The Annals of Statistics
1988, Vol. 16, No. 2, 483–520

ON THE PRESERVATION OF LOCAL ASYMPTOTIC NORMALITY UNDER INFORMATION LOSS[1]

BY LUCIEN LE CAM AND GRACE L. YANG

University of California, Berkeley and University of Maryland

1. Introduction. In the present paper we consider a situation where there are unobservable random variables $X_{j,n}$; $j = 1, \ldots, k_n$; $n = 1, 2, \ldots,$ and where what is actually seen are other variables $Y_{j,n}$ that are less informative than the $X_{j,n}$. For instance, the $Y_{j,n}$ may be functions of the $X_{j,n}$.

It can then readily happen that the family of probability measures that governs the behavior of the $X_{j,n}$ is simple and easily studied but that the corresponding family for the $Y_{j,n}$ is more untractable. We shall show that if the $\{X_{j,n}\}$ satisfy certain conditions, such as the LAN conditions, then the $Y_{j,n}$ will also satisfy the same requirements. This means that certain methods of estimation and testing can be carried out with the $Y_{j,n}$ with some assurance of success without having to verify that their distributions satisfy the necessary requirements. Actual computation of the estimates may still be a complex task, but not nearly as difficult as the analytical effort needed in the verification of assumptions.

The results are proved under suitable negligibility conditions imposed on the $X_{j,n}$ and mostly for the case where the $X_{j,n}$ are independent for each fixed n. However, we point out that a similar phenomenon can also be expected for certain nonindependent double arrays $\{X_{j,n}\}$.

The problem was brought to our attention by several special examples. One of them is in a neurophysiological study where the underlying model involved independent variables U_ν, $X_{1,\nu}, \ldots$. The variables U_ν had a binomial distribution $B(m, p)$ and the $X_{j,\nu}$ were positive i.i.d. variables whose common distribution depended smoothly on a one-dimensional parameter θ. Standard textbook results on the asymptotic behavior of the system were readily applicable to the variables $\{U_\nu, X_{1,\nu}, \ldots\}$ but what could be observed were only the sums

$$Y_\nu = \sum_k \{X_{k,\nu}; 0 \le k \le U_\nu\}, \qquad \nu = 1, 2, \ldots, n.$$

Another example comes from what is called the use of counting processes for life history data (see [1], [3] and [15]). One has a large number n of different individuals behaving independently of each other. During the observation period $[0, L]$, certain events A_i, $i = 1, 2, \ldots, I$, can happen to them. At time t the instantaneous intensity for the occurrence of event A_i for individual k is a variable $\lambda_{k,i}(t)$. As functions of t these $\lambda_{k,i}(t)$ are predictable random processes in the sense that $\lambda_{k,i}(t)$ depends only on the history of individual k in the

Received November 1985; revised August 1987.
[1]Research partially supported by National Science Foundation Grant MCS-84-03239.
AMS 1980 *subject classifications*. 62F05, 62F12, 62G20, 62M05.
Key words and phrases. Asymptotic normality, information loss, LAN, one step estimates.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics*.
www.jstor.org

half-open interval $[0, t)$. In such problems there is a well-developed theory applicable to the situation where each individual is monitored continuously throughout the entire period $[0, L]$. Its elaboration, reviewed in [3], relies materially on the fact that likelihood ratios can be written in a simple explicit form.

The question arises whether the theory can be made applicable to a situation where each individual is seen only at a few isolated times, say $0 \le t_0 < t_1 < \cdots < t_8 \le L$. Even in a purely Markovian system, assumptions made on the intensities $\lambda_{k,i}$ do not translate into simple statements on the likelihood ratios for isolated observation times. The fact that our theorems give some information on the behavior to be expected of likelihood ratios allows one to direct numerical computations in an efective manner.

Our main results depend very strongly on the negligibility assumptions placed on the $X_{j,n}$. It is a trivial fact that the results are not valid without some restrictions, but, except for a modification of a result of Davies [9] given in Section 6, we do not know whether extensions to other situations are possible.

Section 2 gives the essential notation and assumptions with a statement of the problem in a formal mathematical framework. Section 3 shows that contiguity is preserved under loss of information. It also contains related inequalities that will be used in the following sections. Section 4 is about the preservation of asymptotic normality for what we call bounded infinitesimal arrays. These are double arrays of independent experiments where the total information (defined in a suitable manner) remains bounded and where each individual experiment contributes an asymptotically negligible amount of information to the total.

Asymptotic normality has been used mostly in connection with the so-called LAN conditions (see [10], [20], [25] and [34]). These involve not only the local asymptotic normality that gave them their name but also linear relations between vector parameters and logarithms of likelihood ratios. It is shown in Section 5 that such linear relations are also preserved. The technique of proof involves the use of quadratic forms that "control" the asymptotic behavior of likelihood ratios. According to Davies [9], such "quadratic control" may be preserved under certain information losses that do not actually preserve the local Gaussian behavior. It is the main relation needed for the improvement of auxiliary estimates, described later in the same section.

Section 5 ends in a description of a method of construction of estimates. One starts with a good auxiliary estimate and adds to it a correction calculated from likelihood ratios. This gives asymptotically efficient estimates. The general method has been described elsewhere by one of the present authors. It is pointed out that, for our infinitesimal arrays, the method admits a number of variants that are often easily applicable in practice.

An application to the neurophysiological problem that motivated us is described in some detail in Example 1, Section 9.

Section 6 departs from the general theme. It does not use independence or negligibility assumptions. What it shows is that joint asymptotic normality of logarithms of likelihood and of estimates available from the restricted information together with a condition that these estimates be "distinguished" leads to preservation of the LAN conditions. For that LAN situation the result is an improvement of a result of Davies who dealt in [8] and [9] with preservation of what we call "quadratic control."

Section 7 is an aside on the independent identically distributed case. It shows that differentiability in quadratic mean (for square roots of likelihood ratios) is preserved under information loss. This is not surprising.

Section 8 sketches a possible extension of the results of Sections 4 and 5 to cases where the observations are not independent but where the information is still acquired by infinitesimal amounts. It is clearly applicable to certain Markov models where the information is lost by observing only at isolated times instead of continuously. Some of the difficulties that arise in the non-Markovian case are pointed out.

Section 9 contains several examples intended as illustrations. The first is the already mentioned neurophysiological problem that started our investigation. The second is intended to show in an i.i.d. situation that one should not expect that good behavior of maximum likelihood estimates would be preserved. It is also a warning against iterative procedures often used to compute approximations to the maximum likelihood estimates.

Example 3 explains how to deal with homogeneous Markov processes. It also contains a warning about loss of identifiability for the instantaneous transition intensities.

Example 4 shows that asymptotic normality may be preserved even though the rates of convergence of estimates are altered. This is to emphasize the warnings of Section 2.

Example 5 discusses the method of moments used in estimation.

Example 6 shows that the techniques proposed here apply to the loss of information incurred by grouping data.

Finally, Example 7 allows one to pass from ordinary data to censored data. Here again the identifiability problems may interfere. Otherwise solutions are reasonably simple. Each section, except Section 2, begins by a short description of its content. The reader who is not interested in the techniques of proofs may skip them. It is possible to read the examples first, referring as needed to the statements of the theorems.

2. Notation and assumptions. The main results of the present paper refer to a situation describable as follows. For each integer n, let Θ_n be a parameter set. For each pair of integers (j, n); $j = 1, 2, \dots, k_n$, $n = 1, 2, \dots$, let $\mathscr{E}_{j,n}$ be an experiment indexed by Θ_n, that is, a family $\mathscr{E}_{j,n} = \{p_{j,n}(\theta);\ \theta \in \Theta_n\}$ of probability measures on a σ-field $\mathscr{A}_{j,n}$. Let $\bar{\mathscr{E}}_n$ be the direct product of the $\mathscr{E}_{j,n}$, that is, $\bar{\mathscr{E}}_n$ is the experiment where one observes independent random elements $X_{j,n}$ in such a manner that, when θ is true, the distribution of $X_{j,n}$ is given by $p_{j,n}(\theta)$. The joint distribution is given by the product measure

$$P_n(\theta) = \prod_j p_{j,n}(\theta).$$

Consider another double array of experiments $\mathscr{F}_{j,n} = \{q_{j,n}(\theta);\ \theta \in \Theta_n\}$ and the corresponding direct product $\bar{\mathscr{F}}_n$. Assume that $\mathscr{F}_{j,n}$ is less informative than the corresponding $\mathscr{E}_{j,n}$ in the sense of Blackwell [6], that is, for every decision problem, any risk function possible on $\mathscr{F}_{j,n}$ is also possible in $\mathscr{E}_{j,n}$. This can happen, for instance, if $q_{j,n}(\theta)$ is the restriction of $p_{j,n}(\theta)$ to a sub-σ-field $\mathscr{B}_{j,n}$ of $\mathscr{A}_{j,n}$.

In such a situation one may inquire whether certain asymptotic properties of the sequence $\bar{\mathscr{E}}_n$ remain valid for the weaker sequence $\bar{\mathscr{F}}_n$. For instance, let s_n and t_n be two points of Θ_n. Consider the product measures $P_n(s_n)$ and $P_n(t_n)$ and the corresponding products $Q_n(s_n)$ and $Q_n(t_n)$ relative to the experiments $\bar{\mathscr{F}}_n$.

(a) If $\{P_n(s_n)\}$ and $\{P_n(t_n)\}$ are contiguous sequences, is the same true of $\{Q_n(s_n)\}$ and $\{Q_n(t_n)\}$?
(b) If the pair $[P_n(s_n), P_n(t_n)]$ is an asymptotically Gaussian experiment, is the same true of $[Q_n(s_n), Q_n(t_n)]$?
(c) If Θ_n is a Euclidean space R^k and the $\bar{\mathscr{E}}_n$ satisfy the LAN conditions, is the same true of the $\bar{\mathscr{F}}_n$?

The answer to question (a) is easily seen to be positive. This will be proved in Section 3. The answer to (b) and (c) is more complex. We shall show that it is also positive under a "negligibility" requirement imposed on the components $\mathscr{E}_{j,n}$. However, the positivity of the answer must be qualified in an important manner. To indicate the qualifications, note that, for such product experiments, there are some natural metrics definable as follows. Let

$$h^2_{j,n}(s,t) = \tfrac{1}{2}\int\left[\sqrt{dp_{j,n}(s)} - \sqrt{dp_{j,n}(t)}\right]^2$$

be the square Hellinger distance for $\mathscr{E}_{j,n}$. Let $H^2_n(s,t) = \sum_j h^2_{j,n}(s,t)$. Using the alternate measures $q_{j,n}(s)$ in a similar manner, define from $\bar{\mathscr{F}}_n$ a distance K_n analogous to H_n.

For the experiments $\{P_n(\theta);\ \theta \in \Theta_n\}$, the metric H_n is a very natural one. The local asymptotic theories are often carried out on neighborhoods of the type

$$V_n(\theta, b) = \{t;\ t \in \Theta_n,\ H_n(t,\theta) \le b\},$$

for a presumed true value θ. This is what one does, more or less, for the LAN assumptions. This is also what one does in the i.i.d. case for many nonparametric studies. (See [31], for instance.)

The corresponding neighborhoods for the $\bar{\mathscr{F}}_n$ would be

$$W_n(\theta, b) = \{t;\ t \in \Theta_n,\ K_n(t,\theta) \le b\}.$$

Note that since $K_n \le H_n$, one has $V_n(\theta, b) \subset W_n(\theta, b)$ but that $W_n(\theta, b)$ may be immensely larger than $V_n(\theta, b)$ (in terms of the distance H_n).

What we shall show is that certain properties, such as asymptotic normality, are inherited by the $\{Q_n(\theta);\ \theta \in \Theta_n\}$ *on the sets of the type* V_n. Nothing can be said in general about the larger W_n, although Section 7 does contain a result of that nature.

For the same reason, existence of estimates that converge at a given speed on the $\{P_n(\theta);\ \theta \in \Theta_n\}$ is not a property that transfers to the weaker $\{Q_n(\theta);$

$\theta \in \Theta_n\}$, since, for instance, the measures $Q_n(\theta)$; $\theta \in \Theta_n$ might be all equal to a single one, $Q_n(\theta_0)$. Section 7 again, contains some further elaborations on this point. The reader may have noted that, in the description of the experiments $\bar{\mathscr{E}}_n = \{P_n(\theta);\ \theta \in \Theta_n\}$, the index set is allowed to vary with n. This again because of the local nature of the results. It is sometimes possible to stabilize the sets Θ_n by various transformations, such as using $\theta = \sqrt{n}(t - t_0)$ or more elaborate matrix multiplication. Then the Θ_n may be replaced by a fixed set Θ independent of n. However such stabilizations are essentially irrelevant to our main purposes and they would make matters more complex since the stabilizing transformation needed for $\bar{\mathscr{F}}_n$ may be different from that needed for $\bar{\mathscr{E}}_n$.

3. Preservation of contiguity and some other inequalities. The purpose of this section is to show that contiguity is always preserved under loss of information. This can readily be established directly, but we use a more specific approach yielding inequalities on likelihood ratios. They will be used in the subsequent sections.

One first easy remark is as follows.

PROPOSITION 1. *Let $\{P_n(s_n)\}$ and $\{P_n(t_n)\}$ be contiguous. Then the sequences $\{Q_n(s_n)\}$ and $\{Q_n(t_n)\}$ are also contiguous.*

This is a special case of a more general lemma (see [27], Chapter 6, Section 2).

LEMMA 1. *Let $\mathscr{E}_n = \{P_{n,0}, P_{n,1}\}$ and $\mathscr{F}_n = \{Q_{n,0}, Q_{n,1}\}$ be two binary experiments. Assume that* (1) *the deficiency $\delta(\mathscr{E}_n, \mathscr{F}_n)$ of $\mathscr{E}_n$ with respect to $\mathscr{F}_n$ tends to* 0 *and* (2) *the sequences $\{P_{n,0}\}, \{P_{n,1}\}$ are contiguous. Then the sequences $\{Q_{n,0}), \{Q_{n,1}\}$ are contiguous.*

This lemma itself is an immediate consequence of an inequality that will be used in Section 4. To state it, let S be a set of probability measures on a σ-field $\mathscr{A}$. Let λ be another positive finite measure on $\mathscr{A}$. Let us say that the set S is $(\alpha, c\lambda)$ limited if $\|P - P \wedge c\lambda\| \le \alpha$ for every $P \in S$.

Here the norm is the total variation norm and the measure $P \wedge c\lambda$ is defined in the following manner. One takes the Radon–Nikodym density $dP/d\lambda$ of the part of P that is dominated by λ. One takes for $P \wedge c\lambda$ the measure whose density with respect to λ is the minimum of $dP/d\lambda$ and c.

LEMMA 2. *Let $\mathscr{E} = \{P_\theta;\ \theta \in \Theta\}$ and $\mathscr{F} = \{Q_\theta;\ \theta \in \Theta\}$ be two experiments with the same set of indices Θ. Assume*

(i) *$\{P_\theta;\ \theta \in \Theta\}$ is $(\alpha, c\lambda)$ limited;*
(ii) *the deficiency $\delta(\mathscr{E}, \mathscr{F})$ does not exceed $\varepsilon/2$.*

Then there is a positive linear transformation T that sends probability measures into probability measures and is such that

(iii) *$\|Q_\theta - TP_\theta\| \le \varepsilon$ for all θ;*
(iv) *if $\mu = T\lambda$, then $\mathscr{F}$ is $(\alpha + \varepsilon, c\mu)$ limited.*

PROOF. The existence of T is part of a general theorem of [21]. Once this is established note that

$$T(P_\theta \wedge c\lambda) \le (TP_\theta) \wedge (Tc\lambda) = (TP_\theta) \wedge (c\mu).$$

Also

$$\|T(P_\theta - P_\theta \wedge c\lambda)\| \le \|P_\theta - P_\theta \wedge c\lambda\| \le \alpha.$$

Therefore,

$$\|TP_\theta - (TP_\theta) \wedge (c\mu)\| \le \alpha.$$

Now write $Q_\theta - TP_\theta = D = D^+ - D^-$. Then $Q_\theta \le TP_\theta + D^+$ and $Q_\theta \wedge c\mu \ge (TP_\theta - D^-) \wedge c\mu$. Therefore,

$$\|Q_\theta - Q_\theta \wedge c\mu\| \le \|TP_\theta - (TP_\theta) \wedge (c\mu)\| + \|D\|.$$

This gives the desired result. □

Applying this to the case where $\delta(\mathscr{E}, \mathscr{F}) = 0$, one sees that truncating the densities of the measures Q_θ results in a modification of these measures that is smaller than the modification made by the same truncation on the P_θ.

One should expect that some other inequalities would also be preserved. Many arguments involving maximum likelihood estimates use bounds on expresssions of the type

$$\phi(x, V) = \sup_{t \in V} \left[\frac{f(x,t)}{f(x,\theta_0)} - 1 \right]^+,$$

for densities $f(x,t) = dP_t/d\lambda$. A bound on an integral $\int \phi(x, V)\, dP_{\theta_0}$ will carry over to the similar expression defined on the weaker $\mathscr{F}$. However, some precautions must be taken. See the examples of Section 9.

4. Preservation of asymptotic normality. In this section we consider experiments that are weakly asymptotically normal in the sense of the Gaussian approximability described in Definition 2. It is first shown that this property is equivalent to approximability of distributions of log-likelihood by multivariate Gaussian distributions at least whenever Hellinger affinities remain bounded away from 0.

This being done, we consider independent observations where the individual experiments form what we call bounded infinitesimal arrays (Definition 3). For these it is shown that, when loss of information occurs on each individual component, asymptotic normality is preserved.

The approximations by Gaussian experiments involve quadratic forms that "control" the behavior of likelihood ratios (Definition 4). The forms that control the weaker experiments are smaller than the initial ones in the sense that the difference is positive semidefinite.

Explicitly, the situation can be described as follows.

An experiment $\mathscr{G} = \{G_\theta;\ \theta \in \Theta\}$ is called Gaussian if it satisfies the following two conditions:

(i) For any pair (s, t) of elements of Θ the measures G_s and G_t are mutually absolutely continuous.

(ii) Let $\Lambda(t, s) = \log(dG_t/dG_s)$. The stochastic process $t \rightsquigarrow \Lambda(t, s)$, $t \in \Theta$, is a Gaussian process for the distributions induced by G_s. Here s is an arbitrary point of Θ.

It is easily checked that, under (i), if condition (ii) is satisfied for some choice of s, it is satisfied for all.

Definition 1. Let $\bar{\mathscr{E}}_n = \{P_n(\theta);\ \theta \in \Theta_n\}$ be a sequence of experiments. One says that $\bar{\mathscr{E}}_n$ admits strong Gaussian approximations if there are Gaussian experiments $\mathscr{G}_n = \{G_n(\theta);\ \theta \in \Theta_n\}$ such that the distance $\Delta(\bar{\mathscr{E}}_n, \mathscr{G}_n)$ tends to 0.

(The distance Δ is the one defined in [21] or [27].)

Definition 2. One says that $\bar{\mathscr{E}}_n$ admits weak Gaussian approximations if there are Gaussian $\mathscr{G}_n$ such that for any subsets $S_n \subset \Theta_n$, which have a cardinality bounded independently of n, the distance between $\bar{\mathscr{E}}_n$ and $\mathscr{G}_n$ restricted to S_n tends to 0.

The condition of Definition 2 is similar to a relation often used in asymptotics under circumstances where Θ_n is a fixed set Θ independent of n. Then one can speak of weak convergence of $\bar{\mathscr{E}}_n$ to a Gaussian limit $\mathscr{G} = \{G_\theta;\ \theta \in \Theta\}$. This is convergence for the distance Δ for $\bar{\mathscr{E}}_n$ restricted to *fixed* finite subsets $S \subset \Theta$. Note that the weak approximability of Definition 2 is considerably stronger than what would be obtained from weak convergence to a limit $\mathscr{G}$. Our sets $S_n \subset \Theta$ have bounded cardinality, just as a fixed finite S, but they *are allowed to vary arbitrarily with n*. Thus our weak approximability is uniform on all sets of a given finite cardinality. It would be inconvenient here to use weak convergence to limits since our successive Θ_n need not be related at all.

Definitions 1 or 2 do not put any additional restrictions on the Gaussian $\mathscr{G}_n$. Under some restrictions the weak approximability can be checked on the behavior of distributions of likelihood ratios as follows.

Proposition 2. *Let $\bar{\mathscr{E}}_n = \{P_n(\theta);\ \theta \in \Theta_n\}$ be a sequence of experiments. Assume that there is some $\varepsilon > 0$ such that*

$$\inf_{n, s, t} \int \sqrt{dP_n(s)\, dP_n(t)} \geq \varepsilon.$$

Then the $\mathscr{E}_n$ admit weak Gaussian approximations if and only if for every fixed k and for every subset $\{s_{0,n}, s_{1,n}, \ldots, s_{k,n}\}$ of elements of Θ_n the joint distributions F_n under $s_{0,n}$ of the logarithms

$$\log \frac{dP_n(s_{j,n})}{dP_n(s_{0,n})}, \qquad j = 1, 2, \ldots, k,$$

are approximable by Gaussian distributions in the sense that the Lévy (or Prohorov) distance between F_n and a suitable k-variate Gaussian distribution tends to 0 *as* $n \to \infty$.

NOTE. By $dP(t)/dP(s)$ is meant the density with respect to $P(s)$ of the part of $P(t)$ dominated by $P(s)$.

PROOF. Consider first two point sets $\{s_{0,n}, s_{1,n}\}$ and the corresponding binary experiments $\mathscr{E}_n' = \{P_n(s_{0,n}), P_n(s_{1,n})\}$. Taking a subsequence if necessary, one can assume convergence to a limit $\mathscr{E}' = \{R_0, R_1\}$. The lower bound imposed on affinities implies that the R_i, $i = 0, 1$, cannot be disjoint. Let X_n be distributed as

$$\log \frac{dP_n(s_{1,n})}{dP_n(s_{0,n})},$$

under $P_n(s_{0,n})$. There are numbers $a < b$ such that $\Pr[a \le X_n \le b]$ remains bounded away from 0. Let Y_n be a normal variable with a distribution $\mathscr{N}(\mu_n, \sigma_n^2)$ approximating that of X_n. Then $\Pr[a - 1 \le Y_n \le b + 1]$ stays bounded away from 0. It follows that σ_n must remain bounded away from $+\infty$.

If so, $|\mu_n|$ must also remain bounded. This means that cluster points of the sequence of distributions $\mathscr{L}(e^{X_n})$ cannot have masses at 0. Thus R_1 dominates R_0. This implies that the sequences $\{P_n(s_{0,n})\}$ and $\{P_n(s_{1,n})\}$ must be contiguous. It then follows by a standard argument that the conditions given are sufficient to imply the weak Gaussian approximability for an arbitary k. The result in the opposite direction is also the consequence of a standard argument for which see for instance [22], page 14. □

Keeping in mind this result, let us return to the case of product experiments $\bar{\mathscr{E}}_n = \prod_j \mathscr{E}_{j,n}$ described in Section 2.

DEFINITION 3. The double array $\{\mathscr{E}_{j,n}\}$ with $\mathscr{E}_{j,n} = \{p_{j,n}(\theta), \theta \in \Theta_n\}$ will be called bounded and infinitesimal if it satisfies the following two requirements for all pairs (s_n, t_n) extracted from Θ_n:

$$\text{(A)} \qquad \sup_n \sum_j h_j^2(s_n, t_n) < \infty,$$

$$\text{(B)} \qquad \lim_n \sup_j h_j^2(s_n, t_n) = 0$$

(where h_j is the Hellinger distance defined in Section 2).

Note that when (B) holds, condition (A) is equivalent to the affinity restriction used in Proposition 2.

For such bounded infinitesimal arrays, weak Gaussian approximability can be shown to be equivalent to any one of a large number of other properties. The one that will be most convenient here is as follows.

Proposition 3. *Let $\{\mathscr{E}_{j,n}\}$ be a bounded infinitesimal array. Then the product $\bar{\mathscr{E}}_n$ admits weak Gaussian approximations if and only if for every choice of pairs (s_n, t_n) of elements of Θ_n and every $\varepsilon > 0$,*

$$\sum_j \|p_{j,n}(t_n) - p_{j,n}(t_n) \wedge (1+\varepsilon)p_{j,n}(s_n)\|$$

tends to 0.

Proof. Let

$$\Lambda_{j,n} = \log \frac{dp_{j,n}(t_n)}{dp_{j,n}(s_n)}.$$

The condition as stated is equivalent to the fact that, for every $\varepsilon > 0$, $\Sigma \Pr[\Lambda_{j,n}(t_n, s_n) > \varepsilon | t_n] \to 0$. Computing probabilities under s_n instead of t_n one obtains that $\Sigma_j \Pr[\Lambda_{j,n}(t_n, s_n) < -\varepsilon | s_n] \to 0$. However $\Pr[\Lambda_{j,n}(t_n, s_n) < -\varepsilon | t_n] \le \Pr[\Lambda_{j,n}(t_n, s_n) < -\varepsilon | s_n]$. Hence the result follows by the n.a.s.c. for the C.L.T. □

From this result one can immediately obtain the following.

Theorem 1. *Let $\{\mathscr{E}_{j,n}\}$ be a bounded infinitesimal array of experiments indexed by Θ_n. Let $\{\mathscr{F}_{j,n}\}$ be another array such that $\mathscr{F}_{j,n}$ is weaker than the corresponding $\mathscr{E}_{j,n}$. Let $\bar{\mathscr{E}}_n = \prod_j \mathscr{E}_{j,n}$ and $\bar{\mathscr{F}}_n = \prod_j \mathscr{F}_{j,n}$. If $\bar{\mathscr{E}}_n$ admits weak Gaussian approximations, so does $\bar{\mathscr{F}}_n$.*

Proof. It is clear that $\{\mathscr{F}_{j,n}\}$ is also a bounded infinitesimal array. By Lemma 2, Section 3, there are transformations $T_{j,n}$ such that the measures $q_{j,n}(t)$ of $\mathscr{F}_{j,n}$ satisfy $q_{j,n}(t) = T_{j,n}p_{j,n}(t)$ and also

$$\begin{aligned} &\|q_{j,n}(t_n) - q_{j,n}(t_n) \wedge (1+\varepsilon)q_{j,n}(s_n)\| \\ &\quad \le \|p_{j,n}(t_n) - p_{j,n}(t_n) \wedge (1+\varepsilon)p_{j,n}(s_n)\|. \end{aligned}$$

Hence the result. □

Note that Theorem 1 refers to weak approximability. At the time of the present writing, we do not know whether strong Gaussian approximability is preserved. However, a conjecture stated in [26] would imply that, for bounded infinitesimal arrays, strong and weak Gaussian approximability are equivalent.

The structure of a Gaussian experiment is well defined by certain quadratic forms that may be introduced as follows. Let $\mathscr{G}_n = \{G_{\theta,n}; \theta \in \Theta_n\}$ be a Gaussian experiment indexed by Θ_n. Let $\mathscr{M}_n$ be the linear space of finite signed measures μ that have finite support on Θ_n and are such that $\mu(\Theta_n) = 0$. Let $\Lambda'_n(t,s) = \log(dG_{t,n}/dG_{s,n})$ and let $\Gamma(\mu)$ be the variance of $\int \Lambda'_n(t,s)\mu(dt)$. This is a positive semidefinite quadratic form on $\mathscr{M}_n$. It can be used to complete a quotient of $\mathscr{M}_n$ to obtain a Hilbert space.

The quadratic form Γ controls the behavior of the process Λ'_n in the following sense: The integrals $\int [\Lambda'_n(t,s) + \frac{1}{2}\Gamma(\delta_t - \delta_s)]\mu(dt)$ are almost surely 0 if and only if $\Gamma(\mu) = 0$. In other words, linearity relations satisfied by the random part of Λ'_n are those describable through Γ.

For experiments that are only approximately Gaussian, the situation is not so neat but one can describe an analog as follows.

Let $\Lambda_n(t, s) = \log(dP_n(t)/dP_n(s))$. Let Γ_n be a positive semidefinite quadratic form on $\mathscr{M}_n$.

Definition 4. The sequence of experiments $\bar{\mathscr{E}}_n$ is under the control of the quadratic forms Γ_n if, for elements $\mu_n \in \mathscr{M}_n$ that have bounded mass ($\sup_n \|\mu_n\| < \infty$) and supports with bounded cardinality, convergence of $\Gamma_n(\mu_n)$ to 0 implies that

$$\int \left\{ \Lambda_n(t, s_n) + \tfrac{1}{2}\Gamma_n(\delta_t - \delta_{s_n}) \right\} \mu_n(dt)$$

tends to 0 in $P_n(s_n)$ probability.

This is obviously a way to say that, if $\Gamma_n(\mu_n)$ tends to 0, the recentered Λ_n satisfy approximately the corresponding linear relation. The literature contains several conditions meant to express that log-likelihood admit approximate linear–quadratic expansions. This is usually done assuming that Θ_n is a finite-dimensional vector space. One writes that $\Lambda_n(t + \theta, \theta)$ is approximately $t'V_n - \frac{1}{2}t'M_n t$. In the LAN conditions M_n is nonrandom. In the LAMN conditions M_n is random, the pair (V_n, M_n) has a limiting distribution $\mathscr{L}(V, M)$, where, conditionally given M, the variable $t'V$ is $\mathscr{N}[0, t'Mt]$. The case where this restriction on conditional distributions is omitted has been considered by Davies [9] as the case "when the amount of information is random." For these conditions see Basawa and Prakasa Rao [4], Basawa and Scott [5], Feigin [11] and Jeganathan [17].

Note that all these conditions imply a relation between the linear structure of Θ_n and that of the log-likelihood. Our quadratic control condition is different. It does not rely on any particular structure, linear or otherwise on Θ_n. To understand it, let us consider experiments $\mathscr{E} = \{P_\theta;\ \theta \in \Theta\}$ that do not depend on n at all. Suppose that Θ is the real line and that P_θ is the distribution of a two-dimensional vector (X, Y), jointly normal with covariance matrix the identity and expectations $E_\theta X = \theta$, $E_\theta Y = \sinh\theta$. Then $\mathscr{E}$ does not satisfy the LAN or LAMN conditions on Θ, yet it is controlled by a quadratic Γ such that $\Gamma(\delta_t - \delta_s) = |t - s|^2 + |\sinh t - \sinh s|^2$ for the Dirac masses δ_s and δ_t. This is typical of Gaussian shift experiments where the shift parameters are not linearly related to the original θ.

However, the quadratic control places a strong restriction on the log-likelihood. For instance, let P_θ be the ordinary gamma distribution with density $(\beta^\alpha/\Gamma(\alpha))e^{-\beta x}x^{\alpha-1}$. The log-likelihood is a linear function of the vector $\{\beta, \alpha, \log\Gamma(\alpha)\}$ but the experiment $\mathscr{E} = \{P_\theta;\ \theta \in \Theta\}$ with $\theta = (\alpha, \beta)$ does not admit quadratic control.

The following results say that linear relations are preserved by going from the product $\prod_j \mathscr{E}_{j,n}$ to the weaker $\prod_j \mathscr{F}_{j,n}$.

Theorem 2. *Let $\{\mathscr{E}_{j,n}\}$ be a bounded infinitesimal array of experiments $\mathscr{E}_{j,n} = \{p_{j,n}(\theta);\ \theta \in \Theta_n\}$. Let $\mathscr{F}_{j,n}$ be weaker than $\mathscr{E}_{j,n}$.*

If the products $\bar{\mathscr{E}}_n$ admit weak Gaussian approximations, they are under the control of quadratic forms Γ_n and the weaker products $\bar{\mathscr{F}}_n$ are under the control of quadratic forms Γ_n^* such that $\Gamma_n - \Gamma_n^*$ is positive semidefinite.

PROOF. Select a point $s_n \in \Theta_n$ and let

$$X_{j,n}(t) = \frac{dp_{j,n}(t)}{dp_{j,n}(s_n)} - 1.$$

Let $Y_{j,n}(t) = 1 \wedge X_{j,n}(t)$. It is easy to see that, under the conditions given, for any sequence t_n, the difference between $\Lambda_n(t_n, s_n)$ and $\Sigma_j Y_{j,n}(t_n) - \frac{1}{2}\Sigma_j \operatorname{var} Y_{j,n}(t_n)$ tends to 0 in $P_n(s_n)$ probability. (See for instance [22] or [27].)

On the space of measures $\mathscr{M}_n$, define $\Gamma_n(\mu)$ as the variance under s_n of $\int [\Sigma_j Y_{j,n}(t)]\mu(dt)$. Then the integral

$$\int \left\{ \Lambda_n(t, s_n) + \tfrac{1}{2}\Gamma_n\left[\delta_t - \delta_{s_n}\right]\right\}\mu_n(dt)$$

differs from $\int [\Sigma_j Y_{j,n}(t)]\mu_n(dt)$ by a quantity that tends to 0 in probability as long as $\|\mu_n\|$ remains bounded and as long as the supports of the μ_n have bounded cardinality. Thus the assertion that $\bar{\mathscr{E}}_n$ is under the control of the quadratic form Γ_n is equivalent to the statement that $\int [\Sigma_j Y_{j,n}(t)]\mu_n(dt)$ tends to 0 in probability whenever its variance tends to 0. However, the contiguity restrictions involved in the Gaussian approximability imply that the distribution of a term such as $\Sigma_j Y_{j,n}(t)$ is approximated by a Gaussian distribution with expectation 0. Hence the first assertion.

To obtain the second result, let us use again the transformations $T_{j,n}$ with $T_{j,n} p_{j,n}(t) \equiv q_{j,n}(t)$. Under some technical regularity conditions, these $T_{j,n}$ can be represented by Markov kernels. If we look only at finite subsets of Θ_n and only at distributions of likelihood ratios for the $\mathscr{F}_{j,n}$, these technical conditions can be assumed to be satisfied. Thus the situation can be described as follows. Let $\mathscr{A}_{j,n}$ and $\mathscr{B}_{j,n}$ be, respectively, the σ-fields of $\mathscr{E}_{j,n}$ and $\mathscr{F}_{j,n}$.

Let $K_{j,n}(B,\cdot)$ be the Markov kernel of $T_{j,n}$ evaluated at B in $\mathscr{B}_{j,n}$. Define measures by $M_{j,n}[t; A \times B] = \int_A K_{j,n}(B,\cdot)\, dp_{j,n}(t)$. Their marginals on $\mathscr{B}_{j,n}$ are the $q_{j,n}(t)$ and their likelihood ratios are $\mathscr{A}_{j,n}$-measurable, equivalent to those of the $p_{j,n}(t)$. Now replace our previous $Y_{j,n}(t)$ by the equivalent

$$\frac{dM'_{j,n}(t)}{dM_{j,n}(s_n)} - 1,$$

with $M'_{j,n}(t) = M_{j,n}(t) \wedge 2M_{j,n}(s_n)$. Let $Z_{n,j}(t)$ be the conditional expectation (under s_n) of $Y_{j,n}(t)$ given $\mathscr{B}_{j,n}$. This differs from

$$\frac{dq_{j,n}(t)}{dq_{j,n}(s_n)} - 1,$$

by an amount that can be neglected. It is thus possible to show that

$$\sum_j Z_{n,j}(t) - \tfrac{1}{2}\operatorname{var}\sum_j Z_{n,j}(t)$$

is an approximation to the logarithm of

$$\frac{dQ_n(t)}{dQ_n(s_n)},$$

for the product $\bar{\mathscr{F}}_n$. For the same reasons given previously, the $\bar{\mathscr{F}}_n$ are under the control of the quadratic forms defined by $\Gamma_n^*(\mu) = \operatorname{var} \int [\Sigma_j Z_{n,j}(t)] \mu(dt)$. However, since $Z_{n,j}(t)$ is the conditional expectation of $Y_{j,n}(t)$, the variances are smaller so that $\Gamma_n^*(\mu) \le \Gamma_n(\mu)$ for all $\mu \in \mathscr{M}_n$. This completes the proof of the theorem. □

REMARK 1. The forms Γ_n and Γ_n^* used here are nonrandom. We shall see in Section 8 experiments that are under the control of *random* quadratic forms. Under some supplementary restrictions (contiguity for pairs $\{P_n(s_n)\}, \{P_n(t_n)\}$, convexity of Θ_n if imbedded in the Hilbert space), one can show that control by nonrandom quadratic forms is equivalent to weak Gaussian approximability.

REMARK 2. The assumption that we have products with bounded infinitesimal arrays $\{\mathscr{E}_{j,n}\}$ and $\{\mathscr{F}_{j,n}\}$ and with $\mathscr{F}_{j,n}$ weaker than $\mathscr{E}_{j,n}$ has been used very forcibly in the preceding proofs. The results might be extendable to some other cases, as shown, for instance, in Section 6. However, they are not true without restriction. One can have an experiment $\mathscr{F}$ weaker than a Gaussian $\mathscr{E}$ with $\mathscr{F}$ remote from Gaussian. For instance, let $\mathscr{E}$ be the family $\mathscr{N}(\theta, 1)$ of normal distributions with variance unity on the line. If X is the observable variable let Y be the integer closest to X. It does give a family of distributions $\mathscr{F} = \{Q_\theta;\ \theta \in \Theta_n\}$ weaker than $\mathscr{E}$. However, the log-likelihood ratios for the Q_θ have discrete distributions that are far from Gaussian.

5. Preservation of the LAN conditions, part I. The contents of the present section bear on two separate questions. The first is that, for the bounded infinitesimal arrays of Section 4, the LAN conditions are preserved under loss of information if this loss occurs on the individual components. The second refers to the construction of estimates by adding to our auxiliary estimate a correction calculable from likelihood ratios. Here many different procedures may be used. We describe some that can often be easily carried out in practice. An explicit example is given in Section 9, Example 1.

The first recorded instance of the conditions called LAN that has come to our attention is that of [20]. Since then conditions called LAN have been used by many authors. (See [9], [10], [13], [14] and [34], for instance.) Unfortunately, the statements used by various authors differ in some aspects that are sometimes incidental and sometimes important. Because of this we shall first give two sets of assumptions, and show that they are equivalent if the index sets Θ_n are sufficiently rich. Then we shall show that they are preserved by passages from a product $\bar{\mathscr{E}}_n$ to the weaker $\bar{\mathscr{F}}_n$ of Section 2. We end by a description of a method of construction of estimates.

The LAN (for "locally asymptotically normal") assumptions used in the references listed previously differ considerably from the asymptotic Gaussian approximability conditions of Definitions 1 and 2 of Section 4. They involve a finite-dimensional vector space V and a sequence of norms $|\cdot|_n$ on V. The sequence of norms is used for two purposes: (a) to indicate the size of sets on which Gaussian approximability is contemplated; and (b) to relate the linear structure of V to that of the Gaussian experiments used as approximations. The relations will be described in detail in the following discussion.

For the typical LAN assumptions one considers a fixed finite-dimensional real vector space V and a sequence $\{|\cdot|_n\}$ of norms on V. The parameter set Θ_n that indexes the experiments $\bar{\mathscr{E}}_n = \{P_n(\theta);\ \theta \in \Theta_n\}$ is mapped into V by some function, say τ_n. The assumptions usually refer to some particular $\theta_{0,n} \in \Theta_n$ called the true value. The norms are often obtained from a single norm $|\cdot|$ on V by using either multiplication by some numerical factors or more complicated matrix manipulation. Think of the fairly common renormalizations $|\theta - \theta_{0,n}|_n = \sqrt{n}\,|\theta - \theta_{0,n}|$ or $|\theta - \theta_{0,n}|_n = |M_n(\theta - \theta_{0,n})|$ for matrices M_n. These M_n are often selected to "stabilize" the image of Θ_n in V making it tend to some limit.

These renormalization and stabilization operations may be convenient in practice, but they distract attention from the main statistical arguments. Thus we shall not consider them.

In fact, for simplicity, *we shall just assume that* Θ_n *is a subset of* V *and that* $\theta_{0,n}$ *is the origin of* V.

Consider then experiments $\bar{\mathscr{E}}_n$ subject to the following requirements.

(R.1) Each Θ_n is a subset of the fixed space V and the origin of V belongs to Θ_n.

(R.2) If $|\theta_n|_n$ remains bounded, the sequences $\{P_n(\theta_n)\}$ and $\{P_n(\theta_{0,n})\}$ are contiguous.

(R.3) For any set $S_n \subset \Theta_n$ such that $\sup\{|s|_n;\ s \in S_n,\ n = 1,2,\dots\} < \infty$, the experiments $\{P_n(\theta);\ \theta \in S_n\}$ admit weak Gaussian approximations according to Definition 2, Section 4.

To state the remaining conditions proceed as in Section 3 and introduce the set $\mathscr{M}_n$ of finite signed measures μ carried by finite subsets of Θ_n and satisfying $\mu(\Theta_n) = 0$.

For subsets $S_n \subset \Theta_n$, let $\mathscr{M}_n(S_n)$ be the subspace of $\mathscr{M}_n$ formed by measures μ whose support is in S_n [with $\mu(S_n) = 0$].

The existence of Gaussian approximations $\mathscr{G}_n = \{G_{\theta,n};\ \theta \in S_n\}$ implies the existence of corresponding quadratic forms Γ_{n,S_n} defined on $\mathscr{M}_n(S_n)$. For $\mu \in \mathscr{M}_n(S_n)$ and for

$$\Lambda_n^*(t, s_n) = \log \frac{dG_{t,n}}{dG_{s_n,n}},$$

the value $\Gamma_{n,S_n}(\mu)$ is the variance of $\int \Lambda_n^*(t, s_n)\mu(dt)$.

The link between the linear structure of V and that of the Gaussian approximations is as follows.

(R.4) Consider sets $S_n \subset \Theta_n$ such that $\sup\{|s|_n;\ s \in S_n,\ n = 1,2,\dots\} < \infty$ and such that the cardinality $\operatorname{card} S_n$ remains bounded. Then there are Gaussian approximations with the following property:

If $\mu_n \in \mathscr{M}_n(S_n)$ is such that $\sup_n \|\mu_n\| < \infty$ and such that $|\int s\mu_n(ds)|_n$ tends to 0, then $\Gamma_{n, S_n}(\mu_n) \to 0$.

It should be clear that the role of the norms in (R.3) is to indicate the size of sets on which Gaussian approximability is contemplated. The role of (R.4) is to link the vector structure of V to that of the Gaussian approximations.

This set (R.1)–(R.4) of assumptions is often described in a manner that looks very different as follows:

Let

$$\Lambda_n(t) = \log \frac{dP_n(t)}{dP_n(0)}.$$

(L.1) There are random vectors W_n (with values in V) and Euclidean norms $\|\cdot\|_n$ such that if $\sup_n |t_n|_n < \infty$, then

$$\Lambda_n(t_n) - t_n' W_n + \tfrac{1}{2}\|t_n\|_n^2$$

tends to 0 in $P_n(0)$ probability.

(L.2) Let F_n be the distribution under $P_n(0)$ of the vector W_n. There are joint Gaussian distributions G_n such that the dual Lipschitz norm $\|F_n - G_n\|_D$ induced by $|\cdot|_n$ tends to 0. [The dual Lipschitz norm is

$$\|F_n - G_n\|_D = \sup_f \left| \int f\, dF_n - \int f\, dG_n \right|,$$

for a supremum taken over functions f such that $|f| \le 1$ and

$$|f(x) - f(y)| \le |x - y|_n.]$$

The sets (R.1), (R.2), (R.3), (R.4) and (R.1), (R.2), (L.1), (L.2) are technically very different. To relate them, one encounters technical difficulties due to the fact that the sets Θ_n may be very sparse subsets of V. However, the two systems are known to be equivalent (see [27] for instance) under the following restriction.

(R.5) Let k be the dimension of V. There are sets $\{s_{0,n}, s_{1,n}, \dots, s_{k,n}\} \subset \Theta_n$ with the following properties:

(i) $\sup_{i,n} |s_{i,n}|_n < \infty$.

(ii) If $t_n \in \Theta_n$ is such that $\sup_n |t_n|_n < \infty$, then there are numbers $c_{n,i}$; $i = 1, 2, \dots, k$, such that

$$t_n - s_{0,n} = \sum_{i=1}^{k} c_{n,i}(s_{i,n} - s_{0,n})$$

and such that $\sum_{i=1}^{k} |c_{n,i}|$ remains bounded.

[If one requires more, for instance, that the simplex spanned by the $s_{i,n}$ be contained in Θ_n, then condition (L.2) is already a consequence of (R.1), (R.2), (L.1).]

As seen in Section 4, it is not true that replacing an experiment $\bar{\mathscr{E}}_n$ by a weaker one will preserve the Gaussian approximability. Thus the LAN conditions (R.1)–(R.4) are not preserved either. However, for product experiments one can assert the following.

THEOREM 3. *Let* $\{\mathscr{E}_{j,n}\}$ *be a bounded infinitesimal array of experiments and let* $\{\mathscr{F}_{j,n}\}$ *be another double array all indexed by sets* Θ_n. *Assume that* $\mathscr{F}_{j,n}$ *is less informative than* $\mathscr{E}_{j,n}$ *and consider the products* $\bar{\mathscr{E}}_n = \prod_j \mathscr{E}_{j,n}$ *and* $\bar{\mathscr{F}}_n = \prod_j \mathscr{F}_{j,n}$.

If $\bar{\mathscr{E}}_n$ *satisfies the conditions* (R.1), (R.2), (R.3), (R.4), *so does* $\bar{\mathscr{F}}_n$. *If, in addition, the requirement* (R.5) *is satisfied, one can replace the conditions* (R.1), (R.2), (R.3), (R.4) *by* (R.1), (R.2), (L.1), (L.2).

PROOF. The preservation of the conditions (R.1), (R.2), (R.3), (R.4) is an immediate consequence of Theorems 1 and 2, Section 4: Contiguity, Gaussian approximability are preserved and so are the linear relations satisfied by the Gaussian approximations. Under (R.1), (R.2), (R.5) the pairs (R.3), (R.4) and (L.1), (L.2) are always equivalent. Hence the result. □

To be complete, we should prove the stated equivalence of (R.3), (R.4) and (L.1), (L.2). However, for bounded infinitesimal arrays, one can use a number of other relations. Since these are important for the possible construction of asymptotically sufficient estimates, we shall now proceed to state some of them.

Suppose, for instance, that (R.1)–(R.5) are all satisfied. For $i = 1, 2, \dots, k$ let $X_{j,n,i}$ be the random variable

$$X_{j,n,i} = \left[\frac{dp_{j,n}(s_{i,n})}{dp_{j,n}(s_{0,n})} - 1\right] \wedge 1.$$

Let $Y_{n,i}$ be the sum $\sum_j X_{j,n,i}$. Suppose that $t_n \in \Theta_n$ is such that $\sup_n |t_n|_n < \infty$. Then, by (R.5), there are numbers such that

$$t_n - s_{0,n} = \sum_i c_{n,i}(s_{i,n} - s_{0,n})$$

and such that $\sup_n \sum_i |c_{n,i}| < \infty$. Let $D_n(t_n)$ be the difference

$$D_n(t_n) = \sum_j \left[\frac{dp_{j,n}(t_n)}{dp_{j,n}(s_{0,n})} - 1\right] - \sum_{i=1}^{k} c_{n,i} Y_{n,i}.$$

We claim that, under (R.1)–(R.5), the difference $D_n(t_n)$ will tend in probability to 0. Indeed, let μ_n be the measure that assigns mass $(-c_{n,i})$ to $s_{i,n}$ for $i = 1, 2, \dots, k$, mass 1 to t_n and mass $(\sum_{i=1}^k c_{n,i}) - 1$ to $s_{0,n}$. By definition of the $c_{n,i}$ one has $\int t \mu_n(dt) = 0$. Consider then the logarithms

$$\Lambda_n(t) = \log \frac{dP_n(t)}{dP_n(0)}$$

and the integrals $\int \Lambda_n(t) \mu_n(dt)$. According to (R.1)–(R.3), these integrals have a distribution that is approximately Gaussian with a certain variance $\Gamma_n(\mu_n)$. Since

$\int t\mu_n(dt) = 0$ this variance $\Gamma_n(\mu_n)$ must tend to 0, by (R.4). However, $D_n(t_n)$ is approximately equal to $\int\Lambda_n(t)\mu_n(dt)$ recentered by subtracting the expectation of the Gaussian approximation. Hence the assertion.

The same conclusion can be obtained if one uses (L.1), (L.2) instead of (R.3), (R.4).

To see how this may be used for the construction of estimates, consider the weaker $\mathscr{F}_{j,n} = \{q_{j,n}(\theta);\ \theta \in \Theta_n\}$ and the same sets $S_n = \{s_{0,n}, s_{1,n}, \dots, s_{k,n}\}$.

An estimation technique that is often successful is to maximize a smooth approximation to the logarithms

$$L_n(t) = \log\frac{dQ_n(t)}{dQ_n(s_{0,n})}.$$

Note that we said to maximize a *smooth approximation*, not the $L_n(t)$ themselves. Let

$$Z_{j,n,i} = \frac{dq_{j,n}(s_{i,n})}{dq_{j,n}(s_{0,n})} - 1$$

and let $Z^*_{j,n,i} = Z_{j,n,i} \wedge 1$.

Under the conditions of Theorem 3 for $t_n - s_{0,n} = \sum_{i=1}^k v_{n,i}(s_{i,n} - s_{0,n})$, a possible approximation to $L_n(t)$ is

$$L_n^{(1)}(t_n) = \sum_j\sum_i v_{n,i}Z^*_{j,n,i} - \tfrac{1}{2}\operatorname{var}\left[\sum_j\sum_i v_{n,i}Z^*_{j,n,i}\right].$$

This consists of a random term that is linear in the vector $v_n' = \{v_{n,i};\ i = 1, \dots, k\}$ and a nonrandom term that is quadratic in that vector. In many cases, the variances are not readily computable. Then one may prefer to use the approximation

$$L_n^{(2)}(t_n) = \sum_j\sum_i v_{n,i}Z_{j,n,i} - \tfrac{1}{2}\sum_j\left[\sum_i v_{n,i}Z_{j,n,i}\right]^2.$$

In matrix notation this may be written as follows. Let Z_n be the column vector whose coordinates are the $\sum_j Z_{j,n,i}$. Let M_n be the matrix whose (α, β) entry is $\sum_j Z_{j,n,\alpha}Z_{j,n,\beta}$. Then

$$L_n^{(2)}(t_n) = v_n'Z_n - \tfrac{1}{2}v_n'M_nv_n.$$

When M_n is invertible this can also be written in the form

$$L_n^{(2)}(t_n) = -\tfrac{1}{2}\left\{\left(v_n - M_n^{-1}Z_n\right)'M_n\left(v_n - M_n^{-1}Z_n\right) - Z_n'M_n^{-1}Z_n\right\},$$

from which it follows that the maximum of this approximation is reached at the point $\hat{v}_n = M_n^{-1}Z_n$.

To carry out the preceding operations, one needs to know the sets $\{s_{0,n}, s_{1,n}, \dots, s_{k,n}\}$ (and to be able to obtain good approximations to the likelihood ratios). Of course, if the purpose is to estimate θ, one will not know what set $\{s_{0,n}, \dots, s_{k,n}\}$ to use.

The technique proposed in [20] and variously expounded in [22] and [27] is to replace the set $\{s_{0,n}, \dots, s_{k,n}\}$ by a suitable estimate of it.

As explained in [20], the validity of the technique seemed to depend on curious cancellations involving the difference between $s_{0,n}$ and an estimate $\hat{s}_{0,n}$. However, the basic reason is very simple: The points that maximize $L_n(t_n)$ are the same as those that maximize $L_n(t_n) - L_n(\hat{s}_{0,n})$. This remains approximately true for the approximations of the type $L_n^{(1)}$ or $L_n^{(2)}$.

Note, however, that for the validity of the argument the following must hold.

(1) The auxiliary estimate $\hat{s}_{i,n}$ must with large probability be in the range where the approximation holds.

(2) Similarly, the correction term $\hat{v}_n = M_n^{-1}Z_n$ must be in the range where the approximation holds.

In the present situation this calls for the following comments. The existence of estimates that converge rapidly enough for $\bar{\mathscr{E}}_n$ does not imply existence of similar estimates for $\bar{\mathscr{F}}_n$. The fact that, in Theorem 2, the difference $\Gamma_n - \Gamma_n^*$ is positive semidefinite implies that the matrix M_n will typically be smaller than the corresponding matrix calculated on $\bar{\mathscr{E}}_n$. The fact that this latter would not degenerate is no guarantee that M_n will not. These properties must be checked for $\bar{\mathscr{F}}_n$ itself.

The matrix M_n can be adjusted to a certain extent by modification of the basis $\hat{s}_{i,n} - \hat{s}_{0,n}$; $i = 1, 2, \ldots, k$. It may be wise to check that a change of basis does not modify the end result too substantially.

In spite of all these warnings, we shall describe, in Section 7, a situation where the necessary verification can be carried out rather easily.

To terminate this section, note that the approximations $L_n^{(1)}$ and $L_n^{(2)}$ are not the only possible ones. There are many more. In [27] one of the present authors suggested estimating what takes the place of M_n by parallelogram differences calculated on the log-likelihood. For the situation covered by Theorem 3, one can use instead of the

$$\frac{dq_{j,n}(s_{i,n})}{dq_{j,n}(s_{0,n})}$$

a variety of functions of them, in particular, their logarithms or their square roots. For instance, if

$$W_{j,n,i} = \left\{ \frac{dq_{j,n}(s_{i,n})}{dq_{j,n}(s_{0,n})} \right\}^{1/2} - 1,$$

then

$$L_n^{(3)}(t_n) = \sum_j \sum_i c_{i,n} W_{j,n,i} - \tfrac{1}{2} \sum_j \left[\sum_i c_{i,n} W_{j,n,i} \right]^2$$

is an approximation to $\frac{1}{2}L(t_n)$.

The estimates $\{\hat{s}_{0,n}, \ldots, \hat{s}_{k,n}\}$ used in [20], for instance, were assumed to satisfy certain discreteness conditions. This is for technical reasons: They should not seek for singularities in the likelihood functions. That can be achieved by methods other than discretization. If the likelihood functions are sufficiently regular, one does not need to worry about this particular point. A referee pointed out that the estimates described here may display a severe lack of robustness. This is particularly true if the approximation called $L_n^{(2)}$ is used with untruncated variables. Fortunately, one can often detect easily forms of misbehavior of the variables $Z_{j,n,i}$. Possible methods for doing so and for correcting the system are too complex to be described here. We shall return to this matter elsewhere.

Finally, instead of using differences one can sometimes use derivatives and the Newton–Raphson procedure. However, to show that this particular technique will work, one needs assumptions that are much more restrictive than the ones used here.

The reader will note that in Sections 4 and 5 we replaced *each component* $\mathscr{E}_{j,n}$ by a weaker $\mathscr{F}_{j,n}$. This is essential in the proofs. The next section expounds on a result of Davies in which the loss of information occurs in a very different manner.

6. Preservation of the LAN conditions, part II. Davies ([8] and [9]) has observed that the LAN conditions are preserved under circumstances that are rather different from the ones described in Section 5: No product structure is assumed. Even if it is there, the loss of information occurs by passage to sub-σ-fields that do not preserve the product structure. To describe the situation, we shall consider a sequence $\{\mathscr{E}_n\}$ of experiments $\{P_n(\theta);\ \theta \in \Theta\}$, where the $P_n(\theta)$ are measures on some σ-field $\mathscr{A}_n$. The weaker experiment $\mathscr{F}_n = \{Q_n(\theta);\ \theta \in \Theta_n\}$ is obtained by taking the restrictions $Q_n(\theta)$ of the $P_n(\theta)$ to some subfield $\mathscr{B}_n$.

We shall show that the LAN conditions are preserved if the weaker experiments admit "distinguished" statistics whose joint distributions with the log-likelihood of the initial experiment are asymptotically normal. The relations with Davies' work are complex. They are described at the end of the section.

We shall assume that the experiments $\mathscr{E}_n$ satisfy the requirements (R.1), (R.2) and (L.1), (L.2) of Section 5. Recall that (L.1) involves certain random vectors W_n and Euclidean norms $\|\cdot\|_n$.

Consider also statistics T_n defined on the experiment $\mathscr{F}_n$ and taking values in a fixed Euclidean space R^q. Recall that the sequence $\{T_n\}$ is called *distinguished* for the sequence $\{\mathscr{F}_n\}$ if the following property holds. Let $F_n(\theta)$ be the distribution of T_n for $Q_n(\theta)$. Take pairs (s_n, t_n) of elements of Θ_n. This gives a pair $\{F_n(s_n), F_n(t_n)\}$ of measures on R^q and a binary experiment $\mathscr{F}_n' = \{Q_n(s_n), Q_n(t_n)\}$.

Compactifying R^q by adjunction of points at ∞, one can extract subsequences $\{\nu\} \subset \{n\}$ such that (i) the experiments $\mathscr{F}_\nu'$ have a limit $\mathscr{F}'$, and (ii) the measures $F_\nu(s_\nu)$ and $F_\nu(t_\nu)$ tend in the usual weak sense to certain limits, say F_0 and F_1.

The sequence $\{T_n\}$ is distinguished if for all such subsequences the experiment formed by the limiting distributions $\{F_0, F_1\}$ is as strong as the limit experiment $\mathscr{F}'$. (The definition of [24] involves more than pairs, but it is shown in [27] that it is enough to look at pairs.) For the purposes of our next theorem, we shall metrize $V \times R^q$ by the square norms $\|\cdot\|_n^2 + |\cdot|^2$.

THEOREM 4. *Consider experiments* $\mathscr{E}_n = \{P_n(\theta);\ \theta \in \Theta\}$ *given by measures on a* σ*-field* $\mathscr{A}_n$ *and their restrictions* $\mathscr{F}_n = \{Q_n(\theta);\ \theta \in \Theta\}$ *to subfields* $\mathscr{B}_n \subset \mathscr{A}_n$. *Assume the following*:

(i) *The* $\mathscr{E}_n$ *satisfy the LAN conditions* (R.1), (R.2), (L.1) *and* (L.2).

(ii) *The joint distributions* $\mathscr{L}[W_n, T_n | P_n(\theta_{0,n})]$ *admit Gaussian approximations in the sense of condition* (L.2) *and, in this approximation, the variable* T_n *is centered at* 0.

(iii) *For the sequence* $\{\mathscr{F}_n\}$ *the sequence* $\{T_n\}$ *is distinguished.*

Then the $\mathscr{F}_n$ *satisfy conditions* (R.1), (R.2), (L.1) *and* (L.2). *Furthermore, there are nonrandom matrices* A_n *and* B_n *such that*

$$\log \frac{dQ_n(t_n)}{dQ_n(\theta_{0,n})}$$

is approximable by $t_n' A_n T_n - \frac{1}{2} t_n' B_n t_n$ *(for the inner products of the initial norms* $\|\cdot\|_n$).

PROOF. Take a Gaussian approximation to the joint distribution $\mathscr{L}[W_n, T_n | P_n(\theta_{0,n})]$. Let C_n be the covariance matrix of T_n and W_n in that approximation. (Take C_n in the form $C_n = ET_n W_n'$.) If $G_n(\theta_{0,n}) = \mathscr{N}(0, K_n)$ is a centered approximation to $\mathscr{L}[T_n | Q_n(\theta_{0,n})]$, then, by [20], $G_n(t_n) = \mathscr{N}[C_n t_n, K_n]$ is an approximation to $\mathscr{L}[T_n | Q_n(t_n)]$.

The assumption that T_n is distinguished insures that $\mathscr{F}_n$ satisfies (R.1), (R.2) and (R.3) for the approximating Gaussian experiments $\mathscr{G}_n = \{G_n(\theta);\ \theta \in \Theta_n\}$. The contiguity condition of (R.2) shows that, for n large, the $G_n(\theta)$, $\theta \in \Theta_n$, are mutually absolutely continuous. Thus they are all carried by the smallest linear subspace L_n of R^q that carries $G_n(\theta_{0,n})$. Also, restricted to that subspace, K_n must be nonsingular and the values $C_n t_n$, $t_n \in \Theta_n$ must all be in L_n. It may happen, of course, that L_n is reduced to the origin of R^q. In that case the final statement of the theorem is true with matrices $A_n = B_n = 0$. If L_n is not reduced to 0, then the log-likelihood

$$\log \frac{dG_n(t_n)}{dG_n(\theta_{0,n})}$$

has on L_n the form

$$g_n(T_n, t_n) = -\tfrac{1}{2}\{[T_n - C_n t_n]' K_n^{-1} [T_n - C_n t_n] - T_n' K_n^{-1} T_n\}$$

and the $t_n' C_n' K_n^{-1} C_n t_n$ must remain bounded. Here again, if for all choices of the $t_n \in \Theta_n$ the $t_n' C_n' K_n^{-1} C_n t_n$ tend to 0, then the final assertion of the theorem holds with $A_n = B_n = 0$. Thus we shall assume that some of these sequences do not tend to 0.

Now return to R^q. It was assumed to carry a Euclidean norm, say $|\cdot|$, used, in particular, to describe the approximability of distributions $\mathscr{L}[T_n | Q_n(\theta_{0,n})]$ by Gaussian ones. Let Π_n be the orthogonal projection of R^q onto L_n for that norm. Note that the difference between $\mathscr{L}[T_n | Q_n(\theta_{0,n})]$ and $\mathscr{L}[\Pi_n T_n | Q_n(\theta_{0,n})]$

must tend to 0. However, a theorem of [25] shows that in such a case the difference between $g_n(\Pi_n T_n, t_n)$ and

$$\log \frac{dQ_n(t_n)}{dQ_n(\theta_{0,n})}$$

must also tend to 0. Hence the result. □

The result presented here is related to a result of Davies ([8] and [9]). His assumptions differ considerably from ours. Davies is interested in the situation where the restricted experiments $\mathscr{F}_n$ do not necessarily satisfy the LAN condition but only what we have called "quadratic control" by *random* quadratic forms. This corresponds to the approximability of the log-likelihood as in condition (L.1) of Section 5 but with random norms $\| \quad \|_n$. He assumes that the *conditional* distributions $\mathscr{L}\{W_n|\mathscr{B}_n, P_n(\theta_{0,n})\}$ are approximately Gaussian. To prove this in particular cases, he uses local limit theorems and an argument previously used by Steck [37].

We have limited ourselves here to the case of control by *nonrandom* quadratic forms. For that particular case our assumptions are weaker than those made by Davies in that joint asymptotic normality of W_n and T_n does not necessarily imply asymptotic normality of the conditional distributions $\mathscr{L}\{W_n|T_n, P_n(\theta_{0,n})\}$.

For instance, it is highly visible that the vectors Z_n used in Theorem 3 and further on to describe the construction of estimates are distinguished. Thus one could prove the required joint asymptotic normality as in Section 4 and then apply our present Theorem 4 to obtain an alternate proof of Theorem 3, Section 5. This does not require examination of conditional distributions.

However, it can also be argued that, under the conditions of our Theorem 4, Davies' conditions are almost satisfied. Indeed, it can be shown (see [27], Chapter 7, Section 3) that the assumption that the sequence $\{T_n\}$ is distinguished is equivalent to an assumption to the effect that the likelihood ratios

$$\frac{dQ_n(t_n)}{dQ_n(\theta_{0,n})}$$

can be approximated by smooth functions of T_n. To state that more specifically, take a fixed k independent of n and sets $\{t_{n,i};\ i = 1, 2, \dots k\}$ in Θ_n. Let Z_n be the k-dimensional vector formed by the densities of the $Q_n(t_{n,i})$ with respect to their sum μ_n. It is shown in [27], Chapter 7, Section 3, that $\{T_n\}$ is a distinguished sequence if and only if for each set $\{t_{n,i};\ i = 1, 2, \dots k\}$ there is a fixed finite set $\{\gamma_\nu;\ \nu = 1, 2, \dots r\}$ of continuous functions from R^q to the unit simplex of R^k such that

$$\inf_\nu \int |Z_n - \gamma_\nu(T_n)|\, d\mu_n$$

tends to 0. (Here the norm is the maximum coordinate norm.) The proof of that statement is rather involved. Under the contiguity restrictions used here, one can replace the sum μ_n by $Q_n(\theta_{0,n})$. Suppose then, in addition, that the sets Θ_n are sufficiently rich in that, for instance, they contain a fixed open subset of V.

It is then possible, using Laplace transforms, to conclude that the conditional distributions of W_n given T_n are approximable in $Q_n(\theta_{0,n})$ probability by equicontinuous functions of the T_n. If the joint distributions converge to normal limits, Steck's theorems will then imply that these equicontinuous approximations will also converge to normal limits. This is all that would be required for the application of Davies' theorem.

Thus, modulo a rather lengthy and complex argument and for sufficiently rich sets Θ_n, one can say that our assumptions imply those made by Davies. It should be noted, however, that to derive Theorems 2 and 3 from Theorem 4, one still needs to carry out a good part of the arguments previously described in Sections 4 and 5. As already pointed out, to derive Theorem 4 from Davies' results, one needs a considerable amount of work. This is also true for Theorems 2 and 3.

A further remark is that the proof of Theorem 4 does not actually depend on the fact that the spaces V and R^q have a fixed finite dimensionality. One can carry out a similar proof in infinite-dimensional spaces. This is useful in nonparametric situations and will be published elsewhere.

7. Differentiability in quadratic mean and the i.i.d. case. Let Θ be a subset of a Euclidean space and let $\{p(\theta): \theta \in \Theta\}$ be a family of probability measures on a σ-field $\mathscr{A}$. A condition often used to study the i.i.d. case is a condition of differentiability in quadratic mean expressible as follows. At a point θ_0 let $\xi^2(s)$, $\xi(s) \geq 0$, be the density $dp(s)/dp(\theta_0)$ of the part of $p(s)$ that is dominated by $p(\theta_0)$.

(DQM). (i) *There are random vectors X such that as $|t| \to 0$,*

$$E_{\theta_0}\frac{1}{|t|^2}\left|\xi(\theta_0+t)-\xi(\theta_0)-t'X\right|^2 \to 0.$$

(ii) *Let $\beta(\theta_0+t)$ be the mass of $p(\theta_0+t)$ that is $p(\theta_0)$ singular. Then*

$$\frac{1}{|t|^2}\beta(\theta_0+t) \to 0, \quad \textit{as } |t| \to 0.$$

Now replace the σ-field $\mathscr{A}$ by a smaller σ-field $\mathscr{B}$ and let $q(\theta)$ be the restriction of $p(\theta)$ to $\mathscr{B}$. We shall show that, if $\{p(\theta);\ \theta \in \Theta\}$ satisfies (DMQ) at θ_0, so does $\{q(\theta);\ \theta \in \Theta\}$. For this we shall use an indirect proof, using the asymptotic properties of the likelihood ratios when the number of observations increases indefinitely. A direct proof can be carried out, but it is less informative than the indirect proof for the standard i.i.d. case that can be described as follows.

Let $\mathscr{E}$ be the system of measures $\{p(\theta);\ \theta \in \Theta\}$ on a σ-field $\mathscr{A}$ carried by a set $\mathscr{X}$. For each integer j, let $\{\mathscr{X}_j, \mathscr{A}_j, \mathscr{E}_j\}$ be a copy of $\{\mathscr{X}, \mathscr{A}, \mathscr{E}\}$. The experiment carried out at stage n is the direct product $\tilde{\mathscr{E}}_n = \prod_{j=1}^n \mathscr{E}_j$ of n copies of $\mathscr{E}$. That is, $\tilde{\mathscr{E}}_n$ is the experiment that consists of observing n independent variables $\omega_1, \omega_2, \ldots, \omega_n$ with common distribution equal to a certain $p(\theta_0)$,

$\theta_0 \in \Theta$. As n tends to ∞, we assume that the "true" θ_0 stays fixed independent of n. (For certain purposes, for instance, to prove uniformity of convergences, one makes the "true" value of the parameter depend on n. If so, the situation one encounters is more restricted than the general situation covered by the infinitesimal arrays of Section 4. However, the restrictions are not as drastic as one could presume.)

For the experiments $\mathscr{E}_{j,n} = \{p_{j,n}(\theta);\ \theta \in \Theta_n\}$ of Sections 2–5, we shall take the copy $\mathscr{E}_{j,n} = \{p(\theta);\ \theta \in \Theta_n\}$ with θ restricted to lie in a subset Θ_n of Θ. Finally, we shall consider only the case where the Θ_n are of the form $\Theta_n(b) = \{\theta \in \Theta;\ \sqrt{n}\,|\theta - \theta_0| \le b\}$ or perhaps of the form $\Theta_n'(b) = \{\theta \in \Theta;\ a_n|\theta - \theta_0| \le b\}$ for some other coefficients a_n that tend to ∞.

It is well known that, in such situations, the condition (DQM) implies that the products $\tilde{\mathscr{E}}_n$ satisfy the LAN conditions on sets of the form $\Theta_n(b) = \{\theta \in \Theta;\ \sqrt{n}\,|\theta - \theta_0| \le b\}$. It is also known (see, for instance, Example 4, Section 9) that cases occur where (DQM) is not satisfied but where the LAN conditions are satisfied, for instance, on sets $\{\theta \in \Theta;\ \sqrt{n \log n}\,|\theta - \theta_0| \le b\}$. Even if one sticks with the sets $\Theta_n(b)$ and the speed of convergence $\sqrt{n}$, the validity of the LAN conditions does not imply that of (DQM). (An example is given in [23], page 816.) Thus the preservation of the LAN conditions by passage from the σ-field $\mathscr{A}$ to a smaller σ-field $\mathscr{B}$ does not logically imply that (DQM) will also be preserved. We shall now show that it is preserved.

It is known (see [27], Chapter 17, Section 3) that the (DQM) requirement is equivalent to the following. For a pair (s, t) let $\Lambda_n(t, s)$ be the logarithm of likelihood ratio

$$\log \frac{dP_n(t)}{dP_n(s)},$$

for the product measures $P_n(s) = \prod_j p_{j,s}$, where $p_{j,s}$ is the jth copy of p_s.

(DQM$'$). *There are random vectors X_j defined on $(\mathscr{X}_j, \mathscr{A}_j)$ with the following properties*:

(i) *The X_j are copies of a given X_1.*
(ii) *$E_{\theta_0}X_j = 0$ and for $t \in V$, $E_{\theta_0}(t'X_j)^2 < \infty$.*
(iii) *If $|t_n|$ remains bounded, then*

$$\Lambda_n\left(\theta_0 + \frac{t_n}{\sqrt{n}}, \theta_0\right) - t_n'\left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} X_j\right) + \tfrac{1}{2}E_{\theta_0}(t_n'X_1)^2$$

tends to 0 *in $P_n(\theta_0)$ probability.*

In other words, for any fixed b let

$$\Theta_n(b) = \left\{\theta \in \Theta;\ \sqrt{n}\,|\theta - \theta_0| \le b\right\}$$

and let $\tilde{\mathscr{E}}_n$ be the product experiment indexed by $\Theta_n(b)$. The (DQM$'$) condition is equivalent to the requirement that the $\tilde{\mathscr{E}}_n$ satisfy the LAN conditions (R.1), (R.2), (L.1) and (L.2) of Section 5 with the further specification that the linear term $t_n'W_n$ of condition (L.1) have the special form $t_n'((1/\sqrt{n})\Sigma_{j=1}^n X_j)$.

Instead of using the logarithms $\Lambda_n(\theta_0 + t_n/\sqrt{n}, \theta_0)$, one can use the sum

$$S_n(t_n) = \sum_{j=1}^{n} \left[\frac{dp_{j,\theta_n}}{dp_{j,\theta_0}} - 1 \right],$$

where $\theta_n = \theta_0 + t_n/\sqrt{n}$. The condition is then that

$$S_n(t_n) - t_n'\left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} X_j \right)$$

tends to 0 in probability.

Now let us consider experiments $\mathscr{F}_{j,n}$ that are weaker than our $\mathscr{E}_{j,n}$. Passage to weaker experiments need not preserve the i.i.d. structure. For instance, it could be obtained by passing from the σ-field $\mathscr{A}_j$ to a σ-field $\mathscr{B}_{j,n} \subset \mathscr{A}_j$ but in such a way that the $\mathscr{B}_{j,n}$ are not copies of each other. An example can be readily constructed by looking at the counting processes described in the Introduction. They might be i.i.d. on the total observation period $[0, L]$ but on the weaker experiment the isolated times at which the patients are observed may depend on the patient.

We shall consider only the situation where there is a σ-field $\mathscr{B} \subset \mathscr{A}$ with copies $\mathscr{B}_j \subset \mathscr{A}_j$ and where the jth experiment $\mathscr{F}_j$ is obtained by restricting the p_θ of $\mathscr{E}$ to $\mathscr{B}$ and copying the result on $\mathscr{B}_j$.

One immediate result is

Proposition 4. *If the p_θ; $\theta \in \Theta$, satisfy (DQM) at θ_0 so do their restrictions q_θ to the σ-field $\mathscr{B} \subset \mathscr{A}$.*

Proof. Use the variant of (DQM′) that says that $S_n(t_n) - t_n'((1/\sqrt{n})\Sigma_{j=1}^n X_j)$ tends to 0. According to the argument of Theorem 2, the corresponding sums $S_n^*(t_n)$ for the weaker experiment will be such that

$$S_n^*(t_n) - t_n'\left((1/\sqrt{n}) \sum_{j=1}^{n} X_j^* \right) \to 0 \quad \text{for } X_j^* = E_{\theta_0}(X_j | \mathscr{B}_j). \qquad \square$$

(The preservation of (DQM) is no surprise in view of its relation to Lipschitz conditions or to the rate of separation $\sqrt{n}$. See [23].)

In the present case, since the rate of convergence $\sqrt{n}$ for estimates is about as bad as can be, one can also hope that the weaker experiments will still give suitable estimates. This need not be so. However, here is a usable result.

Consider the expectation $E_{\theta_0}(t'X_1)^2$ used in (DQM′). It may be written $t'\Gamma(\theta_0)t$ for a certain matrix $\Gamma(\theta_0)$ called the Fisher information matrix.

Proposition 5. *Assume that the p_θ satisfy the following requirements*:

(i) *If $\theta_k \to \theta$, then the total variation $\|p_{\theta_k} - p_\theta\|$ tends to* 0.
(ii) *If $t \neq s$, then $\|p_t - p_s\| > 0$.*

Then there exist estimates T_n based on n observations such that $\sqrt{n}(T_n - \theta)$ stays bounded in $P_n(\theta)$ probability for any $\theta \in \Theta$ that satisfies (DQM) with a nonsingular information matrix $\Gamma(\theta)$.

PROOF. This is a known result. See for instance [22]. □

Now consider the weaker experiments with measures restricted to the copies $\mathscr{B}_j$ of a σ-field $\mathscr{B} \subset \mathscr{A}$. Let q_θ be the restriction of p_θ to $\mathscr{B}$.

It is obvious that whenever condition (i) of Proposition 5 is satisfied by the p_θ, it is also satisfied by the q_θ. Condition (ii) need not be preserved. That $p_t \neq p_s$ does not necessarily imply $q_t \neq q_s$. However, that $s \neq t$ implies $q_t \neq q_s$ is often easy to check.

Thus we are essentially left with the question of nonsingularity of Fisher information matrices. A technique that often works is the following. If the q_θ are complicated, there may be σ-fields $\mathscr{C} \subset \mathscr{B}$ on which their restrictions r_θ are quite manageable. It is sufficient to verify the nonsingularity of information matrices for the r_θ.

8. An extension to a dependent situation. The proofs of Sections 4 and 5 depend very strongly on the independence assumptions made there. However, they suggest a possible extension to a dependent situation.

It often happens that an experiment has the following structure: It is given by measures $P_n(\theta)$ on a σ-field $\mathscr{A}_n$ but there is a filtration $\mathscr{A}_{1,n} \subset \mathscr{A}_{2,n} \subset \cdots \subset \mathscr{A}_{j,n} \subset \mathscr{A}_{j+1,n} \subset \cdots \subset \mathscr{A}_n$ formed by a finite or infinite increasing family of sub-σ-fields. Since our purpose is not to attain maximum generality, we shall assume that the filtration is a finite sequence with $\mathscr{A}_n$ being the last one of the sequence.

Except for technical details that are annoying but not so important, one can describe such a situation in a more intuitive manner as follows.

One starts by performing the experiment with the P_θ restricted to $\mathscr{A}_{1,n}$. If everything has been carried out up to and including the jth step, one performs an additional experiment $\mathscr{E}_{j+1,n}$, *where the probability measures depend on the observations seen up to and including the jth step*. One continues until the end of the sequence has been reached.

To retain a notation similar to that used previously, we shall let $p_{j,n}(\theta)$ be the conditional distribution used at the jth stage when the value of the parameter is θ. It is a function of the previous observations, measurable with respect to $\mathscr{A}_{j-1,n}$. Just as in Section 2, one can define quantities $h_{j,n}$ by

$$h^2_{j,n}(s,t) = \tfrac{1}{2}\int\left[\sqrt{dp_{j,n}(s)} - \sqrt{dp_{j,n}(t)}\right]^2,$$

for integrals taken conditionally given $\mathscr{A}_{j-1,n}$. Now the $h_{j,n}$ are $\mathscr{A}_{j-1,n}$ measurable functions. The joint distributions will be denoted $P_n(\theta)$ as before.

By analogy with the requirements imposed in Section 4, we shall assume that the following conditions are satisfied for all possible choices of pairs (s_n, t_n) in Θ_n.

(A) As $n \to \infty$ the quantities $\sup_j h_{j,n}(s_n, t_n)$ tend to 0 in $P_n(s_n) + P_n(t_n)$ probability.

(B) There is a number $b < \infty$, independent of (s_n, t_n), such that

$$\sum_j h_{j,n}^2(s_n, t_n) \leq b.$$

(C) For every fixed $\varepsilon > 0$ the sums

$$\sum_j \| p_{j,n}(t_n) - \left[(1+\varepsilon) p_{j,n}(s_n)\right] \wedge p_{j,n}(t_n) \|$$

tend to 0 in probability.

Note that here conditions (A)–(C) are no longer sufficient to imply asymptotic normality of the logarithms

$$\Lambda_n(t_n, s_n) = \sum_j \log \frac{dp_{j,n}(t_n)}{dp_{j,n}(s_n)}.$$

As far as we have been able to determine, necessary and sufficient conditions for the asymptotic normality of $\Lambda_n(t_n, s_n)$ are not known, even under the restrictive conditions (A) and (B). In particular, it is not known whether (C) is necessary. We have written it by analogy with the conditions that are necessary and sufficient in the independent case. The literature does contain necessary and sufficient conditions for the asymptotic normality of the likelihood process considered as a function of θ and of a time parameter corresponding to the index j of $\mathscr{A}_{j,n}$. These are the so-called "invariance principles," for which see Aalen [1], Rebolledo ([32], [33]) or Shiryayev [35]. However, that is a different matter. Here we are interested only in the behavior of the likelihood ratios on the final σ-field $\mathscr{A}_n$.

Under conditions (A)–(C), the Taylor expansion argument that gives an approximation for the log-likelihood $\Lambda_n(t_n, s_n)$ is still valid. Therefore, it is still true that the difference

$$\Lambda_n(t_n, s_n) - \sum_j \left[\frac{dp_{j,n}(t_n)}{dp_{j,n}(s_n)} - 1\right] + \tfrac{1}{2}\sum_j \left[\frac{dp_{j,n}(t_n)}{dp_{j,n}(s_n)} - 1\right]^2$$

tends in probability to 0. This suggests the introduction of quadratic forms Γ_n as follows. Let $\mathscr{M}_n$ be the space of finite signed measures with finite support on Θ_n and total algebraic mass 0 as before. Select some $s_n \in \Theta_n$ and let

$$\Gamma_n(\mu) = \sum_j \left| \int \left[\frac{dp_{j,n}(t)}{dp_{j,n}(s_n)} - 1\right] \mu(dt) \right|^2.$$

Consider also the forms

$$K_n(\mu) = -4 \sum_j \iint h_{j,n}^2(s, t) \mu(ds) \mu(dt).$$

Theorem 5. *Let the $\mathscr{E}_{j,n}$ satisfy conditions* (A)–(C). *Then the experiment $\bar{\mathscr{E}}_n$ defined by the $\mathscr{E}_{j,n}$ is controlled by the quadratic forms K_n. Also if the* *cardinality of the supports of measures* μ_n *remains bounded and if* $\sup_n \|\mu_n\| < \infty$, *then* $K_n(\mu_n) - \Gamma_n(\mu_n)$ *tends to* 0 *in probability.*

PROOF. According to condition (C), there is no real loss of generality in assuming that the conditional distributions $p_{j,n}(\theta)$, $\theta \in \Theta$, are mutually absolutely continuous. A standard truncation argument shows also that one can proceed as if the martingale differences

$$X_{j,n}(t_n) = \frac{dp_{j,n}(t_n)}{dp_{j,n}(s_n)} - 1$$

were all bounded by unity. The integrals $\int X_{j,n}(t)\mu_n(dt)$ are also martingale differences under $P_n(s_n)$. Their sum will tend to 0 in probability if their quadratic variation $\Gamma_n(\mu_n)$ tends to 0. This quadratic variation can also be replaced by the sum

$$\sum_j E_{j-1}\left[\int X_{j,n}(t)\mu_n(dt)\right]^2,$$

where E_{j-1} denotes an expectation taken conditionally given the past at $j-1$. (For these relations, see Neveu [29].)

A Taylor expansion argument shows that if $Y_{j,n}(t) = \sqrt{1 + X_{j,n}(t)} - 1$, then $X_{j,n}(t)$ can be replaced by $2Y_{j,n}(t)$ in the preceding formula. Thus we are led to consider the expressions

$$\begin{aligned} 4E_{j-1}\left[\int Y_{j,n}(t)\mu_n(dt)\right]^2 &= 4E_{j-1}\iint Y_{j,n}(s)Y_{j,n}(t)\mu_n(ds)\mu_n(dt) \\ &= 4\iint\left[1 - h_{j,n}^2(s,t)\right]\mu_n(ds)\mu_n(dt) \\ &= K_n(\mu_n), \end{aligned}$$

yielding the desired result. □

The fact that the experiments $\bar{\mathscr{E}}_n$ are quadratically controlled can often be used to construct estimates that possess asymptotic minimaxity or asymptotic sufficiency properties. A theory to that effect has been expounded in [27], Chapter 11. The technique of [27] requires for its validity a number of additional restrictions such as existence of well-behaved auxiliary estimates and dimensionality restrictions on the parameter spaces Θ_n. For such conditions we can only refer the reader to [27] and to the work of Jeganathan ([16] and [17]).

Now let us pass to the situation where there is another filtration $\mathscr{B}_{j,n}$, $j = 1, 2, \ldots$, with $\mathscr{B}_{j,n} \subset \mathscr{B}_{j+1,n} \subset \cdots \subset \mathscr{B}_n$. If $\mathscr{B}_{j,n} \subset \mathscr{A}_{j,n}$ for every j, it may happen that the conditional experiment $\mathscr{F}_{j,n}$ carried out at the jth stage is always weaker than the corresponding $\mathscr{E}_{j,n}$. This is far from automatic. It is also plain that conditions such as (A) or (C) have no reason to be inherited by the conditional distributions $q_{j,n}(\theta)$ used on the weaker filtration. Note that the conditional measures $q_{j,n}(\theta)$ are $\mathscr{B}_{j-1,n}$-measurable, whereas the $p_{j,n}(\theta)$ are only $\mathscr{A}_{j-1,n}$-measurable. We shall restrict our attention to cases *where the experi-*

ments $\{q_{j,n}(\theta);\ \theta \in \Theta_n\}$ *are always weaker than the* $\{p_{j,n}(\theta);\ \theta \in \Theta_n\}$ *no matter what may have happened in the passage from* $\mathscr{A}_{j-1,n}$ *to* $\mathscr{B}_{j-1,n}$. This may seem to be a very strong restriction, and it really is. However, there do exist interesting situations of that type.

For an example, consider a finite state Markov process $\{X(t)\colon\ t \in [0, L]\}$ observed during the entire interval $[0, L]$. One can obtain the type of structure described previously by dividing the interval $[0, L]$ by times t_j such that $0 = t_0 < t_1 < \cdots < t_j < t_{j+1} < \cdots < t_k = L$. The conditional experiment $\mathscr{E}_{j,n}$ is then obtained by observing the process during the entire interval $(t_{j-1}, t_j]$. One obtains a weaker $\mathscr{F}_{j,n}$ by observing $X(t)$ only at the end point t_j.

Another example is that of "aggregated Markov chains," for which see [19] and the references therein. It occurs as follows. Let $\{X_s(t_j);\ j = 1, 2, \dots k;\ s = 1, 2, \dots m\}$ be m independent identically distributed finite state Markov chains.

At time t_j the fully informed observer sees the actual states for all the individual processes $X_s(t_j)$. The restricted observer looks at each possible state i and is given the number $N(t_j, i)$ of processes that are in state i. Here, in spite of the fact that $\mathscr{A}_{j-1,n}$ can be considerably larger than $\mathscr{B}_{j-1,n}$, the fully informed observer can carry out his conditional experiment $\mathscr{E}_{j,n}$ using only the information in $\mathscr{B}_{j-1,n}$, but he keeps track of who goes where, whereas the restricted observer is given only summary information.

In view of such examples the following result may be useful.

THEOREM 6. *Let the* $\mathscr{E}_{j,n}$ *satisfy conditions* (A)–(C). *Assume that no matter what was observed in the* $\mathscr{E}_{r,n}$, $r \leq j-1$, *the conditional experiment* $\mathscr{F}_{j,n}$ *is always weaker than* $\mathscr{E}_{j,n}$. *Then the* $\mathscr{F}_{j,n}$ *also satisfy conditions* (A)–(C). *The global experiment* $\bar{\mathscr{F}}_n$ *defined by the* $\mathscr{F}_{j,n}$ *is under the control of quadratic forms* K_n^* *such that the differences* $K_n - K_n^*$ *are positive semidefinite.*

PROOF. This can be proved exactly as for the independent case, using Lemma 2 of Section 3 on the conditional distributions. □

A corollary is that if the $\bar{\mathscr{E}}_n$ satisfy the LAMN conditions, then, under the restrictions of Theorem 6, the log-likelihoods for $\bar{\mathscr{F}}_n$ will possess the same kind of linear–quadratic expansions. All the arguments of Sections 4 and 5 remain applicable, *except* that we have said nothing at all about asymptotic normality. The construction of estimates described in Section 5 remains feasible. The resulting estimates will still be asymptotically Bayes, asymptotically minimax, etc., as shown by Jeganathan [16]. Together with estimates of the quadratic forms, they will be asymptotically sufficient.

As to the matter of asymptotic normality, standard martingale limit theorems (see [33] or [35]) show that the sums $\sum_j \int X_{j,n}(t)\mu_n(dt)$ used in Theorem 5 will be asymptotically normally distributed if the quadratics $\Gamma_n(\mu_n)$ or $K_n(\mu_n)$ differ from nonrandom quantities by amounts that tend to 0. It would be convenient if that property was inherited by the weaker $\mathscr{F}_{j,n}$. However, it need not be. To give an example, consider the case where one has n independent identically distributed variables, $\xi_1, \xi_2, \dots, \xi_n$ and where $\mathscr{E}_{j,n}$ consists of observing ξ_j. Let N_n be a stopping time of the sequence $\xi_1, \xi_2, \dots, \xi_n$ with $N_n \le n$. Let $\mathscr{F}_{j,n} = \mathscr{E}_{j,n}$ if $j \le N_n$. If $j > N_n$ let $\mathscr{F}_{j,n}$ be the trivial experiment where nothing is observed.

Suppose that the $\mathscr{E}_{j,n}$ satisfy the differentiability in quadratic mean condition of Section 7 with nondegenerate derivatives. Then, if the N_n/n have distributions that do not degenerate as $n \to \infty$, the $\bar{\mathscr{F}}_n$ will not be asymptotically normal.

We have already observed that conditions such as (A) and (C) have no reason to be inherited in the passage from σ-fields $\mathscr{A}_{j,n}$ to smaller $\mathscr{B}_{j,n}$. Condition (B) is of a different nature. Something like it is inheritable. Indeed, conditions (A)–(C) imply the contiguity of pairs $\{P_n(s_n)\}, \{P_n(t_n)\}$. It has been shown by Greenwood and Shiryayev ([12], Theorem 4, page 48) that contiguity is equivalent to the conjunction of two conditions, one of which is that the sum $\sum_j h_{j,n}^2(s_n, t_n)$ remains bounded in $P_n(s_n) + P_n(t_n)$ probability. Since contiguity is inherited by the weaker experiments, if (A)–(C) holds for the $\mathscr{E}_{j,n}$, then the corresponding sum of conditional square Hellinger distances for the weaker experiments $\mathscr{F}_{j,n}$ will still be bounded in probability whether or not they satisfy (A), (C) or the conditions of Theorem 6.

In Section 6 we mentioned the work of Davies [9]. He considers a supercritical Galton–Watson branching process, where the observer sees for each $j \le n$ the size Z_j of the jth generation. This can correspond to our experiment $\mathscr{F}_{j,n}$. Davies studies the properties of the experiment $\mathscr{F}_n$ obtained from the $\mathscr{F}_{j,n}$ by introducing stronger experiments $\mathscr{E}_{j,n}$ as follows. Label the Z_{j-1} individuals of the $(j-1)$st generation by integers $i = 0, 1, 2, \dots, Z_{j-1}$. For individual i, let ξ_i be the size of his progeny in the jth generation. Record the ξ_i for $i = 0, 1, \dots, Z_{j-1}$. Letting j vary from 1 to n, this yields our experiment $\bar{\mathscr{E}}_n$. The situation so described is analogous to the situation we mentioned earlier for Markov processes except that here the loss of information from $\mathscr{E}_{j,n}$ to $\mathscr{F}_{j,n}$ is due to a different cause.

However, the theorems given in the present section do not apply directly to Davies' situation. The negligibility condition called (A) here is not satisfied. A fortiori (C) is not satisfied. To see this, it is enough to note that for n large, the passage from the $(n-1)$st generation to the nth one gives a large fraction of the information contained in the entire set of observations from 1 to n. This is so because of the exponential increase in the size of the population.

The reasons for the validity of Davies' procedure were described in Section 6.

9. Illustrative examples.

EXAMPLE 1. The problem that motivated us to write the present paper arose from stochastic modelling of the activity of nerve cells [39]. The membrane of such a cell contains numerous channels responsible for exchanges between the interior of the cell and the surrounding medium. In the experiments reported in [2], a nerve cell was electrically stimulated every second for a total of approximately 500 s and the behavior of sodium channels was monitored on a microscopic patch of the membrane. Opening of a sodium channel results in an influx of sodium ions and in a measurable change of the electric potential between the interior and the exterior of the cell.

Experimenters try to isolate patches containing very few channels. The actual number m of such channels on a patch is a number that can be ascertained by neurophysiological means.

The channels can be in several different states. In the experiments described they always start in a closed resting state and end in an inactivated absorbing state. The actual states are unobservable; however, every passage through an open state results in a measurable voltage change. For each stimulus, the number of passages through the open state is recorded. The actual activity lasts about 15 ms. It has been modelled by Markov processes as follows.

It is assumed that the various sodium channels in the patch behave independently of one another. For a given stimulus, one channel has a probability p of responding. If it responds, it moves through various states according to a homogeneous Markov process, the total number of passages through the open state being random, distributed according to a geometric distribution. This gives a variable W such that $P[W = 0] = 1 - p$ and $P[W = k] = p\theta^{k-1}(1 - \theta)$ for $k \geq 1$. Since there are m channels on the patch, the result actually observed at the jth stimulus is a sum $Y_j = \sum_{\nu=1}^{m} W_{\nu,j}$ of m independent replicates of W.

The relevant density is given by $P[Y_j = 0] = (1 - p)^m$ and for $k = 1, 2, \ldots$

$$P[Y_j = k] = \sum_{l=1}^{m} \binom{m}{l}\binom{k-1}{k-l} p^l (1-p)^{m-l} \theta^{k-l}(1-\theta)^l,$$

which is a mixture of negative binomials with binomial weights.

There are three parameters p, θ and m. However, for the data described in [2], m was supposed to be equal to 4. In addition, p had been estimated from other considerations and could be taken equal to 0.4. Thus we were left with the problem of estimating θ. In view of technical complications that arose in the experiments, it was decided to use the conditional probabilities

$$q(k, \theta) = P[Y_j = k \mid Y_j \neq 0].$$

It is a simple matter to obtain an auxiliary estimate θ_n^* by the method of moments and to compute $q(k, \theta_n^*)$ and $q(k, \theta_n^* + 1/\sqrt{n})$ for all values of k actually observed in the sample. In the particular set we used, k never exceeded 7.

Let N_k be the number of times the value k was recorded and let Z_k be equal to

$$Z_k = \frac{q(k, \theta_n^* + 1/\sqrt{n})}{q(k, \theta_n^*)} - 1.$$

Compute the sums $S_1 = \sum_k N_k Z_k$ and $S_2 = \sum_k N_k Z_k^2$. The recipe suggested in Section 5 yields the estimated value $T_n = \theta_n^* + (1/\sqrt{n}) S_1 / S_2$.

The operation was duly carried out. Unfortunately, a chi-square test showed that the fit was not acceptable. The fit from maximum likelihood was even worse. Thus we carried out the procedure described in Section 5 estimating both

p and θ from a data set of $n = 560$ observations. The method of moments gave starting values $p^* = 0.457$ and $\theta^* = 0.159$. To compute the approximation $L_n^{(2)}$ of Section 5, one needs to select incremented values $p^* + h_1/\sqrt{n}$ and $\theta^* + h_2/\sqrt{n}$. This was done for various values from $h_i = -1$ to $h_i = +1$. The estimates $\tilde{p}$ so obtained varied only in their fourth decimals. The estimates $\tilde{\theta}$ varied in their third decimals, from 0.1635 to 0.1665. This value $\tilde{\theta} = 0.1665$ with $\tilde{p} = 0.4535$ gave the best fit, as judged by chi-square.

For comparison, the maximum-likelihood estimates were $\hat{p} = 0.4538$ and $\hat{\theta} = 0.1649$, giving a slightly larger, but acceptable chi-square.

The technique also provides automatically approximate values of the covariance matrix of $(\tilde{p}, \tilde{\theta})$ from the estimated M_n^{-1} of Section 5. Here the estimated covariance matrix is

$$10^{-4}\begin{pmatrix} 1.88 & -1.42 \\ -1.42 & 4.32 \end{pmatrix}.$$

EXAMPLE 2. This is similar to Example 1 but intended to show that, often, one must proceed with due caution. Consider a two-dimensional parameter $\theta = (\mu, \sigma)$ with $\mu \in (-\infty, +\infty)$ and $\sigma \in (0, \infty)$. Let $f(x, \theta)$ and $g(x, \theta)$ be two densities on the real line and let α be a known number $\alpha \in (0, 1)$, say $\alpha = 1/10$. (The case where α is also unknown is very important. It can be treated by extensions of the arguments given here. However, many complications can occur. They would take us too far afield.)

Let us suppose that a refined experimenter could observe n independent pairs (X_j, I_j), where I_j takes value 0 with probability $1 - \alpha$ and value 1 with probability α.

Assume that if $I_j = 0$, then X_j has the density $f(x, \theta)$. If $I_j = 1$, then X_j has the density $g(x, \theta)$. For this observer, the likelihood function takes the form

$$L_n(\theta) = \prod_{j=1}^{n} \left[(1 - \alpha) f(x_j, \theta)\right]^{1 - I_j} \left[\alpha g(x_j, \theta)\right]^{I_j}.$$

If both f and g satisfy the (DQM) condition of Section 7 with respect to θ, so does the density of (X_j, I_j).

Now consider a restricted observer who sees only the variables X_j. His likelihood function takes the form

$$L_n^*(\theta) = \prod_{j=1}^{n} \left\{(1 - \alpha) f(x_j, \theta) + \alpha g(x_j, \theta)\right\}.$$

According to Section 7 if (DQM) is satisfied for the refined observer, it is also satisfied for the restricted one. Thus one can proceed to the construction of estimates as in Section 5 at least whenever the Fisher information matrix does not become singular.

Now consider the two special cases where $f(x, \theta)$ is the ordinary normal,

$$f(x, \theta) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x - \mu)^2\right\},$$

and where, for case 1,

$$g_1(x, \theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\},$$

whereas for case 2, $g_2(x, \theta)$ is the Cauchy,

$$g_2(x, \theta) = \frac{1}{\sigma\pi} \frac{1}{1 + \sigma^{-2}(x - \mu)^2}.$$

In both cases, $\sup_\theta L_n(\theta) < \infty$ unless $\sum_{j=1}^n I_j < 2$. Thus $\sup_\theta L_n(\theta) < \infty$ except for cases having a total probability at most

$$[1 + (n - 1)\alpha](1 - \alpha)^{n-1}.$$

By contrast (see Kiefer and Wolfowitz [18]), $\sup_\theta L_n^*(\theta) = \infty$ always. The infinite value can be achieved by taking μ equal to any one of the observed x_j and taking $\sigma = 0$. This does not contradict the inequalities of Section 3, since $E\{\sup_\theta L_n(\theta)|x_1, \dots, x_n\} = \infty$ even though the probability that $\sup_\theta L_n(\theta)$ be large tends to 0 exponentially fast as $n \to \infty$.

As a result of this state of affairs, one can assert the following. Any procedure that attempts to select, iteratively or not, a value $\hat{\theta}_n$ such that $L_n^*(\hat{\theta}_n) = \sup_\theta L_n^*(\theta)$ will either get trapped at or near a local maximum or it will be such that the $\hat{\sigma}_n$ of $\hat{\theta}_n = (\hat{\mu}_n, \hat{\sigma}_n)$ tends to 0, or it will not achieve any maximum local or not.

This is particularly visible in the Cauchy contamination case (called case 2 in the preceding discussion). There, an appealing auxiliary estimate of μ is the median m_n of the observations. However, if n is odd, $\sup_\sigma L_n^*[(m_n, \sigma)] = \infty$. Thus an iterative procedure that starts with an estimate (m_n, σ_n^*) of $\theta = (\mu, \sigma)$ may readily be trapped into a path where L_n^* goes to ∞. In spite of this, the arguments of Section 5 remain applicable. One can prevent the auxiliary estimates from searching for singularities in the likelihood function by a variety of procedures: discretization, smoothing, etc. If so, the one-step estimates of Section 5 are guaranteed to work "asymptotically." (Contrary to common advice, it may not be safe to iterate!) This, of course, does not guarantee good behavior for a fixed finite n. As is the case for most statistical problems, we do not have any general recipe that always avoids all troubles for all finite n. The only general recommendation that comes to mind is to exert due caution and at least check that the estimated model fits the data.

The general theme of the present example can be applied to many cases where the densities are mixtures of smooth families. This remains true even if the mixing proportions are unknown. However, in this more general situation, due care must be exercised in the selection of auxiliary estimates. Also identifiability problems may make life even more complicated (as is the case in the mixture of Gaussian distributions of our case 1. If $\sigma = 1$, the mixing proportion α is estimable for the refined observer but not for the restricted one. Neighborhoods of the form $\{\alpha: \sqrt{n}\,|\alpha - \alpha_0| \leq b\}$ are then inappropriate. We shall return elsewhere to what happens on the neighborhoods described in Section 2).

EXAMPLE 3. Consider, for each integer $j \le n$, an homogeneous Markov process $\{Z_j(t), t \in [0,1]\}$ with three states S_1, S_2, S_3 and an infinitesimal generator A. Let the experiment $\mathscr{E}_{j,n}$ consist of observing the $Z_j(t)$ for the entire interval $[0,1]$, the Z_j being mutually independent. Let $\mathscr{F}_{j,n}$ be the experiment in which the values of $Z_j(t)$ are seen only at times 0 and 1. Assume, for instance, that the initial distribution of $Z_j(0)$ is uniform over the three states. The problem is to estimate the generator A.

The transition probabilities $P[Z_j(1) = S_k | Z_j(0) = S_1]$ are the entries in the matrix exponential $\exp\{A\}$. For any fixed matrix A, it is fairly easy to compute the exponential $\exp\{A\}$, but carrying out a maximum likelihood procedure is another matter. However, one can look at the matrix M_n that gives the observed frequencies of transitions, take its logarithm and take the generator A_n^* closest to that logarithm. If A is unrestricted, there are six parameters. Application of the method of Section 5 will necessitate computation of $\exp\{A_n^*\}$ and of six exponentials of the type $\exp\{A_n^* + u_{n,i}\}$ for suitably selected deviations $u_{n,i}$, $i = 1, \dots, 6$.

Often the generator A is assumed to belong to a parametrized family. For instance, one could consider a one-dimensional family $A(\theta)$ of the type $A(\theta) = (1-\theta)A_0 + A_1$ for fixed generators A_0 and A_1 and for $\theta \in [0,1]$. In such a case, one would take the $A(\theta_n^*)$ closest to $\log M_n$ in the model. It would be enough to compute two exponentials of the type $\exp\{A(\theta_n^*)\}$ and $\exp\{A(\theta_n^* + u_n)\}$.

Here several complications may occur.

The matrices A may have complex eigenvalues. Thus the determination of $\log M$ is an uncertain affair. It may also be that several different generators yield the same transition probabilities. (See [7] and [36].) Here is an example of a parametrized family $A(\theta) = (1-\theta)A_0 + A_1$, where $\exp\{A_0\} = \exp\{A_1\}$ but where $\exp\{A(r)\} \ne \exp\{A(s)\}$ for any pair (r,s), $r \ne s$ in $(0,1)$. Let $A_0 = -aI + aM - (3\pi/2)\Phi$ and $A_1 = -aI + aM + (3\pi/2)\Phi$, where a is the number $a = 3\pi\sqrt{3}/2$, the matrices I and M are, respectively, the identity matrix and the matrix whose entries are all equal to $1/3$ and the matrix Φ is

$$\Phi = \begin{pmatrix} 0 & 1/2 & -1/2 \\ -1/2 & 0 & 1/2 \\ 1/2 & -1/2 & 0 \end{pmatrix}.$$

The same type of example can be used to illustrate the results of Section 8. Instead of observing n independent replicates $Z_j(t)$, $t \in [0,1]$, observe just one process $Z(t)$, $t \in [0,n]$. Let $\mathscr{E}_{j,n}$ be the observation of the process in $(j-1, j]$ conditionally on the value at $j-1$. The restricted experiment $\mathscr{F}_{j,n}$ consists of observing only $Z(j)$. The same kind of analysis will apply as long as all the states communicate with one another.

EXAMPLE 4. Let X_j, $j = 1, 2, \dots, n$, be i.i.d. with density $[1 - |x-\theta|]^+$ on the line. Let U_j, $j = 1, \dots, n$, be i.i.d. Independent of the X_j with a fixed known distribution, say $N(0,1)$. Let $\mathscr{E}_{j,n}$ consist of observing X_j and let $\mathscr{F}_{j,n}$ consist of observing $X_j + U_j$.

For the $\mathscr{E}_{j,n}$ the LAN conditions are satisfied in neighborhoods of the type $V_n = \{\sqrt{n \log n}\,|\theta - \theta_0| \le b\}$. Thus, by Section 5, the $\mathscr{F}_{j,n}$ will also satisfy the LAN conditions in these same neighborhoods. However, this is no great consolation since in these V_n the product $\bar{\bar{\mathscr{F}}}_n$ degenerates. The $\bar{\bar{\mathscr{F}}}_n$ do satisfy the LAN conditions in neighborhoods of the type $\{\sqrt{n}\,|\theta - \theta_0| \le b\}$. This is immediate from the fact that the distributions satisfy the (DQM) condition of Section 7. However, it is not a consequence of the results of Sections 4–7. It could be deduced from the arguments of Section 6 if one already knew of the existence of distinguished sequence of estimates for the local problems.

EXAMPLE 5 (Method of moments). Let $\omega_1, \omega_2, \dots, \omega_n$ be independent random elements with values in a set Ω and a common distribution p_θ; $\theta \in \Theta_n$. The method of moments consists of selecting a measurable function ϕ_n from Ω to a Euclidean space R^m and estimating θ by minimizing the Euclidean norm

$$\left\| \sum_{j=1}^{n} \left[\phi_n(\omega_j) - E_\theta \phi_n(\omega_j) \right] \right\|.$$

We shall assume that Θ_n is a convex subset of a Euclidean space R^k and that the origin of R^k belongs to Θ_n.

The passage from ω_j to the vector $\phi_n(\omega_j)$ can result in a loss of information. It is subject to the results of Sections 4 and 5. Let

$$P_{\theta,n} = \mathscr{L}[\omega_1, \omega_2, \dots, \omega_n | \theta],$$
$$Q_{\theta,n} = \mathscr{L}[\phi_n(\omega_1), \dots, \phi_n(\omega_n) | \theta].$$

If $\mathscr{E}_n = \{P_{\theta,n};\ \theta \in \Theta_n\}$ is asymptotically Gaussian, so is $\mathscr{F}_n = \{Q_{\theta,n};\ \theta \in \Theta_n\}$. If $\mathscr{E}_n$ satisfies the LAN conditions, so will $\mathscr{F}_n$.

The passage from the n-tuple $\{\phi_n(\omega_1), \dots, \phi_n(\omega_n)\}$ to the sum $T_n = \sum_{j=1}^n \phi_n(\omega_j)$ can also result in a further loss of information. Here the results of Sections 4 and 5 do not apply. The situation is complex. We shall discuss only a special case for which we need an appropriate affine invariant distance between measures on R^m. We shall use the half-space distance. A half-space is any set of the form $\{x\colon\ t'x \ge a\}$ or $\{x\colon\ t'x > a\}$ with a real and with t' in the dual of R^m. The half-space norm of a signed measure μ is

$$\|\mu\|_H = \sup_H |\mu(H)|,$$

for supremum taken over all half-spaces.

We shall use the following restrictions.

(a) *The experiments $\mathscr{F}_n$ satisfy the LAN conditions.*

Let us call "outliers" any one of the $\phi_n(\omega_j)$ selected by the statistician after inspection of the n-tuple $\{\phi_n(\omega_j);\ j = 1, 2, \dots, n\}$ by any measurable criterion he pleases.

(b) *Let T_n^* be the sum T_n with one outlier removed. Then no matter what the choice of outlier the difference*

$$\mathscr{L}(T_n^* | 0) - \mathscr{L}(T_n | 0)$$

tends to 0 *for the half-space distance.*

This is a commendable form of robustness. By a result of Lévy, it is equivalent to the assertion that there exist Gaussian measures $G_{0,n}$ such that $\|\mathscr{L}(T_n|0) - G_{0,n}\|_H$ tends to 0.

The following condition can always be insured by restricting oneself to an affine subspace.

(c) *The support of* $\mathscr{L}(T_n|0)$ *is not contained in any proper affine subspace of* R^m.

If (b) and (c) are satisfied, one can renormalize T_n so that $G_{0,n}$ is the standard $\mathscr{N}(0, I)$ distribution on R^m. We shall do so.

Let $\Lambda_n(\theta) = \log(dQ_{\theta,n}/dQ_{0,n})$. Condition (a) implies the existence of random variables $V_{n,j}$, functions of $\phi_n(\omega_j)$ and the existence of matrices M_n such that

$$\Lambda_n(\theta) - \theta' \sum_{j=1}^{n} V_{n,j} + \tfrac{1}{2}\theta' M_n \theta$$

tends to 0 in probability. Furthermore, let $\bar{V}_n = \sum_{j=1}^{n} V_{n,j}$. Under conditions (a)–(c) with the renormalization indicated, the joint distribution $\mathscr{L}[\bar{V}_n, T_n|0]$ admits a Gaussian approximation for the Lévy or Prokorov distance. Indeed, $\mathscr{L}(\bar{V}_n|0)$ and $\mathscr{L}(T_n|0)$ admit separate Gaussian approximations and the joint distribution $\mathscr{L}(\bar{V}_n, T_n|0)$ admits an infinitely divisible approximation. It follows that there are matrices B_n such that if $\theta_n \in \Theta_n$, then

$$\| \mathscr{L}[T_n|\theta_n] - \mathscr{N}[B_n\theta_n, I] \|_H$$

tends to 0. These matrices can be obtained as covariance matrices in the normal approximations to $\mathscr{L}[\bar{V}_n, T_n|0]$.

This suggests that the experiments $\bar{\mathscr{F}}_n = \{\bar{Q}_{\theta,n};\ \theta \in \Theta_n\}$ with $\bar{Q}_{\theta,n} = \mathscr{L}[T_n|\theta]$ will also be asymptotically Gaussian. Indeed, we do not presently have any examples to the contrary. It does not follow, however, that if the $\bar{\mathscr{F}}_n$ are asymptotically Gaussian they are approximable by the $\mathscr{G}_n = \{G_{\theta,n};\ \theta \in \Theta\}$ with $G_{\theta,n} = \mathscr{N}[B_n\theta, I]$, as will be shown.

To enforce such a property, one needs to add further restrictions. To state them consider the following possible modifications of T_n. Let ε_n be independent random vectors independent of T_n. Let $T_n' = T_n + \varepsilon_n$. Also, consider an integer $\nu(n)$ and let T_n'' be obtained as follows. Write the coordinates of T_n in decimal expansions. Then T_n'' has the same decimals as T_n up to $\nu(n)$. The remaining ones are all put equal to 0. Consider binary experiments $\mathscr{B}_n = \{\bar{Q}_{0,n}, \bar{Q}_{\theta_n,n}\}$, $\mathscr{B}_n' = \{Q_{0,n}', Q_{\theta_n,n}'\}$ and $\mathscr{B}_n'' = \{Q_{0,n}'', Q_{\theta_n,n}''\}$ with $\theta_n \in \Theta_n$ and with $Q_{\theta,n}' = \mathscr{L}[T_n'|\theta]$, $Q_{\theta,n}'' = \mathscr{L}[T_n''|\theta]$.

(d) *For any arbitrary choice of noises* ε_n *if* $\|\bar{Q}_{0,n} - Q_{0,n}'\|_H$ *tends to* 0 *so does the Lévy distance between the distributions* $\mathscr{L}[d\bar{Q}_{\theta_n,n}/d\bar{Q}_{0,n}|0]$ *and* $\mathscr{L}[dQ_{\theta_n,n}'/dQ_{0,n}'|0]$.

(d′) *Assume that* $\mathscr{L}(T_n|0)$ *has been normalized as mentioned previously. Then, for any choice of cut-off* $\nu(n)$ *such that* $\nu(n) \to \infty$, *the Lévy distance between* $\mathscr{L}[d\bar{Q}_{\theta_n,n}/d\bar{Q}_{0,n}|0]$ *and* $\mathscr{L}[dQ_{\theta_n,n}''/dQ_{0,n}''|0]$ *tends to* 0.

Conditions (d) or (d′) can be taken as expressing the fact that the information contained in T_n occurs there in a robust sort of a way.

Under conditions (a)–(d), the T_n form a distinguished sequence for the experiments $\bar{\mathscr{F}}_n$. Thus Theorem 4 of Section 6 becomes applicable. The same holds true if (d) is replaced by (d′).

This is not difficult to prove. If (d) is used, one can find noises ε_n such that $\mathscr{L}[T_n'|0] - \mathscr{N}(0, I)$ tends to 0 in total variation. The same will then be true of $\mathscr{L}[T_n'|\theta_n] - \mathscr{N}(B_n\theta_n, I)$. If (d′) is used instead, one can obtain a similar type of effect by considering discretizations of $\mathscr{N}(0, I)$ that match $\mathscr{L}(T_n''|0)$ in total variation norm except for a difference that tends to 0 as $n \to \infty$. The argument is given in [27], Chapter 7, Section 3, Theorem 2. Condition (d′) suggests a construction to obtain sums T_n that satisfy (a)–(c) and are such that the corresponding experiments $\bar{\mathscr{F}}_n$ are asymptotically Gaussian but significantly more informative than the $\mathscr{G}_n$.

To do this, suppose for simplicity that the ϕ_n are real valued and that they and the the $V_{n,j}$ have been truncated so that $|\phi_n(\omega_j)| \le 1$ and $|V_{n,j}| \le 1$. Under (a)–(c) and the normalization described earlier, this can be done without modifying the asymptotic behavior of the distributions or experiments concerned.

Let $\tilde{\phi}_n(\omega_j)$ be $\phi_n(\omega_j)$ with all its decimals beyond the nth one put equal to 0. Let $\phi_n^+(\omega_j) = \tilde{\phi}_n(\omega_j) + 10^{-2n}V_{n,j}$.

Then the sum $T_n^+ = \sum_{j=1}^n \phi_n^+(\omega_j)$ defines an experiment $\mathscr{F}_n^+$ that admits a Gaussian approximation $\mathscr{G}_n^+$ such that the difference between $\mathscr{F}_n$ and $\mathscr{G}_n^+$ tends to 0. Indeed, the sum $\bar{V}_n$ is asymptotically sufficient and distinguished for $\mathscr{F}_n$. The asymptotic behavior of distributions is not changed for the half-space distance. However, $\mathscr{G}_n^+$ can be considerably more informative than $\mathscr{G}_n$. This can be seen, for instance, in the gamma density $[\Gamma(\alpha)]^{-1}e^{-x}x^{\alpha-1}$ if one estimates α by $(1/n)\sum_{j=1}^n X_j$.

A further remark is as follows. In the method of moments, one estimates θ globally, whereas our conditions (a) and (b) are meant to be applied locally in small sets Θ_n. In the local situation it is tempting to replace the expectation $E_\theta T_n$ by the centers $B_n\theta$ of the Gaussian approximations. This can be done under suitable integrability requirements. Then minimizing the Euclidean norm $\|T_n - B_n\theta\|$ becomes a linear problem. It will have a point solution $\hat{\theta}_n$ only if the dimension m of T_n is at least equal to the dimension k of θ and if B_n has full rank. The solution is then $\hat{\theta}_n = (B_n'B_n)^{-1}B_n'T_n$. If $m > k$, passage from T_n to $\hat{\theta}_n$ will often lose information. If $m = k$, no information is lost. However, unless conditions such as (d) or (d′) are satisfied, information may be lost in the passage to approximate solutions. This can happen in particular in the minimization of $\|T_n - E_\theta T_n\|$.

Example 6 (Grouping data). Let X_j, $j = 1, 2, \dots n$, be i.i.d. with densities $f(x, \theta)$, $\theta \in \Theta$, on the real line. For $\mathscr{E}_{j,n}$ one observes X_j itself. For $\mathscr{F}_{j,n}$ one sees only the integer Y_j nearest to X_j. (If X_j happens to be equal to $k + \frac{1}{2}$ one takes k or $k + 1$ by tossing a coin.)

Even for fairly simple functions $f(x, \theta)$, it is usually obnoxious to deal with integrals of the type $\int_{k-1/2}^{k+1/2} f(x, \theta)\, dx$. However, let us assume, for instance, that $\Theta = R^1$ and that the $f(x, \theta)$ satisfy (DQM) of Section 7. It is often fairly easy to verify that the $\mathscr{F}_{j,n}$ satisfy the identifiability conditions of Proposition 5, Section 7. Assuming this, and assuming that one can find suitable auxiliary estimates θ_n^*, the method described in Section 5 will require only computation of the integrals $\int_{k-1/2}^{k+1/2} f(x, s)\, dx$ for $q + 1$ values of s and for those k that have been seen in the sample. Note that the results remain applicable to situations where the grouping classes are made dependent on the number n of observations.

For grouping in classes that are random, dependent on the observations, other arguments are required.

EXAMPLE 7 (Censoring). Consider independent pairs (X_j, Y_j) of positive random variables with a joint density $f(x, y, \theta)$ that depends on a Euclidean parameter θ. The refined experiment $\mathscr{E}_{j,n}$ consists of observing the pair (X_j, Y_j). In the restricted experiment $\mathscr{F}_{j,n}$, one sees only $\min(X_j, Y_j)$ and an indicator variable I_j equal to 0 if $X_j > Y_j$ and to unity if $X_j \leq Y_j$. Assume that the $\mathscr{E}_{j,n}$ yield a product experiment $\tilde{\mathscr{E}}_n$ that satisfies the LAN conditions on sets of the form $\{\theta: \sqrt{n}\,|\theta - \theta_0| \leq b\}$. Then the restricted product $\tilde{\mathscr{F}}_n$ will also satisfy the same LAN condition. If, in addition, the densities $f(x, y, \theta)$ satisfy the differentiability in quadratic mean condition of Section 7, the same will be true of the distributions of $\{\min(X_j, Y_j), I_j\}$.

In such a situation it is well known (see [30] and [38]) that a θ that is identifiable on the complete experiment may become unidentifiable on the restricted one.

However, most of the examples discussed in the literature satisfy conditions analogous to the following.

(1) The densities $f(x, y, \theta)$ have the form $f(x, y, \theta) = g(x, \theta)h(y, \theta)$ so that X_j and Y_j are independent.
(2) The supports of X_j and Y_j are the same.
(3) Let $G(x, \theta) = \int_x^\infty g(t, \theta)\,dt$ and $H(y, \theta) = \int_y^\infty h(t, \theta)\,dt$ be the survival functions attached, respectively, to g and h. There are two positive numbers α and β such that $|\theta_1 - \theta_2| > \alpha$ implies $\sup_x |G(x, \theta_1) - G(x, \theta_2)| > \beta$.
(4) The family $\{g(x, \theta);\ \theta \in \Theta\}$ satisfies (DQM) with nonsingular covariance matrices.

Under such conditions, θ is identifiable. Also one can get auxiliary estimates θ_n^* that converge at the $\sqrt{n}$ rate. A possibility is to start with some nonparametric estimate $\hat{G}_n$ of the survival function, for instance, the Kaplan–Meier estimate or the Nelson–Aalen estimate from cumulative hazard functions. Then apply a minimum distance technique as in the procedure that yields Proposition 5, Section 7.

A direct verification that nonsingularity of Fisher information matrices is inherited by the weaker experiment seems awkward. However, this nonsingularity is implied by the properties of the estimate just described.

(In the most standard case, where h does not depend on θ, verification is trivial.)

Thus the results of Section 5 can be applied here. The density of $Z_j = \min(X_j, Y_j)$ has the form

$$g(z, \theta)H(z, \theta), \quad \text{on } I_j = 1$$

and

$$h(z, \theta)G(z, \theta), \quad \text{on } I_j = 0.$$

Assuming that these formulas can be computed for the observed values (Z_j, I_j) and a few values of the parameter θ, Section 5 will yield asymptotically efficient estimates.

Acknowledgments. We are indebted to the referees who pointed out a gap in our discussion of the method of moments and the possibility of applying the results of Section 8 to the aggregated Markov chains situation.

REFERENCES

[1] AALEN, O. (1978). Nonparametric estimation of partial transition probabilities in multiple decrement models. *Ann. Statist.* **6** 534–545.

[2] ALDRICH, R. W., COREY, D. P. and STEVEN, C. F. (1983). A reinterpretation of mammalian sodium channel gating based on a single channel recording. *Nature* **306** 436–441.

[3] ANDERSEN, P. K. and BORGAN, O. (1985). Counting process models for life history data: A review. *Scand. J. Statist.* **12** 97–158.

[4] BASAWA, I. and PRAKASA RAO, B. L. S. (1980). *Statistical Inference for Stochastic Processes.* Academic, New York.

[5] BASAWA, I. V. and SCOTT, D. J. (1983). *Asymptotic Optimal Inference for Non-Ergodic Models. Lecture Notes in Statist.* **17**. Springer, New York.

[6] BLACKWELL, D. (1953). Equivalent comparisons of experiments. *Ann. Math. Statist.* **24** 265–272.

[7] CUTHBERT, J. R. (1972). On uniqueness of the logarithm for Markov semi-groups. *J. London Math. Soc.* (2) **4** 623–630.

[8] DAVIES, R. B. (1984). Asymptotic inference in the frequency domain. In *Handbook of Statistics* **3**: *Time Series in the Frequency Domain* (D. R. Brillinger and P. R. Krishnaiah, eds.). North-Holland, Amsterdam.

[9] DAVIES, R. B. (1985). Asymptotic inference when the amount of information is random. In *Proc. of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. M. Le Cam and R. A. Olshen, eds.) **2** 841–864. Wadsworth, Monterey, Calif.

[10] FABIAN, V. and HANNAN, J. (1985). *Introduction to Probability and Mathematical Statistics.* Wiley, New York.

[11] FEIGIN, P. D. (1986). Asymptotic theory of conditional inference for stochastic processes. *Stochastic Process. Appl.* **22** 89–102.

[12] GREENWOOD, P. E. and SHIRYAYEV, A. N. (1985). *Contiguity and the Statistical Invariance Principle.* Gordon and Breach, New York.

[13] HÁJEK, J. and ŠIDÁK, Z. (1967). *Theory of Rank Tests.* Academia, Prague.

[14] IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation, Asymptotic Theory.* Springer, New York.

[15] JACOBSON, M. (1982). *Statistical Analysis of Counting Processes. Lecture Notes in Statist.* **12** Springer, New York.

[16] JEGANATHAN, P. (1980). Asymptotic theory of estimation when the limit of the log likelihood is mixed normal. Thesis, Indian Statistical Institute.

[17] JEGANATHAN, P. (1981). On a decomposition of the limit distribution of a sequence of estimators. *Sankhyā Ser. A* **43** 26–36.

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

[19] LAWLESS, J. F. and MCLEISH, D. L. (1984). The information in aggregate data from Markov chains. *Biometrika* **71** 419–430.

[20] LE CAM, L. (1960). Locally asymptotically normal families of distributions. *Univ. California Publ. Statist.* **3** 37–98.

[21] LE CAM, L. (1964). Sufficiency and approximate sufficiency. *Ann. Math. Statist.* **35** 1419–1455.

[22] LE CAM, L. (1969). *Théorie asymptotique de la décision statistique.* Les Presses de l'Université de Montréal, Montréal.

[23] LE CAM, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.* **41** 802–828.

[24] LE CAM, L. (1972). Limits of experiments. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **1** 245–261. Univ. California Press.

[25] LE CAM, L. (1977). On the asymptotic normality of estimates. *Proc. Symp. to Honor J. Neyman* 203–217. PAN, Warsaw.

[26] LE CAM, L. (1985). Sur l'approximation de familles de mesures par des familles Gaussiennes. *Ann. Inst. H. Poincaré Sect. B.* (*N.S.*) **21** 225–287.

[27] LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

[28] MCLEISH, D. L. (1985). Estimation for aggregate models: The aggregate Markov chain. *Canad. J. Statist.* **12** 265–282.

[29] NEVEU, J. (1972). *Martingales à temps discret*. Masson, Paris.

[30] PETERSON, A. V. (1975). Nonparametric estimation in the competing risk problem. Technical Report 13, Dept. Statistics, Stanford Univ.

[31] PFANZAGL, J. and WEFELMEYER, W. (1982). *Contributions to a General Asymptotic Statistical Theory. Lecture Notes in Statist.* **13** Springer, New York.

[32] REBOLLEDO, R. (1978). Sur les applications de la théorie des martingales à l'étude statistique d'une famille de processus ponctuels. *Lecture Notes in Math.* **636** 27–70. Springer, Berlin.

[33] REBOLLEDO, R. (1980). Central limit theorem for local martingales. *Z. Wahrsch. verw. Gebiete* **51** 269–286.

[34] ROUSSAS, G. G. (1972). *Contiguous Probability Measures: Some Applications in Statistics*. Cambridge Univ. Press, Cambridge.

[35] SHIRYAYEV, A. N. (1981). Martingales: Recent developments, results and applications. *Internat. Statist. Rev.* **49** 199–234.

[36] SPEAKMAN, J. O. (1967). Two Markov chains with a common skeleton. *Z. Wahrsch. verw. Gebiete* **7** 224.

[37] STECK, G. P. (1957). Limit theorems for conditional distributions. *Univ. California Publ. Statist.* **2** 237–284.

[38] TSIATIS, A. (1975). A non-identifiability aspect of the problem of competing risks. *Proc. Nat. Acad. Sci. U.S.A.* **72** 20–22.

[39] YANG, G. L. (1988). Stochastic modelling of sodium channel gating phenomena and estimation of parameters. In preparation.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
COLLEGE PARK, MARYLAND 20742