

Admissible Predictive Density Estimation ¹

By Lawrence D. BROWN, Edward I. GEORGE and Xinyi XU ²

University of Pennsylvania, University of Pennsylvania, Ohio State University

February 20, 2006

Abstract

Let $X | \mu \sim N_p(\mu, v_x I)$ and $Y | \mu \sim N_p(\mu, v_y I)$ be independent p -dimensional multivariate normal vectors with common unknown mean μ . Based on observing $X = x$, we consider the problem of estimating the true predictive density $p(y | \mu)$ of Y under expected Kullback-Leibler loss. Our focus here is the characterization of admissible procedures for this problem. We show that the class of all generalized Bayes rules is a complete class, and that the easily interpretable conditions of Brown and Hwang (1982) are sufficient for a formal Bayes rule to be admissible.

Keywords: ADMISSIBILITY; BAYESIAN PREDICTIVE DISTRIBUTION; COMPLETE CLASS; PRIOR DISTRIBUTIONS.

¹Supported by NSF grant DMS-0130819.

²Lawrence D. Brown and Edward I. George are Professors, Statistics Department, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340, lbrown@wharton.upenn.edu, edgeorge@wharton.upenn.edu. Xinyi Xu is Assistant Professor, Department of Statistics, The Ohio State University, Cockins Hall, Columbus, OH 43210-1247 and xinyi@stat.ohio-state.edu.

1 Introduction

Let $X | \mu \sim N_p(\mu, v_x I)$ and $Y | \mu \sim N_p(\mu, v_y I)$ be independent p -dimensional multivariate normal vectors with a common unknown mean $\mu \in R^p$. We assume that $v_x > 0$ and $v_y > 0$ are known. We let $p(x | \mu)$ and $p(y | \mu)$ denote the conditional densities of X and Y , suppressing the dependence on v_x and v_y throughout.

Based on observing only $X = x$, we consider the problem of estimating the density $p(y | \mu)$ of Y . The natural action space \mathcal{A}_0 consists of all proper densities on R^p , i.e.

$$\mathcal{A}_0 = \{g : R^p \rightarrow R \text{ such that } g(y) \geq 0 \text{ and } \int g(y) dy = 1\}. \quad (1)$$

For each observation $x \in R^p$, a (nonrandomized) decision procedure $\hat{p}(\cdot | x) : R^p \rightarrow \mathcal{A}_0$ chooses a $g \in \mathcal{A}_0$.

We measure the goodness of fit of $g(y)$ to $p(y | \mu)$ by Kullback-Leibler (KL) loss

$$L(\mu, g) = \begin{cases} \int p(y | \mu) \log \frac{p(y | \mu)}{g(y)} dy & \text{if } g(y) > 0 \text{ a.e.} \\ \infty & \text{otherwise} \end{cases} \quad (2)$$

and evaluate a procedure $\hat{p}(\cdot | x)$ by its risk function

$$R_{KL}(\mu, \hat{p}) = \int L(\mu, \hat{p}(\cdot | x)) p(x | \mu) dx. \quad (3)$$

For the comparison of two (nonrandomized) procedures, we say that \hat{p}_1 dominates \hat{p}_2 if $R_{KL}(\mu, \hat{p}_1) \leq R_{KL}(\mu, \hat{p}_2)$ for all μ and with strict inequality for some μ . A procedure $\hat{p}(\cdot | x)$ is called admissible if it cannot be dominated by any other procedures.

Two widely used methods to obtain predictive densities are “plug-in” rules and Bayes rules. A plug-in rule

$$\hat{p}_{\hat{\mu}}(y | x) = p(y | \mu = \hat{\mu}(x)) \quad (4)$$

simply substitutes an estimate $\hat{\mu}$ for μ in $p(y | \mu)$. In contrast, a Bayes rule integrates μ out with respect to a non-negative and locally finite prior measure M to obtain

$$\hat{p}_M(y | x) = \frac{\int p(x | \mu) p(y | \mu) M(d\mu)}{\int p(x | \mu) M(d\mu)} = \int p(y | \mu) M(d\mu | x). \quad (5)$$

When writing an expression such as (5), we implicitly assume that the denominator in the middle expression is finite for all x , and hence all terms in (5) are finite for all x .

We use the symbol π to denote the density of M when it exists, and will write either \hat{p}_π or \hat{p}_M in that case.

Aitchison (1975) showed that for proper M , $\hat{p}_M(y | x)$ minimizes the average KL risk

$$B_{KL}(M, \hat{p}) = \int R_{KL}(\mu, \hat{p})M(d\mu). \quad (6)$$

Aitchison also showed that the (formal) Bayes rule (5) under the uniform prior density $\pi_U(\mu) = 1$, namely $\hat{p}_{\pi_U}(y | x)$, dominates the plug-in rule $p(y | \hat{\mu}_{MLE})$, which substitutes the maximum likelihood estimate $\hat{\mu}_{MLE} = x$ for μ . Indeed, as will be seen in section 3, all the admissible procedures for multivariate normal density prediction under KL loss are Bayes rules in the sense of (5).

The constant risk Bayes rule \hat{p}_{π_U} is best invariant, minimax, admissible when $p = 1$, (Murray 1977, Ng 1980 and Liang and Barron 2003), and as we shall show in Section 2, admissible when $p = 2$. However, it is inadmissible when $p \geq 3$. This was first established by Komaki (2001) who showed that \hat{p}_{π_U} is dominated by the Bayes rule under the (nonconstant) harmonic prior when $p \geq 3$. Liang (2002) further showed that \hat{p}_{π_U} is dominated by proper Bayes rules under Strawderman priors when $p \geq 5$.

It is interesting to note the parallels between our predictive density estimation problem and the problem of estimating a multivariate normal mean under quadratic loss. Based on observing $Z | \mu \sim N_p(\mu, vI)$ with v known, this latter problem is to estimate μ under quadratic risk

$$R_Q^v(\mu, \hat{\mu}) = E_\mu \|\hat{\mu} - \mu\|^2, \quad (7)$$

where the dependence of R_Q^v on v is indicated by the superscript v . Here the maximum likelihood estimator $\hat{\mu}_{MLE}$, which is best invariant, minimax and admissible when $p = 1$ or 2, is dominated when $p \geq 3$ by the Bayes rules $\hat{\mu}_\pi = \int \mu \pi(\mu | x) d\mu$ under the harmonic prior (Stein 1974) and under the Strawderman priors (Strawderman 1971). Note that in the KL risk problem $\hat{p}_{\pi_U}(y | x)$, rather than $\hat{p}(y | \hat{\mu}_{MLE})$, plays the same role as $\hat{\mu}_{MLE}$ in the quadratic risk problem. Recall that $\hat{\mu}_{MLE}$ can also be motivated as the Bayes rule under $\pi_U(\mu) = 1$ in the quadratic risk problem.

George, Feng and Xu (2005) recently drew out these parallels between the KL risk and quadratic risk problems, and found that they could be explained by connections between unbiased estimates of risk. These connections were shown to yield analogous

sufficient conditions for the minimaxity of Bayes rules in both problems. In this paper, we establish further parallels concerning the characterization of admissibility in both problems. As proper Bayes rules are easily shown to be admissible in the KL setting, see Berger (1985), our focus will be on improper π under which $\hat{p}_\pi(y | x)$ is sometimes more precisely called a formal or generalized Bayes rule. In section 2, we establish analogous sufficient conditions for the admissibility of Bayes rules $\hat{p}_\pi(y | x)$ under KL loss, by extending the approach of Brown(1971) and Brown and Huang (1982). In section 3, we prove that all admissible procedures for the KL risk problems are Bayes rules, a direct parallel of the complete class theorem of Brown (1971) for quadratic risk.

It might be of interest to note that when $v_y \rightarrow 0$, $p(y | \mu)$ degenerates to a point mass $I\{y = \mu\}$ and that by (5),

$$\hat{p}_\pi(y | x) = \int p(y | \mu)\pi(\mu | x)d\mu \rightarrow \pi(y | x).$$

Therefore, the limiting KL risk of a Bayes rule \hat{p}_π is

$$\lim_{v_y \rightarrow 0} R_{KL}(\mu, \hat{p}_\pi) = E_\mu \left[I\{y = \mu\} \log \frac{I\{y = \mu\}}{\pi(y | X)} \right] = -E_\mu \log \pi(\mu | X),$$

where the right hand side can be viewed as the KL risk for “estimating a point mass at μ ” by a posterior density. Thus, our setup can provide a decision theoretic framework for evaluating a prior by the extent to which $E_\mu \log \pi(\mu | X)$ is large for all μ .

2 Sufficient Conditions for Admissibility

For $Z \sim N(\mu, I)$, Brown (1971) and Brown and Hwang (1982) developed general sufficient conditions for the admissibility of formal Bayes rules for the quadratic risk problem. To utilize their results and obtain analogous sufficient conditions for the KL risk problem, we first establish a relationship between KL risk and quadratic risk. In this section, we assume that the prior measure M has a density π and that $R_{KL}(\mu, \hat{p}_\pi) < \infty$ for all $\mu \in R^p$. Let

$$m_\pi(z; v) = \int p(z | \mu)\pi(\mu)d\mu \tag{8}$$

be the marginal density of $Z \sim N(\mu, vI)$ under π .

Theorem 1. *Let π be a prior density on μ such that $m_\pi(z; v_x)$ is finite for all z . Then*

$$R_{KL}(\mu, \hat{p}_{\pi_U}) - R_{KL}(\mu, \hat{p}_\pi) = \frac{1}{2} \int_{v_w}^{v_x} \frac{1}{v^2} [R_Q^v(\mu, \hat{\mu}_{MLE}) - R_Q^v(\mu, \hat{\mu}_\pi)] dv \tag{9}$$

where $v_w = v_x v_y / (v_x + v_y) < v_x$.

Proof. Let $m_\pi(w; v_w)$ denote the marginal density under π of

$$W = \frac{v_y X + v_x Y}{v_x + v_y} \sim N_p(\mu, v_w I). \quad (10)$$

By Lemmas 2 and 3 of George, Liang and Xu (2005),

$$R_{KL}(\mu, \hat{p}_{\pi_U}) - R_{KL}(\mu, \hat{p}_\pi) = E_{\mu, v_w} \log m_\pi(W; v_w) - E_{\mu, v_x} \log m_\pi(X; v_x), \quad (11)$$

$m_\pi(z; v)$ is finite for any $v_w \leq v \leq v_x$, and

$$\frac{\partial}{\partial v} E_{\mu, v} \log m_\pi(Z; v) = E_{\mu, v} \left(2 \frac{\nabla^2 \sqrt{m_\pi(Z; v)}}{\sqrt{m_\pi(Z; v)}} \right) \quad (12)$$

where $E_{\mu, v}(\cdot)$ stands for expectation with respect to the $N(\mu, vI)$ distribution. Furthermore, Stein (1974, 1981) showed that for the quadratic risk problem

$$R_Q^v(\mu, \hat{\mu}_{MLE}) - R_Q^v(\mu, \hat{\mu}_\pi) = -4v^2 E_\mu \frac{\nabla^2 \sqrt{m_\pi(Z; v)}}{\sqrt{m_\pi(Z; v)}}. \quad (13)$$

Combining (11), (12) and (13), the lemma follows. \ddagger

Now let $B_{KL}(\pi, \hat{p}) = \int R_{KL}(\mu, \hat{p}) \pi(\mu) d\mu$ and $B_Q^v(\pi, \hat{p}) = \int R_Q^v(\mu, \hat{p}) \pi(\mu) d\mu$ be the average KL and quadratic risks over π . The following relationship between the average KL risk difference and the average quadratic risk difference of Bayes rules follows from (9) and averaging over a finite prior measure π_n .

Corollary 1. *Let π and π_n be priors on μ such that $m_\pi(z; v_x)$ and $m_{\pi_n}(z; v_x)$ are finite for all z . Furthermore, assume π_n is a finite measure. Then*

$$B_{KL}(\pi_n, \hat{p}_\pi) - B_{KL}(\pi_n, \hat{p}_{\pi_n}) = \frac{1}{2} \int_{v_w}^{v_x} \frac{1}{v^2} [B_Q^v(\pi_n, \hat{\mu}_\pi) - B_Q^v(\pi_n, \hat{\mu}_{\pi_n})] dv \quad (14)$$

Corollary 1 enables us extend the approach of Brown and Hwang (1982) to establish conditions for the admissibility of formal Bayes rules in the KL risk problem. As in Brown and Hwang (1982), we use Blyth's method which can be extended to any statistical estimation problem with a strictly convex loss function (Brown 1971).

Lemma 1. Let \hat{p} be such that $R_{KL}(\mu, \hat{p}) < \infty$ for all $\mu \in R^p$. If there exists a sequence of finite densities $\{\pi_n\}$ such that $\int_{\|\mu\| \leq 1} \pi_n(\mu) d\mu > c$ for some positive constant c , and

$$B_{KL}(\pi_n, \hat{p}) - B_{KL}(\pi_n, \hat{p}_{\pi_n}) \rightarrow 0 \quad (15)$$

then \hat{p} is admissible.

Proof. Suppose \hat{p} is not admissible. Then there is a \hat{p}' such that $R_{KL}(\mu, \hat{p}') \leq R_{KL}(\mu, \hat{p})$ with strict inequality for some μ . Let $\hat{p}'' = (\hat{p} + \hat{p}')/2$. Thus

$$\begin{aligned} R_{KL}(\mu, \hat{p}'') &= \int p(x | \mu) p(y | \mu) \left[\log \frac{p(y | \mu)}{\hat{p}''(y | x)} \right] dx dy \\ &= \int p(x | \mu) p(y | \mu) \left[\log p(y | \mu) - \log \left(\frac{1}{2} \hat{p}(y | x) + \frac{1}{2} \hat{p}'(y | x) \right) \right] dx dy \\ &< \int p(x | \mu) p(y | \mu) \left[\log p(y | \mu) - \frac{1}{2} (\log \hat{p}(y | x) + \log \hat{p}'(y | x)) \right] dx dy \\ &= \frac{1}{2} (R_{KL}(\mu, \hat{p}) + R_{KL}(\mu, \hat{p}')) \leq R_{KL}(\mu, \hat{p}) \end{aligned}$$

Since $R_{KL}(\mu, \hat{p})$ and $R_{KL}(\mu, \hat{p}'')$ are both continuous in μ , there exists an $\varepsilon > 0$ such that for all $\mu \in \{\mu : \|\mu\| \leq 1\}$,

$$R_{KL}(\mu, \hat{p}) - R_{KL}(\mu, \hat{p}'') \geq \varepsilon > 0.$$

Therefore, we have

$$B_{KL}(\pi_n, \hat{p}) - B_{KL}(\pi_n, \hat{p}_{\pi_n}) \geq B_{KL}(\pi_n, \hat{p}) - B_{KL}(\pi_n, \hat{p}'') \geq \varepsilon \cdot c > 0$$

which contradicts (15). The admissibility of \hat{p} follows. \ddagger

We assume without loss of generality that the coordinate system is chosen so that $\int_{\|\mu\| \leq 1} \pi(\mu) d\mu > 0$. Using Lemma 1, we extend the approach of Brown and Hwang (1982) to obtain the following.

Theorem 2. A formal Bayes rule \hat{p}_π is admissible under KL loss if for every $v \in [v_w, v_x]$, the improper π satisfies both

(i) the growth condition:

$$\int_{R^p - S} \frac{\pi(\mu)}{\|\mu\|^2 \ln^2(\|\mu\| \vee 2)} d\mu < \infty \quad (16)$$

where $S = \{\mu : \|\mu\| \leq 1\}$ and $a \vee b = \max\{a, b\}$, and

(ii) the asymptotic flatness condition:

$$\int \int \pi(\mu) \left\| \frac{m_{\nabla\pi}(z; v)}{m_{\pi}(z; v)} - \frac{\nabla\pi}{\pi} \right\|^2 p(z | \mu) d\mu dz < \infty, \quad (17)$$

Proof. For $v = 1$, Brown and Hwang showed that when the prior density π satisfies the growth condition (16) and the asymptotic flatness condition (17), there exists a sequence of densities $\{\pi_n\}$ such that $\int_{\|\mu\| \leq 1} \pi_n(\mu) d\mu = \int_{\|\mu\| \leq 1} \pi(\mu) d\mu > 0$ and that $B_Q(\pi_n, \hat{\mu}) - B_Q(\pi_n, \hat{\mu}_{\pi_n}) \rightarrow 0$. Furthermore, they showed that an explicit construction of such a sequence $\{\pi_n\}$ is obtained by defining

$$j_n(\mu) = \begin{cases} 1 & \|\mu\| \leq 1 \\ 1 - \frac{\ln(\|\mu\|)}{\ln(n)} & 1 \leq \|\mu\| \leq n \\ 0 & \|\mu\| \geq n \end{cases} \quad (18)$$

for $n = 2, 3, \dots$, and letting

$$\pi_n(\mu) = j_n^2(\mu) \pi(\mu). \quad (19)$$

It is straightforward to show that the above construction also works for general v . That is, for any v , if π satisfies conditions (16) and (17), then for the sequence $\{\pi_n\}$ obtained by (18) and (19), $\Delta_{n,v} \equiv [B_Q(\pi_n, \hat{\mu}_\pi) - B_Q(\pi_n, \hat{\mu}_{\pi_n})]_v \rightarrow 0$. It thus follows that if π satisfies conditions (16) and (17) for every $v \in [v_w, v_x]$, then by Corollary 1 and by the continuity in v of $\Delta_{n,v}$,

$$B_{KL}(\pi_n, \hat{p}_\pi) - B_{KL}(\pi_n, \hat{p}_{\pi_n}) = \frac{1}{2} \int_{v_w}^{v_x} \frac{1}{v^2} \Delta_{n,v} dv \rightarrow 0. \quad (20)$$

That \hat{p}_π is admissible now follows immediately from Lemma 1. ‡

Example 1. (Uniform prior) Let $\pi(\mu) = 1$ for any μ , then $\nabla\pi = 0$. In this case, the conditions of Theorem 2 are easy to verify when $p = 1$ or 2 . Therefore, the formal Bayes rule \hat{p}_{π_U} is admissible when $p = 1$ or 2 .

It was pointed out in Brown and Hwang (1982) that if

$$\pi(\mu) \leq \|\mu\|^{2-p}, \quad \frac{\nabla\pi(\mu)}{\pi(\mu)} = o(\|\mu\|^{-1}), \quad \text{and} \quad \left| \frac{\partial^2\pi(\mu)}{\partial\mu_i\partial\mu_j} \right| = o(\|\mu\|^{-2}), \quad (21)$$

then (16) is easy to check and (17) can be verified with some difficulty (extending Lemma 3.4.1 of Brown (1971)). Hence, by Theorem 2, the corresponding \hat{p}_π is admissible under KL loss.

Example 2. (Harmonic prior) Let $\pi_H(\mu) = \|\mu\|^{-(p-2)}$ for $p \geq 3$. Because this prior satisfies (21), the formal Bayes rule \hat{p}_{π_H} is admissible when $p \geq 3$.

The following corollary is similarly a straightforward extension from Brown and Hwang (1982). It replaces condition (17) of Theorem 2 with a condition that is slightly less general, but more transparent and easier to verify.

Corollary 2. *If an improper density π satisfies $\pi(\{\mu : \|\mu\| \leq 1\}) \geq 0$, and if π satisfies (16) and*

$$\int \frac{\|\nabla\pi(\mu)\|^2}{\pi(\mu)} d\mu < \infty, \quad (22)$$

then the formal Bayes rule \hat{p}_π is admissible under KL loss.

Finally, it was also pointed out in Brown and Hwang (1982) that if

$$\pi(\mu) \leq \|\mu\|^{2-p-\varepsilon} \text{ for some } \varepsilon > 0, \text{ and } \frac{\nabla\pi(\mu)}{\pi(\mu)} = o(\|\mu\|^{-1}), \quad (23)$$

then (16) and (22) are easy to check. Hence, by Corollary 2, the corresponding \hat{p}_π is admissible under KL loss.

3 A Complete Class Theorem

We now turn to establishing that all (generalized) Bayes rules form a complete class for the KL loss problem. In Section 3.1, we begin by first establishing properties of some modified action spaces and the KL loss function. We then make use of these properties in Section 3.2 where we prove our main complete class results.

3.1 Preliminary Lemmas

Because the true density $p(y|\mu)$ is bounded by a constant $C = (2\pi v_y)^{-p/2}$ for any μ , it will eventually be useful to restrict attention to bounded densities estimates. Let

$$\mathcal{A} = \{g : R^p \rightarrow R \text{ such that } 0 \leq g(y) \leq C \text{ a.e. and } \int g(y)dy = 1\}. \quad (24)$$

Obviously, $\mathcal{A} \subset \mathcal{A}_0$.

Furthermore, note that if $g \in \mathcal{A}$, then

$$L(\mu, g) \geq \int p(y|\mu) \log \frac{p(y|\mu)}{C} dy \geq -\frac{p}{2}$$

The fact that $L(\mu, g)$ is bounded below is important for justifying the use of Fubini's theorem that is implicit in some of the results established later.

The following lemma, which is proved in appendix, shows that no admissible actions are lost by restricting the action space to \mathcal{A} .

Lemma 2. *Suppose $g_0(\cdot) \in \mathcal{A}_0$. If $g_0 \notin \mathcal{A}$, i.e. $g_0 \geq C$ on a set $S \subset R^p$ with positive measure, then there exists a $g \in \mathcal{A}$ that dominates g_0 in the sense that $L(\mu, g_0) > L(\mu, g)$ for all μ .*

It will also be useful to consider extending \mathcal{A} to its closure

$$\mathcal{A}^* = \{g : R^p \rightarrow R \text{ such that } 0 \leq g(y) \leq C \text{ a.e. and } \int g(y)dy \leq 1\}, \quad (25)$$

and then to make use of the topological properties of \mathcal{A}^* . Because \mathcal{A}^* is a subset of the Banach space \mathcal{L}_∞ , we will consider the topology on \mathcal{A}^* induced by the weak* topology on \mathcal{L}_∞ . Under this weak* topology, a sequence $\{g_i\} \in \mathcal{A}^*$ converges to a $g \in \mathcal{A}^*$ if

$$\int f(y)g_i(y)dy \rightarrow \int f(y)g(y)dy, \quad \forall f \in \mathcal{L}_1 \quad (26)$$

We will eventually make use of the following properties of \mathcal{A}^* under the weak* topology.

Lemma 3. *Define the action space \mathcal{A}^* as in (25), then*

(i) \mathcal{A}^* is weak* compact.

(ii) The weak* topology on \mathcal{A}^* is metrizable by

$$\rho(g, h) = \sum_{k=1}^{\infty} 2^{-k} \left| \int [g(y) - h(y)] f_k(y) dy \right|, \quad \text{for any } g, h \in \mathcal{A}^* \quad (27)$$

where $\{f_k, k = 1, 2, \dots\}$ is a countable dense subset of \mathcal{L}_1 . And \mathcal{A}^* is separable and second countable under this metric (27).

(iii) Suppose $g^*(\cdot) \in \mathcal{A}^*$. If $g^* \notin \mathcal{A}$, then there exists a $g \in \mathcal{A}$ that dominates g^* in the sense that $L(\mu, g^*) > L(\mu, g)$ for all μ . Thus, the extension from \mathcal{A} to \mathcal{A}^* doesn't incur any new admissible actions.

Finally, we also need to make use of the following properties of the Kullback-Leibler loss function.

Lemma 4. For the KL loss function $L(\mu, \cdot)$ in (2),

(i) $L(\mu, \cdot)$ is lower semi-continuous, i.e. if $g_i \rightarrow g$ weak*, then

$$\liminf_{i \rightarrow \infty} L(\mu, g_i) \geq L(\mu, g), \quad \forall \mu \in R^p \quad (28)$$

(ii) $L(\mu, \cdot)$ is strictly convex on

$$\mathcal{A}_+^* = \{g : g \in \mathcal{A}^* \text{ and } L(\mu, g) < \infty \text{ for } \forall \mu\} \quad (29)$$

for any $\mu \in R^p$.

3.2 The Main Theorems

Having established Lemma 2, 3 and 4 in Section 3.1, we are now ready to prove that all admissible procedures for the normal density prediction problem under KL loss are (generalized) Bayes rules. This proof consists of three steps:

(i) All the admissible procedures are non-randomized.

(ii) For any admissible procedure $\hat{p}(\cdot | x)$, there exists a sequence of priors $M_i(\mu)$ such that $\hat{p}_{M_i}(\cdot | x) \rightarrow \hat{p}(\cdot | x)$ for almost every x under the weak* topology (26).

(iii) We can find a subsequence $\{M_{i''}\}$ and a limit prior M such that $\hat{p}_{M_{i''}}(\cdot | x) \rightarrow \hat{p}_M(\cdot | x)$ weak* for almost every x . Therefore, $\hat{p}(\cdot | x) = \hat{p}_M(\cdot | x)$ for a.e. x , i.e. $\hat{p}(\cdot | x)$ is a (generalized) Bayes rule.

Theorem 3. *All non-randomized procedures form a complete class.*

Proof. Let $\delta : R^p \rightarrow P(\mathcal{A}_0)$ be an admissible and randomized procedure, where $P(\mathcal{A}_0)$ denotes the space of probability distributions over \mathcal{A}_0 . It follows from Lemma 2 that $\delta(x) \in P(\mathcal{A}) \subset P(\mathcal{A}^*)$ for a.e. x .

Let $\hat{p}^*(y | x) = E^{\delta(\cdot | x)}(\hat{p}(y))$, then by Lemma 4 (ii) and Jensen's inequality,

$$L(\mu, \hat{p}^*(y | x)) \leq E^{\delta(\cdot | x)}(L(\mu, \hat{p}(y))) = L(\mu, \delta(y | x)), \quad \text{for } \forall \mu \quad (30)$$

Furthermore, strict inequality holds in (30) unless either $\delta(\cdot | x)$ is non-randomized with probability 1 or $L(\mu, \delta(y | x)) = \infty$, which implies that δ can be dominated by a finite-risk non-randomized procedure. Therefore, it contradicts that δ is admissible and randomized. It then follows that the non-randomized procedures are a complete class. \ddagger

Theorem 3 shows that we can restrict attention to nonrandomized procedures $\hat{p}(\cdot | x)$. Next we prove that for a.e. x , all admissible procedures are limits of Bayes rules (5). Since the Bayes rules are also nonrandomized, this convergence can be evaluated with respect to the weak* topology for each x .

Theorem 4. *For any admissible procedure $\hat{p}(\cdot | x)$, there exists a sequence of priors $\{M_i\}$, supported on a finite set, such that $\hat{p}_{M_i}(\cdot | x) \rightarrow \hat{p}(\cdot | x)$ weak* for a.e. x under the topology (26).*

Proof. This is essentially Theorem 4A.12 of Brown (1986). There are some minor differences between the formulations there and here which we now note in order to clarify how that Theorem 4A.12 yields the current Theorem 4. The principal difference is that the action space \mathcal{A}^* in Brown (1986) was assumed to be Euclidean whereas here it is merely compact, separable, and metrizable. Because the space \mathcal{A}^* , here is compact, the one point compactification $\{i\}$ introduced in Brown (1986) is not needed. This simplifies the proof of Proposition 4A.11 there, which in our context becomes Theorem

3. The remainder of the proof proceeds as discussed in the text of proof of Theorem 4A.12. ‡

Theorem 4 establishes that any admissible procedure $\hat{p}(\cdot | x)$ is a limit of Bayes rules for a.e. x . To prove $\hat{p}(\cdot | x)$ itself is also a (generalized) Bayes rule, we need to find a (possibly improper) prior M such that $\hat{p}_M(\cdot | x) = \hat{p}(\cdot | x)$ for a.e. x .

Theorem 5. *The set of all generalized Bayes procedures is a complete class of procedures.*

Proof. Suppose $\hat{p}(\cdot | x)$ is an admissible procedure. Then by Theorem 4, there exists a sequence of measures M_i supported on finite sets such that $\hat{p}_{M_i}(\cdot | x) \rightarrow \hat{p}(\cdot | x)$ for a.e. x under the weak* topology (26).

Let

$$r_i = \int_{\|x\| \leq 1} \int p(x | \mu) M_i(d\mu) dx > 0,$$

and let

$$M'_i = M_i / r_i.$$

Note that $\hat{p}_{M_i} = \hat{p}_{M'_i}$ and also that

$$\int_{\|x\| \leq 1} \int p(x | \mu) M'_i(d\mu) dx = 1, \tag{31}$$

so by 2.16 (iv) of Brown (1986), there is a finite limiting measure M such that $M'_i \rightarrow M$ weak* on compact sets. Moreover, there exists a convex set S such that $\liminf_{i \rightarrow \infty} \sup_{x \in S} \int p(x | \mu) M'_i(d\mu) < \infty$. It then follows from Theorem 2.17 in Brown (1986) that for any $x \in S^0$,

$$\int p(x | \mu) M'_i(d\mu) \rightarrow \int p(x | \mu) M(d\mu). \tag{32}$$

We will prove that the closure $\bar{S} = R^p$.

Suppose $x \notin \bar{S}$. Then

$$\int p(x | \mu) M'_i(d\mu) \rightarrow \infty,$$

and

$$\begin{aligned}
\int_{\|y\| \leq 1} p_{M'_i}(y | x) dy &= \int_{\|y\| \leq 1} \frac{\int p(x | \mu) p(y | \mu) M'_i(d\mu)}{\int p(x | \mu) M'_i(d\mu)} \\
&\leq (2\pi v_x)^{-p/2} \frac{\int_{\|y\| \leq 1} p(y | \mu) M'_i(d\mu)}{\int p(x | \mu) M'_i(d\mu)} \\
&\rightarrow 0,
\end{aligned}$$

This implies $\hat{p}(\cdot | x) = 0$ for any $x \notin \bar{S}$. If $\bar{S} \neq R^p$, then the measure of \bar{S}^c is positive and $R_{KL}(\mu, \hat{p}) = \infty$. This would contradict the assumed admissibility of \hat{p} . Hence $\bar{S} = R^p$.

It then follows from (32) and the dominated convergence that for any x and y ,

$$\begin{aligned}
\hat{p}_{M'_i} &= \frac{\int p(x | \mu) p(y | \mu) M'_i(d\mu)}{\int p(x | \mu) M'_i(d\mu)} \\
&\rightarrow \frac{\int p(x | \mu) p(y | \mu) M(d\mu)}{\int p(x | \mu) M(d\mu)} \\
&= \hat{p}_M(y | x).
\end{aligned}$$

Hence $\hat{p} = \hat{p}_M$ is a generalized Bayes procedure.

Appendix

In this appendix, we provide the proofs of Lemma 2, 3 and 4 from Section 3.1.

Proof of Lemma 2:

- (i) Suppose $g_0 = 0$ on a set with positive measure. Then by definition $L(\mu, g_0) = \infty$ for any μ . So any $g \in \mathcal{A}$ with finite risk dominates it and thus g_0 is inadmissible.
- (ii) Suppose $g_0 > 0$ almost everywhere. If $g_0 \geq C$ on a set S with Lebesgue measure $\nu(S) > 0$, then a g can be constructed by truncating g_0 on S and lifting it in the other areas. Notice that $\int_{S^c} g_0(y) dy > 0$, so we can define

$$c = \frac{1 - C\nu(S)}{\int_{S^c} g_0(y) dy}, \quad (33)$$

where S^c is the complementary set of S . It is easy to check $c > 1$. Let

$$g(y) = \begin{cases} c g_0 & y \in S^c \\ C & y \in S \end{cases} \quad (34)$$

Obviously, $g \in \mathcal{A}$. For any μ , the difference between the loss functions of g_0 and g is

$$\begin{aligned} & L(\mu, g_0) - L(\mu, g) \\ &= \int p(y | \mu) \log g(y) dy - \int p(y | \mu) \log g_0(y) dy \\ &= \int_S p(y | \mu) \log C dy + \int_{S^c} p(y | \mu) \log(c g_0(y)) dy - \int p(y | \mu) \log g_0(y) dy \\ &= \int_S p(y | \mu) \log C dy + \log c \int_{S^c} p(y | \mu) dy + \int_{S^c} p(y | \mu) \log g_0(y) dy - \int p(y | \mu) \log g_0(y) dy \\ &= \int_S p(y | \mu) \log C dy + \log c \int_{S^c} p(y | \mu) dy - \int_S p(y | \mu) \log g_0(y) dy \\ &= \log c - \int_S p(y | \mu) \log \frac{c g_0(y)}{C} dy \\ &\geq \log c - \log \int_S p(y | \mu) \frac{c g_0(y)}{C} dy \quad (\text{Jensen's Inequality}) \\ &\geq \log c - \log \int_S c g_0(y) dy \\ &> 0 \end{aligned}$$

The last strict inequality holds because $\int_S g_0(y) dy = 1 - \int_{S^c} g_0(y) dy < 1$. Therefore, g dominates g_0 . †

Proof of Lemma 3:

- (i) By the Banach-Alaoglu Theorem, the \mathcal{L}_1 unit ball $\{g : R^p \rightarrow R \mid \int g(y) dy \leq 1\}$ is weak* compact. Also, it is easy to check that the bounded set $\{g : R^p \rightarrow R \mid 0 \leq g(y) \leq C\}$ is closed and thus compact. So their intersection \mathcal{A}^* is compact.
- (ii) Because \mathcal{L}_1 a separable normed space, the weak* topology on the closed ball of its dual space \mathcal{L}_∞ can be metrized by (27). And since every compact metric space is separable and second countable, (ii) follows immediately from (i).

- (iii) Suppose $g^* \in \mathcal{A}^*$ but $g^* \notin \mathcal{A}$, then $\int g^*(y)dy < 1$. If $\int g^*(y)dy = 0$, its loss function $L(\mu, g^*) = \infty$ for any μ and thus g^* is inadmissible. Otherwise let $g' = g^* / \int g^*(y)dy$, then $\int g'(y)dy = 1$ and it is easy to check that g' dominates g^* . Truncate g' as in (34) if necessary, it yields a $g \in \mathcal{A}$ that dominates g' and therefore dominates g^* . ‡

Proof of Lemma 4:

- (i) Suppose $g_i \rightarrow g$ weak*.

- (a) If $g = 0$ on a set K with positive measure, then $L(\mu, g) = \infty$ for any μ .

Also,

$$\begin{aligned}
L(\mu, g_i) &= \int p(y | \mu) \log \frac{p(y | \mu)}{g_i(y)} dy \\
&= \int p(y | \mu) \log p(y | \mu) dy - \int_{K^c} p(y | \mu) \log g_i(y) dy - \int_K p(y | \mu) \log g_i(y) dy \\
&\geq \int p(y | \mu) \log p(y | \mu) dy - \int_{K^c} p(y | \mu) \log g_i(y) dy - \log \int_K p(y | \mu) g_i(y) dy \\
&\geq \int p(y | \mu) \log p(y | \mu) dy - \int_{K^c} p(y | \mu) \log g_i(y) dy - \log \int_K C g_i(y) dy \\
&\rightarrow \int p(y | \mu) \log p(y | \mu) dy - \int_{K^c} p(y | \mu) \log g(y) dy - \log \int_K C g(y) dy \\
&= \infty \quad \text{as } i \rightarrow \infty
\end{aligned}$$

The last equation holds since $\int_K C g(y) dy = C \int_K g(y) dy = 0$. Therefore,

$\liminf_{i \rightarrow \infty} L(\mu, g_i) = L(\mu, g) = \infty$ for any $\mu \in R^p$.

- (b) If $g > 0$ almost everywhere, then for any $\mu \in R^p$,

$$\begin{aligned}
L(\mu, g) - L(\mu, g_i) &= \int p(y | \mu) \log \frac{g_i(y)}{g(y)} dy \\
&\leq \int p(y | \mu) \left(\frac{g_i(y)}{g(y)} - 1 \right) dy \\
&= \int p(y | \mu) \frac{g_i(y)}{g(y)} dy - 1 \\
&= \int_{L(\varepsilon)} p(y | \mu) \frac{g_i(y)}{g(y)} dy + \int_{L(\varepsilon)^c} p(y | \mu) \frac{g_i(y)}{g(y)} dy - 1
\end{aligned}$$

where $L(\varepsilon) = \{y \mid g(y) \geq \varepsilon\}$ for a fixed $\varepsilon > 0$. When $y \in L(\varepsilon)$, $\frac{p(y \mid \mu)}{g(y)} \leq \frac{p(y \mid \mu)}{\varepsilon}$ is a \mathcal{L}_1 function, so

$$\int_{L(\varepsilon)} p(y \mid \mu) \frac{g_i(y)}{g(y)} dy \rightarrow \int_{L(\varepsilon)} p(y \mid \mu) dy$$

Moreover, as $\varepsilon \downarrow 0$, $L(\varepsilon) \rightarrow R^p$ and therefore

$$\int_{L(\varepsilon)} p(y \mid \mu) dy \uparrow 1 \quad \text{and} \quad \int_{L(\varepsilon)^c} \frac{p(y \mid \mu)}{g(y)} g_i(y) dy \downarrow 0$$

So $L(\mu, g) - \liminf_{i \rightarrow \infty} L(\mu, g) = \limsup_{i \rightarrow \infty} (L(\mu, g) - L(\mu, g_i)) \leq 0$. The lower semi-continuity of $L(\mu, \cdot)$ is verified.

(ii) To prove the strict convexity of $L(\mu, \cdot)$ on \mathcal{A}_+^* , suppose $g_1, g_2 \in \mathcal{A}_+^*$ and $g_\lambda(y \mid x) = \lambda g_1 + (1 - \lambda)g_2$ with $0 < \lambda < 1$, then

$$\begin{aligned} L(\mu, g_\lambda) &= \int p(y \mid \mu) \log \frac{p(y \mid \mu)}{g_\lambda(y)} dy \\ &= \int p(y \mid \mu) \log p(y \mid \mu) dy - \int p(y \mid \mu) \log [\lambda g_1(y) + (1 - \lambda)g_2(y)] dy \\ &< \int p(y \mid \mu) \log p(y \mid \mu) dy - \int p(y \mid \mu) [\lambda \log g_1(y) + (1 - \lambda) \log g_2(y)] dy \\ &= \lambda \int p(y \mid \mu) \log \frac{p(y \mid \mu)}{g_1(y)} dy + (1 - \lambda) \int p(y \mid \mu) \log \frac{p(y \mid \mu)}{g_2(y)} dy \\ &= \lambda L(\mu, g_1) + (1 - \lambda)L(\mu, g_2) \end{aligned}$$

The strict convexity follows immediately. ‡

References

- Aitchison, J. (1975). Goodness of Prediction Fit. *Biometrika*. 62, 547-554.
- Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis, Second Edition*. Springer, New York.
- Brown, L.D. (1971). Admissible Estimators, Recurrent Diffusions, and Insoluble Boundary Value Problems. *Annals of Mathematical Statistics*, 42, 855-903.

- Brown, L.D. (1986). Fundamentals of Statistical Exponential Families with applications in Statistical Decision Theory. *Institute of Mathematical Statistics, Lecture Notes-Monograph Series, Volume 9*, Shanti S. Gupta, series editor.
- Brown, L and Hwang, J (1982). A Unified Admissibility Proof. *Statistical Decision Theory and Related Topics, III. S. S. Gupta and J. O. Berger (eds.)* Academic Press, New York, 1, 205-230.
- George, E.I., Liang, F. and Xu, X. (2005). Improved Minimax Prediction Under Kullback-Leibler Loss. (Accepted by the *Annals of Statistics*).
- Komaki, F. (2001). A Shrinkage Predictive Distribution for Multivariate Normal Observations, *Biometrika*. 88, 859-864.
- Liang, F. (2002). *Exact Minimax Procedures for Predictive Density Estimation and Data Compression*. Ph.D. dissertation, Department of Statistics, Yale University.
- Liang, F. and Barron, A. (2003). Exact Minimax Strategies for Predictive Density Estimation, Data Compression and Model Selection. *IEEE Information Theory Transactions*, to appear.
- Murray, G.D. (1977), A Note on the Estimation of Probability Density Functions. *Biometrika*. 64, 150-152.
- Nussbaum, M. (1999) Minimax risk: Pinsker bound. In *Encyclopedia of Statistical Sciences*, S. Kotz, editor, Wiley, New York, 451460.
- Stein, C. (1974). Estimation of the Mean of a Multivariate Normal Distribution. In *Proceedings of the Prague Symposium on Asymptotic Statistics*, Ed. J. Hajek, pp. 345-81. Prague: Universita Karlova.
- Stein, C. (1981). Estimation of a Multivariate Normal Mean. *Ann. Statist.* **9**, 1135-51.
- Strawderman, W.E. (1971). Proper Bayes Minimax Estimators of the Multivariate Normal Mean. *Annals of Mathematical Statistics*. 42, 385-388.