ROCKY MOUNTAIN
JOURNAL OF MATHEMATICS
Volume 20, Number 1, Winter 1990

MAXIMUM LIKELIHOOD AND BEST APPROXIMATIONS

A. EGGER

ABSTRACT. That least squares approximation is an appropriate method in the presence of normally distributed errors is a consequence of the fact that the Maximum Likelihood Estimate and Best Approximation Problems coincide in this setting. It is shown that such a relationship holds only for exponentially distributed errors and the ℓ^p norms. Thus there is no norm which is similarly suited for curve fitting or data smoothing in the presence of errors distributed according to, for example, the Cauchy distribution.

Natural sources of approximation problems include curve fitting, signal filtering, data smoothing and parameter estimation. In the discrete setting, the object to be approximated is a vector $z \in \mathbf{R}^n$. Given a set $K \subset \mathbf{R}^n$ of approximating vectors, the Best Approximation Problem is to determine $k \in K$ as close as possible to z. When the distance function is given by a norm $||\ ||$, this is equivalent to minimizing $||k - z||$ over K. We shall say that $x^* \in K$ is a Best Approximation (BA) from K to z if $||x^* - z|| = \inf_{k \in K} ||k - z||$. There are infinitely many norms on $\mathbf{R}^n$ and, although they are all equivalent, they generate distinct Best Approximation Problems. Among the most frequently considered norms are the ℓ^p norms, $||x||_p = (\Sigma |x_i|^p)^{1/p}$, $1 \leq p < \infty$, and $||x||_\infty = \max\{|x_i| : 1 \leq i \leq n\}$. In a specific problem, the choice of a norm is typically influenced by computational considerations. Often, however, it is known or assumed that the residual $r = z - x^*$ is distributed according to some probability density function ρ. In this setting, there is sometimes a norm which is particularly appropriate.

Suppose that the approximating set K is a subspace of $\mathbf{R}^n$ with basis $V^1, V^2, \ldots, V^k$, where $V^i = (v_1^i, v_2^i, \ldots, v_n^i)$. Assume that $z = \overline{x} + \overline{\varepsilon}$, where $\overline{x} \in K$ and $\overline{\varepsilon} = (\overline{\varepsilon}_1, \overline{\varepsilon}_2, \ldots, \overline{\varepsilon}_n)$ are such that the $\overline{\varepsilon}_i$ are independent random errors distributed according to a probability density ρ. Specifying $x \in K$ to approximate z is equivalent to specifying a vector $\nu = (\nu_1, \ldots, \nu_k)$ of coefficients that generates a residual

Received by the editors on April 6, 1987 and in revised form on June 25, 1987.

$\varepsilon = z - x$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$, where $x = \sum_{i=1}^{k} \nu_i V^i$. In this setting, given ρ and z, the likelihood of x is given by $L(x) = L(\nu) = \prod_{i=1}^{k} \rho(\varepsilon_i)$ [**2**]. Then $x^* \in K$ is a Maximum Likelihood Estimate (MLE) for z if $L(x^*) = \sup_{x \in K} L(x)$. Since log is monotone, for any $\rho > 0$, this is equivalent to maximizing $\log(L(x))$.

Now, if ρ is an exponential distribution, $\rho(w) = \beta e^{-|w|^p/\alpha}, 1 \leq p < \infty$, we have that $\log(L(x)) = n \log(\beta) - (1/\alpha) \sum_j |\nu_1 v_j^1 + \nu_2 v_j^2 + \cdots + \nu_k v_j^k - z_j|^p$, which yields the following theorem.

THEOREM 1. *If the ε_i are assumed to be independently distributed according to the density $\rho(w) = \beta e^{-|w|^p/\alpha}, 1 \leq p < \infty$, then x^* is an* MLE *for z from K if and only if it is a* BA *to z from K in the ℓ^p norm* [**2**].

REMARKS. In the case $p = 2$, Theorem 1 justifies the use of the mean square norm when the errors are normally distributed. Note that, for any two exponential distributions $\rho_1(w)$ and $\rho_2(w)$ corresponding to exponents p_1 and p_2, with $p_1 > p_2 \geq 1$, we have that $\rho_1(w) < \rho_2(w)$ for large w. Thus the associated error curves would tend to be more spikey for p near 1. This is why ℓ^p approximation, $1 \leq p < 2$, is sometimes considered for Robust Estimization Problems. Furthermore, since the net of ℓ^p Best Approximations, $1 \leq p < \infty$, converges to a Best Approximation with respect to $\| \ \|_\infty$, this theorem also justifies Uniform Approximation when the residuals are to be uniformly distributed. Finally, the theorem easily generalizes to the case that the i-th error is distributed according to the density $\rho_i(w) = \beta_i e^{-|w|^p/\alpha_i}$. Here $\log(L(x)) = \sum \log(\beta_j) - \sum (1/\alpha_j) |\nu_1 V_j^1 + \cdots + \nu_k V_j^k - z_j|^p$, so determining MLEs is equivalent to finding weighted ℓ^p BAs.

If such a MLE-BA pairing were to hold for arbitrary error distributions, then it would be possible to construct a norm which was appropriate for a specific problem or data set. That this is not possible is easily seen by considering any density ρ such that $\rho(0) = 0$. Then, for $z \in K$, z need not be a MLE for itself, but must be a BA in every norm. Even if $\rho(0) > \rho(w) > 0$ for all $w \neq 0$, there need not exist a MLE-BA correspondence, as the following example shows.

EXAMPLE 1. Consider the Cauchy distribution, $\rho(w) = 1/(\pi(1+w^2))$. Let $K \subset \mathbf{R}^4$ be the subspace of constant vectors. Set $z = (a, a, -a, -a)$ and $x = (a, a, a, a,)$. Then $L(0) = 1/\left(\pi^4(1+a^2)^4\right)$ and $L(x) = 1/\left(\pi^4(1+4a^2)^2\right)$. For large a, $L(0) < L(a)$, so 0 is not a MLE in this case. Observe that if y is a MLE for z with respect to ρ, then so is $-y$. If the BA-MLE pairing were to hold for $||\ ||$ and ρ, then both y and $-y$ would be BA's. By the convexity of norms, 0 would be a BA, which is impossible. Thus, in this case, there is no norm for which the BA-MLE pairing holds.

Although Example 1 shows that a BA-MLE pairing is not possible for every norm, perhaps such a pairing holds for some additional norms. In view of Theorem 1, it is natural to consider generalizations of the ℓ^p norms. Perhaps the simplest such norms are the Luxemburg norms, defined as follows. Let M be a continuous, convex, non-decreasing function defined for $t \geq 0$ with $M(0) = 0$ and $\lim_{t\to\infty} M(t) = \infty$. The Luxemburg norm $||\ ||_M$, is defined by $||x||_M = \inf\{\lambda > 0 : \sum M[|x_i|/\lambda] \leq 1\}$ [**4**]. Let ρ be a continuous, symmetric, probability density which is strictly decreasing for $w > 0$. Define $M(w) = -\log(\rho(w)/\rho(0))$. Assume that M is convex. Then M satisfies the conditions above and we may form $||\ ||_M$. Then, with the notation of Theorem 1, we have

THEOREM 2. *For each $z \in \mathbf{R}^n$, there exists a norm, $||\ ||_N$ such that x^* is a* BA *to z with respect to this norm if and only if x^* is a* MLE *for z with respect to ρ.*

PROOF. In this setting, the existence of at least one MLE is immediate. Let $\beta = \sum M(w_i - z_i)$ for any w a MLE. Then $\beta \leq \sum M(y_i - z_i)$ for all $y \in K$. If $\beta = 0$, then $z \in K$, and the result is immediate in this case. If $\beta \neq 0$ let $N(w) = 1/\beta M(w)$. Then $||x^* - z||_N \leq ||w - z||_N$. If, for any $\alpha > 0$, $\sum N[(w_i - z_i)/\alpha] \leq 1$, then $\sum N[(x_i^* - z_i)/\alpha] \leq 1$ as well. For $\lambda = 1$, $\sum N(w_i - z_i) = 1/\beta \sum M(w_i - z_i) = 1$. Thus $\sum M(x_i^* - z_i) \leq \beta$ and x^* is a MLE. Now suppose that x^* is a MLE. Then $\sum N(x_i^* - z_i) = \beta$ and $\sum N(x_i^* - z_i) = 1$. Since there must exist at least one BA with respect to $||\ ||_N$, let y be such a BA. Then $\sum N[(y_i - z_i)/\lambda] \leq 1$ for some $\lambda \leq 1$. If x^* is not a BA, then this in-

equality holds for some $\lambda < 1$. Since N and M are strictly increasing, $\sum N(y_i - z_i)$ would be strictly less than 1 which is impossible. Thus x^* must be a BA. □

Theorem 2 is not completely satisfactory, since the norm depends upon z. The following example illustrates this shortcoming.

EXAMPLE 2. Let $\rho(s) = e^{-c(s^2+|s|)}$ be such that $\int_R \rho\, ds = 1$. Then ρ satisfies the conditions of Theorem 2. Let K be the subspace of constant vectors in $\mathbf{R}^3$ and let $z = (0, a, 0)$. If x^* is a BA from K to z with respect to some fixed norm, then λx^* is a BA to λz as well. Finding MLE's with respect to ρ is equivalent to minimizing $2b^2 + 2|b| + (b-a)^2 + |a-b|$ over b in $\mathbf{R}$. For $a = 1$ this occurs uniquely at $b = 1/6$, and for $a = 1/2$ this occurs uniquely at $b = 0$. If BA-MLE pairing were to hold independent of z, then $(1/12, 1/12, 1/12)$ would be a MLE for $(0, 1/2, 0)$ which is not possible.

In view of the above, it is of interest to characterize those probability densities and norms for which a BA-MLE pairing holds independent of the specific problem. Such a characterization follows.

THEOREM 3. *Let $X = \Phi_0$, the vector space of sequences $\{\xi_i\}$ such that $\xi_n = 0$ for large n. Suppose that $||\ ||$ is a norm on every finite dimensional subspace of X. Let ρ be a probability density function satisfying*

(1) $\rho(0) > \rho(w) = \rho(-w) > 0$ *for* $w \neq 0$,

(2) ρ *is continuously differentiable*,

(3) $\log(\rho(w))$ *is strictly decreasing for* $w > 0$.

Suppose that, for each finite dimensional subspace $K \subset X$ and each $z \in X$, we have that $k \in K$ is a BA *to z with respect to $||\ ||$ if and only if k is a* MLE *for z with respect to ρ. Then ρ is one of the exponential distributions of Theorem* 1.

PROOF. For fixed n, m let $K = K_{nm}$ be given by

$$K_{nm} = \{x : x(i) = a,\ 1 \leq i \leq n+m,\ x(i) = 0 \text{ if } i > m+n\},$$

and define $z = z_{nm}$ by

$$z(i) = \begin{cases} 0, & 1 \le i \le n, \\ 1, & n < i \le n+m. \end{cases}$$

If x^* is a BA to z from K, then λx^* is a BA to λz for each $\lambda \in \mathbf{R}$. If the Let BA-MLE pairing holds, then λx^* must be an MLE for λz. Conditions (1)–(3) ensure that x^* is an MLE for z if and only if

$$n \log \rho(a) + m \log \rho(a-1) = \max_{b \in \mathbf{R}} (n \log \rho(b) + m \log \rho(b-1)),$$

where $x^*(i) = a$, $i = 1, \ldots, n+m$. These conditions also ensure this is equivalent to

$$\frac{n\rho'(a)}{\rho(a)} = \frac{-m\rho'(a-1)}{\rho(a-1)}. \tag{1}$$

Similarly λx^* is an MLE for λz if and only if

$$\frac{n\rho'(\lambda a)}{\rho(\lambda a)} = \frac{-m\rho'(\lambda a - \lambda)}{\rho(\lambda a - \lambda)}. \tag{2}$$

Let $A = \{a \in \mathbf{R} : a \text{ satisfies (1) for some } n \text{ and } m\}$. The cases $n = 0, m = 1$ and $n = 1, m = 0$ ensure $0 \in A$ and $1 \in A$. In fact, A is dense in $[0,1]$. To see this let $a \in (0,1)$. Then, for some $s > 0$, $s\frac{\rho'(a)}{\rho(a)} = \frac{\rho'(1-a)}{\rho(1-a)}$, and since increasing a decreases $1-a$, condition (3) ensures that there exist arbitrarily small $\varepsilon > 0$ and m and n positive integers, for which $n/m = \frac{\rho'(a+\varepsilon)}{\rho(a+\varepsilon)} = \frac{\rho'(1-(a+\varepsilon))}{\rho(1-(a+\varepsilon))}$ is a rational number. Thus A is dense in $[0,1]$. Define $F(w) = \frac{\rho'(w)}{\rho(w)}$. Equation (2) implies that, for $a \in A$, $\lambda a = w, F(w(1-a)/a) = C_a F(w)$ for fixed C_a and all w. For $w = 1$, we have $F((1-a)/a) = C_a F(1)$. If $F(1) = 0$, then $F((1-a)/a) = 0$ for all non-zero $a \in A$, which is not possible. So $F(tw) = F(t)F(w)/F(1)$ for $t = (1-a)/a, a \neq 0, a \in A$. $F(1) < 0$, so let $-\alpha = 1/F(1)$. Since A is dense and F is continuous and odd, $F(tw) = -\alpha F(t)F(w)$ for all t and w. Set $G(w) = -\alpha F(w)$. Then $G(wt) = G(w)G(t)$. Then $G(w) = |w|^c \text{sign}(w)$ [**1**]. Hence $F(w) = |w|^c \text{sign}(w)/(-\alpha)$ and $\rho(w) = \beta e^{-|w|^\delta/\nu}$ as desired. □

Thus problems involving exponentially or uniformly distributed residuals are naturally associated with the corresponding ℓ^p norms $1 \le p \le$

∞, via the principle of maximum likelihood. Any other residual distribution can have no such pairing, and the choice of approximating norm would have to be based on other considerations.

REFERENCES

1. J. Aczel, *Lectures on Functional Equations and Their Applications,* Academic Press, New York, 1966.

2. D.R. Cox and D.V. Hinkley, *Theoretical Statistics,* Chapman and Hall, London, 1974.

3. M.A. Krasnosel'skiĭ and Y.B. Rutickiĭ, *Convex Functions and Orlicz Spaces,* P. Noordhoff, Groningen, 1961.

4. J. Lindenstrauss and L. Tzafriri, *Classical Banach Spaces* I, Springer Verlag, Berlin, 1977.

Department of Mathematics, Idaho State University, Pocatello, ID 83209