Max-Planck-Institut für demografische Forschung
Max Planck Institute for Demographic Research
Konrad-Zuse-Strasse 1 · D-18057 Rostock · GERMANY
Tel +49 (0) 3 81 20 81 - 0; Fax +49 (0) 3 81 20 81 - 202;
http://www.demogr.mpg.de

MPIDR WORKING PAPER WP 2003-018
JUNE 2003

A simulation study of different correlated frailty models and estimation strategies

A. Wienke (wienke@demogr.mpg.de)
K. Arbeev (arbeev@demogr.mpg.de)
I. Locatelli (isabella.locatelli@uni-bocconi.it)
A.I. Yashin (yashin@demogr.mpg.de)

Working papers of the Max Planck Institute for Demographic Research receive only limited review. Views or opinions expressed in working papers are attributable to the authors and do not necessarily reflect those of the Institute.

Summary

Frailty models are becoming more and more popular in the area of multivariate survival analysis. In particular, shared frailty models are often used despite their limitations. To overcome the disadvantages of shared frailty models numerous correlated frailty models were established during the last decade. In the present study we examine correlated frailty models, especially the behavior of the parameter estimates when using different estimation strategies. Three different frailty models are considered: the gamma model and two versions of the lognormal model. The traditional maximum likelihood procedure of parameter estimation in the gamma case with an explicit available likelihood function is compared with maximum likelihood methods based on numerical integration and a Bayesian approach using MCMC methods with the help of a comprehensive simulation study. A strong dependence between the two parameter estimates (variance and correlation of frailties) in the multivariate correlated frailty model is detected and analyzed in detail.

1 Introduction

Frailty models have been used frequently for modeling dependence in multivariate time-to-event data (Clayton, 1978; Oakes, 1982; Yashin et al., 1995; Hougaard, 2000; Wienke et al., 2002). The dependence usually arises because individuals in the same group (family, litter, study center) are related to each other or because of multiple recurrence of an event for the same person. In such cases the traditional proportional hazards model can not be applied. A possible solution to this problem is the use of conditional proportional hazards given the frailty. The variability of lifetimes is formulated as arising from two different sources. The first one is natural variability, which is included in the baseline hazard function, while the second one is explained by the frailty. Lifetimes are conditionally independent given the frailty (as individual random effect), and the frailty term represents unobserved covariates. It is assumed that, given the unobserved frailty, the hazard for each survival time follows a proportional hazards model with the frailty variable and the covariate effect acting multiplicatively on the baseline hazard. Consequently, specification of the baseline hazard and distributional assumptions about the frailty are necessary.

The most common frailty distribution is the gamma distribution. The gamma distribution has been widely applied as a mixture distribution (for example, Clayton, 1978; Vaupel et al., 1979; Oakes, 1982; Yashin and Iachine, 1995; Hougaard, 2000; Wienke et al., 2000, 2001). From a computational and analytical point of view the gamma distribution fits very well to failure data, because it is easy to derive the closed form expressions of survival, density and the hazard function. This is due to the simplicity of the Laplace transform, which is the reason why this distribution has been used in most applications published so far. The simple and explicit available form of the Laplace transform allows for the use of traditional maximum likelihood procedures in parameter estimation.

The second frailty model considered in the present paper is the log-normal model. Again the frailty is acting multiplicatively on the baseline hazard following a log-normal distribution. Especially in multivariate modeling the log-normal approach is much more flexible than the gamma model in creating correlated but different frailties as necessary in the correlated frailty model. Two variants of the log-normal model are analyzed. We assume a normally distributed random variable W to generate frailty as $Z = e^W$. The two variants of the model are given by the restrictions $\mathbf{E}W = 0$ and $\mathbf{E}Z = 1$. Unfortunately, no explicit form of the unconditional likelihood exists. Consequently, estimation strategies based on numerical integration in the maximum likelihood approach are required.

Please note that no biological reason exists which would prefer the use of one frailty distribution over another. All arguments in favor or against a distribution are mathematically based.

Shared frailty models explain correlations within groups (family, litter, or clinic) or for recurrent events from the same individual. However, this approach does have some limitations. First, it forces the unobserved factors to be the same within the cluster, which is not generally reasonable. For example, sometimes it may be inappropriate to assume that both partners in a twin pair share all their unobserved risk factors. Second, the dependence between survival times within the cluster is based on marginal distributions of survival times. To see this, when covariates are present in a proportional hazards model with gamma-distributed frailty, the dependence parameter and the population heterogeneity are confounded (Clayton and Cuzick, 1985), implying that the joint distribution can be identified from the marginal distributions (Hougaard, 1986a). Elbers and Ridder (1982) show that this problem exists for any univariate frailty distribution with a finite mean. Third, in most cases, shared frailty will only induce positive association within the group. However, there are some situations in which survival times for subjects within the same cluster are negatively associated. For example, if animals live in the same litter with a limited food supply, their growth rates are probably negatively associated.

To avoid all these limitations, correlated frailty models are developed for the analysis of multivariate failure time data, in which associated random variables are used to characterize the frailty effect for each cluster. For example, in twin pairs one random variable is assigned for twin 1 and one for twin 2 so that they are no longer be constrained to having a common frailty. These two variables are associated and jointly distributed, therefore, knowing one of them does not necessarily imply the other. Also, these two variables can certainly be negatively associated, which would induce negative association between survival times.

Consequently, correlated frailty models provide not only variance parameters of the frailties as in shared frailty models, they contain additional parameters for modeling the correlation between frailties in each group.

After working for a long time with correlated frailty models of different types, we recognized a strange linkage between the parameter estimates of the variance and the correlation of the frailties. To check whether this dependence is related to the estimation strategy or to the choice of the frailty distribution we use the above mentioned models and three different estimation strategies. First, we perform a traditional maximum likelihood estimation procedure (only possible in the gamma model), second we use a maximum likelihood approach based on numerical integration and finally we utilize MCMC methods from a Bayesian approach.

The paper is organized as follows. In section 2 we introduce the correlated frailty models considered. In section 3 we describe the different estimation strategies. Section 4 deals with the simulation studies and their results. Finally, the paper ends with a discussion of our findings in section 5.

2 Statistical Models

2.1 General bivariate frailty model

Consider some bivariate observations, e.g., the life spans of twins, or age at onset of a disease in spouses, etc. We are assuming that the frailties are acting multiplicatively on the baseline hazard function and that the observations in a pair are conditionally independent given the frailties. Hence, the hazard of individual j $(j = 1, 2)$ in pair i $(i = 1, \ldots, n)$ has the form

$$\mu(t, Z_{ij}, X_{ij}) = Z_{ij}\mu_0(t)e^{\beta X_{ij}}, \tag{1}$$

where t denotes age, X_{ij} a vector of observable covariates, β is a vector of unknown regression coefficients describing the effect of the covariates X_{ij}, $\mu_0(t)$ is some baseline hazard function and Z_{ij} are unobserved (random) effects or frailties. Bivariate frailty models are characterized by the joint distribution of a two-dimensional vector of frailties (Z_{i1}, Z_{i2}). The form of the baseline hazard is important because all methods described below are parametrical. In principle, any parametric formula for a hazard rate is possible (e.g., Gompertz, Gompertz-Makeham, Weibull, exponential, piecewise constant, etc.). The methods reviewed in the following were developed mainly for some specific baseline hazard rate, e.g., exponential or piecewise constant. However, these methods are general and can be modified in order to incorporate any baseline hazard. A vast literature on human mortality suggests using the Gompertz hazard rate to describe the mortality. Correlated frailty models with the Gompertz baseline hazard have been used quite frequently (Yashin et al., 1995; Iachine et al., 1998; Wienke et al., 2001;

among others). For that reason and to save space we investigate only bivariate frailty models with the Gompertz baseline hazard rate:

$$\mu_0(t) = ae^{bt}. \tag{2}$$

Any method in this context is based on likelihood functions. In order to derive a marginal likelihood function, the facilitating assumption of conditional independence of life spans given frailty is always used. Denote by θ the vector of all parameters of the model. Let δ_{ij} be a censoring indicator for an individual j $(j = 1, 2)$ in pair i $(i = 1, \dots, n)$. Indicator δ_{ij} is 1 if the individual has experienced the event of interest and 0 otherwise. According to (1), the conditional survival function of the j-th individual in the i-th pair is

$$S(t|Z_{ij}, X_{ij}) = e^{Z_{ij} H_0(t) e^{\beta X_{ij}}}, \tag{3}$$

where $H_0(t)$ is the cumulative baseline hazard function. Here and in the following S is used as a generic symbol for a survival function. Given (2),

$$H_0(t) = \frac{a}{b}(e^{bt} - 1). \tag{4}$$

The contribution of the j-th individual in the i-th pair of the conditional likelihood is given by

$$L(t_{ij}, \delta_{ij}|Z_{ij}, X_{ij}) = \Big(Z_{ij}\mu_0(t_{ij})e^{\beta X_{ij}}\Big)^{\delta_{ij}} e^{Z_{ij} H_0(t_{ij}) e^{\beta X_{ij}}}, \tag{5}$$

where t_{ij} stands for age at death or the censoring time of the individual. Then, assuming the conditional independence of life spans given frailty and integrating out the random effects, we obtain the marginal likelihood function:

$$\begin{aligned} L(t, \delta|X) = \prod_{i=1}^{n} \iint\limits_{R^2} & \Big(z_{i1}\mu_0(t_{i1})e^{\beta X_{i1}}\Big)^{\delta_{i1}} e^{z_{i1} H_0(t_{i1}) e^{\beta X_{i1}}} \\ & * \Big(z_{i2}\mu_0(t_{i2})e^{\beta X_{i2}}\Big)^{\delta_{i2}} e^{z_{i2} H_0(t_{i2}) e^{\beta X_{i2}}} f_Z(z_{i1}, z_{i2}, \theta)\, dz_{i1}\, dz_{i2}, \end{aligned} \tag{6}$$

where $t = (t_1, \dots, t_n)$, $t_i = (t_{i1}, t_{i2})$, $\delta = (\delta_1, \dots, \delta_n)$, $\delta_i = (\delta_{i1}, \delta_{i2})$, $X = (X_1, \dots, X_n)$, $X_i = (X_{i1}, X_{i2})$ and $f_Z(\cdot, \cdot|\theta)$ is the p.d.f. of the corresponding frailty distribution.

2.2 Gamma model

The gamma distribution (we use notation $\Gamma(k, \lambda)$ for the two parameter distribution with shape parameter k and scale parameter λ) is one of the most popular frailty distributions. Frailty cannot be negative. The gamma distribution is, along with the log-normal distribution, one of the most commonly used distributions to model variables that are necessarily positive. Furthermore, it turns out that the assumption that frailty at birth is gamma-distributed yields some useful mathematical results, including

- Frailty among the survivors at any time t is gamma-distributed with the same value of the shape parameter k as at birth. The value of the second parameter, however, is now given by $\lambda(t) = \lambda + H_0(t)$, where $H_0(t)$ denotes the cumulative baseline hazard function.
- Frailty among those who die at any age t is also gamma-distributed, with the same parameter $\lambda(t)$ as among those surviving to age t but with shape parameter $k + 1$.
- The Laplace transform of a gamma-distributed random variable $Z \sim \Gamma(k, \lambda)$ is of a very simple form: $\mathbf{L}_Z(s) = \mathbf{E}e^{-Zs} = (1 + \frac{s}{\lambda})^{-k}$.

To make sure that the model is identifiable, it makes sense to use the parameter restriction $\mathbf{E}Z = 1$, which results in $k = \lambda$ for the gamma distribution. Denoting the variance of the frailty variable by $\sigma^2 := \frac{1}{\lambda}$, the univariate survival function is represented by

$$S(t) = \mathbf{L}(H_0(t)) = (1 + \sigma^2 H_0(t))^{-\frac{1}{\sigma^2}},$$

where $H_0(t)$ denotes the cumulative baseline hazard function. For more detailed information regarding the univariate gamma-frailty model see Vaupel et al. (1979).

The correlated gamma-frailty model (Yashin and Iachine, 1994; Pickles et al., 1994; Petersen 1998) is developed for the analysis of multivariate failure time data, in which two associated random variables are used to characterize the frailty effect for each cluster. For example, one random variable is assigned for twin 1 and another for twin 2 so that they are no longer constrained to having a common frailty as in the shared frailty model. To be more specific, let k_0, k_1 be some real positive variables. Set $\lambda = k_0 + k_1$ an let Y_0, Y_1, Y_2 be independent gamma-distributed random variables with $Y_0 \sim \Gamma(k_0, \lambda)$, $Y_1 \sim \Gamma(k_1, \lambda)$, $Y_2 \sim \Gamma(k_1, \lambda)$. Consequently,

$$\begin{aligned} Z_1 &= Y_0 + Y_1 \sim \Gamma(k_0 + k_1, \lambda) \sim \Gamma(\lambda, \lambda) \qquad (7) \\ Z_2 &= Y_0 + Y_2 \sim \Gamma(k_0 + k_1, \lambda) \sim \Gamma(\lambda, \lambda) \end{aligned}$$

are the frailties of individual 1 and 2 in a pair. The bivariate survival function of this model is given by

$$S(t_1, t_2) = \begin{cases} S(t_1)^{1-\rho} S(t_2)^{1-\rho} (S(t_1)^{-\sigma^2} + S(t_2)^{-\sigma^2} - 1)^{-\frac{\rho}{\sigma^2}} & \text{if } \sigma^2, \rho > 0 \\ S(t_1)S(t_2) & \text{if } \sigma^2 = 0 \text{ or } \rho = 0 \end{cases} \qquad (8)$$

where $S(t)$ denotes the marginal univariate survival function, assumed to be equal for both partners in a twin pair and $0 \leq \rho \leq 1$ holds. For simplicity, we drop the dependence of the survival functions from observed covariates. Obviously, the shared gamma-frailty model by Clayton (1978) is a special case of (8) when $\rho = 1$. We will refer to model (8) as Model 1.

2.3 Log-normal model

The log-normal model is much more flexible than the gamma model, because it is not based on the additive composition of the two frailties as used in (7). On the other hand, the log-normal distribution does not allow an explicit representation of the likelihood function, which requires more sophisticated estimation strategies. We assume that the two frailties of individuals in a pair are given by

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \text{LogN}\left(\begin{pmatrix} m \\ m \end{pmatrix}, \begin{pmatrix} s^2 & rs^2 \\ rs^2 & s^2 \end{pmatrix} \right), \tag{9}$$

where LogN denotes the (bivariate) log-normal distribution. Here m, s^2 and r denote the mean, variance and correlation of the respective normal distribution. Mean, variance and correlation of the frailties are related to these parameters as follows:

$$\mu = \mathbf{E}Z_{ij} = e^{m+\frac{s^2}{2}} \tag{10}$$

$$\sigma^2 = \mathbf{V}(Z_{ij}) = e^{2m+s^2}(e^{s^2} - 1) \tag{11}$$

$$\rho = \mathbf{corr}(Z_{i1}, Z_{i2}) = \frac{e^{rs^2} - 1}{e^{s^2} - 1}. \tag{12}$$

Two different types of log-normal frailty models arise from two restrictions on the parameters of frailty distribution. First, one can use the restriction $m = 0$. This means that the logarithm of frailty has a mean of zero. In this case a ”standard” individual has the logarithm of hazard rate which is equal to $\log \mu_0(t)$. Any individual in a population has the logarithm of hazard rate distorted by some random variables $W_{ij} = \log Z_{ij}$. This value is added to the ”true” logarithm of hazard rate $\log \mu_0(t)$ to provide the logarithm of hazard rate of the individual. In this interpretation it is natural to assume that the distortions W_{ij} have a normal distribution with mean of zero. Such a model will be called Model 2 throughout the text. Second, following the usual definition of frailty used in demography (Vaupel et al., 1979; Clayton, 1978) one can use $\mu = 1$. It follows from (10)-(12) that in this case

$$m = \mathbf{E} \log Z_{ij} = -\frac{1}{2} s^2 \tag{13}$$

$$s^2 = \mathbf{V}(\log Z_{ij}) = \ln(1 + \sigma^2) \,. \tag{14}$$

In this model a ”standard” individual has the hazard rate $\mu_0(t)$. Individual j in the i-th pair has the hazard rate of a ”standard” individual multiplied by the frailty Z_{ij}. The above restriction on μ means that the average frailty in a population equals 1 (at the beginning of the follow-up). We shall refer to this model as Model 3.

The shared log-normal model was applied in the papers by, for example, McGilchrist and Aisbett (1991) and McGilchrist (1993), while bivariate log-normal models were analyzed by Ripatti and Palmgren (2000, 2002).

3 Estimation strategies

Parameter estimation in the gamma model is straightforward. The frailty term can be integrated out and an explicit representation of the unconditional bivariate survival function exists (8), which can be used to derive the likelihood function.

Unfortunately, in the log-normal model the integrals in (6) have no explicit solution. Consequently, several estimation methods for bivariate log-normal frailty models have been suggested within a non-Bayesian framework. Various modifications of the maximum likelihood procedure are applicable to the bivariate frailty models. Ripatti and Palmgren (2000) derived an estimating algorithm based on the penalized partial likelihood (PPL). Xue and Brookmeyer (1996) suggested a modified EM algorithm for the bivariate log-normal frailty models. Sastry (1997) developed the modified EM algorithm for the multiplicative two-level gamma frailty model. The same method can be applied to the bivariate log-normal frailty models (see Arbeev and Yashin, 2003). Ripatti et al. (2002) present one more method to deal with EM-like algorithms in a bivariate log-normal frailty model.

In the present paper we use numerical integration procedures. Integrals over the univariate and multivariate normal distributions can be approximated in different ways. One possibility is to use Gauss-Hermite quadratures (Naylor and Smith, 1982; Smith et al., 1987). Similar ideas are used in various applications of random effect models in event history analysis (Lillard, 1993; Lillard at al., 1995; Panis and Lillard, 1995; among others). The methods are implemented in the aML software package (aML version 1, see Lillard and Panis, 2000). Both methods were used to estimate parameters of the bivariate log-normal frailty models for both simulated and real data.

Several papers on the application of Bayesian methods to multivariate frailty models exist. The application of a Bayesian approach to the gamma frailty model can be found in Bolstad and Manda (2001). Gibbs' sampling scheme for the bivariate log-normal frailty model with an exponential baseline hazard is given in Xue and Ding (1999). Korsgaard et al. (1998) present Bayesian inference in the log-normal frailty model with semi-parametric hazard.

To apply MCMC methods, we assume that, conditional on explanatory variables and on the entire set of parameters, observations are independent and prior distributions for all parameters are mutually independent. The conditional probability of data given the parameters is

$$\begin{aligned} L(t,\delta,Z|\theta) = \prod_{i=1}^{n} & \Big(Z_{i1}\mu_0(t_{i1})e^{\beta X_{i1}}\Big)^{\delta_{i1}} e^{Z_{i1}H_0(t_{i1})e^{\beta X_{i1}}} \\ & * \Big(Z_{i2}\mu_0(t_{i2})e^{\beta X_{i2}}\Big)^{\delta_{i2}} e^{Z_{i2}H_0(t_{i2})e^{\beta X_{i2}}} f_Z(Z_{i1},Z_{i2}|\theta), \end{aligned} \tag{15}$$

with $t = (t_1,\dots,t_n)$, $t_i = (t_{i1},t_{i2})$, $\delta = (\delta_1,\dots,\delta_n)$, $\delta_i = (\delta_{i1},\delta_{i2})$, $X = (X_1,\dots,X_n)$, $X_i = (X_{i1},X_{i2})$, $Z = (Z_1,\dots,Z_n)$, $Z_i = (Z_{i1},Z_{i2})$ and $f_Z(\cdot,\cdot|\theta)$ is the p.d.f. of the corresponding frailty distribution.

In the Bayesian framework, the parameters as well as frailties are viewed as random variables with some prior distributions. By definition of the model, the prior distribution of frailty is the bivariate log-normal distribution. We assume the following priors for the parameters: uniform priors over the intervals [1e-7, 0.005], [0.05, 0.15] and [-1, 1] for the Gompertz parameters a, b and the correlation ρ , correspondingly; log-normal priors with mean 0.5 and variance 0.25 for the variance σ^2; multivariate normal priors for β. These prior distributions cover the reasonable interval of the parameter values. Given the distribution (15) and the priors, all full conditional distributions of the parameters can be calculated. These full conditional distributions are used in a Gibbs sampling procedure. We perform the Gibbs sampling in WinBUGS software (see Gilks et al., 1994; Spiegelhalter et al., 2000). The results are presented in the next section.

4 Simulations

We estimated Model 1 using a maximization procedure in a Gauss program, Model 2 in aML software and also in Matlab using a numerical integration procedure (Gauss-Hermite quadrature) from a self-written program. We generated data sets with different frailty distributions. First, we used $\sigma^2 = 1$ and $\rho = 0.7$. Second, we used $\sigma^2 = 0.3$ and $\rho = 0.5$. In both cases $a = 0.003$, $b = 0.07$, $\beta = (\beta_1, \beta_2)$, $\beta_1 = 0.1$ and $\beta_2 = 0.2$. The observed covariates were generated as

$$X_{ij1} = \begin{cases} 1 & \text{if } i \leq \frac{n}{2} \\ 0 & \text{if } i > \frac{n}{2} \end{cases} \tag{16}$$

and $X_{ij2} \sim N(0,1)$. We used sample sizes of 500 and 5,000 pairs. We simulated 500 data sets in each case. In Model 2, the same data sets were estimated in aML and Matlab. As aML does not allow estimating such an analysis, for Model 3 only the Matlab program was applied. The results are shown in Tables 1 - 3.

Method	Sample size	a	b	σ^2	ρ	β_1	β_2
	true values	3.00e-3	0.070	0.300	0.500	0.100	-0.200
Gauss	500	3.02e-3	0.070	0.292	0.528	0.100	-0.199
		(3.89e-4)	(0.005)	(0.085)	(0.221)	(0.080)	(0.043)
Gauss	5000	3.00e-3	0.070	0.300	0.498	0.100	-0.200
		(1.23e-4)	(0.001)	(0.026)	(0.065)	(0.026)	(0.013)
	true values	3.00e-3	0.070	1.000	0.700	0.100	-0.200
Gauss	500	2.99e-3	0.070	1.001	0.699	0.107	-0.202
		(4.03e-4)	(0.005)	(0.141)	(0.080)	(0.105)	(0.054)
Gauss	5000	3.01e-3	0.070	1.000	0.700	0.098	-0.199
		(1.34e-4)	(0.002)	(0.044)	(0.023)	(0.034)	(0.017)

Table 1: Bivariate gamma frailty model with Gompertz baseline hazard and two covariates (Model 1): parameters estimated by authors' program in Gauss, simulated data, 500 data sets (means of estimates, sample standard deviations are in parentheses)

All three models show the same pattern of estimations. As expected, the estimations for the larger sample size are far more accurate. The more striking effect is that the same strong correlation exists between estimates of ρ and σ^2, independently of the model and the estimation procedure.

Method	Sample size	a	b	σ^2	ρ	β_1	β_2
	true values	3.00e-3	0.070	0.300	0.500	0.100	-0.200
aML	500	3.02e-3	0.070	0.328	0.547	0.097	-0.206
		(6.59e-4)	(0.008)	(0.247)	(0.299)	(0.088)	(0.045)
Matlab	500	2.97e-3	0.071	0.339	0.522	0.097	-0.207
		(6.07e-4)	(0.007)	(0.243)	(0.264)	(0.088)	(0.044)
aML	5000	2.98e-3	0.070	0.307	0.502	0.102	-0.200
		(2.01e-4)	(0.002)	(0.058)	(0.100)	(0.026)	(0.013)
Matlab	5000	2.98e-3	0.070	0.308	0.503	0.102	-0.200
		(2.00e-4)	(0.002)	(0.058)	(0.099)	(0.026)	(0.013)
	true values	3.00e-3	0.070	1.000	0.700	0.100	-0.200
aML	500	2.81e-3	0.075	1.283	0.689	0.113	-0.211
		(1.06e-3)	0.075	(0.731)	0.689	0.113	(0.059)
Matlab	500	2.65e-3	0.078	1.551	0.667	0.120	-0.221
		(1.09e-3)	(0.020)	(1.348)	(0.172)	(0.127)	(0.072)
aML	5000	3.07e-3	0.069	0.977	0.720	0.098	-0.199
		(3.44e-4)	(0.004)	(0.173)	(0.072)	(0.034)	(0.017)
Matlab	5000	2.98e-3	0.071	1.028	0.703	0.099	-0.201
		(3.56e-4)	(0.004)	(0.192)	(0.071)	(0.034)	(0.017)
MCMC	5000	3.12e-3	0.069	0.981	0.726	0.091	-0.198
		(3.93e-4)	(0.004)	(0.193)	(0.074)	(0.031)	(0.015)

Table 2: Bivariate log-normal frailty model (Model 2): parameters estimated by aML and authors' program in Matlab, simulated data, 500 data sets (MCMC 50 data sets)

Method	Sample size	a	b	σ^2	ρ	β_1	β_2
	true values	3.00e-3	0.070	0.300	0.500	0.100	-0.200
Matlab	500	2.95e-3	0.072	0.350	0.508	0.010	-0.204
		(4.16e-4)	(0.007)	(0.209)	(0.263)	(0.088)	(0.042)
Matlab	5000	3.00e-3	0.070	0.302	0.504	0.099	-0.200
		(1.30e-4)	(0.002)	(0.056)	(0.098)	(0.026)	(0.013)
	true values	3.00e-3	0.070	1.000	0.700	0.100	-0.200
Matlab	500	3.00e-3	0.075	1.323	0.683	0.107	-0.212
		(4.35e-4)	(0.015)	(0.998)	(0.160)	(0.117)	(0.064)
Matlab	5000	3.00e-3	0.070	1.022	0.701	0.099	-0.201
		(1.46e-4)	(0.004)	(0.179)	(0.067)	(0.034)	(0.018)
MCMC	5000	3.02e-3	0.070	1.000	0.713	0.102	-0.199
		(1.30e-4)	(0.003)	(0.134)	(0.058)	(0.034)	(0.015)

Table 3: Bivariate log-normal frailty model (Model 3): parameters estimated by authors' program in Matlab, simulated data, 500 data sets (MCMC 50 data sets)

Model	Method	Sample size	**corr**(ρ, σ^2)	Parameter
1	Gauss	500	-0.412**	$\sigma^2 = 0.3$, $\rho = 0.5$
1	Gauss	5000	-0.405**	
1	Gauss	500	-0.431**	$\sigma^2 = 1$, $\rho = 0.7$
1	Gauss	5000	-0.396**	
2	aMl	500	-0.516**	$\sigma^2 = 0.3$, $\rho = 0.5$
2	Matlab	500	-0.453**	
2	aMl	5000	-0.662**	
2	Matlab	5000	-0.659**	
2	aMl	500	-0.789**	$\sigma^2 = 1$, $\rho = 0.7$
2	Matlab	500	-0.740**	
2	aMl	5000	-0.875**	
2	Matlab	5000	-0.886**	
2	MCMC	5000	-0.921**	
3	Matlab	500	-0.507**	$\sigma^2 = 0.3$, $\rho = 0.5$
3	Matlab	5000	-0.682**	
3	Matlab	500	-0.717**	$\sigma^2 = 1$, $\rho = 0.7$
3	Matlab	5000	-0.862**	
3	MCMC	5000	-0.854**	

Table 4: Bivariate frailty models (Models 1-3): correlation between parameters estimated by authors' program in Matlab and aMl, simulated data, 500 data sets (MCMC 50 data sets). ** indicates significance at the p=0.01 level.

As Bayesian methods proved to be very time-consuming, we generated only 50 data sets with 5,000 pairs each. We run two parallel chains from different starting points and considered the first 4,000 iterations for each chain as a ”burn-in” interval. The quality of convergence was checked by Gelman-Rubin statistics (see Brooks and Gelman, 1998). The simulated values of parameters of random effects have auto-correlation close to unity. In this case convergence is very slow. Altogether 10,000-60,000 iterations per chain were generated after a ”burn-in” interval for each data set. The values of the Gelman-Rubin statistics in this case are quite close to one.

5 Discussion

Because of their simplicity, multivariate frailty models have become very popular over the last decade. A wide range of papers have been published, dealing with the following: different structures of multivariate models (shared vs. correlated frailty models), different distributions of frailty (gamma, log-normal, stable etc.), different assumptions about the parametric structure (parametric vs. semi-parametric models) and different estimation strategies (traditional maximum likelihood procedures, maximum likelihood procedures based on numerical integration, EM algorithm, MCMC methods). After dealing with correlated frailty models for a long time, we recognized a strange correlation between the estimates of variance and the correlation of the frailties. The present study is the first one to analyze this kind of correlation. Thus, the aim of this paper has been to draw attention to this problem, to elaborate possible reasons for this effect and to present (very preliminary) suggestions on how to overcome this problem.

The first main idea was to test whether the correlation of the estimates is dependent of the distribution of the frailty. We used three very popular frailty distributions to answer this question: the gamma distribution and two log-normal distributions (Models 1 - 3).

The second main idea was to test whether the observed effect was caused by the estimation strategy. That is why we used four different estimation strategies: traditional maximum likelihood estimation (using a self-written Gauss code), maximum likelihood estimation based on numerical integration (using routines in aML and a self-written code in Matlab) and MCMC methods in WinBugs.

The results of the simulations are very clear. The observed effect is stable over different frailty distributions and different estimation strategies. Moreover, different choices of parameters and sample sizes did not change the correlation.

High correlation of parameter estimates could be a sign of identifiability problems in the model. Correlated frailty models were investigated in order to overcome the problems of the shared frailty models, which provide only one parameter to model variance and correlation. One idea was to include observed covariates into the models to improve identifiability characteristics. That is why all models were run with and without observed covariates. The results in both cases are very similar, consequently, we dropped the results for models without observed covariates. We decided to use two covariates in our model, a dichotomous one and a continuous one. No effect of the observed covariates could be detected.

The present study focuses on parametric models, which implies parametric specification of the baseline hazard and the distribution of the frailty. In a separate simulation of the correlated gamma-frailty model (without observed covariates) with an unspecified baseline hazard (semi-parametric model) we found a similar correlation between the estimates as in the parametric models (results are not shown).

Regarding identifiability aspects there is one fact that should be kept in mind. Heterogeneity (measured by variance of frailty) and correlation between frailties (implying correlations between lifetimes) are not completely independent in a frailty model based on conditional independence. To see this, assume that the variance of the frailty tends to zero. This implies zero correlation. This conditional independence assumption could be the reason for the correlation linking the estimates of variance (heterogeneity) and correlation between frailties. This would explain why the observed effect is stable over different models and estimation procedures.

The conclusion to draw is that researchers should be very careful and aware of the presented problem in applying correlated frailty models. On the other hand, this study shows that the models perform well and that there is nearly no bias in the parameter estimates. We did not find any correlation between the estimates of the regression coefficients β_1 and β_2. This supports the use of correlated frailty models for obtaining accurate estimates of covariate effects.

More detailed analyses of the properties of correlated frailty models are needed. Starting from the hypothesis that the problems are caused by a conditional independence assumption, it would be extremely useful to extend correlated frailty models to allow for dependence between lifetimes independent of the frailty.

Acknowledgments

The authors wish to thank Susann Backer for help in preparing this paper for publication. The research was partly supported by NIH/NIA grant 7PO1AG08761-09.

References

[1] Arbeev, K.G., and Yashin, A.I. (2003) Bivariate Lognormal Frailty Models: Estimation Methods, Simulation Studies and Application to Danish Twins Data, MPIDR Working Paper Series, to appear

[2] Bolstad, W.M., Manda, S.O. (2001) Investigating Child Mortality in Malawi Using Family and Community Random Effects: A Bayesian Analysis. Journal of the American Statistical Association 96, 12 - 19

[3] Brooks, S.P., Gelman, A. (1998) Alternative methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7, 434 - 455

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

[5] Gilks, W.R., Thomas, A., Spiegelhalter, D.J. (1994) A language and program for complex Bayesian modelling. The Statistician 43; 169 - 178

[6] Hougaard, P. (2000) Analysis of Multivariate Survival Data. Springer New York

[7] Iachine, I.A., Holm, N.V., Harris J.R., Begun, A.Z., Iachina, M.K., Laitinen, M., Kaprio, J., Yashin, A.I. (1998) How heritable is individual susceptibility to death? The results of an analysis of survival data on Danish, Swedish and Finnish twins. Twin Research 1, 196 - 205

[8] Korsgaard, I.R., Madsen, P., Jensen, J. (1998) Bayesian inference in the semiparametric log normal frailty model using Gibbs sampling. Genetics, Selection, Evolution 30, 241 - 256

[9] Lillard, L.A. (1993) Simultaneous equations for hazards: marriage duration and fertility timing. Journal of Econometrics 56, 189 - 217

[10] Lillard, L.A., Brien, M.J., Waite, L.J. (1995) Premarital Cohabitation and Subsequent Marital Dissolution: A Matter of Self-Selection? Demography 32, 437 - 457

[11] Lillard, L.A., Panis, C.W.A. (2000) aML User’s Guide and Reference Manual. Los Angeles: EconWare.

[12] McGilchrist, C.A., Aisbett, C.W. (1991) Regression with Frailty in Survival Analysis.
Biometrics 47, 461 - 466

[13] Naylor, J.C., Smith, A.F.M. (1982) Applications of a Method for the Efficient Computation of Posterior Distribution. Applied Statistics 31, 214 - 225

[14] Oakes, D. (1982) A Concordance Test for Independence in the Presence of Censoring.
Biometrics 38, 451 - 455

[15] Panis, C.W.A., Lillard, L.A. (1995) Child Mortality in Malaysia: Explaining Ethnic Differences and the Recent Decline. Population Studies 49, 463 - 479

[16] Ripatti, S., Palmgren, J. (2000) Estimation of multivariate frailty models using penalized partial likelihood. Biometrics 56, 1016 - 1022

[17] Ripatti, S., Palmgren, J. (2002) Maximum likelihood inference for multivariate frailty models using an automated Monte Carlo EM algorithm. Lifetime Data Analysis 8, 349 - 360

[18] Sastry, N. (1997) A Nested Frailty Model for Survival Data, With an Application to the Study of Child Survival in Northeast Brazil. Journal of the American Statistical Association 92, 426 - 435

[19] Smith, A.F.M., Skene, A.M., Shaw, J.E.H., Naylor, J.C. (1987) Progress with Numerical and Graphical Methods for Practical Bayesian Statistics. The Statistician 36, 75 - 82

[20] Spiegelhalter, D., Thomas, A., and Best, N. (2000) WinBUGS: User manual, http://www.mrc-bsu.cam.ac.uk/bugs

[21] Vaupel, J.W., Manton, K.G., Stallard, E. (1979) The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality. Demography 16, 439 - 454

[22] Wienke, A., Christensen, K., Holm, N., Yashin, A. (2000) Heritability of death from respiratory diseases: an analysis of Danish twin survival data using a correlated frailty model.
In: Medical Infobahn for Europe. A. Hasman et al. (Eds.), IOS Press, Amsterdam, 407 - 411

[23] Wienke, A., Holm, N., Skytthe, A., Yashin A. (2001) The heritability of mortality due to heart diseases: a correlated frailty model applied to Danish twins. Twin Research 4, 266 - 274

[24] Wienke, A., Christensen, K., Skytthe, A., Yashin, A.I. (2002) Genetic analysis of cause of death in a mixture model with bivariate lifetime data. Statistical Modelling 2, 89-102

[25] Xue, X., Brookmeyer, R. (1996) Bivariate frailty model for the analysis of multivariate survival time. Lifetime Data Analysis 2, 277 - 290

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

[27] Yashin, A.I., Iachine, I.A. (1994) Environment determines 50 % of variability in individual frailty: results from Danish twin study. Research Report, Population Studies of Aging, 10, Odense University, Odense

[28] Yashin, A.I., Iachine, I.A. (1995) How long can humans live? Lower bound for biological limit of human longevity calculated from Danish twin data using correlated frailty model. Mechanisms of Ageing and Development 80, 147 - 169

[29] Yashin, A.I., Vaupel, J.W., Iachine, I.A. (1995) Correlated Individual Frailty: An Advantageous Approach to Survival Analysis of Bivariate data. Mathematical Population Studies 5, 145 - 159