11-1-2004

A Modification Of The EM Algorithm To Estimate An Andersen-Gill Gamma Frailty Model For Multivariate Failure Time Data

Maria Antònia Barceló
University of Girona, Spain

Marc Saez
Universitat de Girona, Spain, marc.saez@udg.es

A Modification Of The EM Algorithm To Estimate An Andersen-Gill Gamma Frailty Model For Multivariate Failure Time Data

Maria Antònia Barceló   Marc Saez
Research Group on Statistics, Applied Economics and Health (GRECS)
University of Girona, Spain

A modification of the Andersen-Gill gamma shared frailty model is presented. The variance of the frailty is directly modeled by means of a generalized linear model, the EM algorithm is modified in order to simultaneously estimate a semiparametric model for the failure times and a model for the variance of the frailty. A simulation study is conducted to evaluate the performance of the proposed algorithm (EMB algorithm) and compared with other methods, a marginal model, and a conditional model. Multivariate data from a nosocomial infection study is used to illustrate the methods. The EMB fit turned out to be better than the fit obtained from a marginal model or from a conditional model. The EMB provided the best fit (being the least over-dispersed and having the highest AIC and the highest pseudo-R square) and estimated the parameters most efficiently. The proposed method is able to capture and to take into account unobservable random effects in semiparametric models.

Key words: Frailty, marginal and conditional models, generalized linear models, EM, nosocomial
            infections

Introduction

Patients admitted to intensive care units (ICU) run a high risk of contracting a nosocomial infection due not only to the susceptibility associated with the severity of their conditions, but also to medical procedures that the ICUs use.

Barceló and Saez (2001) analysed the factors that determine the occurrence of nosocomial infections in the ICU of a tertiary-level hospital in Girona, Spain, during the second quarter of 1999 (March-June, 1999). The authors tried to determine which factors, those associated to patients (such as their immunodeficiency) or those related to ICU (such as invasive medical procedures or the inappropriate use of antimicrobial agents), were the most relevant in the explanation of the occurrence of nosocomial infections in the ICU. They were interested in analysing the factors that determine both the occurrence of an infection and also the time leading up to the onset of the infection.

In that context, the standard approach to obtain adjusted risk (hazard) factors for the infection would be the Cox model (Cox, 1972). The problem was that a patient could have several episodes of infection during her/his ICU stay. As a consequence the data set had multiple events per subject, i.e. recurrent events. As is well known the main problem of the Cox model with multivariate data is that the observations are not independent, implying, among other things, the violation of the proportionality hypothesis. It is also known that standard Cox

models ignore such dependence, leading to estimates that are inefficient and biased.

The Andersen-Gill approximation (AG) to the Cox model (Andersen & Gill, 1982; Andersen et al., 1993) overcomes, in part, this problem. The AG model is a counting process approach in which each patient is represented as a set of rows with time intervals of (entry time, first infection], (first infection, second infection], ... , (nth infection, last follow-up]. Each row is treated as a different patient and, therefore, risk proportionality is not violated. However, the underlying hypothesis in AG, called the hypothesis of independent increments, is very restrictive and may be untenable. Under this hypothesis the multiple observations of an individual are independent, although conditioned on the explanatory variables. Therefore, a suitable alternative is needed.

In choosing a model for the time to recurrent infection one needs to consider the biological process of disease. It was very likely that after experiencing the first infection, the risk (hazard) of subsequent infections would increase. This could happen if each infection permanently compromised the ability of the immune system to combat subsequent infection. If this were the case one would use a model containing separate strata for each episode of infection (Therneau & Hamilton, 1997). In this sense, the first choice was the Prentice, Williams and Peterson (PWP) model (Prentice, Williams & Peterson, 1981). The PWP is a marginal model with respect to the estimation of the parameters, which treats the dependence between event times as a nuisance to control for, without explicitly specifying models for this dependence.

Conditional methods, in contrast, explicitly model the dependence between recurrences. Amongst them, frailty models (Clayton & Cuzick, 1985) have become the most popular for analysing multivariate survival data. In those models the dependence between the events is accounted for by the introduction of frailties or unobservable random effects into the marginal hazards (Klein, 1992). The frailties are shared among recurrences from the same individual. Maximum likelihood estimation in the AG shared frailty model (with gamma-distributed frailties) is usually performed using

the EM algorithm as suggested by Gill (1985) and further discussed by Nielsen et al. (1992) and Klein (1992).

In particular, the estimation of the model using the EM algorithm is carried out by fixing through the variance of the frailty until its convergence. Then, the algorithm iterates between the E and the M steps. In the E step the frailties are replaced in the complete data log likelihood by their conditional expectation. The M step consists of computing the Nelson-Aalen estimator as if the frailties had been observed. This procedure is repeated for other arbitrary values of the variance obtaining in each case the log incomplete data profile likelihood as a function of the variance. Finally, the estimate of the variance is computed either numerically or graphically. The EM algorithm, however, could converge arbitrarily and slowly and, furthermore, the final estimate of the variance obviously depends on the initial choices for that parameter.

Here, the directly modelling of the variance is proposed. In this sense, Clayton (1988) and Lindsey (1999) were followed. The former proposes the possibility of extending the EM algorithm by simultaneously estimating the variance of the frailty. Lindsey (1999) pointed out that "dispersion varying with the explanatory variables is surprisingly common" (Lindsey, 1999, p. 2230) and suggests estimating a separate regression equation for the dispersion parameter. Besides the extension of the model, a modification of the EM algorithm is also proposed, which is called EMB, to simultaneously estimate such a two-equation model.

An alternative to the frailty models can be found in the penalised likelihood models (Behrman et al.,1991; Therneau & Grambsch, 1998). The idea is to use a penalty function for a constrained solution, equal to the log gamma density. The penalty function captures the local variability underlying the joint density of data. The problem is that such variability is in fact approaching two different things, frailty (false contagion) and serial correlation or dependence (true contagion). In addition, the choice of the shrinkage parameter used in the penalisation is a controversial question in survival analysis.

Methodology

Suppose that there is a random sample of I individuals from an underlying group population and that each individual can have J observations. For this framework the most straightforward mathematical notation derives from the theory of counting processes (Fleming & Harrington, 1991; Andersen et al., 1993).

Let i (I = 1,…,I) denote individual and (i,j) denote the jth observation in the ith individual. For each observation (i,j), where i = 1,…, I and j = 1,…, J, let $N_{ij}(t)$ be an observed multivariate counting process. $N_{ij}(t)$ is the cumulative number of events observed for the i-th subject. A process $Y_{ij}(t)$ is further observed, indicating whether individual i is observed to be at risk for experiencing an jth event at time t-. Finally covariates $X_{ij}(t)$ (possibly time-dependent) are observed (Andersen, 1992). The multivariate counting process $N_{ij}(t)$ has an intensity process given by,

$$\lambda_{ij}(t) = Y_{ij}(t)\alpha_0(t)e^{\beta'X_{ij}(t)} = \lambda_0(t)e^{\beta'X_{ij}(t)} \quad [1]$$

where $\lambda_0(t)$ denotes an unknown baseline intensity; and β is a vector of unknown parameters.

The Prentice, Williams and Peterson (PWP) model

Although the PWP is a marginal model, it is conditional in relation to the construction of the risk set. In this sense, the model allows the baseline risk to vary between recurrences, i.e. $\lambda_{oj}, j = 1,...,J$,

$$\lambda_{ij}(t) = \lambda_{0j}(t)e^{\beta'X_{ij}(t)} \quad (2)$$

Thus, it is actually an AG model with time dependent strata. That is to say, the risk set for the recurrence j, for instance, only contains those individuals who experienced j-1 recurrences. Such a strategy makes it possible to control dependence between the recurrences, stratifying according to them.

Although estimates obtained from PWP models are consistent (Prentice, Williams & Peterson, 1981), the dependence between observations remains in fact uncontrolled. As a consequence, standard errors are biased (usually overstated). For this reason we propose to robustly estimate the standard errors of the parameters. In particular we chose a grouping jackknife estimate (Therneau & Hamilton, 1997). The idea is to compute the i change in the estimates of the parameters with all the observations for the i-th subject removed from data set. This will result in a matrix D, each row i of which will be an estimate of the leverage, i.e., average change, of the i-th subject. The matrix D'D will approximate the grouped jackknife estimate of variance and will be an asymptotically unbiased estimate of the variance of the robust parameter estimates (Therneau & Hamilton, 1997; Lin & Wei, 1989).

The AG Gamma frailty model.

Following Nielsen et al. (1992), we formulate now an intensity process λ satisfying,

$$\lambda_{ij}(t\,|\,\varpi_i) = \varpi_i Y_{ij}(t)\alpha_{0j}(t)e^{\beta'X_{ij}(t)} \quad (3)$$

where ϖ_i denotes subject specific frailties independently drawn from a gamma(ν,η) distribution. Note that here we also allow the baseline hazards to vary between recurrences, i.e. α_{oj}.

The gamma density of the frailties is

$$f(\varpi;\nu,\eta) = \frac{\varpi^{\nu-1}e^{-\eta\varpi}\eta^{\nu}}{\Gamma(\nu)}$$

where η is the scale parameter and ν the shape parameter.

Due to identification problems it is usually assumed that $\nu = \eta$, i.e. that the distribution of the frailties has unitary mean and variance equal to $1/\eta$, *say* θ (Clayton, 1978;

Vaupel et al., 1979; Nielsen et al., 1992; Klein, 1992).

Let Z be a nxq design matrix that describes how the frailties applied to individuals subjects, $\varpi_i = e^{Z_i \omega}$. It is also assumed that the frailty consists of independent clusters of observations, i.e. $Z_{ij} = 1$ iff recurrence j belongs to individual i. Let us define

$$D_i = \sum_{j=1}^{J} Z_{ij} \delta_j$$

$$E_i^* = \sum_{j=1}^{J} Z_{ij} \Lambda_0(t_j) e^{\beta'X_j}$$

where δ_i is an indicator equal to one in a failure time case and zero otherwise; Λ_0 is the cumulative baseline hazard.

D$_i$ is the number of events in the i-th individual and $E_i = E_i^* e^{\omega_i}$ is the expected number of events in the individual based on the covariates and the model.

The full log likelihood, when ω is observed, is then,

$$L_f = \sum_{i=1}^{I} \left[\left(\frac{1}{\theta} \omega_i - \frac{1}{\theta} e^{\omega_i} \right) + \frac{1}{\theta} \log\left(\frac{1}{\theta}\right) - \log \Gamma\left(\frac{1}{\theta}\right) \right]$$

$$+ \sum_{i=1}^{I} \left[D_i \omega_i - E_i^* e^{\omega_i} \right] + \sum_{j=1}^{J} \delta_j \left[\log(\lambda_{0j}(t_j)) + X_j \beta \right]$$

$$(4)$$

It can be shown (Therneau & Grambsch, 1998) that as a function of any single ω_i, [4] is proportional to,

$$\left(\frac{1}{\theta} + D_i \right) \omega_i - \left(\frac{1}{\theta} + E_i^* \right) e^{\omega_i}$$

therefore, conditional on the data, the ϖ_i are distributed as gamma variates with shape $\frac{1}{\theta} + D_i$ and scale $\frac{1}{\theta} + E_i^*$. In this sense, we can write,

$$E\left(e^{\omega_i} \middle| \tilde{\beta}, \frac{1}{\tilde{\theta}}, y \right) = \frac{D_i + \frac{1}{\tilde{\theta}}}{E_i^* + \frac{1}{\tilde{\theta}}} \equiv e^{\hat{\omega}_i} \qquad (5)$$

where y denotes the observed data and the tilde denotes either provisional or definitive estimates.

The maximisation of the log likelihood (4) can be done using the EM algorithm. Therneau and Grambsch (1998) suggested how to use only the quantities returned by an ordinary Cox model program. Starting with the case of a fixed variance, the quantities D$_i$ and M$_i$ = D$_i$-E$_i$ can be obtained by summing over the input data and the returned martingale residuals, respectively. E_i^* is obtained from E$_i$ and the current estimates of $\hat{\omega}_i$ (E-step). The next estimates of $\hat{\omega}_i$ are obtained from equation (5) and, finally, $Z_{ij} \hat{\omega}_i$ is used as a prior in the next invocation of the Cox model (M-step).

One problem with the EM algorithm is that variance estimates for the estimated parameters are not immediately provided (Louis, 1982). It was suggested by Gill (1989) and further discussed by Nielsen et al. (1992) and, above all, Andersen et al. (1997), that a non parametric information calculation was likely to provide consistent variance estimators. A simpler possibility lies in using the robust estimate of the covariance matrix of the estimated parameters described above (Therneau & Hamilton, 1997).

The Penalised Cox model

As mentioned above, Behrman et al. (1991) proposed to alternatively use the penalised log likelihood,

$$\log L_\alpha(f) = \sum_i \log(f(x_i)) - \alpha R(f)$$

where αR(f) is a penalty term that takes account of the roughness or local variability in the joint

density of the data. The smoothing parameter α, which controls the balance between smoothness and goodness of fit, must be typically chosen by cross-validation.

Therneau and Grambsch (1998) suggested using the log gamma density as the penalty function for the constrained solution,

$$-\alpha\, R(f) = \frac{1}{\theta}(\omega - e^{\omega}) + \frac{1}{\theta}\log\left(\frac{1}{\theta}\right) - \log\Gamma\left(\frac{1}{\theta}\right)$$

Therefore, in our case,

$$PLL = Cox\, PL + \sum_{i=1}^{I}\frac{1}{\theta}(\omega_i - e^{\omega_i}) \qquad (6)$$

$$+\frac{1}{\theta}\log\left(\frac{1}{\theta}\right) - \log\Gamma\left(\frac{1}{\theta}\right)$$

where PLL denotes the penalised log likelihood and by Cox PL we mean the numerical value returned as the partial likelihood by a standard Cox model program for the given values of β and ω, ω having been entered as an offset term.

Therneau and Grambsch (1998) pointed out that for any fixed value of the variance of the frailty the EM algorithm and the constrained minimisation of the penalised likelihood have the same solution.

A modification of the EM algorithm for the estimation of the AG gamma frailty model: The EMB algorithm.

In both the AG gamma frailty and the penalised Cox models, frailty is assumed to be constant between individuals and also within each individual, i.e. between recurrences. As an alternative, and following Wassell and Moeschberger (1993), we propose to directly model the variance of the frailty,

$$\theta_{ij} = e^{c + \gamma' X_{ij}} \qquad (7)$$

where γ is a vector of parameters, X_{ij} denotes the covariates for individual i .

To model the variance we propose to use a generalised linear model, GLM (Nelder &

Pregibon, 1987; Nelder & Lee, 1991, 1996; Nelder, 1998). In particular,

Link function : $\log\left(\theta_{ij}\right) = c + \gamma' X_{ij}$

$$(8)$$

Variance function : $Var\left(\theta_{ij}\right) = 2\theta_{ij}^2$

The link function is simply a transformation of (7). For the variance function we have chosen the deviance transformation. This is because it is close to the optimal normalising transform for the GLM distributions irrespective of the distribution chosen for the link (Pierce & Schafer, 1986).

Note that we allow the frailty to differ between different individuals. It is also possible that the frailty may vary through the recurrences. In this sense we have introduced flexibility into the gamma frailty model.

In the estimation of the model we propose a modification of the EM algorithm, which we called EMB. In particular a new step (step 1) is introduced in the algorithm,

0.- From the provisional value of $\tilde{\theta} = 1$, estimate a standard AG model and compute $\tilde{\theta}_i$.

1.- Estimate a model for the variance and obtain the fitted values of $\tilde{\theta}_i$.

2.- Use the values of the variance computed in step 1 to fit the AG gamma frailty model using the standard EM algorithm.

3.- Compute $\tilde{\theta}_i$ and return to step 1.

The EMB algorithm is iterated until convergence. The complete EMB algorithm is shown in the appendix.

An obvious starting value for $\tilde{\theta}$ would be 0, i.e. no frailty, the problem is that in this case the frailty remains fixed at zero in the update formula. For this reason, we have preferred here $\tilde{\theta} = 1$.

Results

A simulation study was conducted to evaluate the performance of the proposed EMB algorithm and to compare it with other methods, the PWP model and the penalized Cox model in particular.

Multivariate failure times were generated from an AG gamma frailty model with the following hazard function, where i denoted individuals and j denoted repeated measures, i.e. recurrences, within the same individual. In particular, we considered I = 100 individuals and J = 2 recurrences.

$$\lambda_{ij}(t) = \lambda_{0j}(t)\, e^{X_{ij}\beta + Z_{ij}\omega}$$

We simulated two Weibull baseline hazards,

1) $\lambda_{0j}(t) = 0.5\,(e^{0.1X1_i + 0.3X2_i}{}_t)^{-0.5}$

2) $\lambda_{0j}(t) = 0.5\,(e^{0.1X1_i + 0.3X2_i + 0.2X3_{ij}}{}_t)^{-0.5}$

Two fixed variables (X_1, X_2) and one time-varying explanatory variable (X_3) were first simulated, although they were maintained fixed throughout the simulation. In particular, X1~binomial (1, p = 0.6), X2~normal (32.639, 12.967) and X3~binomial (4, p = 0.25).

Let $\varpi = \exp(\omega)$ follow a gamma distribution with parameters ν and η. Without loss of generality, we will assume that $\nu = \eta$, i.e. that the subject-specific ϖ_i has mean one and variance $(1/\nu) = \theta_i$, where θ dictates the heterogeneity across individuals. Summing up, $\varpi_i \sim \Gamma\!\left(\dfrac{1}{\theta_i}, \dfrac{1}{\theta_i}\right)$. Furthermore, we will assume that the random effect consists of independent clusters of observations, i.e. $Z_{ij} = 1$ if recurrence j belongs to individual i, and zero otherwise.

From (8), we simulated two cases: Case A. Var (ϖ_i) fixed over time ($\theta_{ij} = 1$); Case B. Var (ϖ_i) time-dependent ($\theta_{ij} = 0.5 + 0.1\,ID + 0.25X3_{ij}$), where ID=1,2,...,100. Finally, we compute $t_{ij} = \lambda_{ij}/\varpi_i$.

Simulated failure times for units (i,j) were independently censored by three uniform variables across all datasets to achieve overall censoring levels of 95%, 80% and 40%. Summing up, 500 datasets were simulated for twelve possible designs, 1A, 2A, 1B and 2B with 95%, 80% and 40% censoring for each one.

Three methods were used to fit the simulated data sets, PWP, penalised Cox (PC) and our proposed modification (EMB). In all cases baseline hazards were allow to vary between recurrences. Only in the penalised Cox case the variances of the parameters were not robustly estimated. Furthermore, in this latter case, the design matrix for the frailty was set equal to a diagonal matrix, each element of the diagonal corresponding to a different individual.

In Table 1, we show the results of the simulation. It was expected that failure to model existing frailty would result in biased estimates of parameters and reduced efficiency (Wassell & Moeschberger, 1993). In fact, it is possible to see a different pattern for the estimates of the parameters and for the estimates of the standard errors. With respect to the parameter estimation, PC and EMB fits were more similar to one another than to the PWP fit. Lower levels of censoring provided the most similar results for PC and EMB.

Note also that these two methods were more similar in the case of non-constant variance of frailty than in the constant. With respect to the estimates of the standard errors EMB seemed to provide the most efficient estimates. In fact, the PC fits were always more inefficient than the rest, even in relation to the model that did not explicitly model the frailty, i.e. PWP. Again, EMB was more efficient in lower censoring and in non-constant variance of frailty cases. In addition, although the S-plus macro for the estimation of the PC gave an estimation of what it called the variance of the frailty (Therneau & Grambsch, 1998), we are not sure that it was in fact such variance. In this sense, note the discrepancies with the EMB results of the estimation of such variance.

Table 1. Results of the simulation. 500 data-sets of four possible designs (constant frailty and fixed covariates 1A; constant frailty and time-varying covariates 1B; non-constant frailties and fixed covariates 2A; non-constant frailties and time-varying covariates 2B). 100 individuals and two recurrences. Three levels of censoring (60%, 80% and 95%).

	n=100, censoring = 60%					
	PWP1A	PWP1B	PC1A	PC1B	EMB1A	EMB1B
Beta1	0,00984	0.06272	0.09305	0.16261	0.15457	0.17256
Beta2	0,01491	0.03160	0.01490	0.03278	0.01081	0.02789
s.e. Beta1	0,18456	0.19418	0.19418	0.19832	0.18281	0.18649
s.e Beta2	0,00672	0.00691	0.00726	0.00745	0.00652	0.00668
Var frailty			0.00626	0.04133	0.92705	0.91868
	PWP1A	PWP1B	PC1A	PC1B	EMB1A	EMB1B
Beta1	0.00130	0.07330	0.00865	0.17326	0.06499	0.13064
Beta2	0.01378	0.03408	0.01367	0.03474	0.04596	0.03441
Beta3	0.01799	0.02471	0.01783	0.02530	0.06693	0.06126
s.e. Beta1	0.18670	0.19338	0.01946	0.19689	0.19305	0.18157
s.e. Beta2	0.00671	0.00691	0.00726	0.00739	0.00678	0.00663
s.e. Beta3	0.01063	0.01050	0.01073	0.01081	0.01060	0.01007
Var Frailty			0.00455	0.02825	0.95122	0.91837
	n=100. censoring = 80%					
	PWP1A	PWP1B	PC1A	PC1B	EMB1A	EMB1B
Beta1	0.08614	0.14580	0.11143	0.20428	0.10390	0.16684
Beta2	0.00553	0.02439	0.00597	0.03056	0.00400	0.05279
s.e. Beta1	0.34517	0.35449	0.36676	0.39950	0.30799	0.29621
s.e Beta2	0.01443	0.01435	0.01327	0.01415	0.01262	0.01217
Var frailty			0.29096	0.69528	0.90931	0.91146
	PWP1A	PWP1B	PC1A	PC1B	EMB1A	EMB1B
Beta1	0.07811	0.15113	0.10246	0.21342	0.09290	0.17243
Beta2	0.00682	0.02737	0.00597	0.03428	0.00247	0.02815
Beta3	-0.12416	-0.10825	-0.11837	-0.10280	-0.11970	-0.10655
s.e. Beta1	0.34398	0.35390	0.36665	0.39631	0.30853	0.31750
s.e. Beta2	0.01476	0.01479	0.01338	0.01414	0.01293	0.01273
s.e. Beta3	0.01883	0.01854	0.02037	0.02143	0.01711	0.01691
Var Frailty			0.29377	0.67242	0.90686	0.87601
	n=100. censoring = 95%					
	PWP1A	PWP1B	PC1A	PC1B	EMB1A	EMB1B
Beta1	-0.48964	-0.37113	-0.49018	-0.37202	-0.50812	-0.62366
Beta2	0.02062	0.04942	0.02063	0.04946	0.01756	0.11219
s.e. Beta1	0.55103	0.54926	0.58883	0.57913	0.53947	0.53414
s.e Beta2	0.01934	0.01936	0.02241	0.02228	0.01897	0.02875
Var frailty			0.00449	0.00495	0.93732	0.93564
	PWP1A	PWP1B	PC1A	PC1B	EMB1A	EMB1B
Beta1	-0.51248	-0.34955	-0.53836	-0.36148	-0.52390	-0.35898
Beta2	0.02647	0.05148	0.02510	0.05211	0.02469	0.06804
Beta3	0.50120	0.48832	0.52027	0.49929	0.50609	0.10052
s.e. Beta1	0.57865	0.56923	0.61556	0.59845	0.56670	0.53548
s.e. Beta2	0.01964	0.01918	0.02277	0.02226	0.01936	0.01810
s.e. Beta3	0.02754	0.02750	0.03108	0.03038	0.02667	0.02822
Var Frailty			0.16467	0.10965	0.92469	0.94274

PWP denotes Prentice, Williams and Peterson model; PC denotes penalised Cox model and EMB the AG gamma frailty model fitted using the EMB algorithm. The variances of the parameters in the PWP and the EMB were robustly estimated.

Application to nosocomial infection in an intensive care unit study

As mentioned above, in Barceló and Saez (2001) we tried to determine which factors, those associated to patients (intrinsic risk factors for the nosocomial infection, NI) or those related to the intensive care unit, ICU (extrinsic risk factors), were more relevant in the explanation of the occurrence of nosocomial infections in a tertiary-level hospital in Girona, Spain, during the second quarter of 1999.

The dependent variable (episode of infection hereinafter) consisted of either the time from the admission to the ICU to the onset of the infection (originated by a micro-organism, bacteria or fungi) or the time between the onset and the end of the infection. It was possible for patients to be infected more than once during their ICU stay. In the definition of our dependent variable, only ICU nosocomial infections were considered. Community-acquired infections, infections from other hospitals and infections from other hospital areas were not included under this definition. The beginning of the study did not always coincide with the patients' admission to the ICU, but with their admission in the hospital. Therefore, delayed entry was allowed. We also considered the possibility of (right) censoring because some patients would not get infected during their ICU stay.

Following the medical literature possible risk factors considered were classified as either intrinsic or extrinsic risk factors for NI. The former contained those directly related to the patient, such as gender and age, as well as those originating outside the ICU, such as previous infections (either community-acquired, or from another hospital or from another hospital area), severity of disease at admission and urgent surgery. Extrinsic risk factors for NI considered were location, mechanical ventilation, catheterism (central venous intravascular and arterial), tracheotomy, probes (urinary and nasogastric) and antibiotic treatment (antimicrobial used, duration and dosage). The effect of extrinsic risk factors was evaluated using the days of exposure to such a risk factor in a particular patient. The exposure was limited to the days prior to the onset of infections. Further details on data, variables and

additional results can be found in Barceló and Saez (2001).

Results of the fit of the model by PWP, PC and EMB are shown in Table 2. PWP was used here for comparative purposes. In this regard, note in Figure 1 that deviance residuals were not symmetrically distributed around zero and, above all, did not present a constant dispersion. Furthermore, in Figure 2, we show the estimates of the variance of the frailty (computed as shown in [A1]). The variance was not fixed between or within individuals. Some explanatory variables could explain such variability. See for instance in Figure 3 the relationship between the variance and one intrinsic factor (community-acquired previous infection) and between the variance and an extrinsic factor (mechanical ventilation).

In the implementation of the S-plus macro for fitting the PC, we used a design matrix consisting of independent clusters of observations, i.e., $Z_{ij} = 1$ if recurrence j belongs to individual i, and zero otherwise. With respect to EMB we needed to specify a model for the variance of the frailty (see equation (8)). We tried a forward stepwise strategy. We started with a single explanatory variable and included another one only if the AIC diminished. When we had a preliminary specification, and in order to check its robustness, a backward strategy with all the variables included was also tried. The final model for the variance is shown in Table 2.

From Table 2 we can see that the best fit was obtained from the EMB. In this sense, compare the EMB and the PC fits. The overdispersion (47.9% in EMB and 71.6% in PC); the AIC (162.41 and 180.42, respectively) and the pseudo R-square (Nagelkirke, 1991) (0.529 and 0.442, respectively) were lower in the case of EMB. Furthermore, the estimates obtained using our proposed modification were the most efficient. Note, in addition, that the estimate of the variance of the frailty provided by the PC S-plus macro was close to zero, meaning that there was no frailty in the model. This result contradicted the variability shown in Figures 1 to 3. Note also the varying behaviour of the variance of the frailties in Figure 4 (derived from the EMB fit).

Figure 1. Plot of the deviance residuals of the PWP model.
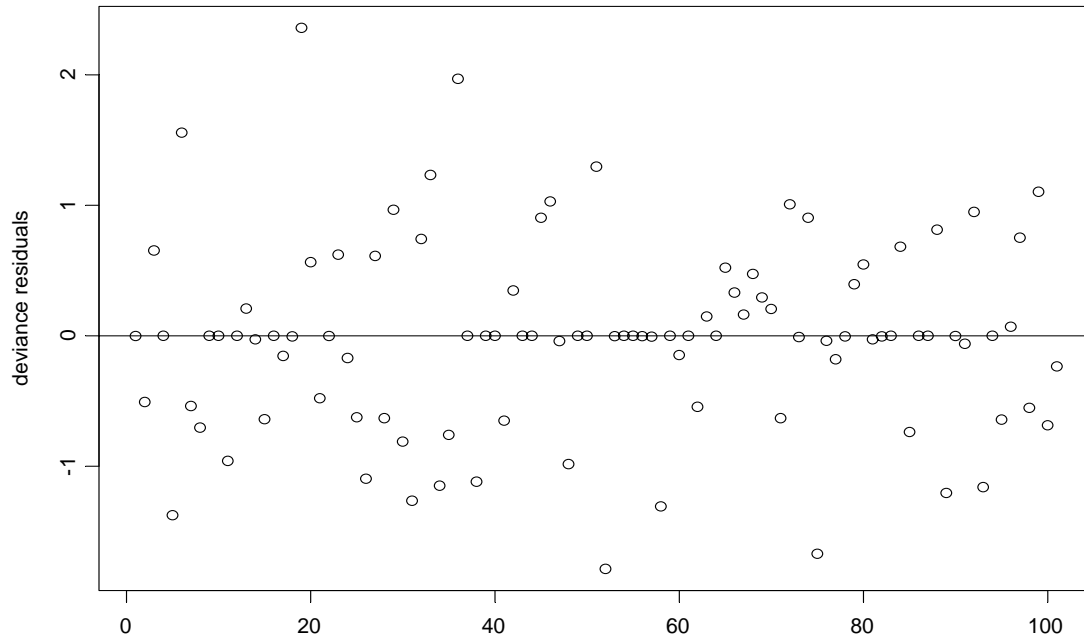

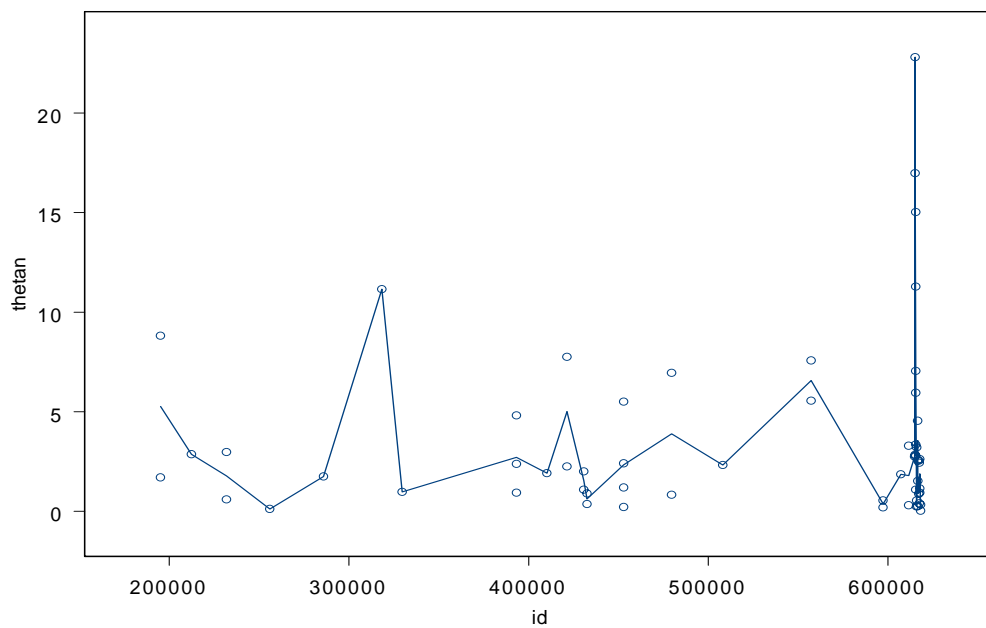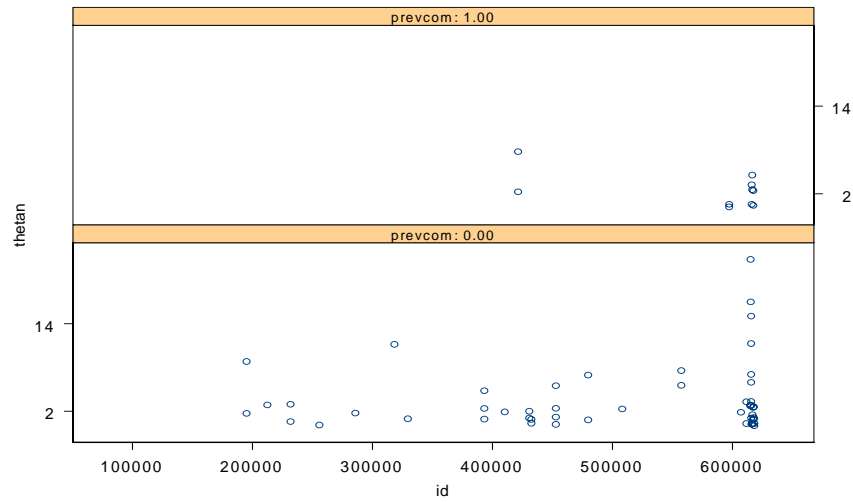Figure 2. Plot of the estimate of the variance of the frailty. PWP model.

Figure 3. Plot of the estimate of the variance of the frailty vs. an intrinsic risk factor (community-acquired previous infection) and an extrinsic risk factor (mechanical ventilation). PWP model.

Community-acquired previous infection
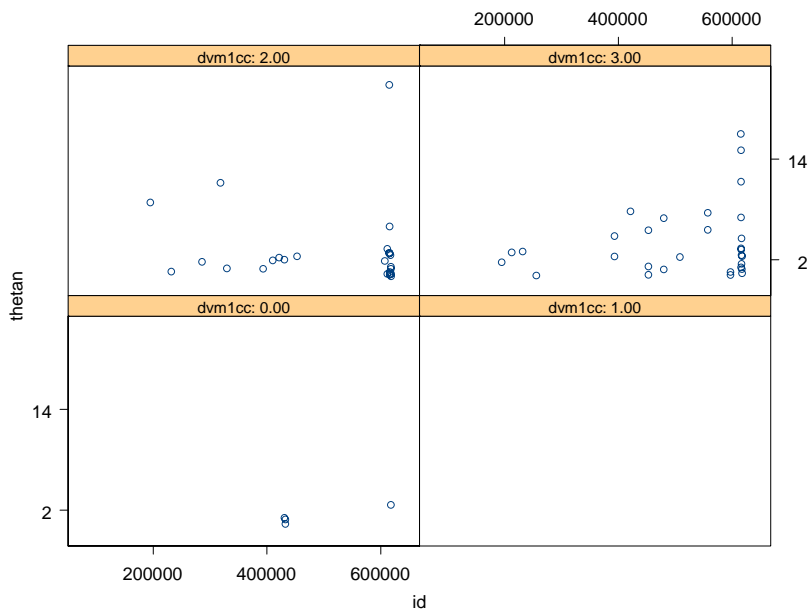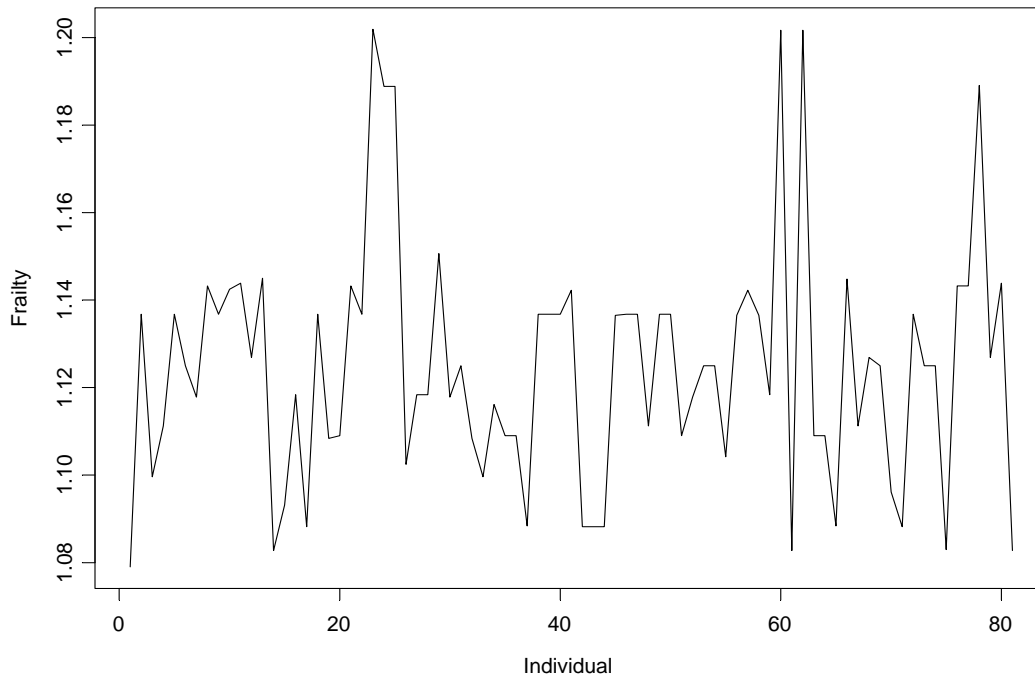

Mechanical ventilation

Table 2: Results of the estimation of the Prentice, Williams and Peterson, PWP; Penalised Cox and EMB models. PWP and EMB with robust estimation of the variance.

	PWP			Penalised Cox			EMB		
	Hazard Rate	lower .95	upper .95	Hazard Rate	lower .95	upper .95	Hazard Rate	lower .95	upper .95
Intrinsic risk factors									
Gender (male)	0.631	0.222	1.788	0.631	0.133	2.999	0.578	0.221	1.512
Age	1.009*	0.985	1.033	1.009*	0.981	1.037	1.009*	0.990	1.028
Previous infections (non)									
Community-acquired	4.56**	1.435	14.477	4.558*	0.954	21.772	6.923**	2.578	18.594
Other infections	0.56	0.201	1.557	0.560	0.118	2.648	0.418*	0.162	1.077
CDC (stable)									
Unstable intens. Care	54.468*	0.786	3776.406	8025.27**	1740	36968.77	49.012**	1.111	2162.135
Unstable shock	219.967**	3.187	15182.609	NA	NA	NA	257.040**	5.997	11017.03
Urg. surg. (non)	0.668	0.258	1.731	0.668	0.221	2.021	0.579	0.268	1.251
Extrinsic risk factors									
Location (rest of beds)									
Bed 4	2.255	0.545	9.340	2.255	0.346	14.707	2.247	0.621	8.135
Bed 5, 10, 11	0.946	0.381	2.348	0.946	0.257	3.478	0.784	0.369	1.667
Mechanical Vent. (non)									
≤ 3 days	0.011*	0.000	1.175	NA	NA	NA	0.006**	0.000	0.314
4-10 days	61.929**	1.325	2894.463	64.406**	1.35	3065.471	59.981**	2.105	1709.126
>10 days	129.534**	4.005	4189.427	134.714**	3.06	5934.619	98.713**	4.260	2287.525
Venous catheter (≤ 3d)	1.065	1.007	1.775	NA	NA	NA	1.082	1.011	1.822
Arterial catheter (non)									
> 0 days	1.833	0.579	5.805	1.833	0.495	6.782	1.629	0.638	4.162
Tracheotomy (non)									
≤ 6 days	0.231	0.028	1.924	0.231	0.030	1.768	0.243*	0.044	1.355
> 6 days	0.351	0.072	1.705	0.351	0.057	2.171	0.303*	0.075	1.224
Urinary probe (non)									
≤ 4 days	2.453	0.018	334.750	2.453	0.024	251.814	3.922	0.066	231.811
5-12 days	6.355	0.306	132.032	6.355	0.137	294.936	7.942*	0.630	100.190
> 12 days	1.566	0.079	31.014	1.566	0.035	70.518	1.998	0.159	25.104
Nasogastric probe (non)									
≤ 9 days	1.066	0.310	3.660	1.066	0.285	3.990	1.063	0.377	3.000
> 9 days	0.133**	0.023	0.748	0.133**	0.027	0.649	0.129**	0.030	0.550
Antibiotic treat. (non)									
≤ 7 days	0.415	0.112	1.539	0.415	0.093	1.858	0.329**	0.122	0.883
> 7 days	0.038**	0.004	0.385	0.038**	0.005	0.327	0.028**	0.004	0.216
Antibiotic dose-DDD	1.035	0.948	1.129	1.035	0.972	1.101	1.037	0.982	1.096
Deviance (degrees freedom)	130.427 (76)			130.422 (76)			112.4127 (76)		
Overdispersion	1.716144737			1.716078947			1.479114474		
AIC	180.427			180.422			162.4127		
Pseudo-R²	0.441886			0.441889			0.529338327		
Var frailty				5e-007			1.124881		
Model for the variance							β	s.e.(β)	
Mechanical Vent. (non)									
≤ 3 days							0.051	0.03602	
4-10 days							0.056	0.03617	
>10 days							0.086	0.03210	
Tracheotomy (non)									
≤ 6 days							0.008	0.026	
> 6 days							-0.006	0.023	
Urinary probe (non)									
≤ 4 days							0.038	0.045	
5-12 days							-0.011	0.042	
> 12 days							-0.019	0.046	
Nasogastric probe (non)									
≤ 9 days							-0.0002	0.021	
> 9 days							-0.0059	0.021	

Figure 4. Plot of the estimate of the variance of the frailty. EMB model.

Summarizing the results, both, intrinsic and extrinsic factors were predictors of NI. In this sense an intrinsic factor such as CDC classification (unstable patients) and an extrinsic one like mechanical ventilation (more than 3 days) presented the highest hazard rates. For any type of infection all the (statistically significant) intrinsic variables were risk factors for NI. In decreasing order of importance we could mention CDC classification (unstable) and previous community-acquired infections. Most of the extrinsic factors were also risk factors. In this sense, and again in decreasing order of importance, we can list mechanical ventilation (more than 3 days), urinary probe (5-12 days), location (bed 4), the presence of an arterial catheter, and central venous catheter. Only three of the extrinsic factors were protective, presence of tracheotomy, nasogastric probe (more than 9

days), and, in particular, antibiotic treatment (days of treatment).

The interpretation of the model for the variance of the frailty is also worth while. Note that only extrinsic factors (mechanical ventilation, tracheotomy and probes, urinary and nasogastric) explained the variance of the frailty. In this sense, the sources of heterogeneity, both between and within individuals, could be attributed to the medical procedures that the ICUs use, whereas the effect of those factors related to the susceptibility of the patients could only be marginal.

Conclusion

Our purpose was to present a modification of the AG gamma frailty model. In particular we proposed to directly model the variance of the

frailty by means of a GLM and also to modify the EM algorithm, using the EMB algorithm, in order to simultaneously estimate a semiparametric model for the failure times and a model for the variance of the frailty.

In both the simulation and in the application to multivariate data from a nosocomial infection study, the EMB fit turned out to be better than the fit obtained from a marginal model (PWP) and from a conditional one (penalized Cox model). In this sense, the EMB provided the best fit (being the least overdispersed and having the highest AIC and the highest pseudo-R square) and estimated the parameters most efficiently. We think, therefore, that our proposed method is able to take into account and to estimate unobservable random effects in semiparametric models.

Two shortcomings, however, should be mentioned. First, as in the rest of frailty models, we introduce frailties into the marginal hazards in order to explicitly model the dependence between recurrences. The problem is that frailties are in fact capturing two different, although related, sources of variation, that is heterogeneity (or false contagion) and serial dependence (or true contagion) (Aalen, 1994). The former, the original use of frailty, is a consequence of unobserved individual covariates that are not included in the study either because of practical circumstances or because they are not known to be risk factors. The latter is in fact a consequence of unobserved common covariates that are integrated out (Petersen, 1998).

It seems, at any rate, that frailty models successfully capture heterogeneity but permit a considerable amount of non-controlled serial dependence. A possible but partial solution tried here is to stratify according to the recurrences, as in the PWP model, thus allowing the hazard to vary between them. The second shortcoming, also shared with the rest of frailty models, is the lack of methods with which to assess the goodness-of-fit of our method. At any rate, we are sure that these shortcomings deserve further research.

References

Aalen, O. O. (1994). Effects of frailty in survival analysis. Scandinavian Methods in *Medical Research 3*, 227-243.

Andersen, P. K. &Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics*, *10*, 1100-1120.

Andersen, P. K. (1992). Repeated assessment of risk factors in survival analysis. *Statistical Methods in Medical Research*, *1*, 297-315.

Andersen, P. K., Borgan, O. Gill, R. D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York, Springer-Verlag.

Andersen, P. K Klein, J. P., Knudsen, K. M. & Tabanera-Palacios,R. (1997). Estimation of variance in Cox's regression model with shared gamma frailties. *Biometrics, 53*, 1475-1484.

Barceló, M. A. & Saez, M. (2001). Factors determining the incidence of nosocomial infections in an Intensive Care Unit of a public hospital in Girona, Spain. Research Group on Statistics, Applied Economics and Health (GRECS), Univesity of Girona, Spain

Behrman, J. R., Sickles, R., Taubman, P., & Yazbeck, A. (1991). Black-white mortality inequalities. *Journal of Econometrics*, *50*, 183-203.

Clayton, D. (1978). A model for association in bivariate life tables and its applications in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, *65*, 141-151.

Clayton, D. & Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *Journal of the Royal Statistical Society*, Series A, *148*, 82-117.

Clayton, D. (1988). The analysis of event history data: a review of progress and outstanding problems. *Statistics in Medicine*, *7*, 819-841.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society*, Series B *34*, 187-220.

Fleming, T. R., & Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. New York, John Wiley & Sons, Inc.

Gill, R. D. (1985). Discussion of multivariate generalizations of the proportional hazards model, by Clayton and Cuzick. *Journal of the Royal Statistical Society*, Series A *148*, 82-117.

Gill, R. D. (1989). Non- and semi-parametric maximum likelihood estimators of the von Mises method (Part 1). *Scandinavian Journal of Statistics, 16*, 97-128.

Klein, J. P. (1992). Semiparametric estimation of random effect using the Cox model based on the EM algorithm. *Biometrics*, *48*, 795-806.

Lee, Y. & Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). J*ournal of the Royal Statistical Society*, Series B, *58*, 619-678.

Lin, D. & y Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, *84*, (408), 1074-1078.

Lindsey, J. (1999). A review of some extensions to generalized linear models. *Statistics in Medicine*, *18*, 2223-2236.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Association*, Series B *44*, 226-233.

Nagelkirke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*, 691-692.

Nelder, J. A. (1998). A large class of models derived from generalized linear models. *Statistics in Medicine*, *17*, 2747-2753.

Nelder, J. A., & Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, *72*, 221-232.

Nelder, J. A., & Lee, Y. (1991). Generalized linear models for the analysis of Taguchi-type experiments. *Applied Stochastic Models and Data Analysis*, *7*, 107-120.

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. *Scandinavian Journal of Statistics*, *19*, 25-43.

Petersen, J. H. (1998). An additive frailty model for correlated life times. *Biometrics*, *54*, 646-661.

Pierce, D. A., & Schafer, D. W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association*, *81*, 396, 977-986.

Prentice, R. L., Williams, B. J., & Peterson, A. L. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, *68*, 2, 373-79.

Therneau, T. M., & Hamilton, S. A. (1997). rhDNase as an example of recurrent event analysis. *Statistics in Medicine*, *16*, 2029-2047.

Therneau, T. M., & Grambsch, P. (1998). *Penalized Cox Models*. Mayo Clinic.

Vaupel, J. W., Manton, K., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, *16*, 439-454.

Wassell, J. T., & Moeschberger, M. L. (1993). A bivariate survival model with modified gamma frailty for assessing the impact of interventions. *Statistics in Medicine*, *12*, 241-48.