

# STAT 830

## Bayesian Estimation

Richard Lockhart

Simon Fraser University

STAT 830 — Fall 2011



# Purposes of These Notes

- Discuss Bayesian Estimation
- Motivate posterior mean via Bayes quadratic risk.
- Discuss prior to posterior.
- Define admissibility, minimax estimates



- Focus on problem of estimation of 1 dimensional parameter.
- Mean Squared Error corresponds to using

$$L(d, \theta) = (d - \theta)^2.$$

- Risk function of procedure (estimator)  $\hat{\theta}$  is

$$R_{\hat{\theta}}(\theta) = E_{\theta}[(\hat{\theta} - \theta)^2]$$

- Now consider prior with density  $\pi(\theta)$ .
- Bayes risk of  $\hat{\theta}$  is

$$\begin{aligned} r_{\pi} &= \int R_{\hat{\theta}}(\theta)\pi(\theta)d\theta \\ &= \int \int (\hat{\theta}(x) - \theta)^2 f(x; \theta)\pi(\theta)dx d\theta \end{aligned}$$



## Posterior mean

- Choose  $\hat{\theta}$  to minimize  $r_{\pi}$ ?
- Recognize that  $f(x; \theta)\pi(\theta)$  is really a joint density

$$\int \int f(x; \theta)\pi(\theta) dx d\theta = 1$$

- For this joint density: conditional density of  $X$  given  $\theta$  is just the model  $f(x; \theta)$ .
- Justifies notation  $f(x|\theta)$ .
- Compute  $r_{\pi}$  differently by factoring joint density a different way:

$$f(x|\theta)\pi(\theta) = \pi(\theta|x)f(x)$$

where now  $f(x)$  is the marginal density of  $x$  and  $\pi(\theta|x)$  denotes the conditional density of  $\theta$  given  $X$ .

- Call  $\pi(\theta|x)$  the **posterior density**.
- Found via Bayes theorem (which is why this is Bayesian statistics):

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\phi)\pi(\phi) d\phi}$$



## The posterior mean

- With this notation we can write

$$r_{\pi}(\hat{\theta}) = \int \left[ \int (\hat{\theta}(x) - \theta)^2 \pi(\theta|x) d\theta \right] f(x) dx$$

- Can choose  $\hat{\theta}(x)$  separately for each  $x$  to minimize the quantity in square brackets (as in the NP lemma).
- Quantity in square brackets is quadratic function of  $\hat{\theta}(x)$ ; minimized by

$$\hat{\theta}(x) = \int \theta \pi(\theta|x) d\theta$$

which is

$$E(\theta|X)$$

and is called the **posterior mean** of  $\theta$ .



## Example

- **Example:** estimating normal mean  $\mu$ .
- Imagine, for example that  $\mu$  is the true speed of sound.
- I think this is around 330 metres per second and am pretty sure that I am within 30 metres per second of the truth with that guess.
- I might summarize my opinion by saying that I think  $\mu$  has a normal distribution with mean  $\nu = 330$  and standard deviation  $\tau = 10$ .
- That is, I take a prior density  $\pi$  for  $\mu$  to be  $N(\nu, \tau^2)$ .
- Before I make any measurements best guess of  $\mu$  minimizes

$$\int (\hat{\mu} - \mu)^2 \frac{1}{\tau\sqrt{2\pi}} \exp\{-(\mu - \nu)^2 / (2\tau^2)\} d\mu$$

- This quantity is minimized by the prior mean of  $\mu$ , namely,

$$\hat{\mu} = E_{\pi}(\mu) = \int \mu\pi(\mu)d\mu = \nu.$$



## From prior to posterior

- Now collect 25 measurements of the speed of sound.
- Assume: relationship between the measurements and  $\mu$  is that the measurements are unbiased and that the standard deviation of the measurement errors is  $\sigma = 15$  which I assume that we know.
- So model is: given  $\mu$ ,  $X_1, \dots, X_n$  iid  $N(\mu, \sigma^2)$ .
- The joint density of the data and  $\mu$  is then

$$\frac{\exp\{-\sum (X_i - \mu)^2 / (2\sigma^2)\}}{(2\pi)^{n/2} \sigma^n} \times \frac{\exp\{-(\mu - \nu)^2 / \tau^2\}}{(2\pi)^{1/2} \tau}.$$

- Thus  $(X_1, \dots, X_n, \mu) \sim MVN$ .
- Conditional distribution of  $\theta$  given  $X_1, \dots, X_n$  is normal.
- Use standard MVN formulas to get conditional means and variances



## Posterior Density

- Alternatively: exponent in joint density has form

$$-\frac{1}{2} [\mu^2/\gamma^2 - 2\mu\psi/\gamma^2]$$

plus terms not involving  $\mu$  where

$$\frac{1}{\gamma^2} = \left( \frac{n}{\sigma^2} + \frac{1}{\tau^2} \right) \text{ and } \frac{\psi}{\gamma^2} = \frac{\sum X_i}{\sigma^2} + \frac{\nu}{\tau^2}.$$

- So: conditional of  $\mu$  given data is  $N(\psi, \gamma^2)$ .
- In other words the posterior mean of  $\mu$  is

$$\frac{\frac{n}{\sigma^2} \bar{X} + \frac{1}{\tau^2} \nu}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

which is weighted average of prior mean  $\nu$  and sample mean  $\bar{X}$ .

- Notice: weight on data is large when  $n$  is large or  $\sigma$  is small (precise measurements) and small when  $\tau$  is small (precise prior opinion).





## Improper priors

- When the density does not integrate to 1 we can still follow the machinery of Bayes' formula to derive a posterior.
- **Example:**  $N(\mu, \sigma^2)$ ; consider prior density

$$\pi(\mu) \equiv 1.$$

- This “density” integrates to  $\infty$ ; using Bayes' theorem to compute the posterior would give

$$\pi(\mu|X) = \frac{(2\pi)^{-n/2} \sigma^{-n} \exp\{-\sum (X_i - \mu)^2 / (2\sigma^2)\}}{\int (2\pi)^{-n/2} \sigma^{-n} \exp\{-\sum (X_i - \nu)^2 / (2\sigma^2)\} d\nu}$$

- It is easy to see that this cancels to the limit of the case previously done when  $\tau \rightarrow \infty$  giving a  $N(\bar{X}, \sigma^2/n)$  density.
- I.e., Bayes estimate of  $\mu$  for this improper prior is  $\bar{X}$ .



# Admissibility

- Bayes procedures corresponding to proper priors are admissible.
- It follows that for each  $w \in (0, 1)$  and each real  $\nu$  the estimate

$$w\bar{X} + (1 - w)\nu$$

is admissible.

- That this is also true for  $w = 1$ , that is, that  $\bar{X}$  is admissible is much harder to prove.
- **Minimax estimation:** The risk function of  $\bar{X}$  is simply  $\sigma^2/n$ .
- That is, the risk function is constant since it does not depend on  $\mu$ .
- Were  $\bar{X}$  Bayes for a proper prior this would prove that  $\bar{X}$  is minimax.
- In fact this is also true but hard to prove.



## Binomial( $n, p$ ) example

- Given  $p$ ,  $X$  has a Binomial( $n, p$ ) distribution.
- Give  $p$  a Beta( $\alpha, \beta$ ) prior density

$$\pi(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1}$$

- The joint “density” of  $X$  and  $p$  is

$$\binom{n}{X} p^X (1-p)^{n-X} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1};$$

posterior density of  $p$  given  $X$  is of the form

$$c p^{X+\alpha-1} (1-p)^{n-X+\beta-1}$$

for a suitable normalizing constant  $c$ .

- This is Beta( $X + \alpha, n - X + \beta$ ) density.



## Example continued

- Mean of Beta( $\alpha, \beta$ ) distribution is  $\alpha/(\alpha + \beta)$ .
- So Bayes estimate of  $p$  is

$$\frac{X + \alpha}{n + \alpha + \beta} = w\hat{p} + (1 - w)\frac{\alpha}{\alpha + \beta}$$

where  $\hat{p} = X/n$  is the usual mle.

- Notice: again weighted average of prior mean and mle.
- Notice: prior is proper for  $\alpha > 0$  and  $\beta > 0$ .
- To get  $w = 1$  take  $\alpha = \beta = 0$ ; use improper prior

$$\frac{1}{p(1 - p)}$$

- Again: each  $w\hat{p} + (1 - w)p_0$  is admissible for  $w \in (0, 1)$ .
- Again: it is true that  $\hat{p}$  is admissible but our theorem is not adequate to prove this fact.



## Minimax estimate

- The risk function of  $w\hat{p} + (1 - w)p_0$  is

$$R(p) = E[(w\hat{p} + (1 - w)p_0 - p)^2]$$

which is

$$\begin{aligned} w^2 \text{Var}(\hat{p}) + (wp + (1 - w)p - p)^2 \\ = \\ w^2 p(1 - p)/n + (1 - w)^2 (p - p_0)^2 \end{aligned}$$

- Risk function constant if coefficients of  $p^2$  and  $p$  in risk are 0.
- Coefficient of  $p^2$  is

$$-w^2/n + (1 - w)^2$$

so  $w = n^{1/2}/(1 + n^{1/2})$ .

- Coefficient of  $p$  is then

$$w^2/n - 2p_0(1 - w)^2$$

which vanishes if  $2p_0 = 1$  or  $p_0 = 1/2$ .



## Minimax continued

- Working backwards: to get these values for  $w$  and  $p_0$  require  $\alpha = \beta$ .
- Moreover

$$w^2/(1-w)^2 = n$$

gives

$$n/(\alpha + \beta) = \sqrt{n}$$

or  $\alpha = \beta = \sqrt{n}/2$ .

- Minimax estimate of  $p$  is

$$\frac{\sqrt{n}}{1 + \sqrt{n}} \hat{p} + \frac{1}{1 + \sqrt{n}} \frac{1}{2}$$

- Example:**  $X_1, \dots, X_n$  iid  $MVN(\mu, \Sigma)$  with  $\Sigma$  known.
- Take improper prior for  $\mu$  which is constant.
- Posterior of  $\mu$  given  $X$  is then  $MVN(\bar{X}, \Sigma/n)$ .



- Finite population called  $\theta_1, \dots, \theta_B$  in text.
- Take simple random sample (with replacement) of size  $n$  from population.
- Call  $X_1, \dots, X_n$  the indexes of the sampled data. Each  $X_i$  is an integer from 1 to  $B$ .
- Estimate

$$\psi = \frac{1}{B} \sum_{i=1}^B \theta_i.$$

- In STAT 410 would suggest Horvitz-Thompson estimator

$$\hat{\psi} = \frac{1}{n} \sum_{i=1}^n \theta_{X_i}$$

- This is the sample mean of the observed values of  $\theta$ .



## Mean and variance of our estimate

- We have

$$E(\hat{\psi}) = \frac{1}{n} \sum_{i=1}^n E(\theta_{X_i}) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^B \theta_j P(X_i = j) = \psi.$$

- And we can compute the variance too:

$$\text{Var}(\hat{\psi}) = \frac{1}{n} \frac{1}{B} \sum_{j=1}^B (\theta_j - \psi)^2.$$

- This is the population variance of the  $\theta$ s divided by  $n$  so it gets small as  $n$  grows.





# Bayesian Analysis

- Text motivates the parameter  $\psi$  in terms of non-response mechanism.
- Analyzed later.
- Put down prior density  $\pi(\theta_1, \dots, \theta_B)$ .
- Define  $W_i = \theta_{X_i}$ . Data is  $(X_1, W_1), \dots, (X_n, W_n)$ .
- Likelihood is

$$P_{\theta}(X_1 = i_1, \dots, X_n = i_n, W_1 = w_1, \dots, W_n = w_n)$$

- Posterior is just conditional of  $\psi$  given

$$(\theta_{i_1} = w_1, \dots, \theta_{i_n} = w_n).$$

- This is

$$\frac{\pi(\theta_1, \dots, \theta_n)}{\pi_{i_1, \dots, i_n}(\theta_{i_1}, \dots, \theta_{i_n})}$$

- Denominator is supposed to be marginal density of the observed  $\theta$ s



## Where's the problem?

- The text takes it for granted that the conditional law of unobserved  $\theta$ s is not changed.
- Suppose that our prior says  $\theta_1, \dots, \theta_n$  are independent.
- Let's say iid with prior density  $p(\theta_j)$  – same  $p$  for each  $i$ .
- Then the posterior density of all the  $\theta_j$  *other* than  $\theta_{i_1}, \dots, \theta_{i_n}$  is

$$\prod_{j \notin \{i_1, \dots, i_n\}} p(\theta_j).$$

- This is the same as the prior!
- So except for learning the  $n$  particular  $\theta$ s in the sample you learned nothing.
- So Bayes is a flop, right?
- Wrong: the message is **the prior matters**.



# Realistic Priors

- If your prior says you are *a priori* sure of something stupid, your posterior will be stupid too.
- In this case: if I tell you the sampled  $\theta_i$  you *do* learn about the  $\theta_i$ .
- Try the following prior:
  - ▶ There is a quantity  $\mu$ . Given  $\mu$  the  $\theta_i$  are iid  $N(\mu, 1)$ .
  - ▶ The quantity  $\mu$  has a  $N(0, 1)$  prior.
- So  $\theta_1, \dots, \theta_B$  has a multivariate normal distribution with mean vector 0 and variance matrix

$$\Sigma_B = \mathbf{I}_{B \times B} + \mathbf{1}_B \mathbf{1}_B^t$$

- This is a *hierarchical* prior – specified in two layers.



## Posterior for hierarchical prior

- Notationally simpler if we imagine our sample happened to be the first  $n$  elements.
- So we observe  $\theta_1, \dots, \theta_n$ .
- Posterior is just conditional density of  $\theta_{n+1}, \dots, \theta_B$  given  $\theta_1, \dots, \theta_n$ .
- The density of  $\theta_1, \dots, \theta_n$  is multivariate normal with mean vector 0 and variance covariance matrix

$$\Sigma_n \mathbf{1}_{n \times n} + \mathbf{1}_n \mathbf{1}_n^t$$

- so get posterior by dividing two multivariate normal densities.



## Posterior for a reasonable prior

- To get specific formulas need matrix inverses and determinants.
- We can check:

$$\det \Sigma = B$$

$$\det \Sigma_n = n$$

$$\Sigma^{-1} = \mathbf{I}_{B \times B} - \mathbf{1}_B \mathbf{1}_B^t / (B + 1)$$

$$\Sigma_n^{-1} = \mathbf{I}_{n \times n} - \mathbf{1}_n \mathbf{1}_n^t / (n + 1)$$

- Get posterior density of unobserved  $\theta$ s from joint over marginal.

$$\frac{(2\pi)^{-B/2} B^{-1/2} \exp\left(-\left[\sum_1^B \theta_i^2 - (\sum_1^B \theta_i)^2 / (B + 1)\right] / 2\right)}{(2\pi)^{-n/2} n^{-1/2} \exp\left(-\left[\sum_1^n \theta_i^2 - (\sum_1^n \theta_i)^2 / (n + 1)\right] / 2\right)}$$

- Can simplify but I just want Bayes estimate

$$E(\psi | \theta_1, \dots, \theta_n) = B^{-1} \left( \sum_1^n \theta_i + \sum_{n+1}^B E(\theta_j | \theta_1, \dots, \theta_n) \right).$$



## Better prior details

- Calculate the individual conditional expectations using MVN conditionals.
- Find, denoting  $\bar{\theta} = \sum_1^n \theta_i/n$ ,

$$E(\theta_j | \theta_1, \dots, \theta_n) = n\bar{\theta}/(n+1)$$

- This gives the Bayes estimate

$$\bar{\theta}(1 - 1/(n+1))(1 + 1/B).$$

- Compare this to Horvitz-Thompson estimator  $\bar{\theta}$ .
- Not much different!
- The formula for the Bayes estimate is right regardless of sample drawn.



## The example in the text

- In the text you don't observe  $\theta_{X_i}$  but a variable  $R_{X_i}$  which is Bernoulli with success probability  $\xi_{X_i}$ , given  $X_i$ .
- Then if  $R_i = 1$  you observe  $Y_i$  which is Bernoulli with success probability  $\theta_{X_i}$ , again conditional on  $X_i$ .
- This leads to a more complicated Horvitz-Thompson estimator and means you don't directly observe the  $\theta_i$ .
- But the hierarchical prior means you believe that learning about some  $\theta$ s tells you about others.
- The hierarchical prior says the  $\theta$ s are correlated!
- In the example in the text the priors appear to be independence priors.
- So you can't learn about one  $\theta$  from another.
- In my independence prior as  $B \rightarrow \infty$  the prior variance of  $\psi$  goes to 0!
- So you are saying you *know*  $\psi$  if you specify an independence prior.

