Goodness-of-fit tests for the frailty distribution in proportional hazards models with shared frailty

CANDIDA GEERDENS

Center for Statistics, Hasselt University, Agoralaan Building D, B-3590 Diepenbeek, Belgium

GERDA CLAESKENS*

ORSTAT and Leuven Statistics Research Center, KU Leuven, Naamsestraat 69, B-3000 Leuven, Belgium

gerda.claeskens@kuleuven.be

PAUL JANSSEN

Center for Statistics, Hasselt University, Agoralaan Building D, B-3590 Diepenbeek, Belgium

SUMMARY

Frailty models account for the clustering present in event time data. A proportional hazards model with shared frailties expresses the hazard for each subject. Often a one-parameter gamma distribution is assumed for the frailties. In this paper, we construct formal goodness-of-fit tests to test for gamma frailties. We construct a new class of frailty models that extend the gamma frailty model by using certain polynomial expansions that are orthogonal with respect to the gamma density. For this extended family, we obtain an explicit expression for the marginal likelihood of the data. The order selection test is based on finding the best fitting model in such a series of expanded models. A bootstrap is used to obtain $p$-values for the tests. Simulations and data examples illustrate the test’s performance.

Keywords: Frailty model; Gamma distribution; Goodness-of-fit; Order selection test; Orthogonal polynomials.

1. Introduction

In many experimental settings event data arise in groups (clustered data). As an illustration consider the following two examples from veterinary medicine. In mastitis studies, times to infection of udder quarters are observed. Since each udder quarter is separated from the three other quarters, one quarter might be infected with the other quarters being infection free. The grouping is at the cow level; the cluster size is four (Goethals and others, 2009). In artificial insemination programs for dairy cows, times from parturition to first insemination are observed as this is one of the main factors determining the length of the calving interval that lies between 12 and 13 months (Duchateau and others, 2005). For such data the clustering is at the dairy farm level. The cluster size is typically large and varies per cluster. The mastitis and the artificial insemination data are used to illustrate the goodness-of-fit methodology proposed in this paper.

*To whom correspondence should be addressed.

© The Author 2012. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
Frailty models account for the clustering present in grouped event time data. For a subject \( j (j = 1, \ldots, n_s) \) in cluster \( s (s = 1, \ldots, S) \) denote \( Y_{sj} = \min(T_{sj}, C_{sj}) \), where \( T_{sj} \) is the event time for this subject, and \( C_{sj} \) is the censoring time. The indicator \( \delta_{sj} = I(T_{sj} \leq C_{sj}) \) is one for a subject where the event has taken place, while \( \delta_{sj} = 0 \) for a censored observation. Event times are assumed to be independent of the censoring times. A proportional hazards model with shared frailties expresses the hazard for subject \( j \) in cluster \( s \) at time \( t \) by

\[
h_{sj}(t) = h_0(t) \exp(x_{sj}^\top \beta + W_s) = h_0(t)U_s \exp(x_{sj}^\top \beta),
\]

where \( h_0(t) \) is the baseline hazard function, \( x_{sj} \) is the covariate value of subject \( j \) in cluster \( s \), and \( U_s = \exp(W_s) \) is the frailty for cluster \( s \) that is common to all members of that cluster. The frailties \( U_1, \ldots, U_S \) are assumed to be independent and identically distributed copies of a generic frailty \( U \). A commonly used distribution for \( U \) is the one-parameter gamma distribution with a mean equal to 1. Other often used distributions for \( U \) are inverse Gaussian and positive stable distributions. For detailed information about frailty models, see Duchateau and Janssen (2008).

The choice of a particular frailty distribution is, most of the time, based on the mathematical tractability and/or availability of software, rather than on the way it fits the data. It is therefore important that we can test for the appropriateness of the distribution. Given that frailty models have three different components (the baseline hazard; the linear form of the covariates and the exponential impact of the covariates on the hazard; the frailty distribution), the goodness-of-fit problem for frailty models is a difficult but challenging topic. We address, for frailty models with a parametric baseline hazard, testing the null hypothesis of a gamma frailty distribution. In Section 7, we further discuss in detail these challenges and give guidelines for further extensions.

We construct tests that are non-parametric in spirit and that are in the style of the order selection tests for the normality of random effects in linear mixed models (Claeskens and Hart, 2009), using ideas of flexible modeling of random effects by series expansions inspired by Zhang and Davidian (2001). The estimators and test statistics in those papers were constructed for testing for normality of the random effect distribution; it is therefore natural to consider Hermite expansions. Indeed Hermite polynomials possess orthogonality properties with respect to the normal distribution. To test for the gamma frailty distribution, other choices of basis expansions should be considered. In particular, using polynomials that are orthogonal with respect to the gamma density, we define a new class of frailty densities. We show that each continuous density on the positive half-line can be approximated using members of this new class of “extended gamma densities”. In that sense the proposed class contains any alternative to the gamma density. For this extended family we obtain an explicit expression for the marginal likelihood of the data.

Non-parametric order selection tests (Eubank and Hart, 1992; Hart, 1997) are naturally phrased for estimators defined via series expansions. A data-driven model selection method such as an adapted version of Akaike’s information criterion (AIC) (Akaike, 1973) is used to find the truncation point of the series. The test rejects the null hypothesis when at least one term is selected. Aerts and others (1999, 2000) obtained the asymptotic distribution of the order selection tests for testing parametric hypotheses in likelihood regression models with independent observations. Our work extends some of these ideas to gamma frailty models and is more in line with the order selection tests in linear mixed models (Claeskens and Hart, 2009).

Goodness-of-fit for frailty models needs further exploration (see Duchateau and Janssen, 2008, p. 197). A pioneering paper is Shih and Louis (1995). To see whether the frailty term follows a gamma density, they develop, for frailty models with a parametric baseline, a graphical test based on the posterior expectation of the frailties given the observable data over time. They show that, for the gamma frailty model, this posterior expectation is constant over time; extensions of this test are given in Glidden (1999). For an unspecified baseline Cui and Sun (2004) study the supremum over time of the Shih and Louis test process.
For the simple case of bivariate data, goodness-of-fit tests based on copulas are available. Relevant references include Shih (1998), Wang and Wells (2000), and Andersen and others (2005). For quadruple data Massonnet and others (2009) use a pseudo-likelihood ratio test to select an appropriate copula from the power variance family (see Duchateau and Janssen, 2008).

2. A CLASS OF EXTENDED GAMMA FRAILTY DENSITIES

For identifiability reasons (see below), we standardize the frailties such that these have expectation equal to 1. Define \( U_m = \tilde{U}_m / E(\tilde{U}_m) \), where \( \tilde{U}_m \) is the unstandardized frailty for the model indexed by \( m \). Instead of assuming that the frailties have a one-parameter gamma density we assume that the frailties \( \tilde{U}_m \) have a density of form

\[
\tilde{f}_{U_m}(u) = \frac{f_U(u)}{c(d_{(m)})} \left( \sum_{j=0}^{m} d_j v_j(u) \right)^2,
\]

(2.1)

where \( f_U \) is the one-parameter gamma density given by \( f_U(u) = u^{1/\theta - 1} \exp(-u/\theta) \theta^{-1}/\Gamma(1/\theta) \), \( v_j \) are polynomials orthonormal to \( f_U \), defined in Lemma 2.1, and \( c(d_{(m)}) = d_{(m)}^T d_{(m)} = \sum_{j=0}^{m} d_j^2 \) is a normalization constant, with \( d_0 = 1 \) and \( d_{(m)} = (d_0, \ldots, d_m)^T \). The density \( \tilde{f}_{U_m} \) depends on an unknown parameter value \( \theta \) that will be estimated. To keep the notation simple, we omit the dependence on \( \theta \). Note that \( m = 0 \) corresponds to the regular one-parameter gamma density family and that \( U_0 = U \). By fitting models with a different series cutoff value \( m \) using maximum likelihood estimation in each such model and using a model selection method to determine the appropriate value of \( m \), one can test whether or not the null model describes the data accurately.

Similar to the generalized Laguerre polynomials that are orthogonal with respect to the weight function \( u^\alpha \exp(-u) \), it follows via an application of the Gram–Schmidt orthogonalization procedure (the proof is straightforward and provided in Appendix A of the supplementary materials available at Biostatistics online) that the following polynomial functions \( v_n \) are orthonormal with respect to the density \( f_U \).

**Lemma 2.1** For \( n = 0, 1, 2, \ldots \), the polynomial functions \( p_n : \mathbb{R}^+ \to \mathbb{R} : u \to p_n(u) \) with

\[
p_n(u) = (-\theta)^n u^{-(1/\theta - 1)} \exp^{1/\theta} \frac{d^n}{du^n} [e^{-u/\theta} u^{1/\theta + n - 1}] = \sum_{i=0}^{n} (-\theta)^n \binom{n}{i} \Gamma(1/\theta + n) \Gamma(1/\theta + i) u^i
\]

are orthogonal with respect to the inner product \( \langle g_1, g_2 \rangle = \int_0^\infty g_1(u) g_2(u) f_U(u) \, du \), where \( f_U \) is the one-parameter gamma density function. The set of functions \( \{v_n; n = 0, 1, 2, \ldots\} \) is orthonormal, where \( v_n(u) = p_n(u) / \|p_n\|^{1/2} \) with \( \|p_n\| = n! \theta^{2n} \Gamma(1/\theta + n) / \Gamma(1/\theta) \).

Concretely, \( v_0(u) = 1 \), \( v_1(u) = (u - 1) \theta^{-1/2} \), \( v_2(u) = (u^2 - u(2 + 2 \theta) + \theta + 1) / (\theta(2 + 2 \theta)^{1/2}) \), \( v_3(u) = (u^3 - 3(1 + 2 \theta)u^2 + 3(1 + 3 \theta + 2 \theta^2)u - (1 + 3 \theta + 2 \theta^2)) / (6 \theta^3(1 + 3 \theta + 2 \theta^2))^{1/2} \). A graphical representation is given in Figure 1.

Since \( \int_0^\infty e^{-u} f_U(u) \, du < \infty \) for any \( 0 < c < 1/\theta \), the system of polynomials \( \{v_n; n = 0, 1, \ldots\} \) is closed with respect to continuous functions \( g \) on \( (0, \infty) \) that satisfy

\[
\int_0^\infty g^2(u) f_U(u) \, du < \infty.
\]

(2.2)

From this it follows that (Nikiforov and Uvarov, 1988, Section 8) the polynomials are complete in the sense that \( \lim_{m \to \infty} \int_0^\infty \left\{ g(u) - \sum_{n=0}^{m} c_n v_n(u) \right\}^2 f_U(u) \, du = 0 \) for all continuous functions \( g \) on
Fig. 1. Polynomials orthonormal to the one-parameter gamma density function with \( \theta = 1 \). Polynomial \( v_0 \) as a dot-dashed line, \( v_1 \) dotted, \( v_2 \) dashed, and \( v_3 \) as a solid line.

(0, \( \infty \)) that satisfy (2.2). Take \( \tilde{f} \) any continuous density function on (0, \( \infty \)), and define \( g(u) = (\tilde{f}(u)/f_U(u))^{1/2} \). It is readily verified that (2.2) holds and hence the series expansion (2.1) is able to describe any continuous density on the positive real line, where \( c_n = \int_0^\infty (\tilde{f}(u)f_U(u))^{1/2} v_n(u) \, du \). Moreover, if \( \int_0^\infty (g'(u))^2 u f_U(u) \, du < \infty \), then the series expansion \( g(u) = \sum_{n=0}^\infty c_n v_n(u) \) converges uniformly on every interval \([u_1, u_2] \subset (0, \infty)\) (Nikiforov and Uvarov, 1988, p. 59, Theorem 1). On the basis of this finding it is important to highlight that the proposed extended gamma frailty density offers an excellent tool of modeling the unobserved heterogeneity since the family is able to describe any continuous density on the positive half-line.

We define the coefficients \( a_{ij} = (-\theta)^{-i-j} (i+j) \Gamma(1/\theta + j)/\Gamma(1/\theta + i) \), \( a_{ij}^* = a_{ij}/\|p_j\|^{1/2} \), \( b_{ij}^* = a_{ij}^* + 2 \sum_{k+l=i,k+l<j} a_{ik}^* a_{lj}^* \), and \( c_{ij}^* = 2 \sum_{k+l=i+1,k+l<j} a_{ik}^* a_{lj}^* \). The proof of Lemma 2.2 is straightforward (see Appendix A of the supplementary material available at Biostatistics online).

**Lemma 2.2** For \( \tilde{U}_m \) a random variable with density \( f_{\tilde{U}_m} \) as in (2.1),

\[
E_m = E(\tilde{U}_m) = \left[ \sum_{j=0}^{m} d_j^2 \left\{ \sum_{i=0}^{j} b_{ij}^* \theta^{2i+1} \Gamma(2i + 1/\theta + 1) + \sum_{i=0}^{j-1} c_{ij}^* \theta^{2i+2} \Gamma(2i + 1/\theta + 2) \right\} \right. \\
+ \left. 2 \sum_{j=0}^{m-1} \sum_{k=j+1}^{m} d_j d_k \left\{ \sum_{i=0}^{j} \sum_{l=0}^{k} a_{ij}^* a_{lk}^* \theta^{i+l+1} \Gamma(i + l + 1/\theta + 1) \right\} \right] \frac{1}{c(d(m)) \Gamma(1/\theta)}. 
\]

Since \( E(\tilde{U}_m) \) depends on \( d(m) \) and on the value of \( m \), and because the frailties act in a multiplicative way on the baseline hazard in model (1.1), a standardization of the frailties to have a mean equal to 1 guarantees that model (1.1) is well defined. Indeed, otherwise the baseline hazard would change from one model to
another. It readily follows that the density of the standardized frailty random variable \( U_m = \tilde{U}_m / E(\tilde{U}_m) \), where \( \tilde{U}_m \) has density \( f_{\tilde{U}_m}(u) = f_U(u E_m) |E_m/c(d_m)| \left\{ \sum_{j=0}^m d_j v_j(u E_m) \right\}^2 \).

3. The marginal log-likelihood

A nice feature of frailty models with extended gamma frailty densities is that the marginal likelihood of the data has a closed form. With \( H_0 \) the cumulative baseline hazard and \( x_{ij} \) the covariate value of subject \( j \) in cluster \( s \), the following quantities appear in the marginal likelihood: for cluster \( s = 1, \ldots, S \) with size \( n_s \), \( A_s = \prod_{j=1}^{n_s} (h_0(y_{ij}) \exp(x_{ij}^T \beta))^y_j \), \( B_s = \sum_{j=1}^{n_s} H_0(y_{ij}) \exp(x_{ij}^T \beta) \), \( D_s = \sum_{j=1}^{n_s} \delta_{sj} \). We use \( \xi \) as notation for the parameter vector used to model the baseline hazard, and use \( \zeta_m \) as notation for the vector \((\xi, \beta, \theta, d(m))\). For cluster \( s \), the likelihood of the data conditional on the frailties is \( L_s(\xi, \beta | U_{m,s} = u_s) = u_s^{D_s} A_s \exp(-u_s B_s) \).

**Theorem 3.1** The marginal log-likelihood of the data in the extended gamma frailty model with frailty density \( f_{U_m} \) is, with the above notation, given by

\[
\ell_{m,\text{marg}}(\zeta_m) = \sum_{s=1}^S \log \int_0^\infty L_s(\xi, \beta | U_{m,s} = u_s) f_{U_m}(u_s) du_s
\]

\[
= \sum_{s=1}^S \log \left( \frac{A_s E_m^{1/\theta}}{c(d_m) \theta^{1/\theta} \Gamma(1/\theta)} \left\{ \sum_{j=0}^m d_j \right\} \left\{ \sum_{i=0}^{j-1} c_i^* E_i^{2i+1} \Gamma(2i + D_s + 1/\theta) \right\}\left(\frac{b_{ij}^* E_i^{2i+1} \Gamma(2i + D_s + 1/\theta + 1)}{(B_s + E_m/\theta)^{2i+D_s+1/\theta+1}}\right)\right)
\]

\[
+ 2 \sum_{j=0}^{m-1} \sum_{k=j+1}^m d_j d_k \sum_{i=0}^j \sum_{l=0}^k a_i^* a_k^* E_i^{l+i+1} \Gamma(i + l + D_s + 1/\theta) \left(\frac{b_{ij}^* E_i^{2i+1} \Gamma(i + l + D_s + 1/\theta)}{(B_s + E_m/\theta)^{i+l+D_s+1/\theta}}\right).
\]

The proof of Theorem 3.1 is given in Appendix A of the supplementary material available at Biostatistics online.

4. Order selection tests for gamma frailties

4.1 Null and alternative hypotheses

In model (1.1), we wish to test the null hypothesis that the frailties come from a one-parameter gamma density, with a mean equal to 1 and variance given by the parameter \( \theta \),

\[
\mathcal{H}_0 : U \sim f_U(u) = \frac{u^{1/\theta - 1} \exp(-u/\theta)}{\theta^{1/\theta} \Gamma(1/\theta)}, \quad \text{for some value } \theta > 0.
\] (4.1)

The alternative hypothesis \( \mathcal{H}_a \) states that the frailty follows a distribution different from the gamma distribution. Using the extended gamma family, fitting models with a different series cutoff value \( m = 0, 1, \ldots, M_N \) (with \( N = \sum_{s=1}^S n_s \)), the hypotheses are rephrased as

\( \mathcal{H}_0 : m = 0 \), which is equivalent with: for all \( j = 1, \ldots, M_N : d_j = 0 \).

\( \mathcal{H}_a : m > 0 \), which is equivalent with: there exists a \( j \in \{1, \ldots, M_N\} : d_j \neq 0 \).
Different values of $m = 1, 2, \ldots, M_N$ lead to different extensions of the gamma density. Models with a large value of the order $m$ contain the models with a smaller value of $m$ as special cases. In other words, a nested model sequence is constructed by letting $m$ grow. Note that a natural upper bound for $M_N$ is $N$ minus the number of parameters different from $d_j$, $j = 1, \ldots, M_N$. In practical settings $M_N$ is chosen fixed and not too large, since deviations from $\mathcal{H}_0$ will typically be detected for fixed small values of $M_N$.

4.2 Likelihood ratio order selection tests

For likelihood-based models, Aerts and others (1999) defined the order selection (OS) statistic, rephrased to this setting, by

$$T_{N,OS} = \max_{1 \leq m \leq M_N} 2 \frac{\ell_{m,\text{marg}}(\hat{\xi}_m) - \ell_{0,\text{marg}}(\hat{\xi}_0)}{m},$$

where $\xi_m$ is the parameter vector of the model using density $f_{U_m}$, which is estimated under the alternative models by the maximum likelihood estimator $\hat{\xi}_m$, while $\hat{\xi}_0$ is the estimator under the null model with a gamma frailty effect. The length of $\xi_m$ is denoted by $q_m$.

Expanding the extended gamma frailty model when $m$ increases with one unit implies adding one polynomial $v_{m+1}$ to the series, thus adding $d_{m+1}$ to the vector of coefficients. The denominator of $T_{N,OS}$ gives the difference in the number of parameters of model $m$ as compared to the null model, $q_m - q_0 = m$.

The omnibus nature of the test becomes clear. The test is not a single likelihood ratio test, but a maximum of weighted likelihood ratio statistics. By taking a maximum, the statistic $T_{N,OS}$ combines several separate likelihood ratio statistics and is a way of handling the multiple testing issues. The weights take the complexity of the models into account, and down-weight large models. The use of a non-parametric series expansion avoids the specification of a parametric alternative model to the hypothesized null model.

The name “order selection” becomes clear by rewriting the test that rejects the null hypothesis when $T_{N,OS}$ is larger than a critical value $C_\alpha$ at significance level $\alpha$ in an equivalent form in terms of a modified version of AIC (Akaike, 1973) to select the order $m$. When modifying the traditional AIC by changing the penalty constant $2$ to the value $C_\alpha, \text{AIC}_{C_\alpha}(m) = 2\ell_{m,\text{marg}}(\hat{\xi}_m) - C_\alpha q_m$, it is immediately clear that rejecting $\mathcal{H}_0$ when $T_{N,OS} > C_\alpha$ is equivalent to rejecting $\mathcal{H}_0$ when $m = \arg \max_{m=0,\ldots,M_N} \text{AIC}_{C_\alpha}(m) > 0$. To obtain the critical value $C_\alpha$, information on the (asymptotic) distribution of $T_{N,OS}$ is required. Aerts and others (1999) show, in the context of parametric likelihood models, that $T_{OS} = \max_{m \geq 1} (V_m/m)$, where $V_m = \sum_{j=1}^{m} Z_j^2$ with $Z_1, Z_2, \ldots$ independent $N(0, 1)$ distributed random variables, is the limiting form of $T_{N,OS}$ for $N \to \infty$.

Our numerical study (see Figure B.1 in the supplementary material available at Biostatistics online) has revealed that the dependence of the orthogonal polynomials on the value of $\theta$ (see Lemma 2.1) is observable in the finite sample null distribution of the test statistic. We further observed slow convergence. For these reasons, we suggest a bootstrap resampling scheme to get more adequate results. For bootstrap results to be valid for testing, bootstrap data are generated under the null hypothesis. The parameter estimates obtained under the assumption of a one-parameter gamma frailty distribution are used to generate new event times and censoring times. More details are given in Section 5.

5. A numerical performance study

To illustrate the performance of the proposed methodology, we use a simulation study. The simulation settings are along the lines of the extensive simulation on the performance of the gamma frailty model in Section 5.2.3 of Duchateau and Janssen (2008). R code is provided in Appendix C of the supplementary material available at Biostatistics online.

Event times. The $T_{sj}$’s (in years) are generated using the shared frailty model (1.1). We take a Weibull baseline hazard with $\xi = (\lambda, \rho) = (0.22, 1)$ (exponential) and, for a binary covariate, $\beta = \log(1.3)$, i.e. a
hazard rate of 1.3. The frailty densities we use are the following:

(i) \( f_1 \) is one-parameter gamma with \( \theta = 0.3 \);
(ii) \( f_2 \) is inverse Gaussian, i.e. \( f_2(u) = (\alpha/2\pi)^{1/2}u^{-3/2} \exp((-\alpha/2u\mu^2)(u - \mu)^2) \) with \( \mu = 1 \) and \( \alpha = 10/3 \);
(iii) \( f_3 \) is inverse Gaussian with \( \mu = 2 \) and \( \alpha = 5 \);
(iv) \( f_4 \) is positive stable, i.e. \( f_3(u) = (1/\pi)\sum_{k=1}^{\infty} (-1)^{k+1}(\Gamma(kv+1)/k!)(u^{-kv-1}\sin(vk\pi)) \) with \( v = 0.85 \).

The frailty distributions in (ii)–(iv) serve as alternative frailty distributions. The variance of an inverse Gaussian density is \( \mu^3/\alpha \) (0.3 for \( f_2 \) and 1.6 for \( f_3 \)); therefore \( f_2 \) is close to \( f_1 \), \( f_3 \) is away from \( f_1 \) in a moderate way, and \( f_4 \) is far away from \( f_1 \).

**Censoring times** \( C_{sj} \). Think about a trial where patients enter the study in a uniform way over an accrual period of 5 years and a follow-up period of 3 years. The censoring time (time at risk) for a subject thus consists of the time at risk before the end of the accrual period plus the follow-up time.

**Number of clusters.** \( S = 150 \) or \( S = 300 \). **Number of observations per cluster.** \( n_s \equiv n = 4 \). We further assume that we observe the minimum of the event time and the time at risk, \( Y_{sj} = \min(T_{sj}, C_{sj}) \). For simulating under the gamma frailty model, this resulted in about 33% censored observations. For the inverse Gaussian and positive stable frailty distributions, the percentages of censored observations are approximately 32 (\( f_2 \)), 15 (\( f_3 \)), and 1.4 for \( f_4 \). For a trial with uniform censoring times it is natural that the percentage of censored observations decreases for frailties that induce higher heterogeneity. Equal censoring percentages can be obtained by adapting the censoring mechanism used in the simulation setting.

Then, we of course have an intake mechanism that is different from uniform staggered entry.

The concrete settings we consider for \((S,n,\text{frailty})\) (see Table 1) are \((150, 4, f_1)\), \((300, 4, f_1)\), \((150, 4, f_2)\), \((300, 4, f_2)\), \((150, 4, f_3)\), \((300, 4, f_3)\), \((150, 4, f_4)\), \((300, 4, f_4)\). To compute the \( p \)-values in the simulation study we use a parametric bootstrap algorithm (Massonnet and others, 2006).

Since the maximization of the likelihood is numerically difficult and time-consuming, especially when the value of \( m \) is large, which combined with a bootstrap algorithm costs even more time, we took the number of simulated datasets \( R = 200 \) and \( B = 150 \). While for a single data example the accuracy can be taken higher, this was not feasible in a simulation study.

Table 1 lists the percentage of times that the null hypothesis is rejected.

<table>
<thead>
<tr>
<th>((S, n, \text{frailty}))</th>
<th>( \alpha = 0.05 )</th>
<th>( \alpha = 0.10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(150, 4, ( f_1 ))</td>
<td>0.09</td>
<td>0.14</td>
</tr>
<tr>
<td>(300, 4, ( f_1 ))</td>
<td>0.03</td>
<td>0.09</td>
</tr>
<tr>
<td>(150, 4, ( f_2 ))</td>
<td>0.10</td>
<td>0.15</td>
</tr>
<tr>
<td>(300, 4, ( f_2 ))</td>
<td>0.08</td>
<td>0.15</td>
</tr>
<tr>
<td>(150, 4, ( f_3 ))</td>
<td>0.14</td>
<td>0.22</td>
</tr>
<tr>
<td>(300, 4, ( f_3 ))</td>
<td>0.20</td>
<td>0.28</td>
</tr>
<tr>
<td>(150, 4, ( f_4 ))</td>
<td>0.91</td>
<td>0.92</td>
</tr>
<tr>
<td>(300, 4, ( f_4 ))</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Column 1: simulation setting; Column 2: percentage of simulated datasets for which the null hypothesis is rejected at level \( \alpha = 0.05 \); Column 3: percentage of simulated datasets for which the null hypothesis is rejected at level \( \alpha = 0.10 \). The first two rows correspond to the null hypothesis (simulated significance levels).
Algorithm 1 Given one of the six concrete settings

**Step 1.** Fit the log-likelihood in Theorem 3.1 with $m = 0, \ldots, M_N$ and obtain the actual value of the order selection test statistic: $T_{act}$ (given that a deviation from $H_0$ is typically detected for a fixed small value of $M_N$ we take $M_N = 3$; the latter also to limit computation time). Let $\hat{\lambda}_0, \hat{\rho}_0, \hat{\beta}_0$ and $\hat{\theta}_0$ denote the parameter estimates under the null hypothesis ($m = 0$).

**Step 2.** Generate $B$ resamples in the following way:

**Step 2.1.** Sample $u_1^*, \ldots, u_N^*$ from a one-parameter gamma distribution with parameter $\hat{\theta}_0$.

**Step 2.2.** For $j = 1, \ldots, n_s$; $s = 1, \ldots, S$, generate event times $T_{sj}^*$ from the estimated survival function $\hat{S}_{sj}(t) = \{\exp(-\hat{H}_0(t))\}^{u_{sj}^*\exp(u_{sj}^*/\hat{\beta}_0)}$ with $\hat{H}_0(t)$ the estimated cumulative Weibull baseline hazard.

**Step 2.3.** If $\delta_{sj} = 0$ set $C_{sj}^* = Y_{sj}$; if $\delta_{sj} = 1$ generate $C_{sj}^*$ from a uniform distribution with the same accrual and follow up period as for the original data.

**Step 2.4.** Set $Y_{sj}^* = \min(T_{sj}^*, C_{sj}^*)$ with $\delta_{sj}^* = 1$ if $Y_{sj}^* = T_{sj}^*$; and $\delta_{sj}^* = 0$ otherwise.

**Step 2.5.** Obtain the bootstrap version of the order selection test statistic: $T_{act}^*$.

**Step 3.** Obtain the bootstrap version of $P_{H_0}(T > T_{act})$, i.e., $p^* = \#\{b : T_b^* > T_{act}\}/B$.

Table 1 shows that the positive stable alternative model ($f_4$) gives the highest simulated rejection probability. For the inverse Gaussian alternative model ($f_3$) the test is also able to pick up the deviation from the null hypothesis model, but with less pronounced simulated rejection probability. For the inverse Gaussian alternative model defined as $f_3$, a large number of clusters is needed to detect the deviation from the null hypothesis. Intuition for the simulated rejection probabilities in Table 1 is obtained from the information that is contained in Figure 2.

In the left panel, we generated data from a shared positive stable frailty model ($f_4$) with Weibull baseline and one binary covariate. Given the generated data, we fit a shared gamma frailty model; this gives $\hat{\theta} = 0.20$ (se = $0.21 \cdot 10^{-2}$) for the estimated scale parameter of the gamma frailty density. This gamma density is depicted together with the positive stable frailty used to generate the data. The estimated gamma density is in some sense the best possible “null” frailty for the given dataset. It is clear from the picture that the “best possible null frailty” is far from the “real” positive stable frailty; this results in the high values for the simulated rejection probabilities. In the right panel, we do the same now, replacing the positive stable frailty ($f_3$) by an inverse Gaussian frailty ($f_3$). For this situation the “best possible null frailty” ($\hat{\theta} = 0.29$, se = $0.27 \cdot 10^{-2}$) is closer to the “real” inverse Gaussian frailty. In this case, the test is still able to detect the deviation from the null hypothesis model but, compared with the positive stable alternative model, with lower simulated rejection probability. As expected, the simulated rejection probabilities of the test increases with an increased sample size for data generated under the alternative hypothesis.

6. Two illustrative examples

We use the proposed order selection statistic to test the null hypothesis that the frailties come from a one-parameter gamma density. First, we fit the log-likelihood in Theorem 3.1 with $m = 0, \ldots, 5$ and obtain the actual value of the order selection test statistic. To obtain the bootstrapped $p$-value, we follow steps 2 and 3 of the algorithm described in Section 5, without the assumption that the censoring times come from a uniform distribution. Rather, we use a non-parametric Kaplan–Meier estimator of the censoring distribution. Particularly, we replace step 2.3 by

**Step 2.3(bis).** If $\delta_{sj} = 0$, set $C_{sj}^* = Y_{sj}$; if $\delta_{sj} = 1$, generate $C_{sj}^*$ from the conditional censoring distribution given that $C_{sj} > Y_{sj}$, namely $\{\hat{G}(t) - \hat{G}(Y_{sj})\}/\{1 - \hat{G}(Y_{sj})\}$, with $\hat{G}$ the Kaplan–Meier estimate of the censoring distribution (assume that $G$ is independent of the covariate).
6.1 Analysis of the mastitis data

Mastitis, an infection of the udder, is a disease in the dairy cow sector with an important economic impact. A cluster consists of the four udder quarters of a cow. In this example 100 cows are followed up for infections. We consider parity as the single (binary) covariate. The parity of a cow is the number of calvings that the cow has already experienced. Parity is often converted into a binary covariate, grouping all the cows with more than one calving in the group of multiparous cows (parity = 0) compared with the group of primiparous cows, cows with only one calving (parity = 1). Each data line contains the cow identification number, the minimum of the infection time (in days) and the censoring time for each of the four udder quarters, the corresponding censoring indicators and the parity; e.g. for the first cow the data information is: {1, (308.5, 308.5, 308.5, 308.5), (0, 0, 0, 0), 1}. We assume a Weibull baseline hazard, i.e. $\xi = (\lambda, \rho)$.

When taking $M_N = 5$ the value of the order selection test statistic $T_{N,\text{OS}}$ is 4.82 and the bootstrapped $p$-value (we use a bootstrap algorithm that takes into account that, for the mastitis data, the censoring is at the cow level; we use 300 resamples) equals 0.03. We therefore reject the null hypothesis. The value of $M_N$ is not that important; also for, e.g. $M_N = 3$ the same conclusion holds. For the given dataset, the extended gamma frailty density with $m = 2$ was preferred: with $C_{0.05}$ estimated to be 4.15, AIC$_{C_{0.05}}$ for the null model equals $-841.31$ while $m = 2$ corresponds to an AIC$_{C_{0.05}}$ of $-839.98$. The parameter estimates of the chosen density are given by $\hat{\lambda} = 8.31 \cdot 10^{-5}$ (se = $4.14 \cdot 10^{-5}$), $\hat{\rho} = 2.05$ (se = 0.11), $\hat{\beta} = 0.34$ (se = 0.29),
Fig. 3. Graphical representation of the estimated null density of the log-frailty, \( \log(U) \) (solid line) and the estimated selected density of \( \log(U) \) (dashed line) for both the mastitis (left panel) and the insemination data (right panel).

\[ \hat{\theta} = 0.92 \text{ (se = 0.25)}, \hat{d}_1 = 0.09 \text{ (se = 0.09)}, \text{ and } \hat{d}_2 = 0.48 \text{ (se = 0.16)}. \] Figure 3 shows the estimated null density together with the estimated preferred frailty density.

For the mastitis data we plot, in the left panel of Figure 4, the centered average of the posterior mean frailty over time together with the 95% confidence limits (see Shih and Louis (1995) and Section 4.2.6 in Duchateau and Janssen (2008) for technical details). The relatively late constancy over time provides some evidence against the gamma frailty model. This is in line with our finding.

6.2 Analysis of the insemination data

Time from parturition to first insemination is one of the main factors determining the calving interval (the time between two calvings) which is optimally between 12 and 13 months. One objective of artificial insemination programs is to look for cow factors that can be used to determine the appropriate moment for first insemination (where the considered event time is the time from parturition to first insemination). The concrete dataset contains data from 181 dairy farms (= herds), the number of cows within a herd varies from 1 to 174, with an average of 58 cows. As in the previous example we consider parity as the single binary covariate. Each data line contains the herd identification number as well as, for each cow in that herd, the minimum of the time from parturition to first insemination and the censoring time, the censoring indicator and the parity, e.g. for cow 1 and cow 51 in herd 1 (herd 1 has 51 cows) we have \( \{1, 68.5, 1, 0\} \) and \( \{1, 70.5, 1, 1\} \).

We assume a Weibull baseline hazard, i.e. \( \xi = (\lambda, \rho) \). For this example, we obtain 51.51 as \( T_{N,\text{os}} \), which is much larger than the largest of the 300 bootstrap samples (8.41), resulting in \( p \)-value zero using \( M_N = 5 \), i.e. there is no support for the null hypothesis that the frailty terms follow a one-parameter gamma distribution. Also, here the choice of \( M_N \) is not crucial, the same \( p \)-value is obtained when setting \( M_N = 3 \).
For this dataset, the extended gamma frailty density with $m = 4$ was preferred: with $C_{0.05}$ estimated to be 3.93, $\text{AIC}_{C_{0.05}}$ for the null model equals $-30538.06$ while for $m = 4$ $\text{AIC}_{C_{0.05}}$ takes a value of $-30435.46$. The parameter estimates of the chosen density are given by $\hat{\lambda} = 0.04 \cdot 10^{-2}$ (se = 0.03 · $10^{-3}$), $\hat{\rho} = 1.76$ (se = 0.01), $\hat{\beta} = -0.16$ (se = 0.02), $\hat{\theta} = 6.02$ (se = 5.32), $\hat{d}_1 = 1.31$ (se = 0.29), $\hat{d}_2 = 2.50$ (se = 0.51), $\hat{d}_3 = 2.46$ (se = 0.46), and $\hat{d}_4 = 1.77$ (se = 0.56). Figure 3 gives a graphical representation of the estimated null density and the estimated frailty density with $m = 4$.

The plot in the right panel in Figure 4 for the time to first insemination data is far from constant, i.e. the null model (gamma frailty model) is clearly not able to fit the data. Again, this conclusion is in line with our finding.

7. Discussion and possible extensions

7.1 Score-based tests

A disadvantage of this likelihood-based test is that, for each value of $m = 0, \ldots, M_N$, the model needs to be refit in order to find the maximum of the likelihood. An alternative to a likelihood ratio statistic is a score statistic. This has the advantage that the model only needs to be fit under the null hypothesis. For linear mixed models, the use of a score statistic for testing the distribution of random effects has been suggested by Thas (2009) in a discussion of Claesknens and Hart (2009). A limited simulation comparison
has shown good power behavior in comparison with the nested models based order selection test. For the shared frailty models, the score statistic requires the computation of the matrix of second partial derivatives with respect to all unknown parameters. If the fitting times are an issue, this might be worthwhile to derive.

7.2 Singleton tests

Another option to reduce computation times of a full likelihood-based test is to consider the so-called singleton expansions in which case we construct extensions of the gamma density that contain only a single term of the series expansion in the following way, \( \tilde{f}_{Uj}(u) = f_{Uj}(u) v_j^T(u) \). Since the polynomials \( v_j \) are orthonormal with respect to the gamma density, the integral of \( \tilde{f}_{Uj} \) is 1, and the square guarantees that the function is positive. Such singleton alternative models were considered by Aerts and others (2004) in a regression setting. A difference with the above construction is that the different models are no longer nested, with the exception of the null model that is nested within each other model. On the basis of the set of singleton alternatives, we consider a test statistic that combines the likelihood values of the different singleton expansions of the gamma density. Define \( T_{N,\text{singleton}} = \max_{m=1,...,M_N} \{ \log \hat{L}_m(\hat{\xi}_m, \hat{\beta}_m, \hat{\theta}_m) - \log \hat{L}_0(\hat{\xi}_0, \hat{\beta}_0, \hat{\theta}_0) \} \), where \( \hat{L}_m \) (\( m = 1, \ldots, M_N \)) denotes the likelihood of the model using the singleton alternative density for the frailty effect in the model. Also, for this test a bootstrap procedure would be advised rather than working with the asymptotic distribution.

7.3 Parametric frailty models: the choice of the baseline hazard and linear form of covariates

As mentioned in the introduction, a frailty model has several components. In this paper, we focus on the frailty (distribution) component. Although the Weibull baseline hazard is used in the previous sections, the proposed methodology is valid for any parametric choice of the baseline hazard, e.g. a piecewise constant baseline hazard. The parametric baseline component can thus be modeled in a flexible way. To relax the linear form of the covariates, we can rely on spline ideas used in Duchateau and Janssen (2004).

7.4 Semi-parametric frailty models

As mentioned in the introduction, assuming a parametric baseline hazard is restrictive in the sense that detected lack of fit does not ensure that this is a direct consequence of the misspecification of the gamma frailty density. It can indeed also be caused by misspecification of one of the other components in the proposed frailty model, e.g. the baseline hazard. It therefore would be interesting to have a method that still uses the family of generalized gamma frailties but leaves the baseline unspecified. A first approach could be to profile out the baseline hazard by using the EM algorithm to fit frailty models. The main issue there is that we need expressions for the expected values of the unobserved frailties (Duchateau and Janssen, 2008, Section 5.1); numerical methods will be instrumental in tackling the problem this way. A second way is to adapt the penalized likelihood algorithm (Duchateau and Janssen, 2008, Section 5.2) to the case of generalized gamma frailty densities and to develop an appropriate order selection test. For this approach the starting point is the algorithm given in Figure 5.5 of Duchateau and Janssen (2008). A third approach could be based on the following relation between the marginal survival function, \( S(t|x) \), and the cumulative hazard, \( H_0(t \exp(x^T\beta)) \), i.e. \( S(t|x) = L_f(H_0(t \exp(x^T\beta))) \) with \( L_f \) the Laplace transform of the frailty density under consideration. Using this identity, a two-step iterative approach could provide a solution. We then need a consistent non-parametric estimator for the marginal cumulative hazard \( H(t|x) \), since \( S(t|x) = \exp(-H(t|x)) \). Given that estimator and using the given relation between \( H(t|x) \) and \( H_0(t) \), a (numerical) procedure should be developed to estimate the cumulative baseline hazard. We then replace in the likelihood expression the parts that involve the baseline hazard by their estimated counterparts and estimate, in a second step, the parameters of the generalized gamma frailty density. To further explore the
latter idea, the papers by Glidden (1999, 2000), Cui and Sun (2004), and Spiekerman and Lin (1998) are instrumental. This is left for future research.

7.5 Testing for other frailty distributions

The proposed orthogonal polynomials are specifically constructed for the gamma frailty distribution. Calculations using a similar type of arguments could be done for other distributions, but this is beyond the scope of the current paper. While the orthogonalization might still be explicit, there is no guarantee that, for other distributions, the marginal likelihood can be explicitly obtained.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

The authors wish to thank Prof. A. Davison for helpful discussions and suggestions as well as Dr H. Laevens (Catholic University College Sint-Lieven, Sint-Niklaas, Belgium) for permission to use the mastitis data for this research. We further thank all reviewers for their constructive remarks. Conflict of Interest: None declared.

FUNDING

This work was supported by the IAP Research Network P6/03 of the Belgian State (Belgian Research Policy). For the simulations we used the infrastructure of the VSC—Flemish Supercomputer Center, funded by the Hercules Foundation and the Flemish Government—department EWI. The time to first insemination data are retrieved from: http://www.vetstat.ugent.be/education/survivalpartim2/0008insem.dat.

REFERENCES


[Received July 17, 2012; revised November 9, 2012; accepted for publication November 9, 2012]