Lifetime Data Analysis, 11, 265–284, 2005
© 2005 Springer Science+Business Media, Inc. Printed in The Netherlands.

Small Sample Bias in the Gamma Frailty Model for Univariate Survival

PETER BARKER*
Department of Mathematics and Statistics, Lancaster University

ROBIN HENDERSON
Department of Mathematics and Statistics, Lancaster University

Received October 17, 2003; Revised July 29, 2004; Accepted August 16, 2004

Abstract. The gamma frailty model is a natural extension of the Cox proportional hazards model in survival analysis. Because the frailties are unobserved, an E-M approach is often used for estimation. Such an approach is shown to lead to finite sample underestimation of the frailty variance, with the corresponding regression parameters also being underestimated as a result. For the univariate case, we investigate the source of the bias with simulation studies and a complete enumeration. The rank-based E-M approach, we note, only identifies frailty through the order in which failures occur; additional frailty which is evident in the survival times is ignored, and as a result the frailty variance is underestimated. An adaption of the standard E-M approach is suggested, whereby the non-parametric Breslow estimate is replaced by a local likelihood formulation for the baseline hazard which allows the survival times themselves to enter the model. Simulations demonstrate that this approach substantially reduces the bias, even at small sample sizes. The method developed is applied to survival data from the North West Regional Leukaemia Register.

Key words: bias, censoring, E-M algorithm, gamma frailty, local likelihood, life history data, proportional hazards model, smoothing

1. Introduction

Frailty models can be used in survival analysis to represent random effects or unexplained heterogeneity between individuals or groups. Suppose we have a sample of n individuals (t_i, δ_i, x_i), where t_i is the failure or censoring time, δ_i is the censoring indicator and x_i is the vector of covariates corresponding to individual i. In this paper, we consider univariate survival data where each individual i has an independent frailty z_i which acts multiplicatively on the hazard. This is of course equivalent to assuming the data arise from the non-proportional hazard, marginal distribution.

Frailty is easily introduced to the Cox proportional hazards model. Given the frailty z_i and the covariates x_i, the conditional hazard and survivor functions for individual i are

$$\alpha_i(t|z_i, x_i) = z_i \alpha_0(t) \exp(\beta' x_i), \tag{1}$$

$$S_i(t|z_i, x_i) = \exp(-z_i A_0(t) \exp(\beta' x_i)), \tag{2}$$

*Present address: TE1 J/3, Parklands, Alderley Park, Macclesfield, Cheshire, SK10 4TF, U.K. E-mail: Peter.N.Barker@astrazeneca.com

where β is the vector of regression parameters corresponding to covariates x_i, $\alpha_0(t)$ is the baseline hazard function, which can remain unspecified, and $A_0(t)$ is the cumulative baseline hazard function $\int_0^t \alpha_0(u)du$. Hougaard (2000) outlines a number of distributions suitable for the frailty. One of the most commonly used is gamma, written here as $Z \sim \Gamma(k, \lambda)$, to imply mean k/λ. The assumption that $E(Z) = 1$ is usually adopted to avoid an unidentifiable scale factor in (1) and (2). We take $Z \sim \Gamma(\xi^{-1}, \xi^{-1})$, so that $\mathrm{E}[Z] = 1$ and $\mathrm{Var}(Z) = \xi$. Unlike the shared gamma frailty model for multivariate survival data, where frailty and time dependence are identifiable in the absence of covariates, the model for univariate survival data requires a sufficiently variable covariate to guarantee identifiability. Both Elbers and Ridder (1982) and Heckman and Singer (1984) prove the identifiability of the gamma frailty model under moderate assumptions.

The shared gamma frailty model has been discussed by a number of authors. Because the frailties are unobserved, the E-M algorithm is the approach most commonly used. Proposed by Dempster et al. (1977), the E-M algorithm was applied to lifetime data with frailty by Nielsen et al. (1992), with a similar approach derived independently by Klein (1992). A penalized likelihood approach was also studied by Therneau and Grambsch (2000), and is the method implemented in R, although this has been shown to be equivalent to the E-M procedure for gamma frailty. These methods all use modified revisions of the non-parametric Breslow estimate, denoted $\tilde{A}_0(t)$ to estimate $A_0(t)$, where $\tilde{A}_0(t)$ is a discrete estimate of $A_0(t)$ with mass at the distinct failure times.

Nielsen et al. conducted simulation studies to investigate the performance of the E-M procedure applied to bivariate lifetime data. They found that the frailty variance ξ was underestimated in small or medium sized samples. Many authors, including Henderson and Oman (1999), observe that underestimating the frailty variance ξ results in the regression coefficients β also being underestimated. Because these are used to explain treatment or covariate effects, the underestimation of ξ is a serious problem.

Recently Rondeau et al. (2003) have used a penalized version of the marginal likelihood to fit the shared gamma frailty model with a continuous non-parametric baseline hazard. The method, illustrated on bivariate survival data, performs significantly better than the E-M procedure of Nielsen et al. (1992), practically removing the bias at all sample sizes considered. However, bias in the univariate case is not considered and the authors provide no insight into the reasons for the improvement in estimation.

In this paper, we investigate the source of the bias for univariate survival data. The format of the paper is as follows. In Section 2, we briefly outline the E-M procedure of Nielsen et al. and demonstrate the bias problem for simulated univariate survival data. The source of this bias is investigated with the aid of further simulation studies and a complete enumeration. We discover the bias is greatly affected by the estimate of the baseline hazard used in the E-M procedure. In Section 2, we consider the effect that the ordering of the failures and the failure times themselves have on the

estimation of the frailty variance ξ by considering a complete enumeration. In Section 3, we outline how continuous non-parametric estimates of the baseline and cumulative baseline hazards may be obtained using local likelihood techniques, and we apply these in the E-M procedure of Nielsen et al. We demonstrate the superiority of this method over the standard E-M approach of Nielsen et al. using simulated data. In Section 4, we apply the new method to data from the North West regional leukaemia register. We conclude the paper in Section 5 with a discussion of the methods introduced and suggestions for further work.

2. The E-M Procedure of Nielsen et al.

If the frailties were observed, the complete data log-likelihood would be

$$l(\beta) = \sum_{i=1}^{n} \delta_i \log(\alpha_0(t_i)e^{\beta' x_i}) - z_i e^{\beta' x_i} A_0(t_i) + \delta_i \log(z_i) - \frac{1}{\xi}\log(\xi) + \left(\frac{1}{\xi} - 1\right)\log(z_i) - \frac{z_i}{\xi} - \log\left(\Gamma\left(\frac{1}{\xi}\right)\right). \tag{3}$$

Nielsen et al. propose maximizing the observed data likelihood using an E-M algorithm on the above. They suggest first considering ξ to be fixed, in which case the last five terms above, which do not involve the unknown parameters β and A_0, can be ignored in the first place.

Initial estimates of β are obtained from the standard Cox proportional hazards model, and Breslow's estimate, $\tilde{A}_0(t)$, of the cumulative baseline hazard is obtained,

$$\tilde{A}_0(t) = \sum_{t_j \le t} \frac{d_j}{\sum_{k \in R(t_j)} \exp(\beta' x_k)}, \tag{4}$$

where d_j is the number of failures at time t_j, and $R(t_j)$ is the risk set of individuals known to be at risk immediately before time t_j.

The following steps are then iterated until convergence:

1. *E step:* For univariate gamma frailty, we have

$$\mathrm{E}[z_i | t_i, \delta_i, x_i] = \frac{1 + \xi\delta_i}{1 + \xi \tilde{A}_0^{(c)}(t_i) e^{\beta^{(c)\prime} x_i}}, \tag{5}$$

where $\tilde{A}_0^{(c)}$ and $\beta^{(c)}$ are the current estimates of A_0 and β, respectively.

2. *M step:* We maximize

$$l_c(\beta) = \sum_{i=1}^{n} \delta_i \log(\alpha_0(t_i)e^{\beta' x_i}) - z_i e^{\beta' x_i} A_0(t_i)$$

using standard Breslow/partial likelihood methods with offset $\mathrm{E}[z_i | t_i, \delta_i, x_i]$.

At this stage, the final estimates $\hat{\beta}_\xi$, $\hat{A}_{0\xi}$ say, are substituted into the marginal log-likelihood after frailties are integrated out

$$l_m(\xi|\hat{\beta}_\xi, \hat{A_{0\xi}}) = \sum_{i=1}^{n} \delta_i \log(\hat{\alpha}_{0\xi}(t_i) e^{\hat{\beta}'_\xi x_i}) - \left(\frac{1}{\xi} + \delta_i\right) \log(1 + \xi e^{\hat{\beta}'_\xi x_i} \hat{A_{0\xi}}(t_i)), \tag{6}$$

where the increment of the Breslow estimate $\hat{A_{0\xi}}$ is used for $\hat{\alpha}_{0\xi}$.

An outer loop then consists of a numerical search over ξ to maximize (6). A similar procedure was suggested by Klein, although ξ is estimated at the M step alongside A_0 and β, using all the terms in (3) and hence also requiring $E[\log Z|t, \delta, x]$ as well as $E[Z|t, \delta, x]$.

2.1. Simulation Study

Nielsen et al. performed a simulation study to demonstrate the E-M procedure for bivariate survival data. They found that $\hat{\xi}$ was negatively biased, and that this bias decreased with increasing sample size n. Furthermore, they found that the bias of $\hat{\xi}$ was worse for uncensored data than when individuals were right censored at time $\tau = 2$.

We perform a similar study for univariate survival data. Data were simulated from the gamma frailty model with an exponential baseline hazard, one standard normal covariate, with regression effect $\beta = 1$ and various amounts of frailty. We only provide details of the no censoring case because this situation yielded the most biased estimates of $\hat{\xi}$ in our simulation study. The gamma frailty model was fitted to the simulated data sets using the method described above. Following Nielsen et al. we allow small negative values of $\hat{\xi}$ so long as all expectations, $E[Z|t, \delta, x]$ remain positive. The results of 500 such repetitions are summarised in Table 1.

We obtain similar results to those of Nielsen et al. for bivariate survival data; the frailty variance is consistently underestimated, and as a result, the regression parameter β is also underestimated. However, the bias in our simulation study for univariate survival data appears worse than the results obtained by Nielsen et al. for bivariate data. If we compare the results for sample size 500, for example, we find that when the true value of ξ is 0.4, the mean estimate of $\hat{\xi}$ is 0.3321 in the univariate case, compared to 0.3811 in the Nielsen et al. paper. However, the sample size n in the bivariate simulation study refers to the number of pairs in the data, not the number of individuals. In the univariate case, for the same value of ξ, we find that at sample size 1000, the mean $\hat{\xi}$ was 0.3521, which is still noticeably less than the Nielsen et al. estimate at $n = 500$. The finite sample bias in $\hat{\xi}$ is worse for univariate than bivariate survival data.

Table 1 also demonstrates that the bias problem increases with frailty and decreasing sample size. Right truncation of the survival times does reduce the bias a little, but results (not shown) are still substantially biased, especially at small sample sizes.

Further simulation studies were performed to investigate the cause of the bias. When ξ was fixed at its true value, the E-M procedure produced unbiased estimates for β. We therefore consider the effect that fixing β or $\alpha_0(t)$ at their true values has on the estimation of the frailty variance ξ. Figure 1 shows the mean estimate of ξ based on 500 repetitions at sample size 200 when: (a) all parameters in the model are

Table 1. Overall simulation results based on 500 repetitions at various sample sizes.

n	True ξ	Mean($\hat{\xi}$)	sd($\hat{\xi}$)	mse($\hat{\xi}$)	Mean($\hat{\beta}$)	sd($\hat{\beta}$)
1000	0.0	−0.0239	0.0626	0.0045	0.9866	0.0584
1000	0.2	0.1715	0.0733	0.0062	0.9841	0.0639
1000	0.4	0.3521	0.0988	0.0121	0.9764	0.0702
1000	0.6	0.5414	0.1232	0.0186	0.9698	0.0763
1000	0.8	0.7208	0.1584	0.0314	0.9649	0.0843
1000	1.0	0.8999	0.1710	0.0393	0.9671	0.0831
500	0.0	−0.0418	0.0893	0.0097	0.9723	0.0884
500	0.2	0.1418	0.1167	0.0170	0.9659	0.1028
500	0.4	0.3321	0.1304	0.0216	0.9694	0.0957
500	0.6	0.4975	0.1663	0.0382	0.9509	0.1106
500	0.8	0.6732	0.2107	0.0605	0.9508	0.1191
500	1.0	0.7981	0.2255	0.0917	0.9292	0.1218
200	0.0	−0.0846	0.1170	0.0209	0.9600	0.1224
200	0.2	0.0572	0.1695	0.0492	0.9182	0.1442
200	0.4	0.2443	0.2153	0.0706	0.9236	0.1564
200	0.6	0.3971	0.2559	0.1067	0.9029	0.1702
200	0.8	0.5275	0.2709	0.1478	0.8874	0.1773
200	1.0	0.6103	0.3023	0.2435	0.8538	0.1830

estimated; (b) β is fixed at its true value of 1; (c) $\alpha_0(t)$ and $A_0(t)$ are fixed at their true value; (d) β, $\alpha_0(t)$ and $A_0(t)$ are all fixed at their true values.

We see that knowing the true values of any of the parameter values improves the estimation of ξ, but when the true baseline and cumulative baseline hazards are known, the estimates of ξ are almost unbiased. If these functions were estimated more accurately, we should expect improved estimation of ξ.

If the true parametric form of $\alpha_0(t)$ is known, we may estimate the model directly from the marginal likelihood, and a simulation study, not shown, demonstrates that apparently unbiased estimates of ξ and β may be obtained even at small sample sizes. Henceforth, we restrict ξ to be positive in order to keep the frailty interpretation, and as a result, we do not include simulations for $\xi = 0$ in the remainder of the paper.

2.2. A Complete Enumeration

In the simulation studies above, we have seen the importance that the estimation of $A_0(t)$ has on the effective estimation of the frailty variance ξ. When fixed at its true values in the E-M procedure, estimates of ξ, and therefore β are virtually unbiased.

In this section we consider a complete enumeration of three individuals to illustrate how frailty is identified from the data. Suppose our data consist of three individuals with covariate values -1, 0 and 1, respectively. In the following, we refer to these individuals by their covariate value as well as their order of failure. If we assume that all individuals fail in continuous time, so there are no ties in the survival times, there are six possible orderings of the failures: (−1 0 1), (−1 1 0), (0 −1 1), (0 1 −1), (1 −1 0) and (1 0 −1).

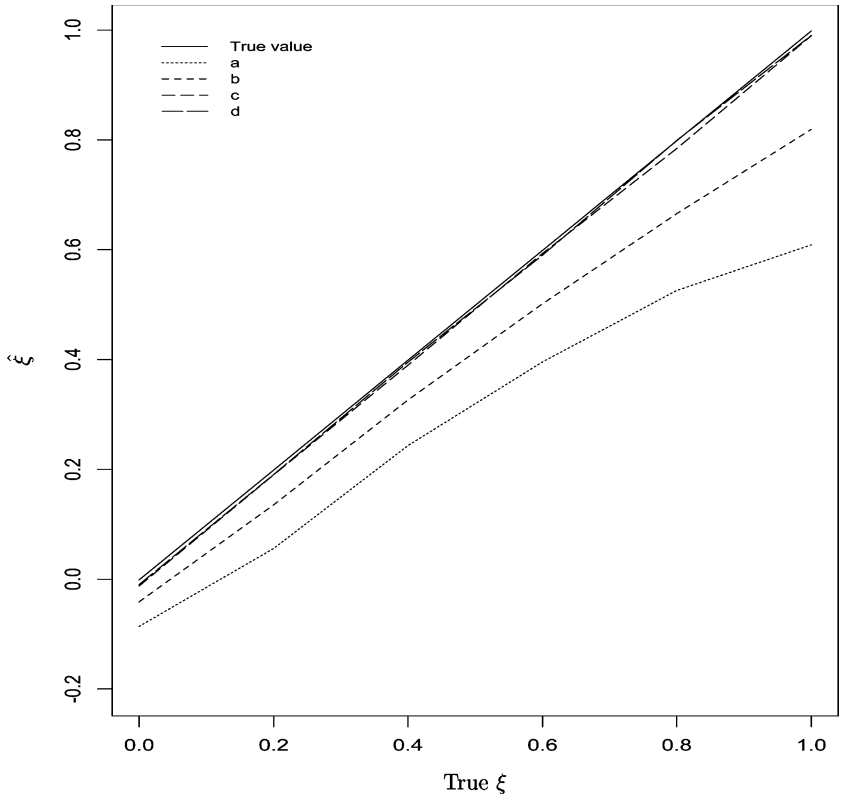

Figure 1. Plots of the mean $\hat{\xi}$'s for sample size 200 and no censoring, when the following parameters are fixed at their true values in the E-M procedure: (a) all parameters in the model are estimated; (b) β is fixed at its true value of 1; (c) $\alpha_0(t)$ and $A_0(t)$ are fixed at their true value; (d) β, $\alpha_0(t)$ and $A_0(t)$ are all fixed at their true values.

We begin by assuming a piecewise constant form for the cumulative baseline hazard with one parameter for each distinct failure time. Denote γ to be the vector $(\gamma_1, \gamma_2, \gamma_3)$. The cumulative baseline hazard is

$$A_0(t) = \sum_{t_i \leq t} \gamma_i.$$

We refer to this as the non-parametric estimate of $A_0(t)$. If at each failure time, γ_i coincides with the corresponding increment of the Breslow estimate, Equation (4), then the two estimates are equivalent.

We assume β, the regression effect, is fixed at $\beta = 3$. We may estimate γ directly from the marginal likelihood. Like the E-M approach outlined in Section 2, this formulation of $A_0(t)$ leads to a rank-based model; the orders of the failures affect the

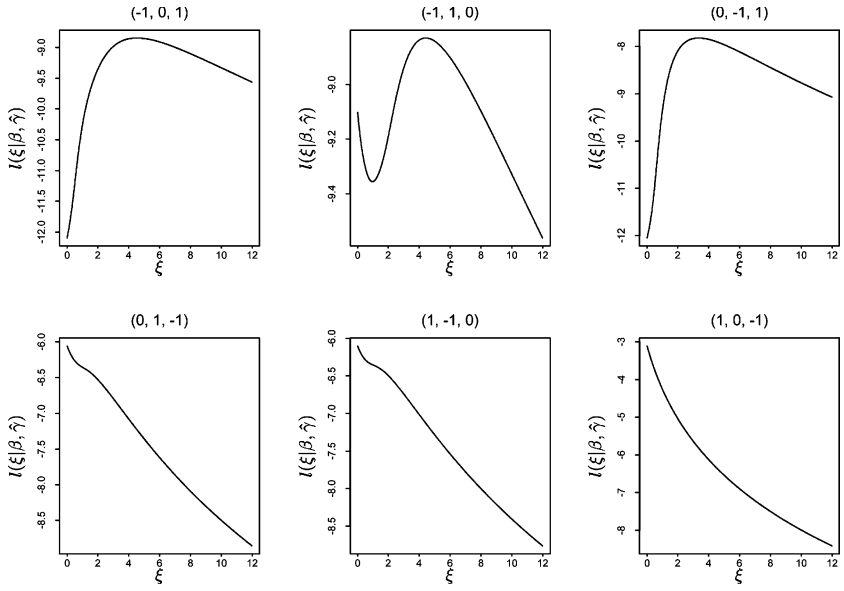

Figure 2. Profile likelihood over ξ for the 6 orderings using a rank-based estimate of $A_0(t)$.

likelihood, but the failure times themselves do not. We plot the profile likelihood over ξ in Figure 2 for the six orderings.

We are able to identify frailty in the top three plots, where individual -1 fails before individual 1, but we are unable to identify frailty in any of the other plots. In a similar plot for $\beta = 5$ (not shown), we are able to identify frailty for all orderings with the exception of (1, 0, -1). This is consistent with Elbers and Ridder (1982) who proved identifiability for the gamma frailty model. They suggest that in practice, it is important to have large variation in $\exp(\beta' x)$ to ensure that the model is identified. When the order of failure is (1, 0, -1), the failure order matches the covariate order from high to low risk, and we will never be able to identify frailty using this formulation for $A_0(t)$.

As we have emphasised, the formulation above leads to rank-based estimation; the survival times will not affect the profile likelihoods in Figure 2. Neither will the true value of ξ, although this will determine the probabilities of each of the six orderings occurring.

If the above estimate of $A_0(t)$ is replaced with the correct continuous parametric form, the model is no longer rank-based and the failure times do contribute to the likelihood. For the scenario above, we simulated one million sets of failure times using an exponential baseline hazard with rate $\lambda = 1$ at various amounts of frailty. Because the failure times themselves affect the shape of the likelihood, we can no longer represent the profile likelihood over ξ by six distinct curves. Instead, for each of the orderings, we may plot a mean profile likelihood over ξ. We do this in

Figure 3 for survival times simulated with $\xi = 0$ and $\xi = 2.5$. The percentage of times each ordering occurs in the simulations is given above each plot. Figure 3.1, the mean profile likelihood curves when the survival times were simulated with no frailty, looks very similar to Figure 2. We identify frailty in the top three plots, which account for just 0.2% of the simulated data sets, but we are unable to identify frailty in any of the other plots. However, if we simulate with $\xi = 2.5$, Figure 3.2 shows that we identify frailty in all of the orderings, even (1, 0, -1). By specifying an appropriate form for the baseline hazard, we are able to identify frailty from the failure times themselves as well as the order in which the failures occur.

To illustrate how the non-parametric and parametric profile likelihood curves compare, we produce an overall mean profile likelihood for the two approaches in Figure 4 for survival times simulated with $\xi = 2.5$. The non-parametric approach fails to identify frailty even though the true value of ξ was 2.5. If we use the parametric form for the baseline hazard, we are able to identify the presence of frailty, even when there are only three individuals in the data.

3. The Local Likelihood E-M Procedure

In the previous section, we demonstrated how frailty is identified by both the ordering of the failures and the survival times themselves. As the frailty variance increases, selection is increasingly influenced by the frailty terms z rather than the covariate values alone. We are more likely to see individuals failing in orders which do not correspond to the ordering of their relative risks. Similarly with increasing frailty, we expect a wider range of survival times. Individuals with a combination of high risk covariates and large frailties tend to fail extremely early, whereas those with low risk covariates and small frailties survive much longer.

The E-M procedure of Nielsen et al. (1992), described in Section 2, identifies frailty from the ordering, but not from the survival times themselves. We suggest that ξ is underestimated because the additional frailty that could have been identified from the actual failure times is ignored. If an appropriate parametric form can be found for $A_0(t)$, we obtain unbiased estimates for ξ but the model is no longer semi-parametric.

There are several methods for obtaining continuous non-parametric estimates of the baseline hazard which allow information on the failure times themselves to enter the model. One possible approach, outlined in Klein and Moeschberger (1997), is to kernel smooth the Breslow increments. The baseline hazard at time t is a weighted sum of the Breslow increments at times close to t; the failure times enter the model through the kernel function which determines the weighting. In simulation studies we found kernel smoothing methods to be inferior to a local likelihood procedure which will be outlined below.

Local likelihood is an extension of local regression and scatter-plot smoothing techniques, such as the loess procedure of Cleveland (1979). Such methods apply global regression techniques locally in a *smoothing window* around a point t defined by a bandwidth b. Local regression techniques were extended to likelihood based

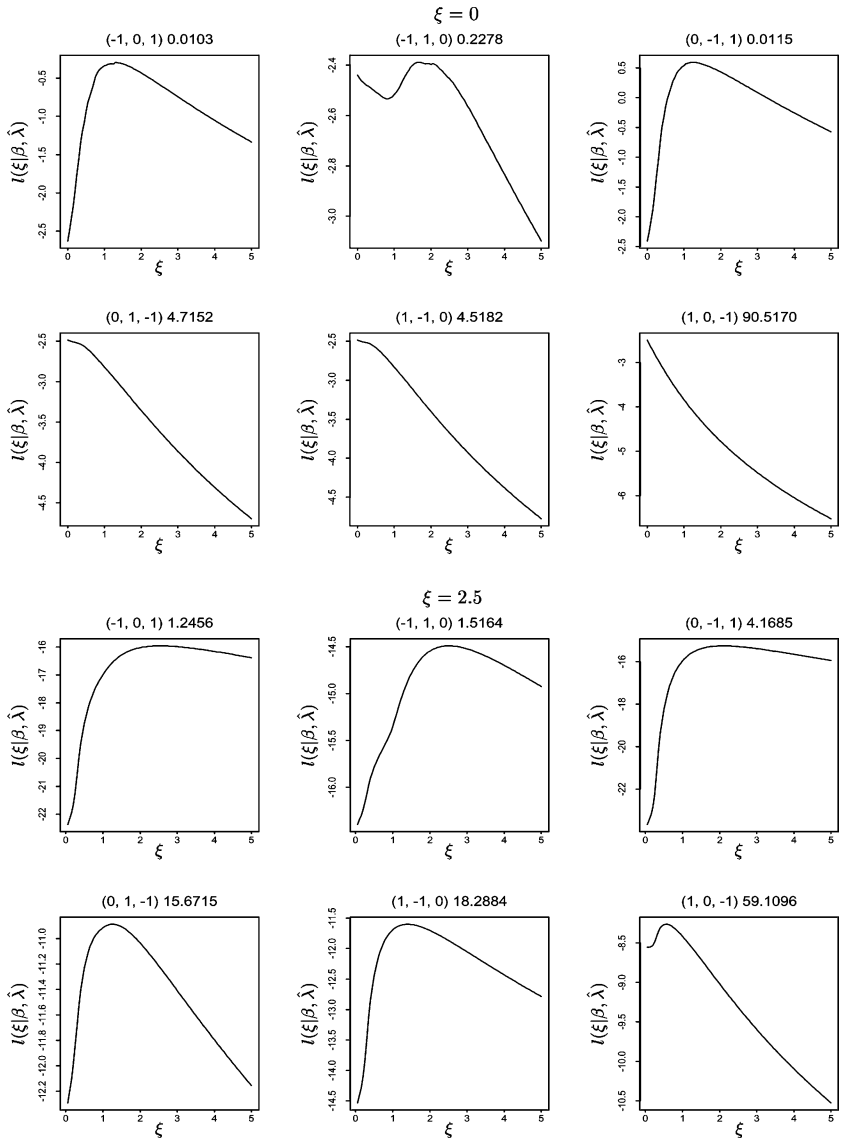

Figure 3. Profile likelihood over ξ for the 6 orderings with an exponential baseline hazard (the percentage of times each combination occurred in the simulation study is given above each plot).

models by Tibshirani and Hastie (1987), and local likelihood estimation in a proportional hazards model with right censored data is reviewed in Betensky et al. (2002). Frailty is easily introduced into this approach.

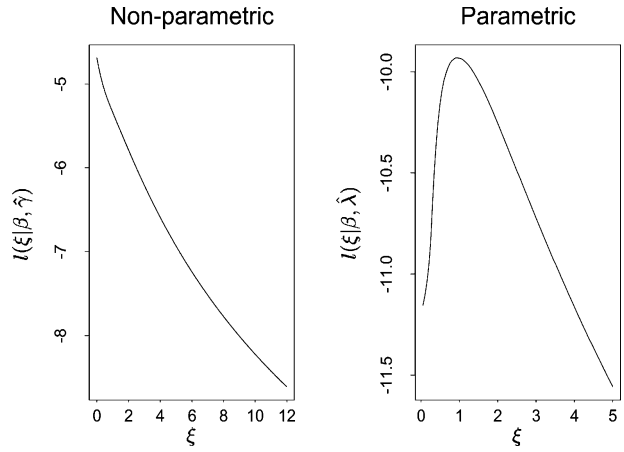

Figure 4. Overall mean profile likelihood over ξ for data simulated with $\xi = 2.5$ using the non-parametric and parametric baseline hazards.

We begin by considering the conditional log-likelihood of a proportional hazards frailty model, assuming frailties z_i are observed

$$l_f = \sum_{i=1}^{n} \left\{ \delta_i \log\left(\alpha_0(t_i)\right) + \delta_i \beta' x_i + \delta_i \log(z_i) - \int_0^{t_i} z_i e^{\beta' x_i} \alpha_0(u) du \right\}. \tag{7}$$

The local likelihood approach consists of approximating the log of the baseline hazard as a local polynomial in the smoothing window defined by a parameter b. For a particular failure time t, the polynomial approximation of degree p is given by

$$\log(\alpha_0(t)) \approx \lambda_{0t} + \lambda_{1t}(s-t) + \cdots + \lambda_{pt}(s-t)^p \quad \text{for} \quad |s-t| \leq b, \tag{8}$$

where s denotes a failure time within the bandwidth of t. For convenience, we denote the vector $(\lambda_{0t}, \lambda_{1t}, \ldots, \lambda_{pt})'$ as λ_t.

We can use the polynomial approximation to form a *local log-likelihood* component at t

$$l_t = \sum_{i=1}^{n} \left\{ \delta_i \beta' x_i + \delta_i \log(z_i) + \delta_i K\Big(\frac{t_i - t}{b}\Big)\{\lambda_{0t} + \cdots + \lambda_{pt}(t_i - t)^p\} \right.$$
$$\left. - \int_0^{t_i} z_i e^{\beta' x_i} e^{\lambda_{0t} + \cdots + \lambda_{pt}(u-t)^p} K\Big(\frac{u-t}{b}\Big) du \right\}, \tag{9}$$

where $K(.)$ is a kernel function defined on $|t - t_i| \leq b$. The kernel $K(.)$ gives progressively greater weight to points closer to t within the smoothing window, and gives no weight to points outside this window. The local likelihood component at time t is therefore just a weighted form of the complete data log-likelihood; failures occurring close to t contribute more to the local likelihood component than those further away.

The local likelihood component, l_t, may be treated in the same way as the complete data log-likelihood; the maximum local likelihood estimate $\hat{\lambda}_t$ is the vector which maximises (9), or alternatively $\hat{\lambda}_t$ is the solution to the $p+1$ score equations

$$\sum_{i=1}^{n}\left\{\delta_i K\Big(\frac{t_i-t}{b}\Big)(t_i-t)^j-\int_0^{t_i} z_i e^{\beta' x_i}(u-t)^j e^{\lambda_{0t}+\cdots+\lambda_{0p}(u-t)^p}K\Big(\frac{u-t}{b}\Big)du\right\}=0$$

for $j=1,\ldots,p$. The estimate of the baseline hazard at time t is obtained as $\hat{\alpha}_0(t)=\exp(\hat{\lambda}_{0t})$. In order to fit the gamma frailty model, we require estimates of $\alpha_0(t)$ for every observation, so therefore a local likelihood component is maximized at each failure or censoring time in the data set.

Betensky et al. (2002) point out that the estimation of the baseline hazard may also be based upon approximating $\alpha_0(t)$ as a local polynomial, rather than $\log(\alpha_0(t))$; the second approach is sensible, they argue, because it guarantees that $\hat{\alpha}_0(t)$ remains positive, which is essential in a survival setting.

We return to the local likelihood component (9) and consider the individual contributions to the local likelihood component at time t. If we disregard the $\delta_i\beta' x_i$ and $\delta_i \log(z_i)$ terms, which do not involve λ_t, we see that the local log-likelihood component comprises of two parts, the second involving an integral between 0 and t_i.

We recall that the kernel $K(.)$ is defined on $|t-t_i|\leq b$, and is zero elsewhere; therefore, only individuals who fail within the bandwidth of t contribute to the first part of the local log-likelihood component for λ_t. The individual contributions to the second part of l_t are determined by the value of the integrated kernel function. Only individuals who fail or are censored within the smoothing window, or beyond it contribute to the second component. This result should be unsurprising; because $\alpha_0(t)$ measures the baseline instantaneous failure rate at t, we should expect individuals known to be alive at t to contribute to this. If we consider the Breslow estimate of the cumulative hazard, we note that individuals who have not yet failed or been censored are members of the risk set R_t and contribute to the denominator of the Breslow increment at time t as well.

3.1. Issues Relating to Bandwidth Choice

A key issue surrounding any smoothing technique is how to choose the bandwidth. A variable bandwidth which is wider in areas where the data are sparse is preferable to a fixed bandwidth. Unlike a kernel smoothing approach where the estimate of $\alpha_0(t)$ is essentially just a weighted sum of the Breslow increments, the local likelihood estimate is obtained by maximising the local likelihood components at the failure times. There must therefore be a sufficient amount of data contributing to the local likelihood components to enable us to obtain maximum likelihood estimates. Because of the discussion above, for small t there will be more data contributing to the likelihood component than for larger t (because both t_i within the bandwidth and t_i greater than $t+b$ contribute to the likelihood). We need a variable bandwidth which is wide enough to provide enough data for us to maximize the likelihood

components even for large failure times. As a result, we use a bandwidth proportional to the distance to the kth nearest neighbour, k being chosen to cover a proportion of the data. Betensky et al. (2002) use a nearest neighbour bandwidth which encompasses 40% of the data in the smoothing window. Throughout this paper, we use the distance to the $n/4$th nearest neighbour as the bandwidth and a local linear approximation for $\log(\alpha_0(t))$ within the smoothing window. Simulation results, discussed below, suggest this is a suitable choice.

It may be possible to derive a method based upon cross validation to determine the value of k best suited to the data. However, such an approach would be extremely computer intensive and would therefore be of limited use in practice. We suggest some form of sensitivity analysis be done to ensure that the estimates of β and ξ are not overly influenced by the value of k used to fit the model.

3.2. Implementation and Computational Issues

We may simply replace the Breslow estimate $\tilde{A}_0(t)$ and its increment $\Delta\tilde{A}_0(t)$ in the E-M procedure of Nielsen et al. (1992) with local likelihood estimates. We obtain the estimates of $A_0(t)$ from the local likelihood estimates of the baseline hazard using the trapezium rule to approximate the integrals.

An important consideration, when using the local likelihood method of estimating $\alpha_0(t)$, is the degree of local polynomial to use. Betensky et al. (1999) state that degree 0, 1 or 2 polynomials are usually used, although evidence suggests that the use of degree 1 or 2 polynomials are preferable to degree 0 ones as they have less boundary bias. We have chosen to use a local linear approximation in the local E-M implementation; when an Epanechnikov kernel along with a local linear approximation is used in the local likelihood component, (9), we may evaluate the integral explicitly which greatly speeds the computation.

We use a Newton Raphson procedure to estimate the $\hat{\lambda}_t$ terms. The choice of suitable starting values is essential if the procedure is to converge. We would expect the estimates $\hat{\lambda}_{t_i}$ to be similar to $\hat{\lambda}_{t_{i-1}}$, so it is sensible to use the value of the previous $\hat{\lambda}_t$ as starting values for the current one. To estimate $\hat{\lambda}_{t_0}$, we suggest using the estimate from the previous iteration in the E-M algorithm as starting values. Convergence is easier for small t, as more data contribute to the local likelihood components. For λ_{t_1}, sensible starting values must be chosen, although the choice is not as important as that for larger t for the reasons discussed above.

3.3. Simulation Study

We compare the standard E-M procedure with the local E-M procedure outlined above. Data were simulated as before, although we encountered problems with convergence in the local E-M procedure when simulated survival times became excessively large and sparse in the right tail of the distribution; to resolve these problems, we have chosen to right censor the final 5% of failure times in this simulation study. The results of this simulation study for 500 convergent repetitions at

sample sizes 200, 500 and 1000 are given in Table 2. Convergence problems occurred mainly when ξ was large, with at worst about 4% of simulations not converging to a global maximum likelihood when $\xi = 1$ and $n = 1000$.

In all cases, the mean local likelihood estimate of ξ is less biased than the standard E-M, although, for larger amounts of frailty and small sample sizes, the mean squared errors can be slightly larger for the local likelihood method due to increased variability in $\hat{\xi}$. The mean, mean squared error and standard deviation are not ideal summary statistics for skewed data. We therefore provide box plots of the estimates of $\hat{\xi}$ for the standard E-M and local E-M methods in Figure 5. The medians of the local likelihood estimates of $\hat{\xi}$ are much closer to the true values than the standard E-M medians. We also note that although the interquartile ranges do appear slightly wider for the local likelihood estimates, there does not appear to be too much difference. Using a local likelihood formulation for the baseline hazard does appear to have solved the bias in estimating ξ and the regression parameters β.

We may also consider transforming $\hat{\xi}$ to account for the skew in our estimates. If we consider $\sqrt{\hat{\xi}}$, which has a distribution which is considerably less skewed than $\hat{\xi}$, and calculate mean squared errors, we find that for all combinations of frailty and sample size, the local likelihood estimates have smaller mean squared errors than the standard E-M.

In the simulation study above, we used a local linear polynomial to approximate the true log baseline hazard from an exponential distribution. It is important to note that an exponential baseline hazard is log-linear. To emphasise the flexibility of the local E-M procedure, we repeat the simulation study with the baseline hazard simulated from the following Weibull mixture of survival functions:

$$S(t) = 0.5 \exp(-2t^{0.9}) + 0.5 \exp(-5t^4).$$

We present box-plots of the estimates of ξ for various sample sizes and values of the frailty variance in Figure 6. The top left plot shows the mixed Weibull baseline hazard used to simulate the data. The results in Figure 6 look very similar to the ones we saw for the exponential baseline hazard. Whereas the median estimates for the standard E-M procedure fall below the true values, those from the local likelihood procedure coincide with the true values far more. Again there is some increased variability in the local likelihood estimates, but the interquartile ranges do not differ greatly in width. The mean $\hat{\beta}$ is within 0.03 of the true value even for sample size 200.

In summary, the problem of consistently underestimating ξ and as a result, the regression parameters β in the standard E-M procedure can be resolved by replacing the discrete Breslow estimate of the cumulative hazard, and its increments, with smooth estimates obtained using local likelihood techniques.

4. Application to Acute Myeloid and Acute Lymphoblastic Leukaemia Survival

In this section, we apply the local likelihood E-M procedure to data collected from the UK North West Regional Leukaemia Register, supplied by Dr. David Gorst of

Table 2. Simulation results from fitting a gamma frailty model with a non-parametric and a continuous local likelihood baseline hazard to simulated data with an exponential baseline hazard, sample sizes 500 and 200, with 500 convergent repetitions and 5% right censoring – the bandwidth is distance to the $n/4$th nearest neighbour.

		Standard E-M procedure					Local E-M procedure				
n	True ξ	Mean($\hat{\xi}$)	sd($\hat{\xi}$)	mse($\hat{\xi}$)	Mean($\hat{\beta}$)	sd($\hat{\beta}$)	Mean($\hat{\xi}$)	sd($\hat{\xi}$)	mse($\hat{\xi}$)	Mean($\hat{\beta}$)	sd($\hat{\beta}$)
1000	0.2	0.1678	0.0998	0.0110	0.9816	0.0764	0.1909	0.0889	0.0080	0.9950	0.0708
1000	0.4	0.3727	0.1251	0.0164	0.9826	0.0805	0.3999	0.1136	0.0129	0.9955	0.0742
1000	0.6	0.5643	0.1342	0.0193	0.9868	0.0782	0.5955	0.1385	0.0192	0.9989	0.0783
1000	0.8	0.7542	0.1550	0.0261	0.9893	0.0846	0.7976	0.1601	0.0256	1.0042	0.0839
1000	1.0	0.9397	0.2018	0.0444	0.9804	0.0928	0.9967	0.2045	0.0418	0.9973	0.0903
500	0.2	0.1608	0.1325	0.0191	0.9769	0.0998	0.1973	0.1316	0.0173	0.9974	0.0976
500	0.4	0.3368	0.1840	0.0379	0.9630	0.1145	0.3894	0.1753	0.0309	0.9886	0.1072
500	0.6	0.5391	0.1846	0.0378	0.9656	0.1057	0.5988	0.2002	0.0401	0.9880	0.1090
500	0.8	0.7282	0.2357	0.0607	0.9696	0.1168	0.8161	0.2555	0.0655	0.9988	0.1173
500	1.0	0.8877	0.2578	0.0791	0.9642	0.1237	0.9996	0.2838	0.0806	0.9969	0.1267
200	0.2	0.1324	0.1610	0.0305	0.9659	0.1378	0.2023	0.1930	0.0373	1.0031	0.1462
200	0.4	0.2744	0.2459	0.0763	0.9281	0.1606	0.3956	0.2725	0.0743	0.9844	0.1635
200	0.6	0.4434	0.2765	0.1010	0.9230	0.1771	0.5921	0.3338	0.1115	0.9790	0.1816
200	0.8	0.6221	0.3252	0.1375	0.9209	0.1779	0.8386	0.3977	0.1596	0.9924	0.1835
200	1.0	0.7459	0.3550	0.1907	0.9113	0.1864	1.0230	0.4451	0.1987	0.9961	0.1928

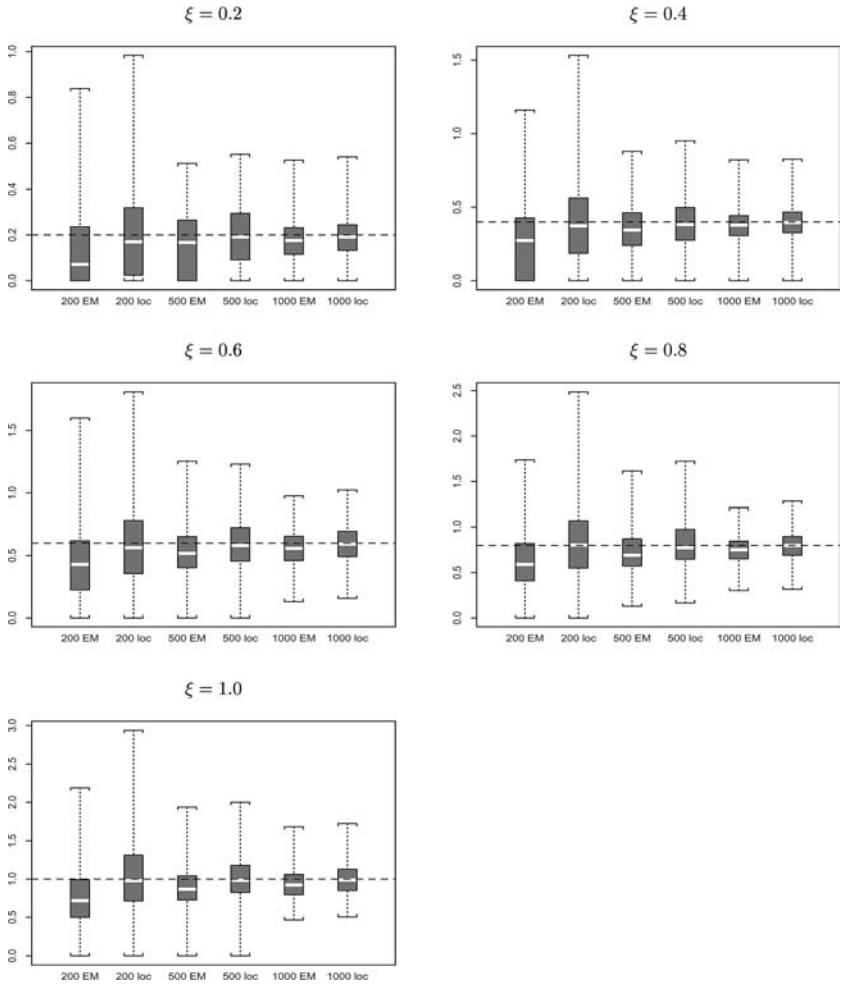

Figure 5. Box plots of $\hat{\xi}$ for the standard and local E-M procedures (abbreviated as EM and loc, respectively) based on 500 repetitions at sample sizes 200, 500 and 1000.

Lancaster Royal Infirmary. The data consist of 1043 patients diagnosed with acute myeloid leukaemia (AML) and 206 patients diagnosed with acute lymphoblastic leukaemia (ALL). We have the following patient information

- survival time in days: there is approximately 16% and 27% censoring for AML and ALL patients respectively;
- age: the median age was 65 (range 14–92) for the AML data and 38 (range 14–94) for the ALL data;

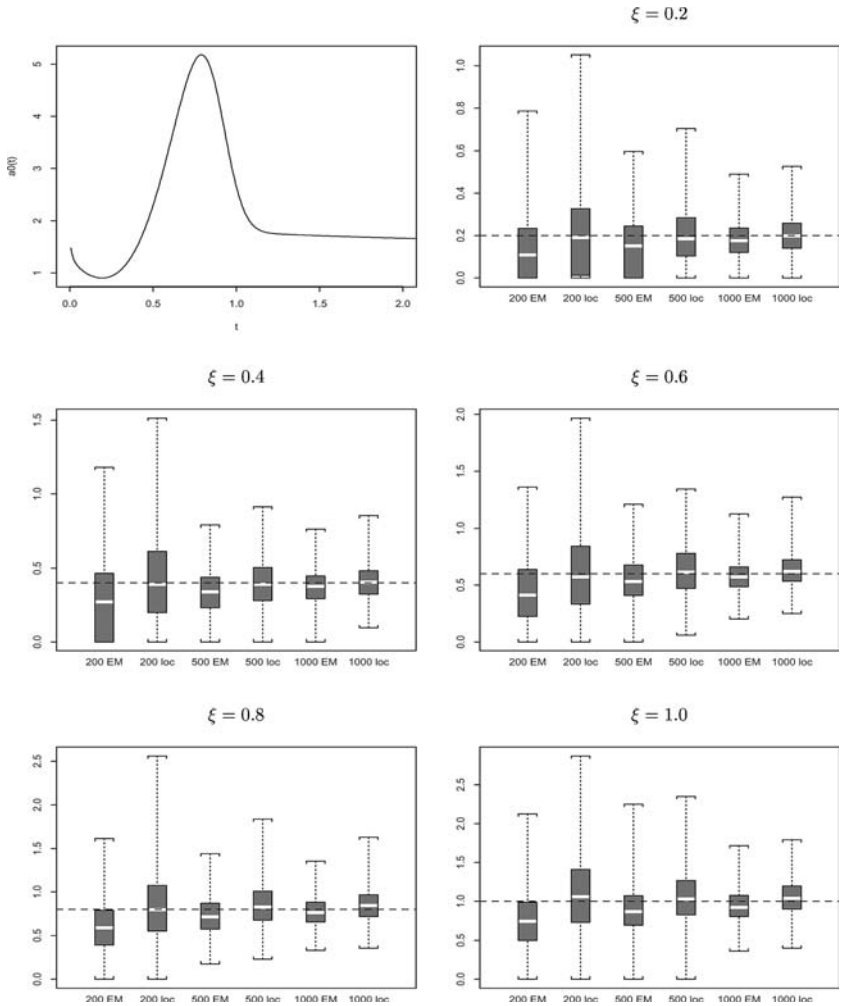

Figure 6. Box plots of $\hat{\xi}$ for the standard and local E-M procedures (abbreviated as EM and loc, respectively) fitted to simulated gamma frailty data with a mixed Weibull baseline hazard. Top-left plot shows true baseline hazard.

- sex: 47.6% and 44.2% of patients are male in the AML and ALL data, respectively;
- white blood cell count (WBC): ranging from 0 to a truncated value of 500 with a median of 7.9 and 8.7 in the AML and ALL groups, respectively;
- Townsend deprivation score (deprivation): this is a measure of deprivation in the area in which an individual lives, a higher score corresponding to a more deprived area.

Table 3. Parameter estimates for the gamma frailty model applied to the AML and ALL data using the standard and local E-M approaches.

	AML Standard E-M			AML Local E-M		
Effect	Coeff	Exp(Coeff)	SE	Coeff	Exp(Coeff)	SE
Age	0.0466	1.0478	0.0042	0.0520	1.0534	0.0041
WBC	0.0053	1.0053	0.0008	0.0063	1.0063	0.0010
Deprivation	0.0549	1.0564	0.0145	0.0604	1.0623	0.0145
ξ	0.7876		0.1799	1.1030		0.2073
L.R. teststatistic		34.8926			49.1416	
Age	0.0597	1.0615	0.0122	0.0647	1.0665	0.0111
WBC	0.0049	1.0049	0.0024	0.0055	1.0054	0.0034
Deprivation	0.0319	1.0324	0.0378	0.0385	1.0389	0.0407
ξ	1.4550		0.6052	1.8416		0.5790
L.R. teststatistic		10.9042			14.5500	

The L.R. test statistic compares the fitted model with standard proportional hazards.

We fit the gamma frailty model to the AML and ALL data sets using both the Nielsen et al. E-M procedure and the local E-M procedure described in Section 3. For both data sets, likelihood ratio tests revealed that the sex effect was not significant when either the standard or local E-M methods were used. As a result, this covariate has been removed from the analysis. The effect of deprivation was statistically significant in the AML data, but not in the ALL data, however, we have chosen to retain this covariate in both analyses.

Results obtained are given in Table 3. For the local E-M approach, bootstrap standard error estimates for the parameters are given, based on 100 re-samples. For the standard E-M approach, estimates of the standard errors were obtained from the variance matrix, as outlined in Andersen et al. (1997).

We cannot compare the maximized log-likelihoods of the standard E-M and local E-M gamma frailty models directly because the two approaches assume different forms for the baseline hazard. Therefore, in Table 3, we provide the likelihood ratio test statistics obtained when either method is compared to the standard Cox proportional hazards regression model. The full likelihood of the Cox model, containing either the Breslow or local likelihood estimate of $\alpha_0(t)$ and $A_0(t)$, is used to obtain these test statistics, rather than the more familiar partial log-likelihood. Although the AML and ALL standard E-M gamma frailty models produce highly significant likelihood ratio test statistics, the corresponding test statistics for the local E-M approach are even more statistically significant.

We plot the estimates of $A_0(t)$ for the standard and local E-M gamma frailty models in Figure 7. The Breslow and local likelihood estimates from the Cox proportional hazards model have been added for reference. For both data sets, the local estimate of $A_0(t)$ for the Cox model is extremely close to the corresponding Breslow

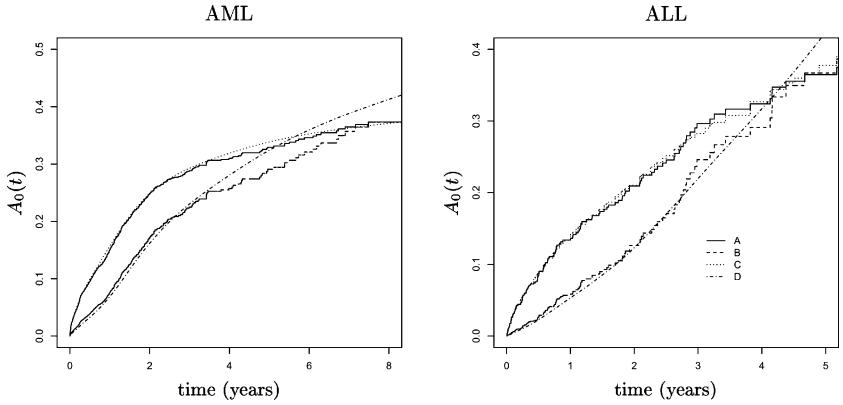

Figure 7. The AML and ALL cumulative baseline hazard estimates from (A) The Cox model with Breslow $A_0(t)$, (B) The standard E-M gamma frailty model, (C) The Cox model with local $A_0(t)$ and (D) The local E-M gamma frailty model.

estimate. For gamma frailty, the local likelihood estimate of $A_0(t)$ is also reasonably close to the Breslow estimate in both data sets, although after 4 years of follow up in the AML data, the local estimate is clearly bigger.

We briefly discuss how estimates of ξ and β differ between approaches. For both the AML and ALL data, the estimate of ξ is very different for the two methods, with the local E-M procedure producing a substantially larger parameter estimate; in the AML data, the local E-M $\hat{\xi}$ is 40% larger than the standard E-M estimate, whereas in the ALL data, the local $\hat{\xi}$ is 26.5% larger. As a result of the larger $\hat{\xi}$, we find that the covariate effect estimates are noticeably larger in the local E-M approach for both data sets.

5. Discussion

In this paper, we have investigated the bias of the Nielsen et al. (1992) E-M procedure for fitting the semi-parametric gamma frailty model. Such an approach is rank-based, and we suggest that although frailty is partly identified through the order in which failures occur, we ignore additional frailty that could have been identified from the survival times themselves, if the Nielsen et al. procedure is used. When the baseline hazard was estimated using local likelihood, the failure times enter the model and the bias problem is removed. It remains for a formal proof to be found to establish the cause of the bias. However, since such proofs usually rely on asymptotic theory, it may be difficult to identify the source of the bias analytically, because the standard E-M procedure has been shown to produce consistent estimates for ξ and β (Parner, 1998).

In Section 3, we used a nearest neighbour bandwidth. The use of a variable bandwidth is essential with frailty data because data get more sparse with time. This may not be an optimal choice for the bandwidth. Cross validation techniques are often used for bandwidth selection, but in the local E-M procedure, we estimate the baseline hazard within the E-M algorithm, and such an approach will be very time consuming. This could be done near the end of the estimation once we are close to the maximum likelihood estimate $\hat{\xi}$.

In Section 4, we applied both the standard and local E-M procedures to the leukaemia data sets. Standard errors of the parameters were estimated using the bootstrap. We have not discussed inference in the local likelihood E-M approach, and this remains as further work. Andersen et al. (1997) have discussed inference for the standard E-M procedure. The cumulative baseline hazard is modelled parametrically with a parameter corresponding to each failure time (this is the same formulation for $A_0(t)$ that we used in Section 2.2); standard errors are obtained from the observed Fisher information matrix in the usual way. This approach gives the same parameter estimates as the Nielsen et al. E-M procedure. Such an approach is not feasible for the local likelihood method, so we recommend estimating the standard errors using the bootstrap, although this is very computer intensive.

An alternative method of fitting the gamma frailty model via a penalized version of the marginal likelihood was suggested recently by Rondeau et al. (2003). This method, which was demonstrated on simulated bivariate gamma frailty data, performs significantly better than the E-M procedure of Nielsen et al. (1992). Like the local E-M procedure outlined in this article, the approach of Rondeau et al. (2003) introduces the survival times themselves into the model. As a result, additional frailty which is ignored by the rank-based E-M approach, is identified by the penalised method. It would be interesting to compare the performance of the local E-M and penalised marginal likelihood approaches on univariate survival data, where the finite sample bias of the standard E-M procedure is more extreme than in the bivariate case, but as yet, publicly available software for the penalized likelihood method is not available.

Acknowledgments

We thank Dr. David Gorst of Lancaster Royal Infirmary for providing the leukaemia data and Hans van Houwelingen and Niels Keiding for useful discussions. Peter Barker was sponsored by the UK Medical Research Council. We are grateful for the constructive comments of an anonymous referee.

References

P. Andersen, J. Klein, K. Knudsen, and R. Palacious, "Estimation of variance in Cox's regression model with shared gamma frailties", *Biometrics* vol. 53 pp. 1475–1484, 1997.

R. Betensky, L.M. Lindsey, J. Ryan, and M. Wand, "A local likelihood proportional hazards model for interval censored data", *Statistics in Medicine* vol. 21 pp. 263–275, 2002.

R. Betensky, J. Lindsey, L. Ryan, and M. Wand, "Local EM estimation of the hazard function for interval-censored data", *Biometrics* vol. 55 pp. 238–245, 1999.

W. Cleveland, "Robust locally weighted regression and smoothing scatterplots", *Journal of the American Statistical Association* vol. 74 pp. 829–836, 1979.

A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society Series B* vol. 39 pp. 1–38, 1977.

C. Elbers, and G. Ridder, "True and spurious duration dependence: The identifiability of the proportional hazard model", *Review of Economic Studies* vol. 49 pp. 403–409, 1982.

J. Heckman, and B. Singer, "The identifiability of the proportional hazard model", *Review of Economic Studies* vol. 51 pp. 231–241, 1984.

R. Henderson, and P. Oman, "Effect of frailty on marginal regression estimates in survival analysis", *Journal of the Royal Statistical Society Series B* vol. 61 pp. 367–379, 1999.

P. Hougaard, *Analysis of Multivariate Survival Data*, Springer-Verlag: New York, 2000.

J. Klein, "Semiparametric estimation of random effects using the Cox model based on the E-M algorithm", *Biometrics* vol. 48 pp. 795–806, 1992.

J. P. Klein and M. L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*, Springer-Verlag: New York, 1997.

G. G. Nielsen, R. G. Gill, and P. K. Andersen, "A counting process approach to maximum likelihood estimation in frailty models", *Scandinavian Journal of Statistics*, vol. 19 pp. 25–43, 1992.

E. Parner, "Asymptotic theory for the correlated gamma-frailty model", *The Annals of Statistics* vol. 26 pp. 183–214, 1998.

V. Rondeau, D. Commenges, and P. Joly, "Maximum penalized lieklihood estimation in a gamma frailty model", *Lifetime data analysis* vol. 9 pp. 139–153, 2003.

T. Therneau, and P. Grambsch, *Modelling Survival Data*, chap. 9. Springer-Verlag: New York, 2000.

R. Tibshirani, and T. Hastie, "Local likelihood estimation", *Journal of the American Statistical Society* vol. 82 pp. 559–567, 1987.