Discussion of "Combining Predictive Densities using Nonlinear Filtering with Applications to US Economics Data"

Dimitris Korobilis[1,2]

[1]University of Glasgow
[2]Rimini Center for Economic Analysis

Workshop on *Uncertainty and Forecasting in Macroeconomics*
Frankfurt
June 1 & 2, 2012

Contribution of Billio et al. (2011)

The authors begin with a combination scheme for whole densities of the form

$$p(y_t|\bullet) = \sum p(w|\bullet)p(\tilde{y}_t|\bullet) \tag{1}$$

where

- $p(y_t|\bullet)$ is the target weighted density for the variable(s) of interest y
- $p(w|\bullet)$ is the density of the weights (or of any nonlinear function of them)
- $p(\tilde{y}_t|\bullet)$ is the density of the predictors/predictions of y

The contribution is to model weights which vary with time ($w = w_t$).

> *Interestingly, however, many of the combination methods that attempt to build in time-variations in the combination weights (either in the form of discounting of past performance or time-varying parameters) have generally not proved to be successful, although there have been exceptions. [Timmermann, 2006]*

How to model time-varying weights

The authors consider issues such as "learning" and "correlated weights". Begin with

$$w_{i,t} = g(x_{i,t}), \textit{ for predictor/prediction } i = 1, .., L \quad (2)$$

$$x_t = x_{t-1} + \triangle \varepsilon_t \quad (3)$$

where $x_t = (x_{1,t}, ..., x_{L,t})$, and

- $g(x_{i,t}) = \frac{exp(x_{i,t})}{\sum_i exp(x_{i,t})}$
- x_t is a latent process which we model as time-varying
- $\triangle \varepsilon_t = \varepsilon_t - \varepsilon_{t-1}$, where ε_t is the vector of forecast errors

The specification above implies that since $x_{i,t}$ is a function of the forecast errors $(y_t - \tilde{y}_{i,t})$, hence also $w_{i,t}|y_t, \tilde{y}_t)$.

Time-varying weighting schemes

Francesco has presented a state-space representation of the weights, and a particle filter to draw w_t (and any other parameters θ.

Since the authors use an Exponentially Weighted Moving Average specification of the form

$$\varepsilon_t = \lambda \varepsilon_{t-1} + (1 - \lambda)(y_t - \tilde{y}_t)^2 \tag{4}$$

to obtain the [variance of the] forecast errors, I would also suggest fast approximate updating schemes for the weights, some of which are based on exponential decay.

That way, a larger set of predictors (and their weights) could be used in combination forecasting. This would potentially balance the effect of an increased estimation error from using approximations.

Linear Forgetting (Kulhavý & Kraus, 1996)

$$w_{t|t-1}^{(i)} = \frac{\mu\left(w_{t-1|t-1}^{(i)}\right) + (1-\mu)\, p(\omega_i)}{\sum_{i=1}^{L}\left[\mu\left(w_{t-1|t-1}^{(i)}\right) + (1-\mu)\, p(\omega_i)\right]} \tag{5}$$

$$w_{t|t}^{(i)} \propto w_{t|t-1}^{(i)} f\left(\varepsilon_t\right) \tag{6}$$

where μ is a decay/forgetting factor (similar to λ in the EWMA specification shown previously), and $f\left(\varepsilon_t\right)$ is a measure of acuracy as function of the forecast errors. For instance, Koop and Korobilis (2012, IER) set $f\left(\varepsilon_t\right) = p(y_t|y^{t-1})$, i.e. each predictor's predictive likelihood.

Finally $p(\omega_i)$ is a prior for each predictor's i probability $w_t^{(i)}$ which can be helpful in light of prior beliefs. $\longrightarrow$ Uninformative prior is $p(\omega_i) = 1/L$.

Exponential Forgetting (Kulhavý & Kraus, 1996)

Similarly, we can use exponential forgetting for the update of the weights:

$$w_{t|t-1}^{(i)} = \frac{\left(w_{t-1|t-1}^{(i)}\right)^{\mu} + p(\omega_i)^{(1-\mu)}}{\sum_{i=1}^{L}\left[\left(w_{t-1|t-1}^{(i)}\right)^{\mu} + p(\omega_i)^{(1-\mu)}\right]} \tag{7}$$

$$w_{t|t}^{(i)} \propto w_{t|t-1}^{(i)} f\left(\varepsilon_t\right) \tag{8}$$

Additionally $f\left(\varepsilon_t\right)$ can be any function of forecast errors, for instance following Kapetanios, Labhard and Price (2008, JBES) we can set

$$f\left(\varepsilon_t\right) = \frac{\exp\left(-1/2\Psi^{(i)}\right)}{\sum_{i=1}^{L}\exp\left(-1/2\Psi^{(i)}\right)} \tag{9}$$

where $\Psi^{(i)} = MSFE_{t-1}^{(i)} - \min_j MSFE_{t-1}^{(j)}$, where $MSFE$ is the mean squared forecast error.

Shrinkage

- Approximations are fast, however they do not allow enough modelling flexibility.
- Additionally, error in the estimation of weights can be very important $\rightarrow$ Timmermann (2006) explains that the "equal weights" $(1/L)$ approach works better sometimes just because it is error free.
- Hence it is important to incorporate shrinkage.

Shrinkage could be applied in the state-space representation for x_t that the authors use (see Frühwirth Schnatter and Wagner, 2010, JoE).

Additionally, the authors also suggest alternative Dirichlet process and mixture/ Markov-Switching models for x_t, for which shrinkage representations do exist in the Bayesian literature; see Dunson et al. (2008, JASA) and Tadese et al. (2005, JASA), respectively.

Shrinkage 2

Example 1: Belmonte, Koop and Korobilis (2012)

- Develop Bayesian least absolute and selection operator (LASSO) shrinkage prior for state-space / time-varying parameters models
- Model can be shrunk towards different directions, for instance: 1) constant parameters, 2) time-varying parameters, 3) slowly moving parameters, and 4) parameters shrunk to zero.

Example 2: Covariance selection models (Smith and Kohn, 2002, JASA)

$\rightarrow$ Shrink covariance matrix of weights to zero (applies to stochastic covariance matrix as well)

Example 3: Shrinkage for factor models (Korobilis, 2012, OBES)

- Assume factor stochastic volatility structure on the weights (more parsimonious), and apply shrinkage
- Can also be used when nonlinearities (e.g. structural breaks) are present in the loadings, or other coefficients (see also Korobilis, 2012, JAE)

Some other thoughts for the future

- The time-varying setting is ideal for dealing with "missing" predictors (for instance, combination of nowcasts, or non-model-based forecasts which might be missing randomly)
- The time-varying setting is also ideal for dealing with predictors measured at different frequencies
- Both of these could be summarized in an exercise which would involve real-time data, or nowcasting
- Other important questions to be asked: Some predictors may improve the mean/median combined forecast, however other predictors might improve the uncertainty (variance) of the combined forecast
- Develop a decomposition of the MSE of the combined density that could provide this information (think of the Brier score for probabilistic forecasts, which is a decomposition of the MSE into *calibration-refinement*, or *uncertainty-reliability-resolution*)