Model-Acentric, Focused Bayesian Prediction

David Frazier, Ruben Loaiza Maya and Gael Martin

Department of Econometrics and Business Statistics

Monash University, Melbourne

BIRS workshop, Oaxaca, Nov. 2018

Note that this is a modified version of the talk given

Bayesian Prediction

- Distribution of interest is:

$$
\begin{aligned}
p(y_{T+1}|y_{1:T}) &= \int_{\boldsymbol{\theta}} p(y_{T+1}, \boldsymbol{\theta}|y_{1:T})\, d\boldsymbol{\theta} \\
&= \int_{\theta} p(y_{T+1}|y_{1:T}, \boldsymbol{\theta})\, p(\boldsymbol{\theta}|y_{1:T})\, d\boldsymbol{\theta} \\
&= E_{\boldsymbol{\theta}|\mathbf{y}}\left[p(y_{T+1}|y_{1:T}, \boldsymbol{\theta})\right]
\end{aligned}
$$

- **(Marginal)** predictive = expect. of **conditional** predictive
- **Conditional** predictive reflects the **assumed DGP**
- As does the **posterior**: $p(\boldsymbol{\theta}|y_{1:T}) \propto p(y_{1:T}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})$

Implementing Bayesian Prediction

- In the usual case where $E_{\boldsymbol{\theta}|y_{1:T}}\left[p(y_{T+1}|y_{1:T},\boldsymbol{\theta})\right]$ cannot be evaluated **analytically**
- Take M draws from $p(\boldsymbol{\theta}|y_{1:T})$ (via a Markov chain Monte Carlo algorithm, say)
- And **estimate** $p(y_{T+1}|y_{1:T})$ as
  1. either:
  $$\widehat{p}(y_{T+1}|\mathbf{y}_{1:T}) = \frac{1}{M}\sum_{i-1}^{M} p(y_{T+1}|y_{1:T},\boldsymbol{\theta}^{(i)})$$
  2. or: $\widehat{p}(y_{T+1}|y_{1:T})$ constructed from draws of $y_{T+1}^{(i)}$ simulated from $p(y_{T+1}|y_{1:T},\boldsymbol{\theta}^{(i)})$
- i.e. MCMC $\Rightarrow$ **exact Bayesian prediction**
  - (up to simulation error)

Achilles Heels!

1. What happens when we can't generate an MCMC chain because $p(\boldsymbol{\theta}|y_{1:T})$ is inaccessible?
   - $\Rightarrow$ **exact** Bayesian prediction not feasible
   - **Frazier, Maneesoonthorn, Martin and McCabe: "Approximate Bayesian Forecasting", IJF, 2018**
2. What happens when we acknowledge that the **DGP** used to construct $p(y_{T+1}|y_{1:T})$ **misspecified**?
   - This impinges on $p(y_{T+1}|y_{1:T})$ via its two components:

$$p(y_{T+1}|y_{1:T}) = \int_{\theta} p(y_{T+1}|y_{1:T}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|y_{1:T}) d\boldsymbol{\theta} \text{ and}$$

   - The **conditional** predictive: $p(y_{T+1}|y_{1:T}, \boldsymbol{\theta})$
   - and $p(\boldsymbol{\theta}|y_{1:T}) \propto p(y_{1:T}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})$
   - In what sense does $p(y_{T+1}|y_{1:T})$ remain the gold standard?

A New Paradigm for Bayesian Prediction

- Appropriate for the realistic setting in which the **true DGP is unknown**
- **The ideas are still evolving!**
- Define $\mathcal{P}$ as the class of **conditional predictives** that we believe **could** have generated the data
- With elements:

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

- where $P(y_{T+1}|y_{1:T}, \cdot)$ conditions on data: $y_{1:T}$, and on some unknowns

A New Paradigm for Bayesian Prediction

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

Focused Bayesian Prediction (FBP)

- Update the **prior**:

$$\Pi[P(y_{T+1}|y_{1:T},\cdot)]$$

  to a **posterior**:

$$\Pi[P(y_{T+1}|y_{1:T},\cdot)|y_{1:T}]$$

- According to **predictive performance** over some 'test' set, $\mathcal{T}$
- $\Rightarrow \Pi[P(y_{T+1}|y_{1:T},\cdot)|y_{1:T}]$ is '**focused' on** elements of $\mathcal{P}$ with **high predictive accuracy** $\Leftrightarrow$ **small loss**
- Different (problem-specific) measures of **loss** $\Rightarrow$ different **posteriors**

Focused Bayesian Prediction (FBP)

- First attempt....
- Define **a proper scoring rule:** $S(P(y_{T+1}|y_{1:T},\cdot), y_{T+1})$
- with expectation, under the **truth,** $F(y_{T+1}|y_{1:T})$, as:

$$\mathcal{S}(P,F) = \mathbb{E}_F\left[S(P(y_{T+1}|y_{1:T},\cdot), y_{T+1})\right]$$

- The map $P \mapsto -\mathcal{S}(P,F)$ defines a **loss function** over the models in $\mathcal{P}$
- Aim is to **focus** on the elements of $\mathcal{P}$ that **minimize this loss**

Focused Bayesian Prediction (FBP)

- Partition the sample: $y_1, y_2, ..., y_T$ into:
  - A **training** set: $\mathcal{D} = \{y_t;\, 1 \leq t \leq \tau\}$
  - A **test** set: $\mathcal{T} = \{y_t;\, \tau + 1 \leq t \leq \tau + n = T\}$
- **Fit** P on $\mathcal{D} \Rightarrow \widehat{P}(y_{t+1}|y_{1:t}, \cdot)$ (when necessary)
- Use $\mathcal{T}$ (and **expanding** $\mathcal{D}$) to **compute**:

$$S_n(P, F) = \frac{1}{n} \sum_{i=0}^{n-1} S(\widehat{P}(y_{(\tau+i)+1}|y_{1:(\tau+i)}, \cdot), y_{(\tau+i)+1})$$

- as an estimate of $\mathcal{S}(P, F)$

Focused Bayesian Prediction (FBP)

- Using short-hand:

$$P = P(y_{T+1}|y_{1:T}, \cdot) \in \mathcal{P}; \; F = F(y_{T+1}|y_{1:T});$$

- **Simplest form of FBP Algorithm:**

1.	Draw P^i from $\Pi[P]$,	$i = 1, 2, ..., N$
2.	Compute $\widehat{P}^i$	using $\mathcal{D}$ and P^i
3.	Compute $s = S_n(\widehat{P}^i, F)$	over test set $\mathcal{T}$
3.	For each $i = 1, 2, ..., N$	accept $\widehat{P}^i$ if $s \geq \varepsilon_n$

- Different choices for $\varepsilon_n \Rightarrow$ different **aversion to loss**

Focused Bayesian Prediction

- This **likelihood-free algorithm** produces *i.i.d.* draws from a **'posterior'** for P, **given** $y_{1:T} : \Pi_{\varepsilon_n}[P|y_{1:T}]$
- where the replacement of a **likelihood** function with an alternative **loss** function
- And - hence - the use of '**posterior**'
- Is similar in spirit to **Bissiri et al. (JRSS(B), 2016)**:
- "*A general framework for updating belief distributions*"
- But applied to **prediction** rather than **inference**

Focused Bayesian Prediction

- Further refinements certainly possible
- E.g. via addition of an **approximate Bayesian computation (ABC)** step
- $\Rightarrow$ draws P^i (s.t. $s \geq \varepsilon_n$) are **weighted** according to their ability to produce **simulated** values (z_{T+1}) that '**match**' the **observed** values (y_{T+1}) in test period
- according the given score (or loss)
- or, maybe, according to an **additional score** (or loss)

Preliminary Theoretical Results

- **Theorem 1: 'Posterior' Concentration:**
- Define:

$$P^* = \arg\max_{P \in \mathcal{P}} \mathcal{S}(P, F) \text{ with } \varepsilon^* = \mathcal{S}(P^*, F)$$

- For $\varepsilon_n \to \varepsilon^*$; $\delta_n \to 0$ (and under other conditions):

$$\Pi_{\varepsilon_n}[|\mathcal{S}(P, F) - \mathcal{S}(P^*, F)| > \delta_n | y_{1:T}] \underset{n\to\infty}{\longrightarrow} 0$$

- $\Rightarrow$ **distribution** of the expected score of $P \in \mathcal{P}$ **concentrates onto** the **maximum** expected score possible under F

Preliminary Theoretical Results

- ‘**Posterior**’ **concentration** (in terms of P) *would* then be defined as:

$$\Pi_{\varepsilon_n}[\rho(P, P^*) > \delta_n | y_{1:T}] \underset{n\to\infty}{\to} 0$$

- For some functional metric, ρ, (like total variation)
- $\Rightarrow$ **posterior** of P **concentrates onto** element of $\mathcal{P}$ that:
- **maximizes the expected score** $\Leftrightarrow$ **minimizes loss** in $\mathcal{P}$
- Proof on the drawing board.....

Preliminary Theoretical Results

- So the distribution of $\mathcal{S}(P, F)$ concentrates onto $\mathcal{S}(P^*, F)$
- ($\Rightarrow$ 'loosely speaking' that P concentrates onto P^*)
- with P^* determined by the choice of score (or loss) function, the choice of $\mathcal{P}$, and by the **true** F
- How does the 'posterior' of P relate to the true F?
- Define:

$$\begin{aligned} E_{\varepsilon_n}[P|y_{1:T}] &= \int_{\mathcal{P}} P d\Pi_{\varepsilon_n}[P|y_{1:T}] \\ &= \text{the 'posterior' mean of } P \end{aligned}$$

Preliminary Theoretical Results

- **Theorem 2: Predictive Merging.** As $n \to \infty$ and $\varepsilon_n \to \varepsilon^*$

(a) If $F \in \mathcal{P}$ (i.e. when the **true predictive** is in the class) we **do recover it**:

$$\rho^2_{TV}\left(E_{\varepsilon_n}[P|y_{1:T}], F\right) \underset{n\to\infty}{\longrightarrow} 0$$

- i.e. (squared) total variation distance of $E_{\varepsilon_n}[P|y_{1:T}]$ from the true predictive $\to 0$

Preliminary Theoretical Results

- **Theorem 2: Predictive Merging.** As $n \to \infty$ and $\varepsilon_n \to \varepsilon^*$

(b) If $F \notin \mathcal{P}$ (so under **mis-specification**):

$$\lim_{n\to\infty} \rho_{TV}\left(E_{\varepsilon_n}[P|y_{1:T}], F\right) \leq 2\rho_{Hellinger}\left(P^*, F\right)$$

- $P^* =$ predictive distribution that **maximizes the expected score** $\Leftrightarrow$ **is closest to** F in this sense
- $\Rightarrow$ the bound is the (H) distance between F and the P^* that is closest to F in this score
- **Actual magnitude** of the bound is (of course) affected by $\mathcal{P}$ and the chosen score (or loss)

Illustrative Example 1: Financial Asset Return

- Let $\ln S_t$ = log of an asset price
- Let $\mathcal{P}$ define a class of **parametric predictives**, $P_{\boldsymbol{\theta}}$, associated with a **stochastic volatility** model

$$d\ln S_t = \sqrt{V_t}dB_t^S$$
$$dV_t = (\theta_1 - \theta_2 V_t)\,dt + \theta_3\sqrt{V_t}dB_t^v$$

- with $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)'$
- The **true DGP**, F, is a stochastic volatility model with random **jumps**:

$$d\ln S_t = \sqrt{V_t}dB_t^S + \underbrace{Z_t dN_t}_{=g(\theta_{0,4},\theta_{0,5}\dots)}$$
$$dV_t = (\theta_{0,1} - \theta_{0,2} V_t)\,dt + \theta_{0,3}\sqrt{V_t}dB_t^v$$

  - $\boldsymbol{\theta}_0 = (\theta_{0,1}, \theta_{0,2}, \theta_{0,3}, \dots)'$ = **true parameter** (vector)

Exact but mis-specified predictive?

- If we **were** to simply adopt the (implied) **mis-specified SV** model for

$$y_t = \ln S_t - \ln S_{t-1} = \textbf{return} \text{ at time } t$$

- and produce the conventional exact Bayesian predictive: $p(y_{T+1}|y_{1:T})$
- What would we find?
- $p(\boldsymbol{\theta}|y_{1:T})$ (under regul.) concentrates onto **pseudo-true** $\boldsymbol{\theta}$, $\boldsymbol{\theta}^*$
- where $\boldsymbol{\theta}^*$ is close to $\boldsymbol{\theta}_0$ (in KL-based sense)
- $\Rightarrow$

$$\lim_{T\to\infty} p(y_{T+1}|y_{1:T}) = p(y_{T+1}|y_{1:T}, \boldsymbol{\theta}^*) = \textit{what??}$$

Exact but mis-specified predictive?

- P is misspecified
- $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}_0$
- **Minimizing** KL divergence $\equiv$ **maximizing log score** *in sample*
- **No guarantee** of *out-of-sample* performance
- In particular, with respect to some other score/loss
- **FBF ensures (in principle)** accurate *out-of-sample* performance according to **any given score/loss**

Focused Bayesian Prediction

- **Five** loss functions considered:
  - **Three scores:**
    1. Log score
    2. Continuous rank probability score (CRPS)
    3. CRPS for lower tail (appropriate for a financial return)
  - **Two 'auxiliary predictive'**-based losses
  - Adopting the flavour of **auxiliary model-based** ABC
  - **Drovandi et al. (2011, 2015, 2018); Creel and Kristensen (2015); Drovandi (2018); Martin, McCabe, Frazier, Maneesoonthorn and Robert (2018)**

Auxiliary predictive-based loss function

- What do we know about **prediction**?
- **Simple parsimoneous** models often forecast better than **complex, highly parameterized (but incorrect)** models....
- $\Rightarrow$ Pick a **simple parsimoneous 'auxiliary predictive':** $q(y_{T+1}|y_{1:T}, \boldsymbol{\beta})$
- And **select** $p(y_{T+1}|y_{1:T}, \boldsymbol{\theta}^i)$ (from $\mathcal{P}$) such that their predictive performance closely **matches** that of $q(y_{T+1}|y_{1:T}, \boldsymbol{\beta})$ over the test period

Auxiliary predictive-based loss function

- i.e. **select** $p(y_{T+1}|y_{1:T}, \boldsymbol{\theta}^i)$ such that:

$$\frac{1}{n}\sum_{i=0}^{n-1}\left|p(y_{(\tau+i)+1}|y_{1:(\tau+i)}, \boldsymbol{\theta}^i) - q(y_{(\tau+i)+1}|y_{1:(\tau+i)}, \widehat{\boldsymbol{\beta}})\right|$$

$<$ the **lowest** ($\alpha\%$, say) quantile

- i.e. such that **loss** (defined by this predictive difference) is **small**
- Choose $q(y_{T+1}|y_{1:T}, \boldsymbol{\beta})$ to be a **generalized autoregressive conditionally heteroscedastic (GARCH)** model
  - with Student t errors (work-horse of empirical finance)
  - with normal errors (expected to be a poorer 'benchmark')

Numerical results

- For each of the 5 posteriors:
- Estimate: $E_{\varepsilon_n}[P|y_{1:T}] = \int_{\mathcal{P}} P d\Pi_{\varepsilon_n}[P|y_{1:T}]$
- by taking the sample average of the selected P
- Roll the whole process forward (with expanding T)
- Compute, over 200 (truely) out-of-sample periods:
- Median:
  - **log scores; CRPS scores; tail-weighted CRPS scores**
- Compare with results for **exact (MCMC) mis-specified:** $p(y_{T+1}|y_{1:T})$

Numerical results

- The loss function based on matching the Student t GARCH (auxiliary) predictive **yields the most accurate predictive** - according to all measures of predictive accuracy
- The loss function based on the (raw) CRPS score is **second best** - according to all measures of predictive accuracy
- The loss function based on matching the normal GARCH (auxiliary) predictive does not - as anticipated - perform well
- The **exact but mis-specified** predictive is **beaten by FBP** in all cases.....
- So we *are* gaining in terms of predictive accuracy via **FBP**
- Numerical results influenced (however) by simulation error (in **MCMC** and the **particle filtering** used to produce $\widehat{P}$)

Illustrative Example 2: No Simulation Error

- **True** model (F): Gaussian $AR(4)$ with **stochastic** volatility
- **Predictive class** ($P_{\boldsymbol{\theta}} \in \mathcal{P}$): Gaussian $AR(1)$ with **constant** volatility
- Exact (misspecified) $p(y_{T+1}|y_{1:T})$ has **closed-form**
- As does $P_{\boldsymbol{\theta}}$
- $\Rightarrow$ has enabled large values for:
  - Draws from $\Pi\left[P\right]$ (50,000)
  - Test period, $n\,(5000+)$
  - Out-of-sample evaluations (5000)
- **Very clear (and significant) ranking of CRPS-based FBP over exact (mis-specified) Bayes**
- According to (the mean of) all three out-of-sample scores

Probability Integral Transform (PIT)

- Defining the **cumulative predictive distribution** evaluated at (observed) y^o_{T+1} as:

$$u_{T+1} = \int_{-\infty}^{y^o_{T+1}} p\left(y_{T+1} | y_{1:T}\right) dy_{T+1}$$

- for exact (mis-specified) $p\left(y_{T+1} | y_{1:T}\right)$
- Under H_0 : "$p\left(y_{T+1} | y_{1:T}\right)$ matches the true F" :
- $$u^i_{T+1},\ \ i = 1, 2, ..., 5000,\ \ \text{are}\ i.i.d.U\left(0, 1\right)$$

- H_0 **rejected** for **exact Bayes**
- H_0 **rejected** for **LS-based FBP**
- H_0 **not rejected** for **CRPS-based FBP**
- Early days....more theoretical and numerical results to come......