*Sankhyā : The Indian Journal of Statistics*
2018, Volume 80-A, Part 1, pp. 152-167

# On Being Bayes and Unbiasedness

## Bayes and Unbiasedness

Siamak Noorbaloochi
*University of Minnesota, Minneapolis, USA*
Glen Meeden
*University of Minnesota, Minneapolis, USA*

**Abstract**

Assuming squared error loss, we show that finding unbiased estimators and Bayes estimators can be treated as using a pair of linear operators that operate between two Hilbert spaces. We note that these integral operators are adjoint and then investigate some consequences of this fact. An extension to loss functions that can be defined via an inner product is also presented.

*AMS* (2000) *subject classification.* Primary: 62F10, 62F15; Secondary: 47A05, 46C07.
*Keywords and phrases.* Unbiasedness, Bayes estimators, Squared error loss, Consistency

## 1 Introduction

Statistical inference is used to produce "plausible" data-based assertions and rules about a partially unknown population. Two well-adopted and seemingly adverse inference plausibility criteria are unbiasedness and being Bayes. The task of the current note is to explore further the relationship between these two approaches by treating them as operators that represent our inference procedures.

When estimating a real-valued function of an unknown parameter, a popular choice for the loss function is squared error. Lehmann (1951) noted that the usual definition of unbiasedness which requires that the expectation of the estimator to be equal to the value of the function being estimated for all possible parameter values is closely tied to squared error loss. He then proposed a more general definition of unbiasedness which depends on the loss function being used. Noorbaloochi and Meeden (1983) proposed a generalization of Lehmann's definition which depends not only on loss function but also on a prior distribution $\pi$ for $\theta$. If $r(\delta, \gamma; \pi)$ denotes the

Bayes risk of the estimator $\delta$ for estimating $\gamma$, then $\delta$ is said to be unbiased for estimating $\gamma'$ if $\gamma'$ minimizes the Bayes risk over all possible $\gamma$ for this fixed $\delta$. They noted that, under some rather weak assumptions, the resulting $\gamma'$ does not depend on the prior $\pi$ and so in some sense, unbiasedness can be thought of as the dual of being Bayes.

In this note, restricting attention to squared error loss, corresponding to a given prior $\pi$, we construct a Hilbert space of square integrable real-valued functions defined on the product of the sample space and the parameter space. Given the induced inner products and norms, we observe that unbiasedness and being Bayes are adjoint operators. From this fact, we derive some orthogonality relationships between Bayes and unbiased estimators and the functions they are estimating.

## 2 Notation

Let $X$ denote the sample space and $\Theta$ the parameter space for our experiment. For each $\theta \in \Theta$, let $p_\theta$ be a probability density function on $X$ and let $\pi$ be a fixed prior distribution on $\Theta$. Let

$$\mathcal{H}_{p,\pi} = \{h(x,\theta) : \int\int h^2(x,\theta)p_\theta(x)\pi(\theta)dx\, d\theta < \infty\}$$

be the space of all square integrable real-valued functions of $(x,\theta)$.

Note $\mathcal{H}_{p,\pi}$ becomes a Hilbert space when it is equipped with the inner product

$$(h_1,h_2)_{p,\pi} = \int\int h_1(x,\theta)h_2(x,\theta)p_\theta(x)\pi(\theta)dx\, d\theta.$$

Let $\|h\|_{p,\pi} = \sqrt{(h,h)}$ denote the norm of $h$. We include the subscripts $p$ and $\pi$ to remind us that $\mathcal{H}_{p,\pi}$ does depend on the $p_\theta$s, our model, and $\pi$. We assume that $p_\theta(x)$ is in the above space.

There are two linear subspaces of $\mathcal{H}_{p,\pi}$ which are of particular interest. The first is

$$\Gamma_\pi = \{\gamma(\theta) : \int \gamma^2(\theta)\pi(\theta)d\theta\ < \infty\}$$

and the second is

$$\Delta_m = \{\delta(x) : \int \delta^2(x)m(x)dx\ < \infty\}$$

where $m(x) = \int p_\theta(x)\pi(\theta)\, d\theta$.

The set $\Gamma_\pi$ is a Hilbert subspace of $\mathcal{H}_{p,\pi}$ with the induced weighted inner product $(\gamma_1,\gamma_2)_\pi = \int_\Theta \gamma_1(\theta)\gamma_2(\theta)\pi(\theta)\, d\theta$. Similarly, $\Delta_m$ is a Hilbert subspace with the induced weighted inner product $(\delta_1,\delta_2)_m = \int_{\mathcal{X}} \delta_1(x)\delta_2(x)m$ $(x)dx$.

We also notice that, provided the interchange of order of integration is permitted, for any $\delta \in \Delta_m$, we have $\mathrm{Var}_\theta \delta(X) < \infty$ and hence, $E_\theta \delta(X) < \infty$ for all $\theta$ in the support of $\pi$. If the support of $\pi$ is all of $\Theta$, all members of $\Delta_m$ are unbiased estimators of their expectations.

Also, the assumed square integrability of the likelihoods and the Holder inequality imply that for any $\gamma \in \Gamma_\pi$,

$$\int_\Theta \gamma(\theta) p_\theta(x) \pi(\theta)\, d\theta < \infty$$

and hence, all members of $\Gamma_\pi$ have Bayes estimators. With the above notation, the Euclidian distance between any $\gamma \in \Gamma_\pi$ and $\delta \in \Delta_m$ is $\|\delta - \gamma\|^2_{p,\pi} = r(\delta, \gamma; \pi)$, which is the Bayes risk associated with the pair.

Let us define the operator, $\mathcal{U}$

$$\mathcal{U} : \Delta_m \to \Gamma_\pi$$

as the *unbiasedness operator*, if for a given $\gamma$, $\mathcal{U}\delta = \gamma$ if and only if

$$r(\delta, \gamma; \pi) = \inf_{\gamma' \in \Gamma_\pi} r(\delta, \gamma'; \pi).$$

For mean-unbiasedness, $\mathcal{U}$ can be defined through the integral transform:

$$\mathcal{U} : \Delta_m \to \Gamma_\pi \text{ if and only if } E_\theta(\delta(X)) = \gamma(\theta) \text{ for all } \theta.$$

Hence, in this case, the operator $\mathcal{U}$ is a linear operator. So even though $\Gamma_\pi$ and $\delta_m$ depend on $\pi$, the linear operator $\mathcal{U}$ does not.

The usual decomposition, $\mathrm{Var}_m \delta(X) = E_\pi[\mathrm{Var}_\theta \delta(X)] + \mathrm{Var}_\pi[E_\theta \delta(X)]$, and finiteness of $\mathrm{Var}_m \delta(X)$ imply $E_\pi[E_\theta \delta(X) - E_m \delta(X)]^2 < \infty$. Hence, by definition of $\Gamma_\pi$, the function $E_\theta \delta(X) - E_m \delta(X)$ and all constant functions are in $\Gamma_\pi$. This implies $\Gamma_\pi$ contains $E_\theta \delta(X)$'s for all $\delta \in \Delta_m$ with $E_\theta \delta^2(X) < \infty$. In other words, the set of estimable functions of estimators with finite variance, which we will denote by $\Gamma_e$, is a subset of $\Gamma_\pi$. Now, in order to minimize $r(\delta, \gamma'; \pi) = E_\pi E_\theta (\delta(X) - \gamma'(\theta))^2$ over $\Gamma_\pi$, as in Noorbaloochi and Meeden (1983), fix $\theta$ and find the $\gamma^*$, such that for the given $\theta$,

$$E_\theta(\delta(X) - \gamma^*)^2 \leq E_\theta(\delta(X) - \gamma'(\theta))^2$$

but this is achieved if $\gamma^* = E_\theta \delta(X)$. Averaging the risk function (here MSE), with respect to *any* prior will preserve the inequality for this prior-free minimizer. The important conditions for $\mathcal{U}$ to be independent of $\pi$, then, are that $E_\theta \delta(X)$ be in $\Gamma_\pi$ and that $\Theta$ be the support of $\pi$.

Generally, for complete inner product spaces, $r(\delta, \gamma; \pi) = \|\delta - \gamma\|_{p,\pi}^2$, and for any given $\gamma$, $\mathcal{U}\delta = \gamma$, if and only if,

$$(\gamma', \delta - \mathcal{U}\delta)_{p,\pi} = 0 \text{ for all } \gamma' \in \Gamma_\pi \tag{2.1}$$

that is, $\gamma = \mathcal{U}\delta$ is the orthogonal projection of $\delta$ onto $\Gamma_\pi$. We say a function in $\Gamma_\pi$ is *estimable* if it is the projection of some function in $\Delta_m$.

Similarly, the Bayes operator, $\mathcal{B}_\pi$, may be defined as

$$\mathcal{B}_\pi : \Gamma_\pi \to \Delta_m$$

where for each $\gamma \in \Gamma_\pi$, $\mathcal{B}_\pi\gamma = \delta_\pi$, if and only if,

$$r(\delta_\pi, \gamma; \pi) = \inf_{\delta' \in \Delta_m} r(\delta', \gamma; \pi).$$

For squared error loss, the Bayes operator corresponds to the linear operator:

$$\mathcal{B}_\pi : \Gamma_\pi \to \Delta_m \text{ if and only if } E(\gamma(\theta)|x) = \delta(x) \text{ for all } x \in \mathcal{X}.$$

For the Bayes risks that are defined through inner products of the underlying Hilbert spaces, for a given $\gamma$, $\mathcal{B}_\pi\gamma = \delta_\pi$, if and only if,

$$(\delta', \gamma - \mathcal{B}_\pi\gamma)_{p,\pi} = 0 \text{ for all } \delta' \in \Delta_m. \tag{2.2}$$

Therefore, $\delta_\pi = \mathcal{B}_\pi\gamma$ is the orthogonal projection of $\gamma$ onto $\Delta_m$.

It should be mentioned that, in symmetry to the unbiasedness operator, had the joint distribution over $(\theta, X)$ been specified through an inverse-probability, $\pi_x(\theta)$ and a marginal distribution for $X$, rather than derivation of posterior through the traditional likelihood-prior pair, the linear Bayes operator, $\mathcal{B}_\pi$, would have been independent from the marginal distribution of $X$ under some broad assumptions.

We are now ready to state the main observation. Throughout, we always assume that we are dealing with a fixed prior $\pi$.

## 3 The Relationship Between Unbiasedness and Being Bayes

Given the above setup, we now show that the Bayes and unbiasedness operators, $\mathcal{B}_\pi$ and $\mathcal{U}$, are the adjoint operators of each other. That is, for any $\gamma \in \Gamma_\pi$ and any $\delta \in \Delta_m$, we have

$$(\gamma, \mathcal{U}\delta)_\pi = (\mathcal{B}_\pi\gamma, \delta)_m. \tag{3.1}$$

The proof is easy:

$$\begin{aligned}(\gamma,\mathcal{U}\delta)_\pi &= \int_\Theta \gamma(\theta)E_\theta\delta(X)\pi(\theta)\,d\theta = \int_\mathcal{X}\delta(x)[\int_\Theta \gamma(\theta)f_\theta(x)\pi(\theta)\,d\,\theta]\,dx\\ &= \int_\mathcal{X}\delta(x)[\int_\Theta \gamma(\theta)\frac{f_\theta(x)\pi(\theta)}{m(x)}\,d\theta]m(x)\,dx = (\delta, E(\gamma(\theta)\mid x))_m\\ &= (\delta,\mathcal{B}_\pi\gamma)_m.\end{aligned}$$

So even though $\Gamma_\pi$ and $\Delta_m$ depend on $\pi$, we see that the unbiased operator is independent of the chosen prior and that $\mathcal{U}$, simultaneously, is the adjoint of all $\mathcal{B}_\pi$ for *all* priors with $\Theta$ support.

It is interesting to note that the adjointness property plus either being Bayes or unbiased implies the other. To see this, assume $\mathcal{A}$ is the adjoint of $\mathcal{U}$, we note that any $\delta$ in $\Delta_m$ is an unbiased estimator for $\mathcal{U}\delta$, if and only, according to the above definition, $\mathcal{U}\delta$ is the orthogonal projection of $\delta$ onto $\Gamma_\pi$, that is, if and only if, for all $\gamma \in \Gamma_\pi$,

$$(\delta - \mathcal{U}\delta, \gamma)_{p,\pi} = 0.$$

Now this and adjointness of $\mathcal{A}$ imply

$$(\delta, \mathcal{A}\gamma)_m = (\mathcal{U}\delta,\gamma)_\pi = (\delta,\gamma)_{p,\pi}$$

which further implies $(\delta, \gamma - \mathcal{A}\gamma)_{p,\pi} = 0$, for all $\delta \in \Delta_m$, or equivalently, $\mathcal{A}\gamma$ is the orthogonal projection of $\gamma$.

A similar argument shows that the adjoint operator of the Bayes operator is the unbiased operator. We denote the closure of range of $\mathcal{U}$ by $\overline{\mathcal{R}(\mathcal{U})}$. This is the set of all functions in $\Gamma_\pi$ which have an unbiased estimator or are limits (with respect to $\|\gamma\|_\pi^2$) of a sequence of unbiasedly estimable functions. We denote the closure of the range of $\mathcal{B}_\pi$ by $\overline{\mathcal{R}(\mathcal{B}_\pi)}$. This is the set of all Bayes estimators and their limits (with respect to $\|\delta\|_m^2$) in $\Delta_m$. In addition, we have the null spaces of these two operators. The space $\mathcal{N}(\mathcal{U})$, is the set of all unbiased estimators of zero. If the model is complete, this will contain just one function. The space $\mathcal{N}(\mathcal{B}_\pi)$ is the set of all functions with zero as their Bayes estimator. It is a basic result of functional analysis (see Rudin 1991) that from Eq. 3.1, we can write

$$\Gamma_\pi = \overline{\mathcal{R}(\mathcal{U})}\bigoplus\mathcal{N}(\mathcal{B}_\pi) \tag{3.2}$$

and

$$\Delta_m = \overline{\mathcal{R}(\mathcal{B}_\pi)}\bigoplus\mathcal{N}(\mathcal{U}). \tag{3.3}$$

The first equation implies that every member of $\Gamma_\pi$ can be orthogonally decomposed into a function with an unbiased estimator or limit of estimable functions plus a function whose Bayes estimator is zero. The second equation implies that every member of $\Delta_m$ can be orthogonally decomposed into a Bayes estimator (of some $\gamma$) or limit of Bayes estimators plus an unbiased estimator of zero. These two decompositions are unique. As far as we know, this has never been noted before and shows that the notions of being unbiased and being Bayes are more closely entwined than previously thought.

The Eq. 3.2 shows that given any $\gamma \in \Gamma_\pi$, there exist a unique $\gamma_e \in \overline{\mathcal{R}(\mathcal{U})}$ and a unique $\alpha \in \mathcal{N}(\mathcal{B}_\pi)$ such that

$$\gamma(\theta) = \gamma_e(\theta) + \alpha(\theta) \quad \text{for } \theta \in \Theta.$$

So every function will have an unbiased estimator if and only if the only function whose Bayes estimator is the zero function is zero function. Furthermore, when $\alpha$ is not the trivial function, we see that the Bayes estimator of $\gamma$ must also be the Bayes estimator of $\gamma_e$.

The Eq. 3.3 shows that given $\delta \in \Delta_m$, there exists a unique $\delta_\pi \in \overline{\mathcal{R}(\mathcal{B}_\pi)}$ and a unique $\delta_0 \in \mathcal{N}(\mathcal{U})$ such that

$$\delta(x) = \delta_\pi(x) + \delta_0(x) \quad \text{for } x \in \mathcal{X}.$$

So every decision function will be a Bayes estimator for some $\gamma$ if and only if the only unbiased estimator of the zero function is the zero function. Also, the orthogonal decompositions imply that

$$\|\delta\|_m^2 = \|\delta_\pi\|_m^2 + \|\delta_0\|_m^2 \qquad (3.4)$$

$$\|\gamma\|_\pi^2 = \|\gamma_e\|_\pi^2 + \|\alpha\|_\pi^2. \qquad (3.5)$$

An important special case is when $\Gamma_\pi = \overline{\mathcal{R}(\mathcal{U})}$, which will be the case if $\mathcal{N}(\mathcal{U}) = \{0\}$, that is,

$$\mathcal{U}\delta = 0 \Rightarrow \delta = 0 \qquad (3.6)$$

or, equivalently, the zero element of $\Delta_m$ is the only element that can be projected to the zero element of $\Gamma_\pi$. For squared error loss, if the family is complete,

$$E_\theta(\delta(X)) = 0 \Rightarrow \delta = 0$$

then Eq. 3.6 holds.

If the posterior family (as indexed by $x$) is complete, we also have $\mathcal{N}(\mathcal{B}) = \{0\}$ and hence, $\Delta_m = \overline{\mathcal{R}(\mathcal{B})}$. Therefore, any square integrable($m$) data function is either a Bayes estimator or is a $L^2(m)$ limit of Bayes estimators of a sequence of estimands for the given prior.

## 4 Some Consequences

**Theorem 1.** *Suppose $\overline{\mathcal{R}(\mathcal{U})}$ is a proper subset of $\Gamma_\pi$. Let $\delta_\pi \in \Delta_m$ be the Bayes estimator for some function $\gamma \in \Gamma_\pi$ which does not belong to $\overline{\mathcal{R}(\mathcal{U})}$. Then there exists a unique $\gamma_e \in \overline{\mathcal{R}(\mathcal{U})}$ and an unique $\alpha \in \mathcal{N}(\mathcal{B}_\pi)$ such that $\delta_\pi$ is the Bayes estimator of $\gamma_e$. In addition, the Bayes risk of $\delta_\pi$ when estimating $\gamma_e$ is strictly less than its Bayes risk when estimating $\gamma$.*

Proof. A given $\gamma$, by Eq. 3.2, can be written as

$$\gamma = \gamma_e + \alpha,$$

where $\gamma_e \in \overline{\mathcal{R}(\mathcal{U})}$ and $\alpha \in \mathcal{N}(\mathcal{B}_\pi)$. Then since $\mathcal{B}_\pi$ is linear, $\mathcal{B}_\pi\gamma = \mathcal{B}_\pi\gamma_e + \mathcal{B}_\pi\alpha$ but $\mathcal{B}_\pi\alpha = 0$ and hence, $\mathcal{B}_\pi\gamma = \mathcal{B}_\pi\gamma_e$. Indeed, $\gamma_e$ is the projection of $\gamma$ onto $\overline{\mathcal{R}(\mathcal{U})}$, that is,

$$\|\gamma(\theta) - \gamma_e(\theta)\|_\pi = \min_{\gamma'_e \in \overline{\mathcal{R}(\mathcal{U})}} \|\gamma(\theta) - \gamma'_e(\theta)\|_\pi.$$

To prove the second part, we note that

$$\|\delta_\pi - \gamma\|^2_{p,\pi} = \|\delta_\pi - \gamma_e\|^2_{p,\pi} + \|\alpha\|^2_\pi - 2(\delta_\pi - \gamma_e, \alpha)_{p,\pi}.$$

From the first part, we have $\delta_\pi = \mathcal{B}_\pi\gamma_e$ and hence,

$$(\delta_\pi - \gamma_e, \alpha)_{p,\pi} = (\mathcal{B}\gamma_e - \gamma_e, \alpha)_{p,\pi}$$

but this is zero by Eq. 2.2 and by choosing $\gamma'_e(\theta) = \alpha(\theta)$. Therefore,

$$\|\delta_\pi - \gamma\|^2_{p,\pi} = \|\delta_\pi - \gamma_e\|^2_{p,\pi} + \|\alpha\|^2_\pi.$$

This completes the proof.

The next theorem shows that the minimum Bayes risk, that is, the Bayes risk of the Bayes rule, is a linear function of the bias. Let us define the *bias* of a rule $\delta$ in estimating a given $\gamma(\theta)$ to be $b = \gamma - \mathcal{U}\delta$. For the inner product $(h_1, h_2)_{p,\pi} = E(h_1 h_2)$, this reduces to the usual notion of bias.

**Theorem 2.** *Let $\gamma$ be a member of $\Gamma_\pi$ and let $\delta_\pi = \mathcal{B}_\pi\gamma$ be its Bayes estimator and $b = \gamma - \mathcal{U}\delta_\pi$. Then the Bayes risk of $\delta_\pi$ is $(b, \gamma)_\pi$.*

Proof. The Bayes risk of the Bayes rule is

$$r(\delta_\pi, \gamma, \pi) = \|\mathcal{B}_\pi\gamma - \gamma\|^2_{p,\pi} = \|\mathcal{B}_\pi\gamma\|^2_m + \|\gamma\|^2_\pi - 2(\mathcal{B}_\pi\gamma, \gamma)_{p,\pi}.$$

In Eq. 2.2, setting $\delta' = \mathcal{B}_\pi\gamma$ yields $(\mathcal{B}_\pi\gamma, \gamma)_{p,\pi} = (\mathcal{B}_\pi\gamma, \mathcal{B}_\pi\gamma)_m = \|\mathcal{B}_\pi\gamma\|^2_m$. Also, Eq. 2.1 implies

$$(\mathcal{B}_\pi\gamma, \gamma)_{p,\pi} = (\mathcal{U}\mathcal{B}_\pi\gamma, \gamma)_\pi.$$

Therefore,

$$r(\delta_\pi, \gamma, \pi) = \|\mathcal{B}_\pi\gamma - \gamma\|^2_{p,\pi} = \|\gamma\|^2_\pi - (\mathcal{U}\mathcal{B}_\pi\gamma, \gamma)_\pi = (\gamma - \mathcal{U}\delta_\pi, \gamma)_\pi = (b, \gamma)_\pi$$

as was claimed.

**Corollary.** *Let $\gamma$ be a member of $\Gamma_\pi$ and let $\delta_\pi$ be its Bayes estimator with $b(\theta) = \gamma(\theta) - E_\theta\delta_\pi(X)$, its bias. Then the Bayes risk of $\delta_\pi$ is $(b, \gamma)_\pi$.*

Proof.

The Bayes risk of the Bayes rule $\delta_\pi$ for estimating $\gamma$ is

$$\begin{aligned} r(\delta_\pi, \gamma; \pi) &= E_\pi E_\theta(\delta_\pi(X) - \gamma(\theta))^2 \\ &= E_m\delta^2_\pi(X) + E_\pi\gamma^2(\theta) - E_\pi E_\theta(\delta_\pi(X)\gamma(\theta)) - E_\pi E_\theta(\delta_\pi(X)\gamma(\theta)) \end{aligned}$$

but

$$E_\pi E_\theta(\delta_\pi(X)\gamma(\theta)) = E_m(\delta_\pi(X)E(\gamma(\theta)|X)) = E_m\delta^2_\pi(X)$$

and similarly, conditioning on $\theta$,

$$E_\pi E_\theta(\delta_\pi(X)\gamma(\theta)) = E_\pi(\gamma(\theta)E(\delta_\pi(X)|\theta)).$$

Substituting these into the previous equation and simplifying, we have

$$\begin{aligned} r(\delta_\pi, \gamma; \pi) &= E_\pi\gamma^2(\theta) - E_\pi(\gamma(\theta)E_\theta\delta_\pi(X)) \\ &= E_\pi\gamma^2(\theta) + E_\pi(\gamma(\theta)(b(\theta) - \gamma(\theta))) \\ &= E_\pi\gamma(\theta)b(\theta) \\ &= (b, \gamma)_\pi \end{aligned}$$

which completes the proof.

Note that an immediate corollary is the well known fact that if an estimator is both unbiased and Bayes for some $\gamma$, then its Bayes risk is zero. Indeed, more generally, if $\mathcal{U}\delta_\pi = \gamma$, then $(b, \gamma)_\pi = (0, \gamma) = 0$. Hence, the above result is valid for all loss functions that can be defined via inner products.

Given a model, a prior and a function to be estimated, the Bayes risk of the Bayes rule,

$$\inf_{\delta' \in \Delta_m} r(\delta', \gamma; \pi) = (b, \gamma)_\pi$$

is a measure of the informativeness of our inferences. The smaller the size of the Bayes risk, the more informative is our “best” estimator about the function being estimated. If this minimum is large then the Bayes estimator is not very informative. (For further discussion on this point, see Raiffa and Schlaiffer (2000), DeGroot (1962, 1984) and Ginebra (2007).) The previous theorem quantifies the relationship between the Bayes risk and bias and shows that there will only be a "good" Bayes estimator when its bias is small.

**Theorem 3.** *For a sample of size one, let $U \in \Delta_m$ be an unbiased estimator of $\gamma \in \Gamma_\pi$ and $U_n = \sum_{i=1}^n U(X_i)/n$. Let $\delta_{\pi,n}$ be the Bayes estimator of $\gamma$ based on the sample of size $n$. Then*

*i.* $\|\delta_{\pi,n} - \gamma\|^2_{p,\pi} \to 0$ *as* $n \to \infty$.

*ii.* $\|U_n - \delta_{\pi,n}\|^2_m \to 0$ *as* $n \to \infty$.

Proof. a. Let $\tau$ be the Bayes risk of $U$ for estimating $\gamma$ when $n = 1$. Then the Bayes risk of $U_n$ is just $\tau/n$. But the Bayes risk for $\delta_{\pi,n}$ for estimating $\gamma$ is no greater than the Bayes risk of $U_n$ so part $i$ follows.

It is easy to see that

$$\|U_n - \gamma\|^2_{p,\pi} = \|U_n - \delta_{\pi,n}\|^2_m + \|\delta_{\pi,n} - \gamma\|^2_{p,\pi}.$$

by adding and subtracting $\delta_{\pi,n}$ inside the lefthand side and then multiplying out and observing that the cross product term is zero by conditioning on the data. Now part $ii$ follows from part $i$ and the above equation because both of the terms involving $\gamma$ go to zero as $n \to \infty$ and the proof is complete.

The second part of this theorem implies that for large $n$, a Bayesian whose prior is $\pi$ believes with high probability that their estimator will be close to the unbiased estimator. Now another Bayesian with a different prior believes the same thing. So they both expect agreement as the sample size increases. This result is somewhat in the spirit of one in Blackwell and Dubins (1962).

It is well known that Bayes estimators are usually consistent and tend to agree as the sample size increases. The classical Bernstein-von Mises theorem on the asymptotic normality of the posterior distribution and consistency of the Bayes estimators are well-studied subjects. Most of the standard arguments for the consistency of Bayes estimators start with considering the asymptotic distribution of the posterior distribution, see for example Jonhson (1970) and Diaconis and Freedman (1986). These arguments are closely related to the asymptotic behavior of the maximum likelihood estimator and can be technically quite complex.

Indeed, under our setup, the central limit theorem implies that $\sqrt{n}(U_n - \gamma(\theta))$ asymptomatically has a normal distribution with mean 0 and $\mathrm{Var}(U(X_1))$.

This fact and the mean convergence in the second part of the previous theorem imply that for all $\theta$ in the support of the prior, $\sqrt{n}(\delta_{\pi,n} - \gamma(\theta))$ has the same asymptotic normal distribution regardless of the chosen prior. The existence assumption of an unbiased estimator based on a single sample is the price we had paid for the simplicity of the argument.

## 5 Three Simple Examples

In this section, we consider three simple examples to demonstrate in more detail the relationships between the two concepts.

**Example 1.** *Let $X$ be a random variable, with a family of possible mass functions given by $p_\theta(x) = \frac{1}{2}$, for $x = 1$, $p_\theta(x) = \frac{1-\theta}{4}$, for $x = 2$, and$p_\theta(x) = \frac{1+\theta}{4}$,for $x = 3$, with $0 < \theta \leq 1$. For the uniform prior, the posterior is $\pi(\theta|x) = 1$, when$x = 1, \pi(\theta|x) = 2(1-\theta)$, for $x = 2$, and $\pi(\theta|x) = \frac{2(1+\theta)}{3}$, for $x = 3$, with $m(x) = \frac{1}{2}, \frac{1}{8}$ and $\frac{3}{8}$, respectively, for $x = 1, 2,$ and$3$. For this example, $\Gamma_\pi = \{\gamma(\theta) : \int_0^1 \gamma^2(\theta)\, d\theta < \infty\}$ and $\Delta_m = \mathcal{R}^3$. It is easy to check that $\mathcal{R}(U) = \{l(\theta) : l(\theta) = a + b\theta \; a, b \in \mathcal{R}\}$ and $\mathcal{N}(U) = \mathcal{Z} = \{(a, -a, -a) : a \in \mathcal{R}\}$. Letting $c = \int_0^1 \gamma(\theta)\, d\theta$ and $d = \int_0^1 \theta\gamma(\theta)\, d\theta$, we see that $\mathcal{R}(\mathcal{B}_\pi) = \{(c, 2(c-d), \frac{2(c+d)}{3}) : \; c, d \in \mathcal{R}\}$. We also have that*

$$\mathcal{N}(\mathcal{B}_\pi) = \{\gamma(\theta) : \int_0^1 \gamma(\theta)\, d\theta = 0 \;\textit{and} \int_0^1 \theta\gamma(\theta)\, d\theta = 0\}.$$

*Note that $(\delta, z)_m = \sum_1^3 \delta_\pi(x) z(x) m(x) = \frac{1}{2}ac - \frac{2(ac-d)}{8} - \frac{3}{8}\frac{2(ac+d)}{3} = 0$ confirming the orthogonality of Bayes rules and unbiased estimators of zero. Also, for any $\gamma(\theta) \in \mathcal{N}(\mathcal{B}_\pi)$ and any $l(\theta) = a + b\theta \in \mathcal{R}(\mathcal{U})$:*

$$(\gamma, l)_\pi = \int_0^1 \gamma(\theta)(a + b\theta)\, d\theta = a\int_0^1 \gamma(\theta) + b\int_0^1 \theta\gamma(\theta) = 0$$

*and thus confirming the orthogonality of functions which have an unbiased estimator and those whose Bayes estimate is zero.*

**Example 2.** *Let $X$ be a Bernoulli random variable with success probability $\theta$ and with the uniform distribution over $(0, 1)$ as the prior. The class of linear functions in $\theta$ is the subspace of estimable functions, $\Gamma_e$. The function $e^\theta$ does not have an unbiased estimator but its Bayes estimator is $E(e^\theta|x) = 2(e-2)$, when $x = 0$ and is equal to two when $x = 1$. However, projecting $e^\theta$ onto this subspace using the least square procedure by solving*

$$\begin{cases} a_0 + a_1 E_\pi\theta = E_\pi e^\theta \\ a_0 E_\pi\theta + a_1 E_\pi\theta^2 = E_\pi \theta e^\theta \end{cases}$$

*yields the function* $\gamma_e(\theta) = (4e - 10) + (18 - 6e)\theta$*. This function has the same Bayes estimator as* $e^\theta$ *and the Bayes estimator of the approximation error,* $(4e-9)+(17-6e)\theta+\sum_{i=2}^{\infty}\theta^k/k!$*, is zero. However, the same Bayes estimator of* $e^\theta$ *and* $(4e-10)+(18-6e)\theta$ *has different Bayes risks:* $1/2(3e^2-16e+21) \approx 0.1626$ *and* $2e^2 - 12e + 18 \approx 0.1587$*, respectively.*

*Suppose now we wish to estimate the function* $1-\theta^2$*. Its Bayes estimator, say* $\delta^*$*, estimates 5/6 when* $X = 0$ *and 1/2 when* $X = 1$*. It is easy to check that the only polynomials of degree 2 whose Bayes estimator is the zero function are of the form* $a/6 - a\theta + a\theta^2$ *for some real number* $a$*. Hence, the decomposition*

$$1 - \theta^2 = (7/6 - \theta) + (-1/6 + \theta - \theta^2)$$

*breaks it up into the sum of a function with an unbiased estimator and a function whose Bayes estimator is the zero function. So* $\delta^*$ *is also the Bayes estimator of the function* $7/6 - \theta$*.*

*More generally, in the binomial case of size n, if we let* $P_0(\theta) = 1$ *and for* $k = 1, \ldots, n$ *let* $P_k(\theta)$ *be orthonormal polynomials with respect to the inner product of* $\Gamma_\pi$*, which together forms a basis for the space of the functions which have unbiased estimators, then the Bayes estimator of any* $\gamma \in \Gamma_\pi$ *is*

$$E(\gamma(\theta)|x) = \sum_{k=0}^{n}(\gamma, P_k)_\pi E(P_k(\theta)|x) = E_\pi\gamma(\theta) + \sum_{k=1}^{n}(\gamma, P_k)_\pi E(P_k(\theta)|x).$$

*This shows that all of the Bayes estimates belong to the linear space spanned by the Bayes estimates of this basis and moreover,* $\gamma(\theta) - \sum_{i=0}^{n} (\gamma, P_k)_\pi P_k(\theta)$ *has zero as its Bayes estimator. Note that the space of unbiasedly estimable functions is independent of the prior, but the projection of a given function* $\gamma$ *(without an unbiased estimator) onto this subspace depends on the prior through the induced inner product.*

**Example 3.** *Let* $X$ *be a normal random variable with mean* $\theta$ *and variance one where* $\Theta$ *is the set of real numbers. This family is complete and hence,* $\mathcal{N}(\mathcal{U}) = \{0\}$*. Suppose the prior distribution for* $\theta$ *is the normal distribution with mean equal to zero and variance equal to one. Then the posterior distribution is normal with mean* $x/2$ *and variance* $1/2$*. As a function of* $x$*, this family of possible posterior distributions is also complete and hence* $\mathcal{N}(\mathcal{B}_\pi) = \{0\}$*. Therefore,*

$$\Gamma_\pi = \overline{\mathcal{R}(\mathcal{U})}, \quad \Delta_m = \overline{\mathcal{R}(\mathcal{B}_\pi)}.$$

*That is, any square integrable(*$\pi$*) parametric function is either unbiasedly estimable or is a* $L^2(\pi)$ *limit of unbiasedly estimable functions. We note* *that all members of $\mathcal{R}(\mathcal{U})$ are infinitely differentiable functions. Also, any square integrable $(m)$ function of $X$ is either a Bayes estimator of some function (and hence infinitely differentiable with respect to $x$) or is a $L^2(m)$ limit of a sequence of Bayes estimators. For example, $|X|$ cannot be a Bayes estimator of any member of $\Gamma_\pi$*

*For the standard normal prior, the Hermite orthogonal polynomials constitute a complete orthogonal basis for $L^2(\pi)$. This implies all members of $L^2(\pi)$ can be represented as $\sum_{i=0}^{\infty} a_i H_i(\theta)$. The parametric function $|\theta|$ which is in $\Gamma_\pi$ is not unbiasedly estimable; hence, it has to be in the boundary of $\mathcal{R}(\mathcal{U})$. Indeed, we note that since $|\theta|$ is an even function, it can be approximated by the unbiasedly estimable sequence*

$$\gamma_k(\theta) = \sum_{i=0}^{k} a_i H_i(\theta), \qquad k = 0, 1, 2, \ldots.$$

*For any $j$, $E_\theta(H_j(X)) = \theta^j$ and $H_i(\theta) = \sum_{j=0}^{i} b_{ij}\theta^{2j}$ and hence,*

$$U_k(X) = \sum_{i=0}^{k} a_i \sum_{j=0}^{i} b_{ij} H_j(X)$$

*is an unbiased estimator for $\gamma_k(\theta)$. The limiting estimator is not unbiased but can be approximated by an unbiased estimator. For a detailed study of these estimators, their rate of convergence, and minimax error bounds, see Cai and Low (2011). Also, Lepski et al. (1999) using Fourier series approximation of $|\theta|$ construct unbiased estimates of individual terms in the approximation.*

*In this example, all terms of the sequence of parametric orthogonal polynomials are estimable but the limiting function is not. However, this is not true for all orthogonal basis of $\Gamma_\pi$. For example, for a one-parameter exponential family, where its parameter space contains an interval, all estimable functions can be expanded by the Haar wavelets where none of the approximating terms are unbiasedly estimable!*

*Note that as we have seen in this example, if the posterior model is complete then all of the Bayes estimators are estimators of unbiasedly estimable functions or the limit of estimable functions.*

## 6 Some Generalizations

The essential requirement of the above development is the ability to write the Bayes risk as a metric or distance derived from a norm. This suggests that one can use Banach spaces to explore inference procedures as operators acting between the two subspaces $\Gamma_\pi$ and $\Delta_m$. Although the notion of adjointness of operators has been extended to Banach spaces, it is mainly defined for operators acting on inner product spaces. Therefore, complete inner product spaces are the natural spaces for possible generalizations. To this end, assume the inner product has been defined as

$$(h_1, h_2)_{p,\pi} = E_\pi E_\theta C(h_1(X,\theta), h_2(X,\theta))$$

where the bivariate function $C$ is continuous and chosen so that $(h_1, h_2)_{p,\pi}$ is an inner product.

Examples of such functions for $Z = (X, \theta)$ are $C(h_1, h_2) = h_1(z)h_2(z)$, which, as we have seen, yields the Bayes risk for squared error loss. For such a function and $\lambda > 0$ one can define the inner product

$$(h_1, h_2)_{p,\pi} = E_\pi E_\theta h_1(Z)h_2(Z) + \lambda E_\pi E_\theta[(h_1(Z) - Eh_1(Z))(h_2(Z) - Eh_2(Z))]$$

which results in the penalized Bayes risk when $h(x, \theta) = \delta(x) - \gamma(\theta)$ of

$$\|\delta - \gamma\|^2_{p,\pi} = E_\pi E_\theta(\delta(X) - \gamma(\theta))^2 + \lambda E_m(\delta(X) - E_m\delta(X))^2.$$

Note that the dimensions of $\theta$ and $X$ could be large and the resulting Bayes or unbiased operators constructed based on minimization of this penalized Bayes risk could be used when finding procedures which reduce the dimensionality of the problem.

Another group of inner product Bayes risks is the Sobolev-type inner products with

$$C_j(h_1, h_2) = \sum_{\alpha_1+\alpha_2<j} \frac{\partial^{|\alpha_1+\alpha_2|}h_1(z)}{\partial^{\alpha_1}x\partial^{\alpha_2}\theta} \frac{\partial^{|\alpha+\alpha_2|}h_2(z)}{\partial^{\alpha_1}x\partial^{\alpha_2}\theta}$$

which are defined on the Sobolev spaces of bivariate functions having the needed derivatives.

A special case of this type is the inner product

$$(h_1, h_2)_{p,\pi} = E_\pi E_\theta(h_1(Z)h_2(Z)) + \lambda E(\frac{\partial h_1(Z)}{\partial X} + \frac{\partial h_1(Z)}{\partial \theta})(\frac{\partial h_2(Z)}{\partial X} + \frac{\partial h_2(Z)}{\partial \theta})$$

which induces the distance or Bayes risk

$$\|\delta - \gamma\|^2_{p,\pi} = E_\pi E_\theta(\delta(X) - \gamma(\theta))^2 + \lambda E_\pi E_\theta(\frac{\partial \delta(X)}{\partial X} - \frac{\partial \gamma(\theta)}{\partial \theta})^2.$$

This can be used to construct restricted smooth Bayes and unbiased operators.

Although constructing the corresponding operators can be difficult, being able to frame the problem of deriving Bayes and unbiased rules within an inner product space makes it possible to use the rich machinery of Hilbert space theory when studying the inference procedures.

For the proportional weighted inner product case where

$$C(h_1(x,\theta), h_2(x,\theta)) = h_1(x,\theta)h_2(x,\theta)w_1(x)w_2(\theta),$$

the derivation of the operators is straight forward. The induced inner product for $\Gamma_\pi$ is

$$(\gamma_1,\gamma_2)_\pi = E_\pi(\gamma_1(\theta)\gamma_2(\theta)E_\theta w_1(X))$$

while the induced inner product for $\Delta_m$ is

$$(\delta_1,\delta_2)_m = E_m(\delta_1(X)\delta_2(X)E(w_2(\theta)|X)).$$

The associated Bayes risk is the distance between members of these two subspaces as measured by the metric constructed from the corresponding norm. For this weighted inner product situation, our operators are given by

$$\mathcal{U} : \Delta_m \to \Gamma_\pi \text{ if and only if } \frac{E_\theta(\delta(X))}{E_\theta(w_1(X))} = \gamma(\theta) \text{ for all } \theta$$

and

$$\mathcal{B}_\pi : \Gamma_\pi \to \Delta_m \text{ if and only if } \frac{E(\gamma(\theta)w_2(\theta)|x)}{E(w_2(\theta)|x)} = \delta_\pi(x) \text{ for all } x \in \mathcal{X}.$$

Note that the theorems of the last section remain true for the above operators, as well as, any pair of Bayes and unbiasedness operators constructed within a chosen Hilbert space.

Small and McLeish (1994) present a novel approach to statistical inference that is based on projections in a Hilbert space. They discuss unbiased estimation from this new point of view but make no use of a prior distribution. At the present time, any possible connection to their work and what is presented here is unclear to us.

It should be emphasized that not all Bayes risks can be derived from an inner product. Indeed, a classical necessary and sufficient condition is that the induced norm should satisfy the parallelogram law:

$$\begin{aligned}2E_mC(\delta(X),\delta(X))+2E_\pi(C(\gamma(\theta),\gamma(\theta)) &= E_\pi E_\theta C(\delta(X)+\gamma(\theta),\delta(X)+\gamma(\theta))\\ &\quad + E_\pi E_\theta C(\delta(X)-\gamma(\theta),\delta(X)-\gamma(\theta)).\end{aligned}$$

The Bayes risk based on the absolute error loss, for example, does not satisfy the parallelogram law.

## 7 Final Remarks

We see from Eq. 3.2 that a function $\gamma$ has an unbiased estimator in our more general sense if and only if it is orthogonal to every function of the parameter which has the zero function as its Bayes estimator. From Eq. 3.3, we see that a function $\delta$ is a Bayes estimator of some $\gamma$ if and only if it is orthogonal to all unbiased estimators of the zero function.

This becomes clearer if we think about the situation where both $\Theta$ and $\mathcal{X}$ are finite and the distance is squared error loss. Let $K$ be the number of elements in $\Theta$ and $N$ be the number of elements in $\mathcal{X}$. Then $\gamma$ has an unbiased estimator $\delta$ if

$$\sum_{x\in\mathcal{X}}(\delta(x)-\gamma(\theta))p_\theta(x)=0 \quad \text{for } \theta\in\Theta$$

while $\delta$ is Bayes for $\gamma$ if

$$\sum_{\theta\in\Theta}(\delta(x)-\gamma(\theta))\pi(\theta|x)=0 \quad \text{for } x\in\mathcal{X}.$$

These are two systems of linear equations. The first has $K$ equations in $N$ unknowns, the $\delta(x)$s, while the second has $N$ equations in $K$ unknowns, the $\gamma(\theta)$s. We may arrange the members of the statistical model into the $K\times N$ stochastic matrix having as its rows the $K$ probability mass functions, $P=(p_{\theta_i}(x_j))$. The collection of posteriors yields a $N$ by $K$ stochastic matrix $\Pi=(\pi(\theta_i|x_j))$. Here $\mathcal{U}$ is the matrix $P$ and $\Pi$ is $\mathcal{B}_\pi$. It is the relationship between ranks of $P$ and $\Pi$ which determines the form of the four linear subspaces described above and results in the close relationship between being Bayes and being unbiased.

Here we have considered loss functions that can be represented as a distance function induced by the norm of an inner product space. For such problems, we have seen that operator theory from Hilbert spaces can lead to new insights about the relationship between being Bayes and unbiasedness. This suggests that there could be further insights that arise from this perspective.

*Acknowledgments.* Research was supported in part by IIR 12-340 –HSR&D grant, Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development.

## References

BLACKWELL, D. and DUBINS, L. (1962). Merging of opinions with increasing information. *Ann. Math. Stat.,* **33**, 882–886.

Cai, T.T. and Low, M. (2011). Testing composite hypotheses, hermite polynomials, and optimal estimation of a nonsmooth functional. *Ann. Stat.,* **39**, 1012–1041.

DeGroot, M.H. (1962). Uncertainty, information, and sequential experiments. *Ann. Math. Statist.,* **33**, 404–419.

DeGroot, M.H. (1984). Changes in utility as information. *Theory Decis.,* **17**, 3, 287–303.

Diaconis, P. and Freedman, D. (1986). On the consistency of bayes estimates. *Ann. Stat.,* **14**, 1–26.

Ginebra, J. (2007). On the measure of the information in a statistical experiment. *Bayesian Anal.,* **2**, 1, 167–211.

Jonhson, R.A. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Stat.,* **41**, 851–854.

Lehmann, E.L. (1951). A general concept of unbiasedness. *Ann. Math. Stat.* **22**, 587–592.

Lepski, O., Nemirovski, A. and Spokoiny, V. (1999). On estimation of the $l_r$ norm of a regression function. *Probab. Theory Related Fields,* **113**, 221–253.

Noorbaloochi, S. and Meeden, G. (1983). Unbiasedness as the dual of being Bayes. *J. Amer. Stat. Assoc.,* **78**, 619–623.

Raiffa, H. and Schlaifer, R. (2000). *Applied statistical decision theory.* Wiley Classics Library. Wiley-Interscience, New York. With a foreword by Bertrand Fox, Reprint of the 1961 original.

Rudin, W. (1991). *Functional analysis*, 2nd edn. McGraw-Hill, Boston.

Small, C.G. and McLeish, D.L. (1994). *Hilbert Space Methods in probability and statistical inference.* John Wiley, New York.

Siamak Noorbaloochi
Center for Chronic Diseases
Outcomes Research, VA Medical
Center and Department of Medicine,
University of Minnesota,
Minneapolis, MN, USA
E-mail: siamak.noorbaloochi@va.gov

Glen Meeden
School of Statistics,
University of Minnesota,
313 Ford Hall, 224 Church ST S.E.,
Minneapolis, MN 55455-0460, USA
E-mail: gmeeden@umn.edu

Paper received: 26 May 2016.