The Annals of Mathematical Statistics
1970, Vol. 41, No. 3, 1011–1015

EQUIVALENCE OF GAUSS'S PRINCIPLE AND MINIMUM DISCRIMINATION INFORMATION ESTIMATION OF PROBABILITIES[1]

BY L. L. CAMPBELL

Queen's University at Kingston

1. Introduction. Suppose that X is a random variable with unknown probability distribution function $F(x)$. Prior to an experiment, let the experimenter make the initial guess that the distribution function is $F_0(x)$. Now let the experimenter observe the number.

$$\hat{a} = n^{-1} \sum_{k=1}^{n} T(X_k), \tag{1.1}$$

where T is a known function, and $X_1, X_2, \cdots, X_n$ are the values of X at n independent trials. How should the experimenter modify his guess about $F(x)$ in order to take account of the new information provided by this number?

There are, of course, many possible answers to this question. In this paper, we examine two attractive answers and show that they lead to the same distribution. The first method, which is an extension of Gauss's derivation of the normal probability distribution ([1] page 107), assumes that the probability distribution F depends on the parameter

$$a = E[T(X)] = \int_{-\infty}^{\infty} T dF. \tag{1.2}$$

It is then assumed that the functional form of F is such that the arithmetic mean, $\hat{a}$, is also the maximum likelihood estimate of a, given the observations $X_1, X_2, \cdots, X_n$. Finally, it is assumed that F depends on the parameter a in such a way that there is a value, a^0, of a for which $F = F_0$. These assumptions, together with the assumption $a = \hat{a}$ determine F, under reasonable assumptions about F_0 and T. Hosszú and Vincze [6] have recently studied extensions of this method.

The second method which we consider consists in choosing the distribution function F which minimizes $I(F, F_0) = \int_{-\infty}^{\infty} \log(dF/dF_0)\, dF$ subject to the constraint that $\int_{-\infty}^{\infty} T dF = \hat{a}$. Here, dF/dF_0 is a Radon–Nikodym derivative. If F and F_0 have density functions f and f_0 respectively, then $dF/dF_0 = f/f_0$. If F and F_0 correspond to discrete probabilities p_i and q_i respectively, then $I(F, F_0) = \sum p_i \log p_i/q_i$.

This is the method of minimum discrimination information estimation of probabilities. If $f_0 = 1$ or $q_i = 1$, it becomes the maximum entropy method. Maximum entropy estimation of probabilities has been proposed for use in statistical mechanics [7], operations research [2], and in the formation of hypotheses [4]. Dutta [3], Good [5] and Jaynes [8] have recently considered some of

Received May 27, 1969; revised January 5, 1970.

[1] The research for this paper was supported in part by the Defense Research Board of Canada, Grant number 2801-29.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

the properties of maximum entropy estimates of probabilities. The functional $I(F, F_0)$ is Kullback's [9] "directed divergence," or Rényi's [12] "information gain."

As stated above, it will be shown that these two methods lead to the same guess about F. For discrete distributions and the case $T(x) = x$, Dutta [3] has obtained the same result. For continuous distributions where a is a location parameter, McBride [11] showed that Gauss's principle leads to density functions which are exponential functions; it is known [9] that exponential densities maximize entropy. In a somewhat different setting, Good [4] noted a connection between maximum likelihood estimates and maximum entropy estimates.

In the present paper we extend and unify these results. We allow T to be a vector-valued function and we do not necessarily assume that a is a location parameter. In addition, we examine the role of the initial guess, F_0, in a way which does not seem to have been done before, and we exhibit the directed divergence as a sort of potential function.

2. Gauss's principle. By Gauss's principle we shall mean that a distribution should be chosen so that the maximum likelihood estimate of the parameter a in (1.2) is the same as the arithmetic mean estimate given by (1.1). Let $T(x)$ be a vector $(T_1(x), T_2(x), \cdots, T_s(x))$, corresponding to s possible measurements which can be made at each trial and suppose that the functions T_j are differentiable. Let $\hat{a}$ be the vector defined by (1.1) and let the distribution function $F(x; a)$ depend on the vector parameter $a = (a_1, a_2, \cdots, a_s)$ in such a way that (1.2) holds. Let there be a value a^0 which is such that $F(x; a^0) = F_0(x)$, the initial guess.

We restrict our attention to the two cases for which maximum likelihood estimates are usually considered: the case where $F(x; a)$ has a density function $f(x; a)$ and the case where $F(x; a)$ is a step function. In the latter case, we assume that the range of values of the random variable X is some finite set $\{x^1, x^2, \cdots, x^m\}$ and that the probabilities are $P(X = x^k) = f(x^k; a)$. We assume, finally, that the function $f(x; a)$ has continuous mixed second partial derivatives with respect to x and the components of a.

If $X_1, \cdots, X_n$ are n independent observations, the maximum likelihood estimate of a satisfies the equations

$$\frac{\partial}{\partial a_i} \log L = 0, \qquad (i = 1, 2, \cdots, s),$$

where $\log L = \sum_{k=1}^{n} \log f(X_k; a)$. Put $\phi(x, a) = \log f(x; a)$ and denote partial derivatives by subscripts. The likelihood equation is then

$$\sum_{k=1}^{n} \phi_{a_i}(X_k, \hat{a}) = 0, \qquad (i = 1, 2, \cdots, s). \tag{2.1}$$

If we apply Gauss's principle to choose f, then we require that $\hat{a}$ in (1.1) and (2.1) be the same. Put another way, Gauss's principle implies that the $(n-s)$-dimensional hypersurface defined by the equations

$$\sum_{k=1}^{n} [T_i(x_k) - a_i] = 0, \qquad (i = 1, 2, \cdots, s), \tag{2.2}$$

is the same as the $(n-s)$-dimensional hypersurface defined by

$$\sum_{k=1}^{n} \phi_{a_i}(x_k, a_1, \cdots, a_s) = 0, \qquad (i = 1, 2, \cdots, s). \tag{2.3}$$

We assume here that $s < n$ and that the derivatives of the functions T_i are linearly independent. Then a normal to any of the $(n-1)$-dimensional surfaces (2.3) must be a linear combination of normals to the $(n-1)$-dimensional surfaces (2.2). That is, there are numbers μ_{ij} such that (equating kth components of normals)

$$\phi_{a_i x}(x_k, a_1, \cdots, a_s) = \sum_{j=1}^{s} \mu_{ij} T_j'(x_k), \qquad (i = 1, 2, \cdots, s;\ k = 1, 2, \cdots, n), \tag{2.4}$$

where T_j' is the derivative of T_j. Since μ_{ij} does not depend on k, (2.4) can be written

$$\phi_{a_i x}(x, a) = \sum_{j=1}^{s} \mu_{ij} T_j'(x), \qquad (i = 1, 2, \cdots, s),$$

where μ_{ij} is independent of x. We can now integrate these equations with respect to x and substitute back in (2.2) and (2.3) to evaluate the constants of integration. The result is

$$\phi_{a_i}(x, a) = \sum_{j=1}^{s} \mu_{ij}[T_j(x) - a_j], \qquad (i = 1, 2, \cdots, s). \tag{2.5}$$

In order that (2.5) be consistent, we must have $\phi_{a_i a_j} = \phi_{a_j a_i}$. This leads to the equations

$$\sum_{k=1}^{s} \left(\frac{\partial \mu_{ik}}{\partial a_j} - \frac{\partial \mu_{jk}}{\partial a_i} \right) [T_k(x) - a_k] - \mu_{ij} + \mu_{ji} = 0.$$

Since the derivatives T_k' form a linearly independent set of functions, the set of functions $\{T_1, \cdots, T_n, 1\}$ is linearly independent. Thus $\partial \mu_{ik}/\partial a_j = \partial \mu_{jk}/\partial a_i$ and $\mu_{ij} = \mu_{ji}$. The first of these equations implies the existence of functions $\lambda_k(a)$ which are such that $\mu_{ik} = \partial \lambda_k / \partial a_i$ and $\lambda_k(a^0) = 0$. The equations $\mu_{ij} = \mu_{ji}$ then imply the existence of some "potential" function V which is such that $\lambda_j = \partial V / \partial a_j$ and $V(a^0) = 0$. We can now write (2.5) in the form

$$\phi_{a_i}(x, a) = \frac{\partial}{\partial a_i} \left\{ \sum_{j=1}^{s} \lambda_j [T_j(x) - a_j] + V \right\}, \qquad (i = 1, 2, \cdots, s). \tag{2.6}$$

Each side of (2.6) is now expressed as a component of a gradient vector which we integrate from a^0 to a, getting

$$\phi(x, a) = \phi(x, a^0) + \lambda(a) \cdot [T(x) - a] + V(a)$$

or

$$f(x; a) = f(x; a^0) \exp\{\lambda(a) \cdot [T(x) - a] + V(a)\}, \tag{2.7}$$

where $\lambda \cdot [T - a]$ is the inner product of the vectors $T - a$ and $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_s) =$ grad V. Thus, if f satisfies Gauss's principle $f(x; a)/f(x; a^0)$ is an exponential function of $T(x)$. The parameters λ and V are determined by the conditions that f be a density function or a set of probabilities and that $E[T(x)] = a$.

One way to determine λ and V is to introduce the partition function Z by

$$Z(\lambda) = \int_{-\infty}^{\infty} f(x; a^0)\, e^{\lambda \cdot T(x)}\, dx \tag{2.8}$$

or

$$Z(\lambda) = \sum_{k=1}^{m} f(x^k; a^0)\, e^{\lambda \cdot T(x^k)}, \tag{2.9}$$

when F is respectively a continuous function or a step function. The normalization $\int f(x; a)dx = 1$ or $\sum f(x^k; a) = 1$ implies that

$$Z(\lambda) = e^{\lambda \cdot a - V}, \tag{2.10}$$

so that

$$f(x; a) = f(x; a^0)\, e^{\lambda \cdot T(x)}/Z(\lambda). \tag{2.11}$$

The condition $E[T(X)] = a$ now implies that $\lambda_1, \cdots, \lambda_s$ must be a solution of the equations

$$\frac{\partial Z}{\partial \lambda_i} = a_i Z(\lambda_1, \cdots, \lambda_s), \qquad (i = 1, 2, \cdots, s). \tag{2.12}$$

We do not examine here the conditions under which the integral in (2.8) converges or the system (2.12) has a unique solution.

It should be remarked that the assumption of differentiability of T can be relaxed. We do not attempt to obtain a very general result of this nature, but content ourselves with a discussion of one special case. If $s = 1$ and T is monotonic, with inverse function T^{-1}, let $y_k = T(x_k) - a$. Then (2.2) and (2.3) can be written

$$\sum_{k=1}^{n} y_k = 0$$

$$\sum_{k=1}^{n} \phi_a(T^{-1}(y_k + a), a) = 0.$$

This functional equation for ϕ_a has the solution ([1], page 47)

$$\phi_a(T^{-1}(y + a), a) = cy = c[T(x) - a],$$

which is (2.5) with $s = 1$.

Finally, we note that the potential function, $V(a)$, is just the directed divergence between $f(x; a)$ and $f(x; a^0)$. Let

$$I(F, F_0) = \int_{-\infty}^{\infty} f(x; a) \log \frac{f(x; a)}{f(x; a^0)}\, dx$$

or

$$I(F, F_0) = \sum_{k=1}^{m} f(x^k; a) \log \frac{f(x^k; a)}{f(x^k; a^0)}$$

when F is respectively a continuous function or a step function. It now follows from (2.7) and the fact that f is chosen to satisfy (1.2) that

$$I(F, F_0) = V(a). \tag{2.13}$$

3. Minimum discrimination information estimation. We now show that $f(x; a)$, defined by (2.7) or (2.11), minimizes $I(F, F_0)$ subject to the constraint (1.2). In fact, this result is an immediate consequence of the minimum discrimination information theorem of Kullback.

Let G be any other probability distribution which is such that $\int_{-\infty}^{\infty} T dG = a$. Then

$$\int_{-\infty}^{\infty} (\lambda \cdot T)\, dG = \lambda \cdot a, \tag{3.1}$$

where λ is a solution of (2.12). But the minimum discrimination information theorem [10] shows that $I(G, F_0) \geqq I(F, F_0)$ for any G satisfying (3.1).

Hence, the distribution, F, which was produced by an application of Gauss's principle is also the distribution which minimizes $I(F, F_0)$ subject to the constraint (3.1).

REFERENCES

[1] ACZEL, J. (1966). *Lectures on Functional Equations and their Applications*. Academic Press, New York.

[2] CLOUGH, D. J. (1964). Application of the principle of maximizing entropy in the formulation of hypotheses. *CORS J.* **2** 53–70.

[3] DUTTA, M. (1966). On maximum (information-theoretic) entropy estimation. *Sankhyā Ser. A.* **28** 319–328.

[4] Good, I. J. (1963). Maximum entropy for hypothesis formation, especially for multidimensional contingency tables. *Ann. Math. Statist.* **34** 911–934.

[5] GOOD, I. J. (1966). How to estimate probabilities. *J. Inst. Math. Appl.* **2** 364–383.

[6] HOSSZÚ, M. and VINCZE, E. (1963). Über den wahrscheinlichsten Wert. *Acta Math. Acad. Sci. Hungar.* **14** 131–136.

[7] JAYNES, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev.* **106** 620–630.

[8] JAYNES, E. T. (1968). Prior probabilities. *IEEE Trans. Systems Science and Cybernetics* **SSC-4** 227–241.

[9] KULLBACK, S. (1959). *Information Theory and Statistics*. Wiley, New York.

[10] KULLBACK, S. and KHAIRAT, M. A. (1966). A note on minimum discrimination information. *Ann. Math. Statist.* **37** 279–280.

[11] MCBRIDE, W. J. (1968). A natural law. *Proc. IEEE* **56** 1713–1715.

[12] RÉNYI, A. (1962). *Wahrscheinlichkeitsrechnung, mit einem Anhang uber Informationstheorie*. Deutscher Verlag der Wissenschaften, Berlin.