Statistica Sinica **15**(2005), 1003-1014

MAXIMUM SMOOTHED LIKELIHOOD ESTIMATION

E. L. Ionides

University of Michigan

Abstract: Looking myopically at the larger features of the likelihood function, absent some fine detail, can theoretically improve maximum likelihood estimation. Such estimators are, in fact, used routinely, since numerical techniques for maximizing a computationally expensive likelihood function or for maximizing a Monte Carlo approximation to a likelihood function may be unable to investigate small scale behavior of the likelihood. A class of maximum smoothed likelihood estimators is introduced and shown to be asymptotically efficient for models possessing local asymptotic normality. This theoretical result corresponds to good finite sample properties in two examples, with a likelihood that is smooth but multimodal, and a likelihood that is not smooth.

Key words and phrases: Likelihood, local asymptotic normality, maximum likelihood estimation, maximum smoothed likelihood estimation, smoothing.

1. Introduction

Estimating parameters by seeking the maximum of a smoothed likelihood function was first proposed by Daniels (1960). Kernel smoothing of the likelihood using a scale parameter decreasing to zero more rapidly than $n^{-1/2}$ was suggested as a way to find asymptotically efficient estimators under conditions weaker than the classical result for maximum likelihood estimation of Cramér (1946). Barnett (1966) investigated finite sample properties of Daniels' estimator and discovered through simulation that, when estimating the location parameter of a Cauchy distribution from a small sample, a maximum smoothed likelihood estimator (MSLE) could be 10% more efficient than the maximum likelihood estimator (MLE). The motivation for maximizing a smoothed likelihood function in Daniels (1960) is that the MLE can pay too much attention to small-scale features of the likelihood. Smoothing has also been recognized as a tool for general stochastic optimization problems (Kreimer and Rubinstein (1988)). Small, Wang and Yang (2000) discuss smoothing as a method to resolve multiple roots of estimating equations.

An alternative motivation arises when the likelihood function can only be approximated by Monte Carlo methods. If independent Monte Carlo estimates of the likelihood are available at a range of parameter values a natural approach,

investigated by Diggle and Gratton (1984), is to fit a smooth curve approximating the likelihood. Diggle and Gratton (1984) were motivated by inference for spatial point process patterns. Similar issues arise for sequential Monte Carlo methods (Doucet, de Freitas and Gordon (2001)). In this case, Hürzeler and Künsch (2001) show that one may do better by constructing dependent Monte Carlo estimates of the likelihood over a range of parameter values. The procedure of Hürzeler and Künsch (2001) is an adaptation to sequential Monte Carlo of the trick of fixing the seed of the Monte Carlo random number generator while comparing different parameter values. Independent sequential Monte Carlo likelihood estimates are still attractive for their reduced algorithmic complexity.

A computational issue motivating consideration of a smooth approximation to the likelihood arises when the cost of evaluating the likelihood function is high, and one may wish to approximate the likelihood from relatively few function evaluations.

A final motivation for maximum smoothed likelihood (discussed in more detail in Section 3) is that the MSLE methods introduced in this paper bear a close resemblance to practical procedures carried out under the name of MLE. Those who maximize a smoothed likelihood out of practical necessity might claim to be finding an approximation to the MLE, when in fact MSLE has its own theoretical appeal.

There are many ways in which one might smooth the likelihood. A wide class of smoothers is introduced in Section 2. Section 4 gives some concrete constructions and, in particular, shows that certain implementations of locally weighted regression (Cleveland (1979)) fall in this class. Section 5 demonstrates MSLE in a situation where the likelihood is smooth but multimodal — estimating the location of spectral peaks. Section 6 gives some relevant asymptotic results. Section 7 presents a second example, motivated by the asymptotic results for non-smooth likelihoods, where MSLE is 15% more efficient than MLE. Section 8 is a concluding discussion.

2. A Class of Maximum Smoothed Likelihood Estimators

Definition 1. A *smoother*, S, takes real-valued functions on an arbitrary finite subset $G \subset \mathbb{R}^d$ to functions on $\mathbb{R}^d$: for each $g : G \to \mathbb{R}$, we have $S(g) : \mathbb{R}^d \to \mathbb{R}$.

Since the likelihood and log likelihood are real valued functions on $\mathbb{R}^d$, it might appear natural that an appropriate smoother should map the space of such functions into itself. However, Definition 1 suggests that the (log) likelihood should first be sampled on a finite grid. The discrete sampling is required for technical reasons, but is not a major limitation since numerical implementations will necessarily be of this kind. We also require that S satisfy the following quadratic approximation property.

(S1). Let q be a second degree polynomial on $\mathbb{R}^d$, and suppose G specifies a quadratic, as in Definition 2 below. There is a positive constants C, depending on G but not on q, such that for any $g : G \to \mathbb{R}$,

$$|S(g)(t) - q(t)| < C(1 + t^T t) \times \max_{t^* \in G} |g(t^*) - q(t^*)|.$$

Definition 2. We say that G *specifies* a quadratic if there exists a subset G' of G with $(d+1)(d+2)/2$ elements such that G' has no more than $(r+1)(r+2)/2$ elements lying on a linear subspace of dimension r for $r = 1, \ldots, d$.

This means that a function g' on G' may be uniquely interpolated by a polynomial of degree ≤ 2.

The condition (S1) formalizes a requirement that if g can be well approximated by a second degree polynomial q, on G, then the smoother produces a function $S(g)$ close to q. In Section 4 we develop ways of constructing and modifying smoothers to satisfy (S1).

Now suppose we have a sequence of statistical experiments corresponding to families of measures $P_{\theta,n}$ with $\theta \in \Theta \subset \mathbb{R}^d$ and $n = 1, 2, \ldots$, giving rise to the log likelihood ratio

$$\Lambda_n(\theta, \phi) = \log\Big(\frac{dP_{\theta,n}}{dP_{\phi,n}}\Big).$$

Formally, we set $\Lambda_n(\theta, \phi) = 0$ if $dP_{\theta,n}/dP_{\phi,n}$ is zero or undefined. We will apply the smoother to a rescaled log likelihood function evaluated on a set of points which, with high probability, lie in a neighborhood of the true parameter value θ_0. Specifically, let $\tilde{\theta}_n$ be a sequence of preliminary estimators taking values on a grid $n^{-1/2}(z_1\mathbb{Z} \times \ldots \times z_d\mathbb{Z})$ for some constants $z_1, \ldots, z_d > 0$. A discretization is required for technical reasons in the proof of Theorem 1, and will rarely be an issue in practice. It prevents $\tilde{\theta}_n$ from selecting an unusual feature of the likelihood, such as the maximum. For the asymptotic analysis we require that $\tilde{\theta}_n$ is $n^{1/2}$-consistent. With G a finite subset of $\mathbb{R}^d$, as before, we define the rescaled log likelihood, $\lambda_n : G \to \mathbb{R}$, by

$$\lambda_n(t) = \Lambda_n(\tilde{\theta}_n + tn^{-\frac{1}{2}}, \tilde{\theta}_n). \tag{1}$$

An MSLE corresponding to the smoother S is then

$$\hat{\theta}_n = \tilde{\theta}_n + n^{-\frac{1}{2}} \times \arg\max_t S(\lambda_n)(t). \tag{2}$$

The maximum need not be unique. The estimate is undefined when a maximum does not exist. For the remainder of this article we will mean by MSLE an estimator of the form (2). Note that $\tilde{\theta}_n$ plays two roles in (1), defining the origin of the local coordinates, and specifying the measure with respect to which

the likelihood is calculated. In the situation where each measure $P_{\theta,n}$ has a positive density with respect to a measure μ_n, typically Lebesgue or counting measure, this second role becomes unimportant. We can then replace (1) by $\lambda_n(t) = \log(dP_{\tilde{\theta}_n + tn^{-1/2}}/d\mu_n)$ without changing the estimator $\hat{\theta}_n$. In the case of an independent sample this becomes $\lambda_n(t) = \sum_{i=1}^{n} \log f(x_i \mid \tilde{\theta}_n + tn^{-1/2})$, where $x_1, \ldots, x_n$ are drawn from a density $f(x|\theta)$.

3. MSLE Corresponds to Accepted Statistical Practice

A practical procedure for likelihood-based parameter estimation from a complicated likelihood function might include the following steps.

(P1) Take several starting values, θ_k, $k = 1, \ldots, K$. Hopefully knowledge of the particular application will suggest some reasonable values of θ_k. These might also come from some more readily available estimator, such as maximum pseudo-likelihood or the method of moments.

(P2) For each θ_k, run a numerical optimization procedure starting at θ_k to attempt to find the maximum of the likelihood function. Hopefully this algorithm will terminate under a reasonable convergence criterion to give an estimate $\hat{\theta}_k$.

(P3) If all the $\hat{\theta}_k$ are close, use their common value $\hat{\theta}$ for an estimate of $\hat{\theta}_0$. An estimate of the error on $\hat{\theta}$ can come from numerical calculation of the second derivative of the likelihood function at $\hat{\theta}$, using asymptotic properties of the likelihood function.

(P4) If the values of $\hat{\theta}_k$ for $1 \leq k \leq K$ vary considerably, try to use knowledge of the subject matter, the form of the likelihood function and the numerical algorithm used to understand why. Possibly one or more of the estimates may be rejected as unreasonable.

(P5) In the event of either (P3) or (P4), plot the region of interest of the likelihood function, or try to find some graphical representation such as marginal plots when the parameter space is of too high dimension to allow a standard plot.

The MSLE is similar to the method carried out in (P1)−(P5), although a statistician following these steps might well claim to be calculating an MLE. In particular, plotting the likelihood function in a region around the estimated value gives an approximation to the likelihood function based on evaluations on a grid of points, as used for the MSLE. Thinking of the procedure as MSLE instead of MLE has some consequences. One is encouraged to spend most of the computational effort investigating features of the likelihood on the same scale

$(n^{-1/2})$ as the error of the estimation, and the importance of the initial values is made explicit.

4. Constructing an MSLE

First we show that a least squares quadratic approximation results in a smoother satisfying (S1). This result is used to investigate the use of widely used smoothing methods, such as those described by Hastie and Tibshirani (1990).

Lemma 1. *Let $q : \mathbb{R}^d \to \mathbb{R}$ be a quadratic polynomial, $q(t) = \sum_{i\le j} a_{ij}t_it_j + \sum_i b_it_i + c$, and write $\alpha = (a_{11}, \ldots, a_{dd}, b_1, \ldots, b_d, c)^T$. Define the least squares quadratic smoother, S_{LS}, to map $g : G \to \mathbb{R}$ to the quadratic $\tilde{q}$ with coefficients $\tilde{\alpha} = (\tilde{a}_{11}, \ldots, \tilde{c})^T$ minimizing $\sum_{t^*\in G}(\tilde{q}(t^*) - g(t^*))^2$. The smoother S_{LS} satisfies* (S1).

Proof. Write $\tilde{\gamma}$ for a vector in $\mathbb{R}^{|G|}$ listing $\{g(t^*) : t^* \in G\}$, and let γ be the equivalent vector of $\{q(t^*) : t^* \in G\}$. Since G specifies a quadratic, there is a unique least squares quadratic fit to g whose coefficients satisfy a linear equation, say $\tilde{\alpha} = H\tilde{\gamma}$, with H depending only on G. Since q is the least squares quadratic fit to itself, we have $\alpha = H\gamma$ and so $\tilde{\alpha} - \alpha = H(\tilde{\gamma} - \gamma)$. This gives a bound on the coefficients of the quadratic $(S_{LS}(g) - q)$ which is linear in $\max_k(|\tilde{\gamma}_k - \gamma_k|)$, *i.e.*, one can find a vector β with $|\tilde{\alpha}_i - \alpha_i| < \beta_i \max_{t^*\in G} |g(t^*) - q(t^*)|$. One can then choose C to make $C(1 + t^Tt)$ greater than the quadratic with coefficient vector β, demonstrating that (S1) is satisfied.

Locally weighted polynomial smoothers (Cleveland (1979)) form a widely used class of smoothers. The implementation of these methods employed in the examples of Sections 5 and 7 is the `loess` algorithm written by B. D. Ripley, available in the `modreg` package for the R language. By default, `loess` calculates a local quadratic surface by least squares and so, following Lemma 1, satisfies (S1). Fitting local linear surfaces via robust M-estimation, as originally suggested by Cleveland (1979), is available using `loess` via non-default options.

We now consider how to coerce more general smoothers into satisfying (S1). For a smoother, S, define the modification S' by

$$S'(g) = S_{LS}(g) + S(g - S_{LS}(g)|_G),$$

where $S_{LS}(g)|_G$ is the restriction to G of the least squares quadratic smoother from Lemma 1. S' satisfies (S1) as long as S maps small functions to small functions, *i.e.*, if there is a constant D with $|S(\epsilon)(t)| < D \max_{t^*\in G} |\epsilon(t^*)|$.

The condition (S1) has two main effects: it allows the smoother to extrapolate in an (asymptotically) reasonable way, and it ensures that the smoother does not destroy the (asymptotic) quadratic structure of the log likelihood. The

modified smoother S', which essentially smooths the residuals from a quadratic approximation, provides a simple way to enjoy the flexibility of general smoothing methods without suffering (asymptotic) disadvantages.

5. An Example: Maximum Smoothed Likelihood Estimation of Spectral Peaks

Identifying spectral peaks and estimating relevant parameters is critical for analysis of protein structure by nuclear magnetic resonance spectroscopy (Wüthrich (1995)). It is necessary to attempt identification of peaks with intensities only slightly over the noise level, so efficient methods with appropriate uncertainty estimates are required. The spectra may contain hundreds of peaks and may have up to four frequency dimensions.

We consider a simplified caricature of a peak estimation problem, with just one peak, in one dimension, observed with white noise. Suppose that $2n$ observations are made from a stationary process with spectral density given by

$$f(\nu|\theta) = 1 + \frac{h}{1 + (\frac{\nu-\theta}{w})^2}. \tag{3}$$

Inference may be based on the Whittle model for the periodogram (Shumway and Stoffer (2000, Section 3.7)),

$$I_k = I(\nu_k) = f(\nu_k|\theta)E_k. \tag{4}$$

Here, $\nu_k = k/2n$ for $k = 1, \ldots, n-1$, and the E_k are independent Exponential random variables with density e^{-x} for $x > 0$.

Fixing the height and width of the peak results in a location estimation problem for θ. The MLE is asymptotically well behaved for large n, but in practice the likelihood has many local maxima. A worked example is presented in Figure 1 and the following caption. MSLE was found to deliver a 6% efficiency improvement over MLE, when both correctly identified the peak. MSLE also identified the peak a little more often.

As well as demonstrating one situation where MSLE outperforms MLE, it is worth noting that MSLE was found to be comparable to MLE for a wide variety of parameter values and amounts of smoothing (results not shown). The practical advantages of smoothing the likelihood to simplify maximization in more complex situations may be more compelling than small differences in efficiency.

In this example the likelihood is smoothed globally. The initial estimator does not play a role, and so there is no reasonable implementation of MQLE. However, we have demonstrated here that the more general class of MSLE estimators is nevertheless useful. For larger problems, when exhaustive maximization is not possible, maximization of the likelihood or smoothed likelihood may

require a reasonable initial estimate and/or the use of numerical optimization techniques.

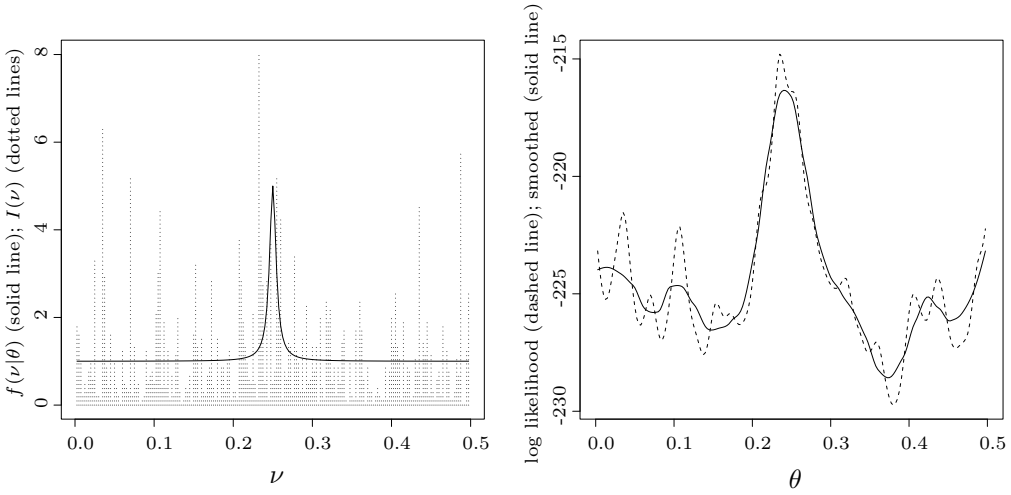

Figure 1. (a) The true spectral density (solid line) for (3) with $\theta_0 = 0.25$, $h = 4$ and $w = 0.05$, and a realization from (4) with $n = 200$ (dotted lines); (b) The log likelihood for this realization (dashed line) and the smoothed likelihood (solid line). Smoothing was carried out by applying `loess` to the log likelihood evaluated at $\theta = k/2n$, $k = 1, \ldots, n-1$. MSLE (using `loess` with smoothing parameter `span` $= 0.1$) was 6% more efficient than MLE, measured from the relative variance of MLE and MSLE conditional on both methods correctly identifying the peak (i.e., when both estimates were within $\theta_0 \pm 5w$). This occurred 92.9% of the time. MSLE was also slightly more likely to identify the peak (93.35% compared to 93.19%, a difference of 0.16%). Results are based on a simulation of size $200,000$.

6. Asymptotic Properties of Maximum Smoothed Likelihood Estimators

Following Le Cam (1986), a family of measures $\{P_{\theta,n}, \theta \in \Theta\}$ with Θ an open subset of $\mathbb{R}^d$ is said to have *local asymptotic normality* (LAN) at θ_0 if there exists a positive definite matrix K and a sequence of random variables $\{\Delta_n\}$ such that, for any bounded sequence $\{t_n\}$ in $\mathbb{R}^d$,

$$\begin{aligned} \Lambda_n\left(\theta_0 + t_n n^{-\frac{1}{2}}, \theta_0\right) = \; & t_n^T \Delta_n - \frac{1}{2} t_n^T K t_n + o_p(1; \theta_0), \\ & \Delta_n \xrightarrow{d} N(0, K) \text{ under } \; P_{\theta_0,n}. \end{aligned} \tag{5}$$

Here, $\xrightarrow{d}$ indicates convergence in distribution, and $\zeta_n = o_p(\alpha_n; \theta)$ means that $\zeta_n/\alpha_n \to 0$ in probability under $P_{\theta,n}$. LAN implies that the probability of

$dP_{\theta,n}/dP_{\theta_0,n}$ being zero or undefined tends to zero. The matrix K is thought of as the asymptotic information rate concerning θ; it coincides with the Fisher information under regularity conditions. Hájek's convolution theorem justifies calling an estimator $\hat{\theta}_n$ of θ asymptotically efficient if LAN holds and $n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, K^{-1})$. More lengthy discussions of LAN can be found in Bickel, Klaassen, Ritov and Wellner (1993) and Van der Vaart (2002). A generalization of LAN that could be similarly used to motivate MSLE is the local asymptotic quadratic condition of Le Cam and Yang (2000, Section 6.2).

To demonstrate the asymptotic efficiency of the MSLE $\hat{\theta}_n$, we show that it is asymptotically equivalent to Le Cam's efficient one step estimator $\bar{\theta}_n$, in the sense that $|\hat{\theta}_n - \bar{\theta}_n| = o_p(n^{-1/2}; \theta_0)$. The one-step estimator of Le Cam and Yang (2000, Section 6.3) is the special case of MSLE which we have called the maximum quadratic likelihood approximation estimator (MQLE). The MQLE may be defined as an MSLE where G is a set of $(d+1)(d+2)/2$ points specifying a quadratic, and S returns the interpolating polynomial of degree ≤ 2.

Theorem 1. *If $\{P_{\theta,n}\}$ has LAN, then any two MSLE estimators, $\hat{\theta}_n$ and $\hat{\theta}'_n$, are asymptotically equivalent in that $|\hat{\theta}_n - \hat{\theta}'_n| = o_p(n^{-1/2}; \theta_0)$.*

Heuristically, Theorem 1 holds because any smoother satisfying (S1) will asymptotically return a smoothed log likelihood close to a quadratic approximation guaranteed by LAN. Lemma 2 gives a formal approximation result.

Lemma 2. *Let $\tilde{\theta}_n$ be the $n^{1/2}$-consistent estimator in* (2)*, and define*

$$q_n(t, u) = t^T(\Delta_n - Ku) - \frac{1}{2}t^T K t, \tag{6}$$

with K and Δ_n identified in (5)*. Then $\lambda_n(t)$ in* (1) *and $q_n(t, u)$ satisfy*

$$\max_{t^* \in G}(\lambda_n(t^*) - q_n(t^*, n^{\frac{1}{2}}(\tilde{\theta}_n - \theta_0))) = o_p(1; \theta_0). \tag{7}$$

Proof of Lemma 2. We can choose M so that $P_{\theta_0}(|\tilde{\theta}_n - \theta_0| \leq Mn^{-1/2}) < \varepsilon$ for a given $\varepsilon > 0$ and sufficiently large n. Now let $\tau_n = n^{1/2}(\tilde{\theta}_n - \theta_0)$ if $n^{1/2}|\tilde{\theta}_n - \theta_0| \leq M$ and $\tau_n = 0$ otherwise, so $\{\tau_n\}$ is a bounded sequence. Define the truncated rescaled log likelihood, $\lambda'_n(t)$, as

$$\begin{aligned}
\lambda'_n(t) &= \Lambda_n(\theta_0 + n^{-\frac{1}{2}}(t + \tau_n), \theta_0 + n^{-\frac{1}{2}}\tau_n) \\
&= \Lambda_n(\theta_0 + n^{-\frac{1}{2}}(t + \tau_n), \theta_0) - \Lambda_n(\theta_0 + n^{-\frac{1}{2}}\tau_n, \theta_0) + \alpha_n \\
&= (t^T + \tau_n^T - \tau_n^T)\Delta_n - \frac{1}{2}(t + \tau_n)^T K(t + \tau_n) + \frac{1}{2}\tau_n^T K \tau_n + \beta_n \\
&= t^T(\Delta_n - K\tau_n) - \frac{1}{2}t^T K t + \beta_n \\
&= q_n(t, \tau_n) + \beta_n.
\end{aligned} \tag{8}$$

The term $\alpha_n = o_p(1;\theta_0)$ allows for a possible lack of absolute continuity. The LAN assumption gives $\beta_n = o_p(1;\theta_0)$ since τ_n takes values on a set of bounded size, namely

$$\tau_n \in \{\theta_0 n^{\frac{1}{2}} + z : z \in \ (z_1\mathbb{Z} \times \cdots \times z_d\mathbb{Z}) \text{ with } |\theta_0 n^{\frac{1}{2}} + z| \leq M\} \cup \{0\}.$$

From (8), recalling the construction of $\{\tau_n\}$, we see that $\lambda_n(t) - q_n(t, n^{1/2}(\tilde{\theta}_n - \theta_0)) = o_p(1;\theta_0)$. Finally, to prove the Lemma, notice that G is a finite set.

Proof of Theorem 1. The maximum of $q_n(t, n^{1/2}(\tilde{\theta}_n - \theta_0))$, defined in (6), occurs at $t = K^{-1}\Delta_n - n^{1/2}(\tilde{\theta}_n - \theta_0)$. Now (7) together with (S1) gives

$$\arg\max_t \ S(\lambda_n(t)) = K^{-1}\Delta_n - n^{1/2}(\tilde{\theta}_n - \theta_0) + o_p(1;\theta_0).$$

We can now write a bound for the MSLE,

$$\begin{aligned}\hat{\theta}_n &= \tilde{\theta}_n + n^{-\frac{1}{2}}\left(K^{-1}\Delta_n - n^{\frac{1}{2}}(\tilde{\theta}_n - \theta_0) + o_p(1;\theta_0)\right)\\ &= \theta_0 + n^{-\frac{1}{2}}K^{-1}\Delta_n + o_p(n^{-\frac{1}{2}};\theta_0). \qquad (9)\end{aligned}$$

The result (9) does not depend on the form of the MSLE used, proving the theorem.

Theorem 1 gives the asymptotic distribution of the MSLE when K is known. If $S(\lambda_n(t))$ is sufficiently smooth that a condition analogous to (S1) holds for the matrix of second derivatives,

$$\nabla^2 S(g)(t) = \Big[\frac{\partial^2}{\partial t_i \partial t_j} S(g)(t)\Big],$$

then K may be estimated by $\hat{K}_n = -\nabla^2 S(\lambda_n)(n^{1/2}(\hat{\theta}_n - \tilde{\theta}_n))$. Confidence intervals and hypothesis tests may then be constructed following the methods described by Le Cam and Yang (2000, Section 6.8) for MQLE.

7. MSLE for Non-Smooth Likelihoods

The framework of LAN has been convenient to study asymptotic properties of maximum smoothed likelihood estimators. To compare MSLE to MLE conceptually, it is helpful to compare LAN to the property that the maximum likelihood estimator is consistent, asymptotically efficient and asymptotically normal, which we call MLCEN. Widely used Cramér (1946) type conditions ensuring MLCEN for independent observations imply LAN (Le Cam and Yang (2000, Section 7.2)).

Non-smooth likelihoods, sufficiently pathological for Cramér type conditions to fail, provide an opportunity for MSLE to out-perform MLE. Consider the shift family with densities on $\mathbb{R}$ given by

$$f(x|\theta) \propto \exp(-|x-\theta|^{\alpha}). \qquad (10)$$

For $\alpha > 1/2$, this family possesses LAN, though for $\alpha \leq 1$ it does not satisfy Cramér type conditions (Le Cam and Yang (2000, Section 7.2)).

Table 1. The relative efficiency of MSLE, MQLE and the median compared to MLE for model (10) with $\alpha = 0.6$, $\theta = 0$, and a sample of size $n = 100$. This was estimated by simulation, as the ratio of the variance of the MLE to that of the estimator. The MSLE was calculated by evaluating the log likelihood at 12 points, equally spaced on an interval of width 1.6, centered on the median. The `loess` smoother, as programmed in R with default parameters, was then applied. The MQLE was calculated from the likelihood evaluated at the median and points distant 0.4 each side. The quadratic was maximized over this interval of width 0.8, to handle rare cases when the estimator would otherwise be badly behaved.

	median	MQLE	MSLE
Efficiency	1.01	1.15	1.15

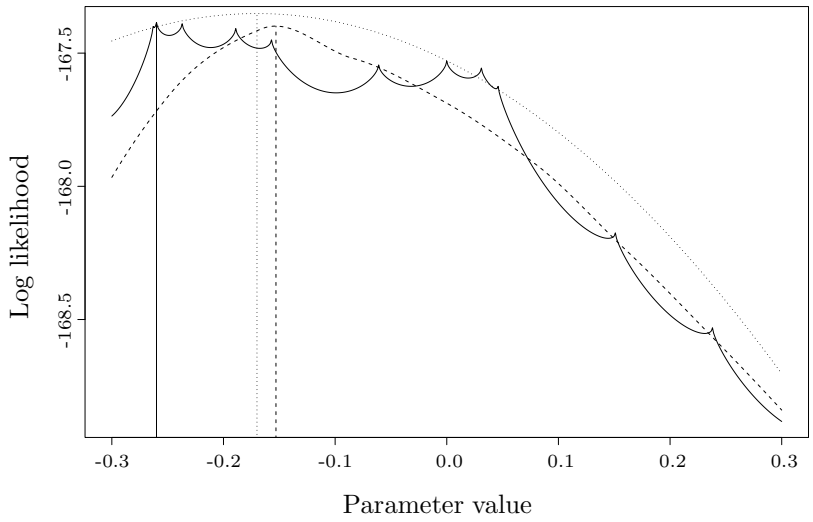

Figure 2. The log likelihood and MLE (solid lines) compared with the smoothed likelihood and MSLE (dashed lines) and the quadratic likelihood approximation and MQLE (dotted lines) for one realization from (10) with $\alpha = 0.6$, $\theta = 0$, and a sample of size 100.

The MLE, MSLE, median and MQLE are compared in Table 1. In this case, the MSLE and MQLE share a 15% increase in efficiency over the MLE. A sample likelihood function is presented in Figure 2. The example illustrates the difference between MQLE and a single iteration of Newton-Raphson maximization: the

quadratic approximation does not evaluate the second derivative of the likelihood, but looks for a larger scale approximation. This suggests one answer to the question "why not repeat the quadratic approximation until convergence?" The MSLE was usually closer to the MLE than was the MQLE. One could argue that the success of MQLE on this example may be due to symmetry. MQLE imposes a symmetric likelihood approximation which in this particular case is appropriate.

8. Discussion

Many theoretical developments have taken place in the framework of LAN, and in related situations such as local asymptotic mixed normality (Van der Vaart (2002)). To take advantage of these advances requires estimators based on LAN. The family of maximum smoothed likelihood estimators introduced in this paper fills much of the gap between the rather crude one-step estimator and the widely accepted MLE, helping to resolve theoretical and practical difficulties that may arise with likelihood based estimation methods. Even when Cramér type conditions for the MLE hold, the weaker LAN condition may be considerably easier to check — compare, for example, Bickel and Ritov (1996) with Bickel, Ritov and Ryden (1998).

This paper has taken a frequentist viewpoint, but has some connections to Bayesian methods. MLE may be compared to maximum a posteriori estimation. A posterior distribution, in the same way as a likelihood function, may have small scale features which raises similar issues to those discussed in this paper. Another relationship is that the posterior mean and MSLE both provide an averaging over the likelihood function. Choosing a smoother is then loosely analogous to choosing a prior distribution.

The asymptotic properties of MSLE suggest explanations for observed finite sample properties. Asymptotically, the smoother used is relatively unimportant as long as one smooths on a scale of $n^{-1/2}$ (which is more smoothing than previously suggested in the literature). Cross-validation techniques, or simulation studies, can be used to compare different smoothers for particular applications. Asymptotically, MSLE can work in situations where MLE fails. Our examples suggest that, when the likelihood is multimodal or is not smooth, this can translate to an increased finite sample efficiency for MSLE over MLE.

Acknowledgement

The author thanks an associate editor and an anonymous referee for their careful reading of the paper and their constructive suggestions.

References

Barnett, V. D. (1966). Evaluation of the maximum likelihood estimator where the likelihood equation has multiple roots. *Biometrika* **53**, 151-165.

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

Bickel, P. J. and Ritov, Y. (1996). Inference in hidden Markov models, I: local asymptotic normality in the stationary case. *Bernoulli* **2**, 199-228.

Bickel, P. J., Ritov, Y. and Ryden, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann. Statist.* **26**, 1614-1635.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74**, 829-836.

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

Daniels, H. E. (1960). The asymptotic efficiency of a maximum likelihood estimator. In *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1**, 151-163. University of California Press, Berkeley.

Diggle, P. J. and Gratton, R. J. (1984). Monte Carlo methods of inference for implicit statistical models. *J. Roy. Statist. Soc.* **46**, 193-227.

Doucet, A., de Freitas, N. and Gordon, N. J. (2001). *Sequential Monte Carlo Methods in Practice*. Springer, New York.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hürzeler, M. and Künsch, H. R. (2001). Approximating and maximising the likelihood for a general state-space model. In *Sequential Monte Carlo Methods in Practice* (Edited by A. Doucet, N. de Freitas and N. J. Gordon), 159-175. Springer, New York.

Kreimer, J. and Rubinstein, R. Y. (1988). Smoothed functionals and constrained stochastic approximation. *SIAM J. Num. Anal.* **25**, 470-487.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

Le Cam, L. and Yang, G. L. (2000). *Asymptotics in Statistics*. 2nd edition. Springer, New York.

Shumway, R. H. and Stoffer, D. S. (2000). *Time Series Analysis and Its Applications*. Springer, New York.

Small, C. G., Wang, J. and Yang, Z. (2000). Eliminating multiple root problems in estimation. *Statist. Science* **15**, 313-341.

Van der Vaart, A. (2002). The statistical work of Lucien Le Cam. *Ann. Statist.* **30**, 631-682.

Wüthrich, K. (1995). *NMR in Structural Biology*. World Scientific, Singapore.

Department of Statistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

E-mail: ionides@umich.edu

(Received August 2003; accepted October 2004)