Comment on: 'Predictive Inference: A Path Towards Objectivity'

B. Clarke

Dept. of Statistics
U. Nebraska-Lincoln

September 10, 2022

Outline

Key Points

- Forward predictive sampling is a new technique for finding an objective posterior.
- In this sort of predictive modeling the dialog between Statistician and Scientist is how to update a predictive model rather than selecting prior and likelihood.
- In this conceptualization of the Statistician-Scientist dialog, modeling means ensuring that predictive updates don't drift from the data generator rather than finding a likelihood (to eliminate bias) or prior selection (to summarize pre-experiemental information).
- Uncertainty quantification is derived from the unseen data.

Setting

1. Fundmental Equation:

$$\pi(\theta \mid y_{\text{obs}}) = \int \pi(\theta \mid y_{\text{comp}}) p(y_{\text{mis}} \mid y_{\text{obs}}) \mathrm{d} y_{\text{mis}}.$$

   The focus is the posterior predictive $p(y_{\text{mis}} \mid y_{\text{obs}})$.
2. To model the data, it is enough to specify $p(y_{\text{mis}} \mid y_{\text{obs}})$ directly; the prior does not appear.
3. Therefore seek objective predictives not objective priors.
4. Use EDF to get a one-step-ahead objective predictive:

$$P(Y_{n+1} | y_{1:n}) = (1/n) \sum_{i=1}^{n} \delta_{y_i}$$

   Predictives for Y_{n+2}, Y_{n+3} etc. similar.

Procedure

- The outcomes from the one-step ahead predictives give

$$Y_{n+1:\infty} \sim p(y_{n+1:\infty}|y_{1:n}) = p(y_{\text{mis}} \mid y_{\text{obs}})$$

- Feed these into the Fundamental Equation to find the posterior $\pi(\theta \mid y_{\text{obs}})$.
- Theory: If we have exchangeable data $y_{1:n}$ from density m_n. De Finetti tells us $\exists\ p_\theta, \pi(\cdot)$ so that

$$m_n(y_{1:n}) = \int \pi(\theta) p_\theta(y_1) \cdots p_\theta(y_n) \mathrm{d}\theta.$$

- Now, $\pi(\theta|y_{\text{obs}}) = \pi(\theta|y_{1:n})$ is well-defined.

Algorithm I

- Given π and p_θ we can form $m(y_{n+1}|y_{1:n})$:

$$m(y_{n+1}|y_{1:n}) = \int p_\theta(y_{n+1})\pi(\theta|y_{1:n})\mathrm{d}\theta$$

- Draw a y_{n+1} from $m(y_{n+1}|y_{1:n})$. Now we we have $y_{1:n+1}$ and $m(y_{1:n+1}) = m(y_{n+1}|y_{1:n})m(y_{1:n})$.
- So, we could in principle form (but we don't)

$$\pi(\theta|y_{1:n+1}) = \pi(\theta)p_\theta(y_{1:n+1})/m(y_{1:n+1}).$$

- In fact, for objectivity, we use the predictive EDF in place of $m(y_{n+1}|y_{1:n})$'s to generate y_{n+1}.

Algorithm II

- Using the posterior we could find find

$$\bar{\theta}_{n+1} = E(\Theta|y_{1:n+1}) = \int \theta\pi(\theta|y_{1:n+1})\mathrm{d}\theta$$

  (but we don't). We find $\bar{\theta}$ using the outomes of the predictive EDF's.
- Repeat this procedure N times for $n+1$, $n+2$, $n+3$, and so on up to $n+N$ to get $\bar{\theta}_{n+N}$.
- Write $\bar{\theta}_{n+N} = \bar{\theta}_{n+N,1}$ and repeat the above procedure M times to get the sequence $\bar{\theta}_{n+N,1}, \ldots, \bar{\theta}_{n+N,M}$.
- Use the sequence of length M to form $\hat{\pi}(\theta|y_{1:n})$.
- Can show $\hat{\pi}(\theta|y_{1:n}) \to \pi(\theta|y_{1:n})$ in various modes – in m.

Missing data

- Original paper on reference priors (JRSSB 1979) described missing data as the result of infinite repetitions of an experiment that we didn't do.
- In particular, asymptotically maximizing

$$E_m D(w(\cdot) \| w(\cdot | Y^n))$$

  over all the missing data for each n gives the prior w that makes the prior and posterior as far apart as possible in expected (in m) KL distance.
- w_{opt} is defined asymptotically and ensures the missing data is maximally informative, in m.
- This is the same concept of missing information and mode as used here. Maybe we should think this way more often.

Martingales

- Easy to see that $E_m\pi(\theta|Y_{1:n}) = \pi(\theta)$. So, updating adds no information under m.
- More is true: $E(\pi(\theta|Y_{1:n+1})|Y_{1:n}) = \pi(\theta|Y_{1:n})$. So the posterior density is a martingale under m.
- So is any predictive: $E(m_n(\cdot|Y_{1:n+1})|Y_{1:n}) = m(\cdot|Y_{1:n})$.
- This is typical for conditioned quantities that have a limit, e.g., have finite absolute moments.
- Thus: Same is true if we replace $\pi(\theta)$ by $\pi(\theta|y_{1:n})$ and adjust the conditioning accordingly.
- Not true under IID models like p_θ.
- Thus, $E(\Theta|y_{1:n})$ is a martingale under m and converges as $n \to \infty$ – to what? Spoiler: Θ.

Let's look at convergences under *m*

- If $\theta \in A$ then under P_θ,

$$\Pi(A|y_{1:n}) = \frac{\int_A \pi(\theta)p(y_{1:n}|\theta)\mathrm{d}\theta}{\int_\Omega \pi(\theta)p(y_{1:n}|\theta)\mathrm{d}\theta} \to 1.$$

- If $\theta \in A^c$ then under P_θ,

$$\Pi(A|y_{1:n}) = \frac{\int_A \pi(\theta)p(y_{1:n}|\theta)\mathrm{d}\theta}{\int_\Omega \pi(\theta)p(y_{1:n}|\theta)\mathrm{d}\theta} \to 0.$$

- In m, Lijoi et al. (2004) Theorem 1 gives $\exists\ \hat{g}$ random

$$\Pi(A|y_{1:n}) \to I_{\hat{g}}(A).$$

Getting back Θ

- Recall for any A,
  $m(y_{1:n}) = \int_A \pi(\theta) p(y_{1:n}|\theta) \mathrm{d}\theta + \int_{A^c} \pi(\theta) p(y_{1:n}|\theta) \mathrm{d}\theta$.
- So, mixing over θ with w to get convergence in m lets us see that the limit is

$$I_{\hat{g}}(A) = I_\Theta(A) = \begin{cases} 1 & \Theta = \theta \in A \\ 0 & \Theta = \theta \in A^c \end{cases}$$

- Since $\Pi(I_\Theta(A) = 1) = \Pi(A)$, under m, as $n \to \infty$

$$\Pi(A|y_{1:n}) \to \Pi(A)$$

- It looks like we're nowhere. But:

Getting the posterior

- Take $\pi(\theta)$ to be the unknown $\pi(\theta|y_{1:n})$. Then

$$\Pi(A|y_{1:n}, y_{n+1:N}) \xrightarrow{m} \Pi(A|y_{1:n})$$

  as $N \to \infty$.
- Want analogous results for posterior density, posterior mean, posterior predictives and predictive EDF's. Especially to justify the forward predictive sampling that generates the missing data for the algorithm.

Best guesses

- Using standard asymptotics and martingale convergence:

$$E_m(\Theta|y_{1:n}, Y_{n+1:N}) \begin{cases} \xrightarrow{p_\theta} \theta \\ \xrightarrow{m} (\Theta|y_{1:n}). \end{cases} .$$

  This convergence is why posterior means work to give the posterior.

- Similar results for $\pi(\theta|y_{1:n}, Y_{n+1:N})$, $m(y_{n+i+1}|y_{1:n+i})$, and $\hat{F}(y_{n+i+1}|y_{1:n+i})$.
- Doob's theorem gives convergences to random variables that appear unrelated to the sequence converging.

In context

- Everything is going to a function of Θ under m.
- Thus: The algorithm is an implementation of martingale convergence under m by repeated sampling from EDF predictives over vectors $y_{n+1:N}$ for large N.
- The 'missing' data generated from the $m(y_{n+i+1}|y_{1:n+i})$'s gives M independent copies of $\bar{\theta}_N$ that can generate a consistent estimate of the posterior.

Intuition for Doob's Theorem

- Theorem 6.10 from Ghosal and van der Vaart (2017): Under regularity conditions,

$$\exists\, f : \mathcal{Y}^\infty \longrightarrow \Omega$$

  so that $\forall \theta \in \Omega$, $f(y^\infty) = \theta$, a.s., in P_θ.
- Nice estimators like posterior means $\bar{\theta} = E(\Theta | y_{1:n})$ have this property.
- This gives a 'foliation' of $\mathcal{Y}^\infty$ under the p_θ's:

$$\mathcal{Y}^\infty = \dot{\cup}_{\theta\in\Omega}\{y^\infty | \theta(\hat{y}^\infty) = \theta\} \equiv \dot{\cup}_{\theta\in\Omega} V_\theta$$

  with $V_\theta \cap V_{\theta'} = \phi$ and $P_\theta(V_{\theta'}) = 0$, for $\theta \neq \theta'$; $P_\theta(V_\theta) = 1$.

Strings of data

- Note the V_θ's are big sets – in particular they are closed under permutation and finite dimensional perturbation.
- But: Under M we have $M(V_\theta) = 0$ even as $M(\dot{\cup}_{\theta \in \Omega} V_\theta) = 1$.
- Loosely, Chen (1985) explains how Bayes convergences are functions of strings of data, i.e., which V_θ has the data.
- So, convergences in M necessarily give random variables as limtis because they mix over the V_θ's.

In context

- Many strings of data $y_{n+1:N}$ are generated, from many V_θ's so the $\bar{\theta}$'s fill out the range of $\pi(\cdot|y_{1:n})$ as a representative sample of the posterior.
- So, in $\Omega \times \mathcal{Y}^\infty$, we can have conceptually a data point $(\theta, y_{1:\infty})$ and a 'density' value $\pi(\theta, y_{1:\infty})$ for it.
- Maybe better not to write densities on $\mathcal{Y}^\infty$ (since it's not clear what dominating measure to use) and think only in terms of distributions. Thus use M not m.
- In fact, θ and $y_{1:\infty}$ have to match i.e., $y_{1:\infty} \in V_\theta$.
- Thus $\theta_\infty = \theta(Y_{1:\infty})$ makes sense as does $\theta(y_{1:n}, Y_{n+1:\infty}) \overset{m}{\sim} \pi(\theta|y_{1:n})$.

Summary

- This is a timely paper.
- It gives a predictive technique (using future sampling or 'missing' data) to compute a finite n posterior.
- This technique qualitatively changes the Statistician-Scientist dialog by focusing on $m(y_{i+1}|y_{1:i})$. Remains to be done in practice more broadly.
- The intuition changes dramatically when you change the mode from p_θ to M. Central to Bayesian thinking.
- Some differences/convergences have been worked out..But systematically? Common knowledge?
- Predictive techniques are not just for prediction.