

# STATS300A - Lecture 9

Dominik Rothenhaeusler  
Scribed by Michael Howes

10/18/21

## Contents

<b>1 Overview</b>	<b>1</b>
<b>2 Stengths of the Bayesian approach</b>	<b>1</b>
2.1 Bayesian recursion	2
2.2 Hierarchical models and empirical Bayes	2
2.3 Why Bayes estimators?	3
<b>3 Minimax estimators</b>	<b>4</b>

## 1 Overview

We have been studying optimal estimation. We have tried multiple things:

- (a) Uniform comparisons.
- (b) Restricting the class of estimators.
- (c) Collapsing the risk.
  - i. Bayesian estimators.
  - ii. Minimax estimators.

Today we will discuss:

- Some strengths of Bayesian techniques.
- Minimax risk estimation.

## 2 Stengths of the Bayesian approach

**Lemma 1.** [TPE 4.14] *Let  $Q$  be the marginal distribution of  $X$ . That is*

$$Q(A) = \int_{\Omega} \mathbb{P}_{\theta}(X \in A) d\Lambda.$$

*If the loss is strictly convex in  $d$ , then the Bayes estimator  $\delta_{\Lambda}$  is unique a.s.  $\mathbb{P}_{\theta}$  for all  $\theta \in \Omega$  if*

- (a)  $r(\Lambda, d_{\Lambda}) < \infty$ .
- (b) *If  $A \subseteq \mathcal{X}$  and  $Q(A) = 0$ , then  $\mathbb{P}_{\theta}(A) = 0$  for all  $\theta \in \Omega$ .*

See the textbook for a proof. Note that if the following all hold, then we can conclude that (b) holds above.

- (a)  $\Omega$  is an open subset of  $\mathbb{R}^k$ .
- (b) The map  $\theta \rightarrow \mathbb{P}_\theta(A)$  is continuous for all  $A$ .
- (c)  $\pi(\theta) > 0$  for all  $\theta \in \Omega$ .
- (d)  $\theta \rightarrow \pi(\theta)$  is continuous.

## 2.1 Bayesian recursion

Suppose  $\theta \sim \Lambda$  and  $X_i \stackrel{\text{iid}}{\sim} \mathbb{P}_\theta$ . We can then update the prior sequentially. Note that the posterior for  $m < n$  observations is

$$p(\theta|X_1, \dots, X_n) \propto \text{likelihood} \times \text{prior} = p(X_1, \dots, X_m|\theta)\pi(\theta).$$

The posterior for the full sample of  $n$  observations is

$$\begin{aligned} p(\theta|X_1, \dots, X_n) &\propto \text{likelihood} \times \text{prior} \\ &= p(X_1, \dots, X_n|\theta)\pi(\theta) \\ &= p(X_1, \dots, X_m|\theta)p(X_{m+1}, \dots, X_n|\theta)\pi(\theta) \\ &\propto p(X_{m+1}, \dots, X_n|\theta)p(\theta|X_1, \dots, X_m). \end{aligned}$$

Thus we can think of  $X_{m+1}, \dots, X_n$  as new data and the posterior  $p(\theta|X_1, \dots, X_m)$  as a new prior. It follows that we can compute posteriors recursively by changing the prior. This allows for speedy calculations.

**Example 1.** Suppose  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2)$  where  $\sigma^2$  is known and  $\theta \sim \mathcal{N}(\mu, b^2)$ . We saw that

$$\text{posterior} \propto \exp \left\{ -\frac{1}{2}\theta^2 w_n + \theta \bar{w}_n \right\},$$

where

$$\begin{aligned} w_n &= \frac{n}{\sigma^2} + \frac{1}{b^2}, \\ \bar{w}_n &= \frac{\sum_{i=1}^n x_i}{\sigma^2} + \frac{\mu}{b^2}. \end{aligned}$$

The weights  $w_n$  satisfy the recursion

$$\begin{aligned} w_n &= \frac{1}{\sigma^2} + w_{n-1}, \\ \bar{w}_n &= \frac{x_n}{\sigma^2} + \bar{w}_{n-1}. \end{aligned}$$

Thus our updates are quick linear calculations. This has many applications.

## 2.2 Hierarchical models and empirical Bayes

We can use Bayesian ideas to model problems with repeat structure and pool information across observations. Suppose we have  $\theta_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \tau^2)$  for  $i = 1, \dots, p$  where  $\tau^2 > 0$  is known (for now).

Suppose that we also have  $X_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\theta_i, 1)$ . For example  $\theta_i$  might be the effect of an experiment and  $X_i$  is the measured effect and our measurement errors are i.i.d.  $\mathcal{N}(0, 1)$ . Consider the loss

$$L(\theta, \delta) = \sum_{i=1}^p (\theta_i - \delta_i(X))^2.$$

One can show (see Keener section 11) that the Bayes estimator is given by

$$\delta_i(X) = \left(1 - \frac{1}{1 + \tau^2}\right) X_i = \frac{\tau^2}{1 + \tau^2} X_i.$$

That is we shrink our data towards 0. The amount of shrinkage depends on  $\tau^2$ . If  $\tau^2$  is small we shrink a lot, if it is large we shrink less. We will not prove that  $\delta$  is the Bayes estimator but we will show that it has the lowest average risk of estimators of the form  $\hat{\theta}_\alpha = \alpha X$  when  $p = 1$ . The risk of such an estimator is

$$\mathbb{E}[(\theta - \alpha X)^2] = \mathbb{E}[\theta^2] - 2\alpha\mathbb{E}[\theta X] + \alpha^2\mathbb{E}[X^2] = \tau^2 - 2\alpha\tau^2 + \alpha^2(\tau^2 + 1),$$

which is minimized when  $\alpha(\tau^2 + 1) = \tau^2$ , that is  $\alpha = \frac{\tau^2}{1 + \tau^2}$ . The shrinkage term is the ratio of the experiment noise and the measurement noise.

What if we don't know  $\tau$ ? We can still estimate it. This is the idea behind empirical Bayes. We can estimate the variance of  $X_i$  since

$$X_i \stackrel{\text{iid}}{\sim} N(0, 1 + \tau^2),$$

Thus  $\frac{1}{p} \sum_{j=1}^p X_j^2$  is unbiased for  $1 + \tau^2$ . This gives us the new estimator

$$\delta'_i(X) = \left(1 - \frac{p}{\sum_{j=1}^n X_j^2}\right) X_i.$$

This is an example of empirical Bayes when the prior is estimated from the data. We will revisit this estimator in the frequentist setting later in the course.

### 2.3 Why Bayes estimators?

- (a) Every admissible estimator is a Bayes estimator or a limit of Bayes estimators. That is for a sequence of priors  $\Lambda_n$ ,  $\delta_{\Lambda_n}(x) \rightarrow \delta(x)$  a.e.  $\mathbb{P}_\theta$ .
- (b) Can incorporate prior experience and beliefs.
- (c) Quantification of uncertainty are sometimes easier to interpret.
- (d) Encode complex data structures.

How do we choose priors?

- (a) Subjective: previous knowledge.
- (b) Objective: Select an “uninformative prior” such as a Jeffery’s prior or a uniform prior.
- (c) Empirical: Estimate prior parameters from data.
- (d) Computational: Use conjugate priors.

As  $n \rightarrow \infty$ , the posterior is independent of the prior - Bernstein-von-Mises.

### 3 Minimax estimators

Given  $X \sim \mathbb{P}_\theta$ ,  $\theta$  fixed and unknown. Our goal is to find an estimator  $\delta$  that minimizes

$$\sup_{\theta \in \Omega} R(\theta, \delta).$$

Such a  $\delta$  is called a minimax estimator. How do we find such estimators? Note that

$$\sup_{\theta} R(\theta, \delta) \geq \int_{\Omega} R(\theta, \delta) d\Lambda,$$

for all probability distributions  $\Lambda$ . Thus our goal is to find the “worst prior.”

**Definition 1.** A prior  $\Lambda$  is called *least favourable* if  $r_{\Lambda} \geq r_{\Lambda'}$  for any other prior  $\Lambda'$ .

Recall that  $r_{\Lambda} = r(\Lambda, \delta_{\Lambda})$  where  $\delta_{\Lambda}$  is the Bayes estimator for  $\Lambda$ .

**Theorem 1.** [TPE 5.1.4] Suppose  $\delta_{\Lambda}$  is a Bayes estimator with  $r_{\Lambda} = \sup_{\theta} R(\theta, \delta_{\Lambda})$ , then

- (a)  $\delta_{\Lambda}$  is minimax.
- (b)  $\Lambda$  is least favourable.
- (c) If  $\delta_{\Lambda}$  is a unique Bayes estimator (a.e  $\mathbb{P}_{\theta}$  for all  $\theta \in \Omega$ ), then  $\delta_{\Lambda}$  is the unique minimax estimator.

*Proof.* For (a) let  $\delta$  be another estimator, then

$$\begin{aligned} \sup_{\theta} R(\theta, \delta) &\geq \int_{\Omega} R(\theta, \delta) d\Lambda \\ &\geq \int_{\Omega} R(\theta, \delta_{\Lambda}) d\Lambda \\ &= r_{\Lambda} \\ &= \sup_{\theta} R(\theta, \delta_{\Lambda}) \end{aligned}$$

Thus  $\delta_{\Lambda}$  is minimax. For (b), let  $\Lambda'$  be another prior, then

$$\begin{aligned} r_{\Lambda'} &= \int_{\Omega} R(\theta, \delta_{\Lambda'}) d\Lambda' \\ &\leq \int_{\Omega} R(\theta, \delta_{\Lambda}) d\Lambda' \\ &\leq \sup_{\theta \in \Omega} R(\theta, \delta_{\Lambda}) \\ &= \int_{\Omega} R(\theta, \delta_{\Lambda}) d\Lambda \\ &= r_{\Lambda}. \end{aligned}$$

Thus  $\Lambda$  is least favourable. For (c), suppose that  $\delta$  is a minimax estimator, then

$$\begin{aligned} r(\Lambda, \delta) &= \int_{\Omega} R(\theta, \delta) d\Lambda \\ &\leq \sup_{\theta \in \mathcal{L}} R(\theta, \delta) \\ &= \sup_{\theta \in \mathcal{L}} R(\theta, \delta_{\Lambda}) \\ &= r_{\Lambda}. \end{aligned}$$

Thus  $\delta$  is the Bayes estimator of the prior  $\Lambda$ . It follows that if  $\delta_{\Lambda}$  is the unique Bayes estimator for  $\Lambda$ , then  $\delta_{\Lambda}$  is the unique minimax estimator.  $\square$

**Example 2.** Suppose  $X \sim \text{Bin}(n, \theta)$  and we want to find the minimax estimator of  $\theta$  under squared error loss. Our goal is to find a Bayes estimator that has constant risk. This would automatically give

$$r_\Lambda = \sup_{\theta \in \Omega} R(\theta, \delta_\Lambda).$$

We previously say that if  $\Theta$  has a  $\text{Beta}(a, b)$  prior, then the Bayes estimator has the form

$$\delta_\Lambda(x) = \frac{x + a}{a + b + n}.$$

We wish to find  $a, b$  such that this estimator has constant risk. Note that

$$\begin{aligned} R(\theta, \delta_\Lambda) &= \mathbb{E}_\theta \left[ \left( \frac{x + a}{n + a + b} - \theta \right)^2 \right] \\ &= \frac{1}{(n + a + b)^2} \mathbb{E}_\theta \left[ (X + a - \theta(n + a + b))^2 \right] \\ &= \frac{1}{(n + a + b)^2} \mathbb{E}_\theta \left[ (X - n\theta + a(1 - \theta) - b\theta)^2 \right] \\ &= \frac{1}{(n + a + b)^2} [\text{Var}_\theta(X - n\theta) + (a(1 - \theta) - b\theta)^2] \\ &= \frac{1}{(n + a + b)^2} [n\theta(1 - \theta) + (a(1 - \theta) - b\theta)^2]. \end{aligned}$$

Thus we wish to find  $a, b$  such that  $n\theta(1 - \theta) + (a(1 - \theta) - b\theta)^2$  is constant in  $\theta$ . The solution is  $a = b = \frac{\sqrt{n}}{2}$ . Thus

$$\delta_\Lambda(x) = \frac{x + \frac{\sqrt{n}}{2}}{n + \sqrt{n}},$$

is a minimax estimator for  $\theta$ .