THE ECONOMIC AND SOCIAL RESEARCH INSTITUTE

Memorandum Series No. 174.

LIKELIHOOD AND ESTIMATION

by

Denis Conniffe

September 1986

IR£2

LIKELIHOOD AND ESTIMATION

by

Denis Conniffe

The Economic and Social Research Institute, Dublin 4, Ireland

LIKELIHOOD AND ESTIMATION*

Abstract

This paper argues that inferences about the true values of parameters logically proceed via the expectation of the (log) likelihood. In particular, the true values maximise the expected log likelihood. So the estimation problem is to estimate these maximising values. Only for a single parameter, or for special forms of the likelihood function, is this the same as maximising the likelihood. A modification to the maximum likelihood procedure is proposed and simple examples are used to suggest that the modification has advantages, besides being intuitively plausible.

* I thank John FitzGerald and Gary Keogh of the ESRI for useful discussions, and Professor John Spencer of Queen's University for helpful comments. The faults and omissions remain my own.

1 Introduction

It has been recognised that the M.L. procedure is not a self-evident principle. For example, Kendall (1940, 1949) remarked, it is not immediately or intuitively acceptable to proceed on the assumption that the most likely event has happened. Fisher, of course, felt differently and went much further when he said (Fisher, 1973, Ch. 3) "The likelihood supplies a natural order of preference among the possibilities under consideration". Obviously the likelihood could be used to order a set of values of a parameter θ but "natural order of preference" suggests that the resulting order is some way associated with order of closeness to the true value. Fisher went on to say "It is not surprising, therefore, though independently demonstrable, that in the Theory of Estimation, all rational criteria of what is to be desired in an estimate coverage on the particular value for which the likelihood is maximised".

I will argue that the desirable properties of the maximum likelihood estimator, when it is the appropriate estimator, follow from behaviour "on average", as related to the expected likelihood. It will be suggested that as a consequence ML estimators may be less appropriate than an alternative approach in the multiple parameter case. The arguments, which are solely frequentist in nature, are aimed as establishing the intuitive plausibility of the approach and rely on fairly simple examples. The objective is that the paper shall serve to raise discussion on the possibilities of the alternative approach.

2 A single parameter

An intuitive illustration may help introduce the arguments. Take a small sample x_1, where x_1 is the vector of observations, from a normal distribution with known variance. Let μ_o be the true mean. Then

$$\log L(\mu_o,x_1) < \log L(\bar{x}_1,x_1)$$

If the experiment is repeated, drawing the same size sample x_2, it is also true that

$$\log L(\mu_o,x_2) < \log L(\bar{x}_2, x_2)$$

But, on the other hand, it could well be true that

$$\log L(\bar{x}_2,x_1) < \log L(\mu_o,x_1) \text{ and } \log L(\bar{x}_1,x_2) < \log L(\mu_o,x_2).$$

So if drawings are repeated, there will be an $\bar{x}$ in any one repetition with a higher log likelihood than μ_o, but averaged over repetitions one could expect that μ_o would outperform any of the $\bar{x}_i$. Then the true value would maximise the expected log likelihood and the problem of estimating the true value would become that of estimating that maximising value. That the estimate might also be the value that maximises the sample likelihood, if it does so, could be thought of as coincidental.

Let θ represent an unknown parameter with true value θ_o and x be a vector of observations.

$$E[L(\theta, x)/L(\theta_o, x)] = 1$$

and since by the Arithmetic-Geometric Mean inequality,

$$E\{\log [L(\theta, x)/L(\theta_o,x)]\} < \log E [L(\theta, x)/L(\theta_o, x)],$$

it follows that for $\theta = \theta_o$

$$E [\log L(\theta, x)] < E [\log L(\theta_o, x)]$$

So, given regularity conditions, it follows that, at $\theta = \theta_o$,

$$\frac{\partial}{\partial\theta} [E(\log L)] = E (\frac{\partial \log L}{\partial\theta}) = 0 \qquad (1)$$

This result also follows starting from

$$\int L(\theta, x)\, dx = 1,$$

and differentiating both sides with respect to θ giving

$$\int \frac{\partial L}{\partial \theta}\, dx = \int \frac{\partial \log L}{\partial \theta}\, L dx = 0$$

which, for $\theta = \theta_0$ is equation (1) again. Of course, there is nothing original about equation (1) or in the methods of deriving it, but it can be interpreted in a different way than it usually is. The fact that the true value maximises the expected log likelihood leads to equation (1). Therefore θ_0 can be estimated by equating the derivative of log L to the expectation it would have (zero) at the true value. So the equation

$$\frac{\partial \log L}{\partial \theta} = 0$$

is an estimating equation for θ_0, as well as giving the value that maximises the sample likelihood. This coincidence of estimating true values and maximising likelihood will not, usually, carry over to the multi-parameter case.

A broader question in the single parameter case is if equal, or almost equal, values of the log-likelihood for parameter values θ_1 and θ_2 imply that these are equally supported by the data as contenders for the true value θ_0. Although θ_0 cannot be directly related to log L (θ, x), it can to $E\{\log L(\theta, x)\}$ since, as has just been seen, this takes its maximum at θ_0. By definition, log L(θ_1, x) and log L(θ_2, x) are unbiased estimators of $E\{\log L(\theta_1, x)\}$ and $E\{\log L(\theta_2, x)\}$ respectively, so if the log likelihoods are almost equal, the expected values could well be also. Some notion of closeness to the true value is implicit in questions of relative support for possible parameter values, and indeed also for Fisher's "natural order of preference". Suppose one is indifferent between θ_1 and θ_2 if $|\theta_0 - \theta_1| = |\theta_0 - \theta_2|$. The issue then is if this is implied by equal expected log likelihoods. Conversely, does $|\theta_o - \theta_1| < |\theta_o - \theta_2|$ imply $E[\log L(\theta_1, x)] > E[\log L(\theta_2, x)]$, so that the observed log likelihoods probably differ similarly?

The simple case of a sample of size one from a Normal illustrates the situation. First, suppose the variance is known. Apart from a constant,

$$\log L(\mu, x) = -\tfrac{1}{2} \log \sigma_o^2 - \frac{1}{2\sigma_o^2} (x - \mu)^2,$$

and so

$$E[\log L(\mu,x)] = -\tfrac{1}{2} [1 + \log \sigma_o^2 + \frac{1}{\sigma_o^2} (\mu_o - \mu)^2]$$

Clearly, if $|\mu_o - \mu_1| = |\mu_o - \mu_2|$ the expected log likelihoods are equal since there is symmetry about μ_o. Now suppose the mean is known rather than the variance. Apart from a constant

$$\log L(\sigma^2, x) = -\tfrac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (x - \mu_o)^2,$$

and so

$$E[\log L(\sigma^2, x)] = -\tfrac{1}{2}(\log \sigma^2 + \frac{\sigma_o^2}{\sigma^2}).$$

This is not symmetric about σ_o^2 and if σ_1^2, and σ_2^2 were equidistant on each side of σ_o^2, the larger would have the greater expected log likelihood.

Of course, the foregoing assumed that closeness to θ_o is measured by Euclidean distance. Defining distance as the difference between expected log likelihoods at θ and θ_o would ensure that equal log likelihoods implied equal "closeness". This definition might sometimes produce sensible distance functions - and in the immediate neighbourhood of the true value might always do so, since the expected log likelihood would be approximately quadratic about its maximum - but it often will not.

Then there is no general justification for assuming that equal likelihoods imply equal proximity to the true value. In the case of a single value from a

Normal with known mean, an observed equality of log $L(\sigma_1^2, x)$ and log $L(\sigma_2^2, x)$, with σ_1^2 and σ_2^2 very different, would suggest that the smaller is closer to the true value. Knowing all the values of $L(\theta, x)$ for given x and varying θ, is not generally enough for inference. The distributions, including the ranges of x matter.

3 Multiple parameters

As before it may be useful to commence with a simple illustration. Given a small sample of size n from a Normal with known mean μ_o, it will again be true that

$$\log L(\sigma_o^2, x) < \log L(S^2, x),$$

where
$$S^2 = \frac{1}{n}\sum_1^n (x_i - \mu_o)^2.$$

But if repetitions of the sample are considered, one expects that σ_o^2 will, on average, give a higher log likelihood than any other σ^2. If both mean and variance are unknown it is plausible to expect, and can be easily proved, that the maximum of the expected likelihood should occur at μ_o, σ_o^2. But in any sample the values of μ and σ^2 that maximise the likelihood are $\bar{x}$ and S^2 where

$$S^2 = \frac{1}{n} \sum_1^n (x_i - \bar{x})^2.$$

In any sample

$$\log L(\mu_o\ \sigma_o^2, x) < \log L(\bar{x}\ S^2, x),$$

but because the data are more closely grouped around the sample mean than around μ_o, and will be in every sample, one does not expect that an averaging over repetitions will reveal σ_o^2 to have the highest average log likelihood. Instead, a somewhat smaller value than σ_o^2 would have. The implication is that averaging over repetitions, with $\bar{x}$ and S^2 for μ and σ^2, does not lead to the maximum of the expected likelihood.

The log likelihood is

$$\text{Constant} - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\ \Sigma(x - \mu)^2,$$

with expectation

$$\text{Constant} - \frac{n}{2} [\log \sigma^2 + \frac{\sigma_o^2}{\sigma^2} + \frac{(\mu - \mu_o)^2}{\sigma^2}]$$

This has its maximum at μ_o and σ_o^2 . Having replaced μ by $\bar{x}$ the log likelihood is

$$\text{Constant} - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \Sigma(x - \bar{x})^2$$

with expectation

$$\text{Constant} - \frac{n}{2} (\log \sigma^2 + \frac{n-1}{n} \frac{\sigma_o^2}{\sigma^2}),$$

which has its maximum at $\sigma^2 = \sigma_o^2 (n-1)/n$.

The arguments of the previous section that proved that the true value maximises the expected log likelihood easily generalise, given regularity conditions, to multiple parameters so that, if θ is a vector of parameters with true value θ_o

$$E(\frac{\partial \log L}{\partial \theta_i}) = 0, \text{ for } \theta = \theta_o \qquad (3)$$

Note that the validity of (3) requires that all parameters, not cancelling out of the equation, be at their true values. If

$$\frac{\partial \log L}{\partial \theta_i} = G(\theta)\ H(\theta_i, x) \qquad (4)$$

equation (3) implies $E(H) = 0$ at $\theta_i = \theta_{io}$ so that equating H to zero both equates to the mean at the true value and determines the value of θ_i corresponding to the maximum of log L.

A little more generally, suppose

$$\Sigma\, a_i \frac{\partial \log L}{\partial \theta_i} = G(\theta)\ H(\theta_i, x) \qquad (5)$$

where the a_i are perhaps functions of θ . Then H may again be equated to zero since the expectation of the left hand side is zero. If H reduces to the form $\hat{\theta}_i(x) - \theta_i$ then $\hat{\theta}_i$ is an unbiased estimator and a variation of the Cramer-Rao bound argument shows that its variance is a lower bound for that of any unbiased estimator. For example, given a sample of size n from a bivariate normal:

$$\frac{1}{n}\,\frac{\partial \log L}{\partial \mu_1} = \sigma^{11}(\bar{x} - \mu_1) + \sigma^{12}(\bar{x}_2 - \mu_2),$$

and
$$\frac{1}{n}\,\frac{\partial \log L}{\partial \mu_2} = \sigma^{12}(\bar{x}_1 - \mu_1) + \sigma^{22}(\bar{x}_2 - \mu_2).$$

so
$$\frac{1}{n}\left[\frac{\partial \log L}{\partial \mu_1} - \frac{\sigma^{12}}{\sigma^{22}}\,\frac{\partial \log L}{\partial \mu_2}\right] = \frac{1}{\sigma_{11}}(\bar{x}_1 - \mu_1)$$

and the left hand side has expectation zero at the true value of θ.

However, if a linear combination of the derivatives of the log likelihood is insufficient so that, for example

$$\Sigma\,\Sigma\, a_{ik}\left(\frac{\partial \log L}{\partial \theta_i}\right)^{C_{ik}}\left(\frac{\partial \log L}{\partial \theta_k}\right)^{D_{jk}} = G(\theta)\; H(\theta_i x), \tag{6}$$

then the expectation of the left hand side may be non-zero at the true value of θ in small samples and setting H to zero may not be equating an expression to its expected value. Returning to the univariate normal as an example:

$$\frac{1}{n}\,\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2}(\bar{x} - \mu) \tag{7}$$

and
$$\frac{1}{n}\,\frac{\partial \log L}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2n\sigma^4}\,[(x - \mu)^2,$$
$$= -\frac{1}{2\sigma^2} + \frac{1}{2n\sigma^4}\,\Sigma\,(x-\bar{x})^2 + \frac{1}{2\sigma^4}\,(\bar{x} - \mu)^2$$

so
$$\frac{1}{n}\,\frac{\partial \log L}{\partial \sigma^2} - \left(\frac{1}{n}\,\frac{\partial \log L}{\partial \mu}\right)^2 = -\frac{1}{2\sigma^4}\left[\sigma^2 - \frac{1}{n}\,\Sigma(x-\bar{x})^2\right] \tag{8}$$

The expectation of the left hand side is $-1/(n\sigma^2)$ and not zero (where there is no danger of confusion σ^2 will be used to mean σ_o^2) and one could argue that the right hand side should be equal to this and not to zero. It will be shown in Section 5 that in at least some cases of the form (6) equating to expectations gives minimum variance unbiased estimators.

It should be said at this point that all problems cannot be reduced to the form (6). There need not always even be a closed form algebraic solution for θ_i from a set of non-linear equations. When there is, the left hand side corresponding to (6) could be a complicated function of x as well as of the θ and derivatives. Also the final estimating equations may be simultaneous, that is, taking the expectation of a left hand side may not cancel out all parameters except θ_i. However, it is not the purpose of this paper to present a general computational approach, but to raise discussion on the possibility of improving on maximum likelihood. In subsequent examples it will usually be most convenient to derive the equation of the form (6) by starting from the derivative of the likelihood with respect to θ_i, expanding it as two sets of terms with only the second involving parameters other than θ_i and then replacing these by functions of other likelihood derivatives. Of course, this is not being claimed to be a generally applicable method.

Maximum likelihood would equate the right hand sides of (6) and (7) to zero by making all derivatives zero. An alternative procedure would equate to the expectations of left hand sides, using the knowledge that

$$E\left(\frac{\partial \log L}{\partial \theta_i}\right) = 0 \qquad \text{for } \theta = \theta_o .$$

These equations express the fact that the true values maximise the expected likelihoods, which seem to be the only link between true values and likelihood in small samples. For the remainder of the paper I will call these alternatives, which seem to me more intuitively acceptable, the estimated maximum expected likelihood (EMEL) estimators.

In the normal example the ML and EMEL approaches differ in how equation (7) is interpreted and subsequently used. The EMEL interpretation does not mean $\bar{x} - \mu_o = 0$ but $E(\bar{x} - \mu) = 0$, at $\mu = \mu_o$. So $(\bar{x} - \mu_o)^2$ is non-zero and μ should be eliminated from the equation for σ^2 by replacing the term by the expectation of $(\bar{x} - \mu_o)^2$ rather than setting it to zero. In the normal example and more generally in equation (6) the EMEL procedure could also be regarded as equating H to its own expectation. Indeed, if H reduces to $\hat{\theta}_i(x) - \theta_i$ the procedure is equivalent to trying to correct the bias of the ML estimator.

4 Some examples

EXAMPLE 1: A single sample value from a Normal: This is perhaps a rather trivial case, but it puzzled me in the past. Maximum likelihood gives:

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2}(x - \mu), \text{ leading to } \hat{\mu} = x,$$

and $$\frac{\partial \log L}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x - \mu)^2, \text{ leading to } \hat{\sigma}^2 = 0,$$

when μ is replaced by its estimate. That is, the likelihood is increased to infinity by choosing the location parameter equal to the value and the spread zero. This is an absurd result. The EMEL approach would be to note that

$$\left(\frac{\partial \log L}{\partial \mu}\right)^2 = \frac{1}{\sigma^4}(x - \mu)^2$$

so that $$\frac{\partial \log L}{\partial \sigma^2} - \frac{1}{2}\left(\frac{\partial \log L}{\partial \mu}\right)^2 = -\frac{1}{2\sigma^2}$$

Taking the expectation of the left hand side gives the plausible, if uninformative, result

$$-\frac{1}{2\sigma^2} = -\frac{1}{2\sigma^2}$$

Other cases in which there are absurd results from maximum likelihood can also be resolved by the alternative procedure.

EXAMPLE 2: Standard Multiple Regression: In this case the log likelihood, apart from a constant, is

$$-\frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}(Y - XB)'(Y - XB)$$

The well known maximum likelihood estimators are:

$$\hat{B} = (X'X)^{-1} X'Y \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}(Y-X\hat{B})'(Y-X\hat{B}).$$

Now

$$\frac{\log L}{\partial \sigma 2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(Y-XB)'(Y-XB)$$
$$= \left[-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(Y-X\hat{B})'(Y-X\hat{B})\right] + \frac{1}{2\sigma^4}(\hat{B}-B)' X'X(\hat{B}-B)$$

The second term is $\frac{1}{2}\left(\frac{\partial \log L}{\partial B}\right)^1 (X'X)^{-1} \frac{\partial \log L}{\partial B}$ with expectation

$$\frac{1}{2\sigma^4} T_r (X'X. \operatorname{var} \hat{B}) = \frac{1}{2\sigma^2} p,$$

where p is the number of columns in X. So the estimating equation is

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(Y-X\hat{B})'(Y-X\hat{B}) + \frac{1}{2\sigma^2}p = 0,$$

or
$$\sigma^2 = \frac{1}{n-p}(Y-X\hat{B})'(Y-X\hat{B})$$

EXAMPLE 3: Missing values of an explanatory variable in simple linear regression

The model is $y_i = b\,x_i + e_i, \quad i = 1,2,\ldots,n,$

where the x_i for i = r + 1, r + 2,, . . . , n have been lost and will be treated as unknown parameters.

$$\log L = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_1^r(y_i - bx_i)^2 - \frac{1}{2\sigma^2}\sum_{r+1}^n(y_i - bx_i)^2$$

$$\frac{\partial\log L}{\partial x_i} = \frac{b}{\sigma^2}(y_i - bx_i), \quad i = r+1,\ r+2,\ \ldots,\ n.$$

$$\frac{\partial\log L}{\partial b} = \frac{1}{\sigma^2}\sum_1^r x_i(y_i - bx_i) + \frac{1}{\sigma^2}\sum_{r+1}^n x_i(y_i - bx_i)$$

$$\frac{\partial\log L}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_1^r(y_i - bx_i)^2 + \frac{1}{2\sigma^4}\sum_{r+1}^n(y_i - bx_i)^2$$

Maximum likelihood leads to the estimators:

$$\hat{b} = \sum_1^r x_iy_i/\sum_1^r x_i^2, \quad x_i = y_i/\hat{b}, \quad \text{for } i > r \text{ and } \hat{\sigma}^2 = \frac{1}{n}\sum_1^r(y_i - \hat{b}x_i)^2.$$

So the greater the number of missing values, the smaller the estimate of the variance. Using EMEL instead of ML leads to the same estimators of b and x_i $(i > r)$ because the expectation of the second term of the b equation, a linear combination of derivatives w.r.t. x_i is zero, while the first equation could be written

$$\frac{b}{\sigma^2}(y_i - \hat{b}\,x_i) + \frac{x_i}{\sum_1^r x^2}\left[b\frac{\partial\log L}{\partial b} - \sum_{j=r+1}^n x_j\frac{\partial\log L}{\partial x_j}\right]$$

and the expectation of the second term is again zero. The equation for σ^2 may be written

$$\frac{\partial\log L}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_1^r(y_i - \hat{b}x)^2 + \frac{1}{2\sigma^4}(\hat{b} - b)^2\sum_1^r x^2 + \frac{1}{2\sigma^4}\sum_{r+1}^r(y_i - bx_i)^2$$

The last term on the right has expectation $(n-r)/2\,\sigma^2$ and the third has expectation $1/\sigma^2$, being functions of squares of derivatives w.r.t. x_i and b, leading to

$$\sigma^2 = \frac{1}{r-1}\sum_1^r(y_i - \hat{b}x_i)^2$$

In all three of these examples the procedure was to expand the derivatives of the log likelihoods and then eliminate the other parameters by replacing terms by functions of derivatives with respect to other parameters. Taking expectations then led to the EMEL estimators.

5 Properties of estimated maximum expected likelihood

Many of the optimal properties associated with maximum likelihood are large sample properties. If all observations are independent and come from the same distribution f(θ x), where x is a scalar,

$$\frac{1}{n}\frac{\partial \log L}{\partial \theta_j} = \frac{1}{n}\frac{\partial}{\partial \theta_j}\sum_{i=1}^{n}[\log f(\theta, x_i)] = \frac{1}{n}\sum_{i=1}^{n}\frac{f_j(\theta, x_i)}{f(\theta, x_i)} \qquad (9)$$

where f_j denotes differentiation with respect to θ_j . So (9) is a sample mean and therefore by the strong law of large numbers tends to the expectation of each term for large n. This expectation is

$$E\left\{\frac{f_i(\theta, x_i)}{f(\theta, x_i)}\right\} = \int \frac{f_i(\theta, x_i)}{f(\theta, x_i)} f(\theta_o, x_i)\,dx_i = \int f_j(\theta_o, x_i)\,dx_i = 0, \text{ for } \theta = \theta_o.$$

So, for large n the true value θ_o satisfies the equations

$$\frac{1}{n}\frac{\partial \log L}{\partial \theta_i} = 0, \text{ as well as } E\left(\frac{\partial \log L}{\partial \theta_i}\right) = 0$$

and the former follows from the latter. If derivatives divided by n, and powers of them, can be treated as zero then so can the left hand sides of equations such as (5) and (7) and ML and EMEL coincide. So EMEL estimators have the same asymptotic properties as ML estimators.

More interestingly, there are cases where the EMEL estimators satisfy the criterion of consistency when the corresponding ML estimators do not. It is well known that consistency difficulties can arise with maximum likelihood when the sample values are from different distributions, for example, if the number of parameters increase with sample size. Suppose there are n Normal distributions with different means, and two observations from each. The case of one observation from each gives another example very similar to the first of Section 4. Apart from a constant the log likelihood is:

$$-\frac{2n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{2}(x_{ij}-\mu_i)^2$$

$$\frac{\partial \log L}{\partial \mu_i} = \frac{1}{\sigma^2}(\bar{x}_i - \mu_i)$$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\sum_{j=1}^{2}(x_{ij}-\mu_i)^2$$

Maximum likelihood gives

$$\hat{\sigma}^2 = \frac{1}{4}\left[\frac{1}{n}\sum_{i=1}^{n}(x_{i1}-x_{i2})^2\right]$$

Since the strong law implies that the sample mean tends to the expectation of $(x_{i1}-x_{i2})^2$, that is $2\sigma^2$, for large n, the estimator tends to $\frac{1}{2}\sigma^2$ and so is inconsistent. But writing

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\sum_{j=1}^{2}(x_{ij}-\bar{x}_i)^2 + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\sum_{j=1}^{2}(\bar{x}_i-\mu_i)^2$$

and replacing the last term on the right, which is half the sum of squared derivatives w.r.t. the μ_i, by its expectation gives

$$\sigma^2 = \frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n}(x_{1i}-x_{2i})^2\right],$$

which is consistent. Perhaps it is worthwhile taking a closer look at the equation for σ^2.

$$\frac{\partial \log L}{\partial \sigma^2} = \sum_{i=1}^{n}\left[-\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{j=1}^{2}(x_{ij}-\bar{x}_i)^2 + \frac{1}{\sigma^4}(\bar{x}_i-\mu_i)^2\right].$$

The right hand side is an unbiased estimator of the expectation of the derivatives and remains so if each $(\bar{x}_i-\mu_i)^2$ is replaced by its expectation. For the true value, $\mu_i = \mu_{io}$ and $\sigma^2 = \sigma_o^2$, the expectation of each term in the summation is zero and so the strong law of large numbers ensures consistency. If the $(\bar{x}_i - \mu_i)^2$ are set to zero, however, each term has a non-zero expectation and consistency does not follow.

It may be going too far to claim that estimated maximum expected likelihood will always produce consistent estimators when samples are drawn from different distributions. The strong law may not even hold. But supposing for a sample of size r, a derivative equation can be written, having replaced appropriate terms by expectations, so that the expectation of the right hand side is zero at the true value. Then it seems plausible that arguments based on n repetitions of samples of size r will prove consistency.

The optimal properties of a maximum likelihood estimators for θ_j in small samples depend on the attainment of the Cramér-Rao bound or on the existence of a single sufficient statistic for θ_j. But in these circumstances, equations (4) or (5) are valid and then ML and EMEL coincide. It is interesting to see if extra optimal properties can be established. Suppose

$$\frac{\partial \log L}{\partial \theta_i} = P(\theta_i\ x) + Q(\theta_i\ \phi\ x)$$

where ϕ is the vector of other parameters occurring in the equation. As in previous examples Q may be expressible as a function of powers of the derivatives w.r.t. ϕ so that the equations may be of the form (6). At the true value

$$E(P + Q) = E[P + E(Q)] = 0$$

so

$$E\{\ \tau\ (\theta_i)[P + E(Q)]\} = 0 \tag{10}$$

where $\tau(\theta_i)$ is a function of θ_i. Now suppose that t is an unbiased estimator of $\tau(\theta_i)$

$$E(t) = \tau(\theta_i)$$

Assuming regularity conditions and differentiating gives

$$E[t(P + Q)] = \tau'(\theta_i) \tag{11}$$

Now assume E(tQ) = E(t)E(Q). There will be log likelihoods for which this is true for any t that is unbiased. For example, if the likelihood can be written

$$g(\theta_i \phi\ \hat{\phi})\ \ h(\theta_i,\ x) \tag{12}$$

where g is the distribution of $\hat{\phi}$. Then

$$E(t) = \underset{\hat{\phi}}{E}\ [\ E_{x/\hat{\phi}}\ (t)] = \int [g(\theta_i\ \phi\ \hat{\phi})\ \ E_{x/\hat{\phi}}(t)\ \ d\ \hat{\phi}\ ].$$

If E(t), conditional on $\hat{\phi}$, is a function of $\hat{\phi}$ as well as of θ_i then the integral must be a function of ϕ as well as of θ_i and so t could not be unbiased for all θ_i and ϕ. Therefore

$$\underset{x/\hat{\phi}}{E}(t) = \tau(\theta_i),$$

and if the likelihood can be written in the form (12), then Q in (11) is a function of x only through $\hat{\phi}$. So

$$E(tQ) = \underset{\hat{\phi}}{E}\ \underset{x/\hat{\phi}}{E}(tQ) = \underset{\hat{\phi}}{E}(Q)\ \underset{x/\hat{\phi}}{E}(t) = E(t)\ E(Q).$$

Then (10) and (11) give

$$E\{[t - \tau(\theta_i)]\ [P + E(Q)]\} = \tau'(\theta_i),$$

and, using the Cauchy-Schwarz inequality,

$$\text{var}\ (t) \geqslant \frac{[\tau'\ (\theta_i)]^2}{E\ \{[P + E(Q)]^2\}} \tag{13}$$

This is an analogue of the Cramér-Rao bound and it is attained when

$$P + E(Q) = A[t - \tau\ (\theta_i)]. \tag{14}$$

So one procedure for seeking EMEL estimators sometimes reduces to equating one statistic t to its expectation while attaining the lower bound (13). Obviously t is very analogous to a sufficient statistic and one could continue to consider the families of distributions that permit equation (14). However, for the present the important point is that optimality properties can be obtained. A sample from a Normal with unknown mean and variance is one example where the likelihood can be written in the form (12). Then

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_1^n (x_i - \mu)^2$$

$$= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_1^n (x_i - \bar{x})^2 + \frac{n}{2\sigma^4}(\bar{x} - \mu)^2 ,$$

$$= \qquad P \qquad + \qquad Q.$$

So

$$P + E(Q) = -\frac{n-1}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_1^n (x_i - \bar{x})^2,$$

or

$$P + E(Q) = \frac{n-1}{2\sigma^4}\left[\frac{\Sigma(x_i - \bar{x})^2}{n-1} - \sigma^2\right] ,$$

which is of the form (14), so that the estimator

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$$

is the unbiased estimator of σ^2 of minimum variance attaining the lower bound (13).

It is fairly obvious that similar estimators of σ^2, with this small sample optimal property, will be obtainable in regression and analysis of variance models where the degrees of freedom will occur in denominators. It is tempting to speculate that the property also generalises to multiple variance component estimation and may relate to the restricted maximum likelihood procedure, advocated by Patterson and Thompson (1971) for analysis of incomplete block designs. I realise this discussion of EMEL's properties draws on suggestive examples rather than general proofs, but perhaps this is sufficient to at least generate interest.

6 Summary and discussion

The argument of this paper is that since the expected log likelihood is maximised at the true value, it is the co-ordinates of this maximum that ought to be estimated. For any known θ, the expected log likelihood is a function of θ and θ_o and is therefore unknown, but the log likelihood provides an estimator, unbiased by definition, of it. Since under regularity conditions

$$E(\frac{\partial \log L}{\partial\theta}) = \frac{\partial}{\partial\theta} E(\log L) = 0, \text{ for } \theta = \theta_o,$$

this is generally equivalent to equating expectations of derivatives to zero. From the expressions for the derivatives it is possible, at least in some cases, to equate a function of θ_i and the data to a function of the derivatives. If this function has a non-zero expectation the EMEL estimator, which equates to the expectation, differs from the ML estimator which equates to zero. The EMEL procedure, like MLin the case of a single sufficient statistic, sometimes gives on unbiased estimator of a function of θ_i while attaining a lower variance bound.

The notion that the likelihood contains all the information in the sample goes back to Fisher and has been variously interpreted. The extreme view that the likelihood alone should suffice for inference is incompatible with the development here, because even the choice of expectation of log likelihood rather than log likelihood as a criterion would contradict this. Some of the implications have been discussed in Section 2 for the case of a single parameter. However, even the more common interpretation of information as the reciprocal of the Cramér-Rao lower bound, could perhaps be questioned given (13). If no unbiased estimator can have a variance below the bound given by (13) and if the bound is attained for one estimator, is it reasonable to think the sample provides any more information ?

Fisher treated unbiasedness, at least in the context of maximum likelihood, as a relatively unimportant property. Yet the optimal properties of maximum likelihood in small samples in the case of sufficient statistics depends on an unbiased estimate of a function of the parameter attaining the Cramér-Rao bound. Unbiasedness, though not necessarily of the final estimator, plays an important part in the whole EMEL argument. The equation (15) equates quantities to their expectation and, simple though this is, it is the essential step. This unbiasedness may be carried through to equation (14) when that simplification is possible.

Large sample properties are the same for ML and EMEL, at least for independent observations from the same distribution. I would suggest that the optimal large sample properties of ML occur because it is then tending to coincidence with EMEL rather than vice versa. There is a direct connection between expected log likelihood and the true values of parameters that gives EMEL some prior plausibility which, to me at least, seems lacking in ML procedures.

The aim of this paper was to introduce and argue a case for an alternative approach to maximum likelihood and the examples chosen were capable of straightforward solution. I have not tried to present a general computational scheme and no doubt explicit algebraic solutions may be sometimes unachievable, demanding reliance on numerical methods. One approach in such a situation might be to start with the ML solution, assuming that consistent, and try to move towards the EMEL solution. The fact the EMEL is sometimes equivalent to correcting the bias of the ML estimator suggests that the variety of bias reduction techniques might repay investigation. On the other hand, the fact the EMEL estimators may be consistent when ML estimators are not would hint towards a wider approach. However, it would seem premature to investigate general computational methods before obtaining acceptance of the idea that it is the maximum of the expected likelihood that ought to be estimated, rather than the maximum of the sample likelihood.

References

FISHER, R.A. (1973) Statistical Methods and Scientific Inference, 3rd ed., (Edinburgh, Oliver and Boyd).

KENDALL, M.G. (1940) On the method of maximum likelihood, Journal of the Royal Statistical Society, 103, pp. 388-399.

KENDALL, M.G. (1949) On the reconciliation of theories of probability, Biometrika, 36, pp. 101-116.

PATTERSON, H.D. and THOMPSON, R. (1971) Recovery of inter-block information when block sizes are unequal. Biometrika, 58, pp. 545-554.