The Annals of Statistics
1998, Vol. 26, No. 1, 183–214

ASYMPTOTIC THEORY FOR THE CORRELATED GAMMA-FRAILTY MODEL

BY ERIK PARNER

University of Aarhus

The frailty model is a generalization of Cox's proportional hazard model, where a shared unobserved quantity in the intensity induces a positive correlation among the survival times. Murphy showed consistency and asymptotic normality of the nonparametric maximum likelihood estimator (NPMLE) for the shared gamma-frailty model without covariates. In this paper we extend this result to the correlated gamma-frailty model, and we allow for covariates. We discuss the definition of the nonparametric likelihood function in terms of a classical proof of consistency for the maximum likelihood estimator, which goes back to Wald. Our proof of the consistency for the NPMLE is essentially the same as the classical proof for the maximum likelihood estimator. A new central limit theorem for processes of bounded variation is given. Furthermore, we prove that a consistent estimator for the asymptotic variance of the NPMLE is given by the inverse of a discrete observed information matrix.

0. Introduction. One of the standard assumptions in the analysis of survival data is that the individuals under observation are independent. This assumption can in many cases be questionable. A simple model for dependent survival times, which is a generalization of the proportional hazard model, is via the concept of frailty. This was first proposed by Vaupel, Manton and Stallard (1979). The motivation for the frailty model is that shared unobserved risk factors not included in the model give a dependence among a group of related survival times. Typical 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.

The frailty is usually modeled as an unobserved random variable acting multiplicatively on the baseline hazard functions. So if the hazard function for an individual with frailty 1 is $\alpha(u)$, then the hazard function of an individual with frailty value z is $z\alpha(u)$. If individuals in a group share the same value of z, then this is called a shared frailty model. Usually, it is assumed that the frailty follows a gamma distribution with mean 1 and unknown variance θ. The value $\theta = 0$ corresponds to independence and a high value of θ should preferably correspond to a high correlation between the survival times. The choice of the gamma distribution is made mostly for

Received June 1996; revised August 1997.
AMS 1991 *subject classifications*. 62M09, 62G05.
Key words and phrases. Nonparametric maximum likelihood estimation, survival data, heterogeneity, correlated frailty, central limit theorem, semiparametric models.

mathematical convenience. Other choices for the distribution of the frailty have been discussed in a series of papers by Hougaard [see Hougaard (1987), and the references therein]. Covariates are incorporated in the model by assuming that the conditional hazards, given the frailty, follow a Cox regression model, that is, $\alpha(u) = \exp(\boldsymbol{\beta}^{\top}\mathbf{x})\alpha_0(u)$, where $\mathbf{x}$ is a covariate and $\alpha_0(\cdot)$ is the baseline hazard.

It was shown in Elbers and Ridder (1982) that, with covariates in the model, frailty distributions with finite mean can be identified from marginal data, as for example the gamma distribution [see also Kortram, Van Rooij, Lenstra and Ridder (1995) for a constructive proof of the identification]. For example, for twin data this means that we can estimate the parameter θ knowing only one of the twins. Thus the parameter θ in this model describes something more than just the correlation between the survival times. This was one of the reasons which lead Yashin, Vaupel and Iachine (1995) to split the frailty for an individual, j, say, into two components, $z^{(j)} = z_0 + z_j$, where z_0 is a common component for all individuals and z_j is an individual component. The two frailties are independent, gamma-distributed random variables with the same scale parameters, but different shape parameters. Therefore also $z^{(j)}$ is gamma distributed. Now, the correlation between $z^{(j)}$ and $z^{(l)}$, $j \neq l$, cannot be identified from marginal data.

In this paper we use a counting process approach to the frailty model. This approach was first introduced by Gill (1985) in a discussion of the paper by Clayton and Cuzick (1985). He suggested estimating both the Euclidean part and the infinite-dimensional part of the parameter by nonparametric maximum likelihood estimation and, moreover, that this in application could be carried out by an EM-algorithm. This approach was further developed in Nielsen, Gill, Andersen and Sørensen (1992). Murphy (1994, 1995) showed consistency and asymptotic normality of the NPMLE in the shared frailty model without covariates, that is, where the parameters are the integrated hazard function and the variance of the gamma-distributed frailty. In this paper we extend these results to the correlated frailty model and we allow for covariates.

The paper is organized as follows: in Section 1 we define the correlated frailty model and derive the likelihood function. In Section 2 we discuss the definition of the nonparametric likelihood function in connection with a classical proof of consistency, which goes back to Wald (1949). We show in Section 3 that Wald's (1949) technique to prove consistency of the maximum likelihood estimator can be extended to prove the consistency of the NPMLE. In Section 4 we show that the NPMLE is asymptotically normal and prove that a consistent estimator for the asymptotic variance for the NPMLE is given by the inverse of a discrete information matrix.

1. The model. We shall use a counting process approach to the correlated frailty model. The model is a generalization of the shared frailty model presented by Nielsen, Gill, Andersen and Sørensen (1992).

Let $\mathbf{N} = (N_1, \ldots, N_m)$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ be a multivariate counting process and intensity process associated with a unit (e.g., family, pair of twins, group). Let $\mathbf{Y} = (Y_1, \ldots, Y_m)$ be a vector of left-continuous processes with right-hand limits (caglad), taking values in $\{0, 1\}$ and indicating (by the value 1) if the jth component of the unit is under observation. We consider $Y_j(\cdot)$ to be decreasing, corresponding to right censoring. So, $N_j(\cdot) \in \{0, 1\}$ and $N_j(\cdot)$ only jumps when $Y_j(\cdot)$ is equal to 1. Let $\mathbf{X}_j$ be a column vector of d caglad covariate processes for the jth component and let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_m)$. We assume that the intensity is given by

$$\lambda_j(\cdot|Z^{(j)}) = Z^{(j)} Y_j(\cdot) \exp\left(\boldsymbol{\beta}^{\mathsf{T}} \mathbf{X}_j(\cdot)\right) \alpha(\cdot), \qquad j = 1, \ldots, m, \tag{1}$$

where $Z^{(j)} = Z_0 + Z_j$ and $\mathbf{Z} = (Z_0, Z_1, \ldots, Z_m)$ are independent, unobservable, gamma-distributed random variables with parameters $(\nu, \eta), (\nu^*, \eta)$, $\ldots, (\nu^*, \eta)$, respectively. Here $\boldsymbol{\beta}$ is a column vector of d regression parameters and $\alpha(\cdot)$ is an unknown baseline hazard function. In this case $Z^{(j)}$ is gamma distributed with parameters $(\nu + \nu^*, \eta)$. In order to be able to identify both the parameters in the distribution of $\mathbf{Z}$ and $\alpha(\cdot)$, we restrict the mean of $Z^{(j)}$ to 1, which means that $\eta = \nu + \nu^*$. We assume that (1) is valid conditional on the value of the covariates and that the covariates follow some arbitrary distribution. By doing this, the covariates becomes exogenous.

It turns out to be convenient to use another parametrization. Instead of ν, ν^* we shall use the parameters $\theta = \operatorname{var}(Z_0) = \nu/\eta^2$ and $\theta^* = \operatorname{var}(Z_j) = \nu^*/\eta^2$. One could also reparametrize with $\theta_{\cdot} = \theta + \theta^*$ and $\delta = \operatorname{corr}(Z^{(j)}, Z^{(l)})$ as is Yashin, Vaupel and Iachine (1995), but this parametrization has the disadvantage that the parameters are not variation independent (if $\theta_{\cdot} = 0$ then $\rho = 1$). For $\theta^* = 0$ the model becomes the shared frailty model. In the following, we let ψ denote the parameters $(\theta, \theta^*, \boldsymbol{\beta}, \int \alpha\, du)$.

Assuming as in Nielsen, Gill, Andersen and Sørensen (1992) that the censoring is independent and noninformative of $\mathbf{Z}$ and $(\boldsymbol{\beta}, \alpha)$, the likelihood for the full but unobserved data set $(\mathbf{N}, \mathbf{Y}, \mathbf{Z})$ is given by

$$\prod_{j=1}^{m} \left\{ \prod_u \lambda_j(u|z^{(j)})^{\Delta N_j(u)} \exp\left(-z^{(j)} \Lambda_j\right) p\left(z_j; \theta^* \theta_{\cdot}^{-2}, \theta_{\cdot}^{-1}\right) \right\} p\left(z_0; \theta \theta_{\cdot}^{-2}, \theta_{\cdot}^{-1}\right),$$

where $\Lambda_j(u) = \int_0^u Y_j(s; \boldsymbol{\beta}) \alpha(s)\, ds = \int_0^u Y_j(s) \exp(\boldsymbol{\beta}^{\mathsf{T}} \mathbf{X}_j(s)) \alpha(s)\, ds$, $\Lambda_j = \Lambda_j(\tau)$ and $p(\cdot; \nu, \eta)$ denotes the gamma density with parameters (ν, η) [see also Parner, (1996a)]. Using the binomial formula, we get [$N_j(u) \in \{0, 1\}$, hence all the binomial coefficients are equal to one]

$$\begin{aligned} \prod_{j=1}^{m} (z_0 + z_j)^{N_j(\tau)} &= \sum_{k_1=0}^{N_1(\tau)} \cdots \sum_{k_m=0}^{N_m(\tau)} \prod_{j=1}^{m} z_0^{N_j(\tau) - k_j} z_j^{k_j} \\ &= \sum_{\mathbf{k} \in K(\tau)} \prod_{j=1}^{m} \left\{ z_j^{k_j} \right\} z_0^{N_{\cdot}(\tau) - k_{\cdot}}, \end{aligned}$$

where $\mathbf{k} = (k_1, \ldots, k_m)$, $K(u) = \{\mathbf{k} | k_j \in \{0, N_j(u)\}, j = 1, \ldots, m\}$, $k_\cdot = \sum_{j=1}^m k_j$ and $N_\cdot(u) = \sum_{j=1}^m N_j(u)$. For each term of the form $\prod_{j=1}^m \{z_j^{k_j}\} z_0^{N_\cdot(\tau)-k_\cdot}$, using the structure of the gamma density, it is easy to integrate the $\mathbf{z}$'s out in the full likelihood function. This gives that the likelihood function for $(\mathbf{N}, \mathbf{Y})$ observed up to a fixed time τ is

$$L_\tau^{N,Y}(\psi) = \prod_{j=1}^m \prod_{u \le \tau} \{Y_j(u; \boldsymbol{\beta}) \alpha(u)\}^{\Delta N_j(u)} \sum_{\mathbf{k} \in K(\tau)} a(\mathbf{k}, \tau), \tag{2}$$

where $a(\mathbf{k}, u) = a(\mathbf{k}, u; \psi)$ is given by

$$\prod_{j=1}^m \left\{ \frac{\Gamma(\theta^* \theta_\cdot^{-2} + k_j)}{\Gamma(\theta^* \theta_\cdot^{-2})} \frac{1}{(1 + \theta_\cdot \Lambda_j(u))^{\theta^* \theta_\cdot^{-2} + k_j}} \right\}$$
$$\times \frac{\Gamma(\theta \theta_\cdot^{-2} + N_\cdot(u) - k_\cdot)}{\Gamma(\theta \theta_\cdot^{-2})} \frac{\theta_\cdot^{N_\cdot(u)}}{(1 + \theta_\cdot \Lambda_\cdot(u))^{\theta \theta_\cdot^{-2} + N_\cdot(u) - k_\cdot}}$$

and $\Lambda_\cdot(u) = \sum_{j=1}^m \Lambda_j(u)$.

The conditional expectations of Z_0, Z_j given the σ-algebra $\mathcal{F}_u = \sigma(\mathbf{X}(s), \mathbf{N}(s), \mathbf{Y}(s): 0 \le s \le u)$ are

$$\hat{Z}_0(u) = \hat{Z}_0(u; \psi) = E_\psi[Z_0 | \mathcal{F}_u] = \frac{\sum_{\mathbf{k} \in K(u)} a(\mathbf{k}, u) b_0(\mathbf{k}, u)}{\sum_{\mathbf{k} \in K(u)} a(\mathbf{k}, u)},$$

$$\hat{Z}_j(u) = \hat{Z}_j(u; \psi) = E_\psi[Z_j | \mathcal{F}_u] = \frac{\sum_{\mathbf{k} \in K(u)} a(\mathbf{k}, u) b_j(\mathbf{k}, u)}{\sum_{\mathbf{k} \in K(u)} a(\mathbf{k}, u)},$$

where

$$b_0(\mathbf{k}, u) = \frac{\theta \theta_\cdot^{-1} + \theta_\cdot (N_\cdot(u) - k_\cdot)}{1 + \theta_\cdot \Lambda_\cdot(u)}, \qquad b_j(\mathbf{k}, u) = \frac{\theta^* \theta_\cdot^{-1} + \theta_\cdot k_j}{1 + \theta_\cdot \Lambda_j(u)}.$$

There is another way of deriving the likelihood for $(\mathbf{N}, \mathbf{Y})$. This is to use the innovation theorem [Bremaud (1981)] to the observed history $(\mathcal{F}_t)$. The $(\mathcal{F}_t)$-intensity of $\mathbf{N}$ is given by replacing $\mathbf{Z}$ by its conditional expectation with respect to the history $(\mathcal{F}_{t-})$, that is,

$$\lambda_j^{\mathcal{F}}(u) = \hat{Z}^{(j)}(u-) Y_j(u) \exp(\boldsymbol{\beta}^\top \mathbf{X}_j(u)) \alpha(u), \tag{3}$$

where $\hat{Z}^{(j)}(u) = \hat{Z}_0(u) + \hat{Z}_j(u)$. This implies that the likelihood function for $(\mathbf{N}, \mathbf{Y})$ can be written as

$$\prod_{j=1}^m \prod_{u \le \tau} \lambda_j^{\mathcal{F}}(u)^{\Delta N_j(u)} \exp\left(- \int_0^\tau \lambda_j^{\mathcal{F}}(u)\, du \right). \tag{4}$$

It was shown in Gill (1992) that (2) and (4) are equal if the censoring is noninformative of $\mathbf{Z}$.

Let $\{(\mathbf{N}_i, \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)\}_i$ be a sequence of i.i.d. replicates of $(\mathbf{N}, \mathbf{Y}, \mathbf{X}, \mathbf{Z})$. This will induce an additional index i on all the quantities defined in this section. Let ψ_0 denote the true value of ψ. We impose the following regularity conditions on the model.

CONDITION 1. (a) θ_0, θ_0^* lie in a known interval $[0, M[$.
(b) $\boldsymbol{\beta}_0$ is an interior point in a known compact subset $\mathscr{B} \subset \mathscr{R}^d$.
(c) The covariates $\mathbf{X}_j$ are uniformly bounded in variation norm.
(d) $\alpha_0(\cdot) > 0$, $\int_0^\tau \alpha_0(u)\, du < \infty$ and $\alpha_0(\cdot)$ caglad.

In Section 2 we shall show that it is natural to impose at least some bound on the variance parameters in order to ensure consistency of the estimators. Note that the second requirement in 1(d) implies that we can work with the supermums norm on the space of integrated hazard functions and also that $\tau < \infty$.

Let $P_0 = P_{\psi_0}$ and $Y_\cdot(u) = \sum_{j=1}^m Y_j(u)$. To ensure identifiability of the parameters we assume the next condition.

CONDITION 2. (e) $P_0(Y_\cdot(u) \geq 1 \text{ for all } u \in [0, \tau]) > 0$.
(f) $P_0(Y_\cdot(0) \geq 2) > 0$.
(g) If either $P_0(\mathbf{c}^\top \mathbf{X}_j(u) Y_j(u) = c_0 Y_j(u) \text{ for all } u \in [0, \tau],\ j = 1, \ldots, m) = 1$ or $P_0(\mathbf{c}^\top \mathbf{X}_j(0+) Y_j(0) = c_0 Y_j(0) \text{ for } j = 1, \ldots, m) = 1$ then $\mathbf{c} = \mathbf{0}$.
(h) For $\theta_0 = 0$, there exists a j and a $u \in [0, \tau]$ such that $Y_j(u)\exp(\boldsymbol{\beta}_0^\top \mathbf{X}_j(u))$ attains at least two different values other than zero.

Condition 2(e) ensures that we can observe failures on the entire interval and therefore be able to estimate A on the entire interval. This is also assumed in Andersen and Gill's (1982) treatment of the standard Cox regression model (for $m = 1$ and $\theta = \theta^* = 0$). Condition 2(f) is of course necessary, since we otherwise cannot identify the correlation $\theta\theta_\cdot^{-1}$ between Z_0 and Z_j. Note that we do not require that $Y_j(0) = 1$ with probability 1 for all j; that is, the number of components in the group could be random. For example, in litters of mice the size of the different litters are not necessarily the same and could even be equal to 1. The number m should therefore be seen as the maximum number of components in the group. It is useful to think of the model as being constructed given the size of the group, $S = s$, and then letting S follow some distribution on $\{1, \ldots, m\}$ with $P(S \geq 2) > 0$.

Condition 2(g) is to avoid colinearity among the covariates. Note that the second assumption in 2(g) is indeed necessary, otherwise the model is not identifiable. For the shared frailty model, the second assumption in 2(g) is not necessary. If Condition 2(e) is replaced by

$$P_0\big(Y_j(u) = 1 \quad \text{for all } u \in [0, \tau] | \mathbf{Z} = \mathbf{z}, \mathbf{X} = \mathbf{x}\big) > 0 \quad \text{for all } \mathbf{z} \text{ and } \mathbf{x},$$

then using that the covariates are exogeneous, the first probability in 2(g) is

$$\begin{aligned}
&P_0\big(\mathbf{c}^\top \mathbf{X}_j(u) Y_j(u) = c_0(u) Y_j(u) \text{ for all } u \in [0,\tau],\, j = 1,\ldots,m\big) \\
&\quad = \int P_0\big(\mathbf{c}^\top \mathbf{x}_j(\cdot) Y(\cdot) = c_0(\cdot) Y(\cdot) \\
&\qquad\qquad \text{for } j = 1,\ldots,m | \mathbf{Z} = \mathbf{z}, \mathbf{X} = \mathbf{x}\big)\, dP_{0ZX}(\mathbf{z},\mathbf{x}) \\
&\quad = \int P_0\big(\mathbf{c}^\top \mathbf{x}_j(\cdot) = c_0(\cdot) \text{ for } j = 1,\ldots,m | \mathbf{Z} = \mathbf{z}, \mathbf{X} = \mathbf{x}\big)\, dP_{0ZX}(\mathbf{z},\mathbf{x}) \\
&\quad = P_0\big(\mathbf{c}^\top \mathbf{X}_j(u) = c_0(u) \text{ for all } u \in [0,\tau],\, j = 1,\ldots,m\big),
\end{aligned}$$

which means that the covariates should, as processes, be affine independent.

The type of condition in 2(g) does not immediately appear in Andersen and Gill (1982). They assume instead that the observed (partial) information matrix converges to a strictly positive definite matrix, $\mathbf{\Sigma}$ say. Since Cox's partial likelihood function is a profile likelihood of the nonparametric likelihood function, it is relatively easy to show that if the Fisher information operator is one-to-one, then $\mathbf{\Sigma}$ is positive definite. For the Cox regression model, the Fisher information operator is one-to-one if and only if the equality $P_0(\mathbf{c}^\top \mathbf{X}_1(u) Y_1(u) = c_0(u) Y_1(u) \text{ for all } u \in [0,\tau]) = 1$ implies $\mathbf{c} = \mathbf{0}$. If the covariates are exogenous and we assume

$$P_0\big(Y_1(u) = 1 \text{ for all } u \in [0,\tau] | \mathbf{X}_1 = \mathbf{x}_1\big) > 0 \quad \text{for all } \mathbf{x}_1$$

then the probability can be rewritten as above. If the covariates are time independent then the condition becomes: if $P(\mathbf{c}^\top \mathbf{X}_1 = c_0) = 1$ holds then $\mathbf{c} = \mathbf{0}$. Surprisingly enough, this fact has not been noticed before.

For $\theta_0 = 0$, the components are independent and, as noted by Elbers and Ridder (1982), the model without covariates is not identifiable. Therefore condition 2(h) is assumed.

We shall now argue that it is possible to extend the model for strictly negative θ, θ^*. This allows for formal testing of whether the correlated frailty is appropriate, that is, if $\theta = 0$ or $\theta^* = 0$. For fixed θ, respectively θ^*, we have

$$\lim_{\theta^* \to 0} \hat{Z}^{(j)}(u;\psi) = \frac{1 + \theta N_\cdot(u)}{1 + \theta \int_0^u Y_\cdot(\boldsymbol{\beta})\, dA},$$

$$\lim_{\theta \to 0} \hat{Z}^{(j)}(u;\psi) = \frac{1 + \theta^* N_j(u)}{1 + \theta^* \int_0^u Y_j(\boldsymbol{\beta})\, dA}$$

and further

$$\lim_{\theta \to 0,\, \theta^* \to 0} \hat{Z}^{(j)}(u;\psi) = 1.$$

For θ^* negative, we can find an $\varepsilon^* = \varepsilon^*(\theta, A) > 0$ such that $\hat{Z}^{(j)}(u; \psi)$ is positive for $\theta^* \geq -\varepsilon$. Furthermore, we can find an ε sufficiently close to zero such that the likelihood function is sufficiently close to the limit

$$\prod_{j=1}^{m} \prod_{u} \{Y_j(u; \boldsymbol{\beta})\, \alpha(u)\}^{\Delta N_j(u)} \frac{\prod_u \{1 + \theta N_\cdot(u-)\}^{\Delta N_\cdot(u)}}{(1 + \theta \Lambda_\cdot(\tau))^{\theta^{-1} + N_\cdot(\tau)}},$$

for example, such that the relative difference is between 0.9 and 1.1. Similarly for θ negative, gives an $\varepsilon = \varepsilon(\theta^*, A) > 0$. The parameter space Ψ is then

$$\{\psi | (\boldsymbol{\beta}, A) \in \mathscr{B} \times \mathscr{A},\ \theta \in [-\varepsilon(\theta^*, A), M],\ \theta^* \in [-\varepsilon(\theta, A), M]\},$$

where $\mathscr{A}$ is the space of integrated hazard functions.

2. Nonparametric maximum likelihood estimation. In this section we discuss how the nonparametric maximum likelihood estimator (NPMLE) is defined. For simplicity we consider only the shared frailty model. The correlated frailty model is treated in exactly the same way. The log-likelihood function is

$$\begin{aligned} n \log L_n^{N,Y}(\theta, \boldsymbol{\beta}, A) = \sum_{i=1}^{n} \Bigg\{ & \sum_{j=1}^{m} \int_0^\tau \log\big(Y_{ij}(u; \boldsymbol{\beta})\, \alpha(u)\big)\, dN_{ij}(u) \\ & + \int_0^\tau \log(1 + \theta N_i.(u-))\, dN_i.(u) \\ & - \big(\theta^{-1} + N_i.(\tau)\big) \log\Big(1 + \theta \int_0^\tau Y_i.(u; \boldsymbol{\beta})\, dA(u)\Big) \Bigg\}. \end{aligned}$$

Fixing θ and $\boldsymbol{\beta}$, we see that the log-likelihood function tends to infinity as A tends to a discrete integrated hazard function with strictly positive jumps only at the observed failure times, $J_n = \{u | \exists i \leq n\colon \Delta N_i.(u) > 0\}$. Hence, the maximum likelihood estimator does not exist. This is similar to the ordinary Cox regression model, that is, the case where $\theta = 0$. This suggests that we should look for estimators where the integrated hazard function is discrete. One could try to extend the original model to allow for discrete integrated hazard functions and then use maximum likelihood estimation according to Kiefer and Wolfowitz (1956). However, there are many extensions of the frailty model that seem reasonable, depending on which aspect of the model one focuses on. These extensions need not all give maximum likelihood estimators which have nice asymptotic properties. So this way of defining the NPMLE will not in general give good estimators. In the ordinary Cox regression model, such an extension was constructed by Johansen (1983). A discussion of NPMLE in survival analysis is given in Andersen, Borgan, Gill and Keiding (1993). We shall motivate the NPMLE by making the connection to a classical method for proving consistency of the maximum likelihood estimator which goes back to Wald (1949). A more recent treatment of this

method is given in Hoffmann-Jørgensen (1994), Chapter 13. See also Groeneboom (1991) and Wijers (1995) for applications in the nonparametric setting.

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, $\hat{\psi}_n$, say, then

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

where μ denotes a measure dominating $P_{\hat{\psi}_n}$ and P_{ψ_0}. Assume that for any subsequence of $\{n\}$ we can find a further subsequence, $\{n_k\}$, such that $\hat{\psi}_{n_k} \to \psi$ 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

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

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 $\hat{\psi}_n \to \psi_0$.

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 shared frailty model $\psi = (\theta, \boldsymbol{\beta}, A)$, where A is an absolutely continuous integrated hazard function. To define the NPMLE we simply extend this 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, $\psi_n = (\theta_0, \boldsymbol{\beta}_0, A_n)$, 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 $\hat{\psi}_{n_k} \to \psi = (\theta, \boldsymbol{\beta}, A)$ 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

$$\int \log \frac{dP_\psi}{d\mu}\, dP_{\psi_0} - \int \log \frac{dP_{\psi_0}}{d\mu}\, dP_{\psi_0}. \tag{8}$$

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 $\hat{\psi}_n \to \psi_0$.

This way of motivating the NPMLE is new. A similar way of motivating the NPMLE was given in Gill (1989). He motivates the NPMLE by extending score functions for a class of one-dimensional submodels in as smooth a way as possible. In this way he gives an explanation why the NPMLE is asymptotically normal in cases where it is known in advance that the NPMLE is consistent. In practice, the smooth extension of the log-likelihood difference and the smooth extension of the score functions are the same. This way of proving the consistency was in principle also applied in Murphy (1994), but without explicitly making the connection to the classical proof of consistency of the maximum likelihood estimator. By making this connection, we have been able to further simply the proof of consistency of the NPMLE in the frailty model.

For the shared frailty model, assuming A_1 is absolutely A_2-continuous, the log-likelihood difference is

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\Bigg\{&\sum_{j=1}^{m}\left[\int_0^\tau \log\bigl(Y_{ij}(u;\boldsymbol{\beta}_1)\bigr)\,dN_{ij}(u)-\int_0^\tau \log\bigl(Y_{ij}(u;\boldsymbol{\beta}_2)\bigr)\,dN_{ij}(u)\right]\\&+\sum_{j=1}^{m}\int_0^\tau \log\bigl(dA_1(u)/dA_2(u)\bigr)\,dN_{ij}(u)\\&+\int_0^\tau \log\bigl(1+\theta_1 N_{i\cdot}(u-)\bigr)\,dN_{i\cdot}(u)-\int_0^\tau \log\bigl(1+\theta_2 N_{i\cdot}(u-)\bigr)\,dN_{i\cdot}(u)\\&-\bigl(\theta_1^{-1}+N_{i\cdot}(\tau)\bigr)\log\left(1+\theta_1\int_0^\tau Y_{i\cdot}(u;\boldsymbol{\beta}_1)\,dA_1(u)\right)\\&+\bigl(\theta_2^{-1}+N_{i\cdot}(\tau)\bigr)\log\left(1+\theta_2\int_0^\tau Y_{i\cdot}(u;\boldsymbol{\beta}_2)\,dA_2(u)\right)\Bigg\},\end{aligned}$$

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 in J_n, 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

$$\begin{aligned}L_n(\psi)=\frac{1}{n}\sum_{i=1}^{n}\Bigg\{&\sum_{j=1}^{m}\int_0^\tau \log\bigl(Y_{ij}(u;\boldsymbol{\beta})\,\Delta A(u)\bigr)\,dN_{ij}(u)\\&+\int_0^\tau \log\bigl(1+\theta N_{i\cdot}(u-)\bigr)\,dN_{i\cdot}(u)\\&-\bigl(\theta^{-1}+N_{i\cdot}(\tau)\bigr)\log\left(1+\theta\int_0^\tau Y_{i\cdot}(u;\boldsymbol{\beta})\,dA(u)\right)\Bigg\}.\end{aligned}$$

Using this expression, it is easily seen that the NPMLE for the integrated hazard function is discrete with jumps only at time points where we observe failures.

This type of extension works well for transformation models [the correlated frailty model is also a transformation model. For a definition of transformation models see Bickel, Klaassen, Ritov and Wellner (1993)]. By "smooth" extension, we really only mean that the extended log-likelihood difference should converge to minus the Kullback–Leibler information for n tending to infinity. For the shared frailty model, an extensive simulation study was done in Morsing (1994), showing in general good small sample properties for the NPMLE. In Pedersen (1995) a simulation study showed that these good small sample properties carries over to the correlated frailty model.

Let us consider the nonparametric log-likelihood function for the correlated frailty model

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{m}\int_0^\tau \log(\Delta A)\,dN_{ij} + \log\left(\sum_{\mathbf{k}\in K_i(\tau)} a_i(\mathbf{k},\tau)\right)\right\}. \tag{9}$$

We shall in the following give an argument for why there should at least be some bound on the variance parameter θ, θ^*. For ΔA tending to zero as n tends to infinity, the nonparametric log-likelihood function tends to minus infinity. If instead of ΔA in (9) we considered $n\,\Delta A$, then, for $n\,\Delta A(\cdot)$ converging to $\alpha(\cdot)\{E_0(ZY_\cdot(\cdot))\alpha_0(\cdot)\}^{-1}$, this normed nonparametric log-likelihood function evaluated at $(\theta, \theta^*, \boldsymbol{\beta}, \int n\,\Delta A\,d\overline{N}_\cdot)$ is asymptotically equal to the true log-likelihood function evaluated at $(\theta, \theta^*, \boldsymbol{\beta}, \int \alpha\,dt)$ (except for a constant which only depends on the true parameter). The normed nonparametric log-likelihood function is bounded below by

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{m}\int_0^\tau \log(n\,\Delta A)\,dN_{ij} + \log\big(a_i\big((N_{i1}(\tau),\ldots,N_{im}(\tau)),\tau\big)\big)\right\}\\
&\quad=\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{m}\int_0^\tau \log(n\,\Delta A)\,dN_{ij}\right.\\
&\qquad -\big(\theta^*\theta_\cdot^{-2} + N_{ij}(\tau)\big)\log\big(1+\theta_\cdot\Lambda_{ij}(\tau)\big)\\
&\qquad \left.+N_{ij}(\tau)\log\big(\theta^*\theta_\cdot^{-1}\big) - \theta\theta_\cdot^{-2}\log\big(1+\theta_\cdot\Lambda_{i\cdot}(\tau)\big)\right\}.
\end{aligned} \tag{10}$$

Consider values θ, θ^* such that $\theta^*/\theta_\cdot = p$. For $n\,\Delta A(\cdot)$ converging to the same function as above, (10) converges to

$$\begin{aligned}
&E_0\left\{\sum_{j=1}^{m}\int_0^\tau \log(\alpha)\,dN_j - \big(p\theta_\cdot^{-1} + N_j(\tau)\big)\log\left(1+\theta_\cdot\int_0^\tau Y_j\alpha\,du\right)\right.\\
&\qquad \left.+N_j(\tau)\log(p) - (1-p)\,\theta_\cdot^{-1}\log\left(1+\theta_\cdot\int_0^\tau Y_\cdot\alpha\,du\right)\right\}.
\end{aligned} \tag{11}$$

Reparametrizing with the observed integrated hazard function $\Lambda(u) = \theta_\cdot^{-1}\log(1+\theta_\cdot A(u))$ and the hazard function $\lambda(u) = \partial\Lambda(u)/\partial u$, then for Λ

fixed and $\theta_.$ conversing to infinity, (11) converges to

$$E_0\Bigg\{\sum_{j=1}^{m}\int_0^\tau \log(\lambda)\,dN_j - p\int_0^\tau Y_j\,d\Lambda + N_j(\tau)\log(p) - (1-p)\int_0^\tau 1\{Y_.>0\}\,d\Lambda\Bigg\}. \tag{12}$$

Thus, there is an asymptotic lower bound on the (normed) nonparametric log-likelihood function even for $\theta_.$ converging to infinity. Furthermore, for p close to 1, (12) is close to

$$E_0\Bigg(\sum_{j=1}^{m}\int_0^\tau \log(\lambda)\,dN_j - \int_0^\tau Y_j\,d\Lambda\Bigg),$$

the asymptotic log-likelihood function in the case where the components are independent. This means that in the limit we cannot necessarily rule out parameter values on the boundary of the unrestricted parameter space. It is, therefore, natural to impose a bound on the variance parameters.

3. Consistency. The plan for proving the consistency is as follows. In Proposition 1 we state that the NPMLE exists as a maximizer of the nonparametric likelihood function. In Proposition 2 we show that under Condition 2 the Kullback–Leibler information is strictly positive for $\psi \neq \psi_0$.

The next step is, for any subsequence of $\{\hat{\psi}_n\}$, to find a further convergent subsequence. First, we show that the sequence $\{\hat{\psi}_n\}$ stays bounded. Using this we can immediately find a convergent subsequence of the finite-dimensional part of the parameter. To find a convergent subsequence of the integrated hazard functions we write it as $\hat{A}_n(\cdot) = \int_0^\cdot (d\hat{A}_n/d\bar{N}_.)(s)\,d\bar{N}_.(s)$, where $\bar{N}_.(s) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}N_{ij}(s)$. Since $\{\hat{\psi}_n\}$ stays bounded, it follows that also $(d\hat{A}_n/d\bar{N}_.)(\cdot)$ stays bounded. Using that $\bar{N}_.(\cdot)$ converges by the law of large numbers, an application of the Helly's selection theorem [Hildebrandt (1963)] gives us a convergent subsequence of $\hat{A}_n(\cdot)$. This is done in the first part of Theorem 1.

Now we are finished if we can show that in the limit the inequality (5) is equal to the inequality (6). To derive the strong consistency we need to make sure that all the convergences take place on the same set of probability 1. If Ψ were compact, a standard application of the uniform law of large numbers would give the result. In our case, however, Ψ is not compact. Instead we use a version of the uniform law of large numbers which takes into account that the set of possible limit points is separable (Proposition 3 in Appendix A).

The proof of the following proposition is similar to the proof of NPMLE stays bounded, which is done in the proof of Theorem 1, and is therefore omitted [for details see Parner (1996a)].

PROPOSITION 1. *If $N_.(\tau) \geq 1$, then the supremum of L_n exists and is achieved.*

It follows from Proposition 1 that the maximum of the nonparametric likelihood function must be attained at a point where the partial derivatives for $\Delta A(u_l)$ are equal to zero, where u_l denotes a time of a failure. Let $\boldsymbol{\xi}$ denote the finite-dimensional part of the parameter. Taking the derivative of L_n with respect to $\Delta A(u_l)$ and setting it equal to zero, yields, for fixed $\boldsymbol{\xi}$, the following equation for $\hat{A}_n(\cdot) = \hat{A}_n(\cdot; \boldsymbol{\xi})$:

$$\hat{A}_n(u) = \int_0^u \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{Z}_i^{(j)}\big(\tau; \boldsymbol{\xi}, \hat{A}_n\big) Y_{ij}(s; \boldsymbol{\beta}) \right)^{-1} d\bar{N}_{\cdot\cdot}(s). \tag{13}$$

PROPOSITION 2. *Under Condition* 2, *the Kullback–Leibler information is strictly positive for* $\psi \neq \psi_0$.

PROOF. The Kullback–Leibler information is nonnegative and if it is equal to zero then P_ψ and P_{ψ_0} are equal on the σ-algebra $\mathscr{F}_\tau$. This means that if the model for the observed data is identifiable then the Kullback–Leibler information is strictly positive for $\psi \neq \psi_0$.

Let $\alpha_j(u) = \exp\{\boldsymbol{\beta}^\top \mathbf{X}_j(u)\}\alpha(u)$ and $A_j(u) = \int_0^u \alpha_j(s)\, ds$. Further let L_{Z^*}, L_{Z_0}, L_{Z_j} and L_{0Z^*}, L_{0Z_0}, L_{0Z_j} denote the Laplace transforms of $\mathbf{Z}^* = (Z^{(1)}, \ldots, Z^{(m)}), Z_0, Z_j$ under P_ψ and P_{ψ_0}. Dabrowska (1988) shows (for fixed group size) that the observed survival function can be identified under independent right censoring. However, we only know the observed survival function is of the form

$$L_{Z^*}\{A_1(s_1), \ldots, A_m(s_m)\} = L_{Z_0}\{A_1(s_1) + \cdots + A_m(s_m)\} \sum_{j=1}^{m} L_{Z_j}\{A_j(s_j)\} \tag{14}$$

for θ, θ^* nonnegative. In the following, we let L_{Z^*}, L_{Z_0}, L_{Z_j} denote the Laplace transforms extended to allow for negative values of θ, θ^*. Since P_ψ and P_{ψ_0} are equal on the σ-algebra $\mathscr{F}_\tau$ then the stochastic intensities with respect to $(\mathscr{F}_t)$ are equal. For a vector of nonnegative integers $\mathbf{a}$, we let

$$L^{\mathbf{a}}_{Z^*}(s_1, \ldots, s_m) = \frac{\partial^{a_1}}{\partial s_1^{a_1}} \cdots \frac{\partial^{a_m}}{\partial s_m^{a_m}} L_{Z^*}(s_1, \ldots, s_m)$$

[the derivative of $L_{Z^*}(s_1, \ldots, s_m)$, a_1 times with respect to s_1, a_2 times with respect to s_2 and so forth]. The stochastic intensity can be written in the form

$$\lambda_j^{\mathscr{F}}(u; \psi) = -\frac{L^{\mathbf{a}}_{Z^*}\{\Lambda_1(u), \ldots, \Lambda_m(u)\}}{L^{\mathbf{b}}_{Z^*}\{\Lambda_1(u), \ldots, \Lambda_m(u)\}} Y_j(u)\, \alpha_j(u), \tag{15}$$

where

$$\mathbf{a} = \big(N_1(u), \ldots, N_{j-1}(u), N_j(u) + 1, N_{j+1}(u), \ldots, N_m(u)\big),$$
$$\mathbf{b} = \big(N_1(u), \ldots, N_m(u)\big).$$

This formula for the stochastic intensity holds also for θ, θ^* negative. We shall show that from equality of the intensities we have equality of the

extended expression of the survival functions in (14). The proof of this step does not depend on the correlated gamma-frailty model in any way and holds for general frailty models. The final step is to identify the parameters from the observed survival function.

In the following we consider the conditional distribution given the censoring times, $c_1, \ldots, c_m$ say, and the covariates. Suppose that all the failure times are larger than τ. Summation over all the intensities gives

$$\frac{\partial}{\partial u} \log L_{Z^*}\{\Lambda_1(u), \ldots, \Lambda_m(u)\} = \frac{\partial}{\partial u} \log L_{0Z^*}\{\Lambda_{01}(u), \ldots, \Lambda_{0m}(u)\}$$

and therefore

$$L_{Z^*}\{A_1(u \wedge c_1), \ldots, A_m(u \wedge c_m)\} = L_{0Z^*}\{A_{01}(u \wedge c_1), \ldots, A_{0m}(u \wedge c_m)\}$$

for $u \in [0, \tau]$. (The formula is first seen to hold for all $u \in [0, \tau]$ except at the censoring times and by continuity for all $u \in [0, \tau]$.) From (15) it therefore follows that

$$\begin{aligned} &L^d_{Z^*}\{A_1(u \wedge c_1), \ldots, A_m(u \wedge c_m)\}\alpha_j(u) \\ &\qquad = L^d_{0Z^*}\{A_{01}(u \wedge c_1), \ldots, A_{0m}(u \wedge c_m)\}\alpha_{0j}(u) \end{aligned} \tag{16}$$

for $u \le c_j$, where $\mathbf{d}_k = 0$ for $k \neq j$ and $\mathbf{d}_j = 1$.

Suppose that the first component has a failure at time $v < c_1$ and the other failure times are larger than τ. For $u > v$, using a similar argument as above, we derive

$$\begin{aligned} &L^d_{Z^*}\{A_1(v), A_2(u \wedge c_2), \ldots, A_m(u \wedge c_m)\} \\ &\qquad = C(v) L^d_{0Z^*}\{A_{01}(v), A_{02}(u \wedge c_2), \ldots, A_{0m}(u \wedge c_m)\} \end{aligned}$$

for a constant $C(v)$. This holds for all $v \le c_1$ and all $v \le u \le \tau$. choosing $u = v$, we see from (16) that $C(v) = \alpha_0(v)/\alpha(v)$ and therefore

$$\begin{aligned} &L_{Z^*}\{A_1(v \wedge c_1), A_2(u \wedge c_2), \ldots, A_m(u \wedge c_m)\} \\ &\qquad = L_{0Z^*}\{A_{01}(v \wedge c_1), A_{02}(u \wedge c_2), \ldots, A_{0m}(u \wedge c_m)\} \end{aligned}$$

for all $v \le u \le \tau$. By a simple proof of induction we get that

$$\begin{aligned} &L_{Z^*}\{A_1(s_1 \wedge c_1), \ldots, A_m(s_m \wedge c_m)\} \\ &\qquad = L_{0Z^*}\{A_{01}(s_1 \wedge c_1), \ldots, A_{0m}(s_m \wedge c_m)\} \end{aligned} \tag{17}$$

for $0 \le s_1 \le s_2 \le \cdots \le s_m \le \tau$ and, by symmetry, (17) is valid for all $s_1, \ldots, s_m \in [0, \tau]$.

Let us first consider the case where there are no covariates. Let κ_0, κ_1 denote the extended versions of the cumulant transforms of Z_0, Z_1 under P_ψ. From the joint survival function we can by Condition 2(f) clearly identify

$$L_{Z_0}\{A(u)\} L_{Z_1}\{A(u)\},\ L_{Z_0}\{2A(u)\} L_{Z_1}\{A(u)\}^2$$

for $u \in [0, \delta]$, for some $\delta > 0$, and therefore also

$$\kappa_0\{2A(u)\} - 2\kappa_0\{A(u)\}. \tag{18}$$

We can also without loss of generality assume that A, A_0 are twice differentiable. Let $\kappa_0^{(l)}$, $\kappa_1^{(l)}$ denote the lth derivative of κ_0, κ_1. The second and third derivatives of (18) are

$$2\kappa_0^{(2)}(0)\,\alpha(0)^2, \tag{19}$$

$$6\kappa_0^{(3)}(0)\,\alpha(0)^3 + 6\kappa_0^{(2)}(0)\,\alpha(0)\,\alpha^{(1)}(0). \tag{20}$$

Similarly, from the first derivative of the marginal survival function we find $\alpha(0)$ and the second derivative is

$$\{\kappa_0^{(2)}(0) + \kappa_1^{(2)}(0)\}\alpha(0)^2 + \{\kappa_0^{(1)}(0) + \kappa_1^{(1)}(0)\}\alpha^{(1)}(0). \tag{21}$$

From (19) we find $\kappa_0^{(2)}(0) = -\theta$, whence $\theta = \theta_0$. From (20) we find

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

and using (21), we can identify

$$\frac{\kappa_0^{(3)}(0)}{\kappa_0^{(2)}(0)} - \kappa_0^{(2)}(0) - \kappa_1^{(2)}(0) = \theta_\cdot, \tag{22}$$

whence $\theta^* = \theta_0^*$.

The marginal survival functions are

$$L_{Z^{(j)}}\left(\int_0^u Y_j \alpha\, ds\right) = L_{Z^{(j)}}\left(\int_0^u Y_j \alpha_0\, ds\right).$$

Inverting the Laplace transforms gives $\int_0^u Y_\cdot \alpha\, ds = \int_0^u Y_\cdot \alpha_0\, ds$, which by condition 2(e) means that $A(\cdot)$, $A_0(\cdot)$ are equal with positive probability, and therefore that $\alpha(\cdot) = \alpha_0(\cdot)$.

Now we turn to the case where there are covariates in the model. The derivative of the logarithm of the marginal survival functions are

$$\begin{aligned} &\frac{1}{1 + \theta_\cdot \int_0^u Y_j(\boldsymbol{\beta})\, dA} Y_j(\boldsymbol{\beta}; u)\,\alpha(u) \\ &\quad = \frac{1}{1 + \theta_{0\cdot} \int_0^u Y_j(\boldsymbol{\beta}_0)\, dA_0} Y_j(\boldsymbol{\beta}_0; u)\,\alpha_0(u). \end{aligned} \tag{23}$$

For u tending to zero we get

$$\exp\left(\boldsymbol{\beta}^\top \mathbf{X}_j(0+)\right)\alpha(0+)Y_j(0) = \exp\left(\boldsymbol{\beta}_0^\top \mathbf{X}_j(0+)\right)\alpha_0(0+)Y_j(0).$$

By Condition 2(g) this implies that $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Let $\gamma(u) = \alpha(u)/\alpha_0(u)$. We can without loss of generality assume that $\gamma(\cdot)$ is differentiable. If $\theta_{0\cdot} = 0$ then from (23) we get

$$\gamma^{(1)}(u)Y_j(u) = \theta_\cdot \exp\left(\boldsymbol{\beta}_0^\top \mathbf{X}_j(u)\right)\alpha(u)Y_j(u)$$

and from Condition 2(g) it follows that $\theta_{\cdot} = 0$ and vice versa. Now consider the case where both $\theta_{\cdot}, \theta_{0\cdot}$ are not equal to zero. It follows from (23) that

$$\gamma(u)Y_j(u) = \frac{1 + \theta_{\cdot}\int_0^u Y(\boldsymbol{\beta}_0)\, dA}{1 + \theta_{0\cdot}\int_0^u Y(\boldsymbol{\beta}_0)\, dA_0} Y_j(u). \tag{24}$$

Integrating (23) gives

$$\theta_{\cdot}^{-1} \log\left(1 + \theta_{\cdot}\int_0^u Y_j(\boldsymbol{\beta}_0)\, dA\right) = \theta_{0\cdot}^{-1} \log\left(1 + \theta_{0\cdot}\int_0^u Y_j(\boldsymbol{\beta}_0)\, dA_0\right).$$

If $\theta_{\cdot} = \theta_{0\cdot}$, this equation implies that $\alpha(\cdot) = \alpha_0(\cdot)$. So suppose that $\theta_{\cdot} \neq \theta_{0\cdot}$. Using (24) we have

$$\theta_{0\cdot}\theta_{\cdot}^{-1} \log\{\gamma(u)\}Y_j(u) = \left(1 - \theta_{0\cdot}\theta_{\cdot}^{-1}\right)\log\left(1 + \theta_{0\cdot}\int_0^u Y_j(\boldsymbol{\beta}_0)\, dA_0\right)Y_j(u),$$

which implies that

$$\frac{\partial}{\partial u}\left\{\gamma(u)^{\theta_{0\cdot}/(\theta_{\cdot} - \theta_{0\cdot})}\right\}Y_j(u) = \theta_{0\cdot} \exp\left(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_j(u)\right)\alpha_0(u)Y_j(u).$$

By Condition 2(g) this gives a contradiction. Therefore $\theta_{\cdot} = \theta_{0\cdot}$ and hence $\alpha(\cdot) = \alpha_0(\cdot)$. The parameters θ and θ^* can be identified from the simultaneous intensities in the same way as above. This shows that the parameters are identifiable. □

Using the norm $\|\psi\| = |\boldsymbol{\xi}| \vee \sup_{u \in [0,\tau]}|A(u)|$ we have:

THEOREM 1. *Under Conditions* 1 *and* 2, $\hat{\psi}_n \to \psi_0$, P_0-*a.s.*

PROOF. The proof will be for ω fixed in a set of probability 1. The set is defined as a intersection of sets, each of probability 1, where the strong law of large numbers holds for some average. Hence in the proof we shall make sure that we only use the law of large numbers at most countably many times.

The first step is to show that $\{\hat{\psi}_n\}$ stays bounded, that is, $\limsup_n \hat{A}_n(\tau) < \infty$. If this is not the case, we can find a subsequence, $\{n_k\}$, such that $\hat{\boldsymbol{\xi}}_{n_k}$ converges to some $\boldsymbol{\xi} = (\theta, \theta^*, \boldsymbol{\beta})$ and $\hat{A}_{n_k}(\tau)$ tends to infinity. We shall show that if this is the case, then the nonparametric log-likelihood difference asymptotically becomes negative. Let us for simplicity call this subsequence $\{n\}$.

First consider the case where $\theta_{\cdot} = 0$. Assume we have chosen K large enough such that $K^{-1} \leq \exp(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_j(\cdot)) \leq K$ for all $\boldsymbol{\beta} \in \mathscr{B}$ and $j = 1, \ldots, m$. For n large we have

$$\hat{A}_n(\tau) \leq \left\{1 + \hat{\theta}_{n\cdot}K\hat{A}_n(\tau)\right\}K\left(\frac{1}{n}\sum_{i=1}^{n} 1\{Y_{i\cdot}(u) \geq 1 \text{ for all } u \in [0,\tau]\}\right)^{-1} \bar{N}_{\cdot\cdot}(\tau).$$

However, $\hat{\theta}_{n\cdot}$ converges to zero by assumption and the right-hand side is asymptotically smaller than the left-hand side, which gives us a contradiction.

Now consider the case where $\theta_\cdot > 0$. Using the inequality $\prod_{j=1}^m (1 + \theta_\cdot \Lambda_{ij})^{-1} \le (1 + \theta_\cdot \Lambda_{i\cdot})^{-1}$ we dominate $\alpha_i(\mathbf{k}, \tau)$ by

$$\prod_{j=1}^m \left\{ \frac{1}{\left(1 + \theta_\cdot \Lambda_{ij}\right)^{N_{ij}(\tau)}} \right\} \frac{\theta_\cdot^{N_{i\cdot}(\tau)}}{\left(1 + \theta_\cdot \Lambda_{i\cdot}\right)^{\theta_\cdot^{-1}}} \frac{\Gamma\left(\theta_\cdot^{-1} + N_{i\cdot}(\tau)\right)}{\Gamma\left(\theta_\cdot^{-1}\right)}.$$

From the inequality

$$K^{-1} \le \frac{1 + \theta_\cdot K^{-1} \int_0^\tau Y_{i\cdot}\, dA}{1 + \theta_\cdot \int_0^\tau Y_{i\cdot}\, dA} \le 1,$$

we can bound the log-likelihood difference $L_n(\hat{\psi}_n) - L_n(\psi_n)$, for $\psi_n = (\theta_0, \theta_0^*, \boldsymbol{\beta}_0, \overline{N}_{\cdot\cdot}(\cdot))$, above by

$$\begin{aligned} &O(1) + \sum_{j=1}^m \frac{1}{n} \sum_{i=1}^n \int_0^\tau \log\left(n\, \Delta \hat{A}_n\right) dN_{ij} \\ &\quad - \left(T^{-1} + N_{ij}(\tau)\right) \log\left(1 + \hat{\theta}_n \cdot \int_0^\tau Y_{ij}\, d\hat{A}_n\right), \end{aligned} \tag{25}$$

where $T = Mm$. Define $S_{ij} = \inf\{u | Y_{ij}(u) = 0\}$ with infimum over the empty set being defined as infinity. For a sequence $0 = x_0 \le x_1 \le \cdots \le x_{N-1} \le x_N = \tau$, split the jth term in (25) up in the following way:

$$\begin{aligned} &O(1) + \frac{1}{n} \sum_{i=1}^n \left\{ \int_0^\tau \log\left(n\, \Delta \hat{A}_n\right) dN_{ij}\, 1\{S_{ij} \in [\, x_{N-1}, \infty]\} \right. \\ &\qquad \left. - \left(T^{-1} + N_{ij}(\tau)\right) \log\left(1 + \hat{\theta}_n \cdot \int_0^\tau Y_{ij}\, d\hat{A}_n\right) 1\{S_{ij} \in [\, \tau, \infty]\} \right\} \\ &\quad + \sum_{k=1}^{N-1} \frac{1}{n} \sum_{i=1}^n \left\{ \int_0^\tau \log\left(n\, \Delta \hat{A}_n\right) dN_{ij}\, 1\{S_{ij} \in [\, x_{k-1}, x_k\,]\} \right. \\ &\qquad \left. - \left(T^{-1} + N_{ij}(\tau)\right) \log\left(1 + \hat{\theta}_n \cdot \int_0^\tau Y_{ij}\, d\hat{A}_n\right) 1\{S_{ij} \in [\, x_k, x_{k+1}]\} \right\}. \end{aligned}$$

For $N_{ij}^k(u) = N_{ij}(u) 1\{S_{ij} \in [\, x_{k-1}, x_k]\}$ we get from Jensen's inequality that

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \int_0^\tau \log\left(n\, \Delta \hat{A}_n\right) dN_{ij}\, 1\{S_{ij} \in [\, x_{k-1}, x_k\,]\} \\ &\quad \le \overline{N}_{\cdot\cdot}^k(\tau) \log\left(\overline{N}_{\cdot\cdot}^k(\tau)^{-1} \int_0^{x_k} n\, \Delta \hat{A}_n\, d\overline{N}_{\cdot\cdot}^k \right) \\ &\quad \le O(1) + \frac{1}{n} \sum_{i=1}^n N_{ij}(\tau) \log\left(\hat{A}_n(x_k)\right) 1\{S_{ij} \in [\, x_{k-1}, x_k\,]\}. \end{aligned}$$

Hence the jth term in (25) is dominated by

$$O(1) + \log\bigl(\hat{A}_n(\tau)\bigr)\frac{1}{n}\sum_{i=1}^{n}\Bigl\{N_{ij}(\tau)1\{S_{ij} \in [x_{N-1}, \infty]\} - \bigl(T^{-1} + N_{ij}(\tau)\bigr)1\{S_{ij} \in [\tau, \infty]\}\Bigr\} + \sum_{k=1}^{N-1}\log\bigl(\hat{A}_n(x_k)\bigr)\frac{1}{n}\sum_{i=1}^{n}\Bigl\{N_{ij}(\tau)1\{S_{ij} \in [x_{k-1}, x_k]\} - \bigl(T^{-1} + N_{ij}(\tau)\bigr)1\{S_{ij} \in [x_k, x_{k+1}]\}\Bigr\}.$$

We choose the sequence $\{x_k\}$ such that

$$E_0\bigl(N_j(\tau)1\{S_j \in [x_{N-1}, \tau[\}\bigr) < 1/2T^{-1}E_0\bigl(1\{S_j \geq \tau\}\bigr),$$
$$E_0\bigl(N_j(\tau)1\{S_j \in [x_{k-1}, x_k]\}\bigr) < E_0\Bigl(\bigl(N_j(\tau) + T^{-1}\bigr)1\{S_j \in [x_k, x_{k+1}]\}\Bigr).$$

This can be done by choosing $E_0(N_j(\tau)1\{S_j \in [x_k, x_{k+1}]\}) = \varepsilon$ for some $\varepsilon > 0$. [Since $u \to E_0(N_j(\tau)1\{S_j \in [0, u]\})$ need not be continuous, we may have to split these point masses into different groups.] Therefore the likelihood difference converges to minus infinity, which gives us a contradiction. Hence the NPMLE stays bounded.

Let $\{\tilde{n}_k\}$ be an arbitrary subsequence. Using that the NPMLE stays bounded it is straightforward to see that

$$\frac{d\hat{A}_n(u)}{d\overline{N}_{\cdot\cdot}(u)} \leq \text{const.}\left\{n^{-1}\sum_{i=1}^{n}Y_{i\cdot}(u)\right\}^{-1}.$$

Since $n^{-1}\sum_{i=1}^{n}Y_{i\cdot}(\cdot)$ converges to $E_0[Y_{\cdot}(\cdot)]$, P_0-a.s., [see, e.g., Rao (1963)] in supremum norm we get that

$$\limsup_n \sup_u d\hat{A}_n(u)/d\overline{N}_{\cdot\cdot}(u) < \infty.$$

From the Helly selection theorem [Hildebrandt (1963)] we can find a subsequence $\{n_k\} \subseteq \{\tilde{n}_k\}$ and an increasing function, A say, such that $\hat{A}_{n_k} \to A$ pointwise. Since $\overline{N}_{\cdot\cdot}(\cdot)$ converges to a continuous function, it follows that A is continuous and, using that $\{\hat{A}_n\}_n$ are all nondecreasing, we find that the convergence is also uniform. Furthermore, we can assume that along this subsequence, $\hat{\boldsymbol{\xi}}_{n_k}$ converges to some $\boldsymbol{\xi}$.

Define

$$A_n(\cdot) = \int_0^{\cdot}\left\{n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{Z}_i^{(j)}(u - ; \psi_0)Y_{ij}(u; \boldsymbol{\beta}_0)\right\}^{-1} d\overline{N}_{\cdot\cdot}(u).$$

An application of the Helly–Bray lemma gives that $A_n(\cdot)$ converges to $A_0(\cdot)$ in the supremum norm [see, e.g., Gill (1989)] and using Proposition 3 we get

$$\frac{d\hat{A}_{n_k}}{dA_{n_k}}(\cdot) \to \frac{E_0\left(\sum_{j=1}^m \hat{Z}^{(j)}(\cdot - ; \psi_0) Y_j(\cdot; \boldsymbol{\beta}_0)\right)}{E_0\left(\sum_{j=1}^m \hat{Z}^{(j)}(\tau; \psi) Y_j(\cdot; \boldsymbol{\beta})\right)} =: \gamma(\cdot), \tag{26}$$

in the supremum norm, P_0-a.s. Further, another application of Proposition 3 gives that

$$\begin{aligned}
0 &\le L_{n_k}\left(\hat{\psi}_{n_k}\right) - L_{n_k}\left(\psi_{n_k}\right) \\
&= \frac{1}{n}\sum_{i=1}^n \left\{ \sum_{j=1}^m \int_0^\tau \log\left(\frac{Y_{ij}\left(u; \hat{\boldsymbol{\beta}}_{n_k}\right) \Delta \hat{A}_{n_k}(u)}{Y_{ij}(u; \boldsymbol{\beta}_0)\, \Delta A_{n_k}(u)} \right) dN_{ij}(u) \right. \\
&\qquad \left. + \log\left(\sum_{\mathbf{k} \in K_i(\tau)} a_i\left(\mathbf{k}, \tau; \hat{\psi}_{n_k}\right) \right) - \log\left(\sum_{\mathbf{k} \in K_i(\tau)} a_i\left(\mathbf{k}, \tau; \psi_{n_k}\right) \right) \right\} \qquad (27) \\
&\to E_0\left\{ \sum_{j=1}^m \int_0^\tau \log\left(\frac{Y_j(u; \boldsymbol{\beta})}{Y_j(u; \boldsymbol{\beta}_0)} \gamma(u) \right) dN_j(u) \right. \\
&\qquad \left. + \log\left(\sum_{\mathbf{k} \in K(\tau)} a(\mathbf{k}, \tau; \psi) \right) - \log\left(\sum_{\mathbf{k} \in K(\tau)} a(\mathbf{k}, \tau; \psi_0) \right) \right\}, \quad P_0\text{-a.s.},
\end{aligned}$$

which is minus the Kullback–Leibler information.

To see an application of Proposition 3, we show that

$$\frac{1}{n_k}\sum_{i=1}^{n_k} \int_0^\tau \log\left(Y_{ij}\left(u; \hat{\boldsymbol{\beta}}_{n_k}\right) \right) dN_{ij}(u) \to E_0\left\{ \int_0^\tau \log\left(Y_j(u; \boldsymbol{\beta}) \right) dN_j(u) \right\} \quad P_0\text{-a.s.}$$

First note that, since the class of continuous function is separable in the uniform metric, the set of possible limit points is separable. Using that $\mathbf{X}_j$ is uniformly bounded, we have for all $\varepsilon > 0$ that there exists a $\delta > 0$ such that

$$\left| \int_0^\tau \log\left(Y_j(u; \boldsymbol{\beta}_1) \right) dN_j(u) - \int_0^\tau \log\left(Y_j(u; \boldsymbol{\beta}_2) \right) dN_j(u) \right| \le \varepsilon$$

for $|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2| \le \delta$. Therefore condition (46) in Proposition 3 is satisfied.

The Kullback–Leibler information is strictly positive (Proposition 2) and therefore we have that $\psi = \psi_0$. To summarize: for any given subsequence $\{\tilde{n}_k\}$, we have found a further subsequence $\{n_k\}$ such that $\hat{\psi}_{n_k} \to \psi_0$, whence $\hat{\psi}_n \to \psi_0$. □

4. Asymptotic normality. To calculate the score operator, we consider submodels of the form $t \to \psi_t := \psi + t(\mathbf{h}_\xi, \int_0 \cdot h_A\, dA)$, where $\mathbf{h}_\xi$ is a $(d+2)$-dimensional vector and h_A is a function of bounded variation. The score

operator is given by

$$S_n(\psi)(h) = \frac{\partial}{\partial t} L_n(\psi_t)|_{t=0},$$

and can be found in Appendix B. We shall consider h in the space $H_p = \{h = (\mathbf{h}_\xi, h_A) | \|h\|_H = |\mathbf{h}_\xi| + \|h_A\|_v \le p\}$, where $\|\cdot\|_v$ denotes the variation norm. Further, we shall restrict us to h_A which are either caglad or cadlag (cadlag means that at all time points it is continuous from the right and the limit from the left exists). This assumption is not necessary but it simplifies the arguments, because in this case the direction $(\mathbf{h}_\xi, \int_0^\cdot h_A\, dA) = 0$ if and only if $h = 0$. We can consider the parameter ψ as a functional on H_p given by $\psi(h) = \mathbf{h}_\xi^\top \boldsymbol{\xi} + \int_0^\tau h_A\, dA$ and the parameter space Ψ as a subset of $l^\infty(H_p)$, the space of bounded functionals on H_p, equipped with the supremum norm $\|\psi\|_p = \sup_{h \in H_p} |\psi(h)|$. In the following, we let $S(\psi)(h)$ denote the expectation of $S_1(\psi)(h)$.

In general we should choose enough submodels such that the information operator becomes invertible. The submodels $t \to \psi_t$ makes it very easy to show that the NPMLE is efficient (with respect to the tangent space generated by the submodels $t \to \psi_t$). Theorem 1 in van der Vaart (1995) gives that the NPMLE is a regular estimator sequence and it is proved in Theorem 2 that it is an asymptotically linear estimator sequence with influence function contained in the closed linear span of the tangent space. From Proposition 1 in van der Vaart (1995) it follows that the NPMLE is efficient.

The Fréchet derivative of $S(\psi)$ at ψ_0 is calculated by differentiating the score operator in the submodels $t \to \psi_0 + t\psi$. It can be written in the following form:

$$\begin{aligned}
-\dot{S}_{\psi_0}(\psi)(h) &= \iint i_{\xi_0\xi_0}(s,u) h_\xi(s) \#(ds)\, \xi(u) \#(du) \\
&\quad + \iint i_{A_0\xi_0}(s,u) h_A(s)\, ds\, \xi(u) \#(du) \\
&\quad + \iint i_{\xi_0 A_0}(s,u) h_\xi(s) \#(du)\, dA(u) \\
&\quad + \iint i_{A_0A_0}(s,u) h_A(s)\, ds\, dA(u) \\
&= \int \{\sigma_{\xi_0\xi_0}(h)(u) + \sigma_{A_0\xi_0}(h)(u)\} \xi(u) \#(du) \\
&\quad + \int \{\sigma_{\xi_0 A_0}(h)(u) + \sigma_{A_0A_0}(h)(u)\}\, dA(u) \\
&= \int \sigma_{\xi_0}(h)(u)\, \xi(u) \#(du) + \int \sigma_{A_0}(h)(u)\, dA(u),
\end{aligned}$$

where $\#(du)$ is the counting measure on the integers $1, \ldots, d+2$, $i_{\xi_0\xi_0}(k,l) = (E_0 \mathbf{L}_{\xi_0\xi_0})_{kl}$, $\xi(k) = \boldsymbol{\xi}_k$ for $k, l = 1, \ldots, d+2$ and so forth ($d+2$ is the

dimension of the finite part of the parameter space), and $\mathbf{L}_{\xi_0\xi_0}$ is the second derivative of the log-likelihood function with respect to $\boldsymbol{\xi}$. The operator $\sigma = (\sigma_{\xi_0}, \sigma_{A_0})$ is called the Fisher information operator. The operator $\dot{S}_{\psi_0}$ uniquely determines the Fisher information operator σ. As a shorthand notation, we write the right-hand side as

$$\iint i(s,u)h(s)\begin{pmatrix}\#(ds)\\ ds\end{pmatrix}\begin{pmatrix}\xi(u)\#(du)\\ dA(u)\end{pmatrix}.$$

We shall prove the asymptotic normality of the NPMLE by verifying the conditions in Theorem 3.3.1 of van der Vaart and Wellner (1996), here stated as Theorem 4 in Appendix A. The asymptotic normality of the score operator follows from a new central limit for processes of bounded variation (Lemma 2 in Appendix A). We expect this result to be useful in transformation models. To show that the Fisher information operator is continuously invertible, we shall show that it is one-to-one and that it can be written as a sum of a continuously invertible operator and a compact operator. This gives that it is continuously invertible [Rudin (1973)]. The continuously invertible operator is found by using the missing information principle [Woodbury (1971)]. This principle stats that we have the following relationship between the expected information matrices:

$$i_{N,Y} = i_{N,Y,Z} - E_0 i_{Z|N,Y}$$

when evaluating at the true parameter.

THEOREM 2. *Under Conditions* 1 *and* 2, *we have* $\sqrt{n}(\hat{\psi}_n - \psi_0) \Rightarrow \mathcal{G}$, *where* $\mathcal{G}$ *is a tight Gaussian process on* $l^\infty(H_p)$ *with zero mean and covariance process*

$$\operatorname{Cov}(\mathcal{G}(h), \mathcal{G}(g)) = \mathbf{h}_\xi^\top \boldsymbol{\sigma}_{\xi_0}^{-1}(g) + \int_0^\tau h_A \sigma_{A_0}^{-1}(g)\, dA_0,$$

and $\sigma = (\sigma_{\xi_0}, \sigma_{A_0})$ *is a continuously invertible linear operator from* H_∞ *onto* H_∞ *with inverse* $\sigma^{-1} = (\sigma_{\xi_0}^{-1}, \sigma_{A_0}^{-1})$. *The form of* σ *is given above.*

PROOF. We remind the reader that the algebraic form of the scores is given in Appendix B.

We start by verifying condition (d) in Theorem 4. In Proposition 2 it was shown that the Kullback–Leibler information is strictly positive for $\psi \neq \psi_0$, that is, for $\psi_{0t} = \psi_{0t}(h) = \psi_0 + t(\mathbf{h}_\xi, \int_0^\cdot h_A dA_0)$ the function $t \to E_0\{\log L_\tau^{N,Y}(\psi_{0t})\}$ has a maximum at zero. Using that ψ_0 is an interior point and that we can interchange expectation and differentiation, we derive $S(\psi_0) = 0$. Since $\hat{\psi}_n$ asymptotically is an interior point, we similarly get $S_n(\hat{\psi}_n)(h) = 0$ (asymptotically). In Theorem 1, $\hat{\psi}_n$ was shown to be consistent.

Now consider condition (b). From above it follows that $\sqrt{n}\{S_n(\psi_0) - S(\psi_0)\} = \sqrt{n}\, S_n(\psi_0)$. This we can rewrite as

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

where $\mathbf{V}_{\xi n} = \sqrt{n}\,\mathbf{L}_{\xi n}(\psi_0)$ and

$$V_{An}(u) = n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{m} N_{ij}(u) - \int_0^u \hat{Z}_i^{(j)}(\tau; \psi_0) Y_{ij}(s; \boldsymbol{\beta}_0)\, dA_0(s).$$

Here V_{An} is a sum of i.i.d. processes of uniformly bounded variation and it follows from Lemma 2 that V_{An} converges in distribution to a tight process. Therefore, $V_n = (\mathbf{V}_{\xi n}, \mathbf{V}_{An})$ converges to a tight process, V, say. Now consider the function Φ: $\mathscr{R}^{d+2} \times BV \to l^\infty(H_p)$ given by

$$\Phi(\mathbf{v}_\xi, v_A)(h) = \mathbf{h}_\xi^\top \mathbf{v}_\xi + \int_0^\tau h_A\, dv_A,$$

where BV denotes the space of functions with finite variation and Φ is continuous. From the continuous mapping theorem [see, e.g., van der Vaart and Wellner (1996)] we get that $\sqrt{n}\, S_n(\psi_0)$ converges in distribution to a tight Gaussian process, which we denote by $\mathscr{G}$.

The submodel $t \to \psi_{0t}$ is a regular parametric submodel, and we can calculate the asymptotic variance as

$$\begin{aligned} \operatorname{Var}(\mathscr{G}(h)) &= E_0\left(-\frac{\partial^2}{\partial t^2} \log L_\tau^{N,Y}(\psi_{0t})|_{t=0} \right) \\ &= -\dot{S}_{\psi_0}\left(\mathbf{h}_\xi, \int_0^{\cdot} h_A\, dA_0 \right)(h), \end{aligned} \tag{28}$$

where $\dot{S}_{\psi_0}$ is the Fréchet derivative of S at ψ_0 (see below). Similarly, the asymptotic covariance can be calculated by considering two-dimensional submodels $(s,t) \to (\boldsymbol{\xi}_0 + s\mathbf{h}_\xi + t\mathbf{g}_\xi, A_0(\cdot) + s\int_0^{\cdot} h_A\, dA_0 + t\int_0^{\cdot} g_A\, dA_0)$ and differentiating at $(s,t) = (0,0)$, which gives

$$\operatorname{Cov}(\mathscr{G}(h), \mathscr{G}(g)) = -\dot{S}_{\psi_0}\left(\mathbf{g}_\xi, \int_0^{\cdot} g_A\, dA_0 \right)(h).$$

We shall now show that S is Fréchet differentiable. First note that S is Gâteaux differentiable and, since $(\mathbf{N}, \mathbf{Y}, \mathbf{X})$ is uniformly bounded, it follows that the derivative is continuous. It is relatively easy to see that

$$\sup\left\{ \left\| \frac{\partial}{\partial t} S(\psi_0 + t\psi) \right\|_p : \|\psi\|_p \le 1, |t| \le \varepsilon \right\} < \infty$$

for an $\varepsilon > 0$. It follows from Bickel, Klaassen, Ritov and Wellner [(1993), Proposition 1, page 455], that S is Fréchet differentiable and that the derivative is given by

$$\dot{S}_{\psi_0}(\boldsymbol{\xi}, A)(h) = -\left\{ \boldsymbol{\sigma}_{\xi_0}(h)\boldsymbol{\xi} + \int_0^\tau \sigma_{A_0}(h)\, dA \right\}. \tag{29}$$

Thus, the variance in (28) is given by $\boldsymbol{\sigma}_{\xi_0}(h)\mathbf{h}_\xi + \int_0^\tau \sigma_{A_0}(h) h_A\, dA_0$.

Continuous invertibility of $\dot{S}_{\psi_0}$ for some p is equivalent to the fact that for some $\varepsilon > 0$,

$$\inf_{\psi \in \operatorname{lin} \Psi} \frac{\|\dot{S}_{\psi_0}(\psi)\|_p}{\|\psi\|_p} > \varepsilon \tag{30}$$

[see, e.g., Bickel, Klaassen, Ritov and Wellner (1993), Proposition 7, page 418]. To prove (30) we shall show that σ, viewed as an operator from H_∞ to H_∞, is onto and continuously invertible. This means that for all $p > 0$ there exists a $q > 0$ such that $\sigma^{-1}(H_q)$ is contained in H_p. In this case the term on the left-hand side of (30) is bounded below by

$$\inf_{\psi \in \operatorname{lin} \Psi} \frac{\sup_{h \in \sigma^{-1}(H_q)} |\boldsymbol{\sigma}_{\xi_0}(h)\boldsymbol{\xi} + \int_0^\tau \sigma_{A_0}(h)\, dA|}{\|\psi\|_p}$$

$$= \inf_{\psi \in \operatorname{lin} \Psi} \frac{\sup_{h \in H_q} |\mathbf{h}_\xi^\top \boldsymbol{\xi} + \int_0^\tau h_A\, dA|}{\|\psi\|_p},$$

which is larger than $q/3p$. To verify that σ is continuously invertible, we show that σ is one-to-one and write σ as sum of a continuously invertible operator, Σ, and a compact operator, C. This implies that σ is continuously invertible [see, e.g., Rudin (1973)].

That σ is one-to-one means that if $\|h\| > 0$, then $\|\sigma(h)\| > 0$. Suppose this is not the case, that is, $\sigma(h) = 0$ for some h. Then we trivially have

$$0 = \boldsymbol{\sigma}_{\xi_0}(h)\mathbf{h}_\xi + \int_0^\tau \sigma_{A_0}(h) h_A\, dA_0 = E_0\left(\frac{\partial}{\partial t} \log L_\tau^{N,Y}(\psi_{0t})|_{t=0} \right)^2,$$

therefore

$$0 = \frac{\partial}{\partial t} \log L_\tau^{N,Y}(\psi_{0t})|_{t=0}, \tag{31}$$

P_0-a.s. or, equivalently,

$$0 = \frac{\partial}{\partial t} L_\tau^{N,Y}(\psi_{0t})|_{t=0}, \tag{32}$$

P_0-a.s. Let Ω denote the probability space on which all the random functions are defined. Integrating (32) over the set $\Omega_j = \{\omega \in \Omega | (N_j, Y_j)(\omega) = (n_j, y_j)\}$ yields

$$0 = \int_{\Omega_j} \frac{\partial}{\partial t} L_\tau^{N,Y}(\psi_{0t})(\omega)|_{t=0}\, dP_0(\omega)$$

$$= \frac{\partial}{\partial t} \int_{\Omega_j} L_\tau^{N,Y}(\psi_{0t})(\omega)\, dP_0(\omega)|_{t=0}$$

$$= \frac{\partial}{\partial t} L_\tau^{N_j,Y_j}(\psi_{0t})|_{t=0},$$

P_0-a.s., that is, also the marginal score functions are identically equal to zero. Similarly, integrating over the set $\{\omega \in \Omega | (N_j(s), Y_j(s))(\omega) = (n_j(s), y_j(s)), s \in [u, \tau]\}$ we derive

$$0 = \frac{\partial}{\partial t} L_u^{N_j, Y_j}(\psi_{0t})|_{t=0}, \tag{33}$$

P_0-a.s. Let us first consider the case where there are covariates in the model. From (33) we have the following set of marginal score equations with $h_{\theta.} = h_\theta + h_{\theta^*}$:

$$\begin{aligned} 0 = h_{\theta.} \Bigg\{ & \theta_0^{-2} \cdot \log\Bigg(1 + \theta_0 . \int_0^u Y_j \, dA_j\Bigg) \\ & - \frac{1 + \theta_0 . N_j(u)}{1 + \theta_0 . \int_0^u Y_j \, dA_j} \theta_0^{-1} . \int_0^u Y_j \, dA_j \Bigg\} \\ & + \int_0^u \left(h_A + \mathbf{h}_\beta^\top \mathbf{X}_j\right) dN_j - \frac{1 + \theta_0 . N_j(u)}{1 + \theta_0 . \int_0^u Y_j \, dA_j} \int_0^u \left(h_A + \mathbf{h}_\beta^\top \mathbf{X}_j\right) Y_j \, dA_j, \end{aligned} \tag{34}$$

P_0-a.s. For simplicity we shall assume that $\theta_0 . \neq 0$. Let $T_j = \inf\{u | N_j(u) > 0\}$. Then (34) is valid $P_0(\cdot | \mathbf{X}_j = \mathbf{x}_j)$-a.s. for P_0-a.a. $\mathbf{x}_j$. Since we can choose u arbitrary close to zero and since $P_0(T_j \leq u | \mathbf{X}_j = \mathbf{x}_j) > 0$ for all $u > 0$, we get $h_A(0) + \mathbf{h}_\beta^\top \mathbf{x}_j(0 +) = 0$ with positive probability, and hence with probability 1, with respect to $P_0(\cdot | \mathbf{X}_j = \mathbf{x}_j)$ for P_0-a.a. $\mathbf{x}_j$. This implies that $h_A(0) + \mathbf{h}_\beta^\top \mathbf{X}_j(0 +) = 0$, P_0-a.s. By Condition 2(g) we get that $\mathbf{h}_\beta = \mathbf{0}$.

Let us consider $u < T_j$ and again the conditional distribution given $\mathbf{X}_j = \mathbf{x}_j$. From (34) we derive

$$h_A(u) = h_{\theta.} \log\Bigg(1 + \theta_0 . \int_0^u \exp\left(\boldsymbol{\beta}_0^\top \mathbf{x}_j\right) dA_0\Bigg)$$

for all u with positive probability, and therefore with probability 1. Hence

$$h_A(u) = h_{\theta.} \log\Bigg(1 + \theta_0 . \int_0^u \exp\left(\boldsymbol{\beta}_0^\top \mathbf{X}_j\right) dA_0\Bigg)$$

for all u, P_0-a.s. This equation implies that $h_A = 0$ and $h_{\theta.} = 0$. To conclude that $h_\theta = h_{\theta^*} = 0$, we get in a similar way as before that the score function for $\{\mathbf{N}(s), \mathbf{Y}(s) : 0 \leq s \leq u\}$ is identically equal to zero. For $u < \inf\{s | N.(s) > 0\}$ we therefore have

$$0 = h_\theta \Bigg(- \theta_0^{-2} \cdot \sum_{j=1}^m \log\left(1 + \theta_0 . \Lambda_j(u)\right) + \theta_0^{-2} \cdot \log\left(1 + \theta_0 . \Lambda .(u)\right)\Bigg).$$

The quantity inside the brackets, however, is strictly negative and therefore $h_\theta = h_{\theta^*} = 0$. So in the case where there are covariates, the information operator is one-to-one.

We turn to the case where there are no covariates. Consider again the marginal score (34) for $u < T_j$. Taking the derivative of (34), we get

$$0 = h_{\theta\cdot}\frac{A_0(u)}{1 + \theta_0 \cdot A_0(u)} + \frac{\theta_0\cdot}{1 + \theta_0 \cdot A_0(u)}\int_0^u h_A\, dA_0 - h_A(u). \tag{35}$$

From this we see that $h_A(\cdot)$ is differentiable. Letting u tend to zero we get $h_A(0) = 0$. Taking the derivative of (35) we have

$$\begin{aligned} 0 = {} & h_{\theta\cdot}\left\{\frac{-\theta_0 \cdot A_0}{(1 + \theta_0 \cdot A_0)^2} + \frac{1}{1 + \theta_0 \cdot A_0}\right\} \\ & - \frac{\theta_0^2\cdot}{(1 + \theta_0 \cdot A_0)^2}\int_0^{\cdot} h_A\, dA_0 + \frac{\theta_0 \cdot h_A}{1 + \theta_0 \cdot A_0} - h_A^1 \end{aligned} \tag{36}$$

evaluated at u, where $h_A^1(u)$ is the derivative of $h_A(u)$ divided by $\alpha_0(u)$. For u tending to zero this implies that $h_{\theta\cdot} - h_A^1(0) = 0$. From (36) we see that $h_A^1(u)$ is differentiable, and taking the derivative of (36) we find that

$$\begin{aligned} 0 = {} & h_{\theta\cdot}\left\{2\frac{\theta_0^2 \cdot A_0}{(1 + \theta_0 \cdot A_0)^3} - 2\frac{\theta_0\cdot}{(1 + \theta_0 \cdot A_0)^2}\right\} \\ & + 2\frac{\theta_0^3\cdot}{(1 + \theta_0 \cdot A_0)^3}\int_0^{\cdot} h_A\, dA_0 \\ & - 2\frac{\theta_0^2\cdot}{(1 + \theta_0 \cdot A_0)^2}h_A + \frac{\theta_0\cdot}{1 + \theta_0 \cdot A_0}h_A^1 - h_A^2 \end{aligned} \tag{37}$$

evaluated at u, where $h_A^2(u)$ is the derivative of $h_A^1(u)$ divided by $\alpha_0(u)$. For u tending to zero we derive

$$-2h_{\theta\cdot}\theta_0^2\cdot + \theta_0 \cdot h_A^1(0) - h_A^2(0) = -h_{\theta\cdot}\theta_0^2\cdot - h_A^2(0) = 0. \tag{38}$$

Let $N_1(\cdot)$ and $N_2(\cdot)$ denote the counting processes for the two components which from Condition 2(g) is present at time zero with positive probability. With some tedious, but straightforward, calculation it can be shown that taking the second derivative of the score function of $\{N_j(s), Y_j(s)\colon 0 \le s \le u,\ j = 1, 2\}$, with respect to u, and thereafter letting u tend to zero, one obtains

$$2h_\theta + h_{\theta^*} - h_A^1(0) = 0, \tag{39}$$

which implies $h_\theta = 0$. Similarly, the third derivative of the score function of the data $\{N_j(s), Y_j(s)\colon 0 \le s \le u,\ j = 1, 2\}$, evaluated at zero, gives

$$-(10\theta_0 + 4\theta_0^*)h_{\theta^*} + 3(2\theta_0 + \theta_0^*)h_A^1(0) - h_A^2(0) = 0, \tag{40}$$

which together with equations (38) implies that $h_{\theta^*} = 0$. From the marginal score function it is straightforward to see that $h_A = 0$. Hence the Fisher information operator is one-to-one.

For the correlated frailty model Σ is of the form $\Sigma(h) = \Sigma_{\xi_0}(\mathbf{h}_\xi) + \Sigma_{A_0}(h_A)$, and therefore continuously invertible if and only if both Σ_{ξ_0} and Σ_{A_0} are continuously invertible. We choose $\Sigma_{\xi_0}(\mathbf{h}_\xi) = \mathbf{h}_\xi^\top E_0 \mathbf{L}_{\xi_0\xi_0}$. Since Σ_{ξ_0} is one-to-one and a finite-dimensional operator, it is also continuously invertible. To define Σ_{A_0} we use the missing information principle [Woodbury (1971)] for the one-dimensional submodels $t \to \psi_{0t}(\mathbf{0}, h_A)$. This gives

$$\sigma^{N,Y}_{A_0,A_0}(h_A) = \sigma^{N,Y,Z}_{A_0,A_0}(h_A) - E_0\,\sigma^{Z|N,Y}_{A_0,A_0}(h_A),$$

where $\sigma^{N,Y,Z}_{A_0,A_0}(h_A)$ is the expected information operator in the Cox regression model. It is well known [see, e.g., Bickel, Klaassen, Ritov and Wellner (1993)], and also very easy to verify, that this operator is continuously invertible. Then we define $\Sigma_{A_0}(h_A) = \sigma^{N,Y,Z}_{A_0,A_0}(h_A)$. To show that $\sigma(h) - \Sigma(h)$ is compact, we show that for an arbitrary sequence, $\{h_n\}_{n\geq 1}$, there exists a convergent subsequence of $\{\sigma(h_n) - \Sigma(h_n)\}$. It follows from Helly's selection theorem that there exists a subsequence, $\{n_k\}_{k\geq 1}$, and a function, h_A, such that h_{An_k} converges pointwise to h_A. We can choose the subsequence such that $\mathbf{h}_{\xi n_k}$ converges to $\mathbf{h}_\xi$ for some vector $\mathbf{h}_\xi$. Using the dominated convergence theorem, it is easy to see that along this subsequence $\sigma(h_{n_k}) - \Sigma(h_{n_k})$ converges to $\sigma(h) - \Sigma(h)$.

To prove the approximation condition in (a) we use Lemma 1 in Appendix A. To verify (47), we show that the difference of score functions is built up by functions which are P_0-Donsker and then use that, under regularity conditions, sums, products and functions of P_0-Donsker classes again form P_0-Donsker classes [see van der Vaart and Wellner (1996), Examples 2.10.8 and 2.10.9 and Theorem 2.10.6]. To demonstrate how it works, we shall show that the term (for $\theta_\cdot \neq 0$)

$$\theta\theta_\cdot^{-2}\,\log\left(1 + \theta_\cdot \int_0^\tau y_\cdot(\boldsymbol{\beta})\,dA\right), \tag{41}$$

where $y_\cdot(u;\boldsymbol{\beta}) = \sum_{j=1}^m y_j(u;\boldsymbol{\beta}) = \sum_{j=1}^m y_j(u)\exp(\boldsymbol{\beta}^\top \mathbf{x}_j(u))$ is a P_0-Donsker class for $\|\psi - \psi_0\| < \delta$, $h \in H_p$ and $\delta > 0$ chosen small enough. First, using Lemma 2, have that $\mathbf{x}_j(\cdot)$ is P_0-Donsker and the class $\{\boldsymbol{\beta}\colon \|\psi - \psi_0\| < \delta\}$ is trivially P_0-Donsker, so by multiplying two P-Donsker classes we get that $\{\boldsymbol{\beta}^\top \mathbf{x}_j(\cdot)\colon \|\psi - \psi_0\| < \delta\}$ is P_0-Donsker. The exponential function is Lipschitz on compact sets of the real line. (This follows from a first-order Taylor expansion.) Since $\mathbf{x}_j$ is uniformly bounded, we get from van der Vaart and Wellner (1996), Theorem 2.10.6, that the class $\{\exp(\boldsymbol{\beta}^\top \mathbf{x}_j(\cdot))\colon \|\psi - \psi_0\| < \delta\}$ is P_0-Donsker. Lemma 2 gives that $y_j(\cdot)$ is P_0-Donsker, so by taking product over two Donsker classes, we find that the class $\{y_j(\cdot;\boldsymbol{\beta})\colon \|\psi - \psi_0\| < \delta\}$ and hence $\{y_\cdot(\cdot;\boldsymbol{\beta})\colon \|\psi - \psi_0\| < \delta\}$ is P_0-Donsker. Let $\mathscr{B}'$ denote the set $\{\boldsymbol{\beta}\colon |\boldsymbol{\beta} - \boldsymbol{\beta}_0| \leq \delta\}$ and let BV_p denote the set of uniformly bounded function with variation norm smaller than p. Define the function ϕ from $l^\infty([0,\tau] \times \mathscr{B}')$ to $l^\infty(\mathscr{B}' \times BV_p)$ by $\phi(y)(\boldsymbol{\beta}, f) = \int_0^\tau y(u,\boldsymbol{\beta})\,df(u)$. The function ϕ is continuous and from the continuous mapping theorem it follows that $\{\int_0^\tau y_\cdot(\boldsymbol{\beta})\,df\colon \boldsymbol{\beta} \in \mathscr{B}',\ f \in BV_p\}$ is P_0-Donsker for all p and hence $\{\int_0^\tau y_\cdot(\boldsymbol{\beta})\,dA\colon \|\psi - \psi_0\| < \delta\}$ is P_0-Donsker.

Using that the sets $\{\theta\theta_\cdot^{-2}: \|\psi - \psi_0\| < \delta\}$ and $(\theta_\cdot: \|\psi - \psi_0\| < \delta\}$ are P_0-Donsker and the function $x \to \log(1 + x)$ satisfies

$$|\log(1 + x) - \log(1 + y)| \le |x - y|,$$

it follows from Theorem 2.10.6 in van der Vaart and Wellner (1996) that $\{\theta\theta_\cdot^{-2} \log(1 + \theta_\cdot \int_0^\tau y_\cdot(\boldsymbol{\beta})\, dA)$: $\|\psi - \psi_0\| < \delta\}$ is P_0-Donsker. In a similar way, the other terms are shown to be P_0-Donsker classes and condition (47) is satisfied.

Considering the condition in (48), for ψ converging to ψ_0, we have that

$$\theta\theta_\cdot^{-2} \log\left(1 + \theta_\cdot \int_0^\tau Y_\cdot(\boldsymbol{\beta})\, dA\right) - \theta_0 \theta_{0\cdot}^{-2} \log\left(1 + \theta_{0\cdot} \int_0^\tau Y_\cdot(\boldsymbol{\beta}_0)\, dA_0\right)$$

converges to zero pointwise, and by the dominated convergence theorem this convergence is also valid in L^2. In a similar way, the other terms are shown to converge in L^2. Hence, condition (a) in Theorem 4 is satisfied.

According to Theorem 4, the asymptotic distribution of $\sqrt{n}(\hat{\psi}_n - \psi_0)(h)$ is

$$-\dot{S}_{\psi_0}^{-1}(\mathscr{G}(h)) = -\mathscr{G}(\sigma^{-1}(h)),$$

and the result follows. □

Let us consider the problem of calculating the variance of $\mathscr{G}$. According to Theorem 2, the asymptotic variance of $\sqrt{n}(\hat{\psi}_n - \psi_0)(g) = \sqrt{n}\{\mathbf{g}_\xi^\mathsf{T}(\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0) + \int_0^\tau g_A\, d(\hat{A}_n - A_0)\}$ is $\mathbf{g}_\xi^\mathsf{T} \boldsymbol{\sigma}_{\xi_0}^{-1}(g) + \int_0^\tau g_A \sigma_{A_0}^{-1}(g)\, dA_0$. A natural estimate for this is obtained by estimating $\sigma = \sigma(\psi_0)$ by the observed information operator $\hat{\sigma}_n = \sigma_n(\hat{\psi}_n)$ and then inverting $\hat{\sigma}_n$. Below we show that this is a consistent estimator for σ. Let $\{u_l\}_{l \ge 1}$ denote the points where $N_{\cdot\cdot}$ jumps. Define the observed discrete information matrix, $\mathbf{j}_n(\hat{\psi}_n)$, as minus the matrix of second-order derivatives with respect to $\hat{\boldsymbol{\xi}}_n$ and the jumps of $\hat{A}_n$, that is,

$$-\frac{\partial}{\partial(\boldsymbol{\xi}, \Delta\mathbf{A})\, \partial(\boldsymbol{\xi}, \Delta\mathbf{A})^\mathsf{T}} L_n(\psi)|_{\psi = \hat{\psi}_n},$$

where $(\boldsymbol{\xi}, \Delta\mathbf{A})^\mathsf{T} = (\boldsymbol{\xi}^\mathsf{T}, \{\Delta A(u_l)\}_l)$.

THEOREM 3. *The solution $h = \hat{\sigma}_n^{-1}(g)$ to the equation $g = \hat{\sigma}_n(h)$ exists with a probability going to* 1, *as n tends to infinity. Furthermore,*

$$\mathbf{g}_\xi^\mathsf{T} \hat{\boldsymbol{\sigma}}_{\xi_0 n}^{-1}(g) + \int_0^\tau g_A \hat{\sigma}_{A_0 n}^{-1}(g)\, d\hat{A}_n \to \mathbf{g}_\xi^\mathsf{T} \boldsymbol{\sigma}_{\xi_0}^{-1}(g) + \int_0^\tau g_A \sigma_{A_0}^{-1}(g)\, dA_0 \tag{42}$$

in probability. If $\hat{\sigma}_n$ is invertible, then $\mathbf{j}_n(\hat{\psi}_n)$ is also invertible and the left-hand side is with $\mathbf{g}_d^\mathsf{T} = (\mathbf{g}_\xi^\mathsf{T}, \{g(u_l)\}_l)$ equal to

$$\mathbf{g}_d^\mathsf{T} \mathbf{j}_n\left(\hat{\psi}_n\right)^{-1} \mathbf{g}_d. \tag{43}$$

This result has been stated in Gill (1989) and Murphy (1995) without proof. It is worth noting that in the proof of Theorem 3 we are not using any specific structure of the correlated frailty model. Therefore, the result should

also hold for general transformation models where similar regularity conditions are fulfilled.

PROOF. Using Proposition 3 in Appendix A, it is straightforward to show that $\hat{\sigma}_n \to \sigma$ in probability in $l^\infty(H_p)$ for all p. Therefore, with a probability going to 1, we can write $\hat{\sigma}_n$ as a sum of a continuously invertible operator and a compact operator. Further, $\hat{\sigma}_n$ must be one-to-one with a probability going to 1: otherwise, using the linearity of σ we can find a bounded sequence $\{h_{n_k}\}_k$ converging to some $h_0 \neq 0$ such that $\sigma_{n_k}(h_{n_k}) = 0$ converges to $\sigma(h_0) = 0$. By the argument in the proof of Theorem 2, this gives us a contradiction. Hence $\hat{\sigma}_n$ is continuously invertible with a probability going to 1.

Let $h_n = \hat{\sigma}_n^{-1}(g)$. Since σ^{-1} is continuous, there exists a constant K such that

$$\begin{aligned}\|\hat{\sigma}_n^{-1}(g) - \sigma^{-1}(g)\| &= \|\sigma^{-1}(\sigma(h_n)) - \sigma^{-1}(\hat{\sigma}_n(h_n))\| \\ &\leq K\|\sigma(h_n) - \hat{\sigma}_n(h_n)\|.\end{aligned}$$

If $\{h_n\}$ is bounded then the right-hand side converges to zero as n tends to infinity. Suppose that $\{h_n\}$ is not bounded. Then we can find a subsequence $\{n_k\}_k$ and real numbers $\{c_{n_k}\}$ satisfying $c_{n_k} \to 0$, $\{c_{n_k} h_{n_k}\}$ tends to $h_0 \neq 0$ and $\hat{\sigma}_{n_k}(c_{n_k} h_{n_k})$ tends to $\sigma(h_0) \neq 0$. On the other hand, using the linearity of $\hat{\sigma}_n$ we have

$$\lim_k \hat{\sigma}_{n_k}(c_{n_k} h_{n_k}) = \lim_k c_{n_k} g = 0,$$

which gives us a contradiction. Therefore $\hat{\sigma}_n^{-1}(g)$ converges to $\sigma^{-1}(g)$ in probability and formula (42) follows.

Assume now that $\hat{\sigma}_n$ is invertible. We write the Fréchet derivative on the form

$$\dot{S}_{\psi_0}(g) = -\iint j(s,u) g(s) \begin{pmatrix} \#(ds) \\ dA_0(s) \end{pmatrix} \begin{pmatrix} \xi(u)\#(du) \\ dA(u) \end{pmatrix},$$

where $j(s,t) = i(s,t)/\alpha_0(s)$. This we approximate with the operator

$$-\iint j_n(\hat{\psi}_n)(s,u) g(s) \begin{pmatrix} \#(ds) \\ d\hat{A}(s) \end{pmatrix} \begin{pmatrix} \xi(u)\#(du) \\ dA(u) \end{pmatrix} = \int \hat{\sigma}_n(g)\, d\psi, \tag{44}$$

where $j_n(\hat{\psi}_n)(\cdot,\cdot)$ is the empirical version of j evaluated at $\hat{\psi}_n$. Using the chain rule, it is easy to verify that the corresponding matrix $\mathbf{j}_n(\hat{\psi}_n)$ is minus the matrix of second-order derivatives with respect to $\hat{\boldsymbol{\xi}}_n$ and the jumps of $\hat{A}_n$. We shall invert this operator on the subspace of Ψ where the integrated hazards are discrete and only jump when $\hat{A}_n$ jumps. We can write (44) as

$$-\mathbf{g}_d^{\mathsf{T}} \operatorname{Diag}(1, \Delta\hat{A}_n) \mathbf{j}_n(\hat{\psi}_n)(\boldsymbol{\xi}, \boldsymbol{\Delta}\mathbf{A})^{\mathsf{T}},$$

where $\mathrm{Diag}(1, \Delta\hat{A})$ is the diagonal matrix consisting of $d + 2$ times 1 and the elements $\{\Delta\hat{A}(u_l)\}_l$. This is a finite-dimensional operator with inverse

$$-\mathbf{g}_d^{\mathsf{T}}\mathbf{j}_n^{-1}\big(\hat{\psi}_n\big)\mathrm{Diag}\big(1, \Delta\hat{A}_n^{-1}\big)(\boldsymbol{\xi}, \boldsymbol{\Delta}\mathbf{A})^{\mathsf{T}}.$$

Evaluating at $\psi = (\mathbf{g}_\xi, \int_0^\cdot g_A\, d\hat{A}_n)$, we get formula (43). $\square$

For smaller data sets, inverting the discrete observed information matrix is in practice feasible. For larger data sets, this may not be possible, since inverting a general $N \times N$ matrix takes $O(N^3)$ operations. The number N is here the number of Euclidean parameters plus the number of observed survival times. For larger data sets, this can be a very big number!

We shall now propose an estimate for the asymptotic variance of the NPMLE which is less time-consuming to calculate. The frailty model is a transformation model, which means that we observe $(Y, \mathbf{X})$, where $Y = A^{-1}(T)$, A: $\mathscr{R} \to \mathscr{R}$ is a unknown transformation, $\mathbf{X}$ is a covariate and the distribution of T is assumed to lie in a parametric model. For the frailty model, A is the integrated hazard function. If A is absolutely continuous with derivative $\alpha \geq 0$, then the density p of Y given $\mathbf{X} = \mathbf{x}$ is

$$p(y; \mathbf{x}, \boldsymbol{\xi}, A) = p_0(A(y); \mathbf{x}, \boldsymbol{\xi})\,\alpha(y). \tag{45}$$

For the gamma-frailty model with only one component in each group, $p_0(u; \mathbf{x}, \boldsymbol{\xi}) = \exp(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x})(1 + \theta\exp(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x})u)^{-1-\theta^{-1}}$. It was shown in Bickel (1985) [see also Bickel, Klaassen, Ricov and Wellner (1993)] that the least favorable direction for A is the unique solution of a second-order Sturm–Liouville problem. Theorem 3 tells us that it is consistent to make discrete information calculations when estimating the asymptotic variance of the NPMLE. Now, solving a discrete version of a second-order Sturm–Liouville equation can be done by inverting a tridiagonal matrix. This only involves $O(N)$ operations! Once the least favorable direction is computed, the information calculation is essentially parametric with the dimension equal to the dimension of $\boldsymbol{\xi}$. We are currently investigating how to extend this algorithm to the frailty model with more than one component in each group.

APPENDIX A

PROPOSITION 3. *Let $\xi, \xi_1, \xi_2, \ldots$ be i.i.d. random elements defined on a probability space $(\Omega, \mathscr{F}, P)$, taking values in some set Ξ. Consider a metric space (L, d) with a separable subset M. Let f: $L \times \Xi \to \mathscr{R}$ be a measurable function satisfying the following at all points $x \in M$: for all $\varepsilon > 0$, there is a $\delta > 0$ such that*

$$|f(x, \xi) - f(\tilde{x}, \xi)| \leq \varepsilon \quad \text{for } \tilde{x}\text{: } d(x, \tilde{x}) \leq \delta \text{ and all } \xi. \tag{46}$$

Then the following holds with a probability equal to one: *for any sequence* $\{a_n\} \subset L$ *such that* $d(a_n, a) \to 0$, *for* $a \in M$, *we have that*

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(a_n, \xi_i) - E\{f(a, \xi)\} \right| \to 0.$$

The result is also valid if f take value in the space of caglad function equipped with the supremum norm.

The proof of Proposition 3 (and the conditions appearing in the proposition) is similar to the ordinary uniform law of large numbers in Hoffmann-Jørgensen (1994), Theorem 9.15, and is therefore omitted.

THEOREM 4 [van der Vaart and Wellner (1996), Theorem 3.3.1]. *Let* S_n *and* S *be random maps and a fixed map*, *respectively*, *from* Ψ *into a Banach space such that*:

(a) $\sqrt{n}\,(S_n - S)\big(\hat{\psi}_n\big) - \sqrt{n}\,(S_n - S)(\psi_0) = o_P^*\big(1 + \sqrt{n}\,\|\hat{\psi}_n - \psi_0\|\big)$.

(b) *The sequence* $\sqrt{n}\,(S_n - S)(\psi_0)$ *converges in distribution to a tight random element* Z.

(c) *The function* $\psi \to S(\psi)$ *is Fréchet differentiable at* ψ_0 *with a continuously invertible derivative* $\dot{S}_{\psi_0}$ (*on its range*).

(d) $S(\psi_0) = 0$ *and* $\hat{\psi}_n$ *satisfies* $S_n(\hat{\psi}_n) = o_P^*(n^{-1/2})$ *and converges in outer probability to* ψ_0.

Then $\sqrt{n}\,(\hat{\psi}_n - \psi_0) \Rightarrow -\dot{S}_{\psi_0}^{-1} Z$.

LEMMA 1 [van der Vaart and Wellner (1996), Lemma 3.3.5]. *Suppose* $S_n(\psi)(h)$ *in Theorem* 4 *is of the form* $P_n \phi(\psi, h)$, *where* P_n *is the empirical measure. Assume that the class of functions*

$$\{\phi(\psi, h) - \phi(\psi_0, h) : \|\psi - \psi_0\| < \delta, h \in H\} \tag{47}$$

is P_0*-Donsker for some* $\delta > 0$ *and that*

$$\sup_{h \in H} E_0\{\phi(\psi, h) - \phi(\psi_0, h)\}^2 \to 0, \qquad \psi \to \psi_0. \tag{48}$$

If $\hat{\psi}_n$ *converges in outer probability to* ψ_0, *then condition* (a) *is satisfied.*

LEMMA 2. *Let* W *be a caglad process on* $[0, \tau]$ *which is uniformly bounded in variation. Let* W_i *be i.i.d. replicates of* W. *Then the functional central limit theorem is valid for* $n^{-1/2} \sum_{i=1}^{n} \{W_i(\cdot) - EW(\cdot)\}$ *in* $l^\infty([0, \tau])$.

PROOF. Let P denote the distribution of W and let $\mathscr{F}$ denote the class of projections $x \to x(u)$, for $u \in [0, \tau]$. Then Lemma 2 states that $\mathscr{F}$ is P-Donsker. We shall prove the lemma in two steps. First assume that W is nondecreasing, then the result is given in van der Vaart and Wellner (1996), Example 2.10.27; see also Parner (1996b), for another proof.

Consider the general case where W takes value in the space of a caglad function which is uniformly bounded in variation norm. Given a function in this space, x, we can write it as a difference of two monotone decreasing functions, $x(u) = x^1(u) - x^2(u)$, which are both uniformly bounded. We can choose $x^1(u) = x(0) + p(u)$ and $x^2(u) = n(u)$, where p and n is the Jordan decomposition of x [see, e.g., Hildebrandt (1963), page 38]. Now put $f_u^i(x) = x^i(u)$ for $i = 1, 2$. If we equip the space of caglad functions with the projection σ-algebra, then f_u^i are measurable functions. From above it follows that $\mathscr{F}_i = \{f_u^i | u \in [0, \tau]\}$ $i = 1, 2$ are both P-Donsker classes. Then also $\mathscr{F}_1 - \mathscr{F}_2$ is a P-Donsker class [Example 2.10.7 in van der Vaart and Wellner (1996)], and since $\mathscr{F} = \{f_u^1 - f_u^2 | u \in [0, \tau]\} \subseteq \mathscr{F}_1 - \mathscr{F}_2$ it follows that $\mathscr{F}$ is also a P-Donsker class. □

APPENDIX B

Define

$$\mathbf{L}_{\beta n}(\psi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \int_0^\tau \mathbf{X}_{ij}\, dN_{ij} - \hat{Z}_i^{(j)}(\tau) \int_0^\tau Y_{ij}(u; \boldsymbol{\beta}) \mathbf{X}_{ij}(u)\, dA,$$

$$L_{An}(\psi)(h_A) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \int_0^\tau h_A\, dN_{ij} - \hat{Z}_i^{(j)}(\tau) \int_0^\tau Y_{ij}(\boldsymbol{\beta}) h_A\, dA.$$

Further, let

$$L_{\theta n}(\psi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{\mathbf{k} \in K_i(\tau)} a_i(\mathbf{k}, \tau) a_i^\theta(\mathbf{k}, \tau)}{\sum_{\mathbf{k} \in K_i(\tau)} a_i(\mathbf{k}, \tau)},$$

where

$$\begin{aligned} a_i^\theta(\mathbf{k}, \tau) &= \frac{\partial}{\partial \theta} \log\big(a_i(\mathbf{k}, \tau)\big) \\ &= \sum_{j=1}^{m} \left\{ \sum_{h=1}^{k_j} \frac{-2}{\theta_\cdot} + 2\theta^* \theta_\cdot^{-3} \log\big(1 + \theta_\cdot \Lambda_{ij}(\tau)\big) \right. \\ &\qquad \left. - \big(\theta^* \theta_\cdot^{-2} + k_j\big) \frac{\Lambda_{ij}(\tau)}{1 + \theta_\cdot \Lambda_{ij}(\tau)} \right\} \\ &\quad + \sum_{h=1}^{N_i.(\tau) - k_\cdot} \frac{\theta_\cdot - 2\theta}{\theta\theta_\cdot + \theta_\cdot^3 (h - 1)} \\ &\quad + N_i.(\tau) \theta_\cdot^{-1} - \big(\theta_\cdot^{-2} - 2\theta\theta_\cdot^{-3}\big) \log\big(1 + \theta_\cdot \Lambda_i.(\tau)\big) \\ &\quad - \big(\theta\theta_\cdot^{-2} + N_i.(\tau) - k_\cdot\big) \frac{\Lambda_i.(\tau)}{1 + \theta_\cdot \Lambda_i.(\tau)}. \end{aligned}$$

Similarly, let

$$L_{\theta^* n}(\psi) = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{\mathbf{k}\in K_i(\tau)} a_i(\mathbf{k},\tau)\, a_i^{\theta^*}(\mathbf{k},\tau)}{\sum_{\mathbf{k}\in K_i(\tau)} a_i(\mathbf{k},\tau)},$$

where

$$\begin{aligned} a_i^{\theta^*}(\mathbf{k},\tau) &= \frac{\partial}{\partial\theta^*}\log\bigl(a_i(\mathbf{k},\tau)\bigr) \\ &= \sum_{j=1}^{m}\Biggl\{\sum_{h=1}^{k_j}\left(\frac{1}{\theta^*}-\frac{2}{\theta_\cdot}\right) - \bigl(\theta_\cdot^{-2}-2\theta^*\theta_\cdot^{-3}\bigr)\log\bigl(1+\theta_\cdot\Lambda_{ij}(\tau)\bigr) \\ &\qquad\qquad -\bigl(\theta^*\theta_\cdot^{-2}+k_j\bigr)\frac{\Lambda_{ij}(\tau)}{1+\theta_\cdot\Lambda_{ij}(\tau)}\Biggr\} \\ &\quad + \sum_{h=1}^{N_i.(\tau)-k_\cdot}\frac{-2\theta}{\theta\theta_\cdot+\theta_\cdot^3(h-1)} + N_i.(\tau)\,\theta_\cdot^{-1} \\ &\quad + 2\theta\theta_\cdot^{-3}\log\bigl(1+\theta_\cdot\Lambda_i.(\tau)\bigr) \\ &\quad -\bigl(\theta\theta_\cdot^{-2}+N_i.(\tau)-k_\cdot\bigr)\frac{\Lambda_i.(\tau)}{1+\theta_\cdot\Lambda_i.(\tau)}. \end{aligned}$$

The score operator is, with $\mathbf{L}_{\xi n}^{\mathsf{T}}(\psi) = (L_{\theta n}, L_{\theta^* n}, \mathbf{L}_{\beta n}^{\mathsf{T}})(\psi)$, given by

$$S_n(\psi)(h) = \mathbf{h}_\xi^{\mathsf{T}}\mathbf{L}_{\xi n}(\psi) + L_{An}(\psi)(h_A).$$

Acknowledgments. This work was done while visiting Richard Gill at the University of Utrecht. The author appreciates very much his advice and encouragement. The author also thanks Susan Murphy for many helpful discussions and the referees for many useful comments and suggestions which improved the presentation of this work.

REFERENCES

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, Berlin.

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10** 1100–1120.

BICKEL, P. (1985). Efficient testing in a class of transformation models. *Bull. Int. Statist. Inst.* **51** 63–81, Meeting 23. Amsterdam.

BICKEL, P., KLAASSEN, C., RITOV, Y. and WELLNER, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press.

BREMAUD, P. (1981). *Point Processes and Queues*. Springer, New York.

CLAYTON, D. G. and CUZICK, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *J. Roy. Statist. Soc. Ser. A* **148** 82–117.

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. *Rev. Econom. Stud.* **XLIX** 403–409.

GILL, R. D. (1985). Discussion of the paper by D. Clayton and J. Cuzick. *J. Roy. Statist. Soc. Ser. A* **148** 108–109.

GILL, R. D. (1989). Non- and semi-parametrix maximum likelihood estimation and the von Mises method, 1. *Scand. J. Statist.* **16** 97–128.

GILL, R. D. (1992). Marginal partial likelihood. *Scand. J. Statist.* **19** 133–137.

GROENEBOOM, P. (1991). Nonparametric maximum likelihood estimators for interval censoring and deconvolution. Technical Report 378, Dept. Statistics, Stanford Univ.

HILDEBRANDT, T. H. (1963). *Introduction to the Theory of Integration*. Academic Press, New York.

HOFFMANN-JØRGENSEN, J. (1994). *Probability with a View to Statistics*. Chapman and Hall, London.

HOUGAARD, P. (1987). Modelling multivariate survival. *Scand. J. Statist.* **14** 291–304.

JOHANSEN, S. (1983). An extension of Cox's regression model. *Internat. Statist. Rev.* **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.

KORTRAM, R. A., VAN ROOIJ, A. C. M., LENSTRA, A. J. and RIDDER, G. (1995). Constructive identification of the mixed proportional hazards model. *Statist. Neerlandica* **49** 269–281.

MORSING, T. (1994). Competing risks in cross-over designs. Dept. Mathematics, Chalmers Univ. Technology. Preprint 20.

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.

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.

PARNER, E. (1996a). Consistency in the correlated Gamma-frailty model. Research Report 345, Dept. Theoretical Statistics, Inst. Mathematics, Univ. Aarhus.

PARNER, E. (1996b). Asymptotic normality in the correlated Gamma-frailty model. Research Report 346, Dept. Theoretical Statistics, Inst. Mathematics, Univ. Aarhus.

PEDERSEN, J. (1995). A litter frailty model. Research Report 95/13, Dept. Biostatistics, Univ. Copenhagen.

RAO, R. R. (1963). The law of large numbers for $D[0, 1]$-valued random variables. *Theory Probab. Appl.* **8** 70–74.

RUDIN, W. (1973). *Functional Analysis*. McGraw-Hill, New York.

VAN DER VAART, A. W. (1995). Efficiency of infinitely dimensional estimators. *Statist. Neerlandica* **49** 9–30.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

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.

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

WIJERS, B. J. (1995). Consistent non-parametric estimation for a one-dimensional line segment process observed in an interval. *Scand. J. Statist.* **22** 335–360.

WOODBURY, M. A. (1971). Discussion of paper by Hartley and Hocking. *Biometrics* **27** 808–817.

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

DEPARTMENT OF THEORETICAL STATISTICS
UNIVERSITY OF AARHUS
NY MUNKEGADE
DK-8000 AARHUS C
DENMARK
E-MAIL: erik@mi.aau.dk
parner@biostat.aau.dk