The Annals of Statistics
1986, Vol. 14, No. 1, 88–123

THEORY OF PARTIAL LIKELIHOOD[1]

BY WING HUNG WONG

University of Chicago

A general asymptotic theory is developed for the maximum likelihood estimator based on a partial likelihood. Conditions are given for consistency and asymptotic normality, and a method is provided for the calculation of the asymptotic efficiency of the estimator. The implications of the general theory are examined in special cases such as inference in stochastic processes, Cox regression models, and AR processes with missing segments.

1. Introduction. Let y be a realization of the random vector Y with density $f_Y(y; \phi)$ depending on a vector parameter $\phi = (\theta; \eta)$. Consider the situation where Y, perhaps after transformation, can be partitioned into two components, $Y = (V, W)$. Correspondingly, the full likelihood will then factorize into a marginal likelihood and a conditional likelihood:

$$f_Y(y; \phi) = f_V(v; \phi) f_{W|V}(w|v; \phi). \tag{1.1}$$

We will suppose we are interested only in inference about θ; η will play the role of a nuisance parameter. In complicated problems, the dimension of η may be high, and the application of maximum likelihood estimation may lead to misleading results. If in such situations there is a factorization (1.1) such that only one of the two factors involves θ, then it is often helpful to use just that factor and disregard the other (which involves η). Examples and development of marginal and conditional likelihood methods can be found in Kalbfleisch and Sprott (1970) and Andersen (1973).

It is clear that if one uses only one factor in (1.1) while the ignored factor involves both θ and η, then one has not used the full information (about θ) contained in the observations. In exchange for the loss of information we achieve simplicity in analysis resulting from the elimination of nuisance parameters. There is also a gain in robustness of validity because the ignored factor in the likelihood does not have to be specified precisely. In applications these tradeoffs must be weighed carefully.

In the development of methods based on appropriate factorization of the full likelihood, the introduction by Cox (1975) of the concept of partial likelihood is an important milestone. Consider the case when Y can be transformed into a sequence,

$$\mathbf{y} = (w_1, x_1, \ldots, w_n, x_n, \ldots, w_N, x_N), \tag{1.2}$$

the partition being chosen so that the density of x_n conditional on all the

Received March 1984; revised June 1985.

[1]Support for this research was provided by National Science Foundation Grant No. MCS-8301459.

AMS 1980 *subject classifications*. Primary 62A10, 62F12; secondary 62M10, 62P10.

Key words and phrases. Maximum likelihood estimator, nuisance parameter, Minimal Fisher Information, martingale limit theorem, missing values, generalized autoregression, nonstationarity, proportional hazard model, conditional score.

previous variables depends only on θ. The full likelihood factorizes accordingly into

$$f_\phi(\mathbf{y}) = \left[\prod_{n=1}^{N} f_\phi(w_n|d_n)\right]\left[\prod_{n=1}^{N} f_\theta(x_n|c_n)\right], \tag{1.3}$$

where $d_n = (w_1, x_1, \ldots, w_{n-1}, x_{n-1})$, $c_n = (w_1, x_1, \ldots, w_{n-1}, x_{n-1}, w_n)$. Cox called the second product in the right-hand side of (1.3) the partial likelihood of θ based on X in the sequence $(w_n, x_n)_{n=1,\ldots,N}$.

EXAMPLE. Suppose we observe J disconnected segments of a Markov process $[z_{n_1} \cdots z_{m_1}], [z_{n_2} \cdots z_{m_2}], \ldots, [z_{n_J} \cdots z_{m_J}]$, so z's between z_{m_1} and z_{n_2}, etc. are "missing." The joint density for the data is given by (taking z_{m_0} = constant)

$$\prod_{j=1}^{J}\left[f(z_{n_j}|z_{m_{j-1}}) \prod_{n=n_j+1}^{m_j} f(z_n|z_{n-1})\right]. \tag{1.4}$$

Suppose that within the observed segments the one-step transition probabilities are given by $p_\theta(x, y)$ with a parameter θ. If we let $w_1 = z_{n_1}$, $x_1 = [z_{n_1+1}, \ldots, z_{m_1}]$, $w_2 = z_{n_2}$, $x_2 = [z_{n_2+1}, \ldots, z_{m_2}]$, etc., then the partial likelihood based on X is simply $\prod_{j=1}^{J}[\prod_{n=n_j+1}^{m_j} p_\theta(z_{n-1}, z_n)]$. In many situations the missing segments may have a different probabilistic structure from the observed ones. In such cases, the remaining products in (1.4), namely $\prod_{j=1}^{J} f(z_{n_j}|z_{m_{j-1}})$, will typically be difficult to handle because of nuisance parameters. This example is discussed further in Section 6.1.

The purpose of the present paper is to develop an asymptotic theory for maximum likelihood estimation based on a partial likelihood. The advantage of being able to use a partial likelihood is clear: one then has great flexibility in obtaining a factorization (1.3) such that the loss of information due to the ignored factor is small.

An equivalent way to define a partial likelihood is as a product $\prod_{n=1}^{N} f_\theta(x_n|c_n)$ of the densities of the conditional experiments $x_1|c_1, x_2|c_2, \ldots, x_n|c_n$, where the σ-field generated by c_{n+1} contains that generated by c_n, $n = 1, 2, \ldots$. This *nested conditioning requirement* plays a key role in our development of the asymptotic theory: it implies that the scores constructed from the conditional densities form a martingale differences sequence. This means that the information contained in the different conditional experiments are not redundant. Now, the likelihood for a stochastic process $x_1, \ldots, x_N$ can always be written as a partial likelihood. Thus the MLE theory for stochastic processes is a nontrivial special case of the partial likelihood setting. This area has, of course, been studied extensively; see Billingsley (1961), Silvey (1961), Rao (1966), Bhat (1974), Crowder (1976), Caines (1975), Basawa, Feigin, and Heyde (1976), Hall and Heyde (1980), and Basawa and Rao (1980). Billingsley (1961) had already recognized the importance of the martingale differences structure of the conditional scores and proved asymptotic normality of the MLE using a martingale central limit theorem developed by himself. Our treatment of asymptotic normality in Section 4 for the partial likelihood MLE is an extension of the treatment used in the above mentioned

works. On the other hand, our approach to the consistency problem is quite different from the one used in the above works. To draw an analogy with the two main approaches to consistency in classical MLE theory in the i.i.d. setting, the approach used in the above works (except for Caines, who treats only a finite parameter space) is the Cramér approach (Cramér, 1946), namely, to exhibit a consistent sequence of solutions to the likelihood equation, while the approach adopted in the present paper is that of Doob (1934) and Wald (1949). The unified treatment for finite and continuum parameter spaces and the avoidance of differentiability and uniqueness conditions are two main advantages of the Doob–Wald approach. This consistency theory is developed in Section 2. The theory is illustrated in Section 3 by several examples. The example of the Cox model is included for obvious reasons. The generalized autoregression models are studied partly because of their potential usefulness in modelling nonnormal time series data.

Section 5 deals with the theory of efficiency. Our starting point is the classical work of Bahadur (1964), Hajek (1970), and others on the Fisher lower bound for regular estimators under the locally asymptotically normal (LAN) condition of Le Cam (1960). This theory is nicely summarized in the recent monograph of Ibragimov and Hásminskii (1981). The main problem we face is how to handle the nuisance parameter, particularly if it is infinite dimensional. The criterion for efficiency adopted in this paper is an extension of the classical one for the finite dimensional case. It is also used by some previous authors, e.g., Stein (1956), Lindsay (1980), and Begun, Hall, Huang, and Wellner (1983), mainly in the case of i.i.d. observations. In Section 5.2 we present a simple argument showing why the minimal Fisher information (Lindsay, 1980) provides a lower bound for the asymptotic variance of regular estimates in the general case. This argument represents a precise formulation of Stein's argument which predates the rigorous theories of Bahadur and Hajek. We also study the method of calculating the minimal Fisher information by taking the limit in a sequence of finite parameterizations and provide conditions under which the method works. At present this seems to be the only systematic method of calculation in the general case. Although it is believed that in most cases the minimal Fisher information is an appropriate benchmark for efficiency comparison, this has never been rigorously established. We have only been able to provide a partial justification, as discussed at the end of Section 5.2.

The efficiency theory is illustrated in Section 6 by an in-depth study of the example of the segmented AR process. Most of the general points made in Section 5 find concrete representation in this example. The calculation shows that the partial likelihood is not fully informative in the random level shift case, even if the shift density is assumed unknown. This result is, at least initially, rather unexpected. Another example studied in Section 6 is the Cox regression model.

Consistency and asymptotic normality in the Cox regression have been intensively studied in recent years: Tsiatis (1981), Liu and Crowley (1978), Andersen and Gill (1982), Slud (1982), Bailey (1983), and Prentice and Self (1983). In this paper we use the Cox model mainly as an illustration for the general theory. We

assume that the covariates are nonrandom, time independent, and can take only a finite number of values. In this respect our treatment is more restricted than some earlier ones. On the other hand, an arbitrary risk of the form $w(\theta, z)$ is allowed rather than the usual choice $e^{\theta \cdot z}$ used by all authors except Prentice and Self (1983), who allow a form $r(\theta \cdot z)$ where r is a general function. Within our model we give explicit calculations of the asymptotic distributions. It is hoped that the ordinary differential equations derived in the appendix are a useful addition to the literature.

Our calculation of efficiency in the Cox model reproduces some of the results of Efron (1977) and Oakes (1977), but our method is entirely different from theirs. Although implicit in Efron (1977), the rigorous derivation of asymptotic efficiency of the partial likelihood in the Cox model has been given only recently by Begun et al. (1983) under the i.i.d. covariate assumption; see also Pfanzagl (1982). Our derivation in this paper is under the assumption of nonrandom covariates. The discussion also appears to be the first systematic treatment of efficiency for general risk forms.

2. Consistency.

2.1. Finite parameter space. In the classical framework of i.i.d. observations where $x_1, x_2, \ldots, x_N$ are i.i.d. with common density $f(x; \theta_0)$; for any fixed $\theta \neq \theta_0$, if $r_n(\theta) = \log(f_{\theta_0}(x_n)/f_\theta(x_n))$, then the Kullback–Leibler discriminatory information between θ_0 and θ contained in x_n is given by

$$i_n(\theta) = E_{\theta_0}(r_n(\theta)) = \int f_{\theta_0} \log(f_{\theta_0}/f_\theta)\, dx > 0.$$

The variance $j_n(\theta) = \mathrm{Var}_{\theta_0}(r_n(\theta))$ is also independent of n. Hence for each $\theta \neq \theta_0$, we have

$$I_N(\theta) = \sum_1^N i_n(\theta) = N i_1(\theta) \to \infty, \tag{2.1}$$

$$J_N(\theta) = \sum_1^N j_n(\theta) = N j_1(\theta) = o\big(I_N^2(\theta)\big). \tag{2.2}$$

To see the meaning of these two conditions, denote by $R_N(\theta)$ the logarithm of the likelihood-ratio, i.e., $R_N(\theta) = \log((\prod_1^N f_{\theta_0}(x_n))/(\prod_1^N f_\theta(x_n))) = \sum_1^N r_n(\theta)$, then the Kullback–Leibler discriminatory information contained in $x_1, \ldots, x_N$ is $E_{\theta_0} R_N(\theta) = I_N(\theta)$, and the variance of $R_N(\theta)$ is just $J_N(\theta)$. Clearly, (2.1) and (2.2) imply the divergence (to infinity) of $R_N(\theta)$, which in turn implies the consistency of the MLE if the parameter space Θ is finite.

Now the basic structure of the partial likelihood framework is a sequence of conditional experiments $x_1|c_1, \ldots, x_N|c_N$, where the σ-field generated by c_{n+1} contains that generated by c_n. In each experiment, c_n is regarded as fixed and x_n has a conditional density $f_\theta(x_n|c_n)$, the partial likelihood is nothing but the product of such conditional likelihoods (defined up to a multiplicative constant independent of θ), i.e., $\mathrm{PL} \propto \prod_{n=1}^N f_\theta(x_n|c_n)$.

To extend the above argument, let

$$r_n(\theta) = \log\big(f_{\theta_0}(x_n|c_n)/f_\theta(x_n|c_n)\big), \qquad R_N = \sum_1^N r_n,$$

$$i_n(\theta) = E_{\theta_0}(r_n(\theta)|c_n), \qquad I_N = \sum_1^N i_n,$$

$$j_n(\theta) = \operatorname*{Var}_{\theta_0}(r_n(\theta)|c_n), \qquad J_N = \sum_1^N j_n,$$

$$m_n(\theta) = r_n(\theta) - i_n(\theta), \qquad M_N = \sum_1^N m_n(\theta).$$

Note that R_N is the logarithm of the partial-likelihood ratio, i.e., $R_N(\theta) = \log(\mathrm{PL}(\theta_0)/\mathrm{PL}(\theta))$. Conditional on c_n, the discriminatory information (between θ_0 and θ) contained in x_n is just $i_n(\theta)$; for this reason the sum $I_N = \Sigma_1^N i_n$ will be called the *accumulated Kullback–Leibler information* in the sequence of conditional experiments $x_1|c_1, x_2|c_2, \ldots, x_n|c_n$. In general, both the conditional information $i_n(\theta)$ and the conditional variance $j_n(\theta)$ are random variables; they reduce to constants when x_n is independent of c_n.

When the parameter space Θ is finite, the divergence of $R_N(\theta)$ for every $\theta \neq \theta_0$ implies the consistency of $\hat{\theta}$, the value that maximizes the partial likelihood. The following theorem provides sufficient conditions for the divergence of R_N.

THEOREM 2A. *Suppose $\mathcal{F}_1, \mathcal{F}_2, \ldots$ is a sequence of increasing σ-fields, $R_N = \Sigma_1^N r_n$ and, for $n = 1, 2, \ldots$, r_n is measurable with respect to $\mathcal{F}_n$, $i_n = E(r_n|\mathcal{F}_{n-1})$, $j_n = \mathrm{Var}(r_n|\mathcal{F}_{n-1})$. If there exist constants $\delta > 0$, $\alpha_n \uparrow \infty$ such that*

$$P(I_N/\alpha_N > \delta) \to 1 \tag{2.3}$$

$$J_N/\alpha_N^2 \to_P 0 \tag{2.4}$$

then $R_N/I_N \to_P 1$. If only (2.4) holds, then $\alpha_N^{-1}M_N = \alpha_N^{-1}(R_N - I_N) \to_P 0$.

PROOF. Let $A_N = \{I_N/\alpha_N > \delta\}$, then

$$P(|R_N - I_N|/I_N > \varepsilon) \leq P(|R_N - I_N| > \varepsilon\delta\alpha_N) + P(A_N^c).$$

The second term goes to zero by (2.3). To estimate the first term, write

$$R_N - I_N = \sum_{n=1}^N m_n,$$

and let $m_{Nn}^* = m_n\chi(J_n/\alpha_N^2 < 1)$ where $\chi(\cdot)$ is the indicator function. Since $P\{m_{Nn}^* = m_n, \ \forall \ n = 1, \ldots, N\} \geq P\{J_N/\alpha_N^2 < 1\} \to 1$, it suffices to estimate

$P(|\sum_{n=1}^{N} m_{Nn}^*| > \varepsilon\delta\alpha_N)$. Now,

$$E(m_{Nn}^*|\mathscr{F}_{n-1}) = \chi\big(J_n/\alpha_N^2 < 1\big)E(m_n|\mathscr{F}_{n-1}) = 0,$$

$$\operatorname{Var}(m_{Nn}^*|\mathscr{F}_{n-1}) = \chi\big(J_n/\alpha_N^2 < 1\big)\operatorname{Var}(m_n|\mathscr{F}_{n-1}),$$

$$J_N^* = \sum_{n=1}^{N} \operatorname{Var}(m_{Nn}^*|\mathscr{F}_{n-1}) \le J_n.$$

Hence, $\alpha_N^{-2}J_N^* \le \alpha_N^{-2}J_N \to_P 0$. Thus, using boundedness of J_N^*/α_N^2, one obtains

$$P\left(\left|\sum_{n=1}^{N} m_{Nn}^*\right| > \varepsilon\delta\alpha_N\right) \le \frac{1}{\varepsilon^2\delta^2}\operatorname{Var}\left(\sum_{n=1}^{N} m_{Nn}^*/\alpha_N\right) = \frac{1}{\varepsilon^2\delta^2}E\big(J_N^*/\alpha_N^2\big) \to 0. \quad \square$$

In the above theorem, since $\alpha_N \to \infty$, (2.3) and (2.4) implies

$$I_N \to_P \infty \tag{2.5}$$

and

$$J_N = o_P\big(I_N^2\big). \tag{2.6}$$

The conditions (2.5) and (2.6) seem to be the natural extension of conditions (2.1) and (2.2); however, until now a proof of the divergence of R_N under (2.5) and (2.6) has not been obtained. In any case, the meaning of the conditions is clear: (2.1), (2.3), or (2.5) represent "accumulation of information," (2.2), (2.4), or (2.6) represent "stability of variance." Note that for $\hat{\theta}$ to be consistent, the conditions in Theorem 2A must be satisfied by $I_N(\theta)$, $J_N(\theta)$ for each $\theta \neq \theta_0$ in the finite parameter space. The constants α_N may depend on θ as long as $\alpha_N(\theta) \uparrow \infty$ for each θ.

In some applications, such as the Cox regression model discussed below in Section 3, it is necessary to formulate the partial likelihood in terms of triangular arrays, i.e., for each N, there are conditional experiments $x_1^{(N)}|c_1^{(N)}$, $x_2^{(N)}|c_2^{(N)}, \ldots, x_N^{(N)}|c_N^{(N)}$, but for the same n, $x_n^{(N)}$ and $c_n^{(N)}$ need not be the same as $x_n^{(N')}$ and $c_n^{(N')}$ if $N \neq N'$. For the discussion of weak consistency or asymptotic distribution, it does not matter whether the array is single or triangular. For simplicity we will always write x_n and c_n, with the understanding that in the triangular array case x_n and c_n may depend on N.

In the single array case, the following result of Neveu (1965, page 148) is relevant for strong consistency.

LEMMA 2B. *With R_N, I_N, J_N as in Theorem* 2A, $(R_N - I_N)/I_N \to 0$ *almost surely on the set* $\Omega_1 = \{I_N \uparrow \infty, \sum_{n=1}^{\infty}(j_n/I_n^2) < \infty\}$.

To apply this to partial likelihood, we must take $r_n = r_n(\theta) = \log(f_{\theta_0}(x_n|c_n)/f_\theta(x_n|c_n))$, hence Ω_1 in Lemma 2B may depend on θ. If $P(\Omega_1(\theta)) = 1$ for all $\theta \neq \theta_0$ in the finite parameter space, then by Lemma 2B we have $P(\min_{\theta \neq \theta_0} R_N(\theta) \to \infty) = 1$, from which strong consistency of $\hat{\theta}_N$ follows readily.

In the classical case of independent observations, both i_n, j_n are constants, so $P(\Omega_1(\theta))$ is either 1 or 0. In the general case of partial likelihood, the whole range of values in $[0,1]$ is possible for $P(\Omega_1(\theta))$. Likewise, the set $\Omega_2 = \{\hat{\theta}_N \to \theta_0\}$ may have probability other than 0 or 1. We call Ω_2 the *consistency set*, and its probability $P(\Omega_2)$ the *level of consistency*. By Lemma 2B, a lower bound for the level of consistency is $P(\bigcap_{\theta \in \Theta} \Omega_1(\theta))$.

2.2. Compact parameter space. In this section, we assume the parameter space to be compact. Extension to a general parameter space will be taken up in the next section.

THEOREM 2C. *Suppose Θ is compact, and suppose that for any $\theta \neq \theta_0$, there exists an open neighborhood O_θ of θ whose closure G_θ does not contain θ_0, and that there are constants $\delta > 0$, $\alpha_N \uparrow \infty$ (which may depend on θ) such that*

$$P\Big(\inf_{\theta' \in G_\theta} I_N(\theta')/\alpha_N > \delta\Big) \to 1, \tag{2.7}$$

$$J_N(\theta')/\alpha_N^2 \to_P 0 \quad \text{for all } \theta' \in G_\theta, \tag{2.8}$$

(2.9) *The distribution of $\alpha_N^{-1} M_N(\theta')$ is tight in $C(G_\theta)$, where $M_N = R_N - I_N$ and $C(G_\theta)$ is the space of continuous functions on G_θ.*

Then $\hat{\theta}_N \to \theta_0$.

PROOF. (i) First we show that (2.7)–(2.8) implies that

$$P\Big(\inf_{\theta' \in G_\theta} R_N(\theta') \leq 0\Big) \to 0.$$

To see this, observe that by the argument in the proof of Theorem 2A, (2.7) and (2.8) together imply that the finite dimensional distributions of $\alpha_N^{-1} M_N(\cdot)$ converge to those of the random function degenerate at 0. Hence under the tightness condition (2.9), $\alpha^{-1} M_N(\cdot) \to 0$ weakly in $C(G_\theta)$, yielding the desired result.

(ii) To prove the theorem, let O_{θ_0} be any open neighborhood of θ_0, and consider the compact set $\Theta \setminus O_{\theta_0}$. By compactness $\Theta \setminus O_{\theta_0}$ can be covered by a finite number of open sets $O_{\theta_1}, \ldots, O_{\theta_k}$, each of which satisfies conditions (2.7)–(2.9). Hence by part (i) of this proof,

$$P\Big(\inf_{\theta' \in G_{\theta_i}} R_N(\theta') \leq 0\Big) \to 0 \quad \text{for } i = 1, \ldots, k,$$

whence

$$P\Big(\inf_{\theta' \notin O_{\theta_0}} R_N(\theta') \leq 0\Big) \leq \sum_{i=1}^{k} P\Big(\inf_{\theta' \in O_{\theta_i}} R_N(\theta') \leq 0\Big) \to 0 \tag{2.10}$$

as $N \to \infty$, giving the desired result. □

In typical situations the parameter space Θ is also endowed with a metric or a linear structure. If Θ has a natural metric, it is convenient, and we will always do

so, to take O_θ in Theorem 2C to be an open ball centered at θ. If further, Θ is a subset of a normed linear space, then the following criterion for tightness is useful.

LEMMA 2D. *If Θ is a compact subset of a normed linear space, then condition* (2.9) *in Theorem* 2C *can be replaced by*

(2.9a) *With probability* 1, $M_N(\theta)$ *has Frechet derivative* $\nabla M_N(\theta)$ *such that for some constant* $K > 0$, $P\left(\sup_{\theta' \in G_\theta} \alpha_N^{-1} \|\nabla M_N(\theta')\| > K \right) \to 0.$

PROOF. Since G_θ is taken to be a closed ball, any intermediate value between θ_1 and θ_2 also lies in G_θ if θ_1 and θ_2 are in G_θ. Hence under our assumptions tightness follows readily from the intermediate value theorem and Theorem 8.2 of Billingsley (1968). □

2.3. General parameter space. To cover more general parameter spaces, a typical approach, introduced first in Wald (1949), is to consider conditions that guarantee that $\hat{\theta}$ will eventually be confined to a compact subset of Θ. These types of conditions can be called conditions of "essentially compact parameter space." In this section we consider the condition

(2.11) There exists a compact subset K of Θ, such that $\theta_0 \in$ interior of K, and $P\left(\inf_{\theta \notin K} R_N(\theta) \le 0 \right) \to 0.$

It is clear that under this condition, if the local conditions (2.7)–(2.9) of Theorem 2C are satisfied for every $\theta \neq \theta_0$ in K, then we still have $\hat{\theta}_N \to_P \theta_0$. The proof is a straightforward extension of that of Theorem 2C.

The following theorem concerns a special case, covering a variety of applications, in which essential compactness is automatically satisfied.

THEOREM 2E. *Let Θ be a convex set in R^p, $\theta_0 \in$ interior of Θ, and $L_n(\theta)$ the logarithm of the partial likelihood. If the local conditions* (2.7)–(2.9) *are satisfied for all $\theta \neq \theta_0$, and*

(2.12) $P(L_N(\theta)$ *is strictly concave in* $\theta) = 1$ *for all* N,

then (2.11) *is also true, and hence $\hat{\theta}_N \to_P \theta_0$.*

PROOF. Let O_1 and O_2 be open balls centered at θ_0 with radius ρ_1 and ρ_2, respectively; ρ_1 and ρ_2 are chosen such that $\rho_1 < \rho_2$ and $O_2 \subset$ interior of Θ. Let G_2 denote the closure of O_2 and $\tilde{\theta}_N$ denote the θ in G_2 that maximizes $L_N(\theta)$.

Since $R_N(\theta) = L_N(\theta_0) - L_N(\theta)$, we have

$$\left\{ \inf_{\theta \notin G_2} R_N(\theta) \le 0 \right\} \subset \{L_N(\theta^*) \ge L_N(\theta_0) \text{ for some } \theta^* \notin G_2\}$$

$$\overset{\text{by concavity}}{\subset} \{L_N(\theta^{**}) \ge L_N(\theta_0) \text{ for some } \theta^{**} \in G_2 \setminus O_1\}.$$

This last set is seen to have probability tending to zero by applying the result (2.10) with G_2 as the parameter space. □

Two remarks about the conditions: (i) the condition (2.12) can obviously be relaxed to require only that $P(L_N$ strictly concave$) \to 1$; (ii) it is clear that the same proof will go through if Θ is a convex set in a topological vector space and θ_0 belongs to the interior of a compact subset G_2 in Θ. However, such a generalization is only superficial—if every point in a topological vector space is required to have an open neighborhood with compact closure, then the topological vector space must be finite dimensional.

An important class of models where the concavity condition (2.12) holds is the class of natural exponential families

$$f_\theta(x_n|c_n) = h_n(x_n)e^{x_n\cdot\theta - b_n(\theta)} \quad \text{w.r.t. a measure } \nu_n. \tag{2.13}$$

The functions h_n, b_n, the measure ν_n, and even the range of x_n may all depend on c_n. But for any given c_n,

$$b_n(\theta) = \log\int h_n(x)e^{x\cdot\theta}\,d\nu_n(x)$$

is clearly a strictly convex function of θ if $h_n(x)\,d\nu_n(x)$ is not a degenerate distribution. Hence $L_N(\theta) = (\Sigma_1^N x_n)\cdot\theta - \Sigma_1^N b_n(\theta)$ is strictly concave if $h_n(x)\,d\nu_n(x)$ is not degenerate for at least one $n \le N$.

To apply Theorem 2E, one must also check the local conditions (2.7)–(2.9). As will be seen shortly, the verification of (2.7)–(2.8) for nonstationary cases can involve considerable work in each specific model. We now argue that for the natural exponential family model, the tightness condition (2.9) is automatically satisfied whenever condition (2.8) is satisfied. To see this, first use properties of exponential families to check that $E((\partial/\partial\theta_i)m_n(\theta)|c_n) = 0$ and $\mathrm{Var}((\partial/\partial\theta_i)m_n(\theta)|c_n) = \mathrm{Var}(x_{n_i}|c_n)$, where x_{n_i} is the ith component of x_n. If (2.8) is true then certainly $\alpha_N^{-2}\Sigma_{n=1}^N \mathrm{Var}(x_{n_i}|c_n) \to_p 0$, which is sufficient for $\alpha_N^{-1}(\partial/\partial\theta_i)M_N(\theta) = \alpha_N^{-1}\Sigma_{n=1}^N(\partial/\partial\theta_i)m_n(\theta) \to_P 0$. Thus condition (2.9a), and hence condition (2.9), are satisfied.

3. Examples.

3.1. Generalized autoregression. The normal theory linear model with dependent variable x and regressors $z^1,\ldots,z^p$ can be written as $x_1,\ldots,x_N \sim$ independent normals, $Ex_n = \gamma_n$, $\mathrm{Var}\,x_n = \sigma^2$, where $\gamma_n = \Sigma_{i=1}^p \theta_i z_n^i$. To handle time series data, let the regressors be lagged variables. Then we have the normal autoregressive model: given $x_1,\ldots,x_{n-1}$, x_n is normal with mean $\gamma_n = \Sigma_{i=1}^p \theta_i x_{n-i}$ and variance σ^2.

The distributional assumption of the linear model can be relaxed, normality may be substituted by any location scale family (with second moments) without affecting the asymptotic distribution of the least-squares estimates for θ. The constant variance assumption is, however, quite crucial; for this reason the linear model is not appropriate for most discrete data. For example, for binary data x it

is found useful to consider models in which the Probit or Logit of x_n depend linearly on the regressor $\mathbf{z}_n$. By allowing a suitable parameter of the distribution of x_n to depend linearly on $\mathbf{z}_n$, the scope of the model can be extended to cover diverse types of data. Such "generalized linear models" are typically applied to situations where $x_1, \ldots, x_n$ are independent, and parameters are typically estimated by maximum likelihood [see McCullagh and Nelder (1983) for developments of these models]. It seems natural that, to handle time series data, we can choose the regressors to be lagged values of x_n. The resulting model will be called a "generalized autoregressive model." It bears the same relation to the generalized linear model as that of the (normal) autoregressive model to the (normal) linear model.

If we restrict attention to cases where the conditional distribution of x_n given $c_n = (x_1, \ldots, x_{n-1})$ belongs to an exponential family, the generalized autoregressive process can be written in the form:

$$f_\theta(x_n|c_n) = h_n(x_n)\exp\{x_n \cdot \gamma_n(\theta) - b_n(\theta)\},$$

where

$$\gamma_n(\theta) = \gamma(\eta_n(\theta)), \qquad b_n(\theta) = b(\gamma_n(\theta)),$$

and

$$\eta_n(\theta) = \theta_0 + \sum_{i=1}^{p} \theta_i x_{n-i}; \tag{3.1}$$

γ, b are known functions. Some special cases are listed in Table 3.1.

The conditions for ergodicity in Table 3.1 will be derived in Appendix A.1. Under ergodicity it is possible to obtain fairly general conditions for consistency, as we now discuss. By familiar results for exponential families,

$$\begin{aligned} &\text{(a)} \quad r_n = -x_n \Delta\gamma_n + (b_n - b_n^0), \\ &\text{(b)} \quad i_n = E(r_n|c_n) = b_n - b_n^0 - b'(\eta_n^0)\Delta\gamma_n, \\ &\text{(c)} \quad j_n = \operatorname{Var}(r_n|c_n) = b''(\eta_n^0)(\Delta\gamma_n)^2, \\ &\text{(d)} \quad m_n = -(x_n - b'(\gamma_n^0)) \cdot \Delta\gamma_n, \end{aligned} \tag{3.2}$$

where $\Delta\gamma_n = \gamma_n - \gamma_n^0$ and the superscript ^{0} denotes evaluation at the true value.

TABLE 3.1

Conditional distribution	$b(\gamma)$	$b''(\gamma)$	$\gamma(\eta)$	Region of ergodicity ($p = 2$ case)
1. Normal	$\frac{1}{2}\gamma^2$	1	η	$\theta_0 = 0, 1 - \theta_1 B - \theta_2 B^2$ must have roots outside unit circle
2. Bernoulli	$\ln(1 + e^\gamma)$	$\left(\frac{e^\gamma}{1+e^\gamma}\right)\left(1 - \frac{e^\gamma}{1+e^\gamma}\right)$	$\ln\left(\frac{\eta}{1-\eta}\right)$	a bounded polygon (see Appendix A.1)
3. Poisson	e^γ	e^γ	$\ln(1 - e^{-\eta})$	$\theta_0, \theta_1, \theta_2 > 0$ is sufficient

We assume the following smoothness condition on γ and b:

(3.3) $|\gamma'(\cdot)|$ and $b''(\gamma(\cdot))$ are uniformly bounded away from 0 and ∞, for all possible values of $\eta = \theta_0 + \sum_1^p \theta_i x_{n-i}$.

Under this condition the asymptotic behavior of I_N and J_N depends only on that of $\Sigma_1^N(\Delta\eta_n)^2$. To satisfy conditions (2.7)–(2.8) in the consistency theorem (Theorem 2C), it suffices to find constants $\alpha_N \to \infty$ such that, for $\theta \neq \theta^0$,

(3.4) $\alpha_N^{-2}\sum_1^N(\Delta\eta_n)^2 \to_P 0$ but $\alpha_N^{-1}\sum_1^N(\Delta\eta_n)^2$ is locally uniformly bounded away from zero.

If we define $\mathbf{Y}_n = (x_{n-p}, \ldots, x_{n-1})$, then $\{\mathbf{Y}_n\}_{n=1,2,\ldots}$ is clearly a Markov process with stationary transition function. The process $\{x_n\}$ is said to be *ergodic* if $\{\mathbf{Y}_n\}$ is indecomposable and admits a strictly positive probability density *invariant* under the transition function.

LEMMA 3A. *Suppose that the generalized autoregressive process $\{x_n\}$ is ergodic and that $E^*x_{n-i}x_{n-j}$ exist for $1 \le i \le j \le p$, where E^* denotes expectation with respect to the invariant distribution, then condition* (3.4) *holds for any initial probability density.*

PROOF. (i) First we show that a law of large numbers applies: i.e., for any measurable function $g(\mathbf{Y}) = g(x_1, \ldots, x_p)$ such that $E^*|g(\mathbf{Y})|$ exists, we have $P(A) = 1$ where $A = \{N^{-1}\Sigma_1^N g(\mathbf{Y}_n) \to E^*g(\mathbf{Y})\}$. This follows from Birkhoff's ergodic theorem by the following amusing argument (pointed out to the author by R. R. Bahadur). Define $G(\mathbf{y}_1) = P(A|\mathbf{Y}_1 = \mathbf{y}_1)$. From the Markov property $G(\mathbf{y}_1)$ does not depend on the initial density p_0, and $P(A) = \int G(\mathbf{y}_1)p_0(\mathbf{y}_1)\,d\mathbf{y}_1$. Now if $p_0 = p^*$, the invariant density, then $\{Y_n\}_{n=1,2,\ldots}$ is ergodic as a strictly stationary process. Birkhoff's theorem then implies that $\int G(\mathbf{y}_1)p^*(\mathbf{y}_1)\,d\mathbf{y}_1 = 1$. Since p^* is strictly positive, this equation can be true only if $G(\mathbf{y}_1) = 1$ a.e. Hence $P(A) = \int G(\mathbf{y}_1)p_0(\mathbf{y}_1)\,d\mathbf{y}_1 = 1$ for any initial density p_0.

(ii) Now we turn to the main proof. Define $a(\theta) = E^*(\Delta\eta_n)^2$, $a(\theta, \rho) = E^*[\inf_{|\tilde{\theta}-\theta|\le\rho}(\Delta\tilde{\eta}_n)^2]$ where $\Delta\eta_n = \eta_n(\theta) - \eta_n(\theta^\circ) = (\theta_0 - \theta_0^\circ) + \Sigma_{i=1}^P(\theta_i - \theta_i^0)x_{n-i}$ and $\Delta\tilde{\eta}_n = \eta_n(\tilde{\theta}) - \eta_n(\theta^0)$. These expectations are independent of n since they are taken with respect to the invariant distribution. It is easy to see that by the monotone convergence theorem, $a(\theta, \rho) \to a(\theta)$ as $\rho \to 0$, and in part (iii) of this proof we will show that $a(\theta) > 0$ for all $\theta \neq \theta_0$. Hence there exist a $\varepsilon > 0$ such that $a(\theta, \varepsilon) > 0$. The law of large numbers in (i) then gives $N^{-1}\Sigma_1^N[\inf_{|\tilde{\theta}-\theta|\le\varepsilon}(\Delta\tilde{\eta}_n)^2] \to a(\theta, \varepsilon) > 0$ a.e., and $N^{-1}\Sigma_1^N(\Delta\eta_n)^2 \to a(\theta)$ a.e. Condition (3.4) follows immediately.

(iii) It remains to show that $a(\theta) = E^*(\Delta\eta_n)^2$ exists and is strictly positive. To see this, write $\delta = \theta_0 - \theta_0^0$, $\mathbf{d} = (\theta_i - \theta_i^0)_{i=1,\ldots,p}$, $\boldsymbol{\mu} = E^*(\mathbf{Y})$, and $\Sigma = \text{cov}^*(\mathbf{YY}')$. Then direct calculation gives $a(\theta) = (\delta + \mathbf{d}'\boldsymbol{\mu})^2 + \mathbf{d}'\Sigma\mathbf{d}$. Since the

strict positivity of the invariant density implies that Σ is strictly positive definite, it is easy to see that $a(\theta) > 0$ unless $\theta = \theta^0$. □

By the same arguments used in part (ii) of the above proof, it is easy to see that condition (2.9a) is also satisfied under ergodicity. Thus if the parameter space is taken to be any compact subset in the ergodicity region, Theorem 2C can be applied to yield consistency of θ.

When the state space of $\{\mathbf{Y}_n\}$ is decomposable into several ergodic classes, the above theory can still be applied to each ergodic class. Another type of non-ergodicity is much more difficult to handle, namely, when the process exhibits no steady state behavior, such as the nonstationary normal AR process investigated in the next section.

3.2. Nonstationary normal autoregressive process. We now return to the first model of Table 3.1, i.e., the conditional distribution of x_n given previous values is normal with mean $\eta_n = \theta_1 x_{n-1} + \theta_2 x_{n-2} + \cdots + \theta_p x_{n-p}$ and variance 1, but we no longer require θ to lie in the region of ergodicity.

LEMMA 3B. *For the normal AR process, ergodic or otherwise, condition* (3.4) *always holds.*

PROOF. It is easy to check that if $\{x_n\}$ is an AR process then η_n must be an ARMA process with the same autoregressive polynomial. Consider the unique factorization of this AR polynomial

$$\left(1 - \theta_1^0 B - \cdots - \theta_p^0 B^p\right) = \prod_{j=1}^{J} (1 - \lambda_j B)^{m_j}.$$

The asymptotic behavior of the ARMA process depends on the positions of the λ_j's and their multiplicities. To simplify notations, write λ_j in polar form, i.e., $\lambda_j = \rho_j e^{i\omega_j}$ (here $i = \sqrt{-1}$), and order the λ_j's so that $\rho_1 = \rho_2 = \cdots = \rho_{J_0} > \rho_{J_0+1} \geq \cdots > \rho_J$ and $m_1 \geq m_2 \geq \cdots > m_{J_0}$. Let $\rho = \rho_1 = \max_{j \leq J} \rho_j$ and $m = \max_{j \leq J_0} m_j$, and consider three cases:

(i) $\rho < 1$: in this case, the process is ergodic, there is no difficulty.

(ii) $\rho > 1$: this is the so called "explosive" case, in which the variance increases exponentially. The problem is to determine the exact rate of increase. By rather elaborate calculation, it can be shown that

$$\operatorname{Var}\left(\rho^{-N} N^{-(m-1)} \eta_N\right) \sim \sum_{i=1}^{J_0} \sum_{j=1}^{J_0} r_{ij} \cos\left(\nu_{ij} + (\omega_i - \omega_j)N\right),$$

where the amplitudes r_{ij} and the phases ν_{ij} are continuous functions of θ and θ^0. From this it follows that (3.4) holds with, say, $\alpha_N = (\rho^N N^{(m-1)})^{3/2}$.

(iii) $\rho = 1$. This is the nonexplosive nonstationary case. It can be shown that $\operatorname{Var}(N^{-(m-1/2)} \eta_N) \to c$, where c depends continuously on θ and θ^0. Hence (3.4) holds with $\alpha_N = (N^{m-1/2})^{3/2}$. The most complete result on this case can be found in Tiao and Tsay (1983). □

For normal AR processes, the tightness condition (2.9) is trivial to verify, since it has a natural exponential family structure. It is also easy to see that the log-likelihood is strictly concave. Thus by Theorem 2E, $\hat{\theta}$ is consistent for θ^0, without requiring any ergodicity or compactness condition.

3.3. Proportional hazard models. This is the model which led Cox (1972, 1975) to formulate the general idea of partial likelihood. It is thus of interest to examine it in light of the preceding discussions. In the Cox model one observes failure times of a group of individuals subjected to censoring. Suppose the (uncensored) failures occur at distinct times $t_{(1)} < \cdots < t_{(N)}$. Let R_n be the risk set at time $t_{(n)}^-$, i.e., the set of individuals who have not failed or been censored by that time. Furthermore, suppose that for each individual one also observes a set of explanatory variables $z = (z_1, \ldots, z_p)$. The distinctive assumption of the proportional hazard model is that the hazard function for an individual at risk at age t is

$$\lambda(t, z) = \lambda_0(t) w(\theta, z), \tag{3.5}$$

where λ_0 is the base line hazard measuring the hazard at $\theta \equiv 0$, $w(\theta, z)$ is a weighting function (or relative risk factor). The interest is usually in the estimation of the "regression coefficient" $\theta = (\theta_1, \ldots, \theta_p)$, which characterizes how the explanatory variable z affects the hazard, with $\theta = 0$ corresponding to the case of no effect. If $w(\theta, z)$ depends only on the inner product $\theta \cdot z$, i.e., $w(\theta, z) = w(\theta \cdot z)$, then the model can be called a Cox linear regression model. If, further, $w(\theta \cdot z) = e^{\theta \cdot z}$, then we have the natural Cox model.

To obtain a partial likelihood for θ, let x_n specify the covariate value associated with the individual who fails at $t_{(n)}$, and let c_n denote all death and censoring times up to and including time $t_{(n)}$. If $p_n(z)$ is defined to be the fraction of individuals in R_n having covariate value equal to z and $\mathbb{Z}_n$ = the set of covariate values of individuals in R_n, then the conditional likelihood of x_n given c_n is,

$$f_\theta(x_n | c_n) = \frac{p_n(x_n) w(\theta, x_n)}{\sum_{z \in \mathbb{Z}_n} p_n(z) w(\theta, z)}. \tag{3.6}$$

The partial likelihood based on $x_n | c_n$, $n = 1, \ldots, N$ is just the product of these conditional likelihoods.

In this paper we will only study the important though special case when the explanatory variable z is discrete, i.e., $z \in \mathbb{Z} = \{z^{(1)}, \ldots, z^{(k)}\}$, note that each $z^{(i)}$ in $\mathbb{Z}$ is a p-vector. Under this assumption the conditional likelihood of x_n given c_n is a function of only p_n and θ, i.e., $f_\theta(x_n | c_n) = f(x_n; p_n, \theta)$. Similarly, there are well-defined functions r, i, j, and m, such that,

$$r_n(\theta) = r(x_n; p_n, \theta), \qquad m_n(\theta) m_n = m(x_n; p_n, \theta),$$
$$i_n(\theta) = i(p_n, \theta), \qquad j_n(\theta) = j(p_n, \theta).$$

Provided that $w(\theta, z) > 0$ is continuous in $\theta \in \Theta$ for each $z \in \mathbb{Z}$, the functions r, m, i, and j are each continuous in its domain. The domain of r or m is

$\mathbb{Z} \times S \times \Theta$, and the domain of i or j is $S \times \Theta$, where S is the simplex $S = \{p \in R^k\colon\ p_i \geq 0, \Sigma_{i=1}^k p_i = 1\}$.

It is now clear that the asymptotic behavior of I_N and J_N depends on that of $\{p_n,\ n = 1, \ldots, N\}$. To develop the asymptotic theory, consider a sequence of experiments where N, the number of deaths in the experiment, increases without bound. In general, for any fixed n, the conditioning variables $c_n^{(N)}$ and $c_n^{(N')}$ will be different if $N < N'$ due to the fact that in the larger experiment the risk set $R_{(n)}$ contains more individuals. To avoid confusion, the superscript (N) will be used if necessary.

Let $h^{(N)}(t)$ be the vector of relative proportions of individuals in each covariate stratum who are still at risk at relative time t, where time is scaled by mortality experience, i.e., the kth component of $h^{(N)}$ is defined by $h_k^{(N)}(t) = p_n^{(N)}(z^{(k)})$ if $t = n/N$ and linear in between. Then

$$N^{-1}I_N(\theta) = N^{-1}\sum_{n=1}^{N} i\big(h^{(N)}(n/N), \theta\big) \sim \int_0^1 i\big(h^{(N)}(t), \theta\big)\, dt,$$

$$N^{-1}J_N(\theta) = N^{-1}\sum_{n=1}^{N} j\big(h^{(N)}(n/N), \theta\big) \sim \int_0^1 j\big(h^{(N)}(t), \theta\big)\, dt.$$

In Appendix A.2 it will be shown that under the regularity conditions stated, there exists a (nonrandom) differentiable function $\bar{h}$ to which $h^{(N)}$ converges weakly in $C[0, A]^k$ for any $0 < A < 1$. It then follows that

$$N^{-1}I_N(\theta) \to_P \int_0^1 i(\bar{h}(t), \theta)\, dt,$$

$$N^{-1}J_n(\theta) \to_P \int_0^1 j(\bar{h}(t), \theta)\, dt.$$

The function $\bar{h}$ is determined by a system of ordinary differential equations which can be solved numerically. The use of these differential equations will be further discussed in Section 4. Similarly,

$$N^{-1}\inf_{|\theta'-\theta|\leq\rho} I_N(\theta) \geq N^{-1}\sum_{n=1}^{N} \inf_{|\theta'-\theta|\leq\rho} i\big(p_n^{(N)}, \theta'\big) \to_P \int_0^1 \inf_{|\theta'-\theta|\leq\rho} i(\bar{h}, \theta')\, dt.$$

If $i(\bar{h}(t), \theta) > 0$ for any (t_0, θ_0), then by continuity $i(\bar{h}(t), \theta) > 0$ for all (t, θ) near (t_0, θ_0), i.e., there exists ρ small enough s.t. $\inf_{|\theta'-\theta|\leq\rho} i(\bar{h}(t), \theta') > 0$ for all t near t_0. The above limit is thus strictly positive.

In the case when the parameter space is compact, the above results imply that conditions (2.7)–(2.8) are satisfied with $\alpha_N = N$. The tightness condition (2.9) is also easy to verify using similar arguments. The consistency of the partial likelihood MLE then follows from Theorem 2C.

For the natural Cox model, $w(\theta, z) = e^{\theta \cdot z}$ and (3.6) becomes a natural exponential family. If the distribution in the original population with respect to covariate stratum is not degenerate, then Theorem 2E implies the consistency of $\hat{\theta}$, without requiring compactness of the parameter space.

4. Asymptotic normality. For obvious reasons, it is convenient to assume $\Theta \subset R^p$. Assuming that the following derivatives exist almost everywhere, we will write

$$l_n(\theta) = \log f_\theta(x_n|c_n); \qquad L_N = \sum_1^N l_n,$$

$$\dot{l}_n(\theta) = Dl_n(\theta), \qquad \ddot{l}_n(\theta) = D^2 l_n(\theta), \qquad \dddot{l}_n(\theta) = D^3 l_n(\theta),$$

where D^i is the ith order differential operator. For example, $D^3 l_n(\theta_0)$ is the triple array of third-order derivatives $\{D_{ijk} l_n(\theta)|_{\theta=\theta_0}\}_{i,j,k=1,\ldots,p}$, and for any p-vector e, $D^3 l_n(\theta_0) \cdot (e)^3 = \Sigma_{ijk}(D_{ijk} l_n(\theta_0))e_i e_j e_k$.

Let $u_n = \dot{l}_n(\theta_0)$ be the conditional score for the experiment $x_n|c_n$, then under standard conditions for the conditional densities, we have

$$E(U_n|c_n) = 0, \qquad v_n = \mathrm{Cov}(u_n|c_n) = E(-\ddot{l}_n(\theta_0)|c_n). \tag{4.1}$$

With $U_N = \Sigma_1^N u_n$, $V_N = \Sigma_1^N v_n$, the main asymptotic normality result is

THEOREM 4A. *Suppose $\hat{\theta}$ is consistent for $\theta_0 \in$ interior of $\Theta \subset R^p$, and for each n, l_n has third-order derivatives almost surely and* (4.1) *holds. Assume also that there are constants $a_N \uparrow \infty$ and a neighborhood O of θ_0, such that*

$$a_N^{-1} V_N \to_P \text{ some p.d. matrix } Q, \tag{4.2}$$

$$a_N^{-1}(-\ddot{L}_N(\theta_0)) \to_P \text{ some p.d. matrix } Q_1, \tag{4.3}$$

$$P\Big(a_N^{-1} \sup_{\theta \in O} |\dddot{L}_N(\theta)| < M\Big) \to 1 \quad \text{for some constant } M, \tag{4.4}$$

$$a_n^{-3/2} \sum_1^N E\big(\|u_n\|^3 | c_n\big) \to_P 0. \tag{4.5}$$

Then

$$a_N^{1/2}(\hat{\theta}_N - \theta_0) \to_D N\big(0, Q_1^{-1} Q Q_1^{-1}\big).$$

REMARK. In many cases, if (4.2) holds then (4.3) also holds with $Q_1 = Q$; for example, a sufficient condition for this is

$$a_N^{-2} \sum_1^N \mathrm{Var}(e' \ddot{l}_n(\theta_0) e | c_n) \to_P 0 \quad \text{for all unit vectors } e. \tag{4.6}$$

PROOF OF THEOREM. By definition of $\hat{\theta}$ and Taylor expansion,

$$0 = \dot{L}_N(\hat{\theta}) = U_N + \ddot{L}_N(\theta_0) \cdot (\hat{\theta} - \theta_0) + \dddot{L}_N(\theta^*) \cdot (\hat{\theta} - \theta_0)^2/2, \tag{4.7}$$

where θ^* lies between $\hat{\theta}$ and θ_0. Let $B_N = -[\ddot{L}_N(\theta_0) + 1/2 \dddot{L}_N(\theta^*) \cdot (\hat{\theta} - \theta_0)]$. Since $\hat{\theta} \to \theta_0$, by (4.3)–(4.4), $a_N^{-1} B_N$ becomes positive definite with probability tending to 1. Hence from (4.7) we have

$$a_N^{1/2}(\hat{\theta} - \theta_0) = \big[a_N^{-1} B_N\big]^{-1} \big(a_N^{-1/2} U_N\big). \tag{4.8}$$

First note that $[a_N^{-1}B_N]^{-1} \to_P Q_1^{-1}$, and thus the theorem will be proved if we can show that $a_N^{-1/2}U_N \to_D N(0, Q)$. For any constant unit vector e, define $t_n = e \cdot u_n$, then by (4.1) $\{t_n\}_{n=1,2,\ldots}$ are martingale differences with respect to the σ-fields generated by $\{c_n\}_{n=1,2,\ldots}$, and

$$a_N^{-1}\sum_1^N \operatorname{Var}(t_n|c_n) = a_N^{-1}e'V_N e \to_P e'Qe.$$

Furthermore, it follows easily from (4.5) that

$$a_N^{-3/2}\sum_1^N E\big(|t_n|^3|c_n\big) \to_P 0,$$

$$a_N^{-1}\sum_1^N E\big[t_n^2 I\big(|t_n| > \varepsilon a_N^{1/2}\big)|c_n\big] \le \varepsilon^{-1}a_N^{-3/2}\sum_1^N E\big(|t_n|^3|c_n\big) \to_P 0.$$

Since all conditions for the martingale central limit theorem (Brown, 1971) are verified for the martingale $a_N^{-1/2}\Sigma_1^N t_n$, it follows that

$$e \cdot \big(a_N^{-1/2}U_N\big) = a_N^{-1/2}\sum_1^N t_n \to_D N(0, e'Qe).$$

Since this is true for any unit vector e, $a_N^{-1/2}U_N \to_D N(0, Q)$. □

Let us illustrate the theory with the proportional hazard model of Section 3.3. From (3.6), we have

$$l_n(\theta) = \log f_\theta(x_n|c_n) = \text{constant} + c(\theta, x_n) - b(\theta, p_n), \tag{4.9}$$

where

$$c(\theta, x_n) = \log w(\theta, x_n) \qquad \text{and} \qquad b(\theta, p_n) = \log\Big(\sum_z p_n(z)e^{c(\theta, z)}\Big).$$

We will also assume that

(4.10) for each z, $c(\cdot, z)$ is three times continuously differentiable around θ_0,

(4.11) p_1 is nondegenerate,

(4.12) $\operatorname{cov}(\dot{c}(\theta_0, x_n)|c_n)$ is p.d. whenever p_n is nondegenerate.

Then, using the results of Appendix A.2, conditions (4.2)–(4.6) can be verified, and hence by Theorem 4A,

$$N^{1/2}(\hat{\theta}_N - \theta_0) \to_{\mathscr{D}} N(0, Q^{-1}).$$

Furthermore, the (i, j)th component of the (normalized) Fisher-information matrix Q based on the partial likelihood can be calculated in the following manner: with $w^k = w(\theta_0, z^{(k)})$ and $\dot{c}_{ik}$ = the ith partial derivative of $c(\theta, z^{(k)})$ w.r.t. θ at $\theta = \theta_0$,

$$Q_{ij} = \int_0^1 v_{ij}(h(t))\,dt, \tag{4.13}$$

where

(a) $$v_{ij}(h) = \sum_{k=1}^{K} \dot{c}_{ik}\dot{c}_{jk}g_k - \Big(\sum_k \dot{c}_{ik}g_k\Big)\Big(\sum_k \dot{c}_{jk}g_k\Big).$$

(b) The vectors g and h are related by $g_k = h_k w^k/(\Sigma_l h_l w^l)$.

(c) The vector function $h(t)$ is determined by the system of ordinary differential equations specified in (A.13) of the appendix, with initial values $h_k(0)$ = the proportion, at the beginning of the trial, of individuals having explanatory variate value $z^{(k)}$.

The form of the differential equation is particularly simple when there is no censoring:

$$\frac{d}{dt}h_k(t) = \frac{1}{1-t}(h_k(t) - g_k(t)), \qquad k = 1,\ldots,K, \qquad 0 \le t < 1. \tag{4.14}$$

In the case of the two sample problem the explanatory variable z is 0 or 1 and the hazard is $\lambda_0(t)$ of $\lambda_0(t)e^{\theta}$, depending on the sample to which the individual belongs. Let $\alpha = e^{\theta}$, q = fraction of individuals in sample 0, then from (4.13)–(4.14) the asymptotic Fisher information is

$$\lim_{N\to\infty} [N \operatorname{Var}(\hat{\theta})]^{-1} \simeq Q = \int_0^1 g(t)(1-g(t))\,dt, \tag{4.15}$$

where

$$g(t) = h(t)/(h(t) + \alpha(1-h(t))),$$

$$\frac{d}{dt}h(t) = \frac{1}{1-t}(h(t) - g(t)), \qquad h(0) = q.$$

The differential equation in (4.15) is easy to solve numerically. Values of Q calculated from (4.15) for various values of q and α are presented in Table 4.1.

The two-sample problem is also studied in Efron (1977) and Oakes (1977). Efron has derived a formula for the asymptotic Fisher information for the special case when the two samples have different exponentially distributed lifetimes, i.e., the baseline hazard $\lambda_0(t)$ is a constant function. This is formula (4.9) in his paper, reproduced here as

(4.16) (Efron formula): $$[N \operatorname{Var}(\hat{\theta})]^{-1} \simeq \int_0^1 \frac{q(1-q)\,du}{q + (1-q)\alpha u^{(\alpha-1)/\alpha}}.$$

TABLE 4.1
Values of asymptotic information Q

q	α	Q calculated using (4.15)	(4.16)
0.5	2	0.225345	0.225345
0.5	5	0.159402	0.159402
0.25	5	0.114305	0.114305

It is a byproduct of our derivation of (4.13)–(4.14) that in the uncensored case the asymptotic Fisher information Q does not depend on the baseline hazard $\lambda_0(t)$. Therefore, Efron's formula, although derived under the constant $\lambda_0(\cdot)$ case, is actually applicable to the wider case of arbitrary $\lambda_0(\cdot)$, and must therefore agree with formula (4.15). The values of Q calculated using (4.16) are also presented in Table 4.1. There is little doubt that the two formulae are equivalent, although we have not found a direct analytical deduction of one from the other.

5. Efficiency. In this paper we consider relative asymptotic efficiency as the ratio of asymptotic variances of estimators [see e.g., Le Cam (1953) and Bahadur (1964)]. Fisher-information is to be interpreted only through its relationship with asymptotic variances. Other interpretations exist and may be important, but are not discussed here.

Let $l_n^{(w)}(\phi) = \log f_\phi(w_n|d_n)$, $l_n^{(x)}(\theta) = \log f_\theta(x_n|c_n)$, then by (1.3) the logarithm of the full likelihood can be decomposed as

$$L_N^{(y)}(\phi) = \sum_{n=1}^{N} l_n^{(w)}(\phi) + \sum_{n=1}^{N} l_n^{(x)}(\theta) = L_N^{(w)}(\phi) + L_N^{(x)}(\theta), \tag{5.1}$$

the second term being the logarithm of the partial likelihood.

There are three basic situations to consider.

5.1. Finite dimensional nuisance parameter (i.e., $\eta \in \mathbb{R}^r$). Suppose that each of the partial likelihoods $L_N^{(x)}$ and $L_N^{(w)}$ satisfies the regularity conditions of Theorem 4A. (Note that for $L_N^{(w)}$, the parameter $\phi = (\theta, \eta)$ is $p + r$ dimensional.) Denote the MLE of ϕ based on $L_N^{(y)}$ by $\hat{\phi}^{(y)}$ and the MLE of θ based on $L_N^{(x)}$ by $\hat{\theta}^{(x)}$, then

$$\begin{aligned} a_N^{1/2}\big(\hat{\phi}^{(y)} - \phi_0\big) &\to_{\mathscr{D}} N(0, S^{-1}), \\ a_N^{1/2}\big(\hat{\theta}^{(x)} - \theta_0\big) &\to_{\mathscr{D}} N(0, Q^{-1}), \end{aligned} \tag{5.2}$$

where

$$S = \left(\begin{array}{c|c} S_{\theta\theta} & S_{\theta\eta} \\ \hline S_{\eta\theta} & S_{\eta\eta} \end{array}\right)\begin{matrix} p \\ r \end{matrix} = \left(\begin{array}{c|c} Q + H_{\theta\theta} & H_{\theta\eta} \\ \hline H_{\eta\theta} & H_{\eta\eta} \end{array}\right),$$

the $p \times p$ matrix Q and the $(p + r) \times (p + r)$ matrix H are defined by

$$\begin{aligned} Q_{ij} &= -p \lim a_N^{-1}\left(\frac{\partial^2 L_N^{(x)}}{\partial\theta_i\,\partial\theta_j}\right)_0 = p \lim a_N^{-1}\sum_1^N E\left(\left(\frac{\partial l_n^{(x)}}{\partial\theta_i}\right)_0\left(\frac{\partial l_n^{(x)}}{\partial\theta_j}\right)_0\middle| c_n\right), \\ H_{ij} &= -p \lim a_N^{-1}\left(\frac{\partial^2 L_N^{(w)}}{\partial\phi_i\,\partial\phi_j}\right)_0 = p \lim a_N^{-1}\sum_1^N E\left(\left(\frac{\partial l_n^{(w)}}{\partial\phi_i}\right)_0\left(\frac{\partial l_n^{(w)}}{\partial\phi_j}\right)_0\middle| d_n\right), \end{aligned} \tag{5.3}$$

and the subscript 0 indicates evaluation at the true value $\phi = \phi_0$.

Thus, marginally, $a_N^{1/2}(\hat{\theta}^{(y)} - \theta_0) \to_{\mathscr{D}} N(0, S_{\theta\cdot\eta}^{-1})$, where $S_{\theta\cdot\eta} = Q + H_{\theta\cdot\eta}$, $H_{\theta\cdot\eta} = H_{\theta\theta} - H_{\theta\eta}H_{\eta\eta}^{-1}H_{\eta\theta}$. It is appropriate to call the matrix $S_{\theta\cdot\eta}$ the "marginal

(Fisher) information" for estimating θ in the presence of η, and the matrix $H_{\theta \cdot \eta}$ the "loss of information" in using $L_N^{(x)}$ instead of $L_N^{(y)}$.

Why is it sensible to compare $\hat{\theta}^{(x)}$ to $\hat{\theta}^{(y)}$? One justification is the classical result that $\hat{\theta}^{(y)}$ achieves the minimal asymptotic variance among all "regular" estimators. We now describe this result briefly, in preparation for the discussion of the infinite dimensional nuisance parameter case.

A sequence of parameter values $\{\phi_N\}$ is called a "regular sequence" if $a_N^{1/2}(\phi_N - \phi_0) \to e$ for some $e \in \mathbb{R}^{p+r}$. An estimate T_N (of θ) is called

(i) regular in the Hájek sense if for all regular sequences $\{\phi_N\}$, the distribution of $a_N^{1/2}(T_N - \theta_N)$ under ϕ_N converges to a distribution independent of $\{\phi_N\}$;

(ii) regular in the Bahadur sense if the distribution of $a_N^{1/2}(T_N - \theta_0)$ under ϕ_0 converges to a normal distribution, and for all regular sequences $\{\phi_N\}$, $P(T_N \geq \theta_N | \phi_N) \to \frac{1}{2}$.

By a regular estimate we mean an estimate regular either in Hájek or Bahadur sense.

LEMMA 5A. *If $L_N^{(y)}$ satisfies the regularity conditions* (4.2)–(4.6) *of Theorem* 4A (*with ϕ as the parameter and suitable definition of the conditioning fields $\mathcal{F}_1, \mathcal{F}_2 \cdots$*), *then any regular estimate $\{T_N\}$ has asymptotic variance larger than or equal to that of $\hat{\theta}^{(y)}$.*

The Bahadur part of the above result is in Bahadur (1967). The Hájek part can be obtained by a modification of Hájek's original proof (Hájek, 1970) for the case with no nuisance parameter. Both authors make use of the LAN condition (Le Cam, 1960) which is satisfied under the hypothesis of the lemma.

With respect to the calculation of $H_{\theta \cdot \eta}$, the following is a natural question at this point. It is plain that $H_{\theta \cdot \eta}$ is the limiting residual covariance matrix of the regression of $a_N^{-1/2} D_\theta L_N^{(w)}$ on $a_N^{-1/2} D_\eta L_N^{(w)}$. Now $L_N^{(w)} = \Sigma_1^N l_n^{(w)}$, if we regress $D_\theta l_n^{(w)}$ on $D_\eta l_n^{(w)}$ for each n and denote the residual covariance by $h_{\theta \cdot \eta, n}$, will it be true that $a_N^{-1} \Sigma_1^N h_{\theta \cdot \eta, n} \to H_{\theta \cdot \eta}$? The answer is *no*. In general, the limit will only be a lower bound for $H_{\theta \cdot \eta}$, although that bound is sharp in the i.i.d. case.

5.2. Infinite dimensional nuisance parameter. In this section we change the notations slightly: λ will denote the infinite dimensional nuisance parameter and η will denote a finite dimensional parameter to be defined below. The nuisance parameter space Γ is assumed to be an infinite dimensional manifold. For concreteness we take Γ to be a submanifold of a Banach space $\mathcal{H}$. We will derive lower bounds for the variances of regular estimates of θ. These can then be used to provide upper bounds for the loss of information due to using a partial likelihood. The definition of a regular estimate is the same as that given in Section 5.1, but a regular sequence of parameter values $\{\phi_N = (\theta_N, \lambda_N)\}$ is now defined by the property that $a_N^{1/2}(\phi_N - \phi_0)$ converges in the product topology to (e_1, e_2) for some $e_1 \in \mathbb{R}^r$, $e_2 \in \mathcal{H}$.

To obtain lower variance bounds, consider the true nuisance value λ_0 as imbedded in a smooth parametric subfamily $\{\lambda = \lambda(\eta)\colon \eta = (\eta_1, \ldots, \eta_r) \in$ some

neighborhood O_r in $\mathbb{R}^r$}. For the parametric problem involving (θ, η), one can calculate, as in Section 5.1, the lower variance bound $S^{-1}_{\theta\cdot\eta}$ for estimating θ in the presence of $(\eta_1, \ldots, \eta_r)$. Intuition suggests that this should also be a lower variance bound for estimating θ in the presence of λ. We now formulate this more precisely. We will use differential calculus in Banach spaces, see, e.g., Lang (1972).

By a smooth finite dimensional parameterization of λ we mean a differentiable map $\eta \to \lambda(\eta)$ from some neighborhood in $\mathbb{R}^r$ to Γ. Of special interest below is the one-dimensional parameterizations $\xi \to \lambda(\xi)$, $\xi \in \mathbb{R}^1$. These can be regarded as curves in Γ. To each curve is associated a tangent vector based on λ_0, $t = \lim_{\xi \to \xi_0}(\xi - \xi_0)^{-1}(\lambda(\xi) - \lambda(\xi_0))$. The set $\mathscr{T}$ of all possible vectors tangent to Γ at λ_0 constitute a linear subspace of $\mathscr{H}$, called the tangent space at λ_0. The dimension of a parameterization $\eta \to \lambda(\eta)$ is the dimension of the subspace $\mathscr{T}$ spanned by the partial derivatives of the map. For simplicity we will always assume that the parameterization is nonsingular, i.e., its dimension is equal to that of η. To obtain concrete results, consider the following regularity conditions.

(5.4) For any smooth curve $\xi \to \lambda(\xi)$, the full likelihood $L_N^{(\gamma)}$ with the $p+1$ dimensional parameter $\phi = (\theta, \xi)$ satisfies conditions (4.2)–(4.6).

(5.5) The elements of the limiting information matrix Q in (4.2) are uniformly bounded for all curves with tangent vector lengths ≤ 1.

THEOREM 5B. *Suppose* (5.4) *holds and let* $S^{-1}_{\theta\cdot\eta}$ *be the lower variance bound for estimating* θ *in the* $p + r$ *dimensional problem with parameter* $\phi = (\theta, \eta)$, *where* $\eta \to \lambda(\eta)$ *is a smooth* r*-dimensional parameterization of* λ. *If* $\{T_N\}$ *is a regular estimate for* θ *in the infinite dimensional problem with parameter* $\phi = (\theta, \lambda)$, *then* $\{T_N\}$ *has asymptotic variance* $\geq S^{-1}_{\theta\cdot\eta}$.

PROOF. (i) First we show that $\{T_N\}$ must also be a regular estimate for θ in the $p + r$ dimensional problem. To see this, let $\{(\theta_N, \eta_N)\}$ be a regular sequence of parameter values in the $p + r$ dimensional problem, then by differentiability of $\eta \to \lambda(\eta)$,

$$a_N^{1/2}(\lambda(\eta_N) - \lambda(\eta_0)) = a_N^{1/2}\left[\left(D_\eta\lambda(\eta^0) \cdot (\eta_N - \eta_0)\right) + o(|\eta_N - \eta_0|)\right].$$

This converges to a linear combination of the components of the derivative of the map $\eta \to \lambda(\eta)$, since $a_N^{1/2}(\eta_N - \eta_0)$ converges to a vector in $\mathbb{R}^r$. Hence $\{(\theta_N, \lambda(\eta_N))\}$ is a regular sequence of parameter values in the infinite dimensional problem. The desired conclusion now follows directly from the definition of regular estimates.

(ii) It is easy to check that if (4.2)–(4.6) are satisfied for $\phi = (\theta, \xi)$ for all one-dimensional parameterizations $\xi \to \lambda(\xi)$, then they are also satisfied for $\phi = (\theta, \eta)$ for any finite dimensional parameterization $\eta \to \lambda(\eta)$. Hence under (5.4), Lemma 5A can be applied to conclude that $\{T_N\}$ has asymptotic variance larger than $S^{-1}_{\theta\cdot\eta}$. □

In the case when θ is a scalar parameter, it is an elementary fact, first pointed out by Stein (1956), that for any smooth r-dimensional parameterization $\eta \to \lambda(\eta)$, there is a smooth one-dimensional parameterization $\xi \to \lambda(\xi)$ which gives the same lower variance bound as in the r-dimensional case. Thus, to obtain the best lower variance bound for regular estimates of θ, it suffices to consider only one-dimensional parameterizations of λ. A curve $\xi^* \to \lambda(\xi^*)$ which yields the greatest lower variance bound is called a least favorable curve, the corresponding marginal information for θ is called the "minimal Fisher information" by Lindsay (1980, 1983).

Although easy to define, the minimal Fisher information may be difficult to compute. For the i.i.d. case, Lindsay (1980) and Begun, Hall, Huang, and Wellner (1983) give geometric insights as well as examples of computation. In the non i.i.d. case, very few results are available. We now investigate the following general method for computing the minimal Fisher information: consider an increasing sequence of parameterizations $\{\eta^{(r)} \to \lambda(\eta^{(r)})\}_{r=1,2,\ldots}$, here r denotes the dimension of the parameterization. Clearly, the upper information bound $S_{\theta \cdot \eta}(r)$ calculated using the parameterization $\eta^{(r)} \to \lambda(\eta^{(r)})$ will become smaller as r increases. By choosing the sequence of parameterizations appropriately, we hope that the limit of these bounds, $\lim_{r \to \infty} S_{\theta \cdot \eta}(r)$, will provide the minimal Fisher information. For which sequences can the minimal information be calculated by this method?

THEOREM 5C. *If θ is scalar, (5.4)–(5.5) hold, and a least favorable curve $\xi^* \to \lambda(\xi^*)$ exists, then a sufficient condition for $S_{\theta \cdot \eta}(r)$ to converge to the minimal Fisher information as $r \to \infty$ is the following:*

(5.6) *For any $\varepsilon > 0$, there exist $r_0 > 0$ such that $r > r_0$ entails that $\left| t^* - \sum_{i=1}^{r} \alpha_i t_i^{(r)} \right| < \varepsilon$ for some $\alpha_1, \ldots, \alpha_r$; here t^* denotes the tangent vector of the curve $\xi^* \to \lambda(\xi^*)$, and $t_i^{(r)}$ denotes the ith partial derivative of the map $\eta^{(r)} \to \lambda(\eta^{(r)})$.*

PROOF. To each tangent $t \in \mathscr{T}$, let us associate with it the numbers $A(t)$ and $B(t)$ as follows: suppose ξ is any curve with tangent equal to t, then

$$A(t) = \lim a_N^{-1} \sum_1^N E\left(\left(\frac{\partial l_n^{(y)}}{\partial \theta} \right)_0 \left(\frac{\partial l_n^{(y)}}{\partial \xi} \right)_0 \middle| \mathscr{F}_n \right),$$

$$B(t) = \lim a_N^{-1} \sum_1^N E\left(\left(\frac{\partial l_n^{(y)}}{\partial \xi} \right)_0^2 \middle| \mathscr{F}_n \right).$$

Under (5.5), there is a constant K such that

$$|A(t)| \le K|t|, \qquad |B(t)| \le K|t|^2.$$

It can also be checked that $A(\cdot)$ is a linear map and $B(\cdot)$ is a quadratic map. Thus, both A and B are continuous maps.

Now consider the two-dimensional problem with $\phi = (\theta, \xi)$. Here the corresponding bound for the marginal information of θ is given by

$$S_{\theta \cdot \xi} = S_{\theta\theta} - B(t)^{-1}A(t)^2. \tag{5.7}$$

Note that $S_{\theta \cdot \xi}$ is a continuous function of t.

In the $1 + r$ dimensional problem with $\phi = (\theta, \eta^{(r)})$, the corresponding bound for the marginal information of θ is given by

$$S_{\theta \cdot \eta} = S_{\theta\theta} - S_{\theta\eta}S_{\eta\eta}^{-1}S_{\eta\theta}. \tag{5.8}$$

The bound (5.8) can be shown to be smaller than the bound (5.7) when t in (5.7) is any linear combination of $t_i^{(r)}$, $i = 1, \ldots, r$. Combining this fact, condition (5.6) and the continuity of $S_{\theta \cdot \xi}$ in t, the result follows immediately. □

REMARK. Typically, the least favorable tangent t^* is difficult to calculate, and one verifies (5.6) by checking that the span of the partial derivatives $\{t_i^{(r)}, i = 1, \ldots, r\}$ becomes dense in the tangent space $\mathscr{T}$.

We now discuss the case when θ is a p-dimensional parameter, it is now no longer possible to find a one-dimensional parameterization which is as difficult as a given r-dimensional parameterization $\eta \to \lambda(\eta)$. Instead, it is only possible to find a p-dimensional parameterization $\xi \to \lambda(\xi)$ which is as difficult as the r-dimensional parameterization. Thus we must search for the most difficult p-dimensional parameterization $\xi^* \to \lambda(\xi^*)$, which gives the greatest lower bound (among p-dimensional parameterizations) for asymptotic variances of regular estimators of θ. Geometrically, the map $\xi^* \to \lambda(\xi^*)$ gives rise to a p-dimensional surface in Γ which we will call the least favorable surface for the estimation of θ. Any curve tangent to the least favorable surface at $\lambda = \lambda_0$ is the least favorable curve for estimating a particular scalar function of θ. The marginal information for θ given by the least favorable surface is thus the least upper bound for any finite parameterization. We will still call this the minimal Fisher information for the estimation of θ. The general method of calculation outlined above can still be applied: simply calculate the $p \times p$ matrix $S_{\theta \cdot \eta}$ for each $\eta^{(r)}$ and pass to the limit. An obvious extension of Theorem 5C then guarantees that the limit is equal to the minimal Fisher information provided $\operatorname{span}\{t_i^{(r)}, i = 1, \ldots, r\}$ becomes dense in the tangent space as r increases.

Returning finally to the partial likelihood situation, if we have a sequence of parameterizations to which Theorem 5C applies, the marginal information for estimating θ is $Q + H_{\theta \cdot \eta}$ where Q is the same for any parameterization of the nuisance parameter, and $H_{\theta \cdot \eta}$ decreases as the parametric subfamily is enlarged. To get an upper bound for the loss of information of the partial likelihood, we calculate $H_{\theta \cdot \eta}$ from $L_N^{(w)}$ for the subfamilies and pass to the limit. Some illustrative examples are given in Section 6.

We end this discussion with a remark on a difficulty of the minimal Fisher information as a criterion for efficiency comparison in the presence of nuisance

parameters. If an estimate achieves the minimal Fisher information bound, then there can be no other regular estimate with smaller asymptotic variance, and the estimate can, justifiably, be regarded as efficient. On the other hand, if the estimate has asymptotic variance larger than the inverse minimal Fisher information, should it then be regarded as inefficient? This question remains largely unresolved in the case of infinite dimensional nuisance parameters, since in this case it is not known whether there is any regular estimate which can achieve the minimal Fisher information bound. Results on some special cases of this problem can be found in Pfanzagl (1982) and Bickel (1982).

5.3. Incidental nuisance parameter. There are examples, in the i.i.d. case, where "new components" of the nuisance parameter arise as new observations are made, in such a way that none of the components of the nuisance parameter can be estimated with diminishing error. In such cases the MLE is often inconsistent. Neyman and Scott (1948) called such nuisance parameters "incidental parameters."

Similar phenomena of inconsistency also occur in partial likelihood situations. Specifically, if η_n denotes the incidental parameter which appears only in the conditional likelihood $f(w_n|d_n)$, then the full log-likelihood is

$$L_N^{(y)}(\theta, \eta_1, \ldots, \eta_N) = L_N^{(x)}(\theta) + \sum_{n=1}^{N} \log f_{\theta,\eta_n}(w_n|d_n),$$

and the likelihood equations are

$$\frac{\partial L_N^{(x)}}{\partial \theta}(\theta) + \sum_{n=1}^{N} \frac{\partial l_n^{(w)}}{\partial \theta}(\theta, \eta_n) = 0 \quad \text{and} \quad \frac{\partial l_n^{(w)}}{\partial \eta_n}(\theta, \eta_n) = 0. \tag{5.9}$$

Given any θ, the value for η_n can be obtained from the second equation, giving $\eta_n = \hat{\eta}_n(\theta)$, a random variable whose distribution is unaffected by the collection of further data $w_{n+1}, x_{n+1}, w_{n+2}, x_{n+2} \ldots$. Substituting $\hat{\eta}(\theta)$ back in the first equation, we obtain the equation for $\hat{\theta}$:

$$0 = \frac{\partial L_N^{(x)}}{\partial \theta}(\theta) + \sum_{n=1}^{N} g_n(\theta) \quad \text{where } g_n(\theta) = \frac{\partial l_n^{(w)}}{\partial \theta}(\theta, \hat{\eta}_n(\theta)).$$

Now, although $E[(\partial l_n^{(w)}/\partial\theta)_0|d_n] = 0$, because of the distribution of $\hat{\eta}(\theta)$, it is generally true that $E(g_n(\theta_0)|d_n) \neq 0$, and hence the equation for $\hat{\theta}$ would lead to inconsistent estimates.

On the other hand, if the conditions of Theorem 4A are satisfied, the use of only the partial likelihood $L_N^{(x)}(\theta)$ will of course produce consistent and asymptotically normal estimates. Godambe (1976), Andersen (1973), and Lindsay (1980, 1982) have given some conditions, based on extensions of the concepts of sufficiency and ancillarity, under which conditional likelihoods are fully informative.

6. Examples of efficiency calculation.

6.1. Missing values in AR processes. As a first example, suppose we observe J disconnected sequences of a time series: $[z_{n_1}, \ldots, z_{m_1}]$, $[z_{n_2}, \ldots, z_{m_2}]$, $\ldots, [z_{n_J}, \ldots, z_{m_J}]$, where $n_1 < m_1 < n_2 < m_2 < \cdots < n_J < m_J = N$. Suppose that within the segments the series follows a AR(1) model, i.e., $z_t = \theta z_{t-1} + a_t$ provided z_t and z_{t-1} are in the same segment, where $-1 < \theta < 1$, and the a_t's are i.i.d. $N(0,1)$. A partial likelihood can be set up based on the conditional densities $f_\theta(z_n|z_{n-1})$ of those z_n whose predecessor z_{n-1} is also observed. Then

$$L_N^{(x)}(\theta) = -\frac{1}{2}\sum_{j=1}^{J}\sum_{t=n_j+1}^{m_j}(z_t - \theta z_{t-1})^2,$$

$$U_N = -\sum_{j=1}^{J}\sum_{t=n_j+1}^{m_j} a_t z_{t-1} \quad \text{and} \quad V_N = \sum_{j=1}^{J}\sum_{t=n_j+1}^{m_j} z_{t-1}^2.$$

The conditions for consistency and asymptotic normality of the MLE $\hat{\theta}^{(x)}$ based on the partial likelihood are easy to verify if $N^{-1}V_N$ converges to some positive constant. For simplicity consider the regular case when the length of each observed segment is $m_j - n_j = k$, and the length of each missing segment is $n_j - m_{j-1} = l$. We now discuss three different models for the missing values, leading to different comparisons of the partial likelihood MLE $\hat{\theta}^{(x)}$ to the full MLE $\hat{\theta}^{(y)}$.

(i) The whole series $z_1, \ldots, z_N$ follows an AR(1) model: in this case clearly $N^{-1}V_N \to_P (k/(k+l))(1/(1-\theta^2))$ as $N \to \infty$; on the other hand, the full information in the complete data $z_1, \ldots, z_N$ is $1/(1-\theta^2)$. Hence the asymptotic efficiency of $\hat{\theta}^{(x)}$ is bounded from below by $k/(k+l)$; this bound is close to 1 if k is much larger than l.

When l/k is not negligible, the information lost by using only the partial likelihood is contained in the conditional densities of z_{n_j} given $z_{m_{j-1}}$, $j = 2, \ldots, J$. (We are ignoring z_{n_1} but this does not affect the asymptotics.) Write z_{n_j} as

$$z_{n_j} = a_{n_j}^* + \theta^l z_{m_{j-1}},$$

where

$$a_{n_j}^* = a_{n_j} + \theta a_{n_{j-1}} + \cdots + \theta^{l-1} a_{m_{j-1}+1} \text{ is } N(0, \sigma^2)$$

with

$$\theta^2 = (1 - \theta^{2l})/(1 - \theta^2).$$

Let $l_j^{(w)} = \log f(z_{n_j}|z_{m_j})$ and $v_j^{(w)} = \mathrm{Var}[((\partial/\partial\theta) l_j^{(w)})_0 | z_{m_{j-1}}]$. By direct calculation, the (normalized) information contained in $L_N^{(w)}$ is

$$N^{-1}\sum_{j=1}^{J} v_j^{(w)} \to_P \frac{1}{k+l}\left[l^2\theta^{2(l-1)}\frac{1}{1-\theta^{2l}} + \frac{1}{2\sigma^4}\left(\frac{\partial\sigma^2}{\partial\theta}\right)^2\right].$$

Thus

$$\text{(asymp. efficiency of } \hat{\theta}^{(x)}) = \frac{k/(1-\theta^2)}{k/(1-\theta^2) + \left[l^2\theta^{2(l-1)}/(1-\theta^{2l}) + (1/2\sigma^4)(\partial\sigma^2/\partial\theta)^2\right]}. \tag{6.1}$$

Note that as $l \to \infty$, $1/2\sigma^4(\partial\sigma^2/\partial\theta)^2 \to \theta^2/(1-\theta^2)^2$ and hence (asymp. efficiency of $\hat{\theta}^{(x)}$) $\to k(k + (\theta^2/(1-\theta^2)))^{-1}$. This limit is close to 0 if θ is close to 1, a result quite contrary to intuition. This seeming contradiction is due to the oversimplifying assumption that the variance of the random errors are known to be 1. In fact, if $\gamma = \mathrm{Var}(a_n)$ is introduced as a parameter, then the last term in the denominator of the efficiency expression (6.1) disappears, giving the limit 1 for the efficiency as $l \to \infty$.

This is the only context where the assumption of constant error variance makes a qualitative difference; in the following discussion we will continue to assume constant variance for the sake of simplicity.

(ii) Incidental parameters: Suppose that at the end of each observed segment, the level of the series is shifted by an unknown amount μ_j, causing the series to be unobserved in the next l subsequent units of time. That is, we assume $z_{m_j+1} = \theta(\mu_j + z_{m_j}) + a_{m_j+1}$, and that the rest of the series follows an AR(1) model. Thus, each of the conditional densities $f(z_{n_j}|z_{m_{j-1}})$ involves a different incidental parameter μ_j. We will assume also that the sequence $\mu_1, \mu_2, \ldots$ has enough regularity so that

$$P(\delta_1 < N^{-1}V_N < \delta_2) \to 1 \quad \text{for some constants } \delta_1, \delta_2 > 0. \tag{6.2}$$

It is clear that some regularity assumptions for the μ_js are necessary for discussion of asymptotics; condition (6.2) is, in fact, quite mild, being satisfied, for example, if the μ_js are uniformly bounded. Under this condition, $\hat{\theta}^{(x)}$ is consistent for θ; in contrast, the MLE $\hat{\theta}^{(y)}$ turns out to be an inconsistent estimator for θ, a result we now proceed to establish. Using the fact that $z_{n_{j+1}} = a^*_{n_{j+1}} + \theta^l(\mu_j + z_{m_j})$ where $a^*_{n_{j+1}}$ is a $N(0, \sigma^2)$ random variable as defined in (i), we obtain

$$\frac{\partial}{\partial\theta} l_j^{(w)} = -\frac{1}{2}\left[\sigma^{-2}\frac{\partial\sigma^2}{\partial\theta} + \sigma^{-2}2\left(z_{n_{j+1}} - \theta^l(\mu_j + z_{m_j})\right)\left(-l\theta^{l-1}(\mu_j + z_{m_j})\right) + \left(-\sigma^{-4}\frac{\partial\sigma^2}{\partial\theta}\right)\left(z_{n_{j+1}} - \theta^l(\mu_j + z_{m_j})\right)^2\right],$$

$$\frac{\partial}{\partial\mu_j} l_j^{(w)} = \frac{\theta^l}{\sigma^2}\left(z_{n_{j+1}} - \theta^l(\mu_j + z_{m_j})\right).$$

The value of μ_j that makes $(\partial/\partial\mu_j)l_j^{(w)}$ zero will also make the second and third terms of $(\partial/\partial\theta)l_j^{(w)}$ vanish. Recalling the form of the likelihood equations (5.9), we see that the MLE $\hat{\theta}^{(y)}$ must satisfy the equation

$$\frac{\partial L_N^{(x)}}{\partial\theta} - \frac{J}{2}\sigma^{-2}\frac{\partial\sigma^2}{\partial\theta} = 0. \tag{6.3}$$

It follows from an easy calculation that the solution $\hat{\theta}_N^{(y)}$ of (6.3) satisfies

$$\hat{\theta}_N^{(y)} = \theta_0 - \frac{1}{2}\frac{1}{k+l}\frac{1}{N^{-1}V_N}\left(\sigma^{-2}\frac{\partial\sigma^2}{\partial\theta}\right)_{\hat{\theta}_N^{(y)}} + o_p(1).$$

Thus, if $(\sigma^{-2}(\partial\sigma^2/\partial\theta))_{\theta=\theta_0} \neq 0$ then it is impossible for $\hat{\theta}_N^{(y)}$ to converge to θ_0. Finally, it can be checked that for $l \geq 2$ the only real roots of the equation $\sigma^{-2}(\partial\sigma^2/\partial\theta) = 0$ are 0 and 1.

(iii) Random level shifts: When there are many nuisance parameters we may try to model them also. In the present example, we will illustrate the efficiency calculation for the simple model in which the level shifts are random variables with a common unknown density, independent of each other and independent of the a_ns. The appropriateness of such a model, of course, depends on the particular application and should be, just like the original autoregressive assumption, subjected to careful scrutiny.

The above model is equivalent to

$$z_{n_{j+1}} = a^*_{n_{j+1}} + \theta^l z_{m_j} + U_j, \tag{6.4}$$

where U_j's are i.i.d. random variables with a common unknown density g, with respect to a given finite measure μ, and are independent of the errors a_n's. The $N(0, \sigma^2)$ random variable $a^*_{n_{j+1}}$ is as defined in (i).

Suppose that the true density g_0 is positive a.e. (μ). Then without loss of generality we can take μ to be the measure induced by g_0, and take $g_0 \equiv 1$. We will consider $h = \sqrt{g}$ as our nuisance parameter and consider as the nuisance parameter space, $\Gamma = \{h: \int h^2\, d\mu = 1\}$. Let $h_1, h_2, \ldots$ be such that $\{1, h_1, h_2, \ldots\}$ is an orthonormal basis in $L^2(\mu)$, and consider the sequence of parameterizations:

$$h(\cdot|\eta) = \left\|1 + \sum_1^r \eta_i h_i\right\|^{-1}\left(1 + \sum_1^r \eta_i h_i\right) = \left(1 + \sum_1^r \eta_i^2\right)^{-1/2}\left(1 + \sum_1^r \eta_i h_i\right). \tag{6.5}$$

Here $\|\cdot\|$ denotes the norm in $L^2(\mu)$, and we use $\langle \cdot, \cdot \rangle$ to denote the corresponding inner product. To verify the crucial condition (5.6), it suffices to make the following elementary observations:

(a) the tangent space (of Γ) at $h_0 \equiv 1$ is $\mathscr{T} = \{h: \langle h, 1\rangle = 0\}$,
(b) the ith partial derivative of $h(\cdot|\eta)$ at h_0 is simply h_i, $i = 1, 2, \ldots$,
(c) $\{h_1, h_2, \ldots\}$ forms an orthonormal basis of $\mathscr{T}$.

Therefore, we can proceed to calculate the minimal Fisher information by the method of Section 5.2. By (6.4),

$$l_j^{(w)} = \log f(z_{n_{j+1}}|z_{m_j}) = \log\left[\int k_\sigma(z_{n_{j+1}} - \theta^l z_{m_j} - u)g(u)\, d\mu(u)\right],$$

$$k_\sigma(u) = \frac{1}{\sqrt{2\pi}\,\sigma}e^{-u^2/2\sigma^2}.$$

Let $y_j = z_{n_{j+1}} - \theta^l z_{m_j} = a^*_{n_j} + U_j$. Then by direct calculation,

$$\left(\frac{\partial l_j^{(w)}}{\partial\theta}\right)_0 = \alpha_0 + \alpha_1 z_{m_j} f_1(y_j) + \alpha_2 f_2(y_j), \tag{6.6}$$

where

$$\alpha_0 = -\frac{1}{\sigma_0}\frac{\partial \sigma}{\partial \theta_0}, \qquad \alpha_1 = \frac{1}{\sigma_0^2} l\theta_0^{l-1}, \qquad \alpha_2 = \frac{1}{\sigma_0^4}\frac{\partial \sigma^2}{\partial \theta_0},$$

$$f_i(y_j) = \frac{\int (y_j - u)^i k_{\sigma_0}(y_j - u)\, d\mu(u)}{\int k_{\sigma_0}(y_j - u)\, d\mu(u)}, \qquad i = 1, 2;$$

and

$$\left(\frac{\partial l_j^{(w)}}{\partial \eta_i}\right)_0 = 2(Gh_i)(Y_j),$$

where G is the integral operator defined by

$$(Gh)(y) = \frac{\int k_{\sigma_0}(y - u) h(u)\, d\mu(u)}{\int k_{\sigma_0}(y - u)\, d\mu(u)}.$$

These and many expressions derived below can be given a probabilistic interpretation if we introduce three abstract random variables Z, U, and Y, where Z denotes a r.v. having a distribution equal to the marginal distribution of z_{m_j}, U denotes a r.v. having density $g_0(u)$, and Y denotes a r.v. such that the distribution of Y given $U = u$ is $N(u, \sigma_0^2)$. It is also assumed that Z is independent of (U, Y). The operator G can now be interpreted as a conditional expectation $(Gh)(y) = E(h(U)|Y = y)$. Furthermore, it can be easily checked that

$$J^{-1} \sum_{j=1}^{J} \operatorname{Var}\left(\left(\frac{\partial l_j^{(w)}}{\partial \theta}\right)_0 \middle| z_{m_j}\right)$$
$$\to \alpha_1^2 E(Z^2)\operatorname{Var}(f_1(Y)) + 2\alpha_1\alpha_2 E(Z)\operatorname{Cov}(f_1(Y), f_2(Y)) + \alpha_2^2 \operatorname{Var}(f_2(Y)),$$

$$J^{-1} \sum_{j=1}^{J} \operatorname{Cov}\left(\left(\frac{\partial l_j^{(w)}}{\partial \theta}\right)_0, \left(\frac{\partial l_j^{(w)}}{\partial \eta_i}\right)_0 \middle| z_{m_j}\right)$$
$$\to \operatorname{Cov}[\alpha_1 E(Z) f_1(Y) + \alpha_2 f_2(Y), (Gh_i)(Y)].$$

The first limit is the variance of

$$(6.7) \quad \left[\alpha_1(\operatorname{Var}(f_1(Y)))^{1/2} Z\right] + f_3(Y), \quad \text{where } f_3(Y) = \alpha_1 E(Z) f_1(Y) + \alpha_2 f_2(Y).$$

Thus the marginal information $H_{\theta \cdot \eta}$ for θ contained in $L_N^{(w)}$ is obtained by setting $(k + l)H_{\theta \cdot \eta}$ = the residual variance in the regression of the random variable (6.7) on the variables $(Gh_i)(Y)$, $i = 1, \ldots, r$.

Since Z is independent of Y it is clear that $H_{\theta \cdot \eta}$ cannot be made to vanish by increasing r. We now show that the variance of the second term $f_3(Y)$ in (6.7) can be explained arbitrarily well by increasing the number of explanatory variables in the regression. Consider the Hilbert spaces $H(Y) = \{f(Y): Ef^2(Y) < \infty\}$, and $H(U) = \{h(U): Eh^2(U) < \infty\}$. It is clear that $f_3(Y) \in H(Y)$. To prove the result, it suffices to show that $f_3(Y)$ is in the closure of the subspace of $H(Y)$ spanned by $\{1, Gh_1, Gh_2, \ldots\}$. Since $\{1, h_1, h_2, \ldots\}$ is a basis of $H(U)$ and $G1 = 1$, it is enough to show that G: $H(U) \to H(Y)$ is an isomorphism. The continuity of G follows from the inequality $E\{[(Gh)(Y)]^2\} =$

$E\{[E(h(U)|Y)]^2\} \le Eh^2(U) < \infty$. That this map is 1–1 can be seen by checking that

$$(Gh)(y) = 0 \quad \forall\, y \Rightarrow \int k_\sigma(y-u)g_0(u)h(u)\,du = 0 \quad \forall\, y \Rightarrow h(U) = 0 \quad \text{a.e.}$$

That this map is also onto can be seen by the following argument: if $f(Y) \perp Gh(Y)$ for all $h \in H(Y)$, then we have $E(h(U)f(Y)) = 0$ for all h; hence $E(f(Y)|U = u) = 0$ for all $u \in$ support (g_0); it then follows from the completeness of the normal family that $f(Y) = 0$.

Thus the minimal Fisher information contained in the supplementary partial likelihood $L^{(w)}$ is $\alpha_1^2 \mathrm{Var}(f_1(Y))\mathrm{Var}(Z)$. Can this information for θ really be utilized? The following argument convinces us that there must be some way to make use of at least part of this information. Consider the simplest case $l = 1$; then $z_{n_{j+1}} = \theta z_{m_j} + \varepsilon_j$ where $\varepsilon_j = a_{n_{j+1}} + U_j$. Suppose z_{m_j} can take only two values, say δ and $-\delta$. Then conditional on the z_{m_j}'s, the observations z_{n_j}'s have a two-sample location shift structure. Though the density for ε_i is unknown, it is possible to estimate the shift $2\delta\theta$. The larger the value of δ [and thus $\mathrm{Var}(Z)$], the easier it is to estimate θ. In our model the z_{m_j}'s can of course take values other than $\pm\delta$, but even if we throw away all the information in $f(z_{n_{j+1}}|z_{m_j})$ except for those z_{m_j}'s with values close to $\pm\delta$, the above argument implies that we still can use the remaining ones to estimate θ.

How can one make use of this information? A natural approach might be some kind of adaptive estimation involving the estimation of the mixture density g. However, such a method would be very complicated and its properties are largely unknown. The loss of information, $\alpha_1^2\mathrm{Var}(f_1(Y))\mathrm{Var}(Z)$, serves as a guide in choosing between the partial likelihood or more complicated methods. Note that α_1 decreases exponentially as the length l of the missing sequences increases. Thus if θ_0 is not too close to 1 and l is large, we can be sure that the partial likelihood is nearly fully informative.

REMARK. There is a special structure in this example which, at least heuristically, allows an easier computation of the minimal Fisher information. For fixed θ and z_{m_j}, $y_j = z_{n_{j+1}} - \theta^l z_{m_j}$ is a complete sufficient statistics for the unknown shift u_j. In the above we have, by exploiting this completeness and sufficiency, essentially showed that the affine subspace generated by the scores of the nuisance parameters, $(\partial l_j^{(w)}/\partial\eta_i)_0$, $i = 1, 2, \ldots$, is the same as the subspace generated by L^2 functions of y_j. Since projection to this subspace is the same as conditional expectation, the minimal Fisher information can be calculated more easily in the following way:

(a) calculate the "conditional θ-score" $s_j - E(s_j|y_j)$ where s_j denotes the θ-score given in (6.6), i.e.,

$$s_j - E(s_j|y_j) = \alpha_1\big(z_{m_j} - E(z_{m_j})\big)f_1(y_j);$$

(b) calculate the "conditional score information,"

$$\begin{aligned} i_c &= \mathrm{Var}\big(s_j - E(s_j|y_j)\big) \\ &= \alpha_1^2\mathrm{Var}(Z)\mathrm{Var}(f_1(Y)). \end{aligned}$$

This gives the same value as the minimal Fisher information derived above. The approach of studying information by conditioning on a complete sufficient statistics (which may depend on θ) of the incidental parameter is studied in an insightful paper of Lindsay (1983). The definition of the conditional score information (i_c) as given in that paper is not entirely correct—being defined in terms of the density conditional on the incidental parameter (rather than the density with the incidental parameter integrated out); which, in the context of the current example, leads to $i_c = \alpha_1^2 \mathrm{Var}(Z) E(Y - U)^2$, a value always too large as compared to that given above in (b), namely, $\alpha_1^2 \mathrm{Var}(Z) E(Y - E(U|Y))^2$. Nevertheless, Lindsay's main theorem which asserts that in the i.i.d. exponential family mixture setting i_c is the same as the minimal Fisher information, seems to be correct after appropriate modification on the definition of i_c. The above example indicates that Lindsay's theorem might be expected to hold in more general settings.

6.2. Proportional hazard model. As another illustration of the general theory, we treat the proportional hazard model under the assumption of no censoring. In this case the information missed by the partial likelihood $L_N^{(x)}$ is contained in the conditional distribution of t_n given $d_n = (t_{n-1}, \mathbb{R}_n^-)$ where $\mathbb{R}_n^- = \mathbb{R}_n$ = risk set after t_{n-1}. Now, given d_n, t_n is a random variable with hazard function $\lambda_0(t) b_n(\theta)$ for $t > t_{n-1}$, where $b_n(\theta) = \Sigma_{z \in \mathbb{R}_n} w(\theta, z) = (N - n + 1) \Sigma_{z \in \mathbb{Z}} p_n(z) w(\theta, z)$. Thus,

$$l_n^{(w)} = \log f(t_n | d_n) = \log \lambda_0(t_n) + \log b_n(\theta) - \int_{t_{n-1}}^{t_n} \lambda_0(t) b_n(\theta)\, dt,$$

and

$$\left(\frac{\partial l_n^{(w)}}{\partial \theta_i} \right)_0 = a_{ni} - a_{ni} T_n^* \quad \text{where } T_n^* = \int_{t_{n-1}}^{t_n} \lambda_{00}(t) b_n(\theta_0)\, dt,$$

(6.8)

$$a_{ni} = \left(\frac{\partial b_n}{\partial \theta} \Big/ b_n \right)_0 = \frac{\Sigma_{z \in \mathbb{Z}} p_n(z) \dfrac{\partial}{\partial \theta_i} w(\theta_0, z)}{\Sigma_{z \in \mathbb{Z}} p_n(z) w(\theta_0, z)}.$$

Here $\lambda_{00}(\cdot)$ denotes the true value of the base line hazard $\lambda_0(\cdot)$.

Embedding $\lambda_0(t)$ in the parametric family $\lambda_0(t) = \lambda_{00}(t) \exp\{\Sigma_{i=1}^r \eta_i g_i(t)\}$, we have

$$\frac{\partial}{\partial \eta_j} \lambda_0(t) = \lambda_{00}(t) e^{\Sigma_1^r \eta_i g_i(t)} \cdot g_j(t),$$

(6.9)

$$\begin{aligned} \left(\frac{\partial l_n^{(w)}}{\partial \eta_j} \right)_0 &= g_j(t_n) - \int_{t_{n-1}}^{t_n} \lambda_{00}(t) b_n(\theta_0) g_j(t)\, dt \\ &= -g_j(t_{n-1})(T_n^* - 1) + O_p(t_n - t_{n-1}). \end{aligned}$$

From the distribution of t_n given d_n, the distribution of T_n^* conditional on d_n is

seen to be an exponential distribution with expectation 1; using this fact, we have

$$\operatorname{Cov}\left[\left(\frac{\partial l_n^{(w)}}{\partial \theta_i}\right)_0, \left(\frac{\partial l_n^{(w)}}{\partial \theta_j}\right)_0 \middle| d_n\right] = a_{ni}a_{nj},$$

$$\operatorname{Cov}\left[\left(\frac{\partial l_n^{(w)}}{\partial \theta_i}\right)_0, \left(\frac{\partial l_n^{(w)}}{\partial \eta_j}\right)_0 \middle| d_n\right] \doteq a_{ni}g_j(t_{n-1}),$$

$$\operatorname{Cov}\left[\left(\frac{\partial l_n^{(w)}}{\partial \eta_j}\right)_0, \left(\frac{\partial l_n^{(w)}}{\partial \eta_{j'}}\right)_0 \middle| d_n\right] \doteq g_j(t_{n-1})g_{j'}(t_{n-1}).$$

We now discuss the asymptotic behavior of these conditional covariances as $N \to \infty$, $n \to \infty$, $n/N \to x \in (0,1)$; rigorous proof for every step will not be provided, but sufficient details are given so that the limiting values can be obtained numerically in each model specification. The limiting value of a_{ni} depends on that of $p_n(z)$. Hence,

$$a_{ni} \to a_i(x) = \frac{\Sigma_{z \in \mathbf{Z}} \bar{h}_z(x) \dfrac{\partial}{\partial \theta_i} w(\theta_0, z)}{\Sigma_{z \in \mathbf{Z}} \bar{h}_z(x) w(\theta_0, z)} \quad \text{as } n/N \to x, \tag{6.10}$$

where $\bar{h}_z$ is obtained by solving an O.D.E. as given in Appendix A.2. To investigate the asymptotic behavior of t_n, note that from $E(T_{n+1}^* | d_{n+1}) = 1$, we have

$$E\left[\lambda_{00}(t_n)(t_{n+1} - t_n) | d_{n+1}\right] \cong \frac{1}{N(1-x)} \frac{1}{\Sigma_{z \in \mathbf{Z}} \bar{h}_z(x) w(\theta_0, z)}.$$

Suppose $\bar{t}(x)$ is the limiting value of t_n as $n/N \to x$,

$$\Lambda(t) = \int_0^t \lambda_{00}(s)\, ds \quad \text{and} \quad \gamma(x) = \Lambda(\bar{t}(x)),$$

then the above expression suggests that γ satisfies the O.D.E.

$$\frac{d}{dx}\gamma(x) = \frac{1}{1-x} \frac{1}{\Sigma_{z \in \mathbf{Z}} \bar{h}_z(x) w(\theta_0, z)} \quad \text{with } \gamma(0) = 0. \tag{6.11}$$

Thus $\bar{t}(x) = \Lambda^{-1}(\gamma(x))$ can always be solved numerically for any specified w and λ_{00}; under obvious conditions, $\bar{t}(x)$ is a strictly increasing function of x, $\bar{t}(x) \to \infty$ if $x \to 1$. If we define $\tilde{g}_j(x) = g_j(\bar{t}(x))$, then $g_j(t_{n-1}) \to \tilde{g}_j(x)$ as $n/N \to x$.

Finally, returning to the conditional covariance, we see that

$$\frac{1}{N}\sum_{n=1}^{N} \operatorname{Cov}\left[\left(\frac{\partial l_n^{(w)}}{\partial \theta_i}\right)_0, \left(\frac{\partial l_n^{(w)}}{\partial \theta_j}\right)_0 \middle| d_n\right] \to \int_0^1 a_i(x)a_j(x)\, dx =_{\text{def}} H_{ij},$$

$$\frac{1}{N}\sum_{n=1}^{N} \operatorname{Cov}\left[\left(\frac{\partial l_n^{(w)}}{\partial \theta_i}\right)_0, \left(\frac{\partial l_n^{(w)}}{\partial \eta_j}\right)_0 \middle| d_n\right] \to \int_0^1 a_i(x)\tilde{g}_j(x)\, dx =_{\text{def}} H_{i,p+j},$$

$$\frac{1}{N}\sum_{n=1}^{N} \operatorname{Cov}\left[\left(\frac{\partial l_n^{(w)}}{\partial \eta_j}\right)_0, \left(\frac{\partial l_n^{(w)}}{\partial \eta_{j'}}\right)_0 \middle| d_n\right] \to \int_0^1 \tilde{g}_j(x)\tilde{g}_{j'}(x)\, dx =_{\text{def}} H_{p+j,p+j'}.$$

It is easy to see that under suitable conditions the functions $a_i(\cdot)$, $i = 1, \ldots, p$, are square integrable functions, i.e., the regression of $(\partial l^{(w)}/\partial\theta)_0$ on $(\partial l^{(w)}/\partial\eta)_0$ is equivalent to a projection in the Hilbert space $L^2[0,1]$ of $a_1, \ldots, a_p$ on the subspace spanned by $\tilde{g}_1, \ldots, \tilde{g}_r$. Since the g_j's are arbitrary, the $\tilde{g}_j$'s are also arbitrary, and hence the residual variance can be made arbitrarily small by choosing g_j's appropriately. In the natural Cox model, Efron (1977), and Oakes (1977) have given formulae for the efficiency of the partial likelihood. But the calculations from these formulae can become very complicated in special cases. In the theory presented here, the parameterization of the base line hazard is general enough to include the special cases studied by the above authors. Explicit steps are provided for the numerical computation of the matrix of loss of information. The computation is straightforward if an O.D.E. solver is available. The above discussion also appears to be the first systematic account of efficiency calculation for the case of the general form of the relative risk. However, further effort is needed to remove the present restriction of no censoring.

APPENDIX

A.1. Ergodicity of some generalized AR processes. For the Bernoulli case, the transition matrix for $\{Y_n\}$ is

y_1 \ y_2	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	$1-\theta_0$	θ_0	0	0
(0,1)	0	0	$1-(\theta_0+\theta_2)$	$(\theta_0+\theta_2)$
(1,0)	$1-(\theta_0+\theta_1)$	$(\theta_0+\theta_1)$	0	0
(1,1)	0	0	$1-(\theta_0+\theta_1+\theta_2)$	$(\theta_0+\theta_1+\theta_2)$

The region of ergodicity is clearly determined by requiring all eight nonzero entries in the matrix to be strictly positive. This defines a nonempty polygon in R^3.

For the Poisson case, let $\mu_n = (1 - e^{-\eta_n}) = E(x_n|x_{n-1}, \ldots)$, and

$$g(x, y|n, m) = P(x_i = x, x_{i-1} = y|x_{i-1} = n, x_{i-2} = m),$$

then

$$\begin{aligned} g(x, y|m, n) &= \delta_{ny} e^{-\mu_n} \frac{(\mu_n)^x}{x!} \\ &= \delta_{ny} \exp\left[-(1 - e^{-(\theta_0+\theta_1 y+\theta_2 m)})\right] \\ &\quad \times (1 - \exp[-(\theta_0 + \theta_1 y + \theta_2 m)])^x / x!, \end{aligned}$$

where δ_{ny} is the Kronecker delta symbol. If $\theta_0 > 0$, it is clear that all states are reachable from each other. Hence the chain is irreducible. It is clearly aperiodic. To see that it is ergodic, let $g^2(x, y|n, m)$ be the two-step transition matrix.

Then $g^2((0,0)|(n,m)) > \varepsilon > 0$ for some $\varepsilon > 0$. Hence

$$P(x_{2i} = 0, x_{2i-1} = 0 | x_{2(i-1)} = n, x_{2(i-1)-1} = m) > \varepsilon,$$

so,

$$\begin{aligned} &P_{00}(\text{first return to } (0,0) \geq 2n+1) \\ &\quad \leq P_{00}((x_2, x_1) \neq (0,0), (x_4, x_3) \neq (0,0) \cdots (x_{2n}, x_{2n-1}) \neq (0,0)) \\ &\quad \leq (1-\varepsilon)^n. \end{aligned}$$

It follows that mean recurrence time to $(0,0)$ is finite, and the chain is ergodic.

A.2. Stochastic development of proportional hazard systems. The model is described in Section 3.3. First consider the case when there is no censoring, then conditional on the death times $t_{(1)}, \ldots, t_{(N)}$, the stochastic evolution of the system is equivalent to the following: a population of N individuals is sequentially sampled, at each draw the probability for any individual to be selected is proportional to the weight $w(\theta_0, z)$ where z is the explanatory variable value associated with that individual, the selected individual is then removed before the next draw. Throughout we suppose that the set of possible values of the explanatory variable is finite, i.e., $\mathbb{Z} = \{z^{(1)}, \ldots, z^{(K)}\}$. An individual will be called a type k individual if the associated explanatory variable value is $z^{(k)}$. Let $w_k = w(\theta_0, z^{(k)})$ and

(A.1) $x_n = k$ if the individual selected at the nth draw is of type k,

(A.2) $M_{nk}^{(N)}$ = the number of type k individuals just before the nth draw,

(A.3) $h_{nk}^{(N)} = M_{nk}^{(N)}/(N-n+1)$ = proportion of type k before the nth draw,

(A.4) $$g_{nk}^{(N)} = w_k h_{nk}^{(N)} \Big/ \left(\sum_{j=1} w_j h_{nj}^{(N)} \right) = P(x_n = k | x_1, \ldots, x_{n-1}).$$

The question is: With the initial proportions $q_1, \ldots, q_K$ of the K types of individuals fixed and the population size $N \to \infty$, is there any nontrivial limiting behavior in the evolution of the system? If so, how to calculate the limits?

A simple simulation showed that the quantities that become stable are the values of $h_{nk}^{(N)}$ and $g_{nk}^{(N)}$ when n increases with N in such a way that $n/N \to t$. If we define the random vector functions $h^{(N)}(\cdot)$ and $g^{(N)}(\cdot)$ by

(A.5) $$h_k^{(N)}(t) = \begin{cases} h_{nk}^{(N)} & \text{if } t = n/N \\ \text{linear interpolate} & \text{otherwise} \end{cases} \quad \text{for } k = 1, \ldots, K,$$

(A.6) $$g_k^{(N)}(t) = w_k h_k^{(N)}(t) \Big/ \left(\sum_{j=1}^{K} w_j h_j^{(N)}(t) \right),$$

the simulation result suggested that there is a function $\bar{h}(\cdot)$ such that, as $N \to \infty$, $h^{(N)}(t) \to_p \bar{h}(t)$ for all $0 \leq t < 1$.

How does the limiting function $\bar{h}(\cdot)$ depend on the sampling weights w_k's and the initial proportions q_k's, and how can one actually calculate it? We will only

briefly indicate the resolution to these questions here: from (A.1)–(A.4) it is clear that

$$h_k^{(N)}\left(\frac{n+1}{N}\right) = \frac{1}{N-n}\left(M_{nk}^{(N)} - \sum_{j=1}^{K}\delta_{kj}I_{\{x_n=k\}}\right) \tag{A.7}$$
$$= h_k^{(N)}\left(\frac{n}{N}\right)\cdot\frac{N-n+1}{N-n} - \frac{1}{N-n}\sum_{j=1}^{K}\delta_{kj}I_{\{x_n=k\}},$$

$$E\left[h_k^{(N)}\left(\frac{n+1}{N}\right) - h_k^{(N)}\left(\frac{n}{N}\right)\middle| x_1,\ldots,x_n\right] \tag{A.8}$$
$$= \frac{1}{N}\cdot\frac{1}{1-n/N}\left(h_k^{(N)}(n/N) - g_k^{(N)}(n/N)\right).$$

Writing $t = n/N$, $\Delta t = 1/N$, and passing to the limit as $N \to \infty$ in (A.8), we obtain

$$\bar{h}(t+\Delta t) - \bar{h}(t) \simeq \Delta t\cdot(\bar{h}(t) - \bar{g}(t))/(1-t). \tag{A.9}$$

In other words, to determine $\bar{h}(t)$, it is only necessary to solve the system of ordinary differential equations:

$$\frac{d}{dt}\bar{h}(t) = \frac{1}{1-t}(\bar{h}(t) - \bar{g}(t)), \tag{A.10}$$

with

$$\bar{g}_k(t) = w_k\bar{h}_k(t)\Big/\left(\sum_{j=1}^{K} w_j\bar{h}_j(t)\right)$$

and

$$\bar{h}_k(0) = q_k \quad \text{(initial values)}.$$

For example, suppose the initial population size is $N = 5120$, divided equally into two types, and the sampling weights are $w_1 : w_2 = 4:1$. A computer Monte Carlo experiment of 200 replications is performed. In each replication, we record the proportion $h(0.4)$ of the type 1 individuals among those still surviving just before the 2048th death ($5120 \times 0.4 = 2048$). The mean of these 200 proportions is 0.33592 and the SD is 0.0055. The normal score plot is given in Figure A.1 below; there is clearly no evidence of nonnormality. By solving the O.D.E. (A.10) for $\bar{h}$, we obtain $\bar{h}(0.4) = 0.33597$. Using the $\bar{h}$ solved from (A.10), we solve (A.11) to get $\bar{v} = \mathrm{Var}[\sqrt{N}(h - \bar{h})]$; at $t = 0.4$, we obtain $\bar{v} = 0.1491$, giving $\mathrm{SD}(h) = 0.0054$. In fact, much more is true: for any $A < 1$, $\sup_{0\le t\le A}|h^{(N)}(t) - \bar{h}(t)| \to_P 0$; furthermore, for any $t < 1$, $\sqrt{N}(h^{(n)}(t) - \bar{h}(t))$ is asymptotically normal with well-determined variances and covariances. For example, in the simplest case, $K = 2$, if $\bar{v}(t) = \mathrm{Var}[\sqrt{N}(h_1^N(t) - \bar{h}_1(t)]$ and $\bar{S}(t) = \Sigma_1^2 w_k\bar{h}_k(t)$, then

$$\frac{d}{dt}\bar{v} = \frac{2\bar{v}}{(1-t)\bar{S}}[(\bar{S} - w_1) + \bar{g}_1(w_1 - w_2)] + \frac{1}{(1-t)^2}\bar{g}_1(1-\bar{g}_1) \tag{A.11}$$

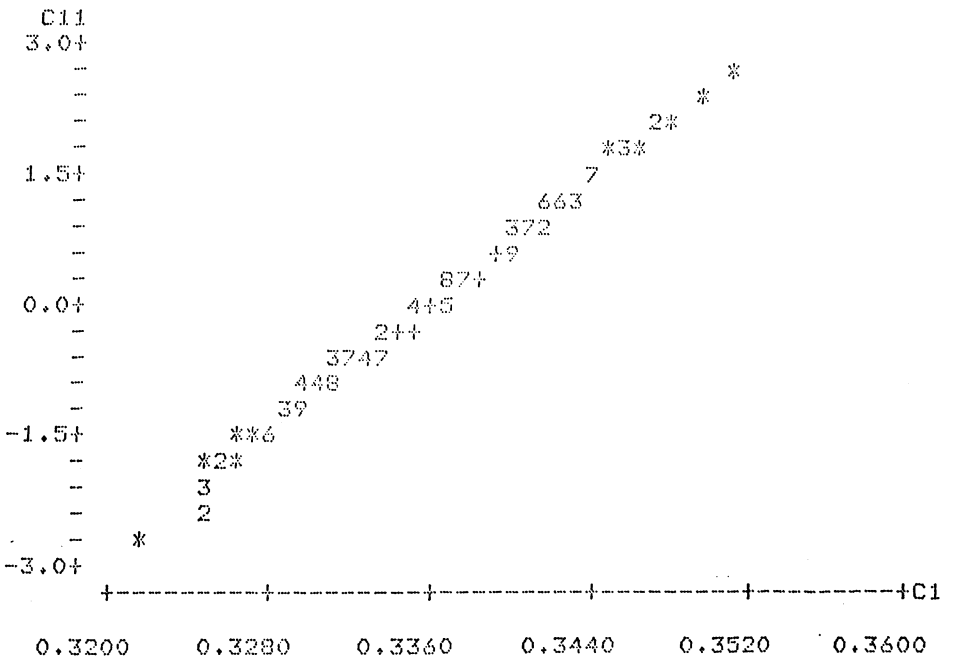

FIG. A.1. *Normal score plots of h(0.4) for 200 replications.*

with $\bar{v}(0) = 0$. The rigorous proofs are not of central interest in this paper and are omitted. These results are in excellent agreement with those from simulation.

To extend these results to the case when censoring is possible, let N_D = total number of deaths, N_c = total number of censors, then $N = N_D + N_c$ = population size. Let $M_{nk}^{(N)}$, $h_{nk}^{(N)}$, $g_{nk}^{(N)}$, $h^{(N)}(\cdot)$, $g^{(N)}(\cdot)$ be as before, in addition, let $l_{nk}^{(N)}$ = number of censors during the interval $(t_{(n)}, t_{(n+1)})$, i.e., between the nth and the $(n+1)$th deaths, and $L_{nk}^{(N)} = \sum_{j=1}^{n} l_{jk}^{(N)}$. Assume that, as $N \to \infty$, N_D/N = proportion of deaths $\to 1 - \alpha$, and as $n/N_D \to x$, the proportion of various types of censors stabilizes, i.e.,

$$\text{(A.12)} \qquad L_{nk}^{(N)}/N_c \to \text{some limit } \gamma_k(x), \quad \text{as } n/N_D \to x.$$

Then by the same reasoning that led to (A.10), we can derive the ordinary differential equation for $\bar{h}$:

$$\text{(A.13)} \qquad \bar{h}_k' = \frac{(1-\alpha)^2}{1 - \alpha\gamma_+ - (1-\alpha)x}\left[(1-\alpha)(h-g)_k + \alpha(h\gamma_+' - \gamma_k')\right],$$

where $\gamma_+ = \sum_1^K \gamma_k$.

Acknowledgments. The author would like to express his deepest appreciation to Professor R. R. Bahadur, who read a major part of the manuscript and provided substantial help, both conceptual and technical. He thanks Referee A

for providing insightful suggestions which greatly improve the sections on efficiency. The valuable comments from Professor Stephen M. Stigler and David L. Wallace are also gratefully acknowledged.

REFERENCES

ANDERSEN, E. B. (1973). *Conditional Inference and Models for Measuring*. Mentalhygiejnisk Forlag, Copenhagen.

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10** 1100–1120.

BAHADUR, R. R. (1964). On Fisher's bound for asymptotic variances. *Ann. Math. Statist.* **35** 1545–1552.

BAHADUR, R. R. (1967). Rates of convergence of estimates and test statistics. *Ann. Math. Statist.* **38** 303–324.

BAILEY, K. R. (1983). The asymptotic joint distribution of regression and survival parameter estimates in the Cox regression model. *Ann. Statist.* **11**, 39–58.

BASAWA, I. V., FEIGIN, P. D. and HEYDE, C. C. (1976). Asymptotic properties of maximum likelihood estimators for stochastic processes. *Sankhyā Ser. A* **38** 259–270.

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

BEGUN, J. M., HALL, W. J., HUANG, W. M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11** 432–452.

BHAT, B. R. (1974). On the method of maximum likelihood for dependent observations. *J. Roy. Statist. Soc. Ser. B* **36** 48–53.

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

BILLINGSLEY, P. (1961). *Statistical Inference for Markov Processes*, University of Chicago Press, Chicago.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

BROWN, B. M. (1971). Martingale central limit theorems. *Ann. Math. Statist.* **42** 59–66.

CAINES, P. E. (1975). A note on the consistency of maximum likelihood estimates for finite families of stochastic processes. *Ann. Statist.* **3** 539–546.

COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–202.

COX, D. R. (1975). Partial likelihood. *Biometrika* **64** 269–276.

CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

CROWDER, M. J. (1976). Maximum likelihood estimation for dependent observations. *J. Roy. Statist. Soc. Ser. B* **38** 45–53.

DOOB, J. S. (1934). Probability and statistics. *Trans. Amer. Math. Soc.* **36** 759–775.

EFRON, B. (1977). The efficiency of Cox's likelihood function for censored data. *J. Amer. Statist. Assoc.* **72** 557–565.

GODAMBE, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika* **63** 277–284.

HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Applications*. Academic, New York.

HÁJEK, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. verw. Gebiete*. **14** 323–330.

IBRAGIMOV, I. A. and HÁSMINSKII, R. Z. (1981). *Statistical Estimation, Asymptotic Theory*. Springer, New York.

KALBFLEISCH, J. D. and SPROTT, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *J. Roy. Statist. Soc. Ser. B* **32** 175–208.

LANG, S. (1972). *Differential Manifolds*. Addison-Wesley, Reading, Mass.

LE CAM, L. (1953). On the asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. Calif. Publ. Statist.* **1** 277–329.

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

LINDSAY, B. G. (1980). Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators. *Phil. Trans. Roy. Soc. London Ser. A* **296** 639–665.

LINDSAY, B. G. (1982). Conditional score functions: some optimality results. *Biometrika* **69** 503–512.

LINDSAY, B. G. (1983). Efficiency of the conditional score in a mixture setting. *Ann. Statist.* **11** 486–497.

LIU, P. Y. and CROWLEY, J. (1978). Large sample theory of the MLE based on Cox's regression model for censored data. Technical report 1, Wisconsin Clinical Cancer Center, Madison.

MCCULLAGH, P. and NELDER, J. A. (1983). *Generalized Linear Models.* Chapman and Hall, London.

NEVEU, J. (1965). *Mathematical Foundations of the Calculus of Probability.* Holden-Day, San Francisco.

NEYMAN, J and SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32.

OAKES, D. (1977). The asymptotic information in censored survival data. *Biometrika* **59** 472–474.

PFANZAGL, J. (1982). *Contributions to a General Asymptotic Statistical Theory.* Springer, New York.

PRENTICE, R. L. and SELF, G. (1983). Asymptotic distribution theory for Cox-type regression model with general relative risk form. *Ann. Statist.* **11** 804–813.

RAO, M. M. (1966). Inference in stochastic processes II. *Z. Wahrsch. verw. Gebiete.* **5** 317–335.

SILVEY, S. D. (1961). A note on maximum likelihood in the case of dependent random variables. *J. Roy. Statist. Soc. Ser. B* **19** 444–452.

SLUD, E. C. (1982). Consistency and efficiency of inferences with partial likelihood. *Biometrika* **69** 547–552.

STEIN, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1** 187–196. Univ. California Press.

TIAO, G. C. and TSAY, R. S. (1983). Consistency properties of least squares estimates of autoregressive parameters in ARMA models. *Ann. Statist.* **11** 856–871.

TSIATIS, A. A. (1981). A large sample study of Cox's regression model. *Ann. Statist.* **9** 93–108.

WALD, A. (1949). Note on the consistency of maximum likelihood estimate. *Ann. Math. Statist.* **20** 595–601.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
5734 UNIVERSITY AVENUE
CHICAGO, ILLINOIS 60637