COMMUN. STATIST.-THEORY METH., 16(5), 1241-1251 (1987)

NEW DERIVATIONS OF THE MAXIMUM LIKELIHOOD ESTIMATOR AND THE LIKELIHOOD RATIO TEST

Andrey Feuerverger

Department of Statistics
University of Toronto
Toronto, Canada M5S 1A1

Key Words and Phrases: maximum likelihood estimator; likelihood ratio test; asymptotic optimality; Gauss-Markov theorem; Neyman-Pearson lemma; exact and approximate Bahadur slope; large deviation probabilities.

ABSTRACT

This paper consists of two distinct components. First, we show that the MLE may be considered to be an asymptotic consequence of the Gauss-Markov theorem. Second, we examine whether asymptotic optimization based on Bahadur slopes leads to the 'correct' result in the Neyman-Pearson context.

1. INTRODUCTION

This paper consists of two separate components: to present a new derivation for the maximum likelihood estimator, and to present a new derivation of the Neyman-Pearson likelihood ratio test. The aims and methods of the two components however are essentially distinct.

For maximum likelihood estimation, our motivation is partly pedagogic. Considering the importance of this method, it is rather surprising how few widely known expositions exist which provide *a priori* reasons for studying the MLE and for *anticipating* its theoretically important properties. Indeed most texts introduce maximum likelihood

estimation in more or less *ad hoc* fashion as being a 'reasonable' procedure and then proceed to discuss its optimality properties. The reader may gauge for himself this assertion by reviewing his own experience, or by examining the discussions in some major texts. Thus students often acquire an impression that the 'real reasons' for the MLE's asymptotic optimalities must somehow be obvious provided one can 'look at things' in just the right way. Ultimately, *through familiarity,* the student 'accepts' the result as natural, but usually cannot, if pressed, identify the source of optimality. In section 2 we give a development in which the MLE emerges in a non-ad-hoc manner out of a natural optimization problem, and specifically as an asymptotic consequence of the Gauss Markov theorem.

In the case of hypothesis testing, for a simple null hypothesis and a simple alternative, the Neyman-Pearson lemma tells the whole story. Nevertheless many important problems can only be approached through asymptotic methods involving criteria such as the Bahadur slope, Chernoff index or Hodges-Lehmann ARE. (A discussion of these may be found, for example, in Serfling (1980) chapter 10. We do not mention here the Pitman and Rubin-Sethuraman criteria, as we are interested only in fixed alternatives.) A natural question then arises: do these asymptotic optimality criteria lead to the 'correct' result in the simple Neyman-Pearson context? In addition to the fact that its answer is of intrinsic interest, and also says something about the criterion, this question is important also to justify employing the criteria in the non-standard problems to which they are applied. As well, the optimization method for resolving this question is of independent interest within the statistical context considered. In section 3 we resolve these questions rather completely for the Bahadur slopes, both approximate and exact; these criteria are by far the most important for applications. Finally, we state an open problem in the application of the method to the problem of testing for dependence.

Of course the properties of the MLE and LRT (likelihood ratio test) have been extensively studied and two particular references that bear on this work are Bahadur (1965) and Godambe (1960). However

the aims and methods of this work differ considerably from those of the references cited.

2. DERIVATION OF THE MLE

We shall show that the maximum likelihood method is an asymptotic consequence of the Gauss-Markov theorem and can be motivated in this way as the solution of a natural optimization problem.

Firstly, in the Gauss-Markov context we are concerned with the model

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{Z}}\,\underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\mathbf{e}}\,, \tag{2.1}$$

and if $\mathbf{e}$ is multivariate normal with mean $\mathbf{o}$ and nonsingular covariance Σ then the optimal estimator $\hat{\beta}$ of the true β_o is the solution of the linear equations

$$\mathbf{Z}'\Sigma^{-1}(\mathbf{Y} - \mathbf{Z}\hat{\beta}) = \mathbf{0}\,. \tag{2.2}$$

Next, let $X_1, X_2, \ldots, X_n$ be iid with density in $\{f_\theta\}$ and let F_θ denote the corresponding cdf's and $F_n(x) = n^{-1}\sum_{i=1}^{n} I(X_i \le x)$ denote the empirical cdf. The unknown true value of θ is denoted by θ_o. For simplicity here we shall take θ to be a real univariate parameter, and correspondingly we shall take the dimension of β in (2.1) to be $p = 1$; the arguments below generalize easily to the multiparameter case. We note the following correspondences between the two contexts:

(i) $\beta \longleftrightarrow \theta,\ \beta_o \longleftrightarrow \theta_o$

(ii) $\mathbf{Y} \longleftrightarrow F_n(x)$

(iii) $\mathbf{Z}\beta \longleftrightarrow F_\theta(x),\ \mathbf{Z}\beta_o \longleftrightarrow F_{\theta_o}(x)$

(iv) $\mathbf{e} = \mathbf{Y} - \mathbf{Z}\beta_o \longleftrightarrow F_n - F_{\theta_o}$

In (ii) the function $F_n(x)$ is being thought of as an infinite dimensional vector while for (iii) $F_\theta(x)$ is regarded as linear in a small neighbourhood of θ_o. By the correspondence (iv) is meant that $F_n - F_{\theta_o}$ is asymptotically zero mean and Gaussian. Now Σ also has a natural analogue, but the situation for Σ^{-1} is more involved. Nevertheless, corresponding to (2.2) we may write down the analogous 'linear' estimating equation

$$\int \big(F_n(x) - F_\theta(x)\big)\, dH(x) = 0 \tag{2.3}$$

except that $H(x)$ is not given and will have to be determined.

To this end rewrite (2.3) using integration by parts in the form

$$\int H(x)\, d\big(F_n(x) - F_\theta(x)\big) = 0 \tag{2.4}$$

and then Taylor-expand F_θ about θ_o to obtain

$$\begin{aligned} &\int H(x) d\big(F_n(x) - F_{\theta_o}(x)\big) - (\theta - \theta_o) \int H(x) \frac{\partial f_{\theta_o}(x)}{\partial \theta}\, dx \\ &\quad - \frac{1}{2}(\theta - \theta_o)^2 \int H(x) \frac{\partial^2 f_{\theta_*}(x)}{\partial \theta^2}\, dx = 0 \end{aligned} \tag{2.5}$$

for some θ_* between θ and θ_o. Then, to second order, the approximate solution to (2.4) is

$$\hat{\theta} = \theta_o + \frac{\int H(x) d\big(F_n(x) - F_{\theta_o}(x)\big)}{\int H(x) \dfrac{\partial f_{\theta_o}(x)}{\partial \theta}\, dx} \tag{2.6}$$

and this has variance

$$n \cdot \mathrm{VAR}_{\theta_o}(\hat{\theta}) = \frac{\mathrm{VAR}_{\theta_o}\big(H(x)\big)}{\left(\int H(x) \dfrac{\partial f_{\theta_o}(x)}{\partial \theta}\, dx\right)^2} . \tag{2.7}$$

The choice for $H(x)$ in (2.3) is taken as that which minimizes (2.7). But since this expression is clearly invariant under the changes $H(x) \to a + bH(x)$ we may take $H(x)$ as that function which minimizes $VAR_{\theta_o}(H(X))$ subject to $E_{\theta_o}H(x) = 0$ and $\int H(x) \frac{\partial f_{\theta_o}(x)}{\partial \theta} dx = 1$. Now since, in section 3, we will solve an analogous problem, we shall omit details here, but the reader may verify that the solution for $H(x)$, to within additive and multiplicative constant terms, is given by

$$H(x) = H_{\theta_o}(x) = \frac{\partial \log f_{\theta_o}(x)}{\partial \theta} \tag{2.8}$$

and that when substituted in (2.7) this gives $n \cdot VAR_{\theta_o}(\hat{\theta}) = I^{-1}(\theta_o)$ where $I(\theta)$ is Fisher's information.

Finally we need to take stock of the fact that (2.8) depends on the unknown θ_o. The obvious approach is to use H_{θ_1} where θ_1 is some consistent estimate, and then obtain, via (2.3), the estimate θ_2. We can then iterate this procedure using H_{θ_2} to obtain θ_3 and so on. Now suppose (as indeed happens under general conditions) that the solutions converge: $\theta_j \to \theta_M$. The limiting θ_M then clearly must satisfy

$$\int H_{\theta_M}(x)\, d\big(F_n(x) - F_{\theta_M}(x)\big) = 0\,, \tag{2.9}$$

i.e.

$$\int \frac{\partial \log f_{\theta_M}(x)}{\partial \theta}\, d\big(F_n(x) - F_{\theta_M}(x)\big) = 0\,. \tag{2.10}$$

It may readily be seen, however, that (2.10) is none other than the likelihood equation itself.

3. DERIVATION OF THE LRT

In this section we suppose that we have an iid sample $X_1, X_2, \ldots, X_n$ and seek to test a simple null hypothesis $H_o : f_o(x)$

versus a simple alternative $H_1 : f_1(x)$ (where f_o and f_1 are densities) on the basis of a test statistic having the form

$$T_n = \int U(x)dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} U(X_i) \ . \tag{3.1}$$

We shall first show that the criterion of the exact Bahadur slope may be used to derive the likelihood ratio statistic.

Taking large values of T_n as evidence against H_o we introduce the constraints

$$E_o T_n = \int U(x)f_o(x)dx = 0 \tag{3.2}$$

$$E_1 T_n = \int U(x)f_1(x)dx = 1 \ . \tag{3.3}$$

Then by a theorem of Bahadur (see Serfling (1980), p. 337) the exact slope of T_n is

$$\lim_{n\to\infty}\left[-\frac{2}{n}\log P_{H_o}\left(\frac{1}{n}\Sigma U(X_j) \geq 1\right)\right] \tag{3.4}$$

and by Chernoff's large-deviations theorem (ibid pp. 326-328) this

$$= -2\log\inf_z \int e^{z(U(x)-1)}f_o(x)\,dx \tag{3.5}$$

so that we need to minimize

$$\inf_z \int e^{z(U(x)-1)}f_o(x)\,dx \tag{3.6}$$

subject to (3.2) and (3.3). The presence of the $\inf_z$ in (3.6) makes this a seemingly formidable problem. However the difficulty may be circumvented by a simple device. In place of (3.6) we may extremize instead

$$\int e^{z(U(x)-1)}f_o(x)\,dx \tag{3.7}$$

while introducing the additional constraint

$$\int e^{z(U(x)-1)}\big(U(x)-1\big) f_o(x)\,dx = 0 \tag{3.8}$$

obtained from taking the derivative in z. Thus let z, $U(x)$ be the required solution and introduce variants δz and $\delta U(x)$. By (3.2) and (3.3) we must have

$$\int \delta U(x) \cdot f_o(x)\,dx = 0 \tag{3.9}$$

$$\int \delta U(x) \cdot f_1(x)\,dx = 0 \tag{3.10}$$

while (3.8) leads (ignoring second order terms) to

$$\int e^{z(u-1)} \Big[\delta z(U-1)^2 + \delta U\big(z(U-1)+1\big)\Big] f_o\,dx = 0\,. \tag{3.11}$$

This last equation may be regarded as giving δz in terms of δU; it plays no further role below.

Finally, introducing the variants in (3.7), ignoring second order terms, and noting (3.7) must already be at a maximum, we have by standard arguments that the first order term must be zero:

$$\delta z \cdot \int e^{z(U-1)}(U-1) f_o\,dx + z \cdot \int e^{z(U-1)} \cdot \delta U\, f_o\,dx = 0 \tag{3.12}$$

or, using (3.8)

$$\int e^{z(U-1)} \delta U\, f_o\,dx = 0\,. \tag{3.13}$$

Now (3.9), (3.10) and (3.13) mean $e^{z(U(x)-1)} f_o(x)$ is orthogonal to every function orthogonal to $f_o(x)$ and $f_1(x)$ suggesting that

$$e^{z(U(x)-1)} f_o(x) = \alpha f_o(x) + \beta f_1(x)\,. \tag{3.14}$$

On substituting (3.14) in (3.8), and then using (3.2) and (3.3), we find $\alpha = 0$ so that (3.14) becomes

$$U(x) = 1 + z^{-1} \log \beta \frac{f_1(x)}{f_0(x)} \tag{3.15}$$

and substituting in (3.2), (3.3) we find

$$z = \int \left(f_1(x) - f_0(x)\right) \log \frac{f_1(x)}{f_0(x)} dx \tag{3.16}$$

and

$$\beta = \exp\left[- \int f_1(x) \log \frac{f_1(x)}{f_0(x)} dx \right]. \tag{3.17}$$

Note that $z = J(0,1)$ and $\beta = \exp\left(-K(0,1)\right)$ where $K(0,1)$ is the Kullback-Leibler information number and $J(0,1)$ is the Kullback-Leibler divergence (see for example Kullback (1959)) and that the maximum attained in (3.5) is

$$\text{slope} = 2K(0,1) . \tag{3.18}$$

The inconsequential constraints (3.2) and (3.3) can be dropped to allow the solution (3.15) to be reexpressed, on linear transformation, in the form

$$U(x) = \log \frac{f_1(x)}{f_0(x)} \tag{3.19}$$

which in (3.1) gives the Neyman-Pearson test.

Consider secondly the criterion of the so-called approximate Bahadur slope. This is defined as

$$c = \lim_{n \to \infty} \left[- \frac{2}{n} \log (O\hat{L}S) \right] \tag{3.20}$$

where $\hat{OLS}$ is the observed level of significance of (3.1) computed from the asymptotic normal distribution $N\left(E_0 U(X), \frac{1}{n} VAR_0(U(X))\right)$ of T_n under H_0. Denoting this distribution by $P^*_{H_0}$ we have

$$c = \lim_{n\to\infty} \left[-\frac{2}{n} \log P^*_{H_0}\left(T_n > T_n(\text{obs})\right) \right] \tag{3.21}$$

$$= \lim \left[-\frac{2}{n} \log P^*_{H_0}\left(\frac{T_n - E_0 U}{\sqrt{VAR_0(U)/n}} > \frac{T_1(\text{obs}) - E_0 U}{\sqrt{VAR_0(U)/n}} \right) \right] \tag{3.22}$$

$$= \lim \left[-\frac{2}{n} \log \left(1 - \Phi \left(\frac{\sqrt{n}\left(T_1(\text{obs}) - E_0 U\right)}{\sqrt{VAR_0(U)}} \right) \right) \right] \tag{3.23}$$

$$= \frac{\left(E_1 U(X) - E_0 U(X)\right)^2}{VAR_0\left(U(X)\right)} \,. \tag{3.24}$$

In these calculations $T_n(\text{obs})$ denotes an observed value of T_n taken from the distribution of H_1. The step from (3.23) to (3.24) uses that $T_n = E_1 U(X) + o_P(1)$ under H_1, and the well known result for the normal cdf $\Phi(x)$ that

$$\log\left(1 - \Phi(x)\right) = -\frac{x^2}{2}\left(1 + o(1)\right) \qquad \text{as } x \to \infty \,. \tag{3.25}$$

We thus seek $U(x)$ to maximize (3.24). This problem is invariant under the changes $U \to a + bU$. Therefore we may require $E_0 U(X) = 0$, $VAR_0\left(U(X)\right) = 1$ and seek to maximize $E_1 U(X)$. We again omit details of the solution; the result (except for constants) is given by

$$U(x) = \frac{f_1(x)}{f_0(x)} \tag{3.26}$$

which, unlike (3.19) is not efficient. We may note however that for H_1 near H_0 in the sense that $\frac{f_1(x)}{f_0(x)}$ is near unity, we will have $\log \frac{f_1(x)}{f_0(x)} = \frac{f_1(x)}{f_0(x)} - 1$ approximately, so that the test which maximizes the approximate Bahadur slope will be nearly optimal. The maximum attained by (3.24) is seen to be $\int \left(f_1^2(x)/f_0(x)\right)dx - 1$.

We close with an open problem that underscores the nontrivial nature of the asymptotic theory for testing. Consider the problem of testing for dependence in the joint distribution of (X,Y) where (X_i, Y_i), i= 1,2,...,n is an iid sample, based on a *generalized covariance*-type statistic of form

$$\begin{aligned} S_n &= \iint V(x,y)\, d\left(F_n^{XY}(x,y) - F_n^X(x)F_n^Y(y)\right) \\ &= \frac{1}{n}\sum_i V(X_i,Y_i) - \frac{1}{n^2}\sum_i\sum_j V(X_i,Y_j) \end{aligned} \tag{3.27}$$

and consider a simple $H_0 : f^X(x)f^Y(y)$ and simple nonfactoring $H_1 : f^{XY}(x,y)$. What is the function $V(x,y)$ which maximizes the exact Bahadur slope? This natural problem, whose solution seems not to appear in the literature, is related in part to an important, unsolved problem concerning the large deviation probabilities for Hoeffding U-statistics and related forms. For (3.27) it is tempting to speculate that $V(x,y) = \log \frac{f^{XY}(x,y)}{f^X(x)f^Y(y)}$ is in some sense 'close' to the optimum.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The MLE result was obtained while the author was visiting Tel Aviv University whose facili-

ties and hospitality are gratefully acknowledged. It is a special pleasure to thank B. Kedem for stimulating conversations on this topic. On the occasion of his seventy-fourth birthday, I dedicate this work to Ben, with love.

BIBLIOGRAPHY

Bahadur, R.R. (1965). An optimal property of the likelihood ratio statistic. Proc. Fifth Berkeley Symp. Math. Statist. Prob., 1, 13-26.

Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. Ann. Math. Statist. 31, 1208-1211.

Kullback, S. (1959). **Information Theory and Statistics.** J. Wiley and Sons, New York.

Serfling, R.J. (1980). **Approximation Theorems of Mathematical Statistics.** J. Wiley and Sons, New York.

Received by Editorial Board member July, 1986; Revised December, 1986.

Recommended by Emanuel Parzen, Texas A&M University, College Station, TX.