Contents

Much of this content comes from Lehmann and Casella [1998] Chapter 4.

---

1 Motivation

Motivating example:

$$X_1, \ldots, X_n \sim \text{i.i.d. } N(\theta, 1)$$
$$\bar{X} \sim N(\theta, 1/n)$$

$L(\theta, d) = (\theta - d)^2 \Rightarrow R(\theta, \bar{X}) = \text{E}[(\theta - \bar{X})^2|\theta] = 1/n$ so the risk of $\bar{X}$ is constant in θ.

What if we suspect that $\theta \approx 0$?

Regularized estimator:
$\delta_a(X) = a\bar{X}$, for some $a \in (0, 1)$.

$$\begin{aligned} R(\theta, \delta_a) = \mathrm{E}[(a\bar{X} - \theta)^2] &= \mathrm{E}[(a\bar{X} - a\theta - (1-a)\theta)^2] \\ &= \mathrm{E}[(a(\bar{X} - \theta)^2 - 2a(\bar{X} - \theta)(1-a)\theta + (1-a)^2\theta^2] \\ &= a^2/n + (1-a)^2\theta^2 \end{aligned}$$

Where does δ_a beat $\bar{X}$?
Some algebra shows that δ_a beats $\bar{X}$ when

$$\theta \in \pm\sqrt{\frac{1}{n}\frac{1+a}{1-a}}.$$

Draw the picture of the risk functions.

- The range of $\sqrt{\frac{1}{n}\frac{1+a}{1-a}}$ as a function of a is $(1/\sqrt{n}, \infty)$.
- If we are confident that $|\theta| < \sqrt{\frac{1}{n}\frac{1+a}{1-a}}$, we might want to use δ_a instead of $\bar{X}$.
- $\delta_a(X) = a\bar{X}$ doesn't beat $\bar{X}$ everywhere, just where we think θ is likely to be.
- $\delta_a(X)$ minimizes a weighted average of the risk function, the Bayes risk, under a $N(0, \tau^2)$ prior distribution on θ, where $\tau^2 = \frac{\sigma^2}{n}\frac{a}{1-a}$.

---

2 Bayes rules

2.1 Defining Bayes estimators

Suppose we have

- a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$;
- a loss function $L(\theta, d)$;
- a probability measure π on Θ.

Definition (Bayes risk)**.** *The Bayes risk of an estimator δ under a prior distribution π is*

$$R(\pi, \delta) = \int \int L(\theta, \delta(x)) P_\theta(dx) \pi(d\theta)$$

Note that unlike the real-valued risk function $R(\theta, \delta)$, $R(\pi, \delta)$ is a positive real scalar.

Definition (Bayes estimator)**.** *A Bayes estimator is a minimizer of the Bayes risk.*

The minimizer is specific to the prior π being used. We will often write a minimizer of $R(\pi, \delta)$ in δ as δ_π, so that

$$R(\pi, \delta_\pi) \leq R(\pi, \delta) \ \forall \delta \in \mathcal{D}.$$

Notes:

1. Bayes estimators do not always exist.
2. Bayes estimators are not always unique.
3. Existence and uniqueness are most easily shown on a case by case basis.

Nevertheless, we will obtain a few general results on existence and uniqueness of Bayes estimators.

2.2 Finding Bayes estimators

Bayesian calculations are most easily done in terms of densities instead of distributions. With this in mind, let $\{P_\theta : \theta \in \Theta\}$ be a set of distributions dominated by a common measure μ on $\mathcal{X}, \sigma(\mathcal{X})$, so

$$P_\theta(A) = \int_A p(x|\theta)\mu(dx) \;\; \forall \theta \in \Theta, \; A \in \sigma(\mathcal{X}).$$

We will sometimes refer to $p(x|\theta)$ as a "sampling density" for X. As a function of θ and with fixed x, $L(\theta : x) = p(x|\theta)$ is often referred to as the likelihood function.

Let ν be a measure that dominates the prior measure π, so that

$$\pi(B) = \int_B \pi(\theta)\nu(d\theta) \; \forall B \in \sigma(\Theta).$$

Note that we are overloading our notation, so that π denotes both a probability measure and density on Θ. How it is being used should be clear from the context: $\pi(\theta)$ is the prior density at θ, $\pi(B)$ is the probability of the set B, $\int f\pi(d\theta)$ is the integral of f with respect to π.

The model and prior together define a joint probability distribution on $(\mathcal{X} \times \Theta)$ defined by

$$\Pr(\{X, \theta\} \in A \times B) = \int_B \int_A p(x|\theta)\pi(\theta)\mu(dx)\nu(d\theta).$$

From this distribution we can derive other distributions that are useful for Bayesian calculations. One of these is the marginal distribution of X, also known as the prior predictive distribution:

Definition. *(Prior predictive density and distribution) The prior predictive density and distribution of X are*

$$p_\pi(x) = \int p(x|\theta)\pi(d\theta)$$
$$P_\pi(A) = \int_\Theta \int_A P_\theta(dx)\pi(d\theta) \; , \; \forall A \in \sigma(\mathcal{X})$$

The former is w.r.t. the dominating measure on $\{P_\theta : \theta \in \Theta\}$.

Interpretation:

- If π describes where you think θ is, $p_\pi(x)$ describes where you think your data will be.
- P_π is a *predictive* distribution - it doesn't depend on any unknown parameters.

  A distribution that depends on unknown parameters can't make predictions.

Definition 1. *(Posterior density and distribution) The posterior density and distribution of θ given $\{X = x\}$ is*

$$\pi(\theta|x) = p(x|\theta)\pi(\theta)/p_\pi(x)$$
$$\pi(B|x) = \int_A \pi(\theta|x)\nu(d\theta)$$

Note that we again are overloading notation: $\pi(\cdot|x)$ may refer to a density or a measure, depending on the context. If you are worried about the existence of the class of conditional distributions $\{\pi(\cdot|x) : x \in \mathcal{X}\}$, see Hoffmann-Jørgensen [1971].

Interpretation:

- $\pi(\theta)$ is where you think θ is.
- P_θ is where you think the data will be, if θ is true.
- P_π is where you think the data will be.
- $\pi(\theta|x)$ is where you think θ is, having seen $X = x$.

2.3 Minimizing posterior and Bayes risk

Having seen the data, observing $\{X = x\}$, what should your decision be?

Definition (Posterior risk)**.** *Under prior π and having observed $\{X = x\}$, the posterior risk of decision d is*

$$\begin{aligned} R(\pi, d|x) &= \mathrm{E}[L(\theta, d)|X = x] \\ &= \int L(\theta, d)\pi(d\theta|x) \end{aligned}$$

Since $\pi(\theta|x)$ represents where we think θ is after seeing $\{X = x\}$, $R(\pi, d|x)$ is our expected loss in taking decision d, where expectation/integral is over our posterior uncertainty in the value of θ.

Intuitively, having observed $\{X = x\}$ the optimal thing to do would be to minimize the posterior risk:

$$d_x = \arg\min_{d \in \mathcal{D}} R(\pi, d|x)$$

This would be the optimal decision, were we to see $\{X = x\}$. We can construct an estimator based on this optimal decision:

$$\delta_\pi(x) = d_x \ , \forall x \in \mathcal{X}.$$

It turns out that $\delta_\pi(x)$ is a Bayes estimator, i.e. it minimizes the Bayes risk.

Theorem 1. *If $\delta_\pi(x)$ minimizes $R(\pi, d|x)$ in d for x a.s. P_π, then $\delta_\pi(x)$ minimizes $R(\pi, \delta)$ and so is a Bayes estimator.*

Proof. Since L is nonnegative, by Tonelli's theorem we can write the risk as

$$\begin{aligned} R(\pi, \delta) &= \int_{\mathcal{X}} \int_{\Theta} L(\theta, \delta) p(x|\theta)\pi(\theta)\mu(dx)\nu(d\theta) \\ &= \int_{\Theta} \int_{\mathcal{X}} L(\theta, \delta)\pi(\theta|x)p_\pi(x)\nu(d\theta)\mu(dx) \\ &= \int_{\mathcal{X}} R(\pi, \delta|x)p_\pi(x)\mu(dx) = \int_{\mathcal{X}} R(\pi, \delta|x)P_\pi(dx). \end{aligned}$$

Therefore, if $\delta_\pi(x)$ minimizes $R(\pi, d|x)$ for x a.s. P_π, and δ is any other estimator, then

$$R(\pi, \delta) = \int_{\mathcal{X}} R(\pi, \delta|x)p_\pi(x)\mu(dx) \geq \int_{\mathcal{X}} R(\pi, \delta_\pi|x)p_\pi(x)\mu(dx) = R(\pi, \delta_\pi)$$

□

The minimizer of the posterior risk is the *Bayes estimate* or *Bayes decision*. The theorem says that the *estimator* or *decision rule* obtained by minimizing the posterior risk for each $x \in \mathcal{X}$ is also a Bayes estimator or rule, i.e. a minimizer of the Bayes risk.

This theorem by itself does not necessarily help us with proving admissibility: If $\delta_\pi(x)$ as the unique minimizer of the posterior risk for each x, the theorem just tells us that δ_π is Bayes, not that it is unique Bayes. What we need to know is whether or not *all* Bayes estimators are minimizers of the posterior risk. The answer is (typically) yes:

Theorem 2. *Suppose $\delta_\pi(x)$ minimizes $R(\pi, \delta)$ and $R(\pi, \delta_\pi) < \infty$. Then $R(\pi, \delta_\pi(x)|x) = \inf_{d \in D} R(\pi, d|x)$ a.s. P_π.*

Proof. By Theorem 3 of Brown and Purves [1973], for every $\epsilon > 0$ there exists a decision rule $\delta_\epsilon(x)$ such that

$$R(\pi, \delta_\epsilon(x)|x) < \epsilon + \inf_{d \in D} R(\pi, d|x).$$

Let $A = \{x : R(\pi, \delta_\pi(x)|x) > R(\pi, \delta_\epsilon(x)|x)\}$, and define the estimator $\delta_0(x)$ as

$$\delta_0(x) = \begin{cases} \delta_\epsilon(x) & \text{if } x \in A \\ \delta_\pi(x) & \text{if } x \in A^c \end{cases}$$

Since δ_π is Bayes with finite Bayes risk, we can write

$$\begin{aligned} 0 &\leq R(\pi, \delta_0) - R(\pi, \delta_\pi) \\ &= \int [R(\pi, \delta_0(x)|x) - R(\pi, \delta_\pi(x)|x)] P_\pi(dx) \\ &= \int_A [R(\pi, \delta_\epsilon(x)|x) - R(\pi, \delta_\pi(x)|x)] P_\pi(dx) \end{aligned}$$

Since the last integrand is negative on A, we must have $P_\pi(A) = 0$. Therefore

$$\begin{aligned} R(\pi, \delta_\pi|x) &\leq R(\pi, \delta_\epsilon|x) \\ &< \epsilon + \inf_{d \in \mathcal{D}} R(\pi, d|x) \text{ a.s. } P_\pi. \end{aligned}$$

Since the last line holds for all $\epsilon > 0$, we have the result. □

To summarize the last two theorems: If there exists an estimator with finite Bayes risk, then an estimator is Bayes if and only if it minimizes the posterior risk a.s. P_π. If there is no estimator with finite Bayes risk, then any estimator is Bayes in that it "minimizes" the Bayes risk, even those that do not minimize the posterior risk. For this reason, some authors define the Bayes estimator as the minimizer of the posterior risk instead of as the minimizer of the Bayes risk.

Interpretation: Assuming an estimator with finite Bayes risk exists, the fact that δ_π is Bayes if and only if it minimizes the posterior risk unites two distinct interpretations of Bayesian inference, what I will call the "subjective" and "pragmatic" schools of Bayesian inference.

Subjective Bayesian:

- $\pi(\theta)$ represents my pre-experimental information about θ.
- $\pi(\theta|x)$ represents my post-experimental information.
- Observing $\{X = x\}$, I should minimize posterior expected loss $R(\pi, d|x)$.
  * My decision is $d_x = \arg\min_{d\in D} R(\pi, d|x)$.

Pragmatic Bayesian:

- $\pi(\theta)$ puts higher weight on θ-values that I think are likely than unlikely.
- θ-values for which $\pi(\theta)$ is large contribute more to the Bayes risk $R(\pi, \delta)$.
- Minimizing Bayes risk minimizes the risk where I think θ might be.
  * My decision rule is $\delta = \arg\min_{\delta\in\mathcal{D}} R(\pi, \delta)$.

The subjective Bayesian should minimize $R(\pi, d|x)$ for each x.
The pragmatic Bayesian should minimize $R(\pi, \delta)$.
According to the theorem, these correspond to the same procedure.

This correspondence is convenient and philosophically pleasing. However, the subjective Bayesian viewpoint leads to a much more complete inferential framework. To the subjective Bayesian, $\pi(\theta|x)$ represents the information about θ having observed $X = x$, from which other procedures can be derived:

- Confidence regions: Find a set $B \subset \Theta : \ \pi(B|x) = 1 - \alpha$.
- Prediction: Forecast future data or experiments with the posterior predictive distribution:
$$\Pr(\tilde{X} \in A|x) = \int_A \int_\Theta p(d\tilde{x}|\theta)\pi(d\theta|x).$$
- Imputation, model selection, hypothesis testing and more.

2.4 Examples

Example (squared error loss):

$$L(\theta, d) = (\theta - d)^2$$
$$\mathrm{E}[L(\theta, d)|x] = \int (\theta - d)^2 \pi(d\theta|x)$$

Suppose $\mathrm{Var}[\theta|x] < \infty$. Then

$$\mathrm{E}[L(\theta, d)|x] = \mathrm{E}[\theta^2|x] - 2\mathrm{E}[\theta d|x] + d^2.$$

Taking derivatives, the minimum in d is attained at $d_x = \int \theta\pi(d\theta|x) = \mathrm{E}[\theta|x]$.
If $\delta_\pi(x) = \mathrm{E}[\theta|x]$ a.s. P_π, then $\delta_\pi(x)$ is a Bayes estimator.
If $\mathrm{Var}[\theta|x] < \infty$ a.s. P_π, then $\delta_\pi(x)$ is the unique a.s. P_π Bayes estimator.

Example (binomial model, beta prior):
$X|\theta \sim \mathrm{bin}(n, \theta) \quad \{p(x|\theta) = \mathrm{dbinom}(x, n, \theta)\}$
$\theta \sim \mathrm{beta}(a, b) \quad \{\pi(\theta) = \mathrm{dbeta}(\theta, a, b)\}$

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(d\theta)} = \frac{p(x|\theta)\pi(\theta)}{p(x)}$$
$$\propto_\theta \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$
$$\propto_\theta \theta^{a+x-1}(1-\theta)^{b+n-x-1} \propto_\theta \text{dbeta}(\theta, a+x, b+n-x).$$

The Bayes estimate of θ under squared error loss is

$$\delta_{ab}(x) = \text{E}[\theta|X=x] = \frac{a+x}{a+b+n}$$
$$= (\frac{a+b}{a+b+n})\frac{a}{a+b} + (\frac{n}{a+b+n})\frac{x}{n}$$

Interpretation:

- $a \approx$ prior number of 1's, $b \approx$ prior number of 0's.
- If $\pi(\theta)$ represents pre-experimental beliefs, $\pi(\theta|x)$ represents rational post-experimental beliefs.
- Regardless of beliefs, δ_{ab} will do better than x/n when θ is near $a/(a+b)$.

Example (absolute loss):

$$L(\theta, d) = |\theta - d|$$
$$\text{E}[L(\theta, d)|x] = \int |\theta - d|\pi(d\theta|x)$$

The minimizer in d is any median $\theta_{1/2}$ of $\pi(\theta|x)$:

$$\Pr(\theta \geq \theta_{1/2}|x) \geq 1/2 \ , \ \Pr(\theta \leq \theta_{1/2}|x) \geq 1/2$$

Prove as an exercise.

2.5 Uniqueness of Bayes rules

The above results can help us identify Bayes rules that are unique a.s. P_π: Suppose for a given prior π

- δ_π has finite Bayes risk, and
- δ_π uniquely minimizes the posterior risk.

Then any other Bayes estimator must also minimize the posterior risk a.s. P_π, and therefore equals δ_π a.s. P_π. In this case, the Bayes estimator is unique a.s. P_π.

Although unique Bayes estimators are admissible, the same is not necessarily true for Bayes estimators that are unique a.s. P_π. The problem is that $\delta_\pi(x) = \delta'_\pi(x)$ a.s. P_π implies that the Bayes risks of δ_π and δ'_π are the same, but does not imply the risk functions are equal, and so one Bayes estimator may dominate the other. This problem is illustrated in the following example:

Example(where P_π matters):
Let $X \sim \text{binomial}(n, \theta)$, $\theta \in [0, 1]$, so $\mathcal{P} = \{\text{dbinom}(x, n, \theta) : \theta \in [0, 1]\}$.
Let $\pi(\{0\}) = \pi(\{1\}) = 1/2$.
In this case $P_\pi(X = 0) = P_\pi(X = n) = 1/2$.
Note that

$$\begin{aligned}\pi(\{0\}|X = 0) &= \frac{P_0(X = 0)\pi(\{0\})}{P_0(X = 0)\pi(\{0\}) + P_1(X = 0)\pi(\{1\})} \\ &= \frac{P_0(X = 0)\pi(\{0\})}{P_\pi(X = 0)} \\ &= \frac{1 \cdot 1/2}{1/2} = 1\end{aligned}$$

Similarly, $\pi(\{1\}|X = n) = 1$.
The posterior risk of an estimator δ, a.e. P_π is given by

$$\begin{aligned}R(\pi, \delta|X = 0) &= \text{E}[L(\theta, \delta)|X = 0] = L(0, \delta(0)) \\ R(\pi, \delta|X = n) &= \text{E}[L(\theta, \delta)|X = n] = L(1, \delta(n)).\end{aligned}$$

We can minimize the posterior risk for each x a.e. P_π by making sure that $\delta(0) = 0$ and $\delta(n) = 1$. The posterior risk $R(\pi, \delta|x)$ for such an estimator is zero a.e. P_π, regardless of what $\delta(x)$ is for $x \in \{1, \ldots, n-1\}$.

Our theorem says this is a Bayes estimator, i.e. the minimizer of the Bayes risk $R(\pi, \delta)$. Let's check this result: The Bayes risk is

$$\begin{aligned} R(\pi, \delta) &= \sum_{\theta \in \{0,1\}} \sum_{x=0}^{n} L(\theta, \delta) p(x|\theta) \pi(\theta) \\ &= \tfrac{1}{2} \left[\sum_{x=0}^{n} L(0, \delta) p(x|0) + \sum_{x=0}^{n} L(1, \delta) p(x|1) \right] \\ &= \tfrac{1}{2} [L(0, \delta(0)) + L(1, \delta(1))] \\ &= 0 \quad \text{if } \delta(0) = 0 \text{ and } \delta(1) = 1 \ . \end{aligned}$$

In this case, every estimator δ for which $\delta(0) = 0$ and $\delta(n) = 1$ minimizes the posterior risk a.s. and is therefore Bayes. The Bayes estimator is unique a.s. P_π but not unique for all $x \in \mathcal{X}$.

In the above example, minimizing the Bayes risk (or the posterior risk a.s. P_π) was not sufficient to uniquely define an estimator for each $x \in \mathcal{X}$, and therefore not sufficient to show admissibility. To study admissibility, we need a notion of uniqueness that can distinguish estimators with different risk functions.

Definition. *An event $A \in \sigma(\mathcal{X})$ happens a.s. P_Θ if $P_\theta(A) = 1 \ \forall \theta \in \Theta$.*

Definition. *δ is unique Bayes if for any other Bayes estimator δ',*

$$\delta(x) = \delta'(x) \text{ a.s. } P_\Theta$$

Exercise: Show that if $\delta = \delta'$ a.s. P_Θ, then $R(\theta, \delta) = R(\theta, \delta') \ \forall \theta \in \Theta$.

Theorem 3. *If $L(\theta, d)$ is strictly convex in d, a Bayes estimator δ_π is unique if*

1. *$R(\pi, \delta_\pi) < \infty$*

2. $P_\theta \ll P_\pi \ \forall \theta \in \Theta$.

Proof. By Theorem 2, any Bayes estimator with finite Bayes risk must minimize the posterior risk a.s. P_π, so

$$R(\pi, \delta_\pi(x)|x) = \inf_{d \in \mathcal{D}} R(\pi, d|x) \text{ a.s. } P_\pi.$$

Since $P_\theta \ll P_\pi \ \forall \theta \in \Theta$,

$$R(\pi, \delta_\pi(x)|x) = \inf_{d \in \mathcal{D}} R(\pi, d|x) \text{ a.s. } P_\Theta.$$

By the strict convexity of L, and that convexity is preserved by averaging, $R(\pi, d|x)$ is convex in d for each x.

Therefore, any minimizer of $R(\pi, d|x)$ is unique (LC exercise 1.7.26), and so any Bayes estimator is unique a.e. P_Θ. □

Condition 1 will be met if there exists any estimator with finite Bayes risk.
Condition 2 will generally be met unless π puts all of its mass on θ-values for which the corresponding P_θ distributions do not cover all of $\mathcal{X}$ (recall the example).

Example (normal mean):

$$\left.\begin{array}{rl} X_1, \ldots, X_n & \sim \text{i.i.d. } N(\theta, 1/n) \\ \theta & \sim N(0, \tau^2) \end{array}\right\} \rightarrow \theta|\bar{x} \sim N(a\bar{X}, a/n)$$

where $a = \frac{n}{n+1/\tau^2}$. Under squared error loss, $\delta(x) = a\bar{x}$ is the unique minimizer of the posterior risk for every x and therefore also X a.s. P_Θ. The estimator is therefore admissible. Also note that

$$\delta(X) \begin{cases} \rightarrow \bar{X} & \text{as } \tau^2 \text{ or } n \rightarrow \infty, \\ \rightarrow 0 & \text{as } \tau^2 \rightarrow 0. \end{cases}$$

Example (predictive loss):

Consider the Kullback-Leibler loss:

$$L(\theta, d) = \int \log \frac{p(x|\theta)}{p(x|d)} p(x|\theta)\ \mu(dx),$$

which measures the predictive accuracy of $p(x|d)$ against the truth $p(x|\theta)$. Note that $L(\theta, d) \geq 0$ unless $p(x|\theta) = p(x|d)$ a.e. μ.

If $\{p(x|\theta) : \theta \in \times\}$ is an exponential family with $\times$ being the natural parameter space, then the convexity of $L(\theta, d)$ follows from the convexity of $A(\theta)$.

Example (where P_π matters):
$X|\theta \sim \text{binomial}(n, \theta),\ \theta \in \Theta = [0, 1]$
$\pi(\theta) = 1(\theta \in \{0, 1\})/2.$

$$\begin{aligned} R(\pi, \delta) &= \sum_{\theta \in \{0,1\}} \sum_{x=0}^{n} L(\theta, \delta) p(x|\theta) \pi(\theta) \\ &= \tfrac{1}{2} \left[\sum_{x=0}^{n} L(0, \delta) p(x|0) + \sum_{x=0}^{n} L(1, \delta) p(x|1) \right] \\ &= \tfrac{1}{2} [L(0, \delta(0)) + L(1, \delta(1))] \\ &= 0 \text{ if } \delta(0) = 0 \text{ and } \delta(1) = 1\ . \end{aligned}$$

As we've discussed,

- $P_\pi(\{0, 1\}) = 1$, so the Bayes estimator is uniquely defined a.e. P_π .
- $P_\theta \not\ll P_\pi\ \forall \theta \in \Theta$, so the Bayes estimator is not uniquely defined a.e. P_Θ.

For convex loss functions there will be unique minimizers of the posterior risk. This will imply uniqueness of Bayes estimators (and therefore admissibility) if $P_\theta \ll P_\pi\ \forall \theta \in \Theta$. For what priors will this condition hold?

Lemma 1. *If*

1. Θ *is open;*

2. *support*$(\pi) = \Theta$*;*

3. $P_\theta(A)$ *is continuous in* θ $\forall A \in \sigma(\mathcal{X})$,

then $P_\theta \ll P_\pi$ $\forall \theta \in \Theta$.

Proof.

Suppose $P_{\theta_0}(A) = \epsilon > 0$. We want to show this implies $P_\pi(A) > 0$.
By continuity there is a ball $B(\theta_0)$ such that $P_\theta(A) > \epsilon/2$ $\forall \theta \in B(\theta_0)$.

$$P_\pi(A) = \int P_\theta(A)\pi(d\theta) \geq \int_{B(\theta_0)} P_\theta(A)\pi(d\theta) > \tfrac{\epsilon}{2}\pi(B(\theta_0)) > 0,$$

where the last inequality holds by the support assumption.
Thus $P_{\theta_0}(A) > 0$ implies $P_\pi(A) > 0$.
Thus $P_\pi(A) = 0$ implies $P_{\theta_0}(A) = 0$. □

Note that we don't really need Θ to be open, just that $\Theta \cap B(\theta_0)$ contains an open subset of Θ for every θ_0.

2.6 Bias

$$\text{Bias}(\delta, \theta) = \text{E}[\delta|\theta] - g(\theta)$$

A biased estimator is systematically "off" from its estimand, which sounds bad.

It seems like an unbiased estimator will be closer to the estimand than a biased one. This is not true, as a biased estimator may have much smaller variances.

"Bias" also suggests subjectivity, i.e. inference that corrupts the data with prior beliefs. This notion is reinforced by the following theorem:

Theorem 4. *Let* $L(\theta, \delta) = (g(\theta) - \delta)^2$. *Then no Bayes estimator can be unbiased unless its Bayes risk is zero.*

Proof. The Bayes risk is

$$\mathrm{E}[(\delta - g)^2]] = \mathrm{E}[\delta_\pi^2 - 2\delta_\pi g + g^2]$$

If δ_π is Bayes and unbiased, then

$$\begin{aligned} \delta_\pi(X) &= \mathrm{E}[g(\theta)|X] \ \text{ since } \delta_\pi \text{ is Bayes.} \\ g(\theta) &= \mathrm{E}[\delta_\pi(X)|\theta] \ \text{ since } \delta_\pi \text{ is unbiased.} \end{aligned}$$

Then

$$\begin{aligned} \mathrm{E}[\delta_\pi(X)g(\theta)] &= \mathrm{E}[\mathrm{E}[\delta_\pi(X)g(\theta)|X]] \\ &= \mathrm{E}[\delta_\pi(X)\mathrm{E}[g(\theta)|X]] = \mathrm{E}[\delta_\pi^2(X)]. \end{aligned}$$

Similarly, you can show $\mathrm{E}[\delta_\pi(X)g(\theta)] = \mathrm{E}[g(\theta)^2]$, which means

$$\begin{aligned} R(\pi, \delta) &= \mathrm{E}[\delta_\pi(X)^2] - 2\mathrm{E}[\delta_\pi(X)g(\theta)] + \mathrm{E}[g(\theta)^2] \\ &= 0. \end{aligned}$$

□

Example (normal mean):
Under a $N(0, \tau^2)$ prior for θ, the Bayes estimator is

$$\delta_{\tau^2} = \frac{n}{n + 1/\tau^2}\bar{X},$$

which is biased (but consistent).

Example (binomial proportion):
Under a beta(a, b) prior for θ, the Bayes estimator is

$$\delta_{ab} = (\frac{a+b}{a+b+n})\frac{a}{a+b} + (\frac{n}{a+b+n})\frac{x}{n},$$

which is biased (but consistent).

The Theorem seems to imply that X (or $\bar{X}$) cannot be a Bayes estimator of its expectation under squared error loss for any prior. This is not quite true - it is only true of priors for which the Bayes risk is not zero:

Example (normal mean and variance):
$X_1, \ldots, X_n \sim$ i.i.d. $N(\mu, \sigma^2)$, $\theta = \{\mu, \sigma^2\} \in \mathbb{R} \times \mathbb{R}^+$.
Can $\bar{X}$ be Bayes for $g(\theta) = \mu$ under squared error loss?
$R(\{\mu, \sigma^2\}, \delta) = \mathrm{E}[(\bar{X} - \mu)^2 | \mu, \sigma^2] = \sigma^2/n > 0 \ \forall \mu \in \mathbb{R}, \sigma^2 > 0$
$R(\pi, \delta) = \mathrm{E}[(\bar{X} - \mu)^2] = \mathrm{E}[\sigma^2/n] > 0$, for any prior over $\sigma^2 > 0$.
(Recall, if $f(x) > 0$ a.s., then $\int f(x) P(dx) > 0$).

What about a prior for which $\pi(\{\sigma^2 = 0\}) = 1$? In this case, $\bar{X}$ has zero Bayes risk, but so does any weighted average of one or more X_i's. Thus $\bar{X}$ is Bayes for this prior, but not unique Bayes, and so admissibility of $\bar{X}$ must be proven some other way.

Finally, keep in mind that the theorem is specific to squared error loss.
Example (posterior mode for a binomial proportion):
Let

- $X \sim \mathrm{bin}(n, \theta)$;
- $\theta \sim \mathrm{beta}(1, 1)$.

Then $\{\theta | X\} \sim \mathrm{beta}(X + 1, n - X + 1)$. The posterior mode of θ is

$$\mathrm{mode}(\theta | X) = \frac{X + 1 - 1}{X + 1 - 1 + n - X + 1 - 1} = X/n.$$

The posterior mode is "sort of" the Bayes estimator under 0-1 loss, where $L(\theta, d) = 1(\theta \neq d)$.

3 Prior distributions

How should the prior distribution be chosen? There are many perspectives on this:

- (subjective Bayes): π should be the "real" prior distribution, a probabilistic representation of real prior knowledge.
- (pragmatic Bayes): Select π from a computationally convenient class of priors (e.g. conjugate priors) that can represent some aspects of real prior information (such as prior expectation and variance).
- (objective Bayes): Use a "noninformative" prior. This approach sounds oxymoronic, but can relate Bayes procedures to procedures that are optimal under different criteria.
  - Jeffreys' prior;
  - improper priors;
  - invariant priors.

3.1 Conjugate priors

Definition. *A class* $\Pi = \{\pi(\theta)\}$ *of prior distributions is called conjugate for a model* $\mathcal{P} = \{p(x|\theta) : \theta \in \Theta\}$ *if*

$$\pi(\theta) \in \Pi \Rightarrow \pi(\theta|x) \in \Pi.$$

A conjugate prior class is "closed under sampling."

Example (binomial-beta family):

$$\begin{aligned} X|\theta &\sim \text{bin}(n,\theta) \quad \{p(x|\theta) = \text{dbinom}(x,n,\theta)\} \\ \theta &\sim \text{beta}(a,b) \quad \{\pi(\theta) = \text{dbeta}(\theta,a,b)\} \end{aligned}$$

$$\begin{aligned} \pi(\theta|x) &= p(x|\theta)\pi(\theta)/p(x) \\ &\propto_\theta p(x|\theta)\pi(\theta) \\ &\propto_\theta \theta^x(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1} \\ &\propto_\theta \theta^{a+x-1}(1-\theta)^{b+n-x-1} \\ &\propto_\theta \text{dbeta}(\theta, a+x, b+n-x) \end{aligned}$$

Also note that

$$\mathrm{E}[\theta|x] = \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{x}{n}.$$

Example (normal-normal family):

$$\begin{array}{rll} \{X_1,\ldots,X_n|\theta,\sigma^2\} & \sim \text{i.i.d. normal}(\theta,\sigma^2) & \{p(x|\theta,\sigma^2) = \prod_{i=1}^n \text{dnorm}(x_i,\theta,\sigma)\} \\ \theta & \sim \text{normal}(\theta_0,\tau_0^2) & \{\pi(\theta) = \text{dnorm}(\theta,\theta_0,\tau_0)\} \end{array}$$

$$\begin{aligned} \pi(\theta|x,\sigma^2) &= p(x|\theta,\sigma^2)\pi(\theta)/p(x|\sigma^2) \\ &\propto_\theta p(x|\theta,\sigma^2)\pi(\theta) \\ &\propto_\theta \exp\{-\textstyle\sum x_i^2/(2\sigma^2) + \theta\sum x_i/\sigma^2 - n\theta^2/(2\sigma^2)\} \times \exp\{-\theta^2/(2\tau_0^2) + \theta_0\theta/\tau_0^2 - \theta_0^2/(2\tau_0^2)\} \\ &\propto_\theta \exp\{-\tfrac{1}{2}\theta^2(n/\sigma^2 + 1/\tau_0^2) + \theta(n\bar{x}/\sigma^2 + \theta_0/\tau^2)\} \\ &\propto_\theta \text{dnorm}(\theta,\theta_n,\tau_n) \end{aligned}$$

where

- $\tau_n^2 = (n/\sigma^2 + 1/\tau_0^2)^{-1}$
- $\theta_n = \tau_n^2(n\bar{x}/\sigma^2 + \theta_0/\tau^2)$

Also note that

$$\mathrm{E}[\theta|x_1,\ldots,x_n] = \theta_n = \frac{1/\tau^2}{n/\sigma^2 + 1/\tau_0^2}\theta_0 + \frac{n/\sigma^2}{n/\sigma^2 + 1/\tau_0^2}\bar{x}.$$

Note that for both examples,

$$\mathrm{E}[\theta|X_1,\ldots,X_n] = w\mathrm{E}[\theta] + (1-w)\bar{X}.$$

Does this hold more generally?

Conjugate priors for exponential families

$p(x|\theta) = \exp\{\theta\cdot t(x) - A(\theta)\}$

$x_1, \ldots, x_n \sim$ i.i.d. $p(x|\theta) \Rightarrow p(\boldsymbol{x}|\theta) = \exp\{\theta \cdot \sum t(x_i) - nA(\theta)\} = \exp\{\theta \cdot n\bar{t} - nA(\theta)\}$

Conjugate prior: Consider $\pi(\theta) \propto_\theta \exp\{\theta \cdot n_0 t_0 - n_0 A(\theta)\}$. Then

$$\pi(\theta|\boldsymbol{x}) \propto \exp\{\theta \cdot (n_0 t_0 + n\bar{t}) - (n_0 + n)A(\theta)\}$$

Then $\pi(\theta)$ and $\pi(\theta|\bar{x})$ have the same form, i.e. they are both members of the class of densities given by

$$\pi_{n_0 t}(\theta) = \frac{\exp\{\theta \cdot n_0 t_0 - n_0 A(\theta)\}}{\int \exp\{\theta \cdot n_0 t_0 - n_0 A(\theta) d\theta\}}$$

What we haven't checked yet is whether or not such densities are actual probability densities, i.e. the integral in the denominator is finite.

Theorem 5 (Diaconis and Ylvisaker [1979]). *If $n_0 > 0$ and $t_0 \in \text{Conv}(\{t(x) : x \in \mathcal{X}\})$, then*

$$\int_{\mathcal{H}} \exp\{\theta \cdot n_0 t_0 - n_0 A(\theta) d\theta\} \equiv c(n_0, t_0)^{-1} < \infty,$$

and $\pi(\theta) = c(n_0, t_0) \exp\{\theta \cdot n_0 t_0 - n_0 A(\theta) d\theta\}$ is a probability density on the natural parameter space $\mathcal{H}$.

Here, $\text{Conv}(A)$ is the convex hull of the set $A \subset \mathbb{R}^s$:

$$\text{Conv}(A) = \{x \in: \mathbb{R}^s : x = \sum_1^K w_k a_k, (a_1, \ldots, a_K) \subset A, \sum w_k = 1, K \in \mathbb{N}\}.$$

The allowable values of t_0 include all of those that could be expressed as a sample average of $t(X)$. Thus we have the following interpretation:

- $n_0 \approx$ prior sample size
- $t_0 \approx$ average of $t(x_1), \ldots t(x_{n_0})$ from the "prior sample".

Let's try to push this interpretation further: Recall that

$$\text{E}[t(X)|\theta] = \nabla A(\theta) \quad \text{Var}[t(X)|\theta] = \nabla^2 A(\theta).$$

Consider estimation of $\mu(\theta) = \text{E}[t(X)|\theta] = A(\theta)$. For example,

- normal model: $\mu = \mathrm{E}[(X, X^2)|\theta] \equiv (\mu, \mu^2 + \sigma^2)$,
- Poisson model: $\mu = \mathrm{E}[X|\theta]$.

Under $\pi(\theta|n_0, t_0)$, what is $\mathrm{E}[\mu] = \mathrm{E}[\nabla A(\theta)]$? Diaconis and Ylvisaker [1979] show that

$$\int \nabla \pi(\theta)\, d\theta = 0.$$

But we also have

$$\begin{aligned} \nabla \pi(\theta) &= \nabla(e^{\theta \cdot n_0 t_0 - n_0 A(\theta)} c(n_0, t_0)) \\ &= [n_0 t_0 - n_0 \nabla A(\theta)] \pi(\theta). \end{aligned}$$

Putting these together gives

$$\int \nabla \pi(\theta)\, d\theta = \int (n_0 t_0 - n_0 \nabla A(\theta)) \pi(\theta)\, d\theta = 0.$$

which implies $n_0 t_0 = n_0 \mathrm{E}[\nabla A(\theta)] = n_0 \mathrm{E}[\mu]$, and so under $\pi(\theta|n_0, t_0)$, $\mathrm{E}[\mu] = t_0$.

Therefore, when selecting a particular member of the conjugate class, you should

- pick t_0 to represent your prior expectation of $\mu = \mathrm{E}[t(X)]$, that is, your prior guess at the population mean of $t(X)$;
- pick n_0 to represent your degree of confidence in your prior guess.

Exercise: Take second derivatives to identify the relationship between n_0 and the prior variance of μ.

Linear Bayes:
Suppose we use the conjugate prior $\pi(\theta|n_0, t_0)$, under which $\mathrm{E}[\mu] = t_0$. Given i.i.d. samples from $p(x|\theta)$, we have

$$\begin{aligned} \pi(\theta|x_1, \ldots, x_n) &\propto \exp\{\theta(n_0 t_0 + n\bar{t}) - (n_0 + n) A(\theta)\} \\ &\propto \pi(\theta|n_1, t_1) \end{aligned}$$

where $n_1 = n_0 + n$ and

$$t_1 = \frac{n_0 t_0 + n\bar{t}}{n_1} = \frac{n_0}{n_0+n} t_0 + \frac{n}{n_0+n}\bar{t}$$

By our calculations above, we must have $\mathrm{E}[\mu] = t_0$ and $\mathrm{E}[\mu|x_1, \ldots, x_n] = t_1$, so

$$\mathrm{E}[\mu|x_1, \ldots, x_n] = \frac{n_0}{n_0+n}\mathrm{E}[\mu] + \frac{n}{n_0+n}\bar{t}.$$

Recall that under squared error loss, this posterior mean is the unique Bayes estimator. Thus conjugate priors generate "linear shrinkage" estimators of $\mu = \mathrm{E}[t(X)|\theta]$ of the form

$$\begin{aligned}\delta(X) &= \frac{n_0}{n_0+n} t_0 + \frac{n}{n_0+n}\bar{t} \\ &= (1-w)t_0 + w\bar{t}\end{aligned}$$

Such estimators allow for

- an amount of shrinkage w;
- a direction of shrinkage t_0.

References

L. D. Brown and R. Purves. Measurable selections of extrema. *Ann. Statist.*, 1: 902–912, 1973. ISSN 0090-5364.

Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families. *Ann. Statist.*, 7(2):269–281, 1979. ISSN 0090-5364.

J. Hoffmann-Jørgensen. Existence of conditional probabilities. *Math. Scand.*, 28: 257–264, 1971. ISSN 0025-5521.

E. L. Lehmann and George Casella. *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 1998. ISBN 0-387-98502-6.