Inference in Semiparametric Frailty Models

Erik Parner

Ph.D. Dissertation

Department of Theoretical Statistics
University of Aarhus

1997

Contents

Preface

During this Ph.D. project I have had the pleasure of visiting Richard Gill at the University of Utrecht a number of times. I am greatly indebted to him for his guidance, help and enthusiasm. I would also like to thank Susan Murphy for stimulating discussions and for providing helpful suggestions.

Summary

One of the standard assumptions in the analysis of survival data is that the individuals under observation are independent. In many cases this is a very realistic assumption because of the way the study was designed. In other cases, such as when studying family aggregation, this assumption is not realistic or even desirable. A simple model for dependent survival times is via the concept of frailty. The motivation for the frailty model is that shared unobserved risk factors not included in the model induces a dependence among a group of related survival times. These unobserved risk factors are denoted frailties. Groups sharing some risk factors might be a family, a pair of twins, mice born in the same litter or repeated measurements on one individual.

A particular area where the frailty models have been used is for the twin- and adoption studies. For the twin- and adoption data, the shared risk factors might be common genes and common environment. Indeed, one of the main purposes of these studies is to separate the effect of environment and genes; is the similarity of the mortality among individuals from the same family due to shared environment or shared genes? The twin- and adoption studies are well suited for answering this question. A higher clustering of survival times for monozygotic twins compared to dizygotic twins is ascribed to a genetic effect and, similarly, a higher clustering of the adoptee and its biological relatives compared to the adoptive relatives is ascribed to a genetic effect. The problem of estimating the degree of association between family members goes back to Galton in the last century.

Semiparametric models arise in situations where we have enough knowledge to model some features of the data parametrically, but are unwilling to assume anything about other features. One of the most used semiparametric models in survival analysis is Cox's proportional hazard model, where, in the two sample case, the treatment effect is modelled multiplicatively, i.e. parametrically, whereas the hazard in each group is left unspecified. The semiparametric frailty models considered here are all extensions of Cox's proportional hazard model.

The likelihood method is one of the most important methods for making inference in statistical models, such as for estimating unknown parameters, constructing confidence regions for the estimand and in hypothesis testing. In parametric models this method is well understood and it is known that the method enjoys a number of nice properties, both small sample properties as well as large sample properties. For the semiparametric frailty models the classical maximum likelihood method fails. This is due to the model containing an unknown infinite dimensional parameter. A development of the method, the nonparametric likelihood method, is in these cases often used.

The main part of this dissertation deals with investigating whether the nonparametric likelihood method shares the same properties as the classical likelihood method. A large part of the statistical inference, like finding confidence region and hypothesis testing, is base on the large sample behaviour of the model. The large sample property of the nonparametric maximum likelihood estimator in a particular frailty model, the correlated frailty model, is investigated. The investigation is based on modern empir-

ical process theory. The semiparametric frailty models are examples of a general class of semiparametric models called transformation models. The extension of the results derived for the correlated frailty model to other transformation models is discussed.

The last part of the dissertation is joint work with Li Hsu, Lue Ping Zhao and Hongzhe Li from the Fred Hutchinson Cancer Research Center, Seattle. In this work, a general class of aggregation models describing correlated survival times within families is proposed and a estimating equation based technique to estimate the relevant parameters is introduced. One of the aggregation models considered in this work is the major gene model. If the interest is in the genetic effect on the general mortality it is often realistic to assume that the effect is a result from a large (infinite) number of genes, each of infinitesimal effect. In this case frailty models based on a continuous distribution of the frailty, often the gamma distribution, is natural. Some diseases, e.g. some cancers, are believed to be controlled by a single or a few genes. In the major gene model the individuals are assumed to be in Hardy-Weinberg equilibrium and the Mendel law is used to describe the effect of a number of genes on the clustering of survival times, resulting in a discrete distribution of the frailty.

Inference in Semiparametric Frailty Models: A Review

Erik Parner

Department of Theoretical Statistics
University of Aarhus
Ny Munkegade
DK-8000 Aarhus C, Denmark
Email: erik@mi.aau.dk

1 Introduction

The frailty model is a multivariate model for survival times where an unobservable latent variable, a frailty, induces a positive correlation between the survival times. In this paper we give a review of some of the different frailty models used in survival analysis. This is done in section 2. A discussion of nonparametric maximum likelihood estimation and a comparison of some of the martingale methods and empirical process methods for studying the large sample properties of the nonparametric maximum likelihood estimator (NPMLE) is presented in section 3 and section 4. Finally, in Section 5 it is shown that the approach of using the EM-algorithm for calculating the NPMLE suggested in Gill (1985) and Nielsen et al. (1992) can be extended to general frailty models for which the Laplace transform of the frailty distribution is known.

2 Frailty models

The notation of frailty was introduced by Vaupel et al. (1979) to model heterogeneity of survival times. Suppose we observe survival times $T_1, ..., T_n$ and that these, conditional on frailty variables $Z_1, ..., Z_n$, have hazards

$$Z_i \alpha(t) \ ,$$

for some baseline hazard $\alpha(\cdot)$. The frailties are assumed to be unobservable random variables and have to be integrated out of the hazard functions. This type of frailty models is denoted individual frailty models. By the innovation theorem (Bremaud, 1981), the observed hazard is given by

$$E(Z_i | T_i > t) \alpha(t) \ .$$

The first term in the above displayed equation is time dependent and therefore the observed hazard can be quantitatively different from the conditional hazards. An example which nicely illustrates the effect of observing a heterogeneous population is taken from Vaupel and Yashin (1985): Divorce rate for the entire population of many countries follows a rise and then a decrease, with the peak at around the 7th year. This phenomenon is called the 7-year crisis of marriage. Does this imply that marriages are shakiest after a few year of marriage? This need not be the case. The effect can also be produced if the entire population consist of two subpopulations; one population which is "immune" or at low risk to divorce and another for which the risk is steadily increasing. In the beginning the risk for the entire population will increase because of the increasing risk in the second subpopulation. After some time the majority of couples in the high risk group are divorced and the risk for the entire population decreases.

In this paper the focus will be on the use of frailty to model multivariate survival times. The motivation for this type of frailty models is that unobservable common risk factors (covariates) not included in the model induce a correlation among a group of related survival times. Clayton (1978) was one of the first to use the idea of an unobservable covariate in the Cox regression model to model association for multivariate survival times. Groups sharing some risk factor might be a family, a pair of twins, mice from the same litter or repeated measurements on one individual. A typical application of frailty models is the twin- and adoption data (see e.g. Nielsen et al., 1992, Yashin et al., 1995, and Petersen, 1996). For the twin- and adoption data, the shared risk factors might be common genes and common environment. Indeed, one of the main purposes of these studies is to separate the effect of environment and genes; is the similarity of the mortality among individuals from the same family due to shared environment or shared genes? The twin- and adoption studies are well suited for answering this question. A higher clustering of survival times for monozygotic twins compared to dizygotic twins is ascribed to a genetic effect and, similarly, a higher clustering of the adoptee and its biological relatives compared to the adoptive relatives is ascribed to a genetic effect.

The frailty is usually modelled as an unobserved random variables acting multiplicative on the baseline hazard. The simplest case, the shared frailty model, is where all individuals in a group share the same value of the frailty. Suppose we observe survival times T_{ij}, $i = 1, ..., n, j = 1, ..., m$, corresponding to the j'th individual in the i'th group. Then the hazards given frailty variables $Z_1, ..., Z_n$ are assumed to be of the form

$$\lambda_{ij}(t|Z_i) = Z_i\alpha(t) \ .$$

The most common choice of distribution for Z_i is a gamma distribution with mean one and an unknown variance θ. The value $\theta = 0$ corresponds to independence and a high value of θ should preferable correspond to a high correlation between the survival times. There is some tradition in using the gamma distribution as the distribution for latent variables, e.g. the negative binomial distribution can be thought of as describing the number of accidents experienced by individuals in a time interval if these are Poisson distributed with parameter λ and λ is assumed to be gamma distributed (Greenwood and Yule, 1920). The choice of the gamma distribution here is made mostly for mathematical convenience. Other choices for the distribution of the frailty have been discussed in a series by Hougaard (1984, 1986, 1987) and Aalen (1992), among these are the positive stable distribution. A common feature of the different distributions considered is that the Laplace transform of the distributions are known, which means that the observed survival function is readily calculated. In Section 5 it is shown that if the Laplace transform of the frailty distribution is known then estimation by means of the EM-algorithm is, in principle, straightforward.

Under various assumptions on the joint distribution of the survival times and the censoring times, several authors have shown identifiability of the multivariate survival function under right censoring (see e.g. Dabrowska, 1988, and Pruitt, 1993). What remains for the frailty model is from the observed joint survival function to identify the baseline hazard function and the distribution of the frailty. In the shared frailty model the joint survival function for a pair of survival times (T_1, T_2) is

$$\begin{aligned} S(t_1,t_2) &= E\exp(-Z\{A(t_1)+A(t_2)\}) \\ &= L_Z\{A(t_1)+A(t_2)\}\ , \end{aligned}$$

where L_Z denotes the Laplace transform of Z and $A(t) = \int_0^t \alpha ds$ is the integrated baseline hazard function. Oakes (1989) showed, using the theory of copula models, that from the observed survival function, a frailty distribution with mean one can be identified. Another way of explaining this result is to note that from the observed survival function one can trivially identify

$$L_Z\{A(t)\}\ ,\ L_Z\{2A(t)\}\ .$$

One can think of this observable pair as coming from an individual frailty model

$$Z\exp(X)\alpha(t)\ ,$$

where X is a covariate that attains two values; 0 and $\log 2$. The identifiability problem for this model has been considered by Elbers and Ridder (1982) and Kortram et al. (1995), and it was shown that frailty distributions with mean one can be identified.

There are, fundamentally, two ways of parametrizing the model; by the conditional hazards and by the observed hazards. These two approaches are called respectively the conditional- and the marginal approach. If the frailty is thought of as an unobserved covariate in the Cox regression model and if the frailty actually were observed then the conditional hazards are what one would be interested in estimating. With this point of view it is natural to parametrize with the conditional hazards. On the other hand, if for the marginal data one would apply a Cox regression model then it is natural to require that the multivariate model be consistent with the marginal model and in this case the observed hazards are of interest. The most common approach is the conditional approach (Clayton, 1978, Nielsen et al., 1992, Yashin et al., 1995), though the marginal approach is also used (Prentice and Cai, 1992). Of course the interpretation of the parameters in the two parametrizations are very different.

To illustrate the difference between the two approaches, consider the shared gamma-frailty model in the bivariate case. The observed joint survival function in the conditional approach is

$$S(t_1, t_2) = (1 + \theta\{A(t_1) + A(t_2)\})^{-\theta^{-1}} .$$

Let the integrated hazard function A be fixed. For θ tending to zero, the observed survival function tends to $\exp\{-A(t_1) - A(t_2)\}$, which corresponds to independence of the two survival times. The observed survival function is, as a function of θ, easily seen to be increasing. For θ tending to infinity, the observed survival function tends to one for all t_1, t_2. The reason for this can be found by looking at the observed marginal hazard function

$$\frac{1}{1 + \theta A(t)} \alpha(t) ,$$

which tends to zero for θ tending to infinity for all t. At the same time as the correlation parameter θ converges to infinity, the probability of observing failure in all finite intervals $[0, t]$ converges to zero. If we want the marginal observed survival function

$$S(t) = \{1 + \theta A(t)\}^{-\theta^{-1}}$$

to be constant then the integrated hazard should, for θ converging to infinity, be of the order

$$A(t) = \theta^{-1}\{S(t)^{-\theta} - 1\} \propto S(t)^{-\theta} ,$$

which converges to infinity. Now, if we instead parametrized with the marginal integrated hazard function, $\Gamma(t) = -\log S(t) = \theta^{-1} \log\{1 + \theta A(t)\}$, then the observed joint survival function can be written as

$$S(t_1, t_2) = (\exp\{\theta\Gamma(t_1)\} + \exp\{\theta\Gamma(t_2)\} - 1)^{-\theta^{-1}} .$$

As θ tends to zero, the observed joint survival function tends to $\exp\{-\Gamma(t_1) - \Gamma(t_2)\}$, corresponding to independence. The survival function is still increasing as a function of θ and for θ tending to infinity the observed survival function tends to $\exp\{-\Gamma(t_1 \vee t_2)\} = S(t_1 \vee t_2)$, corresponding to maximal dependency. The advantage of the marginal approach compared to the conditional approach is that the effect of the dependence parameter and the effect of the hazard is seperated.

For the marginal approach the estimation of the hazard function and the correlation parameter becomes almost orthogonal. This was seen in Oakes and Manatunga (1992) for the share frailty model with positive stable frailty distribution and Weibull marginal hazards, where they found a small correlation between the estimators of the hazard parameters and the dependence parameters. In view of the discussion above, we expect this also to hold for the shared gamma-frailty model.

The difference between the two approaches becomes more apparent when we include covariates. This can be done by assuming that the conditional hazards follow a Cox regression model or that the observed hazards follow a Cox regression model. As a consequence of the former, it follow from a result of Elbers and Ridder (1982) that for the model with covariate, frailty distributions with mean one can be identified from marginal data. For the twin data this means that from data of just one of the twins we can estimate the correlation between the twins, which of course is nonsense. Therefore the parameter θ, in the conditional approach, must describe something more than just the correlation between the individuals. To explain this, consider the marginal observed hazard function

$$\frac{1}{1+\theta\exp(\boldsymbol{\beta}^\top \boldsymbol{x}_j)A(t)}\exp(\boldsymbol{\beta}^\top \boldsymbol{x}_j)\alpha(t) \ .$$

It is seen that for the marginal distribution, θ is a measure of the departure from the proportional hazard model. This interpretation of the parameter is of course inherited in the multivariate model. Thus for the multivariate model, θ seems to have the role as both a correlation parameter and a parameter describing the departure from the proportional hazard model.

For the stable distribution model suggested by Hougaard (1986) the parameter in the positive stable distribution, γ say, cannot be identified from marginal data in the conditional approach. The is due to the fact that the observed hazards

$$\exp(\gamma\boldsymbol{\beta}^\top \boldsymbol{x}_j)A(t)^{\gamma-1}\alpha(t)\gamma \ .$$

are also proportional. Clearly, only $\gamma\boldsymbol{\beta}$ and $A(t)^{\gamma-1}\alpha(t)\gamma$ can be identified.

A consequence of the conditional approach was illustrated in Yashin et al. (1995). When applying the shared gamma-frailty model to the twin data they found a higher value of θ for the monozygotic twins than for the dizygotic twins, as one would expect, but they also found a steeper integrated hazard for the monozygotic twins compared to the dizygotic twins. However, monozygotic twins are as individual like dizygotic twins and have the same mortality. The result is a consequence of parametrizing by the conditional hazards as described above. This was one of the reasons which lead Yashin et al. (1995) to split the frailty for the j'th individual in a group into two component, $Z_{(j)} = Z_0 + Z_j$, where Z_0 is a common shared component and Z_j is an individual component. For the twin data the frailty Z_0 describes the common genes and environment and Z_j models possible heterogeneity between individuals after having accounted for the common genes and the common environment. Yashin et al. (1995) assume that the frailties Z_0, Z_j are independent and gamma distributed with different shape parameters but the same scale parameter. The model is denoted the correlated frailty model or the litter model. Let θ and θ^* denote the variance of Z_0 and Z_j, respectively. Yashin et al. (1995) argues that the correlation between $Z_{(j)}$ and $Z_{(k)}$, i.e. $\theta/(\theta^*+\theta)$, is a proper index of the correlation between the survival times. This index can of course not be identified from the marginal data.

Korsgaard and Andersen (1996) extended the litter model of Yashin et al. (1995) to study genetic effects. Assuming for a family that the father and mother are unrelated, additive frailties for the father (F), the mother (M) and one offspring (O) is given by

$$\begin{aligned} Z_{(F)} &= Z_1 + Z_2 \\ Z_{(M)} &= Z_3 + Z_4 \\ Z_{(O)} &= Z_1 + Z_3 \ . \end{aligned}$$

The frailty Z_1 represent the part of the fathers genome affecting frailty that is transmitted to the offspring, Z_2 the corresponding part of the fathers genome not transmitted to the offspring and so forth. The Z_i's are assumed to be gamma distributed with shape parameter $\eta/2$ and scale parameter η, since father and offspring on average only share half their genes, so the correlation between $Z_{(F)}$ and $Z_{(O)}$ is one half. The model is further extended in Petersen (1996) to allow for environmental effects.

A natural generalization of the litter model presented by Yashin et al. (1995), which also contains the genetic model of Korsgaard and Andersen (1996) and the genetic-environmental model of Petersen (1996), is given by

$$Z_{(j)} = \mathbf{B}_j^\top \boldsymbol{Z} \ , \tag{1}$$

where $\boldsymbol{Z} = (Z_1, ..., Z_p)$ are independent gamma distributed random variable with shape parameter $\boldsymbol{\nu} = (\nu_1, ..., \nu_p)$ and scale parameter $\eta = \nu_1 + ... + \nu_p$ and

$$(\mathbf{B}_j)_k = b_{jk} = \begin{cases} 1 & \text{the } j\text{'th individual has the } k\text{'th component} \\ 0 & \text{otherwise.} \end{cases}$$

This generalization is also discussed in Petersen (1996). In Appendix A.3 the formulas for the observed likelihood function and the conditional expectation of the frailty variable given the observable variables, obtained in Parner (1996a,b), is generalized to the general setting (1). These formulaes are useful for calculating the NPMLE by means of the EM-algorithm in the conditional approach.

The parameters in the correlated gamma-frailty model can be identified. This was shown in Yashin et al. (1994) and in Parner (1996a). In Appendix A.1 the proof of Parner (1996a) is generalized to more general additive gamma-frailty models. One might naturally ask if this is the case for all correlated frailty models. The answer to this question is likely to be negative; the observed joint survival function for a pair of survival times is, for $(Z_{(1)}, Z_{(2)}) = (Z_0 + Z_1, Z_0 + Z_2)$, equal to

$$S(u,v) = L_{Z_{(1)},Z_{(2)}}\{A(u), A(v)\} = L_{Z_0}\{A(u) + A(v)\} L_{Z_1}\{A(u)\} L_{Z_2}\{A(v)\} \ .$$

Assume that $EZ_{(1)} = 1$, that Z_1, Z_2 follow the same distribution and that A is smooth enough such that the functions $u \to S(au, bu)$ can be Taylor expanded on some finite interval for all $a, b \in \boldsymbol{R}$. Then the observed survival function is determined by its derivatives at zero. It is straightforward to see that for this system of equations there are more unknowns than equations. Even though it is not a linear system of equations, it does indicate that the moments cannot be identified and therefore the distribution of Z_0, Z_j neither.

The model of Korsgaard and Andersen (1996) assumes that the effect of the genes on mortality result from a sum of a large (infinite) number of loci, each of infinitesimal effect. The effect of death of a particular decease could also be controlled by a single or few Mendelian locuses. Such models, denoted major gene models, are described in Hsu et al. (1997). In the following we present the simplest example of the major gene model where we study the effect of one gene with two alleles, B and b. Consider again a family consisting of a father (F), a mother (M) and one offspring (O). Let p denote the frequency of the alleles B. Assume that the frequency of the genotypes BB, Bb and bb are p^2, $2p(1-p)$ and $(1-p)^2$. This can be derived in the following way: Let A_{jl}, $l = 1, 2$, $j = F, M, O$, denote the alleles of each of the fathers, mothers and offsprings. We assume that the four alleles of the father and mother are independent and take values B and b with probability p and $(1-p)$. Further, let $A_j = (A_{j1}, A_{j2})$, $j = F, M, O$ and let us for simplicity denote the event $\{A_j = Bb \text{ or } A_j = bB\}$ by $\{A_j = Bb\}$. Then A_j takes values BB, Bb and bb with probability p^2, $2p(1-p)$ and $(1-p)^2$, respectively. We assume that the genotype of the offspring follows a Mendelian transmission, i.e. (A_{O1}, A_F) and (A_{O2}, A_M) are independent and the conditional distribution of A_{O1} given A_F takes the values B, b with probability

$$\begin{aligned} P(A_{O1} = B | A_F = BB) &= 1 \\ P(A_{O1} = B | A_F = Bb) &= 1/2 \\ P(A_{O1} = B | A_F = bb) &= 0 \, , \end{aligned}$$

and similarly for (A_{O2}, A_M). This specifies the joint distribution of (A_F, A_M, A_O). Let $\boldsymbol{G}_j = (G_{j1}, G_{j2})$, where $G_{j1} = 1(A_j = BB)$ and $G_{j2} = 1(A_j = Ba)$. For the major gene model, the frailties for the family are given by $Z_{(j)} = \exp(\boldsymbol{\gamma}^\top \boldsymbol{G}_j)$, $j = F, M, O$.

Because of symmetry in the distribution of $(Z_{(F)}, Z_{(M)}, Z_{(O)})$ we have to assume either (A) $\gamma_1, \gamma_2 \leq 0$ or (B) $p \geq 1/2$ in order to be able to identify the parameters. This corresponds well with intuition; the genotype BB denotes either the genotype with the lowest mortality or the allele B denotes the most common allele. In practice, both assumptions (A) and (B) are often satisfied. Under the hypothesis $\gamma_1 = \gamma_2$, the allele B is dominating in the sense that the mortality of the genotypes BB and Bb are the same. Another hypothesis of interest is whether the individuals are independent or not. Assume for the moment that $\gamma_1 = \gamma_2 = \gamma$. The hypothesis $p = 1$ corresponds to no mutating gene and the hypothesis $\gamma = 0$ corresponds to no effect of the mutating gene. Because we only observe the mortality of the individuals, and not the genes, we cannot distinguish between the two hypotheses. It is therefore natural to incorporate some prior knowledge about the parameters. This could be that from experience one knows that $p \in [1/2, 1-\epsilon]$ or $\gamma \in (-\infty, -\epsilon]$, for some $\epsilon > 0$.

In Appendix A.2 it is shown that if either there are covariates in the model or the baseline hazard function is Weibull then the parameters can be identified. In Appendix A.4 nonparametric maximum likelihood estimation by means of the EM-algorithm is discussed, assuming conditional independence of all individuals given the genes and the observed covariates. In practice this approach is feasible for moderate family sizes.

If for the multivariate survival times the primary interest is in the regression parameter then using the pseudo likelihood function as if the survival times were independent, i.e. the product of the marginal likelihood functions, yields asymptotically normal estimators (Wei et al., 1989). The derivation of the ordinary likelihood function is based on conditional independence of the whole family given the genes. If we only believe this holds locally, e.g. only for one generation up and one generation down in the family tree, then we can use the following generalization of the method of Wei et al. (1989): Let $Y_{ij}(t)$ indicate (by the value one) if the i'th individual in the j'th group is a risk at time $t-$ and define $Y_{ij}(t;\boldsymbol{\beta}) = Y_{ij}(t)\exp(\boldsymbol{\beta}^{\top}\boldsymbol{x}_{ij})$, where $\boldsymbol{x}_{ij}$ is an associated covariate. Let $N_{ij}(t)$ indicate if the ij'th individual has failure before or at time t and let $\lambda_{ij}(u|\boldsymbol{G}_{ij})$ denote the stochastic intensity if the genes actually were observed, $\lambda_{ij}(u|\boldsymbol{G}_{ij}) = Y_{ij}(u;\boldsymbol{\beta})\exp(\boldsymbol{\gamma}^{\top}\boldsymbol{G}_{ij})\alpha(u)$. The log-likelihood function for the full, partially unobserved, data set is

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\int_0^{\tau}\log\{\lambda_{ij}(u|\boldsymbol{G}_{ij})\}\,dN_{ij}(u) - \int_0^{\tau}\lambda_{ij}(u|\boldsymbol{G}_{ij})\,du\ .$$

If we let $\mathcal{F}_u^i$ denote the information in $(N_{ij}, Y_{ij} : j = 1, ..., m)$ up to time u and further let

$$\lambda_{ij}^{\mathcal{F}}(u) = E_{\psi}(\lambda_{ij}(u|\boldsymbol{G}_{ij})|\mathcal{F}_{u-}^{i})\ ,$$

then the observed log-likelihood function can be written

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\int_0^{\tau}\log\{\lambda_{ij}^{\mathcal{F}}(u)\}\,dN_{ij}(u) - \int_0^{\tau}\lambda_{ij}^{\mathcal{F}}(u)\,du\ .$$

Now, let $\mathcal{G}_u^{ij}$ denote the information from the j'th individual, its parents and the offsprings in the i'th group and let $\lambda_{ij}^{\mathcal{G}}(u) = E_{\psi}(\lambda_{ij}(u|\boldsymbol{G}_{ij})|\mathcal{G}_{u-}^{ij})$. Then the following pseudo likelihood function

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\int_0^{\tau}\log\{\lambda_{ij}^{\mathcal{G}}(u)\}\,dN_{ij}(u) - \int_0^{\tau}\lambda_{ij}^{\mathcal{G}}(u)\,du$$

yields consistent estimators (Appendix A.5). Maximizing the pseudo log-likelihood function, however, requires numerical integration. A comparable estimator is proposed in Hsu et al. (1997).

3 Nonparametric maximum likelihood estimation

Martingale methods have traditionally been one of the main tools when studying asymptotics in survival analysis. These methods do, however, not generalize to the frailty models. Murphy (1994, 1995) used modern empirical process theory (see van der Vaart and Wellner, 1996) to prove asymptotic normality and efficiency for the nonparametric maximum likelihood estimator (NPMLE) in the shared gamma-frailty model without covariates. These methods were generalized to the correlated gamma-frailty model, allowing for covariates, in Parner (1996a,b). In this section we discuss nonparametric maximum likelihood estimation with starting point taken in the Cox regression model. This allows for comparing the martingale methods with the empirical process methods. (For details about the martingale methods used in survival analysis the reader is referred to Andersen et al., 1993.) The semiparametric frailty models are examples of transformations models. For a transformation model we observe $\boldsymbol{Y} = (Y_1, ..., Y_d)$, say, where Y_j are real and there exists unknown absolutely continuous transformations ϕ_j such that $Y_j = \phi_j(T_j)$ and the distribution of $(T_1, ..., T_d)$ given some possible covariate $\boldsymbol{x}$ is assumed to follow a parametric model with density $p_0(\cdot; \boldsymbol{x}, \boldsymbol{\xi})$. The density of $\boldsymbol{Y}$ given $\boldsymbol{x}$ is then of the form

$$p_0(\phi_1(Y_1), ..., \phi_d(Y_d); \boldsymbol{x}, \boldsymbol{\xi})\, \phi_1'(Y_1)...\phi_d'(Y_d) \; .$$

(for more details on transformation models see Bickel et al., 1993). Many of the methods developed in Murphy (1994, 1995) and Parner (1996a,b) apply for the nonparametric maximum likelihood estimator in transformation models (see e.g. Murphy et al., 1996c).

Consider independent, possibly censored survival times $T_1, ..., T_n$ with hazard functions

$$\lambda_i(u|\boldsymbol{x}_i) = \exp(\boldsymbol{\beta}^\top \boldsymbol{x}_i)\alpha(u) \; ,$$

where $\boldsymbol{x}_i$ is a covariate vector for the i'th individual, $\boldsymbol{\beta}$ is a column vector of regression parameters and $\alpha(\cdot)$ is the baseline hazard function, i.e. the Cox regression model. As before we let $Y_i(t)$ indicate if the i'th individual is a risk at time $t-$, define $Y_i(u; \boldsymbol{\beta}) = Y_i(u) \exp(\boldsymbol{\beta}^\top \boldsymbol{x}_i)$ and let $N_i(t)$ indicate if the i'th individual has failure before or at time t. The likelihood function in the Cox regression model is

$$\prod_{i=1}^{n} \prod_{u \leq \tau} \{Y_i(u; \boldsymbol{\beta})\alpha(u)\}^{\Delta N_i(u)} \exp(-\int_0^\tau Y_i(u; \boldsymbol{\beta}) dA) \; . \tag{2}$$

For fixed $\boldsymbol{\beta}$, it is straightforward to see that the likelihood function tends to infinity for the integrated hazard function tending to a discrete integrated hazard function with jumps only at the observed failure times, i.e., where $\sum_{i=1}^n N_i(\cdot)$ jumps. Hence the maximum likelihood estimator does not exists. This is a general phenomenon for transformation models. If the interest is in the Euclidean parameters, there are in principle two ways to proceed. The first one is to calculate the efficient score function (Bickel et al., 1993) and then base inference about the parameter of interest on the efficient score function. The efficient score function usually cannot be found on explicit form and has to be approximated, allthough it is possible for the Cox regression model (see e.g. Bickel et al., 1993). The approximation approach was taking in Maguluri (1993) for the shared gamma-frailty model for bivariate survival times with possibly different marginal survival functions. The second way is to use nonparametric maximum likelihood estimation (NPMLE). This is nowadays often referred to as semiparametric maximum likelihood estimation or just maximum likelihood estimation due to its resemblance to the ordinary maximum likelihood method.

For the Cox regression model, Cox (1972, 1975) suggested to base inference about $\boldsymbol{\beta}$ on the partial likelihood function

$$\prod_{i=1}^{n}\prod_{u\leq\tau}\left(\frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{x}_i)}{\sum_{l=1}^{n} Y_l(u)\exp(\boldsymbol{\beta}^\top \boldsymbol{x}_l)}\right)^{\Delta N_i(u)} , \tag{3}$$

and gave a derivation of (3) as a product of conditional probabilities. The integrated hazard function is usually estimated by the Breslow estimator

$$\widehat{A}(t) = \int_0^t \left(\sum_{i=1}^{n} Y_i(u)\exp(\hat{\boldsymbol{\beta}}^\top \boldsymbol{x}_i)\right)^{-1} dN.(u) ,$$

where $\hat{\boldsymbol{\beta}}$ denotes the Cox estimator.

Johansen (1983) gave an explanation for the Cox estimator and the Breslow estimator by means of maximum likelihood estimation. The explanation is based on the observation that if one assumes that the baseline hazard function is piecewise constant on intervals of length ϵ and denotes the maximum likelihood estimator of this submodel for $(\hat{\boldsymbol{\beta}}_\epsilon, \widehat{A}_\epsilon)$, then $(\hat{\boldsymbol{\beta}}_\epsilon, \widehat{A}_\epsilon)$ tends to the Cox estimator and the Breslow estimator for ϵ tending to zero. Johansen (1983) constructed a Poisson-extension of the Cox regression model which allowed for general integrated hazard functions. The maximum likelihood estimator, according to definition of Kiefer and Wolfowitz (1956), was found to be a discrete integrated hazard functions with jumps only at the observed failure times and which maximizes the likelihood function

$$\prod_{i=1}^{n}\prod_{u\leq\tau}\{Y_i(u;\boldsymbol{\beta})\Delta A(u)\}^{\Delta N_i(u)}\exp(-\int_0^\tau Y_i(u;\boldsymbol{\beta})dA) . \tag{4}$$

For fixed $\boldsymbol{\beta}$, the likelihood function is maximized by

$$\widehat{A}(t;\boldsymbol{\beta}) = \int_0^t \left(\sum_{i=1}^{n} Y_i(u)\exp(\boldsymbol{\beta}^\top \boldsymbol{x}_i)\right)^{-1} dN.(u)$$

and the profile likelihood function for $\boldsymbol{\beta}$ is easily seen to be Cox's partial likelihood function. Large sample properties of the Cox estimator and the Breslow estimator was derived using martingale techniques by Andersen and Gill (1982), among others.

Following the line of Johansen (1983), one could define the NPMLE as the maximum likelihood estimator in an extension of the model. However, there are many extensions of the model which seem reasonable depending on which aspect of the model one focuses on. All of these extensions need not produce estimators with good asymptotic properties. Comparing the likelihood function (4) of Johansen (1983) with the likelihood function (2) one sees that $\alpha(u)$ is replaced by $\Delta A(u)$ and $dA(u)$ in the absolutely continuous case is replaced by $dA(u)$ in the discrete case. This procedure is usually taken as a definition of the NPMLE in transformation models. Gill (1989) gave an explanation for the asymptotic normality and efficiency of the NPMLE in cases when it is already know that the NPMLE is consistent. In Parner (1996a) the

NPMLE was motivated by making the connection to a classic method for proving consistency of the maximum likelihood estimator which goes back to Wald (1949) and thereby explaining the consistency of the NPMLE. We shall in the following repeat the argument.

Let P_ψ denote the distribution of a single observation and let P_n denote the empirical distribution of the data. If a maximum likelihood estimator, according to the definition of Kiefer and Wolfowitz (1956), exists, $\widehat{\psi}_n$ say, then

$$\int \log \frac{dP_{\widehat{\psi}_n}}{d\mu}\, dP_n \geq \int \log \frac{dP_{\psi_0}}{d\mu}\, dP_n \ , \tag{5}$$

where μ denotes a measure dominating $P_{\widehat{\psi}_n}$ and P_{ψ_0}. Assume that for any subsequence of $\{n\}$ we can find a further subsequence, $\{n_k\}$, such that $\widehat{\psi}_{n_k} \to \psi$ for some ψ. From the uniform law of large numbers it then follows that the inequality (5) in the limit is

$$\int \log \frac{dP_\psi}{d\mu}\, dP_{\psi_0} \geq \int \log \frac{dP_{\psi_0}}{d\mu}\, dP_{\psi_0} \ . \tag{6}$$

On the other hand, from the positivity of the Kullback-Leibler information we have

$$\int \log \frac{dP_\psi}{d\mu}\, dP_{\psi_0} \leq \int \log \frac{dP_{\psi_0}}{d\mu}\, dP_{\psi_0}$$

with equality if and only if $P_\psi = P_{\psi_0}$ (see e.g. Hoffmann-Jørgensen, 1994, section 8.28). So if the model is identifiable then (6) implies $\psi = \psi_0$. Since the limit is independent of the subsequence we get that $\widehat{\psi}_n \to \psi_0$. One should note that the argument above only depends on the log-likelihood difference $\log(dP_{\psi_1}/d\mu) - \log(dP_{\psi_2}/d\mu)$. In the Cox regression model, $\psi = (\boldsymbol{\beta}, A)$ where A is an absolutely continuous integrated hazard function. To define the NPMLE we simply extend the above difference to allow for a discrete integrated hazard function in as 'smooth' a way as possible, and then define the NPMLE as the value which maximizes the first (extended) term of the difference. Since the true integrated hazard function is absolutely continuous we can no longer compare the NPMLE with the true value. Instead we compare the NPMLE with a sequence converging to the true value, $\psi_n = (\boldsymbol{\beta}_0, A_n)$, where A_n is discrete and $A_n \to A_0$. If for any subsequence we can find a further subsequence, $\{n_k\}$, such that $\widehat{\psi}_{n_k} \to \psi = (\boldsymbol{\beta}, A)$ with A absolutely continuous, and if the extension is 'smooth' enough, then the extended log-likelihood difference still converges to minus the Kullback-Leibler information

$$\int \log \frac{dP_\psi}{d\mu}\, dP_{\psi_0} - \int \log \frac{dP_{\psi_0}}{d\mu}\, dP_{\psi_0} \ . \tag{7}$$

This means that the extension we make should become smaller and smaller as n tends to infinity. Assuming that the parameters are identifiable we get, in the same way as above, that $\widehat{\psi}_n \to \psi_0$.

For the Cox regression model, assuming A_1 is absolutely A_2-continuous, the log-likelihood difference is

$$\frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \log\{\frac{Y_i(u;\boldsymbol{\beta}_1)}{Y_i(u;\boldsymbol{\beta}_2)}\frac{dA_1}{dA_2}(u)\}\, dN_{ij}(u) - \int_0^\tau Y_i(u;\boldsymbol{\beta})dA_1 + \int_0^\tau Y_i(u;\boldsymbol{\beta})dA_2 \ ,$$

since $dA_1(u)/dA_2(u) = \alpha_1(u)/\alpha_2(u)$. The expression is also well defined for A_1, A_2 discrete with mass only at the observed failure times, because then A_1 is absolutely A_2-continuous with derivative $\Delta A_1(u)/\Delta A_2(u)$. In this way we extend the log-likelihood difference to allow for discrete integrated hazard functions. The nonparametric log-likelihood function for a discrete integrated hazard function is then given by

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\{Y_i(u;\boldsymbol{\beta})\Delta A(u)\}\ dN_i(u) - \int_0^\tau Y_i(u;\boldsymbol{\beta})dA\ .$$

Informally, we can write the log-likelihood function as

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\{Y_i(u;\boldsymbol{\beta})dA(u)\}\ dN_i(u) - \int_0^\tau Y_i(u;\boldsymbol{\beta})dA\ ,$$

where in the first term $dA(t) = \alpha(t)$ in the absolutely continuous case and $dA(t) = E_0Y(t;\boldsymbol{\beta}_0)\alpha_0(u)n \times \Delta A(t)$ in the discrete case. The term $E_0Y(t;\boldsymbol{\beta}_0)\alpha_0(u)n$ in the discrete case is used for norming the nonparametric log-likelihood function. In this case if $n\Delta A(t)$ tends to $\alpha(t)\{E_0Y(t;\boldsymbol{\beta}_0)\alpha_0(t)\}^{-1}$ as n tends to infinity, the nonparametric log-likelihood function evaluated at $(\boldsymbol{\beta}, A) = (\boldsymbol{\beta}, \int n\Delta A d\{n^{-1}N.\})$ is asymptotically equal to the true log-likelihood function evaluated at $(\boldsymbol{\beta}, \int \alpha dt)$.

We have demonstrated that if the NPMLE in transformation models stays bounded then, under some regularity condition, it will be consistent. In the next section it is argued that it also will be asymptotically normal and efficient. The boundedness of the NPMLE could of course be obtained by putting a bound on the Euclidean parameters and for the transformations ϕ_j assuming that $n\Delta\phi_j(u)$ is bounded for all n and u. Certainly it is of interest to consider nonparametric maximum likelihood estimation also in the unbounded case. Because the nonparametric likelihood function is an approximation to the real likelihood function, which in some sense is not well behaved, it is clearly not given in advance that the NPMLE in this case should exists nor that it stays bounded, even though the former property usually is straightforward to verify.

The nonparametric likelihood function is not well behaved in all respects. The individual gamma-frailty model with arbitrary baseline hazard is not identifiable. Still, Korsgaard and Andersen (1996) claims that the nonparametric profile likelihood for the variance parameter of the frailty distribution is not a constant function of θ, as would be expected. Further work is needed to give a satisfactory explanation for this phenomenon.

Let us return to the shared gamma-frailty model. Murphy (1994, 1995) proved asymptotic normality and efficiency of the NPMLE in the shared gamma-frailty model, in the conditional approach, under the assumption that the unknown variance parameter θ is known to lie in a bounded set. Since θ is a measure of correlation between the individuals such a bound may in practice be difficult to find. Following the discussion in Section 2, it is also of interest to consider NPMLE in the marginal approach. In appendix A.6 it is shown that the NPMLE in the marginal approach and the unrestricted case stays bounded and hence is asymptotically normal and efficient. By reparametrizing in the conditional approach by the observed discrete hazard function and using some simple algebraic manipulations it turns out that the result also translates to the NPMLE in the conditional approach; the result of Murphy (1994, 1995) is valid also in the case where θ is allowed to vary freely.

4 Asymptotic normality of the NPMLE

As argued in Gill (1989), the NPMLE will also be the maximum likelihood estimator (MLE) in any parametric submodel passing through the point given by the NPMLE. For smooth parametric submodels the MLE solves the likelihood equation. Given the consistency of the NPMLE, we might try to identify the limiting distribution of the NPMLE by making a first order Taylor expansion of each of the score functions. Indeed, this analogue for the proof of the MLE in the finite dimensional case does identify the limiting distribution of the NPMLE if we choose enough submodels so that the score functions identify the NPMLE and thereby asymptotically the true parameter.

Consider submodels of the form $\epsilon \to \psi_\epsilon := \psi + \epsilon(\boldsymbol{h}_\beta, \int_0^{\cdot} h_A dA)$, where $\boldsymbol{h}_\beta$ is a vector with the same size as $\boldsymbol{\beta}$ and h_A is a function of bounded variation. Define $h = (\boldsymbol{h}_\beta, h_A)$ and let L_n denote the logarithm of the likelihood function. The score operator is defined as

$$S_n(\psi)(h) = \frac{\partial}{\partial\epsilon} L_n(\psi_\epsilon)\bigg|_{\epsilon=0} = \boldsymbol{h}_\beta^\top \boldsymbol{L}_{\beta n}(\psi) + L_{An}(\psi)(h_A)\ ,$$

where

$$\begin{aligned} \boldsymbol{L}_{\beta n}(\psi) &= \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \boldsymbol{X}_i\, dN_i - \int_0^\tau \boldsymbol{X}_i Y_i(\boldsymbol{\beta})\, dA \\ L_{An}(\psi)(h_A) &= \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau h_A\, dN_i - \int_0^\tau Y_i(\boldsymbol{\beta}) h_A\, dA\ . \end{aligned}$$

This formula is valid for A absolutely continuous, giving the score operator for the ordinary likelihood function, and for A discrete, giving the score operator for the nonparametric likelihood function, so we may informally write

$$S_n^{NPMLE}(\psi) = S_n^{MLE}(\psi)\ . \tag{8}$$

It is this relationship which really explains the asymptotic normality and efficiency of the NPMLE, assuming that the censoring is noninformative about $(\boldsymbol{\beta}, A)$ (Arjas and Haara, 1984). Gill (1989) takes (8) as the definition of the NPMLE.

Let $S(\psi)$ denote the mean of $S_1(\psi)$. Suppose that $S(\psi)$ is Fréchet differentiable at ψ_0 with derivative $\dot{S}_{\psi_0}$;

$$\begin{aligned} \dot{S}_{\psi_0}(\psi)(h) &= \frac{\partial}{\partial\epsilon} S(\psi_0 + \epsilon\psi)(h)\bigg|_{\epsilon=0} \\ &= -\{\sigma_{\beta\beta}(\boldsymbol{h}_\beta)\boldsymbol{\beta} + \sigma_{\beta A}(\boldsymbol{h}_\beta)A + \sigma_{A\beta}(h_A)\boldsymbol{\beta} + \sigma_{AA}(h_A)A\}\ , \end{aligned}$$

where

$$\begin{aligned}
\sigma_{\beta\beta}(\boldsymbol{h}_\beta)\boldsymbol{\beta} &= \boldsymbol{h}_\beta^\top \int_0^\tau E_0(Y(\boldsymbol{\beta}_0)\boldsymbol{X}\boldsymbol{X}^\top)dA_0\boldsymbol{\beta} \\
\sigma_{\beta A}(\boldsymbol{h}_\beta)A &= \boldsymbol{h}_\beta^\top \int_0^\tau E_0(Y(\boldsymbol{\beta}_0)\boldsymbol{X})dA \\
\sigma_{A\beta}(h_A)\boldsymbol{\beta} &= \int_0^\tau E_0(Y(\boldsymbol{\beta}_0)\boldsymbol{X})h_A dA_0\boldsymbol{\beta} \\
\sigma_{AA}(h_A)A &= \int_0^\tau E_0(Y(\boldsymbol{\beta}_0))h_A dA\ .
\end{aligned}$$

The operator $\sigma = (\sigma_\xi, \sigma_A) = (\sigma_{\xi\xi} + \sigma_{A\xi}, \sigma_{\xi A} + \sigma_{AA})$ is called the Fisher information operator. Let $H_A(t) = \int_0^t h_A ds$. It is useful to think of the Fréchet derivative as being of the form

$$-\dot{S}_{\psi_0}(\psi)(h) = h^\top \cdot i \cdot \psi = h^\top i\psi := (\boldsymbol{h}_\beta, H_A)\begin{pmatrix} i_{\beta\beta} & i_{\beta A} \\ i_{A\beta} & i_{AA}\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta} \\ A\end{pmatrix}, \tag{9}$$

where

$$\begin{aligned}
i_{\beta\beta}(i,j) &= \int_0^\tau E_0(Y(\boldsymbol{\beta}_0)(\boldsymbol{X}\boldsymbol{X}^\top)_{ij})dA_0 \\
i_{\beta A}(i,t) &= E_0(Y(\boldsymbol{\beta}_0)(t)\boldsymbol{X}_i) \\
i_{A\beta}(t,i) &= E_0(Y(\boldsymbol{\beta}_0)(t)\boldsymbol{X}_i)\alpha_0(t) \\
i_{AA}(s,t) &= E_0(Y(\boldsymbol{\beta}_0)(s))1\{s=t\}\ .
\end{aligned}$$

One should think of the operators in (9) as generalized matrices with multiplication for operators A, B, $A \cdot B$ given by

$$A \cdot B = \sum_i A(\cdot, i)B(i, \cdot)$$

in the discrete case and

$$A \cdot B = \int A(\cdot, t)B(t, \cdot)dt$$

in the absolutely continuous case. The transpose mapping is defined as $A(s,t)^\top = A(t,s)$. Note that the operator i is not symmetric.

Due to (8), the Fréchet derivative is calculated with the true likelihood function. Suppose that the Fréchet derivative is continuously invertible. For the Cox regression model the continuous invertibility of the Fréchet derivative is relatively straightforward to verify. For the correlated gamma-frailty model a general argument which uses the mixture construction was used in Parner (1996b). The proof of asymptotic normality is classical in the sence that, by means of empirical process theory, a first order Taylor expansion establishes the relationship

$$\sqrt{n}(\widehat{\psi}_n - \psi_0) = -\dot{S}_{\psi_0}^{-1}\{\sqrt{n}S_n(\psi_0)\} + o_P^*(1)\ . \tag{10}$$

The score operator $\sqrt{n}S_n(\psi_0)$ in the Cox regression model can be written as

$$\boldsymbol{h}_\beta^\top \boldsymbol{V}_{\beta n} + \int_0^\tau h_A \, dV_{An} = \int h \, dV_n \ ,$$

where $\boldsymbol{V}_{\beta n} = \sqrt{n}\boldsymbol{L}_{\beta n}(\psi_0)$ and

$$V_{An}(u) = n^{-1/2} \sum_{i=1}^n N_i(u) - \int_0^u Y_i(s; \boldsymbol{\beta}_0) \, dA_0(s) \ .$$

It is seen that V_{An} is a sum of i.i.d. processes of uniformly bounded variation. The score operator for general transformation models is of this form. It can be shown that the asymptotic normality of the score operator is equivalent to the asymptotic normality of V_{An}. For the Cox regression model (Andersen and Gill, 1982) and the shared gamma-frailty model (Murphy, 1995), V_n can be written as $\int U_n dM_n$, where U_n is predictable and M_n is a martingale. In this case V_n is also a martingale. Central limit theorems for martingales can then be used to verify that V_n is asymptotically normal. In Parner (1996b, Lemma 2) a central limit theorem for processes of bounded variation defined on a finite interval was proved. Using that $\dot{S}_{\psi_0}^{-1}$ is continuous, an application of the continuous mapping theorem (see e.g. van der Vaart and Wellner, 1996) gives that the NPMLE is asymptotically normal.

The equality (8) implies that the NPMLE is efficient; from (10) it follows that the NPMLE is an asymptotically linear estimator sequence. Applying (8) and the argument in van der Vaart (1995, p. 25 line 3-10 from the bottom) we see that the influence function is contained in the closed linear span of the tangent space and from Proposition 1 in van der Vaart (1995) it follows that the NPMLE is efficient.

Let $\boldsymbol{\xi}$ denote the finite-dimensional part of ψ. The asymptotic variance of

$$g^\top\{\sqrt{n}(\widehat{\psi}_n - \psi_0)\} := \sqrt{n}\{\boldsymbol{g}_\xi^\top(\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0) + \int_0^\tau g_A \, d(\widehat{A}_n - A_0)\}$$

is given by

$$g^\top i^{-1} \begin{pmatrix} \boldsymbol{g}_\xi \\ \int_0^\cdot g_A ds \end{pmatrix} := \sigma^{-1}(g) \begin{pmatrix} \boldsymbol{g}_\xi \\ \int_0^\cdot g_A ds \end{pmatrix} = \boldsymbol{g}_\xi^\top \boldsymbol{\sigma}_\xi^{-1}(g) + \int \sigma_A^{-1}(g) g_A dA_0 \ . \qquad (11)$$

For the Cox regression model it is possible to find an explicit expression for the inverse of the Fisher information operator. In general, as for the correlated gamma-frailty model, this is not possible. Because all parameters are estimated at a parametric rate, $\sqrt{n}$, one might expect that the asymptotic variance of the NPMLE can be estimated by the inverse of minus the second derivative of the nonparametric log-likelihood function with respect to the Euclidean parameters and the jumps of the integrated hazard function. This result was stated in Gill (1989) in the general case and in Murphy (1995) for the shared gamma-frailty model. In the Cox regression model the asymptotic variance of the NPMLE was found by martingale methods in Andersen and Gill (1982) and here it is easy to verify that the inverse of minus the second derivative of the nonparametric log-likelihood function is a consistent estimator for the asymptotic variance of the NPMLE. In Parner (1996b) it was proved, for the correlated gamma-frailty model, that minus the second derivative of the nonparametric log-likelihood function with respect to $\boldsymbol{\xi}$ and the jumps of A, evaluated at $\widehat{\psi}_n$, $\boldsymbol{j}_n(\widehat{\psi}_n)$ say, is invertible with probability tending to one and a consistent estimator of the asymptotic variance of the $g^\top\{\sqrt{n}(\widehat{\psi}_n - \psi_0)\}$ is given by

$$\boldsymbol{g}_d^\top \boldsymbol{j}_n(\widehat{\psi}_n)^{-1}\boldsymbol{g}_d \ ,$$

where $\boldsymbol{g}_d^\top = (\boldsymbol{g}_\xi^\top, \{g(u_l)\}_l)$ and $\{u_l\}$ denotes the failure times.

In practice it is not always necessary to invert the hole matrix $\boldsymbol{j}_n$ to obtain estimates of the asymptotic variance of the NPMLE of some of the parameters. Consider for example the shared gamma-frailty and the case where one is interested in estimating the asymptotic variance of the NPMLE of θ. This variance can be used for testing independence of the individuals. The EM-algorithm, as in Nielsen et al. (1992), is used to calculate the profile log-likelihood for θ. According to the above result, the inverse of minus the second derivative of the profile log-likelihood is a consistent estimator of the asymptotic variance of the NPMLE for θ. This derivative may in practice not be calculable. One could take discrete derivatives of the profile log-likelihood function. All such discrete derivatives, under natural regularity conditions, are shown to be consistent estimates of the asymptotic variance in Murphy and van der Vaart (1996b). Finally, one might also use the nonparametric likelihood function to make likelihood ratio inference. Murphy and van der Vaart (1996a) prove that the likelihood ratio statistics for the parametric part for semiparametric models have asymptotic χ^2-distributions.

Let us return to the Cox regression model to see what we have gained, if anything, by the empirical process techniques. Let $\boldsymbol{j}_{pn}(\boldsymbol{\beta})$ denote minus the second derivative of the nonparametric profile log-likelihood function based on the first n observations. Then $\boldsymbol{j}_{pn}(\boldsymbol{\beta})$ converges in probability to a deterministic matrix $\boldsymbol{\Sigma}$ for $\boldsymbol{\beta}$ converging to $\boldsymbol{\beta}_0$ in probability (for details, see Andersen and Gill, 1982). In Andersen and Gill (1982), using martingale methods, the asymptotic distribution of the Cox estimator and the Nelson-Aalen estimator was studied under the assumption that $\boldsymbol{\Sigma}$ is positive definite. In Parner (1996b), using empirical process methods, conditions on the covariates which ensure continuous invertibility of the Fisher information operator σ were identified. If we let σ_p denote the information operator for Cox's partial likelihood, i.e. $\sigma_p(\boldsymbol{h}_\beta) = \boldsymbol{h}_\beta^\top\boldsymbol{\Sigma}$, then the following formula holds

$$\sigma_p = \sigma_{\beta\beta} - \sigma_{\beta A}\sigma_{AA}^{-1}\sigma_{\beta A} \ .$$

In the finite-dimensional case the invertibility of σ would imply that σ_p is invertible. In Parner (1996b) it is shown that for the Cox regression model σ_p is indeed invertible, hence $\boldsymbol{\Sigma}$ is positive definite.

If the covariates are time independent, the condition which ensures invertibility of Fishers information operator states that the covariates should not be collinear in the following sense; if the equality

$$P_0(\boldsymbol{c}^\top \boldsymbol{X} = c_0) = 1$$

holds for a vector $\boldsymbol{c}$ and a constant c_0 then $\boldsymbol{c} = \boldsymbol{0}$. In other words, the condition states that the covariates are affinely independent. If the covariates are time dependent and exogenous, i.e. the Cox regression model is assumed conditional on $\boldsymbol{X} = \boldsymbol{x}$, they should, as processes, be affinely independent; if the equality

$$P_0(\boldsymbol{c}^\top \boldsymbol{X}(u) = c_0(u),\ u \in [0,\tau]) = 1$$

holds for a vector $\boldsymbol{c}$ and a function c_0 then $\boldsymbol{c} = \boldsymbol{0}$. If the covariates are time dependent and not exogenous a modification of the latter condition is needed. These conditions on the covariates are natural and cannot be avoided.

In Appendix A.7 another comment on one of the technical conditions in Andersen and Gill (1982) is given.

5 The EM-algorithm for frailty models

Gill (1985), in a discussion of the paper by Clayton and Cuzick (1985), suggested to use the EM-algorithm for calculating the NPMLE in the shared gamma-frailty model in the conditional approach. This approach was further developed in Nielsen et al. (1992). In this section the approach of Gill (1985) is generalized to shared frailty models for which the Laplace transform of the frailty is known, as for the gamma distribution.

Suppose that we observe independent, possibly censored survival times and that these conditional on a frailty variable Z have hazards given by

$$Z \exp(\boldsymbol{\beta}^\top \boldsymbol{x}_j)\alpha(t) \qquad , \ j = 1, ..., m \ , \tag{12}$$

where $\boldsymbol{x}_j$ is a covariate for the j'th individual. Let $Y_j(t)$ be an indicator for the j'th individual being under observation at time just before t, let $N_j(t)$ indicate if the individual has failure before or at time t and let $\boldsymbol{N} = (N_1, ..., N_m)$ and $\boldsymbol{Y} = (Y_1, ..., Y_m)$. Further, let $(\boldsymbol{N}_1, \boldsymbol{Y}_1, Z_1), ..., (\boldsymbol{N}_n, \boldsymbol{Y}_n, Z_n)$ be i.i.d. replications of $(\boldsymbol{N}, \boldsymbol{Y}, Z)$.

If we denote by $p(\cdot;\theta)$ the density of the frailty distribution then the likelihood function for the full unobservable data set $(\boldsymbol{N}_1, \boldsymbol{Y}_1, Z_1), ..., (\boldsymbol{N}_n, \boldsymbol{Y}_n, Z_n)$ is

$$\begin{aligned}
&\prod_{i=1}^n \prod_{j=1}^m \left\{ \prod_{t\in[0,\tau]} \{Z_i d\Lambda_{ij}(t)\}^{\Delta N_{ij}(t)} \exp(-Z_i \Lambda_{ij}(\tau)) \right\} p(Z_i;\theta) \\
&\quad = \prod_{i=1}^n Z_i^{N_{i\cdot}(\tau)} \exp\{-Z_i \Lambda_{ij}(\tau)\} p(Z_i;\theta) \prod_{i=1}^n \prod_{j=1}^m \prod_{t\in[0,\tau]} \{d\Lambda_{ij}(t)\}^{\Delta N_{ij}(t)} \ ,
\end{aligned} \tag{13}$$

where $\Lambda_{ij}(t) = \int_0^t Y_{ij}(u;\boldsymbol{\beta})dA(u)$ and a dot means summation over the corresponding index. The observed likelihood function is derived by integrating the Z_i's out of (13). Integrating over the i'th term in (13) yields

$$\begin{aligned}
&\int Z_i^{N_{i\cdot}(\tau)} \exp(-Z_i \int_0^\tau Y_{i\cdot}(\boldsymbol{\beta})dA) p(Z_i;\theta) dZ_i \times \prod_{j=1}^m \prod_{t\in[0,\tau]} \{Y_{ij}(t;\boldsymbol{\beta})dA(t)\}^{\Delta N_{ij}(t)} \\
&\quad = (-1)^k L_Z^{(k)}\{\Lambda_{i\cdot}(\tau)\} \times \prod_{j=1}^m \prod_{t\in[0,\tau]} \{Y_{ij}(t;\boldsymbol{\beta})dA(t)\}^{\Delta N_{ij}(t)} \ ,
\end{aligned}$$

where $L_Z^{(k)}(\cdot)$ denotes the k'th derivative of the Laplace transform of Z and $k = N_{i\cdot}(\tau)$.

The logarithm of the full nonparametric likelihood function (13) is

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\int_0^t \log\{Y_{ij}(t;\boldsymbol{\beta})dA(t)\}\ dN_{ij}(t) - Z_i\Lambda_{i\cdot}(\tau) + N_{i\cdot}(\tau)\log Z_i + \log p(Z_i;\theta)\ .$$

Note that the last two terms depend on the parameter θ only. For fixed θ, we can ignore the last two terms in the E-step of the EM-algorithm. The E-step consists therefore of calculating $E_{\psi_k}[Z_i|\boldsymbol{N}_i, \boldsymbol{Y}_i]$ and in the M-step we update $(\boldsymbol{\beta}, A)$ by calculating the NPMLE in the Cox regression model

$$E_{\psi_k}[Z_i|\boldsymbol{N}_i, \boldsymbol{Y}_i]\exp(\boldsymbol{\beta}^\top \boldsymbol{x}_{ij})\alpha(t),\ i = 1, ..., n,\ j = 1, ..., m\ .$$

If θ is one-dimensional we can use the EM-algorithm to calculate the profile likelihood for θ. The NPMLE of θ is then given as the value which maximizes the profile likelihood and the NPMLE of $(\boldsymbol{\beta}, A)$ is found by running the EM-algorithm once more for this value of θ. This estimation scheme was suggested in Nielsen et al. (1992) for the shared gamma frailty model. If θ is higher dimensional this may not be an efficient way of calculating the NPMLE. Instead, we could first fix θ, then use the EM-algorithm to calculate the NPMLE for $(\boldsymbol{\beta}, A)$ for given value of θ. For this value of $(\boldsymbol{\beta}, A)$ we update θ by maximizing the observed nonparametric likelihood function. We repeat this procedure until convergence. This estimation scheme was suggested in Petersen et al. (1995) for the correlated gamma frailty model.

Let f denote the density of $(\boldsymbol{N}_i, \boldsymbol{Y}_i, Z_i)$. The conditional expectation of Z_i given $(\boldsymbol{N}_i, \boldsymbol{Y}_i)$ is given by

$$\begin{aligned} E[Z_i|\boldsymbol{N}_i, \boldsymbol{Y}_i] &= \int Z_i\ f(Z_i|\boldsymbol{N}_i, \boldsymbol{Y}_i)\ dZ_i \\ &= \frac{\int Z_i\ f(\boldsymbol{N}_i, \boldsymbol{Y}_i, Z_i)\ dZ_i}{\int f(\boldsymbol{N}_i, \boldsymbol{Y}_i, Z_i)\ dZ_i} \\ &= -\frac{L^{(k+1)}\{\Lambda_{i\cdot}(\tau)\}}{L^{(k)}\{\Lambda_{i\cdot}(\tau)\}}\ , \end{aligned}$$

with k given as above. For twin data, for example, we need only the first three derivatives of the Laplace transform.

Example 1 The positive stable distribution, as the distribution of the frailty, was proposed by Hougaard (1984). The Laplace transform of the positive stable distribution is given by $L(t) = \exp(-t^\gamma)$, where $\gamma \in (0, 1]$. This distribution is interesting because assuming a proportional hazard model for the conditional hazards then also the observed hazards are proportional. Furthermore, the result of Elbers and Ridder (1982) does not apply here for the positive stable distribution because the mean does not exists. However, as we shall show below, all the conditional expectations given the observed data do exist and can easily be calculated.

To estimate the parameters, Hougaard et al. (1992) proposed the following estimation procedure. First the observed integrated hazard function and the regression parameters are estimated assuming that the individuals are independent. Secondly, keeping the observed hazard and the regression parameter fixed, the likelihood function is maximized as a function of γ. The final step involves a version of Newton's algorithm for maximizing the likelihood over all parameters simultaneously. If only the first two steps are implemented then the estimation procedure is called the two-step procedure.

Alternatively, the EM-algorithm may be used to calculate the NPMLE. The first two derivatives of the Laplace transform of the stable distribution is given by

$$\begin{aligned} L^{(1)}(t) &= -\exp(-t^\gamma)\gamma t^{\gamma-1} \\ L^{(2)}(t) &= \exp(-t^\gamma)\gamma^2 t^{2\gamma-2} - \exp(-t^\gamma)\gamma(\gamma-1)t^{\gamma-2} \end{aligned}$$

For higher order derivatives the following simple recursion formula can be useful. The k'th derivative of a product of two functions f, g can be written as

$$(fg)^{(k)} = \sum_{l=0}^{k} \begin{pmatrix} k \\ l \end{pmatrix} f^{(l)} g^{(k-l)} .$$

By this formula, the k'th derivative of the Laplace transform can be written on the form

$$\begin{aligned} L^{(k)}(t) &= \left(L(t) \frac{\partial}{\partial t} \log L(t) \right)^{(k-1)} \\ &= \sum_{l=0}^{k-1} \begin{pmatrix} k-1 \\ l \end{pmatrix} L^{(l)}(t) \{\log L(t)\}^{(k-l)} , \end{aligned}$$

where

$$\{\log L(t)\}^{(k)} = \alpha(\alpha-1)...(\alpha-k+1)t^{\alpha-k} .$$

This algorithm would be easy to implement on a computer.

This approach to the EM-algorithm can be generalized to more general frailty models than the shared frailty model. However, since identifiability for more general models other than the correlated gamma-frailty model has not been solved, we shall not proceed further with this point.

A Appendix

A.1 Identifiability for the correlated frailty model

In this section we consider the identifiability of the correlated gamma-frailty model, although the following method could also be used for other correlated frailty models where the integrated hazard function A is arbitrary and the distribution of $\boldsymbol{Z}$ only depends on a finite dimensional parameter.

Let $\kappa_0, \kappa_1, \kappa_2$ denote the cumulant transforms of Z_0, Z_1, Z_2. The joint and marginal survival functions are

$$\begin{aligned} L_{Z_{(1)},Z_{(2)}}\{A(u), A(v)\} &= L_{Z_0}\{A(u)+A(v)\}L_{Z_1}\{A(u)\}L_{Z_2}\{A(v)\} \\ L_{Z_{(j)}}\{A(u)\} &= L_{Z_0}\{A(u)\}L_{Z_j}\{A(u)\} \end{aligned}$$

and hence (for $u = v$) the function

$$\kappa_0\{2A(u)\} - 2\kappa_0\{A(u)\} \qquad (14)$$

is known. If either $Z_{(1)}$ or $Z_{(2)}$ have moment of order k then we can without loss of generality assume that A is k-times differentiable. Let $\kappa_0^{(l)}, \kappa_j^{(l)}$ denote the l'th derivative of κ_0, κ_j. Taking derivatives of (14) we derive the following set of equations

$$2\kappa_0^{(2)}(0)\alpha(0)^2 \qquad (15)$$

$$6\kappa_0^{(3)}(0)\alpha(0)^3 + 6\kappa_0^{(2)}(0)\alpha(0)\alpha^{(1)}(0) \qquad (16)$$

$$14\kappa_0^{(4)}(0)\alpha(0)^4 + 36\kappa_0^{(3)}(0)\alpha(0)\alpha^{(1)}(0) + 2\kappa_0^{(2)}(0)\{3\alpha^{(1)}(0)^2 + 4\alpha(0)\alpha^{(2)}(0)\} \qquad (17)$$

and so forth. From the marginal survival function we get

$$\{\kappa_0^{(1)}(0) + \kappa_j^{(1)}(0)\}\alpha(0) \qquad (18)$$

$$\{\kappa_0^{(2)}(0) + \kappa_j^{(2)}(0)\}\alpha(0)^2 + \{\kappa_0^{(1)}(0) + \kappa_j^{(1)}(0)\}\alpha^{(1)}(0) \qquad (19)$$

$$\begin{aligned} &\{\kappa_0^{(3)}(0) + \kappa_j^{(3)}(0)\}\alpha(0)^3 + 3\{\kappa_0^{(2)}(0) + \kappa_j^{(2)}(0)\}\alpha(0)\alpha^{(1)}(0) \qquad (20) \\ &\quad + \{\kappa_0^{(1)}(0) + \kappa_j^{(1)}(0)\}\alpha^{(2)}(0)\ , \end{aligned}$$

$j = 1, 2$, and so forth. Suppose $EZ_{(1)} = 1$. From (18) we then find $\alpha(0)$ and from (15) we find

$$\kappa_0^{(2)}(0) = -\mathrm{Var}(Z_0)\ . \qquad (21)$$

Assume $\mathrm{Var}(Z_0) \neq 0$. From (16) we find

$$\frac{\kappa_0^{(3)}(0)}{\kappa_0^{(2)}(0)}\alpha(0)^2 + \alpha^{(1)}(0)\ .$$

and using (19) we therefore know

$$\frac{\kappa_0^{(3)}(0)}{\kappa_0^{(2)}(0)} - \kappa_0^{(2)}(0) - \kappa_1^{(2)}(0)\ . \qquad (22)$$

For the correlated gamma-frailty it is sufficient to stop here, but one could in principle continue the procedure.

Example 2 For the litter frailty model studied in Yashin et al. (1995) we have $(Z_{(1)}, Z_{(2)}) = (Z_0 + Z_1, Z_0 + Z_2)$, where Z_0, Z_1, Z_2 are gamma distributed with parameter (ν_0, η), (ν_1, η), (ν_2, η), respectively, and $\eta = \nu_0 + \nu_1$. From (21) we find ν_0/η^2. Assuming that $\nu_0 \neq 0$, we find from formula (22) that

$$\frac{2\nu_0/\eta^3}{\nu_0/\eta^2} - \nu_0/\eta^2 - \nu_1/\eta^2 = \eta^{-1}$$

and we can identify both ν_0 and ν_1. From (18) with $j = 2$ we can identify ν_2.

Example 3 Consider the genetic frailty model in Korsgaard and Andersen (1996). Here we have $(Z_{(1)}, Z_{(2)}, Z_{(3)}) = (Z_1 + Z_2, Z_3 + Z_4, Z_1 + Z_3)$, where $\boldsymbol{Z} = (Z_1, ..., Z_4)$ are independent gamma distributed random variables with parameters $(\nu_1, \eta), ..., (\nu_4, \eta)$ and $\eta = \nu_1 + ... + \nu_4$. Assume that the shared components have strictly positive variance, i.e., $\nu_1, \nu_3 \neq 0$. To show that the parameters are identifiable we reparametrize the model. Let $\widetilde{Z} = Z/E(Z_1 + Z_2)$ and $\widetilde{A} = E(Z_1 + Z_2) \times A$. Then $(\widetilde{Z}_1, ..., \widetilde{Z}_4)$ are gamma distributed with parameters $(\nu_1, \tilde{\eta}), ..., (\nu_4, \tilde{\eta})$, where $\tilde{\eta} = E(Z_1 + Z_2) \times \eta = \nu_1 + \nu_2$ and $EZ_{(1)} = Z_1 + Z_2 = 1$. From the observed survival function for first and second component we can identify ν_1, ν_2, ν_3. Similarly, a reparametrization for the second and third component identifies ν_4, which shows that all parameters are identifiable.

Example 4 For the adoption model in Petersen (1996) we have $(Z_{(1)}, Z_{(2)}, Z_{(3)}) = (Z_2 + Z_3, Z_1 + Z_2 + Z_4, Z_1 + Z_5)$, where $Z_1, ..., Z_5$ are independent gamma distributed random variables with parameters $(\nu_1, \eta), ..., (\nu_5, \eta)$ and $\eta = \nu_1 + ... + \nu_5$. Assume again that the shared components have strictly positive variance, i.e., $\nu_2, \nu_3 \neq 0$. Proceeding as in Example 3, a reparametrization for the first and second component identifies $\nu_2, \nu_3, \nu_1 + \nu_4$ and a reparametrization for the second and third component identifies $\nu_1, \nu_2 + \nu_4, \nu_5$. Hence the parameters are identifiable.

A.2 Identifiability for the major gene model

For the identifiability of the observed model we shall treat the case where we have no covariates in the model and the case where we have covariates in the model separately.

Covariates

Assume that $\boldsymbol{\beta}^\top \boldsymbol{X}_j$ attains at least two different values other than zero. From the marginal distribution we can identify

$$L_{Z_j}\{\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_j) A(\cdot)\}$$

on some interval $[0, \gamma]$. From the result of Kortram et al. (1995) it follows that the distribution of

$$H_F := \boldsymbol{\gamma}^\top \boldsymbol{G}_F - \log E_0 \exp(\boldsymbol{\gamma}^\top \boldsymbol{G}_F)$$

can be identified. Define $\eta_1 = \gamma_1 - \log E\{\exp(\boldsymbol{\gamma}^\top \boldsymbol{G}_F)\}$, $\eta_2 = \gamma_2 - \log E\{\exp(\boldsymbol{\gamma}^\top \boldsymbol{G}_F)\}$ and $\eta_3 = -\log E\{\exp(\boldsymbol{\gamma}^\top \boldsymbol{G}_F)\}$.

Consider case (A). From the marginal distribution of

$$\exp(H_F) = \begin{cases} \exp(\eta_1) & , p^2 \\ \exp(\eta_2) & , 2p(1-p) \\ \exp(\eta_3) & , (1-p)^2 \end{cases} ,$$

we can identify η_3 as the largest value and p by its frequency, $(1-p)^2$. We can further identify η_1 and η_2 unless $p^2 = 2p(1-p)$ or $p = 2/3$. In this case it is easily seen that η_1 and η_2 can be identified from the distribution of $\exp(H_F) + \exp(H_O)$. From η_l, $l = 1, 2, 3$, and p we can identify $\boldsymbol{\gamma}$, p and A.

In case (B) we can, since $p^2 < 2p(1-p)$ and $p^2 \leq (1-p)^2$, identify η_1 and p as the value with the smallest frequency and its frequency, p^2. We can further identify η_2 and η_3 unless $2p(1-p) = (1-p)^2$ or $p = 1/3$. In this case η_2 and η_3 can be identified from the distribution of $\exp(H_F) + \exp(H_O)$. Note that under the conditions (A) and (B) all parameters can be identified from the marginal distribution of H_F

No Covariates

Assume that A is a Weibull integrated hazard, i.e. $A(t) = \delta t^\epsilon$ ($\epsilon, \delta > 0$). Let $H_i = \boldsymbol{\gamma}^\top \boldsymbol{G}_i - \log(\delta)$. As noted by Honoré (1990)

$$\epsilon = -\lim_{t \to 0} [\log\{-\log L_{\exp(H_i)}(t^\gamma)\}] / \log t \, .$$

Therefore, we can identify the distribution of $\exp(H_F)$ and $\exp(H_F) + \exp(H_O)$. As shown above this identifies $\boldsymbol{\gamma}$, p and δ.

A.3 NPMLE in the additive gamma-frailty model

Consider the additive frailty model given in (1). If we let $p(\cdot)$ denote the density of the frailty variable $\boldsymbol{Z}$ then the full unobserved likelihood function is

$$\begin{aligned} &\prod_{j=1}^{m} \prod_{t \leq \tau} \left\{ \{Z_{(j)} d\Lambda_j(t)\}^{\Delta N_j(t)} \exp\{-Z_{(j)} \Lambda_j(\tau)\} \right\} p(\boldsymbol{Z}; \boldsymbol{\nu}) \\ &\quad = \prod_{j=1}^{m} \left\{ (\mathbf{B}_j^\top \boldsymbol{Z})^{N_j(\tau)} \exp\{-(\mathbf{B}_j^\top \boldsymbol{Z}) \Lambda_j(\tau)\} p(\boldsymbol{Z}; \boldsymbol{\nu}) \right\} \prod_{j=1}^{m} \prod_{t \leq \tau} d\Lambda_j(t)^{\Delta N_j(t)} \, . \end{aligned}$$

Let $n_j = N_j(\tau)$. For $\boldsymbol{k}_j = (k_{j1}, ..., k_{jp})$ and $K_j = \{\boldsymbol{k}_j | k_{jl} \in \{0, b_{jl} n_j\}, k_{j1} + ... + k_{jp} = n_j\}$ write

$$\begin{aligned} \prod_{j=1}^{m} (\mathbf{B}_j^\top \boldsymbol{Z})^{n_j} &= \prod_{j=1}^{m} (b_{j1} Z_1 + ... + b_{jp} Z_p)^{n_j} \\ &= \sum_{\boldsymbol{k}_1 \in K_1} \begin{pmatrix} n_1 \\ k_{11} \cdots k_{1p} \end{pmatrix} (b_{11} Z_1)^{k_{11}} ... (b_{1p} Z_p)^{k_{1p}} \prod_{j=2}^{m} (b_{j1} Z_1 + ... + b_{jp} Z_p)^{n_j} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{\boldsymbol{k}_1 \in K_1} \cdots \sum_{\boldsymbol{k}_m \in K_m} \prod_{l=1}^{m} (b_{l1} Z_1)^{k_{l1}} \ldots (b_{lp} Z_p)^{k_{lp}} \\
&= \sum_{\boldsymbol{k}_1 \in K_1} \cdots \sum_{\boldsymbol{k}_m \in K_m} Z_1^{k_{\cdot 1}} \ldots Z_p^{k_{\cdot p}} \ .
\end{aligned}
$$

We calculate the contribution of each of the terms $\prod_{l=1}^{p} Z_l^{k_{\cdot l}}$ to the observed likelihood function

$$
\begin{aligned}
&\int \prod_{l=1}^{p} Z_l^{k_{\cdot l}} \exp\{-Z_l \sum_{j=1}^{m} b_{jl} \Lambda_j(\tau)\} p(\boldsymbol{Z}; \boldsymbol{\nu}) d\boldsymbol{Z} \\
&\qquad = \prod_{l=1}^{p} \frac{\Gamma(\nu_l + k_{\cdot l})}{\Gamma(\nu_l)} \frac{\eta^{\nu_l}}{\{\eta + \sum_{j=1}^{m} b_{jl} \Lambda_j(\tau)\}^{\nu_l + k_{\cdot l}}} \ .
\end{aligned}
$$

Therefore the observed likelihood function is given by

$$
\sum_{\boldsymbol{k}_1 \in K_1} \cdots \sum_{\boldsymbol{k}_m \in K_m} \prod_{l=1}^{p} \frac{\Gamma(\nu_l + k_{\cdot l})}{\Gamma(\nu_l)} \frac{\eta^{\nu_l}}{\{\eta + \sum_{j=1}^{m} b_{jl} \Lambda_j(\tau)\}^{\nu_l + k_{\cdot l}}} \prod_{j=1}^{m} \prod_{t \leq \tau} d\Lambda_j(t)^{\Delta N_j(t)} \ . \qquad (23)
$$

The conditional expectation of Z_h becomes a fraction of quantities like the one in (23); the numerator with one added to $k_{\cdot h}$.

A.4 NPMLE in the major gene model

In this section we shall consider nonparametric maximum likelihood estimation in the genetic frailty model for a family consisting of a father, a mother and q offspring. The distribution of the unobservable genes $\boldsymbol{G} = (\boldsymbol{G}_F, \boldsymbol{G}_M, \boldsymbol{G}_{O_1}, \ldots, \boldsymbol{G}_{O_q})$ can be factorized as

$$
\begin{aligned}
P(\boldsymbol{G}) &= P(\boldsymbol{G}_{O_1}, \ldots, \boldsymbol{G}_{O_q} | \boldsymbol{G}_F, \boldsymbol{G}_M) P(\boldsymbol{G}_F) P(\boldsymbol{G}_M) \\
&= \prod_{j=1}^{q} P(\boldsymbol{G}_{O_j} | \boldsymbol{G}_F, \boldsymbol{G}_M) P(\boldsymbol{G}_F) P(\boldsymbol{G}_M) \ .
\end{aligned}
$$

Under a Mendelian model for the genes, the conditional probabilities $P(\boldsymbol{G}_{O_j} | \boldsymbol{G}_F, \boldsymbol{G}_M)$ do not depend on the parameter p and hence the founders, i.e. the parents, are sufficient about p for the complete data. The probability of the parents can be written as

$$
p^{2n_j(BB)} \{2p(1-p)\}^{n_j(Bb)} (1-p)^{2n_j(bb)} \ ,
$$

where $n_j(BB) = 1\{\boldsymbol{G}_j = (1,0)\}$, $n_j(Bb) = 1\{\boldsymbol{G}_j = (0,1)\}$ and $n_j(bb) = 1\{\boldsymbol{G}_j = (0,0)\}$ for $j = F, M$. Let $\lambda_j(\cdot | \boldsymbol{G}_j)$ denote the conditional stochastic intensity of the j'th individual, i.e.,

$$
\lambda_j(u | \boldsymbol{G}_j) = Y_j(u) \exp(\boldsymbol{\gamma}^\top \boldsymbol{G}_j) \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_j) \alpha(u) \ ,
$$

and let $\boldsymbol{N} = (N_1, ..., N_m)$, $\boldsymbol{Y} = (Y_1, ..., Y_m)$ denote the associated counting processes, indicator processes, respectively.

Under conditional independence of the individuals given the unobservable genes, the complete likelihood for n families $(\boldsymbol{N}_1, \boldsymbol{Y}_1, \boldsymbol{G}_1), ..., (\boldsymbol{N}_n, \boldsymbol{Y}_n, \boldsymbol{G}_n)$, i.i.d. replications of $(\boldsymbol{N}, \boldsymbol{Y}, \boldsymbol{G})$, is

$$\begin{aligned}&\prod_{i=1}^{n} L(\boldsymbol{N}_i, \boldsymbol{Y}_i, \boldsymbol{G}_i; \psi)\\ &\qquad = \prod_{i=1}^{n}\prod_{j=1}^{m}\left\{\prod_{u\le\tau} \lambda_{ij}(u|\boldsymbol{G}_{ij})^{\Delta N_{ij}(u)} \exp(-\int_0^\tau \lambda_{ij}(u|\boldsymbol{G}_{ij})du)\right\} P(\boldsymbol{G}_i)\ . \end{aligned}\tag{24}$$

The observed likelihood is obtained by integrating out the genes, i.e.,

$$\prod_{i=1}^{n}\sum_{\boldsymbol{G}_i} L(\boldsymbol{N}_i, \boldsymbol{Y}_i, \boldsymbol{G}_i; \psi)\ .$$

The nonparametric likelihood function is defined by replacing $\alpha(u)$ by $\Delta A(u)$ and $\alpha(u)du$ by $dA(u)$ in (24).

We shall use the EM-algorithm for maximizing the nonparametric log-likelihood function. The logarithm of the nonparametric likelihood function for the complete data set is

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j=1}^{m}\Big\{N_{ij}(\tau)(\boldsymbol{\gamma}^\top\boldsymbol{G}_{ij} + \boldsymbol{\beta}^\top\boldsymbol{X}_{ij}) + \int_0^\tau \log\Delta A(u) dN_{ij}(u)\\ &\qquad -\int_0^\tau Y_{ij}(u)\exp(\boldsymbol{\gamma}^\top\boldsymbol{G}_{ij} + \boldsymbol{\beta}^\top\boldsymbol{X}_{ij})dA(u) + \log P(\boldsymbol{G}_i)\Big\}\ .\end{aligned}\tag{25}$$

In the E-step we calculate

$$\begin{aligned}E_{\psi_k}(\boldsymbol{G}_{ij}|\boldsymbol{N}_i, \boldsymbol{Y}_i) &= \frac{\sum_{\boldsymbol{G}_i} \boldsymbol{G}_{ij} L(\boldsymbol{N}_i, \boldsymbol{Y}_i, \boldsymbol{G}_i; \psi_k)}{\sum_{\boldsymbol{G}_i} L(\boldsymbol{N}_i, \boldsymbol{Y}_i, \boldsymbol{G}_i; \psi_k)}\\ E_{\psi_k}(\exp(\boldsymbol{\gamma}^\top\boldsymbol{G}_{ij})|\boldsymbol{N}_i, \boldsymbol{Y}_i) &= \frac{\sum_{\boldsymbol{G}_i} \exp(\boldsymbol{\gamma}^\top\boldsymbol{G}_{ij}) L(\boldsymbol{N}_i, \boldsymbol{Y}_i, \boldsymbol{G}_i; \psi_k)}{\sum_{\boldsymbol{G}_i} L(\boldsymbol{N}_i, \boldsymbol{Y}_i, \boldsymbol{G}_i; \psi_k)}\\ \hat{n}_{ij}(BB) = E_{\psi_k}(n_{ij}(BB)|\boldsymbol{N}_i, \boldsymbol{Y}_i) &= \frac{\sum_{\boldsymbol{G}_i} n_{ij}(BB) L(\boldsymbol{N}_i, \boldsymbol{Y}_i, \boldsymbol{G}_i; \psi_k)}{\sum_{\boldsymbol{G}_i} L(\boldsymbol{N}_i, \boldsymbol{Y}_i, \boldsymbol{G}_i; \psi_k)}\end{aligned}$$

and similarly for the expected frequencies of Bb and bb.

In the M-step, the estimator for p is easily found to be

$$p_{k+1} = \frac{\sum_{i=1}^{n}\sum_{j=F,M} 2\hat{n}_{ij}(BB) + \hat{n}_{ij}(Bb)}{2n}\ .$$

Fixing $\boldsymbol{\gamma}, \boldsymbol{\beta}$, we find that (25) is maximized for

$$A_{k+1}(u; \boldsymbol{\gamma}, \boldsymbol{\beta}) = \int_0^u \left(\sum_{i=1}^{n} E_{\psi_k}(\exp(\boldsymbol{\gamma}^\top\boldsymbol{G}_{ij})|\boldsymbol{N}_i, \boldsymbol{Y}_i)\exp(\boldsymbol{\beta}^\top\boldsymbol{X}_{ij})Y_{ij}(u)\right)^{-1} dN_{\cdot\cdot}(u)\ .$$

The resulting profile log-likelihood function for $\boldsymbol{\gamma}, \boldsymbol{\beta}$ is given by

$$\sum_{i=1}^{n}\sum_{j=1}^{m}[N_{ij}(\tau)(\boldsymbol{\gamma}^\top \boldsymbol{G}_{ij} + \boldsymbol{\beta}^\top \boldsymbol{X}_{ij})$$
$$- \int_0^\tau \log\{\sum_{h=1}^{n}\sum_{l=1}^{m} E_{\psi_k}(\exp(\gamma^\top \boldsymbol{G}_{hl})|\boldsymbol{N}_h, \boldsymbol{Y}_h)\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{hl})Y_{hl}(u)\}dN_{ij}(u)] \ .$$

The profile log-likelihood function can be maximized f.ex. by a Newton-Raphson algorithm. Determination of the second derivative requires additional evaluation of the following two conditional expectations

$$E_{\psi_k}(\boldsymbol{G}_{ij}\exp(\boldsymbol{\gamma}^\top \boldsymbol{G}_{ij})|\boldsymbol{N}_i, \boldsymbol{Y}_i) \quad E_{\psi_k}(\boldsymbol{G}_{ij}^2\exp(\boldsymbol{\gamma}^\top \boldsymbol{G}_{ij})|\boldsymbol{N}_i, \boldsymbol{Y}_i) \ ,$$

which can be calculated by formulas similar to the above.

If not all the founders are in the data, e.g. if one of the parents is missing, then the estimator for p in the M-step cannot be found on closed form and an iterative algorithm has to be used.

A.5 A pseudo likelihood method

Using the martingale property we can write minus the Kullback-Leibler information as

$$\sum_{j=1}^{m} E_0[\int_0^\tau \log\{\lambda_j^{\mathcal{G}}(u;\psi)\}\ dN_j(u) - \int_0^\tau \lambda_j^{\mathcal{G}}(u;\psi)\ du]$$
$$-E_0[\int_0^\tau \log\{\lambda_j^{\mathcal{G}}(u;\psi_0)\}\ dN_j(u) + \int_0^\tau \lambda_j^{\mathcal{G}}(u;\psi_0)\ du]$$
$$= \sum_{j=1}^{m} E_0[\int_0^\tau \left\{\log\frac{\lambda_j^{\mathcal{G}}(u;\psi)}{\lambda_j^{\mathcal{G}}(u;\psi_0)} - (\frac{\lambda_j^{\mathcal{G}}(u;\psi)}{\lambda_j^{\mathcal{G}}(u;\psi_0)} - 1)\right\} \lambda_j^{\mathcal{G}}(u;\psi_0)\ du] \ .$$

We see that minus the Kullback-Leibler information is non-positive, since $\log(x) \leq x+1$ for all $x > 0$, and therefore the Kullback-Leibler information has maximum in ψ_0. If the Kullback-Leibler information is zero then

$$\sum_{j=1}^{m}\int_0^\tau \left\{\log\frac{\lambda_j^{\mathcal{G}}(u;\psi)}{\lambda_j^{\mathcal{G}}(u;\psi_0)} - (\frac{\lambda_j^{\mathcal{G}}(u;\psi)}{\lambda_j^{\mathcal{G}}(u;\psi_0)} - 1)\right\} \lambda_j^{\mathcal{G}}(u;\psi_0)\ du = 0 \ ,$$

P_0-a.s. Since all quantities involved are left continuous with right hand limit, this implies that

$$\left\{\log\frac{\lambda_j^{\mathcal{G}}(u;\psi)}{\lambda_j^{\mathcal{G}}(u;\psi_0)} - (\frac{\lambda_j^{\mathcal{G}}(u;\psi)}{\lambda_j^{\mathcal{G}}(u;\psi_0)} - 1)\right\} \lambda_j^{\mathcal{G}}(\psi_0;u) = 0 \ , \tag{26}$$

for all $u \in]0,\tau]$, P_0-a.s. Using that $\log(x) = x - 1$ if and only if $x = 1$, then (26) implies that

$$\lambda_j^{\mathcal{G}}(u;\psi) = \lambda_j^{\mathcal{G}}(u;\psi_0) \qquad , \ j = 1, ..., m \ ,$$

for all $u \in]0,\tau]$, P_0-a.s. This identifies the joint distribution of e.g. a offspring and its parents and therefore in the major gene model we have $\psi = \psi_0$. Thus, under suitable regularity conditions, the pseudo likelihood function yields consistent estimators.

A.6 NPMLE in the shared gamma-frailty model

In this section we consider nonparametric maximum likelihood estimation in the shared gamma-frailty model where the model is parametrized by the observed hazards; $\Gamma(t) = \theta^{-1}\log\{1+\theta A(t)\}$, and both Γ and θ are allowed to vary freely. We shall show that if either $\Gamma(\tau)$ or θ tends to infinity then the nonparametric log-likelihood function tends to minus infinity as n tends to infinity. This also shows that the NPMLE exists with a probability tending to one. Finally, we shall treat the case where the model is parametrized with the conditional hazards.

The log-likelihood function is

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\{1+\theta N_{i\cdot}(t-)\}dN_{i\cdot}(t) + \int_0^\tau \log[\exp\{\theta\Gamma(t)\}d\Gamma(t)]dN_{i\cdot}(t)$$
$$-\{\theta^{-1}+N_{i\cdot}(\tau)\}\log(1+\int_0^\tau Y_{i\cdot}(t)d[\exp\{\theta\Gamma(t)\}-1]) \ . \qquad (27)$$

In the following we consider the discrete case, i.e. $d\Gamma(t) = n\Delta\Gamma(t)$. Let T_j denote the survival times in a group, C_j the corresponding right-censoring times, $T_{(1)},...,T_{(m)}$ the ordered values of $T_1,...,T_m$, and $C_{(1)},...,C_{(m)}$ the corresponding censoring times. Suppose that $(T_{ij}, C_{ij}, T_{i(j)}, C_{i(j)} : j = 1,...,m)$ are the survival times and censoring times associated with $(N_{ij}, Y_{ij} : j = 1,...,m)$ and that these are i.i.d. replicates of $(T_j, C_j, T_{(j)}, C_{(j)} : j = 1,...,m)$. We shall need the following two lemmas.

Lemma 1 *Let $a_{ni} > 0$ for $n \geq 1$ and $i = 1,...,n$ be given. Suppose $b \in (0,1)$ and define*

$$c_n = \frac{1}{n}\sum_{i=1}^{n}\log a_{ni} - b\,a_{ni} \ .$$

(a) If $n^{-1}\sum_{i=1}^n a_{ni}$ tends to zero then c_n tends to minus infinity. (b) If $n^{-1}\sum_{i=1}^n a_{ni}$ is bounded away from infinity then c_n is bounded away from infinity. (c) If $n^{-1}\sum_{i=1}^n a_{ni}$ tends to infinity then c_n tends to minus infinity.

Proof. First consider (a). Let $\epsilon > 0$ be given. Write

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} a_{ni} &= \frac{1}{n}\sum_{i=1}^{n} a_{ni}1\{a_{ni}\leq\epsilon\} + \frac{1}{n}\sum_{i=1}^{n} a_{ni}1\{a_{ni}\geq\epsilon\} \\ &\geq \log(\epsilon)\frac{1}{n}\sum_{i=1}^{n}1\{a_{ni}\geq\epsilon\} \ ,\end{aligned}$$

hence $n^{-1}\sum_{i=1}^n 1\{a_{ni}\geq\epsilon\}$ tends to zero. Consider

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\log a_{ni} &= \frac{1}{n}\sum_{i=1}^{n}\log a_{ni}1\{a_{ni}\leq\epsilon\} + \frac{1}{n}\sum_{i=1}^{n}\log a_{ni}1\{a_{ni}\geq\epsilon\} \\ &\leq \log(\epsilon)\frac{1}{n}\sum_{i=1}^{n}1\{a_{ni}\leq\epsilon\} + \frac{1}{n}\sum_{i=1}^{n}(a_{ni}-1)1\{a_{ni}\geq\epsilon\} \\ &\leq \log(\epsilon)\frac{1}{n}\sum_{i=1}^{n}1\{a_{ni}\leq\epsilon\} + \frac{1}{n}\sum_{i=1}^{n}a_{ni} \ .\end{aligned}$$

Therefore the left hand side in the above displayed equation is asymptotically smaller than $\log \epsilon$. Since this is true for all $\epsilon > 0$ then (a) follows. (b) is proved in a similar way.

To prove (c), note that there exists a $x_0 \geq 1$ such that $\log x \leq bx/2$ for $x \in [x_0, \infty)$. Therefore

$$\frac{1}{n}\sum_{i=1}^{n} \log a_{ni} - ba_{ni} \leq \log x_0 + \frac{1}{n}\sum_{i=1}^{n} \frac{b}{2}a_{ni} - ba_{ni} \leq \log x_0 - \frac{b}{2n}\sum_{i=1}^{n} a_{ni} ,$$

which tends to minus infinity and (c) is verified. □

Lemma 2 *The sequence* $n^{-1}\sum_{i=1}^{n}\log(n/i)$ *tends to one.*

Proof. Write

$$\frac{1}{n}\sum_{i=1}^{n}\log(n/i) = \log(\frac{n}{(n!)^{n^{-1}}}) .$$

Define $a_n = n!/n^n$. It is well-known that $(1+n^{-1})^n$ tends to e. Because $a_{n+1}/a_n = (1+n^{-1})^{-n}$ tends to e^{-1}, the sequence $(a_n)^{n^{-1}} = (n!)^{n^{-1}}/n$ is also convergent and has the same limit, e^{-1}, and the result follows. □

For the last term in (27) we have the following inequality

$$\log(1 + \int_0^\tau Y_{i\cdot} d[\exp\{\theta\Gamma\} - 1]) \geq \theta \int_0^\tau 1\{Y_{i\cdot} > 0\} d\Gamma .$$

Therefore the nonparametric log-likelihood function is dominated by

$$\frac{1}{n}\sum_{i=1}^{n}(N_{i\cdot}(\tau) - 1)^{+}\log(1+\theta) - \theta\int_0^\tau U_i d\Gamma + \int_0^\tau \log(n\Delta\Gamma)dN_{i\cdot} - \int_0^\tau 1\{Y_{i\cdot} > 0\}d\Gamma \quad (28)$$

where $U_i(t) = 1\{T_{i(1)} < t \leq T_{i(2)}, C_{i(1)} \geq T_{i(1)}, C_{i(2)} \geq T_{i(2)}\}$.

First consider the case where $\Gamma(\tau)$ tends to infinity and θ is bounded away from infinity. Split the log-likelihood function up into two terms according to whether $\{N_{i\cdot}(\tau) = 0, Y_{i\cdot}(\tau) \geq 1\}$ or not. In the first case (28) is dominated by

$$O(1) - \Gamma(\tau)\frac{1}{n}\sum_{i=1}^{n} 1\{N_{i\cdot}(\tau) = 0, Y_{i\cdot}(\tau) \geq 1\} ,$$

which tends to minus infinity. Now consider the second case. Let $s_{n1} < s_{n2} < \ldots < s_{np_n}$ denote the observed failure times based on the first n observations that fall in the second group. In this case (28) is dominated by

$$O(1) + \frac{1}{n}\sum_{i=1}^{p_n}\log\{n\Delta\Gamma(s_{ni})\} - \sum_{j=1}^{i} m^{-1}\Delta\Gamma(s_{nj}) .$$

The fraction p_n/n converges to some $p \in (0,1)$, P_0-a.s. Consider therefore

$$\frac{1}{p_n}\sum_{i=1}^{p_n} \log\{p_n \Delta\Gamma(s_{ni})\} - \sum_{j=1}^{i} m^{-1}\Delta\Gamma(s_{nj}) \ .$$

By Lemma 2, this is equal to

$$\begin{aligned} O(1) + &\frac{1}{p_n}\sum_{i=1}^{p_n} \log\{(p_n - i + 1)m^{-1}\Delta\Gamma(s_{ni})\} - \sum_{j=1}^{i} m^{-1}\Delta\Gamma(s_{nj}) \\ &= O(1) + \frac{1}{p_n}\sum_{i=1}^{p_n} \log\{(p_n - i + 1)m^{-1}\Delta\Gamma(s_{ni})\} - (p_n - i + 1)m^{-1}\Delta\Gamma(s_{ni}) \\ &\leq O(1) - 1 = O(1) \ . \end{aligned}$$

Hence, in the second case the log-likelihood function stays bounded away from infinity. This concludes the case where $\Gamma(\tau)$ tends to infinity and θ stays bounded.

We turn to the case where θ tends to infinity. This case can be further divided into the following two sub-cases; (A) $\log(\theta)[\theta\{\Gamma(\tau) - \Gamma(x_0)\}]^{-1}$ tends to zero for n tending to infinity for some $x_0 \in (0,\tau)$ and (B) $\log(\theta)[\theta\{\Gamma(\tau) - \Gamma(x_0)\}]^{-1}$ stays bounded away from zero for all $x_0 \in (0,\tau)$. We shall treat the two cases separately.

First consider case (A), i.e., $\log(\theta)[\theta\{\Gamma(\tau) - \Gamma(x_0)\}]^{-1}$ tends to zero for n tending to infinity for some $x_0 \in (0,\tau)$. Using that

$$E_0\{U(t)\} \geq E_0 P_0(T_1 < t \leq T_2, C_1 \geq \tau, C_2 \geq \tau | Z) > 0$$

for all $t \in (0,\tau]$, so that we can find a $\delta > 0$ for which $E_0\{U(t)\} \geq \delta$ for $t \in [x_0,\tau]$, we see that the sum of the first two terms in (28) are asymptotically smaller than

$$\log(\theta)\frac{1}{n}\sum_{i=1}^{n}(N_{i\cdot}(\tau) - 1)^{+} - \theta\delta\{\Gamma(\tau) - \Gamma(x_0)\} \ .$$

Thus the sum of the first two terms in (28) tends to minus infinity. The sum of the last two terms in (28) is

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\int_0^\tau \log\{n\Delta\Gamma\}dN_{ij} - \frac{1}{m}\int_0^\tau 1\{Y_{i\cdot} > 0\}d\Gamma\right\} \ .$$

Let $t_{n1} < t_{n2} < \ldots < t_{nq_n}$ denote the observed failure times based on the first n observations. The sum of the last two terms in (28) is therefore dominated by

$$\frac{1}{n}\sum_{i=1}^{q_n} \log\{n\Delta\Gamma(t_{ni})\} - \sum_{j=1}^{i} m^{-1}\Delta\Gamma(t_{nj}) \ .$$

The fraction q_n/n converges to some $q \in (0,1)$, P_0-a.s. By a similar argument as above we can obtain that the sum of the last two terms in (28) stays bounded from infinity. This finishes case (A).

Now consider case (B), i.e., $\log(\theta)[\theta\{\Gamma(\tau)-\Gamma(x_0)\}]^{-1}$ stays bounded away from zero for all $x_0 \in (0,\tau)$. The log-likelihood function is bounded by

$$\frac{1}{n}\sum_{i=1}^{n} -\log(\theta)1\{N_{i\cdot}(\tau)\geq 1\} + \int_0^\tau \log(\theta n\Delta\Gamma)dN_{i\cdot} - \int_0^\tau 1\{Y_{i\cdot}>0\}d\Gamma \ .$$

By assumption, for all $x_0 \in (0,\tau)$ there exists an $\epsilon = \epsilon(x_0) > 0$ such that asymptotically $\log(\theta)[\theta\{\Gamma(\tau)-\Gamma(x_0)\}]^{-1} \geq \epsilon$ or $-\log\theta \leq -\epsilon\theta\{\Gamma(\tau)-\Gamma(x_0)\}$. Let $\delta = P(N_{\cdot}(\tau)\geq 1)/2$. The log-likelihood function is asymptotically bounded by

$$\begin{aligned}
&-\delta\log\theta - \delta\epsilon\theta\{\Gamma(\tau)-\Gamma(x_0)\} + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log(\theta n\Delta\Gamma)dN_{i\cdot} - \int_0^\tau 1\{Y_{i\cdot}>0\}d\Gamma \\
&\quad\leq -\delta\log\theta - \delta\epsilon\frac{1}{n}\sum_{t_{ni}\in[x_0,\tau]}\theta n\Delta\Gamma(t_{ni}) + \frac{1}{n}\sum_{i=1}^{q_n}\log\{\theta n\Delta\Gamma(t_{ni})\} \\
&\quad\quad -\frac{1}{n}\sum_{i=1}^{q_n}(q_n-i+1)m^{-1}\Delta\Gamma(t_{ni}) \\
&\quad\leq -\delta\log\theta + \log(\theta)\frac{1}{n}\sum_{t_{ni}\in[0,x_0]}1 \\
&\quad\quad -\delta\epsilon\frac{1}{n}\sum_{t_{ni}\in[x_0,\tau]}\theta n\Delta\Gamma(t_{ni}) + \frac{1}{n}\sum_{t_{ni}\in[x_0,\tau]}\log\{\theta n\Delta\Gamma(t_{ni})\} \\
&\quad\quad +\frac{1}{n}\sum_{t_{ni}\in[0,x_0]}\log\{n\Delta\Gamma(t_{ni})\} - \frac{1}{n}\sum_{i=1}^{q_n}(q_n-i+1)m^{-1}\Delta\Gamma(t_{ni}) \\
&\quad\leq O(1) - \delta\log\theta + \log(\theta)\frac{1}{n}\sum_{t_{ni}\in[0,x_0]}1 \\
&\quad\quad -\delta\epsilon\frac{1}{n}\sum_{t_{ni}\in[x_0,\tau]}\theta n\Gamma(t_{ni}) + \frac{1}{n}\sum_{t_{ni}\in[x_0,\tau]}\log\{\theta n\Delta\Gamma(t_{ni})\} \ ,
\end{aligned}$$

where $\sum_{t_{ni}\in[0,x_0]}$, $\sum_{t_{ni}\in[x_0,\tau]}$ is the sum over all $t_{n1},...,t_{nq_n}$ in the interval $[x_0,\tau]$, $[0,x_0]$, respectively. We choose x_0 small enough such that the first two terms tends to minus infinity. The last two terms stay bounded away from infinity according to Lemma 1. This concludes case (B). Thus the nonparametric log-likelihood function tends to minus infinity if either θ or $\Gamma(\tau)$ tends to infinity.

Now consider the nonparametric log-likelihood function in the conditional approach

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\{1+\theta N_{i\cdot}(t-)\}dN_{i\cdot}(t) + \int_0^\tau \log\{n\Delta A(t)\}dN_{i\cdot}(t) \\
-\{\theta^{-1}+N_{i\cdot}(\tau)\}\log(1+\theta\int_0^\tau Y_{i\cdot}(t)dA(t)) \ .
\end{aligned}$$

We parametrize with the observed hazards $\Gamma(t) = \theta^{-1}\log\{1+\theta A(t)\}$, i.e., $A(t) = \theta^{-1}(\exp\{\theta\Gamma(t)\}-1)$ and $\Delta A(t) = \exp\{\theta\Gamma(t)\}\theta^{-1}[1-\exp\{-\theta\Delta\Gamma(t)\}]$. Using that

$1-\exp(-x) \leq x$, for $x \geq 0$, we find that $\theta^{-1}[1-\exp\{-\theta\Gamma(t)\}] \leq \Delta\Gamma(t)$ and the nonparametric log-likelihood function is dominated by

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\{1+\theta N_{i\cdot}(t-)\}dN_{i\cdot}(t)+\int_0^\tau \log[\exp\{\theta\Gamma(t)\}n\Delta\Gamma(t)\}]dN_{i\cdot}(t)$$
$$-\{\theta^{-1}+N_{i\cdot}(\tau)\}\log(1+\int_0^\tau Y_{i\cdot}(t)d[\exp\{\theta\Gamma(t)\}-1]) .$$

This is exactly the nonparametric log-likelihood function in the marginal approach. Therefore, if θ or $\Gamma(\tau)$ tends to infinity the nonparametric log-likelihood function tends to minus infinity as n tends to infinity. If both θ and $\Gamma(\tau)$ are bounded then $A(\tau)$ is also bounded.

In the regression setting, if both the explanatory variables and the regression parameters are bounded then the above results generalize easily. For the k-sample case and in the marginal approach the regression parameters are allowed to vary freely; For each individual, we let

$$\boldsymbol{X}_{ijl} = 1\{ij \text{ belongs to group } l\},\ l=1,...,k-1.$$

Split the nonparametric log-likelihood function up into a sum of k terms according to the k different values of the covariates. For each of the k terms, applying a similar argument as above gives that the nonparametric log-likelihood function tends to minus infinity if either $\exp(\boldsymbol{\beta}_1)\Gamma(\tau),...,\exp(\boldsymbol{\beta}_{k-1})\Gamma(\tau),\Gamma(\tau)$ tends to zero or positive infinity or θ tends to infinity. From this we conclude that the NPMLE for $\boldsymbol{\beta}$, θ and Γ stay bounded.

A.7 On a technical condition in Andersen and Gill (1982)

Andersen and Gill (1982) assume that the following integrability condition for the covariates holds; there exists a neighbourhood around $\boldsymbol{\beta}_0$, $\mathcal{B}$ say, such that

$$E_0\left(\sup_{\boldsymbol{\beta}\in\mathcal{B}}\sup_{t\in[0,\tau]} Y(t)|\boldsymbol{X}(t)|^2\exp\{\boldsymbol{\beta}^\top\boldsymbol{X}(t)\}\right) < \infty .$$

In the following we shall prove that the condition is equivalent to

$$E_0\left(\sup_{t\in[0,\tau]} Y(t)\exp\{\boldsymbol{\beta}^\top\boldsymbol{X}(t)\}\right) < \infty , \tag{29}$$

for all $\boldsymbol{\beta}$ in some neighbourhood of $\boldsymbol{\beta}_0$, $\mathcal{B}_0$ say. Indeed, we shall show that from (29) it follows that there exists a neighbourhood of $\boldsymbol{\beta}_0$, $\mathcal{B}'$ say, such that

$$E_0\left(\sup_{\boldsymbol{\beta}\in\mathcal{B}'}\sup_{t\in[0,\tau]} Y(t)|\boldsymbol{X}(t)|^p\exp(\boldsymbol{\beta}^\top\boldsymbol{X}(t))\right) < \infty \tag{30}$$

for all positive integers p.

Choose δ such that the closed cube $c(\boldsymbol{\beta}_0, \delta)$ with centre $\boldsymbol{\beta}_0$ and side length 2δ is contained in $\mathcal{B}_0$. For $\boldsymbol{\beta} \in c(\boldsymbol{\beta}_0, \delta)$ we have that

$$\boldsymbol{\beta}_j \boldsymbol{X}_j(t) \leq \begin{cases} (\boldsymbol{\beta}_{0j} + \delta)\boldsymbol{X}_j(t) & , \boldsymbol{X}_j(t) \geq 0 \\ (\boldsymbol{\beta}_{0j} - \delta)\boldsymbol{X}_j(t) & , \boldsymbol{X}_j(t) < 0 \end{cases}$$

and hence

$$\exp\{\boldsymbol{\beta}^\top \boldsymbol{X}(t)\} \leq \exp\{\boldsymbol{\beta}_0^\top \boldsymbol{X}(t)\} \left(\max_{(\alpha_1,\dots,\alpha_d)\in\{0,1\}^{\otimes d}} \exp\{\delta \sum_{k=1}^{d} (-1)^{\alpha_k} \boldsymbol{X}_k(t)\} \right) ,$$

where $\{0,1\}^{\otimes d} = \{0,1\} \times \dots \times \{0,1\}$ (d times) and d denotes the size of the vector $\boldsymbol{X}(t)$. We shall call the points in the corner of the cube $c(\boldsymbol{\beta}_0, \delta)$ for $\mathcal{B}_d$. Note that $\mathcal{B}_d$ consists of finitely many points and that from the above displayed inequality it follows that

$$\sup_{\boldsymbol{\beta}\in c(\boldsymbol{\beta}_0,\delta)} \exp\{\boldsymbol{\beta}^\top \boldsymbol{X}(t)\} \leq \sup_{\boldsymbol{\beta}\in \mathcal{B}_d} \exp\{\boldsymbol{\beta}^\top \boldsymbol{X}(t)\} . \tag{31}$$

Taylor expanding $\exp(|y|)$ we derive the following inequality

$$\frac{|y|^n}{n!} \leq e^{-y} + e^{y} .$$

Choose x small such that for $\boldsymbol{h} = (0, \dots, 0, x, 0, \dots, 0)^\top$ (the j'th element is equal to x) we have $\boldsymbol{\beta} \pm \boldsymbol{h} \in \mathcal{B}$. Then for all positive integers p

$$\begin{aligned} & E_0 \left(\sup_{t\in[0,\tau]} Y(t) |\boldsymbol{X}_j(t)|^p \exp\{\boldsymbol{\beta}^\top \boldsymbol{X}(t)\} \right) \\ = \;& \frac{1}{x^p} E_0 \left(\sup_{t\in[0,\tau]} Y(t) |\boldsymbol{h}^\top \boldsymbol{X}(t)|^p \exp\{\boldsymbol{\beta}^\top \boldsymbol{X}(t)\} \right) \\ \leq \;& p! \frac{1}{x^p} E_0 \left(\sup_{t\in[0,\tau]} Y(t) \exp\{(\boldsymbol{\beta}+\boldsymbol{h})^\top \boldsymbol{X}(t)\} + \sup_{t\in[0,\tau]} Y(t) \exp\{(\boldsymbol{\beta}-\boldsymbol{h})^\top \boldsymbol{X}(t)\} \right) \\ < \;& \infty . \end{aligned}$$

Using the simple inequality $|\boldsymbol{a}|^p = |a_1^2 + \dots + a_d^2|^{p/2} \leq d^{p/2} \sum_{i=1}^{d} |a_i|^p$ we get that

$$E_0 \left(\sup_{t\in[0,\tau]} Y(t) |\boldsymbol{X}(t)|^p \exp\{\boldsymbol{\beta}^\top \boldsymbol{X}(t)\} \right) < \infty , \tag{32}$$

for all positive integers p and all $\boldsymbol{\beta} \in c(\boldsymbol{\beta}_0, \delta)$. Combining (32) with (31) we see that (30) holds with $\mathcal{B}' = c(\boldsymbol{\beta}_0, \delta)$.

References

Aalen, O.O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution. *Ann. Appl. Prob.* **2**: 951-72.

Andersen, P.K., and Gill, R.D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10**: 1100-1120.

Andersen, P.K., Borgan Ø, Gill, R.D., and Keiding, N. (1993). *Statistical models based on counting processes.* Berlin: Springer-Verlag.

Arjas, E., and Haara, P. (1984). A marked point process approach to censored failure data with complicated covariates. *Scand. J. Statist.* **11**: 193-209.

Bickel, P., Klassen, C., Ritov, Y., and Wellner, J. (1993). *Efficient and adaptive estimation for semiparametric models.* Johns Hopkins University Press, Baltimore.

Bremaud, P. (1981). *Point Processes and Queues.* New York: Springer Verlag.

Clayton, D.G. (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.

Clayton, D.G., and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *J. Roy. Statist. Soc., series A* **148**: 82-117.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc., series B* **34**: 187-220.

Cox, D.R. (1975). Partial likelihood. *Biometrika* **62**: 269-276.

Dabrowska, D.M. (1988). Kaplan-Meier estimate on the plane. *Ann. Statist.* **16**: 1475-1489.

Elbers, C., and Ridder, G. (1982). True and spurious duration dependence: the identifiability of the proportional hazard model. *Review of Economic Studies* **XLIX**: 403-409.

Gill, R.D. (1985). Discussion of the paper by D. Clayton and J. Cuzick, *J. Roy. Statist. Soc., series A* **148**: 108-109.

Gill, R.D. (1989). Non- and semi-parametric maximum likelihood estimation and the von Mises method (Part 1). *Scand. J. Statist.* **16**: 97-128.

Greenwood, M., and Yule, G.U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *J. R. Statist. Soc.* **83**: 255-279.

Hoffmann-Jørgensen, J. (1994). *Probability with a view to Statistics.* London: Chapman & Hall.

Honoré, B.E. (1990). Simple estimation of a duration model with unobserved heterogeneity. *Econometrica* **58**: 453-473.

Hougaard, P. (1984). Life table methods for heterogeneous populations: distributions describing the heterogeneity. *Biometrika* **71**, 75-83.

Hougaard, P. (1986). A class of multivariate failure time distributions. *Biometrika* **73**: 671-678.

Hougaard, P. (1987). Modelling multivariate survival. *Scand. J. Statist.* **14**: 291-304.

Hougaard, P., Harvald, B., and Holm, N.V. (1992). Models for multivariate failure time data, with application to the survival of twins. In van der Heijden, P.G.M., Jansen, W., Francis, B., and Seeber, G.U.H., editors, *Statistical Modelling*: 159-173. Elsevier Science Publishers B.V.

Hsu, L. Zhao, L.P. Li, H., and Parner, E. (1997). Genetic analysis of censored ages at onset from population based family studies. Preprint.

Jensen, H. (1992). Frailtymodeller, præsentation og anvendelse. Unpublished dissertation in Danish, University of Aarhus, Denmark.

Johansen, S. (1983). An extension of Cox's regression model. *Internat. Statist. Review* **51**: 258-262.

Kiefer, J., and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27**: 887-906.

Korsgaard, I.R., and Andersen, A.H. (1996). The additive genetic gamma frailty model. *Scand. J. Statist.* To appear.

Kortram, R.A., van Rooij, A.C.M., Lenstra, A.J., and Ridder, G. (1995). Constructive identification of the mixed proportional hazards model. *Statistica Neerlandica* **49**, no. 3: 269-281.

Maguluri, G. (1993). Semiparametric inference for association in a bivariate survival function. *Ann. Statist.* **21**: 1648-1662.

Murphy, S.A. (1994). Consistency in a proportional hazard model incorporating a random effect. *Ann. Statist.* **22**: 712-731.

Murphy, S.A. (1995). Asymptotic theory for the frailty model. *Ann. Statist.* **23**: 182-198.

Murphy, S.A., and van der Vaart, A.W. (1996a). Semiparametric likelihood ratio inference. *Ann. Statist.* **25**: To appear.

Murphy, S.A., and van der Vaart, A.W. (1996b). Observed information in semiparametric models. Preprint.

Murphy, S.A., Rossini, A.J., and van der Vaart, A.W. (1996c). MLE in the proportional odds model. *J. Amer. Statist. Assoc.* To appear.

Nielsen, G.G., Gill, R.D., Andersen, P.K., and Sørensen, T.I.A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Statist.* **19**: 25-44.

Oakes, D. (1989). Bivariate survival models induced by frailties. *J. Amer. Statist. Assoc.* **84**: 487-493.

Oakes, D., and Manatunga, A.K. (1992). Fisher information for a bivariate extreme value distribution. *Biometrika* **79**: 827-832.

Parner, E. (1996a). Consistency in the correlated Gamma-frailty model. Research Report 345, Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus.

Parner, E. (1996b). Asymptotic normality in the correlated Gamma-frailty model. Research Report 346, Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus.

Petersen, J.H., Andersen, P.K., and Gill, R.D. (1995). Variance components models for survival data. *Statistica Neerlandica* **50**: 193-211.

Petersen, J.H. (1996). A correlated frailty model. Research Report, Department of Biostatistics, University of Copenhagen.

Prentice, R.L., and Cai, J. (1992). Marginal and conditional models for the analysis of multivariate failure time data. *In* Klein, J.P., and Goel, P.K. (eds), *Survival analysis: state of the art*, Kluwer, Dordrecht: 393-406.

Pruitt, R.C. (1993). Identifiability of bivariate survival curves from censored data. *J. Amer. Statist. Assoc.* **88**: 573-579.

Yashin, A.I., and Iachine, I.A. (1994). Limitations of the shared frailty concept in the survival studies of relatives: correlated frailty as a better alternative. Preprint, Medical School, Odense University.

Yashin, A., Vaupel, J., and Iachine, I. (1995). Correlated individual frailty: an advantageous approach to survival analysis of bivariate data. To appear in *Mathematical Population Studies.*

van der Vaart, A.W. (1995). Efficiency of infinitely dimensional estimators. *Statistica Neerlandica* Vol **49**, no. 1: 9-30.

van der Vaart, A.W., and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes.* New York: Springer Verlag, Inc.

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

Vaupel, J.W., and Yashin, A.I. (1985). Heterogeneity's ruses: some surprising effect of selection on a population dynamics. *The American Statistician* **39**: 176-185.

Wald, A. (1949). Note on the consistency of maximum likelihood estimate. *Ann. Math. Statist.* **20**: 595-601.

Wei, L.J., Lin, D.Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distribution. *J. Amer. Statist. Assoc.* **84**: 1065-1073.

Zhao, L.P. (1994). Segregation analysis of human pedigrees using estimating equations. *Biometrika* **81**: 197-209.