STATISTICAL RESEARCH REPORT No 7
Institute of Mathematics november 1975
University of Oslo

PREDICTION SUFFICIENCY WHEN THE LOSS FUNCTION DOES NOT DEPEND ON THE UNKNOWN PARAMETER.

by

Erik N. Torgersen

Summary

It is shown by Takeuchi and Akahira, 1974, that conditional independence together with a condition of "partial sufficiency" imply "prediction sufficiency" for loss functions not depending on the unknown parameter. We shall here prove that these conditions are necessary as well and thereby obtain a complete description, in terms of conditional expectations, of "prediction sufficiency" for loss functions not depending on the unknown parameter. It turns out that these conditions may be replaced by a condition of conditional independence for prior distributions.

Introduction. Consider the problem of taking a decision t on the basis of our observations X when the loss is determined by t and a non observable variable Y . Consider also a function X_0 of X . It will be assumed that the joint distribution of X and Y is determined by an unknown parameter θ . We are also assuming that the merrit, or the lack of it, of any procedure is to be judged solely on the expected loss, i.e. risk, it incures.

In this context the problem of sufficiency may, somewhat loosely, be phrased: When are we justified in claiming that no information is lost by basing ourselves on X_0 rather than on all of X ? Note that the situation where the loss is determined by t and θ may be regarded as the particular case where $\Pr(Y=\theta|\theta) = 1$ for all θ .

It should be stated at once that we are in this introduction wilfully omitting several qualifications. A rigoruos treatment will be given in the next section.

In order to clarify the scope of this paper, let us for a moment consider the more general situation where the loss depend on θ as well as on t and Y . Considering a non negative function L of (θ,t,Y) as a loss function, we may say that X_0 is L-sufficient for X w.r.t. Y if the set of decision rules based on X_0 is essentially complete.

By theorem 1 in Takeuchi and Akahira [5] (See also theorem 10.2 in Bahadur [1]) , X_0 is L-sufficient for X w.r.t. Y

provided:

C_1: X_0 is sufficient for X

C_2: X and Y are conditionally independent given X_0 for all θ.

If these conditions are satisfied then, following Takeuchi and Akahira [5 page 1019] we shall say that X_0 is prediction sufficient for X w.r.t. Y. This corresponds to X_0 being adequat for X w.r.t. Y in Skibinsky's [4 page 156] terminology.

That prediction sufficiency implies L-sufficiency for any L may be seen directly by a randomization argument. A statistician knowing X_0 only may, by a random mechanisme" construct another variable $\tilde{X}$ so that $(\tilde{X},Y)$ has the same distribution as (X,Y). [Let U be rectangularily distributed on $[0,1]$ and independent of (X,Y). Then there are, for each x_0 in the range of X_0, a function φ_{x_0} so that the distribution of $\varphi_{x_0}(U)$ is equal to the conditional distribution of X given $X_0=x_0$. It is easily checked that we may take $\tilde{X} = \varphi_{X_0}(U)$].

In their paper [5], Takeuchi and Akahira proved that L-sufficiency for sufficiently many loss functions L implies prediction sufficiency. If, however, we restrict attention to loss functions which do not depend on θ then they found that C_1 could be weakened to:

$\bar{C}_1$: There is a set B so that the conditional distribution of X given X_0 does not depend on θ when $X_0 \in B$ while the conditional distribution of Y given X_0 does not depend on θ when $X_0 \notin B$.

Roughly the argument in [5] runs as follows: Let the loss L be determined by Y and the decision taken, and let δ be a decision rule based on X. Choose $\tilde{\delta} = E\delta|X_0$ when $X_0 \in B$ and such that $EL|X_0$ is small when $X_0 \notin B$ and $\tilde{\delta}$ is used. Then, with obvious notations: $E_{\tilde{\delta}}(L|X_0) = E_\delta(L|X_0)$ when $X_0 \in B$ and $E_{\tilde{\delta}}(L|X_0)$ is not much larger than $E_\delta(L|X_0)$ when $X_0 \notin B$. As a particular case consider prediction with squared error loss of some square integrable real valued function Y_0 of X. If $g(X)$ is any predictor with finite risk then $\tilde{g}(X_0)$ given by:

$$\tilde{g}(X_0) = \begin{cases} Eg(X)|X_0 & \text{when } X_0 \in B \\ EY_0|X_0 & \text{when } X_0 \notin B \end{cases}$$

is at least as good.

Consider now a fixed, finite and non trivial decision space T. Denote by $\mathcal{L}$ the class of loss functions $L = L(Y,t)$ which depends only on Y and the decision taken. If X_0 is L-sufficient for X w.r.t Y for all $L \in \mathcal{L}$ then we shall say that X_0 is $\mathcal{L}$-sufficient for X w.r.t. Y.

We shall see in the next section that the conditions C_1 and C_2 can't be reduced without violating $\mathcal{L}$-sufficiency. Situations where we do have $\mathcal{L}$-sufficiency may thus be classified according to the set B appearing in condition $\bar{C}_1$. Prediction sufficiency corresponds to the case where B may be chosen as the whole range of X_0. If the conditional distribution of Y given X depends on (X,θ) only through X_0, then $\bar{C}_1$ and C_2 holds with $B = \emptyset$. As an example of the intermediate situation consider random variables X and Y whose joint distribution is given by the following table of $Pr(X = x, Y = y|\theta)$:

y \ x	1	2	3
1	$(1-\alpha_\theta)(1-\beta)\tau_\theta$	$(1-\alpha_\theta)(1-\beta)(1-\tau_\theta)$	$\alpha_\theta(1-\gamma_\theta)$
2	$(1-\alpha_\theta)\ \beta\tau_\theta$	$(1-\alpha_\theta)\beta(1-\tau_\theta)$	$\alpha_\theta\gamma_\theta$

Here α, γ and τ are functions from Θ to $[0,1]$ while $\beta \in [0,1]$ is a constant. Simple calculations show that $X_0 = \max(X,2)$ is $\mathcal{L}$-sufficient for X w.r.t. Y ; i.e. $\bar{C}_1$ and C_2 are satisfied. X_0 is, however, not prediction sufficient for X w.r.t Y unless τ is constant on $\{\theta : \alpha_\theta < 1\}$.

$\mathcal{L}$-sufficiency is closely related to conditional independence for prior distributions. It will be shown that X_0 is $\mathcal{L}$-sufficient for X w.r.t. Y if and only if X and Y are conditionally independent for all prior distributions with finite support. Actually it suffices to consider the prior distributions which are either degenerate or uniform two point distributions. Utilizing this we prove the existence of "minimum" $\mathcal{L}$- sufficient statistics.

As is usual in this type of discussion, the functional form of the random variables is of minor importance. We shall therefore express our results in terms of algebras of events rather than in terms of random variables.

2. Sufficiency and conditional independence.

Our discussion will be carried out within the following framework. There is given a family $(\chi, \mathcal{A}, P_\theta) : \theta \in \Theta$ of probability spaces and three sub σ algebras, $\mathcal{B}_0$, $\mathcal{B}$ and $\mathcal{C}$, of $\mathcal{A}$. The set Θ is the parameter set of our model. It will be assumed that $\mathcal{B}_0 \subseteqq \mathcal{B}$ and that $\{P_\theta : \theta \in \Theta\}$ is dominated.

Referring to the introduction, $\mathcal{B}_0$, $\mathcal{B}$ and $\mathcal{C}$ may be interpreted as the σ algebras of events induced by, respectively, X_0, X and Y.

We will also assume that we are given a finite set T, with at least two elements, containing all possible decisions.

A decision rule δ is a family $\delta_t : t \in T$ of non negative measurable variables such that $\sum_t \delta_t = 1$. The interpretation of δ is the usual; i.e. $\delta_t(x)$ is the probability of taking decision t given that we have observed x.

A loss function is a non negative function on $\Theta \times \chi \times T$ which is $\mathcal{C}$ measurable in x for fixed (θ, t) in $\Theta \times T$. Denote by $\mathcal{L}$ the class of loss functions which does not depend on θ.

The risk function r_δ of a decision rule δ w.r.t. a loss function L is given by

$$r_\delta(\theta) = E_\theta \sum_t L_\theta(\cdot, t)\delta_t$$

where E_θ denotes expectation w.r.t. P_θ.

The set of all prior distributions on Θ with finite support will be denoted by Λ. The sub set of Λ consisting of the prior distributions which are either degenerate or uniform two point distributions will be denoted by Λ_0.

If $\lambda \in \Lambda$ then $P_\lambda = \sum_\theta \lambda_\theta P_\theta$ and $E_\lambda = \sum_\theta \lambda_\theta E_\theta$.

By Halmos and Savage [2] there is a non negative function c on Θ so that $\Theta_0 = \{\theta : c(\theta) > 0\}$ is countable, $\sum_\theta c(\theta) = 1$ and $\pi = \Sigma c(\theta) P_\theta$ dominates $\{P_\theta : \theta \in \Theta\}$. Put for each $\theta \in \Theta$ and each $\lambda \in \Lambda$, $f_\theta = dP_\theta/d\pi$ and $f_\lambda = dP_\lambda/d\pi$. Expectation w.r.t. π will be denoted by π.

We shall say that $\mathcal{B}_0$ is $\mathcal{L}$-sufficient for $\mathcal{B}$ w.r.t $\mathcal{C}$ if to each loss function L in $\mathcal{L}$ and each decision rule δ corresponds a $\mathcal{B}_0$ measurable decision rule $\tilde{\delta}$ such that:

$$r_{\tilde{\delta}}(\theta) \leqq r_\delta(\theta) ; \quad \theta \in \Theta .$$

Criterions for $\mathcal{L}$-sufficiency are collected in

Theorem

The following conditions are equivalent:

(i) $\mathcal{B}_0$ is $\mathcal{L}$-sufficient for $\mathcal{B}$ w.r.t. $\mathcal{C}$

(ii) $\mathcal{B}_0$ is pairwise $\mathcal{L}$-sufficient for $\mathcal{B}$ w.r.t. $\mathcal{C}$

(iii) $\mathcal{B}$ and $\mathcal{C}$ are conditionally independent given $\mathcal{B}_0$ for each $P_\lambda : \lambda \in \Lambda$

($\tilde{\text{iii}}$) $\mathcal{B}$ and $\mathcal{C}$ are conditionally independent given $\mathcal{B}_0$ for each $P_\lambda : \lambda \in \Lambda_0$

(iv) $\mathcal{B}$ and $\mathcal{C}$ are conditionally independent given $\mathcal{B}_0$ for each θ and there is a set B_0 in $\mathcal{B}_0$ so that:

(a) To each bounded $\mathcal{B}$ measurable function g corresponds a $\mathcal{B}_0$ measurable function s_g so that $E_\theta(g|\mathcal{B}_0) = s_g$ a.e on B_0 for each $\theta \in \Theta$

(b) To each bounded $\mathcal{C}$ measurable function h corresponds a $\mathcal{B}_0$ measurable function t_h so that $E_\theta(h|\mathcal{B}_0) = t_h$ a.e P_θ on B_0^c for each $\theta \in \Theta$.

The implication (iv) => (i) is, essentially, proved in Takeuchi and Akahira [5], while the implication (i) => (iii), and thus (ii) => $(\widetilde{iii})$, follows easily from theorem 2 in their paper.

<u>Proof of the theorem.</u>

The structure of the proof is

$$\begin{tikzcd} (i) \arrow[r, Rightarrow] \arrow[drr, Rightarrow] & (ii) \arrow[r, Rightarrow] & (\widetilde{iii}) \arrow[r, Rightarrow] & (iv) \arrow[r, Rightarrow] & (i) \\ & & (iii) \arrow[u, Rightarrow] & & \end{tikzcd}$$

(i) => (ii): Follows directly from the definition of $\mathcal{L}$-sufficiency.

(i) => (iii): Consider a particular $\lambda \in \Lambda$ and a particular loss function $L \in \mathcal{L}$. If δ is a decision rule then, by (i), there is a $\mathcal{B}_0$ measurable decision rule $\tilde{\delta}$ so that $\int r_{\tilde{\delta}} d\lambda \leqq \int r_\delta d\lambda$.

Thus $\mathcal{B}_o$ is $\mathcal{L}$-sufficient for $\mathcal{B}$ w.r.t. $\mathcal{C}$ when the underlying distribution is known to be P_λ. In this case, however, $\mathcal{L}$ consists of all non negative loss functions. By theorem 2 in [5], $\mathcal{B}_o$ is prediction sufficient for $\mathcal{B}$ w.r.t. $\mathcal{C}$ in this situation. Thus $\mathcal{B}$ and $\mathcal{C}$ are conditionally independent given $\mathcal{B}_o$ under P_λ.

(ii) => ($\widetilde{\text{iii}}$) : This is just a particular case of the statement "(ii) => (iii)": proved above.

(iv) => (i): This is essentially proved in theorem 3 in Takeuchi's and Akahira's paper [5]. For the sake of completeness, however, we include the argument here: Take $L \in \mathcal{L}$ as loss function and let δ be a decision function. By (iv) there are for, for each $t \in T$, $\mathcal{B}_o$ measurable functions φ_t and $M(.,t)$ on, respectively, B_o and B_o^c so that $\varphi_t = E_\theta(\delta_t|\mathcal{B}_o)$; $\theta \in \Theta$ on B_o while $M(.,t) = E_\theta(L(.,t)|\mathcal{B}_o)$; $\theta \in \Theta$ on B_o^c. Define $\tilde{\delta}$ by $\tilde{\delta}_t = \varphi_t$ on B_o while $\tilde{\delta}_\tau = 1$ on B_o^c where $M(\cdot,\tau) = \min_t M(\cdot,t)$. Then:

$$r_\delta(\theta) = E_{\theta,\delta}I_{B_o}L + E_{\theta,\delta}I_{B_o^c}L = \text{(by conditional independence)}$$

$$E_{\theta,\tilde{\delta}}I_{B_o}L + E_{\theta,\delta}I_{B_o^c}M \geq E_{\theta,\tilde{\delta}}I_{B_o}L + E_\theta I_{B_o^c}M(.,\tau) = \text{(by conditional}$$

independence) $E_{\theta,\tilde{\delta}}I_{B_o}L + E_{\theta,\tilde{\delta}}I_{B_o^c}L = r_{\tilde{\delta}}(\theta)$. It remains to prove:

($\widetilde{\text{iii}}$) => (iv): We will in this part of the proof use the notation $\tilde{\mu}$ to denote the restriction of a measure μ to $\mathcal{B}_o$.

Suppose (iii) holds. We must prove the existence of a set B_o with the desired properties. The crucial result needed is:

$$(\S)\quad [E_{\theta_1}(g|\mathcal{S}_0)-E_{\theta_0}(g|\mathcal{S}_0)][E_{\theta_1}(h|\mathcal{S}_0)-E_{\theta_0}(h|\mathcal{S}_0)] = 0$$

almost everywhere $\tilde{P}_{\theta_0} \wedge \tilde{P}_{\theta_1}$ *)

when $\theta_0, \theta_1 \in \Theta$ and g and h are bounded functions on $\mathcal{X}$ which are, respectively, $\mathcal{S}$ measurable and $\mathcal{C}$ measurable.

As only two values, θ_0 and θ_1, of θ are involved we may in the proof of (§) assume that $\Theta = \{0,1\}$, $\theta_0 = 0, \theta_1 = 1$ and $\pi = \frac{1}{2}(P_0+P_1)$. Then **)

$$E(f_i|\mathcal{S}_0) = d\tilde{P}_i|d\pi \ ; \ i = 0,1 \quad \text{and} \quad \bigwedge_{i=0}^{1} E(f_i|\mathcal{S}_0) = d[\tilde{P}_0 \wedge \tilde{P}_1]/d\pi .$$

It follows that we must show that (§) holds a.e. π on the set

$[\bigwedge_{i=0}^{1} E(f_i|\mathcal{S}_0) > 0]$. We restrict ourselves to this set for the remaining part of the proof of "$(\widetilde{iii}) \Rightarrow (iv)$". The qualification "a.e. π" will be omitted.

Note first that

$$E_i(s|\mathcal{S}_0) = E(sf_i|\mathcal{S}_0)/E(f_i|\mathcal{S}_0) \ ; \quad i = 0,1 \quad \text{and}$$

$$E(s|\mathcal{S}_0) = \tfrac{1}{2}\sum_i (f_i|\mathcal{S}_0)E_i(s|\mathcal{S}_0)$$

for any bounded measurable s. It follows, using the Markov property that

---

*) If μ and ν are finite measures on $\mathcal{A}$ then $\mu \wedge \nu$ is the largest measure $\leqq \mu$ and $\leqq \nu$ for the set wise ordering of measures. See Neveu [3 page 107].

**) If a and b are numbers then $a \wedge b = \min(a,b)$.

$$E_i(gh|\mathfrak{B}_o) = E_i(g|\mathfrak{B}_o)E_i(h|\mathfrak{B}_o) \;; \quad i = 0,1$$

and

$$E(gh|\mathfrak{B}_o) = E(g|\mathfrak{B}_o)E(h|\mathfrak{B}_o) \,.$$

The last equation may, using the first equation, be written:

$$\sum_i a_i E_i(g|\mathfrak{B}_o) = 0$$

where

$$a_i = E(f_i|\mathfrak{B}_o)[E_i(h|\mathfrak{B}_o) - \tfrac{1}{2}\sum_j E_j(h|\mathfrak{B}_o)E(f_j|\mathfrak{B}_o)]$$

$\sum_i f_i = 2$ imply

$$a_o = -\, a_1 = \tfrac{1}{2}E(f_o|\mathfrak{B}_o)E(f_1|\mathfrak{B}_o)(E_o(h|\mathfrak{B}_o) - E_1(h|\mathfrak{B}_o)) \,.$$

(§) follows now by inserting these expressions for a_i ; $i=0,1$.

We must now return to the general situation with a dominated family $\{P_\theta : \theta \in \Theta\}$.

We shall first show that

$$(\alpha) \quad [E_{\theta_1}(g|\mathfrak{B}_o)-E_{\theta_2}(g|\mathfrak{B}_o)][E_{\theta_3}(h|\mathfrak{B}_o)-E_{\theta_2}(h|\mathfrak{B}_o)] = 0 \quad \text{a.e.}$$

$$\bigwedge_{i=0}^{3} \tilde{P}_{\theta_i} \quad \text{when} \quad g \in \mathcal{G},\ h \in \mathcal{H}$$

and $\theta_i \in \Theta$; $i=0,1,2,3$. We may - since $d \bigwedge_{i=0}^{3} \tilde{P}_{\theta_i} / d\tilde{\pi} = \bigwedge_{i=0}^{3} E(f_i|\mathfrak{B}_o)$ - restrict attention to the set

$\tilde{B} = [\bigwedge_{i=0}^{3} E(f_i|\mathcal{B}_0) > 0]$. We omit the qualification "a.e. π" in the proof of (α) . By (§) we have:

$$(\beta)\ [E_{\theta_i}(g|\mathcal{B}_0)-E_{\theta_j}(g|\mathcal{B}_0)][E_{\theta_i}(h|\mathcal{B}_0)-E_{\theta_j}(h|\mathcal{B}_0] = 0 \ .$$

Put:

$$\tilde{B}_0 = \tilde{B}\cap[E_{\theta_1}(g|\mathcal{B}_0)-E_{\theta_0}(g|\mathcal{B}_0))(E_{\theta_3}(h|\mathcal{B}_0)-E_{\theta_0}(h|\mathcal{B}_0)) \neq 0]$$

$$\tilde{B}_1 = \tilde{B}\cap[E_{\theta_1}(g|\mathcal{B}_0)-E_{\theta_0}(g|\mathcal{B}_0))(E_{\theta_2}(h|\mathcal{B}_0)-E_{\theta_0}(h|\mathcal{B}_0)) \neq 0]$$

(α) will be proved if we can show that $\pi(\tilde{B}_0) = \pi(\tilde{B}_1) = 0$. On $\tilde{B}_0$ we have - by (β)

$$E_{\theta_1}(h|\mathcal{B}_0) = E_{\theta_0}(h|\mathcal{B}_0) \quad \text{and} \quad E_{\theta_3}(g|\mathcal{B}_0) = E_{\theta_0}(g|\mathcal{B}_0) \ .$$

On the set $\tilde{B}_0\cap[E_{\theta_3}(g|\mathcal{B}_0) \neq E_{\theta_1}(g|\mathcal{B}_0)]$ we will also have

$$E_{\theta_3}(h|\mathcal{B}_0) = E_{\theta_1}(h|\mathcal{B}_0) = E_{\theta_0}(h|\mathcal{B}_0)$$

which is impossible on $\tilde{B}_0$. It follows that $E_{\theta_3}(g|\mathcal{B}_0) = E_{\theta_1}(g|\mathcal{B}_0) = E_{\theta_0}(g|\mathcal{B}_0)$ which is also (π) impossible on $\tilde{B}_0$. Hence $\pi(\tilde{B}_0) = 0$. Similarily $\pi(\tilde{B}_1) = 0$. Thus (α) is proved.

Note next that (α) may be rewritten as (α') $[E(gf_{\theta_1}|\mathcal{B}_o)E(f_{\theta_o}|\mathcal{B}_o)$

$- E(gf_{\theta_o}|\mathcal{B}_o)E(f_{\theta_1}|\mathcal{B}_o)][E(hf_{\theta_3}|\mathcal{B}_o)E(f_{\theta_2}|\mathcal{B}_o)$

$- E(hf_{\theta_2}|\mathcal{B}_o)E(f_{\theta_3}|\mathcal{B}_o)] = 0$; a.e. π .

Multiplying with $c(\theta_o)c(\theta_3)$ and summing over $\theta_o, \theta_3 \in \Theta_o$ we get:

(γ) $[E(gf_{\theta_1}|\mathcal{B}_o)-E(g|\mathcal{B}_o)E(f_{\theta_1}|\mathcal{B}_o)][E(hf_{\theta_2}|\mathcal{B}_o)$

$- E(h|\mathcal{B}_o)E(f_{\theta_2}|\mathcal{B}_o)] = 0$; a.e. π .

Put $V_{\theta,g} = [E(gf_\theta|\mathcal{B}_o) = E(g|\mathcal{B}_o)E(f_\theta|\mathcal{B}_o)]$

and $W_{\theta,h} = [E(hf_\theta|\mathcal{B}_o) = E(h|\mathcal{B}_o)E(f_\theta|\mathcal{B}_o)]$.

Let V and W be sets in $\mathcal{B}_o$ such that

$$I_V = \operatorname{essinf} \{I_{V_{\theta,g}} : \theta \in \Theta , g \in \mathcal{G}\} \quad \text{w.r.t.} \quad \tilde{\pi}$$

and

$$I_W = \operatorname{essinf} \{I_{W_{\theta,h}} : \theta \in \Theta , h \in \mathcal{H}\} \quad \text{w.r.t.} \quad \tilde{\pi} .$$

We will complete the proof by showing that (iv) holds with $B_o = V \cap W^c$.

It follows from (γ) that

$$V^c_{\theta_1,g} \subseteqq W_{\theta_2,h} \quad \text{a.e. } \pi ; \quad \theta_2 \in \Theta , h \in \mathcal{H}$$

Hence $V^c_{\theta_1,g} \subseteqq W$ a.e. π ; $\theta_1 \in \Theta$, $g \in \mathcal{G}$

or $W^c \subseteqq V_{\theta_1,g}$ a.e. π ; $\theta_1 \in \Theta$, $g \in \mathcal{G}$

Hence $W^c \subseteqq V$ a.e. π so that $\pi(V \cup W) = 1$.

Let $\theta \in \Theta$ and $g \in \mathcal{G}$. Then $V_{\theta,g} \subseteqq V$ a.e. π. Hence, by the definition of $V_{\theta,g}$, $E(g|\mathcal{B}_0)$ is a version of $E_\theta(g|\mathcal{B}_0)$ on V. Similarily $E(h|\mathcal{B}_0)$ is a version of $E_\theta(h|\mathcal{B}_0)$ on W. (iv) follows now since $B_0 \subseteqq V$ and $B_0^c \subseteqq W$ a.e. π.

$\square$

Remark 1.

Assume that $\mathcal{B}_0$ satisfies one of (and consequently all) conditions (i)-(iv). Suppose further that there is, for each θ, regular conditional probabilities of $\mathcal{B}$ given $\mathcal{B}_0$ and of $\mathcal{C}$ given $\mathcal{B}_0$. Then these regular conditional probabilities may be specified so that $P_{\theta,x}(B|\mathcal{B}_0)$ does not depend on θ when $x \in B_0$ and $B \in \mathcal{B}$ while $P_{\theta,x}(C|\mathcal{B}_0)$ does not depend on θ when $x \in B_0^c$ and $C \in \mathcal{C}$.

Remark 2.

Consider three arbitrary sub σ algebras $\mathcal{U}$, $\mathcal{V}$ and $\mathcal{W}$ of $\mathcal{A}$. Then $\mathcal{U}$ and $\mathcal{W}$ are conditionally independent given $\mathcal{V}$ if and only if $\mathcal{U} \vee \mathcal{V}$ *) and $\mathcal{W}$ are conditionally independent given $\mathcal{V}$. Thus the theorem may be applied with $\mathcal{B}_0 = \mathcal{V}$, $\mathcal{B} = \mathcal{U} \vee \mathcal{V}$ and $\mathcal{C} = \mathcal{V}$. It follows in particular that conditional independence for all $\lambda \in \Lambda_0$ imply conditional independence for all $\lambda \in \Lambda$.

---

*) $\mathcal{U} \vee \mathcal{V}$ is the smallest σ-algebra containing $\mathcal{U}$ and $\mathcal{V}$.

Remark 3.

Among the equivalence classes of $\mathcal{L}$-sufficient σ-algebras there is a smallest element. In other words there is a sub σ-algebra $\tilde{\mathfrak{B}}$ of $\mathfrak{B}$ such that a sub σ-algebra $\mathfrak{B}_0$ of $\mathfrak{B}$ is $\mathcal{L}$-sufficient if and only if to each $\tilde{B} \in \tilde{\mathfrak{B}}$ corresponds a $B_0 \in \mathfrak{B}_0$ so that $P_\theta(\tilde{B} \Delta B_0) = 0$; $\theta \in \Theta$.

Consider first an arbitrary $\mathcal{L}$-sufficient $\mathfrak{B}_0$. Let $B_0 \in \mathfrak{B}_0$ satisfy (iv). Then

(1) $E(f_\theta|\mathfrak{B}) = E(f_\theta|\mathfrak{B}_0)$ a.e. π on B_0

while

(2) $E_\lambda(h|\mathfrak{B}) = E_\lambda(h|\mathfrak{B}_0)$ a.e. P_λ for all $\lambda \in \Lambda$.

[The last statement follows directly from conditional independence and the first statement follows from the following computations:

Let $B \in \mathfrak{B}$, $B \subseteqq B_0$. Then $\int_B E(f_\theta|\mathfrak{B}_0)d\pi = \int_{B_0} \pi(B|\mathfrak{B}_0)f_\theta d\pi$

$= \text{(by (i,v))} \int_{B_0} P_\theta(B|\mathfrak{B}_0)dP_\theta = P_\theta(B) = \int_B E(f_\theta|\mathfrak{B})d\pi$.]

Define for each $\lambda \in \Lambda_0$ and each bounded $\mathcal{C}$ measurable function h a $\mathfrak{B}$ measurable function $r_\lambda(h)$ by:

$$r_\lambda(h) = \begin{cases} E_\lambda(h|\mathfrak{B}) & \text{when } E(f_\lambda|\mathfrak{B}) > 0 \\ E(h|\mathfrak{B}) & \text{when } E(f_\lambda|\mathfrak{B}) = 0 \end{cases} .$$

Then the sub σ-algebra $\tilde{\mathfrak{B}}$ of $\mathfrak{B}$ which is induced by these functions is "minimum" $\mathcal{L}$-sufficient for $\mathfrak{B}$ w.r.t $\mathcal{C}$.

[By the definition , $\mathcal{B}$ and $\mathcal{C}$ are conditionally independent given $\tilde{\mathcal{B}}$ for each $\lambda \in \Lambda_0$. Hence $\tilde{\mathcal{B}}$ is $\mathcal{L}$-sufficient for $\mathcal{B}$ w.r.t $\mathcal{C}$. The same argument applies to any sub σ algebra of $\mathcal{B}$ containing $\tilde{\mathcal{B}}$. Let $\mathcal{B}_0$ be another $\mathcal{L}$-sufficient σ algebra. It follows then from (1) and (2) that there is, for each (λ,h) where $\lambda \in \Lambda_0$ and h is bounded and $\mathcal{C}$ measurable, a $\mathcal{B}_0$ measurable function $\tilde{r}_\lambda(h)$ so that $r_\lambda(h) = \tilde{r}_\lambda(h)$ a.e. π. Thus $\tilde{\mathcal{B}}$ is, essentially contained in $\mathcal{B}_0$]. The construction of $\tilde{\mathcal{B}}$ may be simplified by noting that we may restrict attention to smaller classes of function h . If, for example, $\tilde{\mathcal{C}}$ is a basis for $\mathcal{C}$ which is closed under finite intersections then if suffices to consider indicators of sets in $\tilde{\mathcal{C}}$.

As an example consider the case where $\Theta = \{1,2\}$ and that the joint distribution of X and Y is given by the table in section 1. Put $\pi = \frac{1}{2}(P_1+P_2)$, $r(x) = \pi(Y=2\,|X=x)$, $r_\theta(x) = \pi(Y=2\,|X=x)$ or $= r(x)$ as $P_\theta(X=x) > 0$ or $= 0$. Then $r_\theta(x) = r(x) = \beta$ when $x \leqq 2$ while $r_\theta(3) = \gamma_\theta$. By the remark above the algebra induced by r, r_1 and r_2 is minimum $\mathcal{L}$-sufficient. Thus $X_0 = \max(X,2)$ is "minimum" $\mathcal{L}$-sufficient provided $\gamma_1 \neq \beta$ or $\gamma_2 \neq \beta$. If, in particular, $\tau_1 = 0$, $\tau_2 = 1$, $\alpha_1 < 1$ and $\alpha_2 < 1$ then $P_\theta(X=\theta) = 0$ and $\pi(X=\theta) > 0$; $\theta=1,2$. It follows that it is essential that r_θ is defined as above on the P_θ singular set $[X=\theta]$.

Remark 4.

It follows from theorem 11.3 in Bahadur [1] (See also Skibinsky [4]) that $\mathcal{B}_0$ is prediction sufficient for $\mathcal{B}$ [i.e. $\mathcal{B}_0$ is sufficient for $\mathcal{B}$ and, $\mathcal{B}$ and $\mathcal{C}$ are conditionally independent given $\mathcal{B}_0$] if and only if $\mathcal{B}_0$ is sufficient for all probability measures on $\mathcal{B}$ of the form $(P_\theta(B|C) : B \in \mathcal{B})$ where $P_\theta(C) > 0$. This yield in particular a description of conditional independence in terms of sufficiency. Combining this with our theorem, the relationship between prediction sufficiency and $\mathcal{L}$-sufficiency may be described as follows:

Let for each pair $(\theta_1,\theta_2) \in \Theta \times \Theta$, k_{θ_1,θ_2} denote the set of probability measures on $\mathcal{B}$ of the

$$\text{form} \left\{ \frac{P_{\theta_1}(BC) + P_{\theta_2}(BC)}{P_{\theta_1}(C) + P_{\theta_2}(C)} \right. \quad ; \quad B \in \mathcal{B} \qquad \text{where}$$

$P_{\theta_1}(C) + P_{\theta_2}(C) > 0$. Then $\mathcal{B}_0$ is prediction sufficient if and only if $\mathcal{B}_0$ is sufficient for $\bigcup_{\theta_1,\theta_2} k_{\theta_1,\theta_2}$, while $\mathcal{B}_0$ is $\mathcal{L}$- sufficient if and only if $\mathcal{B}_0$ is sufficient for each k_{θ_1,θ_2} ; $(\theta_1,\theta_2) \in \Theta \times \Theta$.

References.

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

[2] Halmos, P. and Savage, L.J. (1949). Application of the Radon Nikodym theorem to the theory of sufficient statistics. Ann. Math. Statist. 20, 225-241.

[3] Neveu, J. (1965). Mathematical foundations of the calculus of probability. San Francisco: Holden-Day.

[4] Skibinsky, M. (1967). Adequate subfields and sufficiency. Ann. Math. Statist. 38, 155-161.

[5] Takeuchi, K. and Akahira, M. (1975). Characterizations of prediction sufficiency (adequacy) in terms of risk functions. Ann. Statist. 3, 1018-1024.