On “Encouters with Imprecise Probabilities” by Jim Berger

Chris Holmes

Professor of Biostatistics
University of Oxford

Scientific Director for Health
The Alan Turing Institute

O’Bayes
June 2019

Scene setting for Imprecise Probabilities (IP)

- All probability statements and quantification of beliefs are imprecise
- There's never been a 'fair coin' for which I genuinely believe that $Pr(heads) = 0.5$
  - for which I would be prepared to wager a bet $\$ \to \infty$ on returning a long-run frequency of 1/2
- Jim Berger tackles IP within the framework of Bayesian statistics; although it's worth noting that other learning systems design themselves around this problem
  - Levi (1974) "Indeterminate Probabilities"
  - Dempster–Shafer theory for belief functions

Two general approaches to IP

- JB considers two general approaches to Bayesian updating under IP
- Herman Rubin approach:
  - Undertake an *a priori* sensitivity analysis by constructing a set of models $\mathcal{P}$ to update, and report interesting properties over the class of posteriors
- Jack Good approach:
  - Reduce sensitivity to prior specification via hyper-priors

These objective issues are have strong subjective roots

- with subjectivist connections.....

  "Subjectivists should feel obligated to recognise that any opinion (so much more the initial one) is only vaguely acceptable. . . So it is important not only to know the exact answer for an exactly specified initial problem, but what happens changing in a reasonable neighbourhood the assumed initial opinion." De Finetti, as quoted in Dempster (1975).

  ". . . in practice the theory of personal probability is supposed to be an idealization of one's own standard of behaviour; that the idealization is often imperfect in such a way that an aura of vagueness is attached to many judgements of personal probability..." Savage (1954).

Four motivating applications

- JB presents four motivating applications
    I. Interval probabilities
    II. p-values
    III. Priors for the multivariate normal linear model
    IV. Uncertainty Quantification in computer models

I. Dealing with interval probabilities

- Use the Rubin approach, within the field of *robust Bayesian analysis*, and carry forward a collective set of prior models – note this is not model averaging
- JB treats this in a pure inference setting
- One (small) issue for "objectivists" is that the priors on intervals aren't invariant to transformation
  - for example a uniform probability of rain over the interval $\in [0.75, 0.8]$ will be different to uniform over the log-odds $[\log 3:1, \log 4:1]$

Decision theory for intervals

- An alternative is to explore consequences of imprecise probabilities within decision analysis
- For example, suppose I specify my prior on "rain tomorrow" as 0.4
  - when in truth it was 0.3649274014829063987104
  - or I think that a reasonable interval is $[0.3, 0.5]$
- Should I be worried with using the 0.4 approximation?
  - Maybe yes, but maybe no.....
- It seems to difficult to separate out the impact of prior specification, or specification of intervals, from the decision task
  - If optimal actions change dramatically over the prior interval I would be concerned, but if they are stable then less so
  - This is the approach taken in Watson & Holmes (2016), Statistical Science; and outside of Statistics, Whittle reviewed in his book "Risk optimized control" (1990); and extended by the Nobel Laureate Hansen & Sargent in their book on "Robustness" (2007)

II. Pure testing problems and p-values

It's brilliant!

III. Hierarchical priors for multivariate linear model

We consider the model

$$\boldsymbol{\theta_i} = \boldsymbol{z_i}\boldsymbol{\beta} + \epsilon_i^*, \quad \epsilon_i^* \sim N_k(\cdot|0, \boldsymbol{V})$$

where $\boldsymbol{\theta}$ is a $k \times 1$ multivariate outcome, and new objective priors

$$\begin{aligned} \pi(\boldsymbol{\beta}) &\propto \frac{1}{(1 + ||\boldsymbol{\beta}||^2)^{(p-1)/2}}, \\ \pi(\boldsymbol{V}) &\propto \frac{1}{|\boldsymbol{V}|^{1-1/(2k)} \prod_{i<j}(d_i - d_j)} \end{aligned}$$

for ordered eigenvalues $(d_1, \ldots, d_k)$

- $\pi(\boldsymbol{\beta})$ looks like an improper multivariate Student (with $\nu = -1$ degrees of freedom) and applies global shrinkage
- $\pi(\boldsymbol{V})$ encodes an assumption of white noise (constant frequency spectrum) and equal spread of variance

Table 1 of the accompanying paper suggests for posterior propriety you only need $n \geq 1$ – seems remarkable

III. Hierarchical priors for multivariate linear model

- Simulations in the paper show the advantage (MSE) over twelve other default priors (formed from *independent* combinations of $\pi(\boldsymbol{\beta})\,\pi(\boldsymbol{V})$)
- A few questions:
  - what happens to performance under the least favourable conditions for the prior? e.g. if the real eigenvalues are such that $d_1 \gg d_2 \gg d_3 \ldots$, so that the noise is spread along particular axes
  - what's to be done if $p > n$
  - out of scope: but can we use these priors under model choice?
    - in recent work (Fong & Holmes 2019) we showed that Bayesian marginal likelihood is just exhaustive cross-validation over all of the $(2^n - 1)$ possible held-out test sets using the log marginal predictive as the scoring rule

$$\log p(y|M) = \frac{1}{n}\sum_{k=1}^{n}\binom{n}{k}^{-1}\sum_{j=1}^{\binom{n}{k}}\frac{1}{k}\sum_{i=1}^{k}\log p(y_{j(i)}|y_{j(k+1:n)})$$

it would be interesting to compare models with the objective prior aggregating the cross-validation score after propriety

IV. Uncertainty Quantification

One main question

- What is the advantage of the joint approach

$$y^O(x) = y^M(x,u) + b(x,u) + \epsilon$$

  as compared to a two stage approach

  Stage-1:

$$y^O(x) = y^M(x,u) + \epsilon$$

  without the bias term, followed by model criticism under

  Stage-2:

$$(y^O - y^{\widehat{M}(x,u)}) = b(x,u) + \epsilon$$

- Is there any advantage to consider both?

Conclusions

- This is a wonderful and rich talk (and accompanying paper)
- Thought provoking contribution to dealing with imprecision in statistical inference