The Annals of Statistics
1997, Vol. 25, No. 4, 1471–1509

SEMIPARAMETRIC LIKELIHOOD RATIO INFERENCE

BY S. A. MURPHY[1] AND A. W. VAN DER VAART

Pennsylvania State University and Free University Amsterdam

Likelihood ratio tests and related confidence intervals for a real parameter in the presence of an infinite dimensional nuisance parameter are considered. In all cases, the estimator of the real parameter has an asymptotic normal distribution. However, the estimator of the nuisance parameter may not be asymptotically Gaussian or may converge to the true parameter value at a slower rate than the square root of the sample size. Nevertheless the likelihood ratio statistic is shown to possess an asymptotic chi-squared distribution. The examples considered are tests concerning survival probabilities based on doubly censored data, a test for presence of heterogeneity in the gamma frailty model, a test for significance of the regression coefficient in Cox's regression model for current status data and a test for a ratio of hazards rates in an exponential mixture model. In both of the last examples the rate of convergence of the estimator of the nuisance parameter is less than the square root of the sample size.

1. Introduction. In the past decade considerable progress has been made with the study of maximum likelihood estimators in infinite dimensional statistical models, sometimes called nonparametric maximum likelihood estimators (NPMLE) or semiparametric maximum likelihood estimators. See, for instance, Gill (1989) or van der Vaart (1994a) for reviews of work in this direction and Gu and Zhang (1993), Huang (1996), Murphy (1995a), Van der Laan (1993), van der Vaart (1994c, 1996), Gill, van der Laan and Wijers (1995), Huang and Wellner (1995) and Wijers (1995) for more recent results. Most of this work is directed at proving the asymptotic normality and efficiency of the maximum likelihood estimator of smooth parameters of the model. In contrast, very little progress has been made toward a general likelihood ratio theory for semiparametric models. Here the term 'semiparametric model' is used in a loose sense as a model which is not finite dimensional (as in classical statistics), nor fully nonparametric [cf. Bickel, Klaassen, Ritov and Wellner (1993), pages 1–2]. In this paper we give a general approach for the asymptotic analysis of hypothesis tests and associated confidence regions based on the (semiparametric) likelihood ratio test.

The results of this paper can be viewed as a step in filling the large gap between classical parametric likelihood ratio theory and empirical likelihood as considered by Thomas and Grunkemeier (1975), Owen (1988), Qin and Lawless (1994) and Murphy (1995b) among others. These authors are concerned with the situation where the model for the data is fully nonparametric, in the sense that it contains every possible probability distribution on the sample

Received September 1995; revised September 1996.

[1] Research partially supported by NSF Grant DMS-93-07255.

AMS 1991 *subject classifications*. 62G15, 62G20, 62F25.

Key words and phrases. Least favorable submodel, profile likelihood, confidence interval.

space, or is restrained by finitely many constraints [as in Qin and Lawless (1994) if $r > p$]. Each time the "likelihood" is taken as the product $\prod_i P\{X_i\}$ of the masses given to the observational points, referred to as the empirical likelihood. Qin (1993) and Qin and Wong (1996) extend the above theory to specific semiparametric models. However, in all of the above cases, the likelihood ratio statistic reduces to a function of a vector statistic (often a Lagrange multiplier), simplifying the asymptotic analysis greatly. General semiparametric models, of the type we consider in this paper, appear to require a different approach. For example, this simplification will not occur in the mixture model considered by Roeder, Carroll and Lindsay (1996) in which they invert a likelihood ratio test to form a confidence interval for a regression coefficient.

In the classical parametric case likelihood ratio confidence regions are generally preferred over Wald-type confidence regions, except perhaps from a computational perspective. Advantages mentioned by many authors are small sample coverage probabilities closer to the nominal values, the possibility of asymmetric confidence regions and regions that are transformation respecting [Hall and La Scala (1990)]. Although we do not give a proof in this paper, many of these advantages may be expected to carry over to the semiparametric situation. This appears to be particularly the case when the small sample distribution of the estimator is highly skewed. A concrete example in the nonparametric setting is in the construction of confidence intervals for a survival probability based on the Kaplan–Meier estimator. The Wald confidence interval for a survival probability performs poorly and much work has been done on finding a transformation of the estimator which has an approximate normal distribution for small samples [Andersen et al. (1993)]. In this case inversion of the likelihood ratio test as in Thomas and Grunkemeier (1975), Li (1995) and Murphy (1995b) illustrate how the resulting confidence intervals perform as well as confidence intervals based on widely accepted "best" transformations of the estimator. In their comparison of empirical likelihood confidence intervals with bootstrap confidence intervals, Hall and La Scala argue, that in situations in which both methods can be applied, the empirical likelihood confidence intervals are to be preferred to bootstrap confidence intervals. They state that "the power of the bootstrap resides in the fact that it can be applied to very complex problems and this feature is not available for empirical likelihoods." Usually inference for semiparametric models is a complex problem; however, as we shall demonstrate, likelihood ratio inference will, in general, be available.

The traditional advantage of a Wald confidence interval, ease of computation, does not appear to be valid any more due to the computational difficulty of estimating asymptotic variances. In general the asymptotic variance of infinite dimensional maximum likelihood estimators is not given by a closed formula, or even an expectation of a known function, but can only be characterized as the variance of the efficient influence function. The latter is the solution of an infinite dimensional minimization problem and its computation may require the inversion of an infinite dimensional operator. Even in a discretized form, for instance at observed data points, the inversion may still involve inverting a matrix of high dimension. This is true, for instance, in the semiparametric frailty model considered by Nielsen, Gill, Andersen and Sorensen (1992) and Murphy (1995a), where estimators for the standard error of the estimated frailty variance are found by inverting a matrix which is of the same dimension as the data.

Furthermore, in cases where the efficient influence function can be written down relatively explicitly, the estimation of its variance may involve nonparametric smoothing. This means that the researcher must deal with the difficult choice of a smoothing parameter. This is the case in estimating the regression coefficient in Cox's regression model subject to current status type censoring, considered by Huang (1996), where the efficient influence function depends on a ratio of conditional means.

For the above reasons setting Wald-type tests and associated confidence regions in semiparametric models may be computationally harder than in the classical situation, where one can use a plug-in estimator based on an expression for the Fisher information or the observed information. Therefore, due to the expected gain in quality, likelihood ratio based inference appears doubly attractive in semiparametric settings.

The definition of a likelihood ratio statistic requires the definition of a likelihood function. In classical parametric models this is the density of the observations, while empirical likelihood theory uses the product $\prod P\{X_i\}$. We do not offer a general definition of an infinite dimensional likelihood function in this paper. In some examples the observations have a well-defined density and a likelihood is defined much as in the classical situation. In other examples one uses the empirical likelihood (which, however, is maximized only over the model). Mixtures of these situations occur as well in the literature and in some missing data situations a "partial likelihood" appears appropriate. Some of these possibilities are illustrated in our four examples. With this formulation a semiparametric likelihood estimator is not necessarily discrete (although it can often be taken to be discrete) and it is often not supported on the observed data. Restricting the estimator to a null hypothesis may introduce new support points.

We consider the situation that the observations $X_1, \ldots, X_n$ are a random sample from a distribution P_ψ indexed by a parameter ψ that is known to belong to a set Ψ. Given a parameter (map) $\theta\colon \Psi \to \mathbb{R}$ and a definition of a likelihood $\mathrm{lik}(\psi, X)$ for one observation, the likelihood ratio statistic for testing the null hypothesis $\theta(\psi) = \theta_0$ is given by

$$\begin{aligned}\mathrm{lrt}_n(\theta_0) &= 2\left(\sup_{\psi\in\Psi} \sum_{i=1}^{n} \ln \mathrm{lik}(\psi, X_i) - \sup_{\psi\in\Psi,\, \theta(\psi)=\theta_0} \sum_{i=1}^{n} \ln \mathrm{lik}(\psi, X_i)\right)\\ &= 2n\mathbb{P}_n \ln \mathrm{lik}(\hat{\psi}) - 2n\mathbb{P}_n \ln \mathrm{lik}(\hat{\psi}_0).\end{aligned}$$

Here $\mathbb{P}_n$ is the empirical distribution of the data and $\hat{\psi}$ and $\hat{\psi}_0$ are the maximum likelihood estimators under the full model and the null hypothesis, respectively. In the first example ψ is a distribution function F and $\theta(\psi) = F(t_0)$ is its value at a fixed point. In the remaining three examples the parameter

ψ takes the form $\psi = (\theta, \Lambda)$ or $\psi = (\theta, F)$ for an unknown cumulative hazard function Λ or distribution function F and $\theta(\psi) = \theta$.

For simplicity we restrict ourselves to one-dimensional parameters θ. Then we wish to prove that under the null hypothesis the sequence $\text{lrt}_n(\theta_0)$ converges in distribution to a χ^2-distribution on one degree of freedom. Having proved this for every value of θ_0, the region $\{\theta\colon \text{lrt}_n(\theta) \le z^2_{\alpha/2}\}$ is the associated confidence region of asymptotic level $1 - \alpha$.

The organization of the paper is as follows. In Section 2 we present four rather different examples for which we discuss the meaning of the likelihood and state theorems on the likelihood ratio test. The examples include the double censoring model considered in Chang (1990), regression for current status data considered by Huang (1996), the gamma frailty model of Murphy (1994, 1995a) and a mixture model studied by van der Vaart (1996). Section 3 starts with a discussion of the finite dimensional situation to gain intuition and next gives a general approach to prove the asymptotic validity of the semiparametric likelihood ratio test. The basic scheme given by Theorem 3.1 leaves some nontrivial work for special examples. However, our impression is that it works in the situations where also the asymptotic normality of the maximum likelihood estimator of the parameter of interest can be proved. The last sections contain detailed treatments of our four examples.

2. Examples and results.

This section contains four examples. For each example we give the definition of the likelihood and state a theorem on the likelihood ratio statistics. Proofs are given in Sections 4–7.

EXAMPLE (Doubly censored data). Doubly censored data arise when event times are subject to both right and left censoring. The event time T is observed only if it falls between the left and right censoring times, L and R. Otherwise all that is observed is L and that $T \le L$ in the case of a left censoring or R and that $T > R$ in the case of a right censoring. It is assumed that T is independent of (L, R) and that $L \le R$. Thus the observations are n i.i.d. copies of $X = (U, D)$, where $U = L$ and $D = 1$ if $T \le L$, $U = T$ and $D = 2$ if $L < T \le R$ and $U = R$ and $D = 3$ if $T > R$. If G_L and G_R are the marginal distributions of $L \le R$ and F the distribution of T, then, with lowercase symbols denoting densities, the density of X is given by

$$\begin{aligned} p_F(X) = {} & [F(U)g_L(U)]^{I\{D=1\}}[f(U)(G_L - G_R)(U-)]^{I\{D=2\}} \\ & \times [(1 - F(U))g_R(U)]^{I\{D=3\}}. \end{aligned}$$

When F is completely unknown, the above density is not suitable for use as a likelihood. Instead we use the empirical likelihood $P_F\{X\}$, which is obtained by replacing the densities g_L, f and g_R by the point probabilities $G_L\{U\}$, $F\{U\}$ and $G_R\{U\}$. For inference about F we can drop the terms involving G_L and G_R and define the likelihood to be

$$\text{lik}(F, X) = [F(U)]^{I\{D=1\}}\, \Delta F(U)^{I\{D=2\}}[1 - F(U)]^{I\{D=3\}}.$$

We maximize the above "likelihood" over discrete distribution functions with steps at the U's. The parameter of interest will be $\theta(F) = Fg = \int g\,dF$ for some known function g of bounded variation. Of particular interest is $g(t) = 1\{t > t_0\}$, which leads to a confidence set for $1 - F(t_0)$, the probability of survival longer than t_0. The concavity in F of the density in X along with the continuity of θ in F implies that the confidence set will be a confidence interval (see the Appendix).

We shall prove the following theorem. (Note that $\rightsquigarrow$ denotes convergence in distribution throughout this paper.)

THEOREM 2.1. *Suppose that* $(G_L - G_R)(u-) = P(L < u \le R)$ *is positive on the convex hull* $[\sigma, \tau] \subset [0, \infty)$ *of the support of* F_0. *Furthermore, assume that* F_0, G_L *and* G_R *are continuous, with* $G_L(\tau) = 1$ *and* $G_R(\sigma-) = 0$. *Let* g *be a left continuous function of bounded variation which, on* $[\sigma, \tau]$, *is not* F_0*-almost everywhere equal to a constant. If* $F_0 g = \theta_0$, *then the likelihood ratio statistic for testing that* $Fg = \theta_0$ *satisfies* $\mathrm{lrt}_n(\theta_0) \rightsquigarrow \chi^2_1$ *under* F_0.

The asymptotic consistency and normality of the unrestricted maximum likelihood estimator in this model was proved under stronger conditions by Chang and Yang (1987), Chang (1990) and under weaker conditions by Gu and Zhang (1993).

EXAMPLE (Cox regression for current status data). In current status data, n subjects are examined each at a random observation time and at this time it is observed whether the event time has occurred or not. The event time T is assumed to be independent of the observation time Y given the covariate Z. Then the observations are n i.i.d. copies of $X = (Y, \delta, Z)$, where $\delta = 1$ if $T \le Y$ and zero otherwise. Suppose that the hazard function of T given $Z = z$ is given by Cox's regression model: the hazard at time t is $e^{\theta z}\lambda(t)$. Then the cumulative hazard at time t of T given $Z = z$ is of the form $e^{\theta z}\int_0^t \lambda(s)\,ds = e^{\theta z}\Lambda(t)$ and the density is given by

$$p_{\theta,\Lambda}(X) = \bigl(1 - \exp(-\exp(\theta Z)\Lambda(Y))\bigr)^{\delta}\bigl(\exp(-\exp(\theta Z)\Lambda(Y))\bigr)^{1-\delta} f^{Y,Z}(Y, Z).$$

The parameter of interest is the regression parameter θ; the nuisance parameter Λ is assumed completely unknown. A test of regression would be a test of H_0: $\theta = 0$.

The likelihood $\mathrm{lik}(\theta, \Lambda, X)$ is taken equal to the density, but with the term $f^{Y,Z}(Y, Z)$ omitted. To estimate θ and Λ we maximize the likelihood over θ in a bounded parameter set $\Theta \subset \mathbb{R}$ and over Λ ranging over all nondecreasing cadlag functions taking values in $[0, M]$, for a known M.

We shall prove the following theorem.

THEOREM 2.2. *Let* θ_0 *be an interior point of* Θ. *Let* Y *have a Lebesgue density which is continuous and positive on its support* $[\sigma, \tau]$ *for which* $\Lambda_0(\sigma-) > 0$

and $\Lambda_0(\tau) < M$, *and zero otherwise. Let* Λ_0 *be differentiable on this interval with derivative bounded away from zero. Let* Z *be bounded and* $E\,\mathrm{var}(Z|Y) > 0$. *Finally assume that the function* h^{**} *given by* (5.1) *has a version which is differentiable with a bounded derivative on* $[\sigma, \tau]$. *Then the likelihood ratio statistic for testing* H_0: $\theta = \theta_0$ *satisfies* $\mathrm{lrt}_n(\theta_0) \rightsquigarrow \chi_1^2$ *under* (θ_0, Λ_0).

The asymptotic normality of the maximum likelihood estimator for θ in this model is considered by Huang (1996). The maximum likelihood estimator for the cumulative hazard function converges at an $O(n^{-1/3})$ rate in an L_2-norm. Under the hypothesis H_0: $\theta = 0$ this model reduces to the "case 1 interval censoring" considered by Groeneboom (1987), who obtains the limit distribution of $\hat{F}_0(t)$ for the distribution function corresponding to $\hat{\Lambda}_0$ (with an $n^{-1/3}$-standardization). See Groeneboom and Wellner (1992).

EXAMPLE (Gamma frailty). In the frailty model, subjects occur in groups such as twins or litters. To allow for a positive intragroup correlation in the subjects's event times, subjects in the same group are assumed to share the same frailty Z. In the one-sample problem, we observe n i.i.d. groups where for a given group the observations are J and $(T_j \wedge C_j, D_j)$ for $j = 1, \dots, J$, where T_j is the event time associated with the jth subject in the group, C_j is censoring time, $D_j = 1$ if $T_j \le C_j$ and J is the random group size. The unobserved frailty Z is assumed independent of J and to follow a gamma distribution with mean 1 and variance θ. Given J, $(C_j,\ j = 1, \dots, J)$ is assumed independent of both Z and $(T_j,\ j = 1, \dots, J)$. Given Z and J, the $(T_j, j = 1, \dots, J)$ are independent, with hazards $Z\lambda(\cdot),\ j = 1, \dots, J$.

Put $N(t) = \sum_j I\{T_j \wedge C_j \le t,\ D_j = 1\}$ and $Y(t) = \sum_j I\{T_j \wedge C_j \ge t\}$. So the observation for a group is $X = (N, Y)$. For our statistical inference we shall only use the values of this counting process on a given finite interval $[0, \tau]$. Since the censoring is independent and noninformative of the Z, we have that given $Z = z$, the intensity of N at time t is $zY(t)\lambda(t)$ and the conditional density is proportional to

$$p_\Lambda(X|Z = z) = \prod_{t \le \tau} (zY(t)\lambda(t))^{\Delta N(t)} \exp\left\{-z \int_0^\tau Y\, d\Lambda\right\},$$

where $\Lambda(\cdot) = \int_0^\cdot \lambda(s)\, ds$. [See Andersen et al. (1993), pages 138–150, and Nielsen, Gill, Andersen and Sorensen (1992).] To form the marginal density for a group multiply by the gamma density of Z and integrate over z to get

$$p_{\theta,\Lambda}(X) = \frac{\prod_{t \le \tau}((1 + \theta N(t-))Y(t)\lambda(t))^{\Delta N(t)}}{(1 + \theta \int_0^\tau Y(t) d\Lambda(t))^{1/\theta + N(\tau)}}.$$

When Λ is unknown, the associated likelihood has no maximizer. A convenient extension is

$$\mathrm{lik}(\theta, \Lambda, X) = \frac{\prod_{t \le \tau}\big((1 + \theta N(t-))Y(t)\Delta\Lambda(t)\big)^{\Delta N(t)}}{\big(1 + \theta \int_0^\tau Y(t) d\Lambda(t)\big)^{1/\theta + N(\tau)}}. \tag{2.1}$$

This is not the only possible extension [see Andersen et al. (1993) and Murphy (1995a)]. We are particularly interested in an hypothesis test of zero intragroup correlation, that is, H_0: $\theta = 0$.

The likelihood is also well-defined for negative θ close to zero, even though θ can then not be introduced through a gamma variable as previously. We define the likelihood ratio statistic and the maximum likelihood estimators relative to the parameter set consisting of θ ranging over the interval $[-\varepsilon, M]$ for a small $\varepsilon > 0$, and Λ ranging over all finite nondecreasing functions on $[0, \tau]$.

THEOREM 2.3. *Assume that* $\theta_0 \in [0, M)$ *and that* Λ_0 *is continuous, strictly increasing and finite on* $[0, \tau]$. *Furthermore, assume that* J *has finite support,* $P_0[\bigcup_{j=1}^{J}\{C_j \geq \tau\}] > 0$, *and that the distribution of* $(C_j,\ j = 1, \dots, J)$ *has at most a finite number of discontinuities. Then the likelihood ratio statistic for testing* H_0: $\theta = \theta_0$ *satisfies* $\mathrm{lrt}_n(\theta_0) \rightsquigarrow \chi_1^2$ *under* (θ_0, Λ_0).

The maximum likelihood estimator $(\hat{\theta}, \hat{\Lambda})$ for this model was shown to be asymptotically consistent and normal by Murphy (1994, 1995a) under slightly more general conditions.

EXAMPLE (Mixture model). This is another version of the frailty model. The group size is 2. As before, we allow for intragroup correlation in the event times by assuming that the pair share the same unobserved frailty Z. Given Z, the two event times T_1 and T_2 are assumed to be independent and exponentially distributed with hazard rates Z and θZ, respectively. In contrast to the gamma frailty model, the distribution F of Z is a completely unknown distribution on $(0, \infty)$. The observations are n i.i.d. copies of $X = (T_1, T_2)$ from the density

$$p_{\theta, F}(X) = \int z \exp(-zT_1)\, \theta z \exp(-\theta z T_2)\, dF(z).$$

We use this as the likelihood $\mathrm{lik}(\theta, F, X)$ and are interested in a confidence set for the ratio θ of the hazards and testing that this ratio equals 1.

THEOREM 2.4. *Suppose that* $\int (z^2 + z^{-6.5})\, dF_0(z) < \infty$. *Then the likelihood ratio statistic for testing* H_0: $\theta = \theta_0$ *satisfies* $\mathrm{lrt}_n(\theta_0) \rightsquigarrow \chi_1^2$ *under* (θ_0, F_0).

The maximum likelihood estimator for θ was shown to be asymptotically normal by van der Vaart (1996). The maximum likelihood estimator for the distribution function F is known to be consistent from Kiefer and Wolfowitz (1956). Van der Vaart (1991) proved that the information for estimating $F(t)$ is zero, so that the rate of convergence of the best estimators is less than the square root of n. A reasonable conjecture is that the optimal rate is $n^{-\alpha}$ for some $\alpha > 0$.

3. Intuition.

In the case that the parameter ψ is Euclidean a classical approach to derive the asymptotic χ^2-distribution of the likelihood ratio statistic is to expand the difference

$$2n\mathbb{P}_n\bigl[\ln p_{\hat\psi} - \ln p_{\hat\psi_0}\bigr]$$

in a two-term Taylor expansion around $\hat\psi$. The linear term vanishes and algebraic manipulations involving the joint normal limit distribution of $\hat\psi - \psi_0$ and $\hat\psi_0 - \psi_0$ yield the result. A more insightful derivation can be based on the approximation

$$\hat\psi - \hat\psi_0 = \biggl(\frac{i_{\psi_0}^{-1}\dot\theta_0}{\dot\theta_0^T i_{\psi_0}^{-1}\dot\theta_0} + \varepsilon\biggr)(\hat\theta - \theta_0), \tag{3.1}$$

where $\dot\theta_0^T i_{\psi_0}^{-1}\dot\theta_0$ is the asymptotic variance of $\sqrt{n}(\hat\theta - \theta_0)$; i_ψ is the information matrix, $\dot\theta_0$ is the derivative of $\theta(\psi)$ with respect to ψ; and ε converges to zero in probability. If $\hat\psi$ is multivariate normal and θ is linear, then $\varepsilon = 0$. More concretely, if $\psi = (\theta, \eta)$ and $\theta(\psi) = \theta$, then

$$(\hat\theta, \hat\eta)^T - (\theta_0, \hat\eta_0)^T = \bigl(1, -i_{\eta_0\eta_0}^{-1} i_{\eta_0\theta_0} + \varepsilon\bigr)^T(\hat\theta - \theta_0), \tag{3.2}$$

where the information matrix i_ψ is partitioned into

$$i_{\psi_0} = \begin{pmatrix} i_{\theta_0\theta_0} & i_{\theta_0\eta_0} \\ i_{\eta_0\theta_0} & i_{\eta_0\eta_0} \end{pmatrix}.$$

Under regularity conditions both (3.1) and (3.2) can be justified by Taylor series arguments or by analogy to the case of a multivariate normal observation [cf. Cox and Hinkley (1974), pages 308, 323]. If, as in Cox and Hinkley (1974), we neglect the error term ε, then we can replace $\hat\psi_0$ in the likelihood ratio statistic by a constant times $\hat\theta - \theta_0$, and next perform a two-term Taylor expansion in the one-dimensional parameter $\hat\theta - \theta_0$. This yields the approximation

$$2n\mathbb{P}_n\Bigl[\ln p_{\hat\theta\hat\eta} - \ln p_{\theta_0,\, \hat\eta + (i_{\eta_0\eta_0}^{-1} i_{\eta_0\theta_0})^T(\hat\theta - \theta_0)}\Bigr] \approx -n(\hat\theta - \theta_0)^2\, \mathbb{P}_n \ddot\ell(\cdot; \hat\theta, \hat\eta),$$

where $\ddot\ell(\cdot; t, \hat\eta)$ is the second derivative of the map $t \to \ln p_{t,\, \hat\eta + (i_{\eta_0\eta_0}^{-1} i_{\eta_0\theta_0})^T(\hat\theta - t)}$. The first derivative of this function at $\hat\theta$ can be expressed in the score functions, $(\dot\ell_\theta, \dot\ell_\eta)$, for (θ, η) as

$$\dot\ell(\cdot; \hat\theta, \hat\psi) = \dot\ell_{\hat\theta} - (i_{\eta_0\eta_0}^{-1} i_{\eta_0\theta_0})^T \dot\ell_{\hat\eta}.$$

By the usual identities the expectation of the second derivative should be minus the expectation of the square of the first derivative and, under regularity conditions,

$$-\mathbb{P}_n\ddot\ell(\cdot; \hat\theta, \hat\psi) \to_{\mathrm{P}} P_{\theta_0\eta_0}\bigl(\dot\ell_{\theta_0} - (i_{\eta_0\eta_0}^{-1} i_{\eta_0\theta_0})^T \dot\ell_{\eta_0}\bigr)^2 = i_{\theta_0\theta_0} - i_{\theta_0\eta_0} i_{\eta_0\eta_0}^{-1} i_{\eta_0\theta_0}.$$

This is exactly the $(1,1)$-element of the inverse of i_{ψ_0}, which is the inverse of the asymptotic variance of $\sqrt{n}(\hat\theta - \theta_0)$. The chi-squared limit distribution follows.

We might expect that, at least to the first order, the difference between the full and null maximum likelihood estimator in our semiparametric setting satisfies a generalization of (3.2). Then the difference is finite dimensional (the dimension of θ), and a standard Taylor expansion in θ can be used to prove the asymptotic chi-squared distribution. For example, Murphy [(1995b), equation (6) and the appendix] proves a result of this type for a likelihood ratio test based on the binomial likelihood for right-censored data. There, θ is the probability of survival past time t_0 and ψ is the cumulative hazard function Λ, so that $\theta = \theta(\Lambda) = \prod_{s \le t_0}(1 - d\Lambda(s))$. For $t \le t_0$,

$$\hat\Lambda(t) - \hat\Lambda_0(t) = \frac{\int_0^t \theta_0/(\overline{Y}(s) + \lambda\theta_0)\, d\hat\Lambda(s)}{\operatorname{var}(\sqrt{n}(\hat\theta - \theta_0)) + \varepsilon}(\hat\theta - \theta_0),$$

where $\overline{Y}(s)$ is the number of individuals who have not failed or been censored up to time s, divided by the sample size and both λ and ε converge in probability to zero.

When ψ can be estimated at a square root n rate, then an analog to i_ψ^{-1} exists and (3.2) can be extended appropriately. This approach can be used in both the gamma frailty and the double censoring examples. On the other hand, in many semiparametric models, including our current status and mixture examples, the nuisance parameter is not estimable at square root n rate. Then an extension of (3.2) requires more care. Note for instance that (3.2) implies that the difference $\hat\eta - \hat\eta_0$ is of order $O(n^{-1/2})$, often much smaller than the differences $\hat\eta - \eta_0$ and $\hat\eta_0 - \eta_0$ (depending on the distance). In any case the two-step proof focusing first on the difference $\hat\eta - \hat\eta_0$ and next on expanding the log likelihood requires careful choice of a norm in which the error ε is shown to converge to zero, since semiparametric likelihoods often contain ill-behaved terms.

In view of these potential difficulties our approach to proving the chi-squared limit distribution of semiparametric likelihood ratio statistics will be motivated by the approximation (3.2), but not based on it. Fundamental is the observation that the submodel $t \to p_{t,\eta_0 - (i^{-1}_{\eta_0\eta_0} i_{\eta_0\theta_0})^T(t-\theta_0)}$ is *least favorable* at (θ_0, η_0) when estimating θ in the presence of the nuisance parameter η, in the sense that of all submodels $t \to p_{t,\eta_t}$ this submodel has the smallest information about t. This information is precisely the asymptotic variance of $\sqrt{n}(\hat\theta - \theta_0)$. Thus $(1, -i^{-1}_{\eta_0\eta_0} i_{\eta_0\theta_0})$ is the least favorable direction of approach to (θ_0, η_0) when estimating θ in the presence of the unknown η. Approximation (3.2) shows that $\hat\eta$ approaches $\hat\eta_0$ approximately along the least favorable direction.

The derivative of the logarithm of the least favorable submodel with respect to t at zero is called the *efficient score function* and takes the form

$$\tilde\ell_{\theta_0} = \dot\ell_{\theta_0} - (i^{-1}_{\eta_0\eta_0} i_{\eta_0\theta_0})^T \dot\ell_{\eta_0}.$$

The sequence $\hat\theta$ is asymptotically linear in this function in that

$$\sqrt{n}(\hat\theta-\theta_0)=\sqrt{n}\mathbb{P}_n\big[\tilde\ell_{\theta_0}\big]/\big(i_{\theta_0\theta_0}-i_{\theta_0\eta_0}i_{\eta_0\eta_0}^{-1}i_{\eta_0\theta_0}\big)+o_P(1).$$

The efficient score function can also be seen to be equal to $\dot\ell_{\theta_0}-c^T\dot\ell_{\eta_0}$ for the vector c that minimizes $P_{\theta_0\eta_0}(\dot\ell_{\theta_0}-c^T\dot\ell_{\eta_0})^2$.

The notion of a "least favorable submodel" has been extended to semiparametric models. Given a semiparametric model of the type $\{p_{\theta,\eta}\colon \theta\in\Theta,\ \eta\in\mathscr{H}\}$, the score function for θ is defined, as usual, as the partial derivative with respect to θ of the log density. The *efficient score function* for θ is defined as

$$\tilde\ell_\theta=\dot\ell_\theta-\Pi\dot\ell_\theta,$$

where $\Pi\ell$ minimizes the squared distance $P_{\theta\eta}(\ell-k)^2$ over all functions k in the closed linear span of the score functions for η. The inverse of the variance of $\tilde\ell_\theta$ is the Cramér–Rao bound for estimating θ in the presence of η. A submodel $t\to p_{t,\eta_t}$ with $\eta_\theta=\eta$ is defined to be least favorable at (θ,η) if

$$\tilde\ell_\theta=\frac{\partial}{\partial t}\bigg|_{t=\theta}\ln p_{t,\,\eta_t}.$$

Since a projection $\Pi\ell$ on the closed linear span of the nuisance scores is not necessarily a nuisance score itself, least favorable submodels may not always exist. (Problems seem to arise in particular at the maximum likelihood estimator $(\hat\theta,\hat\eta)$, which may happen to be "on the boundary of the parameter set.") However, in all our examples a least favorable submodel exists or can be approximated sufficiently closely.

In the case that the parameter ψ does not factorize naturally into a parameter of interest θ and a nuisance parameter, an efficient score function can be defined and calculated more elegantly in the following manner. (We use this in the example of doubly censored data with $\psi=F$ the distribution function and $\theta=Fg$.) Assume that score functions for the full model can be written in the form

$$\frac{\partial}{\partial t}\bigg|_{t=0}\ln p_{\psi_t}(x)=\ell_\psi h(x),$$

where h is a "direction" in which ψ_t approaches ψ, running through some Hilbert space H, and $\ell_\psi\colon H\to L_2(P_\psi)$ the "score operator." [In the example of doubly censored data H is the set of all functions in $L_2(F)$ with mean Fh zero.] Furthermore, assume that the parameter $\theta\colon\Psi\to\mathbb{R}$ is differentiable in the sense that, for some $\dot\theta_0\in H$,

$$\frac{\partial}{\partial t}\bigg|_{t=0}\theta(\psi_t)=\langle\dot\theta_0,h\rangle_\psi.$$

Then the Cramér–Rao bound for estimating $\theta(\psi)$ equals

$$\sup_h\frac{\langle\dot\theta_0,\,h\rangle_\psi^2}{P_\psi(\ell_\psi h)^2}.\tag{3.3}$$

This supremum can be given a more concrete form by introducing the adjoint operator $\ell_\psi^*\colon L_2(P_\psi) \to H$, which is characterized by the requirement

$$P_\psi(\ell_\psi h)g = \langle \ell_\psi h, g\rangle_{P_\psi} = \langle h, \ell_\psi^* g\rangle_\psi \quad \text{for every } h \in H,\ g \in L_2(P_\psi).$$

With this notation the efficient influence function for θ is defined as the element g_0 of the closure of $\ell_\psi H$ such that

$$\ell_\psi^* g_0 = \dot\theta_0.$$

If g_0 can be written in the form $\ell_\psi h_0$ for some $h_0 \in H$ (it cannot always), and the "information operator" $\ell_\psi^*\ell_\psi$ is invertible, this readily yields the representation

$$g_0 = \ell_\psi h_0, \qquad h_0 = (\ell_\psi^*\ell_\psi)^{-1}\dot\theta_0. \tag{3.4}$$

In this formulation the variance of the efficient influence function g_0 is the Cramér–Rao bound for estimating θ; the inverse of this variance could be defined as the information about θ. Thus g_0 corresponds to $\tilde\ell_\theta/\operatorname{var}\tilde\ell_\theta$ in the case that $\psi = (\theta, \eta)$ is partitioned. The direction h_0 is the least favorable direction in H; the derivative of the logarithm of a least favorable submodel $t \to p_{\psi_t}$, in t at $t = 0$ is equal to $\ell_\psi h$ with

$$h = \frac{(\ell_\psi^*\ell_\psi)^{-1}\dot\theta_0}{\langle \dot\theta_0, (\ell_\psi^*\ell_\psi)^{-1}\dot\theta_0\rangle_\psi}.$$

Note the similarity to (3.1). The Cramér–Rao bound for the submodel in the least favorable direction gives the supremum in (3.3).

3.1. *A general theorem.* In this section we discuss our approach toward obtaining the asymptotic distribution of the likelihood ratio statistic, which is partly motivated by the preceding discussion. Of course, for any $\tilde\psi$ and any $\tilde\psi_0 \in \Psi_0 = \{\psi\colon \theta(\psi) = \theta_0\}$,

$$\operatorname{lrt}_n(\theta_0) = 2n\mathbb{P}_n\big(\ln \operatorname{lik}(\hat\psi) - \ln \operatorname{lik}(\hat\psi_0)\big) \begin{cases} \geq 2n\mathbb{P}_n\big(\ln \operatorname{lik}(\tilde\psi) - \ln \operatorname{lik}(\hat\psi_0)\big), \\ \leq 2n\mathbb{P}_n\big(\ln \operatorname{lik}(\hat\psi) - \ln \operatorname{lik}(\tilde\psi_0)\big). \end{cases}$$

If both the upper and lower bounds converge to a chi-squared distribution on one degree of freedom, then this also holds for the likelihood ratio statistic. We use (3.2) as motivation to define suitable perturbations $\tilde\psi$ and $\tilde\psi_0$ of $\hat\psi$ and $\hat\psi_0$. We define $\tilde\psi_0$ by perturbing $\hat\psi$ in a least favorable direction [so that in view of (3.2) it can be expected to resemble $\hat\psi_0$]; we define $\tilde\psi$ by perturbing $\hat\psi_0$ in a least favorable direction [so that in view of (3.2) it can be expected to resemble $\hat\psi$]. Thus, the perturbations are constructed as elements of submodels that are approximately least favorable.

The following theorem gives a general framework for our approach. Denote P_{ψ_0} by P_0. We assume that as n tends to infinity, the maximum likelihood estimator $\hat\theta = \theta(\hat\psi)$ satisfies

$$\sqrt{n}(\hat\theta - \theta_0) = \sqrt{n}\mathbb{P}_n\tilde\ell/\tilde I + o_P(1), \qquad \tilde I = P_0\tilde\ell^2, \tag{3.5}$$

for a mean zero function $\tilde{\ell}$ with finite and positive variance $\tilde{I}$ under P_0. In all our examples the maximum likelihood estimator is asymptotically efficient and $\tilde{\ell}/\tilde{I}$ is the efficient influence function for estimating θ.

Furthermore, we assume that there exist "approximately least favorable submodels." For every t and ψ suppose that there exists a curve $t \to \boldsymbol{\psi}_t(\psi)$ in Ψ indexed by the parameter of interest t, and passing through ψ at $t = \theta(\psi)$. Technically this means that

$$\theta(\boldsymbol{\psi}_t(\psi)) = t \quad \text{and} \quad \boldsymbol{\psi}_t(\psi)|_{t=\theta(\psi)} = \psi. \tag{3.6}$$

(The proof below uses these curves only for t close to θ_0 and for $\psi = \hat{\psi}$ or $\psi = \hat{\psi}_0$.) The curve $t \to \boldsymbol{\psi}_t(\psi)$ should be approximately least favorable in that the submodel

$$t \to \ell(x; t, \psi) = \ln \text{lik}(\boldsymbol{\psi}_t(\psi), x) \tag{3.7}$$

should be twice continuously differentiable for every x, with derivatives $\dot{\ell}$ and $\ddot{\ell}$ satisfying

$$-\mathbb{P}_n\ddot{\ell}(\cdot; \tilde{\theta}, \tilde{\psi}) \to_{\mathrm{P}} P_0\tilde{\ell}^2 \quad \text{for any random } \tilde{\theta} \to_{\mathrm{P}} \theta_0,\ \tilde{\psi} \to_{\mathrm{P}} \psi_0, \tag{3.8}$$

$$\sqrt{n}\mathbb{P}_n(\dot{\ell}(\cdot; \theta_0, \hat{\psi}_0) - \tilde{\ell}) \to_{\mathrm{P}} 0. \tag{3.9}$$

The idea is to construct the submodel such that the first derivative $\dot{\ell}(\cdot, \theta_0, \psi_0)$ is equal to the efficient score function $\tilde{\ell}$, whence the expectation of its second derivative should be minus the efficient information for θ.

The preceding conditions presume a topology on the set Ψ, and we assume that the maximum likelihood estimators are consistent with respect to this topology.

THEOREM 3.1. *Suppose that the maps $t \to \ell(x; t, \psi)$ are twice continuously differentiable and satisfy* (3.5)–(3.9), *and suppose that $\hat{\psi}$ and $\hat{\psi}_0$ are consistent. Then* $\text{lrt}_n(\theta_0) \rightsquigarrow \chi_1^2$.

PROOF. Since, by (3.6), $\hat{\psi} = \boldsymbol{\psi}_{\hat{\theta}}(\hat{\psi})$,

$$\begin{aligned}\text{lrt}_n(\theta_0) &= 2n\mathbb{P}_n\big[\ln \text{lik}(\hat{\psi}) - \ln \text{lik}(\hat{\psi}_0)\big] \\ &\le 2n\mathbb{P}_n\big[\ln \text{lik}(\boldsymbol{\psi}_{\hat{\theta}}(\hat{\psi})) - \ln \text{lik}(\boldsymbol{\psi}_{\theta_0}(\hat{\psi}))\big] \\ &= 2n\mathbb{P}_n\big[-\dot{\ell}(\cdot; \hat{\theta}, \hat{\psi})(\theta_0 - \hat{\theta}) - 1/2\ddot{\ell}(\cdot; \tilde{\theta}, \hat{\psi})(\theta_0 - \hat{\theta})^2\big],\end{aligned}$$

for some $\tilde{\theta}$ between θ_0 and $\hat{\theta}$. Here the linear term vanishes, because $t \to \mathbb{P}_n \ln \text{lik}(\boldsymbol{\psi}_t(\hat{\psi}))$ is maximized at $t = \hat{\theta}$. An application of (3.8) and (3.5) shows that the right-hand side converges in distribution to a chi-squared distribution.

Since, by (3.6), $\hat\psi_0 = \boldsymbol{\psi}_{\theta_0}(\hat\psi_0)$,

$$\begin{aligned}\operatorname{lrt}_n(\theta_0) &= 2n\mathbb{P}_n\big[\ln\operatorname{lik}(\hat\psi) - \ln\operatorname{lik}(\hat\psi_0)\big]\\ &\geq 2n\mathbb{P}_n\big[\ln\operatorname{lik}(\boldsymbol{\psi}_{\hat\theta}(\hat\psi_0)) - \ln\operatorname{lik}(\boldsymbol{\psi}_{\theta_0}(\hat\psi_0))\big]\\ &= 2n\mathbb{P}_n\big[\dot\ell(\cdot;\theta_0,\hat\psi_0)\big](\hat\theta-\theta_0) + n\mathbb{P}_n\big[\ddot\ell(\cdot;\tilde\theta,\hat\psi_0)(\theta_0-\hat\theta)^2\big],\end{aligned}$$

for some $\tilde\theta$ between θ_0 and $\hat\theta$. Apply (3.9) to get that the first term on the right is equal to $2\sqrt{n}(\hat\theta-\theta_0)\sqrt{n}\mathbb{P}_n\tilde\ell + o_P(1)$ and apply (3.8) to get that the second term is $-P_0\tilde\ell^2 n(\hat\theta-\theta_0)^2 + o_P(1)$. An application of (3.5) shows that the right-hand side converges in distribution to a chi-squared distribution.

The combination of the preceding two paragraphs yields the theorem. □

We note that it is sufficient that the conditions of the theorem are true "with probability tending to 1." Similarly, it suffices that the "paths" $\{\boldsymbol{\psi}_t(\hat\psi): t \in U\}$ and $\{\boldsymbol{\psi}_t(\hat\psi_0): t \in U\}$ belong to the parameter set with probability tending to 1 for a fixed neighborhood U of θ_0 (not depending on n). We keep this in mind when discussing our examples.

EXAMPLE (Doubly censored data). The theorems are applied with the submodel $\boldsymbol{\psi}_t(\psi) = F_t(\theta, F)$, where

$$\begin{aligned}F_t(\theta,F) &= F + (\theta-t)\int^{\cdot}(g^* - Fg^*)\,dF/(-I_F),\\ I_F &= \int g(g^*-Fg^*)\,dF.\end{aligned}\tag{3.10}$$

(Note that $\theta = Fg$ by definition.) The function g^* is the least favorable direction in the F-space at the true distribution F_0 for estimating Fg and is defined by (4.1). It will be shown to be bounded and of bounded variation. The expression I_{F_0} is the inverse of the information $\tilde I$ and is positive. The expression $F_t(\theta, F)$ does not truly define a probability distribution for every $t-\theta$ and F, although it always has total (signed) mass 1. However, in view of the boundedness of g^*, if $t-\theta$, $Fgg^* - F_0gg^*$, $Fg^* - F_0g^*$ and $Fg - F_0g$ are sufficiently close to zero, then $F_t(\theta,\eta)$ has a positive density $1+(\theta-t)(g^*-Fg^*)/I_F$ with respect to F and hence defines a probability distribution.

EXAMPLE (Cox regression for current status data). The theorems are applied with $\boldsymbol{\psi}_t(\psi) = (t, \Lambda_t(\theta,\Lambda))$, where

$$\Lambda_t(\theta,\Lambda) = \Lambda + (\theta-t)\phi(\Lambda)h^{**}\circ\Lambda_0^{-1}\circ\Lambda.\tag{3.11}$$

Here the function $\Lambda_0 h^{**}$ is the least favorable direction for estimating θ in the Λ-space at the true parameter (θ_0, Λ_0) and is defined by (5.1), and ϕ: $[0, M] \to [0,\infty)$ is a fixed function such that $\phi(y) = y$ on the interval $[\Lambda_0(\sigma), \Lambda_0(\tau)]$, such that the function $y \to \phi(y)/y$ is Lipschitz and such that $\phi(y) \leq c(y \wedge (M-y))$ for a sufficiently large constant c specified below [and depending on (θ_0, Λ_0) only]. [By our assumption that $[\Lambda_0(\sigma), \Lambda_0(\tau)] \subset (0, M)$ such a function exists.] The function $\Lambda_t(\theta, \Lambda)$ is essentially Λ plus a perturbation in the least favorable direction, but its definition is somewhat complicated in order to ensure that $\Lambda_t(\theta, \Lambda)$ really defines a cumulative hazard function within our parameter space, at least for t that are sufficiently close to θ. First, the construction using $h^{**} \circ \Lambda_0^{-1} \circ \Lambda$, rather than h^{**} [taken from Huang (1996)], ensures that the perturbation that is added to Λ is absolutely continuous with respect to Λ; otherwise $\Lambda_t(\theta, \Lambda)$ would not be a nondecreasing function. Second, the function ϕ "truncates" the values of the perturbed hazard function to $[0, M]$.

A precise proof that $\Lambda_t(\theta, \Lambda)$ is a parameter is as follows. Since the function ϕ is bounded and Lipschitz and by assumption, $h^{**} \circ \Lambda_0^{-1}$ is bounded and Lipschitz, so is their product and hence, for $u \leq v$ and $|\theta - t| < \varepsilon$,

$$\Lambda_t(\theta, \Lambda)(v) - \Lambda_t(\theta, \Lambda)(u) \geq \big(\Lambda(v) - \Lambda(u)\big)\big(1 - \varepsilon \|\phi h^{**} \circ \Lambda_0^{-1}\|_{\text{Lipschitz}}\big).$$

For sufficiently small ε the right-hand side is nonnegative. Next, for $|\theta - t| < \varepsilon$,

$$\Lambda_t(\theta, \Lambda) \leq \Lambda + \varepsilon \phi(\Lambda) \|h^{**}\|_\infty.$$

This is certainly bounded above by M (on $[0, \tau]$) if $\phi(y) \leq (M - y)/(\varepsilon \|h^{**}\|_\infty)$ for all $0 \leq y \leq M$. Finally, $\Lambda_t(\theta, \Lambda)$ can be seen to be nonnegative on $[\sigma, \tau]$ by the condition that $\phi(y) \leq cy$.

EXAMPLE (Gamma frailty). The theorems are applied with $\boldsymbol{\psi}_t(\psi) = (t, \Lambda_t(\theta, \Lambda))$, where

$$\Lambda_t(\theta, \Lambda) = \Lambda + (\theta - t) \int^{\cdot} k^* \, d\Lambda. \tag{3.12}$$

The function k^* is the least favorable direction in the Λ-space at the true parameter (θ_0, Λ_0) and is defined by (6.1). This function will be shown to be bounded, so that the density $1+(\theta - t)k^*$ of $\Lambda_t(\theta, \eta)$ with respect to Λ is positive for sufficiently small $|\theta - t|$. In that case $\Lambda_t(\theta, \eta)$ defines a nondecreasing function and is a true parameter of the model.

EXAMPLE (Mixture model). The theorems are applied with $\boldsymbol{\psi}_t(\psi) = (t, F_t(\theta, F))$, where

$$F_t(\theta, F)(B) = F\left(B\left(1 + \frac{\theta - t}{2\theta}\right)^{-1}\right). \tag{3.13}$$

This will be shown to be an exact least favorable submodel at F and is well defined whenever $|\theta - t| < 2\theta$.

Having defined suitable submodels, we next need to check the technical conditions of Theorem 3.1. Regarding condition (3.8), we note that

$$\ddot{\ell}(\cdot; t, \psi) = \frac{\partial^2 \text{lik}(\boldsymbol{\psi}_t(\psi))/\partial t^2}{\text{lik}(\boldsymbol{\psi}_t(\psi))} - \dot{\ell}^2(\cdot; t, \psi).$$

In one of our four examples we use paths such that the maps $t \to \mathrm{lik}(\boldsymbol{\psi}_t)$ are linear in which case the first term on the right-hand side is zero. In the other examples the first term has the form $(\partial^2 pt/\partial t^2)/pt$, which is a mean zero function under p_t and can be shown to give a negligible contribution:

$$\mathbb{P}_n \frac{\partial^2 \mathrm{lik}(\boldsymbol{\psi}_t(\psi))/\partial t^2}{\mathrm{lik}(\boldsymbol{\psi}_t(\psi))}\bigg|_{t=\tilde{\theta}} \to_{\mathrm{P}} 0 \quad \text{for every } \tilde{\theta} \to_{\mathrm{P}} \theta_0. \tag{3.14}$$

This could be proved by classical arguments or with the help of the modern Glivenko–Cantelli theorem: if the functions are contained in a Glivenko–Cantelli class, the empirical measure can be replaced by the true measure.

Conditions (3.8) and (3.9) can often be checked with help of the following lemma, which uses concepts from the theory of empirical processes. See, for instance, van der Vaart and Wellner (1996) for a review and methods to check the conditions.

LEMMA 3.2. *Suppose that there exist neighborhoods U of θ_0 and V of ψ_0 such that the class of functions $\{\dot{\ell}(\cdot; t, \psi)\colon t \in U,\ \psi \in V\}$ is P_0-Donsker with square-integrable envelope function. Furthermore, suppose that $\dot{\ell}(x; t, \psi) \to \tilde{\ell}(x)$ for P_0-almost every x, as $t \to \theta_0$ and $\psi \to \psi_0$. Then for all random sequences $\tilde{\theta}_n$ and $\tilde{\psi}_n$ that converge in probability to θ_0 and ψ_0 we have*

$$\mathbb{P}_n \dot{\ell}^2(\cdot; \tilde{\theta}_n, \tilde{\psi}_n) \to_{\mathrm{P}} P_0 \tilde{\ell}^2,$$

$$\sqrt{n}(\mathbb{P}_n - P_0)\big(\dot{\ell}(\cdot; \tilde{\theta}_n, \tilde{\psi}_n) - \tilde{\ell}\big) \to_{\mathrm{P}} 0.$$

PROOF. Assume without loss of generality that $\tilde{\theta}_n$ and $\tilde{\psi}_n$ take their values in U and V, respectively.

By Theorem 4.6 of Giné and Zinn (1986) or Lemma 2.10.14 of van der Vaart and Wellner (1996) the class of squares f^2 of functions f ranging over a Donsker class with square integrable envelope is Glivenko–Cantelli. It follows that

$$\sup_{t \in U,\, \psi \in V} \big|(\mathbb{P}_n - P_0)\dot{\ell}^2(\cdot; t, \psi)\big| \to_{\mathrm{P}} 0.$$

Thus in the first statement the empirical measure may be replaced by the true underlying measure. By assumption $\tilde{\ell}(x) = \dot{\ell}(x; \theta_0, \psi_0)$ almost surely under P_0. Next the result follows by the dominated convergence theorem.

For the second statement define a stochastic process

$$G_n = \big\{\sqrt{n}(\mathbb{P}_n - P_0)\big(\dot{\ell}(\cdot; t, \psi) - \tilde{\ell}\big)\colon t \in U,\ \psi \in V\big\}.$$

The Donsker assumption (and the square integrability of the envelope function) entails that the sequence G_n is asymptotically uniformly continuous in probability, that is, for every $\varepsilon > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \mathrm{P}^*\Big(\sup_{\rho((t,\psi),(t',\psi')) < \delta} \big|G_n(t, \psi) - G_n(t', \psi')\big| > \varepsilon\Big) = 0,$$

where ρ is the semimetric given by its square

$$\rho^2\big((t, \psi), (t', \psi')\big) = P_0\big(\dot\ell(\cdot; t, \psi) - \dot\ell(\cdot; t', \psi')\big)^2.$$

By dominated convergence and consistency of $(\tilde\theta_n, \tilde\psi_n)$ we have

$$\rho^2\big((\tilde\theta, \tilde\psi), (\theta_0, \psi_0)\big) \to_{\mathrm{P}} 0.$$

Conclude that the sequence $G_n(\tilde\theta_n, \tilde\psi_n) - G_n(\theta_0, \psi_0)$ converges in probability to zero. This is the second assertion of the lemma. □

Once (3.14) is verified and if the preceding lemma is applicable, the conditions of Theorem 3.1 effectively reduce to the condition

$$\sqrt{n} P_0 \dot\ell(\cdot; \theta_0, \hat\psi_0) \to_{\mathrm{P}} 0. \tag{3.15}$$

Whereas the conditions of the lemma can be viewed as regularity conditions, this is a structural condition. An "unbiasedness" condition of this type may be expected in view of results of Klaassen (1987), who shows that if θ can be estimated efficiently, then there must exist (consistent) estimators $\hat\ell$ of the efficient score function such that $\sqrt{n} P_0 \hat\ell \to_{\mathrm{P}} 0$. The preceding display requires that the plug-in estimator $\dot\ell(\cdot; \theta_0, \hat\psi_0)$ has the latter property.

A similar condition (with $\hat\psi$ instead of $\hat\psi_0$) also shows up in proofs of the asymptotic normality of the maximum likelihood estimator $\hat\theta$. [See, e.g., Huang (1996) or van der Vaart (1996).]

At first, condition (3.15) appears to require a rate of convergence of $\hat\psi_0$. This is not true, as in many cases $\dot\ell(\cdot; t, \psi)$ is of a special form. For instance, in semiparametric models in which the density is convex linear in the nuisance parameter, the efficient score function for θ is unbiased in the sense that $P_{\theta\eta}\tilde\ell_\theta(\theta, \eta_0) = 0$ for any θ, η and η_0. In our mixture model example we construct the submodel $t \to \boldsymbol{\psi}_t(\psi)$ in such a way that the derivative is exactly the efficient score function, and the unbiasedness condition is trivially satisfied.

In the worst situation (3.15) should not require more than an $o_P(n^{-1/4})$-rate for the nuisance parameter. For instance, if $\psi = (\theta, \eta)$ and $\tilde\ell(\cdot; \theta_0, \hat\psi_0)$ is the efficient score function, then the expression in (3.15) is equal to

$$\sqrt{n}(P_0 - P_{\hat\psi_0})\tilde\ell(\cdot; \theta_0, \hat\psi_0) \approx -\sqrt{n} P_{\hat\psi_0}\tilde\ell(\cdot; \theta_0, \hat\psi_0)\ell_{\hat\eta_0}(\hat\eta - \eta_0) + \sqrt{n} O\big(\|\hat\eta - \eta_0\|^2\big).$$

Here the first term should vanish, because the efficient score function for θ is orthogonal to all scores for the nuisance parameter.

4. Doubly censored data.

The path defined by $dF_t = (1 + th)\,dF$ for a bounded function with $Fh = 0$ yields a score function of the form

$$\ell_F h(u, d) = \frac{\int_{[0,u]} h\,dF}{F(u)} I\{d = 1\} + h(u)I\{d = 2\} + \frac{\int_{(u,\infty)} h\,dF}{1 - F(u)} I\{d = 3\}.$$

For g and h bounded, we have that $P_F\ell_F(h)\ell_F(g) = \langle \ell_F(h), \ell_F(g)\rangle_{P_F} = \langle h, \ell^*\ell_F g\rangle_F = F(h\,\ell^*\ell_F(g))$. So we may use Fubini's theorem to derive the adjoint operator ℓ^*: $L_2(P_F) \to L_2(F)$,

$$\ell^* g(s) = \int_{[s,\infty)} g(u,1)\,dG_L(u) + g(s,2)(G_L - G_R)(s-) + \int_{[0,s)} g(u,3)\,dG_R(u).$$

Thus the information operator takes the form

$$\ell^*\ell_F h(s) = \int_{[s,\infty)} \frac{\int_{[0,z]} h\,dF}{F(z)}\,dG_L(z) + h(s)(G_L - G_R)(s-) + \int_{[0,s)} \frac{\int_{(y,\tau]} h\,dF}{1-F(y)}\,dG_R(y).$$

Since the function g used to define the null hypothesis $Fg = \theta_0$ is assumed to be bounded and of bounded variation, part (i) of Lemma A.2 shows that the function

$$g^* = (\ell^*\ell_{F_0})^{-1} g \tag{4.1}$$

is well defined, bounded and of bounded variation. We use this function to define the (approximately) least favorable submodel (3.10).

In Lemma A.3, we prove consistency of $\hat F_0$ and that

$$(\hat F - \hat F_0)(h) = F_0\big(h(g^* - F_0 g^*)\big)(\hat\theta - \theta_0)/F_0\big(g(g^* - F_0 g^*)\big) + o_P(n^{-1/2})$$

uniformly for uniformly bounded h of uniformly bounded variation. Since the asymptotic variance of $\sqrt{n}(\hat\theta - \theta_0)$ is $F_0(g(g^* - F_0 g^*))$, this confirms the intuition expressed in Section 3. In the verification of Lemma A.3(ii) we show that $\|\hat F - F_0\|_\infty = O_P(n^{-1/2})$ and $\|\hat F_0 - F_0\|_\infty = O_P(n^{-1/2})$ and that $(\hat\theta - \theta_0) = \mathbb{P}_n \ell_{F_0}(g^* - F_0 g^*) + o_P(n^{-1/2})$ so $\tilde\ell/\tilde I = \ell_{F_0}(g^* - F_0 g^*)$.

Recalling (3.7) and (3.10), we have

$$\dot\ell(u,d;t,F) = \frac{\ell_F(g^* - Fg^*)(u,d)/I_F}{1 + (t - Fg)\ell_F(g^* - Fg^*)(u,d)/I_F}, \qquad I_F = Fg(g^* - Fg^*).$$

Given a bounded, monotone function h the function $\ell_F h$ is composed of three bounded and monotone functions, with the same uniform bound. Since g^* is bounded and of bounded variation, it follows that the class of functions $\ell_F g^*$ with F ranging over all distributions on $[\sigma,\tau]$ consists of uniformly bounded functions of uniformly bounded variation, hence is a Donsker class. For $\|F - F_0\|_\infty$ sufficiently small we have that I_F is close to I_{F_0}, Fg is close to $F_0 g$ and Fg^* is close to $F_0 g^*$. Given Donsker classes $\mathscr{F}_1, \dots, \mathscr{F}_k$ and a Lipschitz function ϕ: $\mathbb{R}^k \to \mathbb{R}$, a uniformly bounded class of functions $x \to \phi\big(f_1(x), \dots, f_k(x)\big)$; $f_i \in \mathscr{F}_i$, $i = 1, \dots, k$, is Donsker by Theorem 2.10.6 of van der Vaart and Wellner (1996). It follows that the class of functions $\dot\ell(u,d;t,F)$ with t sufficiently close to $\theta_0 = F_0 g$ and $\|F - F_0\|_\infty$ sufficiently small is Donsker. As $(t,F) \to (\theta_0, F_0)$ these functions converge a.s.-P_{F_0} to $\tilde\ell = \dot\ell(\cdot;\theta_0,F_0)$. Thus the conditions of Lemma 3.2 are satisfied. Note that (3.14) is trivially satisfied, since the derivative of $\dot\ell(u,d;t,F)$ with respect to

t is $-(\dot{\ell}(u, d; t, F))^2$. For the application of our Theorem 3.1 it suffices to show that

$$\sqrt{n} P_0 \ell_{\hat{F}_0}(g^* - \hat{F}_0 g^*) = \sqrt{n}(P_0 - P_{\hat{F}_0})\ell_{\hat{F}_0} g^* = \sqrt{n}(F_0 - \hat{F}_0)\ell^* \ell_{\hat{F}_0} g^* \to_{\mathrm{P}} 0.$$

Since $(\hat{F}_0 - F_0)g = 0$ the absolute value of this expression is equal to, in view of the definition of g^* and that the support of $\hat{F}_0$ is contained in $[\sigma, \tau]$,

$$\sqrt{n}\big|(F_0 - \hat{F}_0)\big(\ell^* \ell_{\hat{F}_0} g^* - \ell^* \ell_{F_0} g^*\big)\big| \le 2\sqrt{n}\|\hat{F}_0 - F_0\|_\infty \left\|\ell^* \ell_{\hat{F}_0} g^* - \ell^* \ell_{F_0} g^*\right\|_{\mathrm{BV}},$$

where $||\cdot||_{\mathrm{BV}}$ is the sum of the supremum and total variation norms. The first term on the right-hand side is bounded in probability; the second converges to zero in probability by Lemma A.2(ii).

5. Regression for current status data. In this model the score function for θ takes the form

$$\ell_\theta(\theta, \Lambda)(x) = z\Lambda(y)Q(x; \theta, \Lambda),$$

for the function $Q(x; \theta, \Lambda)$ given by

$$Q(x; \theta, \Lambda) = e^{\theta z}\left[\delta \frac{\exp(-\exp(\theta z)\Lambda(y))}{1 - \exp(-\exp(\theta z)\Lambda(y))} - (1 - \delta)\right].$$

Inserting into the log likelihood a submodel $t \to \Lambda_t$ such that $h(y) = -\partial/\partial t|_{t=0}\ \Lambda_t(y)$ exists for every y, and differentiating at $t = 0$ we obtain a score function for Λ of the form

$$\ell_\Lambda(\theta, \Lambda)h(x) = h(y)Q(x; \theta, \Lambda).$$

For every nondecreasing, nonnegative function h the submodel $\Lambda_t = \Lambda + th$ is well defined for $t \ge 0$ and yields a (one-sided) derivative h at $t = 0$. Thus the preceding display gives a (one-sided) score for Λ at least for all h of this type. The linear span of these functions contains $\ell_\Lambda h$ for all bounded functions h of bounded variation. The efficient score function for θ is defined as $\ell_\theta - \ell_\Lambda h^*$ for h^* minimizing the distance $P_{\theta\Lambda}(\ell_\theta - \ell_\Lambda h)^2$. In view of the similar structure of the scores for θ and Λ this is a weighted least squares problem with weight function $Q(y, \delta, z; \theta, \Lambda)$. The solution at the true parameters is given by

$$h^*(Y) = \Lambda_0(Y)h^{**}(Y) = \Lambda_0(Y)\frac{E_{\theta_0\Lambda_0}\big(ZQ^2(X; \theta_0, \Lambda_0)|Y\big)}{E_{\theta_0\Lambda_0}\big(Q^2(X; \theta_0, \Lambda_0)|Y\big)}. \tag{5.1}$$

As the formula shows (and as follows from the nature of the minimization problem) the function h^{**} is unique only up to null sets for the distribution of Y. However, it is an assumption that (under the true parameters) there exists a version of the conditional expectation that is differentiable with bounded derivative. Following Huang (1996) this version is used to define the least favorable submodels (3.11). By the assumption that $P_0 \operatorname{var}(Z|Y) > 0$, the efficient score function $\tilde{\ell} = \ell_\theta(\theta_0, \Lambda_0) - \ell_\Lambda(\theta_0, \Lambda_0)h^*$, the difference between the θ-score and its projection, is nonzero, whence the efficient information $\tilde{I}$ for θ is positive [at (θ_0, Λ_0)].

The consistency of $(\hat{\theta}, \hat{\Lambda})$ and $\hat{\Lambda}_0$ can be proved by a standard consistency proof, where we may start from the inequality $\mathbb{P}_n \ln(\mathrm{lik}(\hat{\psi}) + \mathrm{lik}(\psi_0)) \geq \mathbb{P}_n \ln 2\mathrm{lik}(\psi_0)$, rather than from the more obvious inequality $\mathbb{P}_n \ln \mathrm{lik}(\hat{\psi}) \geq \mathbb{P}_n \ln \mathrm{lik}(\psi_0)$. (The latter inequality implies the first by the concavity of ln.) See, for instance, the consistency proof in Huang and Wellner (1995), or also the Appendix. The identifiability of the parameters is ensured by the assumption that $P_0 \operatorname{var}(Z|Y) > 0$. More precisely, it can be shown under our conditions that on a set of probability 1, $\hat{\Lambda}$ and $\hat{\Lambda}_0$ converge to Λ_0 uniformly on the interval $[\sigma + \varepsilon, \tau - \varepsilon]$, for every $\varepsilon > 0$. Of course, the estimators and the true distribution are not identifiable outside the interval $[\sigma, \tau]$. The consistency of the estimators at σ and τ seems dubious, even though this is used by Huang (1996) in some of his proofs. In the Appendix, we show that the asymptotic normality and efficiency (3.5) of $\hat{\theta}$ remains valid even without the uniform consistency on $[\sigma, \tau]$. We also show that both $\int_\sigma^\tau (\hat{\Lambda} - \Lambda_0)^2(y)\,dy$ [as asserted by Huang (1996)] and $\int_\sigma^\tau (\hat{\Lambda}_0 - \Lambda_0)^2(y)\,dy$ converge to zero at the rate $O_P(n^{-2/3})$. We shall use this to verify condition (3.15).

In view of (3.11), we have

$$\dot{\ell}(x; t, \Lambda, \theta) = \left[z - \frac{\phi(\Lambda)(y)}{\Lambda_t(\theta, \Lambda)(y)} h^{**} \circ \Lambda_0^{-1} \circ \Lambda(y)\right] \Lambda_t(\theta, \Lambda)(y) Q\big(x; t, \Lambda_t(\theta, \Lambda)\big).$$

For (t, Λ, θ) tending to $(\theta_0, \Lambda_0, \theta_0)$ this function converges almost everywhere to its value at $(\theta_0, \Lambda_0, \theta_0)$, which, by construction, is the efficient score function $\tilde{\ell}$ for θ at (θ_0, Λ_0). Furthermore, the class of functions $\dot{\ell}(x; t, \Lambda, \theta)$ with (t, θ) varying over a small neighborhood of (θ_0, θ_0) and with Λ ranging over all nondecreasing cadlag functions with range in $[0, M]$ can be seen to be a Donsker class by repeated application of preservation properties for Donsker classes [cf. van der Vaart and Wellner (1996), Chapter 2.11]. Note here that, since the function $u \to ue^{-u}/(1 - e^{-u})$ is bounded and Lipschitz on $[0, \infty)$, we can write

$$\Lambda(y) Q(x; t, \Lambda) = \psi\big(e^{tz}, \Lambda(y)\big)$$

for a function ψ that is bounded and Lipschitz in its two arguments. Thus, since the classes of functions $z \to e^{tz}$ and $y \to \Lambda(y)$ are Donsker, so is the class of functions $x \to \Lambda(y) Q(x; t, \Lambda)$. Next, since the function $\phi(y)/y$ is bounded and Lipschitz, and $h^* \circ \Lambda_0^{-1}$ is assumed bounded and Lipschitz,

$$\frac{\phi(\Lambda)}{\Lambda_t(\theta, \Lambda)} = \frac{\phi(\Lambda)/\Lambda}{1 + (\theta - t)\phi(\Lambda)/\Lambda\, h^{**} \circ \Lambda_0^{-1} \circ \Lambda} = \chi(\theta - t, \Lambda)$$

for a function χ that is bounded and Lipschitz on an appropriate domain. Next, the product of the two classes of functions in the preceding displays, which are both uniformly bounded and Donsker, is Donsker, and so on.

Thus, the assumptions of Lemma 3.2 are valid, whence the assumptions of Theorem 3.1 have been verified, except for (3.14) and (3.15). For the first of these two conditions we compute that

$$\frac{\partial^2 \mathrm{lik}\big(t, \Lambda_t(\theta,\Lambda)\big)/\partial t^2}{\mathrm{lik}\big(t,\Lambda_t(\theta,\Lambda)\big)} = Q(\cdot\,; t, \Lambda_t(\theta,\Lambda)) \\ \times \Big[z^2\Lambda_t(\theta,\Lambda) - 2\phi(\Lambda)h^{**}\circ\Lambda_0^{-1}\circ\Lambda \\ - e^{tz}\big(z\Lambda_t(\theta,\Lambda) - \phi(\Lambda)h^{**}\circ\Lambda_0^{-1}\circ\Lambda\big)^2\Big].$$

By the same arguments as before these functions are in a Donsker class, hence certainly in a Glivenko–Cantelli class. Thus the empirical measure $\mathbb{P}_n$ in (3.14) can be replaced by the true measure P_0. As (t,θ,Λ) converges to $(\theta_0,\theta_0,\Lambda_0)$ the functions in the preceding display converge almost everywhere to $(\partial^2 p_{t,\Lambda_t(\theta_0,\Lambda_0)}/\partial t^2)/p_{t,\Lambda_t(\theta_0,\Lambda_0)}$ evaluated at $t=\theta_0$. This has mean zero. We can conclude the proof of (3.14) by the dominated convergence theorem.

Abbreviating $\dot\ell(\cdot\,;\theta_0,\Lambda,\theta_0)$ to $\dot\ell(\Lambda)$, we can rewrite the expectation in (3.15) in the form

$$P_0\dot\ell(\hat\Lambda_0) = (P_0 - P_{\theta_0,\hat\Lambda_0})\dot\ell(\Lambda_0) + (P_0 - P_{\theta_0,\hat\Lambda_0})\big(\dot\ell(\hat\Lambda_0) - \dot\ell(\Lambda_0)\big). \tag{5.2}$$

We shall show that both terms on the right-hand side are of the order $O_P(n^{-2/3})$ and hence certainly $o_P(n^{-1/2})$. Since $\dot\ell(\Lambda_0)$ is the efficient score function for θ and hence is orthogonal to every Λ-score, the first term can be rewritten as

$$P_0\dot\ell(\Lambda_0)\big[(p_0 - p_{\theta_0,\hat\Lambda_0})/p_0 - \ell_\Lambda(\theta_0,\Lambda_0)(\Lambda_0 - \hat\Lambda_0)\big].$$

The second term in square brackets is exactly the linear approximation in $\Lambda_0 - \hat\Lambda_0$ of the first. Taking the Taylor expansion one term further shows that the term in square brackets is bounded by a multiple of $(\Lambda_0 - \hat\Lambda_0)^2$ and hence the display is bounded by a multiple of $P_0(\Lambda_0 - \hat\Lambda_0)^2$. The second term in (5.2) can be bounded similarly, since both $\Lambda \to p_{\theta_0,\Lambda}$ and $\Lambda \to \dot\ell(\theta_0,\Lambda,\theta_0)$ are uniformly Lipschitz functions.

6. Gamma frailty.

The natural log of (2.1) is given by

$$\ln \mathrm{lik}(\theta,\Lambda,X) = \int_0^\tau \ln\big(1+\theta N(u-)\big)\,dN(u) \\ - \big(1+\theta N(\tau)\big)\theta^{-1}\ln\Big(1+\theta\int_0^\tau Y d\Lambda\Big) \\ + \int_0^\tau \ln\big(Y(u)\Delta\Lambda(u)\big)\,dN(u),$$

where if $\theta = 0$, $\theta^{-1}\ln(1+\theta\int_0^\tau Y\,d\Lambda)$ is set to its limit, $\int_0^\tau Y\,d\Lambda$. The score function for θ is

$$\ell_\theta(\theta,\Lambda) = \int_0^\tau \frac{N(u-)}{1+\theta N(u-)}\,dN(u) \\ - \theta^{-1}\int_0^\tau Y\,d\Lambda\,\frac{1+\theta N(\tau)}{1+\theta\int_0^\tau Y\,d\Lambda} + \theta^{-2}\ln\Big(1+\theta\int_0^\tau Y\,d\Lambda\Big),$$

where, in the case that $\theta = 0$ the last two terms above are replaced by their limit,

$$-N(\tau)\int_0^\tau Y\,d\Lambda + \tfrac{1}{2}\biggl(\int_0^\tau Y\,d\Lambda\biggr)^2.$$

The path defined by $d\Lambda_t = (1 + th_2)\,d\Lambda$ for a bounded function, h_2, yields a score function for Λ of the form,

$$\ell_\Lambda(\theta, \Lambda)\biggl[\int^{\cdot} h_2\,d\Lambda_1\biggr] = \int_0^\tau h_2\,dN - \frac{1 + \theta N(\tau)}{1 + \theta\int_0^\tau Y\,d\Lambda}\int_0^\tau Y h_2\,d\Lambda_1.$$

Also define three second derivatives. The second derivative $\ell_{\theta\theta}(\theta, \Lambda)$ is obtained by simply differentiating $\ell_\theta(\theta, \Lambda)$ with respect to θ. For bounded h_2 and g_2, the remaining second derivatives are defined as

$$\begin{aligned}
\ell_{\theta\Lambda}(\theta, \Lambda)\biggl[\int^{\cdot} h_2\,d\Lambda_1\biggr] &= -\biggl(\frac{\int_0^\tau Y h_2\,d\Lambda_1}{1 + \theta\int_0^\tau Y\,d\Lambda}\biggr)\\
&\quad\times\biggl(N(\tau) - \frac{1 + \theta N(\tau)}{1 + \theta\int_0^\tau Y\,d\Lambda}\int_0^\tau Y\,d\Lambda\biggr)\\
\ell_{\Lambda\Lambda}(\theta, \Lambda)\biggl[\int^{\cdot} h_2\,d\Lambda_1, \int^{\cdot} g_2 d\Lambda_2\biggr] &= \frac{1 + \theta N(\tau)}{(1 + \theta\int_0^\tau Y\,d\Lambda)^2}\int^{\cdot} h_2 Y\,d\Lambda_1 \int^{\cdot} g_2 Y\,d\Lambda_2\\
&\quad - \frac{1 + \theta N(\tau)}{1 + \theta\int_0^\tau Y\,d\Lambda}\int_0^\tau h_2 g_2 Y\,d\Lambda_2.
\end{aligned}$$

Finally, letting $\mathrm{BV}[0, \tau]$ denote the functions h: $[0, \tau] \to \mathbb{R}$ which are bounded and of bounded variation, equipped with the norm $\|\cdot\|_{\mathrm{BV}} = \|\cdot\|_\infty + \|\cdot\|_{\mathrm{var}}$, define operators $\sigma_{\Lambda\Lambda}$: $\mathrm{BV}[0, \tau] \to \mathrm{BV}[0, \tau]$ and $\sigma = (\sigma_1, \sigma_2)$: $\mathbb{R} \times \mathrm{BV}[0, \tau] \to \mathbb{R} \times \mathrm{BV}[0, \tau]$ by

$$\begin{aligned}
\sigma_{\Lambda\Lambda}[h_2](u) &= P_0\biggl[-Y(u)\theta_0\frac{1 + \theta_0 N(\tau)}{(1 + \theta_0\int_0^\tau Y\,d\Lambda_0)^2}\int^{\cdot} h_2 Y\,d\Lambda_0\biggr]\\
&\quad + h_2(u)P_0\biggl[Y(u)\frac{1 + \theta N(\tau)}{1 + \theta\int_0^\tau Y\,d\Lambda_0}\biggr],\\
\sigma_1(h_1, h_2) &= -h_1 P_0\ell_{\theta\theta}(\theta_0, \Lambda_0) - P_0\ell_{\theta\Lambda}(\theta_0, \Lambda_0)\biggl[\int^{\cdot} h_2\,d\Lambda_0\biggr]\\
\sigma_2(h_1, h_2)(u) &= h_1 P_0\biggl[\frac{Y(u)}{1 + \theta_0\int_0^\tau Y\,d\Lambda_0}\biggl(N(\tau) - \frac{1 + \theta_0 N(\tau)}{1 + \theta_0\int_0^\tau Y\,d\Lambda_0}\int_0^\tau Y\,d\Lambda_0\biggr)\biggr]\\
&\quad + \sigma_{\Lambda\Lambda}[h_2].
\end{aligned}$$

These operators arise in the proof of asymptotic normality of the maximum likelihood estimators given in Murphy (1995a). They also appear in information calculations in this model. Indeed the operator $\sigma_{\Lambda\Lambda}$ is the information operator $\ell_\Lambda^*\ell_\Lambda$ connected to the nuisance parameter Λ; that is,

$$\langle g_2, \sigma_{\Lambda\Lambda}[h_2]\rangle_\Lambda = \left\langle \ell_\Lambda(\theta_0,\Lambda_0)\left[\int^{\cdot} g_2\,d\Lambda_0\right], \ell_\Lambda(\theta_0,\Lambda_0)\left[\int^{\cdot} h_2 d\Lambda_0\right]\right\rangle_{P_0}.$$

This follows from the identity

$$\begin{aligned}\int_0^\tau g_2\sigma_{\Lambda\Lambda}(h)\,d\Lambda_0 &= -P_0\ell_{\Lambda\Lambda}(\theta_0,\Lambda_0)\left[\int^{\cdot} h_2\,d\Lambda_0, \int^{\cdot} g_2\,d\Lambda_0\right]\\ &= P_0\ell_\Lambda(\theta_0,\Lambda_0)\left[\int^{\cdot} h_2\,d\Lambda_0\right]\ell_\Lambda(\theta_0,\Lambda_0)\left[\int^{\cdot} g_2\,d\Lambda_0\right].\end{aligned}$$

Similarly, the operator σ is the information operator $\ell_\psi^*\ell_\psi(\psi_0)$ for the full parameter $\psi = (\theta,\Lambda)$ of the model, for which the score function can be written as $\ell_\psi(\psi)(h_1,h_2)$ and is defined as $h_1\ell_\theta(\theta,\Lambda) + \ell_\Lambda(\theta,\Lambda)[\int^{\cdot} h_2\,d\Lambda]$. The parameter of interest $\theta(\psi) = \theta$ has derivative $\dot\theta_0 = (1,0)$, since

$$\left.\frac{d}{dt}\right|_{t=0}\theta(\theta + th_1, \Lambda_t) = h_1 = \langle (h_1,h_2),(1,0)\rangle_{\mathbb{R}\times\Lambda}.$$

Thus by the general theory [see (3.4)] we can find the influence function for estimating θ, letting $\tilde\sigma = (\tilde\sigma_1,\tilde\sigma_2)$ be the inverse of σ, as

$$\ell_\psi(\psi_0)(\ell_\psi^*\ell_\psi(\psi_0))^{-1}(1,0) = \tilde\sigma_1(1,0)\ell_\theta(\theta_0,\Lambda_0) + \ell_\Lambda(\theta_0,\Lambda_0)\left[\int^{\cdot}\tilde\sigma_2(1,0)\,d\Lambda_0\right].$$

The following lemma shows that this function is well defined. Denote the subset of $\mathrm{BV}[0,\tau]$, with norm bounded above by p, by $\mathrm{BV}_p[0,\tau]$.

LEMMA 6.1. *Under the conditions of Theorem* 2.3 *the following hold*:

(i) $\sigma_{\Lambda\Lambda}$: $\mathrm{BV}[0,\tau] \to \mathrm{BV}[0,\tau]$ *is continuously invertible with inverse* $\tilde\sigma_{\Lambda\Lambda}$;

(ii) σ: $\mathbb{R}\times\mathrm{BV}[0,\tau] \to \mathbb{R}\times\mathrm{BV}[0,\tau]$ *is continuously invertible with inverse* $\tilde\sigma$;

(iii) $\sqrt{n}\|\hat\Lambda - \Lambda_0\|_\infty$ *is* $O_P(1)$;

(iv) $\sqrt{n}(\hat\Lambda_0(\cdot) - \Lambda_0(\cdot)) \rightsquigarrow Z(\cdot)$ *on* $l^\infty(\mathrm{BV}_p[0,\tau])$, *where* Z *is a tight Gaussian process with mean zero and covariance process* $\mathrm{covar}(Z(h_2), Z(g_2)) = \int_0^\tau g_2\tilde\sigma_{\Lambda\Lambda}(h_2)\,d\Lambda_0$;

(v) $\sqrt{n}(\hat\theta - \theta_0) = \tilde\sigma_1(1,0)\mathbb{P}_n[\ell_\theta(\theta_0,\Lambda_0) - \ell_\Lambda(\theta_0,\Lambda_0)[\int^{\cdot} k^* d\Lambda_0]] + o_P(1)$, *where*

$$k^*(u) = -\tilde\sigma_1(1,0)^{-1}\tilde\sigma_2(1,0)(u). \tag{6.1}$$

PROOF. Under assumptions (i)–(iii) of Theorem 2.3 the conditions of Murphy (1994, 1995a) are satisfied, implying that (ii), (iii) and (v) hold. Items (i) and (iv) can be proved by following virtually identical steps to those in Murphy (1994, 1995a). □

The continuous invertibility in (i) and (ii) imply that $\tilde\sigma_1(1,0) > 0$. From (v) we have that $\tilde\ell$ of (3.5) is equal to $\ell_\theta(\theta_0,\Lambda_0) - \ell_\Lambda(\theta_0,\Lambda_0)[\int^{\cdot} k^*\,d\Lambda_0]$ and $\tilde I$ is equal to $\tilde\sigma_1(1,0)^{-1}$.

The least favorable submodels are given by (3.12). The continuous invertibility of σ implies that the function k^* is uniformly bounded on $[0, \tau]$. Thus these submodels define true cumulative hazard functions for $\hat{\theta}$ and t sufficiently close to θ_0.

Recalling (3.7) and (3.12), we see that

$$\dot{\ell}(\cdot; t, \theta, \Lambda) = \ell_\theta\big(t, \Lambda_t(\theta, \Lambda)\big) - \ell_\Lambda\big(t, \Lambda_t(\theta, \Lambda)\big)\left[\int^{\cdot} k^* \, d\Lambda\right]$$

$$\begin{aligned}\ddot{\ell}(\cdot; t, \theta, \Lambda) = {}& \ell_{\theta\theta}\big(t, \Lambda_t(\theta, \Lambda)\big) - 2\ell_{\theta\Lambda}\big(t, \Lambda_t(\theta, \Lambda)\big)\left[\int^{\cdot} k^* \, d\Lambda\right] \\ & + \ell_{\Lambda\Lambda}\big(t, \Lambda_t(\theta, \Lambda)\big)\left[\int^{\cdot} k^* \, d\Lambda, \int^{\cdot} k^* \, d\Lambda\right].\end{aligned}$$

To verify (3.8), note that

$$\begin{aligned}-P_0&\left[\ell_{\theta\theta}(\theta_0, \Lambda_0) - 2\ell_{\theta\Lambda}(\theta_0, \Lambda_0)\left[\int^{\cdot} k^* \, d\Lambda_0\right] + \ell_{\Lambda\Lambda}(\theta_0, \Lambda_0)\left[\int^{\cdot} k^* \, d\Lambda_0, \int^{\cdot} k^* \, d\Lambda_0\right]\right] \\ &= \sigma_1(1, -k^*) + \int \sigma_2(1, -k^*)(u)(-k^*(u)) \, d\Lambda_0 \\ &= \tilde{\sigma}_1(1, 0)^{-1}\left(\sigma_1(\sigma^{-1}(1, 0)) + \int \sigma_2(\sigma^{-1}(1, 0))(u)(-k^*(u)) \, d\Lambda_0\right) \\ &= \tilde{\sigma}_1(1, 0)^{-1}(1 + 0).\end{aligned}$$

Recall that Y is nonincreasing and bounded and N is nondecreasing and bounded. This implies that the derivatives $\ell_\theta(\theta, \Lambda)$, $\ell_\Lambda(\theta, \Lambda)[\int h_2 \, d\Lambda_1]$, $\ell_{\theta\theta}(\theta, \Lambda)$, $\ell_{\theta\Lambda}(\theta, \Lambda)[\int h_2 \, d\Lambda_1]$ and $\ell_{\Lambda\Lambda}(\theta, \Lambda)[\int h_2 \, d\Lambda_1, \int g_2 \, d\Lambda_2]$ are continuous in $(\theta, \Lambda, \Lambda_1, \Lambda_2)$, uniformly in (N, Y), with respect to the Euclidean topology on $\theta \in [-\varepsilon, M]$ and the uniform norm on the cumulative hazard functions, which range over all cumulative hazard functions in $\mathrm{BV}_p[0, \tau]$ for some $p < \infty$. Thus, (3.8) is verified.

The remaining condition (3.9) takes the form

$$\sqrt{n}\mathbb{P}_n\left[\dot{\ell}(\cdot; \theta_0, \theta_0, \hat{\Lambda}_0) - \ell_\theta(\theta_0, \Lambda_0) + \ell_\Lambda(\theta_0, \Lambda_0)\left[\int^{\cdot} k^* \, d\Lambda_0\right]\right]$$

converges to zero in probability. To see this, substitute in for $\dot{\ell}$ and add and subtract $-\ell_{\theta\Lambda}(\theta_0, \Lambda_0)[\hat{\Lambda}_0 - \Lambda_0] + \ell_{\Lambda\Lambda}(\theta_0, \Lambda_0)[\int^{\cdot} k^* \, d\Lambda_0, \hat{\Lambda}_0 - \Lambda_0]$ to get

$$\sqrt{n}\mathbb{P}_n[\ell_\theta(\theta_0, \Lambda_0) - \ell_\theta(\theta_0, \hat{\Lambda}_0) - \ell_{\theta\Lambda}(\theta_0, \Lambda_0)[\hat{\Lambda}_0 - \Lambda_0]] \tag{6.2}$$

$$\begin{aligned}+ \sqrt{n}\mathbb{P}_n&\left[-\ell_\Lambda(\theta_0, \hat{\Lambda}_0)\left[\int^{\cdot} k^* \, d\hat{\Lambda}_0\right] + \ell_\Lambda(\theta_0, \Lambda_0)\left[\int^{\cdot} k^* \, d\Lambda_0\right]\right. \\ &\left. + \ell_{\Lambda\Lambda}(\theta_0, \Lambda_0)\left[\int^{\cdot} k^* \, d\Lambda_0, \hat{\Lambda}_0 - \Lambda_0\right]\right]\end{aligned} \tag{6.3}$$

$$- \sqrt{n}\mathbb{P}_n\left[-\ell_{\theta\Lambda}(\theta_0, \Lambda_0)[\hat{\Lambda}_0 - \Lambda_0] + \ell_{\Lambda\Lambda}(\theta_0, \Lambda_0)\left[\int^{\cdot} k^* d\Lambda_0, \hat{\Lambda}_0 - \Lambda_0\right]\right]. \tag{6.4}$$

We show that each of the three terms converges in probability to zero. Via tedious algebraic arguments, we get that, for $\theta_0 \neq 0$, (6.2) is equal to

$$\sqrt{n}\mathbb{P}_n\left[\theta_0^{-2}\left\{\ln\left[1+\frac{\theta_0\int_0^\tau Y d(\hat{\Lambda}_0-\Lambda_0)}{1+\theta_0\int_0^\tau Y\,d\Lambda_0}\right]-\frac{\theta_0\int_0^\tau Y\,d(\hat{\Lambda}_0-\Lambda_0)}{1+\theta_0\int_0^\tau Y\,d\Lambda_0}\right\}\right]$$
$$+\sqrt{n}\mathbb{P}_n\left[\frac{1+\theta_0 N(\tau)}{(1+\theta_0\int_0^\tau Y\,d\Lambda_0)^2}\frac{(\int_0^\tau Y\,d(\hat{\Lambda}_0-\Lambda_0))^2}{1+\theta_0\int_0^\tau Y\,d\hat{\Lambda}_0}\right].$$

If $\theta_0 = 0$, then (6.2) is $\sqrt{n}2\mathbb{P}_n[\int_0^\tau Y\,d(\hat{\Lambda}_0-\Lambda_0)]^2$. Using more tedious arguments, we have that (6.3) is equal to

$$\sqrt{n}\mathbb{P}_n\left[\theta_0\frac{1+\theta_0 N(\tau)}{(1+\theta_0\int_0^\tau Y\,d\Lambda_0)(1+\theta_0\int_0^\tau Y\,d\hat{\Lambda}_0)}\right.$$
$$\times\left[\int_0^\tau Y\,d(\hat{\Lambda}_0-\Lambda_0)\int_0^\tau Yk^*\,d(\hat{\Lambda}_0-\Lambda_0)\right.$$
$$\left.\left.-\theta_0\int_0^\tau Yk^*\,d\Lambda_0\frac{(\int_0^\tau Yk^*\,d(\hat{\Lambda}_0-\Lambda_0))^2}{1+\theta_0\int_0^\tau Y\,d\hat{\Lambda}_0}\right]\right].$$

Recall that the total variation norms of both Y and k^* are bounded by constants and that $N(\tau)$ is also bounded by a constant. So both (6.2) and (6.3) are $O_P(1)\sqrt{n}\|\hat{\Lambda}_0-\Lambda\|_\infty^2$.

All that is left is to prove that (6.4) converges to zero in probability. This term is equal to $\sqrt{n}\int_0^\tau h_n(u)\,d(\hat{\Lambda}_0-\Lambda_0)(u)$, where h_n is the function

$$h_n(u)=\mathbb{P}_n\left[\frac{1+\theta_0 N(\tau)}{1+\theta_0\int_0^\tau Y\,d\Lambda_0}(k^*(u)-\frac{\theta_0\int_0^\tau Yk^*\,d\Lambda_0}{1+\theta_0\int_0^\tau Y\,d\Lambda_0})\right.$$
$$\left.-\frac{Y(u)}{1+\theta_0\int_0^\tau Y\,d\Lambda_0}\left(N(\tau)-\int_0^\tau Y\,d\Lambda_0\frac{1+\theta_0 N(\tau)}{1+\theta_0\int_0^\tau Y\,d\Lambda_0}\right)\right].$$

Note that the expectation of h_n is $-\sigma_2(1,-k^*)$, which is $-\tilde{\sigma}_1(1,0)^{-1}\times\sigma_2(\sigma^{-1}(1,0))=0$. Apply Rao's (1963) strong law of large numbers to get that $\|h_n\|_\infty$ converges almost surely to zero. Also note that the total variation norm of h_n is uniformly bounded in n by a constant. Put $Z_n(h)=\sqrt{n}\int_0^\tau h\,d(\hat{\Lambda}_0-\Lambda_0)$, for $h\in \mathrm{BV}_p[0,\tau]$. Then (iv) implies that Z_n is asymptotically uniformly ρ_2-equicontinuous in probability [see, e.g., van der Vaart and Wellner (1996)], where $\rho_2(h,g)^2=\int_0^\tau(h(u)-g(u))\sigma_{\Lambda\Lambda}^{-1}(h-g)(u)\,d\Lambda_0(u)$. Note that $\rho_2(h_n,h_n)$ converges to zero in probability. This combined with the asymptotically uniform equicontinuity of Z_n implies that $Z_n(h_n)$ converges in probability to zero.

7. A mixture model.

The score function for θ, the derivative of the log density with respect to θ, is given by

$$\ell_\theta(\theta,F)(u,v)=\frac{\int(\theta^{-1}-zv)z^2e^{-z(u+\theta v)}\,dF(z)}{\int z^2e^{-z(u+\theta v)}\,dF(z)}.$$

In this model the statistic $T_1 + \theta T_2$ is sufficient for F. The *conditional score function* $\tilde{\ell}_\theta(\theta, F)$ is defined as

$$\begin{aligned}\tilde{\ell}_\theta(\theta, F)(u, v) &= \ell_\theta(\theta, F)(u, v) - E_\theta\big(\ell_\theta(\theta, F)(T_1, T_2)|T_1 + \theta T_2 = u + \theta v\big)\\ &= \frac{\int (1/2)(u - \theta v) z^3 e^{-z(u+\theta v)}\, dF(z)}{\int \theta z^2 e^{-z(u+\theta v)}\, dF(z)}.\end{aligned}$$

By an easy calculation we see that, with $F_t(\theta, F)$ the submodel given by (3.13),

$$\tilde{\ell}_\theta(\theta, F)(u, v) = \frac{\partial}{\partial t}\bigg|_{t=\theta} \ln p_{t, F_t(\theta, F)}(u, v).$$

The conditional expectation $E_\theta(\ell_\theta(\theta, F)(T_1, T_2)|T_1 + \theta T_2 = u + \theta v)$ minimizes the distance to $\ell_\theta(\theta, F)(u, v)$ over all functions of $u + \theta v$. Since all score functions for F are functions of $u + \theta v$, and this function is a score function for some submodel, it must be the closest F-score. This, in addition to the preceding equations, implies that the path $t \to F_t(\theta, F)$ indexes a least favorable submodel for θ.

The maximum likelihood estimators $\hat{\theta}$, $\hat{F}$ and $\hat{F}_0$ can be shown to be consistent (for the Euclidean topology and the weak topology) by the method of Wald [cf. Kiefer and Wolfowitz (1956)]. The asymptotic efficiency (3.5) of $\hat{\theta}$ is shown by van der Vaart (1996) [with $\tilde{\ell} = \tilde{\ell}_\theta(\theta_0, F_0)$].

Recalling (3.7) and (3.13), we see that

$$\dot{\ell}(u, v; t, \theta, F) = \frac{\int \dot{\ell}(u, v; t, \theta, F|z)\, z^2 \exp[-z(3\theta - t)(2\theta)^{-1}(u + tv)]\, dF(z)}{\int z^2 \exp[-z(3\theta - t)(2\theta)^{-1}(u + tv)]\, dF(z)},$$

where

$$\dot{\ell}(u, v; t, \theta, F|z) = (\theta - t)\big(3t^{-1}(3\theta - t)^{-1} + zv\theta^{-1}\big) + z(2\theta)^{-1}(u - \theta v).$$

Lemma 7.1 shows that these functions belong to a Donsker class for (t, θ, F) varying over a sufficiently small neighborhood of $(\theta_0, \theta_0, F_0)$. Furthermore, for $(t, \theta, F) \to (\theta_0, \theta_0, F_0)$ these functions converge for every (u, v) to $\tilde{\ell}_\theta(\theta_0, F_0)(u, v)$. The function $\dot{\ell}(u, v; \theta_0, \theta_0, \hat{F}_0)$ equals the efficient score function $\tilde{\ell}_\theta(\theta_0, \hat{F}_0)(u, v)$. From the representation of the efficient score function as a conditional score, we see that $P_{\theta_0 F_0}\tilde{\ell}_\theta(\theta_0, F) = 0$ for every θ_0, F_0 and F; this verifies (3.15). In view of Lemma 3.2 all conditions of Theorem 3.1 are satisfied, except possibly (3.14), which takes the following form: as $(t, \theta, F) \to (\theta_0, \theta_0, F_0)$,

$$\mathbb{P}_n \frac{\partial^2 p_{t, F_t(\theta, F)}/\partial t^2}{p_{t, F_t(\theta, F)}} \to_{\mathrm{P}} 0. \tag{7.1}$$

The second partial derivative in this expression can be written in the form

$$\int R(u, v; t, \theta|z)\, z^2 \exp[-z(3\theta - t)(2\theta)^{-1}(u + tv)]\, dF(z),$$

for

$$R(u, v; t, \theta|z) = \frac{3t-6\theta}{2\theta^2} + z\biggl(\frac{3(t-\theta)}{4\theta^2}(u+tv)(t-3\theta) + \frac{(t-3\theta)^2}{8\theta^3}(5t-3\theta)v + 3\frac{(t-\theta)}{8\theta^3}(t-3\theta)(u-v(3\theta-2t))\biggr) + z^2 t\frac{(t-3\theta)^2}{16\theta^4}\bigl(u-v(3\theta-2t)\bigr)^2.$$

By the lemma below the functions in (7.1) are contained in a P_0–Glivenko–Cantelli class. This implies that it suffices to prove (7.1) with the empirical measure replaced by the true measure P_0. If $(t, \theta, F) \to (\theta_0, \theta_0, F_0)$, then the functions in (7.1) converge for every (u, v) to the function

$$\frac{\int[-3(2\theta_0)^{-1} + zv + z^2(u-\theta_0 v)^2(4\theta_0)^{-1}]\, z^2 \exp[-z(u+\theta_0 v)]\, dF_0(z)}{\int z^2 \exp[-z(u+\theta_0 v)]\, dF_0(z)}.$$

This function has mean zero under P_0. An application of the dominated convergence theorem concludes the proof.

LEMMA 7.1. *Suppose that $\int(z^2 + z^{-6.5})\, dF_0 < \infty$. Then there exists a neighborhood $\mathscr{V}$ of F_0 for the weak topology such that the class of functions*

$$(u, v) \mapsto \frac{\int(a_1 + a_2 zu + a_3 zv + a_4 z^2u^2 + a_5 z^2uv + a_6 z^2v^2)\, z^2\, e^{-z(b_1u+b_2v)}\, dF(z)}{\int z^2\, e^{-z(b_1u+b_2v)}\, dF(z)},$$

where $(a_1, \dots, a_5)$ ranges over a bounded subset of $\mathbb{R}^5$, (b_1, b_2) ranges over a compact subset of $(0, \infty)^2$ and F ranges over $\mathscr{V}$, is P_{θ_0, F_0}-Donsker with square-integrable envelope.

PROOF. By applying Lemma L.23 of Pfanzagl (1990) repeatedly, it follows that there exist constants and a weak neighborhood $\mathscr{V}$ of F_0 such that

$$\sup_{F\in\mathscr{V}} \frac{\int z^{m+2}e^{-zs}\, dF(z)}{\int z^2 e^{-zs}\, dF(z)} \le \begin{cases} C_m s^{-m}|\log s|^m, & \text{if } s \le \frac{1}{2}, \\ C_m, & \text{if } s \ge \frac{1}{2}. \end{cases} \tag{7.2}$$

Since a symmetric convex hull of Donsker classes is Donsker [see, e.g., van der Vaart and Wellner (1996), Example 2.10.7 and Theorem 2.10.3], it suffices to prove for all nonnegative integers k and l with $k + l \le 2$ that the class of functions

$$u^k v^l \frac{\int z^{k+l+2} \exp[-z(b_1u+b_2v)]\, dF(z)}{\int z^2 \exp[-z(b_1u+b_2v)]\, dF(z)},$$

with F ranging over $\mathscr{V}$ and $b = (b_1, b_2)$ over a compact subset of $(0, \infty)^2$, is Donsker. For fixed b let $\mathscr{F}_b$ be the class of these functions with only F varying.

Since b is bounded away from zero, the function

$$\frac{u^k v^l}{(b_1u + b_2v)^{k+l}}$$

is uniformly bounded in u, v and b. The functions in $\mathscr{F}_b$ are the product of this function and the function $h_F(b_1u + b_2v)$ for

$$h_F(s) = s^m \frac{\int z^{m+2} e^{-zs}\, dF(z)}{\int z^2 e^{-zs}\, dF(z)}, \qquad m = k + l.$$

Let Q_b be the distribution of $b_1T_1 + b_2T_2$ under P_0. Then it is not hard to see that the bracketing numbers of $\mathscr{F}_b$ satisfy, for some constant C,

$$N_{[]}\bigl(\varepsilon, \mathscr{F}_b, L_2(P_0)\bigr) \le N_{[]}\bigl(C\varepsilon, \{h_F\colon F \in \mathscr{V}\}, L_2(Q_b)\bigr). \tag{7.3}$$

For a definition of bracketing numbers see, for example, van der Vaart and Wellner [(1996), Definition 2.1.6] or Ossiander (1987). We shall bound the right-hand side by application of Theorem 2.1 in van der Vaart (1994b). Let $\lesssim$ denote less than or equal up to a multiplicative constant. In view of (7.2) we have, for every $1/2 < \alpha < 1$,

$$\begin{aligned} \bigl|h_F(s)\bigr| &\lesssim |\log s|^m, \qquad 0 < s < 1/2,\\ \bigl|h_F(s_1) - h_F(s_2)\bigr| &\lesssim |s_1 - s_2|^\alpha \sup_{s_1 \le s \le s_2} \bigl|h'_F(s)\bigr|^\alpha |\log s_1|^{m(1-\alpha)}\\ &\lesssim |s_1 - s_2|^\alpha\, s_1^{-\alpha} |\log s_1|^{m+\alpha}, \qquad 0 < s_1 < s_2 \le 1/2. \end{aligned}$$

Thus the restrictions of the functions h_F to an interval $[a, b] \subset (0, 1/2]$ belong to the space $C^\alpha_M[a, b]$ for M a multiple of $a^{-\alpha}|\log a|^{m+\alpha}$. Similarly, again in view of (7.2), we have

$$\begin{aligned} \bigl|h_F(s)\bigr| &\lesssim s^m, \qquad s \ge 1/2,\\ \bigl|h_F(s_1) - h_F(s_2)\bigr| &\lesssim |s_1 - s_2|^\alpha s_2^m, \qquad 1/2 \ge s_1 < s_2. \end{aligned}$$

Thus the restrictions of the functions h_F to an interval $[a, b] \subset [1/2, \infty]$ belong to the space $C^\alpha_M[a, b]$ for M a multiple of b^m. Theorem 2.1 of van der Vaart (1994b) applied with the partition $(0, \infty) = \bigcup_i (2^{-i}, 2^{-i+1}] \bigcup_i (i, i+1]$ shows that, for every $V \ge 1/\alpha$,

$$\log N_{[]}\bigl(\varepsilon, \{h_F\colon F \in \mathscr{V}\}, L_2(Q_b)\bigr) \le K K_b^{(V+2)/2} \left(\frac{1}{\varepsilon}\right)^V, \tag{7.4}$$

for a constant K depending only on α, V and K_b defined by

$$K_b = \sum_i \bigl[|\log 2^{-i}|^{2(m+\alpha)} 2^{2i\alpha} Q_b(2^{-i}, 2^{-i+1}]\bigr]^{V/(V+2)} + \sum_i \bigl[i^{2m} Q_b(i, i+1]\bigr]^{V/(V+2)}.$$

By a straightforward calculation we see that Q_b has a Lebesgue density which is bounded above by $\theta_0 s(b_1b_2)^{-1} F_0 z^2$. Thus $Q_b(2^{-i}, 2^{-i+1}]$ is bounded above by a multiple of 2^{-2i}. Furthermore $Q_b(i, i+1] \le Q_b(i, \infty)$ is bounded above by $i^{-l} F_0 z^{-l}$ for any $l \ge 0$. Insert these upper bounds into the definition of K_b to obtain that

$$K_b \lesssim \sum_i \bigl[i^{2(m+\alpha)} 2^{-2i(1-\alpha)}\bigr]^{V/(V+2)} + \sum_i \bigl[i^{2m-l}\bigr]^{V/(V+2)},$$

provided $F_0(z^{-l} + z^2) < \infty$. Here the multiplicative constant depends on F_0, but is uniform in b ranging over our compact region.

For V sufficiently close to 2 this series converges for $l > 2m + 2$. Thus we have proved the existence of $V < 2$ and constants K_b that are uniformly bounded such that (7.4) holds. In view of (7.3) the same estimate is valid for the bracketing numbers of $\mathscr{F}_b$ in $L_2(P_0)$.

Writing the elements of $\mathscr{F}_b$ in the form $f_{F,b}$ we see, using (7.2),

$$\begin{aligned}|f_{F,b} - f_{F,b'}| &\lesssim u^k v^l \biggl[\sup_b \biggl(\frac{|\log(b_1 u + b_2 v)|}{|b_1 u + b_2 v|}\biggr)^{m+1} + 1\biggr](u+v)\|b - b'\| \\ &\lesssim \bigl[|\log(u+v)|^{m+1} + (u+v)u^k v^l\bigr]\,\|b - b'\|.\end{aligned}$$

The function in square brackets, say G, on the far right is square integrable under P_0. We can now form brackets over the class of functions of interest $\bigcup_b \mathscr{F}_b$ by first choosing an ε-net over the set of all b. The number of points in this net can be chosen smaller than $(K/\varepsilon)^2$ for some constant K. For every b_i in the ε-net take a minimal number of brackets $[l_{i,j}, u_{i,j}]$ over $\mathscr{F}_{b_i}$. Then the brackets $[l_{i,j} - \varepsilon G, u_{i,j} + \varepsilon G]$ cover $\mathscr{F} = \cup_b \mathscr{F}_b$ and have size bounded by $\varepsilon(1+2\|G\|_{P_0,2})$. The logarithm of the total number of brackets obtained in this way is bounded by

$$2\log(K/\varepsilon) + \sup_b \log N_{[]}\bigl(\varepsilon, \mathscr{F}_b, L_2(P_0)\bigr) \lesssim \left(\frac{1}{\varepsilon}\right)^V.$$

This is an upper bound for $N_{[]}(\varepsilon', \mathscr{F}, L_2(P_0))$, where $\varepsilon' = \varepsilon(1 + 2\|G\|_{P_0,2})$. Apply the theorem of Ossiander (1987) to conclude that $\mathscr{F}$ is Donsker. □

APPENDIX

A.1. Convexity of the confidence set. In general, a confidence set obtained by inverting the likelihood ratio test is not guaranteed to have a "nice" shape. In simple cases it can be seen to be convex.

LEMMA A.1. *Suppose that $\psi \to \mathrm{lik}(\psi, x)$ is a concave function on a convex subset Ψ of a linear space. Furthermore suppose that for all ψ_1, ψ_2 in Ψ the map $\varepsilon \to \theta(\varepsilon\psi_1 + (1-\varepsilon)\psi_2)$ is continuous on* $[0, 1]$. *Then the confidence region $\{\theta\colon \mathrm{lrt}_n(\theta) \le z^2_{\alpha/2}\}$ is convex.*

PROOF. Define a set A by $A = \{\psi\colon \mathbb{P}_n \ln \mathrm{lik}(\psi) \ge z^2_{\alpha/2} + \mathbb{P}_n \ln \mathrm{lik}(\hat\psi)\}$. Since the map $\psi \to \mathrm{lik}(\psi)$ is concave, the set A is convex. The confidence region is composed of those θ for which there exists a $\psi \in A$ for which $\theta(\psi) = \theta$. If θ_1 and θ_2 are in the confidence region, then there exist ψ_1 and ψ_2 both in A for which $\theta(\psi_1) = \theta_1$ and $\theta(\psi_2) = \theta_2$. Let $\varepsilon' \in (0, 1)$. Consider $\theta(\varepsilon\psi_1 + (1-\varepsilon)\psi_2)$, which for $\varepsilon = 0$ is equal to θ_2 and for $\varepsilon = 1$ is equal to θ_1. The continuity in ε implies that there exists an ε^* for which $\theta(\varepsilon^*\psi_1 + (1-\varepsilon^*)\psi_2) = \varepsilon'\theta_1 + (1-\varepsilon')\theta_2$. Since $\varepsilon^*\psi_1 + (1-\varepsilon^*)\psi_2$ is in A, the convex combination $\varepsilon'\theta_1 + (1-\varepsilon')\theta_2$ is in the confidence region. □

A.2. Double censoring: technical complements. In the following two lemmas, BV$[\sigma, \tau]$ denotes the set of functions $h\colon [\sigma, \tau] \to \mathbb{R}$ that are bounded and of bounded variation, equipped with the norm $||\cdot||_{\mathrm{BV}} = \|\cdot\|_\infty + \|\cdot\|_{\mathrm{var}}$.

LEMMA A.2. *Suppose that $(G_L - G_R)(s-)$ is bounded away from zero for s in $[\sigma, \tau]$ containing the support of F and G_L is continuous at σ:*

(i) *Then the operator $\ell^*\ell_F$:* BV$[\sigma, \tau] \to$ BV$[\sigma, \tau]$ *is one-to-one, onto and continuously invertible.*

(ii) *If F_n are distribution functions, with support contained in $[\sigma, \tau]$ such that $\|F_n - F\|_\infty \to 0$, then $\ell^*\ell_{F_n} \to \ell^*\ell_F$ in operator norm, that is, $\ell^*\ell_{F_n} h \to \ell^*\ell_F h$ in bounded variation norm on $[\sigma, \tau]$, uniformly in uniformly bounded functions h of uniformly bounded variation.*

PROOF. (i) The operator $\ell^*\ell_F$ can be written as the sum $K_1 + A + K_2$ of three operators. The operator $Ah = h(G_L - G_R)$ is continuously invertible [with inverse $A^{-1}h = h/(G_L - G_R)$] under the condition that $G_L - G_R$ is bounded away from zero. Since we can write $\ell^*\ell_F$ in the form $A(A^{-1}K_1 + I + A^{-1}K_2)$ it now suffices by Theorem 4.25 in Rudin (1973) to show that $A^{-1}K_1 + A^{-1}K_2$ is compact and that $K_1 + A + K_1$ is one-to-one. The first is true if both K_1 and K_2 are compact.

Consider K_1. Given a uniformly bounded sequence h_n of monotone functions, the functions

$$g_n = \frac{\int_{[\sigma,u]} h_n \, dF}{F(u)}$$

are a uniformly bounded sequence of monotone functions. By Helly's theorem there exists a subsequence and a monotone function g such that $g_{n'}(u) \to g(u)$ for every u. Then by dominated convergence $\int |g_{n'} - g| \, dG_L \to 0$. This implies that

$$\left\| K_1 h_{n'} - \int_{[\cdot,\,\infty)} g \, dG_L \right\|_\infty \le \sup_u \int_{[u,\,\infty)} |g_{n'} - g| \, dG_L \to 0,$$

$$\begin{aligned}
&\left\| K_1 h_{n'} - \int_{[\cdot,\,\infty)} g \, dG_L \right\|_{\mathrm{var}} \\
&\qquad = \sup \sum \left| \int_{[t_{i+1},\,\infty)} (g_{n'} - g) \, dG_L - \int_{[t_i,\,\infty]} (g_{n'} - g) \, dG_L \right| \\
&\qquad \le \int |g_{n'} - g| \, dG_L \to 0.
\end{aligned}$$

It follows that any given sequence h_n as before has a subsequence along which $K_1 h_{n'}$ converges in BV$[\sigma, \tau]$. Since any uniformly bounded sequence h_n of uniformly bounded variation can be written as the difference of two sequences of this type, it follows that K_1 is compact.

The operator K_2 can be shown to be compact in the same manner.

To see that $K_1 + A + K_2$ is one-to-one, note first that both $A^{-1/2}K_1A^{-1/2}$ and $A^{-1/2}K_2A^{-1/2}$ are nonnegative definite operators on $L_2(F)$ in the Hilbert space sense. Thus the spectrum of these operators is nonnegative and it follows that the spectrum of $A^{-1/2}K_1A^{-1/2} + I + A^{-1/2}K_2A^{-1/2}$ is contained in the interval $[1, \infty)$. Thus this operator is invertible as an operator on $L_2(F)$. The same is true for $K_1 + A + K_2$, so that the equation $(K_1 + A + K_2)h = 0$ (everywhere) for a bounded variation function h, implies that $h = 0$ almost surely under F. Substitute in the definitions of K_1 and K_2 to see that $Ah = 0$, whence $h = 0$ everywhere.

(ii) Since $\|F_n - F\|_\infty \to 0$, we have that $F_nh - Fh \to 0$ uniformly in uniformly bounded functions h of uniformly bounded variation. Thus for any sequence h_n of such functions we have that $\ell_{F_n}h_n - \ell_F h_n \to 0$ pointwise in (u, d). By dominated convergence and the continuity of G_L at σ,

$$\int_{[\sigma,\tau]} \left|\ell_{F_n}h_n(\cdot, 1) - \ell_F h_n(\cdot, 1)\right| dG_L \to 0,$$

$$\int_{[\sigma,\tau)} \left|\ell_{F_n}h_n(\cdot, 3) - \ell_F h_n(\cdot, 3)\right| dG_R \to 0.$$

As in the proof of (i) this implies that $\ell^*\ell_{F_n}h_n - \ell^*\ell_F h_n \to 0$ in bounded variation norm. □

LEMMA A.3. *Assume that, for all u such that $F_0(u) > 0$ and $F_0(u-) < 1$,*

$$(G_L - G_R)(u-) > 0, \tag{A.1}$$

and suppose that g is left continuous, bounded and of bounded variation, and is not identically zero almost surely under F_0. Furthermore, assume that F_0, G_L and G_R are continuous.

(i) *Then $\|\hat{F}_0 - F_0\|_\infty \to 0$ almost surely under P_0.*

Under the more restrictive assumptions of Theorem 2.1, the following holds:

(ii) *As $n \to \infty$ the support of $\hat{F}_0$ belongs to $[\sigma, \tau]$ almost surely under P_0 and*

$$(\hat{F} - \hat{F}_0)h = \frac{F_0(h(g^* - F_0g^*))(\hat{\theta} - \theta_0)}{F_0(g(g^* - F_0g^*))} + o_P(n^{-1/2}),$$

uniformly in $h \in \mathrm{BV}[\sigma, \tau]$ of norm less than or equal to 1.

PROOF. (i) To keep the proof of (i) simple, let $\theta_0 = 0$. The estimator of F under the constraint, $Fg = 0$, is an $\hat{F}_0$ which maximizes $\mathbb{P}_n[\ln p_F(X)]$. There exists a random variable $\hat{\lambda}$ for which

$$\mathbb{P}_n[\ell_{\hat{F}_0}h] - \hat{F}_0h = \hat{\lambda}\int hg\, d\hat{F}_0 \tag{A.2}$$

for all bounded functions h. Note that the form of the likelihood implies that a maximum likelihood estimator $\hat{F}_0$ must have positive jumps at the observed event times.

Since F_0 is continuous, to verify (i), we need only show that $\hat{F}_0(t)$ converges to $F_0(t)$ for all t a.s. Define the processes

$$\begin{aligned}
Q_n^1(t) &= \mathbb{P}_n[1\{d=1\}1\{u \le t\}], & Q^1(u) &= \int_0^u F_0\,dG_L,\\
Q_n^2(t) &= \mathbb{P}_n[1\{d=2\}1\{u \le t\}], & Q^2(u) &= \int_0^u (G_L - G_R)(v-)\,dF_0(v),\\
Q_n^3(t) &= \mathbb{P}_n[1\{d=3\}1\{u \le t\}], & Q^3(u) &= \int_0^u (1-F_0)\,dG_R.
\end{aligned}$$

By the Glivenko–Cantelli theorem each of the sequences Q_n^i converges uniformly to the corresponding function Q^i, almost surely. Fix a sample point within the set of probability 1 on which $\|Q_n^2 - Q^2\|_\infty \to 0$, $\|Q_n^3 - Q^3\|_\infty \to 0$ and $\|Q_n^1 - Q^1\|_\infty \to 0$.

We first prove that all limit points of $\hat{\lambda}$ are finite. Since $F_0 g = 0$ yet g is not a.s. F_0 zero, there exist $\varepsilon > 0$ such that the sets $A = \{g > \varepsilon\}$ and $B = \{g < -\varepsilon\}$ have positive measure. For large n there exists at least one observed event time in the set A [by (A.1)], say x_n. Put $h(u) = 1\{u = t\}$ in (A.2) to get

$$\Delta\hat{F}_0(t)\left(1 - \int_{t-}^{\infty} 1/\hat{F}_0\,dQ_n^1 - \int_{-\infty}^{t-} 1/(1-\hat{F}_0)\,dQ_n^3 + \hat{\lambda} g(t)\right) = \Delta Q_n^2(t). \tag{A.3}$$

Evaluate (A.3) at $t = x_n$. Since $\Delta\hat{F}_0(x_n) > 0$ and $\Delta Q_n^2(x_n) > 0$, we have that $\hat{\lambda}$ cannot diverge to minus infinity. Likewise for large n there exists at least one observed event time in the set B, say y_n. Since $\Delta\hat{F}_0(y_n) > 0$ and $\Delta Q_n^2(y_n) > 0$, we have that $\hat{\lambda}$ cannot diverge to infinity.

Let $h(s)$ be $1\{s \le t\}$ for arbitrary nonnegative t. Write $\hat{F}_0(hg)$ as $g^+(t)\hat{F}_0(t) - \int_{-\infty}^t \hat{F}_0(u)\,dg^+(u)$, where $g^+(u) = g(u+)$ so that g^+ is right continuous. From (A.2) we get

$$\begin{aligned}
&\hat{F}_0(t) + \hat{\lambda}\left(g^+(t)\hat{F}_0(t) - \int_{-\infty}^t \hat{F}_0 dg^+\right)\\
&\quad = \sum_{i=1}^3 Q_n^i(t) + \hat{F}_0(t)\int_t^\infty 1/\hat{F}_0\,dQ_n^1 - (1-\hat{F}_0(t))\int_{-\infty}^t 1/(1-\hat{F}_0)\,dQ_n^3.
\end{aligned} \tag{A.4}$$

Use Helly's selection theorem to get a subsequence of n for which $\hat{F}_0$ converges pointwise, say to $\tilde{F}$, and $\hat{\lambda}$ converges to a finite value, say λ. From (A.3) we see that Q_n^2 is absolutely continuous with respect to $\hat{F}_0$. This along with assumption (A.1) implies that the convex hull of the support of $\tilde{F}$ contains the support of F_0. This allows us to use the dominated convergence theorem along with the uniform convergence of the Q_n^i's, the fact that both $F(t)/F(u)$ as a function of $u > t$ and $(1 - F(t))/(1 - F(u))$ as a function of $u \le t$ have total variation bounded above by 1, to show that

$$\tilde F(t) + \lambda\biggl(g^+(t)\tilde F(t) - \int_{-\infty}^t \tilde F\, dg^+\biggr) = \sum_{i=1}^3 Q^i(t) + \tilde F(t)\int_t^\infty 1/\tilde F\, dQ^1 \\ - (1 - \tilde F(t))\int_{-\infty}^t 1/(1 - \tilde F)\, dQ^3.$$

We would like to write most of the terms in the above equality as integrals with respect to $\tilde F$, but $\tilde F$, although nondecreasing, may not be right continuous. Instead we consider $\tilde F^+$, which is the right-hand limit of $\tilde F$. In the following we derive an equation similar to the above but for $\tilde F^+$.

Since the total mass of $\tilde F$ is bounded above by 1, $\tilde F$ has at most a countable number of discontinuities and therefore $\tilde F^+$ differs from $\tilde F$ at at most a countable number of points. Evaluate the above equality at $t + h$ and let h decrease to zero to get

$$\begin{aligned}&\tilde F^+(t) + \lambda\biggl(g^+(t)\tilde F^+(t) - \int_{-\infty}^t \tilde F\, dg^+\biggr)\\ &\quad = \sum_{i=1}^3 Q^i(t) + \tilde F^+(t)\int_t^\infty 1/\tilde F^+\, dQ^1 - (1 - \tilde F^+(t))\int_{-\infty}^t 1/(1 - \tilde F^+)\, dQ^3.\end{aligned} \tag{A.5}$$

Now $\Delta\tilde F^+(t)$ satisfies

$$\Delta\tilde F^+(t)(1 + \lambda g(t)) + \lambda(\tilde F^+(t) - \tilde F(t))\,\Delta g^+(t) \\ = \Delta\tilde F^+(t)\int_t^\infty 1/\tilde F^+ dQ^1 + \Delta\tilde F^+(t)\int_{-\infty}^t 1/(1 - \tilde F^+) dQ^3.$$

Suppose that, at t, $\tilde F^+(t) - \tilde F(t) > 0$. Then subtracting the equation in $\tilde F$ from the equation in $\tilde F^+$ we get

$$1 + \lambda g^+(t) = \int_t^\infty 1/\tilde F^+\, dQ^1 + \int_{-\infty}^t 1/(1 - \tilde F^+)\, dQ^3. \tag{A.6}$$

Combine the last two equations to get that if $\tilde F^+(t) - \tilde F(t) > 0$, then

$$\Delta g^+(t)\lambda\bigl(-\Delta\tilde F^+(t) + \tilde F^+(t) - \tilde F(t)\bigr) = 0.$$

Let J be the countable set of points t for which $\tilde F^+(t) - \tilde F(t) > 0$. Then, for arbitrary t, $\lambda\int_{-\infty}^t \tilde F\, dg^+ = \lambda\int_{-\infty}^t \tilde F^+\, dg^+ - \lambda\int_{-\infty}^t (\tilde F^+ - \tilde F)1_J\, dg^+$, which by the above argument is equal to $\lambda\int_{-\infty}^t \tilde F^+\, dg^+ - \int_{-\infty}^t \lambda\Delta g^+(u)1_J(u)\, d\tilde F^+(u)$. Equation (A.5) becomes

$$\tilde F^+(t) + \int_{-\infty}^t \bigl(\lambda g(u) + \lambda\,\Delta g^+(u)1_J(u)\bigr)\, d\tilde F^+(u) + \lambda\tilde F^+(-\infty)g^+(-\infty) \\ = \sum_{i=1}^3 Q^i(t) + \tilde F^+(t)\int_t^\infty 1/\tilde F^+\, dQ^1 - (1 - \tilde F^+(t))\int_{-\infty}^t 1/(1 - \tilde F^+)\, dQ^3.$$

For $h(s) = 1\{s \le t\}$, this implies

$$\begin{aligned}&\int h(1+\lambda g + \lambda\,\Delta g^{+} 1_J)\,d\tilde{F}^{+}\\ &\qquad = \int \frac{\int_0^u h\,d\tilde{F}^{+}}{\tilde{F}^{+}(u)}\,dQ^1(u) + \int h(u)(G_L - G_R)(u-)\,dF_0(u)\\ &\qquad\quad + \int \frac{\int_u^\infty h\,d\tilde{F}^{+}}{1-\tilde{F}^{+}(u)}\,dQ^3(u).\end{aligned} \tag{A.7}$$

Indeed the above holds for all bounded h.

Use integration by parts to rewrite (A.7) as

$$\int h(u)(G_L - G_R)(u-)\,dF_0(u) = \int h(u)A(u)\,d\tilde{F}^{+}(u), \tag{A.8}$$

where

$$A(u) = \lambda g(u) + \lambda\,\Delta g^{+}(u)1_J(u) + 1 - \int_{u-}^{\infty} 1/\tilde{F}^{+}\,dQ^1 - \int_0^{u-} 1/(1-\tilde{F}^{+})\,dQ^3,$$

and h is bounded. Note that (A.1) and the above implies that F_0 is absolutely continuous with respect to $\tilde{F}^{+}$. Put h equal to the indicator of the set $\{u\colon A(u) < 0\}$ to see that this set has $\tilde{F}^{+}$ mass zero. Next use approximating simple functions and the monotone convergence theorem to show that (A.8) holds for all nonnegative h. For arbitrary t, put $h(u) = 1\{u \le t\}(G_L - G_R)(u-)^{-1}$ so that $F_0(t) = \int_{-\infty}^{t} A(u)/(G_L - G_R)(u-)\,d\tilde{F}^{+}(u)$ for all $t \ge 0$. We see that a version of the Radon–Nikodym derivative is

$$(dF_0/d\tilde{F}^{+})(u) = A(u)/(G_L - G_R)(u-).$$

Since g is a bounded function, we may put $h(u) = g(u)/(G_L - G_R)(u-)$ to yield $0 = \int g(u)\,dF_0(u) = \int g(u)A(u)/(G_L - G_R)(u-)\,d\tilde{F}^{+}(u)$. In the following we will need that (A.7) holds for $h = (dF_0/d\tilde{F}^{+})$; however, $(dF_0/d\tilde{F}^{+})$ is not necessarily bounded. Now the left-hand side of (A.7) evaluated at $h = (dF_0/d\tilde{F}^{+})$ is finite and is equal to 1. Therefore by approximating $(dF_0/d\tilde{F}^{+})$ by simple functions and using the monotone convergence theorem we get that $P_{F_0}[l_{\tilde{F}^{+}}h^*] = 1$ or, equivalently, $\ln(P_{F_0}[l_{\tilde{F}^{+}}h^*]) = 0$ for $h^* = dF_0/d\tilde{F}^{+}$. Since the natural logarithm is strictly concave, $P_{F_0}[\ln \ell_{\tilde{F}^{+}}h^*] \le 0$. However, $l_{\tilde{F}^{+}}h^* = p_{F_0}/p_{\tilde{F}^{+}}$. However, $P_{F_0}[\ln(p_{F_0}/p_{\tilde{F}^{+}}] \ge 0$ with equality if and only if $p_{F_0}/p_{\tilde{F}^{+}}$ is equal to 1 with P_{F_0} probability 1. So $p_{F_0}/p_{\tilde{F}^{+}}$ is 1 with P_{F_0} probability 1. This means that $\int 1\{\tilde{F}^{+}(u) \ne F_0(u)\}F_0(u)\,dG_L(u) = 0$, and $\int 1\{\tilde{F}^{+}(u) \ne F_0(u)\}(1-F_0(u))\,dG_R(u) = 0$. So, $1-\int_{u-}^{\infty} 1/\tilde{F}^{+}\,dQ^1 - \int_0^{u-} 1/(1-\tilde{F}^{+})\,dQ^3 = (G_L - G_R)(u-)$ for all u. Since $\int 1\{(dF_0/d\tilde{F}^{+})(u) \ne 1\}(G_L - G_R)(u-)\,dF_0(u) = 0$, we get $\int 1\{\lambda g(u) \ne 0\}(G_L - G_R)(u-)\,dF_0(u) = 0$. Then $\lambda = 0$ and $(dF_0/d\tilde{F}^{+})$ is identically 1. So $\tilde{F}^{+}$ is identically equal to F_0. In this case (A.6) yields a contradiction, implying that $\tilde{F}$ is right continuous and identically equal to F_0.

(ii) Finally we prove (ii). Assumption (A.1) made on the interval $[\sigma, \tau]$ implies that for large n the largest U will not have $D = 3$ and the smallest U will not have $D = 1$. Recall that the support points of $\hat{F}_0$ are restricted to the U's. This, along with the additional assumptions that $G_L(\tau) = 1$ and $G_R(\sigma-) = 0$, implies that the convex hull of the support of $\hat{F}_0$ is at most $[\sigma, \tau]$ for large n, P_{F_0}-almost surely.

For a given function $h \in \mathrm{BV}[\sigma, \tau]$ define

$$\hat{h}^* = (\ell^* \ell_{\hat{F}_0})^{-1} h, \qquad \hat{g}^* = (\ell^* \ell_{\hat{F}_0})^{-1} g.$$

Lemma A.2 shows that these functions are well defined and uniformly bounded and of bounded variation. For t close to zero, define the submodel

$$dF_t = d\hat{F}_0\biggl(1 + t\biggl[(\hat{h}^* - \hat{F}_0\hat{h}^*) - \frac{\hat{F}_0(\hat{h}^* g)}{\hat{F}_0(\hat{g}^* g)}(\hat{g}^* - \hat{F}_0\hat{g}^*)\biggr]\biggr).$$

This satisfies the null hypothesis and differentiation with respect to t at zero yields the equation

$$\mathbb{P}_n \ell_{\hat{F}_0}\biggl((\hat{h}^* - \hat{F}_0\hat{h}^*) - \frac{\hat{F}_0(\hat{h}^* g)}{\hat{F}_0(\hat{g}^* g)}(\hat{g}^* - \hat{F}_0\hat{g}^*)\biggr) = 0.$$

Because the information operator preserves expectations we have $\hat{F}_0\hat{h}^* = \hat{F}_0 h$ for every h. So the preceding display can be rewritten as

$$\hat{F}_0 h = \mathbb{P}_n \ell_{\hat{F}_0}\hat{h}^* - \frac{\hat{F}_0(\hat{h}^* g)}{\hat{F}_0(\hat{g}^* g)}\mathbb{P}_n \ell_{\hat{F}_0}\hat{g}^*.$$

Combine this with

$$F_0 h = F_0 \ell^* \ell_{\hat{F}_0}\hat{h}^* = P_{F_0}\ell_{\hat{F}_0}\hat{h}^*$$

and

$$0 = F_0 g = P_{F_0}\ell_{\hat{F}_0}\hat{g}^*$$

to find that

$$(\hat{F}_0 - F_0)h = (\mathbb{P}_n - P_{F_0})\biggl(\ell_{\hat{F}_0}\hat{h}^* - \frac{\hat{F}_0(\hat{h}^* g)}{\hat{F}_0(\hat{g}^* g)}\ell_{\hat{F}_0}\hat{g}^*\biggr).$$

The functions on the right-hand side are in a Donsker class by Lemma A.2(ii). Given consistency, asymptotic normality follows. Asymptotic equicontinuity of the sequence of empirical processes implies that, for $h^* = (\ell^* \ell_{F_0})^{-1} h$,

$$(\hat{F}_0 - F_0)h = (\mathbb{P}_n - P_{F_0})\biggl(\ell_{F_0} h^* - \frac{F_0(h^* g)}{F_0(g^* g)}\ell_{F_0} g^*\biggr) + o_P(n^{-1/2})$$

uniformly for $h \in \mathrm{BV}[\sigma, \tau]$ of norm less than or equal to 1.

In a similar fashion, using the submodel

$$dF_t = d\hat{F}\bigl(1 + t\bigl[(\hat{h}^{**} - \hat{F}\hat{h}^{**})\bigr]\bigr),$$

where $h^{**} = (\ell^*\ell_{\hat{F}})^{-1}h$, we derive $(\hat{F} - F_0)h = (\mathbb{P}_n - P_{F_0})\ell_{F_0}h^* + o_P(n^{-1/2})$ uniformly for $h \in \mathrm{BV}[\sigma, \tau]$ of norm less than or equal to 1. Combining this result with the corresponding result for $\hat{F}_0$, we have

$$(\hat{F} - \hat{F}_0)h = (\mathbb{P}_n - P_{F_0})\left(\frac{F_0(h^*g)}{F_0(g^*g)}\ell_{F_0}g^*\right) + o_P(n^{-1/2})$$

uniformly for $h \in \mathrm{BV}[\sigma, \tau]$ of norm less than or equal to 1. □

A.3. Current status data: technical complements. In this section, we show that $\hat{\Lambda}_0$ and $\hat{\Lambda}$ have a rate of convergence $O_P(n^{-1/3})$ with respect to the L_2-norm on $[\sigma, \tau]$. Furthermore, we sketch the proof of the asymptotic normality of $\hat{\theta}$. These results complement the arguments of Huang (1996).

We shall derive the rate of convergence of the estimators $\hat{\Lambda}$ and $\hat{\Lambda}_0$ from the rate of convergence of the density estimators $p_{\hat{\theta}, \hat{\Lambda}}$ and $p_{\theta_0, \hat{\Lambda}_0}$ with respect to the Hellinger distance. Rates for maximum likelihood estimators in the Hellinger distance were expressed, in general, in the entropy of a model by Birgé and Massart (1993) and Wong and Shen (1995). See van der Vaart and Wellner [(1996), Section 3.4.1] for an exposition. We compute the relevant entropy in the following lemma [cf. Huang (1996), Theorem 3.3].

Recall that we take the parameter set Ψ for (θ, Λ) equal to the product of a compact subset Θ of $\mathbb{R}$ and the set of all nondecreasing, cadlag functions Λ: $[0, \tau] \to [0, M]$. Given the density $p_0 = p_{\theta_0, \Lambda_0}$, the relevant metric in our entropy calculation is h_0 given by

$$h_0^2(p_{\theta_1, \Lambda_1}, p_{\theta_2, \Lambda_2}) = \int\left(\sqrt{p_{\theta_1, \Lambda_1} + p_{\theta_0, \Lambda_0}} - \sqrt{p_{\theta_2, \Lambda_2} + p_{\theta_0, \Lambda_0}}\right)^2 d\mu.$$

This is the Hellinger distance between the densities $p_{\theta, \Lambda} + p_{\theta_0, \Lambda_0}$. [See van der Vaart and Wellner (1996), Theorem 3.4.4. The "ordinary" Hellinger distance can be used as well, and will lead to an upper bound on the rate of convergence, but may yield a suboptimal result. The addition of the term p_0 helps to keep the densities bounded away from zero.]

LEMMA A.4. *Under the conditions of Theorem* 2.2, *there exists a constant* C *such that*, *for every* $\varepsilon > 0$,

$$\log N_{[]}\big(\varepsilon, \{p_{\theta, \Lambda}, (\theta, \Lambda) \in \Psi\}, h_0\big) \le C\left(\frac{1}{\varepsilon}\right).$$

PROOF. First consider the class of densities for a fixed θ. We can write $p_{\theta, \Lambda} + p_0 = \delta\phi_1(\Lambda(y), z) + (1 - \delta)\phi_0(\Lambda(y), z)$ for functions ϕ_i that are monotone in their first argument. Thus a bracket $\Lambda_1 \le \Lambda \le \Lambda_2$ for Λ leads, by substitution, readily to a bracket for $p_{\theta,\Lambda} + p_0$. Since the partial derivatives $(\partial/\partial u)\sqrt{\phi_i(u, z)}$ are uniformly bounded in (u, z, θ) (note that p_0 is bounded away from zero), there exists a constant D such that

$$\int\big(\phi_i^{1/2}(\Lambda_1(y), z) - \phi_i^{1/2}(\Lambda_2(y), z)\big)^2\, dF^{Y,Z}(y, z) \le D\int_\sigma^\tau \big(\Lambda_1(y) - \Lambda_2(y)\big)^2\, dy.$$

Thus, brackets for Λ of L_2-size ε translate into brackets for $p_{\theta,\Lambda}+p_0$ of h_0-size proportional to ε. By, for instance, Theorem 2.7.5 of van der Vaart and Wellner (1996), we can cover the set of all Λ by $\exp C(1/\varepsilon)$ brackets of size ε.

Next, we allow θ to vary freely as well. Since θ is finite-dimensional and $(\partial/\partial\theta)\, p_{\theta,\Lambda}(x)$ is uniformly bounded in (θ,Λ,x), this increases the entropy only slightly (cf. the argument given at the end of the proof of Lemma 7.1). □

In view of the preceding lemma and Theorem 3.4.4 of van der Vaart and Wellner (1996) (see the conclusions after the theorem), the rates of convergence of $p_{\hat\theta,\hat\Lambda}$ and $p_{\theta_0,\hat\Lambda_0}$ in Hellinger distance to p_0 are at least $O_P(n^{-1/3})$. By the following lemma this result implies rates for $\hat\Lambda$ and $\hat\Lambda_0$ in the L_2-norm. (It also implies an upper bound for the rate of $\hat\theta$, but this is suboptimal.)

LEMMA A.5. *Under the conditions of Theorem* 2.2, *there exist constants* $C,\varepsilon>0$ *such that, for all* Λ *and all* $|\theta-\theta_0|<\varepsilon$,

$$\int\bigl[p_{\theta,\Lambda}^{1/2}-p_{\theta_0,\Lambda_0}^{1/2}\bigr]^2\,d\mu\ge C\int_\sigma^\tau(\Lambda-\Lambda_0)^2(y)\,dy+C|\theta-\theta_0|^2.$$

PROOF. The left-hand side of the lemma can be rewritten as

$$\int\frac{[p_{\theta,\Lambda}-p_{\theta_0,\Lambda_0}]^2}{[p_{\theta,\Lambda}^{1/2}+p_{\theta_0,\Lambda_0}^{1/2}]^2}\,d\mu.$$

Since p_{θ_0,Λ_0} is bounded away from zero, and the densities $p_{\theta,\Lambda}$ are uniformly bounded, both by assumption, the denominator can be bounded above and below by positive constants. Thus the Hellinger distance is equivalent to the L_2-distance between the densities. The latter is equal to

$$\int\bigl[\exp[-\exp(\theta z)\Lambda(y)]-\exp[-\exp(\theta_0 z)\Lambda_0(y)]\bigr]^2\,dF^{Y,Z}(y,z).$$

Let $g(t)$ be the function $\exp(-\exp(\theta z)\Lambda(y))$ evaluated at $\theta_t=t\theta+(1-t)\theta_0$ and $\Lambda_t=t\Lambda+(1-t)\Lambda_0$, for fixed (y,z). Then the integrand is equal to $\bigl(g(1)-g(0)\bigr)^2$, and hence, by the mean value theorem, there exists $0\le t=t(y,z)\le 1$ such that the preceding display is equal to

$$P_0\bigl(\exp[-\Lambda_t(y)\exp(\theta_t z)]\exp(\theta_t z)\bigl[(\Lambda-\Lambda_0)(y)\bigl(1+zt(\theta-\theta_0)\bigr)+(\theta-\theta_0)z\Lambda_0(y)\bigr]\bigr)^2.$$

Here the multiplicative factor $\exp[-\Lambda_t(y)\exp(\theta_t z)]\exp(\theta_t z)$ is bounded away from zero. By dropping this term we obtain, up to a constant, a lower bound for the left-hand side of the lemma. Next, since the function $Q(\cdot;\theta_0,\Lambda_0)$ is bounded away from zero and infinity, we may add a factor $Q^2(\cdot;\theta_0,\Lambda_0)$ and obtain the lower bound, up to a constant,

$$P_0\bigl(\ell_\Lambda(\theta_0,\Lambda_0)(\Lambda-\Lambda_0)(y)\bigl(1+zt(\theta-\theta_0)\bigr)+(\theta-\theta_0)\ell_\theta(\theta_0,\Lambda_0)\bigr)^2.$$

Here the function $h = (1 + zt(\theta - \theta_0))$ is uniformly close to 1 if θ is close to θ_0. Furthermore, for any function g,

$$\begin{aligned}\big(P_0\ell_\Lambda(\theta_0, \Lambda_0)g\ell_\theta(\theta_0, \Lambda_0)\big)^2 &= \big(P_0\ell_\Lambda(\theta_0, \Lambda_0)g(\ell_\theta(\theta_0, \Lambda_0) - \tilde{\ell})\big)^2\\ &\le P_0\big(\ell_\Lambda(\theta_0, \Lambda_0)g\big)^2(I_0 - \tilde{I}),\end{aligned}$$

by the Cauchy–Schwarz inequality. Since the efficient information $\tilde{I}$ is positive, the term $I_0 - \tilde{I}$ on the right-hand side can be written $I_0 c$ for a constant $0 < c < 1$. The lemma now follows by application of Lemma A.6. □

LEMMA A.6. *Let h, g_1 and g_2 be measurable functions such that $c_1 \le h \le c_2$ and $(Pg_1g_2)^2 \le cPg_1^2Pg_2^2$ for a constant $c < 1$ and constants $c_1 < 1 < c_2$ close to* 1. *Then*

$$P(g_1h + g_2)^2 \ge C(Pg_1^2 + Pg_2^2),$$

for a constant C depending on c, c_1 and c_2 that approaches $1 - \sqrt{c}$ as $c_1 \uparrow 1$ and $c_2 \downarrow 1$.

PROOF. We may first use the inequalities

$$\begin{aligned}(g_1h + g_2)^2 &\ge c_1g_1^2h + 2g_1hg_2 + c_2^{-1}g_2^2h\\ &= (g_1 + g_2)^2h + (c_1 - 1)g_1^2h + (1 - c_2^{-1})g_2^2h\\ &\ge c_1(g_1^2 + 2g_1g_2 + g_2^2) + (c_1 - 1)c_2g_1^2 + (c_2^{-1} - 1)g_2^2.\end{aligned}$$

Next, we integrate this with respect to P and use the inequality for Pg_1g_2 on the second term to see that the left-hand side of the lemma is bounded below by

$$c_1\Big(Pg_1^2 - 2\sqrt{cPg_1^2Pg_2^2} + Pg_2^2\Big) + (c_1 - 1)c_2Pg_1^2 + (c_2^{-1} - 1)c_2Pg_2^2.$$

Finally, apply the inequality $2xy \le x^2 + y^2$ on the second term. □

Finally, we sketch a proof that the maximum likelihood estimator for θ is asymptotically efficient. Since $\hat{\theta}$ maximizes the function $t \to \mathbb{P}_n \ln \mathrm{lik}(t, \Lambda_t(\hat{\theta}, \hat{\Lambda}))$ over t, we have

$$\mathbb{P}_n\dot{\ell}(\cdot; \hat{\theta}, \hat{\Lambda}, \hat{\theta}) = 0.$$

By the Donsker property of the class of functions $\dot{\ell}(\cdot; t, \Lambda, \theta)$ and the consistency of $(\hat{\theta}, \hat{\Lambda})$, we have, with $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_0)$, the empirical process

$$\mathbb{G}_n\big(\dot{\ell}(\cdot; \hat{\theta}, \hat{\Lambda}, \hat{\theta}) - \dot{\ell}(\cdot; \theta_0, \Lambda_0, \theta_0)\big) \to_{\mathrm{P}} 0.$$

Combining these two equations, we see that

$$-\sqrt{n}P_0\dot{\ell}(\cdot; \hat{\theta}, \hat{\Lambda}, \hat{\theta}) = \mathbb{G}_n\dot{\ell}(\cdot; \theta_0, \Lambda_0, \theta_0) + o_P(1).$$

By the same argument as used for the verification of (3.15), we can use the $O_P(n^{-1/3})$-rate of $\hat\Lambda$ in L_2 to prove that

$$\sqrt{n}P_0\dot\ell(\cdot;\theta_0,\hat\Lambda,\theta_0)\to_{\mathrm{P}} 0.$$

We add this equation to the preceding display and next use the mean value theorem on the left-hand side to obtain that, for some point $\tilde\theta$ between $\hat\theta$ and θ_0 and $\ddot\kappa(\cdot;t,\Lambda) = (\partial/\partial t)\dot\ell(\cdot;t,\Lambda,t)$,

$$-\sqrt{n}(\hat\theta-\theta_0)P_0\ddot\kappa(\cdot;\tilde\theta,\hat\Lambda) = \mathbb{G}_n\dot\ell(\cdot;\theta_0,\Lambda_0,\theta_0)+o_P(1).$$

Here, $P_0\ddot\kappa(\cdot;\tilde\theta,\hat\Lambda)$ converges to $-\tilde I$. [Note that the identity $P_{\theta,\Lambda}\dot\ell(\cdot;\theta,\Lambda,\theta)=0$ implies that $P_0\ddot\kappa(\cdot;\theta_0,\Lambda_0) = -P_0\ell_\theta(\theta_0,\Lambda_0)\dot\ell(\cdot;\theta_0,\Lambda_0,\theta_0)$.] Equation (3.5) follows, since $\dot\ell(\cdot;\theta_0,\Lambda_0,\theta_0)=\tilde\ell$.

REFERENCES

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

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

BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150.

CHANG, M. N. (1990). Weak convergence of a self-consistent estimator of the survival function with doubly censored data. *Ann. Statist.* **18** 391–404.

CHANG, M. N. and YANG, G. L. (1987). Strong consistency of a nonparametric estimator of the survival function with doubly censored data. *Ann. Statist.* **15** 1536–1547.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

GILL, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von-Mises method (part I). *Scand. J. Statist.* **16** 97–128.

GILL, R. D., VAN DER LAAN, M. J. and WIJERS, B. J. (1995). The line segment problem. Preprint.

GINÉ, E. and ZINN, J. (1986). Lectures on the central limit theorem for empirical processes. *Lecture Notes in Math.* **1221** 50–11. Springer, Berlin.

GROENEBOOM, P. (1987). Asymptotics for interval censored observations. Report 87-18, Dept. Mathematics, Univ. Amsterdam.

GROENEBOOM, P. and WELLNER, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser, Basel.

GU, M. G. and ZHANG, C. H. (1993). Asymptotic properties of self-consistent estimators based on doubly censored data. *Ann. Statist.* **21** 611–624.

HALL, P. and LA SCALA, B. (1990). Methodology and algorithms of empirical likelihood. *Internat. Statist. Rev.* **58** 109–127.

HUANG, J. (1996). Efficient estimation for the Cox model with interval censoring. *Ann. Statist.* **24** 540–568.

HUANG, J. and WELLNER, J. A. (1995). Efficient estimation for the Cox model with Case 2 interval censoring. Preprint.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Ann. Math. Statist.* **27** 887–906.

KLAASSEN, C. A. J. (1987). Consistent estimation of the influence function of locally efficient estimates. *Ann. Statist.* **15** 617–627.

LI, G. (1995). On nonparametric likelihood ratio estimation of survival probabilities for censored data. *Statist. Probab. Lett.* **25** 95–104.

MURPHY, S. A. (1994). Consistency in a proportional hazards model incorporating a random effect. *Ann. Statist.* **22** 712–731.

MURPHY, S. A. (1995a). Asymptotic theory for the frailty model *Ann. Statist.* **23** 182–198.

MURPHY, S. A. (1995b). Likelihood ratio based confidence intervals in survival analysis. *J. Amer. Statist. Assoc.* **90** 1399–1405.

NIELSEN, G. G., GILL, R. D., ANDERSEN, P. K. and SORENSEN, T. I. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Statist.* **19** 25–44.

OSSIANDER, M. (1987). A central limit theorem under metric entropy with L_2 bracketing. *Ann. Probab.* **15** 897–919.

OWEN, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249.

PFANZAGL, J. (1990). *Estimation in Semiparametric Models. Lecture Notes in Statist.* **63**. Springer, New York.

QIN, J. (1993). Empirical likelihood in biased sample problems. *Ann. Statist.* **21** 1182–1196.

QIN, J. and LAWLESS, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22** 300–325.

QIN, J. and WONG, A. (1996). Empirical likelihood in a semi-parametric model. *Scand. J. Statist.* **23** 209–220.

RAO, R. R. (1963). The law of large numbers for $D[0,1]$-valued random variables. *Theory Probab. Appl.* **8** 7–74.

ROEDER, K., CARROLL, R. J. and LINDSAY, B. G. (1996). A semiparametric mixture approach to case–control studies with errors in covariables. *J. Amer. Statist. Assoc.* **91** 722–732.

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

THOMAS, D. R. and GRUNKEMEIER, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *J. Amer. Statist. Assoc.* **70** 865–871.

VAN DER LAAN, M. (1993). Efficient and inefficient estimation in semiparametric models. Ph.D. dissertation, Univ. Utrecht.

VAN DER VAART, A. W. (1991). On differentiable functionals. *Ann. Statist.* **19** 178–204.

VAN DER VAART, A. W. (1994a). Infinite dimensional M-estimators In *Proceedings of the 6th International Vilnius Conference* (B. Grigelionis, J. Kubilius, H. Pragarauskas and V. Statulevicius, eds.) 715–734. VSP International Science Publishers, Zeist.

VAN DER VAART, A. W. (1994b). Bracketing smooth functions. *Stochastic Process. Appl.* **52** 93–105.

VAN DER VAART, A. W. (1994c). On a model of Hasminskii and Ibragimov. In *Proceedings of the Kolmogorov Semester at the Euler International Mathematical Institute, St. Petersburg* (A. A. Zaitsev, ed.). North-Holland, Amsterdam. To appear.

VAN DER VAART, A. W. (1996). Efficient estimation in semiparametric models. *Ann. Statist.* **24** 862–878.

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

WIJERS, B. J. (1995). Nonparametric estimation for a windowed line-segment process. Ph.D. dissertation, Univ. Utrecht.

WONG, W. H. and SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23** 339–362.

DEPARTMENT OF STATISTICS
PENNSYLVANIA STATE UNIVERSITY
326 CLASSROOM BUILDING
UNIVERSITY PARK, PENNSYLVANIA 16802
E-MAIL: murphy@stat.psu.edu

DEPARTMENT OF MATHEMATICS
FREE UNIVERSITY
DE BOELELAAN 1081A
1081 HV AMSTERDAM
NETHERLANDS
E-MAIL: aad@cs.vu.nl