Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

M.L. Eaton

Admissibility in fair Bayes prediction problems
I. General theory

Department of Mathematical Statistics Report MS-R8703 April

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

Admissibility in Fair Bayes Prediction Problems

I. General Theory

Morris L. Eaton
Univerisity if Minnesota,
School of Statistics

This paper is concerned with sufficient conditions for the admissibility of posterior predictive distributions which are derived from improper proper distributions. It is argued that the relevant loss functions are the so called fair Bayes loss functions. The admissibility results are valid for a wide class of such loss functions. The main result is applied to the one dimensional translation problem with Lebesgue measure as the improper prior to give sufficient conditions for admissibility of the resulting predictive distribution. A connection with recurrence of an induced Markov chain is described.

1980 Mathematics Subject Classification: Primary: 62C05, Secondary: 62C15.
Key Words & Phrases: prediction, predictive distribution, Bayes rules, improper prior distributions, fair Bayes loss functions.
Note: Portions of this work were completed while the author was a visiting professor at the Department of Statistics, Berkeley and while the author was on sabbatical leave at the Centre for Mathematics and Computer Science, Amsterdam, The Netherlands. This work was supported in part by the National Science Foundation Grant DMS-83-19924.

Report MS-R8703
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

1. Introduction:

The problem of predicting the value of some future observable random quantity on the basis of available data has received considerable attention in the statistical literature. The time series literature abounds with derivations of minimum mean squared error predictions, while the prediction of a future response, given values of covariates, is a classical problem in linear model theory which is ordinarily attacked via mean squared error considerations. No less attention is afforded the prediction problem in the Bayesian world, although the emphasis is somewhat different. Given a probabilistic model and a prior distribution, the Bayesian solution to the prediction problem is just the conditional distribution of the quantity to be predicted, given the data and the prior. This conditional distribution is called the predictive distribution and is discussed at length in the basic text by Aitchison and Dunsmore (1975). It has been argued in the literature that prediction, as opposed to say parametric estimation, is the proper activity of statisticians--partly because prediction is often the scientific question of interest, and partly because the ability of statisticians to predict can actually be checked, unlike the popular parametric estimation--confidence set activity. For an introduction to this point of view, and further references, see Geisser (1980).

In this paper, the prediction problem is studied from a Bayesian-decision theoretic viewpoint. To describe the philosophical underpinnings of the formulation of the problem, first consider the simplest situation in which a random variable Z is to be predicted on the basis of some observed data X, and the joint distribution of X and Z is completely known. In this case the

inferential solution to the prediction problem is given by the conditional distribution of Z given X = x, say $Q(\cdot|x)$, because "$Q(\cdot|x)$ contains everything we known about Z after seeing x". From the Bayesian point of view the above statement is true virtually by definition, while from a decision theory point of view its truth derives from the observation that all decision problems involving Z are solved by choosing the action which minimized the expected loss computed under $Q(\cdot|x)$. Hence $Q(\cdot|x)$ is "sufficient" in the Bayesian sense for all decision problems. That $Q(\cdot|x)$ is the correct solution to the prediction problem is assumed in what follows.

Next assume that the joint distribution of X and Z depends on an unknown parameter θ, but θ has a known prior distribution π. In this case the Bayesian solution is again the conditional distribution of Z given X = x,

$$(1.1) \qquad Q_\pi(\cdot|x) = \int Q(\cdot|x,\theta)\pi(d\theta)$$

where $Q(\cdot|x,\theta)$ is the conditional distribution of Z given X = x and θ. However, a complete solution to the inferential problem is provided by the conditional distribution of (Z,θ) given X = x from which $Q_\pi(\cdot|x)$ is easily obtained. This observation is quite explicit in the proof of the basic Lemma 4.1.

The point of the above discussion is that under certain circumstances, namely when the prior π is known, the solution to the inferential problem regarding the prediction of Z is *a priori* the predictive distribution $Q_\pi(\cdot|x)$. This revealed truth has important consequences for the decision theoretic formulation of the prediction problem to which we now turn.

A common technique for providing predictive distributions which are both analytically tractable and often have frequentist interpretations is the use of improper prior distributions. However the direct Bayesian interpretation is lost and the suitability of such predictive distributions is at issue. One way to get at this issue is to compare proposed predictive distributions in a decision theoretic framework. Because actions, in the decision theory sense, are distributions on the space Φ of possible Z values, an appropriate action space is $M_1(\Phi)$--the set of all probability distributions on Φ. Thus a decision rule (or inference as defined in Eaton (1982)) is a (measurable) function defined on the sample space Ψ taking values in $M_1(\Phi)$. The value of a decision rule δ at $x \in \Psi$ is $\delta_x \in M_1(\Phi)$.

With Θ denoting the parameter space of a probabilistic model for X and Z, let $L(a,x,z,\theta)$ be a loss function where $a \in M_1(\Phi)$, $x \in \Psi$, $z \in \Phi$, and $\theta \in \Theta$. The risk function of a decision rule δ is then

$$R(\delta,\theta) = E_\theta L(\delta_X, X, Z, \theta). \tag{1.2}$$

Given a proper prior π, recall that δ_π is a Bayes rule for π if

$$\int R(\delta,\theta)\pi(d\theta) \geq \int R(\delta_\pi,\theta)\pi(d\theta) \tag{1.3}$$

for all decision rules δ. However, it has already been argued that, *a priori*, the Bayes solution to the prediction problem is the predictive distribution $Q_\pi(\cdot|x)$ in (1.2). Thus, in order to obtain consistency between the decision

theoretic Bayes solution and the "introspective" Bayes solution, it is necessary that

$$(1.4) \qquad \int R(\delta,\theta)\pi(d\theta) \geq \int R(Q_\pi,\theta)\pi(d\theta)$$

for all decision rules δ where Q_π denotes the decision rule whose value at x is $Q_\pi(\cdot|x)$. Loss functions for which (1.4) holds for all proper priors π are called fair Bayes loss functions--fair in the sense that they give the right solution to the Bayes problems whose solutions we claim to know. Examples of a large class of such loss functions are given in Section 2. In this paper, attention is restricted to prediction problems in which decision rules are compared via risk functions obtained from the fair Bayes loss functions defined in the next section.

Here is the main problem with which this paper deals. Let ν be a improper prior distribution on the parameter space Θ. In many standard situations it is possible to define a formal posterior distribution for Z given X = x using the improper prior ν. The existence of such a posterior distribution, say $Q_\nu(\cdot|x)$, is discussed in Section 4. Assuming $Q_\nu(\cdot|x)$, exists, the main result of this paper, Theorem 4.1, provides conditions under which the decision rule Q_ν (whose value at x is $Q_\nu(\cdot|x)$) is almost ν-admissible for a variety of fair Bayes Loss functions. A version of Stein's (1955) sufficient condition for admissibility is coupled with some inequalities discussed in Section 3 to provided the proof of Theorem 4.1.

In Section 5, Theorem 4.1 is applied to the one dimensional translation

parameter problem with $\Theta = R^1$ when the improper prior is Lebesgue measure $d\theta$ on R^1. The results here show that under very weak conditions, the formal posterior obtained from $d\theta$ is almost $d\theta$-admissible. Conditions under which almost ν-admissibility implies admissibility are briefly touched on in Section 6.

The sufficient condition for almost ν-admissibility involves a function defined on $\Theta\times\Theta$ which is rather interesting. To describe this function, let $p(x|\theta)$ be the marginal density of X given θ with respect to a fixed σ-finite measure λ. Given the improper prior ν, set

$$m(x) = \int p(x|\theta)\nu(d\theta). \tag{1.5}$$

Under the assumption that

$$A = \{x|0 < m(x) < +\infty\}$$

has $p(\cdot|\theta)$-measure one for each θ, the function

$$t(\theta,\eta) = \int \frac{p(x|\theta)p(x|\eta)}{m(x)}\,\lambda(dx) \tag{1.6}$$

is well defined. Since $p(x|\theta)/m(x)$ can be interpreted as the conditional density of θ given x obtained from the improper prior ν, $t(\theta,\eta)$ is the expected value of this conditional density at θ when the data X has been sampled from $p(\cdot|\eta)$. The condition for almost ν-admissibility involves the behavior of t interpreted as a linear transformation on the space $L^2(\nu)$ of square integrable

functions on (θ,ν). More precisely, for $h \in L^2(\nu)$, define the linear transformation A by

$$(Ah)(\theta) = \int t(\theta,\eta)h(\eta)\nu(d\eta). \tag{1.6}$$

The condition for almost ν-admissibility of Q_ν can be described as follows. Let $(\cdot,\cdot)$ denote the usual inner product on $L^2(\nu)$ and for each set C of positive ν measure, let

$$H(C) = \{h \in L^2(\nu) \mid h \geq 0, \quad \int_C h^2(\theta)\nu(d\theta) > 0\}.$$

Theorem 4.2 shows that if

$$\inf_{h\in H(C)} \frac{(h,(I-A)h)}{\int_C h^2(\theta)\nu(d\theta)} = 0 \tag{1.7}$$

for each set C of positive ν-measure, then Q_ν is almost ν-admissible for the class of loss functions described in Section 2. In 1.7, I is the identity linear transformation. It is condition (1.7) which leads to a connection between Markov chains and almost ν-admissibility which is discussed briefly in Section 6.

Here is an outline of the paper. Section 2 contains notation and a discussion of a class of fair Bayes loss functions. The inequality in Proposition 2.2 is an important upper bound on the average risk difference of

any decision rule and a Bayes decision rule. A sufficient condition for almost ν-admissibility is given in Section 3. Section 4 contains the main theorem of this paper. In Section 5, the conditions for almost ν-admissibility are applied to the one-dimensional translation problem. Section 6 contains some discussion of the main theorem.

2. Assumptions and Notation:

In this section, notation is set and a decision theoretic prediction problem is formulated. In addition, fair Bayes loss functions are described and some preliminary results are established.

Consider three measurable spaces (Ψ,B_1), (Φ,B_2) and (Θ,B_3). The three spaces are assumed to be Polish and the associated σ-algebras are those generated by the open sets. The problem under consideration is the prediction of the unobserved variable $Z \in \Phi$ on the basis of data $X \in \Psi$. For each parameter value $\theta \in \Theta$, a joint distribution for (X,Z) is given in the form

$$P(dx|z,\theta)S(dz|\theta) \tag{2.1}$$

where $P(\cdot|z,\theta)$ is the conditional distribution of X given $Z = z$ and θ, and $S(\cdot|\theta)$ is the conditional distribution of Z given θ.

Let $M_1(\Phi)$ be the space of all probability measures defined on (Φ,B_2). When equipped with the weak* topology, $M_1(\Phi)$ is a separable metric space. The σ-algebra generated by this topology is denoted by B^*. A <u>decision rule</u> or <u>inference</u> is a measurable function defined on (Ψ,B_1) taking values in

$(M_1(\Phi),B^*)$. This definition of decision rule coincides with the usual notion of a randomized decision rule (see Dubins and Freedman (1964) and Eaton (1982)). The value of δ at $x \in \Psi$ is $\delta_x \in M_1(\Phi)$.

For any probability measure π on (Θ,B_3), the model (2.1) together with π determines a joint distribution on $\Psi\times\Phi\times\Theta$ and thus a marginal distribution on $\Psi\times\Phi$. Since Ψ and Φ are Polish, there then exists a conditional distribution of Z given X = x (see Parthasarathy (1967), Chapter 5). This conditional distribution is denoted by $Q_\pi(\cdot|x)$. Note that $Q_\pi(\cdot|x)$ defines a decision rule, say Q_π, whose value at x is $Q_\pi(\cdot|x)$. As remarked in the previous section, Q_π is the Bayesian solution to the problem of predicting Z when the prior is π.

A loss structure is now introduced into the problem. Let $k(\cdot,\cdot)$ be a bounded symmetric bimeasurable function defined on $\Phi\times\Phi$. The function k defines a bilinear function of bounded signed measures, say ξ_1 and ξ_2, via

$$(2.2) \qquad \langle\xi_1,\xi_2\rangle = \iint k(z_1,z_2)\xi_1(dz_1)\xi_2(dz_2).$$

The bilinear function function $\langle\cdot,\cdot\rangle$ satisfies

$$(2.3) \qquad \langle\xi_1,\xi_2\rangle = \langle\xi_2,\xi_1\rangle$$

because k is symmetric. It is also assumed that $\langle\cdot,\cdot\rangle$ is non-negative definite--that is, $\langle\cdot,\cdot\rangle$ satisfies

$$(2.4) \qquad \langle\xi,\xi\rangle \geq 0$$

for all bounded signed measures ξ. A simple way to construct symmetric k's so that (2.3) and (2.4) hold is to set

$$k(z_1,z_1) = \sum_1^s u_i(z_1)u_i(z_2) \tag{2.5}$$

where $u_1,\ldots, u_s$ are bounded measurable functions defined on Φ.

For $z \in \Phi$ let $\varepsilon_z \in M_1(\Phi)$ denote the probability measure with mass one at z. Also, let $Q_\theta(\cdot|x)$ be the conditional distribution of Z given X = x when the parameter value is θ. Given a bilinear function $\langle\cdot,\cdot\rangle$ as above, define loss functions L_1 and L_2 by

$$\begin{cases} L_1(\alpha,z) = \langle\alpha-\varepsilon_z,\alpha-\varepsilon_z\rangle \\ L_2(\alpha,x,\theta) = \langle\alpha-Q_\theta(\cdot|x),\ \alpha-Q_\theta(\cdot|x)\rangle \end{cases} \tag{2.6}$$

where $\alpha \in M_1(\Phi)$. The loss function L_1 is more appropriate when Z is best regarded as some unknown constant and the problem is to guess the value of Z. For example, if $\Phi = \Theta$, $B_2 = B_3$, and $S(\cdot|\theta)$ is the probability measure degenerate at θ, then we simply have the classical estimation problem. In this case L_1 and L_2 coincide. However, if Z is the future value of some random quantity which we do not know, then the inference $Q_\theta(\cdot|x)$ is most appropriate when we know θ and X = x. It is possible to allow the function k defining $\langle\cdot,\cdot\rangle$ to depend on both θ and x as long as k remains bounded and (2.4) holds. This extension only

complicates the notation and is omitted.

The result below shows that the loss functions L_1 and L_2 define fair Bayes decision problems in the sense that for each proper prior π, the Bayes rule is the posterior $Q_\pi(\cdot|x)$ defined above. See Eaton (1982) for a formal definition and discussion of fair Bayes decision problems.

For any loss function L defined on $M_1(\Phi)\times\Psi\times\Phi\times\Theta$, the risk function of a decision rule δ is defined to be the expected loss under L with θ fixed. In symbols, the risk function is

$$R(\delta,\theta) = E_\theta L(\delta_X,X,Z,\theta)$$

where the expectation is computed under the joint distribution of X and Z with θ fixed. Naturally, sufficient conditions on L are assumed so that the above expression makes sense.

Proposition 2.1: Given any proper prior distribution π on (Θ,B_3), the Bayes rule for the decision problem with loss functions L_1 or L_2 is the posterior distribution $Q_\pi(\cdot|x)$. In other words,

$$\inf_\delta \int R_i(\delta,\theta)\pi(d\theta) = \int R_i(Q_\pi,\theta)\pi(d\theta) \tag{2.7}$$

where the risk function $R_i(\delta,\theta)$ is the expected loss under L_i when the parameter value is θ, i = 1,2.

Proof: The proof is given first for loss function L_2. For any decision rule δ, consider the average risk difference

$$\Delta_2 = \int R_2(\delta,\theta)\pi(d\theta) - \int R_2(Q_\pi,\theta)\pi(d\theta). \tag{2.8}$$

Using the definition of L_2, Δ_2 can be written

$$\begin{aligned}\Delta_2 &= \iiint \langle \delta_x - Q_\pi(\cdot|x),\ \delta_x - Q_\pi(\cdot|x)\rangle P(dx|\theta,z)S(dz|\theta)\pi(d\theta) \\ &\quad + 2\iiint \langle Q_\pi(\cdot|x) - Q_\theta(\cdot|x),\ \delta_x - Q_\pi(\cdot|x)\rangle P(dx|\theta,z)S(dz|\theta)\pi(d\theta) \\ &= \Delta_3 + 2\Delta_4.\end{aligned} \tag{2.9}$$

To show $\Delta_2 \geq 0$, it suffices to show $\Delta_4 = 0$ since $\Delta_3 \geq 0$ due to the non-negative definiteness of $\langle\cdot,\cdot\rangle$. However, Δ_4 is the expectation (over the joint distribution of X, Z, θ) of

$$F(x,\theta) = \langle Q_\pi(\cdot|x) - Q_\theta(\cdot|x),\ \delta_x - Q_\pi(\cdot|x)\rangle. \tag{2.10}$$

Conditioning on X and using the bilinearity of $\langle\cdot,\cdot\rangle$ yields

$$\Delta_4 = E[E(F(X,\theta)|X)] = E[\langle Q_\pi(\cdot|X) - E(Q_\theta(\cdot|X)|X),\ \delta_X - Q_\pi(\cdot|X)\rangle].$$

However, since $Q_\pi(\cdot|X)$ is the conditional distribution of Z given X, it is clear

that

$$Q_\pi(\cdot|X) = E(Q_\theta(\cdot|X)|X).$$

Thus $\Delta_4 = 0$ so (2.7) holds for $i = 2$.

When $i = 1$, the proof is essentially that given above with $Q_\theta(\cdot|X)$ replaced by ε_Z throughout. The verification that $\Delta_4 = 0$ derives from the identity

$$Q_\pi(\cdot|X) = E(\varepsilon_Z(\cdot)|X)$$

which is easily verified. □

Equation (2.9) yields

Corollary 2.1: Given a prior π and a decision rule δ, the difference in average risks under both loss functions L_1 and L_2 is

$$\Delta = \iiint \langle \delta_x - Q_\pi(\cdot|x),\ \delta_x - Q_\pi(\cdot|x) \rangle P(dx|\theta,z)S(dz|\theta)\pi(d\theta). \tag{2.10}$$

We end this section with an upper bound on Δ in (2.10) which is used in the next section. Give two probability measures, say P_1 and P_2, defined on a common probability space, $V(P_1,P_2)$ denotes the variation distance between P_1 and P_2 defined by

$$V(P_1,P_2) = \sup_B |P_1(B)-P_2(B)|.$$

The sup ranges over the relevant σ-algebra. For any σ-finite measure λ_0 which dominates P_1 and P_2, let

$$p_i = dP_i/d\lambda_0$$

be the density of P_i, $i = 1,2$. It is well known that

$$V(P_1,P_2) = \frac{1}{2}\int |p_1(x)-p_2(x)|\lambda_0(dx). \tag{2.11}$$

For a proof, see Billingsley (1968), p. 224. To apply this to the problem at hand, let α_1 and α_2 be two probabilities on (Φ,B_2), and let $p_i = d\alpha_i/d\lambda_0$ where λ_0 dominates α_i, $i = 1,2$. Then

$$\begin{aligned} \langle \alpha_1-\alpha_2, \alpha_1-\alpha_2 \rangle &= \iint k(z_1,z_2)(\alpha_1-\alpha_2)(dz_1)(\alpha_1-\alpha_2)(dz_2) \\ &= \iint k(z_1,z_2)(p_1(z_1)-p_2(z_1))(p_1(z_2)-p_2(z_2))\lambda_0(dz_1)\lambda_0(dz_2) \\ &\leq \iint |k(z_1,z_2)||p_1(z_1)-p_2(z_1)||p_1(z_2)-p_2(z_2)|\lambda_0(dz_1)\lambda_0(dz_2) \\ &\leq 4K[V(\alpha_1,\alpha_2)]^2 \end{aligned} \tag{2.12}$$

where K is an upper bound on $|k|$. Applying this inequality to (2.10) with $\alpha_1 = \delta_x$ and $\alpha_2 = Q_\pi(\cdot|x)$ yields

Proposition 2.2: For loss functions L_1 and L_2, and Δ given by (2.8),

$$\Delta \leq 4K \iiint [V(\delta_x, Q_\pi(\cdot|x))]^2 P(dx|z,\theta) S(dz|\theta) \pi(d\theta). \tag{2.13}$$

Remark 2.1: Inequality (2.13) forms the basis for all the results in succeeding sections. Thus, if L is any fair Bayes loss function yielding an average risk difference Δ in (2.8) which satisfies (2.13), then the results in the remainder of this paper are valid for that L. □

3. Preliminary Results:

The notion of almost admissibility for decision rules was introduced by Stein (1965). Here, we describe one version of a sufficient condition for almost admissibility which is closely related to Stein's (1955) necessary and sufficient condition for admissibility.

Consider a decision theoretic problem with a space of decision rules D and a measurable parameter space (Θ, B_3). The risk function for the problem is $R(\delta,\theta)$, $\delta \in D$ and $\theta \in \Theta$. For each $\delta \in D$, $R(\delta,\cdot)$ is assumed to be bounded and measurable. Fix a σ-finite measure γ on (Θ, B_3).

Definition 3.1: A decision rule δ_0 is almost γ-admissible if for each $\delta \in D$ which satisfies $R(\delta,\theta) \leq R(\delta_0,\theta)$, the set

$$C_0 = \{\theta | R(\delta,\theta) < R(\delta_0,\theta)\}$$

has γ measure zero.

In what follows, g denotes a non-negative measurable function defined on Θ. Given a $g \geq 0$ which satisfies

$$\int g(\theta)\gamma(d\theta) = 1,$$

a decision rule δ_g is a Bayes rule for the prior distribution $\pi(d\theta) = g(\theta)\gamma(d\theta)$ if

$$\int R(\delta,\theta)g(\theta)\gamma(d\theta) \geq \int R(\delta_g,\theta)g(\theta)\gamma(d\theta).$$

for all $\delta \in D$. Bayes rules are assumed to exist for each such density function g.

For each set C of positive γ measure, let

$$G(C) = \{g \mid g \geq 0, \int g(\theta)\gamma(d\theta) = 1, \int_C g(\theta)\gamma(d\theta) > 0\}. \tag{3.1}$$

Here is a sufficient condition for $\delta_0 \in D$ to be almost γ-admissible.

Theorem 3.1 (Stein): Let $\delta_0 \in D$ and for each density function g, let

$$\Delta(g) = \int [R(\delta_0,\theta)-R(\delta_g,\theta)]g(\theta)\gamma(d\theta). \tag{3.2}$$

For each set C of positive Υ-measure, assume that

$$\inf_{g\in G(C)} \frac{\Delta(g)}{\int_C g(\theta)\Upsilon(d\theta)} = 0. \tag{3.3}$$

Then δ_0 is almost Υ-admissible.

Proof: Assume δ_0 is not almost Υ-admissible. Then there exists a decision rule δ_1 which satisfies $R(\delta_1,\theta) \leq R(\delta_0,\theta)$ for all θ and the set

$$C_1 = \{\theta | R(\delta_1,\theta) < R(\delta_0,\theta)\}$$

has positive Υ-measure. Hence there exists and $\varepsilon > 0$ such that the set

$$C_2 = \{\theta | R(\delta_1,\theta) \leq R(\delta_0,\theta)-\varepsilon\}$$

has positive Υ-measure. Thus, for $g \in G(C_2)$,

$$\Delta(g) = \int [R(\delta_0,\theta)-R(\delta_g,\theta)]g(\theta)\Upsilon(d\theta) \geq \int [R(\delta_0,\theta)-R(\delta_1,\theta)]g(\theta)\Upsilon(d\theta)$$
$$\geq \int_{C_2} [R(\delta_0,\theta)-R(\delta_1,\theta)]g(\theta)\Upsilon(d\theta) \geq \varepsilon\int_{C_2} g(\theta)\Upsilon(d\theta).$$

Thus (3.3) cannot hold for $C = C_2$, so δ_0 is almost Υ-admissible. □

In applications, the goal is to find analytically tractable upper bounds for $\Delta(g)$ in (3.3) which yield checkable conditions in concrete problems. The results in the next section are directed to this goal for the fair Bayes prediction problem.

Finally, an inequality relating variation distance and Hellinger distance due to Kraft (1955) is needed.

Lemma 3.1: (Kraft (1955)). Let P_1 and P_2 be probabilities defined on a common space and suppose λ_0 dominates P_1 and P_2. Let $p_i = dP_i/d\lambda_0$, $i = 1,2$. Then

$$[V(P_1,P_2)]^2 \leq 2[1 - \int (p_1 p_2)^{1/2} d\lambda_0]. \tag{3.4}$$

Proof: The proof consists of two applications of the Cauchy-Schwarz inequality. Using (2.11) with $\lambda_0(dx)$ omitted for notational convenience,

$$\begin{aligned}[V(P_1,P_2)]^2 &= [\tfrac{1}{2}\int |p_1 - p_2|]^2 = [\tfrac{1}{2}\int |\sqrt{p_1} - \sqrt{p_2}|\,|\sqrt{p_1} + \sqrt{p_2}|]^2 \\ &\leq \tfrac{1}{4}[\int (\sqrt{p_1} - \sqrt{p_2})^2][\int (\sqrt{p_1} + \sqrt{p_2})^2] = [1 - \int (p_1 p_2)^{1/2}][1 + \int (p_1 p_2)^{1/2}] \\ &\leq 2[1 - \int (p_1 p_2)^{1/2}]. \quad \square\end{aligned}$$

4. A condition for Almost Admissibility:

The notation used in Section 2 is to hold throughout this section. It is assumed that λ is a σ-finite measure on (Ψ, B_1) which dominates $P(\cdot|\theta,z)$. The

density of $P(\cdot|\theta,z)$ with respect to λ is $p(\cdot|\theta,z)$. Hence the conditional density of X given θ is

$$p(\cdot|\theta) = \int p(\cdot|z,\theta)S(dz|\theta). \tag{4.1}$$

We are now ready to discuss the main problem of this paper--namely, to provide some conditions under which the formal posterior distribution of Z derived from an improper prior distribution is an almost admissible decision rule. To give a more precise description of the problem, let ν be a fixed σ-finite measure on (Θ,B_3) such that $\nu(\Theta) = +\infty$. With

$$m(x) = \int p(x|\theta)\nu(d\vartheta), \tag{4.2}$$

let

$$A = \{x|0 < m(x) < +\infty\}.$$

It is assumed that

$$\int_{A^c} p(x|\theta)\lambda(dx) = 0, \quad \theta \in \Theta. \tag{4.3}$$

Under assumption (4.3), the formal posterior distribution of Z given X is defined as follows. First let μ be the σ-finite measure on $(\Phi\times\Theta,B_2\times B_3)$ defined

by

$$\mu(dx,d\theta) = S(dz|\theta)\nu(d\theta). \tag{4.4}$$

Let $r_0(z,\theta)$ be an arbitrary density with respect to μ, and let

$$r(z,\theta|x) = \begin{cases} \dfrac{p(x|z,\theta)}{m(x)} & \text{if } x \in A \\ r_0(z,\theta) & \text{if } x \notin A \,. \end{cases} \tag{4.5}$$

For each x, $r(\cdot,\cdot|x)$ is a density on $\Phi\times\Theta$ with respect to μ. Let $\tilde{Q}_\nu(\cdot|x)$ denote the probability measure on $\Phi\times\Theta$ defined by the density in (4.5) and let $Q_\nu(\cdot|x)$ be the induced marginal probability on Φ obtained from $\tilde{Q}_\nu(\cdot|x)$. The measure $Q_\nu(\cdot|x)$ is what is called the formal posterior distribution of Z given X = x. The main result in this section gives a sufficient condition that the decision rule (inference) Q_ν be almost ν-admissible.

Now, let g be any density on Θ with respect to ν so

$$m_g(x) = \int p(x|\theta)g(\theta)\nu(d\theta) \tag{4.6}$$

is the marginal density of X with respect to λ. With

$$A_g = \{x|0 < m_g(x) < +\infty\}, \tag{4.7}$$

it is clear that

$$(4.8) \qquad r_g(z,\theta|x) = \begin{cases} \dfrac{p(x|z,\theta)g(\theta)}{m_g(x)}, & x \in A_g \\ r_0(z,\theta), & x \notin A_g \end{cases}$$

is a version of the conditional density of (Z,θ) given $X = x$, with respect to μ. This density defines the probability measure $\tilde{Q}_g(\cdot|x)$ on $\Phi\times\Theta$ which in turn induces the marginal distribution $Q_g(\cdot|x)$ on Φ. Thus $Q_g(\cdot|x)$ is the conditional distribution of Z given $X = x$, so the decision rule Q_g is the Bayes rule corresponding to the prior distribution $p(d\theta) = g(\theta)\nu(d\theta)$ (for the loss functions L_1 and L_2). In order to apply Theorem 3.1, let

$$(4.9) \qquad \Delta(g) = \int [R(Q_\nu,\theta)-R(Q_g,\theta)]g(\theta)\nu(d\theta)$$

where the risk function R is computed under either loss function L_1 or L_2. The function

$$(4.10) \qquad t(\theta,\eta) = \int \frac{p(x|\theta)p(x|\eta)}{m(x)}\,\lambda(dx)$$

plays an important role in what follows. Because of assumption (4.3), (4.10) is well defined.

Lemma 4.1: If the density g is strictly positive on Θ, then

(4.11) $$\Delta(g) \leq 8K\rho(g)$$

where

(4.12) $$\rho(g) = 1 - \iint \sqrt{g(\theta)}t(\theta,\eta)\sqrt{g(\eta)}\nu(d\theta)\nu(d\eta)$$

and K is the constant in Proposition 2.2.

Proof: With V denoting variation distance, first observe that

(4.13) $$V(Q_\nu(\cdot|x),Q_g(\cdot|x)) \leq V(\tilde{Q}_\nu(\cdot|x),\tilde{Q}_g(\cdot|x))$$

since $Q_\nu(\cdot|x)$ [resp. $Q_g(\cdot|x)$] is the marginal distribution of $\tilde{Q}_\nu(\cdot|x)$ [resp. $\tilde{Q}_g(\cdot|x)$]. Now, apply Proposition 2.2 with $\Delta = \Delta(g)$, $\delta_x = Q_\nu(\cdot|x)$ and $Q_\pi(\cdot|x) = Q_g(\cdot|x)$. This and (4.13) yield

(4.14) $$\begin{aligned}\Delta(g) &\leq 4K\iiint[V(\tilde{Q}_\nu(\cdot|x),\tilde{Q}_g(\cdot|x))]^2P(dx|z,\theta)S(dz|\theta)g(\theta)\nu(d\theta)\\ &= 4K\iint[V(\tilde{Q}_\nu(\cdot|x),\tilde{Q}_g(\cdot|x))]^2p(x|\theta)g(\theta)\lambda(dx)\nu(d\theta).\end{aligned}$$

For $x \in A \cap A_g$, Lemma 3.1 yields

$$(4.15)\qquad [V(\tilde{Q}_\nu(\cdot|x),\tilde{Q}_g(\cdot|x))]^2 \le 2\Big[1 - \iint \frac{p(x|z,\theta)\sqrt{g(\theta)}}{\sqrt{m(x)m_g(x)}}\,\mu(dz,d\theta)\Big]$$

$$= 2\Big[1 - \int \frac{p(x|\theta)\sqrt{g(\theta)}}{\sqrt{m(x)m_g(x)}}\,\nu(d\theta).$$

However, assumption (4.3) together with the assumption that g is strictly positive implies that

$$(4.16)\qquad \int_{A\cap A_g} p(x|\theta)\lambda(dx) = 1, \qquad \theta \in \Theta.$$

Thus the upper bound in (4.15) may be substituted into the right most expression in (4.14) to yield

$$(4.17)\qquad \Delta(g) \le 8K\Big[1 - \iiint \frac{p(x|\eta)\sqrt{g(\eta)}}{\sqrt{m(x)m_g(x)}}\,p(x|\theta)g(\theta)\nu(d\theta)\nu(d\eta)\lambda(dx)\Big]$$

$$= 8K\Big[1 - \iint \frac{p(x|\eta)\sqrt{g(\eta)}}{m(x)}\sqrt{m(x)m_g(x)}\,\nu(d\eta)\lambda|dx)\Big].$$

However, the Cauchy-Schwarz inequality gives

$$(4.18)\qquad \int \sqrt{g(\theta)}\,p(x|\theta)\nu d(\theta) \le \sqrt{m(x)m_g(x)}.$$

Substitution of the lower bound in (4.18) into the final expression in (4.17) yields

$$(4.19)\qquad \Delta(g) \le 8K\Big[1 - \iiint \sqrt{g(\theta)}\,\frac{p(x|\theta)p(x|\eta)}{m(x)}\,\sqrt{g(\eta)}\;\lambda(dx)\nu(d\theta)\nu(d\eta)\Big] = 8K\rho(g).$$

This completes the proof. □

For a set C of positive ν measure, define G(C) by

$$(4.20)\qquad G(C) = \{g \mid g \ge 0,\ \int g(\theta)\nu(d\theta) = 1,\ \int_C g(\theta)\nu(d\nu) > 0\}.$$

Also, observe that

$$(4.21)\qquad G^{+} = \{g \mid g(\theta) > 0,\ \theta \in \Theta;\ \int g(\theta)\nu(d\theta) = 1\}$$

is a subset of G(C).

Theorem 4.1: If for each set C of positive ν measure,

$$(4.22)\qquad \inf_{g\in G(C)} \frac{\rho(g)}{\int_C g(\theta)\nu(d\theta)} = 0,$$

then Q_ν is almost ν-admissible.

Proof: First consider the condition

(4.23) $$\inf_{g \in G^+} \frac{\rho(g)}{\int_C g(\theta)\nu(d\theta)} = 0.$$

Since $G^+ \subseteq G(C)$, (4.23) implies (4.22). Conversely, if (4.22) holds, then (4.23) holds. To see this, let $\varepsilon > 0$ be given so there exists a $g_0 \in G(C)$ such that

(4.24) $$\frac{\rho(g_0)}{\int_C g_0(\theta)\nu(d\theta)} < \varepsilon/2.$$

Fix $\tilde{g}_0 \in G^+$ and consider

$$g_n = (1 - \frac{1}{n})g_0 + \frac{1}{n}\tilde{g}_0, \quad n = 2,3,\ldots .$$

Then $g_n \in G^+$ for all n and an easy application of the Dominated Convergence Theorem shows that

$$\lim_{n\to\infty} \frac{\rho(g_n)}{\int_C g_n(\theta)\nu(d\theta)} = \frac{\rho(g_0)}{\int_C g_0(\theta)\nu(d\theta)}.$$

Thus the inf in (4.23) is bounded above by ε. Hence (4.23) holds.

For $g \in G^+$, Lemma 4.1 shows that $\Delta(g) \leq 8K\rho(g)$. Therefore,

$$\inf_{g\in G(C)} \frac{\Delta(g)}{\int_C g(\theta)\nu(d\theta)} \leq \inf_{g\in G^+} \frac{\Delta(g)}{\int_C g(\theta)\nu(d\theta)} \leq \inf_{g\in G^+} \frac{8K\rho(g)}{\int_C g(\theta)\nu(d\theta)} = 0.$$

That Q_ν is almost ν-admissible follows from Theorem 3.1. □

In some applications it is useful to have condition (4.22) expressed in terms of element in the Hilbert space $L^2(\nu)$ of square integrable functions on (Θ,B_3,ν). The inner product and norm on $L^2(\nu)$ are $(\cdot,\cdot)$ and $||\cdot||$. For a set C of positive ν-measure, let

$$H(C) = \{h \mid h \geq 0, \quad h \in L^2(\nu), \quad \int_C h^2(\theta)\nu(d\theta) > 0\}. \tag{4.25}$$

Also, for $h \in L^2(\nu)$, $h \geq 0$, define $\phi(h)$ by

$$\phi(h) = ||h||^2 - \iint h(\theta)t(\theta,\eta)h(\eta)\nu(d\theta)\nu(d\eta) \tag{4.26}$$

where t is given in (4.10).

Theorem 4.2: If, for each set C of positive ν measure,

$$\inf_{h\in H(C)} \frac{\phi(h)}{\int_C h^2(\theta)\nu(d\theta)} = 0 \tag{4.27}$$

then Q_ν is almost ν-admissible.

Proof: A routine argument shows that (4.27) and (4.22) are equivalent. □

5. The 1-Dimensional Translation Parameter Case:

In this section, Theorem 3.1 is applied to a special problem involving a one dimensional translation parameter problem. When the improper prior distribution $\nu(d\theta)$ is Lebesgue measure on R^1 (= Θ), sufficient conditions are given so that the induced posterior distribution is almost ν-admissible.

With $\Theta = R^1$ and $\Psi = R^n$, suppose $X \in R^n$ has a marginal density (given θ)

$$p(x|\theta) = f(x-\theta e), \qquad x \in R^n \tag{5.1}$$

where e is the vector of ones in R^n. The dominating measure λ is assumed to satisfy

$$\lambda(B+\theta e) = \lambda(B), \qquad \theta \in R^1 \tag{5.2}$$

for all Borel sets $B \subseteq R^n$. For example, λ might be Lebesgue measure on R^n.

The improper prior under consideration is $\lambda(d\theta) = d\theta$--Lebesgue measure on R^1. Thus

$$m(x) = \int f(x-\theta e)d\theta \tag{5.3}$$

and the set

$$A = \{x \mid 0 < m(x) < +\infty\}$$

is assumed to satisfy

$$\int_{A^c} f(x-\theta e)\lambda(dx) = 0, \quad \theta \in R^1. \tag{5.4}$$

Because

$$m(x) = m(x+\alpha e), \quad \alpha \in R^1, \tag{5.5}$$

(5.5) is equivalent to

$$\int_{A^c} f(x)\lambda(dx) = 0. \tag{5.6}$$

This condition needs to be checked in each example, but is satisfied in all the interesting examples that I know.

Equation (5.5) implies that

$$t(\theta,\eta) = \int \frac{f(x-\theta e)f(x-\eta e)}{m(x)} \lambda(dx) \tag{5.7}$$

satisfies

(5.8) $$t(\theta+\alpha,\eta+\alpha) = t(\theta,\eta), \quad \alpha \in R^1$$

so that

(5.9) $$t(\theta,\eta) = t(\theta-\eta,0).$$

Thus, set

(5.10) $$t_0(u) = t(u,0), \quad u \in R^1$$

so $t_0(\theta-\eta) = t(\theta,\eta)$. Since t is a symmetric function of its arguments, it follows that $t_0(u) = t_0(-u)$. Also, it is easily verified that

(5.11) $$\int t_0(\theta)d\theta) = 1.$$

Theorem 5.1: If

(5.12) $$\int |\theta| t_0(\theta)d\theta) < +\infty,$$

then (4.27) holds.

Proof: Let C be a set of positive Lebesgue measure, and consider the sequence of elements $r_n \in L^2(\nu)$ defined by

(5.13) $$r_n(\theta) = \beta\left(\frac{\theta}{n}\right)$$

where $\beta(u) = (1+u^2)^{-1}$. Since $r_n(\theta) \nearrow 1$ as $n \longrightarrow \infty$, the Monotone Convergence Theorem shows that

$$\lim_{n\longrightarrow\infty} \int_C r_n^2(\theta)d\theta) = \int_C d\theta > 0.$$

Thus, to verify (4.27), it is sufficient to show that

(5.14) $$a_n = ||r_n||^2 - \iint r_n(\theta)t_0(\theta-\eta)r_n(\eta)d\theta d\eta$$

converges to zero. A bit of algebra and a change of variable show that

(5.15) $$a_n = \iint \beta(\theta)t_0(u)n\left[\beta(\theta)-\beta\left(\theta+\frac{u}{n}\right)\right]dud\theta.$$

But for all n, θ, and u, the inequality

$$\left|n\left[\beta(\theta)-\beta\left(\theta+\frac{u}{n}\right)\right]\right| \leq K_0|u|,$$

where K_0 is a constant, is easily verified. Since

$$\lim_{n\longrightarrow\infty} n\left[\beta(\theta)-\beta\left(\theta+\frac{u}{n}\right)\right] = -u\beta'(\theta),$$

the Dominated Convergence Theorem yields $a_n \longrightarrow 0$. This completes the proof. □

Usable sufficient conditions so that (5.12) holds, expressed in terms of f, can be given. For example, suppose $X_1, \ldots, X_n$ are i.i.d. from a one dimensional translation family. Then

$$f(x-\theta e) = \prod_1^n f_0(x_i-\vartheta) \tag{5.16}$$

where f_0 is a density on R^1. Assume that $\lambda(dx) = dx$--Lebesgue measure on R^n. Let k be the greatest integer in (n+1)/2 and let $X_{(k)}$ be the k^{th} coordinate of the order statistic $X_{(1)}, \ldots, X_{(n)}$. Define $W \in R^{n-1}$ by

$$W_i = \begin{cases} X_{(i)} - X_{(k)}, & i = 1, \ldots, k-1 \\ X_{(i+1)} - X_{(k)}, & i = k, \ldots, n-1. \end{cases} \tag{5.17}$$

Let $X_{(k)}$ and $\tilde{X}_{(k)}$ be conditionally independent given W, both with the conditional distribution of $X_{(k)}$ given W. Then, by just looking at the integrals involved,

$$\int |\vartheta| t_0(\vartheta)d\vartheta) = E(E|X_{(k)}-\tilde{X}_{(k)}||W) \tag{5.18}$$

which is bounded above by

(5.19) $$2E(|X_k||W) = 2E|X_{(k)}|.$$

Thus, if $E|X_{(k)}| < +\infty$, then (5.12) holds. For example, if f_0 has a mean then $E|X_{(k)}| < +\infty$. However, f_0 may fail to have a mean, but $E|X_{(k)}|$ may be finite. In particular, if $n \geq 3$ and f_0 corresponds to a Cauchy distribution, then $E|X_{(k)}| < +\infty$.

The application of the above results to prediction problems when $X \in R^n$ and $Z \in R^1$ are dependent runs as follows. Assume the joint density of X and Z is

(5.20) $$\tilde{f}(x-\theta e, z-\theta)$$

with respect to $\lambda x dz$ on $R^n \times R^1$. Here λ is as above and dz is Lebesgue measure on R^1. Now, just apply the previous argument to

$$f(x-\theta e) = \int_{-\infty}^{\infty} \tilde{f}(x-\theta e, z-\theta)dz = \int_{-\infty}^{\infty} \tilde{f}(x-\theta e, z)dz.$$

For example, assume that (X,Z) is jointly multivariate normal, say

(5.21) $$L\left(\begin{pmatrix} X \\ Z \end{pmatrix}|\theta\right) = N(\theta\tilde{e}, \Sigma)$$

where $\theta \in R^1$ is unknown, $\tilde{e}$ is the vector of ones in R^{n+1} and Σ is a known $(n+1)\times(n+1)$ positive definite covariance matrix. To predict $Z \in R^1$ from $X \in R^n$,

take the improper prior $d\theta$. Then, a routine but tedious calculation gives the predictive distribution

$$L(Z|X) = N(\mu(X),\sigma^2) \tag{5.22}$$

where $\mu(x)$ and σ^2 can be calculated as follows. First, let

$$A = \Sigma^{-1} - \frac{\Sigma^{-1}\tilde{e}\tilde{e}'\Sigma^{-1}}{\tilde{e}'\Sigma^{-1}\tilde{e}} : (n+1)\times(n+1) \tag{5.23}$$

and partition A as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \tag{5.24}$$

where A_{11} is $n\times n$ and $A_{22} \in (0,\infty)$. Then,

$$\mu(X) = -\frac{A_{21}X}{A_{22}} \tag{5.25}$$

and

$$\sigma^2 = A_{22}^{-1}. \tag{5.26}$$

The decision rule specified by (5.22) is almost "$d\theta$"-admissible for all of the

loss functions of the type L_1 and L_2 described in Section 2. It follows that for such loss functions, if these loss functions are also translation invariant, then (5.22) is a minimax decision rule (see Kiefer (1957)).

6. Discussion:

There are some special cases of the prediction problem described in Section 2 which are of particular interest. As mentioned earlier, when $(\Phi,\mathcal{B}_2) = (\Theta,\mathcal{B}_3)$ and $S(\cdot|d\theta)$ is the probability measure degenerate at θ, then the prediction problem is just the estimation problem as described in Eaton (1982). In this case, Theorem 4.1 gives a sufficient condition that the posterior distribution on Θ derived from ν be almost ν-admissible for all loss functions of the type L_1.

Now, assume that given θ, Z and X are independent and Z is to be predicted. Then the density of X does not depend on z so the density is (4.1) is just the density of the data, $p(x|\theta)$. Thus, the conditions of Theorem 4.1 depend only on the density of the data and these conditions are the same as in the estimation problem. Hence when X and Z are independent (given θ), the prediction problem is no harder, nor easier, than the estimation problem.

The problem of estimating a function of θ, say $\rho(\theta)$, is cast into our framework by taking Φ to be the range of ρ and $S(\cdot|\theta)$ to be degenerate at $\rho(\theta)$. Again, Theorem 4.1 applies directly.

In a number of interesting examples, the following two conditions hold:

(6.1) The risk function $R(\delta,\cdot)$ is continuous on Θ for all decision rules

δ.

(6.2) The improper prior gives positive measure to all non-empty open subsets of θ.

When (6.1) and (6.2) hold, it is clear that almost ν-admissibility implies admissibility. A sufficient condition that (6.1) and (6.2) hold for loss functions L_1 and L_2 is that X and Z are independent (given θ), and the distributions of X and Z are exponential families.

The successful application of Theorem 4.1 to problems more complicated than those treated in Section 5 depends on obtaining information concerning the behavior of the function $t(\cdot,\cdot)$ defined in (4.10). It is possible that the theory of Markov chains on general state spaces (see Nummelin (1983)) may be of use in this regard. To see this, first observed that T defined on $B_3 \times \Theta$ by

$$T(C|\theta) = \int_C t(\theta,\eta)\nu(d\eta) \tag{6.3}$$

is a probability measure on B_3 for each fixed θ and is a measurable function of θ for each $C \in B_3$. Thus T is a Markov kernel and hence defines a discrete time Markov chain on Θ. Obviously $t(\cdot,\cdot)$ is the transition density of T with respect to ν. Using that fact the $t(\cdot,\cdot)$ is a symmetric function of its arguments, the author has been able to show that under certain regularity conditions, (4.27) is equivalent to the Harris recurrence of the Markov chain (see Nummielin (1983)

for a discussion of recurrence). A report on this work is in preparation (see Eaton (1987)). The relationship between the above observation and the admissibility-Markov process connections for special loss functions established in Brown (1971) and Johnstone (1984) is quite unclear at this point. For example the Markov chain here occurs naturally on the parameter space Θ where as the Markov process which occurs in Johnstone (1984) is constructed on the sample space.

REFERENCES

Aitchison, J. and Dunsmore, I.R. (1975). Statistical Prediction Analysis, Cambridge University Press, New York.

Billingsley, P. (1968). Convergence of Probability Measures. Wiley, New York.

Brown, L.D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. Ann. Math. Statist. **42**, 855-903. Correction, Ann. Statist. 1, 594-596.

Dubins, L. and Freedman, D. (1964). Measurable sets of Measures. Pacific J. Math. **14**, 1211-1222.

Eaton, M.L. (1982). A method for evaluating improper prior distributions. Decision Theorem and Related Topics III, Vol. 1, ed. by Gupta and Berger, p. 329-352.

Geisser, S. (1980). A predictivistic primer. In Bayesian Analysis in Econometrics and Statistics, ed. A. Zellner, North Holland, Amsterdam, 363-381.

Johnstone, I. (1984). Admissibility, difference equations and recurrence in estimating a Poisson mean. Ann. Statist. **12**, 1173-1198.

Kiefer, J. (1957). Invariance, minimax sequential estimation and continuous time processes. Ann. Math. Statist. **28**, 573-601.

Kraft, C. (1955). Some conditions for consistence and uniform consistency of statistical procedures. Univ. California Publications in Statist. Vol. 2, 125-142.

Nummelin, E. (1984). General Irreducible Markov Chains and Non-negative Operators. Cambridge University Press, New York.

Parthassrathy, K.R. (1967). Probability Measures on Metric Spaces. Academic Press, New York.

Stein, C. (1955). A necessary and sufficient condition for admissibility. Ann. Math. Statist. **26**, 518-522.

Stein, C. (1965). Lecture notes on decision theory. Unpublished.