Statistics and Computing (2001) **11,** 57–65

Mean likelihood estimators

A. I. MCLEOD* and B. QUENNEVILLE†

*Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada, N6A 5B7
aim@uwo.ca
†Time Series Research and Analysis Centre, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6
quenne@statcan.ca

Received March 1999 and accepted August 1999

The use of *Mathematica* in deriving mean likelihood estimators is discussed. Comparisons are made between the mean likelihood estimator, the maximum likelihood estimator, and the Bayes estimator based on a Jeffrey's noninformative prior. These estimators are compared using the mean-square error criterion and Pitman measure of closeness. In some cases it is possible, using *Mathematica*, to derive exact results for these criteria. Using *Mathematica*, simulation comparisons among the criteria can be made for any model for which we can readily obtain estimators.

In the binomial and exponential distribution cases, these criteria are evaluated exactly. In the first-order moving-average model, analytical comparisons are possible only for $n = 2$. In general, we find that for the binomial distribution and the first-order moving-average time series model the mean likelihood estimator outperforms the maximum likelihood estimator and the Bayes estimator with a Jeffrey's noninformative prior.

Mathematica was used for symbolic and numeric computations as well as for the graphical display of results. A *Mathematica* notebook which provides the *Mathematica* code used in this article is available: http://www.stats.uwo.ca/mcleod/epubs/mele. Our article concludes with our opinions and criticisms of the relative merits of some of the popular computing environments for statistics researchers.

Keywords: binomial distribution, exponential distribution, first-order moving-average time series model, *Mathematica* in education and research, mean square error criterion, Pitman measure of closeness, simulation comparison of estimators, unit root in MA(1) model

1. Introduction

The maximum likelihood estimator (MLE) is perhaps the most common and widely accepted estimator of a parameter in a statistical model denoted by $(\mathcal{S}, \Omega, f)$, where $\mathcal{S}$, Ω, f denote respectively the sample space, the parameter space and the probability density function (pdf). We will take $\mathcal{S} = \mathcal{R}^n, X = (X_1, X_2, \ldots, X_n) \in \mathcal{S}$, and $f(x, \theta)$. In the standard case of independent and identically distributed observations, $f(x, \theta) = \Pi_{i=1}^n f_1(x_i)$, where $f_1(x)$ is the pdf of X_1. Given data X, the likelihood function is $L(\dot{\theta}) = f(X; \dot{\theta}), \dot{\theta} \in \Omega$ and the MLE of the parameter θ is defined as that value $\dot{\theta}$ which globally maximizes $L(\dot{\theta})$. *Mathematica* (Wolfram 1996) has been widely used in the study of fundamental and general aspects of maximum likelihood estimation – see Andrews and Stafford (1993), Stafford and Andrews (1993) and Stafford, Andrews and Wang (1994). As well *Mathematica* has been used for obtaining symbolically exact maximum likelihood estimators in situations where the use of numerical techniques are less convenient such as with grouped or censored data or logistic regression – see Cabrera (1989) and Currie (1995).

For simplicity we will deal with the case where Ω is one-dimensional. The multidimensional case may in general be reduced to the one-dimensional case by using marginal, conditional or concentrated likelihoods or by integrating over the nuisance parameters whichever is more suitable in a particular situation. Under the usual regularity conditions, the MLE, $\hat{\theta}$, is approximately normally distributed with mean θ and covariance matrix I_θ^{-1}, where I_θ denotes the Fisher information matrix. It is also true that the mean likelihood estimator (MELE) is equally efficient in large samples. In general the MELE $\bar{\theta}$ is defined by

$$\bar{\theta} = \frac{\int_\Omega \dot{\theta} L(\dot{\theta})\, d\dot{\theta}}{\int_\Omega L(\dot{\theta})\, d\dot{\theta}},$$

where $L(\dot{\theta})$ is the likelihood function. It should be noted that although the MELE is identical to the Bayes estimator with a uniform prior, it is not often considered in frequentist settings.

Pitman (1938) showed that when the problem is location invariant, the MELE is the best invariant estimator. Barnard, Jenkins and Winsten (1962) recommended the MELE for time series problems and suggested that it will often have lower MSE than the MLE. Another application of the MELE is to changepoint analysis where the usual regularity conditions for the MLE do not hold. In this situation, the MLE is actually statistically inefficient, even in large samples, however the MELE works well (Ritov 1990, Rubin and Song 1995).

Unlike the MLE the MELE is not invariant under reparameterization. Although the MELE has a Bayesian interpretation, it is not the Bayesian estimator that is usually recommended. In order that the Bayesian estimator share the MLE property of being invariant under parameter transformation, the Jeffrey's noninformative prior is recommended when there is no prior information available (Box and Tiao 1973, §1.3). The Jeffrey's prior is given by $p(\theta) \propto \sqrt{I_\theta}$.

There are situations, such as in the first-order moving-average model, MA(1), where the MLE in finite samples has non-zero probability of lying on the boundary of the parameter region but this phenomenon does not happen with the MELE as can be seen from the following result.

Theorem 1. *Let* $\Omega = [a, b]$ *then* $\Pr\{\bar{\theta} \in (a, b)\} = 1$.

Proof: The likelihood function, $L(\dot{\theta})$, defined below, is easily seen to be continuous and differentiable in the interval $[a, b]$ and non-negative. It then follows from the generalized mean-value theorem (Borowski and Borwein 1991, p. 371) that $\bar{\theta} \in (a, b)$. □

Under suitable restrictions on the prior distribution, Theorem 1 can be extended to Bayesian estimators.

In many cases the MLE is easy to compute using pen and paper. However with *Mathematica* we can now easily obtain the MELE by numerical integration and sometimes symbolically. In fact, for problems where the likelihood function is complicated or difficult to evaluate the MELE may be computationally easier to compute than the traditional MLE. As shown in Theorem 2, both the MLE and MELE are first order efficient.

Theorem 2. *Under the usual regularity conditions for maximum likelihood estimators,* $\bar{\theta} = \hat{\theta} + O_p(1/n)$.

Proof: The likelihood function, $L(\dot{\theta})$, is to $O_p(1/n)$ equal to the normal probability density function with mean θ and variance I_θ^{-1} (Tanner 1993, p. 16). The result then follows directly from this approximation. □

Now consider an estimator $\hat{\theta}_1$ of θ. The mean-square error (MSE) of an estimator $\hat{\theta}_1$ is defined as $\sigma^2(\hat{\theta}_1 \mid \theta) = E\{(\hat{\theta}_1 - \theta)^2\}$. The relative efficiency of $\hat{\theta}_1$ vs $\hat{\theta}$ is defined as $R(\hat{\theta}_1, \hat{\theta} \mid \theta) = \sigma^2(\hat{\theta} \mid \theta)/\sigma^2(\hat{\theta}_1 \mid \theta)$. Clearly, from Theorem 2, as $n \to \infty$, $R(\bar{\theta}, \hat{\theta} \mid \theta) = 1$. Barnard, Jenkins and Winsten (1962) suggested that in many small sample situations the MELE is preferred by the mean-square error criterion and hence at least for some values of θ, $R(\bar{\theta}, \hat{\theta} \mid \theta) > 1$, where $\hat{\theta}$ and $\bar{\theta}$ denote the MLE and MELE respectively.

Pitman (1937) formulated a useful alternative to the MSE in the situation where no explicit loss function is known. Consider two estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, and assume that with probability one, $\hat{\theta}_1 \neq \hat{\theta}_2$ then the Pitman measure of closeness for comparing $\hat{\theta}_1$ vs $\hat{\theta}_2$ is defined as

$$\mathrm{PMC}(\hat{\theta}_1, \hat{\theta}_2 \mid \theta) = \Pr\{|\hat{\theta}_1 - \theta| < |\hat{\theta}_2 - \theta|\}. \tag{1.1}$$

When $\mathrm{PMC}(\hat{\theta}_1, \hat{\theta}_2 \mid \theta) > 1/2$, $\hat{\theta}_1$ is preferred to $\hat{\theta}_2$. The monograph of Keating, Mason and Sen (1993) provides an extensive survey of recent work and applications of the PMC. Additionally, volume 20(11) of *Communications in Statistics: Theory and Methods* contains an entire issue on the PMC.

Unlike the MSE and relative efficiency, the PMC depends on the bivariate distribution of the two estimators. The PMC is more appropriate in many scientific and industrial applications in which the estimator which is closer to the truth is required. Sometimes it is felt that the MSE and other risk criteria give too much weight to large deviations which may seldom occur. Rao and other researchers (Keating, Mason and Sen 1993, §3.3) have found that risk functions such as MSE and mean-absolute-error can often be shrunk but that this shrinkage occurs at the expense of the PMC. The MSE or some other risk function is more appropriate than PMC in the decision theory framework when there is some economic or other loss associated with the estimation error. In practice it is often useful to consider both the PMC and MSE and in many situations there appears to be a high level of concordance between these estimators (Keating, Mason and Sen 1993, §2.5). In complex models, the PMC like the MSE may be evaluated by simulation.

As originally pointed by Pitman (1937) the PMC criterion is intransitive but it is arguable whether this is a practical limitation. This point as well as other limitations and extensions of the PMC are discussed by Keating, Mason and Sen (1993, ch. 3)

Theorem 3. *$\bar{\theta}$ and $\hat{\theta}$ are not necessarily asymptotically equivalent under the PMC.*

Proof: See equation 2.3. □

The next theorem shows that the MELE minimizes the mean likelihood of the squared error.

Theorem 4. *Choosing* $\dot{\theta} = \bar{\theta}$ *minimizes* $\delta(\dot{\theta})$, *where*

$$\delta(\dot{\theta}) = \int_\Omega (\dot{\theta} - \theta)^2 L(\dot{\theta})\, d\dot{\theta}.$$

Proof: Using calculus, the result follows directly. □

Theorem 5. *$\bar{\theta}$ is a function of the sufficient statistic for θ, S, if there is one.*

In general, the MELE is a biased estimator.

Theorem 6. *If Ω has compact support and $0 < Var(\bar{\theta}) < \infty$ then $E\{\bar{\theta}\} \neq \theta$.*

Theorems 5 and 6 are derived in Quenneville (1993). The MELE is formally equivalent to the Bayes estimator under a locally uniform prior with the squared error risk function. Theorems 2–6 have well-known Bayesian analogues.

Frequentist analysis of Bayesian estimators is not often done but Dempster (1998) and Quenneville and Singh (2000) have argued that frequentist considerations are obviously informative even in the Bayesian setting. We are now going to make comparisons between these three estimators each of for three statistical models: Bernouilli trials, exponential lifetimes and MA(1) time series. The symbolic, numeric and graphical computations will all be done using *Mathematica*. The interested reader can reproduce or extend our computations using the *Mathematica* notebooks we have provided (McLeod and Quenneville 1999).

2. Bernoulli trials

We will now examine the performance of these three estimators in the estimation of the parameter p in a sequence of n Bernoulli trials where X is the observed number of successes and p is the probability of success. The probability function is

$$f_x(n, p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

So if X successes are observed in n trials, the likelihood function may be written $L(p) = p^X(1-p)^{(n-X)}$ and the MLE may be derived by calculus, $\hat{p} = X/n$. Using *Mathematica* it is easily shown that the MELE of p is $\bar{p} = (X+1)/(n+2)$ and that $R(\bar{p}, \hat{p} \mid p) > 1$ provided

$$p \in \left(\frac{2n - \sqrt{2n^2+3n+1}+1}{2(2n+1)}, \frac{2n+\sqrt{2n^2+3n+1}+1}{2(2n+1)} \right).$$

As shown in Fig. 1, the MELE is always more efficient over most of the range and the relative efficiency tends to 1 as $n \to \infty$.

It is interesting to compare the MELE with Bayes estimator under a Jeffrey's prior. The Jeffrey's prior for p is (Box and Tiao p. 35), $\pi(p) = 1/\sqrt{p(1-p)}$. Combining with the likelihood we can use *Mathematica* to show that the resulting Bayes estimator is $\tilde{p} = (1+4X)/(2+4n)$. From Fig. 1, we see that the Bayes estimator with Jeffrey's prior tends to have smaller mean-square error over an even slightly larger range of p than the MELE but the gain in efficiency with the MELE can be greater. As with the MELE, the relative efficiency tends to 1 as $n \to \infty$. Once again, using *Mathematica* we can show that $R(\tilde{p}, \hat{p} \mid p) > 1$ provided

$$p \in \left(\frac{1+5n-\sqrt{1+9n+20n^2}}{2(1+5n)}, \frac{1+5n+\sqrt{1+9n+20n^2}}{2(1+5n)} \right).$$

The PMC criterion given in equation (1.1) is not applicable in the case of the binomial since due to the discreteness there

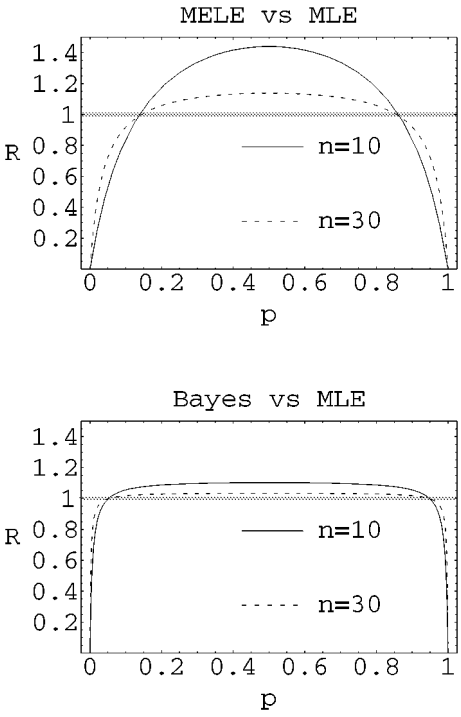

Fig. 1. *Relative efficiency of alternative binomial estimators. Top panel: MELE, relative efficiency, $R(\bar{p}, \hat{p} \mid p)$ for $n = 10, 30$. Bottom panel: Bayes estimator with Jeffrey's prior, relative efficiency, $R(\tilde{p}, \hat{p} \mid p)$ for $n = 10, 30$*

can be ties in the values of the estimators. Keating, Mason and Sen (1993, §3.4.1) and one of the referees have suggested the following modified version of Pitman's measure of closeness,

$$\mathrm{PMC}(\bar{\theta}, \hat{\theta} \mid \theta) = \Pr\{|\bar{\theta} - \theta| < |\hat{\theta} - \theta|\} + \frac{1}{2}\Pr\{|\bar{\theta} - \theta| = |\hat{\theta} - \theta|\}. \quad (2.2)$$

With this modification, PMC is symmetric and reflexive.

Figure 2 suggests the following asymptotic result,

$$\lim_{n\to\infty} \mathrm{PMC}(\bar{p}, \hat{p} \mid p) = \begin{cases} 1 & p = 1/2 \\ \frac{1}{2} & p \neq 1/2, 0, 1 \\ 0 & p = 0, 1 \end{cases} \quad (2.3)$$

In the Appendix, it is shown how, using the Geary-Rao Theorem (Keating, Mason and Sen p. 103), these asymptotic limits may be established. For $\mathrm{PMC}(\tilde{p}, \hat{p})$ the same asymptotic limits hold. These results shows that for discrete distributions, estimators which are asymptotically first-order efficient are not necessarily asymptotically equivalent under the PMC criterion.

3. Exponential lifetimes

Consider a sample of size n denoted by $X_1, \ldots, X_n$ from an exponential distribution with mean μ and let $T = \sum_{i=1}^n X_i$. The likelihood function for μ can be written $L(\mu) = \mu^{-n} e^{-T/\mu}$, the MLE of μ is given by $\hat{\mu} = T/n$ and the MELE of μ is

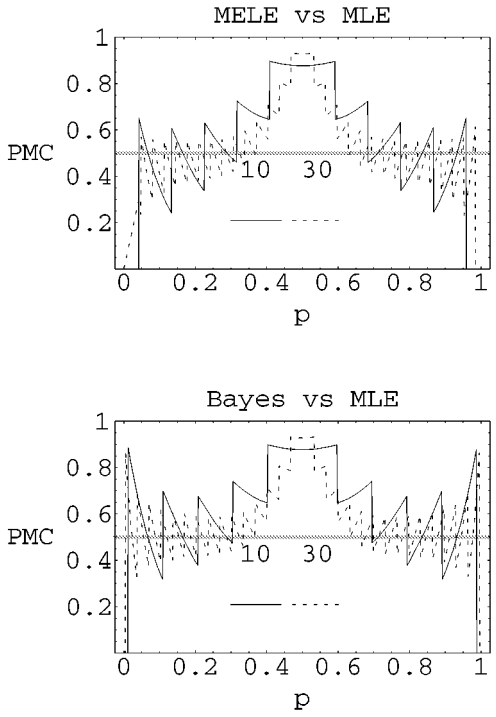

Fig. 2. *Pitman measure of closeness for alternative binomial estimators. Top panel: MELE, PMC($\bar{p}$, $\hat{p} \mid p$) for n = 10, 30. Bottom panel: Bayes estimator with Jeffrey's prior PMC($\tilde{p}$, $\hat{p} \mid p$) for n = 10, 30*

$\bar{\mu} = T/(n-2)$. The Jeffrey's prior for μ can be taken as μ^{-1} which produces a Bayesian estimator, $\tilde{\mu} = T/(n-1)$.

A simple computation with *Mathematica* gives the relative efficiency,

$$R(\bar{\mu}, \hat{\mu}) = \frac{1}{n} + \frac{n-5}{n+4} = 1 - \frac{8}{n} + \frac{36}{n^2} - \frac{144}{n^3} + \frac{576}{n^4} - \frac{2304}{n^5} + \mathcal{O}\left(\frac{1}{n}\right)^6.$$

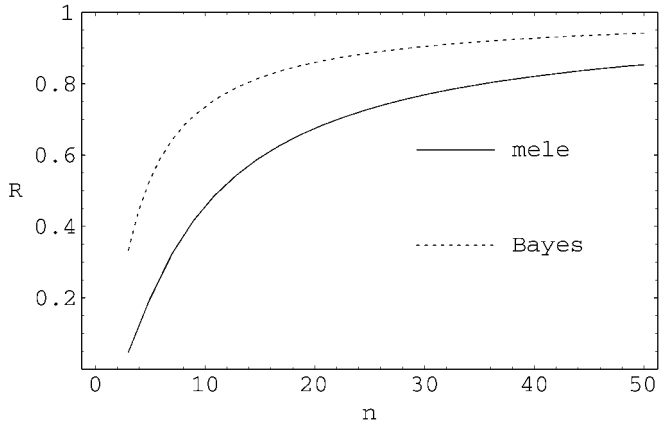

Fig. 3. *Relative efficiency R of the MELE and Bayes estimator vs the MLE of the mean μ in a random sample of size n from an exponential distribution*

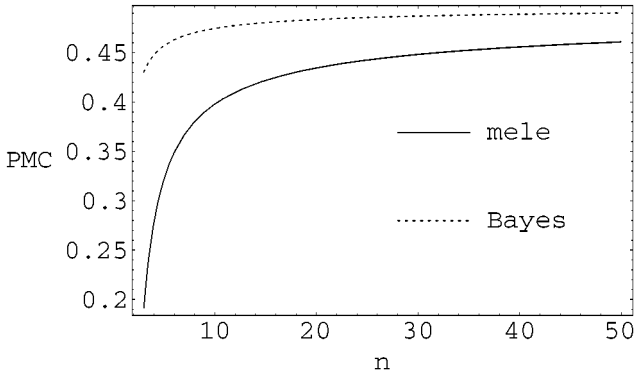

Fig. 4. *Pitman Measure of Closeness, PMC, of the MELE and Bayes estimator vs the MLE of the mean μ in a random sample of size n from an exponential distribution*

Similarly, $R(\tilde{\mu}, \hat{\mu}) = 1 + 1/n + 4/(n+1)$. Figure 3 shows that the MELE and Bayes estimator are less efficient that the MLE.

Since T has a standard gamma distribution with shape parameter n and scale parameter μ, the PMC is easily evaluated using the Geary-Rao Theorem (Keating, Mason and Sen 1993, p. 103). Letting $a = \bar{\mu}$ or $a = \tilde{\mu}$, we can write

$$\mathrm{PMC}(a, \hat{\mu} \mid \mu) = \int_0^{b\mu} \frac{e^{-x/\mu} x^{n-1} \mu^{-n}}{\Gamma(n)}\, dx$$

where $b = n(n-2)/(n-1)$ or $b = 2n(n-1)/(2n-1)$ according as $a = \bar{\mu}$ or $a = \tilde{\mu}$. Notice that without loss of generality we may take $\mu = 1$ since $\mathrm{PMC}(\bar{\mu}, \hat{\mu} \mid \mu) = \mathrm{PMC}(\bar{\mu}, \hat{\mu} \mid 1)$. From Fig. 4, $\mathrm{PMC}(a, \hat{\mu} \mid \mu) < 1/2$ for both $a = \bar{\mu}$ or $a = \tilde{\mu}$.

It is sometimes mistakenly thought that Theorem 4 or its Bayesian analogue guarantees that at least over some region of the parameter space, the MELE or the Bayes estimator will outperform the MLE but, as this example shows, this need not be the case.

4. MA(1) process

4.1. *Introduction*

The MA(1) time series with mean μ may be written $Z_t = \mu + A_t + \theta A_{t-1}$, where Z_t denotes the observation at time $t = 1, 2, \ldots$ and A_t, the innovation at time t, is assumed to be a sequence of independent normal random variables with mean zero and variance σ_A^2. The parameter θ determines the autocorrelation structure of the series and for identifiability we will assume that $|\theta| \leq 1$. When $|\theta| < 1$, the model is invertible (Brockwell and Davis 1991, §3.1). For simplicity we will examine the case where $\mu = 0$. Such MA(1) models often arise in practical applications as the model for a differenced nonstationary time series. The noninvertible case $\theta = 1$ occurs when a series is over-differenced.

In large-samples, standard asymptotic theory suggests that the maximum likelihood estimate for θ, denoted by $\hat{\theta}$, is approximately normal with mean θ and variance $(1-\theta^2)/n$ where n is the length of the observed time series. Cryer and Ledolter (1981)

established the somewhat surprising result that $\Pr\{\hat{\theta} = \pm 1\} > 0$. This result holds for all finite n and for all values of θ. For example when $n = 50$, $\Pr\{\hat{\theta} = 1 \mid \theta = 0\} = 0.002$ and $\Pr\{\hat{\theta} = 1 \mid \theta = 0.8\} = 0.13$ (Cryer and Ledolter 1981, Table 2). Let $\bar{\theta}$ denote the mean likelihood estimator of θ. In view of Theorem 1, this problem does not occur with $\bar{\theta}$.

The standard Bayesian estimator, $\tilde{\theta}$, is derived by Box and Jenkins (1976, p. 250–258) utilizing the Jeffrey's prior, $\pi(\theta) = 1/\sqrt{1-\theta^2}$. Since this prior tends to infinity at the endpoints, we see that the Bayes estimator also cause an undesirable pile-up effect like the MLE. This phenomenon is also investigated in our simulations.

Now we will show that the MELE dominates the MLE and Bayes estimator for both criteria when $n = 2$. When $n = 50$, the MELE is again better than the MLE and Bayes estimator unless the parameter θ is very close to ± 1. But this is due to undesirable pile-up effect of the MLE. We can conclude that MELE is generally a better estimator. Further mean-square error computations which support this conclusion for other values of n are given by Quenneville (1993) and can be verified by the reader using (McLeod and Quenneville 1999).

4.2. *Exact results for n = 2*

Consider a Gaussian time series with $n = 2$ and let Z_1, Z_2, be generated from the first-order moving average, $Z_t = A_t - \theta A_{t-1}$, where A_t are independent normal random variables with mean zero and variance σ_A^2. Let $W = -Z_1 Z_2/(Z_1^2 + Z_2^2)$. Then given data, Z_1, Z_2, the exact concentrated likelihood function for θ is (Cryer and Ledolter 1981, Quenneville 1993),

$$L(\theta \mid W) = \frac{\sqrt{1+\theta^2+\theta^4}}{1+\theta^2-2\theta W}$$

and

$$\hat{\theta} = \begin{cases} -1 & W \in [-0.5, -0.25] \\ \frac{1-\sqrt{1-16W^2}}{4W} & W \in (-0.25, 0.25),\ W \neq 0 \\ 0 & W = 0 \\ 1 & W \in [0.25, 0.5]. \end{cases}$$

Unfortunately $\bar{\theta}$ and $\tilde{\theta}$, cannot be evaluated symbolically. However using **NIntegrate** we can obtain it numerically. From Fig. 5, we see visually the pile-up effect for the MLE and that $\bar{\theta}$ and $\tilde{\theta}$ are either a linear or close to a linear function of W. To speed up our computations for the mean-square error of $\bar{\theta}$, we use the **FunctionInterpolation** in *Mathematica* to construct $\bar{\theta} = \bar{\theta}(W)$. The MSE and PMC for $\bar{\theta}$ and $\hat{\theta}$ are easily evaluated numerically using the pdf of W, $f_W(x)$, derived by Quenneville (1993),

$$f_W(x) = \frac{2\sqrt{1+\theta^2+\theta^4}}{\pi\sqrt{1-4x^2}(1+\theta^2-2\theta x)}, \quad |x| \leq 1/2.$$

From Figs. 6 and 7, it is seen that both the MELE and Bayesian estimator dominate the MLE both for the MSE and PMC criteria. The MELE is slightly better according to the MSE but according to the PMC the Bayes estimator is slightly better than the MELE.

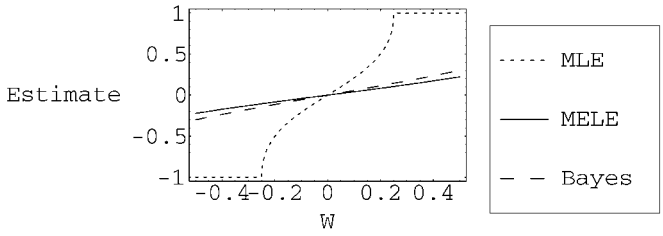

Fig. 5. *MLE, MELE and Bayes estimate for θ as a function of $W = -Z_1 Z_2/(Z_1^2 + Z_2^2)$ when $n = 2$ in MA(1)*

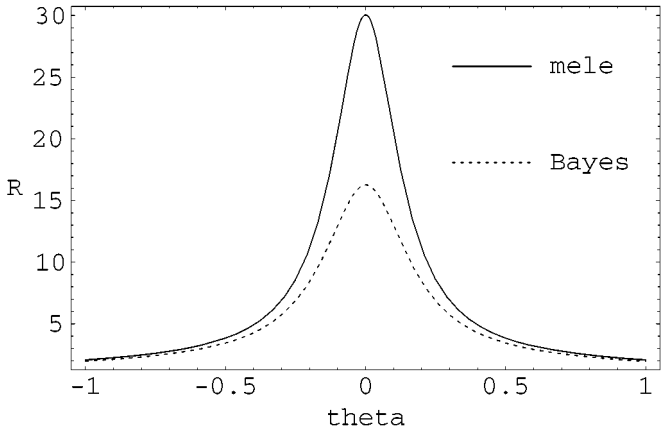

Fig. 6. *Relative efficiency, R, of MELE and Bayes estimator with Jeffrey's noninformative prior in the MA(1) model with $n = 2$*

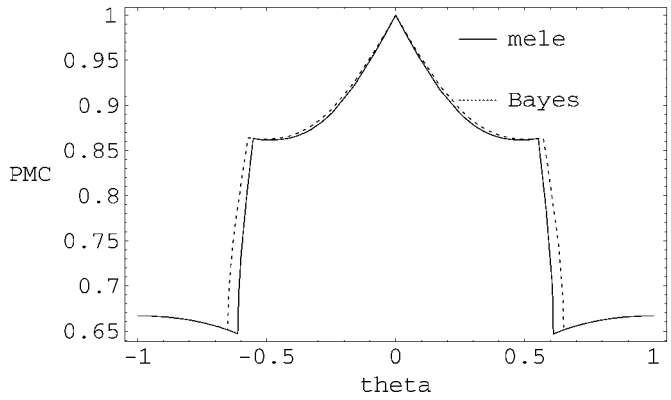

Fig. 7. *Pitman measure of closeness, PMC, of MELE and Bayes estimator with Jeffrey's noninformative prior in the MA(1) model with $n = 2$*

4.3. *Exact symbolic likelihood*

Consider the MA(1) process defined by $Z_t = A_t - \theta A_{t-1}$, where $\theta \leq 1$, A_t is assumed to be normal and independently distributed with mean zero and variance σ_A^2. Given n observations $Z' = (Z_1, \ldots, Z_n)$ the exact log likelihood function of an MA(1) process can be written (Newbold 1974),

$$\log L\left(\theta, \sigma_A^2\right) = -\frac{n}{2}\log\left(\sigma_A^2\right) - \frac{1}{2}\log(D) - \frac{1}{2\sigma_A^2}S(\theta),$$

where $h' = (1, \theta, \theta^2, \ldots, \theta^n)$, $D = h'h$ and

$$S(\theta) = (Lz - hh'Lz/D)'(Lz - hh'Lz/D),$$

where L is the $(n+1)$ by n matrix,

$$L = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ \theta & 1 & 0 & \cdots & 0 & 0 \\ \theta^2 & \theta & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \theta^{n-2} & \theta^{n-3} & \theta^{n-4} & \cdots & 1 & 0 \\ \theta^{n-1} & \theta^{n-2} & \theta^{n-3} & \cdots & \theta & 1 \end{pmatrix}.$$

Maximizing over σ_A^2 the concentrated log likelihood is given by

$$\log L_M(\theta) = -\frac{n}{2}\log[S(\theta)/n] - \frac{1}{2}\log(D).$$

This expression for the concentrated loglikelihood is just as easy to write in *Mathematica* notation as it is in ordinary mathematical notation. Moreover, it can be evaluated symbolically or numerically.

```
LogLikelihoodMA1[t_, z_] :=
    Module[{n = Length[z], Lz, h, detma1, v,
    Sumsq},
      Lz = Join[{0},
           Table[Sum[z[[i]]] t^(j-i), {i, 1, j}],
           {j, 1, n}]];
      h = Table[t^j, {j, 0, Length[z]}];
      detma1 = h . h;
      v = -h . Lz/detma1;
      Sumsq = (Lz + h v). (Lz + h v);
      -n/2 Log[Sumsq/n /. t -> t] -
        1/2 Log[detma1 /. t -> t]
    ];
```

4.4. *Efficient numeric likelihood computations*

Newbold's algorithm can be made much more efficient when only numerical values of the log likelihood are needed by using the *Mathematica* Compiler and by re-writing the calculations involved to make more use of efficient *Mathematica* functions such as **NestList**, **FoldList** and **Apply**. First consider the computation of the vector Lz which is of length $n+1$. After some simplifications, we see that $Lz = (\alpha_j)'$, where $\alpha_0 = 0$ is the first element and the remaining elements are defined recursively by $\alpha_j = \theta\alpha_{j-1} + Z_j$, $j = 1, 2, \ldots, n$, where $Z_0 = 0$. This computation is efficiently performed by *Mathematica*'s **FoldList**. When we are just interested in numerical evaluation we use the compile function to generate code which runs much faster.

```
GetLz=Compile[{{t, _Real},{z, _Real, 1}},
  FoldList[(#1 t + #2)&, 0, z]];
```

The determinant, $D = 1 + \theta^2 + \theta^4 + \cdots + \theta^{2n}$, is efficiently computed using **NestList** to generate the individual terms and then summing.

```
DetMA =Compile[{{t, _Real},{n, _Integer}},
  Apply[Plus,NestList[#1 t &, 1, n]^2]];
```

Next, we evaluate the term hLz/D. Since $hLz = \theta\alpha_1 + \theta^2\alpha_2 + \cdots + \theta^n\alpha_n$ we can use Horner's Rule to efficiently compute this sum. Horner's Rule is implemented in *Mathematica* using the function **Fold**.

```
Getu0 =Compile[{{t, _Real},{Lz, _Real, 1},
{detma, _Real}},
       -Fold[#1 t + #2&, 0, Reverse[Lz]]/detma];
```

The computation of the sum of squares function $S(\theta) = (Lz - hh'Lz/D)'(Lz - hh'Lz/D)$ is straightforward. The *Mathematica* compiler can be used to optimize the vector computations.

```
GetSumSq = Compile[{{t, _Real},{Lz, _Real, 1},
{u, _Real},{n, _Integer}},
  Apply[Plus,(Lz+NestList[#1 t &, 1, n] u)^2]];
```

Finally, the concentrated loglikelihood function is defined. The computation speed is increased by about a factor of 50 times when $n = 50$ and is even larger for larger n.

```
logLMA1F[t_, z_] :=
  Module[{n=Length[z]},
    Lz=GetLz[t,z];
    detma=DetMA[t,n];
    u=Getu0[t,Lz,detma];
    S=GetSumSq[t,Lz,u,n];
    -(1/2) Log[detma]- (n/2)Log[S/n]];
```

The maximum likelihood estimate can be obtained using *Mathematica*'s nonlinear optimization function **FindMinimum**. However as noted by Cryer and Ledholter (1981) the likelihood function can have multiple minima unlike the situation for the AR(1) likelihood function (Minozzo and Azzalini 1993). Thus we found it helpful to use our specialized function which locates the global optiminum in $(-1, 1)$ using a preliminary extensive grid-search followed by quadratic inverse interpolation optimization. Details are given in McLeod and Quenneville (1999).

The mean likelihood estimator $\bar{\theta}$ can be evaluated using **NIntegrate**.

```
Meanle[z_]:=
  NIntegrate[t E^logLMA1F[t, z],{t,-1,1}]/
    NIntegrate[E^logLMA1F[t, z],{t,-1,1}]
```

Notice that in the above expression the loglikelihood function is evaluated separately in both the numerator and denominator. Hence, we can save function evaluations by using our own numerical quadrature routine.

```
SimpsonQuadratureWeights[k_,a_, b_]:=
  With[{h=(2 k)/3},
   {a+(b-a)Range[0,2 k]/(2k),
   Prepend[Append[Drop[Flatten[Table[{4,2},{k}]],
{-1}],1],1]}]
```

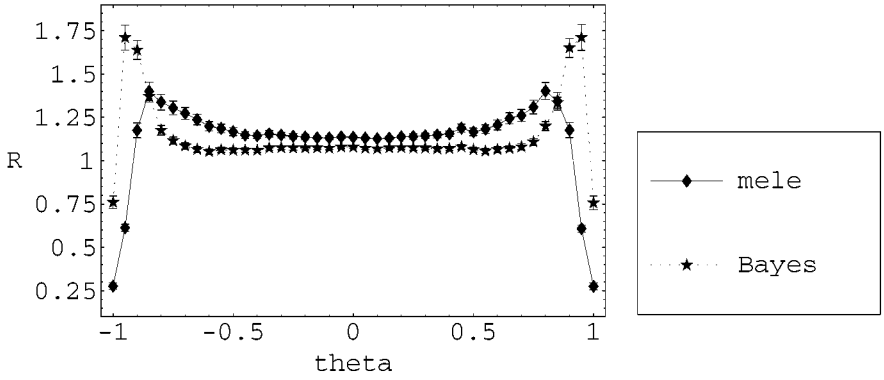

Fig. 8. *Empirical relative efficiency based on 10^4 simulations of the MA(1) with $\mu = 0$ and $n = 50$. In all cases the length of the 99.9% confidence interval for the estimate is less than the size of the plotting symbol*

```
{X,W}=SimpsonQuadratureWeights[100,-1,1];

GETMEANLEF=
  Compile[{{z, _Real, 1},
           {W, _Real, 1},{X, _Real, 1},{f,
           _Real, 1}},
          Plus@@(X f)/Plus@@f];
MEANLEF[z_]:=
  With[{f=Plus@@W E^(logLMA1F[#1,z]&/@X)},
    GETMEANLEF[z, W, X, f]];
```

Our tests indicate acceptable accuracy and about a 70% improvement in speed as compared with *Mathematica*'s more sophisticated **NIntegrate** function.

4.5. *Simulation results for n* = 50

We compared the MELE and Bayesian estimator with the MLE. Using the *Mathematica* algorithms derived above, we determined 99% confidence intervals for the relative efficiency and Pitman measure of closeness of each estimator with respect to the MLE. For each of the 41 parameter values $\theta = -1, -0.95, -0.90, \ldots, 0.95, 1$, 10^4 simulations were done. Complete details are available in the *Mathematica* notebooks which supplement this article (McLeod and Quenneville 1999). Figures 8 and 9 show that the MELE dominates the MLE except for the cases $\theta = \pm 1, \pm 0.95$. The fact that the MLE outperforms near the endpoints is due to the pile-up effect of the MLE (Cryer and Ledholter 1981). We can safely conclude that the MELE is a better overall estimator than the MLE or the usual Bayesian estimator. Of course, as already pointed out another cogent reason for preferring the MELE to the MLE or the Bayes estimator is that it does not produce noninvertible models.

As a check of our MLE computations we compared the probability of getting an estimate equal to ± 1 in our simulations with the probability derived from the approximate formula of Cryer and Ledholter (1981, Table 2). Figure 10 shows the agreement is very close.

If prior information is available then the Bayesian estimator with a suitable informative prior is better. Marriott and Newbold (1998) have developed an approach to the standard unit root problem in autoregressive time series by utilizing this fact.

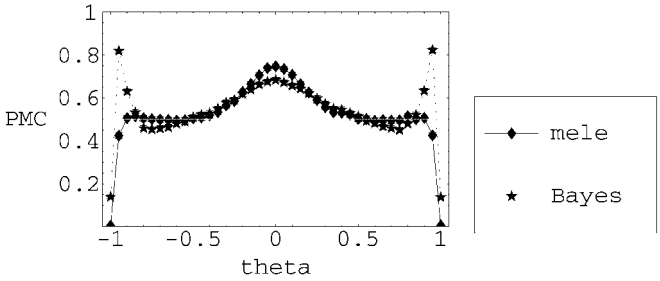

Fig. 9. *Empirical Pitman measure of closeness based on 10^4 simulations of the MA(1) with $\mu = 0$ and $n = 50$. In all cases the length of the 99.9% confidence interval for the estimate is less than the size of the plotting symbol*

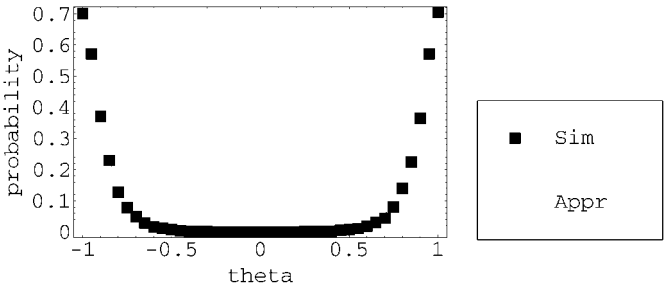

Fig. 10. *Estimated $Pr\{\hat{\theta} = \pm 1\}$ in the simulations compared with the probability from the approximate formula of Cryer and Ledholter (1981)*

The simulations were repeated with the mean μ estimated by the sample average and there was no major change in the results.

5. Concluding remarks

Previously Copas (1966) found that for the first-order autoregression, AR(1), the MELE had lower MSE over much of the

parameter region. Our results show that for the MA(1) the improvement is even better. The MSE is lower over a broader range and the pile-up effect of the MLE is avoided. Quenneville (1993) investigated the small sample properties of the MELE for many other time series models and gave a general algorithm for the MELE in ARMA models and found that in many cases the MELE produced estimates with smaller MSE over most of the parameter region.

Many Bayesian textbooks recommend the Jeffrey's noninformative prior on the grounds of parameter invariance. In view of our results for the MA(1) and the Bernouilli model, it seems that perhaps invariance is an unnecessary purely mathematical requirement if our primary interest is in the natural parameter, θ or p.

We would also like to mention that in our opinion *Mathematica* provides an excellent and indeed unparalleled environment for mathematical statistical research. In comparison, no other computing environment provides such high quality capabilities *simultaneously* in: symbolics, numerics, graphics, typesetting and programming. Typically most researchers need to develop some code to implement their methods. Often the researcher's code will only be executed a few times and the researcher's main consideration is his time and effort as opposed to producing an cpu-efficient stand-alone software product. The importance of a powerful user-oriented programming language for researchers is sometimes lacking in other environments. Iverson (1980) discussed the importance of the programming language to researchers and illustrated the advantages of *APL* over procedural programming languages. Procedural programming languages include Fortran, Cobol, C, Pascal, C, Java and SAS. A study carried out by IBM reported that *APL* programmers were about 15–20 times as efficient as programmers using a COBOL where efficiency is taken as the time needed by the programmer. Similar comments have been made by research statisticians on the ease of programming in *S* and *Splus* as opposed to SAS. If the programming language is a natural extension of mathematical notation, this translates into ease and speed of development. This was found to be true in the past with *APL*, *Splus* and XLISP-STAT. We have found that *Mathematica* provides even more capability.

However, for advanced state-of-the-art research and teaching in applied statistics and data analysis, R, Splus or XLISP-STAT may still be advantageous due to the wide usage by many leading researchers and the high quality functions for standard and advanced statistical methods that are available in the associated infrastructure (Statlib 1999). Furthermore, R and XLISP-STAT are freeware. However from the educational viewpoint, this advantage may not be so important since many students and researchers like to understand the principles involved. With *Mathematica* it is as easy to write out the necessary functions in *Mathematica* notation as it would be to explain the procedures in a traditional mathematical notation. In summary, *Mathematica*'s superior programming language is, in our opinion, one of its key strengths and advantages.

6. Appendix: Asymptotic PMC for binomial estimators

Consider the binomial probability function,

$$f_X(n, p) = \binom{n}{X} p^X (1-p)^{n-X}.$$

and the estimators $\hat{p} = X/n$ and $\bar{p} = (X+1)/(n+2)$. We now show that,

$$\lim_{n\to\infty} \mathrm{PMC}(\bar{p}, \hat{p} \mid p) = \begin{cases} 1 & p = 1/2 \\ \frac{1}{2} & p \neq 1/2, 0, 1 \\ 0 & p = 0, 1 \end{cases} \tag{6.4}$$

Since $\Pr\{\bar{p} = \hat{p}\} \to 0$ as $n \to \infty$, we can work with the simpler original definition,

$$\mathrm{PMC}(\bar{p}, \hat{p} \mid p) = \Pr\{|\bar{p} - p| < |\hat{p} - p|\}. \tag{6.5}$$

At the endpoints $p = 0,1$, equation (6.4) follows from the fact that $\hat{p} = p$ and $\bar{p} \neq p$.

The Geary-Rao Theorem is used to derive equation (6.7) below which is then used to establish equation (6.4) for other values of p. Using the notation of Keating, Mason and Sen (1993, p. 103) for the Geary-Rao Theorem and taking first the case, $0 \leq p \leq 1/2$, we can write

$$\mathrm{PMC}(\bar{p}, \hat{p} \mid p) = \Pr\{R_1\} + \Pr\{R_3\},$$

where

$$R_1 = \left\{ X : \frac{X+1}{n+2} < \frac{X}{n}, \frac{X+1}{n+2} + \frac{X}{n} > 2p \right\}$$

and

$$R_3 = \left\{ X : \frac{X+1}{n+2} > \frac{X}{n}, \frac{X+1}{n+2} + \frac{X}{n} < 2p \right\}.$$

After simplification, $R_1 = \{X : X > x_2\}$ and $R_3 = \{X : X < x_1\}$ where

$$x_1 = \left\lceil \frac{2pn(n+2) - n}{2(n+1)} \right\rceil, \tag{6.6}$$

and $x_2 = \lfloor n/2 \rfloor$, where $\lceil \bullet \rceil$ and $\lfloor \bullet \rfloor$ denote the ceiling and floor functions.

Hence,

$$\mathrm{PMC}(\bar{p}, \hat{p} \mid p) = \sum_{x=0}^{x_1-1} f_x(n, p) + \sum_{x=x_2+1}^{n} f_x(n, p). \tag{6.7}$$

For $p \geq 1/2$, $\mathrm{PMC}(\bar{p}, \hat{p} \mid p) = \mathrm{PMC}(\bar{p}, \hat{p} \mid 1-p)$. Our *Mathematica* notebook (McLeod and Quenneville 1999), provides a check on this formula by computing the PMC directly from equation (6.5) and also by simulation and then comparing with (6.7).

When $p = 1/2$, using equation (6.7) we obtain,

$$\mathrm{PMC}(\bar{p}, \hat{p} \mid p = 1/2) = \begin{cases} 1 & n \quad \text{odd} \\ 1 - f_{n/2}(n, 1/2) & n \quad \text{even} \end{cases} \tag{6.8}$$

Hence PMC($\bar{p}$, $\hat{p} \mid p = 1/2$) $\to 1$ as $n \to \infty$.

Next taking $0 < p < 1/2$ it follows that

$$\sum_{x=0}^{x_1} f_x(n, p) \to 1/2 \text{ as } n \to \infty. \tag{6.9}$$

Since $x_1 = \mu + \mathcal{O}(1)$, where $\mu = np$, equation (6.9) is established by using the standard normal approximation. Finally,

$$\sum_{x=x_2+1}^{n} f_x(n, p) \to 0 \text{ as } n \to \infty, \tag{6.10}$$

since $x_2 = \mu\alpha$, where $\alpha = 1/(2p) > 1$ and so x_2 is arbitrarily far out in the tails as $n \to \infty$. Hence, PMC($\bar{p}$, $\hat{p} \mid p \in (0, 1/2)$) $\to 1/2$ as $n \to \infty$. By symmetry, equation (6.4) follows.

Using the above approach, the same limits may be derived for PMC($\tilde{p}$, $\hat{p} \mid p$) where $\tilde{p} = (1 + 4X)/(2 + 4n)$ is the Bayes estimator under a Jeffrey's prior.

Acknowledgments

The authors wish to thank Jamie Stafford and two anonymous referees for their very helpful and insightful remarks.

References

Andrews D.F. and Stafford J.E. 1993. Tools for the symbolic computation of asymptotic expansions. Journal of the Royal Statistical Society B. 55: 613–627.

Barnard G.A., Jenkins G.M., and Winsten C.B. 1960. Likelihood inference and time series. Journal of the Royal Statistical Society Series A. 125: 321–372.

Borowski E.J. and Borwein J.M. 1991. The HarperCollins Dictionary of Mathematics, HarperCollins, New York.

Box G.E.P. and Jenkins G.M. 1976. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

Box G.E.P. and Tiao G.C. 1973. Bayesian Inference in Statistical Analysis. Addison-Wesley, Reading.

Brockwell P.J. and Davis R.A. 1991. Time Series Theory and Methods. Springer-Verlag, New York.

Cabrera J.F. 1989. Some experiments with maximum likelihood estimation using symbolic manipulations. In: Berk K. (Ed.), Proceedings of the 21st Symposium on the Interface of Statistics and Computer Science.

Copas J.B. 1966. Monte Carlo results for estimation in a stable Markov time series. Journal of the Royal Statistical Society A. 129: 110–116.

Cryer J.D. and Ledolter J. 1981. Small-sample properties of the maximum likelihood estimator in the first-order moving average model. Biometrika 68: 691–694.

Currie I.D. 1995. Maximum likelihood estimation and Mathematica. Applied Statistics 44: 379–394.

Dempster A.P. 1998. Logicist statistics I. Models and modeling. Statistical Science 13: 248–276.

Keating J.P., Mason R.L., and Pranab K.S. 1993. Pitman's Measure of Closeness. SIAM, Philadelphia.

Iverson K.E. 1980. Notation as a tool of thought. Communications of the Association of Computing Machinery 23: 444–465.

Marriott J. and Newbold P. 1998. Bayesian comparison of ARIMA and stationary ARMA models. International Statistical Review 66: 323–336.

McLeod A.I. and Quenneville B. 1999. Mathematica Notebooks to Accompany Mean Likelihood Estimation. http://www.stats.uwo.mcleod/epubs/mele.

Minozzo M. and Azzalini A. 1993. On the unimodality of the exact likelihood function for normal AR(2) series. Journal of Time Series Analysis 14: 497–510.

Newbold P. 1974. The exact likelihood function for a mixed autoregressive-moving average process. Biometrika 61: 423–426.

Pitman E.J.G. 1937. The closest estimates of statistical parameters. Proceedings of the Cambridge Philosophical Society 33: 212–222; Biometrika 30: 391–421.

Pitman E.J.G. 1938. The estimation of the location and scale parameters of a continuous population of any given form. Biometrika 30: 391–421.

Quenneville B. 1993. Mean likelihood estimation and time series analysis. Ph.D. Thesis, University of Western Ontario.

Quenneville B. and Singh A.C. 2000. Bayesian prediction MSE for state space models with estimated parameters. Journal of Time Series Analysis 21: 219–236.

Ritov Y. 1990. Asymptotic efficient estimators of the change point with unknown distributions. The Annals of Statistics 18: 1829–1839.

Rubin H. and Song K.S. 1995. Exact computation of the asymptotic efficiency of the maximum likelihood estimators of a discontinuous signal in a Gaussian white noise. The Annals of Statistics 23: 732–739.

Stafford J.E. and Andrews D.F. 1993. A symbolic algorithm for studying adjustments to the profile likelihood. Biometrika 80: 715–730.

Stafford J.E., Andrews D.F., and Wang Y. 1994. Symbolic computation: a unified approach to studying likelihood. Statistics and Computing 4: 235–245.

Statlib 1999. S-Archive. http://wwwstat.cmu.edu/.

Tanner M.A. 1993. Tools for Statistical Inference. Springer-Verlag, New York.

Wolfram S. 1996. Mathematica. Wolfram Research Inc., Champaign.