Minimum Cross-Entropy Estimation with Inaccurate Side Information

L. Lorne Campbell *

Abstract

Given a prior estimate of a probability, $\mathbf{q}$, and a constraint $\sum p_i x_i = a$, one well-known way of estimating $\mathbf{p}$ is to minimize the cross-entropy $I(\mathbf{p};\mathbf{q})$ subject to the constraint. A modification to this method is proposed for use when the value a is only approximately known. The modification is based on the penalty function method in constrained optimization. It has an interpretation in differential geometry methods in statistics and it sometimes gives a maximum likelihood estimate.

Index Terms — Minimum cross-entropy, maximum entropy, inaccurate data.

---

*This research was supported by the Natural Sciences and Engineering Research Council of Canada, under grant OPG 0002151. Some of these results were reported at the IEEE International Symposium on Information Theory, St.Jovite, Canada, Sept. 1983. The author is with the Department of Mathematics and Statistics, Queen's University, Kingston, Ontario K7L 3N6, Canada.

1 Introduction

In the standard formulation of the minimum cross-entropy method, one has a prior estimate $\mathbf{q} = (q_1, \ldots, q_N)$ of the probabilities of certain events, where $q_i > 0$ and $\sum_1^N q_i = 1$. The "true" probability, $\mathbf{p} = (p_1, \ldots, p_N)$ is unknown, but one knows the the expected value of some random variable X. That is, one has side information of the form

$$\sum_{i=1}^{N} p_i x_i = a, \tag{1}$$

where a and $(x_1, \ldots, x_N)$ are known. (It is assumed throughout that $\min_i\{x_i\} < a < \max_i\{x_i\}$ so that there are many $\mathbf{p}$ satisfying the constraint (1).) If $I(\mathbf{p}; \mathbf{q})$ is the cross-entropy (relative entropy, directed divergence) defined by

$$I(\mathbf{p}; \mathbf{q}) = \sum_{i=1}^{N} p_i \ln \frac{p_i}{q_i}, \tag{2}$$

then the minimum cross-entropy method of estimating $\mathbf{p}$ is to minimize $I(\mathbf{p}; \mathbf{q})$ subject to the constraints (1) and

$$\sum_{i=1}^{N} p_i = 1 \quad p_i > 0. \tag{3}$$

When $\mathbf{q} = (N^{-1}, \ldots, N^{-1})$, *i.e.* the uniform distribution, this is the maximum entropy method. For recent applications of this method see [9] and for a variety of reasons to use the method, see [4, 6, 7, 10, 12, 13, 14, 15].

The method described above would normally be applied when the side information (1) is regarded as being much more reliable than the prior estimate $\mathbf{q}$. In this note, we examine a proposed modification for the case that the value a is only approximately known and one has similar degrees of confidence in the prior estimate $\mathbf{q}$ and in the value a. The method will be developed as a consequence of using a penalty function approach to the minimum cross-entropy method. It will then be shown to have a geometric interpretation. It also agrees with the maximum likelihood method in a special case where both methods are applicable.

For a different approach to the problem of approximately known values of a, scalar or vector, see the work of Csiszár, Gamboa, and Gassiat [8].

2 Penalty function approach

We propose to minimize over $\mathbf{p}$ the function

$$F(\mathbf{p}, \lambda) = I(\mathbf{p}; \mathbf{q}) + \lambda \left| \sum_{i=1}^{N} p_i x_i - a \right| \tag{4}$$

subject to the constraint (3). The parameter $\lambda > 0$, which multiplies the penalty function, determines the weight that is given to the constraint (1). If one has little confidence in the accuracy of (1) then λ should be close to zero, while if one has a great deal of confidence, one should use a large value of λ.

The solution of this minimization problem is easily found by the Lagrange multiplier method and can be shown to be a minimum by a convexity argument. We omit the details since this demonstration so closely resembles the classical minimum cross-entropy calculation. Without loss of generality we assume that $\sum q_i x_i > a$. (If the inequality is reversed we can replace x_i by $-x_i$ and a by $-a$; if there is equality, the minimum is obviously at $\mathbf{p} = \mathbf{q}$.)

Define the partition function Z by

$$Z(\lambda) = \sum_{i=1}^{N} q_i e^{-\lambda(x_i - a)}, \tag{5}$$

put

$$p_i(\lambda) = \frac{q_i e^{-\lambda(x_i - a)}}{Z(\lambda)}, \tag{6}$$

and put

$$\mathbf{p}(\lambda) = (p_1(\lambda), \ldots, p_N(\lambda)).$$

Using the conditions $\min_i\{x_i\} < a < \max_i\{x_i\}$, $\sum q_i x_i > a$, and the positivity of q_i, it is easily shown that $Z(\lambda)$ has a unique minimum at a positive value Λ. When $\sum q_i x_i < a$ the modification necessary is to replace $(x_i - a)$ by $(a - x_i)$ in (5) and (6).

In the notation defined above, the solution of the minimization problem posed at the beginning of this Section is found to be $\mathbf{p}(\lambda)$ for $0 < \lambda \leq \Lambda$ and $\mathbf{p}(\Lambda)$ for all $\lambda \geq \Lambda$. The minimum cross-entropy solution is also $\mathbf{p}(\Lambda)$.

3 Geometric view

In recent years, Amari [1], Čencov [5], and others have used differential geometry to obtain insight into questions of statistics. In particular, one can regard the set of points $\mathbf{p}$ satisfying (3) as a differentiable manifold. It can be shown that there is a Riemannian geometry on this manifold, with line element ds defined by

$$(ds)^2 = \sum_{i=1}^{N} \frac{(dp_i)^2}{p_i},$$

which has specially important properties from a probabilistic point of view. For a development of this point of view that brings out some of the connections to information theory, see [2, 3]. In this view, the exponential family of probability laws $\mathbf{p}(\lambda)$ can be regarded as a curve on the above Riemannian manifold that passes through the point $\mathbf{q}$ and that bears a particularly simple relation to the random variable X [2]. In this Riemannian geometry, the curve is an orthogonal trajectory to the family of surfaces with equations $E[X] = \text{constant}$.

More generally, there is such a curve through each point $\mathbf{q}$ of the manifold. These are geodesic curves for a certain affine connection on the manifold. This family of curves forms a foliation of the manifold. The family of surfaces $E[X]$ =constant forms another foliation of the manifold dual to the first. For information on the importance of foliations in discussing basic statistical properties, see [11] and the references there.

The solution of the minimization problem of the preceding section, in this geometry, interpolates on this curve between the point $\mathbf{p}(0) = \mathbf{q}$ and the intersection of the curve with the surface whose equation is $E[X] = a$. The intersection occurs at the point $\mathbf{p}(\Lambda)$. The distance that one moves along the curve from $\mathbf{q}$ is determined by the degree of confidence one has in the side information (1).

4 Connection with maximum likelihood

We now consider an example for which there is enough additional information that the method of maximum likelihood can, and normally would, be applied. It does illustrate, though, that there is a value of the parameter λ for which the two methods yield the same result. It may be useful also as a guide in the choice of λ when this choice must be made on subjective grounds.

Suppose that M independent observations are made and that outcome i occurs f_i times, where $\sum f_i = M$. The maximum likelihood estimate of p_i given only these observations is known to be $q_i = f_i/M$. Now let L further independent experiments be performed. However, in these experiments, it is only possible to observe whether or not outcome 1 occurs. Let outcome 1 occur g times and some other outcome occur $L - g$ times. The likelihood of these outcomes is

$$(p_1)^{f_1+g}(p_2)^{f_2}\cdots(p_N)^{f_N}(1-p_1)^{L-g},$$

from which we obtain the maximum likelihood estimates

$$\tilde{p}_1 = \frac{q_1 M + aL}{L+M}, \tag{7}$$

and

$$\tilde{p}_k = \frac{q_k}{1-q_1}\frac{(1-q_1)M + (1-a)L}{L+M} \text{ for } k = 2, \ldots, N, \tag{8}$$

where we define $a = g/L$.

A natural way to use the proposed modification of the minimum cross-entropy method in this situation is to use $q_i = f_i/M$ as the initial guess, to let $x_1 = 1$ and $x_i = 0$ for $k = 2, \ldots, N$, and to use as the side information

$$E[X] = g/L = a.$$

As indicated earlier, one would use a small value of λ if L/M is small and a large value of λ if L/M is large. For the case $q_1 > a$, it is easily found that the values given by (6) are

$$p_1(\lambda) = \frac{q_1 e^{-\lambda}}{q_1 e^{-\lambda} + 1 - q_1},$$

and

$$p_k(\lambda) = \frac{q_k}{q_1 e^{-\lambda} + 1 - q_1} \text{ for } k = 2, \ldots, N,$$

and that

$$\Lambda = \ln\left(\frac{(1-a)q_1}{a(1-q_1)}\right).$$

The case $q_1 < a$ is entirely similar.

These values of $p_k(\lambda)$ are the same as the maximum likelihood estimates given by (7) and (8) if

$$\lambda = \ln\left[\frac{1 + \left(\frac{1-a}{1-q_1}\right)\frac{L}{M}}{1 + \frac{a}{q_1}\frac{L}{M}}\right] \quad (a < q_1). \tag{9}$$

Clearly λ given by (9) tends to zero as $L/M \to 0$ and $\lambda \to \Lambda$ as $L/M \to \infty$. Thus the proposed method can yield the maximum likelihood solution with the "correct" choice of λ. Perhaps as important, (9) could serve as a guide to a good choice of λ when there are other parameters that correspond in some way to L and M but where the maximum likelihood method cannot be used.

References

[1] S.. Amari, "Differential-geometric methods in statistics," *Lecture Notes in Statistics*, vol.28, 1985, Springer-Verlag.

[2] L. L. Campbell, "The relation between information theory and the differential geometry approach to statistics," *Information sciences*, vol.35, pp. 195–210. 1985.

[3] L. L. Campbell, "An extended Čencov characterization of the information metric," *Proc. Amer. Math. Soc.* vol. 98, pp. 135–141, Sept. 1986.

[4] L. L. Campbell, "Minimum relative entropy and Hausdorff dimension," *Internat. J. Math. & Stat. Sci.*, vol.1, pp. 35–46, June 1992.

[5] N. N. Čencov, *Statistical Decision Rules and Optimal Inference*, vol. 53 of *Translations of Mathematical Monographs* . Providence RI: Amer. Math. Soc., 1982.

[6] I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *Ann. Statist.* vol.19, pp. 2032–2066, 1991.

[7] I. Csiszár, T. M. Cover, and B. S. Choi, "Conditional limit theorems under Markov conditioning," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 788–801, Nov. 1987.

[8] I. Csiszár, F. Gamboa, and E. Gassiat, "MEM pixel correlated solutions for generalized moment and interpolation problems," *Proc. 1998 IEEE Internat. Symp. on Information Theory*, p. 357, 1998.

[9] G. J. Erickson and C. R. Smith, editors, *Maximum-Entropy and Bayesian Methods in Science and Engineering*, Boston: Kluwer Academic Publishers, 1988.

[10] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proc. IEEE*, vol. 70, pp. 939–952, Sept. 1982.

[11] S. L. Lauritzen, "Statistical manifolds," IMS Lecture Notes - Monograph Series, *Differential Geometry in Statistical Inference*, vol. 10, pp. 163–216, 1987.

[12] J. B. Paris and A. Vencovská, "A note on the inevitability of maximum entropy," *Internat. J. Approx. Reason.* vol. 4, pp. 183–223, 1990.

[13] J. E. Shore and R. W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 26–37, Jan. 1980.

[14] J. E. Shore and R. W. Johnson, "Properties of cross-entropy minimization," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 472–482, July 1981.

[15] J. M. Van Campenhout and T. M. Cover, "Maximum entropy and conditional probability," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 483–489, July 1981.