ST790 — Fall 2022

Imprecise-Probabilistic Foundations of Statistics

Ryan Martin
North Carolina State University
www4.stat.ncsu.edu/~rmartin

Week 10a

This lecture

- Recap the new (partial-prior) IM construction
- Properties:
  - strong validity
  - "near coherence"
- Dimension-reduction to improve efficiency
- Illustration

Introduction

- Usual setup: $(Y, \Theta) \sim \overline{\mathsf{P}}_{Y,\Theta}$
  - precise model $(Y \mid \Theta = \theta) \sim \mathsf{P}_{Y|\theta}$, density $p_\theta(y)$
  - imprecise prior for Θ with contour $q(\theta) = \overline{\mathsf{P}}_\Theta(\{\theta\})$
- Combine the two into a *consonant* IM $(\underline{\Pi}_y, \overline{\Pi}_y)$ with contour

$$\pi_y(\theta) = \overline{\mathsf{P}}_{Y,\Theta}\{\eta(Y, \Theta) \leq \eta(y, \theta)\},$$

where η is a "relative likelihood"

$$\eta(y, \theta) = \frac{p_\theta(y)\, q(\theta)}{\sup_\vartheta p_\vartheta(y)\, q(\vartheta)}$$

Intro, cont.

- Strong validity follows by the construction
- Computation typically is non-trivial
- If $\overline{\mathsf{P}}_\Theta$ is consonant, then

$$\begin{aligned}\pi_y(\theta) &= \int_0^1 \Big[\sup_{\vartheta : q(\vartheta) > \alpha} \mathsf{P}_{Y|\vartheta}\{\eta(Y, \vartheta) \le \eta(y, \theta)\} \Big] \, d\alpha \\ &\approx \int_0^1 \Big[\max_{s : q(\vartheta_s) > \alpha} \frac{1}{M} \sum_{m=1}^{M} 1\{\eta(Y_s^{(m)}, \vartheta_s) \le \eta(y, \theta)\} \Big] \, d\alpha\end{aligned}$$

- At least a naive Monte Carlo strategy is available[1]
- Less-naive strategies would be available depending on specifics of the problem at hand

[1] https://researchers.one/articles/22.05.00001

Properties

- IM construction is based on a suitable "outer consonant approximation" of $\overline{\mathsf{P}}_{Y,\Theta}$
- Strong validity is an immediate consequence

Theorem.

The IM constructed above is strongly valid (wrt $\overline{\mathsf{P}}_{Y,\Theta}$), i.e.,

$$\overline{\mathsf{P}}_{Y,\Theta}\{\pi_Y(\Theta) \le \alpha\} \le \alpha, \quad \alpha \in [0,1]$$

Corollary.

Strong validity implies that IM-based confidence regions, $C_\alpha(y) = \{\theta : \pi_y(\theta) > \alpha\}$, satisfy

$$\overline{\mathsf{P}}_{Y,\Theta}\{C_\alpha(Y) \not\ni \Theta\} \le \alpha, \quad \alpha \in [0,1]$$

Properties, cont.

- Strong validity is "frequentist" in nature
- What about "behavioral" properties, e.g., coherence?
- Think of the IM construction as an updating rule that maps prior $(\underline{\mathsf{P}}_\Theta, \overline{\mathsf{P}}_\Theta)$ to a "posterior" $(\underline{\Pi}_y, \overline{\Pi}_y)$
- Can't get full-blown coherence, but...

Theorem.

The strongly valid constructed above, interpreted as an "updating rule," is *half-coherent* in the sense that

$$\inf_{y \in \mathbb{Y}} \underline{\Pi}_y(A) \leq \underline{\mathsf{P}}_\Theta(A) \quad \text{and} \quad \sup_{y \in \mathbb{Y}} \overline{\Pi}_y(A) \geq \overline{\mathsf{P}}_\Theta(A), \quad \text{all } A$$

Properties, cont.

- As suggested, this falls short of full-blown coherence
- In fact, the above is just one of the two conditions required for coherence, the other condition is

$$\underline{P}_{\Theta}(A) \leq \max\Big\{\underline{\Pi}_y(A), \sup_{x \neq y} \overline{\Pi}_x(A)\Big\}, \quad \text{all } y, \text{ all } A$$

- Intuition: If above condition fails, then you have a strategy that makes me look silly
  1. Determine the (y, A) at which above fails
  2. Sell me a gamble on A for $\underline{P}_{\Theta}(A)$
  3. Then wait for Y to be observed and proceed as follows:
     - if $Y \neq y$, then you buy it back for $\overline{\Pi}_y(A) < \underline{P}_{\Theta}(A)$ and I lose
     - if $Y = y$, then do nothing, forcing me to pay $\underline{P}_{\Theta}(A)$ which is more than my advertised buying price $\underline{\Pi}_y(A)$

Properties, cont.

- There's nothing about the IM construction that ensures the second condition above is satisfied
- But it holds trivially if $y \mapsto \overline{\Pi}_y(A)$ is continuous
- There are cases where this second condition fails
  - in some cases,[2] only gBayes is coherent
  - IM construction is different from gBayes
- So: no general, full-blown coherence result for the new IM
- We get improved efficiency, a good trade (in my opinion)

[2] Walley 1991, Sec. 6.5.4

Efficiency considerations

- A precise, mathematical definition of efficiency is difficult to pin down, currently unavailable
- As described before, want the IM's contours to be as tightly concentrated as possible for each y
- e.g., this makes the $C_\alpha(y)$ confidence region small
- Too much to ask for "tightest concentration" uniformly in y
- We have intuition to guide us, so let's proceed informally

> *If we define mathematics as the art and science of deductive reasoning... then statistics (the art and science of induction) is essentially anti-mathematics. A mathematical theory of statistics is, therefore, a logical impossibility!* —D. Basu

Efficiency, cont.

- Key insight:[3] don't do unnecessary calculations in the Choquet integral that defines contour π_y
  - related to the "curse-of-dimensionality"
  - integrating over non-essential dimensions inflates contour, reduces efficiency
- So, reduce the dimension as much as possible first
- First example: if $U = U(Y)$ is sufficient, then

$$\eta(y, \theta) = \frac{p_\theta(y)\, q(\theta)}{\sup_\vartheta p_\vartheta(y)\, q(\vartheta)} = \frac{p_\theta(u)\, p(y \mid u)\, q(\theta)}{\sup_\vartheta p_\vartheta(u)\, p(y \mid u)\, q(\vartheta)} = \frac{p_\theta(u)\, q(\theta)}{\sup_\vartheta p_\vartheta(u)\, q(\vartheta)} \quad \leftarrow \text{only depends on } (u, \theta)!$$

[3] I'm currently calling this a *Principle of Minimum Complexity*

Efficiency, cont.

- In most cases, $U = U(Y)$ is lower-dim than Y
- If η only depends on (u, θ), then we only need to compute the Choquet integral over $U(\mathbb{Y}) \times \mathbb{T}$, lower-dim!
- This is free — no extra computations or loss of info
- In some cases, further reduction is possible by conditioning on ancillary statistics, but I won't discuss this

Efficiency, cont.

- There are certain "extreme" cases to consider
- One[4] is the case where prior information is *vacuous*
  - here $q(\theta) \equiv 1$, so $\eta(y, \theta) = p_\theta(y) / \sup_\vartheta p_\vartheta(y)$
  - naive application of the construction gives

$$\pi_y^{\text{naive}}(\theta) = \sup_\vartheta \mathsf{P}_{Y \mid \vartheta}\{\eta(Y, \vartheta) \leq \eta(y, \theta)\}$$

  - the integration/supremum over ϑ creates inefficiency
  - a clearly more efficient IM has contour

$$\pi_y(\theta) = \mathsf{P}_{Y \mid \theta}\{\eta(Y, \theta) \leq \eta(y, \theta)\}$$

- How to handle this efficiently?

[4] Another is when the prior is complete, like in the traditional Bayes case

Efficiency, cont.

- Plausibility ordering: $\eta(y,\theta) = p_\theta(y) / \sup_\vartheta p_\vartheta(y)$
- This resembles a *conditional* density, not a joint
- Recall the minimum complexity principle: if dimension can be reduced before integration, then do it
- In this case,
  - can reduce dimension by *fixing* θ
  - work with just the conditional that appears in η
- End result is clearly more efficient:

$$\pi_y(\theta) = \mathsf{P}_{Y|\theta}\{\eta(Y,\theta) \le \eta(y,\theta)\} \le \pi_y^{\text{naive}}(\theta), \quad \text{all } \theta$$

Marginal inference

- It's often the case that the quantity of interest is just a feature of the model parameters
- Express the model parameter Λ as (Θ, Ψ), where
  - Θ is the feature of interest
  - and Ψ is the nuisance parameter
- e.g., suppose Λ is a mean vector, $\Theta = \|\Lambda\|$, and Ψ is the unit vector in direction of Λ
- One (naive) strategy is to construct a strongly valid IM for Λ and then marginalize to Θ via extension principle
- But if one is only interested in Θ, then efficiency can be gained by reducing dimension before IM construction

Marginal inference, cont.

- There are a number of ways this can be handled
- Without any special structure, the best strategy is via *profiling*, i.e., using plausibility order

$$\eta(y, \theta) = \frac{\sup_\psi p_{\theta,\psi}(y)\, q(\theta)}{\sup_{\vartheta,\psi} p_{\vartheta,\psi}(y)\, q(\vartheta)}$$

- Some problems have special structure, however
- Suppose that, e.g.,

$$p_{\theta,\psi}(y) = p_\psi\big(U(y)\big)\ \underbrace{p_\theta\big(V(y) \mid U(y)\big)}_{\text{no dependence on } \psi}$$

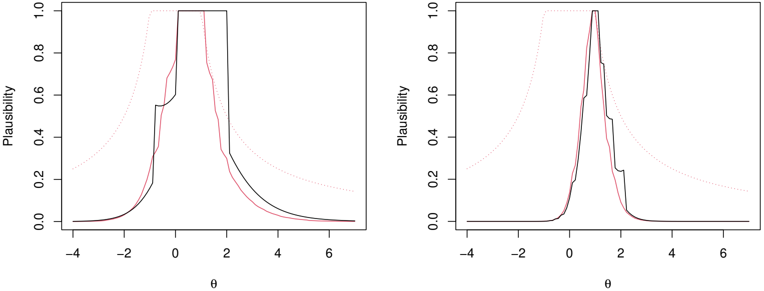
- ψ-dependence cancels out in the profiling step, so we get dimension reduction since only have to integrate over V

Example

- Important practical example:
  - $(Y_i \mid \Lambda_i = \lambda_i) \overset{\text{ind}}{\sim} \text{Bin}(n_i, \lambda_i)$, for $i = 1, 2$
  - interest in the log-odds ratio $\Theta = \log\left(\frac{\Lambda_2}{1-\Lambda_2} \div \frac{\Lambda_1}{1-\Lambda_1}\right)$
  - $\Theta = 0 \iff \Lambda_1 = \Lambda_2$
- Define $V = Y_1$ and $U = Y_1 + Y_2$
- Well-known result is that

$$p_\lambda(v \mid u) \equiv p_\theta(v \mid u) \propto \binom{n_2}{v}\binom{n_1}{u-v} e^{\theta v},$$

  for $v = \max(u - n_1, 0), \ldots, \min(n_2, u)$
- Use this conditional dist to construct a marginal IM for Θ



(a) $Y = (1, 2)$, $n = (43, 39)$ (b) $Y = (4, 11)$, $n = (146, 154)$

Figure: Plausibility contours for the log odds ratio in two mortality data sets: vacuous prior (black) & partial prior (red)

Next lecture

- Wrap-up discussion of IM construction
- Imprecise prob methods for prediction/classification
- ...