Mathematical Statistics II

STA2212H S LEC9101

Week 9

March 17 2021

Start recording!

link

Recap

Are you recording?

- overview of Bayesian inference
- posterior predictive distribution ←
- Bayesian computation
- Example 11.9 AoS x10.8
- Bayesian hierarchical models

$$p(x_{n+1} \mid \underset{\sim}{x}^n) = \int f(x_{n+1} \mid \theta)\, \pi(\theta \mid \underset{\sim}{x}^n)\, d\theta$$

↑ assuming $X_{n+1} \perp\!\!\!\perp \underset{\sim}{X}_n$

$$p(x_{n+1}) = \int f(x_{n+1} \mid \theta)\, \pi(\theta)\, d\theta$$

↑ sometimes used for planning studies

Recap

Are you recording?

- overview of Bayesian inference
- posterior predictive distribution
- Bayesian computation
- Example 11.9 AoS
- Bayesian hierarchical models

Addendum: If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, both parameters unknown, then the conjugate priors for μ and σ^2 are

$$\mu \sim N(\mu_0, \tau^2); \qquad \sigma^2 \sim IG(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$$

Inverse Gamma distribution $IG(\alpha, \beta)$ has density $f(t; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{x^{\alpha+1}} e^{-\beta/x}$

$\propto \quad x^{\alpha-1} e^{-\beta x}$

Today

Start Recording

1. Notes from March 12 (Friday) still to be posted
2. Friday March 19 – Stein's paradox
3. empirical Bayes
4. introduction to decision theory

B-H proof

- Mar 22 3.00 – 4.00 pm EDT
  Data Science ARES
  Jesse Cisewski Kehe
  University of Wisconsin-Madison
  "Astrostatistics: From Exoplanets to
  the Large-scale Structure of the Universe"

Bayesian hierarchical model, recap

SM Ex. 11.25

- $x_i \mid \theta_i \sim N(\theta_i, v_i)$, independent; $\theta_i \mid \mu \sim N(\mu, \sigma^2), i = 1, \ldots, n$ i.i.d; $\mu \sim N(\mu_0, \tau^2)$

$\hat{\theta}_i = x_i$

hyper-par

hyper-prior

$v_i, \sigma^2, \mu_0, \tau^2$ all known

$$E(\theta_i \mid \underset{\sim}{x}^n) = x_i\left(\frac{\sigma^2}{\sigma^2+v_i}\right) + \mu\left(\frac{1}{\sigma^2+v_i}\right)$$

est. $\hat{\theta}_i$ ✓ $= x_i\left(\frac{\sigma^2}{\sigma^2+v_i}\right) + \tilde{\mu}\left(\frac{1}{\sigma^2+v_i}\right)$

$$E(\mu \mid \underset{\sim}{x}^n) = \frac{\sum x_i/(v_i+\sigma^2)}{\sum 1/(v_i+\sigma^2)} = \tilde{\mu}$$

shrinkage estimators

If σ^2 unknown, no explicit sol'n for $\hat{\theta}_i$ & $\tilde{\mu}$: need use MCMC

Bayesian hierarchical model, recap

SM Ex. 11.25

- $x_i \mid \theta_i \sim N(\theta_i, v_i)$, independent; $\theta_i \mid \mu \sim N(\mu, \sigma^2), i = 1, \ldots, n$ i.i.d; $\mu \sim N(\mu_0, \tau^2)$
- $v_i, i = 1, \ldots, n, \sigma^2, \mu_0, \tau^2$ all known

Bayesian hierarchical model, recap

SM Ex. 11.25

- $x_i \mid \theta_i \sim N(\theta_i, v_i)$, independent; $\theta_i \mid \mu \sim N(\mu, \sigma^2), i = 1, \ldots, n$ i.i.d; $\mu \sim N(\mu_0, \tau^2)$
- $v_i, i = 1, \ldots, n, \sigma^2, \mu_0, \tau^2$ all known
-

$$\mathrm{E}(\theta_i \mid \boldsymbol{x}) = x_i \frac{\sigma^2}{\sigma^2 + v_i} + \mathrm{E}(\mu \mid \boldsymbol{x})(1 - \frac{\sigma^2}{\sigma^2 + v_i})$$

Bayesian hierarchical model, recap

SM Ex. 11.25

- $x_i \mid \theta_i \sim N(\theta_i, v_i)$, independent; $\theta_i \mid \mu \sim N(\mu, \sigma^2), i = 1, \ldots, n$ i.i.d; $\mu \sim N(\mu_0, \tau^2)$
- $v_i, i = 1, \ldots, n, \sigma^2, \mu_0, \tau^2$ all known

- $$\mathrm{E}(\theta_i \mid \boldsymbol{x}) = x_i \frac{\sigma^2}{\sigma^2 + v_i} + \mathrm{E}(\mu \mid \boldsymbol{x})(1 - \frac{\sigma^2}{\sigma^2 + v_i})$$
- $$\mathrm{E}(\mu \mid \boldsymbol{x}) = \frac{\mu_0/\tau^2 + \sum x_i/(\sigma^2 + v_i)}{1/\tau^2 + \sum 1/(\sigma^2 + v_i)}$$

Bayesian hierarchical model, recap

SM Ex. 11.25

- $x_i \mid \theta_i \sim N(\theta_i, v_i)$, independent; $\theta_i \mid \mu \sim N(\mu, \sigma^2), i = 1, \ldots, n$ i.i.d; $\mu \sim N(\mu_0, \tau^2)$
- $v_i, i = 1, \ldots, n, \sigma^2, \mu_0, \tau^2$ all known
-
$$\mathrm{E}(\theta_i \mid \boldsymbol{x}) = x_i \frac{\sigma^2}{\sigma^2 + v_i} + \mathrm{E}(\mu \mid \boldsymbol{x})(1 - \frac{\sigma^2}{\sigma^2 + v_i})$$
-
$$\mathrm{E}(\mu \mid \boldsymbol{x}) = \frac{\mu_0/\tau^2 + \sum x_i/(\sigma^2 + v_i)}{1/\tau^2 + \sum 1/(\sigma^2 + v_i)}$$
- If σ^2 unknown, then need to sample from the posterior, no closed form available

Bayesian hierarchical model, recap

SM Ex. 11.25

- $x_i \mid \theta_i \sim N(\theta_i, v_i)$, independent; $\theta_i \mid \mu \sim N(\mu, \sigma^2), i = 1, \ldots, n$ i.i.d; $\mu \sim N(\mu_0, \tau^2)$
- $v_i, i = 1, \ldots, n, \sigma^2, \mu_0, \tau^2$ all known
-
$$\mathrm{E}(\theta_i \mid \boldsymbol{x}) = x_i \frac{\sigma^2}{\sigma^2 + v_i} + \mathrm{E}(\mu \mid \boldsymbol{x})\left(1 - \frac{\sigma^2}{\sigma^2 + v_i}\right)$$
-
$$\mathrm{E}(\mu \mid \boldsymbol{x}) = \frac{\mu_0/\tau^2 + \sum x_i/(\sigma^2 + v_i)}{1/\tau^2 + \sum 1/(\sigma^2 + v_i)}$$
- If σ^2 unknown, then need to sample from the posterior, no closed form available
- Figure 11.11 applies similar ideas, plus sampling from the posterior, in logistic regression

...Bayesian hierarchical model, recap

SM Ex. 11.25

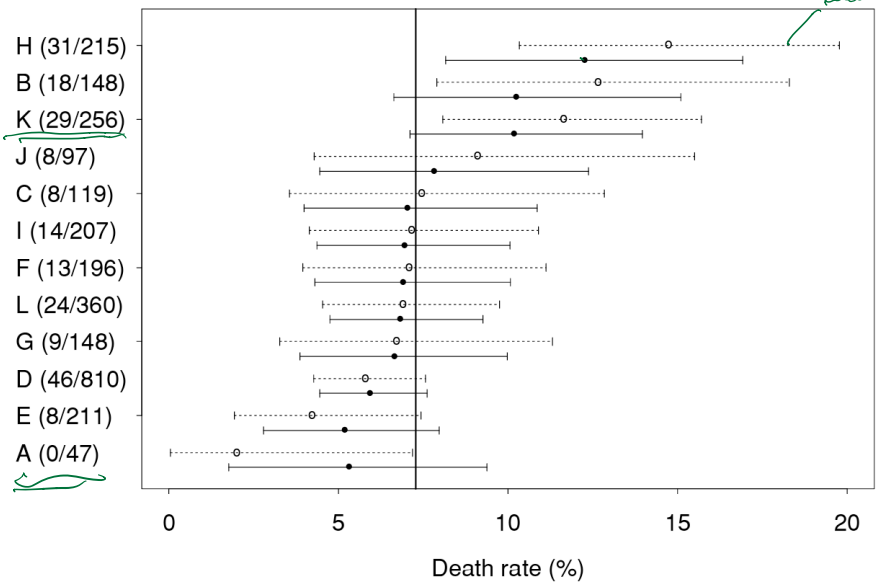

Figure 11.11 Posterior summaries for mortality rates for cardiac surgery data. Posterior means and 0.95 equitailed credible intervals for separate analyses for each hospital are shown by hollow circles and dotted lines, while blobs and solid lines show the corresponding quantities for a hierarchical model. Note the shrinkage of the estimates for the hierarchical model towards the overall posterior mean rate, shown as the solid vertical line; the hierarchical intervals are slightly shorter than those for the simpler model.

As above, $x_i \mid \theta_i \sim N(\theta_i, v_i)$, independent; $\theta_i \mid \mu \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$ i.i.d;

$$\theta_i \mid x_i \sim N\left(x_i \frac{\sigma^2}{\sigma^2 + v_i} + \mu \frac{v_i}{\sigma^2 + v_i}, \frac{\sigma^2 v_i}{\sigma^2 + v_i} \right)$$

$$\pi(\theta_i \mid x_i) = f(x_i \mid \theta_i)\, \pi(\theta_i) \Big/ \underbrace{\int f(x_i \mid \theta_i)\pi(\theta_i)\, d\theta_i}_{\text{marginal}} = \frac{m(x_i)}{f(x_i)}$$

$$\prod_{i=1}^{n} f_{x_i}(x_i \mid \mu) = L(\mu; \underset{\sim}{x}^n)$$

(ntbc ind't)

$$\hat{\mu} = \hat{\mu}(\underset{\sim}{x}^n) = \frac{\sum x_i / (v_i + \sigma^2)}{\sum 1/(v_i + \sigma^2)}$$

SM "x_i marginally ind't, $N[\mu, v_j/(\sigma^2 + v_j)]$"

$$\tilde{\theta}_i = \left(x_i \frac{\sigma^2}{\sigma^2+v_i} + \hat{\mu}\left(\frac{v_i}{\sigma^2+v_i}\right)\right)$$

$$= \hat{\mu} + (1-\xi_i)(x_i - \hat{\mu})$$

$$\xi_i = \frac{v_i}{\sigma^2+v_i} \quad \text{weight} \qquad \begin{matrix}\sigma^2 \to \infty \\ v_i \to \infty\end{matrix}$$

Using marg'l ~~Statistics~~ density for $\underline{x}$ to estimate parameters in the prior is called empirical Bayes

"Large sets of parallel situations carry their own prior information"

$$E\tilde{\theta}_i \neq \theta_i = \theta_i \frac{\sigma^2}{\sigma^2+v_i} + \mu \frac{v_i}{\sigma^2+v_i}$$

bias $\neq 0$ perhaps $\text{mse}(\tilde{\theta}_i)$

$$= E(\tilde{\theta}_i - \theta_i)^2 < \text{Var}(x_i)$$

$x_i|\theta_i \sim N(\theta_i, v_i)$ ☆

Alternative to ridge regression

β_j = change in y per unit change in z_{ij}

$$\min_\beta \left\{ \sum_{i=1}^n (x_i - z_i^T\beta)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}$$

$\sum z_i = 0$
$\sum z_i^2 = 1$ ☆
$\leftarrow \frac{z_i - \bar{z}}{s_z}$

$$\hat{\beta}_{ridge} = (\underline{Z}^T\underline{Z} + \lambda \underline{I})^{-1}\underline{Z}^T\underline{y} \quad \leftarrow \text{shrinkage est.'s}$$

Table 6.1 *Counts y_x of number of claims x made in a single year by 9461 automobile insurance policy holders. Robbins' formula (6.7) estimates the number of claims expected in a succeeding year, for instance 0.168 for a customer in the $x = 0$ category. Parametric maximum likelihood analysis based on a gamma prior gives less noisy estimates.*

Claims x	0	1	2	3	4	5	6	7
Counts y_x	7840	1317	239	42	14	4	4	1
Formula (6.7)	.168	.363	.527	1.33	1.43	6.00	1.75	
Gamma MLE	.164	.398	.633	.87	1.10	1.34	1.57	

$$f_k(x;\theta_k) = \frac{\theta_k^x e^{-\theta_k}}{x!} \qquad x = 0, 1, 2, \dots$$

k - customer k

P_n " " had x claims in the year

$$\theta_k^0 e^{-\theta_k} / 0! = e^{-\theta_k}$$

$$\theta_k \sim \Gamma(\nu, 1/\sigma)$$

density $\quad \theta_k^{\nu-1} e^{-\theta_k/\sigma} / \Gamma(\nu)\sigma^\nu$

marginal density $\quad f(\underset{\sim}{x}) = \prod_{k=1}^{k} \int f(\underset{\sim}{x}|\theta_k)\pi(\theta_k)d\theta_k$

$$= f(\underset{\sim}{x} \mid \nu, \sigma)$$

now use ML to estimate ν, σ

$$L(\nu, \sigma \mid \underset{\sim}{x}) \propto f(\underset{\sim}{x} \mid \nu, \sigma)$$

estimates in the table are

$E(\theta_i \mid \underset{\sim}{x})$ using $\hat{\sigma}, \hat{\nu}$ in the prior

Other option is the prior is nonparametric

Nonparametric empirical Bayes

$$f(x) = \int_0^\infty \frac{e^{-\theta}\theta^x}{x!}\pi(\theta)d\theta$$

$$E(\theta|x) = \int \theta f(\theta|x)\pi(\theta)d\theta \Big/ \underbrace{\int f(\theta|x)\pi(\theta)d\theta}_{f(x)}$$

- Loss function $L(\theta;\hat{\theta}) : \Theta \times \Theta \to \mathbb{R}$ — θ unk. par. in model; $\hat{\theta}$ estimate
- Risk of an estimator — reg^{n}: $L(\theta;\hat{\theta}) = (\theta-\hat{\theta})^2$ sq. err; $|\theta-\hat{\theta}|$ abs. err
- Mean-squared error — class^{n}: $\begin{cases} 0, & \hat{\theta}=\theta \\ 1, & \hat{\theta}\neq\theta \end{cases}$ ("0-1 loss") — expected loss
- Bayes risk — $|\theta-\hat{\theta}|^p$ L_p loss
- maximum risk — Kullback-Leibler loss

$$L(\theta,\hat{\theta}) = \int \log\left\{\frac{f(x;\theta)}{f(x;\hat{\theta})}\right\} f(x;\theta)\,dx$$

"decision" is a choice of estimator $\hat{\theta} = \hat{\theta}(x)$.

- Loss function $L(\theta; \hat{\theta})$ ← choose ↙ unk.
- Risk of an estimator $R(\theta; \hat{\theta}) = \int L(\theta; \hat{\theta}) f(x; \theta) dx = E\{L(\theta, \hat{\theta}(X))\}$ ← θ unknown; $\hat{\theta}$ ↑ est.; $f(x;\theta)$; expected loss
- Mean-squared error ↑ under sq'd error loss
- Bayes risk
- maximum risk

$$R(\theta; \hat{\theta}) = \int \{\theta - \hat{\theta}(x)\}^2 f(x;\theta) dx$$ (L_2)

$$= E\{\hat{\theta}(x) - \theta\}^2 = E_\theta\{\hat{\theta}(x) - E\hat{\theta} + E\hat{\theta} - \theta\}^2$$

$$= \text{var}\{\hat{\theta}(x)\} + \text{bias}^2\{\hat{\theta}(x)\}$$

$\bar{X} \sim N(\theta, 1)$

$\hat{\theta}_1 = \bar{X}$

$\hat{\theta}_2 = 3$

$R(\theta; \hat{\theta}) = E\{\theta - \hat{\theta}(x)\}^2$

R_{h_2}

$1 = E(\theta - \bar{X})^2 = \frac{1}{n}$

$2 = E(\theta - 3)^2 = (\theta - 3)^2$

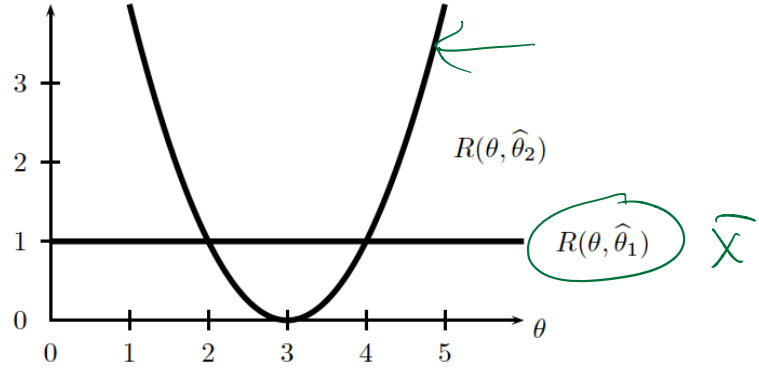

FIGURE 12.1. Comparing two risk functions. Neither risk function dominates the other at all values of θ.

often can't find an estimator for which $R(\theta; \hat{\theta})$ is smallest $\forall\ \theta \in \Theta$

12.3 $X \sim \text{Bin}(n, p)$

$\hat{p}_1 = X/n = \bar{X}$

$R(p, \hat{p}_1) = \text{var}\,\bar{X} = \frac{p(1-p)}{n}$

$\hat{p}_2 = \frac{X+\alpha}{n+\alpha+\beta}$ post. mean if $B(\alpha, \beta)$ prior

$R_{L_2}(p, \hat{p}_2) = \text{var}\,\hat{p}_2 + \{E(\hat{p}_2)\}^2$

$=$ see Eq. below Fig 12.2

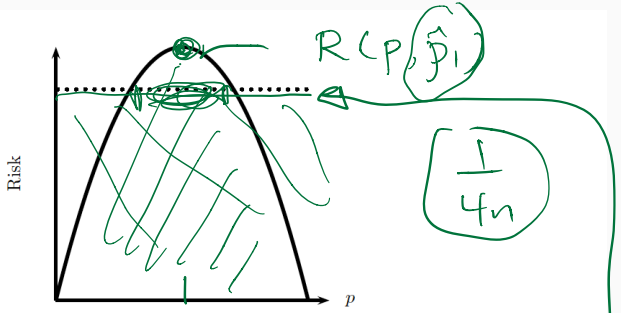

FIGURE 12.2. Risk functions for $\widehat{p}_1$ and $\widehat{p}_2$ in Example 12.3. The solid curve is $R(\widehat{p}_1)$. The dotted line is $R(\widehat{p}_2)$.

$\alpha = \beta = \sqrt{n}/4$

$R_{L_2}(p, \hat{p}_2) = \frac{n}{4(n+\sqrt{n})^2}$

Maximum risk; Bayes risk

$R(\theta, \hat{\theta}(X)) = \int L(\theta, \hat{\theta}(x)) f(x;\theta)\, dx$ ← still dep

$R(\theta, d(x))$

$\hat{p}_1 = \bar{X}; \qquad R(p, \hat{p}_1) = p(1-p)/n$

$\hat{p}_2 = \dfrac{n\bar{X} + \sqrt{n/4}}{n + \sqrt{n}}; \qquad R(p, \hat{p}_2) = \dfrac{n}{4(n+\sqrt{n})^2}$

} Bin

$\bar{R}(\hat{\theta}(x)) = \sup_{\theta \in \Theta} R\{\theta, \hat{\theta}(x)\}$ ~~min~~ max risk

minimax estimator is defined as $\inf_{\tilde{\theta}} \sup_{\theta} R(\theta, \tilde{\theta})$

"minimax lower bd" "minimax decision rule"

Bayes rules

Bayes risk of $\hat{\theta}$: $\int R(\theta, \hat{\theta})\,\pi(\theta)\,d\theta = r_\pi(\hat{\theta})$

$= r(f, \hat{\theta}) = r(\pi, \hat{\theta})$

Bayes est. is the f$^{\underline{n}}$ $\hat{\theta}(x)$ that satisfies

$$\min_{\tilde{\theta} \in \mathcal{C}} r(f, \tilde{\theta}) = \min_{\tilde{\theta} \in \mathcal{C}} r_\pi(\tilde{\theta})$$

$$= \inf_{\tilde{\theta}} r_\pi(\tilde{\theta})$$

Often called Bayes rules (decisions > est'rs)

How to find Bayes rules? (i.e. Bayes estimators)

def: $\inf_{\tilde{\theta}} r_\pi(\tilde{\theta}) = \inf_{\tilde{\theta}} \int_\Theta R(\theta, \tilde{\theta}(x)) \pi(\theta) d\theta$

Theorem 12.7 : note typo in stmt,

$\Theta \leftarrow \hat{\theta}$

Exp'd loss under posterior

$$= \int L(\theta, \hat{\theta}(x))\, \pi(\theta|x)\, d\theta \qquad \text{dep. } \hat{\theta}, x$$

Choose $\hat{\theta}_\pi(x)$ to $\min_{\hat{\theta}} \int L(\theta, \hat{\theta}(x))\, \pi(\theta|x) d\theta$

What's the Bayes risk of $\hat{\theta}_\pi(x)$?

$$= \int R(\theta, \hat{\theta}_\pi(x))\, \pi(\theta) d\theta = r_\pi(\hat{\theta}) \qquad \text{dep. on } \pi$$

$\hat{\theta}$

$$\min_{\tilde{\theta}} \int R(\theta, \tilde{\theta}_\pi(x))\, \pi(\theta) d\theta \longleftarrow$$

$$= \min_{\tilde{\theta}} \int \left\{ \int L(\theta, \tilde{\theta}_\pi(x))\, f(x|\theta) dx \right\} \pi(\theta) d\theta$$

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

$$= \min_{\tilde{\theta}} \int\int L(\theta, \tilde{\theta}_\pi(x))\, \pi(\theta|x) f(x)\, dx\, d\theta$$

$$= \min_{\tilde{\theta}} \int f(x) \int L(\theta, \tilde{\theta}_\pi(x))\, \pi(\theta|x) \quad d\theta\, dx$$

only need to $\min_{\tilde{\theta}} \int L(\theta, \tilde{\theta}_\pi(x))\, \pi(\theta|x) d\theta \longleftarrow$

Conclusion: "Bayes rule" w prior $\equiv$ best posterior estimate

$$\min_{\tilde{\theta}} \int (\theta - \tilde{\theta})^2 \pi(\theta|x) d\theta$$

$$= \ldots = \hat{\theta} = E(\theta|x) \quad \text{answer}$$

admissibility $\leftarrow$ needed w Friday