Focused Bayesian Prediction

Ruben Loaiza-Maya, Gael Martin and David Frazier

Department of Econometrics and Business Statistics

Monash University

One World ABC Seminar, May, 2020

Paper available at: arXiv:1912.12571

Note: there are two sequential sets of slides which have been merged into one .pdf document

References are listed at the end

Bayesian Prediction

- Distribution of interest is:

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

- **(Marginal)** predictive $= E_{\boldsymbol{\theta}|\mathbf{y}} \left[p(y_{n+1}|\mathbf{y}, \boldsymbol{\theta}) \right]$
- **Conditional** predictive reflects the **assumed DGP**
- as does $p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})$

Implementing Bayesian Prediction

- $E_{\boldsymbol{\theta}|\mathbf{y}}[\cdot]$ typically can't be evaluated **analytically** $\Rightarrow$
- Take M draws from $p(\boldsymbol{\theta}|\mathbf{y})$
- via Markov chain Monte Carlo (**MCMC**) say
- And **estimate** $p(y_{n+1}|\mathbf{y})$ as
  1. either:
$$\widehat{p}(y_{n+1}|\mathbf{y}) = \frac{1}{M}\sum_{i-1}^{M} p(y_{n+1}|\mathbf{y}, \boldsymbol{\theta}^{(i)})$$
  2. or: $\widehat{p}(y_{n+1}|\mathbf{y})$ constructed from draws of $y_{n+1}^{(i)}$ simulated from $p(y_{n+1}|\mathbf{y}, \boldsymbol{\theta}^{(i)})$
- 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 $p(\boldsymbol{\theta}|\mathbf{y})$ is 'inaccessible'?

   - DGP $p(\mathbf{y}|\boldsymbol{\theta})$ **unavailable** in closed form
   - **Dimension** of $\mathbf{y}$ too large for evaluation of $p(\mathbf{y}|\boldsymbol{\theta})$ (inside **MCMC**)
   - **Dimension of** $\boldsymbol{\theta}$ too large for efficient exploration of $p(\boldsymbol{\theta}|\mathbf{y})$
   - $\Rightarrow$ exact Bayesian prediction **not feasible**

Achilles Heels!

- Tackled by:
  - **Canale & Ruggiero (2016)**
  - **Frazier, Maneesoonthorn, Martin & McCabe (2019)**
  - **Kon Kam King, Canale & Ruggiero (2019):**
- Using an **ABC** approximation to $p(\boldsymbol{\theta}|\mathbf{y}) : p_{ABC}(\boldsymbol{\theta}|\mathbf{y})$
- $\Rightarrow$ **approximation to** $p(y_{n+1}|\mathbf{y}) : p_{ABC}(y_{n+1}|\mathbf{y})$

Achilles Heels!

- And by:
  - **Park & Nassar (2014)**
  - **Cooper, Frazier, Koo & Martin (2019)**
  - **Koop & Korobilis (2018)**
  - **Quiroz, Nott & Kohn (2020)**
- Using a **variational Bayes** approximation to $p(\boldsymbol{\theta}|\mathbf{y}) : p_{VB}(\boldsymbol{\theta}|\mathbf{y})$
- $\Rightarrow$ **approximation to** $p(y_{n+1}|\mathbf{y}) : p_{VB}(y_{n+1}|\mathbf{y})$

Achilles Heels!

- *One message from both strands of work*:
- As long as **assumed DGP** is **correctly specified**
- (i.e. tallies with the **true DGP**)
- **Crude** approximations to $p(\boldsymbol{\theta}|\mathbf{y}) \Rightarrow$
- **Accurate** approximations to $p(y_{n+1}|\mathbf{y})$

Achilles Heels!

2. What happens when we acknowledge that any **assumed DGP** is **misspecified**?

- Will **any** $p_{approx}(y_{n+1}|\mathbf{y}) \approx p(y_{n+1}|\mathbf{y})$
- when the **predictive model** is wrong?
- *Much more critically*:
- In what sense does:

$$p(y_{n+1}|\mathbf{y}) = \int_{\theta} p(y_{n+1}|\mathbf{y}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \text{ and}$$

- (where **misspecification** impinges on both components)
- remain the gold standard?

A New Paradigm for Bayesian Prediction

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

$$p(y_{n+1}|\mathbf{y}, \cdot) \in \mathcal{P}$$

- where $p(y_{n+1}|\mathbf{y}, \cdot)$ conditions on data: $\mathbf{y}$, and on some unknowns

A New Paradigm for Bayesian Prediction

- In principle, $\mathcal{P}$ may be a class of:
  - distributions, $p(y_{n+1}|\mathbf{y}, \boldsymbol{\theta})$ 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 $\mathcal{P}$: $\Pi[p(y_{n+1}|\mathbf{y}, \cdot)]$
- The **essence** of the idea:

Focused Bayesian Prediction (FBP)

- Update the **prior**:

$$\Pi[p(y_{n+1}|\mathbf{y}, \cdot)]$$

  to a **posterior**:

$$\Pi[p(y_{n+1}|\mathbf{y}, \cdot)|\mathbf{y}]$$

- According to **predictive performance**
- $\Rightarrow \Pi[p(y_{n+1}|\mathbf{y}, \cdot)|\mathbf{y}]$ is '**focused' on** elements of $\mathcal{P}$ with **high predictive accuracy**
- Different measures of **accuracy** $\Rightarrow$ different **posteriors**
- Different methods of **up-dating** $\Rightarrow$ different **posteriors**

Focused Bayesian Prediction (FBP)

- Define **a proper (positively-oriented) scoring rule:**

$$S(p(y_{t+1}|\mathbf{y},\cdot),y_{t+1})$$

- with expectation, under the **truth**, $f(y_{t+1}|\mathbf{y})$, as:

$$\mathbb{E}_f\left[S(p(y_{t+1}|\mathbf{y},\cdot),y_{t+1})\right]$$

- Using short-hand:

$$\begin{aligned} p_{t+1} &= p(y_{t+1}|\mathbf{y},\cdot) \\ p_{n+1} &= p(y_{n+1}|\mathbf{y},\cdot) \end{aligned}$$

- We define the sum of scores over the sample of size n :

$$S(\mathbf{y}) = \sum_{t=0}^{n-1} S(p_{t+1},y_{t+1})$$

Focused Bayesian Prediction (FBP)

- Played around a bit with the **up-dating** step
- Including the use of **ABC** principles
- In the end, we've kept things simple
- $\Rightarrow$ using a **coherent** exponential up-date:

$$\Pi[p_{n+1}|\mathbf{y}] \quad \propto \quad \exp[w_n S(\mathbf{y})] \times \Pi[p_{n+1}]$$

Focused Bayesian Prediction (FBP)

- Takes the spirit of **loss-based Bayesian inference/general Bayesian inference/Gibbs posteriors/PAC-Bayes**:
  - **Jiang and Tanner (2008)**, **Bissiri et al. (2016)**, **Giummole et al. (2017)**, **Holmes & Walker (2017)**, **Guedj (2019)**, **Knoblauch et al. (2019)**, **Lyddon et al. (2019)**
- $\Rightarrow$ into a **prediction** setting
- where 'coherent' $\Rightarrow$

$$\begin{aligned} \Pi[p_{n+1}|\mathbf{y}_1, \mathbf{y}_2] &\propto \exp[w_n S(\mathbf{y}_1, \mathbf{y}_2)] \times \Pi[p_{n+1}] \\ &\equiv \exp[w_n S(\mathbf{y}_2)] \times \Pi[p_{n+1}|\mathbf{y}_1] \end{aligned}$$

Focused Bayesian Prediction (FBP)

- When we choose **log score**:

$$S = \log p(y_{t+1}|\mathbf{y}, \boldsymbol{\theta})$$

- for some **parametric model** $p(y_{t+1}|\mathbf{y}, \boldsymbol{\theta})$, and set $w_n = 1$
- we recover the conventional (potentially **mis-specified**)
- **likelihood-based** Bayesian up-date for $\boldsymbol{\theta}$:

$$\Pi[\boldsymbol{\theta}|\mathbf{y}] \propto p[\mathbf{y}|\boldsymbol{\theta}] \times \Pi[\boldsymbol{\theta}]$$

- (which $\Rightarrow$ a posterior for p: $\Pi[p_{n+1}(\boldsymbol{\theta})|\mathbf{y}]$)
- **Other** choices of $S \Rightarrow$ up-dates
- driven by **other measures of predictive performance**

Focused Bayesian Prediction (FBP)

- Given:

$$\Pi[p_{n+1}|\mathbf{y}] \quad \propto \quad \exp[w_n S(\mathbf{y})] \times \Pi[p_{n+1}]$$

- w_n determines the **weight** of $\exp[w_n S(\mathbf{y})]$ relative to $\Pi[p_{n+1}]$
- Which (in turn) determines the **nature** of $\Pi[p_{n+1}|\mathbf{y}]$
- Including its **variance**
- Need $w_n \Rightarrow$ a sensible $\Pi[p_{n+1}|\mathbf{y}]$
- No one way of doing that
- **Asymptotically** (as $n \longrightarrow \infty$) choice of w_n matters little

Focused Bayesian Prediction (FBP)

- **Proposition:** Under regularity, if $\lim_n w_n = C_w > 0$
- As $n \longrightarrow \infty$
- $\Pi_w[p_{n+1}|\mathbf{y}]$ **concentrates** onto

$$p^*_{n+1} = \arg \max_{p_{n+1} \in \mathcal{P}} \lim_{n \to \infty} \mathbb{E}_f \left[S(\mathbf{y}) / n \right]$$

- i.e. the p^*_{n+1} that maximizes the **expected score**
- whatever the (bounded) value of w_n
- Choice of w_n matters in **finite samples**

Focused Bayesian Prediction (FBP)

- When S is such that

$$\exp[S(\mathbf{y})] = \text{ a probability (density ordinate)}$$

  then we set $w_n = 1$

- Otherwise we set w_n to ensure that

$$tr\{Var_{\Pi[p_{n+1}|\mathbf{y}]}[\boldsymbol{\theta}]\} \approx tr\{Var(\text{some sensible benchmark})\}$$

- like the exact (but misspecified) **likelihood-based** posterior

Mixtures of Predictives

- Remember, we can **define** $\mathcal{P}$ in such a way that the elements of the class are, themselves
- **weighted combinations** of predictives associated with **different** models; **e.g:**

$$p(y_{n+1}|\mathbf{y}, \cdot) = \sum_{k=1}^{K} \theta_k p(y_{n+1}|\mathbf{y}, M_k)$$

- Taking the constituent $p(y_{n+1}|\mathbf{y}, M_k)$ as 'given' $\Rightarrow$
- $p(y_{n+1}|\mathbf{y}, \cdot)$ characterized only by the unknown $\theta_k \in (0, 1)$
- Places inside a **formal coherent Bayesian up-dating scheme**
- the idea of **estimating weighted combinations** of predictives **via predictive criteria**
- (see also **Billio et al., 2013; Pettenuzzo & Ravazzolo, 2016; Casarin et al., 2019**)

Illustration: Simulated data

- Paper contains results for simulated & empirical data
- Will focus on one set of (simulated) results first
- **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 via G^{-1})
- Predicting **extreme returns** accurately is important
- Will **focus** on that goal $\Rightarrow$ use an appropriate S in the up-date

Three predictive classes

- $p(y_{n+1}|\mathbf{y}, \cdot) \in \mathcal{P}$ defined on:

1. $p(y_{n+1}|\mathbf{y}, \boldsymbol{\theta})$ for (normal) ARCH(1)
2. $p(y_{n+1}|\mathbf{y}, \boldsymbol{\theta})$ for (normal) GARCH(1,1)
3. $p(y_{n+1}|\mathbf{y}, \boldsymbol{\theta}) = \sum_{k=1}^{2} \theta_k p(y_{n+1}|\mathbf{y}, M_k)$

- $\Rightarrow$ **combination of predictives** based on:
  - (skewed normal) ARCH(1)
  - (normal) GARCH(1,1)
- $\Rightarrow$ Predictive class **increasingly less mis-specified**

Several scores used in the up-date

- Results for **two types of score** reproduced here:

  1. Log score ($\Rightarrow$ exact Bayes)
  2. **Censored** log score (rewards **predictive accuracy in a tail**)

  (Diks, Panchenko and van Dijk, 2011)

- Use an adaptive **random walk Metropolis Hastings** algorithm to simulate M draws of $p_{n+1}^{(i)}$ from:

$$\Pi[p_{n+1}|\mathbf{y}] \propto \exp[wS(\mathbf{y})] \times \Pi[p_{n+1}]$$

- Estimate: $E[p_{n+1}|\mathbf{y}] = \int_{\mathcal{P}} p_{n+1} d\Pi[p_{n+1}|\mathbf{y}]$ as: $\frac{1}{M} \sum_{i=1}^{M} p_{n+1}^{(i)}$
- Roll the whole process forward (with expanding windows)
- **Assess predictive performance** via the **censored log score**

Questions

Q1. **Does** the (within-sample) up-date based on the **censored score**
$\Rightarrow$

Best **out-of-sample performance** measured by that score?

Q2. **Does** the **degree of mis-specification** of $\mathcal{P}$ matter?

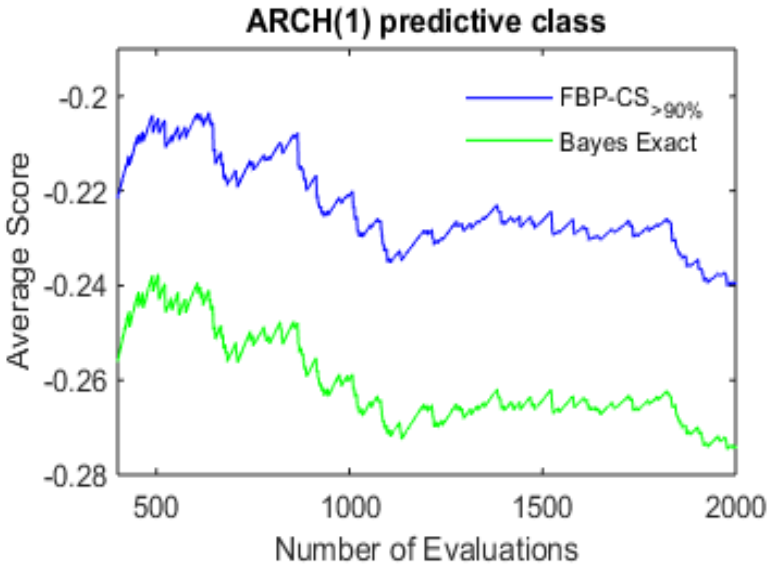
a. Does **less mis-specification**

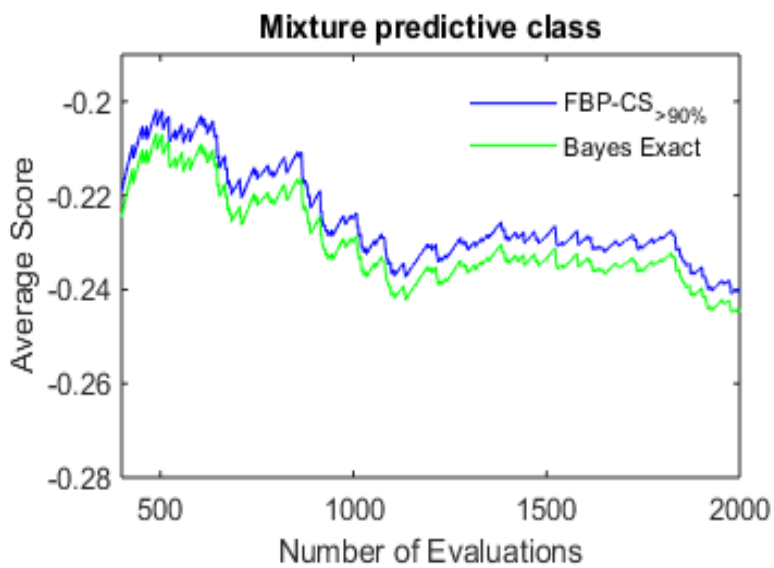
$\Rightarrow$ **less difference** in the out-of-sample performance of the different up-dates?

b. Does **mis-specification** have a **differential impact** on the performance of the different up-dates?

Out-of-sample performance

Average censored score for UPPER 10% tail





Q1. Blue lines **higher** than the green ✓

Q2a. Blue and green lines **closer** as **mis-spec. lessens**

Q2b. Green line ↑ **to meet** the blue as **mis-spec. lessens**

- $\Rightarrow$ **FBP** more **robust** to the choice of **predictive class**

Robustness of FBP makes sense

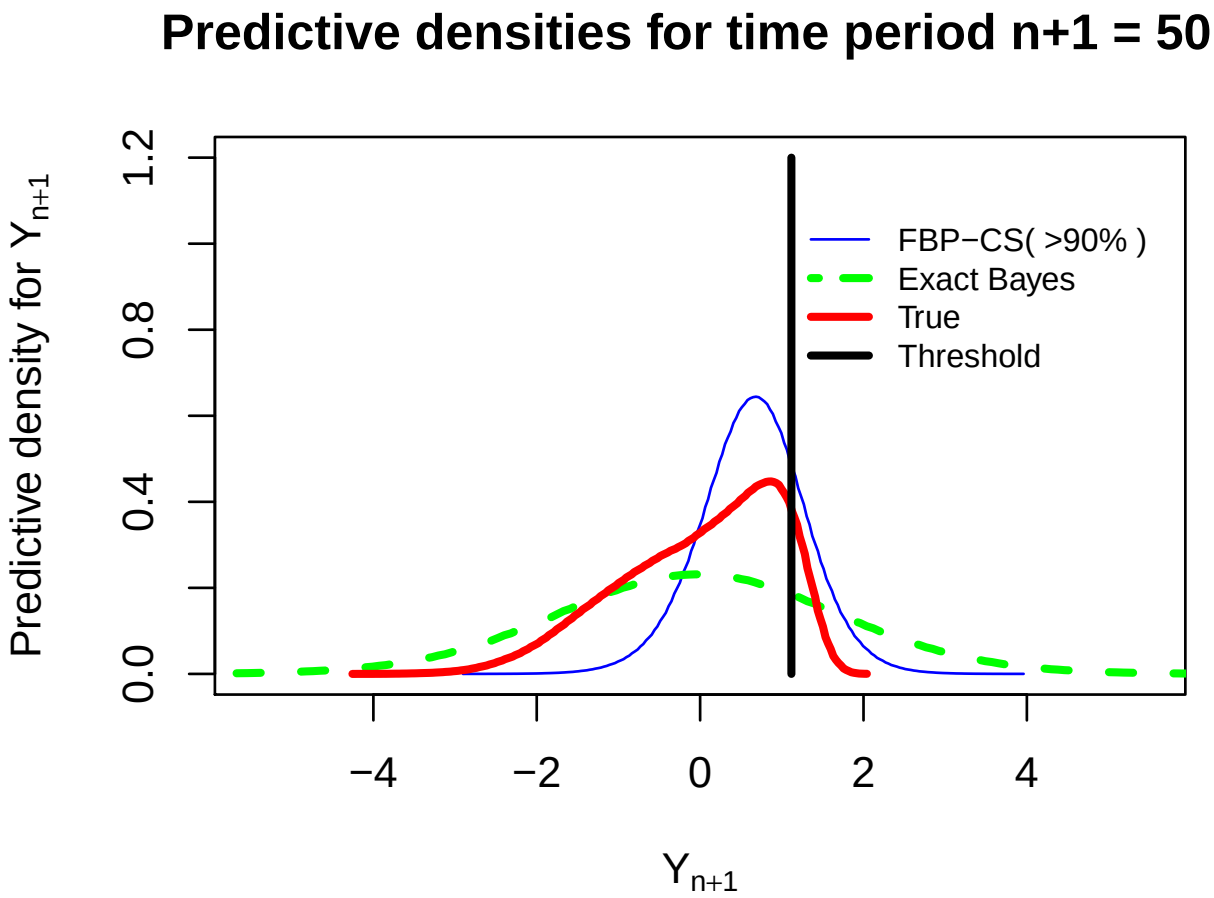
- Focusing on a specific characteristic of the data
- $\Rightarrow$ getting the model wrong matters less
- Is handy!
- A crude, computationally simple, predictive class (like ARCH(1)) does the job
- No need for the predictive combination (to better capture the true DGP)
- Not the aim!
- Aim is (only)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

Animation of Mean Predictives over Time

- Can **see** the dominance of **FBP** over **exact Bayes**
- By visualizing how the **two different** $E[p_{n+1}|\mathbf{y}]$ change over time
- Relative to the **true** predictive distribution
- (available via simulation in this artificial scenario)
- Again: will just focus on **upper tail** accuracy
- Using the **ARCH(1)** predictive class
- $\equiv$ the most **mis-specified** class

Animation of Mean Predictives over Time

- **Upper** tail focus



Posterior variation

- Remember: we have a whole **posterior distribution** of **predictives**, $\Pi\left[p_{n+1}|\mathbf{y}\right]$
- (for **FBP** and **exact Bayes**)
- that we can **exploit** and **visualize** in various ways....
- E.g. we can summarize the draws of p_{n+1} from the two different $\Pi\left[p_{n+1}|\mathbf{y}\right]$
- in terms of a **scalar** that is affected by **upper tail predictive accuracy**
- $\Rightarrow$ produce the **two different** posterior distributions for that **scalar**

Expected shortfall

- **Expected shortfall (for a 'short' portfolio):**

$$ES_{>0.9}^{n+1} = \int_{y_{n+1}>0.9} y_{n+1} d\Pi \left[p_{n+1} | \mathbf{y} \right]$$

- $\Rightarrow$ expected return, **conditional on** $y_{n+1} \in$ **upper** 10% tail
- $\Rightarrow$ expected return in the worst case scenario

Expected shortfall

- Superimpose $p[ES_{>0.9}^{n+1}|\mathbf{y}]$ (based on **exact Bayes** and **FBP**)
- On the **true** $ES_{>0.9}^{n+1}$
- Look at:

1. The location of the two posteriors, relative to the **true** (single) $ES_{>0.9}^{n+1}$ value
2. The degree of concentration of the two posteriors around the **true** $ES_{>0.9}^{n+1}$

- (Using the **ARCH(1)** predictive class)

Expected shortfall: Posterior distributions

Panel B: Short position

Expected shortfall posterior at n+1: $ES_{0.9}^{501}$

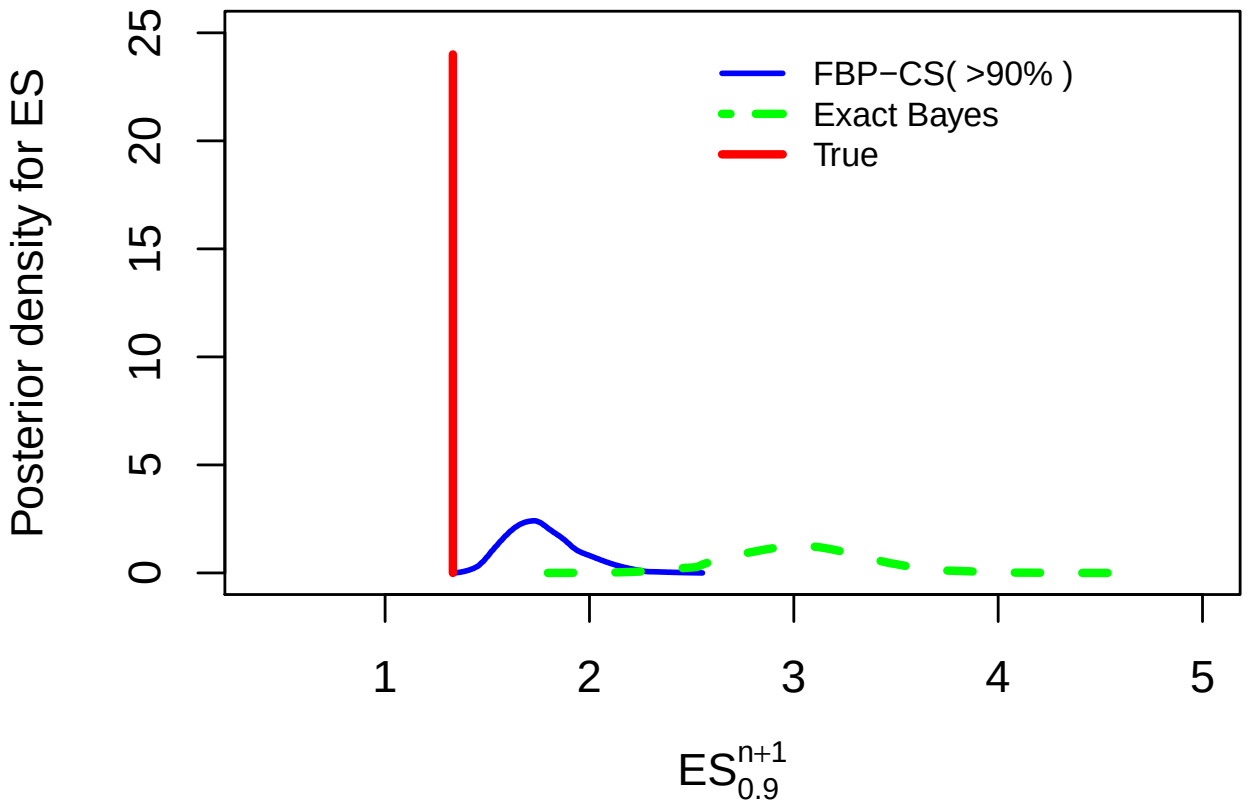


Illustration: Empirical data

2018 M4 Forecasting competition

- The challenge?
- 100-odd different forecast models/methods
- Attempt to accurately forecast **100,000** (!) different y_{n+h}
- Winner: best out-of-sample predictive accuracy
- over all **horizons** $(h = 1, 2, ..., H)$ and all **series**
- We focus on predictive **interval** accuracy measured by the **Mean Scaled Interval Score (MSIS)**
- Penalizes prediction if y_{n+h} falls outside prediction interval
- Rewards narrow prediction interval

M4 Forecasting competition

- Select the **23,000** annual series
- Apply **FBP** with **MSIS** as the up-dating rule:

$$\Pi[p_{n+1}|\mathbf{y}] \propto \exp[wS(\mathbf{y})] \times \Pi[p_{n+1}]$$

- Using a predictive class $p_{n+1} \in \mathcal{P}$ that performed well in **M4**
- (exponential smoothing model with trend and seasonality - **ETS**)
  - **Hyndman et al. (2002)**
- Does **FBP** reap *further* out-of-sample accuracy
- As measured by average **MSIS**?
- Yes!
- Shift the goal posts a little and **FBP** is **No. 1!**

M4 Forecasting competition: 23,000 annual series

MSIS	**M4-1st**	**M4-2nd**	**M4-3rd**	**ETS-4th**	**ETS-FBP**
Mean	**-23.90**	-27.48	-30.20	-34.90	-34.04
Median	-16.18	-16.09	-18.47	-15.49	**-14.70**
No. 'best'	4022	4532	3547	5076	**5823!**

To Conclude.....

- Adapting the Bayesian up-date to **focus** on forecast accuracy
- Yields more accuracy out-of-sample
- Asymptotically
- And in finite samples
- Obviates the need to specify the predictive model correctly
- $\Rightarrow$ Work has practical import

To Conclude.....

- Remember....
- The method applies to general **loss functions** (not just scores)
- E.g:
  - Loss function associated with **optimal portfolio allocation**
  - Loss function associated with under and over **prediction of electricity load**
  - Loss function associated with under and over **prediction of demand for ICU beds**
- Whatever loss function the context requires....

References

- Billio, M., Casarin, R., Ravazzolo, F. and van Dijk, H.K., 2013. 'Time-varying Combinations of Predictive Densities using Nonlinear Filtering', Journal of Econometrics, 177, 213-232.
- Bissiri, P.G., Holmes, C.C. and Walker, S.G., 2016. 'A General Framework for Updating Belief Distributions', Journal of the Royal Statistical Society (Series B), 78, 1103-1130.
- Canale, A. and Ruggiero, M., 2016. 'Bayesian Nonparametric Forecasting of Monotonic Functional Time Series', Electronic Journal of Statistics, 10, 3265-3286.
- Casarin, R., Grassi, S.., Ravazzola, F. and van Dijk, H.K., 2019. 'Forecast Density Combinations with Dynamic Learning for Large Data Sets in Economics and Finance', TI 2019-025/III, Tinbergen Institute Discussion Paper.
- Cooper, A., Frazier, D.T., Koo, B. and Martin, G.M., 2019. 'Variational Forecasting for Observation-driven Time Series Models', In Preparation.

References

- Diks, C., Panchenko, V., and Van Dijk, D., 2011. Likelihood-based Scoring Rules for Comparing Density Forecasts in Tails. Journal of Econometrics, 163, 215-230.
- Frazier, D.T., Maneesoonthorn, W., Martin, G.M. and McCabe, B.P.M., 2019. 'Approximate Bayesian Forecasting', International Journal of Forecasting, 35, 521-539.
- Giummole, F., Mameli, V., Ruli, E., and Ventura, L., 2017. 'Objective Bayesian Inference with Proper Scoring Rules', TEST, 28, 1-28.
- Guedj, B., 2019. 'A Primer on PAC-Bayesian Learning', arXiv preprint arXiv:1901.05353.
- Holmes, C.C. and Walker, S.G., 2017. 'Assigning a Value to a Power Likelihood in a General Bayesian Model', Biometrika, 104, 497-503.

References

- Hyndman, R. J., Koehler, A. B., Snyder, R. D., and Grose, S., 2002. 'A State Space Framework for Automatic Forecasting using Exponential Smoothing Methods', International Journal of Forecasting, 18, 439-454.
- Jiang, W. and Tanner, M.A., 2008. 'Gibbs Posterior for Variable Selection in High-dimensional Classification and Data Mining', Annals of Statistics, 36, 2207-2231.
- Knoblauch, J., Jewson, J., and Damoulas, T., 2019, 'Generalized Variational Inference: Three Arguments for Deriving New Posteriors', arXiv preprint arXiv:1904.02063.
- Kon Kam King, G., Canale, A. and Ruggiero, M., 2019. 'Bayesian Functional Forecasting with Locally-autoregressive Dependent Processes', Bayesian Analysis, 14, 1121-1141.
- Koop, G. and Korobilis, D., 2018. 'Variational Bayes Inference in High-dimensional Time-varying Parameter Models', https://ideas.repec.org/p/esy/uefcwp/22665.html.

References

- Loaiza-Maya, R., Martin, G. M., and Frazier, D. T., 2019. 'Focused Bayesian Prediction', arXiv preprint arXiv:1912.12571.
- Lyddon, S.P., Holmes, C.C. and Walker, S.G., 2019. 'General Bayesian Updating and the Loss-likelihood Bootstrap', Biometrika, 106, 465-478.
- Park, M. and Nassar, M., 2014. 'Variational Bayesian Inference for Forecasting Hierarchical Time Series', Conference paper, International Conference on Machine Learning (ICML), China.
- Pettenuzzo, D. and Ravazzolo, F., 2016. 'Optimal Portfolio Choice under Decision-based Model Combinations', Journal of Applied Econometrics, 31, 1312-1332.
- Quiroz, M., Nott, D.J. and Kohn, R., 2020. 'Gaussian Variational Approximation for High-dimensional State Space Models', arXiv preprint arXiv1801.07873v2.