D. N. Shanbhag and C. R. Rao, eds., *Handbook of Statistics, Vol. 19*

Inference in Stochastic Processes

I. V. Basawa

Large sample properties of estimators and test statistics based on observations from stochastic processes are reviewed. The local asymptotic normality (LAN) is used as a unifying framework. Optimum estimating functions, adaptive estimation for semiparametric models and Bayesian methods are also discussed briefly. Several examples from stochastic processes are presented to illustrate the theory.

1. Introduction

This paper reviews large sample properties of estimators and test statistics based on observations from discrete time stochastic processes. Even though similar results can be obtained for continuous time processes, we limit ourselves to the discrete time for the ease of presentation. The local asymptotic normality (and mixed normality) is used to unify diverse asymptotic results and efficiency properties. Billingsley (1961), Basawa and Prakasa Rao (1980a) and Basawa and Scott (1983) may be consulted for background material for inference in stochastic processes. LeCam (1986) and LeCam and Yang (1990) discuss local asymptotic normality and its applications in a general setting.

If the likelihood function is not known, one can use the theory of optimal estimating functions instead of the likelihood based methods. See Godambe (1991) and Heyde (1997) for the method of estimating functions and its applications.

Section 2 is concerned with likelihood based methods. These include the local asymptotic normality, asymptotic efficiency of the maximum likelihood estimator, efficient tests of both simple and composite hypotheses, and extensions to local asymptotic mixed normality. Optimal estimating functions are introduced in Section 3. Semiparametric models and adaptive estimation are discussed in Section 4. Section 5 reviews Bayes and empirical Bayes methods. Specific applications are discussed in Section 6.

See also Basawa and Prakasa Rao (1980b) and Basawa (1983, 1990) for previous reviews on the topic.

2. Likelihood methods

2.1. The basic framework

Let $\{X_t\}$, $t = 0, \pm 1, \pm 2, \ldots$, denote a discrete time stochastic process defined on a probability space $(\chi, \mathscr{F}, P_\theta)$ where χ is the sample space not depending on θ, $\mathscr{F}$, the corresponding Borel σ-field and P_θ a probability measure indexed by a $(k \times 1)$ vector parameter θ taking values in an open set $\Omega \subset \mathbb{R}^k$. Suppose $\mathbf{X}(n) = (X_1, \ldots, X_n)$ is a vector of n observations defined on the space $(\chi_n, \mathscr{F}_n, P_{n,\theta})$. Let $p_n(x(n); \theta)$ denote the probability density corresponding to $P_{n,\theta}$ defined with respect to an appropriate measure μ_n. It is assumed that the probability measures $\{P_{n,\theta}, \theta \in \Omega\}$ are restrictions of P_θ on $\mathscr{F}_n$, and they are mutually absolutely continuous. Consider the log-likelihood ratio of θ to θ_0,

$$\Lambda_n(\theta, \theta_0) = \log\{p_n(\mathbf{X}(n); \theta)/p_n(\mathbf{X}(n); \theta_0)\} \ , \tag{2.1}$$

where θ_0 is a fixed parameter value and θ ranges in Ω. The asymptotic properties of likelihood based estimators and tests are related crucially to the limiting behaviour of $\Lambda_n(\theta, \theta_0)$ for values of θ close to θ_0. Define a neighborhood $N_n(\theta_0)$ of θ_0 by $N_n(\theta_0) = \{\theta : |\mathbf{C}_n^{-1}(\theta_0)(\theta - \theta_0)| \leq \delta\}$, $\delta > 0$, where $|\mathbf{a}|$, for any column vector $\mathbf{a}$, denotes the vector norm $(\mathbf{a}^{\mathrm{T}}\mathbf{a})^{1/2}$, and $\mathbf{C}_n(\theta_0)$ is a $(k \times k)$ positive definite symmetric matrix (non-random) such that $tr\{\mathbf{C}_n^{\mathrm{T}}(\theta_0)\mathbf{C}_n(\theta_0)\} \to \infty$ as $n \to \infty$. We shall first consider a quadratic approximation for $\Lambda_n(\theta, \theta_0)$ for $\theta \in N_n(\theta_0)$. See the end of Section 2.3 for the various choices of $\mathbf{C}_n(\theta_0)$.

2.2. A quadratic approximation for the log-likelihood ratio

Suppose that for any $\theta \in N_n(\theta_0)$, there exists a random $(k \times 1)$ vector $\mathbf{S}_n(\theta_0)$, and a random $(k \times k)$ (almost surely) positive definite symmetric matrix $\mathbf{\Gamma}_n(\theta_0)$ such that under P_{θ_0}-probability,

$$\Lambda_n(\theta, \theta_0) = (\theta - \theta_0)^{\mathrm{T}}\mathbf{S}_n(\theta_0) - \tfrac{1}{2}(\theta - \theta_0)^{\mathrm{T}}\mathbf{\Gamma}_n(\theta_0)(\theta - \theta_0) + o_p(1) \ , \tag{2.2}$$

where $o_p(1)$ denotes terms that converge to zero as $n \to \infty$ in P_{θ_0}-probability. The quadratic approximation in (2.2) will be used repeatedly in what follows. Denote

$$\mathbf{Q}_n(\theta, \theta_0) = (\theta - \theta_0)^{\mathrm{T}}\mathbf{S}_n(\theta_0) - \tfrac{1}{2}(\theta - \theta_0)^{\mathrm{T}}\mathbf{\Gamma}_n(\theta_0)(\theta - \theta_0) \ . \tag{2.3}$$

Taking vector derivatives with respect to θ, we have

$$\mathbf{Q}_n'(\theta, \theta_0) = \mathbf{S}_n(\theta_0) - \mathbf{\Gamma}_n(\theta_0)(\theta - \theta_0) \ , \tag{2.4}$$

and

$$\mathbf{Q}_n''(\theta, \theta_0) = -\mathbf{\Gamma}_n(\theta_0) \ . \tag{2.5}$$

Consequently, the value of θ that maximizes $\mathbf{Q}_n(\theta, \theta_0)$ is given by

$$\hat{\theta}_0 = \theta_0 + \mathbf{\Gamma}_n^{-1}(\theta_0)\mathbf{S}_n(\theta_0) \ . \tag{2.6}$$

Substituting (2.6) in (2.3) we have

$$\mathbf{Q}_n(\hat{\theta}_0, \theta_0) = \tfrac{1}{2}\mathbf{S}_n^{\mathrm{T}}(\theta_0)\boldsymbol{\Gamma}_n^{-1}(\theta_0)\mathbf{S}_n(\theta_0) \ . \tag{2.7}$$

The maximizer $\hat{\theta}_0$ and the maximum value $\mathbf{Q}_n(\hat{\theta}_0, \theta_0)$ of $\mathbf{Q}_n(\theta, \theta_0)$ play an important role in obtaining efficient estimators and efficient tests respectively.

It may be noted that in most of the applications the approximation in (2.2) can be verified via the Taylor expansion with

$$\mathbf{S}_n(\theta) = \frac{\mathrm{d} \log p_n(\mathbf{X}(n); \theta)}{\mathrm{d}\theta}, \text{ and } \boldsymbol{\Gamma}_n(\theta) = \frac{-\mathrm{d}^2 \log p_n(\mathbf{X}(n); \theta)}{\mathrm{d}\theta \, \mathrm{d}\theta^{\mathrm{T}}} \ , \tag{2.8}$$

the score vector and the sample Fisher information matrix, respectively. From now on, unless otherwise stated, $\mathbf{S}_n(\theta)$ and $\boldsymbol{\Gamma}_n(\theta)$ are chosen as in (2.8). See Basawa and Koul (1988) for further results on quadratic approximation in a more general setting.

2.3. The Local Asymptotic Normality (LAN)

Suppose first that the quadratic approximation for $\Lambda_n(\theta, \theta_0)$ in (2.2) is valid. Further, assume the following conditions, (2.9) and (2.10), are satisfied. There exists a positive definite symmetric non-random matrix $\boldsymbol{\Gamma}(\theta_0)$ such that

$$\mathbf{C}_n^{-1}(\theta_0)\boldsymbol{\Gamma}_n(\theta_0)\mathbf{C}_n^{-1}(\theta_0) = \boldsymbol{\Gamma}(\theta_0) + o_p(1) \ , \tag{2.9}$$

and

$$\mathbf{C}_n^{-1}(\theta_0)\mathbf{S}_n(\theta_0) \xrightarrow{d} N_k(0, \boldsymbol{\Gamma}(\theta_0)) \tag{2.10}$$

both under P_{θ_0}-probability. The family of probability measures $\{P_{n,\theta}\}$ is said to satisfy the local asymptotic normality (LAN) property in $N_n(\theta_0)$ provided (2.2), (2.9) and (2.10) are satisfied. Under the LAN assumption we have, with $\theta_n = \theta_0 + \mathbf{C}_n^{-1}(\theta_0)\mathbf{h}$, where $\mathbf{h}$ is a $(k \times 1)$ vector of real numbers,

$$\Lambda_n(\theta_n, \theta_0) \xrightarrow{d} N(-\tfrac{1}{2}\mathbf{h}^{\mathrm{T}}\boldsymbol{\Gamma}(\theta_0)\mathbf{h}, \mathbf{h}^{\mathrm{T}}\boldsymbol{\Gamma}(\theta_0)\mathbf{h}) \ , \tag{2.11}$$

as $n \to \infty$, under P_{θ_0}-probability.

The LAN property provides a powerful tool in obtaining efficient estimators and tests as will be seen in the next two subsections. The normalizing sequence of matrices $\{\mathbf{C}_n(\theta_0)\}$ can be chosen in a number of ways. More common choices are as follows:

(i) Take $\mathbf{C}_n(\theta_0) = n^{1/2}\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. This choice is appropriate typically when the process $\{X_t\}$ is stationary and ergodic. In particular, when $\{X_t\}$ is a sequence of independent and identically distributed random variables, the above choice leads to the classical large sample theory.

(ii) For nonstationary processes, it is often convenient to take

$$\mathbf{C}_n(\theta_0) = \left[\operatorname{diag}\left\{E_{\theta_0}\left(\frac{-\partial^2 \ln p_n}{\partial\theta_1^2}\right), \ldots, E_{\theta_0}\left(\frac{-\partial^2 \ln p_n}{\partial\theta_k^2}\right)\right\}\right]^{1/2} .$$

Typically, in problems related to regression models the above choice is quite common in the literature.

(iii) A more general choice for $\mathbf{C}_n(\theta_0)$ is $\mathbf{C}_n(\theta_0) = (E_{\theta_0}\boldsymbol{\Gamma}_n(\theta_0))^{1/2}$. With this choice, for ergodic type models typically, $\boldsymbol{\Gamma}(\theta_0)$ in (2.9) to (2.11) is replaced by the identity matrix.

2.4. *Efficient estimation*

A sequence of estimators $\{T_n\}$ of θ is said to be **regular** asymptotically normal if

$$\mathbf{C}_n(\theta_0)(T_n - \theta_n) \xrightarrow{d} N_k(0, \mathbf{V}_T(\theta_0)), \quad \text{under } P_{n,\theta_n}\text{-probability} , \tag{2.12}$$

where $\theta_n = \theta_0 + \mathbf{C}_n^{-1}(\theta_0)\mathbf{h}$, and $\mathbf{V}_T(\theta_0)$ is a positive definite matrix. The LAN property enables one to construct efficient estimators in the class of regular asymptotically normal estimators satisfying (2.12). Several estimators, such as moment, least squares and conditional least squares estimators satisfy (2.12). It can be shown (see, for instance, Hall and Mathiason (1990)) that when the LAN property is satisfied, then for any T_n satisfying (2.12), we have

$$\mathbf{V}_T(\theta_0) \geq \boldsymbol{\Gamma}^{-1}(\theta_0) , \tag{2.13}$$

in the sense that the difference is a non-negative definite matrix. The inequality in (2.13) is the asymptotic analogue of the usual Cramér–Rao inequality for the variance of an unbiased estimator based on a finite sample. Note that the regularity assumption in (2.12) requires the asymptotic normality of T_n under P_{n,θ_n}-probability. The verification of (2.12) can be simplified as follows. Suppose first that $\{T_n\}$ satisfies

$$\begin{pmatrix} \mathbf{Z}_n \\ \boldsymbol{\Delta}_n \end{pmatrix} \xrightarrow{d} N_{2k}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{V}_T & \boldsymbol{\sigma}_{T,S} \\ \boldsymbol{\sigma}_{T,S} & \boldsymbol{\Gamma} \end{pmatrix}\right), \quad \text{under } P_{\theta_0}\text{-probability} , \tag{2.14}$$

where $\mathbf{Z}_n = \mathbf{C}_n(\theta_0)(T_n - \theta_0)$ and $\boldsymbol{\Delta}_n = \mathbf{C}_n^{-1}(\theta_0)\mathbf{S}_n(\theta_0)$. It can then be shown (see Hall and Mathiason (1990)) that, under LAN, any T_n satisfying (2.14) satisfies (2.12) if and only if $\boldsymbol{\sigma}_{T,S} = \mathbf{I}$, the $(k \times k)$ identity matrix. Consequently, in order to verify (2.12) it suffices to show that

$$\begin{pmatrix} \mathbf{Z}_n \\ \boldsymbol{\Delta}_n \end{pmatrix} \xrightarrow{d} N_{2k}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{V}_T & \mathbf{I} \\ \mathbf{I} & \boldsymbol{\Gamma} \end{pmatrix}\right), \quad \text{under } P_{\theta_0}\text{-probability} . \tag{2.15}$$

Note that under appropriate regularity conditions regarding differentiation under the integral sign, the requirement $\boldsymbol{\sigma}_{T,S} = \mathbf{I}$ implies that T_n is asymptotically

unbiased for θ. Suppose that T_n satisfies (2.15). Then (2.13) follows readily. To see this consider $Y_n = \mathbf{Z}_n - \boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}_n$. From (2.15) it follows that, under P_{θ_0}-probability,

$$Y_n \xrightarrow{d} N_k(0, \mathbf{V}_T - \boldsymbol{\Gamma}^{-1}) \ . \tag{2.16}$$

Since $\mathbf{V}_T - \boldsymbol{\Gamma}^{-1}$ is a covariance matrix, the result in (2.13) follows.

A regular estimator is asymptotically efficient if the equality in (2.13) holds. We now consider the problem of constructing efficient estimators which attain the equality in (2.13). First, consider the following assumption on the score function $\mathbf{S}_n(\theta)$. Suppose

$$\mathbf{S}_n(\theta) = \mathbf{S}_n(\theta_0) - \mathbf{C}_n(\theta_0)\boldsymbol{\Gamma}\mathbf{C}_n(\theta_0)(\theta - \theta_0) + o_p(1), \quad \text{under } P_{\theta_0}\text{-probability} \ , \tag{2.17}$$

uniformly in $\theta \in N_n(\theta_0)$. The requirement in (2.17) can usually be verified by a Taylor expansion of $\mathbf{S}_n(\theta)$ at θ_0, and a strengthened version of (2.9), viz.,

$$\mathbf{C}_n^{-1}(\theta_0)\boldsymbol{\Gamma}_n(\theta)\mathbf{C}_n^{-1}(\theta_0) = \boldsymbol{\Gamma}(\theta_0) + o_p(1), \quad \text{under } P_{\theta_0}\text{-probability} \ , \tag{2.18}$$

uniformly in $\theta \in N_n(\theta_0)$. Now consider the estimator $\hat{\theta}_n$ defined by

$$\hat{\theta}_n = \tilde{\theta}_0 + \boldsymbol{\Gamma}_n^{-1}(\tilde{\theta}_0)\mathbf{S}_n(\tilde{\theta}_0) \ , \tag{2.19}$$

where $\tilde{\theta}_0$ is any preliminary estimator of θ such that

$$\mathbf{C}_n(\theta_0)(\tilde{\theta}_0 - \theta_0) = O_p(1), \quad \text{under } P_{\theta_0}\text{-probability} \ , \tag{2.20}$$

where $O_p(1)$ denotes terms bounded in probability. Note that (2.19) is obtained from (2.6) by replacing θ_0 by $\tilde{\theta}_0$. Under LAN, and (2.17), one can verify that $\hat{\theta}_n$ is asymptotically efficient. First, we shall show that

$$\mathbf{C}_n(\theta_0)(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_k(0, \boldsymbol{\Gamma}^{-1}(\theta_0)), \quad \text{under } P_{\theta_0}\text{-probability} \ . \tag{2.21}$$

We have

$$\begin{aligned}
\mathbf{C}_n(\theta_0)(\hat{\theta}_n - \theta_0) &= \mathbf{C}_n(\theta_0)(\tilde{\theta}_0 - \theta_0) + \mathbf{C}_n(\theta_0)\boldsymbol{\Gamma}_n^{-1}(\tilde{\theta}_0)\mathbf{S}_n(\tilde{\theta}_0), \text{ by (2.19)} \\
&= \mathbf{C}_n(\theta_0)(\tilde{\theta}_0 - \theta_0) + \mathbf{C}_n(\theta_0)\boldsymbol{\Gamma}_n^{-1}(\tilde{\theta}_0) \\
&\quad \times \{\mathbf{S}_n(\theta_0) - \mathbf{C}_n(\theta_0)\boldsymbol{\Gamma}\mathbf{C}_n(\theta_0)(\tilde{\theta}_0 - \theta_0) + o_p(1)\}, \text{ by (2.17)} \\
&= \boldsymbol{\Gamma}^{-1}(\theta_0)\mathbf{C}_n^{-1}(\theta_0)\mathbf{S}_n(\theta_0) + o_p(1) \ ,
\end{aligned} \tag{2.22}$$

using (2.18). The result in (2.21) then finally follows from (2.22) and (2.10). It can further be shown (see, for instance, Hall and Mathiason (1990)) that under LAN, $\hat{\theta}_n$ satisfies (2.12) with $\mathbf{V}_T(\theta_0) = \boldsymbol{\Gamma}^{-1}(\theta_0)$. Consequently, $\hat{\theta}_n$ given by (2.19) is asymptotically efficient.

Note that $\hat{\theta}_n$ is the usual one-step solution of the likelihood equation $\mathbf{S}_n(\theta) = 0$, and it requires a preliminary estimator. Typically, moment and least squares estimators can be chosen as preliminary estimators.

2.5. Efficient tests: Simple hypotheses

First, consider the problem of testing a simple hypothesis $H : \theta = \theta_0$, against a simple alternative $K : \theta = \theta_1$. By the Neyman–Pearson lemma, the most powerful test is given by

$$\phi_n = \begin{cases} 1, & \Lambda_n(\theta_1, \theta_0) \geq k_n \\ 0, & \Lambda_n(\theta_1, \theta_0) < k_n \end{cases} , \tag{2.23}$$

where the constant k_n is chosen so that $E_H \phi_n = \alpha_n$, $0 < \alpha_n < 1$. Here ϕ_n takes the value 1 for the rejection of H and the value 0 for the acceptance. Let ϕ_n^* be any other test function such that $E_H \phi_n^* = \alpha_n$. It then follows by the Neyman–Pearson lemma that

$$\beta_{\phi_n}(\theta_1) \geq \beta_{\phi_n^*}(\theta_1) , \tag{2.24}$$

where $\beta_\phi(\theta_1)$ denotes the power of a test ϕ at θ_1, i.e. $\beta_\phi(\theta_1) = E_K \phi$. A test ϕ_n is said to be **consistent** for testing H against K, if $\beta_{\phi_n}(\theta_1) \to 1$ as $n \to \infty$. Several reasonable tests satisfy the consistency requirement. In order to discriminate between consistent tests one may look at the limiting power at a sequence of alternatives $K_n : \theta = \theta_n$, $\theta_n = \theta_0 + \mathbf{C}_n(\theta_0)\mathbf{h}$, rather than at a fixed alternative $\theta = \theta_1$. From (2.24), with θ_1 replaced by θ_n, we have

$$\limsup\{\beta_{\phi_n^*}(\theta_n)\} \leq \lim \beta_{\phi_n}(\theta_n) , \tag{2.25}$$

where the right hand limit in (2.25) can be evaluated under LAN. Note that, under LAN, the limit distribution of $\Lambda_n(\theta_n, \theta_0)$ under H is given by (2.11), viz.,

$$\Lambda_n(\theta_n, \theta_0) \xrightarrow{d} N(-\tfrac{1}{2}\tau^2, \tau^2) , \tag{2.26}$$

where $\tau^2 = \mathbf{h}^{\mathrm{T}} \boldsymbol{\Gamma}(\theta_0)\mathbf{h}$. Letting $\alpha_n \to \alpha$, we find by (2.26) that k_n in (2.23) converges to $k = \tau z_{1-\alpha} - \tau^2/2$, where $\Phi(z_{1-\alpha}) = 1 - \alpha$, and $\Phi(x)$ denotes the distribution function of a standard normal random variable. It can be shown (see, for instance, Basawa and Scott (1983)), under LAN, that

$$\Lambda_n(\theta_n, \theta_0) \xrightarrow{d} N(\tau^2/2, \tau^2), \quad \text{under } K_n . \tag{2.27}$$

We have

$$\begin{aligned} \beta_{\phi_n}(\theta_n) &= P_{n,\theta_n}(\Lambda_n \geq k_n) \\ &\to 1 - \Phi(z_{1-\alpha} - \tau) . \end{aligned} \tag{2.28}$$

The result in (2.28) follows from (2.27), and the fact that $k_n \to \tau z_{1-\alpha} - \tau^2/2$. We finally obtain (under LAN) the following inequality for the limiting power of any size-α_n test ϕ_n^*, $(\alpha_n \to \alpha)$,

$$\lim \sup\{\beta_{\phi_n^*}(\theta_n)\} \leq 1 - \Phi(z_{1-\alpha} - \tau) . \tag{2.29}$$

Any test ϕ_n^* for which the equality in (2.29) is attained is an asymptotically efficient test. Obviously, the Neyman Pearson test statistic $\Lambda_n(\theta_n, \theta_0)$ is asymptotically efficient in the above sense.

For a general alternative hypothesis $K : \theta \neq \theta_0$, it is desirable to find a test which does not depend on the specific direction $\mathbf{h}$ in θ_n. One may consider the score statistic $2\mathbf{Q}_n(\hat{\theta}_0, \theta_0)$ defined by (2.7), i.e.,

$$2\mathbf{Q}_n(\hat{\theta}_0, \theta_0) = \mathbf{S}_n^{\mathrm{T}}(\theta_0)\boldsymbol{\Gamma}_n^{-1}(\theta_0)\mathbf{S}_n(\theta_0) \ .$$

We have

$$\begin{aligned}\mathbf{S}_n^{\mathrm{T}}(\theta_0)\boldsymbol{\Gamma}_n^{-1}(\theta_0)\mathbf{S}_n(\theta_0) &= \boldsymbol{\Delta}_n^{\mathrm{T}}(\theta_0)\{\mathbf{C}_n^{-1}(\theta_0)\boldsymbol{\Gamma}_n(\theta_0)\mathbf{C}_n^{-1}(\theta_0)\}^{-1}\boldsymbol{\Delta}_n(\theta_0) \\ &\xrightarrow{d} \chi^2(k), \quad \text{under } H \ , \end{aligned} \tag{2.30}$$

where $\boldsymbol{\Delta}_n(\theta_0) = \mathbf{C}_n^{-1}(\theta_0)\mathbf{S}_n(\theta_0)$. The result in (2.30) follows readily from (2.9) and (2.10). Under LAN, it can be shown that

$$\boldsymbol{\Delta}_n(\theta_0) \xrightarrow{d} N_k(\boldsymbol{\Gamma}(\theta_0)\mathbf{h}, \boldsymbol{\Gamma}(\theta_0)), \quad \text{under } P_{\theta_n}\text{-probability} \ . \tag{2.31}$$

Also, under LAN, we have, from (2.9),

$$\mathbf{C}_n^{-1}(\theta_0)\boldsymbol{\Gamma}_n(\theta_0)\mathbf{C}_n^{-1}(\theta_0) \xrightarrow{d} \boldsymbol{\Gamma}(\theta_0), \quad \text{under } P_{\theta_n}\text{-probability} \ . \tag{2.32}$$

It follows from (2.31) and (2.32) that, under P_{θ_n}-probability,

$$\mathbf{S}_n^{\mathrm{T}}(\theta_0)\boldsymbol{\Gamma}_n^{-1}(\theta_0)\mathbf{S}_n(\theta_0) \xrightarrow{d} \chi^2(k, \lambda) \ , \tag{2.33}$$

where $\chi^2(k, \lambda)$ denotes a non-central chi-square random variable with k degrees of freedom and non-centrality parameter $\lambda = \mathbf{h}^{\mathrm{T}}\boldsymbol{\Gamma}(\theta_0)\mathbf{h}$. Consider the test function

$$\tilde{\phi}_n = \begin{cases} 1, & 2\mathbf{Q}_n(\hat{\theta}_0, \theta_0) \geq \chi^2_{1-\alpha}(k) \\ 0, & 2\mathbf{Q}_n(\hat{\theta}_0, \theta_0) < \chi^2_{1-\alpha}(k) \end{cases} \ . \tag{2.34}$$

We can show that $\tilde{\phi}_n$ in (2.34) is asymptotically efficient in a certain class of tests. Let T_n be any estimator of θ satisfying (2.12). Consider the test function

$$\tilde{\phi}_n^* = \begin{cases} 1, & (T_n - \theta_0)^{\mathrm{T}}\mathbf{C}_n(\theta_0)\mathbf{V}_T^{-1}(\theta_0)\mathbf{C}_n(\theta_0)(T_n - \theta_0) \geq \chi^2_{1-\alpha}(k) \\ 0, & (T_n - \theta_0)^{\mathrm{T}}\mathbf{C}_n(\theta_0)\mathbf{V}_T^{-1}(\theta_0)\mathbf{C}_n(\theta_0)(T_n - \theta_0) < \chi^2_{1-\alpha}(k) \end{cases} \ . \tag{2.35}$$

We have $\lim E_H\tilde{\phi}_n = \lim E_H\tilde{\phi}_n^* = \alpha$. We shall now compare the limiting powers of $\tilde{\phi}_n$ and $\tilde{\phi}_n^*$. From (2.33) and (2.34), we have,

$$\lim E_{\theta_n}\tilde{\phi}_n = P(\chi^2(k, \lambda) \geq \chi^2_{1-\alpha}(k)) \ . \tag{2.36}$$

From (2.12), we have, under P_{θ_n}-probability,

$$\mathbf{C}_n(\theta_0)(T_n - \theta_0) \xrightarrow{d} N_k(\mathbf{h}, \mathbf{V}_T(\theta_0)) \ . \tag{2.37}$$

Consequently,

$$(T_n - \theta_0)^{\mathrm{T}} \mathbf{C}_n(\theta_0) \mathbf{V}_T^{-1}(\theta_0) \mathbf{C}_n(\theta_0)(T_n - \theta_0) \xrightarrow{d} \chi^2(k, \lambda^*) \ , \tag{2.38}$$

where $\lambda^* = \mathbf{h}^{\mathrm{T}} \mathbf{V}_T^{-1}(\theta_0)\mathbf{h}$. Therefore,

$$\lim E_{\theta_n} \tilde{\phi}_n^* = P(\chi^2(k, \lambda^*) \geq \chi^2_{1-\alpha}(k)) \ . \tag{2.39}$$

From (2.13), we have $\lambda \geq \lambda^*$, and hence by (2.36) and (2.39) we have

$$\lim E_{\theta_n} \tilde{\phi}_n \geq \lim E_{\theta_n} \tilde{\phi}_n^* \ . \tag{2.40}$$

Any test $\tilde{\phi}_n^*$ is said to be asymptotically efficient if its limiting power attains the upper bound in (2.40). We have thus shown that the score statistic $2\mathbf{Q}_n(\hat{\theta}_0, \theta_0)$ is asymptotically efficient in the above sense.

The score test is not the only test which is efficient according to the criterion based on (2.40). The usual likelihood ratio statistic and the so-called Wald statistic are also asymptotically efficient. The likelihood ratio statistic is given by $2\Lambda_n(\hat{\theta}_n, \theta_0)$, where we can use the one-step maximum likelihood estimator $\hat{\theta}_n$ defined by (2.19). We have, from (2.2),

$$2\Lambda_n(\hat{\theta}_n, \theta_0) = 2\mathbf{Q}_n(\hat{\theta}_n, \theta_0) + o_p(1), \quad \text{under } P_{\theta_0}\text{-probability} \ . \tag{2.41}$$

Under LAN, (2.41) remains valid under P_{θ_n}-probability also. We will show that $\mathbf{Q}_n(\hat{\theta}_n, \theta_0)$ and $\mathbf{Q}_n(\hat{\theta}_0, \theta_0)$ have the same limit distributions under both P_{θ_0}- and P_{θ_n}-probabilities. We have

$$\begin{aligned}
\mathbf{Q}_n(\hat{\theta}_n, \theta_0) &= (\hat{\theta}_n - \theta_0)^{\mathrm{T}} \mathbf{S}_n(\theta_0) - \tfrac{1}{2}(\hat{\theta}_n - \theta_0)^{\mathrm{T}} \boldsymbol{\Gamma}_n(\theta_0)(\hat{\theta}_n - \theta_0) \\
&= \boldsymbol{\Delta}_n^{\mathrm{T}}(\theta_0)\boldsymbol{\Gamma}^{-1}(\theta_0)\boldsymbol{\Delta}_n(\theta_0) - \tfrac{1}{2}\boldsymbol{\Delta}_n^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}(\theta_0) \\
&\quad \times (\mathbf{C}_n^{-1}(\theta_0)\boldsymbol{\Gamma}_n(\theta_0)\mathbf{C}_n^{-1}(\theta_0))\boldsymbol{\Gamma}^{-1}(\theta_0)\boldsymbol{\Delta}_n^{\mathrm{T}}(\theta_0) \\
&\quad + o_p(1), \ \text{by (2.22)} \\
&= \tfrac{1}{2}\mathbf{S}_n^{\mathrm{T}}(\theta_0)\boldsymbol{\Gamma}_n^{-1}(\theta_0)\mathbf{S}_n(\theta_0) + o_p(1) \\
&= \mathbf{Q}_n(\hat{\theta}_0, \theta_0) + o_p(1), \quad \text{under } P_{\theta_0} \ ,
\end{aligned} \tag{2.42}$$

by (2.7). Under LAN, the above result also holds under P_{θ_n}-probability via contiguity (see Hall and Mathiason (1990)). Consequently, the likelihood ratio statistic $2\Lambda_n(\hat{\theta}_n, \theta_0)$ has the same limit distributions as that of $2\mathbf{Q}_n(\hat{\theta}_0, \theta_0)$ under both P_{θ_0}- and P_{θ_n}-probabilities, and hence it is asymptotically efficient.

Finally, the Wald statistic is defined by $(\hat{\theta}_n - \theta_0)^{\mathrm{T}} \boldsymbol{\Gamma}_n(\theta_0)(\hat{\theta}_n - \theta_0)$. It is seen that

$$\begin{aligned}
(\hat{\theta}_n - \theta_0)^{\mathrm{T}} \boldsymbol{\Gamma}_n(\theta_0)(\hat{\theta}_n - \theta_0) &= \boldsymbol{\Delta}_n^{\mathrm{T}}(\theta_0)\boldsymbol{\Gamma}^{-1}(\theta_0)(\mathbf{C}_n^{-1}(\theta_0)\boldsymbol{\Gamma}_n(\theta_0)\mathbf{C}_n^{-1}(\theta_0)) \\
&\quad \times \boldsymbol{\Gamma}^{-1}(\theta_0)\boldsymbol{\Delta}_n(\theta_0) + o_p(1), \ \text{by (2.22)} \\
&= \mathbf{S}_n^{\mathrm{T}}(\theta_0)\boldsymbol{\Gamma}_n^{-1}(\theta_0)\mathbf{S}_n(\theta_0) + o_p(1) \\
&= 2\mathbf{Q}_n(\hat{\theta}_0, \theta_0) + o_p(1) \ .
\end{aligned} \tag{2.43}$$

Again, the above result is valid under both P_{θ_0} and P_{θ_n}-probabilities. The Wald statistic is therefore asymptotically efficient.

2.6. Efficient tests: Composite hypotheses

Let $\theta = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $(p \times 1)$ and $(k-p) \times 1$ vectors, respectively. Suppose α is the parameter of interest and β represents a nuisance parameter. Consider the problem of testing a composite hypothesis $H : \alpha = \alpha_0$, when β is unknown. We shall show that, under LAN, the score test, the Wald test and the likelihood ratio test are all asymptotically efficient in the sense of maximizing limiting power at the local alternatives. Partition $\mathbf{h} = (\mathbf{h}_\alpha^{\mathrm{T}}, \mathbf{h}_\beta^{\mathrm{T}})^{\mathrm{T}}$ where $\mathbf{h}_\alpha$ and $\mathbf{h}_\beta$ are of the order $(p \times 1)$ and $(k-p) \times 1$ respectively. Our efficiency criterion will be based on the limiting power at the local alternatives $K_n : \alpha = \alpha_n$, and $\beta = \beta_n$, where $\alpha_n = \alpha_0 + \mathbf{C}_n^{-1}(\alpha_0, \beta)\mathbf{h}_\alpha$, and $\beta_n = \beta + \mathbf{C}_{n,\beta}^{-1}(\alpha_0, \beta)\mathbf{h}_\beta$. Here $\mathbf{C}_n(\theta)$ is taken as a diagional matrix, $\mathbf{C}_{n,\alpha}$ is the $(p \times p)$ diagonal matrix containing the first p diagonal elements and $\mathbf{C}_{n,\beta}$ is the $(k-p) \times (k-p)$ diagonal matrix containing the remaining $(k-p)$ diagonal elements of $\mathbf{C}_n(\theta)$. Note that β is unspecified in both α_n and β_n. Now, partition $\mathbf{S}_n(\theta)$ and $\boldsymbol{\Gamma}_n(\theta)$ as

$$\mathbf{S}_n(\theta) = \begin{pmatrix} \mathbf{S}_{n,\alpha}(\theta) \\ \mathbf{S}_{n,\beta}(\theta) \end{pmatrix}, \quad \boldsymbol{\Gamma}_n(\theta) = \begin{pmatrix} \boldsymbol{\Gamma}_{n,\alpha\alpha}(\theta) & \boldsymbol{\Gamma}_{n,\alpha\beta}(\theta) \\ \boldsymbol{\Gamma}_{n,\alpha\beta}^{\mathrm{T}}(\theta) & \boldsymbol{\Gamma}_{n,\beta\beta}(\theta) \end{pmatrix},$$

where $\mathbf{S}_{n,\alpha}(\theta)$ is of order $(p \times 1)$, $\boldsymbol{\Gamma}_{n,\alpha\alpha}(\theta)$ is a $(p \times p)$ matrix, etc. In analogy with (2.19) we define a one-step likelihood equation estimator for β under the restriction $H : \alpha = \alpha_0$, as

$$\hat{\beta}_{0n} = \tilde{\beta}_0 + \boldsymbol{\Gamma}_n^{\beta\beta}(\alpha_0, \tilde{\beta}_0)\mathbf{S}_{n,\beta}(\alpha_0, \tilde{\beta}_0) \ , \tag{2.44}$$

where $\tilde{\beta}_0$ is any preliminary estimator of β such that

$$\mathbf{C}_{n,\beta}(\alpha_0, \beta)(\tilde{\beta}_0 - \beta) = o_p(1) \ ,$$

and $\boldsymbol{\Gamma}_n^{\beta\beta}$ denotes the lower right hand $(k-p) \times (k-p)$ matrix in $\boldsymbol{\Gamma}_n^{-1}$, i.e.,

$$\boldsymbol{\Gamma}_n^{\beta\beta} = (\boldsymbol{\Gamma}_{n,\beta\beta} - \boldsymbol{\Gamma}_{n,\alpha\beta}^{\mathrm{T}}\boldsymbol{\Gamma}_{n,\alpha\alpha}^{-1}\boldsymbol{\Gamma}_{n,\alpha\beta})^{-1} \ .$$

It follows, as a special case of (2.21), that under H,

$$\mathbf{C}_{n,\beta}(\alpha_0, \beta)(\hat{\beta}_{0n} - \beta) \xrightarrow{d} N_{(k-p)}(0, \boldsymbol{\Gamma}^{\beta\beta}(\alpha_0, \beta)) \ . \tag{2.45}$$

Let $\hat{\theta}_{n,H} = (\alpha_0^{\mathrm{T}}, \hat{\beta}_{0,n}^{\mathrm{T}})^{\mathrm{T}}$. The likelihood ratio statistic for testing the composite hypothesis $H : \alpha = \alpha_0$, is given by

$$T_n^{(1)} = 2\Lambda_n(\hat{\theta}_n, \hat{\theta}_{n,H}) \ , \tag{2.46}$$

where $\hat{\theta}_n = (\hat{\alpha}_n^{\mathrm{T}}, \hat{\beta}_n^{\mathrm{T}})^{\mathrm{T}}$ is given by (2.19). The Wald and the score statistics are given respectively by

$$T_n^{(2)} = (\hat{\alpha}_n - \alpha_0)^{\mathrm{T}} \mathbf{C}_{n,\alpha}(\hat{\theta}_n)\{\boldsymbol{\Gamma}^{\alpha\alpha}(\hat{\theta}_n)\}^{-1} \mathbf{C}_{n,\alpha}(\hat{\theta}_n)(\hat{\alpha}_n - \alpha_0) \ , \tag{2.47}$$

and

$$T_n^{(3)} = \mathbf{S}_{n,\alpha}^{\mathrm{T}}(\hat{\theta}_{n,H})\mathbf{C}_n^{-1}(\hat{\theta}_{n,H})\boldsymbol{\Gamma}^{\alpha\alpha}(\hat{\theta}_{n,H})\mathbf{C}_{n,\alpha}^{-1}(\hat{\theta}_{n,H})\mathbf{S}_{n,\alpha}(\hat{\theta}_{n,H}) \ . \tag{2.48}$$

The limit distributions of the above three statistics $T_n^{(i)}$, $i = 1, 2, 3$, can be shown to be identical under both H and K_n. We have, for $i = 1, 2, 3$,

$$T_n^{(i)} \xrightarrow{d} \begin{cases} \chi^2(p), & \text{under } H \\ \chi^2(p, \lambda), & \text{under } K_n \end{cases} \ , \tag{2.49}$$

where $\lambda = \mathbf{h}_\alpha^{\mathrm{T}}(\boldsymbol{\Gamma}_{(\theta_H)}^{\alpha\alpha})^{-1}\mathbf{h}_\alpha$, with $\theta_H = (\alpha_0, \beta)$. Consider the class of asymptotically similar size-γ tests of H, viz., the class of test functions ϕ_n such that

$$\lim E_H(\phi_n) = \gamma, \quad \text{for all } \beta \ .$$

Now, let U_n be any estimator of α such that, under K_n,

$$\mathbf{C}_{n,\alpha}(\alpha_0, \hat{\beta}_{0n})(U_n - \alpha_0) \xrightarrow{d} N_p(\mathbf{h}_\alpha, \mathbf{B}_U(\alpha_0, \beta)) \ , \tag{2.50}$$

where $\mathbf{B}_U$ is a positive definite matrix. Note that (2.50) is an adaptation of the regularity requirement in (2.12). We then have, as in (2.13),

$$\mathbf{B}_U(\alpha_0, \beta) \geq \boldsymbol{\Gamma}^{\alpha\alpha}(\alpha_0, \beta_0) \ . \tag{2.51}$$

Consider a test statistic T_n based on U_n, defined by

$$T_n = \mathbf{V}_n^{\mathrm{T}}\mathbf{B}_U^{-1}(\alpha_0, \hat{\beta}_{0n})\mathbf{V}_n \ , \tag{2.52}$$

where

$$\mathbf{V}_n = \mathbf{C}_{n,\alpha}(\alpha_0, \hat{\beta}_{0n})(U_n - \alpha_0) \ .$$

It follows from (2.50) that, under K_n,

$$T_n \xrightarrow{d} \chi^2(k, \lambda^*) \ , \tag{2.53}$$

where $\lambda^* = \mathbf{h}_\alpha^{\mathrm{T}}\mathbf{B}_U^{-1}(\alpha_0, \beta)h_\alpha$. From (2.51) it follows that

$$\lambda^* \leq \lambda \ , \tag{2.54}$$

where λ is the non-centrality parameter appearing in (2.49). Now, define a test function ϕ_n,

$$\phi_n = \begin{cases} 1 & \text{if } T_n \geq \chi_{1-\gamma}^2(p) \\ 0 & \text{if } T_n < \chi_{1-\gamma}^2(p) \end{cases} \ . \tag{2.55}$$

It is easily verified that ϕ_n is asymptotically similar size-α. From (2.54) it follows that the Wald statistic $T_n^{(2)}$ is asymptotically efficient in the class of tests given by (2.55). Since the likelihood ratio and the score statistics can both be expressed as

equal to $T_n^{(2)}$ plus a $o_p(1)$ term under both H and K_n, it follows that these two tests are also asymptotically efficient in the same class.

2.7. Neyman and Durbin statistics

Consider the problem of testing discussed in Section 2.6. In addition to the three statistics presented in the previous section, the following two statistics are also efficient. The Neyman $C(\alpha)$-statistic is defined by

$$T_n^{(4)} = Y_n^{\mathrm{T}} \boldsymbol{\Gamma}^{\alpha\alpha}(\hat{\theta}_{n,H}) Y_n \ , \tag{2.56}$$

where

$$Y_n = \mathbf{C}_{n,\alpha}^{-1}(\hat{\theta}_{n,H}) \mathbf{S}_{n,\alpha}(\hat{\theta}_{n,H}) - \boldsymbol{\Gamma}_{\alpha,\beta}(\hat{\theta}_{n,H}) \boldsymbol{\Gamma}_{\beta\beta}^{-1}(\hat{\theta}_{n,H}) \mathbf{C}_{n,\beta}^{-1}(\hat{\theta}_{n,H}) \mathbf{S}_{n,\beta}(\hat{\theta}_{n,H}) \ ,$$

which represents a regression of $\mathbf{S}_{n,\alpha}$ on $\mathbf{S}_{n,\beta}$.

Since $\mathbf{S}_{n,\beta}(\hat{\theta}_{n,H}) = o_p(1)$, it follows that $T_n^{(4)}$ is asymptotically equivalent to the score statistic $T_n^{(3)}$, under both H and K_n.

In order to introduce the Durbin statistic, let $\hat{\beta}_{0n}$ be the estimator defined in (2.44), and let $\hat{\alpha}_{0n}$ be the estimator defined by

$$\hat{\alpha}_{0n} = \tilde{\alpha}_0 + \boldsymbol{\Gamma}_n^{\alpha\alpha}(\tilde{\alpha}_0, \hat{\beta}_{0n}) \mathbf{S}_{n,\alpha}(\tilde{\alpha}_0, \hat{\beta}_{0n}) \ , \tag{2.57}$$

where $\tilde{\alpha}_0$ is any preliminary estimator of α such that $\mathbf{C}_{n,\alpha}(\theta_H)(\tilde{\alpha}_0 - \alpha_0) = O_p(1)$. In other words, $\hat{\beta}_{0n}$ is a one-step solution of the equation $\mathbf{S}_{n,\beta}(\alpha_0, \beta) = 0$ for β, and $\hat{\alpha}_{0n}$ is a one-step solution of the equation $\mathbf{S}_{n,\alpha}(\alpha, \hat{\beta}_{0n}) = 0$ for α. The Durbin statistic for testing $H : \alpha = \alpha_0$ is then defined by

$$T_n^{(5)} = (\hat{\alpha}_{0n} - \alpha_0)^{\mathrm{T}} \mathbf{C}_n(\hat{\theta}_{n,H}) \{\boldsymbol{\Gamma}^{\alpha\alpha}(\hat{\theta}_{n,H})\}^{-1} \mathbf{C}_n(\hat{\theta}_{n,H})(\hat{\alpha}_{0n} - \alpha_0) \ . \tag{2.58}$$

It is easily verified that $T_n^{(5)}$ is asymptotically equivalent to $T_n^{(2)}$ under H as well as under K_n. Consequently, all the five statistics $T_n^{(i)}$, $i = 1, \ldots, 5$, discussed in Sections (2.6) and (2.7) are asymptotically efficient. The choice among these statistics may depend on the simplicity in deriving the statistics in any particular problem.

Note that $T_n^{(1)}$ depends on both the restricted and the unrestricted (maximum) likelihood estimators $\hat{\theta}_{n,H}$ and $\hat{\theta}_n$. The Wald statistic $T_n^{(2)}$ depends only on the unrestricted estimator $\hat{\theta}_n$. The score statistic $T_n^{(3)}$ and the Neyman $C(\alpha)$-statistic $T_n^{(4)}$ both need the restricted estimator $\hat{\theta}_{n,H}$. The Durbin statistic requires the two restricted estimators $\hat{\alpha}_{0n}$ and $\hat{\beta}_{0n}$ obtained in a convenient successive substitution. Even though asymptotically these five statistics are asymptotically equivalent, for small or moderate sample sizes their performance may vary significantly.

2.8. Extension to non-ergodic models

For some models in stochastic processes, it turns out that the limiting Fisher information $\boldsymbol{\Gamma}(\theta)$ in (2.9) is a non-degenerate random matrix. The limit distribution of the maximum likelihood estimator given in (2.21) will then be a mixture

of normals rather than a normal. Typically, such models belong to the local asymptotic mixed normal (LAMN) family rather than the LAN family. See Basawa and Scott (1983) for the theory and applications of the LAMN family. Stochastic models belonging to this class are also referred to as non-ergodic models. See Basawa (1981a, b) and Basawa and Brockwell (1984) for conditional inference for non-ergodic models.

3. Optimal estimating functions

In many cases, the density $p_n(x(n);\theta)$ is either not known, or it may be unwieldy. Consider a class of estimating functions defined by

$$\mathbf{g}_n(\theta) = \sum_{t=1}^{n} \mathbf{W}_t(\theta)(X_t - m_t(\theta)) \ , \tag{3.1}$$

where $\mathbf{g}_n(\theta)$ and $\mathbf{W}_t(\theta)$ are $(p \times 1)$ random vectors such that $\mathbf{g}_n(\theta) \in \mathscr{F}_n$ and $\mathbf{W}_t(\theta) \in \mathscr{F}_{t-1}$, where $\mathscr{F}_m = \sigma(X_m, X_{m-1}, \ldots,)$. If $m_t(\theta) = E(X_t|\mathscr{F}_{t-1})$, $\{\mathbf{g}_n(\theta), \mathscr{F}_n\}$ is a zero-mean martingale. Let $\sigma_t^2(\theta) = \mathrm{Var}(X_t|\mathscr{F}_{t-1})$. Godambe (1985) has shown that an optimum choice of $\mathbf{W}_t(\theta)$ is

$$\mathbf{W}_t^0(\theta) = \left(\frac{\mathrm{d}m_t(\theta)}{\mathrm{d}\theta}\right)\sigma_t^{-2}(\theta) \ , \tag{3.2}$$

where the optimality criterion seeks $\mathbf{W}_t(\theta)$ which maximizes (in the partial order of non-negative definite matrices) the Godambe information matrix

$$\mathbf{I}_{g_n}(\theta) = \left(E\left(\frac{\mathrm{d}\mathbf{g}_n}{\mathrm{d}\theta}\right)\right)\left(E(\mathbf{g}_n\mathbf{g}_n^{\mathrm{T}})\right)^{-1}\left(E\left(\frac{\mathrm{d}\mathbf{g}_n}{\mathrm{d}\theta}\right)\right)^{\mathrm{T}} \ . \tag{3.3}$$

Thus, if

$$\mathbf{g}_n^0(\theta) = \sum_{t=1}^{n}(X_t - m_t(\theta))\mathbf{W}_t^0(\theta) \ ,$$

we have that $(\mathbf{I}_{g_n^0} - \mathbf{I}_{g_n})$ is non-negative definite for all $\mathbf{g}_n$ of the form (3.1) and satisfying some regularity conditions. The optimal estimating function $\mathbf{g}_n^0$ is also referred to as a quasi-score function. See, for instance, Heyde (1997). The quasi-score estimator $\hat{\theta}_q$ is obtained as a solution of the equation $\mathbf{g}_n^0(\theta) = 0$. Under appropriate regularity conditions (see Heyde (1997)) one can show that

$$\mathbf{W}_n^{1/2}(\theta)(\hat{\theta}_q - \theta) \xrightarrow{d} N(0, \mathbf{I}) \ , \tag{3.4}$$

where

$$\mathbf{W}_n(\theta) = \sum_{t=1}^{n}\left(\frac{\mathrm{d}m_t(\theta)}{\mathrm{d}\theta}\right)\left(\frac{\mathrm{d}m_t(\theta)}{\mathrm{d}\theta}\right)^{\mathrm{T}}\sigma_t^{-2}(\theta) \ . \tag{3.5}$$

If the estimating function $\mathbf{g}_n(\theta)$ is not restricted to be of the "linear" form in (3.1), it is well known that the optimum $\mathbf{g}_n(\theta)$ which maximizes $\mathbf{I}_{g_n}(\theta)$ in (3.3) in the unrestricted class of estimating functions is given by the likelihood score function $\mathbf{S}_n(\theta)$. See Godambe (1960).

The theory and applications of quasi-score estimators, confidence sets and test statistics are discussed in Godambe (1991) and Heyde (1997). See also Basawa (1985, 1991) and Basawa et al. (1985) for tests based on estimating functions.

4. Semiparametric models and adaptive estimation

Consider the model

$$X_t = m_t(\theta) + \sigma_t(\theta)\epsilon_t \tag{4.1}$$

where $\{\epsilon_t\}$ is a sequence of independent and identically distributed random errors with $E(\epsilon_t) = 0$ and $\mathrm{Var}(\epsilon_t) = \sigma_\epsilon^2$, $m_t(\theta) = E(X_t|\mathscr{F}_{t-1})$, and $\sigma_t^2(\theta) = \mathrm{Var}(X_t|\mathscr{F}_{t-1})$. Suppose $\boldsymbol{\theta}$ is a $(p \times 1)$ vector of parameters. If the density $f_\epsilon(\cdot)$ of $\{\epsilon_t\}$ is known, one can apply the likelihood methods for inference regarding $\boldsymbol{\theta}$. On the other-hand, if $f_\epsilon(\cdot)$ is unknown, and $m_t(\theta)$ and $\sigma_t^2(\theta)$ are modeled as known functions of θ, measurable with respect to $\mathscr{F}_{t-1}$, the model in (4.1) is an example of a semi-parametric model.

Let $\tilde{\theta}_n$ be a preliminary estimator of θ. For instance, $\tilde{\theta}_n$ may be the least-squares estimator obtained by minimizing

$$\sum_{t=1}^{n} \epsilon_t^2(\theta) = \sum_{t=1}^{n} \frac{(X_t - m_t(\theta))^2}{\sigma_t^2(\theta)} \ . \tag{4.2}$$

Denote $Y_t = \epsilon_t(\tilde{\theta}_n)$. Let $f_n(y)$ denote a kernel density estimator, e.g.,

$$f_n(y) = \frac{1}{n}\sum_{t=1}^{n} \frac{1}{b_n} K\left(\frac{y - Y_t}{b_n}\right) \ , \tag{4.3}$$

where K and b_n are the kernel and the bandwidth respectively. Let $\hat{\theta}_n(f_n)$ denote the estimator obtained as a solution of the estimating equation:

$$\sum_{t=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta} \log f_n\left(\frac{X_t - \mu_t(\theta)}{\sigma_t(\theta)}\right) = 0 \ . \tag{4.4}$$

Note that if f_ϵ is known, the usual likelihood equation is given by

$$\sum_{t=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\epsilon\left(\frac{X_t - \mu_t(\theta)}{\sigma_t(\theta)}\right) = 0 \ . \tag{4.5}$$

If $\hat{\theta}_n(f_\epsilon)$ denotes a solution of (4.5) when f_ϵ is known, one can show, under appropriate regularity conditions, that

$$\sqrt{n}(\hat{\theta}_n(f_n) - \hat{\theta}_n(f_\epsilon)) = o_p(1) \ . \tag{4.6}$$

When (4.6) is satisfied, the estimator $\hat{\theta}_n(f_n)$ is said to be adaptive. See Bickel (1982), Bickel et al. (1993) and Drost et al. (1997) for the theory and applications of adaptive estimation for semiparametric models.

5. Bayes and empirical Bayes methods

Suppose that prior information about the parameter θ is available and that it can be quantified in the form of a density function $\pi(\theta)$, $\theta \in \Omega \subset R^p$. The density $\pi(\theta)$ is referred to as the prior density. Given θ, the observation vector $\mathbf{X}(n)$ has density $p(x(n)|\theta)$. If δ_n is an estimator of θ, and if $l(\delta_n, \theta)$ is a prescribed loss function, the Bayes risk (or average risk) corresponding to δ_n is given by $R(\delta_n) = E(l(\delta_n, \theta))$, where the expectation $E(\cdot)$ is with respect to the joint density, $p(x(n)|\theta)\pi(\theta)$. The Bayes estimator δ_n^0 of θ is such that

$$R(\delta_n^0) \leq R(\delta_n), \text{ for all } \delta_n \in A \ , \tag{5.1}$$

where A is the class of all estimators with finite risk. If the loss function is quadratic, it can be shown that

$$\delta_n^0 = E(\theta|\mathbf{X}(n)) \ . \tag{5.2}$$

Define

$$\mathbf{\Gamma}_n(\theta) = -\frac{\mathrm{d}^2 \log p_n(\mathbf{X}(n)|\theta)}{\mathrm{d}\theta \, \mathrm{d}\theta^{\mathrm{T}}} \ . \tag{5.3}$$

If $\hat{\theta}_n$ is the maximum likelihood estimator, we have seen in Section 2, that under regularity conditions,

$$\mathbf{\Gamma}_n^{1/2}(\theta)(\hat{\theta}_n - \theta) \xrightarrow{d} N_p(0, \mathbf{I}) \ . \tag{5.4}$$

The ML estimator $\hat{\theta}_n$ ignores the prior information $\pi(\theta)$, and uses only the sample information contained in the likelihood function $p_n(x(n)|\theta)$. It is of interest to compare the Bayes estimator δ_n^0 with the ML estimator $\hat{\theta}_n$ for large n. Under regularity conditions (see Basawa and Prakasa Rao (1980, Ch. 10)) one can show that

$$\mathbf{\Gamma}_n^{1/2}(\theta)(\delta_n^0 - \theta) \xrightarrow{d} N_p(0, \mathbf{I}) \ , \tag{5.5}$$

and hence the Bayes estimator δ_n^0 is asymptotically equivalent to the ML estimator $\hat{\theta}_n$.

Often, the prior density depends on an unknown parameter, say α. Denote the prior density as $\pi(\theta; \alpha)$. The Bayes estimator of θ will then depend on α, and denote the Bayes estimator as $\delta_n^0(\alpha)$. Since α is unknown, we may first estimate α from the marginal density

$$p(x(n);\alpha) = \int p(x(n)|\theta)\pi(\theta;\alpha)\mathrm{d}\theta \ . \tag{5.6}$$

Let $\hat{\alpha}_n$ denote the maximum likelihood estimator of α based on the marginal likelihood $p(\mathbf{X}(n);\alpha)$. The estimator, $\delta_n^0(\hat{\alpha}_n)$, obtained from the Bayes estimator by replacing α by $\hat{\alpha}_n$, is known as an empirical Bayes estimator of θ. It can be shown that $\delta_n^0(\hat{\alpha}_n)$ is a good approximation for the Bayes estimator $\delta_n^0(\alpha)$, for large n.

6. Some applications

Here we give some examples to illustrate the inference methods discussed in the previous Sections.

Ex. 1. Markov processes

Let $\{X_t\}$, $t = 1,2,\ldots,$ be a Markov process with a general state space χ and stationary transition measures

$$F_\theta(x,A) = P_\theta(X_{n+1} \in A | X_n = x) \ , \tag{6.1}$$

$\theta \in \Omega \subset R^p$. Suppose these transition measures admit a unique stationary distribution $F_\theta(\cdot)$ defined by

$$F_\theta(A) = \int_\chi F_\theta(x,A)F_\theta(\mathrm{d}x) \ . \tag{6.2}$$

Furthermore, suppose that $F_\theta(x,A)$ admit transition densities $p(y,x;\theta)$, with respect to a measure $\lambda(\cdot)$ defined by the relation

$$F_\theta(x,A) = \int_A p(y,x;\theta)\lambda(\mathrm{d}y) \ . \tag{6.3}$$

The likelihood function based on $\mathbf{X}(n) = (X_1,\ldots,X_n)$ (and conditional on $X_1 = x_1$) is given by

$$p_n(\mathbf{X}(n);\theta) = \prod_{t=1}^{n-1} p(X_t,X_{t+1};\theta) \ . \tag{6.4}$$

Under regularity conditions (see Billingsley (1961)) it can be shown that the model belongs to the LAN family. See also Roussas (1972).

As a specific example, consider a finite state Markov chain with state space $\chi = \{1,2,\ldots,m\}$, and the transition densities

$$p_{ij} = P(X_{t+1} = j | X_t = i), \quad i,j \in \chi \ . \tag{6.5}$$

Let n_{ij} denote the number of transitions $i \to j$ in the sample $\mathbf{X}(n)$. The likelihood function is given by

$$p_n(\mathbf{X}(n);\theta) = \prod_{i,j} p_{ij}^{n_{ij}} \ . \tag{6.6}$$

Here the parameter space for $\theta = \{p_{ij}, i,j \in \chi, \text{ s.t. } \sum_j p_{ij} = 1\}$. The maximum likelihood estimator of p_{ij} is seen to be

$$\hat{p}_{ij} = \frac{n_{ij}}{n_{i.}}, \quad i,j \in \chi \ , \tag{6.7}$$

where $n_{i.} = \sum_j n_{ij}$. It can be shown (see Billinglsey (1961)) that $\sqrt{n}(\hat{p}_{ij} - p_{ij})$, $i,j \in \chi$, are jointly asymptotically normal with mean zero and asymptotic variances and covariances given by

$$\sigma_{i,j,i',j'} = \frac{1}{\pi_i}[\delta_{ii'}(\delta_{jj'}p_{ij} - p_{ij}p_{i'j'})] \ , \tag{6.8}$$

where δ_{uv} denotes the indicator function which takes the value 1 if $u = v$ and zero if $u \neq v$. The asymptotic optimality property of $\{\hat{p}_{ij}\}$ is assured by the LAN property. See Basawa and Prakasa Rao (1980a, Ch. 4) for various problems of inference regarding finite Markov chains.

Ex. 2. Branching processes

Let $X_0 = 1$, X_1, $X_2 \ldots$ be the generation sizes of a Galton–Watson branching process with offspring distribution

$$P(X_1 = j) = p_j, \quad j = 0,1,2,\ldots \ . \tag{6.9}$$

Let $\{Y_{kl}\}$, $k = 0,1,\ldots$, $l = 0,1,\ldots$, denote the lth offspring belonging to the kth generation. We then have

$$X_{n+1} = \sum_{l=1}^{X_n} Y_{nl} \ . \tag{6.10}$$

Note that $\{X_t\}$ is a Markov process with state space $\chi = \{0,1,2,\ldots\}$ and the transition probabilities

$$\begin{aligned} p_{ij} &= P(X_{n+1} = j | X_n = i) \\ &= P\left(\sum_{l=1}^{i} Y_{nl} = j\right) \ . \end{aligned} \tag{6.11}$$

The random variables $\{Y_{kl}\}$ are assumed to be independent and identically distributed, each distributed as X_1, with $E(X_1) = \mu$ and $\text{Var}(X_1) = \sigma^2$. More specifically, suppose the offspring distribution belongs to the power series family, viz.,

$$P(X_1 = x) = a_x\theta^x/A(\theta), \quad x = 0,1,2,\ldots \ .$$

where $A(\theta) = \sum_{x=0}^{\infty} a_x\theta^x$. We have

$$\mu = \theta \frac{\mathrm{d}A(\theta)}{\mathrm{d}\theta} \Big/ A(\theta), \quad \text{and} \quad \sigma^2 = \theta \frac{\mathrm{d}\mu}{\mathrm{d}\theta} \ .$$

The likelihood function based on $(X_1, \ldots, X_n)$ is given by

$$p_n(\mathbf{X}(n); \theta) \alpha \theta^{\sum_1^n X_t} (A(\theta))^{\sum_1^n X_{t-1}} \ . \tag{6.12}$$

We have

$$\begin{aligned} \mathbf{S}_n(\mu) &= \frac{\partial \log p_n(\mathbf{X}(n); \theta)}{\partial \mu} = \left(\frac{\partial \log p_n(\mathbf{X}(n); \theta)}{\partial \theta} \right) \frac{\partial \theta}{\partial \mu} \\ &= \sigma^{-2}(\mu) \left(\sum_1^n X_t - \mu \sum_1^n X_{t-1} \right) \ . \end{aligned} \tag{6.13}$$

The equation $\mathbf{S}_n(\mu) = 0$ gives the ML estimator of μ:

$$\hat{\mu}_n = \sum_1^n X_t \Big/ \sum_1^n X_{t-1} \ . \tag{6.14}$$

The Fisher information is seen to be

$$I_n(\mu) = E\left(\frac{-\partial^2 \log p_n}{\partial \mu^2} \right) = \sigma^{-2}(\mu) \left(\frac{\mu^n - 1}{\mu - 1} \right) \ . \tag{6.15}$$

Assume throughout that $P(Y_1 = 0) = 0$, and $\mu > 1$ to avoid extinction of the process. It can be shown that

$$\left(-\frac{\partial^2 \log p_n}{\partial \mu^2} \Big/ I_n(\mu) \right) \xrightarrow{p} W, \quad \text{as} \quad n \to \infty \ , \tag{6.16}$$

where $W > 0$ is a non-degenerate random variable. This process, therefore, belongs to the LAMN family. See Basawa and Scott (1983) for problems of inference regarding μ. See also Heyde (1975).
Let

$$J_n(\mu) = \sigma^{-2}(\mu) \sum_1^n X_{t-1} \ . \tag{6.17}$$

One can show that

$$J_n^{1/2}(\mu)(\hat{\mu}_n - \mu) \xrightarrow{d} N(0, 1) \ . \tag{6.18}$$

The result in (6.18) is equivalent to

$$I_n^{1/2}(\mu)(\hat{\mu}_n - \mu) \xrightarrow{d} N^*(0, W^{-1}) \ . \tag{6.19}$$

Note that the limit distribution N^* in (6.19) is a mixture of normals rather than a normal.

Guttorp (1991) gives an extensive review of inference problems for branching processes.

Ex. 3. Time series

Consider the model

$$X_t = m_t(\theta) + \sigma_t(\theta)\epsilon_t \ , \tag{6.20}$$

where $\{\epsilon_t\}$ are i.i.d. random errors with mean zero and variance σ_ϵ^2. Also, $m_t(\theta)$ and $\sigma_t(\theta)$ are specified $\mathcal{F}_{t-1}$-measurable functions. The model in (6.20) includes linear time series models such as ARMA, and nonlinear time series such as the threshold autoregressive (TAR) processes. In addition, (6.20) includes the conditionally heteroscedastic autoregressive (ARCH) models. Drost et al. (1997) have established the LAN property for the general class of models in (6.20) when the density of the errors $f_\epsilon(\cdot)$ is specified. The optimality properties of the ML estimator of θ and of related test statistics follow immediately.

Moreover, Drost et al. (1997) have studied adaptive estimation of θ when $f_\epsilon(\cdot)$ is unknown.

Ex. 4. Conditional exponential Markov processes

Let $\{X_t\}$ be a stationary ergodic Markov process with transition densities of the form

$$p(x_t, x_{t+1}; \theta) = h(x_t, x_{t+1}) \exp[\theta^{\mathrm{T}} \mathbf{Z}(x_t, x_{t+1}) - g(\theta, x_t)] \ , \tag{6.21}$$

where θ is a $(p \times 1)$ parameter, $\mathbf{Z}(\cdot)$ is a specified $p \times 1$ vector of statistics, and $g(\cdot)$ is a given real valued function. The likelihood equation based on $\mathbf{X}(n) = (X_1, \ldots, X_n)$ is given by

$$\sum_{t=1}^{n} \left(\frac{\mathrm{d}g(\theta, X_t)}{\mathrm{d}\theta} \right) - \sum_{t=1}^{n-1} \mathbf{Z}(X_t, X_{t+1}) = 0 \ . \tag{6.22}$$

Under regularity conditions, Hwang and Basawa (1994) have established the LAN property for the above model. If $\hat{\theta}_n$ is a consistent solution of the equation (6.22), it can be shown that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N_p(0, \mathbf{\Gamma}^{-1}(\theta)) \ , \tag{6.23}$$

where

$$\mathbf{\Gamma}(\theta) = E\left(\frac{\mathrm{d}^2 g(\theta, X_t)}{\mathrm{d}\theta \mathrm{d}\theta^{\mathrm{T}}} \right) \ , \tag{6.24}$$

the expectation being taken with respect to the stationary distribution. Hwang and Basawa (1994) have also discussed applications of the above model to several nonlinear time series examples.

Ex. 5. Random coefficient autoregressive processes

Suppose $\{X_t\}$ is a sequence of random variables defined by

$$X_t = H_\theta(X_{t-1}, Z_t) + \epsilon_t \ , \tag{6.25}$$

where $\{Z_t\}$ and $\{\epsilon_t\}$ are independent sequences of i.i.d. random variables (unobserved), and $H_\theta(\cdot)$ is a specified function. It then follows that $\{X_t\}$ is a Markov process with transition densities

$$p(x_t, x_{t+1}; \theta) = \int f_\epsilon(x_{t+1} - H_\theta(x_t, z_t)) g_Z(z) \mathrm{dz} \ , \tag{6.26}$$

where $f_\epsilon(\cdot)$ and $g_Z(\cdot)$ denote the densities corresponding to ϵ_t and Z_t respectively.

The above model includes the following special cases:

(i) **Random coefficient AR(1):** $H_\theta(x, y) = (\theta + y)x$.
(ii) **Threshold AR(1):** $H_\theta(x, y) = \theta_1 x^+ + \theta_2 x^-$, where $x^+ = \max\{0, x\}$, and $x^- = \min\{0, x\}$.
(iii) **Exponential AR(1):** $H_\theta(x, y) = [\theta_1 + \theta_2 \exp(-\theta_3 x^2)]x$.
(iv) **Random coefficient exponential AR(1):**

$$H_\theta(x, y) = [(\theta_1 + y_1) + (\theta_2 + y_2)\exp(-x^2)]x \ ,$$

with $y = (y_1, y_2)^{\mathrm{T}}$.
(v) **Random coefficient threshold AR(1):**

$$H_\theta(x, y) = (\theta_1 + y_1)x^+ + (\theta_2 + y_2)x^-, \quad \text{with } y = (y_1, y_2)^{\mathrm{T}} \ .$$

Hwang and Basawa (1993) have established the LAN property for the general class of random coefficient models defined by (6.25) and studied problems of inference regarding θ.

Ex. 6. The pure birth process

Let $\{X_t\}$, $t \geq 0$, be a pure birth process with birth rate θ, and $X_0 = 1$, where X_t denotes the population size at t. This is a continuous time Markov process. The intervals between births, $T_k = t_k - t_{k-1}$, $k = 1, 2, \ldots$ are independent exponential random variables with $E(T_k) = (k\theta)^{-1}$, where t_k denotes the epoch of the kth birth. Suppose we observe the process continuously over the interval $(0, T)$. Let $B(T)$ denote the total number of births occurring in the interval $(0, T)$. Note that $X_T = B(T) + 1$. The likelihood function is given by

$$p_T(x(0, T); \theta) = \left(\prod_{k=1}^{B(T)} k\theta \exp(-k\theta T_k) \right) \exp(-(T - t_{B(T)})\theta X_T) \ . \tag{6.27}$$

We then have

$$\frac{\mathrm{d}\log p_T}{\mathrm{d}\theta} = \frac{B(T)}{\theta} - \left[\sum_{k=1}^{B(T)} kT_k + (T - t_{B(T)})X_T\right],$$
$$= \frac{B(T)}{\theta} - \int_0^{\mathrm{T}} X_t \,\mathrm{d}t \ ,$$

and

$$-\frac{\mathrm{d}^2 \log p_T}{\mathrm{d}\theta^2} = \frac{B(T)}{\theta^2} \ .$$

The maximum likelihood estimator of θ is given by

$$\hat{\theta}_T = B(T) \Big/ \int_0^{\mathrm{T}} X_t \,\mathrm{d}t \ . \tag{6.28}$$

It can be shown that, as $T \to \infty$,

$$B^{1/2}(T)(\hat{\theta}_T - \theta) \xrightarrow{d} N(0, \theta^2) \ . \tag{6.29}$$

Here we have

$$B(T)/E(B(T)) \xrightarrow{p} W \ ,$$

where $W > 0$ is a non-degenerate random variable. See Keiding (1974) for details. Consequently, this example belongs to the LAMN family.

Ex. 7. Optimal estimating functions for longitudinal data

Let $\mathbf{X}_{it}$ denote the observation on the ith individual at time t, $i = 1, \ldots, m$ and $t = 1, \ldots, n_i$. Denote $\mathbf{X}_i = (X_{i1}, \ldots, X_{in_i})^{\mathrm{T}}$, the vector of observations on the ith individual. Assume that X_i are independent with

$$E(\mathbf{X}_i) = \mu_i(\boldsymbol{\beta}), \text{ and } \mathrm{Cov}(\mathbf{X}_i) = V_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) \ , \tag{6.30}$$

where $\boldsymbol{\beta}$ is the parameter vector of interest and $\boldsymbol{\alpha}$ is a vector of nuisance parameters. When α is known, Godambe's optimal estimating function for β is given by

$$g = \sum_{i=1}^{m} \frac{\partial \mu_i}{\partial \beta} V_i^{-1}(y_i - \mu_i) \ . \tag{6.31}$$

If $\hat{\beta}$ is a consistent solution of the equation $g = 0$, one can show, under regularity conditions, that

$$H_n^{1/2}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \mathbf{I}), \quad \text{as } n \to \infty \ , \tag{6.32}$$

where $n = \sum_{i=1}^{m} n_i$, and

$$H_n = \sum_{i=1}^{m} \left(\frac{d\mu_i}{d\beta}\right) V_i^{-1} \left(\frac{d\mu_i}{d\beta}\right)^{\mathrm{T}} . \tag{6.33}$$

See Fahrmeir and Kaufmann (1985) for the application of the above approach to the generalized linear model.

The nuisance parameter α can usually be estimated via ad hoc methods such as the method of moments or the least squares. See Liang and Zeger (1986), Prentice (1988), Liang et al. (1992) and Zhao and Prentice (1990) for various inference problems concerning longitudinal data.

Ex. 8. Bayes and empirical Bayes estimation for autoregressive processes

Let $\mathbf{X}_t(j)$ denote the observation on the jth individual at time t, $t = 1, \ldots, T$ and $j = 1, \ldots, n$. Consider the model

$$\mathbf{X}_t(j) = \phi_j \mathbf{X}_{t-1}(j) + \epsilon_t(j) \ , \tag{6.34}$$

where $\{\phi_j\}$ are assumed to be independent $N(\mathbf{a}_j^{\mathrm{T}}\boldsymbol{\beta}, \sigma_\phi^2)$, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of parameters and $\mathbf{a}_j$ are $(p \times 1)$ vectors of known covariates. It is assumed that $\{\epsilon_t(j)\}$ are independent $N(0, \sigma_\epsilon^2)$ random errors, which are independent of $\{\phi_j\}$. The conditional distribution (posterior distribution) of ϕ_n given $\mathbf{X}(n) = (X_1(n), \ldots, X_T(n))^{\mathrm{T}}$ is seen to be $N(\delta, \sigma_\epsilon^2 \sigma_\phi^2 / a)$, where

$$\delta = (\sigma_\epsilon^2 a^{-1}) a_n^{\mathrm{T}} \beta + (1 - \sigma_\epsilon^2 c^{-1}) \hat{\phi}_n \ , \tag{6.35}$$

$$c = \sigma_\epsilon^2 + \sigma_\phi^2 \sum_{t=1}^{\mathrm{T}} X_{t-1}^2(n) \ ,$$

and

$$\hat{\phi}_n = \frac{\sum_{t=1}^{\mathrm{T}} X_t(n) X_{t-1}(n)}{\sum_{t=1}^{\mathrm{T}} X_{t-1}^2(n)} \ .$$

It is assumed that $X_0(j) = 0$ for each j. If σ_ϵ^2, σ_ϕ^2 and β are known, δ in (6.35) is the Bayes estimator of ϕ_n with respect to the quadratic loss function. The estimator δ is based on the T observations on the nth individual. If σ_ϵ^2, σ_ϕ^2 and β are unknown they may be estimated from the marginal likelihood based on all the nT observations, $\{\mathbf{X}_t(j)\}$, $j = 1, \ldots, n$ and $t = 1, \ldots, T$. Let $\hat{\delta}$ denote δ after σ_ϵ^2, σ_ϕ^2 and β are replaced by their estimates. Then $\hat{\delta}$ is an empirical Bayes estimator of ϕ_n. See Kim and Basawa (1992) for the properties of $\hat{\delta}$.

Acknowledgement

We thank the referee for a careful reading and many useful suggestions.

References

Basawa, I. V. (1981a). Efficient conditional tests for mixture experiments with applications to the birth and branching processes. *Biometrika* **68**, 153–165.

Basawa, I. V. (1981b). Efficiency of conditional maximum likelihood estimators and confidence limits for mixtures of exponential families. *Biometrika* **68**, 515–523.

Basawa, I. V. (1983). Recent trends in asymptotic optimal inference for dependent observations. *Australian J. Statist.* **25**, 182–190.

Basawa, I. V. (1985). Neyman-LeCam tests based on estimating functions. *Proc. Berkeley Conference in Honor of Neyman and Kiefer* (Eds., L. LeCam and R. Olshen), Wadsworth, Belmont, pp. 811–825.

Basawa, I. V. (1990). Large sample statistics for stochastic processes: Some recent developments. *Proc. of R.C. Bose Symposium: Probability, Statistics and Design of Experiments* (Ed., R. R. Bahadur) pp. 107–122. Wiley Eastern, New Delhi.

Basawa, I. V. (1991). Generalized score tests for composite hypotheses. In *Estimating Functions* (Ed., V. P. Godambe) pp. 121–132. Oxford Univ. Press.

Basawa, I. V. and P. J. Brockwell (1984). Asymptotic conditional inference for regular nonergodic models with application to autoregressive processes. *Ann. Statist.* **12**, 161–171.

Basawa, I. V. and H. L. Koul (1988). Large sample statistics via quadratic approximation. *Inter. Statist. Reviews* **56**, 199–219.

Basawa, I. V. and B. L. S. Prakasa Rao (1980a). *Statistical Inference for Stochastic Processes.* Academic Press, London.

Basawa, I. V. and B. L. S. Prakasa Rao (1980b). Asymptotic inference for stochastic processes. *Stoch. Proc. and Appl.* **9**, 291–305.

Basawa, I. V. and D. J. Scott (1983). *Asymptotic Optimal Inference for Nonergodic Models.* Lecture Notes in Statistics, Vol 17. Springer-Verlag, New York.

Basawa, I. V., R. Huggins and R. G. Staudte (1985). Robust tests for time series with an application to first order autoregressive processes. *Biometrika* **72**, 559–571.

Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10**, 647–671.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov and J. A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins Univ. Press, Baltimore.

Billingsley, P. (1961). *Statistical Inference for Markov Processes.* Univ. Chicago Press, Chicago.

Drost, F. C., C. A. J. Klaassen and B. J. M. Werker (1997). Adaptive estimation in time series models. *Ann. Statist.* **25**, 786–817.

Fahrmeir, L. and H. Kaufmann (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.* **13**, 342–368.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31**, 1208–1212.

Godambe, V. P. (1985). The foundations of finite sample estimation in stochastic processes. *Biometrika* **72**, 419–428.

Godambe, V. P. (1991). (Ed.) *Estimating Functions.* Oxford Univ. Press, Oxford.

Guttorp, P. (1991). *Statistical Inference for Branching Processes.* Wiley, New York.

Hall, W. J. and D. J. Mathiason (1990). On large sample estimation and testing in parametric models. *Inter. Statisti. Reviews* **58**, 77–97.

Heyde, C. C. (1975). Remarks on efficiency in estimation for branching processes. *Biometrika* **62**, 49–55.

Heyde, C. C. (1997). *Quasilikelihood and Its Applications.* Springer, New York.

Hwang, S. Y. and I. V. Basawa (1993). Asymptotic optimal inference for a class of nonlinear time series. *Stoch. Proc. and Appl.* **46**, 91–113.

Hwang, S. Y. and I. V. Basawa (1994). Large sample inference for conditional exponential families with applications to nonlinear time series. *J. Statist. Plann. Inf.* **38**, 141–158.

Kim, Y. W. and I. V. Basawa (1992). Empirical Bayes estimation for first order autoregressive processes. *Austral. J. Statist.* **34**, 105–114.

Keiding, N. (1974). Estimation in the birth process. *Biometrika* **61**, 71–80.
LeCam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.
LeCam, L. and G. Yang (1990). *Asymptotics in Statistics: Some Basic Concepts*. Springer-Verlag, New York.
Liang, K. Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.
Liang, K. Y., S. L. Zeger and B. Qaqish (1992). Multivariate regression analysis for categorical data (with discussion). *J. Roy. Statist. Soc. Ser B* **54**, 3–40.
Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033–1048.
Roussas, G. G. (1972). *Contiguity of Probability Measures*. Cambridge Univ. Press, Cambridge.
Zhao, L. P. and R. L. Prentice (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 642–648.