The Annals of Statistics
1975, Vol. 3, No. 4, 1018–1024

CHARACTERIZATIONS OF PREDICTION SUFFICIENCY (ADEQUACY) IN TERMS OF RISK FUNCTIONS

BY KEI TAKEUCHI AND MASAFUMI AKAHIRA

University of Tokyo and University of Electro-Communications

Prediction sufficiency (adequacy), as it is usually defined in terms of conditional expectations, does imply "real" prediction sufficiency; i.e. sufficiency in terms of risk functions. The converse holds provided we permit the loss to depend on the unknown parameter. This is no longer true if we insist on loss functions which do not involve the unknown parameter. Conditional independence still holds but ordinary sufficiency may fail. If, however, we require equivalence of risk functions, then ordinary sufficiency and, consequently, prediction sufficiency follows.

1. Introduction. It has been shown by Bahadur [2] that sufficiency as defined in terms of conditional expectations, under regularity conditions, implies "real sufficiency" i.e. sufficiency in terms of risk functions. Furthermore it follows from Theorem 11.3 in Bahadur's paper [1] that prediction sufficiency is equivalent to ordinary sufficiency w.r.t. a larger class of probability measures. One may therefore expect similar results to hold for prediction sufficiency (adequacy) as well. In the case of prediction problems it may be of interest to consider loss functions which depend only on the decision to be made and the quantity to be predicted. If we insist on this restriction, then prediction sufficiency in terms of risk functions no longer implies prediction sufficiency as it is defined in terms of conditional expectations. Conditional independence holds but ordinary sufficiency may fail. It will, however, be shown that equivalence of risk functions implies ordinary sufficiency and consequently prediction sufficiency. (One of the authors proved this in an earlier work [7].)

We will, essentially, use the framework of Skibinsky [6]. The notion of adequacy in Skibinsky's paper is, however, replaced by the notion of prediction sufficiency.

2. Theorems. We shall assume that we are given a model consisting of a sample space $(\mathscr{X}, \mathscr{A})$ and a family $\{P_\theta : \theta \in \Theta\}$ of probability measures on $\mathscr{A}$. A sub σ-algebra $\mathscr{B}$ of $\mathscr{A}$ summarizes what can and what can not be observed. Similarily, a sub σ-algebra $\mathscr{C}$ of $\mathscr{A}$ describes what we are interested in predicting. Finally, we are given a sub σ-algebra $\mathscr{B}_0$ of $\mathscr{B}$ and our problem is to decide if anything is lost by basing our predictions on $\mathscr{B}_0$ rather than $\mathscr{B}$.

The prediction problem is assumed to be completely described by a decision space $(T, \mathscr{T})$, i.e. a measurable space and a loss function L from $\Theta \times \mathscr{X} \times T$ to $[0, \infty[$. It will always be assumed that L as a function on $\mathscr{X} \times T$ for given

Received January 1973; revised August 1974.

AMS 1970 *subject classifications*. Primary 62B05; Secondary 62C07.

Key words and phrases. Prediction sufficiency, conditional independence, equality of risk functions.

$\theta \in \Theta$ is $\mathscr{C} \times \mathscr{T}$ measurable. This implies that the loss does not depend on all of $x \in \mathscr{X}$, only on the part of x which is to be predicted.

A decision rule δ will here be defined as a Markov kernel $\delta(S \mid x)\colon S \in \mathscr{T}$, $x \in \mathscr{X}$ which is $\mathscr{B}$ measurable when S is fixed and a probability measure on $\mathscr{T}$ when x is fixed.

If δ is a decision rule then its performance characteristic $\mu_\delta(\cdot \mid \theta)$; $\theta \in \Theta$ may be defined by defining—for each θ — $\mu_\delta(\cdot \mid \theta)$ as the probability measure on $\mathscr{C} \times \mathscr{T}$ defined by

$$\mu_\delta(C \times S \mid \theta) = \int_C \delta(S \mid x) P_\theta(dx) .$$

The risk function $r_\delta(\theta)$; $\theta \in \Theta$ of a decision rule δ is given by:

$$r_\delta(\theta) = \int [\int L_\theta(x, t)\delta(dt \mid x)] P_\theta(dx) , \qquad \theta \in \Theta .$$

The risk function is determined by the loss function and the performance characteristic through

$$r_\delta(\theta) = \int L_\theta \, d\mu_\delta(\cdot \mid \theta) , \qquad \theta \in \Theta .$$

A decision rule δ will be called $\mathscr{B}_0$ measurable if $\delta(S \mid \cdot)$ is $\mathscr{B}_0$ measurable for each S.

DEFINITION 1. $\mathscr{B}$ and $\mathscr{C}$ are conditionally independent given $\mathscr{B}_0$ iff

(i) $P_\theta^{\mathscr{C}}(C \mid \mathscr{B}) = P_\theta^{\mathscr{C}}(C \mid \mathscr{B}_0)$ a.e. $[P_\theta]$ for all $C \in \mathscr{C}$ and for all $\theta \in \Theta$.

It is shown in Loève [5], page 351 that (i) and (ii) are equivalent:

(ii) $P_\theta^{\mathscr{A}}(B \cap C \mid \mathscr{B}_0) = P_\theta^{\mathscr{A}}(B \mid \mathscr{B}_0) P_\theta^{\mathscr{A}}(C \mid \mathscr{B}_0)$ a.e. $[P_\theta]$

for all $B \in \mathscr{B}$ and all $C \in \mathscr{C}$ and for all $\theta \in \Theta$.

We define prediction sufficiency (adequacy) and prediction sufficiency in the wide sense as follows:

DEFINITION 2. $\mathscr{B}_0$ is prediction sufficient for $\mathscr{B}$ w.r.t. $\mathscr{C}$ iff $\mathscr{B}_0$ is sufficient for $\mathscr{B}$ and $\mathscr{B}$ and $\mathscr{C}$ are conditionally independent given $\mathscr{B}_0$.

DEFINITION 3. $\mathscr{B}_0$ is prediction sufficient in the wide sense for $\mathscr{B}$ w.r.t. $\mathscr{C}$ iff (a) $\mathscr{B}$ and $\mathscr{C}$ are conditionally independent given $\mathscr{B}_0$ and (b) there exist $\mathscr{B}_0$-measurable sets B_1 and B_2 so that $P_\theta(B_1 \cup B_2) = 1$ for all $\theta \in \Theta$ and $P_{\theta,x}^{\mathscr{C}}(\cdot \mid \mathscr{B}_0)$ is independent of θ if $x \in B_1$ and $P_{\theta,x}^{\mathscr{C}}(\cdot \mid \mathscr{B}_0)$ is independent of θ if $x \in B_2$.

In the following example we shall show that $\mathscr{B}_0$ is prediction sufficient in the wide sense for $\mathscr{B}$ w.r.t. $\mathscr{C}$ but not prediction sufficient for $\mathscr{B}$ w.r.t. $\mathscr{C}$.

We assume that $X_1, X_2, \cdots, X_n$ and Y are random variables such that $X_1, X_2, \cdots, X_n$ are independently and identically distributed as $N(\theta, 1)$ while the conditional distribution of Y given $X_1, X_2, \cdots, X_n$ is $N(0, 1)$ or $N(\theta, 1)$ as $\sum_i X_i > a$ or $\sum_i X_i \leqq a$. Let $\mathscr{B}$, $\mathscr{B}_0$ and $\mathscr{C}$ be the σ-algebras induced by, respectively, $(X_1, X_2, \cdots, X_n)$, $\min(a, \sum_i X_i)$ and Y. Then $\mathscr{B}_0$ is prediction sufficient in the wide sense for $\mathscr{B}$ w.r.t. $\mathscr{C}$ but not sufficient for $\mathscr{B}$.

That sufficiency alone is insufficient in prediction problems may be seen by considering, for example, the situation where P_θ does not depend on θ and

$\mathscr{B}_0 = \{\phi, \mathscr{X}\}$. It is then fairly obvious that prediction of a $\mathscr{C}$ which is not independent of $\mathscr{B}$ should not, in general, be based on $\mathscr{B}_0$.

It follows, as has been pointed out by Skibinsky [6], from Theorem 11.3 in Bahadur [1] that $\mathscr{B}_0$ is prediction sufficient for $\mathscr{B}$ w.r.t. $\mathscr{C}$ if and only if $\mathscr{B}_0$ is sufficient for $\mathscr{B}$ w.r.t. all probability measures on $\mathscr{B}$ of the form:

$$B \cap\!\!\rightarrow P_\theta(B \mid C) ,$$

where $C \in \mathscr{C}$, $\theta \in \Theta$ and $P_\theta(C) > 0$.

In analogy with Theorem 10.2 in Bahadur [1] we get:

THEOREM 1. *Suppose $\mathscr{B}_0$ is prediction sufficient for $\mathscr{B}$* w.r.t. *$\mathscr{C}$ and there exists a regular conditional probability $P^{\mathscr{B}}(\cdot \mid \mathscr{B}_0)$ of $\mathscr{B}$ given $\mathscr{B}_0$ which does not depend on θ. Let δ be any decision rule from $(\mathscr{X}, \mathscr{B})$ to $(T, \mathscr{T})$ and put $\tilde{\delta}(S \mid x) = \int \delta(S \mid x') P_x^{\mathscr{B}}(dx' \mid \mathscr{B}_0)$; $S \in \mathscr{T}$, $x \in \mathscr{X}$.*

Then $\tilde{\delta}$ is $\mathscr{B}_0$ measurable and it has the same performance characteristic as δ. In particular δ and $\tilde{\delta}$ have the same risk functions.

PROOF.

$$\begin{aligned}\mu_\theta(C \times S \mid \tilde{\delta}) &= \int_C \tilde{\delta}(S \mid x) P_\theta(dx) = \int_C E^{\mathscr{B}}(\delta(S \mid \cdot) \mid \mathscr{B}_0)\, dP_\theta \\ &= \int_C E^{\mathscr{B}}(\delta(S \mid \cdot) \mid \mathscr{B}_0, \mathscr{C})\, dP_\theta = \int_C \delta(S \mid \cdot)\, dP_\theta = \mu_0(C \times S \mid \delta) .\end{aligned}$$

REMARK. As is immediately seen from above, we need to allow randomized decision rules. This is not always necessary for the subsequent discussions.

Consequences of "risk prediction sufficiency" for various classes of loss functions. In order to show prediction sufficiency of $\mathscr{B}_0$ we must establish conditional independence and ordinary sufficiency. We will assume that we are given a certain class of loss functions and that to any loss function within that class and to any decision rule δ corresponds a decision rule $\tilde{\delta}$ which is $\mathscr{B}_0$ measurable and has uniformly smaller risk than δ. The problem is to decide whether this suffices to establish conditional independence or ordinary sufficiency. It is clear that conditional independence cannot, in general, be established by only considering loss functions which do not depend on x. Similarly loss functions which do not involve θ will, in general, be insufficient to establish ordinary sufficiency.

Conditional independence may, however, be established by considering only loss functions which do not depend on θ.

Similarly, and this follows from corresponding facts for sufficiency (see Bahadur [2], Blackwell [3] and Le Cam [4]), sufficiency of $\mathscr{B}_0$ may be established by considering loss functions which do not depend on x.

Conditional independence may be established by considering the two decision problem with loss functions not depending on θ as follows:

THEOREM 2. *Consider the decision space $T = \{0, 1\}$ and the set of all loss functions, L, of the form*

$$L_\theta(x, 0) = I_C(x) , \qquad x \in \mathscr{X},\ \theta \in \Theta ,$$

$$L_\theta(x, 1) = p I_{C^c}(x) , \qquad x \in \mathscr{X},\ \theta \in \Theta ,$$

where $p \in]0, 1[$ and $C \in \mathscr{C}$.

Suppose that to each decision rule δ and to each loss function L, of the above form, there corresponds a $\mathscr{B}_0$ measurable decision rule $\tilde{\delta}$ so that

$$r_{\tilde{\delta}}(\theta) \leqq r_\delta(\theta), \qquad \theta \in \Theta.$$

Then $\mathscr{B}$ and $\mathscr{C}$ are conditionally independent given $\mathscr{B}_0$.

Before proving the theorem a few remarks may be in order.

REMARK 1. It follows from the proofs that we may restrict attention to non-randomized decision rules.

REMARK 2. The proofs imply also that much smaller sets of loss functions will do. We may, for example, restrict C to a π-system generating $\mathscr{C}$.

REMARK 3. The parameter space Θ does not play any role in this theorem. We may—and shall—in the proof assume that Θ consists of a single point. Conditional independence is, in this situation, equivalent to prediction sufficiency.

PROOF OF THE THEOREM. We may, by Remark 3, omit the subscript θ. Furthermore a decision rule δ may be identified with the critical function $x \curvearrowright \delta(1 \mid x)$. The risk may then be written:

$$\begin{aligned} r(\delta) &= \int L(\cdot, 0)\, dP + \int [L(\cdot, 1) - L(\cdot, 0)]\delta\, dP \\ &= \int L(\cdot, 0)\, dP + (p+1) \int \left(\frac{p}{p+1} - I_C\right) \delta\, dP \\ &= \int L(\cdot, 0)\, dP + (p+1) \int \left[\frac{p}{p+1} - P^{\mathscr{C}}(C \mid \mathscr{B})\right] \delta\, dP \\ &\geqq \int L(\cdot, 0)\, dP - (p+1) \int \left[P^{\mathscr{C}}(C \mid \mathscr{B}) - \frac{p}{p+1}\right]^+ dP, \end{aligned}$$

where "$=$" is obtained iff $\delta = 0$ a.e. or $\delta = 1$ a.e. as $P^{\mathscr{C}}(C \mid \mathscr{B}) < p/(p+1)$ or $P^{\mathscr{C}}(C \mid \mathscr{B}) > p/(p+1)$.

The same argument applied to $\mathscr{B}_0$ implies, by the assumption of the theorem, that the minimizing δ may be chosen $\mathscr{B}_0$-measurable and such that $\delta = 0$ a.e. or $\delta = 1$ a.e. as $P^{\mathscr{C}}(C \mid \mathscr{B}_0) < p/(p+1)$ or $P^{\mathscr{C}}(C \mid \mathscr{B}_0) > p/(p+1)$. It follows that the event $[P^{\mathscr{C}}(C \mid \mathscr{B}) < p/(p+1)]$ and the event $[P^{\mathscr{C}}(C \mid \mathscr{B}_0) < p/(p+1)]$ are equivalent provided $P^{\mathscr{C}}(C \mid \mathscr{B}) \neq p/(p+1)$ a.e. and $P^{\mathscr{C}}(C \mid \mathscr{B}_0) \neq p/(p+1)$ a.e. This implies that the random variables $P^{\mathscr{C}}(C \mid \mathscr{B})$ and $P^{\mathscr{C}}(C \mid \mathscr{B}_0)$ have the same distribution. Hence, since $P^{\mathscr{C}}(C \mid \mathscr{B}_0) = E^{\mathscr{C}}(P(C \mid \mathscr{B}) \mid \mathscr{B}_0)$: $P^{\mathscr{C}}(C \mid \mathscr{B}_0) = P^{\mathscr{C}}(C \mid \mathscr{B})$ a.e. It follows that $\mathscr{B}$ and $\mathscr{C}$ are conditionally independent given $\mathscr{B}_0$.

REMARK. This form of the proof was suggested by one of the referees.

A criterion based on least squares prediction theory is, as has been pointed out by one of the referees, even simpler to establish. Consider a sufficiently large class of square integrable and $\mathscr{C}$-measurable random variables g. To a

given g we associate the loss function

$$L(x, t) = (g(x) - t)^2, \qquad x \in \mathscr{X},\ t \in]-\infty, \infty[.$$

Then a predictor δ minimizes the risk if and only if it is a version of $E^{\mathscr{C}}(g \mid \mathscr{B})$. If $\mathscr{B}_0$ is assumed to be just as good in this situation then $E^{\mathscr{C}}(g \mid \mathscr{B}) = E^{\mathscr{C}}(g \mid \mathscr{B}_0)$ a.e. This establishes conditional independence if, for example, we admit all functions $g = I_C$ where C runs through a π-system generating $\mathscr{C}$.

For the Lemma and Theorem 3 we assme that $\{P_\theta : \theta \in \Theta\}$ is dominated. Then we get the following lemma.

LEMMA. *If for any $\mathscr{B}$-measurable critical function φ there exists a $\mathscr{B}_0$-measurable critical function ψ such that $E_\theta(\psi) = E_\theta(\varphi)$ for all $\theta \in \Theta$, then $\mathscr{B}_0$ is sufficient for $\mathscr{B}$.*

The proof of the lemma is essentially the same as in Bahadur [2]. The outline is as follows: Let θ_1 and θ_2 be any two points of Θ. Let ϕ be a most powerful test for θ_1 against θ_2. Then for some k

$$\phi(x) = 1 \quad \text{if} \quad \frac{dP_{\theta_2}}{dP_{\theta_1}} > k ,$$

$$= 0 \quad \text{if} \quad \frac{dP_{\theta_2}}{dP_{\theta_1}} < k ,$$

and the set $\{x : (dP_{\theta_2}/dP_{\theta_1})(x) < k\}$ is $\mathscr{B}_0$-measurable. Further for every c (including ∞), the set $\{x : (dP_{\theta_2}/dP_{\theta_1})(x) < c\}$ is $\mathscr{B}_0$-measurable. Hence $\mathscr{B}_0$ is pairwise sufficient for $\mathscr{B}$. Since $\{P_\theta : \theta \in \Theta\}$ is dominated, $\mathscr{B}_0$ is sufficient for $\mathscr{B}$.

The following proposition is an immediate consequence of the lemma:

Suppose that the decision space $T = [0, 1]$ and the loss function L satisfies $L(x, t) = t$, for $0 \leqq t \leqq 1$. If for any $\mathscr{B}$-measurable decision rule δ, there exists a $\mathscr{B}_0$-measurable decision rule $\tilde{\delta}$ such that $r_{\tilde{\delta}}(\theta) = r_\delta(\theta)$ for all $\theta \in \Theta$, then $\mathscr{B}_0$ is sufficient for $\mathscr{B}$.

From the above, we get the following theorem.

THEOREM 3. *If for any loss function L not depending on θ and for any $\mathscr{B}$-measurable decision rule δ, there exists a $\mathscr{B}_0$-measurable decision rule $\tilde{\delta}$ such that $r_{\tilde{\delta}}(\theta) = r_\delta(\theta)$ for all $\theta \in \Theta$, then $\mathscr{B}_0$ is prediction sufficient for $\mathscr{B}$* w.r.t. *$\mathscr{C}$.*

Various criteria for prediction sufficiency may be obtained from these results by considering, in addition to the loss functions described above, loss functions which depend on θ. If we insist on considering only loss functions which do not depend on θ then we will, in general, not be able to conclude prediction sufficiency. This follows by considering the case where $\mathscr{B}_0 = \{\phi, \mathscr{X}\}$, $\mathscr{C}$ is ancillary and independent of $\mathscr{B}$ and $\{P_\theta : \theta \in \Theta\}$ is finite.

Let, in this situation, $(T, \mathscr{T})$ be any decision space and L any loss function which does not depend on θ and δ any decision rule from $(\mathscr{X}, \mathscr{B})$ to $(T, \mathscr{T})$.

Choose a $t_0 \in T$ such that

$$\int L(x, t_0)P_\theta(dx) \leqq \int [\int L(x, t)P_\theta(dx)]\nu_\theta(dt) , \qquad \theta \in \Theta ,$$

where $\nu_\theta(S) = \int \delta(S \mid x) P_\theta(dx)$. Then

$$r_\theta(t_o) \leqq r_\theta(\delta) , \qquad \theta \in \Theta .$$

The necessary condition for prediction sufficiency in the wide sense is obtained as follows:

THEOREM 4. *Suppose that, for each θ, there are regular conditional probabilities $P_\theta(B \mid \mathscr{B}_o) : B \in \mathscr{B}$ and $P_\theta(C \mid \mathscr{B}_o) : C \in \mathscr{C}$. Suppose further that $\mathscr{B}_o$ is prediction sufficient in the wide sense for $\mathscr{B}$* w.r.t. *$\mathscr{C}$.*

Let L be a loss function not depending on θ and assume that there is a $\mathscr{B}_o$-measurable function τ on B_2 and an $\varepsilon \geqq 0$ so that

$$\int L(x', \tau) P(dx' \mid \mathscr{B}_o) \leqq \int L(x', t) P(dx' \mid \mathscr{B}_o) + \varepsilon , \qquad t \in T \text{ on } B_2 .$$

Then there corresponds to any decision rule δ a $\mathscr{B}_o$-measurable decision rule $\tilde{\delta}$ so that

$$r_\theta(\tilde{\delta}) \leqq r_\theta(\delta) + \varepsilon , \qquad \theta \in \Theta .$$

$\tilde{\delta}$ may be defined as τ on B_2 and as $E(\delta(\cdot \mid \cdot) \mid \mathscr{B}_o)$ on B_1.

REMARK. There exist, for each $\varepsilon > 0$, a $\mathscr{B}_o$-measurable τ on B_2 satisfying the desired inequality provided

(i) There exists a countable subset $\{t_1, t_2, \cdots, t_n, \cdots\}$ of T such that for all $t \in T$ and for all $x \in B_2$

$$\inf_n \int L(x', t_n) P_x^{\mathscr{C}}(dx' \mid \mathscr{B}_o) \leqq \int L(x', t) P_x^{\mathscr{C}}(dx' \mid \mathscr{B}_o) ,$$

and

(ii) For every pair i, j and for any $\varepsilon > 0$ the set

$$M_{ij}(\varepsilon) = \{x : \int L(x', t_i) P_x^{\mathscr{C}}(dx' \mid \mathscr{B}_o) < \int L(x', t_j) P_x^{\mathscr{C}}(dx' \mid \mathscr{B}_o) + \varepsilon\}$$

is measurable.

PROOF OF THE THEOREM. Put, for each θ,

$$\delta_\theta^*(S \mid \cdot) = \int \delta(S \mid x') P_\theta(dx' \mid \mathscr{B}_o) .$$

By the proof of Theorem 1:

$$\begin{aligned}
\varepsilon + r_\theta(\delta) &= \varepsilon + \int [\int L(x, t) \delta_\theta^*(dt \mid x)] P_\theta(dx) \\
&= \int_{B_1} [\int L(x, t) \tilde{\delta}(dt \mid x)] P_\theta(dx) \\
&\qquad + \varepsilon + \int_{B_2} \{\int [\int L(x', t) P_x(dx' \mid \mathscr{B}_o)] \delta_\theta^*(dt \mid x)\} P_\theta(dx) \\
&\geqq \int_{B_1} [\int L(x, t) \tilde{\delta}(dt \mid x)] P_\theta(dx) \\
&\qquad + \int_{B_2} \{\int L(x', \tilde{\delta}(x)) P_{\theta, x}(dx' \mid \mathscr{B}_o)\} P_\theta(dx) \\
&= \int_{B_1} [\int L(x, t) \tilde{\delta}(dt \mid x)] P_\theta(dx) + \int_{B_2} L(x, \tilde{\delta}(x)) P_\theta(dx) \\
&= r_\theta(\tilde{\delta}) .
\end{aligned}$$

Acknowledgments. The authors wish to thank Mr. M. Takahashi of Osaka University for valuable suggestions and the referees of the *Annals* for co-operation in completing the final version.

REFERENCES

[1] BAHADUR, R. R. (1954). Sufficiency and statistical decision functions. *Ann. Math. Statist.* **25** 423–462.

[2] BAHADUR, R. R. (1955). A characterization of sufficiency. *Ann. Math. Statist.* **26** 286–293.

[3] BLACKWELL, D. (1953). Equivalent comparisions of experiments. *Ann. Math. Statist.* **24** 265–272.

[4] LE CAM, L. (1964). Sufficiency and approximate sufficiency. *Ann. Math. Statist.* **35** 1419–1455.

[5] LOÈVE, M. (1963). *Probability Theory*, (3rd. ed.). Van Nostrand, Princeton.

[6] SKIBINSKY, M. (1967). Adequate subfields and sufficiency. *Ann. Math. Statist.* **38** 155–161.

[7] TAKEUCHI, K. (1966). On some statistical prediction procedures. (In Japanese). *Keizaigaku Ronshu* **32** 23–31.

FACULTY OF ECONOMICS
UNIVERSITY OF TOKYO
HONGO, BUNKYO-KU
TOKYO, JAPAN

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF ELECTRO-COMMUNICATIONS
CHOFUGAOKA, CHOFU-SHI
TOKYO, JAPAN