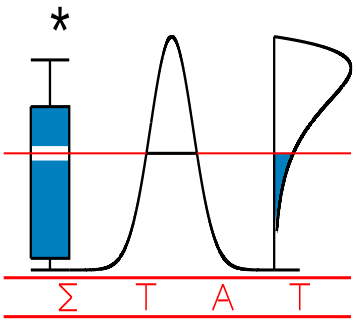
T E C H N I C A L

R E P O R T

0455

A VERSION OF THE EM ALGORITHM FOR PROPORTIONAL HAZARD MODEL WITH RANDOM EFFECTS

CORTINAS ABRAHANTES, J. and T. BURZYKOWSKI



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

A Version of the EM Algorithm for Proportional Hazard Model with Random Effects

José Cortiñas Abrahantes*[1] and Tomasz Burzykowski[1]

[1] Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B3590 Diepenbeek, Belgium.

Received 15 November 2004, revised 30 November 2004, accepted 2 December 2004

Published online 3 December 2004

Summary

Proportional hazard models with multivariate random effects (frailties) acting multiplicatively on the baseline hazard have recently become a topic of an intensive research. One of the main practical problems related to the models is the estimation of parameters. To this aim, several approaches based on the EM algorithm have been proposed. The major difference between these approaches is the method of the computation of conditional expectations required at the E-step. In this paper an alternative implementation of the EM algorithm is proposed, in which the expected values are computed with the use of the Laplace approximation. The method is computationally less demanding than the approaches developed previously. Its performance is assessed based on a simulation study and compared to a non-EM based estimation approach proposed by Ripatti and Palmgren (2000).

Key words: Multivariate failure-time data, frailty model, EM algorithm, Laplace approximation.

* Corresponding author: e-mail: jose.cortinas@luc.ac.be, phone: +0032 11 268 215, fax: +0032 11 268 199.

1 Introduction

Proportional hazard models with random effects acting multiplicatively on the baseline hazard, often called frailty models, have been focus of the research aimed at methods of analyzing multivariate or clustered failure-time data for a long time. Initially, the research concentrated on univariate shared frailty models, with a univariate random effect shared by all the observations from a particular cluster. These models have several limitations, e.g., they generally impose a positive association between the failure-times coming from the same cluster (Xue and Brookmeyer, 1996). For this reason, multivariate random-effects models have recently started to attract some attention. One of the main practical problems related to the latter is the estimation of the parameters. Several approaches have been proposed to deal with the problem. McGilchrist and Aisbett (1991) and McGilchrist (1993, 1994), extending the 'best linear unbiased prediction' argument for normal linear mixed-effects model, used the penalized partial likelihood approach to estimate the fixed effects and the restricted maximum likelihood to estimate the random effects. Xue and Brookmeyer (1996) formulated a bivariate log-normal random-effects model fitted using the EM algorithm, with numerical integration used at the E-step. Xue (1998) developed an alternative fitting method for the same model using estimating equations derived from a Poisson regression formulation, while Xue and Ding (1999) used a Gibbs sampling approach. Ripatti and Palmgren (2000) considered a more general form of the proportional hazard model with random effects and proposed estimation based on a penalized partial likelihood developed by applying the Laplace approximation to the marginal likelihood function. Vaida and Xu (2000), on the other hand, suggested a Monte Carlo EM (MCEM) algorithm, with Monte Carlo Markov Chain (MCMC) sampling used at the E-step. Ripatti, Larsen, and Palmgren (2002), following the ideas by Vaida and Xu (2000) introduced a MCEM algorithm, where the conditional expectations at the E-step were computed by drawing from a posterior distribution of the random effects using the rejection sampling. They also provided a stopping rule based on absolute convergence of the algorithm.

The purpose of this paper is to investigate an alternative implementation of the EM algorithm for the proportional hazard models with random effects. It is based on the use of the Laplace approximation at the E-step. The main advantage of the proposed method is that it is numerically simpler than, e.g., the use of MCMC methods or numerical integration.

The paper is organized as follows. Section 2 briefly recalls the proportional hazard model with random effects. In Section 3 the main features of the EM algorithm are summarized. In Section 4 the use of the Laplace approximation at the E-step is described. In Section 5 we briefly describe the approach proposed by Ripatti and Palmgren (2000). Section 6 presents results of a simulation study in which the performance of the proposed method is evaluated and compared with the approach proposed by Ripatti and Palmgren (2000). Both methods are also applied to a case study and the results are discussed in Section 7. The discussion, presented in Section 8, concludes the paper.

2 The Proportional Hazard Model with Random Effects

We will consider clustered failure-time data with N clusters. The failure-time variable corresponding to subject j ($j = 1, \ldots, n_i$) from cluster i ($i = 1, \ldots, N$) will be denoted by Y_{ij}. It is assumed that observations of Y_{ij} can be right-censored. Thus, for subject j in cluster i we observe $T_{ij} = \min(C_{ij}, Y_{ij})$, where C_{ij} is a random censoring time independent of Y_{ij}. Additionally, a censoring indicator δ_{ij} is observed, with δ_{ij} equal to 1 if $T_{ij} = Y_{ij}$, and 0 if $T_{ij} = C_{ij}$.

In the paper the following mixed-effects proportional hazard model for T_{ij} will be considered:

$$\lambda(t_{ij}|\beta_i, b_i) = \lambda_0(t_{ij}) \exp(x_{ij}^T \beta_i + z_{ij}^T b_i), \tag{1}$$

where $\lambda_0(t)$ is the baseline hazard function, β_i is a vector of cluster-specific fixed-effects corresponding to a vector of covariates x_{ij}, and b_i is a vector of random effects associated with a vector of covariates z_{ij}. The random effects b_i are assumed to be randomly distributed with mean 0 and variance-covariance matrix $D = D(\theta)$, which depends on a d-dimensional vector of parameters

$\theta = (\theta_1, \theta_2, ..., \theta_d)$. The density function of the b_i which, except for θ, is assumed to be known, will be denoted by $f(b_i)$. At this moment we do not need to specify the nature of the distribution in more detail. To simplify formulas, we will also use the baseline cumulative hazard defined as

$$\Lambda_0(t) = \int_0^t \lambda_0(u)du.$$

Model (1) can be seen as a linear mixed-effects model on the log-hazard scale. The estimation of the parameters β_i and θ from the observed data on T_{ij} is our main interest. Assuming the conditional independence of the observations within a cluster given b_i, one might write the (conditional) log-likelihood for the observed data as

$$l^C(\beta, \lambda_0, b) = \sum_{i=1}^{N} l_i^C(\beta_i, \lambda_0, b_i), \tag{2}$$

where

$$l_i^C(\beta_i, \lambda_0, b_i) = \sum_{j=1}^{n_i} \left[\delta_{ij} \left\{\ln \lambda_0(t_{ij}) + x_{ij}^T\beta_i + z_{ij}^T b_i\right\} - \Lambda_0(t_{ij}) \exp\left(x_{ij}^T\beta_i + z_{ij}^T b_i\right)\right] \tag{3}$$

is the (conditional) log-likelihood for the observed data in the ith cluster, and β and b denote the vectors resulting from "stacking" vectors β_i and b_i for all clusters, respectively. The (marginal) likelihood of the observed data for all clusters can then be expressed as

$$L^M(\beta, \theta, \lambda_0) = \prod_{i=1}^{N} \int L_i^A(\beta_i, \theta, \lambda_0, b_i)db_i, \tag{4}$$

where

$$L_i^A(\beta_i, \theta, \lambda_0, b_i) = f(b_i)e^{l_i^C(\beta_i, \lambda_0, b_i)}. \tag{5}$$

Function (5) can be regarded as the likelihood of the "augmented" data for cluster i, treating b_i as additional observations. Consequently,

$$L^A(\beta, \theta, \lambda_0, b) = \prod_{i=1}^{N} L_i^A(\beta_i, \theta, \lambda_0, b_i), \tag{6}$$

is the likelihood of the "augmented" data for all clusters.

One might consider using the likelihood function (4) in the inference on β and θ. There are two major problems with using it for this purpose, however. First, it depends on the baseline hazard function λ_0. Second, the integral in (4) will usually be multi-dimensional, unless a very simple model is considered, and in general will not be available in a closed form. For these reasons, the use of the EM algorithm to estimate the parameters of model (1) has been proposed (Klein, 1992; Xue and Brookmeyer, 1996; Vaida and Xu, 2000). In the following section the basic features of the EM algorithm are reviewed.

3 The EM Algorithm

The EM algorithm consists of two steps: the E-step and the M-step. Starting from initial values of parameters, the algorithm iterates between the two steps until convergence is reached (Dempster, Laird and Rubin 1977). It is important to remark that, under regularity conditions, the algorithm is guaranteed to converge to a stationary point (Dempster, Laird and Rubin 1977; Wu 1983; Vaida 2004). The E-step and the M-step of the algorithm to estimate the parameters of model (1) will be reviewed in more detail now.

3.1 The E-step

In the E-step the expectation of the logarithm of the augmented-data likelihood (6), conditional on the observed data and on the current values $\widetilde{\beta}$, $\widetilde{\theta}$ and $\widetilde{\lambda}_0$ of parameters β, θ and λ_0, respectively, is computed. The expectation will be denoted by $Q(\beta, \theta, \lambda_0)$. It turns out that it can be expressed as (Klein, 1992; Vaida and Xu, 2000)

$$Q(\beta, \lambda_0, \theta) = Q_1(\beta, \lambda_0) + Q_2(\theta), \tag{7}$$

where

$$Q_1(\beta, \lambda_0) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left[\delta_{ij} \left\{ \ln \lambda_0(t_{ij}) + x_{ij}^T \beta_i + z_{ij}^T \mathrm{E}(b_i) \right\} - \Lambda_0(t_{ij}) \exp \left\{ x_{ij}^T \beta_i + \ln \mathrm{E}(e^{z_{ij}^T b_i}) \right\} \right]$$

(8)

and

$$Q_2(\theta) = \sum_{i=1}^{N} \mathrm{E}[\ln f(b_i)], \tag{9}$$

with E(.) denoting conditional expected values given the observed values of the T_{ij} and δ_{ij}. To simplify the notation, the dependence of the expected values in (8) and (9) on the observed data and $\widetilde{\beta}$, $\widetilde{\theta}$ and $\widetilde{\lambda}_0$ has been suppressed.

The set of initial values for β and λ_0 can be obtained using the Cox regression without random effects. The initial values for θ can be specified by taking $D(\theta)$ equal to, e.g., the identity matrix.

3.2 The M-step

In the M-step new estimates $\widetilde{\beta}$ and $\widetilde{\theta}$ are found by maximizing the functions Q_1 and Q_2, respectively. The estimation of β is complicated by the dependence of Q_1 on λ_0. Via the profile-likelihood arguments for λ_0 (Johansen, 1993; Vaida and Xu, 2000) one can arrive at the following estimating function for β_i:

$$Q_1'(\beta) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \delta_{ij} \left[x_{ij}^T \beta_i - \ln \sum_{t_{kl} \geq t_{ij}} \exp\left\{ x_{kl}^T \beta_k + \ln \mathrm{E}(e^{z_{kl}^T b_k}) \right\} \right]. \tag{10}$$

The form of (10) resembles that of the partial log-likelihood for the Cox proportional hazard model with offsets $\ln \mathrm{E}(e^{z_{ij}^T b_i})$. Estimates of parameters β_i, can thus be obtained by maximizing Q_1' using standard software for the Cox model.

If the density f of the random effects b_i belongs to an exponential family, then Q_2 is the log-likelihood of a sample of N observations with sufficient statistics replaced by their conditional expectations. In such a situation, the estimation of θ is generally straightforward and can be achieved by maximizing Q_2. For instance, consider the case where the random effects are multivariate normal with mean 0 and an unconstrained variance-covariance matrix D. Then maximizing

Q_2 would lead to the estimator

$$\hat{D} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}\left(b_i b_i^T\right), \tag{11}$$

where, again, the expectation is conditional on the observed data and $\widetilde{\beta}$, $\widetilde{\theta}$ and $\widetilde{\lambda}_0$.

3.3 Variance estimation for the EM

The variance-covariance matrix of the solution $(\hat{\beta}, \hat{\lambda}_0, \hat{\theta})$ obtained from the EM algorithm, can be estimated using the inverse of an observed information matrix computed from the formula proposed by Louis (1982):

$$I(\beta, \lambda_0, \theta) = \left[\mathrm{E}\left\{-l^{A''}(\beta, \lambda_0, \theta)\right\} - \mathrm{E}\left\{l^{A'}(\beta, \lambda_0, \theta) l^{A'}(\beta, \lambda_0, \theta)^T\right\}\right], \tag{12}$$

where $l^{A'}$ and $l^{A''}$ are the first and the second derivatives with respect to $(\beta, \lambda_0, \theta)$ of the logarithm of the "augmented" likelihood (6). More explicitly, the components of $l^{A'}$ are

$$l^{A'} = \begin{pmatrix} I'_\beta \\ I'_\lambda \\ I'_\theta \end{pmatrix}, \quad I'_\beta = \begin{pmatrix} I'_{\beta_1} \\ I'_{\beta_2} \\ \vdots \\ I'_{\beta_N} \end{pmatrix}, \quad I'_\lambda = \begin{pmatrix} I'_{\lambda_1} \\ I'_{\lambda_2} \\ \vdots \\ I'_{\lambda_r} \end{pmatrix} \text{ and } \quad I'_\theta = \begin{pmatrix} I'_{\theta_1} \\ I'_{\theta_2} \\ \vdots \\ I'_{\theta_d} \end{pmatrix},$$

where

$$I'_{\beta_i} = \frac{\partial \ln L^A}{\partial \beta_i} = \sum_{j=1}^{n_i} x_{ij}\left\{\delta_{ij} - \Lambda_0(t_{ij}) \exp\left(x_{ij}^T \beta_i + z_{ij}^T b_i\right)\right\}, \tag{13}$$

$$I'_{\lambda_m} = \frac{\partial \ln L^A}{\partial \lambda_m} = \frac{1}{\lambda_m} - \sum_{t_{kl} \geq t_m} \exp\left(x_{kl}^T \beta_k + z_{kl}^T b_k\right), \tag{14}$$

$$I'_{\theta_k} = \frac{\partial \ln L^A}{\partial \theta_k} = \frac{\partial \ln f(b_i)}{\partial \theta_k}, \tag{15}$$

with $\lambda_m = \lambda_0(t_m)$, where t_m $(m = 1, \ldots, r)$ are the distinct uncensored failure times.

The components of the second derivative $l^{A''}$ are:

$$l^{A''} = \begin{pmatrix} I''_{\beta\beta} & I''_{\beta\lambda} & I''_{\beta\theta} \\ I''_{\beta\lambda} & I''_{\lambda\lambda} & I''_{\lambda\theta} \\ I''_{\beta\theta} & I''_{\lambda\theta} & I''_{\theta\theta} \end{pmatrix}$$

where

$$I''_{\beta\beta} = \begin{pmatrix} I''_{\beta_1\beta_1} & I''_{\beta_1\beta_2} & \dots & I''_{\beta_1\beta_N} \\ I''_{\beta_1\beta_2} & I''_{\beta_2\beta_2} & \dots & I''_{\beta_2\beta_N} \\ \vdots & \vdots & \ddots & \vdots \\ I''_{\beta_1\beta_N} & I''_{\beta_2\beta_N} & \dots & I''_{\beta_N\beta_N} \end{pmatrix}, \quad I''_{\lambda\lambda} = \begin{pmatrix} I''_{\lambda_1\lambda_1} & I''_{\lambda_1\lambda_2} & \dots & I''_{\lambda_1\lambda_r} \\ I''_{\lambda_1\lambda_2} & I''_{\lambda_2\lambda_2} & \dots & I''_{\lambda_2\lambda_r} \\ \vdots & \vdots & \ddots & \vdots \\ I''_{\lambda_1\lambda_r} & I''_{\lambda_2\lambda_r} & \dots & I''_{\lambda_r\lambda_r} \end{pmatrix},$$

$$I''_{\beta\lambda} = \begin{pmatrix} I''_{\beta_1\lambda_1} & I''_{\beta_1\lambda_2} & \dots & I''_{\beta_1\lambda_r} \\ I''_{\beta_2\lambda_1} & I''_{\beta_2\lambda_2} & \dots & I''_{\beta_2\lambda_r} \\ \vdots & \vdots & \ddots & \vdots \\ I''_{\beta_N\lambda_1} & I''_{\beta_N\lambda_2} & \dots & I''_{\beta_N\lambda_r} \end{pmatrix}, \quad I''_{\theta\theta} = \begin{pmatrix} I''_{\theta_1\theta_1} & I''_{\theta_1\theta_2} & \dots & I''_{\theta_1\theta_d} \\ I''_{\theta_1\theta_2} & I''_{\theta_2\theta_2} & \dots & I''_{\theta_2\theta_d} \\ \vdots & \vdots & \ddots & \vdots \\ I''_{\theta_1\theta_d} & I''_{\theta_2\theta_d} & \dots & I''_{\theta_d\theta_d}, \end{pmatrix}$$

with

$$I''_{\beta_i\beta_{i'}} = \frac{\partial^2 \ln L^A}{\partial\beta_i\partial\beta_{i'}} = \left[-\sum_{j=1}^{n_i} x_{ij}x_{i'j}^T\Lambda_0(t_{ij})\exp\left(x_{ij}^T\beta_i + z_{ij}^Tb_i\right)\right]\mathbf{1}(i=i'), \tag{16}$$

$$I''_{\lambda_m\lambda_{m'}} = \frac{\partial^2 \ln L^A}{\partial\lambda_m\lambda_{m'}} = \frac{1}{\lambda_m^2}\mathbf{1}(m=m'), \tag{17}$$

$$I''_{\beta_k\lambda_m} = \frac{\partial^2 l}{\partial\beta_k\partial\lambda_m} = -\sum_{t_{kl}\geq t_m} x_{kl}\exp[x_{kl}^T\beta_k + z_{kl}^Tb_k)], \tag{18}$$

$$I''_{\theta_k\theta_{k'}} = \frac{\partial^2 \ln L^A}{\partial\theta_k\partial\theta_{k'}} = \frac{\partial^2 \ln f(b_i)}{\partial\theta_k\partial\theta_{k'}}, \tag{19}$$

with $\mathbf{1}(B)$ being the indicator function of event B.

The other off diagonal elements $(I''_{\beta\theta}, I''_{\lambda\theta})$ of $l^{A''}$ are zero (Vaida and Xu 2000).

4 Issues in the Implementation of the EM Algorithm

The use of the EM algorithm, as described above, is complicated by the need to compute the conditional expected values in (8) and (9) at the E-step. Usually, they will not be available in a closed-form. To compute the expected values, Xue and Brookmeyer (1996) proposed to use numerical integration. This solution is feasible, however, only for low-dimensional random vectors b_i. Vaida and Xu (2000) and Ripatti et al. (2002) proposed to use MCMC methods. This approach

is numerically intensive and introduces issues related to the assessment of the convergence of the MCMC algorithm.

An alternative solution, not yet considered in the literature, is to use the Laplace approximation. This is the option we will discuss in more detail now.

4.1 The Laplace approximation

The approximation of multidimensional integrals can be obtained in many ways (see, e.g., Murray 1984; Bleistein and Handelsman 1986; Wong 1989). One of the most common techniques is the Laplace method (Evans and Swartz 2000). In the multivariate context the Laplace theorem states that, under some weak conditions, the following asymptotic equivalence holds:

$$\int_A h(t)e^{-\phi k(t)}\, du \underset{\phi\to+\infty}{\approx} h(\hat{t})e^{-\phi k(\hat{t})}\sqrt{\frac{(2\pi)^d}{|\phi K(\hat{t})|}}, \tag{20}$$

where A is an open subset of $\mathbb{R}^d$, $\phi > 0$ is a real-valued parameter, $K(t)$ is the matrix of the second derivatives of $k(t)$, and $\hat{t}$ is an isolated global minimum of $k(t)$ over A.

4.2 EM and the Laplace approximation

As it was mentioned in section 3.1, at the E-step, we need conditional expectations of functions of the random effects. The conditional expectations involve integrals of the form

$$\mathrm{E}\{g(b_i)\} = \frac{\int g(b_i)e^{l_i^C\left(\tilde{\beta}_i,\tilde{\lambda}_0,b_i\right)+\ln f(b_i)}db_i}{\int e^{l_i^C(\tilde{\beta}_i,\tilde{\lambda}_0,b_i)+\ln f(b_i)}db_i}\ . \tag{21}$$

Using the Laplace formula, it can be shown that

$$\mathrm{E}\{g(b_i)\} \approx g(\hat{b}_i)\ , \tag{22}$$

where $\hat{b}_i$ is an isolated global minimum of

$$k(b_i) = -\left\{l_i^C\left(\tilde{\beta}_i,\tilde{\lambda}_0,b_i\right)+\ln f(b_i)\right\}\ . \tag{23}$$

It is the first-order approximation, as it is based on first-order terms of the Taylor series expansion. The formal asymptotic error order of the approximation is $O(n_i^{-1})$. It is possible to construct

higher-order approximations which involve higher-order terms of the Taylor series expansion (Kass, Tierney, and Kadane, 1990).

By the use of the Laplace approximation, the problem of the computation of the expected values (21) is translated into the need of finding the isolated global minimum $\hat{b}_i$ of the function given by (23). Various numerical procedures are available for this purpose. In most cases, these procedures will require less computation time than, e.g., multi-dimensional numerical integration or MCMC methods.

4.3 Estimation of variance

To estimate the variance-covariance matrix using the information matrix $I(\beta, \lambda_0, \theta)$ defined by (12), one also needs conditional expectations of functions of b_i. Again, the Laplace approximation can be used to compute these expectations. One additional problem is related to the fact that, in order to compute standard error for the parameters of model (1), it would be necessary to invert $I(\beta, \lambda_0, \theta)$. The dimension of the matrix can be very large, since it depends on the number of distinct uncensored failure times. A possible way to tackle this problem is by inverting only the submatrices we are interested in. More specifically, let us partition the information matrix $I(\beta, \lambda_0, \theta)$ as follows:

$$I(\beta, \lambda_0, \theta) = \begin{pmatrix} I_{\beta\beta} & I_{\beta\lambda} & I_{\beta\theta} \\ I_{\beta\lambda} & I_{\lambda\lambda} & I_{\lambda\theta} \\ I_{\beta\theta} & I_{\lambda\theta} & I_{\theta\theta} \end{pmatrix} = \begin{pmatrix} I^{\star}_{\beta\lambda} & I^{\star}_{\beta\lambda\theta} \\ (I^{\star}_{\beta\lambda\theta})^T & I_{\theta\theta} \end{pmatrix},$$

where

$$I^{\star}_{\beta\lambda} = \begin{pmatrix} I_{\beta\beta} & I_{\beta\lambda} \\ I_{\beta\lambda} & I_{\lambda\lambda} \end{pmatrix} \text{ and } I^{\star}_{\beta\lambda\theta} = \begin{pmatrix} I_{\beta\theta} \\ I_{\lambda\theta} \end{pmatrix}.$$

Now, instead of inverting $I(\beta, \lambda_0, \theta)$, one might consider inverting only submatrices $I_{\beta\beta}$ and $I_{\theta\theta}$. In fact, other authors also considered a similar solution: for instance, Therneau and Grambsch (2000) considered the 'sparse' option in S-Plus software to avoid the computation of the inverse

of the full information matrix. In this paper we will follow this simplified strategy and invert only submatrices $I_{\beta\beta}$ and $I_{\theta\theta}$. Its adequacy for the computation of standard error of parameter estimates will be evaluated in the simulations presented in Section 6. In the next section we will briefly describe Ripatti and Palmgren's approach (2000), with which we will compare the estimation method that we are proposing in this article.

5 The Approach of Ripatti and Palmgren (2000)

Using the derivation of a penalized likelihood solution obtained by Breslow and Clayton (1993) for the generalized linear mixed model assuming Gaussian random effects, Ripatti and Palmgren (2000) presented a parallel approximation for model (1). To this aim, they approximated the marginal likelihood (4) using the Laplace approximation. Assuming that the random effects are normally distributed with variance-covariance matrix $D(\theta)$, the marginal likelihood can be expressed as

$$L^M(\beta, \theta, \lambda_0) = c|D(\theta)|^{-\frac{N}{2}} \int e^{-\kappa(b)} db, \tag{24}$$

where

$$\kappa(b) = l^C(\beta, \lambda_0, b) - \frac{1}{2} b^T D(\theta)^{-1} b, \tag{25}$$

with $l^C(\beta, \lambda_0, b)$ given by (2). Using the Laplace theorem, Ripatti and Palmgren (2000) showed that the logarithm of (24) can be approximated by

$$l^M(\beta, \theta, \lambda_0) \approx -\frac{N}{2} |\ln D(\theta)| - \frac{1}{2} \ln |\kappa''(\tilde{b})| - \kappa(\tilde{b}), \tag{26}$$

where κ' and κ'' denote, respectively, the first and the second order partial derivatives of κ with respect to b, and $\tilde{b} = \tilde{b}(\beta, \theta)$ is the solution to $\kappa'(\tilde{b}) = 0$. They further argued that, for fixed θ, the values $\hat{\beta}(\theta)$ and $\hat{b}(\theta)$, which maximize the penalized log-likelihood (25), also maximize the

penalized partial log-likelihood

$$\sum_{i=1}^{N}\sum_{j=1}^{n_i}\delta_{ij}\left[(x_{ij}^T\beta + z_{ij}^T b_i) - \ln\sum_{t_{kl}\geq t_{ij}}\exp\left\{x_{kl}^T\beta + z_{kl}^T b_k)\right\}\right] - \frac{1}{2}b^T D(\theta)^{-1}b. \tag{27}$$

Based on the penalized partial log-likelihood (27), the estimating equations for $\beta(\theta)$ and $b(\theta)$, given θ, can be derived. Once $\hat{\beta}(\theta)$ and $\hat{b}(\theta)$ are computed, θ can be updated by maximizing the approximate profile likelihood derived from (26):

$$l^{PPL}(\hat{\beta}(\theta), \hat{b}(\theta), \theta) \approx -\frac{N}{2}|\ln D(\theta)| - \frac{1}{2}\ln|\kappa''(\hat{b})| - \frac{1}{2}\hat{b}^T D(\theta)^{-1}\hat{b}. \tag{28}$$

Based on empirical evidence, Ripatti and Palmgren (2000) proposed to use in (28) $\kappa''_{PPL}(\hat{b}) = (\partial^2 l^{PPL})/(\partial b \partial b^T)$ instead of $\kappa''(\hat{b})$.

To obtain the standard error of the estimated fixed effects, one can use standard software for the Cox model with the estimated random effects as an offset. To calculate the standard error of the estimates of variance-covariance parameters θ, Ripatti and Palmgren (2000) suggest the computation of the expected value, with respect to b, of the second derivative of (28) with respect to θ. The necesssary formulas are given in Ripatti and Palmgren (2000).

6 Simulation Study

The performance of the EM algorithm with the Laplace approximation at the E-step was evaluated in a set of simulations. The setting of the simulation does not strictly fall under the setup discussed earlier, it is an extension in which we allow different baseline hazards for each of the failure-time. The data were generated using the following proportional hazard model:

$$\lambda_{ij1}(t_{ij1}|\beta_{i1}, b_{i1}) = \lambda_1(t_{ij1})e^{b_{i1}+x_{ij}^T\beta_1}, \tag{29}$$

$$\lambda_{ij2}(t_{ij2}|\beta_{i2}, b_{i2}) = \lambda_2(t_{ij2})e^{b_{i2}+x_{ij}^T\beta_2}, \tag{30}$$

with

$$\begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} \sim N_2 \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right\}. \tag{31}$$

The model corresponds to the setting of data for N clusters (indexed by i), containing n_i observations (indexed by j) for each of two (possibly censored) failure-times. The two failure-times are of potentially different nature, what is reflected in model (29)–(30) by the use of different baseline hazards (λ_1 and λ_2). The random effects b_1 and b_2 are correlated. Note that, conditionally on the random effects b_i, no extra association between the two times is assumed. This setting may be seen as corresponding to, e.g., a multi-center clinical trial with centers as clusters and two different, independent failure-times recorded for each patient. In the model the (fixed) effect of a single binary covariate x_{ij} was considered. This can be regarded as, e.g., a time- and center-specific effect of treatment.

Model (29)–(30) is similar to one of the models considered for simulations by Ripatti and Palmgren (2000). They considered $N = 50$ clusters with $n_i = 2$ observations for each of the two failure-times.

In our simulation study, as compared to the one conducted by Ripatti and Palmgren (2000), a broader range of configurations of the parameters was considered. The aim was to investigate the performance of the proposed version of the EM algorithm for varying numbers of clusters and observations per cluster, percentage of censored observations, and magnitude of the variance and covariance parameters associated with the distribution of the random effects (31). Moreover, a comparison with the performance of the alternative, non-EM based estimation method of Ripatti and Palmgren (2000), was of interest.

More specifically, the number of clusters ranged between 10 and 100 ($N = 10, 20, 50, 100$). The number of bivariate observations (subjects) within the cluster varied from 20 to 100 ($n_i = 20, 50, 100$). (We will slightly abuse the notation now and use n_i to denote the number of pairs

of observed failure-times rather than the total number of observations per cluster.) The baseline hazards were assumed constant, with $\lambda_1(t) = 0.5$ and $\lambda_2(t) = 1$. The effects of covariate x_{ij} were assumed to be equal, $\beta_{i1} = \beta_{i2} \equiv \beta$, with $\beta = 1$. The variances associated with the random effects b_{i1} and b_{i2} were also assumed to be equal, $\sigma_1^2 = \sigma_2^2 \equiv \sigma^2$, with $\sigma^2 = 0.2$ and $\sigma^2 = 1$. Two values of the covariance parameter σ_{12} were considered for each value of σ^2: $\sigma_{12} = 0.1$ and 0.18 for $\sigma^2 = 0.2$, and $\sigma_{12} = 0.5$ and 0.9 for $\sigma^2 = 1$. This is equivalent to assuming, for each value of σ^2, two different values (0.5 and 0.9) of the correlation coefficient ρ for b_1 and b_2. None or 20% censoring was considered. The censoring was induced by using a pair of independent random variables, generated from two different uniform distributions, so that 20% of observations for each of the two failure-times were censored. For each setting of the parameters, 250 simulated datasets were generated.

The EM algorithm was implemented using SAS-IML v8.2 (the code can be obtained from the first author upon request). Both the first- and second-order Laplace approximations were considered. However, in simulations the results for the second-order approximation were essentially the same as for the first-order. Thus, in what follows, the use of the first-order approximation is assumed.

The method proposed by Ripatti and Palmgren (2000) was applied using the S-Plus functions developed by Therneau (2003). The functions do not produce standard errors of the estimated parameters; they were obtained separately using the formulas provided by Ripatti and Palmgren (2000).

For both methods, the common value β of the fixed-effects parameters β_1 and β_2 was estimated using data from both failure-times. On the other hand, although it was assumed that $\sigma_1^2 = \sigma_2^2$, the two parameters were estimated separately. This latter choice was motivated by our interest in the assessment of the ability of both methods to distinguish between different components of variability.

The preliminary set of simulations indicated considerable bias in the estimation of parameters σ_1^2, σ_2^2 and ρ, especially for $\sigma^2 = 0.2$. For univariate shared frailty models Therneau and Grambsch (2000, p. 254) argue that the estimates of random effects should be centered so that the "penalty" term Q_2 in (7) is minimized. In the case of normally-distributed random effects this means their estimates should have zero mean. In fact, this is the solution used, e.g., in the implementation of the univariate shared frailty model in S-Plus software (Therneau, Grambsch, and Pankratz, 2000). For the more general model (1) the argument of Therneau and Grambsch holds only for random intercepts. Nevertheless, we have followed the idea and modified the EM algorithm by centering the estimates of the random effects for all covariates at zero after each E-step.

Tables 1 and 2 present the results of simulations for the four combinations of the values of parameters ρ and σ^2. Only results for 20% censoring are presented, as the results under no censoring are qualitatively similar, but with a slightly smaller bias and variability of the estimated quantities.

One can observe that, in general, the fixed-effect β is estimated well by both methods, with a relative absolute bias less then 8% in any of the considered cases. The bias decreases with the increasing cluster size n_i, but is not substantially influenced by the number of clusters N. Increasing σ^2 from 0.2 to 1 or ρ from 0.5 to 0.9 does not seem to change the magnitude of bias. The estimates obtained by the Ripatti and Palmgren (2000) approach are on average closer to the true value of the parameter ($\beta = 1$). The variability of estimates of β, measured by the empirical standard error, is similar for both methods. In general, the model-based estimates for the proposed version of the EM algorithm adequately estimate this variability, though with a slight underestimation (especially for $n_i = 10$). The model-based estimates for the Ripatti and Palmgren method give plausible values for $n_i = 50$. For smaller cluster sizes, however, they overestimate the empirical variability. The overestimation is substantial especially for $n_i = 10$. Overall, the mean

Table 1: The mean estimates for 250 simulated datasets for the proposed EM algorithm (first row for each N) and the method of Ripatti and Palmgren (second row for each N), when $\sigma_1^2 = \sigma_2^2 = 0.2$ and different values of σ_{12}, with 20% censoring. In parentheses: the mean estimated (first number) and empirical (second number) standard errors.

N	β	σ_1^2	σ_2^2	σ_{12}	ρ
		$\sigma_{12} = 0.1$			
		$n_i = 10$			
10	1.069 (0.248;0.303)	0.269 (0.153;0.122)	0.263 (0.146;0.122)	0.122 (0.109;0.115)	0.458
	1.007 (0.252;0.277)	0.258 (0.074;0.164)	0.235 (0.115;0.156)	0.134 (0.098;0.168)	0.546
50	1.069 (0.105;0.118)	0.210 (0.072;0.053)	0.211 (0.069;0.044)	0.101 (0.053;0.055)	0.481
	0.990 (0.111;0.103)	0.200 (0.039;0.071)	0.192 (0.040;0.064)	0.097 (0.050;0.051)	0.497
100	1.074 (0.077;0.084)	0.206 (0.051;0.039)	0.208 (0.050;0.032)	0.100 (0.038;0.041)	0.486
	0.992 (0.079;0.073)	0.196 (0.020;0.052)	0.194 (0.020;0.047)	0.097 (0.025;0.037)	0.498
		$n_i = 50$			
10	1.015 (0.104;0.107)	0.180 (0.091;0.097)	0.182 (0.089;0.091)	0.087 (0.069;0.077)	0.482
	1.004 (0.108;0.114)	0.181 (0.065;0.105)	0.180 (0.066;0.098)	0.094 (0.037;0.082)	0.518
50	1.010 (0.046;0.044)	0.195 (0.044;0.045)	0.195 (0.044;0.039)	0.097 (0.034;0.035)	0.495
	1.002 (0.048;0.046)	0.197 (0.040;0.048)	0.195 (0.036;0.042)	0.099 (0.026;0.035)	0.504
100	1.012 (0.033;0.032)	0.199 (0.032;0.031)	0.198 (0.032;0.027)	0.099 (0.025;0.025)	0.499
	0.999 (0.034;0.033)	0.199 (0.030;0.033)	0.200 (0.023;0.029)	0.100 (0.021;0.025)	0.501
		$\sigma_{12} = 0.18$			
		$n_i = 10$			
10	1.080 (0.240;0.263)	0.322 (0.108;0.172)	0.301 (0.111;0.156)	0.263 (0.093;0.140)	0.845
	1.020 (0.252;0.269)	0.318 (0.094;0.180)	0.290 (0.114;0.173)	0.263 (0.110;0.177)	0.867
50	1.071 (0.105;0.118)	0.262 (0.053;0.060)	0.255 (0.052;0.054)	0.226 (0.045;0.055)	0.877
	1.000 (0.113;0.104)	0.256 (0.060;0.112)	0.241 (0.067;0.103)	0.221 (0.084;0.110)	0.891
100	1.076 (0.074;0.084)	0.229 (0.037;0.038)	0.222 (0.037;0.033)	0.199 (0.032;0.035)	0.885
	0.997 (0.081;0.075)	0.224 (0.051;0.083)	0.218 (0.048;0.077)	0.198 (0.050;0.081)	0.895
		$n_i = 50$			
10	1.015 (0.104;0.108)	0.191 (0.091;0.104)	0.191 (0.091;0.098)	0.170 (0.081;0.096)	0.889
	1.007 (0.108;0.113)	0.194 (0.070;0.091)	0.195 (0.072;0.085)	0.176 (0.067;0.082)	0.903
50	1.010 (0.046;0.044)	0.196 (0.044;0.048)	0.195 (0.044;0.043)	0.175 (0.040;0.045)	0.896
	1.002 (0.048;0.045)	0.200 (0.055;0.060)	0.197 (0.048;0.053)	0.179 (0.044;0.054)	0.902
100	1.012 (0.033;0.032)	0.200 (0.032;0.033)	0.199 (0.032;0.029)	0.179 (0.029;0.031)	0.899
	0.999 (0.034;0.032)	0.199 (0.030;0.035)	0.199 (0.029;0.031)	0.180 (0.027;0.030)	0.903

squared error is generally smaller (data not shown) for the estimates obtained by the Ripatti and Palmgren approach.

The relative bias for σ_1^2 is presented graphically in Figure 1; the estimates for σ_2^2 show a similar behaviour. One can conclude that for both estimation approaches there is a substantial bias when the number of clusters is small ($N = 10$). The absolute bias decreases with increasing N and n_i, but it remains above 10% even for $N = 100$ if the cluster size is small ($n_i = 10$) and there is low variability in cluster-specific random effects ($\sigma^2 = 0.2$). If the variability is large ($\sigma^2 = 1$), the

Table 2: The mean estimates for 250 simulated datasets for the proposed EM algorithm (first row for each N) and the method of Ripatti and Palmgren (second row for each N), when $\sigma_1^2 = \sigma_2^2 = 1.0$ and different values of σ_{12}, with 20% censoring. In parentheses: the mean estimated (first number) and empirical (second number) standard errors.

N	β	σ_1^2	σ_2^2	σ_{12}	ρ
		$\sigma_{12} = 0.5$			
		$n_i = 10$			
10	1.078 (0.238;0.276)	0.900 (0.535;0.687)	0.895 (0.523;0.617)	0.414 (0.394;0.387)	0.461
	0.994 (0.262;0.284)	0.925 (0.406;0.496)	0.886 (0.396;0.479)	0.444 (0.252;0.364)	0.491
50	1.068 (0.104;0.122)	0.974 (0.264;0.435)	0.974 (0.251;0.285)	0.472 (0.198;0.192)	0.485
	0.995 (0.116;0.109)	0.982 (0.182;0.232)	0.971 (0.184;0.224)	0.490 (0.115;0.176)	0.502
100	1.075 (0.073;0.086)	0.982 (0.185;0.283)	0.984 (0.179;0.207)	0.481 (0.141;0.146)	0.489
	0.995 (0.082;0.077)	0.983 (0.139;0.180)	0.979 (0.125;0.163)	0.493 (0.102;0.134)	0.503
		$n_i = 50$			
10	1.011 (0.104;0.108)	0.893 (0.423;0.487)	0.901 (0.424;0.454)	0.439 (0.329;0.352)	0.489
	1.007 (0.109;0.114)	0.899 (0.362;0.480)	0.911 (0.350;0.443)	0.466 (0.284;0.343)	0.496
50	1.009 (0.046;0.045)	0.976 (0.208;0.284)	0.979 (0.205;0.202)	0.486 (0.162;0.171)	0.498
	1.003 (0.049;0.049)	0.979 (0.204;0.224)	0.976 (0.171;0.195)	0.491 (0.138;0.165)	0.499
100	1.011 (0.032;0.033)	0.991 (0.147;0.160)	1.003 (0.150;0.145)	0.497 (0.117;0.121)	0.498
	0.999 (0.034;0.033)	0.990 (0.127;0.148)	1.004 (0.122;0.138)	0.504 (0.104;0.118)	0.504
		$\sigma_{12} = 0.9$			
		$n_i = 10$			
10	1.081 (0.239;0.278)	0.881 (0.424;0.505)	0.876 (0.438;0.499)	0.780 (0.392;0.412)	0.887
	0.995 (0.257;0.275)	0.886 (0.296;0.411)	0.860 (0.282;0.400)	0.779 (0.246;0.360)	0.893
50	1.069 (0.104;0.122)	0.977 (0.225;0.270)	0.969 (0.224;0.243)	0.869 (0.208;0.202)	0.893
	0.992 (0.117;0.107)	0.979 (0.166;0.225)	0.971 (0.163;0.226)	0.878 (0.124;0.199)	0.900
100	1.076 (0.073;0.086)	0.980 (0.160;0.185)	0.971 (0.158;0.162)	0.873 (0.147;0.141)	0.895
	0.993 (0.083;0.076)	0.985 (0.139;0.175)	0.976 (0.138;0.170)	0.884 (0.114;0.156)	0.901
		$n_i = 50$			
10	1.013 (0.104;0.109)	0.893 (0.424;0.505)	0.902 (0.432;0.476)	0.798 (0.399;0.428)	0.889
	1.005 (0.109;0.111)	0.901 (0.365;0.477)	0.904 (0.377;0.448)	0.815 (0.297;0.436)	0.903
50	1.010 (0.046;0.045)	0.980 (0.205;0.229)	0.974 (0.204;0.216)	0.876 (0.192;0.205)	0.896
	0.999 (0.048;0.048)	0.980 (0.217;0.232)	0.976 (0.197;0.213)	0.875 (0.211;0.223)	0.895
100	1.012 (0.032;0.032)	0.993 (0.148;0.174)	0.997 (0.149;0.151)	0.894 (0.140;0.150)	0.898
	0.998 (0.034;0.033)	0.994 (0.149;0.157)	0.999 (0.140;0.145)	0.898 (0.139;0.146)	0.901

absolute bias is low (around or below 5%) for $N \geq 20$, irrespectively of the cluster size n_i. It is worth noting that both estimation methods produce underestimates for $\sigma^2 = 1$, irrespectively of N and n_i. In general, the estimates for both methods give on average similar results, with a close agreement for $\rho = 0.9$. The empirical variability of estimates of σ_1^2 and σ_2^2 is slightly smaller for the proposed version of the EM algorithm, as compared to the method of Ripatti and Palmgren, for $\sigma^2 = 0.2$. For $\sigma^2 = 1$, the opposite trend seems to be present. In general, the model-based standard errors underestimate the variability for both methods. The estimates for the proposed

version of the EM algorithm are in most cases closer to the empirical standard error than the values obtained for the Ripatti and Palmgren method.

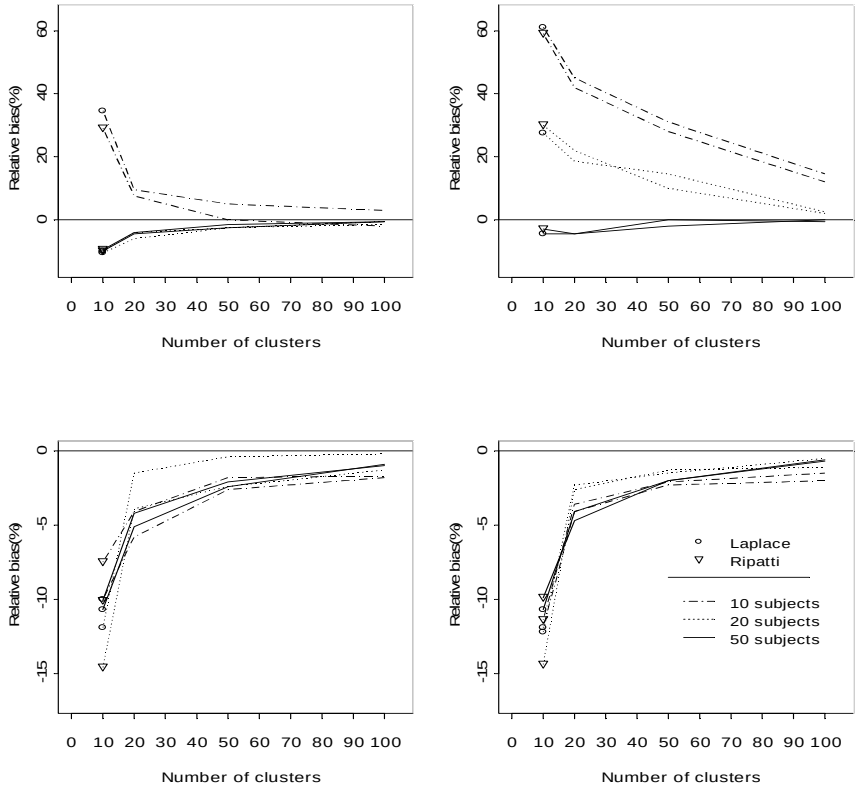


Fig. 1: The relative bias for σ_1^2 for 20% censoring. Left column: $\rho = 0.5$; right column: $\rho = 0.9$. Top row: $\sigma_1^2 = 0.2$; bottom row: $\sigma_1^2 = 1$. The legend in the bottom right panel refers to all plots.

Both estimation methods tend to underestimate the covariance parameter σ_{12} for $\sigma^2 = 1$. For $\sigma^2 = 0.2$, there is no obvious pattern. The absolute bias decreases with increasing N and n_i. In general, it is smaller for the Ripatti and Palmgren method. The empirical variability of the estimates is similar for both methods. The model-based standard errors underestimate the variability for both methods. The estimates for the proposed version of the EM algorithm

are generally closer to the empirical standard error than the values obtained for the Ripatti and Palmgren method.

Overall, for all variance-covariance parameters, the mean squared error (data not shown) of the estimates obtained by the proposed version of the EM algorithm is smaller than for the Ripatti and Palmgren approach when $\sigma^2 = 0.2$, while the opposite trend could be seen for $\sigma^2 = 1$.

A natural measure to assess the association between the two random effects is the correlation coefficient. Note that, since it was not used as a parameter in model (29)–(30), it needs to be computed from the estimated values of σ_2, σ_1 and σ_{12}. Figure 2 presents graphically the relative bias for ρ. The absolute relative bias remains around or below 5% for $\rho = 0.9$; for $\rho = 0.5$, it does so for $n_i \geq 20$. The bias generally decreases with increasing N and n_i, and is smaller when σ^2 increases. One can observe that there are substantial differences between the estimates produced by both methods, especially when there is low variability in cluster-specific random effects ($\sigma^2 = 0.2$). In general, the estimates obtained using the method of Ripatti and Palmgren are closer to the true value of the correlation coefficient. One can also conclude that the proposed version of the EM algorithm tends to underestimate the true value of the coefficient. Since the correlation coefficient was not used in the parametric form of model (29)–(30), its model-based standard error was not directly available. Though it could be computed from the estimated errors for σ_2, σ_1 and σ_{12} by using the delta-method, we did not pursue a more detailed analysis of this aspect of the estimation of ρ.

7 Case Study: Analysis of Survival Data in a Breast Cancer Clinical Trial

In this section we will use a proportional hazard model with multivariate random effects to investigate the between-center variation (heterogeneity) in both the baseline risk and the effectiveness of therapy in a multicenter clinical trial. The variation is of interest because it decreases the power to detect clinically important treatment differences. On the other hand, more heterogeneous trials

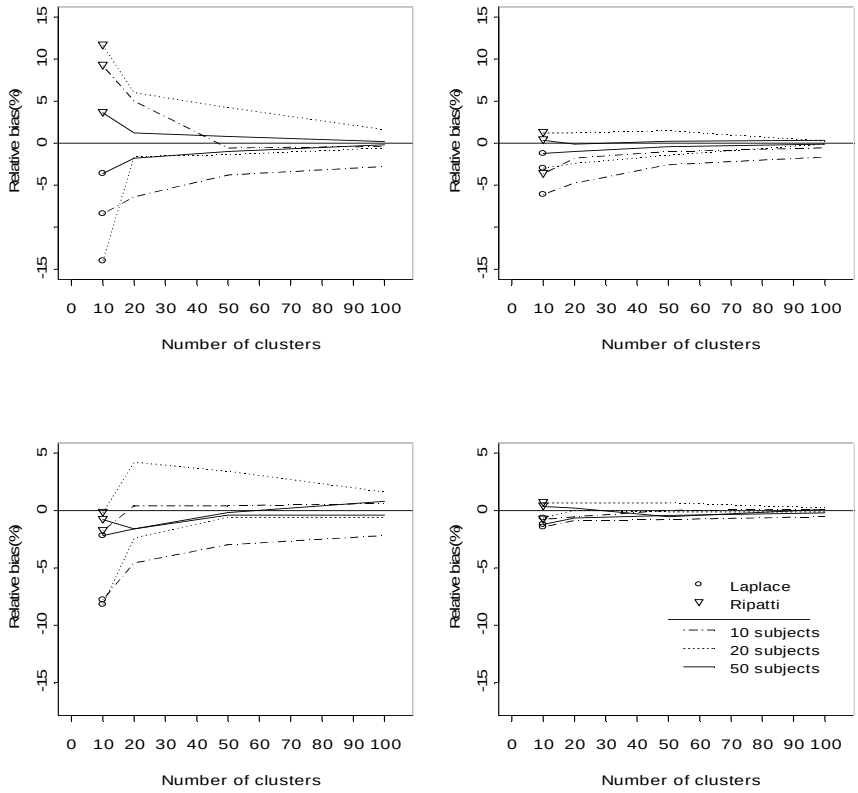


Fig. 2: The relative bias for ρ for 20% censoring. Left column: $\rho = 0.5$; right column: $\rho = 0.9$. Top row: $\sigma_1^2 = 0.2$; bottom row: $\sigma_1^2 = 1$. The legend in the bottom right panel refers to all plots.

lead to more general conclusions as they are based on a wider patient population. Moreover, the differences between centers can be studied to determine whether differences in clinical practice at the center level have an influence on the outcome (Yamaguchi and Ohashi, 2000; Duchateau et al., 2002). Investigation of the heterogeneity is sometimes called "treatment outcome research" (Duchateau et al., 2002).

As an example we will use data on survival time of patients from an European Organization for the Research and Treatment of Cancer (EORTC) early breast cancer clinical trial comparing peri-operative chemotherapy with surgery alone (Clahsen et al., 1996). The trial includes 15 centers,

with the following number of patients per center: 6, 19, 25, 39, 48, 53, 54, 60, 78, 184, 185, 206, 311, 622, 902. Duchateau et al. (2002) used this trial to study the between-center variability in the baseline hazard. To this aim, they applied a shared frailty model with a gamma-distributed frailty to model progression free survival. As a result, they estimated baseline hazard (assumed constant) and the hazard ratio for the surgery-alone treatment to equal 0.07 and 1.16, respectively. The variance of the frailty distribution was estimated to be equal to 0.092.

We will re-analyze the data used by Duchateau et al. (2002), allowing for the variation in both the baseline hazard and the treatment effect. To this aim, we will use the following model:

$$\lambda_{ij}(t_{ij}|\beta, b_{i0}, b_{i1}) \quad = \quad \lambda_0(t_{ij})e^{b_{i0}+x_{ij}(\beta+b_{i1})}\,, \tag{32}$$

where

$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim N_2 \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ & \sigma_1^2 \end{pmatrix} \right\}. \tag{33}$$

A similar model was used by Yamaguchi and Ohashi (2002) in another case study, but with the covariance between the two random effects constrained to 0. They fitted the model using an extension of the penalized partial likelihood approach developed by McGilchrist and Aisbett (1991) and McGilchrist (1993).

The parameter estimates for model (32), obtained by the method of Ripatti and Palmgren (2000) and by the version of the EM algorithm proposed in this paper, are presented in Table 3. The result for both methods are quite comparable. Worth noting are somewhat larger standard errors for the variance-covariance parameters for the Ripatti and Palmgren approach. This is consistent with the results of the simulation study presented in the previous section.

Since Duchateau et al. (2002) used a different model to analyze the data, it is difficult to directly compare their results to those shown in Table 3. Nevertheless, some similarities can be noted. For instance, the estimated value of β presented in the table gives the hazard ratio of 1.17, which is

Table 3: Results for the analysis of the survival data in the breast cancer trial (standard error in parentheses).

Method	β	σ_0^2	σ_1^2	σ_{01}
EM-Laplace	0.160 (0.071)	0.093 (0.034)	0.035 (0.011)	0.022 (0.014)
Ripatti	0.162 (0.072)	0.091 (0.041)	0.036 (0.014)	0.021 (0.017)

very similar to the value of 1.16 obtained by Duchateau et al. (2002). Moreover, the estimated cumulative baseline hazard for model (32) showed a linear trend (data not shown), suggesting a constant baseline hazard equal to 0.067 for the Ripatti and Palmgren method and 0.066 for the EM algorithm. These values are comparable to the value of 0.07 obtained by Duchateau et al. (2002). Finally, the estimated variance of the gamma frailty (0.092) given by Duchateau et al. (2002) implies that, in their analysis, the distribution of the logarithm of the frailty can by approximated by a normal distribution with variance $\psi'(1/0.09) = 0.094$, where $\psi'()$ is the trigamma function (Johnson and Kotz, 1970, p. 181). Thus, the shared gamma-frailty model used by Duchateau et al. (2002) might be approximately equivalent to model (32) without the random treatment effects b_{i1} and with normally-distributed random intercepts b_{i0} with variance 0.094. This value is only slightly higher than the estimates of σ_0^2 given in Table 3.

On the other hand, model (32) provides additional information about the heterogeneity of treatment effects. More specifically, the estimated value of σ_1^2 implies that in 95% of cases the center-specific hazard ratio for treatment should remain in the interval $\exp(0.16 \pm 1.96 \times 0.189)$, i.e., (0.81, 1.70). This range of the variability is thus rather wide. Additionally, the estimates of σ_{01} shown in Table 3 suggest a low correlation (0.37 for the Ripatti and Palmgren method, 0.38 for the EM algorithm) between b_{i0} and b_{i1}.

It is worth noting here that, in the context of "treatment outcome research", the explicit use of random effects in model (32) is of importance, as it allows to quantify the magnitude of between-center heterogeneity in baseline hazards and treatment effects. If one's interest, however, lies only

in, e.g., testing for center effects, alternative methods, not requiring a random effects formulation, can be considered (Gray, 1995).

8 Concluding Remarks

Proportional hazard models with multivariate random effects offer several advantages over univariate shared frailty models (Xue and Brookmeyer, 1996), especially when survival times from the same cluster are negatively associated. The main stumbling block in the use of the former models are estimation methods.

In this paper we have proposed an estimation method based on the EM algorithm. Its main advantage is a lower computational complexity, as compared to the previously developed implementations of the algorithm (Xue and Brookmeyer, 1996; Vaida and Xu, 2000; Ripatti et al., 2002). In the current paper normally-distributed multivariate random effects were considered, but the method might in principle be extended to other types of multivariate distributions. A drawback of the method is the asymptotic nature of the Laplace approximation: to get estimates, one needs a cluster size that cannot be too small. In fact, in our simulation study we did not include the settings with clusters with less than 10 subjects, since for these settings convergence problems were too frequent.

An important issue in the assessment of any estimation method are the statistical properties of the obtained estimates. The consistency of the estimates produced by the EM algorithm has been proven only for the case of the shared frailty model with a univariate, gamma-distributed frailty (Murphy, 1995; Parner, 1998). No formal results are available for other distributions. An empirical study of Ferreira and Garcia (2002) of the EM-algorithm-based estimation method proposed by Nielsen et al. (1992) suggests that the estimates of the variance parameter may be non-consistent when the gamma assumption fails.

Even less is known about the multivariate frailty models. For this reasons we conducted a simulation study, in which we compared our proposal with the non-EM based approach developed by Ripatti and Palmgren (2000). Comparison with other versions of the EM algorithm (Xue and Brookmeyer, 1996; Vaida and Xu, 2000) was not possible due to the numerical complexity and problems with the implementation of these methods. Xue and Brookmeyer (1996) in their paper stated that "although it is computationally feasible for analysis of specific data sets, it is not efficient enough to being considered for computer simulation studies," implying that their approach is fairly computational intensive. Vaida and Xu (2000) reported an MCEM inference approach, where they used Gibbs sampling to draw from the posterior distribution of the random effects where convergence of the algorithm is assessed by visual inspection of the estimates. This is also a limitation if simulations are conducted together with the fact that for large dimensions of the random effects relative to the sample sizes can encountered convergence problems.

In the simulation study both approaches produced on average similar estimates of the variances of random effects. More difference was seen in the estimation of the fixed effects and the covariance/correlation, where the estimates for the method of Ripatti and Palmgren (2000) showed smaller bias. For both methods the bias in the parameter estimates seemed to disappear with the increasing cluster size and (except for the fixed effects) the number of clusters. The empirical variability of the parameter estimates was in general similar for both methods. For the Ripatti and Palmgren method the model-based estimates tended to overestimate the empirical standard error for the fixed-effects parameters (especially when the cluster size was small) and to underestimate the standard error for the variance-covariance parameters. For the proposed version of the EM algorithm the estimates generally underestimated the error, but they were closer to the true value than the estimates for the method of Ripatti and Palmgren. The underestimation was due to the fact, that to reduce numerical complexity, the estimates were computed by inverting only the appropriate sub-matrices of the observed Fisher information matrix. Overall, the mean

squared error of the estimated fixed-effects parameters obtained by the proposed version of the EM algorithm was larger than for the Ripatti and Palmgren approach. For the variance-covariance parameters, the relationship depended on the variability of the random effects: when the variability was small (large), the mean squared error for the proposed version of the EM algorithm was smaller (larger) than for the method of Ripatti and Palmgren. Finally, it is worth mentioning that the computation time needed for the latter method to converge was, in general, shorter. This is not surprising, in view of the linear rate of convergence for the EM algorithm.

The aforementioned simulation results indicate that both the proposed version of the EM algorithm, as well as the method of Ripatti and Palmgren (2000), have some advantages to offer. A more definite evaluation of their merits requires further research.

Acknowledgements The authors would like to thank Terry Therneau for providing the S-Plus functions for the Ripatti and Palmgren approach, the assistance in implementing the software and comments regarding the content of the manuscript. They would also like to thank the Breast Cancer Group of the EORTC for providing the data for the peri-operative chemotherapy trial, and Richard Sylvester and Catherine Legrand from the EORTC for useful comments. Both authors gratefully acknowledge support from Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data".

References

Bleistein, N. and Handelsman, R.A. (1986) Asymptotic Expansions of Integrals. Dover, New York.

Clahsen, P.C., van de Velde, C.J., Julien, J.P., Floiras, J.L., Delozier, T., Mignolet, F.Y., and Sahmoud, T.M. (1996) Improved local control and disease-free survival after preoperative

chemotherapy for early-stage breast cancer. A European Organization for Research and Treatment of Cancer breast cancer cooperative group study. Journal of Clinical Oncology 14, 745–753.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1997) Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B 39, 1–38.

Duchateau, L., Janssen, P., Lindsey, P., Legrand, C., Nguti, R., and Sylvester, R. (2002) The shared frailty model and the power for heterogeneity tests in multicenter trials. Computational Statistics and Data Analysis 40, 603–620.

Evans, M. and Swartz, T. (2000) Approximating Integrals via Monte Carlo and Deterministic Methods. Oxford University Press, Oxford.

Ferreira, A., and Garcia, N.L. (2002) Simulation study for misspecifications on a frailty model. Brazilian Journal of Probability and Statistics 15, 121-134.

Gray, R.J. (1995) Tests for variation over groups in survival data. Journal of the American Statistical Association 90, 198-203.

Johansen, S. (1993) An extension of Cox's regression model. International Statistical Review 51, 258–262.

Johnson, N.L. and Kotz, S. (1970) Continuous Univariate Distributions, Vol. 1. Houghton Mifflin, Boston.

Kass, R.E., Tierney, L., and Kadane J.B. (1990) The validity of posterior expansions based on Laplace's method. Bayesian and likelihood methods in statistics and econometrics (ed. S. Geisser, J.S. Hodges, S.J. Press and A. Zellner). Elsevier Science North Holland, Amsterdam, 473-488.

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

Louis, T.A. (1982) Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society B 44, 190–200.

McGilchrist, C.A. and Aisbett, C.W. (1991) Regression with frailty in survival analysis. Biometrics 47, 461–466.

McGilchrist, C.A. (1993) REML estimation for survival models with frailty. Biometrics 49, 221–225.

McGilchrist, C.A. (1994) Estimation in generalized mixed models. Journal of the Royal Statistical Society B 56, 61–69.

Murray, J.D. (1984) Asymptotic Analysis. Springer-Verlag, New York.

Murphy, S.A. (1995) Asymptotic theory for the frailty model. Annals of Statistics 23, 182–198.

Nielsen, G.G., Gill, R.D., Andersen, P.K., and Sorensen, T.I.A (1992) A counting process approach to maximum likelihood estimation in frailty models. Scandinavian Journal of Statistics 19, 25-43.

Parner, E. (1998) Asymptotic theory for the correlated gamma-frailty model. Annals of Statistics 26, 183–214.

Ripatti, S. and Palmgren, J. (2000) Estimation of multivariate frailty models using penalized partial likelihood. Biometrics 56, 1016–1022.

Ripatti, S., Larsen, K., and Palmgren, J. (2002) Maximum likelihood inference for multivariate frailty models using an automated Monte Carlo EM algorithm. Lifetime Data Analysis 8, 349–360.

Therneau, T. and Grambsch, P.M. (2000) Modeling Survival Data: Extending the Cox model. Springer-Verlag, New York.

Therneau, T., Grambsch, P.M., and Shane Pankratz, V. (2000) Penalized survival models and frailty. Technical Report (June 2000).

Therneau, T. (2003) On mixed effect Cox models, sparse matrices, and modelling data from large pedigree. Technical Report (July 2003).

Vaida, F. (2004) Parameter convergence for EM and MM Algorithms. Statistica Sinica, to appear.

Vaida, F. and Xu, R. (2000) Proportional hazards model with random effects. Statistics in Medicine 19, 3309–3324.

Wong, R. (1989) Asymptotic Approximations of Integrals. Academic Press, San Diego.

Wu, C.-F.J. (1983) On the convergency properties of the EM algorithm. Annals of Statistics 11, 95–103.

Xue, X. (1998) Multivariate survival data under bivariate frailty: an estimating equation approach. Biometrics 54, 1631–1637.

Xue, X. and Brookmeyer, R. (1996) Bivariate frailty model for the analysis of multivariate survival time. Lifetime Data Analysis 2, 277–289.

Xue, X. and Ding, Y. (1999) Assessing heterogeneity and correlation of paired failure times with the bivariate frailty model. Statistics in Medicine 18, 907–918.

Yamaguchi, T. and Ohashi, Y. (1999) Investigating centre effects in a multi-centre clinical trial of superficial bladder cancer. Statistics in Medicine 18, 1961-1971.