SOME RECENT DEVELOPMENTS IN INFERENCE FOR DISCRETE STOCHASTIC PROCESSES

BY

PAUL D. FEIGIN

TECHNICAL REPORT NO. 137
OCTOBER 1979

PREPARED UNDER THE AUSPICES OF
NATIONAL SCIENCE FOUNDATION GRANT MCS77-16974

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
SOME RECENT DEVELOPMENTS IN INERENCE FOR
DISCRETE STOCHASTIC PROCESSES

by
PAUL D. FEIGIN

TECHNICAL REPORT NO. 137
OCTOBER 1979

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MCS77-16974

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
SOME RECENT DEVELOPMENTS IN INERENCE FOR
DISCRETE STOCHASTIC PROCESSES

Paul D. Feigin
Technion - Isreal Institute of Technology

ABSTRACT

We review certain aspects of recent theoretical developments in statistical inference for parametric stochastic process models. Emphasis is focused on models which are essentially nonergodic but nevertheless have certain regular features. The (i) use of martingale theory; and (ii) need to consider various principles of statistical inference are discussed in the framework of an asymptotic theory of inference for stochastic processes.
1. Introduction

The classical approach to modelling and conducting statistical inference for real life processes has been to assume the independence of successive observations and usually that they come from the same distribution. Based on these assumptions, various regularity conditions lead to proofs of the optimality of specific estimation and testing procedures. In many cases the optimality described is for "large samples" or sequences of observations - in other words, asymptotic properties of the inference procedures are analyzed.

An attempt to make the models better approximations to reality often involves dropping the independence assumption and adopting a stochastic process formulation of the mechanism generating the sequence of observations. That many such generalizations are to Markov chain models is not surprising for possibly two reasons: first, the one-step dependence is a natural extension of the complete independence notion; and second, it is harder to formulate or indeed conceive of an alternative to the (possibly m-th order) Markov structure for a real sequence of variables.

Some of the earlier discussions of statistical inference for Markov models have remained close to the independent and identically distributed (i.i.d.) case by dealing with the important stationary and ergodic situations (Billingsley (1961), Roussas (1972)) or by extracting an independent sequence characterizing the stochastic process (Grenander (1951)). Of course, the whole field of stationary time series analysis also falls under the heading of inference for stochastic (although not necessarily Markovian) processes. Some work on essentially ergodic but not necessarily stationary processes goes back to Wald (1948) and is discussed in Basawa, Feigin and Heyde (1976).
Our main interest will be in processes that are not ergodic. As far as inference for such processes is concerned, until recently references have been limited to the analysis of specific models. In this context one may mention, for example, Harris' (1948) work on branching processes, Moran's (1951) and Keiding's (1974, 1975) work on birth and birth-death processes, as well as others.

Recently there has been an attempt to provide a general theory of statistical inference for stochastic processes, particularly the nonergodic ones. This development has its roots in the ideas of Silvey (1961) who recognized the importance of martingales in this theory. It is the aim of this review to discuss the main aspects of this theory for the discrete time case: the continuous time case will be considered elsewhere (a partial discussion is given in Feigin (1976)). We will almost exclusively restrict our attention to the asymptotic theory of inference: the limits being taken as more and more observations are made on a single history of the process.

There are two discernible themes in the present development. The first concerns the relevance of martingale theory to inference for stochastic processes - in this sense we are adding to the case Heyde (1972) made for including martingale theory in the "statistician's repertoire". The second theme concerns the difficulty in directly applying classical statistical concepts to the stochastic process situation with the result that one becomes involved in some controversial issues of statistical theory.

In Section 2 we establish the central role of martingale theory in inference for stochastic processes and in Section 3 we introduce some useful limit theory. Section 4 deals with the special case of conditional
exponential families; and in Section 5 we consider the wider class of so-called regular nonergodic processes. Finally, in Section 6, we investigate the application of the asymptotic results to questions of hypothesis testing and interval estimation.

2. Score Function as a Martingale

Consider the family of probability measures $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$ each defined on the measurable space $(\Omega, \mathcal{F})$ and a sequence of random variables $\{X_1, X_2, X_3, \ldots\}$ defined on the same space. We will largely restrict our discussion to the case in which $\Theta$ is an interval on the real line. Let $P^n_\theta$ be the restriction of $P_\theta$ to $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, the $\sigma$-field generated by $X_1, \ldots, X_n$, and let $\mu$ be any measure dominating the family $\mathcal{P}$, with $\mu^n$ restricted similarly. The likelihood function

$$L_n(\theta) = \frac{dP^n_\theta}{d\mu^n}$$  \hspace{1cm} (2.1)

forms the basis for conducting inference about $\theta$ from the observation of $X_1, \ldots, X_n$. The log likelihood function will be denoted by $\hat{L}_n(\theta)$ and we will consider the maximum likelihood estimate (MLE) of $\theta$ as a solution of

$$\hat{L}_n(\theta) = 0$$  \hspace{1cm} (2.2)

where $\cdot$ denotes differentiation with respect to $\theta$. The function

$$U_n(\theta) = \hat{L}_n(\theta)$$

-the existence of which is assumed - is often called the score function and for $\theta$ fixed it is an $\mathcal{F}_n$-measurable random variable.

In the classical independent identically distributed case $U_n(\theta)$ is a sum of independent identically distributed random variables and it is this fact, together with regularity conditions, that is used to establish
asymptotic properties of inference based on the MLE (via equation (2.2)).

We show below that under virtually the same regularity conditions for generally dependent $X$'s $\{U_n(\theta), \mathcal{F}_n; n \geq 1\}$ is a $\mathbb{P}_\theta$-martingale with zero mean and finite second moments. Hence, to obtain asymptotic properties of the MLE from (2.2) in the general dependent (stochastic process) case the obvious approach is via martingale limit theory. The first result is due essentially to Silvey (1961):

**Proposition 2.1.** If for all $n \geq 1$

(i) $E_\theta [\lvert U_n(\theta) \rvert^2] < \infty$;

(ii) $\frac{d}{d\theta} \{\int_A L_n(\theta) \, d\mu^n\} = \int_A \left\{\frac{d}{d\theta} L_n(\theta) \right\} d\mu^n$ \hspace{1cm} (2.3)

for all $A \in \mathcal{F}_n$ ($\mathcal{F}_0$ being the trivial $\sigma$-field);

then $\{U_n(\theta), \mathcal{F}_n; n \geq 1\}$ is a zero-mean square integrable $\mathbb{P}_\theta$-martingale.

**Proof.** From (i) $E_\theta [\lvert U_n(\theta) \rvert] < \infty$ and from (ii) we can justify the following chain of equalities. For $n \geq 1$ and $A \in \mathcal{F}_{n-1}$

$$\int_A U_n(\theta) \, d\mathbb{P}_\theta = \int_A \left\{\frac{d\mathbb{P}^n}{d\mu^n} \right\}^{-1} \left\{\frac{d}{d\theta} L_n(\theta) \right\} d\mathbb{P}_\theta^n$$

$$= \int_A \frac{d}{d\theta} L_n(\theta) \, d\mu^n$$

$$= \frac{d}{d\theta} \{\int_A \mathbb{P}_\theta^n \} = \frac{d}{d\theta} \{\mathbb{P}_\theta^n(A)\}$$

$$= \frac{d}{d\theta} \{\mathbb{P}^{n-1}(A)\}$$

$$= \int_A U_{n-1}(\theta) \, d\mathbb{P}_\theta$$
and hence \( E_\theta[U_n(\theta)|\mathcal{F}_{n-1}] = U_{n-1}(\theta) \) a.s. \([P_\theta]\). (In statements such as the last we may drop the qualification a.s. in the sequel.) The zero-mean property follows from (2.4) by taking \( A = \Omega \).

The condition (2.3) is not really intended as a practical criterion to be used in applications, although sufficient conditions for (2.3) can readily be obtained. In most cases regular conditional densities can be used together with other properties of the process \( \{X_1, X_2, \ldots \} \) to check the martingale property directly. However the very technical nature of the condition does suggest the almost universal occurrence of the martingale property of the score function for the parametric families of stochastic processes one is likely to encounter. This conclusion is indeed remarkable and shows that the key to extending Cramer's (1946) asymptotic theory to general stochastic processes is martingale limit theory.

We now define some further quantities associated with the score function martingale. Assume that \( L_n(\theta) \) is twice differentiable with respect to \( \theta \) and let \( \hat{\theta}_n \) be a solution of (2.2). Set

\[
I_n(\theta) = -\bar{X}_n(\theta) \tag{2.5}
\]

\[
J_n(\theta) = E_\theta[I_n(\theta)] \tag{2.6}
\]

and define the conditional variance

\[
\xi_n(\theta) = \sum_{l=1}^{n} E_\theta[u_n^2(\theta)|\mathcal{F}_{n-1}] \tag{2.7}
\]

where

\[
u_n(\theta) = U_n(\theta) - U_{n-1}(\theta) .
\]
Under regularity conditions similar to (2.3)

\[ E_\theta \{ u^2_n(\theta) \} = E_\theta \{ z_n(\theta) \} = \delta_n(\theta) ; \quad (2.8) \]

as discussed later each of (2.5), (2.6) and (2.7) define quantities related to the concept of information.

3. Some Limit Theory

Here we discuss some martingale limit theory and will drop the reference to \( \theta \) in the previously defined quantities. One point of departure is the use of \( \{ z_n \} \) as the appropriate (random) normalizing sequence analogous to \( \{ n \} \) in the independent identically distributed case. This particular choice of normalizing sequence gives both a straightforward analogue of a strong law of large numbers as well as a more generally valid central limit theorem. This last reference is specifically important for the so-called nonergodic situation when, under (2.8),

\[ \frac{z_n}{\delta_n} \overset{p}{\rightarrow} W \quad (3.1) \]

and/or

\[ \frac{I_n}{\delta_n} \overset{p}{\rightarrow} W \quad (3.2) \]

where \( W \) is a nondegenerate positive random variable.

The strong law analogue appears in Neveu (1970, IV-6-2) and a proof is also given in Heyde and Feigin (1975).

**Proposition.** If \( \xi_n \overset{\omega}{\rightarrow} \infty \) a.s. then

\[ \frac{\xi^{-1}_n u}{\delta_n} \overset{0}{\rightarrow} \text{a.s.} \quad (3.3) \]

With respect to central limit results, the literature deals both with the case when (3.1) holds with \( W = 1 \) a.s. (Brown (1971), Scott (1973)) and
when it holds for general positive \( W \) (Hall (1977), Rootzen (1977)). However in both these contexts a Lindeberg condition is required. In essence these results seem to be applicable to the stationary nonergodic process situation and not the general nonergodic situation to be discussed in the sequel. The form of the limit result we seek is

\[
\xi_n^{-1/2} U_n \xrightarrow{\mathcal{D}} N(0,1).
\]

(3.4)

In cases when the Lindeberg condition does not hold it is probably inappropriate to call this a central limit result with this term usually referring to the limit of a normalized sum of small components. However we find in some applications that the result (3.4) can be derived from true central limit results by a random time change argument of the following type.

Suppose \( Y_t/\sqrt{t} \) satisfies a conventional central limit result (as \( t \to \infty \)) and that the representation

\[
U_n = Y_a \xi_n
\]

is valid (or at least asymptotically so) for some constant \( a \). Then under condition (3.1) one may be able to show that not only (3.4) but also

\[
(\xi_n^{-1/2} U_n, \xi_n^{-1} \xi_n) \xrightarrow{\mathcal{D}} (Z, W)
\]

(3.5)

holds, where \( Z \sim N(0,1) \) and is independent of \( W \).

Our discussion of asymptotic inference will be based on (3.5) or something similar being true and we will see important examples of this situation in the next section.

4. Markov Case: Conditional Exponential Families

Before considering more general nonergodic processes we will look at the important Markov case and discuss the theory of conditional exponential
families (CEF's) first discussed in Heyde and Feigin (1975) and further developed in Feigin (1979).

Suppose the $X$ process is a time-homogeneous Markov chain with transition probability density, with respect to a measure $\nu$ on $R$, given by $f(y|x; \theta)$ under $P_\theta$. Then we call $\rho$ a conditional exponential family of probability measures for $X$, or call $\{(X, P_\theta); \theta \in \Theta\}$ a conditional exponential family of stochastic processes, if for relevant $x$,

$$f(y|x; \theta) = b(y, x) \exp\{\alpha(\theta) m(y, x) - \beta(\theta, x)\}.$$  \hspace{1cm} (4.1)

This definition is clearly analogous to that for ordinary exponential families. Suppose $P_\theta(X_0 = x_0) = 1$, then

$$L_n(\theta) = \prod_{i=1}^n b(X_i, x_{i-1}) \exp\{\alpha(\theta) \sum_{i=1}^n m(X_i, x_{i-1}) - \sum_{i=1}^n \beta(\theta, x_{i-1})\}$$

with

$$U_n(\theta) = \sum_{i=1}^n \{\dot{\alpha}(\theta) m(X_i, x_{i-1}) - \dot{\beta}(\theta, x_{i-1})\}.$$  \hspace{1cm} (4.2)

These definitions may be extended to the multiparameter case (Feigin (1979)).

Working with the canonical parametrization $\alpha(\theta) = \theta$, we assume that

$$\Theta \subseteq \Theta_x \equiv \{\theta: |\beta(\theta, x)| = |\log \int b(y, x) e^{\theta m(y, x)} \nu(dy)| < \infty\}$$

for all $x$ in the state space. We consider the following conditions:

C1. $E_\theta \beta(\theta, x_{i-1}) < \infty$

C2. $\Theta \equiv \Theta_x$, $\Theta$ open

C3. $\beta(\theta, x) \equiv \gamma(\theta) h(x), h(x) \geq 0$

C4. $m(\cdot, x)$ is an invertible function for each $x$
C5. \exists a sequence \( C_n(\theta) \uparrow \infty \) such that
\[
C_n^{-1}(\theta) \sum_{i=1}^{n} h(X_{i-1}) \overset{P}{\rightarrow} W(\theta) > 0 \text{ a.s. } [\mathbb{P}_\theta].
\]

C6. \( \tilde{\gamma}(\theta) \) is continuous on \( \Theta \).

The results discussed below are proved in Feigin (1979). First, under C1, \( \{U_n(\theta), \mathcal{F}_n; n \geq 1\} \) is a square integrable \( P_\theta \)-martingale and
\[
\xi_n(\theta) = I_n(\theta) = \sum_{i=1}^{n} \beta(\theta, X_{i-1}). \tag{4.3}
\]
In the case of conditionally additive exponential families (CAEF's), that is when C3 holds, we have that under C1,
\[
U_n(\theta) = H_n \{ \tilde{\gamma}(\hat{\theta}_n) - \tilde{\gamma}(\theta) \} = \xi_n(\theta) \{ \tilde{\gamma}(\hat{\theta}_n) - \tilde{\gamma}(\theta) \} / \tilde{\gamma}(\theta) \tag{4.4}
\]
where \( \hat{\theta}_n \) is the (unique) MLE and \( H_n = \sum_{i=1}^{n} h(X_{i-1}) \). Moreover, under C1 to C6, (3.5) holds with the consequence that \( \hat{\theta}_n \) is strongly consistent and
\[
\xi_n^{1/2}(\theta) (\hat{\theta}_n - \theta) \overset{\mathcal{D}}{\rightarrow} N(0,1). \tag{4.5}
\]
For CAEF's the conditions C1 to C4 ensure that the score function can be represented as a randomly stopped additive process \( Y \) as follows:
\[
\mathcal{L}(\{U_n(\theta)\}) = \mathcal{L}(\{\mathcal{Y}_n(\theta)\}). \tag{4.6}
\]
Not only does this representation lead to (3.5) and (4.5), but it also indicates why \( \{H_n\} \) appropriately fulfills the role of \( \{n\} \) in determining the rate at which "information" is accumulating. Specifically, the quantity in (4.3), and not its expectation, is the appropriate analogue of the Fisher information for the CAEF's (see Section 6).
The only condition that may be difficult to check among C1 to C6 is C5. In examples, we find that recognizing that
\[ V_n(\theta) = \prod_{i=1}^{n} \{ m(X_i, X_i-1)/\beta(\theta, X_i-1) \} \]
is also a $P_\theta$-martingale allows one to check C5 via the martingale convergence theorem.

**Example 1.** Consider the simple branching process $\{ 1 = X_0, X_1, \ldots \}$ with offspring distribution $p(\cdot; \theta)$ parametrized by a scalar parameter $\theta$. Here we are dealing with a discrete state space Markov chain and the transition probabilities (i.e. densities with respect to counting measure) are given by

\[ f(y|x; \theta) = P_\theta(X_n = y|X_{n-1} = x) \]
\[ = \sum_{\{v_1:v_1+\cdots+v_x=y\}} \prod_{j=1}^{x} p(v_j; \theta) \]
\[ = \sum_{\{v_1:v_1+\cdots+v_x=y\}} \prod_{j=1}^{x} \log p(v_j; \theta) \quad . \quad (4.7) \]

Comparing (4.7) and (4.1) we may conclude that these branching process models form a conditional exponential family if and only if the part of the exponent in (4.7) which involves $\theta$ is common to all the summands. In other words, $\sum_{\{v_1:v_1+\cdots+v_x=y\}} \prod_{j=1}^{x} \log p(v_j; \theta)$ must have the form

\[ \log p(v; \theta) = \alpha(\theta)v - \lambda(\theta) + \log \kappa(v) \quad (4.8) \]

whereupon $m(y,x) = y$, $\beta(\theta,x) = \lambda(\theta)x$ and

\[ b(y,x) = \sum_{\{v_1:v_1+\cdots+v_x=y\}} \prod_{j=1}^{x} \kappa(v_j) \]
Equation (4.8) corresponds exactly to the power series offspring distributions. We conclude, therefore, that the simple branching process models form a CAEF if and only if the offspring distribution is a one-parameter discrete exponential family or power series distribution—see Johnson and Kotz (1969). Moreover, the separability condition C3 always holds for these models.

This result was obtained in Heyde and Feigin (1975) with a more restrictive definition of CEF's.

**Example 2.** We let \( \{0 = X_0, X_1, \ldots \} \) be a generalized first order autoregressive process satisfying

\[
\Psi(X_i) = \Phi(\theta, X_{i-1}) + \eta_i(X_{i-1})
\]

(4.9)

where the \( \eta_i(X_{i-1}) \) are random variables whose distribution conditional on \( X_{i-1} \) may depend on \( X_{i-1} \). First we note that given \( X_{i-1} = x \),

\( \Phi = \Phi(\theta, x) \) is a location parameter for the distribution of \( \Psi(X_i) \). From a paper of Lindley (1958) we find that the only exponential family possibilities for the conditional distribution of \( X_i \) obtain when either

(i) \( \eta_i \sim N(0, \sigma^2(x)) \); \( \Psi(y) = y \)

or

(ii) \( \exp(\eta_i) \sim \text{gamma}(\lambda(x), 1) \); \( \Psi(y) = \log y \) where a \( \text{gamma}(\lambda, 1) \) variate has density \( u^{\lambda-1} e^{-u}/\Gamma(\lambda) \). Note that (i) corresponds to an additive model while (ii) to a multiplicative one.

If we further require that the conditional exponential form (4.1) is to hold for \( f(y|x; \theta) \) we see that in case (i) \( \Phi(\theta, x) = r(\theta)s(x) \) and in (ii) \( \Phi(\theta, x) = R(\theta) + S(x) \). For example, in (ii) we have, writing \( \lambda = \lambda(x) \),
\[ X_i = e^{\phi(\theta, x_{i-1})} \varepsilon_i ; \quad \varepsilon_i \sim \text{gamma}(\lambda, 1) \]  

(4.10)

and

\[ f(y|x; \theta) = e^{-\lambda \phi(\theta, x)} y^{\lambda-1} e^{-\lambda S(x)}/\Gamma(\lambda) \]  

(4.11)

which has the form (4.1) if and only if \( \phi(\theta, x) = R(\theta) + S(x) \) whereupon

\[ m(y, x) = ye^{-S(x)} , \quad \alpha(\theta) = e^{-R(\theta)} , \]

(4.12)

\[ \beta(\theta, x) = \lambda(x) R(\theta) , \quad b(y, x) = y^{\lambda-1} e^{-\lambda S(x)} . \]

For \( \theta \) to be identifiable, \( R(\theta) \) must be single-valued. We note again that we are in the CAEF case and the conditions Cl to C6 obtain when, for example, \( \lambda(x) = x \) as shown in Feigin (1979).

Heyde and Feigin (1975), when restricting attention to the ordinary autoregressive model, found that the Gaussian error structure was the only one consistent with their more limited definition of conditional exponential families.

5. Regular Nonergodic Processes

Recognizing that the conditional exponential families form a quite restricted class, several authors have looked at more general structures. Below we pick out the underlying ideas which are referred to in the works of Basawa (1977), Basawa and Scott (1977, 1978), Davies (1978), Basawa and Koul (1979) and Heyde (1978). Some of these authors deal with the vector case which may involve some extra complications over the scalar case – see Basawa and Koul (1979), Davies (1978).

Writing \( \Lambda_n(\theta_1, \theta_2) = \mathcal{L}_n(\theta_2) - \mathcal{L}_n(\theta_1) \) the aim is to find an asymptotic expansion of this log-likelihood ratio which is valid for some appropriate sequence \( \{ \Lambda_n(\theta, \theta + k_n^{-1}(\theta)h) \} \) determined by suitably choosing \( k_n(\theta) \). Explicitly, the requirement is:
A. There exists a sequence \( \{K_n(\theta)\} \uparrow \infty \) such that

\[
\Lambda_n(\theta, \theta_n) = h'\Delta_n(\theta) - \frac{1}{2} h' T_n(\theta) h + R_n(\theta, h) \tag{5.1}
\]

where

\[
\theta_n = \theta + K_n^{-1}(\theta) h, \quad R_n(\theta, h) \xrightarrow{p} 0 [p_n, \theta]. \tag{5.2}
\]

In order to use this expansion successfully one needs a satisfactory limiting behavior for the variables \((\Delta_n(\theta), T_n(\theta))\):

B. \((\Delta_n(\theta), T_n(\theta)) \xrightarrow{\mathcal{D}} (\Delta(\theta), T(\theta)) [p_n, \theta] \tag{5.3}\)

We have written the equations (5.1), (5.2) and (5.3) so that they may also be easily interpreted if \( \theta \) is a vector. One particular form in (5.3) is:

C. \( \Delta(\theta) \sim T^{1/2}(\theta) Z; \quad Z \sim N(0, I) \) independent of \( T(\theta) \).

We will call the family a regular nonergodic family if A, B and C all hold. Typically we have

\[
\Delta_n(\theta) = K_n^{-1}(\theta) L_n(\theta) \tag{5.4}
\]

\[
T_n(\theta) = K_n^{-1}(\theta) T_n(\theta) K_n^{-1}(\theta) \tag{5.5}
\]

and (5.5) gives us the key to selecting \( K_n(\theta) \) if (5.3) is to hold. In fact, it is clear that

\[
K_n(\theta) = J_n^{1/2}(\theta)
\]

is a very plausible candidate since it ensures that \( T_n(\theta) \) will have constant expectation \( I \) (unit matrix). Actually, in the scalar case there are no problems with this determination, however for the vector
case checking (5.3) is made easier in practice by choosing \( K_n(\theta) \) to be a diagonal matrix (e.g. the matrix with the square root of the diagonal elements of \( J_n(\theta) \) on its diagonal).

Returning to the scalar case we see that for it to be regular non-ergodic it is sufficient that \( \lambda_n(\theta) \) be twice differentiable and that, when \( \theta_n^* \) is between \( \theta_n \) and \( \theta \),

\[
(1) \quad J_n^{-1}(\theta) \left[ I_n(\theta_n^*) - I_n(\theta) \right] \xrightarrow{p} 0 \quad [P_{n, \theta}]
\]

and

\[
(ii) \quad (J_n^{-1/2}(\theta) \bar{I}_n(\theta), J_n^{-1}(\theta) I_n(\theta)) \xrightarrow{\mathcal{D}} (W^{1/2}(\theta)Z, W(\theta))
\]

hold. While (5.6) is basically a continuity condition, (5.7) often holds due to limit theorems such as those outlined in Section 3 (equation (3.5)) and Section 4.

We indicate briefly some possible conclusions that can be reached for inference in regular nonergodic processes. The first approach is based on establishing the contiguity of \( \{P_{n, \theta}^*\} \) and \( \{P_{n, \theta}\} \), and then showing optimality of estimators by probability of concentration calculations or by computing asymptotic powers (Feigin (1978)) and distributions (Basawa and Koul (1979)) of test statistics. A second approach adopted by Heyde (1978) is to impose continuous convergence in (5.7) to obtain probabilities of concentration asymptotically. However we note that the contiguity follows immediately from A, B and C since it is not hard to show that \( E e^{\Lambda} = 1 \) where \( \Lambda \) is a random variable with distribution equal to the limiting distribution of \( \{A_n(\theta, \theta_n)\} \) under \( \{P_{n, \theta}\} \) (see Feigin (1978), for example).
The results in Heyde (1978) concerning probabilities of concentration show that for symmetric intervals about the origin \( K_n(\theta) (\hat{\theta}_n - \theta) \) has the largest asymptotic such probability against all estimates \( \{\hat{\nu}_n\} \) satisfying the condition

\[
\lim\{P \hat{\nu}_n \text{ } (J_n^{1/2}(\theta)(\hat{\theta}_n - \theta) > -\frac{1}{2} h) - P \hat{\nu}_n \text{ } (J_n^{1/2}(\theta)(\hat{\theta}_n - \theta) > -\frac{1}{2} h)\} = 0
\]

when \( K_n(\theta) = J_n^{1/2}(\theta) \) in (5.2). The result for nonsymmetric intervals is more complicated, although conditional probability statements may be made.

We note that in the conditionally additive exponential family case (5.6) and (5.7) follow from conditions C1 to C6.

6. Implications for Inference

Starting from the more classical notions we look first at an analogue of Fisher's information. In the CAEF case we saw that both the conditional variance of the score function, \( \xi_n(\theta) \), or equivalently the observed information \( I_n(\theta) \), seemed to be logical analogues of Fisher's information. If we want to choose a quantity that will reduce to the ordinary Fisher information in the independent observation case then \( \xi_n(\theta) \) is the appropriate choice and one that can be used generally for stochastic processes. This suggestion was made by Heyde and Feigin (1975) where it was also suggested that this measure provided the appropriate norming for \( (\hat{\theta}_n - \theta) \) in order to obtain asymptotic normality. On the other hand, recommendations for the general use of the observed rather than the expected information (Efron and Hinkley (1978)) would of course lead to regarding \( I_n(\theta) \) as an appropriate measure of information for stochastic processes.
However, the asymptotic normality does not unequivocally resolve questions of how to conduct inference based on asymptotic properties of the MLE. For example, suppose we wish to test $H_0: \theta = \theta_0$ against $H_1: \theta > \theta_0$, and that (5.6) and (5.7) hold. A Taylor expansion argument will reveal that under $H_0$

\begin{equation}
I_n^{1/2}(\theta_0) (\hat{\theta}_n - \theta_0) \overset{D}{\to} N(0,1)
\end{equation}

\begin{equation}
J_n^{-1/2}(\theta_0) I_n(\theta_0)(\hat{\theta}_n - \theta_0) \overset{D}{\to} W^{-1/2}(\theta_0)Z
\end{equation}

and so a different critical region will be obtained by considering each of the three possible limiting distributions when $W$ is nondegenerate; that is, when we are in a strictly nonergodic situation. An attempt to justify that test based on the asymptotic normal distribution was made by considering the Pitman efficiency (Feigin (1978), Sweeting (1978)), but a more convincing approach is one based on an asymptotic ancillarity approach which was suggested in Sweeting (1978) and investigated in Feigin and Reiser (1979).

The basic idea is that, asymptotically, $I_n^{1/2}(\theta)(\hat{\theta}_n - \theta)$ is pivotal since its asymptotic distribution does not involve $\theta$, whereas $J_n^{-1}(\theta)I_n(\theta)$ is asymptotically almost ancillary - it converges to a random variable $W$ with mean 1, but whose distribution may still depend on $\theta$. This would suggest that inference might be conducted conditionally on $J_n^{-1}(\theta)I_n(\theta)$ since it is the finite $n$ substitute for $W$, and by invoking the asymptotic independence of it and the pivotal $I_n^{1/2}(\theta)(\hat{\theta}_n - \theta)$ one may use the latter's asymptotic normality for hypothesis testing and making
confidence statements. These arguments require some more formal justification which is given in Feigin and Reiser (1979); however the idea to treat the finite \( n \) situation as if the limiting joint distribution applied is an axiom of the approach, as is the case for much of classical inference practice.

A further interesting aspect is the application of the notions of curvature and approximate ancillarity as described in Efron (1975) and Efron and Hinkley (1978). First of all, the curvature \( \gamma_n(\theta) \) is given by

\[
\gamma_n^2(\theta) = \{1 - \rho_n^2(\theta)\} \operatorname{var}\{\hat{\theta}_n^{-1}(\theta) I_n(\theta)\}
\]

where \( \rho_n^2(\theta) \) is the correlation between \( \hat{\theta}_n(\theta) \) and \( I_n(\theta) \). From (5.7), it is plausible that \( \rho_n^2(\theta) \to 0 \) since \( W^{1/2}Z \) and \( W \) are uncorrelated. Under a uniform integrability assumption, therefore, we will have \( \gamma_n^2(\theta) \to \operatorname{var}\{W(\theta)\} \). This result, due to Basawa (1977), has intuitive appeal because it shows that the asymptotic curvature is nonzero as a consequence of the nondegeneracy of \( W \). This nonzero curvature of the asymptotic model suggests that even asymptotically there is something to be gained by conditioning on an approximate ancillary - an idea due to Efron and Hinkley (1978).

In fact, if we use their suggestion for constructing a suitable ancillary, we obtain

\[
Q_n = \{1 - I_n(\hat{\theta}_n)/J_n(\hat{\theta}_n)\}/\gamma_n(\hat{\theta}_n)
\]

which will have, under some continuity assumptions, the asymptotic distribution of \((1-W)/\{\operatorname{var} W\}^{1/2}\). This distribution has mean 0 and variance 1 and so is "more ancillary" than \( W \) itself. Moreover, if
we use the asymptotic distributions given in (5.7), conditioning on \( Q_n \), or conditioning on \( J_n^{-1} I_n \) directly, lead to the same conclusions - for large \( n \), inference should be based on the asymptotic normality of \( I_n^{1/2}(\hat{\theta}_n - \theta) \). For a more formal presentation of these arguments we refer the reader to Feigin and Reiser (1979).

If we take a different approach and consider working with a posterior distribution for \( \theta \) which is based on a prior having a continuous nonzero density in a neighborhood of the true value \( \theta_0 \), then Heyde and Johnstone (1979) have shown that the posterior will be asymptotically normal with mean \( \hat{\theta}_n \) and variance \( I_n(\hat{\theta}_n) \). If we use these asymptotics to construct a Bayesian confidence interval for \( \theta \) we find that it will correspond to the classical type based on the asymptotic normality of \( I_n^{1/2}(\hat{\theta}_n)(\hat{\theta}_n - \theta) \).

It seems that both the conditional and the Bayesian approaches point to the same asymptotic inference for regular nonergodic stochastic processes - a somewhat more reassuring situation than that following the first attempts at sorting out the questions associated with the three asymptotic distributions in (6.1).

Acknowledgements

I acknowledge with thanks the receipt of preprints from I. V. Basawa, R. Davies, C. C. Heyde, I. M. Johnstone, H. L. Koul, and D. J. Scott.
References


