LECTURE 7: BAYESIANS VERSUS FREQUENTISTS

Michal Kolesár*

September 23, 2024

REFERENCE CB, Chapter 7.2.3 and 7.3; HMC, Chapter 11.1–11.2 or H22, Chapter 16. For a more thorough discussion of the complete class theorem, a good reference is Ferguson (1967, Chapter 2).

1. FREQUENTIST AND BAYESIAN PARADIGMS

In frequentist statistics, the unknown parameter θ is some fixed number or vector. Given a parameter value θ, we observe data X from a distribution $F(\cdot \mid \theta)$. To estimate the parameter θ, we use an estimator $\hat{\theta}(X)$. Since it depends on the data, the estimator is a random variable. The randomness here is due to sampling. We argued in Lecture 3 that to evaluate the performance of the estimator, we should use expected loss (risk). We also discussed how we may approximate the risk by establishing the estimator's asymptotic properties. We justified this evaluation from an ex-ante perspective: if we have to commit to the estimator before seeing the data, using risk can be justified by the expected utility theorem. This form of evaluation *averages the performance over different draws from the same model, with fixed parameters*.

For instance, if the estimator is consistent, this means that it is very unlikely that we will draw a sample such that our estimator $\hat{\theta}(X)$ is further than some fixed distance δ away from the true parameter value θ, at least if our sample size is large enough. Said differently, if we repeated the sampling experiment many times, $\hat{\theta}(X)$ will be close to θ most of the time.

In contrast, Bayesian theory posits that the uncertainty about the true parameter value θ can be embedded in a probability distribution. Let $\pi(\theta)$ be a probability density function (PDF) describing uncertainty about θ before we see the data, called the *prior distribution*, so that the joint PDF of (X, θ) is given by $f(x \mid \theta)\pi(\theta)$. We never observe the "realized value" of θ, only the realized value of X. Once we observe X, the best thing we can do is to update our beliefs about θ by calculating the conditional distribution of θ given X. This conditional distribution is called the *posterior*. The posterior can be used to make all decisions concerning θ, such as creating an estimator. Since we condition on the data X, the randomness in posterior is only due to uncertainty about θ.

*Email: mkolesar@princeton.edu.

We argued previously that the ex ante perspective of a frequentist makes since in situations where a statistician needs to communicate to an audience with diverse beliefs about θ, or when a committee with diverse beliefs needs to make a decision. In contrast, the Bayesian viewpoint is more appealing when the statistician is the sole decision-maker, in that their payoff directly depends on what $\hat{\theta}(X)$ they report. This is the case in forecasting (think Nate Silver or the Fed), or when you're a CEO of a company that needs to make a decision. In such cases, there is no need to commit to a decision rule ex ante before seeing the data, so it matters little what the ex-ante properties of your procedure are.

2. BAYESIAN UPDATING

Let $\pi(\theta)$ denote our prior. Our statistical model is that X has PDF $f(x \mid \theta)$, that is, the observed data X are drawn from the conditional distribution $f(\cdot \mid \theta)$. The joint PDF of θ and X is $\pi(\theta)f(x \mid \theta)$. The marginal distribution of the data is therefore

$$m(x) = \int_\Theta \pi(\theta) f(x \mid \theta)\, d\theta.$$

Sometimes this is called the *prior predictive distribution*. Using the relationship between joint and conditional distributions, it follows that the posterior can be calculated as

$$\pi(\theta \mid x) = \frac{\pi(\theta) f(x \mid \theta)}{m(x)}.$$

In other words, since $m(x)$ is just a normalizing constant that makes sure $\pi(\theta \mid x)$ integrates to one (over θ), the posterior is proportional to the likelihood times the prior, $\text{posterior}(\theta) \propto \mathcal{L}(\theta) \times \pi(\theta)$.

Example 1. Let $X = (X_1, \ldots, X_n)$ be a random sample from a Bernoulli(p) distribution, and let $S = \sum_{i=1}^n X_i$. Then the joint PDF of the data is

$$f(x \mid p) = p^s (1-p)^{n-s}.$$

To calculate the posterior, we need a prior distribution of p. Suppose we believe that all p are equally likely. Then we have a uniform prior, i.e. $\pi(p) = \mathbb{1}\{0 \le p \le 1\}$. Then the joint PDF of p and X is $p^s(1-p)^{n-s}\,\mathbb{1}\{0 \le p \le 1\}$, which gives

$$m(x) = \int_0^1 p^s (1-p)^{n-s}\, \mathrm{d}p = B(s+1, n-s+1),$$

where, by definition $B(x,y) := \int_0^1 t^{x-1}(1-t)^{y-1}\, \mathrm{d}t$ is the Beta function. The posterior

distribution is

$$\pi(p \mid x) = \frac{p^s(1-p)^{n-s}}{B(s+1, n-s+1)} \mathbb{1}\{0 \leq p \leq 1\} = \frac{p^{(s+1)-1}(1-p)^{(n-s+1)-1}}{B(s+1, n-s+1)} \mathbb{1}\{0 \leq p \leq 1\}.$$

This distribution is called *Beta* with parameters $\alpha = s+1$ and $\beta = n-s+1$, denoted $\mathcal{B}(\alpha, \beta)$. Note that $\mathcal{B}(1,1) \sim \mathcal{U}[0,1]$, and if $p \sim \mathcal{B}(\alpha, \beta)$, then

$$E[p] = \frac{\alpha}{\alpha + \beta}$$

and

$$\text{var}(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

So we start with the prior $p \sim \mathcal{B}(1,1)$, and end up with the posterior $p \mid X \sim \mathcal{B}(S+1, n-S+1)$. Does this make intuitive sense? Note that the posterior mean can be written as a weighted average between the sample mean s/n and the prior mean $1/2$:

$$E[p \mid X] = \frac{s+1}{n+2} = w\frac{s}{n} + (1-w)\frac{1}{2}$$

where $w = n/(n+2)$. One can think of the prior as a belief one would hold after observing two "hypothetical" draws, one of which was a success. w then corresponds to the ratio of the sample size n to the "total" sample size, including these two hypothetical draws. The parameters α and β in the posterior correspond to the "total" number of successes and failures.

With this in mind, suppose our prior is $\mathcal{B}(2,8)$, so the prior mean is $1/5$, and prior variance is $4/275 \approx 0.145$. Suppose we observe 4 observations, 3 of which are successes. Then the posterior is $\mathcal{B}(2+3, 8+1)$, and the posterior variance is $3/196 \approx 0.155$. This is somewhat unusual: typically the posterior variance is lower than the prior variance, as the data provides us with extra information. Since the extra information is at odds with our prior beliefs implies that our uncertainty is actually increased by the extra information. ⊠

2.1. *How to calculate the posterior distribution*

Since $m(x)$ as well as any other parts of the prior and likelihood that don't depend on the data are just normalizing constants that ensure the posterior integrates to 1, one possible way of computing the posterior is to calculate $\pi(\theta)f(x \mid \theta)$ up to all multiplicative terms which do not contain θ, and then integrate it in order to find the normalizing constant.

Example 2. Suppose $X = (X_1, \ldots, X_n)$ is a random sample from $\mathcal{N}(\mu, \sigma^2)$, and that σ^2 is

known. Suppose the prior is $\mathcal{N}(\mu_0, \tau^2)$. Then

$$\pi(\mu \mid X = x) \propto f(x \mid \mu)\pi(\mu) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{1}{2\tau^2}(\mu - \mu_0)^2\right).$$

Thus, by "completing the square",

$$\begin{aligned}
\pi(\mu \mid X = x) &\propto \exp\left(\frac{\mu n \bar{x}_n}{\sigma^2} - \frac{n\mu^2}{2\sigma^2} - \frac{\mu^2}{2\tau^2} + \frac{\mu\mu_0}{\tau^2}\right) \\
&= \exp\left(-\mu^2\left(\frac{n}{2\sigma^2} + \frac{1}{2\tau^2}\right) + 2\mu\left(\frac{n\bar{x}_n}{2\sigma^2} + \frac{\mu_0}{2\tau^2}\right)\right) \\
&= \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)\left[\mu^2 - 2\mu\left(\frac{n\bar{x}_n/\sigma^2 + \mu_0/\tau^2}{n/\sigma^2 + 1/\tau^2}\right)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)[\mu - \tilde{\mu}]^2\right\} = \exp\left\{-\frac{(\mu - \tilde{\mu})^2}{2\tilde{\sigma}^2}\right\},
\end{aligned}$$

where

$$\tilde{\mu} = \tilde{\sigma}^2\left(\frac{n\bar{x}_n}{\sigma^2} + \frac{\mu_0}{\tau^2}\right), \qquad \tilde{\sigma}^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}. \tag{1}$$

We recognize the posterior as the density of the normal $\mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$, so that the missing constant is $C = 1/(\sqrt{2\pi}\tilde{\sigma})$.

Note that the posterior mean is a weighted average of the sample mean $\bar{x}_n$ and prior mean μ_0, weighted by the precision (inverse of the variance) of each. Similarly, the posterior precision is the sum of the precisions of $\bar{x}_n$ and μ_0. Note that as $n \to \infty$, for any fixed τ^2, the sample precision dominates, and the difference between the posterior mean and the sample mean will vanish, $\tilde{\mu} - \bar{x}_n \overset{\text{a.s.}}{\to} 0$ and $n\tilde{\sigma}^2 \to \sigma^2$, so that we have, approximately

$$\mu \mid x \sim_{\text{approximately as } n \to \infty} \mathcal{N}(\bar{x}_n, \sigma^2/n).$$

There is also a less mechanical way of calculating the posterior in this example. In particular, note that by sufficiency, observing the data is equivalent to observing $\overline{X}_n$, which we can split as $\overline{X}_n = \mu + Z$, where $Z \sim \mathcal{N}(0, \sigma^2/n)$, and μ and Z are independent. Therefore,

$$\begin{pmatrix} \overline{X}_n \\ \mu \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_0 \\ \mu_0 \end{pmatrix}, \begin{pmatrix} \tau^2 + \sigma^2/n & \tau^2 \\ \tau^2 & \tau^2 \end{pmatrix}\right).$$

To derive $\mu \mid \overline{X}_n$, we can use the formula Lecture Note 1 on conditional normal distributions, which yields $\mu \mid \overline{X}_n \sim \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$. ⊠

An important issue with Bayesian estimation is that if a prior distribution puts zero probability mass on the true parameter value (jargon for this is that the prior is *dogmatic*), then no matter how large our sample is, posterior distribution will put zero mass on the true parameter value as well.

Let $\mathcal{F} = \{F(\cdot \mid \theta) : \theta \in \Theta\}$. Let $\mathcal{P}$ be the class of prior distributions for θ. Then we say that $\mathcal{P}$ is *conjugate* to $\mathcal{F}$ if whenever data is distributed according to $\mathcal{F}$ and prior distribution is from $\mathcal{P}$, then the posterior distribution is from $\mathcal{P}$ as well.

For example, we have already seen that the class of normal distributions is conjugate to the class of normal distributions with known (fixed) variance. It is also known that the class of Beta distributions is conjugate to the class of binomial distributions. The concept of conjugate classes is introduced because of its mathematical convenience. If the prior lies in the conjugate class, it is possible to calculate the posterior in closed form. For this reason, conjugate priors were almost the only priors used for a long time due to their analytical tractability.

For example, suppose that in Example 2, we specify the prior for μ to be Cauchy, with scale γ, which has density

$$\pi(\mu) = \frac{1}{\pi\gamma} \frac{\gamma^2}{\mu^2 + \gamma^2}.$$

This distribution has heavy tails, so it would better capture our prior beliefs if we want to put a lot of prior mass on extreme values of μ. But then the posterior is

$$\pi(\mu \mid X = x) \propto f(x \mid \mu)\pi(\mu) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2\right) \frac{\gamma^2}{\mu^2 + \gamma^2},$$

which is not a density we recognize. Nowadays, there are numerical techniques such as Markov Chain Monte Carlo methods that allow one to numerically calculate any feature of the posterior that one may be interested in even for priors outside the conjugate family.

3. BAYESIAN DECISIONS

From a Bayesian viewpoint, all there is to learn about θ, the value that actually generates X, is contained in its posterior distribution.

If 99% of the probability mass of the posterior is in some set C, then we can say that is "likely" or "highly probable" that θ is an element of C. One can use this fact to summarize the remaining uncertainty about θ using *credible intervals*. If $\pi(\theta \mid x)$ is the posterior, then for any $\alpha \in [0,1]$, a set $C(x) \subset \Theta$ is called $(1-\alpha)$-credible if $\pi(\theta \in C(x) \mid X = x) \geq 1 - \alpha$. In words, set $C(x)$ contains true parameter value θ with probability of at least $1 - \alpha$. Of course, the whole parameter space Θ is $(1-\alpha)$-credible, which is not particularly useful. Therefore, we should try to choose the smallest $(1-\alpha)$-credible set. The smallest $(1-\alpha)$-credible set contains only points with the *highest posterior density*.

Example 2 (continued). We have $\mu \mid x \sim N(\tilde{\mu}, \tilde{\sigma}^2)$ with $\tilde{\mu}$ and $\tilde{\sigma}^2$ defined in eq. (1). Then $(\mu - \tilde{\mu})/\tilde{\sigma} \sim \mathcal{N}(0,1)$. Let z_α be the α-quantile of the standard normal distribution. Then

$$\pi\left((\mu - \tilde{\mu})/\tilde{\sigma} \in [z_{\alpha/2}, z_{1-\alpha/2}] \mid X = x\right) = 1 - \alpha,$$

or, equivalently,

$$\pi\left(\tilde{\mu} + z_{\alpha/2}\tilde{\sigma} \leq \mu \leq \tilde{\mu} + z_{1-\alpha/2}\tilde{\sigma} \mid X = x\right) = 1 - \alpha.$$

Since the PDF $\phi(x)$ of the standard normal distribution is decreasing on $x \geq 0$ and increasing on $x \leq 0$, $[\tilde{\mu} + z_{\alpha/2}\tilde{\sigma}, \tilde{\mu} + z_{1-\alpha/2}\tilde{\sigma}]$ is the shortest $(1-\alpha)$-credible interval. ⊠

Furthermore, if the aim is to take an action $a \in \mathcal{A}$ that minimizes some loss $L(a, \theta)$, then we can simply choose the action a^* that minimizes the expected loss under the posterior:

$$a^* = \operatorname*{argmin}_{a \in \mathcal{A}} \int_\Theta L(a, \theta)\pi(\theta \mid x)\, \mathrm{d}\theta. \tag{2}$$

Note that different actions (like estimates or decisions to reject a hypothesis) are now judged with respect to their expected performance with respect to the *posterior distribution*, in contrast to frequentist measures.

Consider estimation under quadratic loss, $L(a, \theta) = (a - \theta)^2$, with $\mathcal{A} = \Theta \subseteq \mathbb{R}$. The estimator is then given by the *posterior mean*, as can be seen from plugging the loss into (2) and minimizing (take a first-order condition, and set it to zero). Under absolute value loss, we pick the *posterior median*.[1]

This setup also covers testing. Suppose we have a null hypothesis $\theta \in \Theta_0$, and the alternative is $\theta \in \Theta_1$, with $\Theta = \Theta_1 \cup \Theta_0$, and the action space is $\mathcal{A} = \{0, 1\}$ with 1 meaning "reject". Suppose the loss is $L(a, \theta) = a \cdot \mathbb{1}\{\theta \in \Theta_0\} + \lambda(1 - a) \cdot \mathbb{1}\{\theta \in \Theta_1\}$ (I lose a dollar if I reject a true null and lose λ dollars if I don't reject a false null). The Bayes rule minimizing the average loss minimizes (2), which can in this case be written as

$$\begin{aligned}
&\int_{\Theta_0} a\pi(\theta \mid x)\, \mathrm{d}\theta + \lambda \int_{\Theta_1} (1 - a)\pi(\theta \mid x)\, \mathrm{d}\theta \\
&\qquad = a\pi(\theta \in \Theta_0 \mid X = x) + \lambda(1 - a)\pi(\theta \in \Theta_1 \mid X = x) \\
&\qquad = a\left[\pi(\theta \in \Theta_0 \mid X = x) - \lambda\pi(\theta \in \Theta_1 \mid X = x)\right] + \lambda\pi(\theta \in \Theta_1 \mid X = x).
\end{aligned}$$

This is minimized at

$$a^* = \begin{cases} 1 & \text{if } \frac{\pi(\theta \in \Theta_0 \mid X = x)}{\pi(\theta \in \Theta_1 \mid X = x)} \leq \lambda, \\ 0 & \text{otherwise.} \end{cases}$$

We reject if the posterior puts a large enough mass on Θ_1 relative to the mass it puts on Θ_0.

1. To see this, consider a random variable X, and suppose we're interested in picking m that minimizes $E[|X - m|] = \int_{-\infty}^{m} (m - X)\, \mathrm{d}F(X) + \int_{m}^{\infty} (X - m)\, \mathrm{d}F(X)$. Then by Leibniz rule, the derivative is $P(X \leq m) - P(X \leq m)$, which is set to zero at the median of X.

Example 2 (continued). Suppose we want to test the hypothesis that $\mu \leq 0$. Then we reject the null if the ratio

$$\frac{\pi(\theta < 0 \mid x)}{\pi(\theta > 0 \mid x)} = \frac{\pi(\theta < 0 \mid x)}{1 - \pi(\theta < 0 \mid x)}$$

is small. Since $\pi(\theta < 0 \mid x)$ is decreasing with $\bar{x}_n$, the test will reject (for a fixed prior), if the observed sample mean is sufficiently large. Note that it doesn't make sense to test a particular value θ_0 of θ, since that has posterior probability zero, and the test would always reject it. To allow for such cases, we'd need a prior distribution that puts a positive probability mass on θ_0. ⊠

4. COMPLETE CLASS THEOREM

We now link Bayesian decisions to the frequentist notion of risk.

Recall that in a statistical decision problem $(\Theta, \mathcal{A}, L)$, we evaluate decision rules $\delta(X)$, where $X \sim \mathcal{F}(\cdot \mid \theta)$ based on their risk

$$R(\delta, \theta) = E[L(\delta(X), \theta)] = \int_{\mathcal{X}} L(\delta(x), \theta) f(x \mid \theta) \, \mathrm{d}x.$$

If a decision rule δ is not dominated by another decision, then it is called admissible. What are the risk properties of Bayes rules? Notice that a Bayes decision rule $\delta_\pi(x)$ that minimizes the posterior risk (2) can be seen to also minimize

$$\delta_\pi(x) = \operatorname*{argmin}_{a \in \mathcal{A}} \int_\Theta L(a, \theta) \frac{f(x \mid \theta)\pi(\theta)}{m(x)} \, \mathrm{d}\theta = \operatorname*{argmin}_{a \in \mathcal{A}} \int_\Theta L(a, \theta) f(x \mid \theta) \pi(\theta) \, \mathrm{d}\theta,$$

since the scaling by $m(x)$ doesn't affect the minimum. Therefore, $\delta_\pi(x)$ also minimizes

$$\delta_\pi = \operatorname*{argmin}_{\delta \colon \delta(x) \in \mathcal{A}} \int_{\mathcal{X}} \int_\Theta L(\delta(x), \theta) f(x \mid \theta) \pi(\theta) \, \mathrm{d}\theta \, \mathrm{d}x,$$

since it minimizes the inner integral for each x. Swapping the order of integration, this is equivalent to

$$\delta_\pi = \operatorname*{argmin}_{\delta \colon \delta(x) \in \mathcal{A}} \int_\Theta R(\delta, \theta) \pi(\theta) \, \mathrm{d}\theta.$$

Bayes rules minimize the average frequentist risk, with the prior serving as the weighting function.

Note that, in general (without regularity conditions), δ_π may not be unique, and it may not even exist—same issue arises with the minimax criterion we discussed earlier.

Remark 1. Bayes estimators are by construction guaranteed to have good (the best!) average risk properties. What about other frequentist properties? First, *in general, Bayes estimators are biased*, $E_{X|\theta}[\delta_\pi(X)] \neq \theta$. This can be seen clearly in Example 2. Under

squared error loss the optimal estimator is the posterior mean, which we can write as

$$\tilde{\mu} = \lambda \overline{X}_n + (1-\lambda)\mu_0, \qquad \lambda = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1} \frac{n}{\sigma^2}.$$

Its bias is given by $E_{X|\theta}[\tilde{\mu}] - \mu = (1-\lambda)(\mu_0 - \mu)$, so we only get unbiasedness in the limit as $\lambda \to 1$, which is equivalent to $\tau \to \infty$, but for any "proper" prior with fixed τ and μ_0, the posterior mean is biased. Second, *average risk optimality doesn't necessarily imply good minimax performance*: just because the average risk is low, it doesn't mean the worst-case risk is low (and vice versa, one can criticize the minimax criterion on the grounds that low maximum risk doesn't necessarily imply good average risk). In the normal mean example, the risk is

$$R(\tilde{\mu}, \mu) = (1-\lambda)^2(\mu_0 - \mu)^2 + \lambda^2\sigma^2/n,$$

which blows up as the distance between μ_0 and μ gets large, so that the worst-case risk is infinite. In contrast, the worst-case risk of $\overline{X}_n$ is σ^2/n. On the other hand, the risk of $\tilde{\mu}$ is lower than that of $\overline{X}_n$ if μ_0 is close to μ, so that "on average" (averaged using the prior weights), the risk is better:

$$\begin{aligned}\int R(\tilde{\mu}, \mu)\pi(\mu)\mathrm{d}\mu &= (1-\lambda)^2\tau^2 + \lambda^2\sigma^2/n \\ &= \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1} < \left(\frac{n}{\sigma^2}\right)^{-1} = \int R(\overline{X}_n, \mu)\pi(\mu)\mathrm{d}\mu,\end{aligned}$$

where the first equality follows because the expectation of $(\mu_0 - \mu)^2$ under the prior is τ^2, and the second equality follows by algebra, and the last equality follows because the risk of $\overline{X}_n$ is constant and equal to σ^2/n, so that its average risk is also σ^2/n, no matter what the prior $\pi(\mu)$ is.

The next question we take up is admissibility of Bayes rules. To simplify the analysis, we further simplify the setup by assuming that $\Theta = \{\theta_1, \ldots, \theta_m\}$, and $\mathcal{A} = \{a_1, \ldots, a_k\}$ (i.e. both the parameter space and the action space are finite). Now a couple of definitions:

- A *mixed action* is given by the k-vector $q = (q_1, \ldots, q_k)$, which means that a_i is chosen with probability q_i. Of course, $\sum_{i=1}^k q_i = 1$, and $q_i \geq 0$. The loss of a mixed action is
$$L(\theta, q) = \sum_{i=1}^n q_i L(\theta, a_i) \tag{3}$$
Let $\mathcal{A}^*$ denote the space of all mixed actions.

- A *randomized decision rule* is a function $\delta\colon \mathcal{X} \to \mathcal{A}^*$, and we can denote the set of all randomized rules (which includes all non-randomized rules) by $\mathcal{D}^*$ (there are some measurability constraints in its definition, which we ignore). Note that $\mathcal{D}^*$ is convex: if $\delta_1, \delta_2 \in \mathcal{D}^*$, then so is their average. The point of allowing for

randomized rules is exactly to make the space of all decision rules convex, which simplifies the analysis below.

- The *risk set* $\mathcal{S}$ is the set of risks $(R(\delta,\theta_1),\dots,R(\delta,\theta_m))' \in \mathbb{R}^m$ that can be achieved with some decision rule δ. Note that $\mathcal{S}$ is convex, because if δ_1 and δ_2 have risks $y_1, y_2 \in \mathcal{S}$, then it follows from Equation (3) that $\alpha\delta_1 + (1-\alpha)\delta_2$ has risk $\alpha y_1 + (1-\alpha)y_2$.
- A *complete class* of decision rules is a class $\mathcal{C} \subseteq \mathcal{D}^*$ that contains all admissible decision rules.

We're now ready to state our first big result:

Theorem 2. If a Bayes rule δ_π wrt prior distribution $\pi = \{\pi_1,\dots,\pi_m\}$ on Θ exists, and $\pi_j > 0$ for all j, then it is admissible.

Proof. Suppose δ_π is not admissible. Then there exists a rule δ' such that $R(\delta_\pi,\theta_j) \geq R(\delta',\theta_j)$ for all j, with at least one inequality strict. Because all π_j are positive, this implies that

$$\sum_{j=1}^m \pi_j R(\delta_\pi,\theta_j) > \sum_{j=1}^m \pi_j R(\delta',\theta_j),$$

which contradicts δ_π being Bayes wrt π. □

Digression. The theorem goes through even when Θ is not finite, but we need the risk to be finite. Consider the following counterexample (Berger 1985, p. 254): We observe $X \sim \mathcal{N}(\theta,1)$, and want to estimate θ under the weighed quadratic loss $L(\theta,a) = e^{3\theta^2/4}(\theta-a)^2$ and prior $\theta \sim \mathcal{N}(0,1)$. The posterior is $\theta \mid X \sim \mathcal{N}(X/2,1/2)$, so that if we report the estimate a, the posterior risk is

$$\int w(\theta)(\theta-a)^2\pi(\theta \mid x)d\theta,$$

where $w(\theta) = e^{3\theta^2/4}$. The first order condition then yields the Bayes rule

$$\hat{\theta}(x) = \frac{E_{\theta|x}[\theta w(\theta)]}{E_{\theta|x}[w(\theta)]} = 2x,$$

where the second equality follows by "completing the square": for any k,

$$E_{\theta|X}[\theta^k e^{3\theta^2/4}] = \int \theta^k e^{3\theta^2/4}\frac{\sqrt{2}}{\sqrt{2\pi}}e^{-(\theta-1/2)^2}d\theta = 2e^{3x^2/4}\int \theta^k \frac{1}{\sqrt{2\pi\cdot 2}}e^{-\frac{1}{4}(\theta-2x)^2}d\theta.$$

Here the integral is the expectation of θ^k if $\theta \sim \mathcal{N}(2x,2)$. Thus, the Bayes estimator is $2X$, which has both a larger bias and a larger variance than the usual estimator X. Indeed, the risk of this estimator is $R(\theta,2X) = e^{3\theta^2/4}(4+\theta^2) > e^{3\theta^2/4} = R(\theta,X)$. The issue is that the Bayes risk $\int R(\theta,\delta(X))\pi(\theta)d\theta$ can be shown to be infinite for any δ. ⊠

The second big result is that the class of Bayes rules is complete in our setup. This provides a very strong motivation for a frequentist to use Bayesian methods: if one makes decisions based on any principle other than Bayes risk, then it must be the case that either such decision is not admissible, or else it has a Bayesian interpretation.

Theorem 3 (Complete class theorem). If δ is admissible, then it is a Bayes rule wrt some prior.

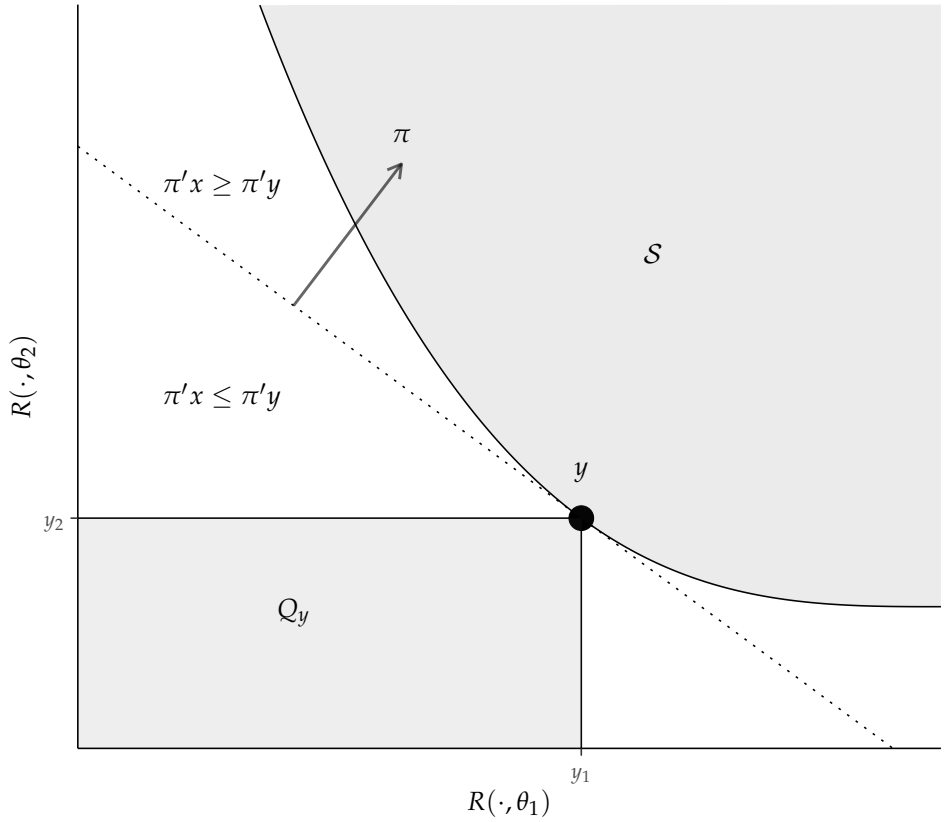

Figure 1: Sets Q_y and $\mathcal{S}$, and a separating hyperplane (dotted) for $M = 2$ in Theorem 3.

Proof. Let $y = (R(\delta, \theta_1), \ldots, R(\delta, \theta_m)) \in \mathbb{R}^m$ denote the risk of δ. Let $Q_y = \{x \in \mathbb{R}^m : x_i \leq y_i$ with at least one inequality being strict$\}$ be the set of all risk points better than y (whether they are feasible). Also, it is a convex set.

Since δ is admissible, $Q_y \cap \mathcal{S} = \emptyset$, so that the two convex sets Q_y and $\mathcal{S}$ are disjoint. By the separating hyperplane theorem, the exists a non-zero vector $\pi = (\pi_1, \ldots, \pi_m)$ such that

$$\pi' x \leq \pi' z \tag{4}$$

for all $x \in Q_y, z \in \mathcal{S}$—see Figure 1. Note that $\pi_j \geq 0$: otherwise, if some π_j were negative, we can always find a sufficiently negative x_j to violate (4). Therefore, we can normalize $\sum_{j=1}^m \pi_j = 1$, so that π is a probability distribution.

It remains to show that δ is a Bayes rule wrt π. Now, eq. (4) implies $\sup_{x \in Q_y} \pi' x \leq \pi' z$. Also, letting $x^n = (y_1 - 1/n, y_2, \ldots, y_m) \in Q_y$, we have $\sup_{x \in Q_y} \pi' x \geq \lim_{n \to \infty} \pi' x^n = \pi' y$, so that

$$\pi' y \leq \pi' z$$

for all $z \in \mathcal{S}$. In other words, for any decision rule δ' in $\mathcal{S}$, we have

$$\sum_{j=1}^m \pi_j R(\theta_j, \delta) \leq \sum_{j=1}^m \pi_j R(\delta', \theta_j),$$

so that δ is a Bayes rule wrt π. □

The complete class theorem in Haiku form,

> Complete class theorem:
> admissible rules are Bayes;
> the converse also. (Keisuke Hirano)

Digression. If the parameter and action spaces are not finite, we also need to consider limits of Bayes procedures, that is procedures $\delta(x)$ such that $\delta_{\pi_k}(x) \to \delta(x)$ for some sequence of priors π_k. See, for example Brown (1986, Theorem 4.14.). ⊠

The following result is an important application of the ideas explored by the complete class theorem that we alluded to in Lecture 3. The proof is optional.

Theorem 4 (Johnstone 2019, Remark 4.3). The maximum likelihood estimator (MLE) in a normal means model $X \sim \mathcal{N}(\theta, 1)$ is admissible under squared error loss.

Proof. Here the parameter space is not finite, so we need to show admissibility indirectly. We argue by contradiction. If X was not admissible, we could find a dominating estimator $\tilde{\theta}$ with $R(\tilde{\theta}, \theta) \leq 1$ for all θ and $R(\tilde{\theta}, \theta) < 1$ for some θ_0. Now, the risk function is continuous (in fact, it's analytic, since by Theorem 2.7.1 in Lehmann and Romano (2005) $\int (\hat{\theta}(y) - \theta)^2 f(y) \mathrm{d}y$ can be expressed in terms of Laplace transforms). Therefore, there must be an interval I containing θ_0 such that $R(\tilde{\theta}, \theta) \leq 1 - \delta$ for $\theta \in I$ and some $\delta > 0$.

Now consider a prior π_τ given by $\theta \sim \mathcal{N}(0, \tau^2)$. The Bayes estimator under this prior has average risk $\tau^2/(\tau^2+1)$ (check!), while the $\tilde{\theta}$ has average risk $\int E_\theta(\tilde{\theta} - \theta)^2 \pi_\tau(\theta)\, \mathrm{d}\theta \leq (1-\delta)\pi_\tau(\theta \in I) + \pi_\tau(\theta \in I^C) = 1 - \delta\pi_\tau(\theta \in I)$. Now, as $\tau \to \infty$ $\tau\pi_\tau(\theta \in I) = \int_I \phi(\theta/\tau) d\theta \to |I|\phi(0)$, where $|I|$ is the interval length. Thus, for large τ, average risk is smaller than $1 - \delta|I|\frac{\phi(0)}{\tau} = 1 - \frac{\delta|I|}{\sqrt{2\pi\tau}}$, plus terms that are of smaller order. For τ large enough, this is smaller than the Bayes risk, so we have a contradiction: no estimator can have risk smaller than $1 - \delta$ over any interval. Hence, X must be admissible. □

REFERENCES

Berger, James O. 1985. *Statistical Decision Theory and Bayesian Analysis.* 2nd ed. New York, NY: Springer. https://doi.org/10.1007/978-1-4757-4286-2.

Brown, Lawrence D. 1986. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory.* Vol. 9. Hayward, CA: Institute of Mathematical Statistics.

Ferguson, Thomas S. 1967. *Mathematical Statistics: A Decision Theoretic Approach.* New York, NY: Academic Press. https://doi.org/10.1016/C2013-0-07705-5.

Johnstone, Iain M. 2019. "Gaussian Estimation: Sequence and Multiresolution Models." Unpublished book draft. https://imjohnstone.su.domains//GE_09_16_19.pdf.

Lehmann, Erich Leo, and Joseph P. Romano. 2005. *Testing Statistical Hypotheses.* 3rd ed. New York, NY: Springer. https://doi.org/10.1007/0-387-27605-X.