Focused Bayesian Prediction

Ruben Loaiza-Maya (Monash University)
David Frazier (Monash University)
Gael Martin (Monash University)

JSM 2019

Motivation

- Goal is to produce a density forecasts of $\{y_t : \Omega \to \mathbb{R} : t \geq 1\}$, defined on $(\Omega, \mathcal{F}, \mathbb{P})$, using a **given** model class

$$\mathcal{P}^t := \left\{P^t_\theta : \theta \in \Theta\right\},\ P^t_\theta(A) := P(A|\theta, \mathcal{F}_t),\ A \subset \Omega,$$

- As accurate as possible in terms of some loss function: $L : \Omega \times \mathcal{P}^t \to \mathbb{R}$.

  1. Most commonly, let L be a **proper scoring rule**
  2. $L(\cdot, \cdot)$ is proper, relative to $\mathbb{P}$, if for $\mathbb{M}(P, Q) := \int_\Omega L(P, y) \mathrm{d}Q(y)$, $\mathbb{M}(Q, Q) \geq \mathbb{M}(P, Q)$, for all $P, Q \in \mathbb{P}$.

- All in a **Bayesian** paradigm...

Standard Bayesian Solution

- The **Bayesian** paradigm expresses uncertainty about

  unknown|known

- **probabilistic**
- For this talk, **unknown** is value of y_{T+1}. **Known** is (data) $\mathbf{y} = (y_1, \ldots, y_T)'$
- Standard Bayesian solution, exact predictive density

$$
\begin{aligned}
p_{exact}(y_{T+1}|\mathcal{F}_T) &= \int p(y_{T+1}, \theta|\mathcal{F}_T)\mathrm{d}\theta \\
&= \int p(y_{T+1}|\theta, \mathcal{F}_T)\pi(\theta|\mathbf{y})\mathrm{d}\theta
\end{aligned}
$$

- $\pi(\theta|\mathbf{y})$: posterior for θ; expression of uncertainty about θ.

A Possible Solution: Approximate Bayesian Forecasting

- **Supremacy of:** $p_{exact}(y_{T+1}|\mathcal{F}_T)$ (over other choices) requires likelihood be correctly specified
- An issue in any empirical setting... Any alternatives?
- Resort to **approximate Bayesian prediction/inference**, by approximating $\pi(\theta|\mathbf{y})$ using Approximate Bayesian Computation (ABC).
- Denote by $\eta(\mathbf{y})$ a vector of summary statistics that “capture” forecasting accuracy, as measured by $L(\cdot,\cdot)$.
- Can then replace exact posterior $\pi(\theta|\mathbf{y})$ by $\pi(\theta|\eta(\mathbf{y}))$ to produce an '**approximate Bayesian predictive**'

$$p_{ABC}(y_{T+1}|\mathcal{F}_T) \quad = \quad \int p(y_{T+1}|\theta, \mathcal{F}_T)\pi(\theta|\eta(\mathbf{y}))d\theta$$

Approximate Bayesian Forecasting

- **However,** generally $p_{ABC}(y_{T+1}|\mathcal{F}_T) \neq p_{exact}(y_{T+1}|\mathcal{F}_T)$.

- Does it matter?

- Tackled in: Frazier, Maneesoonthorn, Martin and McCabe, 2019
  1. Shown that ABF produce **reliable forecasts**
  2. If model is correct: no difference between p_{ABC} and p_{exact} in large samples.
  3. In small samples, difference is negligible, but p_{ABC} requires much simpler computations

- Another alternative: Cooper, Frazier, Koo & Martin, 2019: Variational Bayes approximation to $\pi(\theta|\mathbf{y}) \Rightarrow$ Approximate $p(y_{T+1}|\mathcal{F}_T)$

- See also: Park & Nassar, 2014; Koop & Korobilis; 2018; Quiroz, Nott, & Kohn, 2018

Approximate Bayesian Forecasting

- While ABC does not require us to get $\pi(\theta|\mathbf{y})$ "correct"
- $\pi(\theta|\eta(\mathbf{y}))$ can display concerning behavior when **DGP wrong.**
  1. Frazier, Robert and Rousseau (2019): ABC inference can pay a heavy price in terms of reliability. Worse the more complex the ABC approach.
  2. Different summaries will lead to very different predictions.
  3. How do you match summaries with a chosen loss?
- Does not engender confidence in ABF under model misspecification...

The Object of our Faith?

- ABF based on approximation to

$$p_{exact}(y_{T+1}|\mathcal{F}_T) = \int p(y_{T+1}|\theta, \mathcal{F}_T)\pi(\theta|\mathbf{y})\mathrm{d}\theta,$$

- by approximating $\pi(\theta|\mathbf{y})$
- Model misspecification impinges on $p_{exact}(y_{T+1}|\mathcal{F}_T)$ via two avenues: (1) the posterior $\pi(\theta|\mathbf{y})$; (2) The **conditional** predictive: $p(y_{T+1}|\mathcal{F}_T, \theta)$
- In what sense does $p_{exact}(y_{T+1}|\mathcal{F}_T)$ remain the gold standard?
- Really, want a predictive that it pushed towards optimality in the loss function $L(\cdot, \cdot)$...

A New Paradigm for Bayesian Prediction

- **Goal: Produce accurate density forecasts, in terms of $L(\cdot,\cdot)$,** when **true DGP is unknown**
- **How to do this** given a class of **conditional predictive** , say $\mathcal{P}$, that we believe **could** have generated the data,
- with elements

$$p(y_{T+1}|\mathcal{F}_T,\cdot) \in \mathcal{P}$$

- First, define a prior measure over the elements of $\mathcal{P}$: $\Pi[\cdot]$
- In principle, $\mathcal{P}$ may be a class of:
  - distributions, $p(y_{T+1}|\mathcal{F}_T,\theta)$ say, associated with a **given parametric** model
  - weighted combinations of predictives associated with **different parametric** models
  - **non-parametric** conditional distributions

Focused Bayesian Prediction (FBP)

- The **essence** of the idea
- Update the **prior**:

$$\pi[p(y_{T+1}|\mathcal{F}_T,\cdot)],\ p(y_{T+1}|\mathcal{F}_T,\cdot)\in\mathcal{P}$$

  to a **posterior**:

$$\pi[p(y_{T+1}|\mathcal{F}_T,\cdot)|\mathbf{y}],\ p(y_{T+1}|\mathcal{F}_T,\cdot)\in\mathcal{P}$$

- According to **predictive performance**
- $\Rightarrow \pi[p|\mathbf{y}]$ is '**focused' on** elements of $\mathcal{P}^t$ with **high predictive accuracy**
- Different measures of **accuracy** $\Rightarrow$ different **posteriors**

Focused Bayesian Prediction (FBP)

- Let $L : \Omega \times \mathcal{P} \to \mathbb{R}$ **be a proper (positively-oriented) scoring rule**

$$(p, y) \mapsto L(p, y)$$

- with expected score, under the **truth**, $F(y_{T+1}|\mathbf{y})$, given by

$$\mathbb{M}(P, F) := \mathbb{E}_F\left[L(p, y_{T+1})\right]$$

- Using short-hand:

$$\begin{aligned} p &= p(y_{T+1}|\mathcal{F}_T, \cdot) \\ L &= L(p(y_{T+1}|\mathcal{F}_T, \cdot), y_{T+1}) \end{aligned}$$

- And defining a sample estimate of $\mathbb{M}(\cdot, F)$

$$\widehat{L}_T(p)/T = \sum_{t=0}^{T-1} L(p, y_{t+1})/T$$

Focused Bayesian Prediction (FBP)

- We extend the **inferential** work of **Bissiri *et al.* (2016)**:
- "*A general framework for updating belief distributions*"
- to the **prediction** setting
- $\Rightarrow$ defining a **coherent** Bayesian up-date as:

$$\pi_w[p|\mathbf{y}] \propto \exp[w\widehat{L}_T(p)] \times \pi[p]$$

- $\Longrightarrow$, all else equal, elements $p \in \mathcal{P}$ that yield better in-sample prediction, has higher posterior probability.

Focused Bayesian Prediction (FBP)

- Given:

$$\pi_w[p|\mathbf{y}] \quad \propto \quad \exp[w\widehat{L}_T(p)] \times \pi[p]$$

- w determines the "**weight**" of $\exp[w\widehat{L}_T(p)]$ relative to $\pi[p]$
- Which (in turn) determines the **nature** of $\pi_w[p|\mathbf{y}]$
- But what is a reasonable w?

A w, a w, my kingdom for a w?

- Asymptotically, (for reasonable w) it doesn't really matter for prediction.
- Define the following predictive measures

$$P_w^t(B) = \int_\Theta \mathrm{d}P_\theta^t(B)\mathrm{d}\Pi_w[\theta|\mathbf{y}], \tag{1}$$

$$P_*^t(B) = \int_\Theta P_\theta^t(B)\mathrm{d}\delta_{\theta_*}, \tag{2}$$

where δ_{θ_*} - Dirac measure at $\theta = \theta_*$, and $\theta_* = \arg\max \operatorname{plim} L_T(p)/T$,

Proposition

Under regularity conditions, if $\lim_n w_n = C_w > 0$*, then,*

$$\sup_{B \in \mathcal{F}} |P_w^t(B) - P_*^t(B)| = o_p(1).$$

However, “ Machines never come with any extra parts”...

- “They always come with the exact amount they need...[it] had to be here for some reason.”

- If $L(\cdot,\cdot)$ has a different scale than $\pi[\cdot]$ (e.g., CRPS scores), w is critical for good sampling. Bad $w \Rightarrow$ MCMC chain sticks, poor acceptance rates... bad things happen.

- Several options under exploration:
  1. Currently we choose w to ensure that

$$\mathrm{Tr}\{Var_{\Pi[p|\mathbf{y}]}[\theta]\} \quad = \quad \mathrm{Tr}\{Var_{\Pi[\theta|\mathbf{y}]}[\theta]\}$$

  2. Other options being explored
     - **Holmes & Walker, 2017; Lyddon, Holmes & Walker, 2019**
     - impose a ‘**sandwich-type**’ **var-cov matrix** $\Rightarrow$ mimics the approach to misspecification in likelihood-based settings.
  3. Choose w using k-fold, or time-series, cross validation.

Example: "Indecision is the key to flexibility"

- And predictive accuracy...
- Flexibility to **define** $\mathcal{P}^t$ such that the elements of the class are **weighted combinations** of predictives:

$$p(y_{T+1}|\mathcal{F}_T, \cdot) = \frac{1}{K}\sum_{k=1}^{K} \theta_k p(y_{T+1}|\mathcal{F}_T, M_k)$$

- Taking the constituent $p(y_{T+1}|\mathcal{F}_T, M_k)$ as 'given' $\Rightarrow$
- $p(y_{T+1}|\mathcal{F}_T, \cdot)$ characterized only by the unknown $\theta_k \in (0, 1)$
- $\Rightarrow \pi_w[p|\mathbf{y}]$ produced via **predictive accuracy-based** up-dating
- **without** having to assume that the **true model** lies in the set
- $\Rightarrow \mathcal{M}$-open view of the world (**Bernardo & Smith, 1994**)

Mixtures of Predictives

- Places inside a **formal, coherent** **Bayesian up-dating scheme**
- the practice of using **weighted combinations** of predictives **via predictive criteria**
- Yields a posterior distributions over the different predictives (defined by the θ_k)

Mixtures of Predictives

- A large **frequentist** literature produces **point estimates** of the θ_k based on predictive performance, e.g:
    - **Hall & Mitchell, 2007; Geweke & Amisano, 2011; Gneiting & Ranjan, 2011, 2013; Kapetanios et al. 2015; Claeskens et al., 2016; Ganics, 2017; Opschoor et al., 2017; Aastveit et al., 2018; Post et al., 2019; Pauwels et al., 2019**
- The **Bayesian** literature is still small:
    - **Billio et al., 2013; Pettenuzzo & Ravazzolo, 2016; Casarin et al., 2019**
    - FBP formalizes this from a Bayesian standpoint.

Illustration: Simulated data

- Paper contains results for simulated & empirical data
- Will focus on one set of (simulated) results here
- **True DGP** for a financial return (y_t)

$$\begin{aligned} z_t &= \exp(h_t/2)\varepsilon_t \\ h_t &= \alpha + \beta h_{t-1} + \sigma_h \eta_t \\ y_t &= G^{-1}(F_z(z_t)) \end{aligned}$$

- $\Rightarrow$ Implied copula of a **stochastic volatility** model combined with a **skewed normal marginal,** $g(y_t)$ (imposed through G^{-1})

Three predictive classes

- $p(y_{T+1}|\mathcal{F}_T,\cdot) \in \mathcal{P}$ defined as **combination of predictives** based on
    1. $p(y_{T+1}|\mathcal{F}_T, M_1)$ - ARCH(1) (skewed normal errors)
    2. $p(y_{T+1}|\mathcal{F}_T, M_2)$ - GARCH(1,1) (normal errors)

- $\Rightarrow$ **combination of predictives** based on

$$p(y_{T+1}|\mathcal{F}_T,\theta) = \sum_{k=1}^{2} \theta_k p(y_{T+1}|\mathcal{F}_T, M_k)$$

Forms of updates

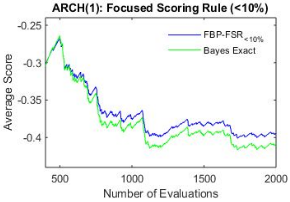
- **Two types of score** used here:
    1. Log score ($\Rightarrow$ exact Bayes)
    2. Focused log score (rewards predictive accuracy in a tail)

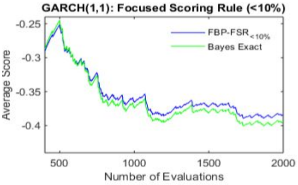
- Estimate: $E[p|\mathbf{y}] = \int_{\mathcal{P}} p d\Pi [p|\mathbf{y}]$ as: $\frac{1}{M} \sum_{i=1}^{M} p^{(i)}$

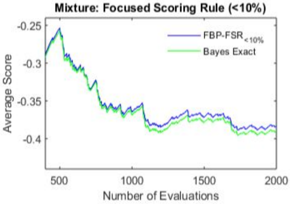
- Roll the whole process forward (with expanding windows)

- **Assess predictive performance** according to **Focused log score**

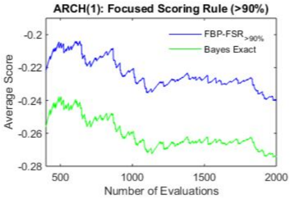
Out-of-Sample Performance: LOWER 10% tail

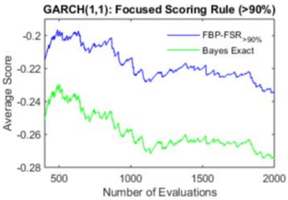


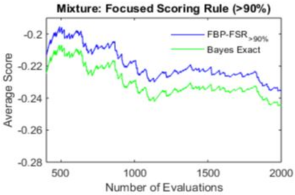




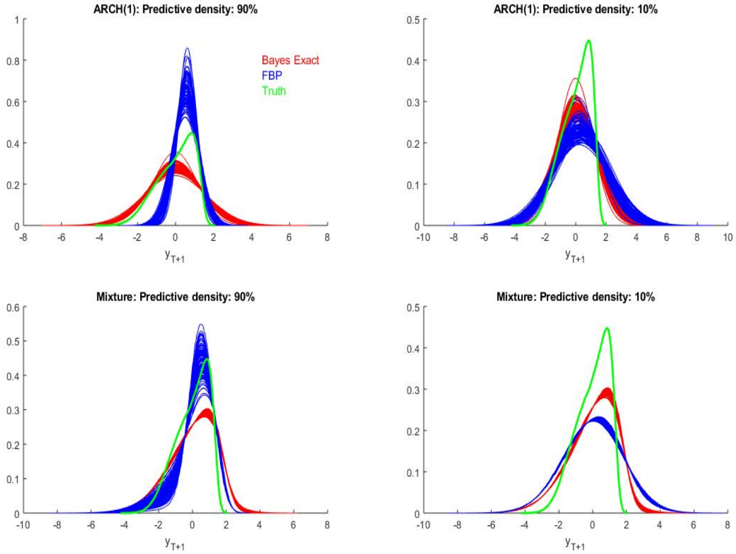
Out-of-Sample Performance: UPPER 10% tail







Predictive Variability: Posterior



Makes sense

- Focusing on a specific characteristic of the data
- $\Rightarrow$ getting the model wrong matters less
- Is handy!
- A crude, computationally simple, predictive class (like ARCH(1)) does the job
- No need for the predictive combination
  (to better capture the true DGP)
- Not the aim!
- Aim is (only) to accurately predict extreme observations
- Aim achieved by using an appropriate up-dating rule

Moving Forward...

- Note: Can also use full draw from $\Pi\left[p|\mathbf{y}\right]$
  - $\Rightarrow$ plus credible bounds on p
  - Alternative definitions of optimality for predictive densities?
- Large dimensional predictives (fully non-parametric classes)...
- Approximate version?
- ...lots to play with!