SEQUENTIAL METHODS IN STATISTICS
BANACH CENTER PUBLICATIONS, VOLUME 16
PWN-POLISH SCIENTIFIC PUBLISHERS
WARSAW 1985

SEQUENTIAL LIKELIHOOD RATIO TESTS

STURE HOLM

Department of Mathematics, Chalmers University of Technology, Göteborg, Sweden

Introduction

In this paper we will discuss the problem of testing hypotheses concerning one main parameter in the presence of a nuisance parameter by sequential tests based on independent identically distributed random variables, available in successive experiments of the same kind. Such problems are often handled by applying invariance principles, but there have also been proposed appealing likelihood methods. We will concentrate here on this second type of methods.

Let the main parameter be denoted by γ, the $(k-1)$-dimensional nuisance parameter by δ and the density of the independent random variables with respect to some measure $\mu(x)$ by $f(x; \gamma, \delta)$, and suppose that the parameter $\theta = (\gamma, \delta^T)^T$ belongs to a parameter set $\Omega \subseteq R^k$.

Bartlett (1946) proposed a sequential likelihood ratio test of $\gamma = \gamma_0$ against $\gamma = \gamma_1$ based on the log-likelihood differences

$$Y_m^{(B)}(\gamma_1) = \sum_{i=1}^{m} \left(\ln f(x_1; \gamma_1, \hat{\delta}_1^{(m)}) - \ln f(x_i; \gamma_0, \hat{\delta}_0^{(m)})\right) \tag{1}$$

where $\hat{\delta}_0^{(m)}$ and $\hat{\delta}_1^{(m)}$ are maximum likelihood estimates of δ under the restrictions $\gamma = \gamma_0$ and $\gamma = \gamma_1$ respectively obtained from the sample $X_1, \ldots, X_m$, and $X_1, X_2, \ldots$ are the successively available random variables. As long as this log-likelihood ratio stays between two limits $b < 0$ and $a > 0$ experimentation is continued, if it exceeds the limit b experimentation terminates and $\gamma = \gamma_0$ is accepted and if it exceeds the limit a experimentation terminates and $\gamma = \gamma_0$ is rejected in favour of $\gamma = \gamma_1$.

Cox (1963) proposed a similar procedure which makes use of the ordinary (unrestricted) maximum likelihood estimator instead of two separate ones. The test of $\gamma = \gamma_0$ against $\gamma = \gamma_1$ is based on the log-likelihood ratio

$$Y_m^{(C1)}(\gamma_1) = \sum_{i=1}^{m} \left(\ln f(x_i, \gamma_1, \hat{\delta}^{(m)}) - \ln f(x_i, \gamma_0, \hat{\delta}^{(m)})\right) \tag{2}$$

where $\hat{\theta}^{(m)} = (\hat{\gamma}^{(m)}, \hat{\delta}^{(m)T})^T$ is the usual maximum likelihood estimate of the parameter $\theta = (\gamma, \delta^T)^T$ calculated for the m first random variables. The rules for terminating experimentation as well as the rules for accepting and rejecting the hypothesis $\gamma = \gamma_0$ are of the same type as in the Bartlett procedure. The simplification by using only one maximum likelihood estimator in the summands is reduced by the fact that the dimension of the estimated parameter is one unit larger here than in the Bartlett test. It is of course possible to use the Cox type of procedure with the simpler restricted maximum likelihood estimator $\hat{\delta}_0^{(m)}$ replacing the ordinary maximum likelihood estimator $\hat{\delta}^{(m)}$ in the summands. But the success of such a replacement depends on the properties of the obtained tests.

For one-parameter problems Berk (1975a) showed that a sequential test based on the sum of derivatives of the log-likelihood function has the property of maximizing the derivative of the power function at the boundary of the hypothesis. This suggests using in the multiparameter case a procedure for testing $\gamma = \gamma_0$ against $\gamma > \gamma_0$ based on the log-likelihood derivative sums

$$Y_m^{(D)} = \sum_{i=1}^{m} \left[\frac{\partial}{\partial \gamma} \ln f(x_i, \gamma, \delta) \right]_{\gamma = \gamma_0,\ \delta = \hat{\delta}^{(m)}} \tag{3}$$

where $\hat{\delta}^{(m)}$ is either the δ-part of the ordinary maximum likelihood estimator $\hat{\theta}^{(m)} = (\hat{\gamma}^{(m)}, \hat{\delta}^{(m)T})^T$ of the parameter $\theta = (\gamma, \delta^T)^T$ or the maximum likelihood estimator of δ under the restriction $\gamma = \gamma_0$. Termination rules, acceptance rules and rejectance rules of the same types as in the Bartlett and Cox tests should be used.

Obviously there are a number of possibilities to construct sequential likelihood procedures in the multiparameter case. A natural question poses itself. What are the properties of the different tests? It is known that the approximate power functions of the Bartlett and Cox tests can be considerably different when the test limits are the same. On the other hand the approximative expected sample sizes can also be considerably different. A fair comparison of the tests ought to be made by holding some of the power or expected sample size properties fixed and make a comparison of others. See for example Berk (1975b). How can such comparisons be made in the multiparameter case? How are the properties changed by reparametrization? How are the properties affected by allowing the test limits to depend on estimates of the nuisance parameter? What are the possibilities of steering the properties of the tests by reparametrization or using test limits depending on estimates of the nuisance parameter? Are there any optimality properties of those tests?

Our aim is to develop the asymptotic theory far enough to answer the above questions in terms of asymptotic properties. We will do this by proving first some results of weak convergence of suitable random processes defined by the likelihood function and then use these results to obtain asymptotic properties for the different tests. The asymptotic properties of the Cox test are also studied by Kohlberger (1978) but the formulations and methods differ from ours.

Breslow (1969) has given a general framework of weak convergence theory in which a number of asymptotic likelihood ratio test problems may be solved. It requires however for each problem a reduction to two sequences of one-dimensional random variables converging weakly to a random process and a nonrandom function. We intend in this paper to get general asymptotic properties without special investigations for each problem.

2. Weak convergence of a log-likelihood process on bounded sets

It was pointed out by Cox (1963) that the test statistic sequence, upon which his test is based, could be approximated by a simple random walk. A proper limit theorem can be formulated for the weak convergence of a normalized sequence of test statistics to a Brownian motion process. And such a limit theorem would give us the asymptotic power function and asymptotic expected sample size function of the same simple form as in a sequential probability ratio test for drift in a Brownian motion. This applies not only to the Cox tests, but also to the Bartlett test and the tests based on the derivative of the log-likelihood.

We will formulate and prove this type of limit theorems later by using the weak convergence of a log-likelihood process to a multidimensional Brownian motion, studied in this section.

The asymptotic properties are obtained when the sample sizes increase and the parameter at the same time approaches a point at the boundary between the hypothesis and the alternative. If n denotes the (increasing) sample size corresponding to the time $t = 1$ for the limiting process, the suitable parameter convergence can be written

$$\theta_n = (\gamma, \delta^T)_n^T = \theta_0 + n^{-1/2} \cdot \Theta = (\gamma_0, \delta_0^T)^T + n^{-1/2} \cdot (\Gamma, \Delta^T)^T \tag{4}$$

where $\theta_0 = (\gamma_0, \delta_0^T)^T$ is a fixed parameter point in the interior Ω^0 of Ω on the boundary between the hypothesis and alternative and $\Theta = (\Gamma, \Delta^T)^T$ is a normalized parameter.

We define a *sequence* of $(k+1)$-dimensional random processes $X_n(t, \Theta)$ for $n = 1, 2, 3, \ldots$ on the set $0 \leqslant t \leqslant \infty$, $\Theta \in R^k$ in the following way. Its first component $X_n^{(1)}(t, \Theta)$ is defined in the time points $t = m/n$ as a sum of differences by

$$X_n^{(1)}\left(\frac{m}{n}, \Theta\right) = \sum_{i=1}^{m} \left(\ln f(X_i, \theta_0 + n^{-1/2}\Theta) - \ln f(X_i, \theta_0)\right) \tag{5}$$

and between those points it is linearly interpolated. Its second component $X_n^{(2)}(t, \Theta)$ is defined in the time points $t = m/n$ as a sum of partial derivatives by

$$X_n^{(2)}\left(\frac{m}{n}, \Theta\right) = \frac{\partial}{\partial \Gamma} X_n^{(1)}\left(\frac{m}{n}, \Theta\right) = \sum_{i=1}^{m} \frac{\partial}{\partial \Gamma} \ln f(X_i, \theta_0 + n^{-1/2}\Theta) \tag{6}$$

and between those points it is linearly interpolated. The remaining $k-1$ components are analogously defined in the time points $t = m/n$ as sums of partial derivatives of the log-likelihood function with respect to the $k-1$ components of Δ and between those points by linear interpolation.

It may seem too complicated to introduce a $(k+1)$-dimensional random process in order to obtain results for one-dimensional processes, but it has the advantage that it easily gives asymptotic results for the different tests in the same framework without voluminous special investigations for the different tests.

For the Bartlett test statistics $Y_m^{(B)}(\gamma_1)$, $m = 1, 2, 3, \ldots$, we can define a suitable sequence $X_n^{(B)}(t)$, $n = 1, 2, 3, \ldots$, of random processes by defining $X_n^{(B)}(t)$ generally by

$$X_n^{(B)}(t) = \sup_{\Gamma = \Gamma_1, \Delta \in R^{k-1}} X_n^{(1)}(t, \Theta) - \sup_{\Gamma = 0, \Delta \in R^{k-1}} X_n^{(1)}(t, \Theta) \tag{7}$$

which for $t = m/n$ reduces to

$$X_n^{(B)}\left(\frac{m}{n}\right) = Y_m^{(B)}(\gamma_0 + n^{-1/2}\Gamma_1). \tag{8}$$

The alternative $\gamma = \gamma_1$ thus here approaches the hypothesis according to

$$\gamma_1 = \gamma_0 + n^{-1/2}\Gamma_1. \tag{9}$$

We call the function generating $X_n^{(B)}(t)$ from $X_n(t, \Theta)$ the *Bartlett test function* or *Bartlett mapping*

The processes $X_n(t, \Theta)$ are defined for $0 < t < \infty$ and for $\Theta \in R^k$. We will later discuss this unbounded definition set. But we begin by considering the processes on a bounded definition set of the type $T \times D$ where $T = [t_0, t_1]$ for some t_0 and t_1 with $0 < t_0 < t_1 < \infty$ and $D = \{\Theta: |\Theta| \leqslant d_1\}$ for some $d_1 > 0$, where $|\Theta|$ is the ordinary Euclidean norm of the k-dimensional vector Θ.

If the log-likelihood function has continuous partial derivatives the process $\{X_n(t, \Theta): t \in T, \Theta \in D\}$ belongs to the set $C^{(k+1)}(T \times D)$ of continuous functions on the set $T \times D$. In $C^{(k+1)}(T \times D)$ we introduce the supremum metric defined as the supremum over $T \times D$ of the Euclidean norm of the function value.

When we consider the process $X_n(t, \Theta)$ only for $t \in T$, $\Theta \in D$ the Bartlett mapping generates

$$X^{(B)}_{n,d_1}(t) = \sup_{\Gamma=\Gamma_1, |\Theta| \leqslant d_1} X^{(1)}_n(t, \Theta) - \sup_{\Gamma=0, |\Theta| \leqslant d_1} X^{(1)}_n(t, \Theta) \tag{10}$$

on $t \in T$.

The possibilities to obtain asymptotic results for the Bartlett test from weak convergence of processes $X_n(t, \Theta)$ depend on the following simple lemma.

LEMMA 1. *The Bartlett test function on $C^{(k+1)}(T \times D)$ has its values in the set $C(T)$ of continuous functions on T. It is also continuous when $C(T)$ is endowed with supremum metric.*

Proof. It is easily seen that the Bartlett mapping applied to a function in $C^{(k+1)}(T \times D)$ gives a continuous function on T.

To see that the Bartlett mapping is also continuous let $X(t, \Theta)$ and $Y(t, \Theta)$ be two functions in $C^{(k+1)}(T \times D)$ with a distance $|Y(t, \Theta) - X(t, \Theta)|$ which is smaller than ε, and let $X^{(B)}(t)$ and $Y^{(B)}(t)$ be their images in $C(T)$. Then obviously the distance $|Y^{(B)}(t) - X^{(B)}(t)|$ between the images satisfies $|Y^{(B)}(t) - X^{(B)}(t)| \leqslant 2\varepsilon$. ∎

The other tests can be treated in a similar way. For the original Cox test we define the maximum likelihood estimator $\hat{\Theta}_{n,t}$ as a value of Θ maximizing $X^{(1)}_n(t, \Theta)$, with some additional rule making it unique. The Cox test 1 mapping is now defined as the function giving for $X_n(t, \Theta)$ the result

$$X^{(C1)}_n(t) = X^{(1)}_n(t, (\Gamma_1, \hat{\Delta}^T_{n,t})^T) - X^{(1)}_n(t, (0, \hat{\Delta}^T_{n,t})^T). \tag{11}$$

where $\hat{\Delta}_{n,t}$ is the Δ-component of $\hat{\Theta}_{n,t}$. For $t = m/n$, $m = 1, 2, 3, \ldots$, this relates to the sequence $Y^{(C1)}_m(\gamma_1)$, $m = 1, 2, 3, \ldots$, in formula (2) by

$$X^{(C1)}_n(t) = Y^{(C1)}_m(\gamma_0 + n^{-1/2}\Gamma_1). \tag{12}$$

For the modified Cox test based on a restricted maximum likelihood estimate we define analogously the Cox test 2 mapping as giving for $X_n(t, \Theta)$ the result

$$X^{(C2)}_n(t) = X^{(1)}_n(t, (\Gamma_1, \hat{\Delta}^T_{n,t})^T) - X^{(1)}_n(t, (0, \hat{\Delta}^T_{n,t})^T) \tag{13}$$

where $\hat{\Delta}_{n,t}$ is a uniquely defined value of Δ maximizing $X^{(1)}(t, (0, \Delta^T)^T)$.

The two Cox test mappings can be defined either for functions $X_n(t, \Theta)$, $0 \leqslant t < \infty$, $\Theta \in R^k$, or for functions $X_n(t, \Theta)$, $t \in T$, $\Theta \in D$, and in the latter case the maximum likelihood estimators are also restricted to the set $|\Theta| \leqslant d_1$.

Asymptotic properties for the two Cox tests can be obtained from weak convergence results for $X_n(t, \Theta)$ by help of the following simple lemma.

LEMMA 2. *The two Cox mappings are continuous and take values in $C(T)$ for each subset of $C^{(k+1)}(T \times D)$ where the second Θ-derivative exist and* *the eigenvalues of the second derivative matrix are bounded above by a negative number.*

Proof. Consider the subset of $C^{(k+1)}(T\times D)$ where the second Θ-derivative is bounded above by $-\lambda < 0$. If $X_1(t, \Theta)$ and $X_2(t, \Theta)$ belong to this set and $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are the corresponding maximum likelihood estimates and $|X_2(t, \Theta) - X_1(t, \Theta)| < \varepsilon$ then we get

$$|\hat{\Theta}_2 - \hat{\Theta}_1|^2 \leqslant \frac{\varepsilon}{\lambda} \tag{14}$$

by considering the derivative components of $X_2(t, \Theta)$ and $X_1(t, \Theta)$.

The statements in the lemma now follow for the Cox test 1 mapping from the continuity of the functions in $C^{(k+1)}(T\times D)$ and of the supremum operations.

If $\hat{\Delta}_1$ and $\hat{\Delta}_2$ are the maximum likelihood estimates of Δ under the restriction $\Gamma = 0$ in $X_1(t, \Theta)$ and $X_2(t, \Theta)$ we have analogously

$$|\hat{\Delta}_2 - \hat{\Delta}_1| \leqslant \frac{\varepsilon}{\lambda} \tag{15}$$

which gives the stated results for the Cox test 2 mapping. ■

Finally for the differentiation test 1 based on the ordinary maximum likelihood estimator we define a mapping generating for $X_n(t, \theta) \in C^{(k+1)}(T\times \times D)$ the function

$$X_n^{(D1)}(t) = X_n^{(2)}\big(t, (0, \hat{\Delta}_{n,t}^T)\big) \tag{16}$$

where $\hat{\Theta}_n = (\hat{\Gamma}_n, \hat{\Delta}_{n,t}^T)^T$ is the maximum likelihood estimator in $C^{(k+1)}(T\times D)$. And for the differentiation test 2 based on the restricted maximum likelihood estimator we introduce a mapping generating for $X_n(t, \Theta) \in C^{(k+1)}(T\times D)$ the function

$$X_n^{(D2)}(t) = X_n^{(2)}\big(t, (0, \hat{\Delta}_{n,t}^T)^T\big) \tag{17}$$

where $\hat{\Delta}_{n,t}^T$ is the restricted maximum likelihood estimator maximizing $X_n^{(1)}\big(t, (0, \Delta^T)^T\big)$ for $|\Delta| \leqslant d_1$.

Observe that in this case the relation to the sequences $Y_m^{(D1)}$, $m = 1, 2, 3, \ldots$, and $Y_m^{(D2)}$, $m = 1, 2, 3, \ldots$, given by formula (3) for the two estimators is given by

$$X_n^{(D1)}\left(\frac{m}{n}\right) = n^{-1/2}\, Y_m^{(D1)} \tag{18}$$

and

$$X_n^{(D2)}\left(\frac{m}{n}\right) = n^{-1/2}\, Y_m^{(D2)}. \tag{19}$$

LEMMA 3. *The two differentiation test mappings are continuous and take values in $C(T)$ for each subset of $C^{(k+1)}(T\times D)$ where the second Θ-derivative exists and the eigenvalues of the second derivative matrix are bounded above by a negative number.*

Proof. From the proof of Lemma 2 we get the continuity of the ordinary and restricted maximum likelihood estimates. The statements in the lemma now follow from the continuity of the functions in $C^{(k+1)}(T\times D)$. ∎

By the above three lemmas we can transfer weak convergence results for the processes $X_n(t, \Theta)$, $n = 1, 2, \ldots$, on $C^{(k+1)}(T\times D)$ to weak convergence results for the different test statistics processes on $C(T)$. See, e.g., Billingsley (1968), Theorem 5.1. Thus we have a good motivation for studying weak convergence of the processes $X_n(t, \Theta)$ on $C^{(k+1)}(T\times D)$. The problems of extending the results to unbounded time and parameter sets will be discussed in the next section.

We begin our study by some general considerations for the space $C^{(k+1)}(T\times D)$.

LEMMA 4. *The space $C^{(k+1)}(T\times D)$ is complete and separable with respect to the supremum metric.*

Proof. A fundamental sequence of functions gives a fundamental sequence in each point. They converge uniformly to a continuous function which proves the completeness. Separability follows from the Stone–Weierstrass theorem (see, e.g., Hewitt and Stromberg (1965), p. 95) which states that the polynomials constitute a dense subset in $C^{(k+1)}(T\times D)$. ∎

Note. Another dense subset which is more closely connected to our convergence problems can be constructed in the following way. For all natural numbers v consider nets with side $1/v$ and functions with rational values in the net points. Within the cubes of the net the value is obtained by summing over the cube corner values multiplied by factors of the type $v_1 \cdot v_2 \cdot \ldots \cdot v_{k+1}$ where v_l are linear functions varying between 0 and 1 over the lth coordinate interval and attaining the value 1 in the corner the factor belongs to.

A sequence P_n, $n = 1, 2, 3, \ldots$, of probability measures on a metric space C is said to *converge weakly* to a probability measure P on C if

$$\lim_{n\to\infty} \int_C g\,dP_n = \int_C g\,dP \tag{20}$$

for every bounded continuous realvalued function g on C.

A set Π of probability measures P on C is said to be *tight* if for every $\varepsilon > 0$ there exist a compact set K_ε such that $P(K_\varepsilon) > 1-\varepsilon$ for each $P\in\Pi$.

We can now prove a fundamental theorem of weak convergence in $C^{(k+1)}(T\times D)$.

THEOREM 1. *Let* P_n, $n = 1, 2, 3, \ldots$, *and* P *be probability measures on* $C^{(k+1)}(T\times D)$. *If the finite dimensional distributions of* P_n, $n = 1, 2, 3, \ldots$, *converge to those of* P *and* $\{P_n\colon n = 1, 2, 3, \ldots\}$ *is tight then* P_n, $n = 1, 2, 3, \ldots$, *converges weakly to* P.

Proof. By Lemma 4 the space $C^{(k+1)}(T\times D)$ is complete and separable. For such spaces tightness implies relative compactness, i.e., each subsequence of $\{P_n\colon n = 1, 2, 3, \ldots\}$ contains a weakly convergent subsequence. See, e.g., Billingsley (1968), p. 37. Let this limit probability measure be denoted by Q. But the finite dimensional distributions are limit determining and thus $P = Q$. The theorem now follows from Theorem 2.3 in Billingsley (1968), p. 16. ∎

The tightness condition in Theorem 1 is possible to substitute by an equivalent condition which is easier to verify.

LEMMA 5. *The sequence* $\{P_n\colon n = 1, 2, 3, \ldots\}$ *is tight if and only if*

(i) *For each* $\eta > 0$ *there exists an* $a > 0$ *such that*

$$P_n(\{x\colon |x(0)| > a\}) \leqslant \eta, \qquad \forall n \geqslant 1.$$

(ii) *For each* $\varepsilon > 0$ *and* $\eta > 0$ *there exists* δ, $0 < \delta < 1$ *such that*

$$P_n(\{x\colon w_x(\delta) \geqslant \varepsilon\}) \leqslant \eta, \qquad \forall n \geqslant 1,$$

where $w_x(\delta)$ *is the continuity modulus of the function* x.

Proof. By the Ascoli–Arzela theorem (see, e.g., Simmons (1963), p. 126) a closed subspace of $C^{(k+1)}(T\times D)$ is compact if and only if it is bounded and equicontinuous. The lemma now follows exactly like the proof of the corresponding lemma for the case $C[0, 1]$ in Billingsley (1968), p. 55. ∎

So far we have discussed weak convergence for a sequence $\{P_n\colon n = 1, 2, 3, \ldots\}$ of probability measures. In our orginal problem we rather have to study a sequence $\{X_n(t, \Theta)\colon n = 1, 2, 3, \ldots\}$ of random processes. Formulated in terms of weak convergence of random processes on the set $T\times D$ we have the following lemma.

LEMMA 6. *Let* $X_n(t, \Theta)$, $n = 1, 2, 3, \ldots$, *and* $X(t, \Theta)$ *be random processes in* $C^{(k+1)}(T\times D)$. *If the finite-dimensional distributions of* X_n *converge to those of* X *and*

$$\lim_{h\to 0}\sup_n P\{\sup_{\left|\binom{t}{\Theta}-\binom{t'}{\Theta'}\right|\leqslant h} |X_n(t, \Theta) - X_n(t', \Theta')| > \varepsilon\} = 0 \tag{21}$$

for every $\varepsilon > 0$ *then* $X_n(t, \Theta)$, $n = 1, 2, 3, \ldots$, *converges weakly to* $X(t, \Theta)$, i.e., *the corresponding probability measures converge weakly.*

Proof. Literally equal to the proof of Theorem 1 in Gikhman and Skorokhod (1969), p. 449, for the $C[0, 1]$ case. ∎

In the final step we are now going to use this lemma for proving weak convergence of our likelihood processes. We need some conditions, which will now be introduced. The first condition involves existence of second order partial derivatives which can hardly be avoided since the final weak convergence result will include those derivatives. The existence of the third order partial derivative and boundedness conditions might be avoided or changed.

CONDITION 1 (existence and boundedness of partial derivatives). *There exist partial derivatives of the first three orders with respect to θ-components of the* log-*likelihood function* $\ln f(x, \theta)$ *for every $\theta \in \Omega$ a.e. with respect to μ.*

Further for each $\theta_0 \in \Omega^0$ on the boundary between the hypothesis and the alternative there exist a neighbourhood Ω_{θ_0} and functions $F_1(x)$, $F_2(x)$ and $F_3(x)$ such that the maximum numerical value of the first, second and third order partial derivatives over Ω_{θ_0} are bounded by $F_1(x)$, $F_2(x)$ and $F_3(x)$. There exist constants K_1, K_2 and K_3 such that

(i) $\int_{-\infty}^{\infty} F_1(x) f(x, \theta) d\mu(x) < K_1, \quad \forall \theta \in \Omega_{\theta_0},$

(ii) $\int_{-\infty}^{\infty} F_2(x) f(x, \theta) d\mu(x) < K_2, \quad \forall \theta \in \Omega_{\theta_0},$

(iii) $\int_{-\infty}^{\infty} F_3(x) f(x, \theta) d\mu(x) < K_3, \quad \forall \theta \in \Omega_{\theta_0}.$

Next condition concerns uniform convergence of means of partial derivatives of the log-likelihood function. We introduce here some notation which will be used throughout. By the notation $\frac{\partial \ln f(x; \theta)}{\partial \theta}$ we mean the k-dimensional vector whose components are the first order partial derivatives with respect to the different components of θ and by $\frac{\partial^2 \ln f(x; \theta)}{\partial \theta^2}$ we mean the $k \times k$ matrix whose (i, j)-element is equal to the second order partial derivative with respect to the ith and jth component of θ.

By Condition 1 the expectations $E\left[\left[\frac{\partial}{\partial \theta} \ln f(X, \theta)\right]_{\theta=\theta_0}\right]$ and $E\left[\left[\frac{\partial^2}{\partial \theta^2} \ln f(X, \theta)_{\theta=\theta_0}\right]\right]$ exist for any true parameter θ^* of the random variable X. Further the covariance matrix of $\left[\frac{\partial}{\partial \theta} \ln f(X, \theta)\right]_{\theta=\theta_0}$ exists and equals $E\left[-\left[\frac{\partial^2}{\partial \theta^2} \ln f(X, \theta)_{\theta=\theta_0}\right]\right]$.

In the case when the parameter in the distribution of X equals θ_0 we denote this expectation by $I(\theta_0)$ and generally

$$I(\theta) = E\left[-\left[\frac{\partial^2}{\partial\theta^2}\ln f(X,\theta)\right]\right]$$

when the true parameter of X is θ and the second derivative matrix is calculated in the same point.

If the true parameter of $X^{(n)}$ is $\theta_0+n^{-1/2}\,\Theta^*$ then from condition 1 and bounded convergence we get

$$\lim_{n\to\infty} E\left[n^{1/2}\left[\frac{\partial}{\partial\theta}\ln f(X^n,\theta)\right]_{\theta=\theta_0}\right] = I(\theta_0)\,\Theta^* \tag{22}$$

and

$$\lim_{n\to\infty} E\left[-\left[\frac{\partial^2}{\partial\theta^2}\ln f(X^{(n)},\theta)\right]_{\theta=\theta_0}\right] = I(\theta_0). \tag{23}$$

The norm of a matrix will be denoted by ordinary modulus signs around the matrix notation and the norm is to be interpreted as the supremum length of the resulting vector when the matrix is multiplied by unit vectors.

CONDITION 2 (uniform convergence). *For each $\theta_0 \in \Omega^0$ on the boundary between the hypothesis and the alternative there exists a neighbourhood Ω_{θ_0} such that $I(\theta)$ is continuous in Ω_{θ_0} and*

$$\lim_{n\to\infty} P_\theta\left(\left|\frac{1}{n}\sum_{j=1}^{n}\frac{\partial^2}{\partial\theta^2}\ln f(X_j;\theta)+I(\theta)\right|>\varepsilon\right)=0$$

uniformly for $\theta\in\Omega_{\theta_0}$ for each $\varepsilon>0$.

When proving the weak convergence of the log-likelihood process we use Lemma 6, i.e., we prove that the tightness condition (21) is satisfied and that the finite-dimensional distributions converge.

LEMMA 7. *If Conditions 1 and 2 are satisfied then for each $\varepsilon>0$*

$$\lim_{h\to 0}\sup_n P\{\sup_{\substack{\left|\binom{t}{\Theta}-\binom{t'}{\Theta'}\right|\le h,\\ t,t'\in T,\,\Theta,\Theta'\in D}} |X_n(t',\Theta')-X_n(t,\Theta)|>\varepsilon\}=0.$$

Proof. Since the X_n functions are continuous the result in the lemma is equivalent to a corresponding result with the supremum over n substituted by limes superior in n. The result will thus follow if we show that for each $\varepsilon>0$

$$\lim_{h\to 0}\overline{\lim_{n\to\infty}}\, P(\sup_{\left|\binom{t}{\Theta}-\binom{t'}{\Theta'}\right|\le h} |X_n(t',\Theta')-X_n(t,\Theta)|>\varepsilon)=0.$$

Now consider first the derivative components of $X_n(t, \Theta)$. We denote by $X'_n(t, \Theta)$ the k-dimensional vector whose components are the k-derivative components of $X_n(t, \Theta)$. We have by the triangle inequality

$$|X'_n(t', \Theta')-X'_n(t, \Theta)| \leqslant |X'_n(t', \Theta')-X'_n(t, \Theta')|+|X'_n(t, \Theta')-X'_n(t, \Theta)|$$

and we will show that for each $\varepsilon > 0$

$$\lim_{h\to 0}\overline{\lim_{n\to\infty}} P\Big(\sup_{|t-t'|\leqslant h, |\Theta'|\leqslant d_1} |X'_n(t', \Theta')-X'_n(t, \Theta')| > \varepsilon\Big) = 0 \tag{24}$$

and

$$\lim_{h\to 0}\overline{\lim_{n\to\infty}} P\Big(\sup_{|\Theta-\Theta'|\leqslant h, t_0\leqslant t\leqslant t_1} |X'_n(t, \Theta')-X'_n(t, \Theta)| > \varepsilon\Big) = 0. \tag{25}$$

We begin by showing formula (24). Supposing that $t' > t$ and denoting

$$m_1 = [nt]+1$$

and

$$m_2 = [nt']$$

we get

$$\begin{aligned} &|X'_n(t', \Theta)-X'_n(t, \Theta')| \\ &\quad\leqslant \left|\sum_{j=m_1}^{m_2} n^{-1/2}\left[\frac{\partial}{\partial\theta}\ln f(X_j^{(n)}, \theta)\right]_{\theta=\theta_0+n^{-1/2}\Theta'}\right|+ \\ &\quad\quad+\left|n^{-1/2}\left[\frac{\partial}{\partial\theta}\ln f(X_{m_1-1}^{(n)}, \theta)\right]_{\theta=\theta_0+n^{-1/2}\Theta'}\right|+ \\ &\quad\quad+\left|n^{-1/2}\left[\frac{\partial}{\partial\theta}\ln f(X_{m_2+1}^{(n)}, \theta)\right]_{\theta=\theta_0+n^{-1/2}\Theta'}\right|. \end{aligned} \tag{26}$$

Here by series expansion

$$\begin{aligned} &\left|\sum_{j=m_1}^{m_2} n^{-1/2}\left[\frac{\partial}{\partial\theta}\ln f(X_j^{(n)}, \theta)\right]_{\theta=\theta_0+n^{-1/2}\Theta'}\right| \\ &\leqslant \left|\sum_{j=m_1}^{m_2} n^{-1/2}\left[\frac{\partial}{\partial\theta}\ln f(X_j^{(n)}, \theta)\right]_{\theta=\theta_0}\right|+\left|\sum_{j=m_1}^{m_2} n^{-1}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X_j^{(n)}, \theta)\right]_{\theta=\theta_0}\right|\cdot|\Theta'|+ \\ &\qquad\qquad+n^{-3/2}|\Theta'|^2 k^3 \sum_{j=m_1}^{m_2} F_3(X_j^{(n)}) \end{aligned} \tag{27}$$

where F_3 is defined in Condition 1.

Now let $J_0, J_1, \ldots,$ be the intervals defined by

$$J_\nu = [\nu h, (\nu+1)h].$$

Then

$$P\Big(\sup_{|t-t'|\leqslant h, |\Theta'|\leqslant d_1} |X'_n(t,\Theta')-X'_n(t',\Theta')|>\varepsilon\Big)$$
$$\leqslant P\Big(\sup_{\substack{|t-t'|\leqslant h, |\Theta'|\leqslant d_1,\\ t,t'\in J_\nu \text{ for some } \nu}} |X'_n(t,\Theta')-X'_n(t',\Theta')|>\tfrac{1}{2}\varepsilon\Big)$$
$$\leqslant P\Big(\sup_{\substack{|t-t'|\leqslant h, |\Theta'|\leqslant d_1,\\ t=[t'/h]\cdot h}} |X'_n(t,\Theta')-X'_n(t',\Theta')|>\tfrac{1}{4}\varepsilon\Big). \qquad (28)$$

We will use this formula together with the estimates in formulas (26) and (27) to obtain (24). Note first that formula (27) applies to all three terms in the right member of (26), by using in the two final terms the special case $m_2 = m_1$.

Taking supremum over $|\Theta'|\leqslant d_1$ in the right member of (27) just means substituting $|\Theta'|$ by d_1 since everything else is independent of Θ'.

Since F_3 is a nonnegative function we get for the third term in the right member of (27)

$$P\Big(\sup_{n_0\leqslant m_1\leqslant m_2\leqslant n_1} n^{-3/2} d_1^2 k^3 \sum_{j=m_1}^{m_2} F_3(x_j^{(n)})>\varepsilon'\Big)$$
$$\leqslant P\left((n_1-n_0+1)^{-1}\sum_{j=n_0}^{n_1} F_3(X_j^{(n)})>\varepsilon' d_1^{-2}k^{-3}\frac{n^{3/2}}{n_1-n_0+1}\right)$$
$$\leqslant K_3(\varepsilon')^{-1} d_1^2 k^3 n^{-3/2}(n_1-n_0+1) \qquad (29)$$

where $n_0=[nt_0]$ and $n_1=[nt_1]+1$. Thus this term tends to 0 as n tends to ∞ for each $\varepsilon'>0$.

Consider now the middle term in the right member of (27) for t' in the subinterval J_0, i.e.,

$$\sup_{n_0+1\leqslant m_2\leqslant n_0+[nh]+1}\left|\sum_{j=n_0+1}^{m_2} n^{-1}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X_j^{(n)},\theta)\right]_{\theta=\theta_0}\right|.$$

There exist a number $M>0$ and for each ε'' a number m'' such that

$$P\left(m^{-1}\left|\sum_{j=1}^{m_2}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X_j^{(n)},\theta)\right]_{\theta=\theta_0}\right|\leqslant M \quad \text{for each } m\geqslant m''\right)\geqslant 1-\varepsilon'' \qquad (30)$$

uniformly for the true parameter θ^* in a neighbourhood of θ_0, i.e.,

$$P\left(n^{-1}\left|\sum_{j=n_0+1}^{m_2}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X_j^{(n)},\theta)\right]_{\theta=\theta_0}\right|\leqslant M\frac{m_2-n_0}{n}\leqslant 2Mh \right.$$
$$\left.\text{for each } m_2,\ n_0+m''\leqslant m_2\leqslant n_0+[nh]+1\geqslant 1\right)\geqslant 1-\varepsilon''. \qquad (31)$$

Further

$$P\Big(n^{-1}\Big|\sum_{j=n_0+1}^{m_2}\Big[\frac{\partial^2}{\partial\theta^2}\ln f(X_j^{(n)},\theta)\Big]_{\theta=\theta_0}\Big|\leqslant 2Mh \quad \text{for each } m_2,\ n_0+1\leqslant m_2\leqslant n_0+m''\Big)$$

$$\geqslant P\Big(\sum_{j=n_0+1}^{n_0+m''} F_2(X_j^{(n)})\leqslant 2Mhnk^{-2}\Big)\geqslant 1-\frac{m''K_2}{2Mhn}k^2. \qquad (32)$$

The same estimates are valid in all subintervals J_v, $v=0, 1, 2, \ldots, \left[\frac{t_1-t_0}{h}\right]$ and thus

$$P\Big(\sup_{|m_2-m_1|\leqslant[nh]+2}\Big|n^{-1}\sum_{j=m_1}^{m_2}\Big[\frac{\partial}{\partial\theta^2}\ln f(X_j^{(n)},\theta)\Big]_{\theta=\theta_0}\Big|>8Mh\Big)$$

$$\leqslant\Big(\Big[\frac{t_1-t_0}{h}\Big]+1\Big)\Big(\varepsilon''+\frac{m''K_2k^2}{2Mhn}\Big). \qquad (33)$$

This gives

$$\overline{\lim_{n\to\infty}}\, P\Big(\sup_{|m_2-m_1|\leqslant[nh]+2}\Big|n^{-1}\sum_{j=m_1}^{m_2}\Big[\frac{\partial^2}{\partial\theta^2}\ln f(X_j^{(n)},\theta)\Big]_{\theta=\theta_0}\Big|>\varepsilon\Big)=0 \qquad (34)$$

for each $\varepsilon>0$ and $h<\frac{\varepsilon}{8M}$ since ε'' can be chosen freely. Thus it will be shown that

$$\lim_{h\to 0}\overline{\lim_{n\to\infty}}\, P\Big(\sup_{|t-t'|\leqslant h, |\Theta'|\leqslant d_1}|X_n'(t',\Theta')-X_n'(t,\Theta')|>\varepsilon\Big)=0 \qquad \text{for each} \quad \varepsilon<0$$

if we show that $X_n'(t, 0)$ is tight on the interval $t_0\leqslant t\leqslant t_1$. The investigation of this problem will be postponed to the end of the proof.

Next we will show that

$$\lim_{h\to 0}\overline{\lim_{n\to\infty}}\, P\Big(\sup_{|\Theta-\Theta'|\leqslant h, t_0\leqslant t\leqslant t_1}|X_n'(t,\Theta')-X_n'(t,\Theta)|>\varepsilon\Big)=0 \qquad (35)$$

for each $\varepsilon>0$. By series expansion and Condition 1 (iii) we get

$$|X_n'(t,\Theta')-X_n'(t,\Theta)|$$

$$\leqslant|\Theta-\Theta'|\,n^{-1}\Big|\sum_{j=n_0+1}^{m}\Big[\frac{\partial^2}{\partial\theta^2}\ln f(X_j^{(n)},\theta)\Big]_{\theta=\theta_0}\Big|+$$

$$+|\Theta-\Theta'|\,n^{-1}\Big|\Big[\frac{\partial^2}{\partial\theta^2}\ln f(X_{m+1}^{(n)},\theta)\Big]_{\theta=\theta_0}\Big|+$$

$$+|\Theta-\Theta'|\,n^{-1}\Big|\Big[\frac{\partial^2}{\partial\theta^2}\ln f(X_{n_0}^{(n)}\theta)\Big]_{\theta=\theta_0}\Big|+4d_1^2n^{-3/2}\sum_{j=n_0+1}^{m+1}F_3(X_j^{(n)}) \qquad (36)$$

where $m=[nt]$. Here the final term does not depend on Θ and Θ',

$$\sum_{j=n_0+1}^{m+1} F_3(X_j^{(n)}) \leqslant \sum_{j=n_0+1}^{[nt_1]+1} F_3(X_j^{(n)}) \tag{37}$$

and

$$P\Big(4d_1^2 n^{-3/2} \sum_{j=n_0+1}^{[nt_1]+1} F_3(X_j^{(n)}) > \varepsilon\Big) \leqslant K_3 \cdot 4d_1^2 n^{-3/2}([nt_1]+1-n_0), \tag{38}$$

i.e., the supremum of the final term tends to 0 when n tends to ∞.

In the first term the part

$$n^{-1}\left|\sum_{j=n_0+1}^{m}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X_j^{(n)}, \theta)\right]_{\theta=\theta_0}\right|$$

is independent of Θ and Θ', and when taking supremum over t (over m) we get like in the previous part of the theorem

$$\overline{\lim_{n\to\infty}} P\left(\sup_{n_0+1\leqslant m\leqslant[nt_1]+1} |\Theta-\Theta'|\, n^{-1}\left|\sum_{j=n_0+1}^{m}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X_j^{(n)}, \theta)\right]_{\theta=\theta_0}\right| > M2d_1(t_1-t_0)\right)=0. \tag{39}$$

The supremum of the middle terms is dominated by twice the supremum of the first term and formula (35) will follow.

Now consider the first component $X_n^{(1)}(t, \Theta)$ of the process $X_n(t, \Theta)$, i.e., the process which is a linear interpolation of $X_n^{(1)}\left(\frac{m}{n}, \Theta\right)$ given by

$$X_n^{(1)}\left(\frac{m}{n}, \Theta\right) = \sum_{j=1}^{m}\left(\ln f(X_j^{(n)}, \theta_0+n^{-1/2}\Theta)-\ln f(X_j^{(n)}, \theta_0)\right).$$

Also in this case we will prove tightness by proving two formulas

$$\lim_{h\to 0}\overline{\lim_{n\to\infty}} P\Big(\sup_{|t-t'|\leqslant h, |\Theta'|\leqslant d_1} |X_n^{(1)}(t', \Theta')-X_n^{(1)}(t, \Theta')| > \varepsilon\Big)=0 \tag{40}$$

and

$$\lim_{h\to 0}\overline{\lim_{n\to\infty}} P\Big(\sup_{t_0\leqslant t\leqslant t_1, |\Theta-\Theta'|\leqslant h} |X^{(1)}(t, \Theta')-X_n^{(1)}(t, \Theta)| > \varepsilon\Big)=0 \tag{41}$$

which together imply tightness.

Observing that $X_n^{(1)}(t, 0)=0$ and making a series expansion we get

$$\begin{aligned}|X_n^{(1)}(t', \Theta')-X_n^{(1)}(t, \Theta')| &= \left|\left(X_n'(t', \Theta'')-X_n'(t, \Theta'')\right)^T\Theta'\right| \qquad (42)\\ &\leqslant |X_n'(t', \Theta'')-X_n'(t, \Theta'')|\cdot|\Theta'|\end{aligned}$$

where $|\Theta''| \leqslant |\Theta'|$. It is thus seen that formula (40) will follow from tightness of $X_n'(t, \Theta)$.

Similarly

$$|X_n^{(1)}(t, \Theta') - X_n^{(1)}(t, \Theta)| = |X_n'(t, \Theta'') \cdot (\Theta' - \Theta)| \leqslant |X_n'(t, \Theta'')| \cdot |\Theta' - \Theta| \tag{43}$$

where $|\Theta''| \leqslant \max(|\Theta'|, |\Theta|)$. The formula (41) will thus also follow from tightness of $X_n'(t, \Theta)$.

It remains to prove tightness of the process $X_n'(t, 0)$ on $t_0 \leqslant t \leqslant t_1$. This process is a linear interpolation of

$$X_n'\left(\frac{m}{n}, 0\right) = n^{-1/2} \sum_{j=1}^{m} \left[\frac{\partial}{\partial \theta} \ln f(X_j^{(n)}, \theta)\right]_{\theta=\theta_0}.$$

By Conditions 1 and 2 the expectation

$$E\left[\left[\frac{\partial}{\partial \theta} \ln f(X_j^{(n)}, \theta)\right]_{\theta=\theta_0}\right]$$

exists and

$$\lim_{n \to \infty} n^{1/2} E\left[\left[\frac{\partial}{\partial \theta} \ln f(X_j^{(n)}, \theta)\right]_{\theta=0}\right] = I(\theta_0) \cdot \Theta^*$$

where $I(\theta_0)$ is positive definite. Further by Conditions 1 and 2 the covariance matrix

$$\mathrm{Cov}\left[\left[\frac{\partial}{\partial \theta} \ln f(X_j^{(n)}, \theta)\right]_{\theta=\theta_0}\right]$$

exists and

$$\lim_{n \to \infty} \mathrm{Cov}\left[\left[\frac{\partial}{\partial \theta} \ln f(X_j^{(n)}, \theta)\right]_{\theta=\theta_0}\right] = I(\theta_0).$$

The tightness of the components of $X_n'(t, 0)$ on $t_0 \leqslant t \leqslant t_1$ now follows from Theorem 1 on page 452 in Gikhman and Skorokhod (1969). With their notations we have for example for the first component

$$\xi_{ni} = \sigma_1^{-1} n^{-1/2} \left\{ \left[\frac{\partial}{\partial \theta} \ln f(X_j^{(n)}, \theta)\right]_{\theta=\theta_0}^{(1)} - E\left[\left[\frac{\partial}{\partial \theta} \ln f(X_j^{(n)}, \theta)\right]_{\theta=\theta_0}^{(1)}\right] \right\} \tag{44}$$

where

$$\sigma_1^2 = \mathrm{Var}\left[\left[\frac{\partial}{\partial \theta} \ln f(X_j^{(n)}, \theta)\right]_{\theta=\theta_0}\right] \tag{45}$$

and the Lindeberg condition is easily seen to be satisfied. The extension of the result for the time interval $0 \leqslant t \leqslant 1$ in Gikhman and Skorokhod (1969) to the time interval $t_0 \leqslant t \leqslant t_1$ is trivial.

The tightness of the whole process $X'_n(t, 0)$ follows directly from the tightness of its components which completes the proof of Lemma 7. ∎

We now turn to the convergence of the finite-dimensional distributions and we begin by proving a lemma relating the limiting finite-dimensional distributions of $X_n(t, \Theta)$ to those of $X'_n(t, 0)$.

LEMMA 8. *If Conditions* 1 *and* 2 *are satisfied then*

$$\operatorname*{limp}_{n\to\infty}\left(X'_n(t, \Theta) - X'_n(t, 0) + I(\theta_0)\Theta t\right) = 0 \tag{46}$$

and

$$\operatorname*{limp}_{n\to\infty}\left(X_n^{(1)}(t, \Theta) - \Theta^T \cdot X'_n(t, 0) + \tfrac{1}{2}\Theta^T I(\theta_0)\Theta t\right) = 0 \tag{47}$$

Proof. By a Taylor series expansion it is seen from Condition 1 that

$$|X_n^{(1)}(t, \Theta) - \Theta^T X'_n(t, 0) - \tfrac{1}{2}\Theta^T X''_n(t, 0)\Theta t| \leqslant \frac{k^2}{6} d_1^3 n^{-3/2} \sum_{j=1}^{m+1} F_3(X_j^{(n)}) \tag{48}$$

where $m = [nt]$ and $X''_n(t, 0)$ is a linear interpolation of

$$n^{-1} \sum_{j=1}^{m} \left[\frac{\partial^2}{\partial\theta^2} \ln f(X_j^{(n)}, \theta)\right]_{\theta=\theta_0}.$$

By Condition 1 (iii)

$$P\left(\frac{k^2}{6} d_1^3 n^{-3/2} \sum_{j=1}^{m+1} F_3(X_j^{(n)}) > \varepsilon\right) \leqslant \frac{K_3}{\varepsilon}\frac{k^2}{6} d_1^3 n^{-1/2} \frac{n}{m+1} \tag{49}$$

and thus by Condition 2 (i)

$$\operatorname*{limp}_{n\to\infty}\left(X_n^{(1)}(t, \Theta) - \Theta^T X'_n(t, 0) + \tfrac{1}{2} n^{-1}\Theta^T I(\theta_0)\Theta t\right) = 0$$

which is formula (47).

Again by a Taylor series expansion

$$|X'(t, \Theta) - X'(t, 0) - X''(t, 0)\Theta t| \leqslant \tfrac{1}{2} k d_1^2 n^{-3/2} \sum_{j=1}^{m+1} F_3(X_j^{(n)}) \tag{50}$$

and Condition 1 (iii) and Condition 2 give

$$\operatorname*{limp}_{n\to\infty}\left(X'_n(t, \Delta) - X'_n(t, 0) + I(\theta_0)\Theta t\right) = 0$$

which is formula (46). ∎

Note. If formula (47) is considered to be the approximation

$$X_n^{(1)}(t, \Theta) \approx \Theta^T \cdot X'_n(t, 0) - \tfrac{1}{2}\Theta^T I(\theta_0)\Theta \tag{51}$$

then formula (46) can be considered to be the approximation

$$X'(t, \Theta) \approx X'(t, 0) - \Theta^T I(\theta_0) \tag{52}$$

obtained by differentiating (51).

Lemma 8 shows that the finite-dimensional distributions of $X_n(t, \Theta)$ are determined by those of $X'_n(t, 0)$ for the same time points. The finite-dimensional distributions of $X_n(t, \Theta)$ degenerate when there are some points with coinciding time values.

The sequence defining the process $X'_n(t, 0)$ has independent increments and its limiting finite-dimensional distributions will be related to the finite-dimensional distributions of a k-dimensional Brownian motion $X'(t, 0)$ with drift $I(\theta_0)\ \Theta^*$ and covariance $I(\theta_0)$ per time unit, i.e., a k-dimensional process with independent increments, the increment $X'(t, 0) - X'(s, 0)$ in the interval $[s, t]$ being normally distributed with expectation $I(\theta_0)\ \Theta^*(t-s)$ and covariance $\mathrm{I}(\theta_0)(t-s)$.

LEMMA 9. *When Conditions* 1 *and* 2 *are satisfied then the finite-dimensional distributions of* $X'_n(t, 0)$ *converge to the corresponding finite-dimensional distributions of a Brownian motion* $X'(t, 0)$ *with drift* $I(\theta_0)\ \Theta^*$ *and covariance* $I(\theta_0)$ *per time unit.*

Proof. Consider first the k-dimensional distribution of $X'_n(t, 0)$. By Chebyshev inequality

$$\operatorname*{limp}_{n\to\infty}\left(X'_n(t, 0) - X'_n\left(\frac{[nt]}{n}, 0\right)\right) = 0 \tag{53}$$

and by the ordinary central limit theorem the limiting distribution of $X'([nt]/n, 0)$ is a normal distribution with mean $I(\theta_0)\Theta^* t$ and covariance $I(\theta_0)t$. Sufficient conditions to obtain this result are included in Conditions 1 and 2.

Next consider two time points s and t, where $s < t$. Then we can obtain the limiting distribution of $(X'_n(s, 0), X'_n(t, 0))$ from the limiting distribution of $(X'_n(s, 0), X'_n(t, 0) - X'_n(s, 0))$.

Again by Chebyshev inequality we can obtain the limiting distribution as the limiting distribution of $(X'_n([ns]/n, 0), X'_n([nt]/n, 0) - X'_n([ns]/n, 0))$. The two components are independent, and by the central limit theorem their distributions converge to those of $(X'(s, 0), X'(t, 0) - X'(s, 0))$. Three or more points can be handled in the same way. ∎

We now have the possibility to easily prove the following main theorem of this section.

THEOREM 2. *If Conditions* 1 *and* 2 *are satisfied then the likelihood process*

$$X_n(t, \Theta) = \begin{pmatrix} \dot{X}_n^{(1)}(t, \Theta) \\ X'_n(t, \Theta) \end{pmatrix},$$

$t \in T$, $\Theta \in D$, converges weakly to the random process

$$X(t, \Theta) = \begin{pmatrix} \Theta^T X'[t, 0] - \frac{1}{2}\Theta^T I(\theta_0)\Theta t \\ X'(t, 0) - I(\theta_0)\Theta t \end{pmatrix}$$

where $X'(t, 0)$ is a k-dimensional Brownian motion with drift $I(\theta_0)\Theta^*$ and covariance $I(\theta_0)$ per time unit.

Proof. Follows directly from Theorem 1 and Lemmas 7, 8 and 9. ∎

In order to see that Conditions 1 and 2 are not too restrictive we will now see that they are satisfied in exponential classes with natural parametrization. With our notation for main and nuisance parameters the density in such a class can be written

$$f(x, \gamma, \delta) = C(\gamma, \delta)\exp\big(\gamma T(x) + \sum_{j=1}^{k-1} \delta_j U_j(x)\big). \tag{54}$$

The function $C(\gamma, \delta)$ is analytic in each of the components $\gamma, \delta_1, \ldots, \delta_{k-1}$ where their real parts are in the interior of the natural parameter space and the partial derivatives of $\ln C(\gamma, \delta)$ give expectations and covariances for $T(X)$, $U_1(X), \ldots, U_{k-1}(X)$. See Lehmann (1959).

The first order partial derivatives of $\ln f(x, \gamma, \delta)$ are bounded by

$$F_1(x) = \sum_{j=0}^{k-1} K'_j + |T(x)| + \sum_{j=1}^{k-1} |U_j(x)| \tag{55}$$

where

$$K'_0 = \sup_{\theta \in \Omega_{\theta_0}} \left|\frac{\partial}{\partial \gamma} \ln C(\gamma, \delta)\right| \tag{56}$$

and

$$K'_j = \sup_{\theta \in \Omega_{\theta_0}} \left|\frac{\partial}{\partial \delta_j} \ln C(\gamma, \delta)\right|. \tag{57}$$

Here

$$E_\theta[|T(X)|] \leq (E_\theta[T^2(X)])^{1/2} = \left(\frac{\partial^2}{\partial \gamma^2} \ln C(\gamma, \delta) + \left(\frac{\partial}{\partial \gamma} \ln C(\gamma, \delta)\right)^2\right)^{1/2} \tag{58}$$

and similar estimates hold for $E_\theta[|U_j(X)|]$, $j = 1, 2, \ldots, k-1$.

Thus the boundedness condition

$$\int_{-\infty}^{\infty} F_1(x) f(x, \theta)\, d\mu(x) < K_1, \qquad \forall \theta \in \Omega_{\theta_0},$$

follows from the continuity of partial derivatives of $\ln C(\theta)$.

The second and third partial derivatives of $\ln f(x, \gamma, \delta)$ are not even random but equal to the same derivatives of $\ln C(\gamma, \delta)$. This means that parts

(ii) and (iii) of Condition 1 as well as Condition 2 follows from the continuity of partial derivatives of $\ln C(\gamma, \delta)$. Thus we have the following corollary.

COROLLARY 1. *Conditions* 1 *and* 2 *and the results of Theorem* 2 *hold for all naturally parametrized exponential classes.*

3. Unbounded parameter set

The random processes corresponding to the different test statistic sequences can be obtained from the log-likelihood process by different mappings. In the statistical applications they must always be mappings of the log-likelihood process on the whole parameter space. In order to be able to use the weak convergence results for bounded normalized parameter sets in statistical applications we thus must consider the problems connected with the transition to unbounded parameter sets. First the definition of the concept of limiting tail inferiority.

DEFINITION 1. A real random process $Y(t, \Theta)$ defined for $t \in T$ and $\Theta \in D \subseteq R^k$ is said to be *tail inferior* in Θ on T if for each $K \in R$ and $\varepsilon > 0$ there exists a $d_1 > 0$ such that

$$P\big(Y(t, \Theta) \leqslant K,\ \forall (t, \Theta)\colon\ t \in T,\ \Theta \in D,\ |\Theta| > d_1\big) \geqslant 1-\varepsilon. \tag{59}$$

DEFINITION 2. A sequence $\{Y_n(t, \Theta)\colon n = 1, 2, \ldots\}$ of real random processes $Y_n(t, \Theta)$, $t \in T$, $\Theta \in D_n \subseteq R^k$, is said to be *limiting tail inferior* in Θ on T if for each $K \in R$ and $\varepsilon > 0$ there exist $d_1 > 0$ and $n_1 > 0$ such that

$$P\big(Y_n(t, \Theta) \leqslant K,\ \forall (t, \Theta)\colon\ t \in T,\ \Theta \in D_n,\ |\Theta| > d_1\big) \geqslant 1-\varepsilon \tag{60}$$

for each $n \geqslant n_1$.

THEOREM 3. *Suppose that Conditions* 1 *and* 2 *are satisfied and that the first components* $X_n^{(1)}(t, \Theta)$ *of* $X_n(t, \Theta)$ *constitute a limiting tail inferior sequence. Then the Bartlett mapping* $X_n^{(B)}(t)$, *the Cox mappings* $X_n^{(C1)}(t)$ *and* $X_n^{(C2)}(t)$ *and the differential mappings* $X_n^{(D1)}(t)$ *and* $X_n^{(D2)}(t)$ *of* $\{X_n(t, \Theta)\colon t \in T,\ \theta_0 + n^{-1/2}\Theta \in \Omega\}$ *converge weakly on* T *to Brownian motion processes*

$$X^{(B)}(t) = \Gamma_1\,[J^T(\theta_0)\,X'(t, 0) - \tfrac{1}{2}\Gamma_1\,J^T(\theta_0)\,I(\theta_0)\,J(\theta_0)\,t], \tag{61}$$

$$X^{(C1)}(t) = \Gamma_1\,(J^T(\theta_0)\,I(\theta_0)\,J(\theta_0))^{-1}\,[J^T(\theta_0)\,X'(t, 0) - \tfrac{1}{2}\Gamma_1\,I_{\gamma\gamma}\cdot t], \tag{62}$$

$$X^{(C2)}(t) = \Gamma_1\,[J^T(\theta_0)\,X'(t, 0) - \tfrac{1}{2}\Gamma_1\,I_{\gamma\gamma}\,t], \tag{63}$$

$$X^{(D1)}(t) = \big(J^T(\theta_0)\,I(\theta_0)\,J(\theta_0)\big)^{-1}\,J^T(\theta_0)\,X'(t, 0) \tag{64}$$

and

$$X^{(D2)}(t) = J^T(\theta_0)\,X'(t, 0) \tag{65}$$

where $I_{\gamma\gamma}$, $I_{\gamma\delta}$, $I_{\delta\gamma}$ and $I_{\delta\delta}$ are submatrices of $I(\theta_0)$ of types 1×1, $1\times(k-1)$, $(k-1)\times 1$ and $(k-1)\times(k-1)$ defined by

$$I(\theta_0)=\begin{pmatrix} I_{\gamma\gamma} & I_{\gamma\delta} \\ I_{\delta\gamma} & I_{\delta\delta}\end{pmatrix}$$

and $J(\theta_0)$ is a $(k\times 1)$-matrix defined by

$$J(\theta_0)=\begin{pmatrix} 1 \\ -I_{\delta\delta}^{-1} I_{\delta\gamma}\end{pmatrix}$$

Proof. Let $X_{n,d}^{(B)}(t)$ denote the Bartlett mapping of

$$\{X_n(t,\Theta)\colon |\Theta|\leqslant d,\ \theta_0+n^{-1/2}\Theta\in\Omega\}$$

and let $X_{\infty,d}^{(B)}(t)$ denote the Bartlett mapping of

$$\{X(t,\Theta)\colon |\Theta|\leqslant d\}.$$

The infima $\inf_{t\in T} X_{n,d}^{(B)}(t)$ and $\inf_{t\in T} X_{\infty,d}^{(B)}(t)$ are continuous functions of $X_{n,d}^{(B)}(t)$ and $X_{\infty,d}^{(B)}(t)$ respectively in $C(T)$. Let further $X_n^{(B)}(t)$ denote the Bartlett mapping of $\{X_n(t,\Theta)\colon \theta_0+n^{-1/2}\Theta\in\Omega\}$ and $X^{(B)}(t)$ denote the Bartlett mapping of $\{X(t,\Theta)\colon \Theta\in R^k\}$. A trite calculation shows that $X^{(B)}(t)$ has the form given by formula (61).

Choose K to be the ε fractile in the distribution of $\inf_{t\in T} X^{(B)}(t)$. By definition

$$X^{(B)}(t)=\sup_{\Gamma=\Gamma_1} X(t,\Theta)-\sup_{\Gamma=0} X(t,\Theta).$$

But here $\sup_{\Gamma=0} X(t,\Theta)\geqslant 0$ since $X(t,0)=0$ and thus $X^{(B)}(t)\geqslant K$ implies $\sup_{\Gamma=\Gamma_1} X(t,\Theta)\geqslant K$.

The process $X^{(1)}(t,\Theta)$ is obviously tail inferior, i.e., for the K and ε above there exists $d_1>0$ such that

$$P(X^{(1)}(t,\Theta)\leqslant K,\ \forall t\in T \text{ and } |\Theta|>d_1)\geqslant 1-\varepsilon. \tag{66}$$

Now that $X^{(B)}(t)\geqslant K$, $\forall t\in T$, $\sup_{\Gamma=\Gamma_1} X(t,\Theta)\geqslant K$, $\forall t\in T$ and the two extremas $\sup_{\Gamma=\Gamma_1} X(t,\Theta)$ and $\sup_{\Gamma=0} X(t,\Theta)$ will be attained for $|\Theta|\leqslant d_1$ if $X^{(1)}(t,\Theta)\leqslant K$, $\forall t\in T$ and $|\Theta|>d_1$. Thus

$$P(X^{(B)}(t)=X_{\infty,d}^{(B)}(t),\ \forall t\in T \text{ and } d\geqslant d_1)\geqslant 1-2\varepsilon. \tag{67}$$

By limiting tail inferiority of $\{X_n^{(1)}(t,\Theta)\colon n=1,2,\ldots\}$ in Θ on T there exists for the above K and ε a $d_2>0$ and a n_2 such that

$$P(X_n^{(1)}(t,\Theta)\leqslant K,\ \forall(t,\Theta)\colon t\in T, |\Theta|>d_2)>d_2)\geqslant 1-\varepsilon \tag{68}$$

for each $n \geqslant n_2$. When $X_n^{(1)}(t, \Theta) \leqslant K < 0$, $\forall (t, \Theta)$: $t \in T, |\Theta| > d_2$ and

$$X_{n,d_2}^{(B)}(t) = \sup_{\Gamma = \Gamma_1, |\Theta| \leqslant d_2} X_n^{(1)}(t, \Theta) - \sup_{\Gamma = 0, |\Theta| \leqslant d_2} X_n^{(1)}(t, \Theta) > K, \qquad \forall t \in T$$

we must have $X_{n,d}^{(B)}(t) = X_n^{(B)}(t)$, $\forall t \in T$ and $d > d_2$ since $X_n^{(1)}(t, 0) = 0$ implies that $\sup_{\Gamma = 0, |\Theta| \leqslant d_2} X_n^{(1)}(t, \Theta) \geqslant 0$ cannot be attained by $X_n^{(1)}(t, \Theta)$ for $|\Theta| > d_2$ and that

$$\sup_{\Gamma = \Gamma_1, |\Theta| \leqslant d_2} X_n^{(1)}(t, \Theta) > K + \sup_{\Gamma = 0, |\Theta| \leqslant d_2} X_n^{(1)}(t, \theta) > K$$

cannot be attained by $X_n^{(1)}(t, \Theta)$ for $|\Theta| > d_2$. This means that with $d_0 = \max(d_1, d_2)$

$$\begin{aligned} &P\big(X_{n,d}^{(B)}(t) = X_n^{(B)}(t), \forall t \in T \text{ and } d > d_0\big) \\ &\geqslant P\big(X_n^{(1)}(t, \Theta) \leqslant K, \forall (t, \Theta)\text{: } t \in T, |\Theta| > d_0 \text{ and } X_{n,d_0}^{(B)}(t) > K, \forall t \in T\big) \\ &\geqslant P\big(X_n^{(1)}(t, \Theta) \leqslant K, \forall (t, \Theta); t \in T, |\Theta| > d_0\big) - \\ &\qquad - P\big(X_{n,d_0}^{(B)}(t) \leqslant K \text{ for some } t \in T\big). \end{aligned} \tag{69}$$

The process $X_n(t, \Theta)$ converges weakly to the process $X(t, \theta)$ on the bounded set $t \in T$, $|\Theta| \leqslant d_0$ when Conditions 1 and 2 are satisfied. Thus by Lemma 1 and the mapping Theorem 5.1 in Billingsley (1968) the process $X_{n,d_0}^{(B)}(t)$ converges weakly to $X_{\infty,d_0}^{(B)}$ on $t \in T$.

This also means that

$$\lim_{n \to \infty} P\big(X_{n,d_0}^{(B)}(t) \leqslant K \text{ for some } t \in T\big) = P\big(X_{\infty,d_0}^{(B)}(t) \leqslant K \text{ for some } t \in T\big), \tag{70}$$

i.e., that $\forall \varepsilon > 0 \; \exists n_3$ such that

$$P\big(X_{n,d_0}^{(B)} \leqslant K \text{ for some } t \in T\big) \leqslant P\big(X_{\infty,d_0}^{(B)}(t) \leqslant K \text{ for some } t \in T\big) + \varepsilon \tag{71}$$

for all $n \geqslant n_3$.

Since $d_1 \leqslant d_0$ we get by (67)

$$P\big(X^{(B)}(t) = X_{\infty,d_0}^{(B)}(t), \forall t \in T\big) \geqslant 1 - 2\varepsilon \tag{72}$$

and thus by (69), (68) and (71)

$$\begin{aligned} &P\big(X_{n,d}^{(B)}(t) = X_n^{(B)}(t), \forall t \in T, d > d_0\big) \\ &\qquad \geqslant 1 - \varepsilon - \big(P\big(X^{(B)}(t) \leqslant K \text{ for some } t \in T\big) + 2\varepsilon + \varepsilon\big) = 1 - 5\varepsilon. \end{aligned} \tag{73}$$

The weak convergence of $X_n^{(B)}(t)$ to $X^{(B)}(t)$ on $t \in T$ now follows from Theorem 2.1 in Billingsley (1968) since ε can be chosen arbitrarily small in (67) and (73).

The proofs of the other cases are similar. At the end of the proofs the mapping theorem is combined with Lemma 2 or Lemma 3 which require a concavity condition to be satisfied. This condition is a direct consequence of

Condition 2, which is supposed to be satisfied. The proofs for these other cases will not be reproduced. ∎

Note. The five processes $X^{(B)}(t)$, $X^{(C1)}(t)$, $X^{(C2)}(t)$, $X^{(D1)}(t)$ and $X^{(D2)}(t)$ are essentially equivalent since they are all transformations of the type

$$\varkappa(\theta_0)\, W(t) + \lambda(\theta_0)\, t$$

of the same Brownian motion process

$$W(t) = J^T(\theta_0)\, X'(t, 0).$$

This fact will be of great importance in the statistical applications to appear later.

For practical applications the weak convergence results on $T = [t_0, t_1]$ obtained for unbounded parameter are sufficient because t_0 can be chosen arbitrarily small and t_1 can be chosen arbitrarily big. And in practice it is reasonable to have some "idle time" in the beginning and a bound on the sample size anyway. The transition from $[t_0, t_1]$ to $[0, \infty]$ may however have a theoretical interest and some formulas for Wiener processes, useful for the original test problems, are simpler in infinite time than in finite time.

Close to 0 the estimates are bad and the log-likelihood process may be "wild". It seems to require severe restrictions on the likelihood function to get tightness properties all the way down to 0.

For big values of t the likelihood process is much more regular and the mappings $X_n^{(B)}(t)$, $X_n^{(C1)}(t)$, $X_n^{(C2)}(t)$, $X_n^{(D1)}(t)$ and $X_n^{(D2)}(t)$ behave approximately like processes with independent increments. Convergence results for functionals of these processes on infinite time intervals can be obtained if some uniform integrability conditions are satisfied. For functionals, whose expectations give asymptotic power functions and asymptotic expected sample size functions for SPR-type tests, these uniform integrability conditions are satisfied if the processes are uniformly exponentially bounded. We are not going to study those problems more extensively here. For (individual) exponential boundedness of similar processes see Wijsman (1977).

The condition of limiting tail inferiority of the first component of $X_n(t, \Theta)$ gives some restrictions of the log-likelihood function. In order to see that those restrictions are not too severe we will prove that this limiting tail inferiority always holds in exponential classes with natural parametrization.

COROLLARY 2. *For exponential classes with natural parametrization the first component $X_n^{(1)}(t, \Theta)$ of the* log-likelihood *process $X_n(t, \Theta)$ is limiting tail inferior in Θ on any interval $[t_0, t_1]$ where $0 < t_0 < t_1 < \infty$ and the results of Theorem* 3 *hold.*

Proof. For exponential classes with natural parametrization the first component of the log likelihood process in time points m/n equals

$$X_n^{(1)}\left(\frac{m}{n}, \Theta\right) = m\ln\frac{C(\theta_n)}{C(\theta_0)} + \sum_{j=1}^{m} (\theta_n - \theta_0)^T T(X_j) \tag{74}$$

where

$$\theta_n = \theta_0 + n^{-1/2}\Theta$$

and $T(X_j)$ is a k-vector valued function of X_j, whose components we denote by $T_\nu(X_j)$, $\nu = 1, 2, \ldots, k$. If the true parameter point is $\theta_n^* = \theta_0 + n^{-1/2}\Theta^*$ we introduce the random vector $Z_n(X_j)$ by

$$Z_n(X_j) = T(X_j) - E_{\theta_n^*}[T(X_j)] = T(X_j) - \left[\frac{\partial}{\partial\theta}\ln C(\theta)\right]_{\theta=\theta_n^*} \tag{75}$$

and write the process $X_n^{(1)}\left(\frac{m}{n}, \Theta\right)$ in the form

$$X_n^{(1)}\left(\frac{m}{n}, \Theta\right) = m\ln\frac{C(\theta_n)}{C(\theta_0)} - \frac{m}{\sqrt{n}}\Theta^T\left[\frac{\partial}{\partial\theta}\ln C(\theta)\right]_{\theta=\theta_n^*} + $$

$$+ n^{-1/2}\Theta^T \sum_{j=1}^{m} Z_n(X_j). \tag{76}$$

Here the components $Z_n^{(\nu)}(X_j)$, $\nu = 1, 2, \ldots, k$ of $Z_n(X_j)$ have expectations 0 and by the Kolmogorov inequality (see, e.g., Billingsley (1968), p. 248)

$$P_{\theta_n^*}\Big(\sup_{t_0 n \leqslant m \leqslant t_1 n} \sum_{j=1}^{m} Z_n^{(\nu)}(X_j) \geqslant \alpha\Big)$$

$$\leqslant \frac{nt_1}{\alpha^2}\operatorname{Var}_{\theta_n^*}(Z_n^{(\nu)}(X_1)) = \frac{nt_1}{\alpha^2}\left[-\frac{\partial^2}{\partial\theta_\nu^2}\ln C(\theta)\right]_{\theta=\theta_n^*} \tag{77}$$

and

$$P_{\theta_n^*}\Big(\sup_{t_0 n \leqslant m \leqslant t_1 n} \Big|\sum_{j=1}^{m} Z_n^{(\nu)}(X_j)\Big| \geqslant \alpha\Big)$$

$$\leqslant P_{\theta_n^*}\Big(\sup_{t_0 n \leqslant m \leqslant t_1 n} \sum_{j=1}^{m} Z_n^{(\nu)}(X_j) \geqslant \alpha\Big) + P_{\theta_n^*}\Big(\sup_{t_0 n \leqslant m \leqslant t_1 n} \sum_{j=1}^{m} -Z_n^{(\nu)}(X_j) \geqslant \alpha\Big)$$

$$\leqslant \frac{2nt_1}{\alpha^2}\left[-\frac{\partial^2}{\partial\theta_\nu^2}\ln C(\theta)\right]_{\theta=\theta_n^*}. \tag{78}$$

Thus for the random part of $X_n^{(1)}\left(\frac{m}{n}, \Theta\right)$ we get the estimate

$$P_{\theta_n^*}\left(\sup_{t_0 n \leqslant m \leqslant t_1 n}\left|\frac{1}{\sqrt{n}}\Theta^T \sum_{j=1}^{m} Z_n(X_j)\right| \geqslant K_1\right) \leqslant \frac{2t_1}{K_1^2}|\Theta|\, k2v_0 \tag{79}$$

for n big enough where

$$v_0 = \sup_{1 \leqslant v \leqslant k}\left[-\frac{\partial^2}{\partial\theta_v^2}\ln C(\theta)\right]_{\theta=\theta_0} \tag{80}$$

since the second order partial derivatives of $\ln C(\theta)$ are continuous.

The nonrandom part of $X_n^{(1)}\left(\frac{m}{n}, \Theta\right)$ equals

$$m\left(\ln\frac{C(\theta_n)}{C(\theta_0)} - \frac{m}{\sqrt{n}}\Theta^T\left[\frac{\partial}{\partial\theta}\ln C(\theta)\right]_{\theta=\theta_n^*}\right)$$
$$= m\left(\ln\frac{\theta_n}{\theta_0} - (\theta_n-\theta_0)^T\left[\frac{\partial}{\partial\theta}\ln C(\theta)\right]_{\theta=\theta_n^*}\right),$$

which is a concave function of θ_n (or Θ) taking its maximal value for $\theta_n = \theta_n^*$ (or $\Theta = \Theta^*$). Because of the continuity of the partial derivatives of $\ln C(\theta)$ we get (uniformly on sets $|\Theta| \leqslant d$)

$$\lim_{n\to\infty} m\left(\ln\frac{\theta_n}{\theta_0} - (\theta_n-\theta_0)^T\left[\frac{\partial}{\partial\theta}\ln C(\theta)\right]_{\theta=\theta_n^*}\right) = t(\Theta^T I\Theta^* - \tfrac{1}{2}\Theta^T I\Theta). \tag{81}$$

Let v_1 and v_2 be the minimal and maximal eigenvalues of I, which both are positive. Then for $|\Theta^*| \leqslant \frac{1}{4}(v_1/v_2)^{1/2}d$ and $|\Theta| = d$ we have

$$t(\Theta^T I\Theta^* - \tfrac{1}{2}\Theta^T I\Theta) \leqslant -\tfrac{1}{4}td^2 v_1. \tag{82}$$

Since

$$m\left(\ln\frac{\theta_n}{\theta_0} - (\theta_n-\theta_0)^T\left[\frac{\partial}{\partial\theta}\ln C(\theta)\right]_{\theta=\theta_n^*}\right)$$

is a concave function of θ_n (a concave function of $\Theta = n^{1/2}(\theta_n-\theta_0)$) we thus get

$$m\left(\ln\frac{\theta_n}{\theta_0} - (\theta_n-\theta_0)^T\left[\frac{\partial}{\partial\theta}\ln C(\theta)\right]_{\theta=\theta_n^*}\right)$$
$$\leqslant -\frac{1}{8}td^2 v_1\frac{|\Theta|}{d} \leqslant -\frac{1}{8}t_1 dv_1|\Theta| \tag{83}$$

for $|\Theta| \geqslant d$ and n big enough.

Thus by (76) and (79) for n big enough

$$P_{\theta_n^*}\left(X_n^{(1)}\left(\frac{m}{n}, \Theta\right) \leqslant K, \forall \frac{m}{n} \in [t_0, t_1] \text{ and } |\Theta| > d_1\right)$$

$$\geqslant P_{\theta_n^*}\left(\sup_{t_0 n \leqslant m \leqslant t_1 n}\left|\frac{1}{\sqrt{n}}\Theta^T \sum_{j=1}^{m} Z_n(X_j)\right| \leqslant K + \frac{1}{8} t_0 dv_1 |\Theta|\right)$$

$$\geqslant 1 - \frac{2t_1 |\Theta| k \cdot 2v_0}{(K + \frac{1}{8} t_1 dv_1 |\Theta|)^2}, \tag{84}$$

which can be made arbitrarily close to 1 by making $|\Theta|$ big enough ■

4. Asymptotic test properties and steering

For a nonsequential test the *asymptotic relative efficiency* (A.R.E.) or *Pitman efficiency* is defined as the limiting ratio of the sample sizes of two tests when they approach the same asymptotic power. When the asymptotic distributions are normal it equals the limiting ratio of the squares of expectation differences divided by standard deviations for the two tests.

The sequential analogue for problems with one-dimensional parameter is given by Lai (1978). When the random processes defined from the sequences of test statistics converge weakly to Brownian motions he defines the A.R.E. as the limiting squared ratio of the scaled change in the asymptotic drift.

For problems with multidimensional parameter the A.R.E. with respect to a main parameter can be defined in the same way for each nuisance parameter point. In general the A.R.E. depends on the nuisance parameter point.

The random processes defined by the test statistics sequences used in the tests B, C1, C2, D1 and D2 all converge weakly to Brownian motion processes of the type

$$\varkappa(\theta_0) W(t) + \lambda(\theta_0) t$$

where

$$W(t) = J^T(\theta_0) X'(t, 0)$$

is a Brownian motion with variance

$$\sigma^2(\theta_0) = J^T(\theta_0) I(\theta_0) J(\theta_0) = I_{\gamma\gamma} - I_{\gamma\delta} I_{\delta\delta}^{-1} I_{\delta\gamma} \tag{85}$$

per time unit and drift

$$J^T(\theta_0) I(\theta_0) \Theta^* = (I_{\gamma\gamma} - I_{\gamma\delta} I^{-1} I_{\delta\gamma}) \Gamma^* = \sigma^2(\theta_0) \Gamma^* \tag{86}$$

per time unit. Thus the difference of the drift for $\Gamma = \Gamma^*$ and $\Gamma = 0$ divided by the standard deviation per time unit for all those test statistic processes equals

$$\frac{\varkappa(\theta_0)\sigma^2(\theta_0)(\Gamma^* - 0)}{\varkappa(\theta_0)\sigma(\theta_0)} = \sigma(\theta_0)\Gamma^*$$

and the A.R.E. for a pair of such test statistics sequences always equals 1 although they do not have the same scale function $\varkappa(\theta_0)$ and drift function $\lambda(\theta_0)$.

Observe however that the A.R.E. properties are obtained for time intervals of the type $[t_0, t_1]$ only, i.e., for sample sizes m in an interval $t_0 n \leqslant m \leqslant t_1 n$ only. The result thus means that if a test with idle time $t_0 n$ and truncation time $t_1 n$ based on one of the test statistic sequences for a fixed nuisance parameter point has some asymptotic properties obtainable from the weak convergence to the Brownian motion process there are also tests with idle time $t_0 n$ and truncation time $t_1 n$ based on the other test statistics sequences with the same asymptotic properties in the same nuisance parameter point. In general the two tests have different asymptotic properties in other nuisance parameter points. But we will soon see that by introducing power steering, where the rejection and acceptance boundaries depend on a consistent estimate of the nuisance parameter, we can get tests based on different test statistics sequences with the same asymptotic properties for all nuisance parameter points.

The original Bartlett and Cox tests are tests of Wald type, i.e., the hypothesis is rejected if the sequence of test statistics reaches a fixed upper boundary and it is accepted if the sequence of test statistics reaches a fixed lower boundary. Since the Wald type of tests have good properties in problems with one-dimensional parameter we will concentrate on this type of tests.

When calculating asymptotic properties for Wald type of tests with idle time nt_0 and truncation time nt_1 we have to use results concerning the hitting of two fixed boundaries in the time interval $[t_0, t_1]$ for a Brownian motion process. In practice the idle time t_0 is usually small and the truncation time t_1 is usually big. In that case the properties for this interval are approximately equal to the corresponding properties for the interval $(0, \infty)$.

Let $\{W(t)\colon 0 \leqslant t < \infty\}$ be a Brownian motion with drift μ and variance σ^2 per time unit. For some $b < 0$ and $a > 0$ let τ be the stopping time

$$\tau = \inf\{t \in [0, \infty)\colon\ W(t) \geqslant a \ \text{or}\ W(t) \leqslant b\}.$$

Then

$$E[\tau] = \frac{a(e^{-2b\mu/\sigma^2} - 1) + b(1 - e^{-2a\mu/\sigma^2})}{\mu(e^{-2b\mu/\sigma^2} - e^{-2a\mu/\sigma^2})} \qquad \text{for} \quad \mu \neq 0, \tag{87}$$

$$E[\tau] = -\frac{ab}{\sigma^2} \qquad \text{for} \quad \mu = 0 \tag{88}$$

and

$$P\big(W(\tau)=a\big)=\frac{e^{-2b\mu/\sigma^2}-1}{e^{-2b\mu/\sigma^2}-e^{-2a\mu/\sigma^2}} \qquad \text{for} \quad \mu \neq 0, \tag{89}$$

$$P\big(W(\tau)=a\big)=-\frac{b}{a-b} \qquad \text{for} \quad \mu=0. \tag{90}$$

See, e.g., Dvoretsky, Kiefer and Wolfowitz (1953). From these formulas we get the approximate expected sample size and power for the five studied likelihood ratio tests when the idle time is small and truncation time is big.

For the simple differentiation test D2 the drift per time unit is $\sigma^2(\theta_0)\cdot\Gamma^*$, and the variance per time unit is $\sigma^2(\theta_0)$. Then the approximate power $P\big(W(\tau)=a\big)$ will be

$$P\big(W(\tau)=a\big)=\frac{e^{-2b\Gamma^*}-1}{e^{-2b\Gamma^*}-e^{-2a\Gamma^*}} \qquad \text{for} \quad \Gamma^*\neq 0 \tag{91}$$

and

$$P\big(W(\tau)=a\big)=-\frac{b}{a-b} \qquad \text{for} \quad \Gamma^*=0, \tag{92}$$

i.e., it will depend only on Γ^* and not on the nuisance parameter δ_0. The same is the case for the Bartlett test while for the others it depends in general on both Γ^* and δ_0.

The general formulas for the approximate expected stopping time τ and power for boundaries $b<0$ and $a>0$ applied to the process $\varkappa(\delta_0)W(t)+\lambda(\delta_0)t$ where $W(t)$ is a process with variance $\sigma^2(\delta_0)$ per time unit and drift $\Gamma^*\sigma^2(\delta_0)$ per time unit are obtained from formulas (87), (88), (89) and (90) by substituting σ^2 with $\varkappa^2(\delta_0)\,\sigma^2(\delta_0)$ and μ with $\varkappa(\delta_0)\,\sigma^2(\delta_0)\Gamma^*+\lambda(\delta_0)$. From these formulas we can thus get approximate expected sample size functions and approximate power function for the different tests by using the appropriate functions $\varkappa(\delta_0)$ and $\lambda(\delta_0)$ from formulas (61)–(65) in Theorem 3.

Already Cox (1963) pointed out that in order to get the Wald approximations for the power of his test (C1) the test limits a and b should not be fixed but should depend on an estimate $\hat{\delta}_0$ of the nuisance parameter δ_0. The test should then also be equivalent to the Bartlett test. Formulas (61) and (62) give the Brownian motion processes

$$W_B(t)=\Gamma_1\big(W(t)-\tfrac{1}{2}\Gamma_1\sigma^2(\delta_0)t\big)$$

and

$$W_{C1}(t)=\Gamma_1\sigma^{-2}(\delta_0)\big(W(t)-\tfrac{1}{2}\Gamma_1\sigma^2(\delta_0)t\big)$$

associated with the tests B and C1. From these it is immediately seen that a B test with limits $b<0$ and $a>0$ is asymptotically equivalent to a C1 test with limits $b \cdot \sigma^2(\delta_0) < 0$ and $a \cdot \sigma^2(\delta_0) > 0$. Then it is also asymptotically equivalent to a C1 test with estimated limits $b \cdot \sigma^2(\hat{\delta}_0) < 0$ and $a \cdot \sigma^2(\hat{\delta}_0) > 0$ on the interval $[t_0, t_1]$ if $\sigma^2(\delta_0)$ is a continuous function of δ_0 and $\hat{\delta}_0$ is a strongly consistent estimate of δ_0.

We will now study more generally ideas of steering properties of likelihood ratio tests of the Wald type by letting the boundaries of the tests as well as an imposed drift depend on an estimate of the nuisance parameter δ_0. Suppose that $\hat{\delta}_0$ is a strongly consistent estimate of δ_0 and let

$$W_1(t) = \varkappa_1(\delta_0)W(t) + \lambda_1(\delta_0)t \tag{93}$$

and

$$W_1(t) = \varkappa_2(\delta_0)W(t) + \lambda_2(\delta_0)t \tag{94}$$

be the Brownian motions associated with any two of the likelihood ratio test statistic sequences. In both cases there are sequences

$$W_{1,n}(t), \qquad n = 1, 2, \ldots,$$

and

$$W_{2,n}(t), \qquad n = 1, 2, \ldots,$$

of sample random processes converging weakly to $W_1(t)$ and $W_2(t)$ respectively in $[t_0, t_1]$.

If $\varkappa_1(\delta_0)$, $\lambda_1(\delta_0)$, $\varkappa_2(\delta_0)$ and $\lambda_2(\delta_0)$ are continuous functions of δ_0 and $\varkappa_1(\delta_0) \neq 0$ and $\varkappa_2(\delta_0) \neq 0$ for all δ_0 then also

$$\left(W_{1,n}(t) - \lambda_1(\hat{\delta}_0)t\right)\frac{\varkappa_2(\hat{\delta}_0)}{\varkappa_1(\hat{\delta}_0)} + \lambda_2(\hat{\delta}_0)t$$

converges weakly to $W_2(t)$ on $[t_0, t_1]$ and

$$\left(W_{2,n}(t) - \lambda_2(\hat{\delta}_0)t\right)\frac{\varkappa_1(\hat{\delta}_0)}{\varkappa_2(\hat{\delta}_0)} + \lambda_1(\hat{\delta}_0)t$$

converges weakly to $W_1(t)$ on $[t_0, t_1]$.

This means that by modifying one of the likelihood ratio test statistics sequences with the help of $\hat{\delta}_0$ we can get a test with the same asymptotic properties for all δ_0 as a test based on another likelihood ratio test statistics sequence. This motivates the given definition of asymptotic relative efficiency for problems with nuisance parameter.

By this method we can not only transform any likelihood ratio test statistics sequence to converge to the Brownian motion process associated with another likelihood ratio test statistics sequence, but also transform it to converge to any Brownian motion process $\varkappa(\delta_0)W(t) + \lambda(\delta_0)t$ where $W(t)$ is a Brownian motion process with drift Γ^* per time unit and variance $\sigma^2(\delta_0)$ per time unit and $\varkappa(\delta_0)$ and $\lambda(\delta_0)$ are continuous functions of δ_0. Thus the asymptotic properties of the test can be steered to a desired dependence of the nuisance parameter δ_0.

The character of the limiting Brownian motion processes for the B, C1 and C2 tests differs from the character of the limiting Brownian motion processes of the D1 and D2 tests in the way that the drift is proportional to $\Gamma^* - \Gamma'/2$ in the first case and proportional to Γ^* in the second case. This depends on the intention to maximize power at $\Gamma^* = \Gamma_1$ in the first three tests and the intention to maximize power locally at $\Gamma^* = 0$ in the last two tests. We will discuss this difference more when studying optimality in next section. In order to demonstrate the principles of steering of power in a simple form we will here only study the steering of the simple D2 tests by a multiplicative factor $\varkappa(\hat{\delta}_0)$. This means that we get a process which converges weakly to a Brownian motion process with drift proportional to Γ^* where the proportionality constant depends on the nuisance parameter δ_0.

In a Wald type of test based on this steered sequence experimentation will stop as soon as $\varkappa(\hat{\delta}_0) \cdot X_n^{(D2)}(m/n)$ reaches one of the boundaries $b < 0$ and $a > 0$, the hypothesis is accepted if $b < 0$ is reached and rejected if $a > 0$ is reached. If the idle time t_0 is small and the truncation time t_1 is big then the approximate power is

$$\beta(\Gamma^*, \delta_0) = \frac{e^{-2b\Gamma^*/\varkappa(\delta_0)} - 1}{e^{-2b\Gamma^*/\varkappa(\delta_0)} - e^{-2a\Gamma^*/\varkappa(\delta_0)}} \quad \text{for} \quad \Gamma^* \neq 0 \tag{95}$$

and

$$\beta(0, \delta_0) = -\frac{b}{a-b} \tag{96}$$

and the approximate expected stopping time is

$$e(\Gamma^*, \delta_0) = \frac{a(e^{-2b\Gamma^*/\varkappa(\delta_0)} - 1) + b(1 - e^{-2a\Gamma^*/\varkappa(\delta_0)})}{\sigma^2(\delta_0)\Gamma^*(e^{-2b\Gamma^*/\varkappa(\delta_0)} - e^{-2a\Gamma^*/\varkappa(\delta_0)})} \quad \text{for} \quad \Gamma^* \neq 0 \tag{97}$$

and

$$e(0, \delta_0) = \frac{ab}{\sigma^2(-\delta_0) \cdot \varkappa^2(\delta_0)} \tag{98}$$

The derivative (with respect to Γ^*) of $\beta(\Gamma^*, \delta_0)$ at $\Gamma^* = 0$ equals

$$\beta'(0, \delta_0) = \frac{-ab}{(a-b)\varkappa(\delta_0)} \tag{99}$$

which can be used as an estimate of the steepness of the power function at δ_0.

If we choose $\varkappa(\delta_0)$ smaller the steepness of the power function will increase but at the same time the expected stopping time will increase even more which is most easily seen in $e(0, \delta_0)$. We illustrate the steering of power by a very simple example.

EXAMPLE 1. Let $X_1, X_2, \ldots$ be independent normally distributed with expectation γ and variance δ. The D2 test of $\gamma \leqslant 0$ against $\gamma > 0$ should be based on the sequence

$$X_n^{(D2)}\left(\frac{m}{n}\right) = n^{-1/2} \sum_{i=1}^{m} \frac{X_i - 0}{\hat{\delta}^{(m)}} \tag{100}$$

where

$$\hat{\delta}^{(m)} = \frac{1}{m} \sum_{i=1}^{m} X_i^2. \tag{101}$$

See formulas (3) and (19). For this test sequence

$$\sigma^2(\delta_0) = J^T(0, \delta_0) I(0, \delta_0) J(0, \delta_0) = \frac{1}{\delta_0}$$

and the test with continuation region

$$b < n^{-1/2} m \sum_{i=1}^{m} X_i / \left(\sum_{i=1}^{m} X_i^2\right) < a \tag{102}$$

(for $nt \leqslant m \leqslant nt_1$) has asymptotic power depending only on $\Gamma^* = \gamma^* \sqrt{n}$ and not on δ_0. The approximate power derivative at $\Gamma^* = 0$ with respect to Γ^* is

$$\beta'(0, \delta_0) = \frac{-ab}{a-b},$$

i.e., the approximate power derivative at $\gamma^* = 0$ with respect to γ^* is

$$\sqrt{n}\, \beta'(0, \delta_0) = \frac{-ab}{a-b} \sqrt{n}$$

independent of δ_0. The approximate expected sample size at the boundary $\gamma^* = 0$ is

$$n \cdot e(0, \delta_0) = n \frac{-ab}{a-b} \delta_0.$$

The power of the test can be steered so that the slope $\beta'(0, \delta_0)$ is proportional to $\delta_0^{-1/2}$ by choosing $\varkappa(\delta_0) = \delta_0^{1/2}$. This means that the test has the continuation region

$$b < n^{-1/2} m \sum_{i=1}^{m} X_i / \left(\sum_{i=1}^{m} X_i^2\right)^{1/2} < a \tag{103}$$

(for $nt_0 \leqslant t \leqslant nt_1$), the approximate power derivative at $\gamma^* = 0$ with respect to γ^* is

$$n^{1/2}\beta'(0, \delta_0) = \frac{-ab}{a-b} n^{1/2} \delta_0^{-1/2}$$

and the approximate expected sample size at the boundary $\gamma^* = 0$ is

$$ne(0, \delta_0) = n \frac{-ab}{a-b} \delta_0 \cdot \delta_0^{-1}.$$

We get here a simple type of sequential t test, where the approximate power is a function of $\dfrac{\Gamma^*}{\varkappa(\delta_0)} = n^{1/2} \dfrac{\gamma}{\delta_0^{1/2}}$. See formula (95) and (96).

One can easily construct a number of examples of sequential likelihood ratio tests with or without steering of power.

5. Locally most powerful unbiased tests and an optimum property

In the previous sections we have studied asymptotic properties of the five sequential likelihood ratio tests and compared them to each other. But we have made no attempt to compare them to other tests or to show some kind of optimality. We will now introduce suitable comparison tests and prove an optimality property of the likelihood ratio tests.

If we knew the exact value δ_0 of the nuisance parameter δ, then we could construct a locally most powerful test of $\gamma \leqslant \gamma_0$ (or $\gamma = \gamma_0$) against $\gamma > \gamma_0$ according to Berk (1975a) which would then give a maximal γ-derivative of the power function at $\gamma = \gamma_0$ among the tests with the same level and expected sample size when $\gamma = \gamma_0$ (and $\delta = \delta_0$). We can never expect a likelihood ratio test (or any other test) with a level α and expected sample size n_0 at $\gamma = \gamma_0$, $\delta = \delta_0$ to have a higher value of the γ-derivative of the power function than such a locally most powerful (LMP) test of $\gamma \leqslant \gamma_0$ against $\gamma > \gamma_0$ at $\delta = \delta_0$ with level α and expected sample size n_0. And we would really be satisfied if we got the same γ-derivative. Usually we do not.

But the comparison between likelihood ratio tests and LMP tests for fixed nuisance parameter points is not fair. The likelihood ratio test is intended to be a reasonably good test (at some level α) for all δ while the LMP is an optimal test for just one fixed δ. It might happen that the level α LMP test in a point δ has a power function whose derivatives with respect to components of δ are not equal to 0 at $\delta = \delta_0$. In this case the level of the test can be much higher than α even in a small neighbourhood of δ_0. It is more fair to compare the likelihood ratio tests to the locally most powerful unbiased (LMPU) tests we are now going to introduce.

DEFINITION 3. Suppose that $X_1, X_2, \ldots$ are i.i.d. random variables whose distribution depends on a one-dimensional main parameter γ and a $(k-1)$ dimensional nuisance parameter δ. A test is said to be a *locally most powerful unbiased* (LMPU) test of $\gamma = \gamma_0$ against $\gamma > \gamma_0$ at $\delta = \delta_0$ of size α with expected sample size $\leqslant n_0$ if

(i) its size is $\leqslant \alpha$,

(ii) its expected sample size at (γ_0, δ_0) is $\leqslant n_0$,

(iii) the derivatives of its power function with respect to components of δ in the point (γ_0, δ_0) are equal to 0

(iv) the γ derivative of its power function in the point (γ_0, δ_0) is greater than or equal to the same derivative for every other test satisfying (i), (ii) and (iii).

In the one parameter case when the random variables $X_1, X_2, \ldots$ have a density $f(x, \theta)$ depending on a one-dimensional θ the test of $\theta = \theta_0$ against $\theta > \theta_0$ based on

$$S_m = \sum_{i=1}^{m} \left[\frac{\partial}{\partial \theta} \ln f(x_i, \theta) \right]_{\theta = \theta_0},$$

with continuation region

$$b < S_m < a$$

is LMP at its size and expected sample size. See Berk (1975a). In the case with multidimensional parameter θ denote $\theta_0 = (\gamma_0, \delta_0)$ and introduce for each unit vector $\lambda \in R^k$ the corresponding one parameter family of densities

$$f(x, \theta_0 + \eta\lambda)$$

where the new parameter is the scalar η. Under our Conditions 1 and 2 the assumptions 1–4 in Berk (1975a) are satisfied and the test with continuation region

$$b < \sum_{i=1}^{m} \left[\frac{\partial}{\partial \eta} \ln f(x_i, \theta_0 + \eta\lambda) \right]_{\eta = 0} < a \tag{104}$$

is LMP at its level and expected sample size for testing $\eta = 0$ against $\eta > 0$. Randomization on the boundaries is allowed.

Here

$$\left[\frac{\partial}{\partial \eta} \ln f(x_i, \theta_0 + \eta\lambda) \right]_{\lambda = 0} = \sum_{j=1}^{k} \lambda_j \cdot \left[\frac{\partial}{\partial \theta_j} \ln f(x_i, \theta) \right]_{\theta = \theta_0} \tag{105}$$

where $\lambda_1, \ldots, \lambda_k$ are the components of λ and $\theta_1, \ldots, \theta_k$ are components of θ. The main parameter γ is supposed to be the first component θ_1 of θ. Using the notation

$$S_m^{(j)} = \sum_{i=1}^{m} \left[\frac{\partial}{\partial \theta_j} \ln f(x_i, \theta) \right]_{\theta = \theta_0} \tag{106}$$

we can write the continuation region (104) for the above LMP test

$$b < \sum_{j=1}^{k} \lambda_j S_m^{(j)} < a.$$

Now let $\mathscr{L}^+(\alpha, n_0)$ be the set of LMP tests of size α with expected sample size $\leqslant n_0$ in directions λ with positive first component λ_1. Then we have the following theorem:

THEOREM 4. *Suppose that Conditions* 1 *and* 2 *are satisfied. If a test in* $\mathscr{L}^+(\alpha, n_0)$ *has a power function with all derivatives with respect to components of* δ *equal to* 0 *in the point* $\theta_0 = (\gamma_0, \delta_0)$ *then it is a locally most powerful unbiased test of* $\gamma = \gamma_0$ *against* $\gamma > \gamma_0$ *on level* α *with expected sample size* $\leqslant n_0$.

Proof. Let λ be a unit vector with $\lambda_1 > 0$ for which the test satisfies the condition that all δ-derivatives of the power functions in $\theta_0 = (\gamma_0, \delta_0)$ equal 0. Denote its stopping time by N. The test has a continuation region

$$b < \sum_{j=1}^{k} \lambda_j S_m^{(j)} < a$$

and possibly a randomization if the limits are hitted exactly. The terminal decision rules (for stopping at n) are given by test functions φ_n depending only on $X_1, X_2, \ldots, X_n$ giving the rejection probability. Then when Conditions 1 and 2 are satisfied the power of the test equals

$$\beta(\theta) = E_\theta[\varphi_N] \tag{107}$$

and the derivative of the power with respect to the jth component of θ in the point θ_0 equals

$$\left[\frac{\partial}{\partial \theta^{(j)}} \right]_{\theta = \theta_0} = E_{\theta_0}[\varphi_N \cdot S_N^{(j)}] \tag{108}$$

This follows directly from Berk (1975a). Thus for our LMP test in the λ direction satisfying the side condition that all δ-derivatives of the power function equal 0 we have

$$E_{\theta_0}[\varphi_N \cdot S_N^{(j)}] = 0 \qquad \text{for} \quad j = 2, 3, \ldots, k \tag{109}$$

and the power function derivative

$$E_\theta[\varphi_N \cdot \sum_{j=1}^{k} \lambda_j S_N^{(j)}]$$

is greater than or equal to the same derivative for any other test with level α and expected sample size $\leqslant n_0$ satisfying the same side conditions. Let M be the stopping time and ψ_n be the test functions for any such test. Then using the just mentioned properties

$$\begin{aligned} E_{\theta_0}[\varphi_N S_N^{(1)}] &= \frac{1}{\lambda_1} E_{\theta_0}[\varphi_N \lambda_1 S_N^{(1)}] \\ &= \frac{1}{\lambda_1} E_{\theta_0}[\varphi_N \sum_{j=1}^{k} \lambda_j S_N^{(j)}] \geqslant \frac{1}{\lambda_1} E_{\theta_0}[\psi_N \sum_{j=1}^{k} \lambda_j S_N^{(j)}] \\ &= \frac{1}{\lambda_1} E_{\theta_0}[\psi_N \lambda_1 S_N^{(1)}] = E_{\theta_0}[\psi_N S_N^{(1)}], \end{aligned} \tag{110}$$

i.e., the LMP test in the λ direction is a LMPU test. ∎

Note. If there exist two λ's for which the LMP tests on level α with expected sample size $\leqslant n_0$ in θ_0 satisfying the side condition that the δ-derivatives of the power function equal 0 in θ_0, then they must have the same power function γ-derivative in θ_0. This is seen by applying (110) to both of them comparing with the other. Observe further that in general the LMPU test is not the LMP test corresponding to $\lambda = (1, 0, 0, \ldots, 0)^T$ but to some other λ.

We will now find an estimate of the γ derivative of the power function in the point $\theta = \theta_0$ for a LMPU test, i.e., an upper boundary of the γ derivative of the power function in the point $\theta = \theta_0$ for any locally unbiased test with size α and expected sample size $\leqslant n_0$. First we consider the case with one-dimensional parameter θ.

LEMMA 10. *Let $X_1, X_2, \ldots$ be i.i.d. random variables whose density $f(x, \theta)$ depend on a one-dimensional parameter θ and suppose that conditions 1 and 2 are satisfied. If $\beta(\theta)$ is the power function of any size α test of $\theta = \theta_0$ against $\theta > \theta_0$ with expected sample size $\leqslant n_0$ in θ_0, then*

$$\left[\frac{\partial \beta}{\partial \theta}\right]_{\theta=\theta_0} \leqslant \alpha^{1/2}(1-\alpha)^{1/2}\sigma n_0^{1/2} \tag{111}$$

where

$$\sigma^2 = E_{\theta_0}\left[\left(\left[\frac{\partial \ln f(X_1, \theta)}{\partial \theta}\right]_{\theta=\theta_0}\right)^2\right]. \tag{112}$$

Proof. Berk (1975a) has shown that the power function derivative $\left[\frac{\partial \beta}{\partial \theta}\right]_{\theta=\theta_0}$ is maximized by a LMP test with continuation region

$$b < S_m < a$$

where

$$S_m = \sum_{i=1}^{m} \left[\frac{\partial}{\partial \theta} \ln f(X_i, \theta) \right]_{\theta = \theta_0}.$$

Denoting the stopping time for the test by N, the random variables $\left[\frac{\partial}{\partial \theta} \ln f(X_i, \theta) \right]_{\theta = \theta_0}$ by Z_i and the test function when stopping at m by φ_m he also shows that the power function derivative at $\theta = \theta_0$ equals

$$\beta' = E_{\theta_0}[\varphi_N \cdot S_N] = \alpha \cdot E_{\theta_0}[S_N \mid S_N \geqslant a]. \tag{113}$$

Further

$$\alpha = E_{\theta_0}[\varphi_N] \tag{114}$$

and by well-known Wald equations (see, e.g., Govindarajulu (1981))

$$\begin{aligned} E_{\theta_0}[N] \cdot E_{\theta_0}[Z_1] &= E_{\theta_0}[S_N] \\ &= \alpha \cdot E_{\theta_0}[S_N \mid S_N \geqslant a] + (1-\alpha) E_{\theta_0}[S_N \mid S_N \leqslant b] \end{aligned} \tag{115}$$

and

$$\begin{aligned} E_{\theta_0}[N] \cdot E_{\theta_0}[Z_1^2] &= E_{\theta_0}[S_N^2] \\ &= \alpha \cdot E_{\theta_0}[S_N^2 \mid S_N \geqslant a] + (1-\alpha) E[S_N^2 \mid S_N \leqslant b]. \end{aligned} \tag{116}$$

Since $E_{\theta_0}[Z_1] = 0$ we have by (115)

$$(1-\alpha) E_{\theta_0}[S_N \mid S_N \leqslant b] = -\alpha E_{\theta_0}[S_N \mid S_N \geqslant a] \tag{117}$$

and thus by (113)

$$(\beta')^2 = \alpha^2 (E_{\theta_0}[S_N \mid S_N \geqslant a])^2 = (1-\alpha)^2 (E_{\theta_0}[S_N \mid S_N \leqslant b])^2.$$

This means that

$$\begin{aligned} (\beta')^2 \left(\frac{1}{\alpha} + \frac{1}{1-\alpha} \right) &= \alpha (E_{\theta_0}[S_N \mid S_N \geqslant a])^2 + (1-\alpha)(E_{\theta_0}[S_N \mid S_N \leqslant b])^2 \\ &\leqslant \alpha E_{\theta_0}[S_N^2 \mid S_N \geqslant a] + (1-\alpha) E_{\theta_0}[S_N^2 \mid S_N \leqslant b] = E_{\theta_0}[N] \cdot E_{\theta_0}[Z_1^2] \end{aligned} \tag{118}$$

and

$$(\beta')^2 \leqslant \alpha(1-\alpha)\sigma^2 n_0. \quad \blacksquare \tag{119}$$

The estimate for the one-dimensional case can now be used to get an analogous estimate for the multidimensional case.

LEMMA 11. *Let $X_1, X_2, \ldots$ be i.i.d. random variables whose density $f(x, \theta)$ depends on a k-dimensional parameter $\theta = (\gamma, \delta^T)^T$ where γ is a main parameter and δ is a nuisance parameter and suppose that conditions 1 and 2 are satisfied. If $\beta(\theta)$ is the power of any size α test of $\theta = \theta_0 = (\gamma_0, \delta_0^T)^T$*

against $\gamma > \gamma_0$ with expected sample size $\leqslant n_0$ in θ_0 *and all δ derivatives of $\beta(\theta)$ in $\theta = \theta_0$ equal to* 0, *then*

$$\left[\frac{\partial \beta}{\partial \gamma}\right]_{\theta=\theta_0} \leqslant \alpha^{1/2}(1-\alpha)^{1/2}\sigma n_0^{1/2} \tag{120}$$

where

$$\sigma^2 = J^T I J,$$

I and J are the $k \times k$ covariance matrix and associated k vector defined in Theorem 3.

Proof. When Conditions 1 and 2 are satisfied the power function of any test is differentiable and for a test satisfying that all δ-derivatives in $\theta = \theta_0$ being equal to 0 we must have

$$\text{grad}_{\theta_0}\beta(\theta) = \beta' \cdot e_1 \tag{121}$$

where e_1 is the unit vector $e_1 = (1, 0, 0, \ldots, 0)^T$.

This means that the derivative with respect to η for $\eta = 0$ of $\beta(\theta_0 + \eta\lambda)$ where λ is a k vector with positive first component λ_1 must be

$$\beta' e_1^T \lambda = \beta' \lambda_1.$$

By Lemma 10 however

$$(\beta' \lambda_1)^2 \leqslant \alpha(1-\alpha)\sigma_\lambda^2 n_0 \tag{122}$$

where

$$\sigma_\lambda^2 = \text{Var}_{\theta_0}\left(\sum_{j=1}^{k} \lambda_1 \left[\frac{\partial}{\partial \theta_j} \ln f(X_1, \theta)\right]_{\theta=\theta_0}\right)$$

and this must be satisfied *for all* unit vectors λ with $\lambda_1 > 0$. Denoting

$$\left[\frac{\partial}{\partial \theta_j} \ln f(X_1, \theta)\right]_{\theta=\theta_0} = W_j \tag{123}$$

we will now find minimum of

$$\frac{1}{\lambda_1^2}\text{Var}_{\theta_0}\left(\sum_{j=1}^{k} \lambda_j W_j\right) = \text{Var}_{\theta_0}\left(W_1 + \sum_{j=2}^{k} \frac{\lambda_j}{\lambda_1} W_j\right). \tag{124}$$

This is just the problem of predicting W by linear combination of $W_2, \ldots, W_k$ with a minimal mean square error. It is easily seen that the best predictor is $-I_{\delta\delta}^{-1} I_{\delta\gamma}(W_2, W_3, \ldots, W_k)^T$ and that the minimal variance is $J^T IJ$.

Thus we have shown that

$$(\beta')^2 \leqslant \alpha(1-\alpha) J^T I J n_0. \qquad \blacksquare \tag{125}$$

When the expected sample size n_0 tends to ∞ the upper estimate of the γ-derivative of the power function also tends to ∞. With a normalization factor $n_0^{-1/2}$ we have

$$n_0^{-1/2} \sup \left(\left[\frac{\partial \beta}{\partial \gamma} \right]_{\theta = \theta_0} \right) \leqslant \alpha^{1/2}/(1-\alpha)^{1/2} (J^T IJ)^{1/2}. \tag{126}$$

It is easily seen from formulas (88) and (91) that for the asymptotic approximation of the D test

$$(E_0[\tau])^{1/2} \left[\frac{\partial P(W(t) = a)}{\partial \Gamma^*} \right]_{\Gamma^* = 0} = \alpha^{1/2}(1-\alpha)^{1/2} \sigma \tag{127}$$

where $\alpha = -b/(a-b)$ and $\sigma^2 = J^T IJ$. The same is the case for the D1 test.

This means that the asymptotic approximations for the D1 and D2 tests have a power function derivative which is equal to the best possible for each value of the nuisance parameter. For the real asymptotic properties taking into consideration idle time and truncation time the result can be formalized as in the following theorem.

THEOREM 5. *Suppose that Conditions* 1 *and* 2 *are satisfied and suppose that* $X_n^{(1)}(t, \theta)$ *is limiting tail inferior in* θ on $[t_0, t_1]$. *Let* $\beta_n(\theta)$, $n = 1, 2, \ldots$, *be the power functions of a sequence of asymptotic level* α *differentation* (D1 *or* D2) *tests of* $\gamma = \gamma_0$ *against* $\gamma > \gamma_0$ *with idle time* nt_0, *truncation time* nt_1 *and continuation region of the type*

$$b < X_n^{(D)}\left(\frac{m}{n}\right) < a$$

for $nt_0 < m < nt_1$, *where the sequence* $X_n^{(D)}\left(\frac{m}{n}\right)$ *is either* $X_n^{(D1)}\left(\frac{m}{n}\right)$ *or* $X_n^{(D2)}\left(\frac{m}{n}\right)$. *Further let* N_n *for* $n = 1, 2, 3, \ldots$ *denote the sequence of stopping times for these tests. Then for each* $\theta_0 = (\gamma_0, \delta_0)^T$

$$\lim_{n \to \infty} (E_{\theta_0}[N_n])^{-1/2} \left[\frac{\partial \beta(\theta)}{\partial \gamma} \right]_{\theta = \theta_0} = \alpha^{1/2}(1-\alpha)^{1/2}(J^T IJ)^{1/2} \cdot g(\alpha, t_0, t_1)$$

where

$$\lim_{\substack{t_0 \to \infty \\ t_1 \to \infty}} g(\alpha, t_0, t_1) = 1$$

for each fixed α, $0 < \alpha < 1$.

Proof. By Theorem 3 the random process $X_n^{(D2)}(t)$ converges weakly to a Brownian motion process $W(t) = J^T X'(t, 0)$ on $t_0 \leqslant t \leqslant t_1$. Thus

$$\lim_{n\to\infty} \frac{1}{n} E_{\theta_0}[N_n] = E_0[\tau] \tag{128}$$

where

$$\tau = \min\big(t_1, \inf_{t_1 > t}\{t\colon W(t) \notin (b, a)\}\big). \tag{129}$$

Further when Conditions 1 and 2 are satisfied

$$\left[\frac{\partial \beta_n(\theta)}{\partial \gamma}\right]_{\theta=\theta_0} = E_{\theta_0}[S_{N_n} \cdot \varphi_N] \tag{130}$$

where

$$S_m = \sum_{j=1}^{m} \left[\frac{\partial}{\partial \gamma} \ln f(X_j, \theta)\right]_{\theta=\theta_0}$$

and φ_m is the indicator of rejection for stopping with m observations and by the weak convergence results in Theorem 2

$$\lim n^{-1/2} \left[\frac{\partial \beta(\theta)}{\partial \gamma}\right]_{\theta=\theta_0} = E_0[e^T X'(\tau, 0) \cdot \varphi(\tau)] \tag{131}$$

where $e^T = (1, 0, 0, \ldots, 0)$ and $\varphi(t)$ is the indicator of rejection for making decision at time t. Now make a decomposition of $e^T X'(t, 0)$ into two orthogonal components

$$e^T X'(t, 0) = cW(t) + W_1(t) \tag{132}$$

where

$$W(t) = J^T X'(t, 0),$$

$$c = \operatorname{Cov}\big(W(t), e^T X'(t, 0)\big)/\operatorname{Var} W(t)$$

and

$$\operatorname{Cov}\big(W(t), W_1(t)\big) = 0.$$

Here

$$c = I_{\gamma\gamma}/J^T IJ \tag{133}$$

is independent of t and the two random processes $\{W(t)\colon 0 \leqslant t < \infty\}$ and $\{W_1(t)\colon 0 \leqslant t < \infty\}$ are independent Brownian motions. Thus we get

$$
\begin{aligned}
E_0[e^T X'(t, 0)\varphi(\tau)] \\
&= c \cdot E_0[W(\tau)\cdot\varphi(\tau)] + E_0[W_1(\tau)\cdot\varphi(\tau)] \\
&= c \cdot E_0[W(\tau)\varphi(\tau)] + E_0[\varphi(\tau)]\cdot E_0[W_1(\tau) \mid W(t), 0 \leqslant t < \infty] \\
&= \frac{I_{\gamma\gamma}}{J^T IJ} E_0[W(\tau)\varphi(\tau)] \qquad (134)
\end{aligned}
$$

since $E_0[W_1(t)] = 0$.

Finally a straightforward calculation for a Brownian motion $W(t)$ with the stopping time τ shows that

$$(E_0[\tau])^{1/2} \frac{I_{\gamma\gamma}}{J^T IJ} E_0[W(\tau)\varphi(\tau)]$$

has the form

$$\alpha^{1/2}(1-\alpha)^{1/2}(J^T IJ)^{1/2} g(\alpha, t_0, t_1)$$

where

$$\lim_{\substack{t_0 \to 0 \\ t_1 \to \infty}} g(\alpha, t_0, t_1) = 1$$

for every α. It is to be observed that the result holds independent of the value of φ (i.e., of the final decision) at truncation time t_1. The proof for the test D1 follows the same lines.

For the sake of simplicity we have concentrated in this section on the utmost local properties of power function derivatives at the boundary between hypothesis and alternative and found that D1 and D2 tests are asymptotically optimal. There might be formulated an asymptotic theory for local alternatives of the type

$$\gamma = \gamma_0 + \Gamma_1 n^{-1/2}$$

where B, C1 and C2 tests are asymptotically optimal in a suitable formulation.

Acknowledgment. Thanks are due to Bo Nilsson and Tommy Norberg for valuable discussions of various parts of this paper.

References

[1] M.S. Bartlett (1946), *The large sample theory of sequential test*, Proc. Cambridge Philos. Soc. **42**, 239–244.

[2] R. H. Berk (1975a)), *Locally most powerful sequential tests*, Ann. Statist. **3**, 373–381.

[3] — (1975b), *Comparing sequential and nonsequential tests*, Ann. Statist. **3**, 991–998.

[4] P. Billingsley (1968), *Convergence of Probability Measures*, Wiley, New York.

[5] — (1979), *Probability and Measure*, Wiley, New York.

[6] N. Breslow (1969), *On large sample sequential analysis with applications to survivorship data*, J. Appl. Probab. **6**, 261–274.

[7] D. R. Cox (1963), *Large sample sequential tests for composite hypotheses*, Sankhyā Ser. A **25**, 5–12.

[8] A. Dvoretsky, J. Kiefer and J. Wolfowitz (1953), *Sequential decision problems for processes with continuous time parameter. Testing hypotheses*, Ann. Math. Statist. **24**, 254–264.

[9] I. I. Gikhman and A. V. Skorokhod (1969), *Introduction to the Theory of Random Processes*, W. B. Saunders, Philadelphia.

[10] Z. Govindarajulu (1981), *The Sequential Statistical Analysis of Hypothesis Testing, Point and Interval Estimation and Decision Theory*, American Sciences Press, Columbus.

[11] E. Hewitt and K. Stromberg (1965), *Real and Abstract Analysis*, Springer-Verlag, Berlin–Heidelberg.

[12] M. Kohlberger (1978), *Asymptotik sequentieller Tests für zusammengesetzte Hypothesen bei Nebenparametern*, Diplomarbeit, Albert-Ludvigs-Universität, Freiburg.

[13] T. L. Lai (1978), *Pitman efficiencies of sequential tests and uniform limit theorems in nonparametric statistics*, Ann. Statist. **6**, 1027–1047.

[14] E. L. Lehmann (1959), *Testing Statistical Hypotheses*, Wiley, New York.

[15] G. F. Simmons (1963), *Introduction to Topology and Modern Analysis*, McGraw-Hill, New York.

[16] R. A. Wijsman (1977), *A general theorem with applications on exponential bounded stopping time without moment conditions*, Ann. Statist. **5**, 292–315.

Presented to the semester
Sequential Methods in Statistics
September 7-December 11, 1981