The Annals of Statistics
1995, Vol. 23, No. 2, 396–421

DUAL LIKELIHOOD[1]

BY PER ASLAK MYKLAND

University of Chicago

This paper introduces the concept of *dual likelihood* as a method of improving accuracy in inference situations depending on martingale estimating equations. Asymptotic results are given for the dual likelihood ratio statistic, and the structure of the family of alternatives is explored. Applications to survival analysis and also to time series, likelihood inference and independent observations are given. Connections to nonparametric likelihood (including empirical likelihood) are established.

1. Introduction. Martingale methods are a powerful tool for dependent variable inference. Estimators in a number of such models have distributions that are (to first order) approximated by the distribution of a martingale, and martingales have asymptotic properties that hold under particularly weak conditions. This is reflected in the wide use of martingale theory to show central limit theorems (CLTs) for estimators based on dependent variables, in particular such longitudinal data as occurs in survival analysis, time series, stochastic differential equations, sequential inference and certain types of stochastic simulation. Assessing the variance of a martingale is also particularly straightforward. There is an extensive literature on the martingale CLTs and their applications; some important references include Aldous (1978, 1989), Hall and Heyde (1980), Rebolledo (1980), Helland (1982), Jeganathan (1982), Jacod and Shiryaev (1987), Fleming and Harrington (1991) and Andersen, Borgan, Gill and Keiding (1993).

A major weakness, however, is that for small samples the quality of the approximation in the martingale CLTs can be quite poor. The purpose of this paper is to propose a way of correcting this problem. We call the approach *dual likelihood* for reasons particularly related to Section 6 and it will permit, in particular, the creation of likelihood ratio statistics in the martingale inference setting.

An overview of the idea is given in Section 4. We shall then argue in detail (Section 5) that the dual likelihood ratio (LR) statistic gives rise to tests and

Received June 1993; revised May 1994.

[1]Research supported in part by NSF grants DMS-92-04504 and DMS-93-05601. This manuscript was prepared using computer facilities supported in part by NSF Grants DMS-89-05292, DMS-87-03942 and DMS-86-01732 awarded to the Department of Statistics at the University of Chicago, and by the University of Chicago Block Fund.

AMS 1991 *subject classifications*. Primary 62E20, 62M09, 62M10, 62P10; secondary 60G42, 60G44, 62J99, 62M99.

Key words and phrases. Accuracy, Bartlett correction, empirical likelihood, likelihood inference, likelihood ratio test, martingale inference, stochastic simulation, survival analysis, time series.

confidence intervals with good accuracy properties, and also that the dual likelihood itself is a reasonably natural construction (Sections 6 and 7). An additional feature is that there are some surprising connections: in the case of independent data, the dual LR statistic coincides with Owen's empirical LR statistic [Owen (1988a, 1990)]. There are similar connections to point process likelihoods in survival analysis. This is discussed in Section 6. In both cases, this connection reduces nonparametric LR statistics to parametric ones, so that, for example, the existence of Bartlett correction becomes "obvious" in a heuristic sense. The relationship to parametric likelihood constructions (ordinary, partial, quasi and projective likelihoods) is discussed in Sections 3.3 and 4.2.

First, however, we give a description of existing technology for martingale inference (Section 2) and a review of the data structures to which martingale methods are most frequently applied (Section 3).

2. The state of the art. A number of data structures are amenable to analysis with the help of martingales. Most of the data types concerned can be characterized as longitudinal (cf. Section 3). Although the inference situations considered are quite diverse, the martingale structure provides a number of unifying characteristics, which makes it fruitful to work on them as a group. For one thing, the current way of setting tests and confidence intervals is mostly the same for all the data types considered. For another, the dual likelihood is also applicable to martingale inference situations in general.

The "baseline" method for martingale-based inference is to find a "score function" $m(\theta) = m_t(\theta)$ which is a martingale for the true value of the parameter θ, and then do inference based on this. The estimate $\hat{\theta}$ of θ is given by

$$m(\hat{\theta}) = 0,$$

and tests and confidence intervals can be based on the asymptotic standard normal distribution of either $m(\theta)/s$ or $\dot{m}(\hat{\theta})(\hat{\theta} - \theta)/s$, where s^2 is some estimate of the variance of the martingale (and similarly in multiparameter problems). The choice of s^2 which works with the greatest generality is the observed quadratic variation, evaluated either at $\hat{\theta}$ ($s^2 = [m(\hat{\theta}), m(\hat{\theta})]$) or under a null hypothesis ($s^2 = [m(\theta), m(\theta)]$). We shall refer to $m(\theta)/\sqrt{[m(\theta), m(\theta)]}$ and its square as martingale score statistics.

The definition of the quadratic variation $[m(\theta), m(\theta)]$ depends on the structure of the martingale $m(\theta)$. In most instances, $m_t(\theta)$ is a compensated sum of jumps,

$$m_t(\theta) = \sum_{0 \le s \le t} \Delta m_s(\theta) - \Lambda_t(\theta). \tag{2.1}$$

This includes the case of discrete time martingales, where the increments can be seen as jumps at fixed times. In this case, $\Lambda_t(\theta) \equiv 0$, so

$$m_n(\theta) = \sum_{i=1}^{n} \Delta m_i(\theta). \tag{2.2}$$

Whenever the martingale is given by (2.1), the quadratic variation is given by

$$[m(\theta), m(\theta)]_t = \sum_{0 \le s \le t} \Delta m_s(\theta)^2. \tag{2.3}$$

If $m_t(\theta) = (m_{1,t}(\theta), \ldots, m_{p,t}(\theta))$ is a vector, one can similarly define a quadratic covariation matrix through $[m(\theta), m(\theta)]_t = \Sigma_{0 \le s \le t} \Delta m_s \, \Delta m_s^{\mathrm{T}}$.

Formula (2.1) assumes that $m(\theta)$ does not have infinite total variation. The results in this paper are not subject to this restriction; our only assumption is that $(m_s(\theta))_{0 \le s \le t}$ should be right continuous with left limits (cadlag). We confine explicit discussion of this case, however, to Section 8, infinite total variation being rare in problems involving real data. Note that we also assume the "usual conditions," in a sense to be discussed in Section 8.

Examples of what these quantities look like are given in the next section. Rigorous conditions for $m(\theta)/\sqrt{[m(\theta), m(\theta)]}$ to be asymptotically standard normal are given in, for example, Theorem 3.2 of Hall and Heyde [(1980), page 58] and Theorem 2 of Rebolledo [(1980), page 273] and Theorem 5.1 of Helland [(1982), page 88]. The conditions one needs to impose on the other statistics mentioned are what is required to make the delta method work.

3. Data structures. Main data types under consideration are discussed in the following subsections. This is by no means an exhaustive review—just a set of motivating examples.

3.1. *Survival data.* Right-censored survival data usually give rise to estimators which can be analyzed with martingales. The most basic problem is the estimation of the survival distribution when there are no covariates. There are currently several competing methods for setting pointwise confidence intervals and global bands. One can use the asymptotic Gaussianity of the Kaplan–Meier [Kaplan and Meier (1958)] or Nelson–Aalen estimators [Nelson (1969); Aalen (1976, 1977, 1978)] or one can consider other estimators such as those in Thomas and Grunkemeier (1975) or the transformation methods considered in Borgan and Liestøl (1990). The reason for the existence of so many approaches is presumably due to a combination of two factors. On the one hand, the problem has substantial practical importance. On the other hand, the CLT tends to not work very well in the presence of certain types of heavy censoring [see Meier (1976) and also Latta (1981) in connection with comparing two survival distributions].

The Nelson–Aalen estimator corresponds to the "baseline" method. Let Λ be the cumulative hazard of patients with i.i.d. lifetimes, and assume for the purpose of this discussion that it is continuous [if the cumulative distribution $F(t)$ is continuous, then $\Lambda_t = -\ln(1 - F(t))$]. If one wishes to estimate $\theta = \Lambda_t$ at a fixed time point t, the martingale which is used in the Nelson–Aalen procedure is given by

$$m_t(\theta) = \int_0^t Y_s^{-1} \, dN_s - \Lambda_t \tag{3.1}$$

up to the time when the last patient ceases to be under observation, where Y_s is the number at risk at time s and the jumps of N represent the observed deaths of patients (so Λ is the compensator of the jumps). Also,

$$\begin{aligned}[m(\theta), m(\theta)] &= \left[m(\hat{\theta}), m(\hat{\theta})\right] \\ &= \int_0^t Y_s^{-2}\, dN_s,\end{aligned} \tag{3.2}$$

which is a commonly used choice for s^2; see Andersen, Borgan, Gill and Keiding [(1993), pages 180–183]. The statistics mentioned in Section 2 then all coincide, and asymptotic normality is guaranteed by, for example, the conditions in Aalen (1977).

In connection with this example, it should be emphasized that in this paper, we only focus on tests and confidence intervals for finitely many parameters. The dual likelihood ratio statistic described in Section 4 may be helpful in constructing confidence bands, too (by evaluating it at each time point rather than taking the whole of Λ as parameter), but we have not investigated this issue.

As far as survival data with covariates are concerned, we shall take as an example the regression model in Aalen (1980, 1989). The model is as follows. Patients 1 to n have survival distributions with cumulative hazards $H_1(t), H_2(t), \ldots$. The vector $H(t) = (H_1(t), \ldots, H_n(t))^T$ is given by

$$H(t) = \int_0^t Y(s)\, d\Lambda_s, \tag{3.3}$$

where $Y(s)$ is an $n \times p$ vector of (possibly time dependent predictable) regressors and Λ_s is a $p \times 1$ "cumulative" coefficient. The most standard case would be coefficients which do not vary over time, in which case the hazard rate $h(t) = H'(t)$ has the form

$$h(t) = Y(t)\alpha, \tag{3.4}$$

where α is a vector (so $\Lambda_t = \alpha t$). In estimating $\theta = \Lambda_t$, the martingale used by Aalen is

$$m_t(\theta) = \int_0^t X(s)\, dN_s - \Lambda_t, \tag{3.5}$$

where $X(s) = (Y(s)^T Y(s))^{-1} Y(s)^T$, up to the time when $Y(s)^T Y(s)$ is no longer of full rank. N_t is an n-dimensional vector whose ith component jumps from 0 to 1 if and when patient $\#i$ dies under observation. The (matrix) quadratic variation is

$$[m(\theta), m(\theta)]_t = \int_0^t X(s)\, d\tilde{N}(s) X(s)^T, \tag{3.6}$$

where $\tilde{N}(t)$ is the matrix with $N(t)$ on the diagonal and zeros everywhere else.

Other models which fall into the martingale framework are the partial likelihoods [Cox (1972, 1975), Andersen and Gill (1982), Wong (1986) and

Efron (1988)], as well as procedures for comparing populations [see, e.g., Aalen (1978) and Latta (1981)]. For books containing broad treatments of the survival analysis/martingale connection, see Fleming and Harrington (1991) and Andersen, Borgan, Gill and Keiding (1993). Also note that point process methods are not restricted in their application to survival data. Other applications include capture–recapture experiments [see, e.g., Becker and Heyde (1990)] and the infection model in Rida (1991).

3.2. *Time series data.* Autoregressive (AR) process inference is amenable to martingale treatment in a general setup where one observes pairs (X_n, Y_n), related by

$$Y_n = \sum_{i=1}^{p} \theta_i X_{i,n} + \varepsilon_n. \tag{3.7}$$

In the case of a linear AR model, $X_{i,n}$ would be Y_{n-i}. For nonlinear models [such as in Priestley (1988) and Tong (1990)], $X_{i,n}$ could be $f(Y_{n-i})$, or it could be something more complicated. The ε's are martingale increments with constant variance σ^2 (i.i.d. being a special case).

Least squares estimation for θ produces a score which is a martingale [see, e.g., Hall and Heyde (1980), Chapter 6.2]. The baseline method is very standard in this context. The score function is the vector $m_n(\theta) = (m_{1,n(\theta)}, \ldots, m_{p,n}(\theta))$ given by

$$m_{i,n}(\theta) = \sum_{k=1}^{n} X_{i,k}\hat{\varepsilon}_k(\theta), \tag{3.8}$$

where

$$\hat{\varepsilon}_k(\theta) = Y_k - \sum_{i=1}^{p} \theta_i X_{i,k}. \tag{3.9}$$

The quadratic variation is $[m_{i,\cdot}(\theta), m_{j,\cdot}(\theta)]_n = \Sigma_{k=1}^{n} X_{i,k}X_{j,k}\hat{\varepsilon}_k(\theta)^2$. Alternatively one can estimate covariance in the quasi-likelihood way by $\hat{\sigma}_n^2 \Sigma_{k=1}^{n} X_{i,k}X_{j,k}$, where $\hat{\sigma}_n^2$ is the variance estimate based on the $\hat{\varepsilon}_k(\theta)$'s (using either the true θ or $\hat{\theta}$).

Asymptotic normality is, for example, assured by the conditions in Chapters 3.2 or 6.3 in Hall and Heyde (1980) [cf. Chan (1990) and Tjøstheim (1990)]. The normal approximation, however, can be arbitrarily bad when one is close to the boundary of the domain of stationarity [indeed, it fails if one approaches the boundary on a suitable triangular array; see, e.g., Chan and Wei (1987)].

3.3. *Parametric and partial likelihood inference.* Though not necessarily longitudinal, the score function in a parametric problem is a martingale ["time" usually representing number of observations; see, e.g., Chapter 6 of Hall and Heyde (1980)].

The dual likelihood concept is somewhat less relevant in this context (it seems less interesting to mimic a likelihood when you already have one). In

instances, however, where there is uncertainty about the validity of the parametric model, a certain measure of robustness can be introduced by basing inference only on the martingale property of the score, rather than on the entire likelihood. In the simple case of a score test, this might mean letting $\hat{\sigma}^2$ be the observed quadratic variation of the score instead of the observed or estimated expected information [as proposed in Royall (1986)].

This approach is closely related to quasi-likelihood inference [see, e.g., Godambe and Heyde (1987) and Chapter 9 of McCullagh and Nelder (1989) and the references cited there] and projective likelihood inference [McLeish and Small (1992)]. In these approaches, however, assumptions concerning the form of the (conditional) second moment are made, leading to an estimate of variance not solely based on the martingale property of the quasi-score. Incorporating overdispersion [see, e.g., McCullagh and Nelder (1989)] into the model can partially or fully offset this, and corresponds to involving the quadratic variation of the martingale.

Likelihood inference is further discussed in Section 4.2.

3.4. *Independent samples.* Suppose one takes independent samples $X_1, \ldots, X_n, \ldots$ and wants to estimate a parameter θ given by

$$E\psi_i(X_i, \theta) = 0.$$

Clearly, the estimating equation

$$m_N(\theta) = \sum_{i=1}^{N} \psi_i(X_i, \theta)$$

is a martingale. For nonrandom N, this may seem quite uninteresting, since one then has the entire arsenal of techniques for independent data at one's disposal. This is belied, however, by the connection between dual and empirical likelihood. If N is based on a stopping rule, the possibility of using martingale methods is even more relevant.

4. Dual likelihood.

4.1. *Definition.* The device by which the likelihood is created is the following. Suppose $m(\theta)$ is a (p-dimensional) martingale at the true value of (the p-dimensional parameter) θ, given by formula (2.1). The dual likelihood is a function $L_\theta(\mu)$ of the parameter θ and a *dual parameter* μ (of the same dimension as θ), so that $l_\theta(\mu) = \ln(L_\theta(\mu))$ is a log likelihood in μ for fixed θ and so that

$$m(\theta)^T = \left(\partial l_\theta(\mu)/\partial\mu_1, \ldots, \partial l_\theta(\mu)/\partial\mu_p\right)|_{\mu=0}. \tag{4.1}$$

One can now go ahead and use likelihood methods on μ instead. For example, a test of $\theta = \theta_0$ can be carried out by doing a likelihood ratio test with $L_{\theta_0}(\mu)$ on the hypothesis that $\mu = 0$. We refer to this procedure as a "dual" LR test.

Similarly, confidence sets for θ can be created by inverting the test (for each θ).

The explicit form of the dual likelihood is the Doléans-Dade multiplicative martingale corresponding to $\mu m_t(\theta)$, alias the product integral of $\mu m_t(\theta)$,

$$L_\theta(\mu) = \exp(-\mu\Lambda_t(\theta)) \prod_{s \le t} (1 + \mu^T \Delta m_s(\theta)), \tag{4.2}$$

so that

$$l_\theta(\mu) = -\mu^T \Lambda_t(\theta) + \sum_{s \le t} \ln(1 + \mu^T \Delta m_s(\theta)). \tag{4.3}$$

References relating to the Doléans-Dade martingale include Doléans-Dade (1970), Jacod and Shirayev (1987) and Gill and Johansen (1990). [The formulas (4.2) and (4.3) assume that $m(\theta)$ does not have infinite total variation; the general formula is given in Section 8.] Computation of the dual LR statistic is discussed in Sections 4.2 and 6.

4.2. *Some whys.* Several questions are immediate. Why is (4.2) a likelihood or why does it have likelihood properties? Why this likelihood, rather than any other ones, and why does one want a likelihood in the first place, anyway?

Most of the rest of this paper is concerned with answering these questions. The following discussion is a summary explanation, with references to later sections.

At the risk of being simplistic, let us divide the desirable properties of likelihood inference into two categories: efficiency and accuracy. Efficiency has to do with terms ranging from "inferentially correct" to "uniformly most powerful." This property, obviously, vanishes when using the wrong model. So dual likelihood is not efficient, in the same way as partial likelihood is not efficient [Wong (1986)].

The accuracy properties, however, remain. Accuracy refers to the quality of the asymptotic approximation and to how close nominal and actual coverage probabilities (and type I errors) are to each other. (Improvement in accuracy is, for example, the main purpose of bootstrapping and Edgeworth correction.) The likelihood ratio statistic, which is our main focus, converges to the χ^2 distribution in a much nicer way than the studentized score statistic converges to the normal distribution. It gives, in other words, rise to asymptotically based tests and confidence intervals which have considerably greater accuracy than the ones derived from score statistics (see the discussion in Section 5.2). It is not necessary for the model to be correct for this to be true. The partial likelihood ratio statistic can generally be expected to have good accuracy properties [see Mykland and Ye (1992)], and the same is true for the dual likelihood ratio statistic (cf. the discussion in Section 5).

In fact, the score statistic from the dual likelihood (4.2) is

$$\frac{\dot{l}_{\theta_0}(0)}{-\ddot{l}_{\theta_0}(0)^{1/2}} = \frac{m(\theta_0)}{[m(\theta_0), m(\theta_0)]^{1/2}}$$

(and similarly in multiparameter problems). Hence, going from the martingale score statistic to the dual LR statistic is, in some sense, like changing to a better statistic in the same parametric problem. In particular, the efficiency of the dual LR procedure is, to first order, the same as for the martingale score test (cf. Section 5).

To sum up, the slogan is *same efficiency, better accuracy.*

In the particular case of martingales which are score functions in an ordinary, partial or projective likelihood model, this means that one can gain robustness without sacrificing accuracy or efficiency (at least in the Pitman sense). In comparison to quasi-likelihood, we conjecture that there will be a gain in accuracy, as quasi-likelihood is not a likelihood in the sense of satisfying Bartlett identities. We emphasize that partial and projective likelihood are likelihoods in this sense.

The reason for using the likelihood (4.2) in particular is that it is the actual likelihood ratio that occurs with a natural family of alternative hypotheses. Also, provided the class of probability distributions considered is sufficiently broad, (4.2) is the only likelihood for testing against this alternative. This issue is further discussed in Sections 7.1 and 7.2.

In addition, (4.2) is, in a sense, the most parsimonious likelihood ratio consistent with $m(\theta)$ being the desired score function (cf. Section 7.3). Admittedly, this argument is an aesthetic one only.

Is (4.2) a likelihood in the first place? The requirements for $L_{\theta_0}(\mu)$ to be a true likelihood (in μ, for fixed θ_0) is that $L_{\theta_0}(\mu)$ integrate to 1 under all probability distributions P for which $\theta(P) = \theta_0$, and that it should be nonnegative [so that $l_{\theta_0}(\mu)$ is defined for a fixed set of μ's]. This is what is needed for $L_{\theta_0}(\mu)$ to be of the form dQ/dP, where Q is a probability distribution.

Both these properties can break down (in which case the alternative hypothesis is a signed measure rather than a probability distribution), but not in ways that typically matter. The set where the dual log likelihood is defined is usually data dependent [we shall take the dual likelihood itself to be defined (and negative) even when $l_{\theta_0}(\mu)$ is not]. In fact, this can also happen to some extent in true likelihood problems. In terms of computing the LR statistic, the data dependent domain can be resolved as follows. There is always a neighborhood around 0 for which $L_{\theta_0}(\mu)$ is defined. Moreover

$$\frac{\partial^2 l_\theta(\mu)}{\partial\mu_i\,\partial\mu_j} = -\sum_{s\le t}\frac{\Delta m_{i,s}(\theta)\,\Delta m_{j,s}(\theta)}{\left(1+\mu^T\Delta m_s(\theta)\right)^2},$$

so $l_\theta(\mu)$ is strictly concave in this neighborhood (unless $[m(\theta), m(\theta)]_t = 0$). Hence the "dual MLE" $\hat{\mu}$ is unique if one restricts consideration to this neighborhood (which seems natural) and, in particular, the LR statistic is given by $2(l_\theta(\hat{\mu}) - l_\theta(0))$ (it is zero or undefined if $[m(\theta), m(\theta)]_t = 0$).

Furthermore, the moment properties of the LR statistic only depend on the likelihood structure through the Bartlett identities. These remain valid in this case [cf. Mykland (1994)], whether or not the likelihood is nonnegative.

The same consideration applies if the dual likelihood does not integrate to 1. The dual likelihood is always a local martingale [cf. page 59 of Jacod and Shiryaev (1987)], so both the Bartlett identities and the property of integrating to 1 remain true up to stopping times. Failure to integrate to 1 is a symptom of "knowing too much," and can be incorporated into the theory (cf. the end of Section 7.1).

Finally, note that we are not trying to improve the point estimate of θ. The assumption is that $m(\theta)$ is the desired score function, and our focus is only on the accuracy of tests and confidence intervals. $l_\theta(\mu)$ is a log likelihood in μ, but we are not estimating μ.

If the dual likelihood concept seems uncomfortable, one can alternatively think of the dual LR statistic as just a transformation of a martingale. It can be given a suitably neutral name, say a generalized Edgeworth–Fisher transform, denoted by T, defined for martingales $(m_s)_{0 \le s \le t}$ as the maximum of (4.3) in the relevant neighborhood of the origin. A technical advantage of this approach is that $T(m)$ is defined without reference to integrability conditions, and even for local martingales [cf., again, page 59 of Jacod and Shiryaev (1987)]. From the likelihood argument we can predict that inference based on $T(m)$ will have high accuracy, but one does not need a likelihood rationale to define $T(m)$.

On balance, we have decided to stay with the likelihood name, as it does convey some intuition. Whether it is correct to call this object a likelihood, however, is clearly debatable.

4.3. *Examples.* Before further discussing the theoretical aspects of this log likelihood, we explain what $L_\mu(\theta)$ looks like for the martingales discussed in Section 3.

EXAMPLE 1. In the case of the Nelson–Aalen estimator (still assuming the cumulative hazard Λ to be continuous), let $\theta = \Lambda_t$. Hence $\Lambda_t(\theta) = \theta$ and $\Delta m_s = \Delta N_s / Y_s$, so that

$$l_\theta(\mu) = -\mu\theta + \int_0^t \ln(1 + \mu/Y_s)\, dN_s. \tag{4.4}$$

One can now use this to set confidence intervals for θ or, equivalently, for

$$P(\text{survival up to time } t) = \exp(-\theta). \tag{4.5}$$

To see the merits of the procedure, consider the following simulation experiment.

Data on survival times of 20 patients were generated from the unit exponential distribution and censoring times were generated independently from the uniform (0, 1) distribution. Pointwise confidence intervals for the survival distribution were generated using (a) the inverted dual LR test and (b) the baseline method, that is, the usual martingale variance estimator setting standard errors for the Nelson–Aalen estimator [see (3.2)]. For comparison, we also included (c) Greenwood's formula, setting standard errors for

TABLE 1
*Simulation of nominal 95% confidence intervals for survival probabilities**

Time	Coverage Error (nominal − actual) (%)			% Intervals Computed
	Dual LR	Kaplan–Meier	Nelson–Aalen	
0.5	1.8	4.3	6.6	100.
0.75	1.9	3.9	6.4	92.6
0.9	5.1	7.4	13.0	59.5
0.95	5.9	7.9	16.7	38.7

*20 patients, survival distribution unit exponential, censoring distribution uniform [0, 1]. Confidence intervals are generated using data up to times 0.5, 0.75, 0.9 and 0.95, respectively. Standard errors for the Kaplan–Meier estimator are set using. Greenwood's formula; for the Nelson–Aalen estimator they are set using the square root of formula (3.2).

the Kaplan–Meier estimator. Note that in view of the discussion in Section 6, the dual LR statistic is the same as the profile LR statistic in this model. Intervals were computed using data up to times 0.5, 0.75, 0.9 and 0.95. The nominal level is 95%. The results are given in Table 1, and it is clear that the dual LR procedure outperforms the two other procedures, at least in this case. [Note that intervals were only included if there were still patients under observation at the relevant time (cf. the last column in Table 1). The number of samples was 1000 at each time.]

To illustrate the situation further, Figure 1 gives the qq plot for the square root of the dual LR statistic and the χ distribution for survival at time 0.5 and the qq plot of the absolute value for the studentized Nelson–Aalen estimator (the "baseline statistic") and the χ distribution at the same point on the time axis. The qq plot for the normalized Kaplan–Meier estimator is very similar (in shape) to the one for the Nelson–Aalen estimator (although the slope is slightly different).

It should be emphasized that we are not trying to improve on the Nelson–Aalen or Kaplan–Meier estimators as point estimates, only on the interval estimators which follow from using the point estimators in conjunction with the usual standard errors.

EXAMPLE 2. Aalen's regression model gives rise to a dual likelihood which is very similar to (4.4). If $\theta = \Lambda_t$, it has the form

$$L_\theta(\mu) = -\mu^T\theta + \sum_{j=1}^{n} \int_0^t \ln\bigl(1 + \mu^T X_j(s)\bigr)\, dN_j(s), \tag{4.6}$$

where $X_j(s)$ is column $\#j$ of the $p \times n$ matrix X and $N_j(s)$ represents patient $\#j$. Note that if one wishes to set a confidence interval for a scalar component θ_i of θ, one can make a dual likelihood for this component only:

$$L_{\theta_i}(\mu) = -\mu\theta_i + \sum_{j=1}^{n} \int_0^t \ln\bigl(1 + \mu X_{ij}(s)\bigr)\, dN_j(s). \tag{4.7}$$

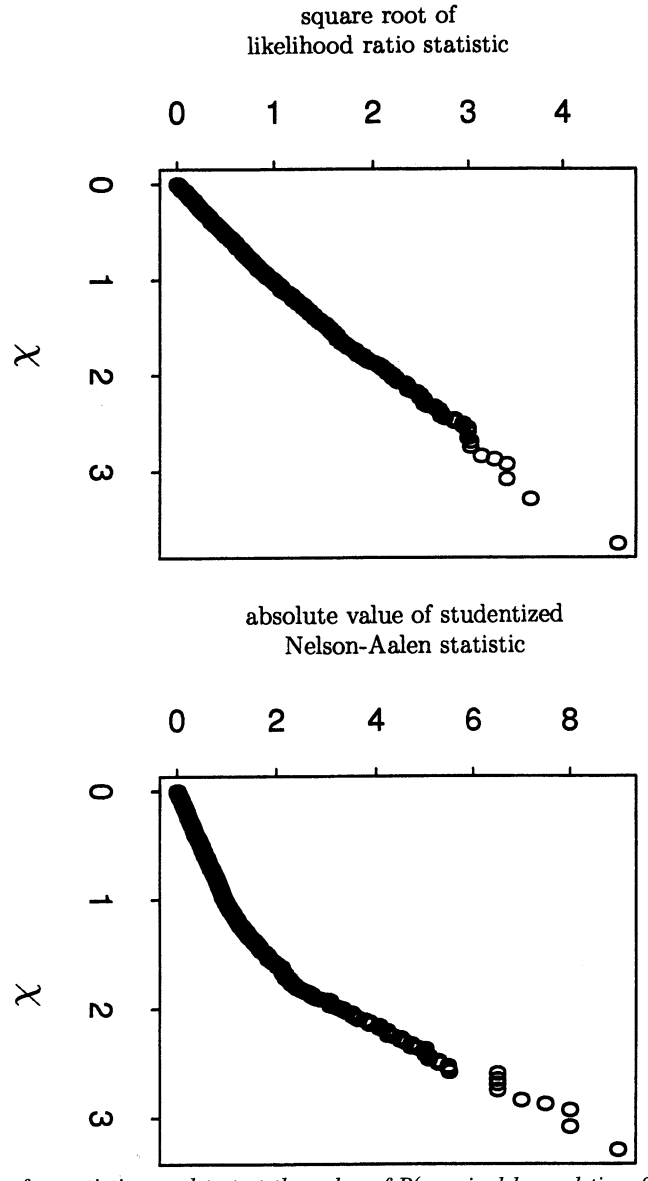

FIG. 1. *qq plots for statistics used to test the value of P(survival beyond time* 0.5*) in the same experiment as in Table* 1.

EXAMPLE 3. If the martingale evolves in discrete time, the compensator $\Lambda_t(\theta) = 0$, whence (4.3) becomes

$$l_\theta(\mu) = \sum_{s \le t} \ln\bigl(1 + \mu^T \Delta m_s(\theta)\bigr). \tag{4.8}$$

This covers the case of AR processes, likelihoods in discrete time and independent data.

4.4. *Alternative likelihood constructions.* Equation (4.3) is not necessarily the only way to achieve (4.1). To state the obvious, if $m(\theta)$ is a score function in a likelihood problem, one can, of course, take L to be the true log likelihood. The same is true for partial likelihoods. Another likelihood construction is the one of McLeish and Small (1992). In some instances, it may also be useful to take

$$l_\theta(\mu) = \mu m(\theta) - \mu^2\kappa(m(\theta), m(\theta))/2! - \mu^3\kappa(m(\theta), m(\theta), m(\theta))/3! - \cdots, \tag{4.9}$$

where the κ's are the cumulant variations defined in Mykland (1994). As far as a general procedure is concerned, however, (4.2) seems to be the most widely applicable construction.

The log likelihoods in (4.3) and (4.9) have previously been used in Mykland (1994) to provide part of the argument to extend the Bartlett identities [Bartlett (1953a, b)] to martingales. This is useful in that it greatly eases the calculation of cumulants for many martingales and it also simplifies theoretical arguments (the martingale CLT is quite easy to show with the Bartlett identities, and the same appears to be true for martingale asymptotic expansions).

5. Asymptotics of the dual likelihood ratio statistic.

5.1. *First order properties.* As far as first order asymptotic behavior is concerned, the dual LR statistic has the same properties as the corresponding score-type statistic, at least as far as asymptotic laws and Pitman efficiency are concerned.

To see this, let $m_t^n = m_t^n(\theta)$, $0 \le t \le t_n$, be a triangular array of martingales (or even local martingales) and let LR_n be the corresponding dual LR statistic, that is, $LR_n = 2\sup_\mu l_\theta(\mu)$, where the supremum is taken in the neighborhood of zero where $L_\theta(\mu)$ is nonnegative, and where $L_\theta(\mu)$ is given by (4.2) or, more generally, by (8.2). Also, let the score statistic be given by

$$S_n = \left(m_{t_n}^n\right)^T [m^n, m^n]_{t_n}^{-1} m_{t_n}^n$$

and let $\underline{\lambda}_n$ be the smallest eigenvalue of $[m^n, m^n]_{t_n}$. We then have the following result.

THEOREM 1. *Suppose that S_n is tight and that* $\sup_{0 \le t \le t_n} \|\Delta m_t^n\|^2/\underline{\lambda}_n = o_p(1)$. *Then*

$$LR_n = S_n + o_p(1). \tag{5.1}$$

The second regularity condition guarantees that the jumps of (m_t^n) are asymptotically negligible. The proof is in Section 8.

Since (5.1) remains true under contiguous alternatives, our remarks about Pitman efficiency also follow from the theorem. As far as limit laws are concerned, one can now apply one's favorite martingale central limit theo-

rem [from, e.g., Hall and Heyde (1980), Rebolledo (1980), Helland (1982), Jeganathan (1982), Jacod and Shiryaev (1987)] to get that

$$LR_n \to \chi_p^2 \tag{5.2}$$

in law. The references cited cover all the data structures discussed in Section 3. In the survival analysis examples, see also Aalen (1977, 1978, 1980, 1989), Fleming and Harrington (1991) and Andersen, Borgan, Gill and Keiding (1993).

Theorem 1 also means, of course, that the dual LR statistic and the score statistic fail to have asymptotic χ^2 distribution in the same instances. This is, for example, the case when $\theta = 1$ in the AR(1) process

$$X_{n+1} = \theta X_n + \varepsilon_{n+1}; \tag{5.3}$$

compare Chan and Wei (1988).

The fact that these two types of statistic fail at the same time does not mean, however, that they behave equally badly at points in the parameter space which are close to singularities. It is usually the hope that the LR statistic behaves better in such cases. This is partially based on numerical experience [cf. Table 1, for example, and also the remarks on page 189 in McCullagh (1987)] and partially on second order asymptotics.

5.2. *Higher order asymptotics.* It is here that the two types of statistics part company. The dual LR statistic will tend to inherit the Bartlett correctability of the ordinary parametric LR statistic [cf. Lawley (1956) and also McCullagh (1987), Chapter 7]. A rigorous statement to this effect can be made in the one parameter case, in view of the calculations in Chapter 7 of McCullagh (1987) and of Theorem 7 in Mykland (1995); see below. We have not investigated rigorous conditions in the multiparameter case, but the discussion in McCullagh (1987) suggests that these would be quite weak (if one is willing to ignore lattice issues).

Even if one does not actually carry out the Bartlett correction, the accuracy properties are quite nice. Arguably, this can be seen from the fact that the Bartlett factor is then the main error term. Also, the χ^2 approximation tends to hold up well in the tail of the distribution, in view of the connection to the saddlepoint approximation and Barndorff-Nielsen's formula; see, for example, Barndorff-Nielsen and Cox (1984), McCullagh (1987) and Reid (1988). We have not, however, investigated the conditions for this property to be inherited by the dual LR statistic.

To state a rigorous result on Bartlett correction, let

$$F_n(x) = P\big(LR_n/E(LR_n) \le x\big) \tag{5.4}$$

and let $\chi_1^2(x)$ be the cdf of the χ^2 distribution with 1 degree of freedom. As is generally the case for Edgeworth expansions for martingales, it is not known how to show general pointwise results [cf. the discussion at the end of Section 1 of Mykland (1992)]. In this instance, we shall show that

$$\int g(x)\, d\big(F_n(x) - \chi_1^2(x)\big) = o(n^{-1}) \tag{5.5}$$

for all sufficiently smooth g [specifically ones that have five continuous derivatives and vanish outside a compact set; the exact requirements on g are, of course, quite irrelevant, the point being that one can assert something like (5.5) at all under broad martingale conditions]. The regularity conditions are those of Theorem 7 of Mykland (1995) and, in addition, that

$$E([m^n, m^n, m^n, m^n]_{t_n}/n)^3 = O(1). \tag{5.6}$$

The theorem is then as follows (the proof is in Section 8).

THEOREM 2. *Under the regularity conditions stated above,* (5.5) *holds for the specified class of functions* g.

A similar argument would show that the signed square root of the dual likelihood ratio statistic [with the sign of $\partial l_\theta(\mu)/\partial\mu$] can be mean and variance adjusted to be $N(0,1) + o(n^{-1})$ (again with a suitable convergence type).

If one practically wishes to use correction factors for the LR statistic or its signed square root, one would normally have to estimate the relevant quantities. We have not investigated how to do this. For the i.i.d. case, estimates can be found in DiCiccio and Romano (1989) in the context of empirical likelihood (cf. the next section).

6. Dual and nonparametric LR statistics. An interesting feature of both (4.4) and (4.8) is that the dual LR statistic coincides with LR statistics derived from nonparametric likelihood considerations. In the case of discrete time, if the increments are independent, the dual and empirical [see, e.g., Owen (1988a, 1990)] LR statistics are the same. Specifically, if there are n observations, the dual LR statistic coincides with the supremum of $-2(\sum \ln(p_i) + n \ln(n))$ subject to $p_i \geq 0$,

$$\sum p_i \, \Delta m_i(\theta) = 0 \tag{6.1}$$

and

$$\sum p_i = 1, \tag{6.2}$$

this being the empirical likelihood for an estimating function with increment $\Delta m_i(\theta)$ [e.g., $\Delta m_i(\theta)$ could be $\psi(X_i, \theta)$ for independent X_i's; for discussions of empirical likelihood with estimating functions, see Kolaczyk (1994) and Qin and Lawless (1994)]. The equality between the empirical and dual LR statistics remains, of course, true for nonindependent martingale increments if one defines the empirical likelihood as if the increments were independent. It is, however, unclear what the likelihood rationale for such a construction would be.

In the case of (4.4), the dual LR statistic is the same as the nonparametric LR statistic based on the point process likelihood from Jacod (1975) [see also Jacod and Memin (1976), Gill and Johansen (1990), Greenwood and Wefelmeyer (1990) and Andersen, Borgan, Gill and Keiding (1993)] under the

assumption that the cumulative hazard is continuous (this is the same likelihood which makes the Nelson–Aalen estimator the nonparametric MLE; the LR statistic corresponding to the likelihood for which the Kaplan–Meier estimator is the MLE is given in Thomas and Grunkemeier (1975)].

Specifically, the nonparametric log likelihood has the form

$$l(\tilde{\Lambda}, \Lambda) = \int_0^t \ln \frac{d\tilde{\Lambda}}{d\Lambda}(s)\, dN_s - \int_0^t Y_s\big(d\tilde{\Lambda}(s) - d\Lambda(s)\big), \tag{6.3}$$

where Λ and $\tilde{\Lambda}$ are two cumulative hazards to be compared. If one lets $\tilde{\Lambda}$ be the Nelson–Aalen estimator and if Λ maximizes l subject to the constraint $\Lambda_t = \theta$, the resulting value of l is the same as the maximum of (4.4) with respect to μ.

In both the above instances, the dual parameter μ is the Lagrange multiplier in the optimization problem arising from the nonparametric likelihood. In the case of (6.3), the Lagrangian is

$$l(\tilde{\Lambda}, \Lambda) - \mu(\Lambda_t - \theta); \tag{6.4}$$

it is readily verified that for given μ, the maximum of (6.4) subject to $\Lambda_t = \theta$ is (4.4). For empirical likelihood, μ is proportional to the multiplier of the constraint (6.1); compare with statements (2.10) and (2.11) in Owen [(1990), page 100].

These connections give rise to a certain mutual validation between the nonparametric and dual likelihoods. They also suggest the possibility of substantial feedback between the two approaches.

An immediately obvious example of this is that the existence of a Bartlett correction for the empirical and point process LR statistics can now be seen to follow as a corollary to the existence of such a correction for the parametric LR statistic [the existence of corrections in the empirical likelihood case has been proved directly in DiCiccio and Romano (1989) and DiCiccio, Hall and Romano (1991); the same result is (as far as we know) previously not known in the point process case].

Another example is that the computation of the empirical LR statistic has been studied, inter alia, in Owen (1988b). These results clearly carry over to the dual LR statistic based on discrete time martingales. Software for empirical likelihood can similarly be used in this case.

A question which naturally presents itself is whether there is a universal "primal" likelihood which in every case gives the same LR statistic as the dual likelihood. It is at the moment not obvious what such a construction would be like.

7. The alternative in dual likelihood.

7.1. *The alternative hypothesis.* Suppose we want to test a null hypothesis that $\theta = \theta_0$ with the dual LR statistic. Set $m_s = m_s(\theta_0)$. The alternative induced by the dual likelihood implies the following. Set $dP_\mu/dP = L_{\theta_0}(\mu)$,

where P is consistent with the null hypothesis [i.e., $\theta(P) = \theta_0$]. Under P_μ, it then holds that

$$m_s - \langle m, m\rangle_s \mu \tag{7.1}$$

is a (local) martingale; $\langle m, m\rangle$ is the predictable quadratic variation, alias the compensator of $[m, m]$ (in particular, it is a $p \times p$ matrix, where p is the dimension of θ and μ).

EXAMPLE 4. If m is given by formula (2.2) (i.e., lives in discrete time), the predictable quadratic variation is given by

$$\langle m, m\rangle_n = \sum_{i=1}^{n} E\left(\Delta m_i \, \Delta m_i^T | \mathcal{F}_{i-1}\right), \tag{7.2}$$

where $(\mathcal{F}_n)$ is the filtration describing the history of the process. In the case of a one-dimensional estimating function ψ based on i.i.d. data $X_1, \ldots, X_n$, that is,

$$\Delta m_i = \psi(X_i, \theta_0), \tag{7.3}$$

it has the form

$$\langle m, m\rangle_n = nE\psi(X, \theta_0)^2. \tag{7.4}$$

Hence, under the alternative,

$$E\psi(X, \theta_0) = \mu E\psi(X, \theta_0)^2, \tag{7.5}$$

which by reparametrization means

$$E\psi(X, \theta_0) = \mu'. \tag{7.6}$$

Similarly, in the case of the time series (3.7),

$$\langle m, m\rangle_n = \sigma^2 \sum_{k=1}^{n} X_{\cdot,k} X_{\cdot,k}^T. \tag{7.7}$$

Statement (7.1) is now the same as $m(\theta + \sigma^2\mu)$, the martingaleness of which is compatible with a model

$$Y_n = \sum_{i=1}^{p} \left(\theta_i + \sigma^2\mu_i\right) X_{i,n} + \varepsilon_n, \tag{7.8}$$

where the ε_n are martingale increments.

In the case of the Nelson–Aalen estimator, we are outside the framework of (7.2). Here,

$$\langle m, m\rangle_t = \int_0^t Y_s^{-1} \, d\Lambda_s. \tag{7.9}$$

There are the "usual" caveats to the description above, relating to integrability and to possible negativity of the dual likelihood. The latter is easy to get around, as the concepts of martingale and local martingale have a natural extension to finite signed measures that integrate to 1. As far as integrability

is concerned, one needs to require that $\langle m, m\rangle_s$ is defined under P. A little stochastic calculus then yields that $(m_t - \langle m, m\rangle_t \mu)L_{\theta_0}(\mu)_t$ is a local martingale under P [cf. Doléans-Dade (1970) and Lenglart (1977); see also Section I.4f of Jacod and Shiryaev (1987), pages 58–61]. Hence, suitably localized, (7.1) is a martingale under P_μ. Formally, one can assert the following theorem.

THEOREM 3. *Suppose that* $(m_s)_{0 \le s \le t}$ *is a martingale under* P. *Then, there is a sequence of stopping times* $\{\tau_n\}$, $P(\tau_n = t) \to 1$, *so that* $dP_\mu^{(n)}/dP = L_{\theta_0}(\mu)_{\tau_n}$ (*the dual likelihood ratio evaluated at time* τ_n) *defines a finite signed measure satisfying* $P_\mu^{(n)}(\Omega) = 1$, *where* Ω *is the entire sample space. If, in addition,* $(\langle m, m\rangle_s)_{0 \le s \le t}$ *is defined under* P, *then* $m_s - \langle m, m\rangle_s \mu$ *is a local martingale under* $P_\mu^{(n)}$ *for* $0 \le s \le \tau_n$ (*i.e.,* $m_{s_1 \wedge \tau_n} - \langle m, m\rangle_{s_1 \wedge \tau_n} \mu$ *is a local martingale*).

The result follows from the above discussion and from further use of Theorem 1 of Doléans-Dade (1970) or Theorem I.4.61 of Jacod and Shiryaev [(1987), page 59].

7.2. *Extremality, and the uniqueness of the alternative.* The interesting fact is that there is a partial converse to the result of the previous section: Let M_0 be the class of probability measures P consistent with the null hypothesis that $\theta = \theta_0$, that is, for which $\theta(P) = \theta_0$. If the class M_0 is sufficiently big, then the dual likelihood (4.2) is the only possible likelihood for testing against the alternative (7.1). To be precise, "sufficiently big" means that M_0 contains extremal elements from the set M of probability measures under which $(m_s)_{0 \le s \le t}$ is a local martingale with $m_0 = 0$ [M and M_0 are defined with reference to the filtration $(\mathcal{F}_s)_{0 \le s \le t}$ generated by the data or, more generally, the history of the process]. Before proceeding with the formal result, here are some examples.

EXAMPLE 5. For a one-dimensional estimating equation based on i.i.d. data, let the martingale increments be given by (7.3). M is given by

$$M = \{P\colon E_P|\psi(X, \theta_0)| < \infty,\ E_P\psi(X, \theta_0) = 0\}. \tag{7.10}$$

Extremal elements of M include the distributions which are degenerate at x's for which $\psi(x, \theta_0) = 0$ and ones which are concentrated at pairs (x_1, x_2) which are not zeros of $\psi(\cdot, \theta_0)$, but for which $\alpha\psi(x_1, \theta_0) + (1 - \alpha)\psi(x_2, \theta_0) = 0$ for some $\alpha \in (0, 1)$.

Hence, if, for example, M_0 is the subset of M for which $E_P\psi(X, \theta_0)^2 < \infty$, then M_0 contains extremal elements from M and is therefore covered by Theorem 4 below.

EXAMPLE 6. For the Nelson–Aalen estimator, let M be the set of probability measures P so that $m_{t \wedge \tau}$ is a martingale, where m_t is given by (3.1) and τ is the time when the last patient ceases to be under observation. Extremal

elements of M include all probabilities P for which the censoring times are nonrandom. This follows by combining Theorem 11.2 in Jacod [(1979), page 338] with Theorem 12.35 in Elliot [(1982), pages 146–147].

We now proceed to the uniqueness result (the proof is in Section 8).

THEOREM 4. *Suppose that $P \in M_0$ is extremal in M and that $\mathcal{F}_0$ only contains sets of P-probability 0 or 1. Let τ be a stopping time and let Q be a finite signed measure on $\mathcal{F}_\tau$, absolutely continuous with respect to P. If $m_s - \langle m, m\rangle_s \mu$ is a local martingale under Q for $0 \le s \le \tau$, then*

$$\frac{dQ}{dP} = L_{\theta_0}(\mu)_\tau.$$

7.3. *Alternative alternatives.* If one wants to consider alternatives other than the ones consistent with (7.1) being a martingale, additional options are obviously available. For simplicity, we confine our discussion to alternatives which are probability measures and to a one-dimensional (m_s).

Suppose one has a family of probabilities $\{Q_\mu\}$ for which $Q_\mu \sim P$ and set $Z_t(\mu) = dQ_\mu/dP$. Let $Z_s(\mu) = E_P(Z_t(\mu)|\mathcal{F}_s)$. One can then set

$$\nu_t(\mu) = \int_0^t Z_{s-}^{-1}(\mu)\, dZ_s. \tag{7.11}$$

The $(\nu_s(\mu))$'s are local martingales under P. $Z_t(\mu)$ is now the Doléans-Dade exponential martingale based on $\nu_t(\mu)$ [cf. Theorem 1 of Doléans-Dade (1970) or Chapter I.4f of Jacod and Shiryaev (1987)] and one can now use $Z_t(\mu)$ as a likelihood, similar to the dual likelihood. Under Q_μ,

$$m_s - \langle m, \nu(\mu)\rangle_t \tag{7.12}$$

is a local martingale [cf. Lenglart (1977)]. Subject to regularity conditions,

$$\frac{d}{d\mu} \ln Z_s(\mu)\bigg|_{\mu=0} = \dot{\nu}_s(0), \tag{7.13}$$

$$\frac{d^2}{d\mu^2} \ln Z_s(\mu)\bigg|_{\mu=0} = -\ddot{\nu}_s(0) - [\dot{\nu}(0), \dot{\nu}(0)]_s \tag{7.14}$$

and

$$\frac{d^3}{d\mu^3} \ln Z_s(\mu)\bigg|_{\mu=0} = \dddot{\nu}_s(0) - 3[\dot{\nu}(0), \ddot{\nu}(0)]_s + 2[\dot{\nu}(0), \dot{\nu}(0), \dot{\nu}(0)]_s, \tag{7.15}$$

where $[\dot{\nu}(0), \dot{\nu}(0), \dot{\nu}(0)]_s$ is the optional cube variation of $\dot{\nu}(0)$ [cf. Section 6 of Mykland (1994)]. Hence, if one wishes a likelihood whose score function at

$\mu = 0$ is m_t, then one must have $\dot{\nu}_s(0) = m_s$. If one further wishes the observed information at 0 to be $-[m, m]_t$, then $\ddot{\nu}_s(0) = 0$. In other words,

$$\nu_t(\mu) = \mu m_t + \frac{1}{3!}\mu^3 \dddot{\nu}_t(0) + O_p(\mu^4) \tag{7.16}$$

and

$$\ln Z_t(\mu) = l_{\theta_0}(\mu) + \frac{1}{3!}\mu^3 \dddot{\nu}_t(0) + O_p(\mu^4). \tag{7.17}$$

In some sense, therefore, $\nu_t(\mu) = \mu m_t$ [which makes $Z_t(\mu)$ the dual likelihood] is a quadratic approximation to an arbitrary alternative consistent with the martingale score statistic.

In particular, it is worth pointing out that if $\nu_t(\mu) = \mu \tilde{m}_t$, then $\tilde{m} = m$. In the discrete time case, for example, one cannot preserve efficiency properties by using a dual likelihood of the form

$$\sum_{i=1}^{n} \ln(1 + \mu c_i \Delta m_i(\theta)), \tag{7.18}$$

where the c_i's are constants that are different from 1. Doing so would, instead, correspond to using an estimating equation of the form

$$\sum_{i=1}^{n} c_i \Delta m_i(\theta), \tag{7.19}$$

which would amount to a reweighting the observations.

8. Of null sets and infinite variation. Before winding up the paper, we propose to entertain the reader with some technical remarks.

First of all, a general local martingale $(m_s)_{0 \le s \le t}$ may have infinite total variation. To see the structure of the dual likelihood in this case, note that there is a unique decomposition (if $m_0 = 0$)

$$m_t = m_t^d + m_t^c, \tag{8.1}$$

where m^c is continuous and m^d is purely discontinuous [cf. Theorem I.4.18 of Jacod and Shiryaev (1987), pages 42 and 43]. The dual likelihood is now in the form

$$\begin{aligned} &\exp(\mu^T m_t^c - \tfrac{1}{2}\mu^T \langle m^c, m^c \rangle \mu) \\ &\quad \times \exp(\mu^T m_t^d) \prod_{s \le t} (1 + \mu^T \Delta m_s) \exp(-\mu^T \Delta m_s) \end{aligned} \tag{8.2}$$

[cf. Doléans-Dade (1970) or Chapter I.4f of Jacod and Shiryaev (1987), pages 58–61]. Except when otherwise explicitly stated, the results and comments of this paper apply equally to (8.2); all the statements in Sections 5 and 7 hold, and if one has occasion to compute the LR statistic in this more general framework, the dual log likelihood is still convex.

The reason why we have not otherwise discussed martingales with continuous components is that such martingales must have infinite total variation.

This is bound to be a highly unusual situation when dealing with real data. Note that also m^d may have infinite total variation, which is the reason why (8.2) looks different from (4.2) even when $m_t^c = 0$. If m^d has finite variation, then it is of the form (2.1), and the dual likelihood has the form (4.2).

A more arcane point is that we are making the "usual assumptions" relating to null sets [cf. Definition I.1.3 of Jacod and Shiryaev (1987), page 2]. Specifically, whenever we are dealing with a probability measure P, we implicitly replace the original filtration $(\mathscr{F}_t)$ with the augmented filtration $(\mathscr{F}_t^P)$ [cf. I.1.4 of Jacod and Shiryaev (1987), page 3]. The usual assumptions are probably not necessary [see Jacod (1979), von Weizsäcker and Winkler (1990) and Andersen, Borgan, Gill and Keiding (1993)], but they make it possible to draw on a larger body of sources relating to stochastic calculus.

We now turn to the proofs.

PROOF OF THEOREM 1. Write $[m^n, m^n]_{t_n} = Q_n^T D_n Q_n$, where D_n is diagonal and Q_n is orthogonal, and set $U_n = D_n^{-1/2} Q_n m_{t_n}^n$. Define

$$\tilde{l}_\theta(\mu) = l_\theta\big(Q_n^T D_n^{-1/2}\mu\big) \tag{8.3}$$

and

$$\nu_n = \Big(\sup_{0 \le t \le t_n} \|\Delta m_t^n\|\Big)/\underline{\lambda}_n^{1/2}. \tag{8.4}$$

Observe that for $\|\mu\| < (\sup_{0 \le t \le t_n} \|\Delta m_t^n\|)^{-1}$,

$$l_\theta(\mu) = \mu^T m_{t_n}^n - \tfrac{1}{2}\mu^T [m^n, m^n]_{t_n} \mu + \tfrac{1}{3}\sum_{s \le t} \big(\mu^T \Delta m_s^n\big)^3 - \cdots, \tag{8.5}$$

whence

$$\begin{aligned} &|l_\theta(\mu) - \mu^T m_{t_n}^n + \tfrac{1}{2}\mu^T [m^n, m^n]_{t_n} \mu| \\ &\qquad \le \mu^T [m^n, m^n]_{t_n} \mu f\Big(\|\mu\| \sup_{0 \le t \le t_n} \|\Delta m_t^n\|\Big), \end{aligned} \tag{8.6}$$

where

$$f(x) = -(\ln(1-x) + x)/x^2 - \tfrac{1}{2}. \tag{8.7}$$

Hence, for $\|\mu\| < \nu_n^{-1}$,

$$|\tilde{l}_\theta(\mu) - \mu^T U_n + \tfrac{1}{2}\mu^T\mu| \le \mu^T\mu f(\|\mu\|\nu_n). \tag{8.8}$$

It is then immediate that $\hat{\mu}_n$ (the MLE) is tight if U_n is tight [by convexity of $l_\theta(\cdot)$ and hence $\tilde{l}_\theta(\cdot)$], and it follows that (5.1) holds. □

PROOF OF THEOREM 2. The real work of this proof is done in Lawley (1956); see also McCullagh [(1987), Chapter 7.4]. What needs to be done to show Theorem 2 is to control the remainder term in formula (7.11) of McCullagh [(1987), page 211] and then to show that the right-hand side of that formula has an Edgeworth expansion of the required type (note that unlike McCullagh, we only consider the one-dimensional case). The latter is a

straightforward consequence of Theorems 5 and 6 of Mykland (1995) and it is derived in much the same way as Theorem 7 of that paper.

To control the remainder term, we shall suppose that we are in a set $\mathscr{E}_n$ of the sample space, which we shall gradually define. $\mathscr{E}_n$ will be made to satisfy

$$P(\mathscr{E}_n) = 1 - o(n^{-1}) \tag{8.9}$$

and it is enough to show that

$$ER_n I_{\mathscr{E}_n} = o(n^{-1}), \tag{8.10}$$

where R_n is the remainder term and $I_{\mathscr{E}_n}$ is the indicator function for $\mathscr{E}_n$.

First suppose that on $\mathscr{E}_n$,

$$|\hat{\mu}_n| \sup |\Delta m_s^n| \le k_n, \tag{8.11}$$

where k_n is nonrandom and $o(1)$. It is then easy to show that, for μ_n^* between 0 and $\hat{\mu}_n$,

$$l_\theta^{(k)}(\mu_n^*) = (-1)^{k-1}(k-1)!\underbrace{[m^n, \dots, m^n]_{t_n}}_{k \text{ times}}(1 + w_n), \tag{8.12}$$

where $|w_n| \le c_n$ and c_n is nonrandom and $o(1)$ (c_n depends on k). Here $[m^n, \dots, m^n]$ is the kth order variation of m^n; see Section 6 of Mykland (1994). Set $\hat{\delta}_n = \hat{\mu}_n \sqrt{n}$. The LR statistic is then given by

$$\begin{aligned} \tfrac{1}{2} LR_n &= \hat{\delta}_n\left(m_{t_n}^n/\sqrt{n}\right) - \tfrac{1}{2}\hat{\delta}_n^2 [m^n, m^n]_{t_n}/n \\ &\quad + \tfrac{1}{3}\hat{\delta}_n^3 [m^n, m^n, m^n]_{t_n}/n^{3/2} \\ &\quad - \tfrac{1}{4}\hat{\delta}_n^4 [m^n, m^n, m^n]_{t_n}(1 + w_n')/n^2, \end{aligned} \tag{8.13}$$

whereas the likelihood equation yields

$$\begin{aligned} m_{t_n}^n/\sqrt{n} - \hat{\delta}_n [m^n, m^n]_{t_n}/n + \hat{\delta}_n^2 [m^n, m^n, m^n]_{t_n}/n^{3/2} \\ - \hat{\delta}_n^3 [m^n, m^n, m^n, m^n](1 + w_n'')/n^2 = 0. \end{aligned} \tag{8.14}$$

For our purpose, it is easiest to work with (8.13) $- \frac{1}{2}\hat{\delta}$(8.14), that is,

$$\begin{aligned} \tfrac{1}{2} LR_n &= \tfrac{1}{2}\hat{\delta}_n\left(m_{t_n}^n/\sqrt{n}\right) - \tfrac{1}{6}\hat{\delta}_n^3 [m^n, m^n, m^n]_{t_n}/n^{3/2} \\ &\quad + \tfrac{1}{4}\hat{\delta}_n^4 [m^n, m^n, m^n, m^n]_{t_n}(1 + w_n'')/n^2. \end{aligned} \tag{8.15}$$

Suppose that we let $\hat{\delta}_n^{(i)}$ be the stochastic expansion of $\hat{\delta}_n$ up to $o(n^{-i/2})$, so that, for example, $\hat{\delta}_n^{(0)} = \mu_{2,n}^{-1}(m_{t_n}^n/\sqrt{n})$, where $\mu_{2,n} = E[m^n, m^n]_{t_n}$. Also suppose that on $\mathscr{E}_n$,

$$c_n'' \ge \frac{[m^n, m^n]_{t_n}}{n} \ge c_n', \tag{8.16}$$

where c_n' and c_n'' are nonrandom and bounded away from 0 and infinity, respectively. Replace $\hat{\delta}_n$ by $\hat{\delta}_n^{(i)}$ in (8.15), where i is chosen minimal in each

term to create an approximation up to $o_p(n^{-1})$, and where the second term on the right-hand side of (8.15) is considered as two terms by decomposing

$$[m^n, m^n, m^n]_{t_n} = [m^n, m^n, m^n]_{t_n} - n\mu_{3,n} + n\mu_{3,n}. \tag{8.17}$$

Then (8.10) holds, in view of our assumptions, provided

$$\|m^n_{t_n}/\sqrt{n}\|_p = O(1), \tag{8.18}$$

$$\left\| n\left(\hat{\delta}_n^{(2)} - \hat{\delta}_n\right) I_{\mathscr{E}_n} \right\|_q = o(1) \tag{8.19}$$

and

$$\left\| n^{1/2}\left(\hat{\delta}_n^{(1)} - \hat{\delta}_n\right) I_{\mathscr{E}_n} \right\|_r = o(1), \tag{8.20}$$

where $p^{-1} + q^{-1} \le 1$, $p^{-1} + 2r^{-1} \le 1$, $r \ge 3$ and $p \ge 12$. This is because, in analogy to (8.14),

$$m^n_{t_n}/\sqrt{n} - \hat{\delta}_n[m^n, m^n]_{t_n}(1 + w_n'''')/n = 0. \tag{8.21}$$

By operating on (8.14), (8.21) and the analogous expansion up to the $[m^n, m^n, m^n]_{t_n}$ term, it is easy to see that (8.20) holds provided

$$\left\| \left(\frac{m^n_{t_n}}{\sqrt{n}} \right)^2 \frac{[m^n, m^n, m^n]_{t_n}}{n} I_{\mathscr{E}_n} \right\|_r = O(1). \tag{8.22}$$

Similarly, (8.19) holds if

$$\begin{aligned} &\left\| \left(\frac{m^n_{t_n}}{\sqrt{n}} \right)^4 \left(\frac{[m^n, m^n, m^n]_{t_n}}{n} \right)^2 I_{\mathscr{E}_n} \right\|_q \\ &+ \left\| \left(\frac{m^n_{t_n}}{\sqrt{n}} \right)^3 \frac{[m^n, m^n, m^n]_{t_n}}{n} \sqrt{n} \left(\frac{[m^n, m^n]_{t_n}}{n} - \mu_{2,n} \right) I_{\mathscr{E}_n} \right\|_q \\ &+ \left\| \left(\frac{m^n_{t_n}}{\sqrt{n}} \right)^3 \frac{[m^n, m^n, m^n, m^n]_{t_n}}{n} I_{\mathscr{E}_n} \right\|_q = O(1). \end{aligned} \tag{8.23}$$

Since, on $\mathscr{E}_n$, $[m^n, m^n, m^n]^2_{t_n} \le nc_n''[m^n, m^n, m^n, m^n]_{t_n}$ [Cauchy–Schwarz and (8.16)], (8.22) and (8.23) follows from our assumptions provided $p \ge 12$.

Suppose we can take $p = 12$. We now argue that one can assume (8.11) and (8.16). Let $c_n' = \mu_{2,n}/2$ and $c_n'' = 3\mu_{2,n}/2$:

$$\begin{aligned} &1 - P\left(c_n'' \ge \frac{[m^n, m^n]_{t_n}}{n} \ge c_n' \right) \\ &\le (c_n')^{-2} E\left[\frac{[m^n, m^n]_{t_n}}{n} - \mu_{2,n} \right]^2 I\left\{ \left| \frac{[m^n, m^n]_{t_n}}{n} - \mu_{2,n} \right| > c_n' \right\} \\ &= o(n^{-1}). \end{aligned} \tag{8.24}$$

Let $\mathscr{E}_n^{(1)}$ be the set where (8.16) holds and where $\sup_t |\Delta m_t^n| \le c_n''' n^{1/3}$. In view of (5.6), c_n''' can be chosen such that it is nonrandom and $o(1)$ and so that $P(\mathscr{E}_n^{(1)}) = 1 - o(n^{-1})$. This is because

$$P\Big(\sup_t |\Delta m_t^n| \ge c_n''' n^{1/3}\Big) \tag{8.25}$$
$$\le (c_n''')^{-12} E\left(\frac{[m^n, m^n, m^n, m^n]_{t_n}}{n}\right)^3 I\Big\{\sup_t |\Delta m_t^n| \ge c_n''' n^{1/3}\Big\}.$$

Clearly, it follows that

$$P\big(|\hat{\mu}_n| \le n^{-1/3} \text{ and } \mathscr{E}_n^{(1)}\big) = 1 - o(n^{-1}), \tag{8.26}$$

since, for example,

$$\begin{aligned} &P\big(\dot{l}(n^{-1/3}) > 0 \text{ and } \mathscr{E}_n^{(1)}\big) \\ &\quad \le P\big(m_{t_n}^n - n^{-1/3}[m^n, m^n]_{t_n}(1 - f(c_n''')) > 0 \text{ and } \mathscr{E}_n^{(1)}\big) \\ &\quad \le n^{-2}\big[c_n'(1 - f(c_n'''))\big]^{-12} E\left(\frac{m_{t_n}^n}{\sqrt{n}}\right)^{12} \\ &\quad = o(n^{-1}). \end{aligned}$$

Hence the set $\mathscr{E}_n$ exists.

It remains to show that one can take $p = 12$ in (8.18). This can be assumed without loss of generality by stopping m_t^n once $[m^n, m^n]_t$ exceeds nc_n''. If this stopping time is called τ_n, then $P(\tau_n \ne t_n) = o(n^{-1})$ by (8.24), and $t_n \wedge \tau_n$ can clearly replace t_n in our entire argument. On the other hand,

$$\begin{aligned} E\left(\frac{[m^n, m^n]_{t_n \wedge \tau_n}}{n}\right)^6 &\le 2(c_n'')^6 + 2E\sup_t |\Delta m_t^n|^{12}/n^6 \\ &= O(1) \end{aligned}$$

by (5.6). Hence (8.18) can be assumed (with $t_n \wedge \tau_n$ replacing t_n) in view of Burkholder's inequality. □

PROOF OF THEOREM 4. Define the process $(Z_s)_{0 \le s \le t}$ through

$$Z_{s \wedge \tau} = \frac{dQ}{dP}\bigg|_{\mathscr{F}_{s \wedge \tau}}. \tag{8.27}$$

(Z_s) is clearly a local martingale under P and we shall use the cadlag version of the process. Since P is extremal, it follows from Theorem 11.2 in Jacod [(1979), page 338] that

$$Z_s = 1 + \int_0^s f_s^T \, dm_s, \tag{8.28}$$

where (f_s) is a predictable p-dimensional process.

On the other hand, $(m_{s\wedge\tau} - \langle m, m\rangle_{s\wedge\tau}\mu)Z_s$ is also a P-local martingale. Using Itô's formula [see, e.g., Theorem I.4.57 of Jacod and Shiryaev (1987), page 57] on this process yields (for $s \le \tau$)

$$\begin{aligned} &d(m_s - \langle m, m\rangle_s \mu)Z_s \\ &\qquad = d(\text{local martingale}) + d\langle Z, m\rangle_s - Z_{s-}d\langle m, m\rangle_s \mu. \end{aligned} \tag{8.29}$$

By Proposition I.4.50 of Jacod and Shiryaev [(1987), page 53],

$$\langle Z, m\rangle_s = \int_0^s Z_{u-}d\langle m, m\rangle_u \mu, \tag{8.30}$$

which combined with the above representation for (Z_s) gives

$$\int_0^{s\wedge\tau} d\langle m, m\rangle_u(\mu Z_{u-} - f_u) = 0. \tag{8.31}$$

Hence, for $s \le \tau$,

$$Z_s = 1 + \int_0^s Z_{u-}d(\mu^T m_s). \tag{8.32}$$

The result now follows from Theorem 1 of Doléans-Dade [(1970), page 183] (cf. also Chapter I.4f of Jacod and Shiryaev [(1987), pages 58–61]. □

Acknowledgments. The author wishes to thank Tom DiCiccio, Peter McCullagh, Art Owen, Wing Wong and two referees for useful discussions and comments, and Mitzi Nakatsuka for typing the manuscript.

REFERENCES

AALEN, O. (1976). Nonparametric inference in connection with multiple decrement models. *Scand. J. Statist.* **3** 15–27.

AALEN, O. (1977). Weak convergence of stochastic integrals related to counting processes. *Z. Wahrsch. Verw. Gebiete* **38** 261–278.

AALEN, O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6** 701–726.

AALEN, O. (1980). *A Model for Nonparametric Regression Analysis of Counting Processes. Lecture Notes in Statist.* **2** 1–25. Springer, New York.

AALEN, O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine* **8** 907–925.

ALDOUS, D. (1978). Stopping times and tightness. *Ann. Probab.* **6** 335–340.

ALDOUS, D. (1989). Stopping times and tightness. II. *Ann. Probab.* **17** 586–595.

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

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

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1984). Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *J. Roy. Statist. Soc. Ser. B* **46** 483–495.

BARTLETT, M. S. (1953a). Approximate confidence intervals. *Biometrika* **40** 12–19.

BARTLETT, M. S. (1953b). Approximate confidence intervals. II. More than one unknown parameter. *Biometrika* **40** 306–317.

BECKER, N. G. and HEYDE, C. C. (1990). Estimating population size from multiple recapture experiments. *Stochastic Process. Appl.* **36** 77–83.

BORGAN, Ø. and LIESTØL, K. (1990). A note on confidence intervals and bands for the survival function based on transformations. *Scand. J. Statist.* **17** 35–41.

CHAN, K. S. (1990). Deterministic stability, stochastic stability, and ergodicity. In *Non-linear Time Series. A Dynamical System Approach* (H. Tong, ed.) 448–466. Oxford Univ. Press.

CHAN, N. H. and WEI, C. Z. (1987). Asymptotic inference for nearly nonstationary AR(1) processes. *Ann. Statist.* **15** 1050–1063.

CHAN, N. H. and WEI, C. Z. (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *Ann. Statist.* **16** 367–401.

COX, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 187–220.

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276.

DICICCIO, T. J., HALL, P. and ROMANO, J. P. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist.* **19** 1053–1061.

DICICCIO, T. J. and ROMANO, J. P. (1989). On adjustments based on the signed root of the empirical likelihood ratio statistic. *Biometrika* **76** 447–456.

DOLÉANS-DADE, C. (1970). Quelques applications de la formule de changement de variable pour les semimartingales. *Z. Wahrsch. Verw. Gebiete* **16** 181–194.

EFRON, B. (1988). Logistic regression, survival analysis, and the Kaplan–Meier curve. *J. Amer. Statist. Assoc.* **83** 414–425.

ELLIOT, R. J. (1982). *Stochastic Calculus and Applications*. Springer, New York.

FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

GILL, R. D. and JOHANSEN, S. (1990). A survey of product-integration with a view toward application in survival analysis. *Ann. Statist.* **18** 1501–1555.

GODAMBE, V. P. and HEYDE, C. C. (1987). Quasi-likelihood and optimal estimation. *Internat. Statist. Rev.* **55** 231–244.

GREENWOOD, P. E. and WEFELMEYER, W. (1990). Efficiency of estimators for partially specified filtered models. *Stochastic Process. Appl.* **36** 353–370.

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

HELLAND, I. S. (1982). Central limit theorems for martingales with discrete or continuous time. *Scand. J. Statist.* **9** 79–94.

JACOD, J. (1975). Multivariate point processes: Predictable projection, Radon–Nikodym derivatives, representation of martingales. *Z. Wahrsch. Verw. Gebiete* **31** 235–253.

JACOD, J. (1979). *Calcul Stochastique et Problèmes de Martingales. Lecture Notes in Math.* **714**. Springer, Berlin.

JACOD, J. and MÉMIN, J. (1976). Caractéristiques locales et conditions de continuité absolue pour les semi-martingales. *Z. Wahrsch. Verw. Gebiete* **35** 1–37.

JACOD, J. and SHIRYAEV, A. N. (1987). *Limit Theorems for Stochastic Processes*. Springer, Berlin.

JEGANATHAN, P. (1982). A solution of the martingale central limit problem, Part I–II. *Sankhyā Ser. A* **44** 299–340.

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481.

KOLACZYK, E. (1994). Empirical likelihood for generalized linear models. *Statist. Sinica* **4** 199–218.

LATTA, R. (1981). A Monte Carlo study of some two-sample rank tests with censored data. *J. Amer. Statist. Assoc.* **76** 713–719.

LAWLEY, D. N. (1956). A general method for approximating the distribution of likelihood ratio criteria. *Biometrika* **43** 295–303.

LENGLART, E. (1977). Transformation des martingales locales par changement absolument continu de probabilites. *Z. Wahrsch. Verw. Gebiete* **39** 65–70.

MCCULLAGH, P. (1987). *Tensor Methods in Statistics*. Chapman and Hall, London.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

McLeish, D. L. and Small, C. G. (1992). A projected likelihood function for semiparametric models. *Biometrika* **79** 93–102.

Meier, P. (1976). Estimation of a distribution function from incomplete observations. In *Perspectives in Probability and Statistics* (J. Gani, ed.) 67–88. Academic Press, New York.

Mykland, P. A. (1992). Asymptotic expansions and bootstrapping-distributions for dependent variables: A martingale approach. *Ann. Statist.* **20** 623–654.

Mykland, P. A. (1994). Bartlett type identities for martingales. *Ann. Statist.* **22** 21–38.

Mykland, P. A. (1995). Embedding and asymptotic expansions for martingales. *Probab. Theory Related Fields*. To appear.

Mykland, P. A. and Ye, J. (1992). Cumulants and Bartlett identities in Cox regression. Technical Report 332, Dept. Statistics, Univ. Chicago.

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology* **1** 27–52.

Owen, A. B. (1988a). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249.

Owen, A. B. (1988b). Computing empirical likelihoods. In *Computer Science and Statistics: Proceedings of the 20th Symposium on the Interface* (E. J. Wegman, ed.) 442–447. Amer. Statist. Assoc., Alexandria, VA.

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18** 90–120.

Priestley, M. B. (1988). *Non-linear and Non-stationary Time Series Analysis*. Academic Press, New York.

Qin, J. and Lawless, G. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22** 300–325.

Rebolledo, R. (1980). Central limit theorems for martingales. *Z. Wahrsch. Verw. Gebiete* **51** 269–286.

Reid, N. (1988). Saddlepoint methods and statistical inference. *Statist. Sci.* **3** 213–227.

Rida, W. N. (1991). Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic model. *J. Roy. Statist. Soc. Ser. B* **53** 269–283.

Royall, R. M. (1986). Model robust confidence intervals using maximum likelihood estimators. *Internat. Statist. Rev.* **54** 221–226.

Thomas, D. R. and Grunkemeier, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *J. Amer. Statist. Assoc.* **70** 865–871.

Tjøstheim, D. (1990). Non-linear time series and Markov chains. *Adv. in Appl. Probab.* **22** 587–611.

Tong, H. (1990). *Non-linear Time Series. A Dynamical System Approach*. Oxford Univ. Press.

von Weizsäcker, H. and Winkler, G. (1990). *Stochastic Integrals. An Introduction*. Vieweg, Braunschweig.

Wong, W. H. (1986). Theory of partial likelihood. *Ann. Statist.* **14** 88–123.

Department of Statistics
University of Chicago
5734 University Avenue
Chicago, Illinois 60637