TI 2018-063/III
Tinbergen Institute Discussion Paper

# Learning to Average Predictively over Good and Bad: Comment on: Using Stacking to Average Bayesian Predictive Distributions

*Lennart (L.F.) Hoogerheide*[1]
*Herman (H.K.) van Dijk*[2]

[1] VU University Amsterdam
[2] Erasmus University, Norges Bank

# Learning to Average Predictively over Good and Bad: Comment on: Using Stacking to Average Bayesian Predictive Distributions*

Lennart Hoogerheide† andHerman K. van Dijk‡

**Abstract.** We suggest to extend the stacking procedure for a combination of predictive densities, proposed by Yao, Vehtari, Simpson, and Gelman(2018), to a setting where dynamic learning occurs about features of predictive densities of possibly misspecified models. This improves the averaging process of good and bad model forecasts. We summarise how this learning is done in economics and finance using mixtures. We also show that our proposal can be extended to combining forecasts and policies. The technical tools necessary for the implementation refer to filtering methods from nonlinear time series and we show their connection with machine learning. We illustrate our suggestion using results from Baştürk, Borowska, Grassi, Hoogerheide, and VanDijk(2018) based on financial data about US portfolios from 1928 until 2015.

## 1 Introduction

A basic practice in macroeconomic and financial forecasting [1] is to make use of a weighted combination of forecasts from several sources, say models, experts and/or large micro-data sets. Let $y_t$ be the variable of interest, and assume that some form of predictive values $\tilde{y}_{1t}, \ldots, \tilde{y}_{nt}$ is available for $y_t$ with a set of weights $w_{1t}, \ldots, w_{nt}$ where $n$ is also a decision variable. Then, basic practice in economics and finance is to make use of:

$$\sum_{i=1}^{n} w_{it} \tilde{y}_{it}. \tag{1.1}$$

This measure is intended as an accurate forecast of the variable of interest.
A major purpose of academic and professional forecasting is to give this practice a probabilistic foundation in order to quantify the uncertainty of such predictive density features as means, volatilities and tail behaviour. A leading example of a forecast density

*This working paper should not be reported as representing the views of Norges Bank. The views expressed are those of the authors and do not necessarily reflect those of Norges Bank. A reduced version of this paper will appear in *Bayesian Analysis*. For expository purposes, we write this more extended version.

†VU University Amsterdam and Tinbergen Institute l.f.hoogerheide@vu.nl

‡Erasmus University Rotterdam, Norges Bank and Tinbergen Institute hkvandijk@ese.eur.nl

[1] In applied statistics and economics the usual terminology is forecasting instead of prediction. In more theoretical work the usage of the word prediction is common. In this comment we use both terminologies interchangeably.

being produced and used in practice is the Bank of England's fan chart for GDP growth and inflation, which has been published each quarter since 1996. For a survey on the evolution of density forecasting in economics, see Aastveit etal.(2018a) and for a related formal Bayesian foundational motivation, see McAlinn and West(2018).
In recent literature and practice in statistics as well as in econometrics, it is shown that Bayesian Model Averaging (BMA) has its limitations for forecast averaging, see the earlier reference for a summary of the literature in economics. Yao, Vehtari, Simpson, and Gelman(2018) focus in their paper on the specific limitation of BMA when the true data generating process is not in the set. The authors also indicate the sensitivity of BMA in case of weakly or non-informative priors. As a better approach in terms of forecast accuracy and robustness, the authors propose the use of *stacking*, which is used in point estimation, and extend it to the case of combinations of predictive densities. A key step in the stacking procedure is that an optimisation step is used to determine the weights of a mixture model in such a way that the averaging method is then relatively robust for misspecified models, in particular, in large samples.
We fully agree that BMA has the earlier mentioned restrictions. However, we argue that a static approach to forecast averaging, as suggested by the authors, will in many cases remain sensitive for the presence of a bad forecast and extremely sensitive to a very bad forecast. We suggest to extend the approach of the authors to a setting where learning about features of predictive densities of possibly incomplete or misspecified models can take place. This extension will improve the process of averaging over good and bad forecasts. To back-up our suggestion, we summarise how this has been developed in empirical econometrics in recent years by Billio etal.(2013), Casarin etal.(2015,2018), Aastveit etal.(2018b) and Baştürk, Borowska, Grassi, Hoogerheide, and VanDijk(2018). Moreover, we show that this approach can be extended to combining not only forecasts but also policies. The technical tools necessary for the implementation refer to filtering methods from the nonlinear time series literature and we show their connection with dynamic machine learning. We illustrate our suggestion using financial data about US portfolios from 1928 until 2015.

## 2 Bayesian and diagnostic learning about misspecified models

**The Fundamental Predictive Density Combination.** Let the predictive probability distribution of the variable of interest $y_t$ of equation (1.1), given the set $\tilde{\boldsymbol{y}}_t = (\tilde{y}_{1t}, \ldots, \tilde{y}_{nt})^{'}$, be specified as a large discrete mixture of conditional probabilities of $y_t$ given $\tilde{\boldsymbol{y}}_t$ coming from $n$ different models [2] with weights $\boldsymbol{w}_t = (w_{1t}, \ldots, w_{nt})^{'}$ that are interpreted as probabilities and form a convex combination. One can then give (1.1) a stochastic interpretation using mixtures. Such a probability model, in terms of densities, is given as:

$$f(y_t|\tilde{\boldsymbol{y}}_t) = \sum_{i=1}^{n} w_{it} f(y_t|\tilde{y}_{it}). \tag{2.1}$$

[2] For convenience, we take as example models, but the approach holds equally for experts opinions etc.

Let the predictive densities from the $n$ models be denoted as $f(\tilde{y}_{it}|I_i), i = 1, ..., n$, where $I_i$ is the information set of model $i$. Given the *fundamental* density combination model of equation(2.1) and the predictive densities from the $n$ models, one can specify, given standard regularity conditions about existence of sums and integrals, that the marginal predictive density of $y_t$ is given as a discrete/continuous mixture,

$$f(y_t|I) \sim \sum_{i=1}^{n} w_{it} \int f(y_t|\tilde{y}_{it}) f(\tilde{y}_{it}|I_i) d\tilde{y}_{it} \tag{2.2}$$

where $I$ is the joint set of information of all models. The numerical evaluation of this equation is simple when all distributions have known simulation properties. An important research line in economics and finance has been to make this approach operational to more realistic environments by allowing for model incompleteness and dynamic learning where the densities have no known simulation properties; see the earlier cited references.

**Mixtures with model incompleteness.** A first step is to introduce, possibly, time-varying model incompleteness by specifying a Gaussian mixture combination model as

$$f(y_t|\tilde{\boldsymbol{y}}_t, \sigma_t^2) \quad \sim \quad \sum_{i=1}^{n} w_{it} f(y_t|\tilde{y}_{it}, \sigma_t^2) \tag{2.3}$$

where $\sigma_t^2$, $t = 1, \ldots, T$, is defined to follow the stochastic volatility process

$$\log \sigma_t^2 \quad \sim \quad f(\log \sigma_t^2 | \log \sigma_{t-1}^2, \sigma_\eta^2) \tag{2.4}$$

The process $\sigma_t^2$ controls the potential size of the misspecification in all models in the mixture. [3] When the value of $\sigma_t^2$ is large, the overall uncertainty and the misspecification of one or more predictor models are substantial. When the uncertainty level tends to zero then the mixture of experts or the smoothly mixing regressions model is recovered as limiting case, see Geweke and Keane(2007), Jacobs etal.(1991). Clearly, the analysis is also conditional upon the choice of the number $n$ and the information specified in the different models. The propose methodology allows to study overall incompleteness as well as incompleteness of each model over time. The importance of this *diagnostic learning* is shown in our empirical illustration.

**Dynamic weight learning.** Let the unrestricted weights stem from a multivariate normal random walk weight process. Other cases with more information are easily incorporated. Next, map the unrestricted weights to the simplex using a logistic transformation so that the weights can be interpreted as a convex set of probabilistic weights of different models which are updated periodically using *Bayesian learning* procedures. The learning aspect can, for instance, be specified by making use of log scores of the different forecasts.

**Representation as a nonlinear state space model with corresponding algorithms.** One can write the model in the form of a nonlinear state space where the measurement equation is a finite mixture of densities with a time-varying volatility to determine model incompleteness and the transition density is a logistic normal density

[3] Note that, for simplicity, $\sigma_t^2$ is constant among models. This assumption can be relaxed.

**Figure1:** Data driven density combinations and machine learning

| | | | |
|---|---|---|---|
| Data driven time series models for asset returns | | | |
| Model forecasts | VAR | SV | DFM |
| | ↓ | ↓ | ↓ |
| Hidden layer 1 of unobserved weights | $w_{11}$ | $w_{12}$ | $w_{13}$ |
| | ↘ | ↓ | ↙ |
| Forecast combination | Mixture of predictive densities | | |
| Data driven portfolio strategies | | | |
| Strategies | Model Momentum | | Residual Momentum |
| | ↓ | | ↓ |
| Hidden layer 2 of unobserved weights | $w_{21}$ | | $w_{22}$ |
| | ↘ | | ↙ |
| Policy combination | Mixture of policies | | |

of the set of mixture weights. The analytic solution of the optimal filtering problem is generally not known. But the representation allows to make use of algorithms based on sequential Monte Carlo methods such as particle filters in order to approximate combination weight and predictive densities.

Table1: Returns and risk for individual models and one strategy using US industry portfolios between 1926M7 and 2015M6.

| | Model Momentum | | | |
|---|---|---|---|---|
| | Mean | Vol. | S.R. | L.L. |
| VAR-N | 0.02 | 5.0 | 0.005 | -24.1 |
| SV | 0.10 | 5.1 | 0.019 | -34.7 |
| DFM-N(4,2) | -0.05 | 5.5 | -0.009 | -27.4 |

**Forecast combinations, policy combinations and machine learning.** To extend the predictive density combination approach to a policy combination one with time-varying learning weights is a very natural step in economics. It is well-known that fiscal and monetary policy should be combined with time-varying weights over the expansionary and recessionary stages of the business cycle in order to be more effective in controlling the amplitude and length of the cycle. The resulting gains in the accuracy of the complete forecast density constitute also very relevant information for events that occur in the tail of densities like the probability of a recession. We summarise in Figure1 how to combine forecasts and policies using a two-layer mixture. That is, we start with a mixture of predictive densities of three data driven time series models, *i.e.* a

Table2: Returns and risks from a mixture of three basic models and two investment strategies as well as from a mixture of two flexible models and two strategies, using US industry portfolios between 1926M7 and 2015M6. 0.11cm 1.2pt

| Model | Strategy | Mean | Vol. | S.R. | L.L. |
|---|---|---|---|---|---|
| *Mixture of basic models and two strategies* | | | | | |
| VAR-N & SV & DFM-N(4,2) | M.M. & R.M. | **0.10** (0.01,0.18) | **3.9** (3.6,4.2) | **0.025** (0.002,0.047) | **-23.0** (-28.8,-17.5) |
| *Mixture of two flexible models and two strategies* | | | | | |
| VAR-SV & DFM-SV(1-4,1-2) | M.M. & R.M. | **0.15** (0.08, 0.22) | **3.7** (3.5, 3.9) | **0.041** (0.021, 0.061) | **-21.6** (-26.4, -16.4) |

Vector-Autoregressive model (VAR), a Stochastic Volatility model (SV) and a Dynamic Factor Model (DFM). These are combined with a mixture of two data driven portfolio strategies that are known as momentum strategies. The basic idea of a momentum strategy is that at each portfolio decision time, past performance of the returns are assessed and one invests in the top set of 'winner' stocks and goes short in the bottom set of 'loser' stocks. For background on the model and residual momentum strategies we refer to to Baştürk, Borowska, Grassi, Hoogerheide, and VanDijk(2018). The mixture of mixtures approach results in this case in $3 \times 2$ possibilities. Figure1 is a graphical representation of equation (2.3). In the top row of Figure1, the predictions of the different models coming from $f(\tilde{y}_{it}|I_i)$ are shown. The next row, labeled as hidden layer of unobserved weights $w_{it}$, contains the information on the weights which have to be integrated out. The third row refers to the combination model $f(y_t|\tilde{\boldsymbol{y}}_{\boldsymbol{t}})$. It is noteworthy that this graphical representation is similar to the one used in machine learning. In our procedure the unobserved weights are integrated out using (particle) filtering methods from nonlinear time series. However, one may also use neural network specifications from machine learning to allow for different flexible functional forms.

**Illustration: Forecast combination, policy combination and risk-management**

We report a selected set of results of forecast density combinations of alternative models *in combination* with a set of portfolio policies from Baştürk, Borowska, Grassi, Hoogerheide, and VanDijk(2018). Table1 reports Means, Volatilities (Vol.), Sharpe Ratios (S.R.) and the Largest Losses (L.L.) for realised returns for the three basic models and one strategy, known as Model Momentum. It is clear that the VAR-N model produces a bad prediction of mean returns and that DFM-N(4,2) produces a very bad prediction. The top panel of Table2 shows return and risk features from a mixture of the three basic models (VAR-N, SV, DFM-N(4,2)) combined with a mixture of two investment strategies. The bottom panel reports these results for the mixture of the two same investment strategies combined with only two, but more flexible, models. The DFM-N(4,2) has been removed. The 90% credible intervals are reported in parentheses. These results lead to the following advice for portfolio strategy of an investment company:

*Conditional upon the information set that consists of US industry portfolios between*

*1926M7 and 2015M6 and conditional upon our set of data driven dynamic models and the two equity momentum strategies:*
*Mixtures of data driven models combined with mixtures of data driven strategies yield better return and risk features than single models combined with a single strategy.*
*In order to obtain higher mean returns and better risk features, it pays to average over flexible models instead of over simple basic models and it is effective to remove a 'bad' model.*
Thus, learning about the choice of a model set in a mixture is important for effective policies.We emphasise that approach from Baştürk, Borowska, Grassi, Hoogerheide, and VanDijk(2018) is fully Bayesian and does not contain an optimisation step as is used in stacking approach. However, the optimisation can be easily made dynamic. For a similar technique used in optimal pooling of forecasts we refer to Geweke and Amisano(2011).

## References

Aastveit, K.A., Mitchell, J., Ravazzolo, F., and van Dijk, H.K. (2018a). "The evolution of forecast density combinations in economics." Forthcoming in the Oxford Research Encyclopaedia in Economics and Finance. 2

Aastveit, K.A., Ravazzolo, F., and van Dijk, H.K. (2018b). "Combined density Nowcasting in an uncertain economic environment." *Journal of Business & Economic Statistics*, 36(1): 131–145. 2

Baştürk, N., Borowska, A., Grassi, S., Hoogerheide, L., and VanDijk, H.K. (2018). "Learning Combinations of Bayesian Dynamic Models and Equity Momentum Strategies." Journal of Econometrics, Forthcoming. 1, 2, 5, 6

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. 2

Casarin, R., Grassi, S., Ravazzolo, F., and van Dijk, H.K. (2015). "Parallel Sequential Monte Carlo for Efficient Density Combination: The Deco Matlab Toolbox." *Journal of Statistical Software*, 68(3). 2

— (2018). "Predictive Density Combinations with Dynamic Learning for Large Data Sets in Economics and Finance." Technical report. 2

Geweke, J. and Amisano, G. (2011). "Optimal prediction pools." *Journal of Econometrics*, 164(1): 130 – 141. 6

Geweke, J. and Keane, M. (2007). "Smoothly mixing regressions." *Journal of Econometrics*, 138: 252–290. 3

Jacobs, R.A., Jordan, M.I., Nowlan, S.J., and Hinton, G.E. (1991). "Adaptive mixtures of local experts." *Journal of Neural Computation*, 3: 79–87. 3

McAlinn, K. and West, M. (2018). "Dynamic Bayesian predictive synthesis for time series forecasting." *Journal of Econometrics*. Forthcoming. 2

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). "Using stacking to average Bayesian predictive distributions." *Bayesian Analysis*. 1, 2