Relationship between Derivatives of the Observed and Full Loglikelihoods and Application to Newton-Raphson Algorithm

Daniel Commenges, Virginie Rondeau

Relationship between derivatives of the observed and full log likelihoods and application to Newton-Raphson algorithm

Daniel Commenges and Virginie Rondeau

INSERM E03 38, ISPED

Université Victor Segalen Bordeaux 2

146 rue Léo Saignat, Bordeaux, 33076, France

Tel: (33) 5 57 57 11 82; Fax (33) 5 56 24 00 81

December 12, 2005

Abstract:

In the case of incomplete data we give general relationships between the first and second derivatives of the log likelihood relative to the full and the incomplete observation set-ups. In the case where these quantities are easy to compute for the full observation set-up we propose to compute their analogue for the incomplete observation set-up using the above mentioned relationships: this involves numerical integrations. Once we are able to compute

these quantities, Newton-Raphson type algorithms can be applied to find the maximum likelihood estimators, together with estimates of their variances. We detail the application of this approach to parametric multiplicative frailty models and we show that the method works well in practice using both a real data and a simulated example. The proposed algorithm outperforms a Newton-Raphson type algorithm using numerical derivatives.

Keywords: likelihood, coarsening, incomplete data, censoring, Radon-Nikodym derivatives, Newton-Raphson algorithm, frailty models.

1 Introduction

In most complex problems we are faced with incomplete data. Even if it is assumed that the mechanism leading to incomplete data is ignorable (Gill et al., 1997) the observed likelihood generally involves integration which can not be solved analytically; this is often the case when observations are interval-censored or when the model includes random effects or frailties. It follows that the first derivative (the score) and the second derivative (the Hessian) of the log likelihood themselves do not have a simple analytical form so that Newton-Raphson algorithm, which is the algorithm of choice when these quantities can be computed, does not seem feasible. This is the main reason of the attractiveness of the EM algorithm (Dempster, Laird and Rubin, 1977) and of the Bayesian approach using the MCMC algorithm (Gilks et al., 1995). However both EM and MCMC algorithms may be time-consuming.

Another reason of the disaffection with the Newton-Raphson algorithm is its lack of robustness when the quadratic approximation is not valid over a large region or when the Hessian is not everywhere positive-definite. This is indeed a problem with the naive version of the Newton-Raphson algorithm; however there exist variants such as the trust region method (see Dennis and Schnabel, 1983) or the Marquardt algorithm (Marquardt, 1963) which are very robust. Also, the score and Hessian can be computed numerically using finite differences (see Overton, 2001, Ch. 11); this possibility has been used successfully in several complex models involving incomplete data. For instance Jacqmin-Gadda et al. (2000) used this technique for a longitudinal analysis with left-censored data and showed that it was more reliable than an EM algorithm. However in complex models there are two problems: computation time and possible loss of accuracy. When the likelihood is computationally demanding and there are many parameters the computation of the Hessian using numerical derivatives is time-consuming. Also, there is a loss of accuracy in the computation of the finite differences due to the cancellation errors (Overton, 2001); when the likelihood itself can not be computed with very high accuracy because it involves numerical integration this loss of accuracy may lead to unacceptable errors in the score and the Hessian and failure of convergence of the algorithm.

The aim of this paper is to use relationships between full and observed scores and Hessians to obtain a faster and more accurate computation of observed score and Hessian which may be used in a Newton-Raphson type algorithm. Such relationships have been given by Louis (1982) for estimating the variances of the estimators when using the EM algorithm; see also Oakes

(1999); Hedeker and Gibbons (1994) implicitly used the relationship for the scores to propose an algorithm using only the scores. In a large part of the statistical literature the observed data are supposed to come from a distribution with probability density function f_X^θ and the likelihood for the observed data x is $\mathcal{L}(\theta; x) = f_X^\theta(x)$. While it is possible to go a long way with such a representation it becomes awkward in complex problems. The limitation of this representation are the following: i) it is not obvious whether the likelihood is a random variable; ii) when speaking of a usual probability density function there is a reference probability which is Lebesgue measure on $\Re^d$ for some d, but it often remains implicit and is given once for all; iii) it is not so easy to represent different amounts of information (this can be done by considering different random variables but this is less flexible than considering σ-fields); iv) it is not clear what is the density when dealing with stochastic processes (even the rather simple case of right-censored survival data can not be treated rigorously in this way).

In this paper we give relationships which are essentially the same as those given by Louis (1982) but we treat the problem in a general framework and using the full power of basic probability theory in which the likelihood is rigorously defined a a Radon-Nikodym derivative, in which we can choose the reference probability and in which information is represented by σ-fields; one particular benefit of this approach is that the likelihood may be relative to the observation of stochastic processes such as in Andersen et al. (1993) or Barndorff-Nielsen and Sorensen (1994); however we restrict to parametric models.

In section 2 we first exhibit two martingales related to the score and the

Hessian processes from which we deduce the relationships between observed and full scores and Hessians. In section 3 we show how these relationships can be exploited to improve Newton-Raphson-type algorithms in the case of incomplete observations: two main situations are considered, the coarsening situation and the random effect situation. The algorithm is applied in section 4 to parametric proportional hazards frailty models. In section 5 it is illustrated on a Weibull model with normal frailties; the proposed algorithm is compared to a Marquardt algorithm using numerical derivatives on simulated data sets and applied to a real data example. We conclude in section 6.

2 Relationship between derivatives of the observed and full log likelihoods

2.1 Martingales related to the derivatives of the log likelihood

Consider a measurable space $(\Omega, \mathcal{F})$ and a family of measures $\{P_\theta\}_{\theta\in\Theta}$, where Θ is a "nice" subset of $\Re^m$ so that this defines a regular parametric model (Bickel et al, 1993); we assume that the P_θ's are absolutely continuous relatively to a dominant measure P_{θ_0} (this means that $P_{\theta_0} \in \{P_\theta\}_{\theta\in\Theta}$ but the results would be unchanged if we took a reference probability P_0 outside of the model). The likelihood ratio on $\mathcal{F}$ is defined by:

$$\mathcal{L}_{\mathcal{F}}^{\theta/\theta_0} = \frac{dP_\theta}{dP_{\theta_0|\mathcal{F}}} \quad \text{a.s.}$$

where $\frac{dP_\theta}{dP_{\theta_0}|_{\mathcal{F}}}$ is the Radon-Nikodym derivative of P_θ relatively to P_{θ_0}. Recall that $\frac{dP_\theta}{dP_{\theta_0}|_{\mathcal{F}}}$ is the $\mathcal{F}$-measurable random variable such that $P_\theta(F) = \int_F \frac{dP_\theta}{dP_{\theta_0}|_{\mathcal{F}}} dP_{\theta_0}, F \in \mathcal{F}$. If the space is equipped with a filtration $(\mathcal{F}_t)$, let us denote $\mathcal{L}_t^{\theta/\theta_0} = \mathcal{L}_{\mathcal{F}_t}^{\theta/\theta_0}$; then we can consider the stochastic process $(\mathcal{L}_t^{\theta/\theta_0})_{t\geq 0}$. The following results apply in continuous time where the filtration is right-continuous, or in discrete time. The process $(\mathcal{L}_t^{\theta/\theta_0})$ is a P_{θ_0}-martingale (Barndorff-Nielsen and Sorensen, 1994) with expectation equal to 1. We further assume as in Barndorff-Nielsen and Sorensen (1994) that for all $t \geq 0$: i) $\mathcal{L}_t^{\theta/\theta_0}$ is twice-continuously differentiable with respect to θ, ii) the class of likelihood gradients $\frac{\partial \mathcal{L}_t^{\theta/\theta_0}}{\partial \theta}$ and random matrices $\frac{\partial^2 \mathcal{L}_t^{\theta/\theta_0}}{\partial \theta^2}$ are locally dominated integrable under P_{θ_0}. Let us denote $L_t^{\theta/\theta_0} = \log \mathcal{L}_t^{\theta/\theta_0}$ the log likelihood at time t. We note that $\frac{\partial L_t^{\theta/\theta_0}}{\partial \theta}$ does not depend on θ_0. This can been seen by starting from $\mathcal{L}_t^{\theta/\theta_0} = \mathcal{L}_t^{\theta/\theta_1}\mathcal{L}_t^{\theta_1/\theta_0}$; taking logs and differentiating we find that $\frac{\partial L_t^{\theta/\theta_0}}{\partial \theta} = \frac{\partial L_t^{\theta/\theta_1}}{\partial \theta}$; we shall denote U_t^θ this common derivative. Let us now consider the Hessian of the log likelihood: $H_t^\theta = \frac{\partial U_t^\theta}{\partial \theta}$. We can now state a theorem defining two interesting martingales.

Theorem 1 *The processes* $(\mathcal{L}_t^{\theta/\theta_0} U_t^\theta)$ *and* $(\mathcal{L}_t^{\theta/\theta_0}(H_t^\theta + U_t^\theta U_t^{\theta T}))$ *are* P_{θ_0}*-martingales with zero mean.*

Proof. Since $(\mathcal{L}_t^{\theta/\theta_0})$ is a P_{θ_0}-martingale with mean equal to 1, its derivatives $\frac{\partial \mathcal{L}_t^{\theta/\theta_0}}{\partial \theta}$ and $\frac{\partial^2 \mathcal{L}_t^{\theta/\theta_0}}{\partial \theta^2}$ are zero-mean P_{θ_0}-martingales. These two martingales can be expressed in terms of U_t^θ and H_t^θ to obtain the two martingales of the theorem. ❑

A particular case of this result has been given by Barndorff-Nielsen and Sorensen (1994): if one takes $\theta_0 = \theta$ we obtain: (U_t^θ) and $(H_t^\theta + U_t^\theta U_t^{\theta T})$ are P_θ-martingales with zero mean.

2.2 Relationship between derivatives of the observed and full log likelihoods

Consider the general statistical model depicted by a measurable space $(\Omega, \mathcal{F})$ and a family of measures $\{P_\theta\}_{\theta\in\Theta}$ absolutely continuous relatively to a dominant measure P_{θ_0}. $\mathcal{F}$ will be interpreted as the full σ-field, in that it contains all the interesting events in the problem we wish to model. We now consider the problem of incomplete observation, that is, the case where all the events of interest are not observed. In that case we represent the observation by the σ-field $\mathcal{O}$ such that $\mathcal{O} \subset \mathcal{F}$ (in a strict sense). Thus our model can be depicted by the 4-uplet $(\Omega, \mathcal{F}, \mathcal{O}, \{P_\theta\})$. Note that if the mechanism leading to incomplete data is random, $\mathcal{F}$ must also include the events which this mechanism generates (see Commenges and Gégout-Petit, 2005). In this framework we can consider both the full likelihood $\mathcal{L}_{\mathcal{F}}^{\theta/\theta_0}$ and the observed likelihood $\mathcal{L}_{\mathcal{O}}^{\theta/\theta_0}$. By direct application of Theorem 1 to the (discrete) filtration $(\Psi, \mathcal{O}, \mathcal{F})$, where $\Psi = \Omega \vee \emptyset$ is the trivial σ-field, we obtain the following result:

Corollary 1 *We have the following relationships between first and second derivatives of the full and observed log likelihood denoted respectively* $U_{\mathcal{F}}^\theta, H_{\mathcal{F}}^\theta$ *and* $U_{\mathcal{O}}^\theta, H_{\mathcal{O}}^\theta$ *:*

i) $\mathcal{L}_{\mathcal{O}}^{\theta/\theta_0} U_{\mathcal{O}}^\theta = \mathrm{E}_{\theta_0}[\mathcal{L}_{\mathcal{F}}^{\theta/\theta_0} U_{\mathcal{F}}^\theta | \mathcal{O}]$

ii) $\mathcal{L}_{\mathcal{O}}^{\theta/\theta_0}(H_{\mathcal{O}}^\theta + U_{\mathcal{O}}^\theta U_{\mathcal{O}}^{\theta T}) = \mathrm{E}_{\theta_0}[\mathcal{L}_{\mathcal{F}}^{\theta/\theta_0}(H_{\mathcal{F}}^\theta + U_{\mathcal{F}}^\theta U_{\mathcal{F}}^{\theta T})|\mathcal{O}]$

iii) $\mathrm{E}_{\theta_0}[\mathcal{L}_{\mathcal{O}}^{\theta/\theta_0}(H_{\mathcal{O}}^\theta + U_{\mathcal{O}}^\theta U_{\mathcal{O}}^{\theta T})] = \mathrm{E}_{\theta_0}[\mathcal{L}_{\mathcal{F}}^{\theta/\theta_0}(H_{\mathcal{F}}^\theta + U_{\mathcal{F}}^\theta U_{\mathcal{F}}^{\theta T})] = 0$.

If we take $\theta_0 = \theta$ we have the simpler equalities: $U_{\mathcal{O}}^\theta = \mathrm{E}_\theta[U_{\mathcal{F}}^\theta|\mathcal{O}]$; $H_{\mathcal{O}}^\theta + U_{\mathcal{O}}^\theta U_{\mathcal{O}}^{\theta T} = \mathrm{E}_\theta[H_{\mathcal{F}}^\theta + U_{\mathcal{F}}^\theta U_{\mathcal{F}}^{\theta T}|\mathcal{O}]$ and $I_{\mathcal{O}}^\theta - \mathrm{E}_\theta[U_{\mathcal{O}}^\theta U_{\mathcal{O}}^{\theta T}] = I_{\mathcal{F}}^\theta - \mathrm{E}_\theta[U_{\mathcal{F}}^\theta U_{\mathcal{F}}^{\theta T}] = 0$,

where $I_{\mathcal{O}}^{\theta} = -\mathrm{E}_{\theta}[H_{\mathcal{O}}^{\theta}]$ and $I_{\mathcal{F}}^{\theta} = -\mathrm{E}_{\theta}[H_{\mathcal{F}}^{\theta}]$. The general formulae of the corollary may seem unnecessarily complicated; however they preserve the choice of the reference probability under which the expectations must be taken: particular choices may render the computations much simpler.

3 Application to Newton-Raphson type algorithms

3.1 General approach

We consider the case where $\mathcal{L}_{\mathcal{F}}^{\theta/\theta_0}$ has a sufficiently simple analytical form so that $U_{\mathcal{F}}^{\theta}$ and $H_{\mathcal{F}}^{\theta}$ can also be computed analytically, while $\mathcal{L}_{\mathcal{O}}^{\theta/\theta_0}$, $U_{\mathcal{O}}^{\theta}$ and $H_{\mathcal{O}}^{\theta}$ do not have analytical form. The Newton-Raphson algorithm is the fastest maximization algorithm but it requires $U_{\mathcal{O}}^{\theta}$ and $H_{\mathcal{O}}^{\theta}$ for constructing an approximate quadratic model at the current value θ; this is also the case for robust versions of the Newton-Raphson algorithm such as the algorithm of Marquardt (1963). The formulae of Corollary 1 allow to compute $U_{\mathcal{O}}^{\theta}$ and $H_{\mathcal{O}}^{\theta}$ by numerical integration with essentially the same precision as $\mathcal{L}_{\mathcal{O}}^{\theta/\theta_0}$. We have:

$$U_{\mathcal{O}}^{\theta} = (\mathcal{L}_{\mathcal{O}}^{\theta/\theta_0})^{-1}\mathrm{E}_{\theta_0}[\mathcal{L}_{\mathcal{F}}^{\theta/\theta_0}U_{\mathcal{F}}^{\theta}|\mathcal{O}], \tag{1}$$

and

$$H_{\mathcal{O}}^{\theta} = -U_{\mathcal{O}}^{\theta}U_{\mathcal{O}}^{\theta T} + (\mathcal{L}_{\mathcal{O}}^{\theta/\theta_0})^{-1}\mathrm{E}_{\theta_0}[\mathcal{L}_{\mathcal{F}}^{\theta/\theta_0}(H_{\mathcal{F}}^{\theta} + U_{\mathcal{F}}^{\theta}U_{\mathcal{F}}^{\theta T})|\mathcal{O}]. \tag{2}$$

In the case where $\mathcal{F} = \mathcal{O} \vee \sigma(\eta)$ (that is $\mathcal{F}$ is the smallest σ-field containing both $\mathcal{O}$ and $\sigma(\eta)$), where η is a random variable, all random variables can be expressed as measurable functions of η and an $\mathcal{O}$-measurable

variable. So, for computing the above expressions our problem is to compute conditional expectations of the form $\mathrm{E}_{\theta_0}[f(\eta, X)|\mathcal{O}]$, where X is an $\mathcal{O}$-measurable variable. Using the disintegration theorem (Kallenberg, 2001) we have: $\mathrm{E}_{\theta_0}[f(\eta, X)|\mathcal{O}] = \int f(s, X)\nu^{\theta_0}_{\eta|\mathcal{O}}(ds)$, where $\nu^{\theta_0}_{\eta|\mathcal{O}}(.)$ is the conditional law of η given $\mathcal{O}$ under the probability P_{θ_0}. It may be more or less difficult to compute this integral, and to begin with, the conditional law $\nu^{\theta_0}_{\eta|\mathcal{O}}(.)$. We shall consider two main situations, the coarsening situation and the random effect situation.

In the typical coarsening situation we are interested in the law of a random element T but we do not observe completely T; that is, the σ-field of all the events generated by T, $\sigma(T)$, is not included in the observed σ-field, $\mathcal{O}$. The following discussion applies to the case where the mechanism is random but ignorable, in which case we can treat the likelihood as if the mechanism was fixed at the observed value. So, considering the mechanism as deterministic, we have $\mathcal{O} \subset \sigma(T)$; η in the above formula is simply T itself. For giving an illustration of this case let us consider the rather simple case where T is a survival time; equivalently the survival situation can be represented by a $0-1$ counting process $N = (N_t)$. Consider the case where this process is observed at discrete times $v_0, \ldots, v_m$, with $v_m \leq C$. In terms of the random variable T this means that we observe in which interval $(v_{l-1}, v_l], l = 1, \ldots, m$ or $(v_m, C]$ T falls. The observed σ-field can be written for instance $\mathcal{O} = \sigma(1_{\{T\in(v_{l-1},v_l]\}}, l = 1, \ldots, m+1)$, with the convention $v_{m+1} = \infty$. Choosing a reference probability under which T has an exponential distribution with parameter 1 Jacod's formula gives us the likelihood for the observation of N

on $[0, C]$:

$$\mathcal{L}^{\theta}_{\mathcal{F}_C} = \lambda^{\theta}(T \wedge C)^{\delta} \exp[-\Lambda^{\theta}(T \wedge C)] e^{T \wedge C}, \tag{3}$$

where $\delta = 1_{T \leq C}$, $\lambda^{\theta}(.)$ (resp. $\Lambda^{\theta}(.)$) is the intensity (resp. the cumulative intensity) of N; Aalen's multiplicative model specifies that $\lambda^{\theta}(t) = 1_{\{N_t=0\}}\alpha^{\theta}(t)$ where $\alpha^{\theta}(t)$ is the hazard function. In case of coarsening it convenient to give the likelihood locally, that is on particular events that can be called atoms or pseudo-atoms (Commenges and Gégout-Petit, 2005b). Here we give the likelihood on the event $(T \in (v_{l-1}, v_l])$. Using the fact that $\nu(ds) = \frac{1_{[v_{l-1}, v_l]}(s)e^{-s}}{(e^{-v_{l-1}} - e^{-v_l})} ds$ gives the law of T given $(T \in (v_{l-1}, v_l])$ for $l \leq m$, we obtain that on this event the likelihood is:

$$\mathcal{L}^{\theta/\theta_0}_{\mathcal{O}} = \frac{1}{e^{-v_{l-1}} - e^{-v_l}} \int_{v_{l-1}}^{v_l} \lambda^{\theta}(s) \exp[-\Lambda^{\theta}(\tilde{T})] ds = \frac{1}{e^{-v_{l-1}} - e^{-v_l}} \int_{v_{l-1}}^{v_l} \alpha^{\theta}(s) \exp[-A^{\theta}(s)] ds,$$

where $A(t) = \int_0^t \alpha^{\theta}(s) ds$.

This can be seen to be proportional to $F(v_l) - F(v_{l-1})$, a result that could have been obtained in a simpler way. We can compute

$$U^{\theta}_{\mathcal{F}} = \frac{\partial \log \mathcal{L}^{\theta/\theta_0}_{\mathcal{F}}}{\partial \theta} = \frac{\frac{\partial \lambda^{\theta}}{\partial \theta}(T \wedge C)}{\lambda^{\theta}(T \wedge C)} - \frac{\partial \Lambda^{\theta}}{\partial \theta}(T \wedge C).$$

Hence applying formula (1) we obtain:

$$U^{\theta}_{\mathcal{O}} = [\mathcal{L}^{\theta/\theta_0}_{\mathcal{O}}]^{-1} \frac{1}{e^{-v_{l-1}} - e^{-v_l}} \int_{v_{l-1}}^{v_l} \Big[\frac{\frac{\partial \lambda^{\theta}}{\partial \theta}(s)}{\lambda^{\theta}(s)} - \frac{\partial \Lambda^{\theta}}{\partial \theta}(s) \Big] \lambda^{\theta}(s) \exp[-\Lambda^{\theta}(s)] ds.$$

The observed Hessian can be computed similarly by applying formula (2). Such a computation is useful only if $F(s)$ does not have an analytical form, which happens in some parametric models or when the distribution is approached on a basis of spline in a semi-parametric approach. In that case this computation may be faster and more accurate than using numerical derivatives. Numerical derivatives still work well in this rather simple

problem (Joly et al., 2001) so this approach may be more interesting in more complex problems involving multivariate counting processes such as in Commenges and Joly (2004) and Commenges and Gégout-Petit (2005b). We do not develop more this application in this paper in which we will rather focus on the second main situation, the random effect one.

In the typical random effect situation we are interested in the conditional distribution of T given a random effect η. Most often random effects are introduced to model dependence within a group of random variables T_j (for modeling clusters or repeated measurements for instance). Most often the best choice of the reference probability will be such that η and the T_j's are independent, so that the T_j's will themselves be mutually independent and we have $\nu_{\eta|\mathcal{O}} = \nu_\eta$. This choice meets what is done in a conventional approach where the reference measure remains implicit; however since in the conventional approach the implicit reference measure is Lebesgue measure which is not a probability measure the term "independence" is generally not used. A particular application within the random effect situation is presented in section 4; in this application the likelihood has the particular structure depicted in section 3.2 which allows savings in computation burden.

The numerical integration can be done by Gaussian quadrature or by simulation. For instance it can be approximated by $M^{-1}\sum_j^M f(g_j, X)$, where g_j are realizations from the conditional distribution of η given $\mathcal{O}$ in the probability P_{θ_0}. As discussed above if P_{θ_0} is chosen in order to make η and $\mathcal{O}$ independent, the g_j's are taken from the marginal distribution of η. For low dimensional integrals Gaussian quadrature is more efficient.

In practice the above formulae are applied to a sample which most often

can be divided into independent parts: for instance the sampled is formed of n independent random variables or of G independent groups (each group may represents clusters or repeated measurements) indexed by i. In that case the observed σ-field can be written $\mathcal{O} = \vee_i \mathcal{O}_i$ and similarly we may consider a full σ-field which can be written $\mathcal{F} = \vee_i \mathcal{F}_i$.

3.2 Cases where the computations can be reduced

Typically the above algorithm involves $m(m+3)/2$ numerical integrals (for each group). In some cases the number of numerical integrals can be reduced. Consider the case where the log likelihood takes the form:

$$L_{\mathcal{F}}^{\theta/\theta_0} = \sum_{k=1}^{K} A_k B_k(\eta),$$

where the A_k's are $\mathcal{O}$-measurable and the B_k's depend on a small number q_k of parameters. By Corollary 1 we have:

$$U_{\mathcal{O}}^{\theta} = \sum_{k=1}^{K} (\mathcal{L}_{\mathcal{O}}^{\theta/\theta_0})^{-1} \Big[A_k \mathrm{E}_{\theta_0}[\mathcal{L}_{\mathcal{F}}^{\theta/\theta_0} \frac{\partial B_k}{\partial \theta} | \mathcal{O}] + \frac{\partial A_k}{\partial \theta} \mathrm{E}_{\theta_0}[\mathcal{L}_{\mathcal{F}}^{\theta/\theta_0} B_k | \mathcal{O}] \Big].$$

The number of numerical integrals for computing the score is thus $K(1 + \sum_{k=1}^{K} q_k)$, which may be less than m. This special structure also allows savings for the computation of the Hessian. In the next section we study in detail the multiplicative frailty models which present such a structure.

4 Score and Hessian for parametric proportional hazards shared frailty models

Consider the case where we have right-censored observations of failure times for n subjects scattered in G groups of size $n_i, i = 1, \ldots, G$, with $\sum_i n_i = n$. For subject j in group i we observe $\tilde{T}_{ij} = \min(T_{ij}, C_{ij})$ and $\delta_{ij} = 1_{\{T_{ij} \leq C_{ij}\}}$, where T_{ij} is the failure time, C_{ij} is a censoring time which will be treated as fixed (if it is random, the mechanism leading to incomplete data is assumed to be ignorable). We assume a parametric frailty model:

$$\alpha_{ij}^{\theta}(t) = \alpha_0^{\gamma}(t) \exp(z_{ij}\beta + \omega\eta_i),$$

where $\alpha_{ij}(t)$ is the risk function for the distribution of T_{ij}, $\alpha_0^{\gamma}(t)$ is the baseline risk function, z_{ij} is a vector of explanatory variables, β a vector of regression coefficients and the η_i's are independently identically distributed random variables (the shared frailties) with $\mathrm{E}_{\theta_0}(\eta_i) = 0$ and $\mathrm{var}_{\theta_0}\eta_i = 1$. We assume a parametric model for $\alpha_0^{\gamma}(t)$ indexed by γ and we denote by $\theta = (\gamma, \beta, \omega)$ the set of all the parameters of the model; the distribution of the η_i does not depend on θ.

We now define observed and "full" σ-fields. The observed σ-field for group i is $\mathcal{O}_i = \sigma(\tilde{T}_{ij}, \delta_{ij}, j = 1, \ldots, n_i)$ and the total observed σ-field is $\mathcal{O} = \vee_i \mathcal{O}_i$. We may define a "full" σ-fields $\mathcal{F}_i$ for group i which includes in addition to the observations the events generated by η_i: that is we will denote $\mathcal{F}_i = \sigma(\tilde{T}_{ij}, \delta_{ij}, j = 1, \ldots, n_i; \eta_i)$; the total full σ-field will be $\mathcal{F} = \vee_i \mathcal{F}_i$. The full likelihood ratio can be written using Jacod formula (Jacod, 1975) and, following and idea of Aalen (1978), we shall choose a particularly simple reference probability: we choose P_{θ_0} such that the T_{ij} are independent

random variables with exponential distributions with parameter κ. With this choice we obtain:

$$\mathcal{L}_{\mathcal{F}_i}^{\theta/\theta_0} = \prod_{j=1}^{n_i} (\alpha_{ij}^{\theta}(\tilde{T}_{ij}) \times \frac{1}{\kappa})^{\delta_{ij}} e^{-A_{ij}^{\theta}(\tilde{T}_{ij})+\kappa\tilde{T}_{ij}},$$

where $A_{ij}^{\theta}(t) = \int_0^t \alpha_{ij}^{\theta}(u)du$. The full likelihood ratio $\mathcal{L}_{\mathcal{F}_i}^{\theta/\theta_0}$ is a function of the unobserved η_i and we can make this explicit by writing $\mathcal{L}_{\mathcal{F}_i}^{\theta/\theta_0}(\eta_i)$. With the choice of the reference probability which makes η_i independent from $\mathcal{O}_i$ we have $\nu_{\eta_i|\mathcal{O}_i}(ds) = \phi_{\eta_i}(s)ds$ where $\phi_{\eta_i}(s)$ is the probability density function of η_i (under P_{θ_0}); the observed likelihood can then be written

$$\mathcal{L}_{\mathcal{O}_i}^{\theta/\theta_0} = \int \mathcal{L}_{\mathcal{F}_i}^{\theta/\theta_0}(s)\phi_{\eta_i}(s)ds.$$

In view of the complexity of $\mathcal{L}_{\mathcal{F}_i}^{\theta/\theta_0}(s)$ it is not surprising that $\mathcal{L}_{\mathcal{O}_i}^{\theta/\theta_0}$ does not have in general an analytical form; this happens if the frailty has a gamma distribution (Nielsen et al., 1992) and this fact explains to a large extent the success of the gamma-frailty model. The present approach is not useful for the gamma-frailty model but may be useful for any other choice.

We proceed by computing the log likelihood:

$$L_{\mathcal{F}_i}^{\theta/\theta_0} = \sum_{j=1}^{n_i} [\delta_{ij}[\log \alpha_0^{\gamma}(\tilde{T}_{ij}) + z_{ij}\beta + \omega\eta_i - \log\kappa] - A_0^{\gamma}(\tilde{T}_{ij})e^{z_{ij}\beta}e^{\omega\eta_i} + \kappa\tilde{T}_{ij}].$$

In order to obtain more synthetic results we rewrite this log-likelihood in separating what is observed and what is not as:

$$L_{\mathcal{F}_i}^{\theta/\theta_0} = a_i + d_i\omega\eta_i - b_i e^{\omega\eta_i} + \kappa\tilde{T}_i - d_i \log\kappa$$

where $a_i = \sum_{j=1}^{n_i} \delta_{ij}[\log \alpha_0^{\gamma}(\tilde{T}_{ij}) + z_{ij}\beta]$, $b_i = \sum_{j=1}^{n_i} A_0^{\gamma}(\tilde{T}_{ij})e^{z_{ij}\beta}$, $\xi = (\gamma, \beta)$, $d_i = \sum_{j=1}^{n_i} \delta_{ij}$ and $\tilde{T}_i = \sum_{j=1}^{n_i} \tilde{T}_{ij}$. With these notations the full score and Hessian are:

$$U_{\mathcal{F}_i,\xi} = \frac{\partial a_i}{\partial \xi} - \frac{\partial b_i}{\partial \xi} e^{\omega \eta_i},$$
$$U_{\mathcal{F}_i,\omega} = d_i \eta_i - b_i \eta_i e^{\omega \eta_i},$$
$$H_{\mathcal{F}_i,\xi\xi} = \frac{\partial^2 a_i}{\partial \xi^2} - \frac{\partial^2 b_i}{\partial \xi^2} e^{\omega \eta_i},$$
$$H_{\mathcal{F}_i,\omega\omega} = -b_i \eta_i^2 e^{\omega \eta_i},$$
$$H_{\mathcal{F}_i,\xi\omega} = -\frac{\partial b_i}{\partial \xi} \eta_i e^{\omega \eta_i}.$$

We can now compute the observed score and Hessian by applying Corollary (1). We note that the desired quantities essentially depend on the conditional expectations of full score and Hessian multiplied by $\mathcal{L}_{\mathcal{F}_i}^{\theta/\theta_0}$ which, because a_i and b_i are $\mathcal{O}_i$-measurable are:

$$\mathrm{E}_{\theta_0}(\mathcal{L}_{\mathcal{F}_i}^{\theta/\theta_0} U_{\mathcal{F}_i,\xi} | \mathcal{O}_i) = \frac{\partial a_i}{\partial \xi} \mathcal{L}_{\mathcal{O}_i}^{\theta/\theta_0} - \frac{\partial b_i}{\partial \xi} \mathrm{E}_{\theta_0}(\mathcal{L}_{\mathcal{F}_i}^{\theta/\theta_0} e^{\omega \eta_i} | \mathcal{O}_i),$$

$$\mathrm{E}_{\theta_0}(\mathcal{L}_{\mathcal{F}_i}^{\theta/\theta_0} U_{\mathcal{F}_i,\omega} | \mathcal{O}_i) = d_i \mathrm{E}_{\theta_0}(\mathcal{L}_{\mathcal{F}_i}^{\theta/\theta_0} \eta_i | \mathcal{O}_i) - b_i \mathrm{E}_{\theta_0}(\mathcal{L}_{\mathcal{F}_i}^{\theta/\theta_0} \eta_i e^{\omega \eta_i} | \mathcal{O}_i),$$

$$\mathrm{E}_{\theta_0}(\mathcal{L}_{\mathcal{F}_i}^{\theta/\theta_0} H_{\mathcal{F}_i,\xi\xi} | \mathcal{O}_i) = \frac{\partial^2 a_i}{\partial \xi^2} \mathcal{L}_{\mathcal{O}_i}^{\theta/\theta_0} - \frac{\partial^2 b_i}{\partial \xi^2} \mathrm{E}_{\theta_0}(\mathcal{L}_{\mathcal{F}_i}^{\theta/\theta_0} e^{\omega \eta_i} | \mathcal{O}_i), \tag{4}$$

$$\mathrm{E}_{\theta_0}(\mathcal{L}_{\mathcal{F}_i}^{\theta/\theta_0} H_{\mathcal{F}_i,\omega\omega} | \mathcal{O}_i) = -b_i \mathrm{E}_{\theta_0}(\mathcal{L}_{\mathcal{F}_i}^{\theta/\theta_0} \eta_i^2 e^{\omega \eta_i} | \mathcal{O}_i), \tag{5}$$

$$\mathrm{E}_{\theta_0}(\mathcal{L}_{\mathcal{F}_i}^{\theta/\theta_0} H_{\mathcal{F}_i,\xi\omega} | \mathcal{O}_i) = -\frac{\partial b_i}{\partial \xi} \mathrm{E}_{\theta_0}(\mathcal{L}_{\mathcal{F}_i}^{\theta/\theta_0} \eta_i e^{\omega \eta_i} | \mathcal{O}_i)., \tag{6}$$

So for computing these terms only five integrals for each group are required, whatever the number of parameters in the model, and it is not necessary to compute them with a very high accuracy because only simple computations lead to the final results used by the Newton-Raphson (or Marquardt) algorithm. This is in contrast to computing the log likelighood for computing numerical first and second derivatives: then the cancellation errors produce an important loss of significant digits, so the initial computations must be very

accurate. The five integrals are: $\mathcal{L}^{\theta/\theta_0}_{\mathcal{O}_i} = \mathrm{E}_{\theta_0}(\mathcal{L}^{\theta/\theta_0}_{\mathcal{F}_i}|\mathcal{O}_i)$, $I_0 = \mathrm{E}_{\theta_0}(\mathcal{L}^{\theta/\theta_0}_{\mathcal{F}_i}\eta_i|\mathcal{O}_i)$, $I_1 = \mathrm{E}_{\theta_0}(\mathcal{L}^{\theta/\theta_0}_{\mathcal{F}_i}e^{\omega\eta_i}|\mathcal{O}_i)$, $I_2 = \mathrm{E}_{\theta_0}(\mathcal{L}^{\theta/\theta_0}_{\mathcal{F}_i}\eta_i e^{\omega\eta_i}|\mathcal{O}_i)$, $I_3 = \mathrm{E}_{\theta_0}(\mathcal{L}^{\theta/\theta_0}_{\mathcal{F}_i}\eta_i^2 e^{\omega\eta_i}|\mathcal{O}_i)$. It may appear difficult to compute these conditional expectations but as already discussed our choice of a simple reference probability pays off: because on P_{θ_0}, η_i is independent from the observations, the conditioning for computing I_j can be removed: we have for instance, considering $\mathcal{L}^{\theta/\theta_0}_{\mathcal{F}_i}$ as a function of η_i, $I_1 = \int \mathcal{L}^{\theta/\theta_0}_{\mathcal{F}_i}(s)e^{\omega s}\phi(s)ds$, where $\phi(s)$ is the p.d.f. of the frailty; so we can use straightforward gaussian quadrature or Monte-Carlo integration techniques for computing them. The final formulae for the score are:

$$U_{\mathcal{O}_i,\xi} = \frac{\partial a_i}{\partial \xi} - \frac{\partial b_i}{\partial \xi}(\mathcal{L}^{\theta/\theta_0}_{\mathcal{O}_i})^{-1} I_1 \tag{7}$$

$$U_{\mathcal{O}_i,\omega} = d_i(\mathcal{L}^{\theta/\theta_0}_{\mathcal{O}_i})^{-1} I_0 - b_i(\mathcal{L}^{\theta/\theta_0}_{\mathcal{O}_i})^{-1} I_2 \tag{8}$$

As for the Hessian, we have from Corollary 1:

$$H^{\theta}_{\mathcal{O}_i} = -U^{\theta}_{\mathcal{O}_i}U^{\theta T}_{\mathcal{O}_i} + (\mathcal{L}^{\theta/\theta_0}_{\mathcal{O}_i})^{-1}[\mathrm{E}_{\theta_0}[\mathcal{L}^{\theta/\theta_0}_{\mathcal{F}_i}H^{\theta}_{\mathcal{F}_i}|\mathcal{O}_i] + \mathrm{E}_{\theta_0}[\mathcal{L}^{\theta/\theta_0}_{\mathcal{F}_i}U^{\theta}_{\mathcal{F}_i}U^{\theta T}_{\mathcal{F}_i}|\mathcal{O}_i]]. \tag{9}$$

The second term in the brackets can be computed numerically after having computed the observed scores from formulae (7) and (8); the blocks of the first matrix are given by formulae (4, 5, 6). It remains for us to compute the blocks of $\mathrm{E}_{\theta_0}[\mathcal{L}^{\theta/\theta_0}_{\mathcal{F}_i}U^{\theta}_{\mathcal{F}_i}U^{\theta T}_{\mathcal{F}_i}|\mathcal{O}_i]$. We find that getting rid of the conditioning by the same argument as above, these quantities involve in addition to I_0, I_1, I_2, I_3 four other integrals $I_4 = \mathrm{E}_{\theta_0}(\mathcal{L}^{\theta/\theta_0}_{\mathcal{F}_i}e^{2\omega\eta_i})$, $I_5 = \mathrm{E}_{\theta_0}(\mathcal{L}^{\theta/\theta_0}_{\mathcal{F}_i}\eta_i^2 e^{2\omega\eta_i})$, $I_6 = \mathrm{E}_{\theta_0}(\mathcal{L}^{\theta/\theta_0}_{\mathcal{F}_i}\eta_i e^{2\omega\eta_i})$ and $I_7 = \mathrm{E}_{\theta_0}(\mathcal{L}^{\theta/\theta_0}_{\mathcal{F}_i}\eta_i^2)$. The formulae are:

$$\mathrm{E}_{\theta_0}[\mathcal{L}^{\theta/\theta_0}_{\mathcal{F}_i}U^{\theta}_{\mathcal{F}_i,\xi}U^{\theta T}_{\mathcal{F}_i,\xi}|\mathcal{O}_i] = \frac{\partial a_i}{\partial \xi}\frac{\partial a_i}{\partial \xi}^T \mathcal{L}^{\theta/\theta_0}_{\mathcal{O}_i} + \frac{\partial b_i}{\partial \xi}\frac{\partial b_i}{\partial \xi}^T I_4 - \frac{\partial a_i}{\partial \xi}\frac{\partial b_i}{\partial \xi}^T I_1 - \frac{\partial b_i}{\partial \xi}\frac{\partial a_i}{\partial \xi}^T I_1,$$

$$\mathrm{E}_{\theta_0}[\mathcal{L}^{\theta/\theta_0}_{\mathcal{F}_i}U^{\theta}_{\mathcal{F}_i,\omega}U^{\theta T}_{\mathcal{F}_i,\omega}|\mathcal{O}_i] = d_i^2 I_7 + b_i^2 I_5 - 2d_i b_i I_3,$$

$$\mathrm{E}_{\theta_0}[\mathcal{L}^{\theta/\theta_0}_{\mathcal{F}_i} U^{\theta}_{\mathcal{F}_i,\xi} U^{\theta T}_{\mathcal{F}_i,\omega} | \mathcal{O}_i] = \frac{\partial a_i}{\partial \xi} d_i I_0 - (\frac{\partial b_i}{\partial \xi} d_i + b_i \frac{\partial a_i}{\partial \xi}) I_2 + b_i \frac{\partial b_i}{\partial \xi} I_6.$$

Thus we have determined formulae for computing observed score and Hessian in the proportional hazards shared frailty models; these formulae depend on nine integrals which must in general be computed numerically. We call this approach semi-analytical.

5 Application to a normal frailty Weibull model

5.1 The normal frailty Weibull model; formulae

There are two reasons for considering normal frailties here. The first is for illustrative purpose: in contrast with the gamma-frailty model (Rondeau and Commenges, 2003), there is no analytical formula for the likelihood. The second is that many models have used normal random effects; the normality assumption is appealing when considering several random effects (although here we will restrict to one), see Ripatti and Palmgren (2000).

In the framework of the proportional hazard model with shared frailty, we will consider a Weibull model for the baseline, $\alpha_0^{\gamma}(t) = \gamma_0 \gamma_1 (t)^{\gamma_1 - 1}$, and normal frailties η_i. Here we have:

$$a_i = d_i \log(\gamma_0 \gamma_1) + (\gamma_1 - 1) \sum_{j=1}^{n_i} \delta_{ij} [\log(\tilde{T}_{ij})] + \sum_{j=1}^{n_i} \delta_{ij} z_{ij} \beta,$$

where $d_i = \sum_{j=1}^{n_i} \delta_{ij}$ and

$$b_i = \gamma_0 \sum_{j=1}^{n_i} e^{z_{ij}\beta} (\tilde{T}_{ij})^{\gamma_1}.$$

The first and second derivatives of these quantities relative to the parameters are given in Appendix. Using these formulae and the more general formulae of section 4 we obtain the semi-analytical score and Hessian for the normal Weibull model. The nine integrals for each group can be computed by gaussian quadrature.

5.2 Simulation

A simulation study was performed in order to compare a Marquardt algorithm using the semi-analytical score and Hessian with the same Marquardt algorithm using numerical derivatives. In order to investigate the effect of increased sample size we considered three values for the number of groups : $G = 50, 500, 5000$ with 2 subjects in each group. The random variables were generated as follows. We generated normal frailties: $\eta_i, i = 1, ..., G,$ i.i.d. $\sim N(0, 1)$. We took $\omega = 0.5$.

Given the η_i, the independent survival times $T_{ij}, j = 1, \ldots, 2$ were generated from a simple Weibull $(\alpha_0(t) = 0.04 \times 2 \times (0.04 \times t)^{2-1} = 0.04^2 \times 2 \times t^{2-1}$, i.e., $\gamma_0 = 0.04^2 = 0.0016, \gamma_1 = 2)$ with $\alpha_{ij}(t|\eta_i) = \alpha_0(t) \exp(\sum_{k=1}^{K} \beta_k X_{ijk} + \omega\eta_i)$, where X_{ijk} were binary covariates generated randomly as independent Bernoulli variables with $P(X_{ijk} = 1) = 0.5$. We generated right-censored data as follows. Right-censoring variable C_{ij} were generated from a uniform distribution on [1,71] producing about 30 % of censoring. The observed samples were $(Y_{i1}, ..., Y_{in_i})$, with $Y_{ij} = \min(T_{ij}, C_{ij})$ and $\delta_{ij} = I_{[T_{ij} \leq C_{ij}]}$. The chosen values of the β's are given in the tables. We performed two simulations, one with $K = 2$ another one with $K = 10$ to examine the effect of increasing the number of explanatory variables on computing time. We used

a Fortran program running with a Linux Redhat 9 system on a Intel Xeon 3.06 GHz processor.

In the two simulations the two algorithms gave identical results with at least five significant digits. The algorithm using numerical derivatives behaves surprisingly well in term of accuracy; much more complicated models would be necessary to enlighten the possible benefit of the new method in term of accuracy over the numerical derivatives. In the tables we give only the common values of the computed estimates. The results are consistent for the two simulations (Tables 1 and 2), with increasing precision as sample size increases. The most interesting result is in term of CPU time for large data sets. In the second simulation ($K = 10$) in the case $G = 5000$ the proposed method took 408 seconds (Table 2); it took 7398 seconds when using numerical derivative, a really different order of magnitude. Note that the increase of time for the proposed algorithm is not very important when going from two (239 seconds) to ten (408 seconds) explanatory variables,

5.3 Application to the catheter example

As an example we fitted the data presented by McGilchrist and Aisbett (1991) on infections in catheters for patients on dialysis. Each observation is the time to infection, at the point of insertion of the catheter for kidney patients using portable dialysis equipment. There are 38 patients each with exactly two observations. Variables retained for illustration are age (in years) and sex (female vs male). The variance of the frailties is a measure of the heterogeneity of the patients. Results are illustrated in Table 3. The program using semi-analytical derivatives took 2.53 seconds of CPU time while using

numerical derivatives it took 18.01 seconds. This illustrates once again the difference in speed of the two methods, although in this case this is not practically significant.

6 Conclusion

We have given general formulae linking full and observed scores and Hessians and have suggested that these formulae could be applied for obtaining efficient Newton-Raphson type algorithms. We have exhibited a particular domain of application which is the inference for parametric proportional hazards shared-frailty models: we have shown that nine numerical integrals had to be computed for each group whatever the number of parameters in the baseline hazard or representing effect of covariates, and we have illustrated this approach using a Weibull model with normal frailty. On a large data set with 5000 groups and ten explanatory variables the proposed algorithm was nearly 20 times faster than using numerical derivatives. We may hope that this approach can be used to treat complex models involving several frailties or hazard represented on a basis of splines such as in Rondeau and Commenges (2003).

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6**, 701-726.

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical*

Models Based on Counting Processes. New-York: Springer-Verlag.

Barndorff-Nielsen, O. E. and Sorensen, M. (1994). A review of some aspects of asymptotic likelihood theory for stochastic processes. *Int. Statist. Rev.*, **62**, 133-165.

Bickel, P; J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. New-York: Springer-Verlag.

Commenges, D. and Gégout-Petit, A. (2005a). Likelihood inference for incompletely observed stochastic processes: ignorability conditions. *arXiv:math.ST/0507151*.

Commenges, D. and Gégout-Petit, A. (2005b). Likelihood for generally coarsened observations from multi-state or counting process models. *Scandinavian Journal of Statistics*, in revision.

Commenges, D. and Joly, P. (2004). Multi-state model for dementia, institutionalization and death. *Communications in Statistics A* **33**, 1315-1326.

Dennis, J. E. and Schnabel R. B. (1983). *Numerical methods for unconstrained optimization and nonlinear equations*. Prentice-Hall (Englewood Cliffs, NJ).

Dempster, A.D., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B* **39**, 1.

Gill, R. D., van der Laan, M. J. and Robins, J.M. (1997). Coarsening at random: characterizations, conjectures and counter-examples, pp. 255-294 in: *State of the Art in Survival Analysis*, D.-Y. Lin and T.R. Fleming (eds),

Springer Lecture Notes in Statistics 123.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1995). *Markov Chain Monte Carlo in practice.* New-York: Chapman & Hall.

Hedeker D. and Gibbons R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50**, 933-944.

Jacqmin-Gadda H, Thiébaut R, Chêne G, Commenges D (2000). Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostatistics* **1**, 355-368.

Jacod, J. (1975). Multivariate point processes: predictable projection; Radon-Nikodym derivative, representation of martingales. *Z. Wahrsheinlichkeitsth.* **31**, 235-253.

Joly, P., Letenneur, L., Alioum, A., and Commenges, D. (1999). PHMPL: a computer program for hazard estimation using a penalized likelihood method with interval-censored and left-truncated data. *Computer Methods and Programs in Biomedicine* **60**, 225-231

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc. B* **44**, 226-233.

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal of Applied Mathematics* **11**, 431-441.

McGilchrist, C. A. and Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics* **49**, 221-225.

Nielsen, G.G., Gill, R. D.,Andersen, P. K., Sorensen, T. I. A. (1992). A

counting process approach to maximum likelihood estimation in frailty models. *Scandinanvian Journal of Statistics* **19**, 25-43.

Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm *Journal of the Royal Statistical Society: Series B:* **61**, 479-482.

Overton, M. (2001). *Numerical computing with IEEE floating point arithmetic.* Philadelphia: SIAM.

Ripatti, S. and Palmgren, J. (2000). Estimation of Multivariate Frailty Models Using Penalized Partial Likelihood *Biometrics* **56**, 1016,1022.

Rondeau, V. and Commenges, D. (2003). Penalized likelihood for frailty models. *Lifetime Data Analysis* **9**, 139-153.

Appendix: Derivatives of a_i and b_i for the Weibull model

$\frac{\partial a_i}{\partial \gamma_0} = \frac{d_i}{\gamma_0}$,
$\frac{\partial a_i}{\partial \gamma_1} = \frac{d_i}{\gamma_1} + \sum_{j=1}^{n_i} \delta_{ij} \log(\tilde{T}_{ij})$,
$\frac{\partial a_i}{\partial \beta} = \sum_{j=1}^{n_i} \delta_{ij} z_{ij}$,
$\frac{\partial b_i}{\partial \gamma_0} = \sum_{j=1}^{n_i} e^{z_{ij}\beta} (\tilde{T}_{ij})^{\gamma_1}$,
$\frac{\partial b_i}{\partial \gamma_1} = \gamma_0 \sum_{j=1}^{n_i} e^{z_{ij}\beta} \log(\tilde{T}_{ij}) (\tilde{T}_{ij})^{\gamma_1}$,
$\frac{\partial b_i}{\partial \beta} = \gamma_0 \sum_{j=1}^{n_i} z_{ij} e^{z_{ij}\beta} (\tilde{T}_{ij})^{\gamma_1}$,
$\frac{\partial^2 a_i}{\partial \gamma_0^2} = -\frac{d_i}{\gamma_0^2}$,
$\frac{\partial^2 a_i}{\partial \gamma_1^2} = -\frac{d_i}{\gamma_1^2}$,
$\frac{\partial^2 a_i}{\partial \beta \partial \beta'} = 0$,

$$\frac{\partial^2 a_i}{\partial \gamma_0 \partial \gamma_1} = 0,$$
$$\frac{\partial^2 a_i}{\partial \gamma_0 \partial \beta} = 0,$$
$$\frac{\partial^2 a_i}{\partial \gamma_1 \partial \beta} = 0,$$
$$\frac{\partial^2 b_i}{\partial \gamma_0^2} = 0,$$
$$\frac{\partial^2 b_i}{\partial \gamma_1^2} = \gamma_0 \sum_{j=1}^{n_i} e^{z_{ij}\beta} (\log(\tilde{T}_{ij}))^2 (\tilde{T}_{ij})^{\gamma_1},$$
$$\frac{\partial^2 b_i}{\partial \beta \partial \beta'} = \gamma_0 \sum_{j=1}^{n_i} z_{ij} z'_{ij} e^{z_{ij}\beta} (\tilde{T}_{ij})^{\gamma_1},$$
$$\frac{\partial^2 b_i}{\partial \gamma_0 \partial \gamma_1} = \sum_{j=1}^{n_i} e^{z_{ij}\beta} \log(\tilde{T}_{ij}) (\tilde{T}_{ij})^{\gamma_1},$$
$$\frac{\partial^2 b_i}{\partial \gamma_0 \partial \beta} = \sum_{j=1}^{n_i} z_{ij} e^{z_{ij}\beta} (\tilde{T}_{ij})^{\gamma_1},$$
$$\frac{\partial^2 b_i}{\partial \gamma_1 \partial \beta} = \gamma_0 \sum_{j=1}^{n_i} z_{ij} e^{z_{ij}\beta} \log(\tilde{T}_{ij}) (\tilde{T}_{ij})^{\gamma_1}.$$

Table 1: Maximum likelihood estimates together with the estimated standard errors (between parentheses) of the parameters in a Weibull normal-frailty model. Simulation 1: 2 explanatory variables; G: number of groups. CPU times are given for the Marquardt algorithm using numerical derivatives and using semi-analytical derivatives.

True parameter value	G = 50	G = 500	G = 5000
$\omega = 0.5$	0.733 (0.356)	0.569 (0.089)	0.507 (0.027)
$\beta_1 =$ -0.5	-0.556 (0.330)	-0.468 (0.088)	-0.486 0.027)
$\beta_2 = 1.5$ (0.368)	1.441 (0.368)	1.434 (0.099)	1.460 (0.032)
Weibull parameters			
$\gamma_0 = 0.0016$	0.0012 (0.0016)	0.0013 (0.00035)	0.0015 (0.00013)
$\gamma_1 = 2.0$	2.14 (0.308)	2.06 (0.082)	2.01 (0.025)
Number of iterations	14	9	9
CPU Time			
Numerical derivatives	117 sec	94 sec	1014 sec
Semi-analytical derivatives	22 sec	14 sec	239 sec

Table 2: Maximum likelihood estimates together with the estimated standard errors (between parentheses) of the parameters in a Weibull normal-frailty model. Simulation 2: 10 explanatory variables; G: number of groups. CPU times are given for the Marquardt algorithm using numerical derivatives and using semi-analytical derivatives.

True parameter values	G = 50	G = 500	G = 5000
$\omega = 0.5$	0.000008 (0.374)	-0.49 (0.086)	0.527 (0.024)
$\beta_1 =$ -0.5	-0.848(0.267)	-0.608(0.085)	-0.517(0.025)
$\beta_2 = 1.5$	0.929 (0.251)	1.357 (0.092)	1.526 (0.030)
$\beta_3 =$ -0.5	-0.368 (0.254)	-0.613 (0.083)	-0.528 (0.026)
$\beta_4 = 1.5$	1.389 (0.270)	1.503 (0.099)	1.475 (0.030)
$\beta_5 =$ -0.5	-0.026 (0.243)	-0.503 (0.080)	-0.502 (0.025)
$\beta_6 = 1.5$	1.344 (0.267)	1.622 (0.099)	1.503(0.030)
$\beta_7 =$ -0.5	-0.576 (0.232)	-0.590 (0.083)	-0.528 (0.026)
$\beta_8 = 1.5$	1.023 (0.269)	1.409 (0.097)	1.462 (0.030)
$\beta_9 =$ -0.5	-0.046 (0.241)	-0.461 (0.083)	-0.506 (0.025)
$\beta_{10} = 1.5$	1.534 (0.287)	1.528 (0.098)	1.483 (0.029)
Weibull parameters			
$\gamma_0 = 0.0016$	0.0012 (0.0019)	0.0022 (0.0005)	0.0017 (0.00015)
$\gamma_1 = 2.0$	1.831 (0.159)	1.916(0.074)	1.998 (0.023)
Number of iterations	9	9	11
CPU Time			
Numerical derivatives	303 sec	660 sec	7398 sec
Semi-analitycal derivatives	13 sec	14 sec	408 sec

Table 3: Analysis of the kidney catheters data set: maximum likelihood estimates for the Weibull normal-frailty model; CPU times are given for the numerical derivatives and semi-analytical approaches.

	Estimate (Standard deviation)
ω	0.770 (0.243)
Age : β_0	0.0596 (0.126)
Sex : β_1	1.63 (0.494)
Weibull parameters	
γ_0	0.00194 (0.00202)
γ_1	1.18 (0.159)
Number of iterations	12
CPU Time	
Numerical derivatives	18.01 sec
Semi-analytical derivatives	2.53 sec