Loss-Based Variational Bayes Prediction

David Frazier, Ruben Loaiza-Maya, Gael Martin and Bonsoo Koo

Department of Econometrics and Business Statistics

Monash University, Australia

Vienna, December, 2021

https://arxiv.org/abs/2104.14054

Standard 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 model/DGP**
- as does $p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})$ via **Bayes theorem**

Standard Bayesian Prediction

- Bayesian model averaging allows for extension to a finite set of K possible **models**:

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

- Bayesian paradigm $\Rightarrow$ a coherent approach to prediction
- **But**...what happens when we acknowledge that any **assumed** model (model set) is **misspecified**?
- In what sense does:

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

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

Focused Bayesian Prediction

- **Loaiza-Maya, Martin and Frazier (JAE, 2021)**
- Appropriate when the **true DGP is unknown**
- Define a class of **conditional predictives** that we believe **could** have generated the data:

$$\mathcal{P}^n \; : \; = \{p(y_{n+1}|\mathbf{y}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

- Elements of $\mathcal{P}^n$ may be:
  - a **single parametric** model with parameters $\boldsymbol{\theta}$
  - weighted combinations of predictives associated with **multiple parametric** models
  - ($\boldsymbol{\theta}$ comprises model-specific parameters and the weights)
- Define a **prior** over the elements of $\mathcal{P}^n$: $\Pi[p(y_{n+1}|\mathbf{y}, \boldsymbol{\theta})]$

Focused Bayesian Prediction

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

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

to a **posterior**:

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

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

Focused Bayesian Prediction

- In the spirit of **loss-based Bayes/generalized Bayes/Gibbs posteriors**
- e.g. **Jiang and Tanner (2008)**, **Bissiri et al. (2016)....**
- Up-date $p(\boldsymbol{\theta})$ to the '**Gibbs**' **posterior**:

$$p_G(\boldsymbol{\theta}|\mathbf{y}) \propto \exp[wS_n(\boldsymbol{\theta})] \times p(\boldsymbol{\theta}); \;\; w_n > 0$$

- via some (pos.) **scoring rule**:

$$S_n(\boldsymbol{\theta}) = \sum_{t=0}^{n-1} S(p(y_{n+1}|\mathbf{y}, \boldsymbol{\theta}), y_{t+1})$$

- that **rewards the predictive accuracy that matters**

Focused Bayesian Prediction

- $\Rightarrow$ (loosely speaking) a posterior over $p(y_{n+1}|\mathbf{y},\boldsymbol{\theta})$ itself.....
- Summarize by e.g. the **mean**:

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

- := '**Gibbs**' **predictive**
- Whilst the **standard** predictive:

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

- is 'trained' using the **log-score** (via $p(\boldsymbol{\theta}|\mathbf{y})$)
- The **Gibbs** predictive is 'trained' by the **score that matters** (via $p_G(\boldsymbol{\theta}|\mathbf{y})$)!

Focused Bayesian Prediction

- And it works!
- **Training** on the measure of predictive accuracy that matters
- (via the Bayesian up-date)
- Produces more accuracy out-of-sample
- (according to that measure)
- Than does a misspecified likelihood (**log-score-based**) update

Loss-based Variational Bayes Prediction

- However.....
- **Numerical computation** scheme is determined by the predictive class
- in **FBP** we adopted *simple* predictive classes (low-dimen. $\boldsymbol{\theta}$)
- $\Rightarrow$ **exact Gibbs posterior,** $p_G(\boldsymbol{\theta}|\mathbf{y})$, was accessible via **MCMC**
- In this paper we '**go big'**
- $\Rightarrow$ **MCMC** is less computationally attractive
- $\Rightarrow$ **approximate** $p_G(\boldsymbol{\theta}|\mathbf{y})$ using **variational Bayes**

Loss-based Variational Bayes Prediction

- Instead of targeting:

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

  - via **MCMC** draws from $p_G(\boldsymbol{\theta}|\mathbf{y})$

- We target:

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

- Where $\widehat{Q}$ (with density $\widehat{q}(\boldsymbol{\theta})$) **minimizes**, in a class $Q \in \mathcal{Q}$:

$$\text{KL}(Q||P_G[\boldsymbol{\theta}|\mathbf{y}]) = \int \log(dQ/P_G[\boldsymbol{\theta}|\mathbf{y}])\, dQ$$

Loss-based Variational Bayes Prediction

- We refer to $p_Q(y_{n+1}|\mathbf{y})$ as the **Gibbs variational predic<u>tive</u> (GVP)**
- And the production and use of $p_Q(y_{n+1}|\mathbf{y})$ as **Gibbs variational predic<u>tion</u> (GVP)**
- (interchangeably with '**loss-based variational prediction...'**)

Gibbs Variational Prediction (GVP)

- **Minimization** of

$$\mathrm{KL}\left(Q||P_G\left[\boldsymbol{\theta}|\mathbf{y}\right]\right) = \int \log\left(dQ/P_G\left[\boldsymbol{\theta}|\mathbf{y}\right]\right) dQ$$

- $\Leftrightarrow$ **maximization** of the **evidence lower bound (ELBO)**:

$$\mathrm{ELBO}[Q||\Pi\left[\cdot|\mathbf{y}\right]] = \mathbb{E}_Q[\log\left\{\exp[wS_n(\boldsymbol{\theta})]p(\boldsymbol{\theta})\right\}] - \mathbb{E}_Q[\log\left\{q(\boldsymbol{\theta})\right\}$$

- Adopting the **mean-field** variational class, $\mathcal{Q}$
- Implemented using **stochastic gradient ascent**

Theoretical Validation

- We show that:

  1. As $n \to \infty$, $\widehat{q}(\boldsymbol{\theta})$ concentrates onto

     $$\boldsymbol{\theta}_* = \arg\max_{\boldsymbol{\theta} \in \times} \lim_{n \to \infty} \mathbb{E}_f \left[S_n(\boldsymbol{\theta}) / n \right]$$

     - i.e. onto the $\boldsymbol{\theta}_*$ that maximizes the **expected score**
     - $\Rightarrow p(y_{n+1} | \mathbf{y}, \boldsymbol{\theta}_*)$ that **is 'optimal' in that scoring rule**

  2. **Rate of concentration** depends on **two terms**:

     - Rate of concentration of $p_G\left(\boldsymbol{\theta} | \mathbf{y}\right)$ onto $\boldsymbol{\theta}_*$
     - Proximity of $\widehat{q}(\boldsymbol{\theta})$ to $p_G\left(\boldsymbol{\theta} | \mathbf{y}\right)$

- **(Related work in: Alquier et al, 2016, Zhang and Gao, 2017, Alquier and Ridgeway, 2020)**

Theoretical Validation

- Viewed through another lense, the **Gibbs variational predictive**: $p_Q(y_{n+1}|\mathbf{y})$
- Is shown to **'merge'** with the **optimal predictive**, $p(y_{n+1}|\mathbf{y}, \boldsymbol{\theta}_*)$
  - **Blackwell and Dubins (1962)**
- To which the **exact Gibbs predictive**: $p_G(y_{n+1}|\mathbf{y})$ also merges
- Hence, in the limit, there is no loss, in terms of predictive accuracy
- By using the variational approximation
- Of course, the variational approximation will *potentially* influence finite sample performance

Numerical Validation

- So we explore the numerical performance of **GVP**
- First, in a **toy example** in which $p_G(y_{n+1}|\mathbf{y})$ is accessible via **MCMC**
  - What do we lose (**in finite samples**) by adopting the variational approximation?
- Then, in **simulation** examples based on **big** predictive models
  - Autoregressive (20-component) mixture model
  - Bayesian neural network
  - (**Both misspecified**)
- Plus an **empirical** example
  - Applying **GVP** to the 4227 daily time series in the M4 forecasting competition
- Will just focus on the **toy eg.** and the **empirical eg.**

Illustration: Simulated data

- **True DGP: stochastic volatility** model for a financial return (y_t)

$$
\begin{aligned}
y_t &= \exp(h_t/2)\varepsilon_t \\
h_t &= \alpha + \rho(h_{t-1} - \alpha) + \sigma_h \eta_t \\
\left[\begin{array}{cc} \varepsilon_t & \eta_t \end{array}\right]' &\sim i.i.d.N(\mathbf{0}, \left[\begin{array}{cc} 1 & -0.35 \\ -0.35 & 0.25 \end{array}\right])
\end{aligned}
$$

- $\Rightarrow$ y_t negatively skewed
- **Predictive model:** (Normal) GARCH(1,1)
- $\Rightarrow$ y_t symmetric
- $\Rightarrow$ **predictive model is misspecified**

Up-dating rule?

- Several (proper) scores used in the up-date:
- All of which reward different forms of predictive accuracy

1. Log-score (**LS**) ($\Rightarrow$ **misspecified** likelihood-based Bayes)
2. **Censored** log score (**CLS**)
    - rewards predictive accuracy **in a tail**
3. Continuously ranked probability score (**CRPS**)
    - rewards predictive mass **near the observed** y_{n+1}
4. Interval score (**IS**)
    - rewards accurate and narrow **prediction intervals**

Predictive Performance

- **Exact Gibbs prediction:** estimate of:

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

- using $M = 20000$ **MCMC** draws from $p_G(\boldsymbol{\theta}|\mathbf{y})$
- **GVP**: estimate of:

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

- using $M = 1000$ $i.i.d.$ draws from $\widehat{q}(\boldsymbol{\theta})$
- Roll the whole process forward (with expanding windows)
- **Assess predictive performance** via the full set of scores

Questions

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

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

- i.e. are the predictions (what we call) **coherent?**
- and does focusing on the form of predictive accuracy that matters yield more accurate forecasts than the **mispecified likelihood-based** up-date

Q2. Are the **exact** and **approximate** results identical?

Q3. And what is the speed gain of **GVP**?

Out-of-sample performance: GVP

- Positively-oriented scores ⇒ large (**in bold**) is good
- **Coherence** ⇒ **in bold** values on the diagonal!

- **Average out-of-sample score**

Up-dating	LS	$CLS_{<20\%}$	$CLS_{>80\%}$	CRPS	IS
LS	**-0.563**	-0.545	-0.354	-0.231	-2.347
$CLS_{<20\%}$	-0.806	**-0.497**	-0.628	-0.286	-2.985
$CLS_{>80\%}$	-0.936	-0.946	**-0.329**	-0.240	-3.325
CRPS	-0.565	-0.563	-0.343	**-0.230**	-2.434
IS	-0.655	-0.611	-0.371	-0.260	**-2.203**

Out-of-sample performance: GVP

- So, despite the approximation of the Gibbs posterior
- **GVP** produces **coherent** predictions
- And.....
- **VB-based** predictive results
- Are qualitatively equivalent to the **MCMC-based** predictive results
- And often numerically equivalent to 2 decimal places
- and are produced in a fraction of the time taken by **MCMC**
- **GVP** in the large (realistic) models still shown to produce **coherent** predictions overall

Illustration: Empirical Data

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 **interval score (IS)**
- Rewards accurate and narrow prediction intervals

Illustration: Empirical Data

M4 Forecasting Competition

- Select the **4227** daily series
- Apply **GVP** with **IS** as the up-dating rule:
- Use a flexible predictive model:
- A 20 component Gaussian autoregressive (AR-1) mixture
- Does **GVP** reap out-of-sample accuracy?
- In terms of out-of-sample **IS**

Illustration: Empirical Data

M4 Forecasting Competition

- As measured by **average IS (**over the **4227** series)
- The answer is **'No'**
- Not too surprising:
- Model is flexible, but probably a poor choice for some daily series
- (e.g. with time-varying volatility)
- The predictive model *still matters*

Illustration: Empirical Data

M4 Forecasting Competition

- As measured by the total number of series (out of **4227**) for which **GVP** is still best
- The answer is '**Yes**'
- **GVP** is the **second-best performer** overall
- *Despite* the shortcomings of the model
- Driving prediction by the **IS** update reaps real benefits
- Using **the appropriate update + a decent model** the ideal option
- *This* is the **new gold standard!**

In Summary....

- If prediction is your goal (rather than inference *per se*)
- And you're interested in a particular form of predictive accuracy
- And your model is too big for **MCMC**
- **GVP** seems to a good way to go.....
- In addition to having **theoretical** validity
- Any inaccuracy in approximating the Gibbs (loss-based) posterior used **VB**
- Has negligible impact on **numerical predictive** results

In Summary....

- This equivalence between **exact** and **approximate** predictions
- Mimics similar qualitative findings in other **VB-prediction** work:
  - **e.g. Quiroz et al. (2018), Koop and Korobilis (2018)**
- Plus earlier work on **ABC-based prediction:**
  - **Frazier, Maneesoonthorn, Martin and McCabe (2019)**
- **GVP** also seen to reap predictive benefits in realistic models for which **MCMC** is not feasible
- *However*, thus far - have only used:

$$\mathcal{P}^n := \{p(y_{n+1}|\mathbf{y}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

- where $p(y_{n+1}|\mathbf{y}, \boldsymbol{\theta})$ is an **observation-driven** predictive model
- If wish to adopt a **state space/hidden Markov** model
- **GVP** requires some extra effort.....

An Epilogue on GVP in SSMs

- Assume:

$$\text{Measurement density:} \qquad p(y_{n+1}|x_{n+1})$$

$$\text{(Markov) Transition density:} \quad p(x_{n+1}|x_n, \boldsymbol{\theta})$$

- Defining $\mathbf{x} = (x_1, x_2, ..., x_n)' \Rightarrow$
- **Exact predictive:**

$$\begin{aligned} & p(y_{n+1}|\mathbf{y}) \\ = \quad & \int_{x_{n+1}} \int_{\mathbf{x}} \int_{\boldsymbol{\theta}} p(y_{n+1}|x_{n+1}) p(x_{n+1}|x_n, \boldsymbol{\theta}) \\ & \times p(x_{n+1}|x_n, \boldsymbol{\theta}) p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} d\mathbf{x} dx_{n+1} \end{aligned}$$

An Epilogue on GVP in SSMs

- **Two** points to note:

1. **Approximate (VB-based) predictive:**

$$
\begin{aligned}
& p_Q(y_{n+1}|\mathbf{y}) \\
= \; & \int_{x_{n+1}} \int_{\mathbf{x}} \int_{\boldsymbol{\theta}} p(y_{n+1}|x_{n+1}) p(x_{n+1}|x_n, \boldsymbol{\theta}) \\
& \times p(x_{n+1}|x_n, \boldsymbol{\theta}) \underbrace{p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})}_{\widehat{q}(\mathbf{x})} \underbrace{p(\boldsymbol{\theta}|\mathbf{y})}_{\widehat{q}(\boldsymbol{\theta})} d\boldsymbol{\theta} d\mathbf{x} dx_{n+1}
\end{aligned}
$$

An Epilogue on GVP in SSMs

- In Frazier, Loaiza-Maya and Martin (2021):
  - **'A Note on the Accuracy of Variational Bayes in State Space Models: Inference and Prediction'**
  - https://arxiv.org/abs/2106.12262
- (Applying **VB** a likelihood-based **SSM** setting, and under **correct specification**)
- We show that:
- Inaccuracy in $\widehat{q}(\mathbf{x})$
  - $\Rightarrow$ lack of **Bayes consistency** for $\widehat{q}(\boldsymbol{\theta})$
  - i.e. $\widehat{q}(\boldsymbol{\theta})$ does not concentrate on $\boldsymbol{\theta}_0$
  - $\Rightarrow$ predictive inaccuracy

An Epilogue on GVP in SSMs

2. **GVP**, in turn, requires:

$$
\begin{aligned}
& p_Q(y_{n+1}|\mathbf{y}) \\
= \; & \int_{x_{n+1}} \int_{\mathbf{x}} \int_{\boldsymbol{\theta}} p(y_{n+1}|x_{n+1}) p(x_{n+1}|x_n, \boldsymbol{\theta}) p(x_{n+1}|x_n, \boldsymbol{\theta}) \\
& \times \underbrace{p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})}_{\widehat{q}(\mathbf{x})} \underbrace{p_G(\boldsymbol{\theta}|\mathbf{y})}_{\widehat{q}(\boldsymbol{\theta})} d\boldsymbol{\theta} d\mathbf{x} dx_{n+1}
\end{aligned}
$$

- where:

$$p_G(\boldsymbol{\theta}|\mathbf{y}) \propto \exp[w S_n(\boldsymbol{\theta})] \times p(\boldsymbol{\theta})$$

- and

$$S_n(\boldsymbol{\theta}) = \sum_{t=0}^{n-1} S(p(y_{t+1}|\mathbf{y}_{1:t}, \boldsymbol{\theta}), y_{t+1})$$

An Epilogue on GVP in SSMs

- In **Frazier, Martin, Loaiza-Maya and Torres-Andrade (2021)**:
  - '**Loss-Based Inference and Prediction in SSMs: A Variational Solution**'
- We implement **GVP** by:

An Epilogue on GVP in SSMs

1. Defining $p_G(\boldsymbol{\theta}|\mathbf{y})$ using $p(y_{n+1}|\mathbf{y},\boldsymbol{\theta})$ from an **approximation** to the **SSM** (e.g. a **LGSSM**) in which $\mathbf{x}$ can be integrated out analytically
2. Approximating this $p_G(\boldsymbol{\theta}|\mathbf{y})$ by $\widehat{q}(\boldsymbol{\theta})$
3. Recognizing that **neither** $p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})$ **nor** $\widehat{q}(\mathbf{x})$ is required for **prediction** in an **SSM**
    - $\Rightarrow$ Only need to access $p(x_n|\mathbf{y},\boldsymbol{\theta})$
    - $\Rightarrow$ Can be achieved **exactly** via particle filtering

- **1.** allows prediction to be driven by the relevant loss
- **2.** and **3.** allow for use of **VB**
    - Without the need for $\widehat{q}(\mathbf{x})$
    - And its inaccuracy impinging on predictive accuracy

Some Preliminary Results

- **True DGP** for a financial return (y_t)

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

- $\Rightarrow$ Implied copula of a **stochastic volatility** model combined with a **skewed normal marginal,** $g(y_t)$ (imposed via G^{-1})
- $\Rightarrow$ negative *skewness* in the **true predictive**
- **Predictive model**:

$$\begin{aligned} y_t &= \exp(h_t/2)\varepsilon_t; & \varepsilon_t &\sim N \\ h_t &= \alpha + \rho(h_{t-1} - \alpha) + \sigma_h \eta_t; & \eta_t &\sim N \end{aligned}$$

- $\Rightarrow$ **(mis-specified)** *symmetric* predictive

Some Preliminary Results

- **Steps:**
  1. **Re-express** the predictive model as:

$$\begin{aligned} y_t^* &= \ln(y_t^2) = h_t + \ln(\varepsilon_t^2) \\ h_t &= \alpha + \rho(h_{t-1} - \alpha) + \sigma_h \eta_t \end{aligned}$$

  2. **Approximate** the predictive model as the **Linear Gaussian SSM**:

$$\begin{aligned} &y_t^* = h_t + e_t; && e_t \sim N \\ &h_t = \alpha + \rho(h_{t-1} - \alpha) + \sigma_h \eta_t; && \eta_t \sim N \end{aligned}$$

  3. Apply the Kalman filter to produce:

$$p(y_{t+1}^* | \mathbf{y}_{1:t}^*, \boldsymbol{\theta})$$

  4. Transform (via the Jacobian) to:

$$\widehat{p}(y_{t+1} | \mathbf{y}_{1:t}, \boldsymbol{\theta})$$

Some Preliminary Results

- Then....

5. Specify the **Gibbs posterior** as:

$$p_G(\boldsymbol{\theta}|\mathbf{y}) \propto \exp[w S_n(\boldsymbol{\theta})] \times p(\boldsymbol{\theta})$$

where:

$$S_n(\boldsymbol{\theta}) = \sum_{t=0}^{n-1} S(\widehat{p}(y_{t+1}|\mathbf{y}_{1:t}, \boldsymbol{\theta}), y_{t+1})$$

and :

1. $S =$ **LS** ($\Rightarrow$ **misspecified** likelihood-based Bayes)
2. $S =$ **CLS** (rewarding predictive accuracy **in a tail**)

6. Produce the **VB** approximation, $\widehat{q}(\boldsymbol{\theta})$, to $p_G(\boldsymbol{\theta}|\mathbf{y})$

Some Preliminary Results

7. Produce a simulation-based estimate of the **GVP**:

$$
\begin{aligned}
& p_Q(y_{n+1}|\mathbf{y}) \\
= & \int_{x_{n+1}} \int_{x_n} \int_{\boldsymbol{\theta}} p(y_{n+1}|x_{n+1}) p(x_{n+1}|x_n, \boldsymbol{\theta}) p(x_{n+1}|x_n, \boldsymbol{\theta}) \\
& \times p(x_n|\mathbf{y}, \boldsymbol{\theta}) \widehat{q}(\boldsymbol{\theta}) d\boldsymbol{\theta} dx_n dx_{n+1}
\end{aligned}
$$

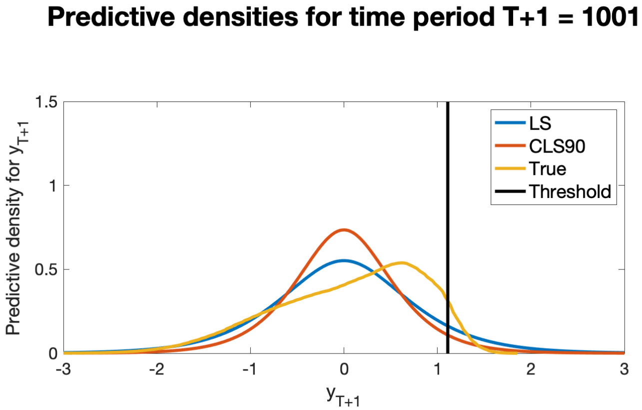
via:

1. draws of $\boldsymbol{\theta}$ from $\widehat{q}(\boldsymbol{\theta})$
2. draws of x_n from $p(x_n|\mathbf{y}, \boldsymbol{\theta})$ via the **bootstrap particle filter**
3. draws of x_{n+1} and y_{n+1} from $p(x_{n+1}|x_n, \boldsymbol{\theta})$ and $p(y_{n+1}|x_{n+1})$

7. Roll the whole process forward (with expanding windows)
8. **Assess predictive performance** via **LS** and (various) **CLS**

Animation of GVP over Time

- **Upper Tail Accuracy: LS** versus **CLS**$_{>90\%}$

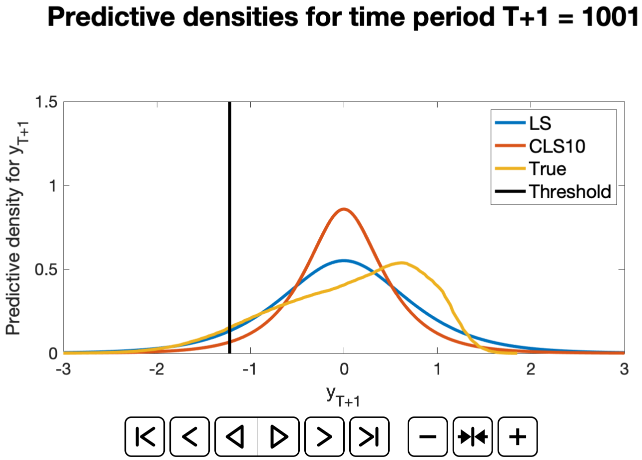


Animation of GVP over Time

- Problem with assumed predictive model is that mean is fixed at zero
- Estimated predictives can't **shift in location** to better pick up the **true predictive tail**
- Even so, designing the loss function to reward accuracy in the upper tail
- Still does what it is meant to do
- Produce a more accurate representation of the true upper tail

Animation of GVP over Time

- **Lower Tail Accuracy: LS** versus **CLS**$_{<10\%}$



Animation of GVP over Time

- The shape of the true predictive
- $\Rightarrow$ less benefit gained by focusing on lower tail accuracy in the up-dating rule
- Than there is in focusing on upper tail accuracy
- And this shows up in numerical out-of-sample results

Out-of-sample performance

- Positively-oriented scores ⇒ large (**in bold**) is good
- **Coherence** ⇒ looking for **bold** values on the diagonal
-

	Average out-of-sample score		
Up-dating	LS	$CLS_{<10\%}$	$CLS_{>90\%}$
LS	**-1.394**	**-0.394**	-0.314
$CLS_{<10\%}$	-1.415	-0.405	-0.302
$CLS_{>90\%}$	-1.473	-0.451	**-0.293**

- You have to pick your poison in this game!

So...a Start....

- To come:
- Predictive **SSMs** that **shift** in location to better pick up the **true predictive tail**
- Alternative **approximations**:

$$\widehat{p}(y_{t+1}|\mathbf{y}_{1:t},\boldsymbol{\theta})$$

- In the construction of the Gibbs posterior
- (E.g. using a Laplace approximation)
- Application of the method to a **large SSM**
- To warrant the use of **VB**

So...a Start....

- **Note though:**
  - Along the way we have provided a method for conducting **loss-based prediction in SSMs**
  - Irrespective of whether the **VB step** is used or not....
- Enough for now....