logrank weight. The first set of simulations considers a single covariate; the second considers the case of two covariates. The studies are designed to evaluate estimator bias and variability, including the quality of the bootstrap standard error estimators. Typically, simulations for evaluating the quality of bootstrap standard error estimates take a prohibitive amount of time. As pointed out in Kosorok *et al.* (2004) and Ma and Kosorok (2005), the centered weighted bootstrap estimates, or $\hat{\psi}^* - \hat{\psi}$, should have the same unconditional distribution as the centered unweighted estimates, or $\hat{\psi} - \psi_0$. Hence the validity of the bootstrap variance estimates were evaluated by computing a single bootstrap estimate per simulated dataset, say $\hat{\psi}^*$, and then computing the resulting Monte Carlo estimate of the variance of $\hat{\psi}^* - \hat{\psi}$.

In the first set of simulations, $\theta_0 = 0.5$ and $\{\Lambda_0(1), \Lambda_0(2)\} = \{1, 2\}$. Given W , the gap times $T \sim V e^{-0.5Z}$, where V is Exponential with rate W , $Z \sim \text{Uniform}(0, 2)$, and $\tau \sim \text{Uniform}(0, 3.5 + Z)$. We consider $\eta_0 = 1/10$ and $\eta_0 = 1/2$; respectively, $\text{var}(W) = 1/10$ and $\text{var}(W) = 1/2$. The Pearson and Spearman correlations between baseline gap times for $\eta_0 = 1/10$ are respectively 0.10 and 0.06. The Spearman correlation between baseline gap times for $\eta_0 = 1/2$ is 0.27; the Pearson correlation does not exist since $E[V_k V_j] = \infty$ for any pair k, j . In this setting, the average number of events per subject (i.e., $E[N_i]$) is approximately equal to 4.01, with respective variances 15.3 and 25.25. Each dataset contains $n = 100$ subjects and the simulation results, reported in Table 1, are based on 1000 simulations.

In the second set of simulations, $\theta_0 = (0.5, -0.1)'$ and $\{\Lambda_0(1), \Lambda_0(2)\} = \{1, 2\}$. Specifically, given W , the gap times $T \sim V e^{-0.5Z_1 + 0.1Z_2}$, where V is as above, $Z_1 \sim \text{Bernoulli}(0.5)$, $Z_2 \sim \text{Uniform}(0, 2)$, $Z_1 \perp Z_2$, and $\tau \sim \text{Uniform}(0, 3.5)$.

We again consider $\eta_0 = 1/10$ and $\eta_0 = 1/2$; the respective baseline gap time correlations are as given above. In this setting, the average number of events per subject (i.e., $E[N_i]$) is approximately equal to 1.75, with respective variances 3.2 and 5.0. We again set $n = 100$; the simulation results, reported in Table 2, are based on 1000 simulations.

Table 1: Simulation results for logrank weight [$\theta_0 = 0.5$, $n = 100$, 1000 data sets]

	$\eta_0 = 1/10$			$\eta_0 = 1/2$		
	Bias	Emp SE	Boot SE	Bias	Emp SE	Boot SE
$\hat{\theta}$	0.009	0.118	0.110	0.007	0.171	0.159
$\hat{\eta}$	0.014	0.054	0.054	0.027	0.136	0.130
$\hat{\Lambda}(1)$	0.000	0.156	0.146	0.019	0.212	0.225
$\hat{\Lambda}(2)$	0.002	0.326	0.313	0.039	0.452	0.444

Table 2: Simulation results for logrank weight [$\theta_0 = (0.5, -0.1)'$, $n = 100$, 1000 data sets]

	$\eta_0 = 1/10$			$\eta_0 = 1/2$		
	Bias	Emp SE	Boot SE	Bias	Emp SE	Boot SE
$\hat{\theta}_1$	0.002	0.157	0.147	0.003	0.213	0.202
$\hat{\theta}_2$	0.002	0.077	0.077	0.006	0.106	0.099
$\hat{\eta}$	0.029	0.087	0.072	0.047	0.174	0.161
$\hat{\Lambda}(1)$	0.005	0.125	0.129	0.003	0.159	0.162
$\hat{\Lambda}(2)$	0.010	0.305	0.305	0.021	0.349	0.378

In Tables 1 and 2, the column 'Bias' is the simulated bias in absolute value, 'Emp SE' is the empirical standard deviation of the simulated estimates, and 'Boot SE' is the empirical standard deviation of the centered weighted bootstrap estimates. For Tables 1 and 2, the algorithm used for obtaining starting values always converged. In Table 1, Algorithm 1 respectively converged for 999 and 1000 of the 1000 original simulated datasets (i.e., for $\eta_0 = 1/10$ and $\eta_0 = 1/2$). Using a single set of random bootstrap weights

per dataset, Algorithm 1 respectively converged for 997 and 996 datasets. In Table 2, Algorithm 1 converged for each of the 1000 original simulated datasets in both cases. Using a single set of random bootstrap weights per dataset, Algorithm 1 respectively converged for 999 and 1000 datasets. Notably, the cases of convergence failure for the original and bootstrap versions of the estimation problem did not overlap, meaning that a failure to converge for the original dataset does not necessarily imply failure in the weighted estimation problem. The results in Tables 1 and 2 are based only on the instances in which the algorithm converged for both the original and bootstrap datasets.

Figure 1: Histograms of $\hat{\eta}$ across the 4 simulated settings

Overall, Tables 1 and 2 demonstrate that the proposed estimation methods exhibit very small bias and that the proposed method for estimating stan-

dard errors performs well. The actual bias of $\hat{\eta}$ is negative in all cases, consistent with known behavior under the CAG frailty model. The table results further suggest that an increase in the bias of $\hat{\eta}$ from $\eta_0 = 1/10$ to $\eta_0 = 1/2$ has at best a modest impact on the bias of the other estimators. Histograms (not shown) of $\hat{\theta}$ all demonstrate normal behavior. Histograms (not shown) of both $\hat{\Lambda}(1)$ and $\hat{\Lambda}(2)$ demonstrate only mild skewness; applying the natural log transformation to each again yields normal behavior. Histograms for $\hat{\eta}$ in the 4 simulated settings may be found in Figure 1. Each demonstrates only mild positive skew, expected behavior due to the existence of a lower bound on the estimated frailty parameter. The frequency of values occurring near $\hat{\eta} = 0$ in the top panel of plots, a problem exacerbated by the negative bias in $\hat{\eta}$, illustrates the difficulty one is likely to face in detecting small departures from independence in moderately sized samples.

5 DATA ANALYSIS

In this section we illustrate the proposed methods using the bladder tumor cancer data of Byar (1980). The data come from a randomized clinical trial conducted by the Veterans Administration Co-operative Urological Group from 1971-1976 and consist of 118 patients with bladder tumors. After these tumors were removed, patients were randomized to one of three treatments: placebo, pyridoxine, or thiotepa. There are 38 patients with 45 total recurrences on the thiotepa arm and 48 patients with 87 total recurrences on the placebo arm. These data are analyzed in Lin *et al.* (1998) and Strawderman (2005) using the baseline variables treatment (1 = placebo, 0 = thiotepa), number of initial tumors removed (integer, range 1-8), and diameter of largest initial tumor (in centimeters, range 1-8) on the average number of recurrences. The AGT model of Strawderman (2005) imposes a renewal structure on each recurrent event process, asserting independence of all gap times. These assumptions permit estimation of both the covariate effects and underlying gap time distribution. The marginal analysis of Lin *et al.* (1998) imposes no such assumption, instead limiting the analysis to covariate effects and the so-called baseline mean function. Under the renewal assumptions of Strawderman (2005), the latter reduces to the renewal function associated with the baseline gap time distribution. Neither methodology permits one to assess whether dependence in the gap time sequence exists, a useful advantage of the DAGT model.

Table 3 summarizes the results for the covariate effects under the DAGT model obtained using the logrank weight function. Also reported are the corresponding estimates from Lin *et al.* (1998) and Strawderman (2005), both for

logrank (i.e., constant) weight functions. The standard errors for the DAGT model are based on 500 randomly weighted bootstrap samples.

Table 3: Bladder Tumor Data Results

Parameter	DAGT model		AGT model		Lin <i>et al.</i> (1998)	
	Estimate	Est SE	Estimate	Est SE	Estimate	Est SE
Treatment	0.572 [†]	0.245	0.681 [†]	0.278	0.542	0.312
Initial #	0.245 [†]	0.069	0.264 [†]	0.067	0.204 [†]	0.066
Initial Size	-0.021	0.089	-0.031	0.096	-0.038	0.084

[†] Significant at 5% level

Each of the models conveys similar information. Specifically, tumor recurrence times are expanded on treatment (0 = thiotepa, 1 = placebo) and contract as the number of initial tumors increase. The treatment effect and initial number of tumors are both statistically significant at the 5% level under the DAGT and AGT models. The estimated standard errors are also generally similar, though appreciably smaller for the treatment effect under the two intensity models compared to the model of Lin *et al.* (1998).

Lin *et al.* (1998) comment that their proposed chi-square goodness of fit test provided “no evidence against the assumed accelerated failure time model” – that is, against their proposed marginal model. The similarities of the regression coefficient estimates under the DAGT and Lin *et al.* (1998) models is therefore interesting, since correspondence would be expected only if the DAGT model is also approximately valid. It is also interesting to note the discrepancy in the estimated treatment effect coefficient under the DAGT and AGT models. The latter is inconsistent under the DAGT model when $\eta_0 \neq 0$, raising some doubt as to the validity of the gap time independence assumption required by the AGT model. A useful feature of the DAGT model is that one may investigate the statistical significance of the dependence parameter η . For these data, $\hat{\eta} = 0.56$, suggesting a modest level of gap time dependence. Using the randomly weighted bootstrap, the estimated standard error is 0.285; the corresponding asymptotic 95% Wald-type confidence interval is (-0.02,1.12). For comparison, an equitailed 95% bootstrap percentile interval is (0.00,1.12). However, the simulation results of the previous section also clearly demonstrate the existence of negative bias in $\hat{\eta}$. Using the randomly weighted bootstrap, the estimated bootstrap bias is -0.066, which is comparable in magnitude to the empirical bias of -0.049 observed in Table 2. A 95% bias-corrected bootstrap percentile interval (e.g., Shao and Tu, 1995, §4.1.3) is (0.15,1.35), providing

stronger evidence that dependence within gap time sequences at the subject level is indeed present.

To further assess the degree of dependence in these data the estimated posterior means (9) are sorted and plotted against their corresponding bootstrap mean values in Figure 2. The diagonal dash-dot line represents perfect agreement; the plot therefore suggests bias is minimal even at the two extremes. The vertical bars represent pointwise 95% bootstrap percentile intervals for each of the $\widehat{W}_i$'s. The horizontal dotted line at 1.0 corresponds to the hypothesis of independence; the fact that most of these confidence intervals lie completely above or below this line provides further evidence for the presence of gap time dependence. Finally, Table 3 above indicates that the greatest impact of incorporating frailty occurs for the treatment variable. The yellow and red circles (and corresponding magenta and blue error bars) respectively break down the aforescribed results by the thiotepa and placebo trial arms. We observe evidence of gap time dependence in both treatment groups, but no obvious pattern in these groups in terms of the sizes of their predicted frailties.

Figure 2: Posterior Means $\widehat{W}_i, i = 1 \dots n$ with 95% Bootstrap Intervals

The DAGT model allows estimation of the marginal recurrence time distribution and quantities derived from it through

$$P\{T > t|Z\} = \left(1 + \eta_0 \Lambda_0(e^{\theta' Z} t)\right)^{-1/\eta_0}.$$

Figure 3 depicts estimates of $P\{T > t|Z\}$ by treatment arm for the “average” patient (i.e., 2.5 initial tumors, the largest being 1.94 centimeters) using the logrank weight. Because the estimated distribution for the placebo group is obtained from that for the thiotepa group by a monotone time transformation, the observed ordering may be partly artifact. Also given are the estimates obtained in Strawderman (2005, Figure 1) under independence (i.e., for $\eta_0 = 0$). The incorporation of frailty yields a somewhat more optimistic view of the recurrence experience. However, this is not entirely unexpected. Specifically, for any fixed value of $\Lambda_0(e^{\theta'Z}t)$,

$$\left(1 + \eta_0 \Lambda_0(e^{\theta'Z}t)\right)^{-1/\eta_0}$$

is a decreasing function of the frailty parameter and the survivor function for $\eta_0 = 0$ is consequently less than the survivor function for any $\eta_0 > 0$. Of course, for $\hat{\eta} \neq 0$, different estimates of $\Lambda_0(\cdot)$ are obtained under the AGT and DAGT models and thus such an ordering is not necessarily guaranteed to occur in practice. The estimates of $\Lambda_0(\cdot)$ obtained under these two models for the bladder tumor data are reasonably similar (not shown), suggesting that the observed ordering in Figure 3 is at least partially attributable to the structural feature described above.

Quantiles of the survivor function are easily estimated under the DAGT model. For example, estimated median times to tumor recurrence for the thiotepa and placebo groups are respectively 22.5 months and 12.75 months with bootstrap standard errors of 4.78 and 1.47. Respectively, 95% bootstrap percentile intervals for median recurrence times are (12.25, 30.25) and (8.25, 14.00). The relationship between median recurrence times is also nicely reflected in the regression coefficient for treatment. Specifically, notice that $22.5/12.75 \doteq 1.77 \doteq \exp(0.572)$.

6 DISCUSSION

This paper has developed an extension of the accelerated failure time model for handling correlated recurrent event data. The regression coefficients under this model, which in the absence of censoring can be thought of as a linear regression model for the logarithm of the gap times with an exchangeable error structure, are relatively easy to interpret. An interesting feature of the model is that the conditional and marginal (i.e., averaged over the frailty distribution) interpretations of the regression coefficients coincide with regard to impact on gap times but not the average number of events. This is not the case for

Figure 3: Estimated Recurrence Time Distributions by Treatment

the CAG frailty model, where the marginal and conditional interpretations of the regression coefficients are known to differ. The proposed model is also an interesting extension of the accelerated gap times (AGT) model considered in Strawderman (2005). In addition to incorporating gap time dependence at the subject level, the DAGT model creates a suitable framework in which to evaluate the validity of the renewal assumption underlying the AGT model. The use of nonlinear least squares as a basis for estimation in rank regression appears to be new. The modified Levenberg-Marquardt algorithm performs well and is easy to modify for use in other rank regression problems. The large sample behavior of the proposed estimators, while technically challenging, is worthy of future research.

The technical issues encountered in developing suitable estimation procedures provide an interesting contrast to the CAG frailty model, where estimation is comparatively straightforward. These differences are not merely a consequence of the difficulty in estimating the unknown weight function $Q(\cdot)$. To see this, suppose $Q(\cdot)$ were indeed available. The resulting efficient full data scores, which do exist in closed form, do not yield the full data MLEs of (θ_0, Λ_0) . The efficient observed data scores, which do not exist in closed form, also do not yield the observed data MLEs of (θ_0, Λ_0) . In fact, it remains to be established whether a regular sequence of MLEs exists for the class of semi-parametric accelerated failure models; if not, then the sequence of estimates

obtained from a version of Algorithm 1 that utilizes the efficient full data scores also will not maximize the observed data likelihood function and, unlike the CAG frailty model, estimator efficiency cannot be deduced from connections to likelihood-based estimation. An interesting practical question is then: if the weight $H(\cdot|\theta)$ happens to be the locally efficient choice in the full data problem (i.e., where frailty is “observed”), does Algorithm 1 yield the locally efficient estimator of ψ_0 in the observed data problem? This question deserves further study because the projection calculations required for computing the efficient score in full data problems are often significantly easier than the corresponding calculations in the observed data problem.

The major computational difficulties encountered in this paper vanish if Λ_0 is assumed to lie in some parametric family of continuous distributions. It is reasonable to expect that this comparative simplicity carries over to the setting in which $\lambda_0(\cdot)$ is assumed to lie in a known space of smooth functions (e.g., a spline). Surprisingly little work exists on smoothing in the accelerated failure time model. Kim and Lai (2000) developed an asymptotically efficient adaptive M -estimator under random censoring by using a spline to approximate the derivative of the log error density. Gray (2000) considers penalized likelihood estimation for estimating the hazard function and hence the efficient weight function. The use of smoothing in the DAGT model is a potentially interesting and useful extension of this work. For example, in addition to facilitating certain computations, the recent work of Barker and Henderson (2005) suggests that smoothing may help to mitigate the bias in the estimated frailty variance $\hat{\eta}$ by making use of the actual gap time data (i.e., as opposed to just using information contained in their ranks).

APPENDIX: WEIGHTED BOOTSTRAP COMPUTATIONS

Let $\mathbf{O}_i$ denote the observed data available on a single subject. Define $\mathbb{P}_n g(\mathbf{O}) = n^{-1} \sum_{i=1}^n g(\mathbf{O}_i)$ for a function $g(\cdot)$ of $\mathbf{O}_i$; similarly, let $\mathbb{P}_n^* g(\mathbf{O}) = n^{-1} \sum_{i=1}^n \varpi_i g(\mathbf{O}_i)$. We desire to represent the equations used to determine $\hat{\psi}$ in terms of $\mathbb{P}_n$. In this regard, it is helpful to note that the “estimated frailty” $\widehat{W}_i$, defined in (9), is really just a function of $\mathbf{O}_i$ and $\psi = (\theta, \Lambda, \eta)$ only. We shall use

$$W_i(\psi) = \frac{1 + \eta N_i}{1 + \eta \sum_{j=1}^{N_i+1} \Lambda(X_{ij}(\theta))} \quad (20)$$

in place of $\widehat{W}_i$ (i.e., in this section only) in order to emphasize this dependence.

The determination of $\hat{\psi}$ can be viewed in terms of solving a fixed point system involving three equations. Let $\hat{\psi}_{[k-1]}$ denote the current iterate at the start of Step 2 in Algorithm 1; Steps 2 and 3 are used to determine $\hat{\psi}_{[k]}$. The equation (14) involves θ only; for $H(u|\theta) = 1$, (14) may be written

$$\mathbb{P}_n ZN - \int_0^\infty \frac{\mathbb{P}_n ZW(\hat{\psi}_{[k-1]})Y(u|\theta)}{\mathbb{P}_n W(\hat{\psi}_{[k-1]})Y(u|\theta)} [\mathbb{P}_n N(du|\theta)].$$

Minimizing the norm of this equation leads to $\hat{\theta}_{[k]}$. In view of (11), the second equation in our system of three equations is

$$\mathbb{P}_n \int_0^\infty \gamma(u) \left[N(du|\theta) - W(\hat{\psi}_{[k-1]})Y(u|\theta)\Lambda(du) = 0 \right],$$

where $\gamma(\cdot)$ is some arbitrary bounded function. For a fixed θ , we obtain $\hat{\Lambda}(t|\theta)$; with $\theta = \hat{\theta}_{[k]}$, we obtain $\hat{\Lambda}_{[k]}(t)$. Finally, observe that $\ell_P(\eta)$ in (16) may be written as $\mathbb{P}_n h(\mathbf{O}, \eta; \theta, \Lambda)$, where

$$h(\mathbf{O}_i, \eta; \theta, \Lambda) = \sum_{j=1}^{N_i} \log [1 + \eta(j-1)] - (1 + \eta N_i) \left\{ \eta^{-1} \log \left[1 + \eta A_i^\dagger(\theta, \Lambda) \right] \right\}$$

for $A_i^\dagger(\theta, \Lambda)$ defined as in (8). The last equation associated with our system is obtained by setting $\theta = \hat{\theta}_{[k]}$ and $\Lambda = \hat{\Lambda}_{[k]}$ and then maximizing the result to obtain $\hat{\eta}_{[k]}$. Applied iteratively, these computations ultimately lead to $\hat{\psi}$.

The randomly weighted bootstrap version of this system is obtained by replacing $\mathbb{P}_n$ with $\mathbb{P}_n^*$ everywhere. As explained in Section 3.5, $\hat{\psi}^*$ is then determined by iteratively solving the system created by (17)-(19). $\square$

REFERENCES

- Aalen, O., Fosen, J., Weedon-Fekjaer, H., Borgan, Ø., and Husebye, E. (2004). Dynamic analysis of multivariate failure time data. *Biometrics*, 60, 764-773.
- Aalen, O. and Husebye, E. (1991). Statistical analysis of repeated events forming renewal processes. *Statistics in Medicine*, 10, 1227-1240.
- Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics* **10**, 1100-1120.
- Andersen, P.K., Borgan, Ø., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.
- Barker, P. and Henderson, R. (2005). Small sample bias in the gamma frailty model for univariate survival. *Lifetime Data Analysis*, 11, 265-284.
- Byar, D. (1980). The Veterans Administration study of chemoprophylaxis for recurrent stage I bladder tumors: comparisons of placebo, pyridoxine, and topical thiotepa. In *Bladder Tumors and Other Topics in Urological Oncology*, Ed. M. Pavone-Macaluso, P.H. Smith and F. Fedsmyn, pp. 363-370. Plenum, New York.
- Chang, S. (2004). Estimating marginal effects in accelerated failure time models for serial sojourn times among repeated events. *Lifetime Data Analysis*, 10, 175-190.
- Cox, D.R. (1972a). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Ser. B* 34, 187-220.
- (1972b). The statistical analysis of dependencies in point processes. In *Stochastic point processes: statistical analysis, theory, and applications*, Ed. P.A.W. Lewis, pp. 55-66. New York: Wiley.
- Daley, D.J. and Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes*. Springer-Verlag, New York.
- Duchateau L., Janssen P., Kezic I., and Fortpied C. (2003). Evolution of recurrent asthma event rate over time in frailty models. *Journal of the Royal Statistical Society, Series C*, 52, 355-363.
- Elashoff M. and Ryan L. (2004). An EM algorithm for estimating equations. *Journal of Computational and Graphical Statistics*, 13, 48-65.
- Fygenon, M., and Ritov, Y. (1994). Monotone estimating equations for censored data. *Annals of Statistics* 22, 732-746.
- Givens, G.H. and Hoeting, J.A. (2005). *Computational Statistics*. Wiley, New York.

- Gray, R. (2000). Estimation of regression parameters and the hazard function in transformed linear survival models. *Biometrics*, 56, 571-576.
- Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer-Verlag, New York.
- Huang, Y. (2002). Censored regression with the multistate accelerated sojourn times model. *Journal of the Royal Statistical Society, Ser. B*, 64, 17-29.
- Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed.. Wiley, New York.
- Kim, C.-K. and Lai, T.L. (2000). Efficient score estimation and adaptive M-estimators in censored and truncated regression models. *Statistica Sinica*, 10, 731-749.
- Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 48, 795-806.
- Kelley, C.T. (1999). *Iterative Methods for Optimization*. SIAM, Philadelphia.
- Kosorok M.R., Lee B.L., and Fine, J.P. (2004). Robust inference for univariate proportional hazards frailty regression models. *Annals of Statistics*, 32, 1448-1491.
- Lambert P., Collett D., Kimber A., and Johnson, R. (2004). Parametric accelerated failure time models with random effects and an application to kidney transplant survival. *Statistics in Medicine*, 23, 3177-3192.
- Lin, D.Y., Wei, L.J., and Ying, Z. (1998). Accelerated failure time models for counting processes. *Biometrika*, 85, 605-618.
- Lin, D.Y., Wei, L.J., Yang, I., and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, Ser. B*, 62, 711-730.
- Liu, C.H. and Rubin, D.B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81, 633-648.
- Ma, S. and Kosorok M.R. (2005). Robust semiparametric M -estimation and the weighted bootstrap. To appear, *Journal of Multivariate Analysis*.
- Madsen, K. and Nielsen, H.B. (1990). Finite algorithms for robust linear regression. *BIT*, 30, 682-699.
- and ——— (1992). A continuation method for linear L_1 -estimation. In *L_1 -Statistical Analysis and Related Methods*, Ed. Y. Dodge, pp. 371-377. Institute of Mathematical Statistics, Hayward CA.

- and ——— (1993). A finite smoothing algorithm for linear L_1 estimation. *SIAM Journal on Optimization*, 3, 223-235.
- Madsen, K. and Nielsen, H.B. and Tingleff, O. (2004). *Methods for Non-Linear Least Squares Problems, 2nd edition*. Lecture notes, Department of Informatics and Mathematical Modelling, Technical University of Denmark.
- McKinnon, K.I.M. (1998). Convergence of the Nelder–Mead simplex method to a nonstationary point. *SIAM Journal on Optimization*, 9, 148-158.
- Miloslavsky M., Keles S., van der Laan M.J., and Butler, S. (2004). Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. *Journal of the Royal Statistical Society, Series B*, 66, 239-257.
- Nielsen, G.G., Gill R.D., Andersen, P.K., and Sorenson, T.I.A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, 19, 25-43.
- Nielsen, H.B. (1998). Computing a minimizer of a piecewise quadratic - implementation. Technical Report IMM-REP-1998-14, Department of Informatics and Mathematical Modelling, Technical University of Denmark.
- Oakes, D. (1992). Frailty models for multiple event times. In *Survival Analysis: State of the Art*, Ed. J. Klein and P. Goel, pp. 371-379. Kluwer, Dordrecht.
- Oakes, D., and Cui, L. (1994). On semiparametric inference for modulated renewal processes. *Biometrika*, 81, 83-90.
- Ortega, J.M., and Rheinboldt, W.C. (2000). *Iterative Solution of Nonlinear Equations in Several Variables*. SIAM, Philadelphia.
- Pan, W. (2001). Using frailties in the accelerated failure time model. *Lifetime Data Analysis*, 7, 55-64.
- Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *Annals of Statistics*, 26, 183-214.
- Peña, E.A., Strawderman, R.L., and Hollander, M. (2001). Nonparametric estimation with recurrent event data. *Journal of the American Statistical Association*, 96, 1299-1315.
- Prentice, R.L., Williams, B., and Petersen, A. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, 68, 373-379.
- Reid, N. (1994). A conversation with Sir David Cox. *Statistical Science*, 9, 439-455.

- Ritov, Y. (1990). Estimation in a linear regression model with censored data. *Annals of Statistics*, 18, 303-328.
- Shao J., and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer-Verlag, New York.
- Strawderman, R.L. (2005). The accelerated gap times model. *Biometrika*, 92, 647-666.
- Tsiatis, A.A. (1990). Estimating regression parameters using linear rank tests for censored data. *Annals of Statistics*, 18, 354-372.
- Wei, L., Lin, D. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84, 1065-1073.