COWLES COMMISSION DISCUSSION PAPER: STATISTICS NO. 355

The Generalised Bayes-Minimax Principle: A Criterion for Decision-Making Under Uncertainty

by Leonid Hurwicz

February 8, 1951

1. Consider the individual I seeking to maximize the expected value $\mathcal{E} u_I$ of his utility u_I and suppose that u_I depends on a complex of circumstances c, so that $u_I = \varphi(c)$. $c \equiv (a, b)$ where a is defined as that component of c which is subject to I's free choice; I has no influence over the choice of b. If $c \in C$ we may regard C as the (Cartesian) product of the two sets A (= I's "decision domain") and B (= the "universe of discourse"), i.e., $C = A \times B$, $a \in A$, $b \in B$.

2. I's problem consists in choosing an "optimal" (in the sense of § 3 below) $\hat{a} \in A$, given the complete knowledge of φ, A, B and the information that $H(b) \in \mathcal{H}_B^{(o)} \subseteq \mathcal{H}_B$, $\mathcal{H}_B^{(o)} \neq 0$ where $\mathcal{H}_B^{(o)}$ is known. H(b) denotes a probability distribution in B.(1) $\mathcal{H}_B$ is the set of all H(b). 0 denotes the empty set. $\mathcal{H}_B^{(o)}$ will be called the "a pribri information area."

3. Definition: $\hat{a} \in A$ is said to be optimal relative to φ, A, B, and $\mathcal{H}_B^{(o)}$ if

$$\inf_{H(b)\in \mathcal{H}_B^{(o)}} \int_B \varphi(\hat{a}, b)\, dH(b) \geq \inf_{H(b)\in \mathcal{H}_B^{(o)}} \int_B \varphi(a, b)\, dH(b)$$

for all $a \in A$.

---

1. Among the elements of $\mathcal{H}_B^{(o)}$ we may find the "singular" distributions $\mathcal{E}(b - b_o)$ with $\text{Prob}(b = b_o) = 1$ (cf. H. Cramér, Mathematical Methods of Statistics 16.1).

4. The following special choices of $\mathcal{H}_B^{(o)}$ are of interest:

4.1. $\mathcal{H}_B^{(o)} = \mathcal{H}_B^{(oo)}$ where $\mathcal{H}_B^{(oo)}$ is the class of all "singular" (cf. footnote 1) distributions, i.e., all $\mathcal{E}(b - b_o)$, $b_o \in B$. Here the principle of optimality given in §3 above reduces to one of <u>simple</u> <u>"maxmin-ing"</u> with regard to B. I.e., here a is optimal if

$$\inf_{b \in B} \varphi(\hat{a}, b) \geq \inf_{b \in B} \varphi(a, b) \text{ for all } a \in A.$$

4.2. $\mathcal{H}_B^{(o)} = \mathcal{H}_B$. If A has the property that $a' \in A$, $a'' \in A$ implies that any randomized mixture of a', a'' also belongs to A, then this case, as is well known from the theory of games and Wald's work, yields the same solution a as that considered in § 4.1 above. Either this case or that in § 4.1 may be said to be that of <u>pure</u> <u>ignorance</u> (with regard to the given universe of discourse).

4.3. $\mathcal{H}_B^{(o)} = (H^{(o)}(b))$, i.e., $\mathcal{H}_B^{(o)}$ is a one-element family. According to the "subjective probability" school, as the writer understands it, this must always be the case, both descriptively and normatively. Here the problem reduces to that of <u>maximizing</u> $\int_B \varphi(a, b)\, dH^{(o)}(b)$ with regard to $a \in A$.

The solutions have been called <u>"Bayesian"</u> (or "Bayes optimal") with regard to $H^{(o)}(b)$.

4.3.1. When B is finite the choice of a one-element $\mathcal{H}_B^{(o)} = (\bar{H}^{(o)}(b))$ has been advocated where $\bar{H}^{(o)}(b)$ assigns equal probabilities to all elements b B. The traditional justification has been called the Principle of Insufficient Reason or the Bayes Postulate. [Not to be confused with the "Bayes optimal" solutions with regard to an arbitrary $H^{(o)}(b)$; cf. § 4.3 above.]

A new justification, based on certain axioms of "rational" behavior, has been given by Chernoff.[2]

---

2. Cowles Commission Discussion Paper, Statistics 326A, 346.

4.3.2. Another special case of a one-element $\mathcal{H}_B^{(o)}$ is that of $\mathcal{H}_B^{(o)} = (H^{(oo)}(b^o))$ where $H^{(oo)}(b^o)$ is "singular" [$\mathcal{E}(b - b_o)$] and assigns probability one to some $b^o \in B$. This is the case of <u>certainty</u>. Here the problem reduces to that of <u>maximizing</u> $\varphi(a, b^o)$ with regard to $a \in A$.

5. Let $B = B_1 \cup B_2$ where $B_1 \cap B_2 = 0$. Consider $\mathcal{H}_B^{(o)} = \mathcal{H}_{B_1}$ defined as the class of all distributions on B_1 (i.e., with zero probabilities for all subsets of B_2).

This is the typical restriction imposed in econometric models when the values of certain parameters are assumed known.

In this case it is clear that I's behavior would have been the same had he started with the set B_1 as the "universe of discourse."

6.1. We may formally reduce the problem to one of simple maxmin [i.e., "pure ignorance" of the § 4.1 type] by writing

$$\mathcal{E}u_I = \Psi(a, H).$$

If it is assumed known that $H \in \mathcal{H}_B^{(o)}$, the "optimal" solution is obtained by taking the supremum of $\inf_H \Psi$. I.e., we have

$$\inf_{H \in \mathcal{H}_B^{(o)}} \Psi(\hat{a}, H) \geq \inf_{H \in \mathcal{H}_B^{(o)}} \Psi(a, H) \text{ for all } a \in A.$$

Here $\mathcal{H}_B^{(o)}$ plays the role of the universe of discourse while the "a priori information area" consists of all the "singular" distributions (cf. § 4.1 above) over $\mathcal{H}_B^{(o)}$.

6.2. In the case of a one-element $\mathcal{H}_B^{(o)} = (H^{(o)}(b))$ [the general "Bayesian" case (cf. § 4.3 above)] the representation used in § 6.1 reduces the problem to the special "Bayesian" case of "certainty" (cf. § 4.3.2), since the "a priori information area" consists of a single element $H^{(o)}(b)$ of what is now the universe of discourse ($\mathcal{H}_B^{(o)}$).

Thus the problem reduces to that of maximizing $\Psi(a, H^o)$ with regard to $a \in A$.

7.0. <u>Examples.</u> In what follows three examples are given of cases which are neither simple maxmin or simple Bayes (unless, of course, the reformulation of § 6 is used). The first two are mainly designed to clarify the principle used; the third seems to be of some interest in application.

7.1. Let $-\varphi(d, b)$ be given as a matrix

	b_1	b_2
d_1	0	2.1
d_2	1	1

[Here the set A is given by all the "mixtures" of d_1 and d_2, i.e., all the values of, say, α_1, where $\alpha_1 \equiv \mathrm{Prob}(d = d_1)$, i.e., all points of the interval $0 \leqq \alpha_1 \leqq 1$.]

Write $\beta_j \equiv \mathrm{Prob}(b = b_j)$.

We may represent subsets of $\mathcal{N}_B$ in terms of sets of values of β_1. Thus $\mathcal{N}_B$ itself corresponds to the interval $0 \leqq \beta_1 \leqq 1$.

(1) Now if $\mathcal{N}_B^{(o)}$ is given by $0 \leqq \beta_1 \leqq 1$ (the case of "pure ignorance") we find that the optimal solution is $\hat{\alpha}_1 = 0$ (i.e., "pure" d_2 should be used).

(2) If $\mathcal{N}_B^{(o)}$ is a one-element family, the following cases are of interest:

(2.1) $\beta_1 = 1$; i.e., certainty that b_1 will occur; $\hat{\alpha}_1 = 1$ (i.e., "pure" d_1).

(2.2) $\beta_1 = \frac{1}{2}$; i.e., (equi-probability for b_1, b_2); $\hat{\alpha}_1 = 0$ (i.e., "pure" d_2).

(2.3) $\beta_1 = \frac{1}{3}$; $\hat{\alpha}_1 = 0$

(2.4) $\beta_1 = 0$; $\hat{\alpha}_1 = 0$.

(3) Let $\mathcal{N}_B^{(o)}$ be given by $0 \leqq \beta_1 \leqq \frac{1}{2}$; then $\hat{\alpha}_1 = 0$.

7.2. Let $-\varphi(d, b)$ be given as a matrix

	b_1	b_2	b_3
d_1	0	2	5
d_2	1	0	0

[Here again A is given by all the mixtures of d_1, d_2.]

Write, as before, $\alpha_i = \text{Prob}(d = d_i)$, $\beta_j = \text{Prob}(b = b_j)$.

Consider the following $\mathcal{H}_B^{(o)}$:

(1) $\beta_3' = 0$;

then $\hat{\alpha}_1' = \frac{1}{3}$

(2) $\beta_2'' = 0$;

then $\hat{\alpha}_1'' = \frac{1}{6}$;

(3) $\beta_2''' = \beta_3'''$;

then $\hat{\alpha}_1''' = \frac{2}{9}$.

It will be noted that $\hat{\alpha}_1'' < \hat{\alpha}_1''' < \hat{\alpha}_1'$.

7.3. An interesting opportunity for applying the principle stated in § 3 arises in connection with the following statistical problem:

Let there be given a sample of size n=1 from a normal bivariate universe, i.e., a pair X_1, X_2 of observations with the likelihood function

$$\frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left\{-\frac{1}{2}[\sigma^{11}(X_1 - \theta_1)^2 + 2\,\sigma^{12}(X_1 - \theta_1)(X_2 - \theta_2) + \sigma^{22}(X_2 - \theta_2)^2]\right\} .$$

The problem is to estimate θ_1 with Σ known.

A special feature of the problem is the assumption that a priori information concerning θ_2 is available. This information is of the following form: θ_2 is considered as a stochastic variable normally distributed with a known mean θ_2^* and a known variance ω_{22}^*.

We shall find a Bayes-minimax estimate $\tilde{\theta}_1^*$ of θ_1 when the weight function is $w = (\tilde{\theta}_1^* - \theta_1)^2$.

Here $\mathcal{H}_B$ consists of all bivariate distributions in θ_1, θ_2 while $\mathcal{H}_B^{(o)}$ consists of all bivariate distributions in θ_1, θ_2 such that the marginal distribution of θ_2 is that given above, i.e., $N(\theta_2 \mid \theta_2^*, \omega_{22}^*)$.

To obtain $\tilde{\theta}_1^*$ we work with a subset[3] $\mathcal{N}_B^{(o)}$ of $\mathcal{N}_B^{(o)}$ given by all bivariate normal distribution in θ_1, θ_2 such that $\mathcal{E}(\theta_1) = 0$ and θ_2 has the marginal distribution $N(\theta_2 | \theta_2^*, \omega_{22}^*)$.

First, let the joint distribution of θ_1, θ_2 be $H_{\omega_{11}}$ [bivariate normal with the respective means 0, θ_2^* and a diagonal covariance matrix $\Omega = \begin{pmatrix} \omega_{11} & 0 \\ 0 & \omega_{22}^* \end{pmatrix}$].

Then the (Bayes) optimal estimate relative to $H_{\omega_{11}}$ is

$$\tilde{\theta}^*_{1(\omega_{11})} = \frac{\begin{vmatrix} \omega_{11} & \sigma_{12} \\ 0 & \sigma_{22} + \omega_{22}^* \end{vmatrix}}{\begin{vmatrix} \sigma_{11} + \omega_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} + \omega_{22}^* \end{vmatrix}} X_1 + \frac{\begin{vmatrix} \sigma_{11} + \omega_{11} & \omega_{11} \\ \sigma_{21} & 0 \end{vmatrix}}{\begin{vmatrix} \sigma_{11} + \omega_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} + \omega_{22}^* \end{vmatrix}} (X_2 - \theta_2^*)$$

By letting $\omega_{11} \to \infty$ we obtain

$$\tilde{\theta}^*_{1(\infty)} = X_1 - \frac{\sigma_{21}}{\sigma_{22} + \omega_{22}^*}(X_2 - \theta_2^*) = X_1 - \left(\frac{\sigma_{21}}{\sigma_{22}}\right) \cdot \frac{1}{1 + \left(\frac{\omega_{22}^*}{\sigma_{22}}\right)} (X_2 - \theta_2^*).^{(4)}$$

We note that

$$\mathcal{E}(\theta^*_{1(\infty)} | \theta_2) = \theta_1 - \left(\frac{\sigma_{21}}{\sigma_{22}}\right) \cdot \frac{1}{1 + \left(\frac{\omega_{22}^*}{\sigma_{22}}\right)} (\theta_2 - \theta_2^*)$$

and

$$\mathcal{E}(\theta^*_{1(\infty)} - \theta_1)^2 = \sigma_{11} - \frac{1}{1 + \left(\frac{\omega_{22}^*}{\sigma_{22}}\right)} \left(\frac{\sigma_{12}}{\sigma_{22}}\right)^2 \sigma_{22}. \qquad (4)$$

Hence[5] $\tilde{\theta}^*_{1(\infty)}$ is a minimax estimate for $\mathcal{N}_B^{(o)}$ since the risk attached to it is independent of θ_1 and it is Bayes optimal (asymptotically as $\omega_{11} \to \infty$).

---

3. The procedure followed here is a modification of that used by Wald and Stein, Wolfowitz, Lehman and Hodges. See references [1] - [4] in Cowles Commission Discussion Paper, Statistics No. 352.

4. Letting ω_{22}^* become 0 and ∞, we obtain the earlier results in Cowles Commission Discussion Paper, Statistics No. 348, in particular equations (6), (8).

5. See Hodges and Lehman, Theorem 2.1, Annals of Mathematical Statistics, Vol. 21, No. 2, June 1950, p. 182.

8. The main point of the present paper is to show how situations characterized by a mixture of knowledge and ignorance can be handled. The method used is that of utilizing information available and reducing the problem to one of "pure ignorance" (over the "a priori information area" redefined as the new universe of discourse). The principle applied in the situation of pure ignorance happens to be that of maxmin (corresponding to the more usual minimax setup). However, any alternative principle of decision-making under ignorance that applies to a sufficiently broad class of cases could have been used instead. (Chernoff's principle mentioned in § 4.3.1 above is not eligible since it only applies to a finite B.)

Among alternatives may be mentioned the principle of minimising the maximum (over $\mathcal{H}_B^{(o)}$) regret and a class of criteria discussed in Cowles Commission Discussion Paper, Statistics No. 356.