Introduction to Empirical Processes and Semiparametric Inference
Lecture 26: Semiparametric Maximum Likelihood Inference

Michael R. Kosorok, Ph.D.

Professor and Chair of Biostatistics

Professor of Statistics and Operations Research

University of North Carolina-Chapel Hill

Maximum Likelihood Estimation

The most common approach to efficient estimation is based on modifications of maximum likelihood estimation that lead to efficient estimates.

These modifications, which we will call "likelihoods," are generally not really likelihoods (products of densities) because of complications resulting from the presence of an infinite dimensional nuisance parameter.

Recall the setting of estimation of an unknown real density $f(x)$ from an i.i.d. sample $X_1, \ldots, X_n$.

The likelihood is $\prod_{i=1}^{n} f(X_i)$, and the maximizer over all densities has arbitrarily high peaks at the observations, with zero at the other values, and is therefore not a density.

This problem can be fixed by using an empirical likelihood $\prod_{i=1}^{n} p_i$, where $p_1, \ldots, p_n$ are the masses assigned to the observations indexed by $i = 1, \ldots, n$ and are constrained to satisfy $\sum_{i=1}^{n} p_i = 1$.

This leads to the empirical distribution function estimator, which is known to be fully efficient.

Consider again the Cox model for right-censored data explored in the previous section.

The density for a single observation $X = (V, d, Z)$ is proportional to

$$\left[e^{\beta' Z}\lambda(V)\right]^d \exp\left[-e^{\beta' Z}\Lambda(V)\right].$$

Maximizing the likelihood based on this density will result in the same problem raised in the previous paragraph.

A likelihood that works is the following, which assigns mass only at observed failure times:

$$L_n(\beta, \Lambda) \quad = \quad \prod_{i=1}^{n} \left[e^{\beta' Z_i} \Delta\Lambda(V_i) \right]^{d_i} \exp\left[-e^{\beta' Z_i} \Lambda(V_i) \right], \quad (1)$$

where $\Delta\Lambda(t)$ is the jump size of Λ at t.

For each value of β, one can maximize or *profile* $L_n(\beta, \Lambda)$ over the "nuisance" parameter Λ to obtain the profile likelihood $pL_n(\beta)$, which for the Cox model is $\exp\left[-\sum_{i=1}^{n} d_i\right]$ times the partial likelihood (3.4).

Let $\hat{\beta}$ be the maximizer of $pL_n(\beta)$.

Then the maximizer $\hat{\Lambda}$ of $L_n(\hat{\beta}, \Lambda)$ is the "Breslow estimator"

$$\hat{\Lambda}(t) = \int_0^t \frac{\mathbb{P}_n dN(s)}{\mathbb{P}_n \left[Y(s) e^{\hat{\beta}' Z} \right]}.$$

We showed in Chapter 4 that $\hat{\beta}$ and $\hat{\Lambda}$ are both efficient.

Another useful class of likelihood variants are *penalized likelihoods*.

Penalized likelihoods add a penalty term (or terms) in order to maintain an appropriate level of smoothness for one or more of the nuisance parameters.

This method was used in the partly linear logistic regression model described in Chapter 1.

Other methods of generating likelihood variants that work are possible.

The basic idea is that using the likelihood principle to guide estimation of semiparametric models often leads to efficient estimators for the model components which are $\sqrt{n}$ consistent.

Because of the richness of this approach to estimation, one needs to verify for each new situation that a likelihood-inspired estimator is consistent, efficient and well-behaved for moderate sample sizes.

Verifying efficiency usually entails demonstrating that the estimator satisfies the efficient score equation described in the previous section.

Approximately Least-Favorable Submodels

One of the key challenges in this setting is to ensure that the efficient score is a derivative of the chosen log likelihood along some submodel.

Something helpful in this setting are *approximately least-favorable submodels*.

The basic idea is to find a function $\eta_t(\theta, \eta)$ such that

- $\eta_0(\theta, \eta) = \eta$, for all $\theta \in \Theta$ and $\eta \in H$, where $\eta_t(\theta, \eta) \in H$ for all t small enough, and such that
- $\tilde{\kappa}_{\theta_0,\eta_0} = \tilde{\ell}_{\theta_0,\eta_0}$, where

$$\tilde{\kappa}_{\theta,\eta}(x) = \left. \frac{\partial l_{\theta+t,\eta_t(\theta,\eta)}(x)}{\partial t} \right|_{t=0},$$

  $l_{\theta,\eta}(x)$ is the log-likelihood for the observed value x at the parameters (θ, η), and where (θ_0, η_0) are the true parameter values.

Note that we require $\tilde{\kappa}_{\theta,\eta} = \tilde{\ell}_{\theta,\eta}$ only when $(\theta, \eta) = (\theta_0, \eta_0)$.

If $(\hat{\theta}_n, \hat{\eta}_n)$ is the maximum likelihood estimator, i.e., the maximizer of $\mathbb{P}_n l_{\theta,\eta}$, then the function

$$t \mapsto \mathbb{P}_n l_{\hat{\theta}_n + t, \eta_t(\hat{\theta}_n, \hat{\eta}_n)}$$

is maximal at $t = 0$, and thus $(\hat{\theta}_n, \hat{\eta}_n)$ is a zero of $\mathbb{P}_n \tilde{\kappa}_{\tilde{\theta},\tilde{\eta}}$.

Now if $\hat{\theta}_n$ and

$$\hat{\ell}_{\tilde{\theta},n} = \tilde{\kappa}_{\tilde{\theta},\hat{\eta}_n}$$

satisfy the conditions of Theorem 3.1 at $(\theta, \eta) = (\theta_0, \eta_0)$, then the maximum likelihood estimator $\hat{\theta}_n$ is asymptotically efficient at (θ_0, η_0).

Consider now the maximum likelihood estimator $\hat{\theta}_n$ based on maximizing the joint empirical log-likelihood

$$L_n(\theta, \eta) \equiv n\mathbb{P}_n l(\theta, \eta),$$

where $l(\cdot, \cdot)$ is the log-likelihood for a single observation.

For now, η will be regarded as a nuisance parameter, and thus we can restrict our attention to the profile log-likelihood

$$\theta \mapsto pL_n(\theta) \equiv \sup_{\eta} L_n(\theta, \eta).$$

Note that L_n is a sum and not an average, since we multiplied the empirical measure by n.

While the solution of an efficient score equation need not be a maximum likelihood estimator, it is also possible that the maximum likelihood estimator in a semiparametric model may not be expressible as the zero of an efficient score equation.

This possibility occurs because the efficient score is a projection, and, as such, there is no assurance that this projection is the derivative of the log-likelihood along a submodel.

This is the main issue that motivates approximately least-favorable submodels.

An approximately least-favorable submodel approximates the true least-favorable submodel to a useful level of accuracy that facilitates analysis of semiparametric estimators.

We will now describe this process in generality: the specifics will depend on the situation.

As mentioned previously, we first need a general map from the neighborhood of θ into the parameter set for η, which map we will denote by $t \mapsto \eta_t(\theta, \eta)$, for $t \in \mathbb{R}^k$.

We require that

$$\begin{aligned} \eta_t(\theta, \eta) &\in \hat{H}, \quad \text{for all } \|t - \theta\| \text{ small enough, and} \qquad (2) \\ \eta_\theta(\theta, \eta) &= \eta \quad \text{for any } (\theta, \eta) \in \Theta \times \hat{H}, \end{aligned}$$

where $\hat{H}$ is a suitable enlargement of H that includes all estimators that satisfy the constraints of the estimation process.

Now define the map

$$\ell(t, \theta, \eta) \equiv l(t, \eta_t(\theta, \eta)).$$

We will require several things of $\ell(\cdot, \cdot, \cdot)$, at various point in our discussion, that will result in further restrictions on $\eta_t(\theta, \eta)$.

Define

$$\dot{\ell}(t, \theta, \eta) \equiv \frac{\partial}{\partial t)} \ell(t, \theta, \eta),$$

and let

$$\hat{\ell}_{\theta,n} \equiv \dot{\ell}(\theta, \theta, \hat{\eta}_n).$$

Clearly, $\mathbb{P}_n \hat{\ell}_{\hat{\theta}_n,n} = 0$, and thus $\hat{\theta}_n$ is efficient for θ_0, provided $\hat{\ell}_{\theta,n}$ satisfies the conditions of Theorem 3.1.

The reason it is necessary to check this even for maximum likelihood estimators is that, as mentioned previously, $\hat{\eta}_n$ is often on the boundary (or even a little bit outside of) the parameter space.

Recall again the Cox model setting for right censored data.

In this case, η is the baseline integrated hazard function which is usually assumed to be continuous.

However, $\hat{\eta}_n$ is the Breslow estimator, which is a right-continuous step function with jumps at observed failure times and is therefore not in the parameter space.

Thus direct differentiation of the log-likelihood at the maximum likelihood estimator will not yield an efficient score equation.

We will also require that

$$\dot{\ell}(\theta_0, \theta_0, \eta_0) \quad = \quad \tilde{\ell}_{\theta_0, \eta_0}. \tag{3}$$

Note that we are only insisting that this identity holds at the true parameter values.

This approximately least-favorable submodel structure is very useful for developing methods of inference for θ.

Example 1: Right Censored Cox Model

The Cox model for right censored data has been discussed previously.

Because of notational tradition, we will use β instead of θ for the regression parameter and Λ instead η for the baseline integrated hazard function.

Recall that an observation from this model has the form $X = (W, \delta, Z)$, where

- $W = T \wedge C$ and $\delta = \mathbf{1}\{W = T\}$,
- $Z \in \mathbb{R}^k$ is a regression covariate,
- T is a right-censored failure time with integrated hazard given Z equal to
$$t \mapsto e^{\beta' Z} \Lambda(t),$$
- and C is a censoring time independent of T given Z and uninformative of (β, Λ).

We assume that there exists a $\tau < \infty$ such that
$P(C \geq \tau) = P(C = \tau) > 0$.

We require H to consist of all monotone increasing, functions
$\Lambda \in C[0, \tau]$ with $\Lambda(0) = 0$.

We define $\hat{H}$ to be the set of all monotone, increasing functions
$\Lambda \in D[0, \tau]$.

As shown previously, the efficient score for β is

$$\tilde{\ell}_{\beta,\Lambda} = \int_0^\tau (Z - h_0(s))dM(s),$$

where

$$M(t) \equiv N(t) - \int_0^t Y(s)e^{\beta' Z}d\Lambda(s),$$

N and Y are the usual counting and at-risk processes respectively, and

$$h_0(t) \equiv \frac{P[Z\mathbf{1}\{W \geq t\}e^{\beta_0' Z}]}{P[\mathbf{1}\{W \geq t\}e^{\beta_0' Z}]},$$

where P is the true probability measure (at the parameter values (β_0, Λ_0)).

Recall that the log-likelihood for a single observation is

$$l(\theta, \Lambda) = (\beta' Z + \log \Delta\Lambda(W))\delta - e^{\beta' Z}\Lambda(W),$$

where $\Delta\Lambda(w)$ is the jump size in Λ at w.

We will now verify that Conditions (2) and (3) are both satisfied.

These results will prove useful to develop valid methods of inference for β

If we let

$$t \mapsto \Lambda_t(\beta, \Lambda)) \equiv \int_0^{(\cdot)} (1 + (\beta - t)' h_0(s)) d\Lambda(s),$$

then $\Lambda_t(\beta, \Lambda) \in \hat{H}$ for all t small enough, $\Lambda_\beta(\beta, \Lambda) = \Lambda$, and

$$\dot{\ell}(\beta_0, \beta_0, \Lambda_0) \quad = \quad \int_0^\tau (Z - h_0(s)) dM(s) = \tilde{\ell}_{\beta_0, \Lambda_0} \qquad (4)$$

(see Exercise 19.5.2).

Thus Conditions (2) and (3) are both satisfied.

Example 2: Current Status Cox Model

Current status data arises when each subject is observed at a single examination time, Y, to determine whether an event has occurred.

The event time, T, cannot be observed exactly.

Including the covariate Z, the observed data consists of n independent and identically distributed realizations of $X = (Y, \delta, Z)$, where $\delta = \mathbf{1}\{T \leq Y\}$.

We assume that the integrated hazard function of T given Z has the proportional hazards form and parameters (β, Λ) as given in Section 19.2.1 above.

We also make the following additional assumptions:

- T and Y are independent given Z.
- Z lies in a compact set almost surely and the covariance of $Z - E(Z|Y)$ is positive definite which guarantees that the efficient information $\tilde{I}_0$ is strictly positive.
- Y possesses a Lebesgue density which is continuous and positive on its support $[\sigma, \tau]$, where $0 < \sigma < \tau < \infty$,
- for which the true parameter Λ_0 satisfies $\Lambda_0(\sigma-) > 0$ and $\Lambda_0(\tau) < M < \infty$, for some known M,
- and is continuously differentiable on this interval with derivative bounded above zero.

We let H denote all such possible choices of Λ_0 satisfying these constraints for the given value of M, and we let $\hat{H}$ consist of all nonnegative, nondecreasing right-continuous functions Λ on $[\sigma, \tau]$ with $\Lambda(\tau) \leq M$.

We can deduce that the log-likelihood for a single observation, $l(\beta, \Lambda)$, has the form

$$\begin{aligned} l(\beta, \Lambda) &= \delta \log[1 - \exp(-\Lambda(Y) \exp(\beta' Z))] \\ &\quad -(1-\delta) \exp(\beta' Z) \Lambda(Y). \end{aligned} \tag{5}$$

From the likelihood, we can deduce that the score function for β takes the form

$$\dot{\ell}_{\beta,\Lambda}(x) = z\Lambda(y)Q(x;\beta,\Lambda),$$

where

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

Inserting a submodel $t \mapsto \Lambda_t$ such that

$$h(y) = -\frac{\partial}{\partial} t \Big|_{t=0} \Lambda_t(y)$$

exists for every y into the log likelihood and differentiating at $t = 0$, we obtain a score function for Λ of the form

$$A_{\beta,\Lambda} h(x) = h(y) Q(x; \beta, \Lambda).$$

The linear span of these functions contains $A_{\beta,\Lambda} h$ for all bounded functions h of bounded variation.

Thus the efficient score function for θ is

$$\tilde{\ell}_{\beta,\Lambda} = \dot{\ell}_{\beta,\Lambda} - A_{\beta,\Lambda} h_{\beta,\Lambda}$$

for the least-favorable direction vector of functions $h_{\beta,\Lambda}$ minimizing the distance

$$P_{\beta,\Lambda} \| \dot{\ell}_{\beta,\Lambda} - A_{\beta,\Lambda} h \|^2.$$

The solution at the true parameter is

$$\begin{aligned} h_0(Y) &\equiv \Lambda_0(Y) h_{00}(Y) \\ &\equiv \Lambda_0(Y) \frac{E_{\beta_0,\Lambda_0}(ZQ^2(X;\beta_0,\Lambda_0)|Y)}{E_{\beta_0,\Lambda_0}(Q^2(X;\beta_0,\Lambda_0)|Y)} \end{aligned} \tag{6}$$

(see Exercise 19.5.4).

As the formula shows, the vector of functions $h_0(y)$ is unique a.s., and $h_0(y)$ is a bounded function since $Q(x; \theta_0, \Lambda_0)$ is bounded away from zero and infinity.

We assume that the function h_0 given by (6) has a version which is differentiable with a bounded derivative on $[\sigma, \tau]$.

An approximately least-favorable submodel can thus be of the form

$$\Lambda_t(\beta, \Lambda)(\cdot) = \Lambda(\cdot) + \phi(\Lambda(\cdot))(\beta - t)' h_{00} \circ \Lambda_0^{-1} \circ \Lambda(\cdot),$$

where $\phi(\cdot)$ is a fixed function we will define shortly that approximates the identity.

Note that we need to extend Λ_0^{-1} so that it is defined on all of $[0, M]$.

This is done by letting $\Lambda_0^{-1}(t) = \sigma$ for all $t \in [0, \Lambda_0(\sigma)]$ and $\Lambda_0^{-1}(t) = \tau$ for all $t \in [\Lambda_0(\tau), M]$.

We take $\phi : [0, M] \mapsto [0, M]$ to be any fixed function

- with $\phi(u) = u$ on $[\Lambda_0(\sigma), \Lambda_0(\tau)]$
- such that $u \mapsto \phi(u)/u$ is Lipschitz
- and $\phi(u) \leq c(u \wedge (M - u))$ for all $u \in [0, M]$ and some $c < \infty$ depending only on (β_0, Λ_0).

Our conditions on the model ensure that such a function exists.

The function $\Lambda_t(\beta, \Lambda)$ is essentially Λ plus a perturbation in the least favorable direction, h_0, but its definition is somewhat complicated in order to ensure that $\Lambda_t(\beta, \Lambda)$ really defines a cumulative hazard function within our parameter space, at least for all t that is sufficiently close to β.

To see this, first note that by using $h_{00} \circ \Lambda_0^{-1} \circ \Lambda$, rather than h_{00}, we ensure that the perturbation that is added to Λ is Lipschitz-continuous with respect to Λ.

Combining this with the Lipschitz-continuity of $\phi(u)/u$, we obtain that for any $0 \leq v < u \leq M$,

$$\Lambda_t(\beta, \Lambda)(u) - \Lambda_t(\beta, \Lambda)(v) \geq \Lambda(u) - \Lambda(v) - k_0 \|\beta - t\| (\Lambda(u) - \Lambda(v)),$$

for some universal constant $0 < k_0 < \infty$.

Since $\Lambda(v) \leq M$, we obtain that for all $\|t - \beta\|$ small enough, $\Lambda_t(\beta, \Lambda)$ is non-decreasing.

The additional constraints on ϕ ensures that

$$0 \leq \Lambda_t(\beta, \Lambda) < M$$

for all $\|t - \beta\|$ small enough.

Hence $t \mapsto \Lambda_t(\beta, \Lambda)$ satisfies both (2) and (3).

Additional details about the construction of the approximately least-favorable submodel for this example can be found in Section 4.1 of Murphy and van der Vaart (2000).

Quadratic Expansion of the Profile Likelihood

The main ideas of this section come from a very elegant paper by Murphy and van der Vaart (2000) on profile likelihood.

The context is where we have maximum likelihood estimators $(\hat{\theta}_n, \hat{\eta}_n)$ based on a i.i.d. sample $X_1, \ldots, X_n$, where one is the finite-dimensional parameter of primary interest ($\hat{\theta}_n$) and the other is an infinite-dimensional nuisance parameter ($\hat{\eta}_n$).

The main result, given formally later as Theorem 1, is that under certain regularity conditions, we have for any estimator $\tilde{\theta}_n \xrightarrow{\mathsf{P}} \theta_0$, that

$$\begin{aligned} pL_n(\tilde{\theta}_n) &= pL_n(\theta_0) + (\tilde{\theta}_n - \theta_0)' \sum_{i=1}^{n} \tilde{\ell}_{\theta_0,\eta_0}(X_i) \qquad (7) \\ &\quad -\frac{1}{2} n (\tilde{\theta}_n - \theta_0)' \tilde{I}_{\theta_0,\eta_0} (\tilde{\theta}_n - \theta_0) \\ &\quad + o_{P_0}(1 + \sqrt{n}\|\tilde{\theta}_n - \theta_0\|)^2, \end{aligned}$$

where $\tilde{I}_{\theta_0,\eta_0}$ is the efficient Fisher information and P_0 the probability measure of X at the true parameter values.

Suppose we can know that the maximum profile likelihood estimator is consistent, i.e., that $\hat{\theta}_n = \theta_0 + o_{P_0}(1)$, and that $\tilde{I}_{\theta_0,\eta_0}$ is positive definite.

Then if (7) also holds, we have that

$$\begin{aligned}
\|\sqrt{n}(\hat{\theta}_n - \theta_0)\|^2 &\leq \sqrt{n}(\hat{\theta}_n - \theta_0)' \left[n^{-1/2} \sum_{i=1}^{n} \tilde{\ell}_{\theta_0,\eta_0}(X_i) \right] \\
&\quad + o_{P_0}(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|)^2 \\
&= O_{P_0}(\sqrt{n}\|\hat{\theta}_n - \theta_0\|) + o_{P_0}(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|)^2,
\end{aligned}$$

since $pL_n(\hat{\theta}_n) - pL_n(\theta_0) \geq 0$.

This now implies that

$$(1+\sqrt{n}\|\hat{\theta}_n-\theta_0\|)^2 = O_{P_0}(1+\sqrt{n}\|\hat{\theta}_n-\theta_0\|)+o_{P_0}(1+\sqrt{n}\|\hat{\theta}_n-\theta_0\|)^2,$$

which yields that

$$\sqrt{n}\|\hat{\theta}_n - \theta_0\| = O_{P_0}(1).$$

Let $K \subset \mathbb{R}^k$ be a compact neighborhood of 0, and note that for any sequence of possibly random points $\theta_n \in \theta_0 + n^{-1/2}K$, we have $\theta_n = \theta_0 + o_{P_0}(1)$ and $\sqrt{n}(\theta_n - \theta_0) = O_{P_0}(1)$.

Thus

$$\sup_{u \in K} |pL_n(\theta_0 + u/\sqrt{n}) - pL_n(\theta_0) - M_n(u)| = o_{P_0}(1),$$

where

$$u \mapsto M_n(u) \equiv u' Z_n - (1/2) u' \tilde{I}_{\theta_0, \eta_0} u$$

and

$$Z_n \equiv \sqrt{n} \mathbb{P}_n \tilde{\ell}_{\theta_0, \eta_0}(X).$$

Since $Z_n \rightsquigarrow Z$, where

$$Z \sim N_k(0, \tilde{I}_{\theta_0,\eta_0}),$$

and since K was arbitrary, we have by the argmax theorem (Theorem 14.1) that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \tilde{I}^{-1}_{\theta_0,\eta_0} Z,$$

which implies that $\hat{\theta}_n$ is efficient, and thus, by Theorem 18.7, we also have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \quad = \quad \sqrt{n}\mathbb{P}_n \tilde{I}^{-1}_{\theta_0,\eta_0} \tilde{\ell}_{\theta_0,\eta_0}(X) + o_{P_0}(1). \qquad (8)$$

These arguments can easily be strengthened to imply the following simple corollary:

COROLLARY 1. *Let the estimator $\check{\theta}_n$ be consistent for θ_0 and satisfy*

$$pL_n(\check{\theta}_n) \geq pL_n(\hat{\theta}_n) - o_{P_0}(1).$$

Then, provided $\tilde{I}_{\theta_0,\eta_0}$ is positive definite and (7) holds, $\check{\theta}_n$ is efficient.

Combining (7) and Corollary 1, we obtain the following:

COROLLARY 2. *Let* $\check{\theta}_n = \theta_0 + o_{P_0}(1)$ *and satisfy* $pL_n(\check{\theta}_n) \geq pL_n(\hat{\theta}_n) - o_{P_0}(1)$.

Then, provided $\tilde{I}_{\theta_0,\eta_0}$ *is positive definite and (7) holds, we have for any random sequence* $\tilde{\theta}_n = \theta_0 + o_{P_0}(1)$,

$$\begin{aligned} pL_n(\tilde{\theta}_n) &= pL_n(\check{\theta}_n) - \frac{1}{2}n(\tilde{\theta}_n - \check{\theta}_n)'\tilde{I}_{\theta_0,\eta_0}(\tilde{\theta}_n - \check{\theta}_n) \\ &\quad + o_{P_0}(1 + \sqrt{n}\|\tilde{\theta}_n - \theta_0\|)^2. \end{aligned} \tag{9}$$

The following two additional corollaries provide methods of using this quadratic expansion to conduct inference for θ_0:

COROLLARY 3. *Assume the conditions of Corollary 1 hold for* $\check{\theta}_n$*.*

Then, under the null hypothesis $H_0 : \theta = \theta_0$*,*

$$2(pL_n(\check{\theta}_n) - pL_n(\theta_0)) \rightsquigarrow \chi^2(k),$$

where $\chi^2(k)$ *is a chi-squared random variable with* k *degrees of freedom.*

COROLLARY 4. *Assume the conditions of Corollary 1 hold for* $\check{\theta}_n$.

Then for any vector sequence $v_n \xrightarrow{\mathsf{P}} v \in \mathbb{R}^k$ *and any scalar sequence* $h_n \xrightarrow{\mathsf{P}} 0$ *such that* $(\sqrt{n}h_n)^{-1} = O_P(1)$, *where the convergence is under* $P = P_0$, *we have*

$$-2\frac{pL_n(\check{\theta}_n + h_n v_n) - pL_n(\check{\theta}_n)}{nh_n^2} \xrightarrow{\mathsf{P}} v' \tilde{I}_{\theta_0,\eta_0} v.$$

Corollary 3 can be used for hypothesis testing and confidence region construction for θ_0, while Corollary 4 can be used to obtain consistent, numerical estimates of $\tilde{I}_{\theta_0,\eta_0}$.

The purpose of the remainder of this section is to present and verify reasonable regularity conditions for (7) to hold.

To begin with, we will need an approximately least-favorable submodel $t \mapsto \eta_t(\theta, \eta)$ that satisfies Conditions (2) and (3).

Define

$$\ddot{\ell}(t, \theta, \eta) \equiv \left(\frac{\partial}{\partial t} \right) \dot{\ell}(t, \theta, \eta)$$

and

$$\hat{\eta}_\theta \equiv \arg\max_\eta L_n(\theta, \eta).$$

Assume that for any possibly random sequence $\tilde{\theta}_n \xrightarrow{\text{P}} \theta_0$, we have

$$\hat{\eta}_{\tilde{\theta}_n} \xrightarrow{\text{P}} \eta \text{ and} \tag{10}$$

$$P_0\dot{\ell}(\theta_0, \tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n}) = o_{P_0}(\|\tilde{\theta}_n - \theta_0\| + n^{-1/2}). \tag{11}$$

We are now ready to present the main theorem:

THEOREM 1. *Assume the following:*

- *Conditions (2), (3), (10) and (11) are satisfied;*
- *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 $(\theta_0, \theta_0, \eta_0)$ *for* P_0*-almost every* X*;*

- *for some neighborhood* V *of* $(\theta_0, \theta_0, \eta_0)$*, the class of functions*

$$\mathcal{F}_1 \equiv \{\dot{\ell}(t, \theta, \eta) : (t, \theta, \eta) \in V\}$$

is P_0*-Donsker with square-integrable envelope function; and*

- *the class of functions*

$$\mathcal{F}_2 \equiv \{\ddot{\ell}(t,\theta,\eta) : (t,\theta,\eta) \in V\}$$

is P_0-Glivenko-Cantelli and bounded in $L_1(P_0)$.

Then (7) holds.

We can readily verify the conditions of this theorem for several models:

- the Cox model for right censored data;
- the Cox model for current status data;
- the proportional odds model under right-censoring;
- the partly-linear logistic regression model;
- a case-control model with a missing covariate;
- a shared gamma-frailty model under right-censoring;
- an interesting semiparametric mixture model.

Some of these examples are in the book and some are in Murphy and van der Vaart (2000).