Discussion of “Predictive Density Combinations with Dynamic Learning for Large Data Sets in Economics and Finance”

by Casarin, Grassi, Ravazzolo, Herman K. van Dijk

Dimitris Korobilis
University of Essex, UK

10th ECB Forecasting Workshop

Why predictive density combinations?

Authors propose a very flexible Bayesian model for predictive density combinations

This paper comes as no surprise, in the sense that these authors have long-standing interest in the topic and many past papers

While point forecasts still stubbornly dominate media, the background of the authors suggests a “many models - many forecasts”, real-time, density forecasting that supports real-time decision-making

How does this paper extend previous work by the authors?

Bilio et al., (2013), JoE; Aastveit et al., (2017), JBES; Casarin et al., (2015), JSS, among many others

How is this paper different?

The proposed approach has several desirable features

- **Scalable Computation:** Combination of information in thousands of predictive densities is possible
- **Dynamic Learning:** Combination weights adapt automatically to a changing environment
- **Regularized Estimation:** This is done using classification via mixtures

An immediate critique that applies to this paper, as with many other “machine learning” papers (incl. my poster yesterday):

- ♣ The methodology is statistical and, hence, atheoretical
- ♣ The model is too flexible and “black box”

→ I replicate these points here not to stress them, rather “get over” them quickly and enjoy what this paper is about

Understanding the methodology

- Let y_t be a univariate variable of interest
- Let $\widetilde{y}_{it}$ for $i = 1, ..., n$ be series of predictions of y_t
- Herman also defines the conditioning information set I (typically $y_{t-1}, ...$ for time-series data/models) – but **ignore this** for notational simplicity

Then the marginal predictive density of y_t is

$$\begin{aligned} f(y) &= \sum_{i=1}^{n} w_{it} f(y_{it}) && (1) \\ &\equiv \sum_{i=1}^{n} w_{it} \int_{\mathbb{R}} f(y_t|\widetilde{y}_{it}) f(\widetilde{y}_{it}) \mathrm{d}\widetilde{y}_{it} && (2) \end{aligned}$$

Model components

$$f\left(y\right)=\sum_{i=1}^{n}w_{it}\int_{\mathbb{R}}f\left(y_{t}|\widetilde{y}_{it}\right)f\left(\widetilde{y}_{it}\right)\mathrm{d}\widetilde{y}_{it} \tag{3}$$

Three components stand out

- $f\left(\widetilde{y}_{it}\right)$: This in practice can be written as $f\left(\widetilde{y}_{it}|I\right)$ or $f\left(\widetilde{y}_{it}|y_{t-1}\right)$ and is given (by model, expert, consumers etc)
- $f\left(y_{t}|\widetilde{y}_{it}\right)$: Authors assume this conditional density is $N\left(\widetilde{y}_{it},\sigma_{t}^{2}\right)$
- w_{it}: Time-varying weights that are updated flexibly using a dynamic Bayesian learning scheme (explained in next slide)

Dynamic learning of weights

Weights are estimated by clustering n weights into m groups ($m \ll n$), and then allowing for a state-evolution of latent quantities

1. Map weights w_{it} into latent variables z and parameters B_{it} via $w_{it} = \phi(z_{it}, B_{it})$
   - → Function $\phi(\bullet)$ maps from $\mathbb{S}^n$ to $\mathbb{S}^m$ with $\mathbb{S}$ the set [0,1]
   - → B_{it} allow correlation of m latent variables
2. Map z_{it} to latent variables v_{it} via a logistic function

$$z_{it} = \frac{\exp(v_{it})}{\sum_{j=1}^{m} \exp(v_{it})} \tag{4}$$

3. Allow v_{it} to evolve as a random walk: $v_{it} = v_{it-1} + \eta_t$

The weights are correlated over time (step 3) and over the cross-section (step 1)

Evaluation

- Very flexible approach, that also leads to efficient parallel (k-means clustering) GPU computation
- "Big Data" approach allows combination of thousands of predictive densities
- Two useful empirical illustrations, one in finance and one in macroeconomics
- Finance exercise reveals good performance in terms of various measures, including VaR-based measures
- Macro exercise also shows good performance using various point and density prediction measures

Remark 1

- A basic assumption is that $f\left(y_t|\widetilde{y}_{it}\right) = N\left(\widetilde{y}_{it}, \sigma_t^2\right)$. While σ_t^2 captures model incompleteness and mixtures of Gaussian specifications can be very flexible, using a Gaussian combination density might not be optimal for capturing tail behaviour of the combined density.
- Instead of specifying a "mean" heteroskedastic regression for y_t on $\widetilde{y}_t$, you could try a quantile heteroskedastic regression for these variables.
- This would imply that you have a different process σ_t^2 for each quantile of the combination density: model incompleteness would become a function of the specific quantile of interest!
- From a Bayesian perspective you could use a Laplace likelihood which can lead to conditionally conjugate inference without excessive additional computational burden

Remark 2

- While your measures of forecast performance are well-established, it would be interested to look at PITs
- ...and follow the analysis of Rossi and Sekhposyan (2014) to find out if they are well-calibrated, uniform, independent etc.
- Rossi and Sekhposyan (2014) also look at combination forecasts with equal weights, and time-constant (non-stochastic) weights estimated using Bayesian shrinkage
- On a separate note, Galbraith and van Norden (2012) assess the probability of the variable of interest exceeding a certain threshold, instead of traditional point or density forecasts. This avenue might be worth exploring given the possible complex forms that your combined density scheme might be able to capture.

Remark 3

- In the beginning of this presentation I mentioned that a standard criticism of such models is that they seem to be “black box”
- Clustering and classification you use has interpretation/labelling issues similar to factor models
- There is no indication if there exist “good” or “bad” sets of densities out of all n densities (with n too large to enumerate)
- This is very important for applied researchers
  - It is important to know if certain individual density forecasts perform better during certain periods
  - There are large maintenance costs associated with having to feed in your model with too many densities at each time period
- One way to dampen such effect is to replace the k-means clustering procedure with shrinkage and model selection (e.g. LASSO)

Remark 4

- An obvious use of your model is forecasting
- However, your model has two important features: combination of huge information sets AND achieving a flexible combined predictive distribution
- Could you use these features to extract other measures of interest to policy-makers such as disagreement, or uncertainty measures?
- Such measures depend on the first few moments of the predictive distribution, e.g. disagreement is the interquantile range of distribution of survey forecasts
- Uncertainty is the volatility of the unforecastable component of a series, and Jurado et al. (2015) measure total macro uncertainty by aggregating individual uncertainty with simple weights
- Can your combined density approach (assuming you can obtain data on thousands of individual densities) improve over simpler measures of disagreement?

Remark 5

- Another issue is that of scalability and usability of the estimation procedure
- Working with GPUs is obviously liberating, but it has obvious hardware limitations and accessibility issues
- An alternative approach would be to focus on approximating iterative algorithms that have a simple structure
- For example, mean field variational Bayes approximations to nonlinear state-space posterior would be trivial to derive
- They would result in computationally efficient variational Bayes Kalman filter algorithms for high-dimensional inference
- For example in Koop and Korobilis (in preparation) we derive a variational Bayes Kalman filter algorithm that allows us to use a **dynamic** spike and slab prior in a TVP regression with hundreds of predictors (similar MCMC algorithms can only handle a handful of predictors)

Conclusions

- This paper gives a great example of inference using flexible combinations of predictive densities and efficient GPU computation
- It paves the way for further developments in the area of combining flexibly a large number of forecasts (either point or density)
- As the availability of relevant micro and survey data increases rapidly, works such as the Casarin et al paper have the potential to provide a good benchmark for understanding what works in forecasting and what doesn't
- In that respect my final advice is to urge you to examine as many special cases of your flexible specification as you can, by switching off certain features (you already do that to some extend)