Discussion of:

# "Answering the Queen: Online Machine Learning and Financial Crises"

by Jérémy Fouliard, Michael Howell, Hélène Rey

Christian Julliard

London School of Economics

# The problem: "Why did nobody notice it?"

We would like to forecast the next (financial) crisis:

1. Without knowing the "true" model of the economy.

⇒ Reduced form forecasting.

2. Using as much information as possible → $N >> T$

⇒ Average forecasts over multiple, low dimension, sub-models – "experts"

3. In a way that is robust to the naïve Lucas critique (Sims (1980, 1987), Sargent (1994)...)

⇒ Need flexibility to accommodate dynamic evolving forecasting ⇒ time varying weights to "experts' opinions"

4. In a computationally feasible way.

⇒ "online" update rather than "batch" estimation.

Puzzle: but the "paper" actually does MLE (I think) for each model considered ... only aggregation is online...

Note: we can only hope to forecast crises types in the convex hull of history.

# Optimal Learning

- generally, there is no *uniformly* optimal estimation strategy.

E.g.:
1) minimax principle: optimizes prediction for worst true density
2) Rao-Cramér efficiency: minimum variance in the unbiased estimators class (or as $T \to \infty$)
3) Bayesian: can define an "average case" optimality (average over both random drawing of data and of true parameters of the DGP)

Remark: "average" optimality implies that no estimator can beat a Bayesian procedure for all true parameters.

Note: quadratic loss function over prediction densities implies that the optimal "on average" is the **mixture of all possible distributions** (in the considered family) **weighted by their posterior probabilities** aka Bayesian Model Averaging.

Bonus: the BMA predictive distribution minimizes the relative entropy, KLIC, relative to the true unknown DGP ⇒ i.e. as close as possible to the unknown truth even if misspecified.

# Bayesian Learning and Model Averaging

$P(D^t|\theta^k)$ : likelihood function of data $D^t := (z_1, ..., z_t)$ in $k$-th model.

$p(\theta^k)$ : prior belief (arbitrarily diffuse) on DGP parameters $\theta^k \in \Theta^k$ in $k$-th model/distribution/expert.

$p(\theta^k|D^t)$ : posterior distribution $\propto P(D^t|\theta^k)p(\theta^k)$

- In any $k$ model can forecast any $f(\theta^k|D^t)$ (e.g. pre-crisis prob.):

$$\widehat{f}_t^k := \int_{\Theta^k} f(\theta^k|D^t)p(\theta^k|D^t)d\theta^k$$

BMA: optimal "on average" forecast

$$\widehat{f}_t := \sum_k \widehat{f}_t^k \pi_t^k$$

combine multiple models'/experts' forecasts using models' posterior probabilities $\pi_t^k$ given by:

$$\frac{\text{prob. of } D^t \text{ in k-th model} \times \pi_0^k}{\sum_k \text{prob. of } D^t \text{ in k-th model} \times \pi_0^k} \equiv \frac{\int_{\Theta^k} P(D^t|\theta^k)p(\theta^k)d\theta^k \times \pi_0^k}{\sum_k \int_{\Theta^k} P(D^t|\theta^k)p(\theta^k)d\theta^k \times \pi_0^k}$$

where $\pi_0^k$ = prior probability of model $k$ (e.g. 1/#models)

# This paper: an "approximated" BMA (hence, I like it! ☺)

With:

1) the class of DGP considered: $P(D_t|\theta^k)$ is logistic.

2) $\widehat{f}_t^k$: posterior mean approximated by the forecast at the MLEs.

$\Rightarrow$ negligible approximation error IF the likelihoods are very sharp. (Are they? AUROC preselection might be helping...)

Note: could replace/mix "batch" MLE with "online" learning too (e.g. gradient descent) $\Rightarrow$ massive computational time gain to be had.

3) $\pi_0^k$: prior on models is $1/\#$number of models

4) $\pi_t^k$: posterior model prob. replaced by a (gradient descent algo) EWA.

$T \to \infty$ "should" converge to weights given by: $\frac{e^{-\frac{1}{2}BIC_t^k}}{\sum_k e^{-\frac{1}{2}BIC_t^k}}$ ...

... and this converges to $\pi_t^k$ IF data are (covariance) stationary.

5) preselect subset of possible models based on performance on sub-sample: $\approx$ 20-25 variable, 1.5-3 mil models per country.

$\Rightarrow$ Compatible with BMA (Occam/principle of parsimony, Madigan and Raftery(1994))

Note: doing proper BMA could construct a Markov Chain over possible models and feasibly work with even more models...

# Q1: but do we need an approximation?

1. - The "online" part of the paper is only (I think) the model averaging, not the individual model/expert MLE... but that's the computationally light part!
   - Posterior evaluation of Bayesian Probit (e.g Lancaster (2003)) is as fast as MLE: 1) Gibbs sampler = sequence of Gaussian draws; 2) "embarrassingly parallel" problem.
   - Given above, computing $\pi_t^k$ is straightforward (e.g. harmonic mean)
   - And it's realistically feasible!
     - Sala-y-Martin AER1997: 2 million models
     - At today's processing time $\approx$ 2 billion models i.e. 5 mil. models per country with "batch" posterior evaluation $\forall t$.
2. Posterior evolution can naturally be tracked "online" since $p(\theta^k|D^{t+1}) \propto P(z_{t+1}|\theta^k)p(\theta^k|D^t)$, i.e. time $t$ posterior = $t+1$ prior $\Rightarrow$ update based on $t+1$ data likelihood only.

Baseline: might need more convincing case for the approximation.

Bonus of proper BMA: can directly assess relevance of individual predictors (BMA of individual coefficients and/or marginal effects).

# Q2: and is EWA the "best" approximation?

- From my understanding of the slides, actually MLE is performed for each model/expert.

$\Rightarrow$ have the log-likelihood and Hessian at the MLE as a freebie.

But: from Laplace's method (second order approx.) we have

$$\int_{\Theta^k} P(D^t|\theta^k)p(\theta^k)d\theta^k \approx (2\pi)^{d_{\theta^k}/2}\left|\hat{\Sigma}_{\theta^k}\right|^{\frac{1}{2}} P\left(D^t|\hat{\theta}^k\right) p\left(\hat{\theta}^k\right)$$

where $d_{\theta^k}$ = dimension of $\theta^k$, $\hat{\Sigma}_{\theta}^{-1}$ is the negative Hessian evaluated at the MLE (i.e. observed Information matrix ) and $\hat{\theta}^k = \hat{\theta}^k_{MLE}$.

- Hence, can compute the posterior probabilities with no additional computational burden as:

$$\pi_t^k = \frac{(2\pi)^{d_{\theta^k}/2}\left|\hat{\Sigma}_{\theta^k}\right|^{\frac{1}{2}} P\left(D^t|\hat{\theta}^k\right)}{\sum_k (2\pi)^{d_{\theta^k}/2}\left|\hat{\Sigma}_{\theta^k}\right|^{\frac{1}{2}} P\left(D^t|\hat{\theta}^k\right)}$$

Note: valid under the same conditions needed to replace the posterior mean $\widehat{f}_t^k$ with its MLE value (as the authors do)

# "No ~~man~~ country is an island"

Financial crises tend to be global, rather than local, phenomena. Or at least to spill over domestic boundaries.

But: the "experts"/models considered are purely domestic.

⇒ should expand the space of models to include foreign states.

⇒ increases dimensionality of the problem...

- need to either replace/mix MLEs with online learning (frequentist or Bayesian) or construct a Markov Chain over the the models for taking draws (or again use Occam's razor)

Note: in Bayesian setting one can also "easily" handle models like:

$$y_{i,t} = \begin{cases} 1 & \text{if } y^*_{i,t} = x_t\beta + \phi \sum_{j \neq i} g_{i,j,t} y^*_{j,t} + \varepsilon_{i,t} < 0 \\ 0 & \text{otherwise} \end{cases}$$

Where $y^*_i$ is the latent state of country $i$ and the weights $g_{i,j,t}$ capture the network considered (i.e. trade links, borrowing/lending relations, etc.)

⇒ could be part of BMA/experts considered.

# Summary

- An ambitious and needed project.
- And the approach proposed does make sense (for a Bayesian at least).

Note: "data mining" is a swear word only in our field...

- I look forward to the paper!
- And therein I'd like to see:
  1. a strong case in favour of the approach proposed vis-à-vis (proper and/or approximated via Laplace's method) Bayesian Model Averaging.
  2. experts/models that allow for cross country linkages and spillovers.