ST790 — Fall 2022

Imprecise-Probabilistic Foundations of Statistics

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

Week 14a

This lecture

- Uncertainty about the model
  - setup, challenges, etc.
  - Occam's razor & regularization
  - partial priors
- Imprecision in the model
  - missing data
  - coarse data
  - ...

Introduction

- In our stat/ML discussions, we've assumed that the statistical model is known & precise
- But this is unrealistic, often there's
  - uncertainty about the model
  - imprecision in the model
- Model uncertainty, I think, is clear/easy to imagine
  - different models being entertained: normal vs heavy-tailed
  - "model" is of direct interest, e.g., variable selection
- Model imprecision isn't as clear/familiar (at least to me)
  - e.g., manifests is when data are missing/coarse
  - can also be "imprecisely specified"

Model uncertainty

- Let $M \in \mathbb{M}$ represent the uncertain model
- Model-specific parameter $\Theta_M \in \mathbb{T}_M$, given M
- Density/mass function for Y depends on (M, Θ_M)
- Examples:
  - sparse normal mean vector
    - $\mathbb{M} =$ power set of $\{1, 2, \ldots, n\}$
    - $\mathbb{T}_M =$ n-vectors with 0's in the entries corresponding to M^c
    - $(Y \mid M, \Theta_M) \sim \mathsf{N}_n(\Theta_M, I_n)$
  - mixture model
    - $\mathbb{M} = \{1, 2, \ldots\}$
    - $\Theta_M = (\omega_1^M, \ldots, \omega_M^M, \lambda_1^M, \ldots, \lambda_M^M)$ for $M \in \mathbb{M}$
    - PDF/PMF: $y \mapsto \sum_{m=1}^M \omega_m^M p_{\lambda_m^M}(y)$
- *Goal:* quantify uncertainty about M, given $Y = y$

Model uncertainty, cont.

- Can identify Θ as the pair (M, Θ_M)
- Then M is the interest parameter, Θ_M is a nuisance
- Goal is marginal inference on M
- Unique aspects of this marginal inference problem:
  - $\mathbb{M}$ is discrete, but could be *very large*
  - data necessarily supports[1] the most complex $M \in \mathbb{M}$
- Last point explains the need[2] for *regularization*
- Classical ways to do this:
  - frequentists use "ad hoc" penalties: AIC, BIC, etc.
  - Bayesians need a precise prior on (M, Θ_M)
- Imprecise probability, i.e., partial priors, seems promising

[1] This is why you can't use R^2 for variable selection in regression!
[2] Some inference can be done w/o regularization (M., *ISIPTA'19*)

Valid partial-prior marginal IMs

- Let $L_y(m, \theta_m)$ denote the likelihood function
- Occam's razor[3]-motivated partial prior:[4]
  - informative about M, vacuous on Θ_M
  - i.e., $q(m, \theta_m) \equiv q(m)$, $m \in \mathbb{M}$
  - assume q is a *possibility contour* on $\mathbb{M}$
  - $q = 1$ at the "simplest M," decreasing in complexity
- Profile relative likelihood for M:

$$\eta(y, m) = \frac{\sup_{\theta_m} L_y(m, \theta_m)\, q(m)}{\sup_\mu \sup_{\theta_\mu} L_y(\mu, \theta_\mu)\, q(\mu)}, \quad m \in \mathbb{M}$$

- Red terms might be easy to compute...

[3] The *principle of parsimony*, i.e., simpler models are preferred to more complex models, `https://en.wikipedia.org/wiki/Occam's_razor`

[4] Of course, this isn't the only option

Valid marginal IMs, cont.

- Follow the general framework I described before:

$$\begin{aligned}\pi_y(m) &= \text{upper probability of } \text{``}\eta(Y, M) \leq \eta(y, m)\text{''} \\ &= \int_0^1 \sup_{(\mu,\theta):q(\mu)>\alpha} \mathsf{P}_{Y|\mu,\theta}\{\eta(Y,\mu) \leq \eta(y,m)\}\, d\alpha\end{aligned}$$

- Computation???
- General validity[5] results apply here, e.g.,
  - valid "probabilistic reasoning" about M
  - certain coherence-like properties hold
  - $\{m : \pi_y(m) > \alpha\}$ is a $100(1-\alpha)\%$ confidence set for M
- To my knowledge, no other results like this are available...

[5] Recall, "validity" here is wrt the imprecise joint dist for (Y, M, Θ_M)

Sparse normal mean

- Partial prior: $q(M) = q_{|M|}$, only depends on cardinality $|M|$
- Relative likelihood[6]

$$\eta(Y, M) = \frac{\exp(-\frac{1}{2\sigma^2}\sum_{i \notin M} Y_i^2)\, q_{|M|}}{\max_{k \in 0:n}\{\exp(-\frac{1}{2\sigma^2}\sum_{i > k} |Y|_{[i]}^2)\, q_k\}}$$

- Distribution of $\eta(Y, M)$ as a function of Y when $\Theta_{M^c} = 0$?
- In particular, how does the distribution depend on $\Theta_M \neq 0$?
- *Conjecture:* $\eta(Y, M)$ is stochastically largest when $\Theta_M = 0$
- Intuitive explanation:
  - a Y_i with large non-zero mean can only make den small
  - smaller denominator makes ratio $\eta(Y, M)$ larger
  - larger $\eta(Y, M)$ makes smaller $x \mapsto \mathrm{P}\{\eta(Y, M) \leq x\}$

[6] $|Y|_{[1]} > |Y|_{[2]} > \cdots > |Y|_{[n]}$, reverse order statistics

Sparse normal mean, cont.

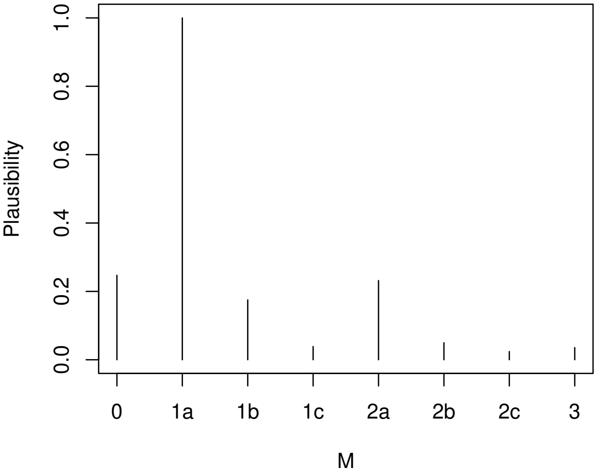
- The conjecture would *drastically* simplify the computation
- In particular, if all Θ's are zero, then
  - can use the same Y samples for all the Monte Carlo evals
  - $\eta(Y, M) \stackrel{\text{D}}{=} \eta(Y, |M|)$
- Form is actually pretty simple:

$$\pi_y(m) = \sum_{k=0}^{n} (q_k - q_{k-1}) \mathsf{P}_{Y|0}\{\eta(Y, k) \leq \eta(y, m)\}, \quad m \in \mathbb{M}$$

- In my examples below, $q^k = 0.2^k$, for $k = 0, 1, \ldots, n$

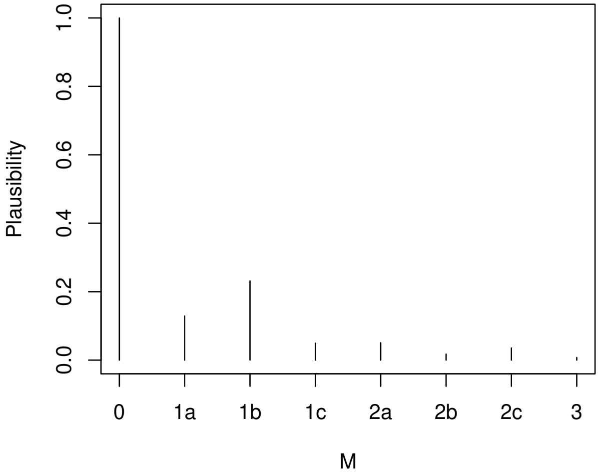
Sparse normal mean, cont.

$$n = 3,\ \sigma = 1, \text{ and } y = (0.1, 1.5, 2)$$



Sparse normal mean, cont.

$$n = 3,\ \sigma = 1,\ \text{and}\ y = (0.1, 1.5, 1)$$



Remarks

- Wouldn't be too hard to scale this up to larger n
- The simplicity is specific to the normal mean problem
- Also depends on a *conjecture*
- But keep in mind:
  - valid UQ about the *model*
  - based on partial prior, no unnecessary assumptions needed
  - neither Bayes nor frequentist can do this
- How far can this be pushed?

Model imprecision

- Previously, "$\overline{\mathrm{P}}_{Y,\Theta}$" was based on
  - a precise model for Y, given $\Theta = \theta$
  - a partial/imprecise prior for Θ
- "$Y \mid \Theta$" could be imprecise too, say, for robustness[7]
- e.g., an ε-contamination nbhd around a given "$\mathrm{P}_{Y|\theta}$"
- Then lower/upper likelihood functions[8] become relevant

$$\underline{L}_y(\theta) = \inf \mathrm{P}_{Y|\theta}(\{y\}) = (1-\varepsilon)p_\theta(y)$$
$$\overline{L}_y(\theta) = \sup \mathrm{P}_{Y|\theta}(\{y\}) = (1-\varepsilon)p_\theta(y) + \varepsilon$$

- I don't have much experience with this...

[7] Huber & Ronchetti's *Robust Statistics*

[8] I'm assuming Y is discrete here...

Model/data imprecision

- A different perspective emerges with imprecise data[9]
- Start with a simple/extreme case of *missing data*
- In observational studies, it's common for data to be missing, e.g., non-response to some/all questions on a survey
- Only safe to ignore missing data under strong assumptions
  - introduces bias if missingness and response are related
  - can't test/check for this because missing data is *missing*
- So great care is needed here...
- Turns out to have some connection to imprecise probability

[9] More generally, *partially identified* models as in Manski's book

Model/data imprecision, cont.

- Y consists of a pair (Y, Δ)
  - Y is the actual response value
  - Δ is the not-missing/missing indicator
- There exists a Y value regardless of Δ, it's just that we don't get to see the value of Y if $\Delta = 0$
- There's a marginal distribution for Y:

$$p_\theta(y) = \underbrace{w_\theta(1)}_{\checkmark} \underbrace{p_\theta(y \mid \Delta = 1)}_{\checkmark} + \underbrace{w_\theta(0)}_{\checkmark} \underbrace{p_\theta(y \mid \Delta = 0)}_{\times}$$

- Some parts are identified, some aren't
- Hence, Manski's *partial identifiability* terminology

Model/data imprecision, cont.

- Unidentified parts can be effectively anything, so the model is really a contamination nbhd w/ upper likelihood, etc.
- Why no imprecise probability in the missing data literature?
- If one *assumes* that missingness is completely random
  - i.e., $\mathsf{P}_\theta(\Delta = 1)$ is constant in θ
  - then likelihood only depends on the observed y values
  - can get MLE etc. directly from this
- Assumption might or might not be justifiable
- The situation is much more complicated/interesting when covariates are involved

Model/data imprecision, cont.

- More generally, data might be *coarse*
- Measurement of Y has limited precision
  - Missing data is an extreme case of zero precision
  - censored data is a common example, a result of not being able to continuously monitor subjects
- Arguably, almost all real problems involve coarse data
- Most natural strategy is a *random set* model
- Why don't you see this approach in the stat literature?
  - might assume data imprecision is negligible compared to...
  - like above, if one *assumes* that coarsening happens randomly, then likelihood only depends on the "precise model"
  - MLE, etc., can be obtained w/o thinking about imprecision
- Again, more complicated/interesting with covariates

Summary

- Uncertainty and/or imprecision can be at the model level
- Existing approaches can deal with model uncertainty, but (IMO) not in a satisfactory way:
  - frequentists can choose $\widehat{M}$, but no UQ
  - Bayesians get UQ, but it requires a (precise and proper) prior and has no validity guarantees[10]
- New framework for strongly valid marginal IMs applies, at least in principle, right off the shelf
- Questions remain about efficient computation
- I didn't really say anything about model/data imprecision
- My very modest goal was just to point out that these issues exist and deserve serious attention

[10] See plots in M. *ISIPTA'19*

Next lecture

- Simpson's paradox
  - general setup & why it's scary
  - connection to imprecise probability
- Miscellany