On Profile Likelihood

S.A. Murphy and A.W. van der Vaart

We show that semiparametric profile likelihoods, where the nuisance parameter has been profiled out, behave like ordinary likelihoods in that they have a quadratic expansion. In this expansion the score function and the Fisher information are replaced by the efficient score function and efficient Fisher information, respectively. The expansion may be used, among others, to prove the asymptotic normality of the maximum likelihood estimator, to derive the asymptotic chi-squared distribution of the log likelihood ratio statistic, and to prove the consistency of the observed information as an estimator of the inverse of the asymptotic variance.

KEY WORDS: Least favorable submodel, maximum likelihood, standard error, nuisance parameter, semiparametric model, likelihood ratio statistic.

1. INTRODUCTION

A likelihood function for a low-dimensional parameter can be conveniently visualized by its graph. If the likelihood is smooth, then this is roughly, at least locally, a reversed parabola with its top at the maximum likelihood estimator. The (negative) curvature of the graph at the maximum likelihood estimator is known as the "observed information" and provides an estimate for the inverse of the variance of the maximum likelihood estimator: steep likelihoods yield accurate estimates.

The use of the likelihood function in this fashion is not possible for higher-dimensional parameters, and fails particularly for semiparametric models. For example, in semiparametric models the observed information, if it exists, would at best be an infinite-dimensional operator. Frequently, this problem is overcome by using a profile likelihood rather than a full likelihood. If the parameter is partitioned as (θ, η), with θ being a low-dimensional parameter of interest and η a higher dimensional nuisance parameter, and $\mathrm{l}_n(\theta, \eta)$ is the full likelihood, then the *profile likelihood* for θ is defined as

$$\mathrm{pl}_n(\theta) = \sup_{\eta} \mathrm{l}_n(\theta, \eta). \tag{1.1}$$

The profile likelihood may be used to a considerable extent as a full likelihood for θ. First, the maximum likelihood estimator for θ, the first component of the pair $(\hat{\theta}_n, \hat{\eta}_n)$ that maximizes $\mathrm{l}_n(\theta, \eta)$, is the maximizer of the profile likelihood function $\theta \mapsto \mathrm{pl}_n(\theta)$. This is just

Susan Murphy is Associate Professor of Statistics and Senior Associate Research Scientist, Institute for Social Research, University of Michigan and Aad Van der Vaart is Professor of Statistics, Free University, Amsterdam. Murphy's research was partially supported by National Institute on Drug Abuse grant A P50 DA 10075, and National Science Foundation grants SBR-9811983 and DMS-9802885. The authors thank Jamie Robins for sharing his general insights and for pointing out the relevance of the paper by Donald and Newey (1994) for the interpretation of the "no bias" condition.

"taking the supremum in two steps". Second, the likelihood ratio statistic for testing the (composite) hypothesis $H_0: \theta = \theta_0$ can be expressed in the profile likelihood function as

$$\frac{\mathrm{pl}_n(\hat{\theta}_n)}{\mathrm{pl}_n(\theta_0)} = \frac{\sup_{\theta,\eta} \mathrm{l}_n(\theta,\eta)}{\sup_{\eta} \mathrm{l}_n(\theta_0,\eta)}.$$

Finally, it is customary to use the curvature of the profile likelihood function as an estimate of the variability of $\hat{\theta}$. For a Euclidean parameter this is justified by Patefield (1977), who shows that in parametric models the inverse of the observed profile information is equal to the θ-aspect of the full observed inverse information. A further discussion in the parametric context is given by Barndorff-Nielsen and Cox (1994, p.90).

Thus, it appears that the profile likelihood can be used and visualized in the same fashion as an ordinary parametric likelihood. This may be considered obvious enough to recommend their use, for computational and inferential purposes alike. However, the usual theoretical justification for use of a (profile) likelihood as an inferential tool, based on a Taylor expansion, fails for semiparametric models for a variety of reasons. First, the maximum likelihood estimator $\hat{\eta}$ for the nuisance parameter may converge at a slower rate than the usual $\sqrt{n}$-rate. Second, semiparametric likelihoods may fail to be differentiable in the nuisance parameter in a suitable sense, in particular when such a likelihood contains "empirical terms". Third, handling the remainder terms in some sort of a Taylor expansion is impossible under naive classical Cramér conditions, by the presence of an infinite-dimensional nuisance parameter.

The purpose of this paper is to give a general justification for using a semiparametric profile likelihood function as an inferential tool. For this we use empirical processes to reduce the infinite dimensional problem to a problem involving the finite dimensional "least favorable submodel." Once this reduction is made, classical Cramér conditions on the low dimensional model can be used.

Some of the issues can be illustrated by the well-known Cox model for survival analysis. Consider the simplest version, without censoring. In this model the hazard function of the survival time T of a subject with covariate Z is given by

$$\lambda^{T|Z}(t) = e^{\theta Z}\lambda(t),$$

for a linear function $z \mapsto \theta z$ (taken one dimensional for simplicity) and an unspecified baseline hazard function λ. The density of the observation (T, Z) (relative to the product of the Lebesgue measure and the marginal distribution of Z, which is assumed free of (θ, Λ)) is equal to

$$e^{\theta z}\lambda(t)e^{-e^{\theta z}\Lambda(t)},$$

where Λ is the cumulative hazard (with $\Lambda(0) = 0$). The usual estimator for (θ, Λ) based on a sample of size n from this model is the maximum likelihood estimator $(\hat{\theta}, \hat{\Lambda})$, where the

likelihood (Andersen et al., 1993) is defined as, with $\Lambda\{t\}$ the jump of Λ at t,

$$\mathrm{l}_n(\theta,\Lambda)=\prod_{i=1}^n e^{\theta z_i}\Lambda\{t_i\}e^{-e^{\theta z_i}\Lambda(t_i)}.$$

The form of this likelihood forces $\hat{\Lambda}$ to be a jump function with jumps at the observed deaths t_i only, and the likelihood depends smoothly on the unknowns $\big(\Lambda\{t_1\},\ldots,\Lambda\{t_n\}\big)$. However, the usual Taylor expansion scheme does not apply to this likelihood, because the dimension of this vector converges to infinity with n, and it is unclear which other device could be used to "differentiate" the likelihood with respect to Λ. The usual solution is to "profile" out the nuisance parameter. Elementary calculus shows that, for a fixed θ, the function

$$(\lambda_1,\ldots,\lambda_n)\mapsto\prod_{i=1}^n e^{\theta z_i}\lambda_i e^{-e^{\theta z_i}\sum_{j:t_j\le t_i}\lambda_j}$$

is maximal for

$$\frac{1}{\lambda_k}=\sum_{i:t_i\ge t_k}e^{\theta z_i}.$$

Thus the profile likelihood is given by

$$\mathrm{pl}_n(\theta)=\sup_{\Lambda}\mathrm{l}_n(\theta,\Lambda)=\prod_{i=1}^n\frac{e^{\theta z_i}}{\sum_{i:t_i\ge t_k}e^{\theta z_i}}e^{-1}.$$

The latter expression is the Cox partial likelihood. Thus the Cox partial likelihood estimator is a maximum likelihood estimator, and a maximum profile likelihood estimator for θ.

We can derive the same results for more practical versions of the Cox model. In particular, we can make one step up in complexity by introducing censoring: rather than observing (T,Z) we only observe $X=(T\wedge C,1\{T\le C\},Z)$, where given Z the variables T and C are independent, and (T,Z) follows the Cox model as before. The density of $X=(Y,\Delta,Z)$ is now given by

$$\Big(e^{\theta z}\lambda(y)e^{-e^{\theta z}\Lambda(y)}\big(F_{C|Z}(y-|\,z)\big)\Big)^{\delta}\Big(e^{-e^{\theta z}\Lambda(y)}f_{C|Z}(y|\,z)\Big)^{1-\delta}. \tag{1.2}$$

We define a likelihood for the parameters (θ,Λ) by dropping the factors involving the distribution of (C,Z) and replacing $\lambda(y)$ by the pointmass $\Lambda\{y\}$,

$$\mathrm{l}_n(\theta,\Lambda)=\prod_{i=1}^n\Big(e^{\theta z_i}\Lambda\{y_i\}e^{-e^{\theta z_i}\Lambda(y_i)}\Big)^{\delta_i}\Big(e^{-e^{\theta z_i}\Lambda(y_i)}\Big)^{1-\delta_i}. \tag{1.3}$$

In the remainder of the paper we shall use the Cox model in this form as an illustrative example. Even though this example is well-known and has been treated by various methods by many authors, our treatment here is actually novel. More importantly, our treatment extends to other models, whereas existing approaches to the Cox model exploit its special features and do not (easily) generalize. The Cox model is a useful example as the formulas for the model are particularly simple.

The Cox estimator was shown to be asymptotically normal by Tsiatis (1981) and Andersen and Gill (1982). This was not a trivial undertaking, because in form the profile likelihood does not resemble an ordinary likelihood at all. Even though it is a product, the terms in the product are heavily dependent. Tsiatis (1981) and Andersen and Gill (1982) show how this particular profile likelihood can nevertheless be elegantly handled using counting processes and martingales.

The Cox profile likelihood has the great advantage that it can be found in closed form by analytic methods. This has been the basis for finding the properties of the Cox estimator, but it turns out to be special to this model. In most other examples of interest the profile likelihood is, and remains, a supremum, which can be calculated numerically, but whose statistical properties must be deduced in a more implicit manner than those of the Cox profile likelihood. In fact, it is not even clear that a general profile likelihood is differentiable, a supremum of differentiable functions not necessarily being differentiable itself. Thus we need to attack the problem in another way. In this paper we use a sandwiching device based on least favorable submodels.

The Cox likelihood as given previously can be termed an "empirical likelihood", as it contains the point masses $\Lambda\{t_i\}$. This terminology is in agreement with that of Owen (1988), as the terms in the product approximate the pointmasses $\mathrm{P}_{\theta,\Lambda}(T = t_i |\, Z = z_i)$. A difference between our setting and that of Owen is that we restrict the point masses to point masses resulting from a given semiparametric model. We also describe our functional of interest directly through a parameterization (θ, η), and not as a functional on the model (such as a mean or a solution of an estimating equation). The latter is special, as we also assume that the parameter spaces for θ and η are fixed and independent of the other parameter ("variation free"). Our main result can be extended to general functionals on the model, but the formulation will be less transparent. Empirical likelihoods in this sense have a long history in survival analysis, where they are sometimes also referred to as "nonparametric likelihoods", leading to "nonparametric maximum likelihood estimators" (NPMLE). A complicating feature of such likelihoods is that they are often not smooth in the nuisance parameter (viewed as an element of an appropriate normed space). This immediately destroys any hope for a general asymptotic result about likelihood under "Cramér-style" conditions.

Several other slightly different choices of likelihood are natural in the Cox model also, although they lead to slightly different estimators. Our present choice has the advantage of leading to Cox's partial likelihood estimator, which is firmly established in survival analysis. See Andersen et al. (1993, pg. 221-226) for a discussion of this issue in the context of nonparametric estimation of the cumulative hazard in survival analysis and Bailey (1994) for a discussion for the Cox model. Asymptotically, these likelihoods give identical results, and the finite sample simulations appear not to lead to a clear preference. In general, there is no unique method to define a likelihood for a semiparametric model, and the

semiparametric likelihoods proposed in the literature take a variety of forms: for instance, ordinary densities, empirical likelihoods and mixtures of these two. In this paper we consider arbitrary "likelihoods" based on a random sample of n observations $X_1, \ldots, X_n$ from a distribution depending on two parameters θ and η. We write the "likelihood" for one observation as $\mathrm{l}(\theta, \eta)(x_i)$, whence the full likelihood $\mathrm{l}_n(\theta, \eta)$ is a product of this expression over i. Of course, the results of the paper may not be valid if the choice of likelihood does not satisfy the conditions made below, and these conditions are motivated by our intuitive ideas of "likelihood". The following two examples illustrate some of the possibilities and are discussed in detail later on.

Example: Cox Regression for Current Status Data. Under "current status" censoring a subject is examined once, at a random observation time Y, and at this time it is observed whether the subject is alive or not. We assume that the pair (T, Z) of time of death T and covariate Z follows a Cox model, but we only observe $X = (Y, 1\{T \leq Y\}, Z)$.

The density of X is given by

$$p_{\theta,\Lambda}(x) = \bigl(1 - \exp(-e^{\theta^T z}\Lambda(y))\bigr)^{\delta} \bigl(\exp(-e^{\theta^T z}\Lambda(y))\bigr)^{1-\delta} f^{Y,Z}(y,z),$$

with parameters, Λ, the cumulative baseline hazard function and θ, the vector of regression coefficients. We take this expression, but with the term $f^{Y,Z}(Y,Z)$ deleted, as the likelihood. Thus in this example the likelihood is an ordinary density.

This asymptotic behavior of the maximum likelihood estimator in this model is discussed in Huang (1996), who also computed the profile log likelihood function for an application to a study of tumor growth in mice. The profile log likelihood for θ based on a sample of $n = 144$ mice with binary covariate value $z \in \{0, 1\}$ for assignment to a germ-free or conventional environment is given on p551 of this paper. The profile log likelihood is markedly quadratic. Even though there is no explicit expression, the profile likelihood function is easy to compute, for instance using the iterative least concave majorant algorithm of Groeneboom and Wellner (1992). The results of our paper show that this graph can be interpreted as usual..

Example: Missing Covariate. Consider a basic random vector (D, W, Z), whose distribution is described in the following way:

- D is a logistic regression on $\exp Z$ with intercept and slope γ and β, respectively;
- W is a linear regression on Z with intercept and slope α_0 and α_1, respectively, and an $N(0, \sigma^2)$-error;
- Given Z the variables D and W are independent;
- Z has a completely unspecified distribution η.

The unknown parameters are $\theta = (\beta, \alpha_0, \alpha_1, \gamma, \sigma)$ and the distribution η of the regression variable. The likelihood for the vector (D, W, Z) takes the form $p_\theta(d, w|\, z)\, d\eta(z)$, with ϕ

denoting the standard normal density,

$$p_\theta(d, w|\, z) = \Big(\frac{1}{1+\exp(-\gamma-\beta e^z)}\Big)^d \Big(\frac{\exp(-\gamma-\beta e^z)}{1+\exp(-\gamma-\beta e^z)}\Big)^{1-d} \frac{1}{\sigma}\phi\Big(\frac{w-\alpha_0-\alpha_1 z}{\sigma}\Big)$$

and $d\eta$ denoting the density of η with respect to a dominating measure.

In a simplified version of a model considered by Roeder et al. (1996), a typical observation $X = (Y_C, Z_C, Y_R)$ consists of a "complete" observation $Y_C = ((D_C, W_C), Z_C)$ and an "reduced" observation $Y_R = (D_R, W_R)$, in which the covariate is missing. The density of X takes the form

$$p_{\theta,\eta}(x) = p_\theta(y_C|\, z_C)\, d\eta(z_C) \int p_\theta(y_R|\, z)\, d\eta(z).$$

The likelihood is constructed out of a mixture of ordinary and empirical likelihoods,

$$\mathrm{l}(\theta,\eta)(x) = p_\theta(y_C|\, z_C)\, \eta\{z_C\} \int p_\theta(y_R|\, z)\, d\eta(z).$$

Note that this "likelihood" is a density if η is assumed to be discrete. However the point masses $\eta\{z\}$ are used even if it is known that η must be continuous.

Algorithms for the computation of the profile likelihood and simulation results (for more realistic and retrospective versions of the model) are reported in Roeder et al. (1996, p728-730), as well as an application to a cholesterol study. A graph of the profile likelihood for one of the parameters in the cholesterol study can be found on p731 of their paper. Our results give a theoretical justification for the likelihood-based confidence interval indicated in this graph. The authors conclude that "simulations and an example indicate the feasibility of the methodology" (p731). They also point out that maximum likelihood performs generally better or as good as alternative methods, even though a sample size of at least 60 appears to be necessary to see a real gain.

Example: Shared Gamma Frailty Model. In the frailty model, possibly censored failure times for subjects occurring in groups such as twins or litters are observed. To allow for a positive intra-group correlation in the subjects' event times, subjects in the same group are assumed to share a common unobservable variable G, called the "frailty". Given G and a vector of covariates, Z, the hazard function for the failure time distribution is assumed to follow a proportional hazards model,

$$G\exp(\beta^T Z_j(\cdot))Y_j(\cdot)\dot{\eta}(\cdot)$$

where β is a d-dimensional vector of regression coefficients and $\dot{\eta}$ is the baseline hazard function. The unobserved frailty, G, follows a gamma distribution with mean one and variance σ.

The unknown parameters are $\theta = (\beta, \sigma)$ and the cumulative baseline hazard, η. The likelihood for an observation of a group of (possibly) censored failure times, censoring indicators, and the covariates for the group, is given in Section 4.3. In a variety of simulated

scenarios, Nielsen et al. (1992) evaluated the practical usefulness of estimated standard errors found by using the curvature of a parabola fitted to the log profile likelihood. Their simulations used samples of 100 to 1000 groups of size two. The curvature of the profile likelihood produced remarkably good estimators of the standard errors. Additionally they fit this model to a data set of 16 groups, with group sizes ranging from 2 to 4.

The main result of this paper is the following asymptotic expansion of the profile likelihood. Under conditions, listed in Section 3, we prove that, for any random sequence $\tilde{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$,

$$\begin{aligned}\log \mathrm{pl}_n(\tilde{\theta}_n) = \log \mathrm{pl}_n(\theta_0) + (\tilde{\theta}_n - \theta_0)^T \sum_{i=1}^n \tilde{\ell}_0(X_i) \\ - \tfrac{1}{2} n (\tilde{\theta}_n - \theta_0)^T \tilde{I}_0 (\tilde{\theta}_n - \theta_0) + o_{P_{\theta_0,\eta_0}}\bigl(\sqrt{n}\|\tilde{\theta}_n - \theta_0\| + 1\bigr)^2.\end{aligned} \tag{1.4}$$

Here $\tilde{\ell}_0$ is the *efficient score function for* θ, the ordinary score function minus its orthogonal projection onto the closed linear span of the score functions for the nuisance parameter. (Cf. Begun et al. (1983).) Furthermore, $\tilde{I}_0$ is its covariance matrix, the *efficient Fisher information matrix*. Under similar conditions, the maximum likelihood estimator is asymptotically normal, and has the asymptotic expansion

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{I}_0^{-1} \tilde{\ell}_0(X_i) + o_{P_{\theta_0,\eta_0}}(1). \tag{1.5}$$

Taking this into account we see that the parabolic approximation to the log profile likelihood given by equation (1.4) is centered, to the first order, at $\hat{\theta}_n$. In other words, it is possible to expand the log profile likelihood function around $\hat{\theta}_n$, in the form

$$\begin{aligned}\log \mathrm{pl}_n(\tilde{\theta}_n) = \log \mathrm{pl}_n(\hat{\theta}_n) - \tfrac{1}{2} n (\tilde{\theta}_n - \hat{\theta}_n)^T \tilde{I}_0 (\tilde{\theta}_n - \hat{\theta}_n) \\ + o_{P_{\theta_0,\eta_0}}\bigl(\sqrt{n}\|\tilde{\theta}_n - \theta_0\| + 1\bigr)^2.\end{aligned} \tag{1.6}$$

The asymptotic expansions (1.4) and (1.6) justify using a semiparametric profile likelihood as an ordinary likelihood, at least asymptotically. In particular, we present three corollaries. (These also show that (1.5)–(1.6) are a consequence of (1.4).) Parametric versions of some of these corollaries have appeared many times; see for instance Chernoff (1954) or Pollard (1984, Ch. VII). We assume throughout the paper that the true parameter θ_0 is interior to the parameter set.

Corollary 1.1. *If (1.4) holds, $\tilde{I}_0$ is invertible, and $\hat{\theta}_n$ is consistent, then (1.5)–(1.6) hold. In particular, the maximum likelihood estimator is asymptotically normal with mean zero and covariance matrix the inverse of $\tilde{I}_0$.*

Proof. Relation (1.6) follows from (1.4)–(1.5) and some algebra. We shall derive (1.5) from (1.4). Set $\Delta_n = n^{-1/2} \sum_{i=1}^n \tilde{\ell}_0(X_i)$ and $\hat{h} = \sqrt{n}(\hat{\theta} - \theta_0)$.

Applying (1.4) with the choices $\tilde{\theta} = \hat{\theta}$ and $\tilde{\theta} = \theta_0 + n^{-1/2}\tilde{I}_0^{-1}\Delta_n$, we find

$$\log \mathrm{pl}_n(\hat{\theta}) = \log \mathrm{pl}_n(\theta_0) + \hat{h}^T\Delta_n - \tfrac{1}{2}\hat{h}^T\tilde{I}_0\hat{h} + o_P\big(\|\hat{h}\| + 1\big)^2,$$
$$\log \mathrm{pl}_n(\theta_0 + n^{-1/2}\tilde{I}_0^{-1}\Delta_n) = \log \mathrm{pl}_n(\theta_0) + \Delta_n^T\tilde{I}_0^{-1}\Delta_n - \tfrac{1}{2}\Delta_n^T\tilde{I}_0^{-1}\Delta_n + o_P(1).$$

By the definition of $\hat{\theta}$, the expression on the left (and hence on the right) in the first equation is larger than the expression on the left in the second equation. It follows that

$$\hat{h}^T\Delta_n - \tfrac{1}{2}\hat{h}^T\tilde{I}_0\hat{h} - \tfrac{1}{2}\Delta_n^T\tilde{I}_0^{-1}\Delta_n \geq -o_P\big(\|\hat{h}\| + 1\big)^2.$$

The left side of this inequality is equal to

$$-\tfrac{1}{2}(\hat{h} - \tilde{I}_0^{-1}\Delta_n)^T\tilde{I}_0(\hat{h} - \tilde{I}_0^{-1}\Delta_n) \leq -c\|\hat{h} - \tilde{I}_0^{-1}\Delta_n\|^2,$$

for a positive constant c, by the nonsingularity of $\tilde{I}_0$. Conclude that

$$\|\hat{h} - \tilde{I}_0^{-1}\Delta_n\| = o_P\big(\|\hat{h}\| + 1\big).$$

This implies first that $\|\hat{h}\| = O_P(1)$, and next, by reinsertion, that $\|\hat{h} - \tilde{I}_0^{-1}\Delta_n\| = o_P(1)$. This completes the proof of (1.5). ■

The second corollary concerns the likelihood ratio statistic, and shows that this behaves as it should. The corollary justifies using the set

$$\left\{\theta: 2\log\frac{\mathrm{pl}_n(\hat{\theta}_n)}{\mathrm{pl}_n(\theta)} \leq \chi^2_{d,1-\alpha}\right\}$$

as a confidence set of approximate coverage probability $1 - \alpha$.

Corollary 1.2. *If (1.4) holds, $\tilde{I}_0$ is invertible, and $\hat{\theta}_n$ is consistent, then under the null hypothesis $H_0{:}\,\theta = \theta_0$, the sequence $2\log\big(\mathrm{pl}_n(\hat{\theta}_n)/\mathrm{pl}_n(\theta_0)\big)$ is asymptotically chi-squared distributed with d degrees of freedom.*

Proof. This is immediate from (1.5)–(1.6) upon choosing $\tilde{\theta}_n = \theta_0$ in (1.6). ■

The third corollary concerns a discretized second derivative of the profile likelihood. This estimator is the square of the numerical derivative of the signed log-likelihood ratio statistic as discussed by Chen and Jennrich (1996). In their Theorem 3.1 they show that, in the parametric setting, the derivative of the signed log-likelihood ratio statistic is the square root of the observed information about θ. Indeed since the first derivative of the profile likelihood at $\hat{\theta}_n$ should be zero, the expression in the following display can be labelled the *observed information* about θ (evaluated in the direction v_n). Thus, this corollary can be used to construct an estimate of the standard error of $\hat{\theta}_n$.

Corollary 1.3. *If (1.4) holds and $\hat{\theta}_n$ is consistent, then, for all sequences $v_n \xrightarrow{\mathrm{P}} v \in \mathbb{R}^d$ and $h_n \xrightarrow{\mathrm{P}} 0$ such that $(\sqrt{n}h_n)^{-1} = O_P(1)$,*

$$-2\frac{\log \mathrm{pl}_n(\hat{\theta}_n + h_n v_n) - \log \mathrm{pl}_n(\hat{\theta}_n)}{nh_n^2} \xrightarrow{\mathrm{P}} v^T \tilde{I}_0 v.$$

Proof. This is immediate from (1.6) upon choosing $\tilde{\theta}_n = \hat{\theta}_n + h_n v_n$. ∎

The restriction on the rate of convergence of h_n to zero stems from the error term in (1.6) being in terms of $\|\tilde{\theta}_n - \theta_0\| + 1$ rather than in $\|\tilde{\theta}_n - \hat{\theta}_n\|$ only. If the model is parametric and $\hat{\eta}_{\tilde{\theta}_n}$ (defined below) is consistent, then Cramér's conditions (Cramér, 1946, pg. 500; Le Cam and Yang, 1990, pg. 101), are sufficient to show that the error is in terms of the latter difference. For this reason, we believe that in semiparametric problems for which the nuisance parameter can be estimated at a parametric (square root n) rate, the error term in (1.6) can be improved.

The organization of the paper is as follows. In Section 2 we review the notion of a least favorable submodel. This concept is useful in understanding why the general scheme to prove expansion (1.4) given in Section 3 should be and is successful. This scheme applies to a wide variety of examples. In Section 4 we discuss four examples, where we refer to the literature for technical details. In the course of describing our main result in Sections 2 and 3, we also discuss the Cox model as introduced earlier as a fifth example. This does not lead to new results on the Cox estimators, but because the formulas for this model are extremely simple this continued discussion may be helpful. Section 4 is followed by a discussion of the main points of this paper.

We use the notations $\mathbb{P}_n$ and $\mathbb{G}_n$ for the empirical distribution and the empirical process of the observations, respectively. Furthermore, we use operator notation for evaluating expectations. Thus, for every measurable function f and probability measure P,

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i), \qquad Pf = \int f\, dP, \qquad \mathbb{G}_n f = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(f(X_i) - P_0 f\big),$$

where $P_0 = P_{\theta_0,\eta_0}$ is the true underlying measure of the observations. We abbreviate (θ_0, η_0) to 0 also at other places. Additionally, for a function $x \mapsto \ell(\theta,\eta)(x)$ indexed by (θ,η), notation such as $P\ell(\hat{\theta},\hat{\eta})$ is an abbreviation for $\int \ell(\theta,\eta)(x)\, dP(x)$ evaluated at $(\theta,\eta) = (\hat{\theta},\hat{\eta})$.

2. LEAST FAVORABLE SUBMODELS

Our primary tool in demonstrating (1.4) is to reduce the high dimensional model to a finite dimensional, random submodel of the same dimension as θ. An example of a submodel is given below. We then use ordinary Taylor series expansions on the submodel, controlling error terms by empirical process techniques. Since this paper concerns maximum likelihood estimators, the facts below will lead us to the use of "approximately least favorable submodels". In this section we review the concepts of "least favorable submodels" and "efficient scores." We also review how one may define a score function for a infinite dimensional parameter. This section may be skipped by readers familiar with (semiparametric) efficiency theory as given in, for instance, Begun et al. (1983) and Bickel et al. (1993).

First consider least favorable submodels in the setting of a smooth parametric model $P_{\theta,\eta}$, where both θ and η are Euclidean. Partition the Fisher information matrix as

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

If $i_{\eta_0\theta_0}$ can be expressed as a linear combination of the columns of $i_{\eta_0\eta_0}$, then $i^-_{\eta_0\eta_0} i_{\eta_0\theta_0}$ can be defined as the solution in x to the equation $i_{\eta_0\eta_0} x = i_{\eta_0\theta_0}$. (This is certainly possible if $i_{\eta_0\eta_0}$ is nonsingular, and then we may read $i^-_{\eta_0\eta_0}$ as the inverse of $i_{\eta_0\eta_0}$, but is possible more generally.) An example of a submodel of the model $(\theta,\eta) \mapsto P_{\theta,\eta}$ is $\theta \mapsto P_{\theta,\eta_\theta}$ where $\eta_\theta = \eta_0 + \left(i^-_{\eta_0\eta_0} i_{\eta_0\theta_0}\right)^T(\theta_0 - \theta)$. The full model is of the dimension of θ plus the dimension of η, whereas the submodel is only of the dimension of θ.

The *efficient score function* for θ at (θ_0,η_0), is given by

$$\tilde{\ell}_{\theta_0,\eta_0} = \ell_{\theta_0,\eta_0} - (i^-_{\eta_0\eta_0} i_{\eta_0\theta_0})^T A_{\theta_0,\eta_0},$$

where ℓ_{θ_0,η_0} and A_{θ_0,η_0} are the score functions for θ and η, respectively, and the function $(i^-_{\eta_0\eta_0} i_{\eta_0\theta_0})^T A_{\theta_0,\eta_0}$ can be shown to be the (componentwise) projection of the score function for θ on the space spanned by the components of the nuisance score, A_{θ_0,η_0}. (Projection means minimizing the quadratic $P_{\theta_0,\eta_0}(\ell_{\theta_0,\eta_0} - k)^2$ over k ranging over the linear span of the nuisance scores.) A sequence of estimators $\hat{\theta}_n$ is best regular (or asymptotically efficient) at (θ_0,η_0) if and only if it is asymptotically linear in the efficient score, that is,

$$\sqrt{n}(\hat{\theta} - \theta_0) = \left(i_{\theta_0\theta_0} - i_{\theta_0\eta_0} i^{-1}_{\eta_0\eta_0} i_{\eta_0\theta_0}\right)^- \sqrt{n}\mathbb{P}_n \tilde{\ell}_{\theta_0,\eta_0} + o_P(1). \tag{2.1}$$

Under regularity conditions the maximum likelihood estimator for θ in the full parametric model $(\theta,\eta) \mapsto P_{\theta,\eta}$ will satisfy this equation.

Consider the submodel $\theta \mapsto P_{\theta,\eta_\theta}$ defined above, where $\eta_\theta = \eta_0 + (i^-_{\eta_0\eta_0} i_{\eta_0\theta_0})^T(\theta_0 - \theta)$. The information for the estimation of θ in this submodel is given by the covariance of the first derivative of the log likelihood with respect to θ and evaluated at $\theta = \theta_0$. The first derivative is $\tilde{\ell}_{\theta_0,\eta_0}$ and its covariance matrix is $i_{\theta_0\theta_0} - i_{\theta_0\eta_0} i^-_{\eta_0\eta_0} i_{\eta_0\theta_0}$. Thus the well-behaved maximum likelihood estimator of θ in the submodel satisfies (2.1). This submodel is

called *least favorable* at (θ_0, η_0) when estimating θ in the presence of the nuisance parameter η, because out of all submodels $\theta \mapsto P_{\theta,\eta_\theta}$, this submodel has the smallest information about θ. The vector $i_{\eta_0\eta_0}^- i_{\eta_0\theta_0}$ is called the *least favorable direction*. In parametric models existence of the efficient score and existence of the least favorable submodel are synonymous.

Similarly in semiparametric models, a best regular (asymptotically efficient) estimator must be asymptotically linear in the efficient score. Thus if the maximum likelihood estimator of θ is efficient in the full model, it is asymptotically linear in the efficient score. Given a semiparametric model of the type $\{P_{\theta,\eta}: \theta \in \Theta, \eta \in \mathcal{H}\}$ the score function for θ is defined, as usual, as the partial derivative with respect to θ of the log likelihood. A score function for η is of the form

$$\frac{\partial}{\partial t}_{|t=0} \ln p_{\theta_0,\eta_t}(x) =: A_{\theta_0,\eta_0} h(x),$$

where h is a "direction" in which $\eta_t \in \mathcal{H}$ approaches η_0, running through some index set H. In information calculations H is usually taken equal to a subset of a Hilbert space $\bar{H}_{\eta_0}$, and each $h \in H$ indexes some path $t \mapsto \eta_t$ for which the derivative in the preceding display exists; the corresponding scores are then denoted $A_{\theta_0,\eta_0} h$. Typically, this then extends to a continuous, linear map $A_{\theta_0,\eta_0}: \bar{H}_{\eta_0} \mapsto L_2(P_{\theta_0,\eta_0})$.

The theory behind such information calculations is fairly well-developed (see e.g. the book Bickel et al. (1983) for many examples). Unfortunately, by themselves they do not yield the right type of local approximations to study semiparametric likelihoods. This is different from parametric models, whose Euclidean parameter spaces allow a complete local reparametrization and corresponding derivatives. For this reason we adopt information-theoretic notation and concepts, but choose to formulate our theorems at a different level. Thus the concepts explained in this section are essential for an understanding of our results, but the mathematical proofs do not rely on them.

Example: Cox model, continued. In the Cox model with right censoring introduced in Section 1, a typical observation (Y, Δ, Z) is distributed according to the density (1.2). The score function for θ is readily computed as

$$\ell_{\theta,\Lambda}(y, \delta, z) = \delta z - z e^{\theta z} \Lambda(y).$$

The nuisance parameter, here denoted by Λ rather than η, is an arbitrary cumulative hazard function (apart from perhaps the restriction that it be absolutely continuous). Given a fixed Λ and a bounded function $h: \mathbb{R} \mapsto \mathbb{R}$ we can define a path Λ_t by $d\Lambda_t(y) = \big(1 + th(y)\big)\, d\Lambda(y)$, or in integral form $\Lambda_t(y) = \int_{[0,y]} \big(1 + th(s)\big)\, d\Lambda(s)$. As, by definition, a cumulative hazard function is a finite measure, $\Lambda_t(y)$ is well-defined for every y. To check that it is itself a cumulative hazard function we must verify that it is nonnegative and nondecreasing on $[0, \infty)$. This need not be the case, but if h is bounded, then $1 + th(y)$ is nonnegative for all $y \geq 0$ at least for every sufficiently small $|t|$ and then Λ_t satisfies both requirements. Thus

Λ_t is a valid element of the parameter set for $|t| \approx 0$ and we can use it to define a submodel. Inserting Λ_t in the density (1.2), taking the logarithm and differentiating at $t = 0$ we find the score function

$$A_{\theta,\Lambda} h(y, \delta, z) := \delta h(y) - e^{\theta z} \int_{[0,y]} z e^{\theta z} h \, d\Lambda.$$

Note that this is linear in h. The natural Hilbert space to work with in this case is $L_2(\Lambda)$, which contains all bounded functions $h: \mathbb{R} \mapsto \mathbb{R}$, but is slightly larger. It can be verified that $A_{\theta,\Lambda}: L_2(\Lambda) \mapsto L_2(P_{\theta,\Lambda})$ as given in the display is a continuous map, but we omit the details. The simplicity of the formulas of both types of scores should be viewed a consequence of the cleverness of Cox's parametrization of this model. We shall see below that other formulas also take pleasantly simple appearances.

Given the set of scores for the nuisance parameter, the *efficient score function* for θ at (θ_0, η_0) is defined as

$$\tilde{\ell}_{\theta_0,\eta_0} = \ell_{\theta_0,\eta_0} - \Pi_{\theta_0,\eta_0} \ell_{\theta_0,\eta_0}$$

where $\Pi_{\theta_0,\eta_0} \ell_{\theta_0,\eta_0}$ minimizes the squared distance $P_{\theta_0\eta_0} (\ell_{\theta_0,\eta_0} - k)^2$ over all functions k in the closed linear span (in $L_2(P_{\theta_0,\eta_0})$) of the score functions for η. The inverse of the variance of $\tilde{\ell}_{\theta_0,\eta_0}$ is the Cramér-Rao bound for estimating θ in the presence of η. A submodel $\theta \mapsto P_{\theta,\eta_\theta}$ with $\eta_{\theta_0} = \eta_0$ is least favorable at (θ_0, η_0) if

$$\tilde{\ell}_{\theta_0,\eta_0} = \frac{\partial}{\partial \theta}_{|\theta=\theta_0} \ln p_{\theta,\eta_\theta}.$$

Since a projection $\Pi_{\theta_0,\eta_0} \ell_{\theta_0,\eta_0}$ on the closed linear span of the nuisance scores is not necessarily a nuisance score itself, existence of the efficient score does not imply existence of a least favorable submodel. (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.

If the projection $\Pi_{\theta_0,\eta_0} \ell_{\theta_0,\eta_0}$ is a nuisance score, then it can be written analogously to the finite dimensional case, which is of help to construct least favorable submodels. The *adjoint operator* $A^*_{\theta_0,\eta_0}: L_2(P_{\theta_0,\eta_0}) \mapsto \bar{H}_{\eta_0}$, is a continuous, linear map that is characterized by the requirement

$$P_{\theta_0,\eta_0}(A_{\theta_0,\eta_0} h) g = \langle A_{\theta_0,\eta_0} h, g \rangle_{P_{\theta_0,\eta_0}} = \langle h, A^*_{\theta_0,\eta_0} g \rangle_{\eta_0}, \qquad \text{every } h \in \bar{H}_{\eta_0}, g \in L_2(P_{\theta_0,\eta_0}).$$

Here the brackets denote the inner products in the Hilbert spaces $L_2(P_{\theta_0,\eta_0})$ and $\bar{H}_{\eta_0}$. The projection of ℓ_{θ_0,η_0} onto the closure of the space spanned by the nuisance scores $A_{\theta_0,\eta_0} h$ is itself a nuisance score (of the form $A_{\theta_0,\eta_0} h$) if and only if the range of $A^*_{\theta_0,\eta_0} A_{\theta_0,\eta_0}$ contains $A^*_{\theta_0,\eta_0} \ell_{\theta_0,\eta_0}$, in which case it is given by

$$\Pi_{\theta_0,\eta_0} \ell_{\theta_0,\eta_0} = A_{\theta_0,\eta_0} h_{\theta_0,\eta_0}, \qquad h_{\theta_0,\eta_0} = (A^*_{\theta_0,\eta_0} A_{\theta_0,\eta_0})^- A^*_{\theta_0,\eta_0} \ell_{\theta_0,\eta_0}$$

where $(A^*_{\theta_0,\eta_0}A_{\theta_0,\eta_0})^-$ is a "generalized inverse" and h_{θ_0,η_0} is called the least favorable direction. In three of our examples we can form the projection in this manner. In the other two examples the projection exists (as it always does), but cannot written in the form as in the preceding display. The split between these cases corresponds roughly to models in which the full nuisance parameter is estimable at $\sqrt{n}$-rate and models in which the rate is slower than $\sqrt{n}$: failure of invertibility of the *information operator* $A^*_{\theta_0,\eta_0}A_{\theta_0,\eta_0}$ indicates that some aspect of η cannot be estimated at $\sqrt{n}$-rate. (For this and earlier claims in this paragraph, see Van der Vaart (1991).) The details of the arguments that need to be used to verify our main result are also different in the two cases, although the two cases can be unified at the level of least favorable submodels.

We illustrate the formulas in the case of the Cox model. This model is unusual in that one can write down explicit formulas for the inverse operator and adjoint operator (first achieved in Begun et al. (1983)). In the other examples we must work with the inverse operator $(A^*_{\theta_0,\eta_0}A_{\theta_0,\eta_0})^-$ without having an explicit representation. This is good enough as long as we can ascertain sufficiently many of its properties.

Example: Cox model, continued. For continuous Λ the information operator takes the form

$$A^*_{\theta,\Lambda}A_{\theta,\Lambda}h(y) = h(y)\mathrm{E}_{\theta,\Lambda}1_{Y\geq y}e^{\theta Z}.$$

To see this, we employ the identity $P_{\theta,\Lambda}(A_{\theta,\Lambda}g)(A_{\theta,\Lambda}h) = \Lambda\big[(g)\,(A^*_{\theta,\Lambda}A_{\theta,\Lambda}h)\big]$. First write the product $(A_{\theta,\Lambda}g)(A_{\theta,\Lambda}h)$ as the sum of four terms:

$$\delta h(y)g(y) - \delta h(y)e^{\theta z}g\,d\Lambda - \delta g(y)e^{\theta z}\int_0^y h\,d\Lambda + e^{2\theta z}\int_0^y g\,d\Lambda\int_0^y h\,d\Lambda.$$

Take the expectation under $P_{\theta,\Lambda}$ and interchange the order of the integrals to write this in the form $\Lambda\big[gI_{\theta,\Lambda}(h)\big]$, where $I_{\theta,\Lambda}(h)$ is also a sum of four terms. In view of the defining identity for the adjoint, $I_{\theta,\Lambda}(h) = A^*_{\theta,\Lambda}A_{\theta,\Lambda}h$. To simplify the definition of $I_{\theta,\lambda}(h)$ partially integrate its fourth term to see that this cancels the second and third terms. Thus $A^*_{\theta,\Lambda}A_{\theta,\Lambda}h$ is the first term of $I_{\theta,\Lambda}(h)$, which is as claimed.

The function $A^*_{\theta,\Lambda}\ell_{\theta,\Lambda}$, which is also part of the formula for the efficient score function, can be obtained by a similar argument, starting from the identity $P_{\theta,\Lambda}\ell_{\theta,\Lambda}(A_{\theta,\Lambda}h) = \Lambda\big[(A^*_{\theta,\Lambda}\ell_{\theta,\Lambda})\,h\big]$. It is given by

$$A^*_{\theta,\Lambda}\ell_{\theta,\Lambda}(y) = \mathrm{E}_{\theta,\Lambda}1_{Y\geq y}Ze^{\theta Z}.$$

Begun et al. (1983) parameterized the model through the survival function corresponding to Λ and obtained similar formulas, which are more complicated because they also contain the operators translating survival functions in hazard functions and conversely.

There is an even faster method of obtaining the same formulas, which is based on a semiparametric version of the fact that the Fisher information is equal to "the expectation of minus the second derivative of a log likelihood". For instance, to obtain the formula for

the information operator we can proceed as follows. Consider the two-dimensional submodel $(s,t) \mapsto P_{\theta,\Lambda_{s,t}}$ for $\Lambda_{s,t}$ defined by $d\Lambda_{s,t} = (1 + sg + th)\, d\Lambda$ and two given functions h and g. The scores for s and t of this submodel at $(s,t) = (0,0)$ are $A_{\theta,\Lambda} g$ and $A_{\theta,\Lambda} h$, respectively. Therefore, by the identity mentioned previously applied to the two-dimensional (sub)model,

$$P_{\theta,\Lambda}(A_{\theta,\Lambda}h)(A_{\theta,\Lambda}g) = -P_{\theta,\Lambda}\frac{\partial^2}{\partial s \partial t}_{|\,s,t=0} \log l(\theta, \Lambda_{s,t}) = -P_{\theta,\Lambda}\frac{\partial}{\partial t}_{|\,t=0} A_{\theta,\Lambda_{0,t}} g.$$

From the formula for $A_{\theta,\Lambda} g$ derived previously, the right side is seen to evaluate as

$$P_{\theta,\Lambda} e^{\theta z} \int_{[0,y]} gh\, d\Lambda = \int g(y) \Big(\mathrm{E}_{\theta,\Lambda} e^{\theta Z} 1_{Y \geq y} h(y) \Big)\, d\Lambda(y).$$

Because this must be equal to $\Lambda\big[(g)\,(A^*_{\theta,\Lambda} A_{\theta,\Lambda} h)\big]$, we easily read off the right expression for the information operator.

The formula for $A^*_{\theta,\Lambda} \ell_{\theta,\Lambda}$ can be found in a similar fashion, now starting from a two-dimensional model varying both θ and Λ.

Having found the information operator, we can now easily find an expression for the efficient score function. The information operator is simply multiplying a given function h by the function $M_{0,\theta,\Lambda}(y) := \mathrm{E}_{\theta,\Lambda} e^{\theta Z} 1_{Y \geq y}$. We need its inverse, which of course should be multiplying a given function with the function $1/M_{0,\theta,\Lambda}$ To make the inverse operation well-defined we assume that the function $M_{0,\theta,\Lambda}$ is strictly positive. A practically relevant assumption that ensures this is that a positive fraction of individuals is censored at some point: there exists a time τ such that $\mathrm{P}(C \geq \tau) = \mathrm{P}(C = \tau) > 0$ while $\mathrm{P}_{\theta,\Lambda}(T > \tau) > 0$. The efficient score function is then,

$$\begin{aligned} \tilde{\ell}_{\theta,\Lambda}(y,\delta,z) &= \ell_{\theta,\Lambda} - A_{\theta,\Lambda}\left(\frac{A^*_{\theta,\Lambda} \ell_{\theta,\Lambda}(y,\delta,z)}{M_{0,\theta,\Lambda}(y)} \right) \\ &= \delta\Big(z - \frac{M_{1,\theta,\Lambda}}{M_{0,\theta,\Lambda}}(y) \Big) - e^{\theta z} \int_{[0,y]} \Big(z - \frac{M_{1,\theta,\Lambda}}{M_{0,\theta,\Lambda}}(t) \Big)\, d\Lambda(t), \end{aligned}$$

where $M_{i,\theta,\Lambda}(y) = \mathrm{E}_{\theta,\Lambda} e^{\theta Z} Z^i 1_{Y \geq y}$ for $i = 0, 1$. The last step requires some clever algebra, but this simplification is irrelevant for our purpose. This formula was obtained in Begun et al. (1983).

3. MAIN RESULT

We assume the existence of an approximately least favorable submodel. That is, we assume that, for each parameter (θ, η), there exists a map, which we denote by $t \mapsto \boldsymbol{\eta}_t(\theta, \eta)$, from a fixed neighborhood of θ into the parameter set for η such that the map $t \mapsto \ell(t, \theta, \eta)(x)$ defined by

$$\ell(t, \theta, \eta)(x) = \log \mathrm{l}\big(t, \boldsymbol{\eta}_t(\theta, \eta)\big)(x)$$

is twice continuously differentiable, for all x. We denote the derivatives by $\dot{\ell}(t, \theta, \eta)(x)$ and $\ddot{\ell}(t, \theta, \eta)(x)$, respectively. The submodel with parameters $\big(t, \boldsymbol{\eta}_t(\theta, \eta)\big)$ should pass through (θ, η) at $t = \theta$:

$$\boldsymbol{\eta}_\theta(\theta, \eta) = \eta, \qquad \text{every } (\theta, \eta). \tag{3.1}$$

The second important structural requirement that should lead the construction of this submodel is that it be least favorable at (θ_0, η_0) for estimating θ in the sense that

$$\dot{\ell}(\theta_0, \theta_0, \eta_0) = \tilde{\ell}_{\theta_0, \eta_0}. \tag{3.2}$$

This means that the score function for the "parameter," t of the model with likelihood $\mathrm{l}\big(t, \boldsymbol{\eta}_t(\theta_0, \eta_0)\big)$ at $t = \theta_0$ is the efficient score function for θ. Note that we assume this only at the true parameter (θ_0, η_0). (In particular, not at every realization of the (profile) estimators, where (3.2) may be problematic due to the rough nature of these estimators.)

For every fixed θ, let $\hat{\eta}_\theta$ be a parameter at which the supremum in the definition of the profile likelihood is taken. (Thus $\mathrm{pl}_n(\theta) = \mathrm{l}_n(\theta, \hat{\eta}_\theta)$ for every θ.) Assume that for any random sequences $\tilde{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$

$$\hat{\eta}_{\tilde{\theta}_n} \xrightarrow{\mathrm{P}} \eta, \tag{3.3}$$

$$P_0 \dot{\ell}(\theta_0, \tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n}) = o_P\big(\|\tilde{\theta}_n - \theta_0\| + n^{-1/2}\big). \tag{3.4}$$

Conditions (3.1)–(3.4) are the structural conditions for the following theorem. Conditions (3.1)–(3.2) say how the least favorable submodel should be defined, leaving considerable freedom to adapt to examples. Condition (3.4) is a "no-bias condition" and is discussed below. Condition (3.3) implicitly assumes that we have a metric or topology defined on the set of nuisance parameters η. There is no need to be concrete about this topology, as it is only used as a technical go-between. However, we should fix a topology and use the same topology also for the other conditions of the following theorem.

In addition, we require continuity of the functions $\dot{\ell}(t, \theta, \eta)$ and $\ddot{\ell}(t, \theta, \eta)$ in (t, θ, η), and conditions that restrict the sizes of the collections of the functions $\dot{\ell}$ and $\ddot{\ell}$, so as to make them manageable. The latter conditions are expressed in the language of empirical processes. See e.g. Van der Vaart and Wellner (1996) for the definition, properties and examples of Glivenko-Cantelli and Donsker classes. In some examples, these tools simplify the proofs greatly while in other examples these tools appear to be unavoidable. Essentially they are used to show that the remainder terms of certain expansions are negligible.

Before stating the theorem, consider this set of conditions for the Cox model.

Example: Cox model, continued. We have previously computed the efficient score function for the Cox model. A convenient approximately least favorable submodel (with $\eta = \Lambda$) is defined by

$$d\Lambda_t(\theta, \Lambda) = \big(1 + (\theta - t)h_0\big)\, d\Lambda,$$

where h_0 is the least favorable direction at the true parameter (θ_0, Λ_0) defined by

$$h_0(y) = (A^*_{\theta_0,\Lambda_0} A_{\theta_0,\Lambda_0})^{-1} A^*_{\theta_0,\Lambda_0} \ell_{\theta_0,\Lambda_0}(y) = \frac{\mathrm{E}_0 1_{Y \geq y} Z e^{\theta_0 Z}}{\mathrm{E}_0 1_{Y \geq y} e^{\theta_0 Z}}.$$

This path is smooth in t and, more importantly, we obtain a smooth function of t if this submodel is substituted in the likelihood. Condition (3.1) is obviously satisfied, and by simple calculus we see that condition (3.2) is satisfied as well, in view of our work in Section 2. More precisely, this submodel is least favorable at the true parameter (θ_0, Λ_0), which is all that is required by (3.2). We could have constructed a submodel that is least favorable at every (θ, Λ) by replacing h_0 by $h_{\theta,\Lambda}$ as given in Section 2, but it is actually technically more convenient to follow the route chosen here.

There is one issue that must be discussed: it is not immediately clear that $\Lambda_t(\theta, \Lambda)$ is a cumulative hazard function for all combinations of (t, θ, Λ). We noted in Section 2 that this type of construction does yield a valid cumulative hazard function, at least for $\theta - t \approx 0$, which is enough, if h_0 is a bounded function. To ensure this we might for instance assume that the covariate Z ranges over a bounded interval. Then the explicit formula for h_0 shows that this is bounded by the upper bound on the range of Z. In other examples we shall have to deduce a similar property from the implicit definition of the information operator.

We discuss conditions (3.3)-(3.4) after stating the main theorem.

Theorem 3.1. *Let (3.1)–(3.4) be satisfied, and suppose that the functions* $(t, \theta, \eta) \mapsto \dot\ell(t, \theta, \eta)(x)$ *and* $(t, \theta, \eta) \mapsto \ddot\ell(t, \theta, \eta)(x)$ *are continuous at* $(\theta_0, \theta_0, \eta)$ *for* P_0*-almost every* x *(or in measure). Furthermore, suppose that there exists a neighborhood* V *of* $(\theta_0, \theta_0, \eta)$ *such that the class of functions* $\big\{\dot\ell(t, \theta, \eta)\colon (t, \theta, \eta) \in V\big\}$ *is* P_0*-Donsker with square-integrable envelope function, and such that the class of functions* $\big\{\ddot\ell(t, \theta, \eta)\colon (t, \theta, \eta) \in V\big\}$ *is* P_0*-Glivenko-Cantelli and is bounded in* $L_1(P_0)$*. Then (1.4) is satisfied.*

Proof. For simplicity assume that θ is one-dimensional. The general proof is similar. Since $\dot\ell(t, \theta, \eta) \to \tilde\ell_0$ as $(t, \theta, \eta) \to (\theta_0, \theta_0, \eta_0)$, and the functions $\dot\ell(t, \theta, \eta)$ are dominated by a square-integrable function, we have by the dominated convergence theorem, for every $(\tilde t, \tilde\theta, \tilde\eta) \xrightarrow{\mathrm{P}} (\theta_0, \theta_0, \eta_0)$, that $P_0\big(\dot\ell(\tilde t, \tilde\theta, \tilde\eta) - \tilde\ell_0\big)^2 \xrightarrow{\mathrm{P}} 0$. Similarly, we have that $P_0\ddot\ell(\tilde t, \tilde\theta, \tilde\eta) \xrightarrow{\mathrm{P}} P_0\ddot\ell(\theta_0, \theta_0, \eta_0)$. Since $t \mapsto \exp \ell(t, \theta_0, \eta_0)$ is proportional to a smooth one-dimensional submodel, its derivatives satisfy the usual identity

$$P_0\ddot\ell(\theta_0, \theta_0, \eta_0) = -P_0\dot\ell^2(\theta_0, \theta_0, \eta_0) = -\tilde I_0. \tag{3.5}$$

These facts, together with the empirical process conditions imply that, for every random sequence $(\tilde{t}, \tilde{\theta}, \tilde{\eta}) \xrightarrow{P} (\theta_0, \theta_0, \eta_0)$,

$$\mathbb{G}_n \dot{\ell}(\tilde{t}, \tilde{\theta}, \tilde{\eta})) = \mathbb{G}_n \tilde{\ell}_0 + o_P(1), \tag{3.6}$$

$$\mathbb{P}_n \ddot{\ell}(\tilde{t}, \tilde{\theta}, \tilde{\eta}) \xrightarrow{P} -\tilde{I}_0, \tag{3.7}$$

By (3.1) and the definition of $\hat{\eta}_\theta$,

$$\frac{1}{n}\Big(\log \mathrm{pl}_n(\tilde{\theta}) - \log \mathrm{pl}_n(\theta_0)\Big) = \mathbb{P}_n \log \mathrm{l}(\tilde{\theta}, \hat{\eta}_{\tilde{\theta}}) - \mathbb{P}_n \log \mathrm{l}(\theta_0, \hat{\eta}_{\theta_0})$$

$$\begin{cases} \geq \mathbb{P}_n \log \mathrm{l}\big(\tilde{\theta}, \boldsymbol{\eta}_{\tilde{\theta}}(\theta_0, \hat{\eta}_{\theta_0})\big) - \mathbb{P}_n \log \mathrm{l}\big(\theta_0, \boldsymbol{\eta}_{\theta_0}(\theta_0, \hat{\eta}_{\theta_0})\big) \\ \leq \mathbb{P}_n \log \mathrm{l}\big(\tilde{\theta}, \boldsymbol{\eta}_{\tilde{\theta}}(\tilde{\theta}, \hat{\eta}_{\tilde{\theta}})\big) - \mathbb{P}_n \log \mathrm{l}\big(\theta_0, \boldsymbol{\eta}_{\theta_0}(\tilde{\theta}, \hat{\eta}_{\tilde{\theta}})\big). \end{cases}$$

To obtain the lower bound, the first log likelihood is replaced by a lower value and the second log likelihood is left alone. For the upper bound, the reverse procedure is carried out. Both the upper and the lower bound are differences $\mathbb{P}_n \ell(\tilde{\theta}, \tilde{\psi}) - \mathbb{P}_n \ell(\theta_0, \tilde{\psi})$ (with $\tilde{\psi}$ equal to $(\theta_0, \hat{\eta}_{\theta_0})$ and $(\tilde{\theta}, \hat{\eta}_{\tilde{\theta}})$, respectively) in the log likelihood at two parameter values (θ_0 and $\tilde{\theta}$) of least favorable submodels. We apply a two-term Taylor expansion to these differences, leaving $\tilde{\psi}$ fixed, and show that the upper and lower bounds agree up to a term of the order $o_P\big(\|\tilde{\theta} - \theta_0\| + n^{-1/2}\big)^2$.

For $\tilde{t}$ a convex combination of $\tilde{\theta}$ and θ_0,

$$\mathbb{P}_n \ell(\tilde{\theta}, \tilde{\psi}) - \mathbb{P}_n \ell(\theta_0, \tilde{\psi}) = (\tilde{\theta} - \theta_0)^T \mathbb{P}_n \dot{\ell}(\theta_0, \tilde{\psi}) + \tfrac{1}{2}(\tilde{\theta} - \theta_0)^T \mathbb{P}_n \ddot{\ell}(\tilde{t}, \tilde{\psi})(\tilde{\theta} - \theta_0).$$

In the second term on the right we can replace $\mathbb{P}_n \ddot{\ell}(\tilde{t}, \tilde{\psi})$ by $-\tilde{I}_0$ at the cost of adding a $o_P\|\tilde{\theta} - \theta_0\|^2$-term, in view of (3.7). The first term on the right is equal to

$$\frac{1}{\sqrt{n}}(\tilde{\theta} - \theta_0)^T \mathbb{G}_n \dot{\ell}(\theta_0, \tilde{\psi}) + (\tilde{\theta} - \theta_0)^T P_0 \dot{\ell}(\theta_0, \tilde{\psi})$$

$$= \frac{1}{\sqrt{n}}(\tilde{\theta} - \theta_0)^T \mathbb{G}_n \tilde{\ell}_0 + (\tilde{\theta} - \theta_0) o_P\big(\|\tilde{\theta} - \theta_0\| + n^{-1/2}\big).$$

Combining the results of the last two displays, we obtain (1.4). ∎

An heuristic reason why the upper/lower bound argument is successful is as follows. If $\hat{\eta}_\theta$ achieves the supremum in (1.1), then the map $\theta \mapsto (\theta, \hat{\eta}_\theta)$ ought to be an estimator of a least favorable submodel for the estimation of θ (See Severini and Wong (1992)). In both the lower and upper bound, we differentiate the likelihood along approximations to the least favorable submodel. By definition, differentiation of the likelihood along the least favorable submodel (if the derivative exists) yields the efficient score function for θ. The efficient information matrix is the covariance matrix of the efficient score function, and, as usual, the expectation of minus the second derivative along this submodel should yield the same matrix. This explains relations (3.6)-(3.7).

An interpretation of condition (3.4) is as follows. Given the construction of the sub-model $t \mapsto \boldsymbol{\eta}_t(\theta, \eta)$, the derivative $\dot{\ell}(\theta_0, \tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n})$ can be considered an estimator of the efficient score function $\tilde{\ell}_0 = \dot{\ell}(\theta_0, \theta_0, \eta_0)$. Condition (3.4) requires that the difference in the mean of this estimator and the mean $P_0\tilde{\ell}_0 = 0$ is negligible to the order $o_P\big(\|\tilde{\theta}_n - \theta_0\| + n^{-1/2}\big)$. This "no-bias" condition has appeared in the literature in several forms and connections, and in that sense appears to be natural. For instance, in his construction of efficient one-step estimators, Klaassen (1987) requires estimators $\hat{\ell}_{n,\theta}$ for $\tilde{\ell}_0$ such that $P_0\hat{\ell}_{n,\theta_0} = o_P(n^{-1/2})$. Klaassen also shows that the existence of such estimators is necessary for asymptotically efficient estimators of $\hat{\theta}_n$ to exist, so that his "no bias" condition does not require too much. (Also cf. Bickel et al. (1993, p 395, equation (19)).) If the "no-bias" condition fails, then the Cramér-Rao bound, which is based on a local, differential analysis of the model near the true parameter, is too optimistic. Thus the "no-bias" condition can be viewed as the condition on the global model that implies that the local information calculations are not misleading.

The condition has appeared also in proofs of asymptotic normality of the maximum likelihood estimator, and of the asymptotic chi-squared distribution of the likelihood ratio statistic. For instance, Huang (1996) and Van der Vaart (1996) impose the condition $P_0\dot{\ell}(\theta_0, \hat{\theta}, \hat{\eta}_{\hat{\theta}}) = o_P(n^{-1/2})$, and in their work on the likelihood ratio statistic Murphy and Van der Vaart (1997) assume that $P_0\dot{\ell}(\theta_0, \theta_0, \hat{\eta}_{\theta_0}) = o_P(n^{-1/2})$. The present condition (3.4) is an extension of these conditions.

The verification of (3.4) may be easy due to special properties of the model, but may also require considerable effort. In the latter case, the verification usually boils down to a linearization of $P_0\dot{\ell}(\theta_0, \theta, \eta)$ in (θ, η) combined with establishing a rate of convergence of $\hat{\eta}_{\tilde{\theta}_n}$.

Here the expansion relative to θ does not cause difficulty, as the partial derivative of $P_0\dot{\ell}(\theta_0, \theta, \eta)$ relative to θ will vanish at (θ_0, η_0). If this is true, then the effect of replacing the first appearance of $\tilde{\theta}_n$ in (3.4) by θ_0 is of the order $o_P\big(\|\tilde{\theta}_n - \theta_0\|\big)$ and hence is negligible in the sense of this condition. That the derivative with respect to θ vanishes can be seen as follows. Because $\dot{\ell}(\theta, \theta, \eta)$ is a score function at the model indexed by (θ, η) we have, under the usual regularity conditions, $P_{\theta,\eta}\dot{\ell}(\theta, \theta, \eta) = 0$ for every (θ, η). Fixing η and differentiating this identity relative to θ gives, with $\ell_{\theta,\eta}$ the ordinary score function for θ,

$$P_{\theta,\eta}\ell_{\theta,\eta}\dot{\ell}(\theta, \theta, \eta) + P_{\theta,\eta}\ddot{\ell}(\theta, \theta, \eta) + \frac{\partial}{\partial t}_{|t=\theta} P_{\theta,\eta}\dot{\ell}(\theta, t, \eta) = 0.$$

Upon rearranging and evaluating this at $(\theta, \eta) = (\theta_0, \eta_0)$, we find

$$-\frac{\partial}{\partial t}_{|t=\theta_0} P_0\dot{\ell}(\theta_0, t, \eta_0) = P_0\dot{\ell}_0\dot{\ell}(\theta_0, \theta_0, \eta_0) + P_0\ddot{\ell}(\theta_0, \theta_0, \eta_0) = \tilde{I}_0 - \tilde{I}_0 = 0, \qquad (3.8)$$

where we use the fact that the efficient score function is a projection and (3.5) in the second last step.

Thus, under regularity conditions, condition (3.4) is equivalent to

$$P_0\dot{\ell}(\theta_0,\theta_0,\hat{\eta}_{\tilde{\theta}_n}) = o_P\big(\|\tilde{\theta}_n-\theta_0\|+n^{-1/2}\big). \tag{3.9}$$

This may be true trivially, or could be further analyzed by a Taylor expansion in η. Since η is infinite dimensional, a Taylor series may be nontrivial. There are at least two general schemes to verify (3.9). The first is to write $P_0\dot{\ell}(\theta_0,\theta_0,\eta)$ as

$$\begin{aligned} P_0\Big[\frac{p_0-p_{\theta_0,\eta}}{p_0}\big(\dot{\ell}(\theta_0,\theta_0,\eta)-\dot{\ell}(\theta_0,\theta_0,\eta_0)\big)\Big] \\ -P_0\dot{\ell}(\theta_0,\theta_0,\eta_0)\Big[\frac{p_{\theta_0,\eta}-p_0}{p_0}-A_0(\eta-\eta_0)\Big], \end{aligned} \tag{3.10}$$

where A_0 is the score operator for η at (θ_0,η_0) (and hence $P_0\tilde{\ell}_0A_0h=0$ for every h by the orthogonality property of the efficient score function). (This presumes that η belongs to a linear space and that A_0 is a derivative of $\log p_{\theta_0,\eta}$ relative to η; in specific examples the Taylor expansion may take a different form.) Now assume that the map $\eta\mapsto p_{\theta_0,\eta}$ is differentiable at η_0 relative to some norm $\|\cdot\|$ on the set of nuisance parameters. If also the difference $\big(\dot{\ell}(\theta_0,\theta_0,\eta)-\dot{\ell}(\theta_0,\theta_0,\eta_0)\big)$ tends to zero as $\eta\to\eta_0$, then the expression in the display should be of order $o_P\big(\|\eta-\eta_0\|\big)$. This suggests that (3.9) is certainly satisfied if

$$\|\hat{\eta}_{\tilde{\theta}_n}-\eta_0\| = O_P\big(\|\tilde{\theta}_n-\theta_0\|\big)+O_P\big(n^{-1/2}\big). \tag{3.11}$$

This should certainly be true in smooth parametric models (η finite dimensional), explaining why condition (3.4) does not show up in the analysis of classical parametric models. In cases where the nuisance parameter is not estimable at $\sqrt{n}$-rate, (3.11) is not satisfied. Then it may be still be possible to carry the expansion into its second order term. If $\eta\mapsto p_{\theta_0,\eta}$ is twice differentiable and $\eta\mapsto\dot{\ell}(\theta_0,\theta_0,\eta)$ is differentiable at η_0, then the expression in (3.10) ought to be of the order $O_P\big(\|\eta-\eta_0\|^2\big)$ and it is still sufficient to have

$$\|\hat{\eta}_{\tilde{\theta}_n}-\eta_0\| = O_P\big(\|\tilde{\theta}_n-\theta_0\|\big)+o_P\big(n^{-1/4}\big).$$

The above method is used in Section 4.1. Special properties of the model may allow to take the expansion even further or make part of the expressions in the display zero, and then a slower rate of convergence of the nuisance parameter may be sufficient. The extreme case that the expression in the left side of (3.9) is identically zero occurs, among others, in certain mixture models. (See Section 4.4 for an example.)

A second general scheme is useful when η is a measure, and $P_0\dot{\ell}(\theta_0,\theta_0,\eta)$ can be written as $P_0\big(\ell_{\theta_0,\eta}-A_{\theta_0,\eta}\hat{h}\big)$ ($\hat{h}$ is a approximation to the least favorable direction). Let η_t be a path in the parameter space so that differentiation of $\log p_{\theta_0,\eta_t}$ at $t=\theta_0$ yields A_0h. Then

$$\frac{\partial}{\partial t}P_0\dot{\ell}(\theta_0,\theta_0,\eta_t)|_{t=\theta_0} = -P_0\ell_{\theta_0,\eta_0}A_0h+P_0A_0\hat{h}A_0h$$

since the expectation of the derivative of the θ score with respect to η is minus the covariance of the θ score with the η score and the expectation of the derivative of the η score with respect to η is minus the variance of the η score. Furthermore by definition of the adjoint,

$$\frac{\partial}{\partial t} P_0 \dot{\ell}(\theta_0, \theta_0, \eta_t)|_{t=\theta_0} = -\int (A_0^* \ell_{\theta_0,\eta_0}) h d\eta + \int (A_0^* A_0 \hat{h}) h d\eta.$$

Thus since $P_0 \dot{\ell}(\theta_0, \theta_0, \eta_0) = 0$, the first term in a linearization in η of $P_0 \dot{\ell}(\theta_0, \theta_0, \eta)$ is zero and the higher order terms should mirror the previous display,

$$-\int A_0^* \ell_{\theta_0,\eta_0} d(\eta - \eta_0) + \int A_0^* A_0 \hat{h} d(\eta - \eta_0) + o_P(||\eta - \eta_0||).$$

See Van der Vaart (1998, pg. 421) for a discussion of this. Recall $\hat{h}$ should be close to h_{θ_0,η_0}. Insert in the place of $\hat{h}$, the formula for h_{θ_0,η_0} as given in Section 2. This yields $P_0 \dot{\ell}(\theta_0, \theta_0, \eta) = o_P(||\eta - \eta_0||)$. Thus combining this with a rate of convergence for $\eta = \hat{\eta}_{\tilde{\theta}_n}$ as given in (3.11) will yield (3.9). (See Section 4.2 and 4.3 for examples.)

Another example of interest is the partially linear regression model considered by Donald and Newey (1994). Here the nuisance parameter is partitioned in an unknown regression function and an unknown error density and the second derivative can be considered a (2×2)-matrix of operators. In this matrix the diagonal elements (which normally would yield positive-definite terms) vanish and there is cancellation in the mean of the cross product term. A simple upper bound of this cross product (by Cauchy-Schwarz) by a product of norms is overly pessimistic. As shown by Donald and Newey (1994) in this example much less than a $n^{-1/4}$-rate on the nuisance parameter suffices. It appears difficult to capture such special situations in a general result, whence we have left the "no-bias" condition as a basic condition.

Methods to deduce rates of convergence of maximum likelihood estimators of infinitely dimensional parameters such as (3.11) were recently developed by Van de Geer (1993, 1995), Wong and Shen (1995), Birgé and Massart (1993, 1997), and were adapted to incorporate nuisance parameters by Murphy and Van der Vaart (1997c). Generally, these authors show how the rate of convergence can be expressed in the entropy of the model. We shall not describe these methods in this paper.

Example: Cox model, continued. Substituting $\theta = t$ and $\Lambda = \Lambda_t(\theta, \Lambda)$ in the Cox likelihood (as in (1.3) for $n = 1$) and differentiating with respect to t gives

$$\begin{aligned}\dot{\ell}(t, \theta, \Lambda)(x) &= \ell_{t,\Lambda_t(\theta,\Lambda)} - A_{t,\Lambda_t(\theta,\Lambda)} h_0(x) \\ &= \delta z - z e^{tz} \Lambda_t(\theta, \Lambda)(y) - \delta h_0(y) - e^{tz} \int_{[0,y]} h_0 \, d\Lambda_t(\theta, \Lambda).\end{aligned}$$

These are well-behaved functions. As functions of y they are bounded and of bounded variation and uniformly so if Λ itself varies over a set of bounded cumulative hazard functions. This is a standard example of a Donsker class. The dependence on (z, δ) is simple and t is

one-dimensional. Given the many known results on empirical processes (cf. Van der Vaart and Wellner (1996)) we can put these facts without further work together to conclude that the set of functions $\dot{\ell}(t,\theta,\Lambda)$ forms a Donsker class as well. Similarly, we can compute the functions $\ddot{\ell}(t,\theta,\Lambda)$ and conclude that they form a Glivenko-Cantelli class. The functions are rather smooth in (t,θ,Λ), where we may, for instance, use the uniform norm on Λ. Therefore, the regularity conditions of Theorem 3.1 are easy to verify using standard theory of empirical processes.

In order to verify (3.4) it suffices to check regularity conditions and (3.9). For the more interesting part, the verification of (3.9), write

$$\begin{aligned}
P_0\dot{\ell}(\theta_0,\theta_0,\Lambda) &= P_0(\ell_{\theta_0,\Lambda} - A_{\theta_0,\Lambda}h_0) \\
&= P_0\big[(\ell_{\theta_0,\Lambda} - A_{\theta_0,\Lambda}h_0) - (\ell_{\theta_0,\Lambda_0} - A_{\theta_0,\Lambda_0}h_0)\big] \\
&= -P_0\left[ze^{\theta_0 z}(\Lambda-\Lambda_0)(y) - e^{\theta_0 z}\int_{[0,y]} h_0\, d(\Lambda-\Lambda_0)\right] \\
&= -\int (M_{1,\theta_0,\Lambda_0} - M_{0,\theta_0,\Lambda_0}h_0)\, d\Lambda_0.
\end{aligned}$$

The right side vanishes by the definition of the least favorable direction h_0. Therefore, the "no bias" condition (3.9) is satisfied in the strongest possible sense. We could have inferred this from the linearity of the score functions in Λ (even though the likelihood is not linear in Λ). Again, the Cox model is as nice as it can be; in other cases we do find a remainder term, and need to establish some rate of convergence.

The only condition that must still be verified is the consistency condition (3.3). This verification is not a trivial matter, but most standard methods of proving consistency of estimators can be adapted to yield the desired result, where we can define the convergence relative to the uniform distance. To be able to apply the results of Section 1 and draw conclusions on the profile likelihood (which is the partial likelihood), we must of course also establish the consistency of the maximum likelihood estimator itself. This requires different methods.

4. EXAMPLES

In this section we provide a number of examples with the aims of:

- showing the variation in definitions of likelihoods;
- showing the different ways that least favorable submodels can be defined;
- indicating the different techniques that can be applied to verify the no-bias condition (3.4).

These examples have been the subject of separate papers, by various authors, one example per paper (or even multiple papers). Most of these papers focused on the asymptotic normality of the maximum likelihood estimator $\hat{\theta}$, or considered computational issues. To justify the quadratic expansion of the present paper, we need to extend the methods used in these papers, but since the basic techniques remain the same, we shall refer to the literature for most of the technical details.

4.1. The Proportional Hazards Model for Current Status Data

It turns out that the efficient score can be easily calculated via a least squares projection as follows. The score function for θ takes the form

$$\ell_{\theta,\eta}(x) = z\eta(y)Q(x;\theta,\eta),$$

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

$$Q(x;\theta,\eta) = e^{\theta^T z}\Big[\delta\frac{e^{-e^{\theta^T z}\eta(y)}}{1-e^{-e^{\theta^T z}\eta(y)}} - (1-\delta)\Big].$$

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

$$A_{\theta,\eta}h(x) = h(y)Q(x;\theta,\eta).$$

For example, for every nondecreasing, nonnegative function h, the submodel $\eta_t = \eta + th$ is well defined if t is positive and yields a (one-sided) derivative h at $t = 0$. Thus the preceding display gives a (one-sided) score for η at least for all h of this type. The linear span of these functions contains $A_{\theta,\eta}h$ for all bounded functions h of bounded variation. The efficient score function for θ is defined as $\tilde{\ell}_{\theta,\eta} = \ell_{\theta,\eta} - A_{\theta,\eta}h_{\theta,\eta}$ for the vector of functions $h_{\theta,\eta}$ minimizing the distance $P_{\theta\eta}\|\ell_{\theta,\eta} - A_{\theta,\eta}h\|^2$. In view of the similar structure of the scores for θ and η, this is a weighted least squares problem with weight function $Q(x;\theta,\eta)$. The solution at the true parameters is given by the vector-valued function

$$h_0(Y) = \eta_0(Y)h_{00}(Y) = \eta_0(Y)\frac{\mathrm{E}_{\theta_0\eta_0}\big(ZQ^2(X;\theta_0,\eta_0)\big|\,Y\big)}{\mathrm{E}_{\theta_0\eta_0}\big(Q^2(X;\theta_0,\eta_0)\big|\,Y\big)}. \tag{4.1}$$

As the formula shows (and as follows from the nature of the minimization problem), the vector of functions $h_0(y)$ is unique only up to null sets for the distribution of Y. We shall

assume that (under the true parameters) there exists a version of the conditional expectation that is differentiable with bounded derivative.

A first guess at $\boldsymbol{\eta}_t(\theta,\eta)$ is $\eta + (\theta - t)h_0$ where h_0 is the least favorable direction; this, however, will not work as $\boldsymbol{\eta}_t(\theta,\eta)$ must lie in the parameter space of cumulative hazard functions. A modification of this guess is, for t a vector in $\mathbb{R}^d$ and ϕ a fixed function that is approximately the identity,

$$\begin{aligned}\boldsymbol{\eta}_t(\theta,\eta) &= \eta + (\theta - t)^T \phi(\eta)(h_{00} \circ \eta_0^{-1} \circ \eta) \\ \ell(t,\theta,\eta) &= \log \mathrm{l}(t, \boldsymbol{\eta}_t(\theta,\eta)).\end{aligned}$$

The function $\boldsymbol{\eta}_t(\theta,\eta)$ is essentially η plus a perturbation in the least favorable direction, h_0, but its definition is somewhat complicated in order to ensure that $\boldsymbol{\eta}_t(\theta,\eta)$ really defines a cumulative hazard function within our parameter space, at least for t that are sufficiently close to θ. First, the construction using $h_{00} \circ \eta_0^{-1} \circ \eta$, rather than h_{00}, ensures that the perturbation that is added to η is absolutely continuous with respect to η; otherwise $\boldsymbol{\eta}_t(\theta,\eta)$ would not be a nondecreasing function. Second, the function ϕ serves a similar purpose. As long as $\phi(\eta_0) = \eta_0$ on the support of the observation times Y, the model $t \mapsto \boldsymbol{\eta}_t(\theta_0,\eta_0)$ will be least favorable at (θ_0,η_0). Below we assume that the parameter space for η is the space of all cumulative hazard functions, bounded above by a constant M and below by zero. Thus $\boldsymbol{\eta}_t(\theta,\eta)$ must satisfy the same constraint. Essentially $\phi(u)$ is smooth approximation to the indicator variable, $I\{0 < u < M\}$. (See below for the precise properties ϕ should have.)

Note that $\boldsymbol{\eta}_\theta(\theta,\eta) = \eta$ and thus (3.1) is satisfied. Furthermore (3.2) is also satisfied since, $\dot{\ell}(t,\theta,\eta) = \big(z\eta(y) - \phi(\eta(y))(h_{00} \circ \eta_0^{-1} \circ \eta)(y)\big)Q(x;t,\boldsymbol{\eta}_t(\theta,\eta))$, evaluated at $t = \theta = \theta_0$ and $\eta = \eta_0$ is the efficient score, $(z\eta_0(y) - h_0(y))Q(x;\theta_0,\eta_0)$.

Precise arguments reveal that the conditions of this paper can be checked under the following assumptions. We restrict the parameter set for θ to a compact in $\mathbb{R}^d$ and assume that the true parameter θ_0 is an interior point. Furthermore, we restrict the parameter set for η to the set of all cumulative hazard functions on an interval $[0,\tau]$ with $\eta(\tau) \le M$ for a given (large) constant M. The observations times Y are in an interval $[\sigma,\tau]$. The true parameter η_0 satisfies $\eta_0(\sigma-) > 0$ and $\eta_0(\tau) < M$, and is continuously differentiable with derivative bounded away from zero on $[\sigma,\tau]$. The covariate vector Z is bounded and $\mathrm{E}\,\mathrm{cov}(Z|\,Y)$ is positive definite. Finally, we assume that the function h_0 given by (4.1) has a version which is differentiable with a bounded derivative on $[\sigma,\tau]$. Then we can take the function $\phi\colon [0,M] \mapsto [0,M]$ to be any fixed function such that $\phi(u) = u$ on the interval $[\eta_0(\sigma),\eta_0(\tau)]$, such that the function $u \mapsto \phi(u)/u$ is Lipschitz and such that $\phi(u) \le c(u \wedge (M-u))$ for a sufficiently large constant c depending on (θ_0,η_0) only. (By our assumption that $[\eta_0(\sigma),\eta_0(\tau)] \subset (0,M)$ such a function exists.)

In this model, the cumulative hazard functions is not estimable at $\sqrt{n}$-rate. However, using entropy methods, Murphy and Van der Vaart (1997), extending earlier results of

Huang (1996), show that

$$\int (\hat{\eta}_{\tilde{\theta}}(y) - \eta_0(y))^2 \, dF^Y(y) = O_P\big(\|\tilde{\theta} - \theta_0\|^2 + n^{-2/3}\big). \tag{4.2}$$

Thus, (3.3) holds with the $L_2(F^Y)$ norm on the nuisance parameter space. This rate of convergence helps in the verification of equation (3.4). Note that $P_0 \frac{\partial}{\partial \theta} \dot{\ell}(\theta_0, \theta, \eta)$ evaluated at $(\theta, \eta) = (\theta_0, \eta_0)$ is zero. Thus,

$$P_0 \dot{\ell}(\theta_0; \theta, \eta) - \dot{\ell}(\theta_0; \theta_0, \eta) = P_0 \left(\frac{\partial}{\partial \theta} \dot{\ell}(\theta_0, \theta^*, \eta) - \frac{\partial}{\partial \theta} \dot{\ell}(\theta_0, \theta_0, \eta_0) \right) (\theta - \theta_0)$$

for θ^* an intermediate point between θ and θ_0. Evaluating the above equation at $\eta = \hat{\eta}_{\tilde{\theta}}$ and $\theta = \tilde{\theta}$, and using the rate of convergence for $\hat{\eta}_{\tilde{\theta}}$, we see that we need only verify (3.9). For instance, we can verify (3.9) by the decomposition as in (3.10) upon noting that, for some constant C and every x (under the above regularity conditions),

$$\begin{aligned} \big| p_0(x) - p_{\theta_0, \eta_{\theta_0}(\theta_0, \eta)}(x) \big| &\le C |\eta - \eta_0|(y), \\ \big| \dot{\ell}(\theta_0, \theta_0, \eta)(x) - \dot{\ell}(\theta_0, \theta_0, \eta_0)(x) \big| &\le C |\eta - \eta_0|(y) \text{ and} \\ \big| p_{\theta_0, \eta}(x) - p_0(x) - A_0(\eta - \eta_0)(x) p_0(x) \big| &\le C |\eta - \eta_0|^2(y). \end{aligned}$$

Since the expressions on the left for a fixed x depend on η only through $\eta(y)$, these inequalities follow from ordinary Taylor expansions and uniform bounds on the first and second derivatives. Thus by the decomposition as in (3.10) and taking expectations, the rate of convergence on $\hat{\eta}_{\tilde{\theta}}$ given in the preceding display readily translates into a rate of convergence of the "bias" term, of the order $O_P\big(\|\tilde{\theta} - \theta_0\|^2\big) + O_P(n^{-2/3})$, better than needed.

4.2. Case-Control Studies with a Missing Covariate

Roeder, Carroll and Lindsay (1996) and Murphy and Van der Vaart (1997b) consider both a prospective and retrospective (or case-control) model. In the prospective model we observe two independent random samples of sizes n_C and n_R from the distributions of (Y_C, Z_C) and Y_R, respectively. In the terminology of Roeder, Carroll and Lindsay (1996), the covariate Z in a full observation is a "golden standard", but, in view of the costs of measurement, for a selection of observations only the "surrogate covariate" W_R in Y_R is available. In their example W is the natural logarithm of total cholesterol, Z is the natural logarithm of LDL cholesterol, and we are interested in heart disease $D = 1$.

We shall concentrate on the regression coefficient, β, considering $\theta_2 = (\alpha_0, \alpha_1, \sigma)$ and η as nuisance parameters. (Thus, the parameter θ in the general results should be replaced by β throughout this section.) The distribution η of the regression variable is assumed to have support in an interval $\mathcal{Z} \subset \mathbb{R}$.

We shall consider the situation that the number of complete and reduced observations are of comparable magnitude. More precisely, our theorem applies to the situation that the fraction n_C/n_R is bounded away from 0 and ∞. For simplicity of notation, we shall

henceforth assume that $n_C = n_R$. Then the observations can be paired and the observations in the prospective model can be summarized as n i.i.d. copies of $X = (Y_C, Z_C, Y_R)$ from the density

$$x = (y_C, z_C, y_R) \mapsto p_\theta(y_C|\, z_C)\, d\eta(z_C) \int p_\theta(y_R|\, z)\, d\eta(z) =: p_\theta(y_C|\, z_C)\, d\eta(z_C)\, p_\theta(y_R|\, \eta).$$

Here we denote the complete sample components by $Y_C = (D_C, W_C)$ and Z_C and the reduced sample components by $Y_R = (D_R, W_R)$. In the complete sample part of the likelihood we use an empirical likelihood with $\eta\{z\}$, the measure of the point $\{z\}$,

$$l(\theta, \eta)(x) = p_\theta(y_C|\, z_C)\, \eta\{z_C\} \int p_\theta(y_R|\, z)\, d\eta(z).$$

Note that the "likelihood" is a density if η is assumed to be discrete; however we shall use the above even if it is known that η must be continuous.

There is no similar notational device to fit the retrospective version of the model in the i.i.d. set-up of this paper. Therefore, we shall restrict ourselves to the prospective model as given in the preceding paragraph. However, since the profile likelihood for β in the prospective and retrospective models are algebraically identical, the arguments can be extended to the retrospective model.

Assuming that η_0 is nondegenerate, Murphy and Van der Vaart (1997b) show that the maximum likelihood estimator $(\hat\theta, \hat\eta)$ is asymptotically normal. Note that the assumption of a known support means that in the maximum likelihood estimation, η is constrained to have support contained in $\mathcal{Z}$. Here we shall verify that the conditions of Theorem 3.1 are satisfied, so that the profile likelihood function is approximated by a quadratic in β.

To define a least favorable submodel, we first calculate scores for θ and η. The score function $\ell_{\theta,\eta}$ for θ is given by

$$\ell_{\theta,\eta}(y_C, z_C, y_R) = \ell_\theta(y_C|z_C) + E_{\theta,\eta}[\ell_\theta(Y_R|Z_R)|Y_C = y_c, Z_C = z_C, Y_R = y_R]$$

where $\ell_\theta(y|z) = \partial/\partial\theta \log p_\theta(y|\, z)$. Furthermore, defining $\eta_t(\cdot) = \eta(\cdot) + t\int^{\cdot} h\, d\eta$ for h a bounded function satisfying $\int h\, d\eta = 0$, we can differentiate $l(\theta, \eta_t)$ at $t = 0$ to form the score for η in the direction h,

$$A_{\theta,\eta} h(x) = h(z_C) + E_{\theta,\eta}[h(Z_R)|Y_C = y_c, Z_C = z_C, Y_R = y_R].$$

Using the theory discussed in Section 2, the Hilbert space adjoint $A^*_{\theta,\eta}$ of this operator is given by

$$A^*_{\theta,\eta} g(z_C, z) = E_{\theta,\eta}[g(Y_R, Z_C)|Z_C = z_c, Z_R = z].$$

In Section 8 of Murphy and Van der Vaart (1997b) it is shown that $A^*_{\theta_0,\eta_0} A_{\theta_0,\eta_0}$ is continuously invertible on the space of Lipschitz continuous functions. An explicit expression for the inverse operator $(A^*_{\theta,\eta} A_{\theta,\eta})^{-1}$ is unknown (and it is unlikely that there is one). The assumptions stated above imply that $A^*_{\theta_0,\eta_0} \ell_{\theta_0,\eta_0}$ is Lipschitz continuous. Thus, as in Section

2, the least favorable direction $h_{\theta,\eta}$ for the estimation of θ in the presence of the unknown η is given by

$$h_{\theta,\eta} = (A^*_{\theta,\eta} A_{\theta,\eta})^{-1} A^*_{\theta,\eta} \ell_{\theta,\eta}$$

and the efficient information matrix for θ when η is unknown is

$$\tilde{I}_{\theta,\eta} = P_{\theta,\eta} \ell_{\theta,\eta} \ell^T_{\theta,\eta} - P_{\theta,\eta} \big(A_{\theta,\eta} (A^*_{\theta,\eta} A_{\theta,\eta})^{-1} A^*_{\theta,\eta} \ell_{\theta,\eta} \big) \ell^T_{\theta,\eta}.$$

Since $A^*_{\theta_0,\eta_0} \ell_{\theta_0,\eta_0}$ is Lipschitz continuous, the function h_0 is also bounded and Lipschitz, so that it can be used to define the submodel η_t as given previously.

The efficient information for β is obtained by inverting the efficient information matrix for θ and taking the $(1,1)$-element. This corresponds to a further projection of the efficient score functions for θ. Partition θ into $\theta = (\beta, \theta_2)$, where $\theta_2 = (\alpha_0, \alpha_1, \gamma, \sigma^2)$, and partition the efficient information matrix $\tilde{I}_0$ for θ in four submatrices, accordingly. The efficient score function for β is the first coordinate of $\tilde{\ell}_{\theta,\eta}$ minus its projection on the remaining coordinates of $\tilde{\ell}_{\theta,\eta}$. Thus we define

$$\begin{aligned} a_0^T &= \big(1, -\tilde{I}_{0,12} (\tilde{I}_{0,22})^{-1}\big), \\ d\boldsymbol{\eta}_t(\theta,\eta) &= \big(1 + (\beta - t) a_0^T (h_0 - \eta h_0)\big)\, d\eta, \\ \boldsymbol{\theta}_t(\theta,\eta) &= \theta - (\beta - t) a_0, \end{aligned}$$

where $\eta h = \int h \, d\eta$ and $\eta_0 h_0 = 0$. Since h_0 is bounded, $\boldsymbol{\eta}_t(\theta,\eta)$ has a positive density with respect to η for every sufficiently small $|\beta - t|$ and hence defines an element of the parameter set for η. Now we use the least favorable path

$$t \mapsto \big(\boldsymbol{\theta}_t(\theta,\eta)_2, \boldsymbol{\eta}_t(\theta,\eta)\big)$$

in the parameter space for the nuisance parameter (θ_2, η). This leads to $\ell(t, \beta, \theta_2, \eta) = \log l\big(\boldsymbol{\theta}_t(\theta,\eta), \boldsymbol{\eta}_t(\theta,\eta)\big)$. Since $d\boldsymbol{\eta}_\beta(\theta,\eta) = d\eta$ and $\boldsymbol{\theta}_\beta(\theta,\eta)_2 = \theta_2$, (3.1) is satisfied. Furthermore, this submodel is least favorable at (θ_0, η_0) in that (3.2) is satisfied in the form

$$\frac{\partial}{\partial t}_{|t=\beta_0} \ell(t, \beta_0, (\theta_0)_2, \eta_0) = a_0^T \tilde{\ell}_0,$$

where

$$\tilde{\ell}_0(x) = \ell_{\theta_0,\eta_0}(x) - A_{\theta_0,\eta_0} h_0(x).$$

The function $\tilde{\ell}_0$ is the efficient influence function for the parameter θ in the presence of the nuisance parameter η, while the function $a_0^T \tilde{\ell}_0$ is the efficient score function for β in the presence of the nuisance parameter (θ_2, η), both evaluated at (θ_0, η_0). (Also cf. Section 7 of Murphy and Van der Vaart (1997b).)

In this model, the distribution function is estimable at $\sqrt{n}$-rate relative to a variety of norms. For instance, for $\mathcal{H}$ the set of functions $h\colon \mathcal{Z} \mapsto \mathbb{R}$ that are uniformly bounded by

1 and uniformly Lipschitz with Lipschitz constant 1, Murphy and Van der Vaart (1997b) show that

$$\sup_{h\in\mathcal{H}}\left|\int h(z)\,d(\hat{\eta}_{\tilde{\beta}}-\eta_0)(z)\right|+\|\hat{\theta}_{\tilde{\beta}}-\theta_0\|=O_P\big(|\tilde{\beta}-\beta_0|+n^{-1/2}\big),$$

for any $\tilde{\beta}$ converging in probability to β_0. This satisfies (3.3) for the bounded Lipshitz norm on distributions.

To verify (3.4) we next study the dependence of $P_0\dot{\ell}(\beta_0,\beta,\theta_2,\eta)$ on (β,θ_2), but this is straightforward, because it only involves Euclidean parameters. (Cf. (3.8).) Next note that $P_0\dot{\ell}(\beta_0,\theta_0,\eta)=P_0a_0^T\ell_{\theta_0,\eta}-A_{\theta_0,\eta}a_0^T(h_0-\eta h_0)$. Furthermore, Murphy and Van der Vaart (Section 6, 1997b) show that

$$P_0\ell_{\theta_0,\eta}-A_{\theta_0,\eta}(h_0-\eta h_0)=-\int A_0^*\ell_{\theta_0,\eta_0}d(\eta-\eta_0)+\int A_0^*A_0h_0\,d(\eta-\eta_0)+o_P(||\eta-\eta_0||).$$

Substituting in for h_0, we see that the first term is zero and thus the rate of convergence for $\hat{\eta}_{\tilde{\beta}}$ given above suffices for (3.9).

4.3. Shared Gamma Frailty Model

For a given group the observations are $X=\{N_j(t),Y_j(t),Z_j(t),\ j=1,\ldots,J,\ 0\le t\le\tau\}$, where $N_j, j=1,\ldots,J$ count the events for each of the, at most J, subjects in the group and form a multivariate counting process. The observation interval, $[0,\tau]$ is assumed finite. The censoring process, Y_j is nonincreasing, is left continuous, with right hand limits (caglad) and takes values in $\{0,1\}$. The jth subject can be observed to experience an event at time t only if $Y_j(t)=1$. The d-dimensional covariate process, Z_j is also caglad and is assumed to be a.s. uniformly bounded and of uniformly bounded variation. Let $\mathcal{F}_t^{com}$ be the "complete data" sigma field generated by $\{G,N_j(s),Y_j(s),Z_j(s),\ j=1,\ldots,J,\ 0\le s\le t\}$. Given G and Z the intensity of N_j with respect to the sequence of complete data sigma fields, indexed by t, is assumed to follow a proportional hazards model,

$$G\exp(\beta^TZ_j(\cdot))Y_j(\cdot)\dot{\eta}(\cdot)$$

where β is a d-dimensional vector of regression coefficients and $\dot{\eta}$ is the baseline hazard function. The unobserved frailty, G, follows a gamma distribution with mean one and variance σ. We assume that given G and Z, the censoring is independent and furthermore both the censoring and covariate processes are noninformative of G. See Nielsen et al. (1992) for more discussion concerning the censoring. The terms independent and noninformative are discussed by Andersen et al. (1993). The key consequence of assumed noninformative censoring and covariate processes is that the likelihood for one group's complete data (X,G), is a function of G only through the partial likelihood,

$$\prod_{j=1}^{J}\left\{\prod_{t\le\tau}\big(G\exp(\beta^TZ_j(t))Y_j(t)\dot{\eta}(t)\big)^{\Delta N_j(t)}\,e^{-G\int_0^\tau\exp(\beta^TZ_j(s))Y_j(s)\dot{\eta}(s)\,ds}\right\}p(G;\sigma),$$

where $p(\cdot;\sigma)$ is a gamma density corresponding to a mean of one and variance of σ.

Since G is not observed, we integrate the compete data partial likelihood over G to form the observed data partial likelihood,

$$\frac{\prod_{j=1}^{J}\prod_{t\le\tau}\left((1+\sigma N_{\cdot}(t-))\exp(\beta^T Z_j(t))Y_j(t)\dot{\eta}(t)\right)^{\Delta N_j(t)}}{\left(1+\sigma\int_0^\tau \tilde{Y}_{\cdot}^{\beta}(s)\dot{\eta}(s)ds\right)^{1/\sigma+N_{\cdot}(\tau)}},$$

where $\tilde{Y}_{\cdot}^{\beta}=\sum_j \exp(\beta^T Z_j)Y_j$ and $N_{\cdot}=\sum_j N_j$. Also given X, G has a gamma distribution with

$$\mathrm{E}(G|\,X)=\frac{1+\sigma N_{\cdot}(\tau)}{1+\sigma\int_0^\tau \tilde{Y}_{\cdot}^{\beta}\,d\eta},\ \text{and}\ \ \mathrm{var}(G|\,X)\ =\ \sigma\frac{1+\sigma N_{\cdot}(\tau)}{\left(1+\sigma\int_0^\tau \tilde{Y}_{\cdot}^{\beta}\,d\eta\right)^2}.$$

This model has been considered previously by Nielsen et al. (1992), Murphy (1994, 1995) and Parner (1998).

We consider estimation of the parameters, $\theta=(\beta,\sigma)$, $\eta(\cdot)=\int_0^{\cdot}\dot{\eta}(s)ds$ based on observation of n i.i.d. copies of X. The parameter space for η is the set of nondecreasing, nonnegative functions on $[0,\tau]$ with $\eta(0)=0$, the parameter space for β is $\mathcal{B}$, a compact subset of $\mathbb{R}^d$, and lastly the parameter space for σ is $[-\epsilon/\eta(\tau),M]$ for known M and $\epsilon=(J\max_{\beta,z}e^{\beta^T z})^{-1}$. Thus the parameter space for σ is not the same for every value of η.

The observed data partial likelihood has no maximizer. A convenient empirical likelihood is given by replacing $\dot{\eta}(t)$ by $\eta\{t\}$,

$$\mathrm{l}(\theta,\eta)(X)=\frac{\prod_{j=1}^{J}\prod_{t\le\tau}\left((1+\sigma N_{\cdot}(t-))\exp(\beta^T Z_j(t))Y_j(t)\eta\{t\}\right)^{\Delta N_j(t)}}{\left(1+\sigma\int_0^\tau \tilde{Y}_{\cdot}^{\beta}(s)d\eta(s)\right)^{1/\sigma+N_{\cdot}(\tau)}}. \tag{4.3}$$

This is not the only possible extension; see Andersen et. al. (1993) and Murphy (1995). The conditions of this paper may be checked under the assumptions, that $\sigma_0\in[0,M)$, β_0 belongs to the interior of the set $\mathcal{B}$, $\eta_0(\tau)<\infty$ and η_0 is strictly increasing on $[0,\tau]$. Furthermore assume that $P(Y_{\cdot}(\tau)>0)>0$, $P(N_{\cdot}(\tau)\ge 2)>0$ and if for a vector, c and a scalar, c_0, $P(c^T Z_j(0+)Y_j(0)=c_0 Y_j(0), j=1,\ldots,J)=1$ then $c=0$.

The maximum likelihood estimator $(\hat{\theta},\hat{\eta})$ for this model was shown to be asymptotically consistent (with respect to the uniform norm) and normal by Parner (1998) in the more general correlated frailty model but under similar conditions.

The score function for θ, the derivative of the log empirical likelihood with respect to θ, is given by $\ell_{\theta,\eta}=(\ell_{1,\theta,\eta}^T,\ell_{2,\theta,\eta})^T$ where

$$\ell_{1,\theta,\eta}(X)=\sum_j\int Z_j\,dN_j-\frac{1+\sigma N_{\cdot}(\tau)}{1+\sigma\int_0^\tau \tilde{Y}_{\cdot}^{\beta}\,d\eta}\int_0^\tau\sum_j e^{\beta^T Z_j}Y_j Z_j\,d\eta$$

and

$$\ell_{2,\theta,\eta}(X) = \int_0^\tau \frac{N_.(s-)}{1+\sigma N_.(s-)}\, dN_.(s)$$
$$+ \sigma^{-2}\left(\log(1+\sigma\int_0^\tau \tilde{Y}_.^\beta\, d\eta) - \frac{1+\sigma N_.(\tau)}{1+\sigma\int_0^\tau \tilde{Y}_.^\beta\, d\eta}\sigma\int_0^\tau \tilde{Y}_.^\beta\, d\eta\right).$$

The score operator for η in the direction, h (a bounded function) is

$$A_{\theta,\eta}h(X) = \int h\, dN_. - \frac{1+\sigma N_.(\tau)}{1+\sigma\int_0^\tau \tilde{Y}_.^\beta\, d\eta}\int_0^\tau \tilde{Y}_.^\beta h\, d\eta.$$

Since the adjoint of $A_{\theta,\eta}$ satisfies the equality,

$$\begin{aligned} P_{\theta,\eta}(A_{\theta,\eta}h)^2 &= \langle A_{\theta,\eta}h,\ A_{\theta,\eta}h\rangle_{P_{\theta,\eta}} \\ &= \langle h, A^*_{\theta,\eta}A_{\theta,\eta}h\rangle_\eta \end{aligned}$$

for h a bounded function, the variance of the score operator for η yields a version of $A^*_{\theta,\eta}A_{\theta,\eta}$. Following Woodbury's (1971) missing information principle we have that the variance of the observed data score is given by the variance of the complete data score minus the expected value of the conditional variance of the complete data score given the observed data. (Also cf. Bickel et al. (1993).) The complete data score is given by $\int h(dN_. - G\tilde{Y}^\beta d\eta)$. So

$$\begin{aligned} P_{\theta,\eta}(A_{\theta,\eta}h)^2 &= \mathrm{var}\left(\int h(dN_. - G\tilde{Y}^\beta d\eta)\right) - P_{\theta,\eta}\left(\mathrm{var}(G|\,X)\left(\int h\tilde{Y}^\beta\, d\eta\right)^2\right) \\ &= P_{\theta,\eta}\left(\int h^2 G\tilde{Y}_.^\beta\, d\eta - \sigma\frac{1+\sigma N_.(\tau)}{\left(1+\sigma\int_0^\tau \tilde{Y}_.^\beta\, d\eta\right)^2}\left(\int h\tilde{Y}^\beta\, d\eta\right)^2\right). \end{aligned}$$

From this we see that,

$$A^*_{\theta,\eta}A_{\theta,\eta}h(t) = P_{\theta,\eta}\left(h(t)\frac{1+\sigma N_.(\tau)}{1+\sigma\int_0^\tau \tilde{Y}_.^\beta\, d\eta}\tilde{Y}_.^\beta(t) - \sigma\frac{1+\sigma N_.(\tau)}{\left(1+\sigma\int_0^\tau \tilde{Y}_.^\beta\, d\eta\right)^2}\int_0^\tau h\tilde{Y}_.^\beta\, d\eta\tilde{Y}_.^\beta(t)\right).$$

Similarly we may derive $A^*_{\theta,\eta}\ell_{\theta,\eta}$ by considering the covariance of $A_{\theta,\eta}$ with $\ell_{\theta,\eta}$ and using Woodbury's missing information principle. This results in

$$A^*_{\theta,\eta}\ell_{1,\theta,\eta}(t) = P_{\theta,\eta}\left(\tilde{Y}_.^{\beta\prime}(t)\frac{1+\sigma N_.(\tau)}{1+\sigma\int_0^\tau \tilde{Y}_.^\beta\, d\eta} - \tilde{Y}_.^\beta(t)\frac{1+\sigma N_.(\tau)}{\left(1+\sigma\int_0^\tau \tilde{Y}_.^\beta\, d\eta\right)^2}\sigma\int_0^\tau \tilde{Y}_.^{\beta\prime}\, d\eta\right)$$

and

$$A^*_{\theta,\eta}\ell_{2,\theta,\eta}(t) = P_{\theta,\eta}\left(\tilde{Y}_.^\beta(t)\frac{N_.(\tau) - \int_0^\tau \tilde{Y}_.^\beta\, d\eta}{\left(1+\sigma\int_0^\tau \tilde{Y}_.^\beta\, d\eta\right)^2}\right)$$

where $\tilde{Y}_\cdot^{\beta\prime}(t) = \sum_j Z_j(t) e^{\beta^T Z_j(t)} Y_j(t)$. Parner (1998) shows that $A^*_{\theta_0,\eta_0} A_{\theta_0,\eta_0}$ is continuously invertible as an operator on the space of bounded functions of bounded variation (this may also be proved via the methods used by Murphy, Rossini and Van der Vaart (1996)). Since $A^*_{\theta_0,\eta_0} \ell_{\theta_0,\eta_0}$ is a vector of bounded functions of bounded variation, the least favorable direction for the estimation of θ in the presence of the unknown η is $h_0 = \left(A^*_{\theta_0,\eta_0} A_{\theta_0,\eta_0}\right)^{-1} A^*_{\theta_0,\eta_0} \ell_{\theta_0,\eta_0}$.

The theorems are applied, for t a vector in $\mathbb{R}^{d+1}$,

$$\boldsymbol{\eta}_t(\theta,\eta) = \eta + (\theta - t)^T \int^{\cdot} h_0 \, d\eta$$
$$\ell(t,\theta,\eta) = \log l(t, \boldsymbol{\eta}_t(\theta,\eta))$$

Since h_0 is bounded, the density $1 + (\theta - t)^T h_0$ of $\boldsymbol{\eta}_t(\theta,\eta)$ with respect to η is positive for sufficiently small $|\theta - t|$. In this case $\boldsymbol{\eta}_t(\theta,\eta)$ defines a nondecreasing function and is a true parameter of the model. Note that (3.1) is satisfied. Also differentiation of $\ell(t,\theta_0,\eta_0)$ with respect to t at θ_0 yields $\ell_{\theta_0,\eta_0} - A_{\theta_0,\eta_0} h_0$, the efficient score ((3.2) is satisfied).

The cumulative baseline hazard, η is estimable at $\sqrt{n}$-rate in the supremum norm. Suppose that $\tilde{\theta} \xrightarrow{P} \theta_0$. Uniform consistency of $\hat{\eta}_{\tilde{\theta}}$, (3.3), may be proved by minor adaption of the consistency proofs given by Murphy (1994) and Parner (1998). To show that

$$\sup_{0 \le t \le \tau} |\hat{\eta}_{\tilde{\theta}}(t) - \eta_0(t)| = O_P(\|\tilde{\theta} - \theta_0\| + n^{-1/2}),$$

we may verify the conditions of Theorem 3.1 in Murphy and Van der Vaart (1997c). These conditions (continuous invertibility of the information operator and that the score functions belong to a Donsker class) were verified by Parner (1998). The verification of (3.4) follows the same steps as in the previous example.

4.4. A Mixture Model

This is another version of the frailty model, with the group size equal to two. The observations are a random sample of pairs of event times. As before, we allow for intra-group correlation by assuming that every pair of event times share the same unobserved frailty G. Given G, the two event times U and V are assumed to be independent and exponentially distributed with hazard rates G and θG, respectively. The parameter, θ is assumed to belong to $(0,\infty)$. In contrast to the gamma frailty model, the distribution η of G is a completely unknown distribution on $(0,\infty)$. The observations are n i.i.d. copies of $X = (U,V)$ from the density

$$p_{\theta,\eta}(X) = \int g e^{-gU} \, \theta g e^{-\theta g V} \, d\eta(g).$$

This forms the likelihood, $l(\theta,\eta)$, for estimation. The conditions of this paper may be checked under the assumption that $\int (g^2 + g^{-6.5}) \, d\eta_0(g) < \infty$. Asymptotic normality and efficiency of $\hat{\theta}$ is shown by Van der Vaart (1996). Murphy and Van der Vaart (1997) consider likelihood ratio tests concerning θ.

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

$$\ell_{\theta,\eta}(x) = \frac{\int (\theta^{-1} - g\,v) g^2 e^{-g(u+\theta v)}\, d\eta(g)}{\int g^2 e^{-g(u+\theta v)}\, d\eta(g)}.$$

The conditional expectation $\mathrm{E}_\theta\big(\ell_{\theta,\eta}(U,V) \big| \, U + \theta V = u + \theta v\big)$ minimizes the distance to $\ell_{\theta,\eta}(u,v)$ over all functions of $u + \theta v$. Since the statistic $U + \theta V$ is sufficient for η, all score functions for η are functions of $u + \theta v$. Therefore, if it is an η-score, then the conditional expectation must be the closest η-score. In that case the efficient score function $\tilde{\ell}_0$ is given by

$$\begin{aligned}\tilde{\ell}_0(u,v) &= \ell_{\theta_0,\eta_0}(u,v) - \mathrm{E}_{\theta_0}\big(\ell_{\theta_0,\eta_0}(U,V) \big| \, U + \theta V = u + \theta_0 v\big) \\ &= \frac{\int \frac{1}{2}(u - \theta_0 v) g^3 e^{-g(u+\theta_0 v)}\, d\eta_0(g)}{\int \theta_0 g^2 e^{-g(u+\theta_0 v)}\, d\eta_0(g)}.\end{aligned}$$

To verify that the conditional expectation is an η-score, it suffices to verify that $\tilde{\ell}_0$ is a score. For B a measurable set in $(0,\infty)$, define,

$$\begin{aligned}\boldsymbol{\eta}_t(\theta,\eta)(B) &= \eta\Big(B\Big(1 + \frac{\theta - t}{2\theta}\Big)^{-1}\Big) \\ \ell(t,\theta,\eta) &= \ln p_{t,\eta_t(\theta,\eta)}.\end{aligned}$$

Whenever $|\theta - t| < 2\theta$, $\boldsymbol{\eta}_t$ is a distribution on the positive real line. By an easy calculation we see that,

$$\tilde{\ell}_{\theta,\eta}(x) = \dot{\ell}(\theta,\theta,\eta)(x).$$

This, in addition to the preceding equations, implies that (3.1) and (3.2) are satisfied, and implies that the path $t \mapsto \boldsymbol{\eta}_t(\theta,\eta)$ indexes a least favorable submodel for θ.

Consistency of $\hat{\eta}_{\tilde{\theta}}$ (with respect to the weak topology) can be shown by minor adaptations to the classical proof of Kiefer and Wolfowitz (1956). From the representation of the efficient score function as a conditional score, we see that $P_{\theta_0,\eta_0}\dot{\ell}(\theta_0,\theta_0,\eta) = 0$ for every θ_0, η_0 and η, thus (3.9) is automatically satisfied. This rather nice property allows us to avoid proving a rate of convergence for the estimator $\hat{\eta}_{\tilde{\theta}}$. (This is fortunate, because it follows from Van der Vaart (1991) that the information for estimating the distribution function $\eta(0,b]$ is zero, so that the rate of convergence of the best estimators is less than the square root of n.) To finish the verification of (3.4) we next study the dependence of $P_0\dot{\ell}(\theta_0,\theta,\eta)$ on θ, but this is straightforward, because it only involves a Euclidean parameter. (Cf. (3.8).)

5. DISCUSSION

For many years, we statisticians have formulated and then implemented semiparametric models. In the numerical evaluation of the maximum likelihood estimators of the model, we discovered that the log profile likelihood for θ appears quadratic. This paper says that it then makes sense to use our intuition garnered from classical parametric likelihood theory. Indeed we can then expect that the maximum likelihood estimator of θ will be asymptotically normal and efficient. Furthermore the curvature of the log profile likelihood can indeed be used to estimate the standard error matrix of $\hat{\theta}$. If we want to demonstrate theoretically that our intuition is correct, we can follow the methods of this paper.

The methodology presented here can also be applied to penalized and sieved likelihood estimators. Furthermore, the results extend to functionals of interest θ on models that are not naturally parametrized by a partitioned parameter (θ, η). We have not investigated such extensions in this paper.

A selection of other settings in which an understanding of semiparametric profile likelihoods is important are the proportional odds model as discussed by Rossini and Tsiatis (1996), Murphy, Rossini and Van der Vaart (1997), and Shen (1999), the linear errors-in-variables model (Murphy and Van der Vaart, 1996), penalized partially linear logistic regression (Mammen and van de Geer, 1997, Murphy and Van der Vaart, 1999), selection bias models models as discussed by Gill, Vardi and Wellner (1988) and their semiparametric extension as in Gilbert, Lele and Vardi (1999), the correlated frailty model treated by Parner (1998), ranked set samples as in Huang (1997), the double censoring model as in Chang and Yang (1987), Chang (1990) and Gu and Zhang, (1993), models for interval-truncated data as discussed by Tsai and Zhang (1995), models restricted by estimating equations as in Qin and Lawless (1994), and censored line segment processes as in Van der Laan (1996). This methodology should apply to the above models, although we do not claim to have solved all problems. The validation of the application of infinitely-dimensional statistical models remains a challenge for theory.

The models which present the greatest challenge are the semiparametric mixture models (see Lindsay, 1995). Our last example is of this type, but it is special in that the least favorable submodel is concretely given. Another example is treated in Susko, Kalbfleisch and Chen (1998), but it is also special in that it concerns a special parametrization of a finite-dimensional model. It appears difficult to derive good approximations to a least favorable path for such models, and given such approximations it is unclear how one would verify the no-bias condition.

REFERENCES

[1] Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N., (1993). *Statistical Models Based on Counting Processes*. Springer Verlag, New York.

[2] Andersen, P.K. and Gill, R.D., (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics* **10**, 1100–1120.

[3] Bailey, K.R., (1984). Asymptotic equivalence between the Cox estimator and the general ML estimators of regression and survival parameters in the Cox mode. *Annals of Statistics* **12**, 730–736.

[4] Barndorff-Nielsen, O.E. and Cox, D.R., (1994). *Inference and Asymptotics*. Chapman & Hall, London.

[5] Begun, J.M., Hall, W.J., Huang, W.-.M. and Wellner, J.A., (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Annals of Statistics* **11**, 432–452.

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

[7] Birgé, L. and Massart, P., (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97**, 113–150.

[8] Birgé, L. and Massart, P., (1997). From model selection to adaptive estimation. *Festschrift for Lucien Le Cam*, 55–87. Springer, New York.

[9] Chang, M.N. and Yang, G.L., (1987). Strong consistency of a nonparametric estimator of the survival function with doubly censored data. *Annals of Statistics* **15**, 1536–1547.

[10] Chang, M.N., (1990). Weak convergence of a self-consistent estimator of the survival function with doubly censored data. *Annals of Statistics* **18**, 391–404.

[11] Chernoff, H., (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics* **25**, 573–578.

[12] Chen, J.S. and Jennrich, R.I., (1996). The signed root deviance profile and confidence intervals in maximum likelihood analysis. *Journal of the American Statistical Association* **91**, 993–998.

[13] Cramér, H., (1946). *Mathematical methods of statistics*. Princeton University Press, Princeton.

[14] Donald, S.G. and Newey, W.K., (1994). Series estimation of semilinear models 50. *Journal of Multivariate Analysis* **50**, 30–40.

[15] Van de Geer, S.A., (1995). The method of sieves and minimum contrast estimators. *Math. Methods Statist.* **4**, 20–38.

[16] Van de Geer, S.A., (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Annals of Statistics* **21**, 14–44.

[17] Gilbert, P.B., Lele, S.R. and Vardi, Y., (1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika*, to appear.

[18] Gill, R.D., Vardi, Y. and Wellner, J.A., (1988). Large sample theory of empirical distributions in biased sampling models. *Annals of Statistics* **16**, 1069–1012.

[19] Groeneboom, P. and Wellner, J.A., (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser, Basel.

[20] Gu, M.G. and Zhang, C.H., (1993). Asymptotic properties of self-consistent estimators based on doubly censored data. *Annals of Statistics* **21**, 611–624.

[21] Huang, J., (1996). Efficient estimation for the Cox model with interval censoring. *Annals of Statistics* **24**, 540–568.

[22] Huang, J., (1997). Asymptotic properties of the NPMLE of a distribution function based on ranked set samples. *Annals of Statistics* **25**, 1036–1049.

[23] Huang, J. and Wellner, J.A., (1995). Asymptotic normality of the NPMLE of linear functionals for interval censored data, case 1. *Statistica Neerlandica* **49**, 153–163.

[24] Kiefer, J. and Wolfowitz, J., (1956). Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Nuisance Parameters. *Annals of Mathematical Statistics* **27**, 887–906.

[25] Klaassen, C.A.J., (1987). Consistent estimation of the influence function of locally efficient estimates. *Annals of Statistics* **15**, 617–627.

[26] van der Laan, M.J., (1996). Efficiency of the NPMLE in the line-segment problem. *Scandinavian Journal of Statistics* **23**, 527–550.

[27] Le Cam, L. and Yang, G.L., (1990). *Asymptotics in statistics*. Springer-Verlag, New York.

[28] Lindsay, B.G., (1995). *Mixture models : theory, geometry, and applications*. Institute of Mathematical Statistics, American Statistical Association, Alexandria, Virginia.

[29] Mammen, E. and van de Geer, S., (1997). Penalized quasi-likelihood estimation in partial linear models. *Annals of Statistics* **25**, 1014–1035.

[30] Murphy, S.A., (1994). Consistency in a proportional hazards model incorporating a random effect. *Annals of Statistics* **22**, 712–731.

[31] Murphy, S.A., (1995). Asymptotic theory for the frailty model. *Annals of Statistics* **23**, 182–198.

[32] Murphy, S.A., Rossini, A.J. and Van der Vaart, A.W., (1997d). MLE in the proportional odds model. *Journal of the American Statistical Association* **92**, 968–976.

[33] Murphy, S.A. and Van der Vaart, A.W., (1996). Likelihood inference in the errors-in-variables model. *Journal of Multivariate Analysis* **59**, 81–108.

[34] Murphy, S.A. and Van der Vaart, A.W., (1997a). Semiparametric likelihood ratio inference. *Annals of Statistics* **25**, 1471–1509.

[35] Murphy, S.A. and Van der Vaart, A.W., (1997b). Semiparametric mixtures in case-control studies. *preprint*.

[36] Murphy, S.A. and Van der Vaart, A.W., (1999). Observed information in semiparametric models. *Bernoulli, to appear.*

[37] 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. *Scandinavian Journal of Statistics* **19**, 25–44.

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

[39] Parner, E., (1998). Asymptotic theory for the correlated gamma-frailty model. *Annals of Statistics* **26**, 183–214.

[40] Patefield, W.M., (1977). On the maximized likelihood function. *Sankhyā, Ser. B* **39**, 92–96.

[41] Pollard, D., (1984). *Convergence of Stochastic Processes*. Springer Verlag.

[42] Qin, J. and Lawless, J., (1994). Empirical likelihood and general estimating equations. *Annals of Statistics* **22**, 300–325.

[43] Roeder, K., Carroll, R.J. and Lindsay, B.G., (1996). A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association* **91**, 722–732.

[44] Rossini, A.J. and Tsiatis, A.A., (1996). A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association* **91**, 713–721.

[45] Shen, X., (1999). Proportional odds regression and sieve maximum likelihood estimation.. *Biometrika*, to appear.

[46] Severini, T.A. and Wong, W.H., (1992). Profile likelihood and conditionally parametric models. *Annals of Statistics* **20**, 1768-1802.

[47] Susko, E., Kalbfleisch, J.D. and Chen, J., (1998). Constrained nonparametric maximum-likelihood estimation for mixture models. *Canadian Journal of Statistics* **26**, 601–617.

[48] Tsai, W.Y. and Zhang, C.H., (1995). Asymptotic properties of nonparametric maximum likelihood estimator for interval-truncated data. *Scandinavian Journal of Statistics* **22**, 361–370.

[49] Tsiatis, A.A., (1981). A large sample study of Cox's regression model. *Annals of Statistics* **9**, 93-108.

[50] Van der Vaart, A.W., (1991). On differentiable functionals. *Annals of Statistics* **19**, 178–204.

[51] Van der Vaart, A.W., (1996). Efficient estimation in semiparametric models. *Annals of Statistics* **24**, 862–878.

[52] Van der Vaart, A.W., (1998). *Asymptotic Statistics*. Cambridge University Press, New York.

[53] Van der Vaart, A.W. and Wellner, J.A., (1996). *Weak Convergence and Empirical Processes*. Springer Verlag, New York.

[54] Wong, W.H. and Shen, X., (1995). A probability ineuqality for the likelihood surface and convergence rate of the maximum likelihood estimator. *Annals of Statistics* **23**, 603–632.

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

S. A. Murphy
Department of Statistics
University of Michigan
1440 Mason Hall
Ann Arbor, MI 48109-1027, USA

A.W. Van der Vaart
Department of Mathematics
Free University
De Boelelaan 1081a
1081 HV Amsterdam, The Netherlands