Time-varying combinations of predictive densities using nonlinear filtering

Monica Billio
University of Venice

Roberto Casarin
University of Venice

Francesco Ravazzolo
BI Norwegian Business School
Norges Bank

Herman K. van Dijk
Erasmus University Rotterdam

This is the authors' manuscript to the article published in

***Journal of Econometrics,* 177(2013)2: 213-232**

DOI: http://dx.doi.org/10.1016/j.jeconom.2013.04.009

Time-varying Combinations of Predictive Densities using Nonlinear Filtering*

Monica Billio† Roberto Casarin†

Francesco Ravazzolo‡ Herman K. van Dijk§**

†University of Venice, GRETA Assoc. and School for Advanced Studies in Venice

‡Norges Bank and BI Norwegian Business School

§Erasmus University Rotterdam, VU University Amsterdam and Tinbergen Institute

October 29, 2012

Abstract

We propose a Bayesian combination approach for multivariate predictive densities which relies upon a distributional state space representation of the combination weights. Several specifications of multivariate time-varying weights are introduced with a particular focus on weight dynamics driven by the past performance of the predictive densities and the use of learning mechanisms. In the proposed approach the model set can be incomplete, meaning that all models can be individually misspecified. A Sequential Monte Carlo method is proposed to approximate the filtering and predictive densities. The combination approach is assessed using statistical and utility-based performance measures for evaluating density forecasts of simulated data, US macroeconomic time series and surveys of stock market prices. Simulation results indicate that, for a set of linear autoregressive models, the combination strategy is successful in selecting, with probability close to one, the true model when the model set is complete and it is able to detect parameter instability when the model set includes the true model that has generated subsamples of data. Also, substantial uncertainty appears in the weights when predictors are similar; residual uncertainty reduces when the model set is complete; and learning reduces this uncertainty. For the macro series we find that incompleteness of the models is relatively large in the 70's, the beginning of the 80's and during

*We benefited greatly from discussions with Concepción Ausín, Marco Del Negro, Frank Diebold, John Geweke, Dimitris Korobilis, Frank Schorfheide, Xuguang Sheng, Michael Wiper. We also thank conference and seminar participants at: the 4th CSDA International Conference on Computational and Financial Econometrics, the 6th Eurostat Colloquium, Norges Bank, the NBER Summer Institute 2011 Forecasting, the 2011 European Economic Association and Econometric Society, the Deutsche Bundesbank and Ifo Institute workshop on "Uncertainty and Forecasting in Macroeconomics", Boston University, Federal Reserve Bank of New York, University of Pennsylvania for constructive comments. The views expressed in this paper are our own and do not necessarily reflect those of Norges Bank.

**Corresponding author: Herman K. van Dijk, hkvandijk@ese.eur.nl. Other contacts: billio@unive.it (Monica Billio); r.casarin@unive.it (Roberto Casarin); francesco.ravazzolo@norges-bank.no (Francesco Ravazzolo).

the recent financial crisis, and lower during the Great Moderation; the predicted probabilities of recession accurately compare with the NBER business cycle dating; model weights have substantial uncertainty attached. With respect to returns of the S&P 500 series, we find that an investment strategy using a combination of predictions from professional forecasters and from a white noise model puts more weight on the white noise model in the beginning of the 90's and switches to giving more weight to the professional forecasts over time. Information on the complete predictive distribution and not just on some moments turns out to be very important, above all during turbulent times such as the recent financial crisis. More generally, the proposed distributional state space representation offers a great flexibility in combining densities.

JEL codes: C11, C15, C53, E37.

Keywords: Density Forecast Combination, Survey Forecast, Bayesian Filtering, Sequential Monte Carlo.

1 Introduction

When multiple forecasts are available from different models or sources it is possible to combine these in order to make use of all relevant information on the variable to be predicted and, as a consequence, to produce better forecasts. One of the first papers on forecasting with model combinations is Barnard [1963], who considered air passenger data, and see also Roberts [1965] who introduced a distribution which includes the predictions from two experts (or models). This latter distribution is essentially a weighted average of the posterior distributions of two models and is similar to the result of a Bayesian Model Averaging (BMA) procedure. See Hoeting et al. [1999] for a review on BMA, with an historical perspective. Raftery et al. [2005] and Sloughter et al. [2010] extend the BMA framework by introducing a method for obtaining probabilistic forecasts from ensembles in the form of predictive densities and apply it to weather forecasting.

Our paper builds on another stream of literature, starting with Bates and Granger [1969] and dealing with the combination of predictions from different forecasting models; see Granger [2006] for an updated review. Granger and Ramanathan [1984] extend Bates and Granger [1969] and propose to combine forecasts with unrestricted regression coefficients as weights. Liang et al. [2011] derive optimal weights in a similar framework. Terui and van Dijk [2002] generalize the least squares model weights by representing the dynamic forecast combination as a state space with weights that are assumed to

follow a random walk process. This approach has been extended by Guidolin and Timmermann [2009], who introduce Markov-switching weights, and by Hoogerheide et al. [2010], who propose robust time-varying weights and accounte for both model and parameter uncertainty in model averaging. Raftery et al. [2010] derive time-varying weights in "dynamic model averaging", following the spirit of Terui and van Dijk [2002], and speed up computations by applying forgetting factors in the recursive Kalman filter updating. Hansen [2007] and Hansen [2008] compute optimal weights by maximizing a Mallow criterion. Hall and Mitchell [2007] introduce the Kullback-Leibler divergence as a unified measure for the evaluation and combination of density forecasts and suggest weights that maximize such a distance, see also Geweke and Amisano [2010b]. Gneiting and Raftery [2007] recommend strictly proper scoring rules, such as the cumulative rank probability score.

In this paper, we assume that the weights associated with the predictive densities are time-varying and propose a general distributional state space representation of predictive densities and combination schemes. For a review on basic distributional state space representations in the Bayesian literature, see Harrison and West [1997]. Our combination method allows for all models to be false and therefore the model set to be misspecified as discussed in Geweke [2010] and Geweke and Amisano [2010b]. In this sense we extend the state space representation of Terui and van Dijk [2002] and Hoogerheide et al. [2010] and the model mixing via mixture of experts (see for example Jordan and Jacobs [1994] and Huerta et al. [2003]) by allowing for the possibility that all models are misspecified or, in other words, the model set is incomplete. Our approach is general enough to include multivariate linear and Gaussian models (e.g., see Terui and van Dijk [2002]), dynamic mixtures and Markov-switching models (e.g., see Guidolin and Timmermann [2009]), as special cases. We represent our combination schemes in terms of conditional densities and write equations for producing predictive densities and not point forecasts (as is often the case) for the variables of interest. Given this general representation, we can estimate (optimal) model weights that minimize the distance between the empirical density and the combination density, by taking into account past performances. In particular, we consider convex combinations of the predictive densities and assume that the time-varying weights associated with the different predictive densities belong to the standard simplex. Under this constraint the weights can be interpreted as discrete probabilities over the set of predictors. Tests for a specific hypothesis on the values of the weights can be conducted due to their random nature. We discuss weighting schemes with continuous dynamics, which allow for a smooth convex combination of the predictive densities.

A learning mechanism is also introduced to enable the dynamics of each weight to be driven by past and current performances of the predictive densities of all models in the combinations.

The constraint that time-varying weights associated with different forecast densities belong to the standard simplex makes the inference process nontrivial and calls for the use of nonlinear filtering methods. We apply simulation based filtering methods, such as Sequential Monte Carlo (SMC), in the context of combining forecasts, see for example Doucet et al. [2001] for a review with applications of this approach and Del Moral [2004] for convergence issues. SMC methods are extremely flexible algorithms that can be applied for inference to both off-line and on-line analysis of nonlinear and non-Gaussian latent variable models, see for example Creal [2009]. Billio and Casarin [2010] successfully applied SMC methods to time-inhomogeneous Markov-switching models for an accurate forecasting of the business cycle of the euro area.

Important features of our Bayesian combination approach have been analyzed in section 5 using a set of Monte Carlo simulation experiments. This yielded the following results. For the case of a set of linear models, the combination strategy is successful in selecting with probability close to one the true model when the model set is complete. High uncertainty levels in the combination weights appear due to the presence of predictors that are similar in terms of unconditional mean and that differ little in terms of unconditional variance. The learning mechanism produces better discrimination between forecast models with the same unconditional mean, but different unconditional variance. The degree of uncertainty in the residuals reduces when the model set is complete. A combination of linear with non-linear models shows that the learning period may be longer than for the case in which only linear models are present. Finally, we consider an example of a set of models containing a true model with structural instability. Here it is shown that the proposed combination approach is able to detect the instability when the model set includes the true model that is generating subsamples of data.

To show practical and operational implications of the proposed approach with real data, this paper focuses on the problem of combining density forecasts using two relevant economic datasets. The first one contains the quarterly series of US real Gross Domestic Product (GDP) and US inflation as measured by the Personal Consumption Expenditures (PCE) deflator. Density forecasts are produced by several of the most commonly used models in macroeconomics. We combine these densities forecasts in a multivariate set-up with model and variable specific weights. For these macro series we find that incompleteness of the models is relatively large in the 70's, the beginning of the 80's and

during the recent financial crisis while it is lower during the Great Moderation. Furthermore, the predicted probabilities of recession accurately compare with the NBER business cycle dating. Model weights have substantial uncertainty attached and neglecting it may yield misleading inference on the model's relevance. To the best of our knowledge, there are no other papers applying this general density combination method to macroeconomic data. The second dataset considers density forecasts on future movements of a stock price index. Recent literature has shown that survey-based forecasts are particularly useful for macroeconomic variables, but there are fewer results for finance. We consider density forecasts generated by financial survey data. More precisely we use the Livingston dataset of six-months ahead forecasts on the Standard & Poor's 500 (S&P 500), combine the survey-based densities with the densities from a simple benchmark model and provide both statistical and utility-based performance measures of the mixed combination strategy. To be specific, with respect to the returns of the S&P 500 series we find that an investment strategy using a combination of predictions from professional forecasters and from a white noise model puts more weight on the white noise model in the beginning of the 90's and switches to giving more weight to the professional forecasts over time. Information on the complete predictive distribution and not just from basic first and second order moments turns out to be very important in all investigated cases and, more generally, the proposed distributional state space representation of predictive densities and of combination schemes demonstrates to be very flexible.

The structure of the paper is as follows. Section 2 introduces combinations of predictive densities in a multivariate context. Section 3 presents different models for the weight dynamics and introduces learning mechanisms. Section 4 describes the nonlinear filtering problem and shows how Sequential Monte Carlo methods could be used to combine predictive densities. Section 5 contains results using simulated data and Section 6 provides results of the application of the proposed combination method to the macroeconomic and financial datasets. Section 7 contains conclusions and presents suggestions for further research. In the Appendices the data sets used are described in detail. Further, alternative combination schemes and the relationships with some existing schemes in the literature are discussed together with the Sequential Monte Carlo method used.

2 Combinations of Multivariate Predictive Densities

Let $\mathbf{y}_t \in \mathcal{Y} \subset \mathbb{R}^L$ be the L-vector of observable variables at time t and $\mathbf{y}_{1:t} = (\mathbf{y}_1, \ldots, \mathbf{y}_t)$ be the collection of these vectors from $1, \ldots, t$. Let $\tilde{\mathbf{y}}_{k,t} = (\tilde{y}^1_{k,t}, \ldots, \tilde{y}^L_{k,t})' \in \mathcal{Y} \subset \mathbb{R}^L$ be the typical one-step ahead predictor for $\mathbf{y}_t$ for the k-th model, where $k = 1, \ldots, K$. For the sake of simplicity we present the new combination method for the one-step ahead forecasting horizon, but our results can easily be extended to multi-step ahead forecasting horizons.

Assume that the L-vector of observable variables is generated from a distribution with conditional density $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$ and that for each predictor $\tilde{\mathbf{y}}_{k,t}$ there exists a predictive density $p_k(\tilde{\mathbf{y}}_{k,t}|\mathbf{y}_{1:t-1})$. To simplify notation, in what follows we define $\tilde{\mathbf{y}}_t = \text{vec}(\tilde{Y}_t')$, where $\tilde{Y}_t = (\tilde{\mathbf{y}}_{1,t}, \ldots, \tilde{\mathbf{y}}_{K,t})$ is the $L \times K$ matrix of predictors and vec is the operator that stacks the columns of this matrix into an KL-vector. We denote with $p(\tilde{\mathbf{y}}_t|\mathbf{y}_{1:t-1})$ the joint predictive density of the set of predictors at time t and let

$$p(\tilde{\mathbf{y}}_{1:t}|\mathbf{y}_{1:t-1}) = \prod_{s=1}^{t} p(\tilde{\mathbf{y}}_s|\mathbf{y}_{1:s-1})$$

be the joint predictive density of the predictors up to time t.

Generally speaking a combination scheme of a set of predictive densities is a probabilistic relationship between the density of the observable variable and the set of predictive densities. This relationship between the density of $\mathbf{y}_t$, conditionally on $\mathbf{y}_{1:t-1}$, and the set of predictive densities from the K different sources is given as:

$$p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) = \int_{\mathcal{Y}^{Kt}} p(\mathbf{y}_t|\tilde{\mathbf{y}}_{1:t}, \mathbf{y}_{1:t-1}) p(\tilde{\mathbf{y}}_{1:t}|\mathbf{y}_{1:t-1}) d\tilde{\mathbf{y}}_{1:t} \tag{1}$$

where the specific dependence structure between the observable and the predictive densities is specified below. This relationship might be misspecified because all models in the combination are false (incomplete model set) and to model this possibly misspecified dependence we consider a parametric latent variable model. We also assume that this model is dynamic to capture time variability in the dependence structure. Modeling the relationship between the observable and the predictive densities allows us to compute combination residuals and their distributions, which is a measure of the incompleteness of the model set. For example, the analysis of the residuals may be used to measure the lack of contribution of each model to the forecast of the variable of interest. The residual analysis may also

reveal the presence of time variation in the incompleteness level, e.g. due to structural change in the Data Generating Process (DGP). In Section 5 we investigate these issues through some Monte Carlo simulation studies.

Among others, Hall and Mitchell [2007], Jore et al. [2010] and Geweke and Amisano [2010b] discuss the use of the log score as a ranking device on the forecast ability of different models. The log score is easy to evaluate and can be used to detect misspecification by studying how model weights change over different vintages. One difference with our approach is that we consider the complete distribution of the residuals. This yields information about a bad fit in the center but also about a bad fit on scale and tails of the distribution; some results are reported in section 5. Therefore, we can contemporaneously study the dynamics of both weight distributions and predictive errors. Furthermore, the log score appears to be sensitive to tail events; see the discussion in Gneiting and Raftery [2007] and Gneiting and Ranjan [2011]. In the empirical macroeocnomic application we compare our method to combination schemes based on log score, see section 6. However, a careful analysis of the relative advantages of using the log score versus the time-varying combinations of predictive densities is a topic for further research.

To specify the latent variable model and the combination scheme we first define the latent space. Let $\mathbf{1}_n = (1, \ldots, 1)' \in \mathbb{R}^n$ and $\mathbf{0}_n = (0, \ldots, 0)' \in \mathbb{R}^n$ be the n-dimensional unit and null vectors respectively and denote with $\Delta_{[0,1]^n} \subset \mathbb{R}^n$ the set of all vectors $\mathbf{w} \in \mathbb{R}^n$ such that $\mathbf{w}'\mathbf{1}_n = 1$ and $w_k \geq 0$, $k = 1, \ldots, n$. $\Delta_{[0,1]^n}$ is called the standard n-dimensional simplex and is the latent space used in all our combination schemes.

Then, we introduce the latent model, that is a matrix-valued stochastic process, with random variable $W_t \in \mathcal{W} \subset \mathbb{R}^L \times \mathbb{R}^{KL}$, which represents the time-varying weights of the combination scheme. Denote with $w^l_{h,t}$ the h-th column ($h = 1, \ldots, KL$) and l-th row ($l = 1, \ldots, L$) elements of W_t, then we assume that the row vectors $\mathbf{w}^l_t = (w^l_{1,t}, \ldots, w^l_{KL,t})$ satisfy $\mathbf{w}^l_t \in \Delta_{[0,1]^{KL}}$. The proposed latent variable modelling framework generalizes previous literature on model combination with exponential weights (see for example Hoogerheide et al. [2010]) by inferring dynamics of positive weights which belong to the simplex $\Delta_{[0,1]^{LK}}$.[1] As the latent space is the standard simplex, the combination weights are [0,1]-valued processes and one can interpret them as discrete probabilities over the set of predictors. Thus, in our framework, the weights on the model set are not given, as in the standard model selection or

[1] Winkler [1981] does not restrict weights to the simplex, but allows them to be negative. It would be interesting to investigate which restrictions are necessary to assure positive predictive densities with negative weights in our framework. We leave this for further research.

BMA frameworks, but are random quantities. In this sense the proposed combination scheme shares some similarities with the dilution and hierarchical model set prior distributions for BMA, proposed in George [2010] and Ley and Steel [2009] respectively. A hierarchical specification of the weights in order to achieve a reduction of the model space by removing redundant weights is a matter of further research.

We assume that at time t, the time-varying process of random W_t has a distribution with density $p(W_t|\mathbf{y}_{1:t-1}, \tilde{\mathbf{y}}_{1:t-1})$. Then we can write Eq. (1) as

$$p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) = \int_{\mathcal{Y}^{Kt}} \left(\int_{\mathcal{W}} p(\mathbf{y}_t|W_t, \tilde{\mathbf{y}}_t) p(W_t|\mathbf{y}_{1:t-1}, \tilde{\mathbf{y}}_{1:t-1}) dW_t \right) p(\tilde{\mathbf{y}}_{1:t}|\mathbf{y}_{1:t-1}) d\tilde{\mathbf{y}}_{1:t} \tag{2}$$

We assume a quite general specification of the transition density, $p(W_t|W_{t-1}, \mathbf{y}_{1:t-1}, \tilde{\mathbf{y}}_{1:t-1})$, that allows the weights to have a first-order Markovian dynamics and to depend on the past values $\mathbf{y}_{1:t-1}$ of the observables and $\tilde{\mathbf{y}}_{1:t-1}$ of the predictors. Under this assumption, the inner integral in Eq. (2) can be further decomposed as follows

$$p(W_t|\mathbf{y}_{1:t-1}, \tilde{\mathbf{y}}_{1:t-1}) = \int_{\mathcal{W}} p(W_t|W_{t-1}, \mathbf{y}_{1:t-1}, \tilde{\mathbf{y}}_{1:t-1}) p(W_{t-1}|\mathbf{y}_{1:t-2}, \tilde{\mathbf{y}}_{1:t-2}) dW_{t-1} \tag{3}$$

The proposed combination method extends previous model pooling by assuming possibly non-Gaussian predictive densities as well as nonlinear weight dynamics that maximize general utility functions.

It is important to highlight that this nonlinear state space representation offers a great flexibility in combining densities. In Example 1 we present a possible specification of the conditional predictive density $p(\mathbf{y}_t|W_t, \tilde{\mathbf{y}}_t)$, that we consider in the applications. In Appendix B we present two further examples that allow for heavy-tailed conditional distributions. In the next section we will also consider a specification for the weights transition density $p(W_t|W_{t-1}, \mathbf{y}_{1:t-1}, \tilde{\mathbf{y}}_{1:t-1})$.

Example 1 - (*Gaussian combination scheme*)
The conditional Gaussian combination model is defined by the probability density function

$$p(\mathbf{y}_t|W_t, \tilde{\mathbf{y}}_t) \propto \exp\left\{ -\frac{1}{2} (\mathbf{y}_t - W_t\tilde{\mathbf{y}}_t)' \Sigma^{-1} (\mathbf{y}_t - W_t\tilde{\mathbf{y}}_t) \right\} \tag{4}$$

where $W_t \in \Delta_{[0,1]^{L\times KL}}$ is the weight matrix defined above and Σ is the covariance matrix.

■

A special case of the previous model is given by the following specification of the combination

$$p(\mathbf{y}_t|W_t,\tilde{\mathbf{y}}_t)\propto\exp\left\{-\frac{1}{2}\left(\mathbf{y}_t-\sum_{k=1}^{K}\mathbf{w}_{k,t}\odot\tilde{\mathbf{y}}_{k,t}\right)'\Sigma^{-1}\left(\mathbf{y}_t-\sum_{k=1}^{K}\mathbf{w}_{k,t}\odot\tilde{\mathbf{y}}_{k,t}\right)\right\} \tag{5}$$

where $\mathbf{w}_{k,t}=(w_{k,t}^1,\ldots,w_{k,t}^L)'$ is a weights vector and $\odot$ is the Hadamard's product. The system of weights is given as $\mathbf{w}_t^l=(w_{1,t}^l,\ldots,w_{L,t}^l)'\in\Delta_{[0,1]^L}$, for $l=1,\ldots,L$. In this model the weights may vary over the elements of $\mathbf{y}_t$ and only the i-th elements of each predictor $\tilde{\mathbf{y}}_{k,t}$ of $\mathbf{y}_t$ are combined in order to have a prediction of the i-th element of $\mathbf{y}_t$.

Special cases of model combinations are given in the Appendix.

3 Weight Dynamics

In this section we discuss the specification of the weight conditional density, $p(W_t|W_{t-1},\mathbf{y}_{1:t-1},\tilde{\mathbf{y}}_{1:t-1})$, appearing in (3). First, we introduce a vector of latent processes $\mathbf{x}_t=\text{vec}(X_t)\in\mathbb{R}^{KL^2}$ where $X_t=(\mathbf{x}_t^1,\ldots,\mathbf{x}_t^L)'$ and $\mathbf{x}_t^l=(x_{1,t}^l,\ldots,x_{KL,t}^l)'\in\mathcal{X}\subset\mathbb{R}^{KL}$. Next, for the l-th predicted variables of the vector $\mathbf{y}_t$, in order to have weights $\mathbf{w}_t^l$ which belong to the simplex $\Delta_{[0,1]^{KL}}$, we introduce the multivariate transform $\mathbf{g}=(g_1,\ldots,g_{KL})'$

$$\mathbf{g}:\left[\begin{array}{ll}\mathbb{R}^{KL} & \to\Delta_{[0,1]^{KL}}\\ \mathbf{x}_t^l & \mapsto\mathbf{w}_t=(g_1(\mathbf{x}_t^l),\ldots,g_{KL}(\mathbf{x}_t^l))'\end{array}\right. \tag{6}$$

Under this convexity constraint, the weights can be interpreted as discrete probabilities over the set of predictors. A hypothesis on the specific values of the weights can be tested by using their random distributions.

In the simple case of a constant-weights combination scheme the latent process is simply $x_{h,t}^l=x_h^l$, $\forall t$, where $x_h^l\in\mathbb{R}$ is a set of predictor-specific parameters. The weights can be written as: $w_h^l=g_h(\mathbf{x}^l)$ for each $l=1,\ldots,L$, where

$$g_h(\mathbf{x}^l)=\frac{\exp\{x_h^l\}}{\sum_{j=1}^{KL}\exp\{x_j^l\}},\quad\text{with}\,h=1,\ldots,KL \tag{7}$$

is the multivariate logistic transform. In standard Bayesian model averaging, $\mathbf{x}^l$ is equal to the marginal likelihood, see, e.g. Hoeting et al. [1999]. Geweke and Whiteman [2006] propose to use the logarithm of the predictive likelihood, see, e.g. Hoogerheide et al. [2010] for further details. Mitchell and Hall [2005] discuss the relationship of the predictive likelihood to the Kullback-Leibler information criterion. We note that such weights assume that the model set is complete and the true DGP can be observed or approximated by a combination of different models.

3.1 Time-varying Weights

If parameters are estimated recursively over time, say, using Kalman Filters then this creates substantial flexibility in dynamic adjustment. Following the same idea we define for the latent random vector $\mathbf{x}_h^l$ a stochastic process that accounts for the time variation of the weight estimates. In our first specification of W_t, we assume that the weights have their fluctuations generated by the latent process

$$\mathbf{x}_t \sim p(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{8}$$

with a non-degenerate distribution and then apply the transform $\mathbf{g}$ defined in Eq. (6)

$$\mathbf{w}_t^l = \mathbf{g}(\mathbf{x}_t^l), \quad l = 1, \ldots, L \tag{9}$$

where $\mathbf{w}_t^l = (w_{1,t}^l, \ldots, w_{KL,t}^l) \in \Delta_{[0,1]^{KL}}$ is the l-th row of W_t. Note that this prior specification is a special case of the transition density, $p(W_t|W_{t-1}, \mathbf{y}_{1:t-1}, \tilde{\mathbf{y}}_{1:t-1})$, appearing in Eq. (3), where we assume the model weights do not depend on the past values $\tilde{\mathbf{y}}_{1:t-1}$ of the predictors and $\mathbf{y}_{1:t-1}$ of the observables.

Example 1 - (*Logistic-Transformed Gaussian Weights*)
We assume that the conditional density function of $\mathbf{x}_t$ is a Gaussian one

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) \propto \exp\left\{-\frac{1}{2}\left(\mathbf{x}_t - \mathbf{x}_{t-1}\right)' \Lambda^{-1} \left(\mathbf{x}_t - \mathbf{x}_{t-1}\right)\right\} \tag{10}$$

where Λ is the covariance matrix and the weights are logistic transforms of the latent process

$$w_{h,t}^l = \frac{\exp\{x_{h,t}^l\}}{\sum_{j=1}^{KL} \exp\{x_{j,t}^l\}}, \quad h = 1, \ldots, KL, \quad l = 1, \ldots, L$$

We note that the density functions of the weights $\mathbf{w}_t^l$ is not of a known form and will be computed by a nonlinear filtering method, see section 4.

■

3.2 Learning Mechanism

We generalize the weight structures given above and in related literature (see for example Hoogerheide et al. [2010]) by including a learning strategy in the weight dynamics and by estimating these weights using nonlinear filtering (see also Branch [2004] for a discussion of learning mechanism in macroeconomic forecasting). Our weights are explicitly driven by the past and current forecast errors and capture the residual evolution of the combination scheme. Instead of choosing between the use of exponential discounting in the weight dynamics or time-varying random weights (see Diebold and Pauly [1987] and for an updated review Timmermann [2006]), we combine the two approaches.

We consider an exponentially weighted moving average of the forecast errors of the different predictors. In this way it is possible to have at the same time a better estimate of the current distribution of the prediction error and to attribute greater importance to the most recent prediction error. We consider a moving window of τ observations and define the distance vector $\mathbf{e}_t^l = (e_{1,t}^l, \ldots, e_{KL,t}^l)'$, where

$$e_{K(l-1)+k,t}^l = (1-\lambda) \sum_{i=1}^{\tau} \lambda^{i-1} f\left(y_{t-i}^l, \tilde{y}_{k,t-i}^l\right), \quad k = 1, \ldots, K, \quad l = 1, \ldots, L \tag{11}$$

is an exponentially weighted average of forecast errors, with $\lambda \in (0,1)$ a smoothing parameter and $f(y, \tilde{y})$ a measure of the forecast error. In this paper we consider the distribution of the quadratic errors, approximated through i.i.d. draws from the predictive density of $y_{k,t}^l$. Note that other forecast measures proposed in the literature, such as utility-based measure or predictive log score, could be used in our combination approach with learning. Define $\mathbf{e}_t = \text{vec}(E_t)$, where $E_t = (\mathbf{e}_t^1, \ldots, \mathbf{e}_t^L)$, then

we specify the following relationship between combination weights and predictors

$$\mathbf{w}_t^l = \mathbf{g}(\mathbf{x}_t^l), \quad l = 1, \ldots, L \tag{12}$$

$$\mathbf{x}_t \sim p(\mathbf{x}_t | \mathbf{x}_{t-1}, \Delta\mathbf{e}_t) \tag{13}$$

where $\Delta\mathbf{e}_t = \mathbf{e}_t - \mathbf{e}_{t-1}$. In this way, we include the exponentially weighted learning strategy into the weight dynamics and estimate the density of $\mathbf{x}_t$ accounting for the density of the conditional square forecast errors $p_\lambda(\mathbf{e}_{h,t}^l | \widetilde{y}_{h,t-\tau:t-1}^l, y_{t-\tau:t-1}^l)$ induced by Eq. (11). We emphasize that for the l-th variable in the model, with $l = 1, \ldots, L$, an increase at time t in the average of the square forecasting errors implies a reduction in the value of the weight associated with the $h-$th predictor in the predictive density for the l-th variables in $\mathbf{y}_t$. Thus in the specification of the weights density we assume that the conditional mean is an increasing function of $\Delta\mathbf{e}_t$. One possible choice of the weight density is given in the following example.

Example 2 - (*Logistic-Gaussian Weights (continued)*)
Let $\mathbf{w}_t^l = \mathbf{g}(\mathbf{x}_t^l)$, with $l = 1, \ldots, L$, we assume that the distribution of $\mathbf{x}_t$ conditional on the prediction errors is

$$p(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{y}_{t-\tau:t-1}, \tilde{\mathbf{y}}_{t-\tau:t-1}) \propto \exp\left\{ -\frac{1}{2} \left(\mathbf{x}_t - \mathbf{x}_{t-1} + \Delta\mathbf{e}_t \right)' \Lambda^{-1} \left(\mathbf{x}_t - \mathbf{x}_{t-1} + \Delta\mathbf{e}_t \right) \right\} \tag{14}$$

■

Note that, the above specification of the weight dynamics with learning leads to a special case of the transition density $p(W_t | W_{t-1}, \mathbf{y}_{1:t-1}, \tilde{\mathbf{y}}_{1:t-1})$ of Eq. (3), where we assume that the weight dynamics depend on the recent values of the predictors and observables, i.e. $p(W_t | W_{t-1}, \mathbf{y}_{1:t-1}, \tilde{\mathbf{y}}_{1:t-1}) = p(W_t | W_{t-1}, \mathbf{y}_{t-\tau:t-1}, \tilde{\mathbf{y}}_{t-\tau:t-1}), \quad \tau > 0$. Under these assumptions, the first integral in Eq. (2) simplifies as it is now defined on the set $\mathcal{Y}^{K(\tau+1)}$ and is taken with respect to the probability measure that has $p(\tilde{\mathbf{y}}_{t-\tau:t} | \mathbf{y}_{1:t-1})$ as joint predictive density. As a final remark, note that the weight dynamics do not include information about the predictive density $p(\tilde{\mathbf{y}}_t | \mathbf{y}_{1:t-1})$, such as the correlation between the predictors, which is available at time t. Our combination approach can be extended to include such a piece of information, when the researcher thinks it plays a crucial role in the forecasting problem.

Summary of the applied combination scheme

In the simulation and empirical exercises we will apply a Gaussian combination scheme with logistic-transformed Gaussian weights with and without learning. The scheme is specified as:

$$p(\mathbf{y}_t|W_t, \tilde{\mathbf{y}}_t) \propto \exp\left\{-\frac{1}{2}\left(\mathbf{y}_t - W_t\tilde{\mathbf{y}}_t\right)' \Sigma^{-1}\left(\mathbf{y}_t - W_t\tilde{\mathbf{y}}_t\right)\right\}$$

where $\mathbf{w}_t^l$, $l = 1, \ldots, L$ elements of W_t; and

$$w_{h,t}^l = \frac{\exp\{x_{h,t}^l\}}{\sum_{j=1}^{KL}\exp\{x_{j,t}^l\}}, \quad \text{with } h = 1, \ldots, KL$$

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) \propto \exp\left\{-\frac{1}{2}\left(\mathbf{x}_t - \mathbf{x}_{t-1}\right)' \Lambda^{-1}\left(\mathbf{x}_t - \mathbf{x}_{t-1}\right)\right\}$$

with $\mathbf{x}_t = \text{vec}(X_t) \in \mathbb{R}^{KL^2}$ where $X_t = (\mathbf{x}_t^1, \ldots, \mathbf{x}_t^L)'$ and extended with learning as:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_{t-\tau:t-1}, \tilde{\mathbf{y}}_{t-\tau:t-1}) \propto \exp\left\{-\frac{1}{2}\left(\mathbf{x}_t - \mathbf{x}_{t-1} + \Delta\mathbf{e}_t\right)' \Lambda^{-1}\left(\mathbf{x}_t - \mathbf{x}_{t-1} + \Delta\mathbf{e}_t\right)\right\}$$

4 Non-linear Filtering and Prediction

As already noted in section 2.3, the proposed general distributional representation allows us to represent the density of observable variables, conditional on the combination scheme, on the predictions and on combination weights, as a nonlinear and possibly non-Gaussian state-space model. In the following we consider a general state space representation and show how Sequential Monte Carlo methods can be used to approximate the filtering and predictive densities.

Let $\mathcal{F}_t = \sigma(\{\mathbf{y}_s\}_{s\leq t})$ be the σ-algebra generated by the observable process and assume that the predictors $\tilde{\mathbf{y}}_t = (\tilde{\mathbf{y}}'_{1,t}, \ldots, \tilde{\mathbf{y}}'_{K,t})' \in \mathcal{Y} \subset \mathbb{R}^{KL}$ stem from a $\mathcal{F}_{t-1}$-measurable stochastic process associated with the predictive densities of the K different models in the pool. Let $\mathbf{w}_t = (\mathbf{w}'_{1,t}, \ldots \mathbf{w}'_{K,t})' \in \mathcal{X} \subset \mathbb{R}^{KL}$ be the vector of latent variables (i.e. the model weights) associated with $\tilde{\mathbf{y}}_t$ and $\boldsymbol{\theta} \in \Theta$ the parameter vector of the predictive model. Include the parameter vector into the state vector and thus define the augmented state vector $\mathbf{z}_t = (\mathbf{w}_t, \boldsymbol{\theta}) \in \mathcal{X} \times \Theta$. The distributional state space form of the

forecast model is

$$\mathbf{y}_t|\mathbf{z}_t,\tilde{\mathbf{y}}_t \sim p(\mathbf{y}_t|\mathbf{z}_t,\tilde{\mathbf{y}}_t) \tag{15}$$

$$\mathbf{z}_t|\mathbf{z}_{t-1},\mathbf{y}_{1:t-1},\tilde{\mathbf{y}}_{1:t-1} \sim p(\mathbf{z}_t|\mathbf{z}_{t-1},\mathbf{y}_{1:t-1},\tilde{\mathbf{y}}_{1:t-1}) \tag{16}$$

$$\mathbf{z}_0 \sim p(\mathbf{z}_0) \tag{17}$$

The hidden state predictive and filtering densities conditional on the predictive variables $\tilde{\mathbf{y}}_{1:t}$ are

$$p(\mathbf{z}_{t+1}|\mathbf{y}_{1:t},\tilde{\mathbf{y}}_{1:t}) = \int_{\mathcal{X}} p(\mathbf{z}_{t+1}|\mathbf{z}_t,\mathbf{y}_{1:t},\tilde{\mathbf{y}}_{1:t})p(\mathbf{z}_t|\mathbf{y}_{1:t},\tilde{\mathbf{y}}_{1:t})d\mathbf{z}_t \tag{18}$$

$$p(\mathbf{z}_{t+1}|\mathbf{y}_{1:t+1},\tilde{\mathbf{y}}_{1:t+1}) \propto p(\mathbf{y}_{t+1}|\mathbf{z}_{t+1},\tilde{\mathbf{y}}_{t+1})p(\mathbf{z}_{t+1}|\mathbf{y}_{1:t},\tilde{\mathbf{y}}_{1:t}) \tag{19}$$

which represent the optimal nonlinear filter (see Doucet et al. [2001]). The marginal predictive density of the observable variables is then

$$\begin{aligned} p(\mathbf{y}_{t+1}|\mathbf{y}_{1:t}) &= \int_{\mathcal{X}\times\mathcal{Y}^{t+1}} p(\mathbf{y}_{t+1}|\mathbf{z}_{t+1},\tilde{\mathbf{y}}_{t+1})p(\mathbf{z}_{t+1}|\mathbf{y}_{1:t},\tilde{\mathbf{y}}_{1:t})p(\tilde{\mathbf{y}}_{1:t+1}|\mathbf{y}_{1:t})d\mathbf{z}_{t+1}d\tilde{\mathbf{y}}_{1:t+1} \\ &= \int_{\mathcal{Y}} p(\mathbf{y}_{t+1}|\mathbf{y}_{1:t},\tilde{\mathbf{y}}_{t+1})p(\tilde{\mathbf{y}}_{t+1}|\mathbf{y}_{1:t})d\tilde{\mathbf{y}}_{t+1} \end{aligned}$$

where

$$p(\mathbf{y}_{t+1}|\mathbf{y}_{1:t},\tilde{\mathbf{y}}_{t+1}) = \int_{\mathcal{X}\times\mathcal{Y}^{t}} p(\mathbf{y}_{t+1}|\mathbf{z}_{t+1},\tilde{\mathbf{y}}_{t+1})p(\mathbf{z}_{t+1}|\mathbf{y}_{1:t},\tilde{\mathbf{y}}_{1:t})p(\tilde{\mathbf{y}}_{1:t}|\mathbf{y}_{1:t-1})d\mathbf{z}_{t+1}d\tilde{\mathbf{y}}_{1:t}$$

is the conditional predictive density of the observable given the predicted variables.

To construct an optimal nonlinear filter we have to implement the exact update and prediction steps given above. As an analytical solution of the general filtering and prediction problems is not known for non-linear state space models, we apply an optimal numerical approximation method, that converges to the optimal filter in Hilbert metric, in the total variation norm or in a weaker distance suitable for random probability distributions (e.g., see Legland and Oudjane [2004]). More specifically we consider a sequential Monte Carlo (SMC) approach to filtering. See Doucet et al. [2001] for an introduction to SMC and Creal [2009] for a recent survey on SMC in economics. Let $\Xi_t = \{\mathbf{z}_t^i, \omega_t^i\}_{i=1}^N$ be a set of particles, then the basic SMC algorithm uses the particle set to approximate the prediction

and filtering densities with the empirical prediction and filtering densities, which are defined as

$$p_N(\mathbf{z}_{t+1}|\mathbf{y}_{1:t}, \tilde{\mathbf{y}}_{1:t}) = \sum_{i=1}^{N} p(\mathbf{z}_{t+1}|\mathbf{z}_t, \mathbf{y}_{1:t}, \tilde{\mathbf{y}}_{1:t})\omega_t^i \delta_{\mathbf{z}_t^i}(\mathbf{z}_t) \tag{20}$$

$$p_N(\mathbf{z}_{t+1}|\mathbf{y}_{1:t+1}, \tilde{\mathbf{y}}_{1:t+1}) = \sum_{i=1}^{N} \omega_{t+1}^i \delta_{\mathbf{z}_{t+1}^i}(\mathbf{z}_{t+1}) \tag{21}$$

respectively, where $\omega_{t+1}^i \propto \omega_t^i p(\mathbf{y}_{t+1}|\mathbf{z}_{t+1}^i, \tilde{\mathbf{y}}_{t+1})$ and $\delta_x(y)$ denotes the Dirac mass centered at x. The hidden state predictive density can be used to approximate the observable prediction density as follows

$$p_N(\mathbf{y}_{t+1}|\mathbf{y}_{1:t}, \tilde{\mathbf{y}}_{1:t+1}) = \sum_{i=1}^{N} \omega_t^i \delta_{\mathbf{y}_{t+1}^i}(\mathbf{y}_{t+1}) \tag{22}$$

where $\mathbf{y}_{t+1}^i$ has been simulated from the measurement density $p(\mathbf{y}_{t+1}|\mathbf{z}_{t+1}^i, \tilde{\mathbf{y}}_{t+1})$ independently for $i = 1, \ldots, N$. For the applications in the present paper we use a regularized version of the SMC procedure given above (see Liu and West [2001] and Musso et al. [2001]). Moreover we assume that the densities $p(\tilde{\mathbf{y}}_s|\mathbf{y}_{1:s-1})$ are discrete

$$p(\tilde{\mathbf{y}}_s|\mathbf{y}_{1:s-1}) = \sum_{j=1}^{M} \delta_{\tilde{\mathbf{y}}_s^j}(\tilde{\mathbf{y}}_s)$$

This assumption does not alter the validity of our approach and is mainly motivated by the forecasting practice, see literature on model pooling, e.g. Jore et al. [2010]. In fact, the predictions usually come from different models or sources. In some cases the discrete prediction density is the result of a collection of point forecasts from many subjects, such as surveys forecasts. In other cases the discrete predictive is a result of a Monte Carlo approximation of the predictive density (e.g. Importance Sampling or Markov-Chain Monte Carlo approximations).

Under this assumption it is possible to approximate the marginal predictive density by the following steps. First, draw M independent values $\tilde{\mathbf{y}}_{1:t+1}^j$, with $j = 1, \ldots, M$ from the sequence of predictive densities $p(\tilde{\mathbf{y}}_{s+1}|\mathbf{y}_{1:s})$, with $s = 1, \ldots, t$. Secondly, apply the SMC algorithm, conditionally on $\tilde{\mathbf{y}}_{1:t+1}^j$, in order to generate the particle set $\Xi_t^{i,j} = \{\mathbf{z}_{1:t}^{i,j}, \omega_t^{i,j}\}_{i=1}^N$, with $j = 1, \ldots, M$. At the last step, simulate $\mathbf{y}_{t+1}^{i,j}$, $i = 1, \ldots, N$ and $j = 1, \ldots, M$, from $p(\mathbf{y}_{t+1}|\mathbf{z}_{t+1}^{i,j}, \tilde{\mathbf{y}}_{t+1}^j)$ and obtain the following empirical

predictive density

$$p_{M,N}(\mathbf{y}_{t+1}|\mathbf{y}_{1:t}) = \frac{1}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N}\omega_t^{i,j}\delta_{\mathbf{y}_{t+1}^{i,j}}(\mathbf{y}_{t+1}) \tag{23}$$

5 Experiments using simulated data

5.1 Complete and incomplete model sets

Using simulated data we start to study the ability of the nonlinear filtering procedure to select the true model, when the model set is complete. Next, we study the behavior of both weights and residuals for an incomplete set. We do consider models that are similar and belong to the class of Gaussian, linear autoregressive models. This class is widely studies in the forecasting literature (e.g., see see Clements and Hendry [1998], Patton and Timmermann [2012] for an extension to testing using inequality constraints and Hoogerheide et al. [2012] to include risk measures).

We run two sets of experiments. In the first set, we have three linear stationary autoregressive (AR) models with different unconditional means (UM), i.e.

$$\mathcal{M}_1: \quad y_{1t} = 0.1 + 0.6y_{1t-1} + \varepsilon_{1t} \tag{24}$$

$$\mathcal{M}_2: \quad y_{2t} = 0.3 + 0.2y_{2t-2} + \varepsilon_{2t} \tag{25}$$

$$\mathcal{M}_3: \quad y_{3t} = 0.5 + 0.1y_{3t-1} + \varepsilon_{3t} \tag{26}$$

with $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$, $t = 1,\ldots,T$, independent for $i = 1,2,3$ and assume $y_{i0} = 0.25$, $i = 1,2,3$ and $\sigma = 0.05$. Note that, as we generate data from model $\mathcal{M}_1$, which is the true model, then in this experiment we have two biased predictors, $\mathcal{M}_2$ and $\mathcal{M}_3$ and one unbiased predictor $\mathcal{M}_1$. Moreover, the three models differ in terms of persistence patterns in the autoregression. The true model has UM=0.25 and the series is moderately autoregressive with root 10/6. Model $\mathcal{M}_2$ has a different intercept, autoregressive coefficient and lag structure. It has UM = 0.375 and the series is more close to normal white noise with a root equal to $\sqrt{10/2}$. Model $\mathcal{M}_3$ has the same lag structure as the true model, but different intercept and autoregressive coefficient. It has UM = 0.56 and the series is really close to white noise: the root is 10.

In the second set of experiments, we consider three stationary autoregressive processes with equal means. The two processes have almost the same roots. Specifically, let $\mathcal{M}_1$ be defined as in the

previous section and

$$\mathcal{M}_2: \quad y_{2t} = 0.125 + 0.5y_{2t-2} + \varepsilon_{2t} \tag{27}$$

$$\mathcal{M}_3: \quad y_{3t} = 0.2 + 0.2y_{3t-1} + \varepsilon_{3t} \tag{28}$$

with $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ independent for $i = 1, 2, 3$. Model $\mathcal{M}_1$ has UM = 0.25 and is moderately autoregressive, with unconditional variance (UV) equal to 0.0039. Model $\mathcal{M}_2$ has UM = 0.25 and is moderately autoregressive with UV=0.0033. Finally, Model $\mathcal{M}_3$ has UM = 0.25 and is close to white noise with UV=0.0026. Models $\mathcal{M}_2$ and $\mathcal{M}_3$ have the same UM as the one of the true model, and are similar to it in terms of unconditional variance. We thus consider three unbiased predictors where two are even almost equal in persistence pattern and close in terms of unconditional variance.

In the two sets of experiments, we generate a random sequence y_{1t}, $t = 1, \ldots, T$, with $T = 100$, from $\mathcal{M}_1$ and set $y_t = y_{1t}$, assume that the model set is complete and apply our density combination method. We specify the following combination scheme

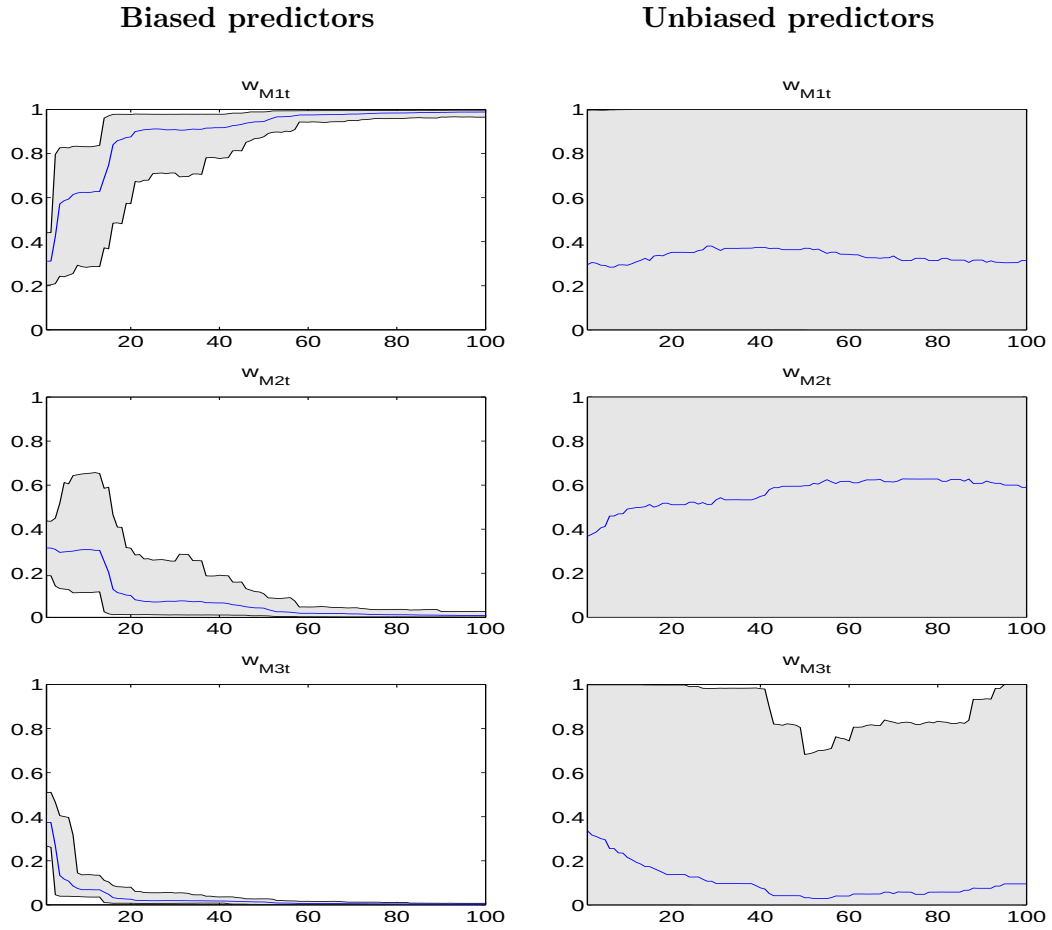
$$p(y_t|\tilde{y}_t) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y_t - \sum_{i=1}^{3} w_{it}\tilde{y}_{it}\right)^2\right\} \tag{29}$$

where $\tilde{y}_{it}$ are forecast for y_t generated at time $t-1$ from the different models and $\tilde{y}_t = (\tilde{y}_{1t}, \tilde{y}_{2t}, \tilde{y}_{3t})'$. As regards the probabilities, w_{it}, for the model index $i = 1, 2, 3$, we assume that the vector $\mathbf{w}_t = (w_{1t}, w_{2t}, w_{3t})'$ is a multivariate logistic transform, φ, of the latent process $\mathbf{x}_t = (x_{1t}, x_{2t}, x_{3t})'$ (see section 3) and consider independent random walk processes for x_{it}, $i = 1, 2, 3$ for updating. We assume the initial value of the weights is known and set it equals to $w_{it} = 1/3$, $i = 1, 2, 3,$.

We apply a sequential Monte Carlo (SMC) approximation to the filtering and predictive densities (see Appendix B) and find optimal weights (see blue lines in the left column of Fig. 1) and their credibility regions (gray areas in the same figure) for the three models.

In the first experiment, after some iterations the weight of the model $\mathcal{M}_1$ converges to one and the weights of the other models converge to zero. The credibility region for w_{1t} does not overlap with the credibility regions of the other weights. This leads us to conclude that it is credible that the weights are different in our simulation experiment. Note that we used different random sequences simulated from the true model and different random numbers for the SMC algorithm and find the same results.

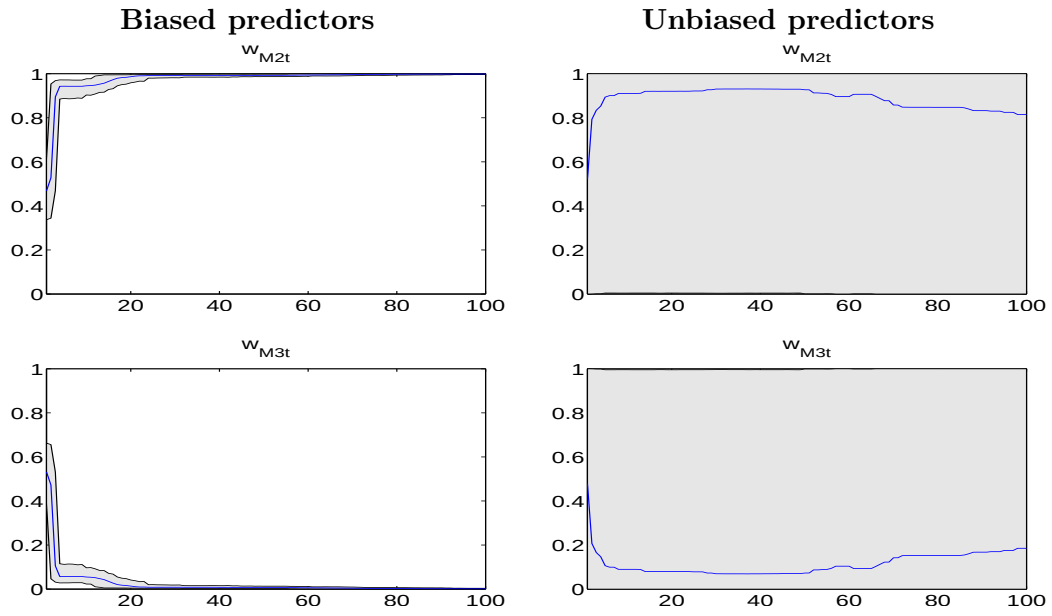
Figure 1: Filtered model probability weights, when the true model is $\mathcal{M}_1 : y_{1t} = 0.1 + 0.6y_{1t-1} + \varepsilon_{1t}$. Left: results for a complete model set in presence of biased predictors: $\mathcal{M}_2 : y_{2t} = 0.3 + 0.2y_{2t-2} + \varepsilon_{2t}$ and $\mathcal{M}_3 : y_{3t} = 0.5 + 0.1y_{3t-1} + \varepsilon_{3t}$, with $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, $t = 1, \ldots, T$. Right: results for a complete model set in presence of unbiased predictors: $\mathcal{M}_2 : y_{2t} = 0.125 + 0.5y_{2t-2} + \varepsilon_{2t}$ and $\mathcal{M}_3 : y_{3t} = 0.2 + 0.2y_{3t-1} + \varepsilon_{3t}$. Model weights (blue line) and 95% credibility region (gray area) for models 1,2 and 3 (different rows).



On the same simulated dataset we apply our optimal combination scheme to an incomplete set of models and find the optimal weights presented in the left column of Fig. 2. The weight of the model $\mathcal{M}_3$ converges to one, while $\mathcal{M}_2$ has weight converging to zero. Note that for the incomplete set the variance of the residuals is larger than the variance for the complete set (see left column of Fig. 3).

In the second experiment the credibility regions of the model weights are given in the right column of Fig. 1 for the complete model set and in the right column of Fig. 2 for the incomplete model set. Both experiments show that the weights have a high variability. This leads us to conclude that

Figure 2: Filtered combination weights for the incomplete model set, in presence of biased (left): $\mathcal{M}_2 : y_{2t} = 0.3 + 0.2y_{2t-2} + \varepsilon_{2t}$ and $\mathcal{M}_3 : y_{3t} = 0.5 + 0.1y_{3t-1} + \varepsilon_{3t}$, with $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, $t = 1, \ldots, T$ and unbiased (right): $\mathcal{M}_2 : y_{2t} = 0.125 + 0.5y_{2t-2} + \varepsilon_{2t}$, $\mathcal{M}_3 : y_{3t} = 0.2 + 0.2y_{3t-1} + \varepsilon_{3t}$, predictors. Model weights (blue line) and 95% credibility region (gray area) for models 2 and 3 (different rows).



the three models in the complete set have the same weights. The same conclusion holds true for the incomplete set.

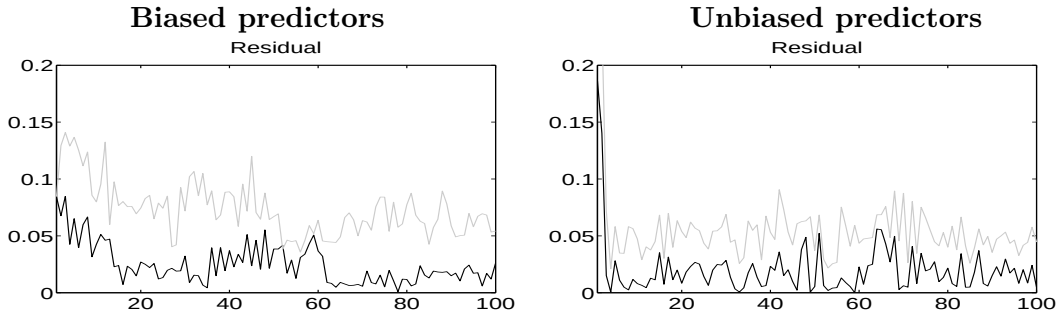
Nevertheless, from the analysis of the residuals it is evident that differences in the fit of the two model combinations exist. In fact, for the incomplete set the residuals have a larger variance than the residuals for the complete set (see right column of Fig. 3).

In conclusion, our simulation experiments enable us to interpret the behavior of the weights and that of the residuals in our density forecast combination approach. More specifically, the high uncertainty level in the weights appear due to the presence of predictors that are similar in terms of unconditional mean and differ a little in terms of unconditional variance. The degree of uncertainty in the residuals reduces when the true model is in the set of combined models.

5.2 Different degrees of persistence

Next, we study the effect of varying the persistence parameter on the results presented above. Further, we show that time-varying weights with learning can account for differences in the unconditional

Figure 3: Standard deviation of the combination residuals for complete (balck line) and incomplete (gray line) model sets in presence of biased (left) and unbiased (right) predictors



predictive distribution of the different models. In our experiments, the learning mechanism produces a better discrimination between forecast models with the same unconditional mean, but with different unconditional variance.

We consider models $\mathcal{M}_2$ and $\mathcal{M}_3$ defined previously and a sequence of models $\mathcal{M}_1$ parameterized by the persistence parameter ϕ, with $\phi \in (0, 1)$. The model set include the following models

$$\mathcal{M}_1 : \quad y_{1t} = 0.1 + \phi y_{1t-1} + \varepsilon_{1t} \tag{30}$$

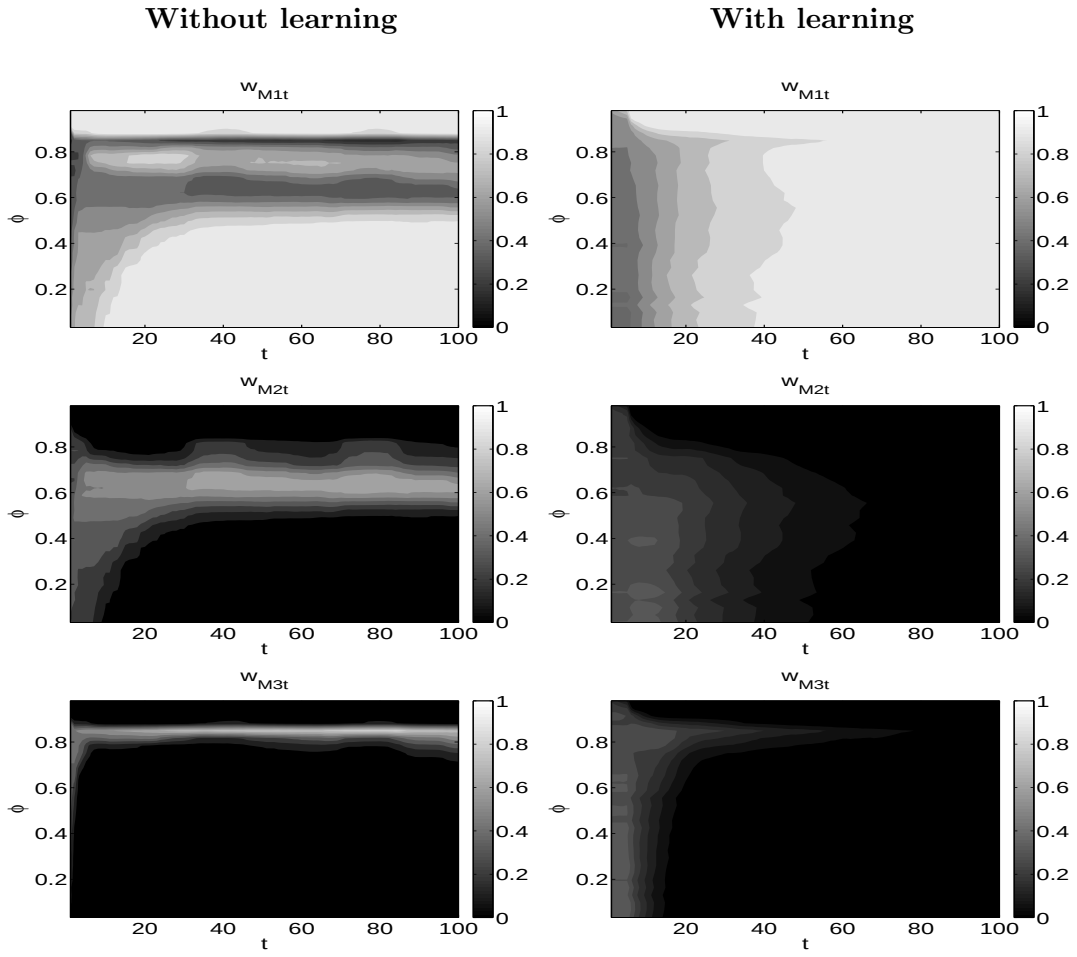
$$\mathcal{M}_2 : \quad y_{2t} = 0.125 + 0.5 y_{2t-2} + \varepsilon_{2t} \tag{31}$$

$$\mathcal{M}_3 : \quad y_{3t} = 0.5 + 0.2 y_{3t-1} + \varepsilon_{3t} \tag{32}$$

with $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, $t = 1, \ldots, T$, independent for $i = 1, 2, 3$. The unconditional mean, $0.1/(1-\phi)$, of model $\mathcal{M}_1$ is closed to the one of model $\mathcal{M}_2$, for $\phi = 0.6$, and to the one of model $\mathcal{M}_3$, for $\phi = 0.84$. For such values of the persistence parameter, the unconditional variance $\sigma^2/(1-\phi^2)$ is 0.0039 and 0.0084 respectively, and is very close to the UV of models $\mathcal{M}_2$ and $\mathcal{M}_3$, i.e. 0.0033 and 0.0026 respectively.

For different values of the persistence parameter and when ϕ is far from 0.6 and 0.84, a combination approach without learning (see filtered weights in the left column of Fig. 4) is able to detect the true model, i.e. model $\mathcal{M}_1$. In fact, the filtered weights are close to one for $\mathcal{M}_1$ and to zero for the other models. However, in that part of the parameter space where these three models share similarities in terms of predictive ability, i.e. $\phi = 0.6, 0.84$, and have the same unconditional mean, then the

Figure 4: Filtered combination weights over time and for different values of the persistence parameter $\phi \in (0,1)$ of the true model $\mathcal{M}_1 : y_{1t} = 0.1 + \phi y_{1t-1} + \varepsilon_{1t}$ with $\varepsilon_{1t} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. Left: results of the combination scheme without learning. Right: results of the combination scheme with learning in the weights dynamics.



combination weights of model $\mathcal{M}_1$ are not close to one and the weights for model $\mathcal{M}_2$ and $\mathcal{M}_3$ are not null.

We repeated the same experiments, while keeping fixed the seed of the simulated series in order to reduce the variability of the results, and apply a combination procedure with learning. The results are given in the right column of Fig. 4. These show that a learning mechanism, with parameters $\lambda = 0.6$ and $\tau = 10$, is able to discriminate between models which have the same UM but differ in terms of UV. In fact, for all values of $\phi \in (0,1)$ the weights of model $\mathcal{M}_1$ are close to one.

5.3 Linear and non-linear predictors

In the following simulation experiments we study the ability of our combination approach to discriminate between an AR with stochastic volatility (AR-SV) and an AR without SV, i.e.

$$\mathcal{M}_1: \quad y_{1t} = 0.01 + 0.02 y_{1t-1} + \sigma_t \varepsilon_{1t} \tag{33}$$

$$\mathcal{M}_2: \quad y_{2t} = 0.01 + 0.02 y_{2t-1} + \sigma \varepsilon_{2t} \tag{34}$$

with $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $t = 1, \ldots, T$, independent for $i = 1, 2$, $\sigma = 0.05$ and $\sigma_t = \exp\{h_t/2\}$, where

$$h_t = \phi + \alpha h_{t-1} + \gamma \eta_t, \quad \eta_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$$

and η_t is independent of ϵ_s, $\forall s, t$. We assume the true model is $\mathcal{M}_1$ and consider two typical parameter settings (see Casarin and Marin [2009]): low persistence in volatility, i.e. $\phi = 0.0025, \gamma = 0.1, \alpha = 0.9$ and high persistence in volatility, i.e. $\phi = 0.0025, \gamma = 0.01, \alpha = 0.99$, which can be usually found in financial applications. For each setting we simulate $T = 1000$ observations and apply the combination scheme presented in Section 2. Figure 5 shows the combination weights (black lines) and their high credibility regions (coloured areas) for the two parameter settings.

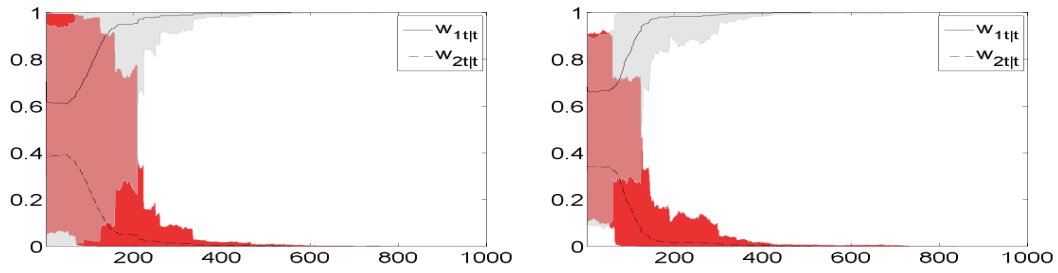
We expect that non-overlapping regions indicate a high probability that the two weights take different values. Our combination procedure is able to detect the true model assigning to it a combination weight with mean equal to one. From a comparison with the results of the previous experiments, notice that the learning period is longer than for the case in which the set includes only linear models. Finally, a comparison between the two dataset, show that in the low-persistence setting the learning about the model weights is slower than for the high-persistence setting.

5.4 Structural instability

We study the behavior of the model weights in presence of a structural break in the parameters of the data generating process. We generate a random sample from the following autoregressive model with breaks

$$y_t = 0.1 + 0.3 \mathbb{I}_{(T_0,T]}(t) + \left(0.6 - 0.4 \mathbb{I}_{(T_0,T]}(t)\right) y_{t-1} + \varepsilon_t \tag{35}$$

Figure 5: Filtered combination weights (dark lines) and high probability density region (coloured areas) for the SV-AR model, $\mathcal{M}_1 : y_{1t} = 0.01 + 0.02y_{1t-1} + \sigma_t\varepsilon_{1t}$, $\sigma_t = \exp\{h_t/2\}$, $h_t = \phi + \alpha h_{t-1} + \gamma\eta_t$, $\eta_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ (solid line) and for the AR model $\mathcal{M}_2 : y_{2t} = 0.01 + 0.02y_{2t-1} + \sigma\varepsilon_{2t}$, (dashed line), when assuming that the true model is $\mathcal{M}_1$. Left: low persistence in volatility, $\phi = 0.0025$, $\gamma = 0.1$, $\alpha = 0.9$. Right: high persistence in volatility, $\phi = 0.0025$, $\gamma = 0.01$, $\alpha = 0.99$



for $t = 1, \ldots, T$ with $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, $\sigma = 0.05$, $T_0 = 50$ and $T = 100$ and where $\mathbb{I}(z)_A$ takes a value 1 if $z \in A$ and equals 0 otherwise. We apply our combination strategy to the following set of prediction models

$$\mathcal{M}_1 : \quad y_{1t} = 0.1 + 0.6y_{1t-1} + \varepsilon_{1t} \tag{36}$$

$$\mathcal{M}_2 : \quad y_{2t} = 0.4 + 0.2y_{2t-1} + \varepsilon_{2t} \tag{37}$$

$$\mathcal{M}_3 : \quad y_{3t} = 0.9 + 0.1y_{3t-1} + \varepsilon_{3t} \tag{38}$$

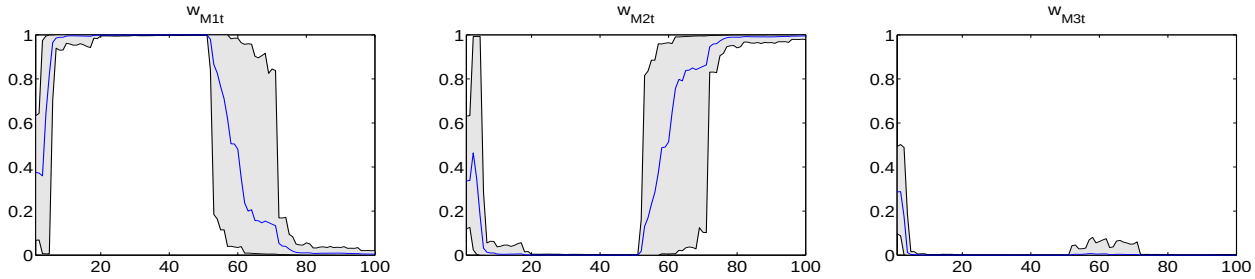
with $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ independent for $i = 1, 2, 3$ and assume $y_{i0} = 0.25$, $i = 1, 2, 3$ and $\sigma = 0.05$. Note that the model set is incomplete, but it includes two models, i.e. $\mathcal{M}_1$ and $\mathcal{M}_2$, that are equivalent stochastic version of the true model in the two parts, $t < T_0$ and $t \geq T_0$ respectively, of the sample. The results in Fig. 6 show that the combination strategy is successful in selecting with probability close to one, model $\mathcal{M}_1$ for the first part of the sample and model $\mathcal{M}_2$ in the second part.

6 Empirical Applications

6.1 Comparing Combination Schemes

To shed light on the predictive ability of individual models, we consider several evaluation statistics for point and density forecasts previously proposed in literature. We compare point forecasts in terms

Figure 6: Filtered combination weights for the three models: $\mathcal{M}_1 : y_{1t} = 0.1 + 0.6y_{1t-1} + \varepsilon_{1t}$, $\mathcal{M}_2 : y_{2t} = 0.4 + 0.2y_{2t-1} + \varepsilon_{2t}$ and $\mathcal{M}_3 : y_{3t} = 0.9 + 0.1y_{3t-1} + \varepsilon_{3t}$, with $\varepsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.05^2)$, independent for $i = 1, 2, 3$, when the parameters of the true model has a structural break at time $T_0 = 50$, i.e. $y_t = 0.1 + 0.3\mathbb{I}_{(T_0,T]}(t) + \left(0.6 - 0.4\mathbb{I}_{(T_0,T]}(t)\right) y_{t-1} + \varepsilon_t$, $t = 1, \ldots, T$ with $T = 100$ and $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.05^2)$.



of Root Mean Square Prediction Errors (RMSPE)

$$RMSPE_k = \sqrt{\frac{1}{t^*} \sum_{t=\underline{t}}^{\bar{t}} e_{k,t+1}}$$

where $t^* = \bar{t} - \underline{t} + 1$ and $e_{k,t+1}$ is the square prediction error of model k and test for substantial differences between the AR benchmark and the model k by using the Clark and West [2007]' statistic (CW). The null of the CW test is equal mean square prediction errors, the one-side alternative is the superior predictive accuracy of the model k.

We evaluate the predictive densities using two relative measures. Firstly, we consider a Kullback Leibler Information Criterion (KLIC) based measure, utilizing the expected difference in the Logarithmic Scores of the candidate forecast densities; see for example Kitamura [2002], Mitchell and Hall [2005], Hall and Mitchell [2007], Amisano and Giacomini [2007], Kascha and Ravazzolo [2010]. The KLIC chooses the model which on average gives higher probability to events that have actually occurred. Specifically, the KLIC distance between the true density $p(y_{t+1}|y_{1:t})$ of a random variable y_{t+1} and some candidate density $p(\tilde{y}_{k,t+1}|y_{1:t})$ obtained from model k is defined as

$$\begin{aligned} \text{KLIC}_{k,t+1} &= \int p(y_{t+1}|y_{1:t}) \ln \frac{p(y_{t+1}|y_{1:t})}{p(\tilde{y}_{k,t+1}|y_{1:t})} dy_{t+1}, \\ &= \mathbb{E}_t[\ln p(y_{t+1}|y_{1:t}) - \ln p(\tilde{y}_{k,t+1}|y_{1:t}))]. \end{aligned} \tag{39}$$

where $\mathbb{E}_t(\cdot) = \mathbb{E}(\cdot|\mathcal{F}_t)$ is the conditional expectation given information set $\mathcal{F}_t$ at time t. An estimate can be obtained from the average of the sample information, $y_{\underline{t}+1}, \ldots, y_{\bar{t}+1}$, on $p(y_{t+1}|y_{1:t})$ and $p(\tilde{y}_{k,t+1}|y_{1:t})$:

$$\overline{KLIC}_k = \frac{1}{t^*}\sum_{t=\underline{t}}^{\bar{t}}[\ln p(y_{t+1}|y_{1:t}) - \ln p(\tilde{y}_{k,t+1}|y_{1:t})]. \tag{40}$$

Even though we do not know the true density, we can still compare multiple densities, $p(\tilde{y}_{k,t+1}|y_{1:t})$. For the comparison of two competing models, it is sufficient to consider the Logarithmic Score (LS), which corresponds to the latter term in the above sum,

$$LS_k = -\frac{1}{t^*}\sum_{t=\underline{t}}^{\bar{t}} \ln p(\tilde{y}_{k,t+1}|y_{1:t}), \tag{41}$$

for all k and to choose the model for which the expression in (41) is minimal, or as we report in our tables, the opposite of the expression in (41) is maximal.

Secondly, we also evaluate density forecasts based on the continuous rank probability score (CRPS). This CRPS circumvents some of the drawbacks of the LS, as the latter does not reward values from the predictive density that are close but not equal to the realization (see, e.g., Gneiting and Raftery [2007]) and it is very sensitive to outliers; see Gneiting and Ranjan [2011], Groen et al. [2012] and Ravazzolo and Vahey [2012] for applications to inflation density forecasts. The CRPS for the model k measures the average absolute distance between the empirical cumulative distribution function (CDF) of y_{t+h}, which is simply a step function in y_{t+h}, and the empirical CDF that is associated with model k's predictive density:

$$\text{CRPS}_{k,t+1} = \int \left(F(z) - \mathbb{I}_{[y_{t+1},+\infty)}(z)\right)^2 \mathrm{d}z \tag{42}$$

$$= \mathbb{E}_t|\tilde{y}_{t+1,k} - y_{t+1}| - \tfrac{1}{2}\mathbb{E}_t|\tilde{y}_{t+1,k} - y'_{t+1,k}|, \tag{43}$$

where F is the CDF from the predictive density $p(\tilde{y}_{k,t+1}|y_{1:t})$ of model k and $\tilde{y}_{t+1,k}$ and $\tilde{y}'_{t+1,k}$ are independent random variables with common sampling density equal to the posterior predictive density $p(\tilde{y}_{k,t+1}|y_{1:t})$. Smaller CRPS implies higher precisions and, as for the log score, we report in tables the average $CRPS_k$ for each model k.

The distribution properties of a statistical test to compare density accuracy performances, both measured in terms of LS and CRPS, are not derived when working with nested models and expanding data window for parameter updating, such as in our exercise. Therefore, following evidence in Clark and McCracken [2012] for point forecasts, we apply the methodology in Groen et al. [2012] and test the null of equal finite sample forecast accuracy, based on either a LS and CRPS measures, *versus* the alternative that a model outperformed the AR benchmark using the Harvey et al. [1997] small sample correction of the Diebold and Mariano [1995] and West [1996] statistic to standard normal critical values.[2]

Finally, following the idea in Welch and Goyal [2008] for cumulative squared prediction error difference, and in Kascha and Ravazzolo [2010] for cumulative log score difference, we compute the cumulative rank probability score difference

$$CRPSD_{k,t+1} = \sum_{s=\underline{t}}^{t} d_{k,s+1}, \tag{44}$$

where $d_{k,s+1} = \text{CRPS}_{AR,s+1} - \text{CRPS}_{k,s+1}$. If $CRPSD_{k,t+1}$ increases at observation $t+1$, this indicates that the alternative to the AR benchmark has a lower CRPS at time $t+1$.

6.2 GDP growth and PCE inflation

We consider $K = 6$ time series models to predict US GDP growth and PCE inflation: an univariate autoregressive model of order one (AR); a bivariate vector autoregressive model for GDP and PCE, of order one (VAR); a two-state Markov-switching autoregressive model of order one (ARMS); a two-state Markov-switching vector autoregressive model of order one for GDP and inflation (VARMS); a time-varying autoregressive model with stochastic volatility (TVPARSV); and a time-varying vector autoregressive model with stochastic volatility (TVPVARSV). Therefore, our model set includes constant parameter univariate and multivariate specification; univariate and multivariate models with discrete breaks (Markov-Switiching specifications); and univariate and multivariate models with continuous breaks. See Appendix A for further details.

First we evaluate the performance of the individual models for forecasting US GDP growth and PCE inflation. Results in Table 1 indicate that the time-varying AR and VAR models with stochastic

[2]We use the left tail p-values for the CRPS based test since we minimize CRPS and right tail for the LS based test since we maximize LS.

volatility produce the most accurate point and density forecasts for both variables. Clark and Ravazzolo [2012] find similar evidence in larger VAR models applied to US and UK real-time data; see also Korobilis [2011] and D'Agostino et al. [2011].

Secondly, we apply four combination schemes. The first one is a Bayesian model averaging (BMA) approach similar to Jore et al. [2010] and Hoogerheide et al. [2010]. Following the notation in the previous section, model predictions are combined by:

$$\mathbf{y}_{t+1} = W_{t+1}\tilde{\mathbf{y}}_{t+1} \tag{45}$$

The combination is usually run independently for each series, $l = 1, .., L$. The weights W_t are computed as in (7) where $x^l_{k,t}$ is equal to the cumulative log score in (41). See, e.g., Hoogerheide et al. [2010] for further details.

The second method (BMA_{opt}) follows intuition in Hall and Mitchell [2007] and derivation in Geweke and Amisano [2010b], and computes optimal log score weights. The method maximizes the log score of the equation (45) to compute W_{t+1}:

$$\sum_{t=\underline{t}}^{\bar{t}} log(W_{t+1}\tilde{\mathbf{y}}_{t+1}) \tag{46}$$

subject to the restrictions that weights for each series $l = 1, \ldots, L$ must be positive and sum to unity.[3] See Geweke and Amisano [2010b] for further details.

The other two methods are derived from our contribution in equations from (1) to (3). We only combine the i-th predictive densities of each predictor $\tilde{\mathbf{y}}_{k,t+1}$ of $\mathbf{y}_{t+1}$ in order to have a prediction of the i-th element of $\mathbf{y}_{t+1}$ as in equation (5). One scheme consider time-varying weights (TVW) with logistic-Gaussian dynamics and without learning (see equation (10)); the other scheme computes weights with learning (TVW(λ, τ)) as in (14). Weights are estimated and predictive density computed as in section 4 using $N = 1000$ particles. Equal weights are used in all three schemes for the first forecast 1970:Q1.[4]

[3] We present results using the multivariate approach, therefore the same weight is given to each model for GDP and inflation forecasts. The multivariate joint predictive densities for the univariate models is assumed to be diagonal. Out-of-sample results are qualitative similar when combining each series independently.

[4] We also investigate a combination scheme based on equal weights but its (point and density) forecast accuracy was always lower than that both of the best individual model and of the four schemes listed above. Results are available upon request.

Table 1: Forecast accuracy for the macro application.

						GDP				
	AR	ARMS	TVPARSV	VAR	VARMS	TVPVARSV	BMA	BMA_{opt}	TVW	TVW(λ,τ)
RMSPE	0.881	0.907	0.850	0.875	1.001	0.868	0.852	0.844	0.649	**0.648**
CW		0.108	0.000	0.054	0.061	0.014	0.000	0.000	0.000	0.000
LS	-1.320	-1.405	-1.185	-1.377	-1.362	-1.225	-1.211	-1.151	-1.129	**-1.097**
p-value		0.713	0.001	0.760	0.846	0.020	0.014	0.037	0.004	0.028
CRPS	0.478	0.472	0.445	0.468	0.523	0.452	0.445	0.447	0.328	**0.328**
p-value		0.342	0.000	0.103	0.984	0.010	0.008	0.000	0.000	0.000
						Inflation				
	AR	ARMS	TVPARSV	VAR	VARMS	TVPVARSV	BMA	BMA_{opt}	TVW	TVW(λ,τ)
RMSPE	0.388	0.386	0.372	0.388	0.615	0.383	0.370	0.367	**0.260**	0.262
CW		0.034	0.001	0.172	0.077	0.053	0.003	0.001	0.000	0.000
LS	-1.541	-1.381	-0.376	-1.277	-1.091	-0.609	-0.400	-0.385	0.252	**0.223**
p-value		0.213	0.147	0.201	0.349	0.160	0.152	0.122	0.058	0.057
CRPS	0.201	0.199	0.196	0.203	0.375	0.201	0.195	0.194	0.120	**0.120**
p-value		0.327	0.166	0.731	1.000	0.480	0.115	0.093	0.000	0.000

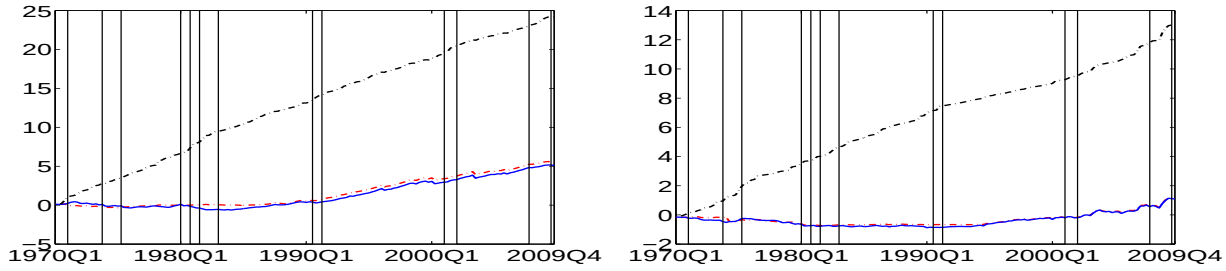
Note: AR, $ARMS$, $TVPARSV$, VAR, $VARMS$, $TVPVARSV$: individual models defined in Section 2. BMA: constant weights Bayesian Model Averaging. BMA: log pooling with optimal log score weights. TVW: time-varying weights without learning. TVW(λ,τ): time-varying weights with learning mechanism with smoothness parameter $\lambda = 0.95$ and window size $\tau = 9$. RMSPE: Root Mean Square Prediction Error. CW: p-value of the Clark and West [2007] test. LS: average Logarithmic Score over the evaluation period. CRPS: cumulative rank probability score. LS p-value and CRPS p-value: Harvey et al. [1997] type of test for LS and CRPS differentials respectively.

The results of the comparison are given in Table 1. We observe that our combination schemes both outperform BMA and the single models. In particular, the TVW(λ,τ), with smoothing factor $\lambda = 0.95$ and window size $\tau = 9$, which we mainly focus on the following analysis, outperforms the TVW model in terms of RMSPE, LS and CRPS. See section 5 for properties of such weights in simulation exercises. The values of λ and τ have been chosen on the basis of the optimal RMSPE as discussed below. Gains are substantial and up to 30%. The top panel of Fig. 10 shows that GDP density forecasts are wider than the inflation forecasts and they track accurately the realizations.[5] When comparing differentials of CRPS as shown in Fig. 7, TVW(λ,τ) outperforms for both GDP and inflation forecasting the benchmark and other density combinations all over the sample and not just for specific episodes. The graphs also show that the two other combination schemes do not always outperform the AR for inflation over the sample and optimal weights do not provide more accurate forecasts.

The optimal values for the smoothing parameters and the window size are estimated via a grid search. We set the grid $\lambda \in [0.1, 1]$ with step size 0.01 and $\tau \in \{1, 2, \ldots, 20\}$ with step size 1 and

[5]Unreported results show that all the densities are correctly specified following a Berkowitz [2001] test on PITs for GDP, but just the densities from our combinations are for inflation.

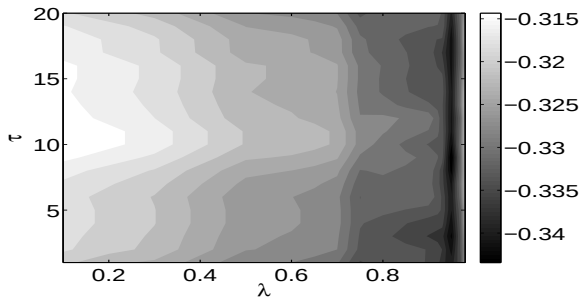
Figure 7: Cumulative Rank Probability Score Differential



GDP Inflation

Note: Left: CRPSD of the TVW(λ, τ) versus the AR model (black dashed line); CRPSD of the BMA versus the AR model (red dashed line); CRPSD of the BMA$_{opt}$ versus the AR model (blue solid line) for forecasting GDP. Right: CRPSD as in left panel for forecasting inflation.

Figure 8: Optimal combination learning parameters



Note: Root mean square prediction error (RMSPE), in logarithmic scale, of the TVW(λ, τ) scheme as a function of λ and τ. We considered $\lambda \in [0.1, 1]$ with step size 0.01 and $\tau \in \{1, 2, \ldots, 20\}$ with step size 1. Dark gray areas indicate low RMSPE.

on the GDP dataset, for each point of the grid we iterate 10 times the SMC estimation procedure and evaluate the RMSPE for forecasting GDP.[6] The level sets of the resulting approximated RMSPE surface are given in Fig. 8. A look at the RMSPE contour reveals that in our dataset, for each τ in the considered interval, the optimal value of λ is 0.95. The analysis shows that the value of τ which gives the lowest RMSPE is $\tau = 9$.

Fig. 9 shows for the TVW(λ, τ) scheme the evolution over time of the filtered weights (the average and the quantiles at the 5% and 95%) conditionally on each one of the 1,000 draws from the predictive densities. The resulting empirical distribution allows us to obtain an approximation of the predictive density accounting for both model and parameter uncertainty. The figures show that the

[6]Other accuracy measures, such as LS or CRPS, and multiple series evaluation is also possible. We leave it for further research.

weight uncertainty is enormous and inference on the model relevance neglecting it may be misleading. PCE average weights (or model average probability) are more volatile and have wider distributions than GDP average probability. The TVPARSV and TVPVARSV models have higher probability and VARMS a lower probability for both series, confirming CRPS ordering in table 1.

The residual 95% HPD plotted in the second panel of Fig. 10 represents a measure of incompleteness of the model set. Above all for GDP, the incompleteness is larger in the 70's, at beginning of 80's and in the last part of the sample during the financial crises, periods when zero does not belong the HPD region. In the central part of our sample period, often defined as the Great moderation period, standard statistical time-series models, such as the set of our models, approximate accurately the data and the incompleteness for both GDP and inflation is smaller; see section 5 for a discussion of the incompleteness properties.

Finally, our combined predictive densities can be used to nowcast recession probabilities at time t, such as those given in the last row of Fig. 10. To define them we follow a standard practice in business cycle analysis and apply the following rule

$$Pr\left(y_{t-3} < y_{t-1}, y_{t-2} < y_{t-1}, y_t < y_{t-1}, y_{t+1} < y_{t-1}\right) \tag{47}$$

where we use as y_t the GDP growth rate at time t. The estimated probabilities are approximated as follow

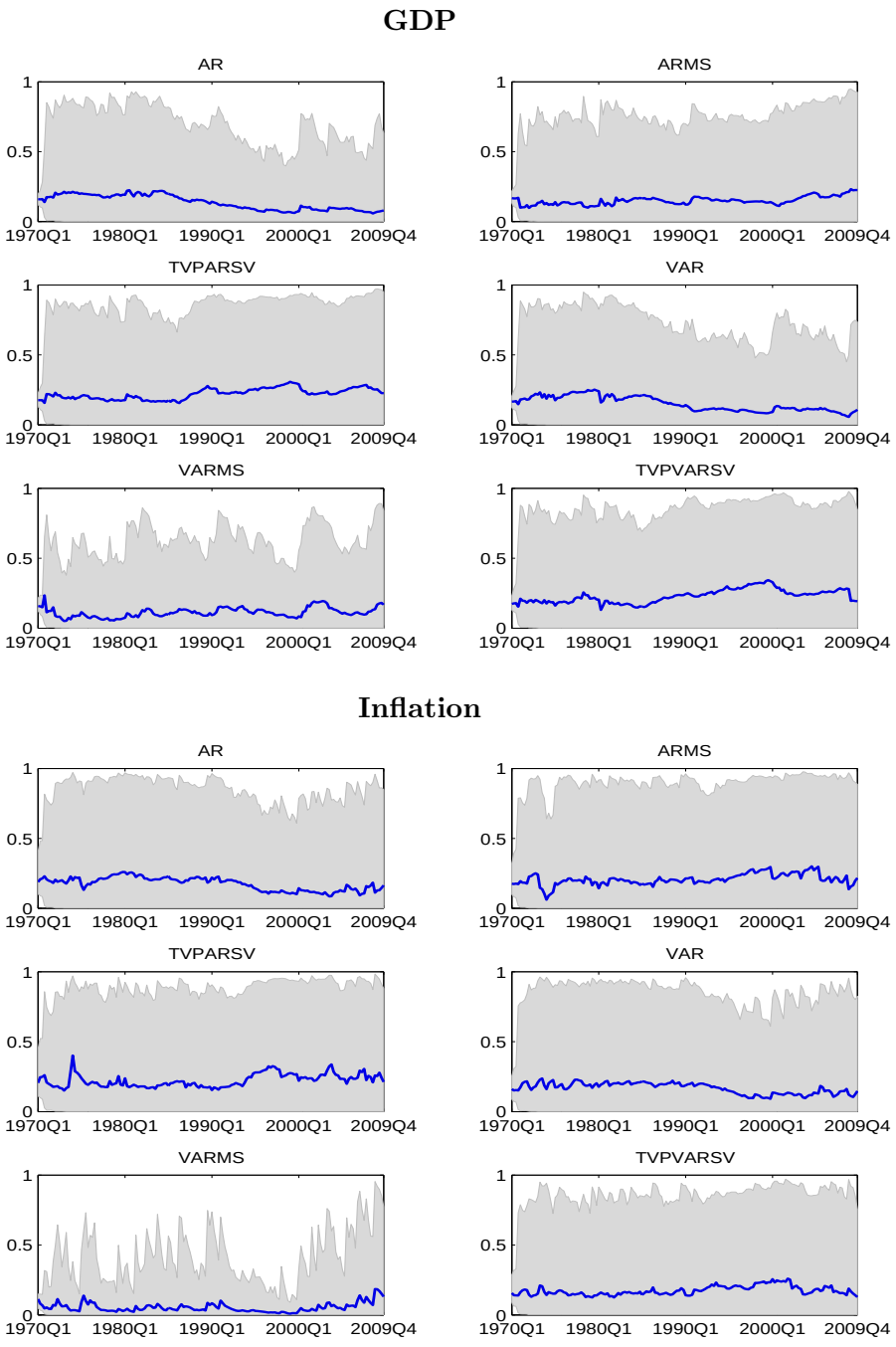
$$\frac{1}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N}\left(\mathbb{I}_{(-\infty,y_{t-1})}(y_{t-3})\mathbb{I}_{(-\infty,y_{t-1})}(y_{t-2})\mathbb{I}_{(-\infty,y_{t-1})}(y_t)\mathbb{I}_{(-\infty,y_{t-1})}(y_{t+1}^{ij})\right)$$

where y_{t+1}^{ij} is drawn by SMC from $p(\mathbf{y}_{t+1}|\mathbf{y}_{1:t})$. The estimated recession probabilities fits accurately the US business cycle and have values higher than 0.5 in each of the recessions identified by the NBER. Anyway, probabilities seems to lag at beginning of the recessions, which might be due to the use of GDP as business cycle indicator. Equation (47) could also be extended to multi-step forecasts to investigate whether timing can improve.

6.3 Returns to Standard & Poor's 500

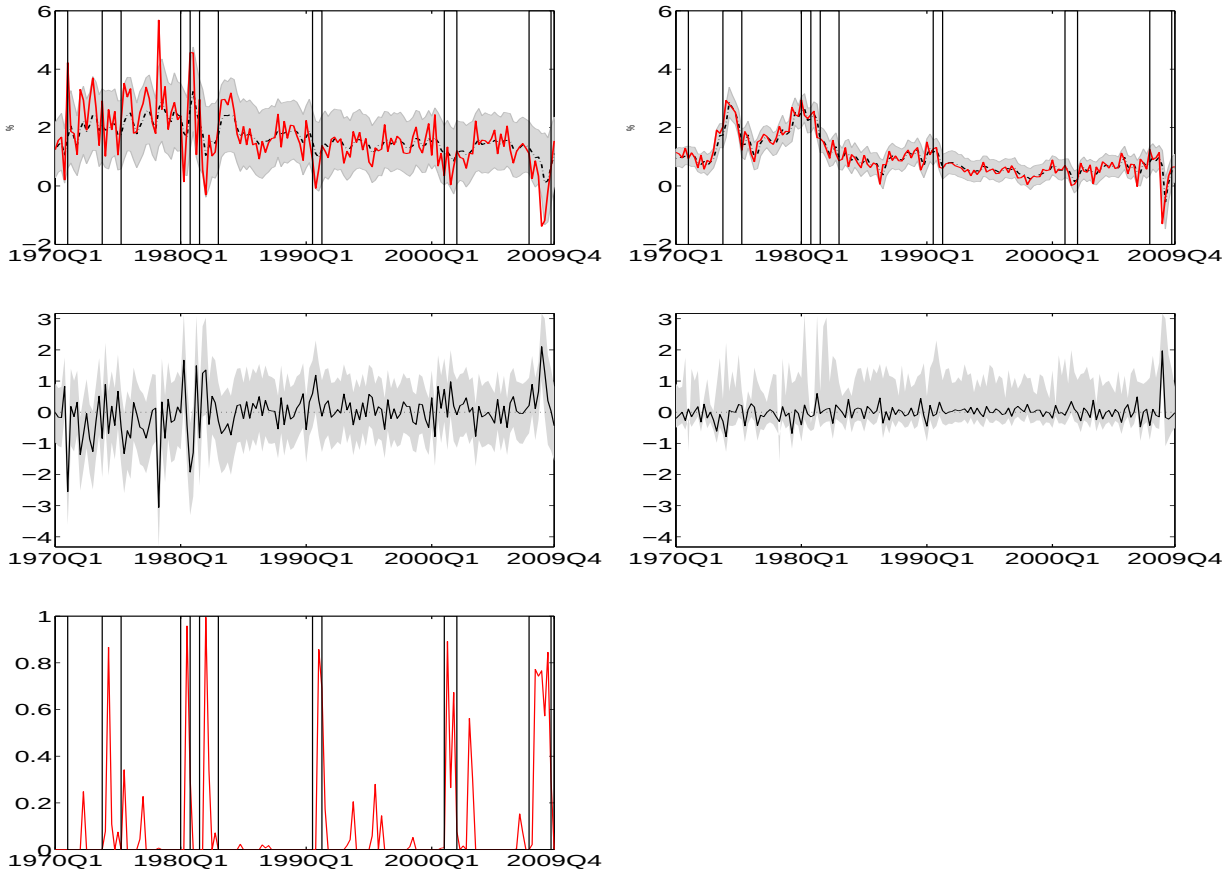
We use stock returns collected from the Livingston survey and consider a nonparametric estimated density forecasts as one possible way to predict future stock returns, see discussion in Appendix A. We

Figure 9: Time-varying weights with learning



Note: Average filtered time-varying weights with learning (solid line) with 2.5% and 97.5% quantiles (gray area). Note that the quintile are obtained using the different draws from the predictive densities.

Figure 10: Combination forecasts for the TVW(λ, τ) . Left column: GDP. Right column: Inflation.



Note: First: estimated mean (dashed line) and 2.5% and 97.5% quintile (gray area) of the marginal prediction density for $\mathbf{y}_{t+1}$. Realizations for y_{t+1} in red solid line. Second: residual mean (solid line) and residual density (gray area) of the combination scheme. Third: estimated recession probability (solid line). Vertical lines: NBER business cycle expansion and contraction dates.

call these survey forecasts (SR). The alternative is a white noise model (WN).[7] This model assumes and thus forecasts that log returns are normally distributed with mean and standard deviation equal to the unconditional (up to time t for forecasting at time $t+1$) mean and standard deviation. WN is a standard benchmark to forecast stock returns since it implies a random walk assumption for prices, which is difficult to beat (see for example Welch and Goyal [2008]). We apply our combination scheme from (1) to (3) with time-varying weights (TVW) with logistic-Gaussian dynamics and learning (see equation (10)).

Following the analysis in Hoogerheide et al. [2010] we evaluate the statistical accuracy of point forecasts, the survey forecasts and the combination schemes in terms of the root mean square error

[7] In the interest of brevity, we restrict this exercise to two individual models.

(RMSPE), and in terms of the correctly predicted percentage of sign (Sign Ratio) for the log percent stock index returns. We also evaluate the statistical accuracy of the density forecasts in terms of the LS and CRPS as in the previous section.

Moreover, as an investor is mainly interested in the economic value of a forecasting model, we develop an active short-term investment exercise, with an investment horizon of six months. The investor's portfolio consists of a stock index and risk free bonds only.[8]

At the end of each period t, the investor decides upon the fraction α_{t+1} of her portfolio to be held in stocks for the period $t+1$, based on the forecast of the stock index return. We constrain α_{t+1} to be in the $[0,1]$ interval, not allowing for short-sales or leveraging (see Barberis [2000]). The investor maximize a power utility function:

$$u(R_{t+1}) = \frac{R_{t+1}^{1-\gamma}}{1-\gamma}, \qquad \gamma > 1, \tag{48}$$

where γ is the coefficient of relative risk aversion and R_{t+1} is the wealth at time $t+1$, which is equal to

$$R_{t+1} = R_t \left((1-\alpha_{t+1})\exp(y_{f,t+1}) + \alpha_{t+1}\exp(y_{f,t+1}+\tilde{y}_{t+1})\right), \tag{49}$$

where R_t denotes initial wealth, $y_{f,t+1}$ the 1-step ahead risk free rate and $\tilde{y}_{t+1}$ the 1-step ahead forecast of the stock index return in excess of the risk free made at time t. Dangl and Halling [2012] apply time-variation directly in the individual models and use a mean-variance approach to infer the economic value of their models.

When the initial wealth is set equal to one, i.e. $R_0 = 1$, the investor's optimization problem is given by

$$\max_{\alpha_{t+1}\in[0,1]} \mathbb{E}_t \left(\frac{((1-\alpha_{t+1})\exp(y_{f,t+1}) + \alpha_{t+1}\exp(y_{f,t+1}+\tilde{y}_{t+1}))^{1-\gamma}}{1-\gamma} \right),$$

This expectation depends on the predictive density for the excess returns, $\tilde{y}_{t+1}$. Following notation in section 4, denoting this density as $p(\tilde{y}_{t+1}|y_{1:t})$, the investor solves the following problem:

$$\max_{\alpha_{t+1}\in[0,1]} \int u(R_{t+1}) p(\tilde{y}_{t+1}|y_{1:t}) d\tilde{y}_{t+1}. \tag{50}$$

[8]The risk free asset is approximated by transforming the monthly federal fund rate in the month the forecasts are produce in a six month rate. This corresponds to buying a future on the federal fund rate that pays the rate for the next six months. We collect the federal fund rate from the Fred database at the Federal Reserve Bank of St Louis.

We approximate the integral in (50) by generating with the SMC procedure MN equally weighted independent draws $\{y_{t+1}^g, w_{t+1}^g\}_{g=1}^{MN}$ from the predictive density $p(\tilde{y}_{t+1}|y_{1:t})$, and then use a numerical optimization method to find:

$$\max_{\alpha_{t+1}\in[0,1]} \frac{1}{MN}\sum_{g=1}^{MN}\left(\frac{((1-\alpha_{t+1})\exp(y_{f,t+1})+\alpha_{t+1}\exp(y_{f,t+1}+\tilde{y}_{t+1}^g))^{1-\gamma}}{1-\gamma}\right) \tag{51}$$

We consider an investor who can choose between different forecast densities of the (excess) stock return y_{t+1} to solve the optimal allocation problem described above. We include three cases in the empirical analysis below and assume the investor uses alternatively the density from the WN individual model, the empirical density from the Livingston Survey (SR) or finally a density combination (DC) of the WN and SR densities. We apply here the DC scheme used in the previous section.

We evaluate the different investment strategies by computing the *ex post* annualized mean portfolio return, the annualized standard deviation, the annualized Sharpe ratio and the total utility. Utility levels are computed by substituting the realized return of the portfolios at time $t+1$ into (48). Total utility is then obtained as the sum of $u(R_{t+1})$ across all $t^* = (\bar{t}-\underline{t}+1)$ investment periods $t = \underline{t}, \ldots, \bar{t}$, where the first investment decision is made at the end of period $\underline{t}$. We compare the wealth provided at time $t+1$ by two resulting portfolios by determining the value of multiplication factor of wealth Δ which equates their average utilities. For example, suppose we compare two strategies A and B.

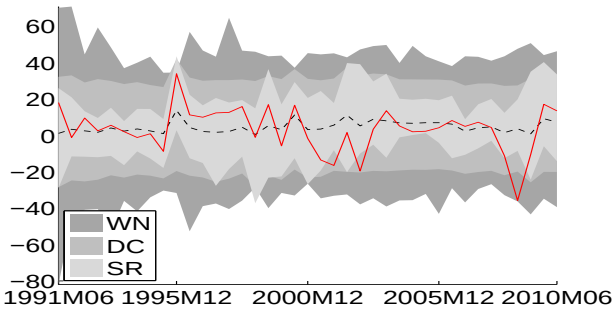
$$\sum_{t=\underline{t}}^{\bar{t}} u(R_{A,t+1}) = \sum_{t=\underline{t}}^{\bar{t}} u(R_{B,t+1}/\exp(r)). \tag{52}$$

where $u(R_{A,t+1})$ and $u(R_{B,t+1})$ are the wealth provided at time $T+1$ by the two resulting portfolios A and B, respectively. Following West et al. [1993], we interpret Δ as the maximum performance fee the investor would be willing to pay to switch from strategy A to strategy B.[9] We infer the added value of strategies based on individual models and the combination scheme by computing Δ with respect to three static benchmark strategies: holding stocks only (r_s), holding a portfolio consisting of 50% stocks and 50% bonds (r_m), and holding bonds only (r_b).

Finally, transaction costs play a non-trivial role since the portfolio weights in the active investment strategies change every period (semester), and the portfolio must be rebalanced accordingly. Rebal-

[9]See, for example, Fleming et al. [2001] for an application with stock returns.

Figure 11: Prediction densities for S&P 500



Note: The figure presents the (99%) interval forecasts given by the White Noise benchmark model (WN), the survey forecast (SR) and our density combination scheme (DC). The red solid line shows the realized values for S&P 500 percent log returns, for each out-of-sample observation.

ancing the portfolio at the start of month $t+1$ means that the weight invested in stocks is changed from α_t to α_{t+1}. We assume that transaction costs amount to a fixed percentage c on each traded dollar. As we assume that the initial wealth R_0 equals to 1, transaction costs at time $t+1$ are equal to

$$c_{t+1} = 2c|\alpha_{t+1} - \alpha_t| \tag{53}$$

where the multiplication by 2 follows from the fact that the investor rebalances her investments in both stocks and bonds. The net excess portfolio return is then given by $y_{t+1} - c_{t+1}$. We apply a scenario with transaction costs of $c = 0.1\%$.

Panel A in Table 2 reports statical accuracy forecasting results. The survey forecasts produce the most accurate point forecasts: its RMSPE is the lowest. The survey is also the most precise in terms of sign ratio. This seems to confirm evidence that survey forecasts contain timing information. Evidence is, however, mixed in terms of density forecasts: the WH has higher log score whether the SR has the lowest CRPS; the highest log score is for our combination scheme. Figure 11 plots density forecasts given by the three approaches. The density forecasts of the survey are too narrow and therefore highly penalized from the LS statistics when missing substantial drops in stock returns as at the beginning of recession periods. The problem might be caused by the lack of reliable answers during those periods. However, this assumption cannot be easily investigated. The score for the WN is marginally lower than for our model combination. However the interval given by the WN is often too large and indeed

Table 2: Active portfolio performance

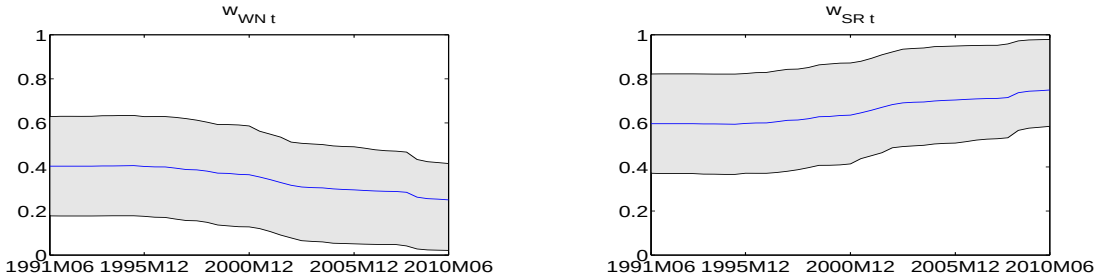
	$\gamma = 4$			$\gamma = 6$			$\gamma = 8$		
	WN	SR	DC	WN	SR	DC	WN	SR	DC
Panel A: Statistical accuracy									
RMSPE	12.62	11.23	11.54	-	-	-	-	-	-
SIGN	0.692	0.718	0.692	-	-	-	-	-	-
LS	-3.976	-20.44	-3.880	-	-	-	-	-	-
CRPS	6.816	6.181	6.188	-	-	-	-	-	-
Panel B: Economic analysis									
Mean	5.500	7.492	7.228	4.986	7.698	6.964	4.712	7.603	6.204
St dev	14.50	15.93	14.41	10.62	15.62	10.91	8.059	15.40	8.254
SPR	0.111	0.226	0.232	0.103	0.244	0.282	0.102	0.241	0.280
Utility	-12.53	-12.37	-12.19	-7.322	-7.770	-6.965	-5.045	-6.438	-4.787
r_s	73.1	157.4	254.2	471.5	234.1	671.6	950.9	254.6	1101
r_m	-202.1	-117.8	-20.94	-114.3	-351.7	85.84	3.312	-693.0	153.5
r_b	-138.2	-53.9	43.03	-131.3	-368.8	68.79	-98.86	-795.1	51.32
Panel C: Transaction costs									
Mean	5.464	7.341	7.128	4.951	7.538	6.875	4.683	7.439	6.136
St dev	14.50	15.93	14.40	10.62	15.62	10.89	8.058	15.40	8.239
SPR	0.108	0.217	0.225	0.100	0.233	0.274	0.098	0.230	0.272
Utility	-12.53	-12.40	-12.21	-7.329	-7.804	-6.982	-5.050	-6.484	-4.799
r_s	69.8	142.2	244.3	468.1	216.6	662.2	948.1	234.0	1094
r_m	-205.5	-133.1	-31.05	-117.7	-369.2	76.36	0.603	-713.5	146.3
r_b	-141.2	-68.81	33.22	-134.5	-385.9	59.62	-101.2	-815.3	44.44

Note: In Panel A the root mean square prediction error (RMSPE), the correctly predicted sign ratio (SIGN) and the Logarithmic Score (LS) for the individual models and combination schemes in forecasting the six month ahead S&P500 index over the sample December 1990 - June 2010. WN, SR and DC denote strategies based on excess return forecasts from the White Noise model, the Livingston-based forecasts and our density combination scheme in equation (1)-(3) and (10). In Panel B the annualized percentage point average portfolio return and standard deviation, the annualized Sharpe ratio (SPR), the final value of the utility function, and the annualized return in basis points that an investor is willing to give up to switch from the passive stock (s), mixed (m), or bond (b) strategy to the active strategies and short selling and leveraging restrictions are given. In Panel C the same statistics as in Panel B are reported when transaction costs $c = 10$ basis points are assumed. The results are reported for three different risk aversion coefficients $\gamma = (4, 6, 8)$.

the realization never exceeds the 2.5% and 97.5% percentiles.

Figure 12 shows the combination weights with learning for the individual forecasts. The weights seem to converge to a $\{0, 1\}$ optimal solution, where the survey has all the weight towards the end of the period even if the uncertainty is still substantial. Changing regulations, increased sophistication of instruments, technological advances and recent global recessions have increased the value added of survey forecasts, although forecast uncertainty must be modeled carefully as survey forecasts often seem too confident. When taking account for such drawback on the forecast uncertainty, we might

Figure 12: Combination weights for the S&P 500 forecasts

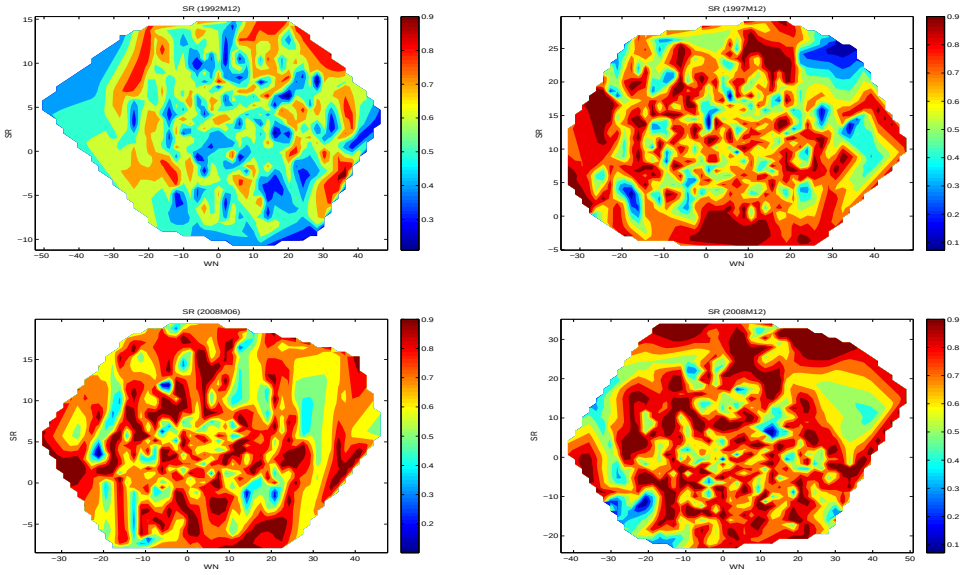


conclude that survey should always be selected. We add further analysis to show this is not always the best strategy.

Figure 13 shows the contours for SR weight in our density combination scheme for four different periods, 1992M12, 1997M12, 2008M6, 2008M12, times when forecasts are made. At beginning of the sample (1992M12), WN has most of the weight in the left tail and the SR in the right tail. However, there is a shift after five years, with SR having most of the mass in the left tail. The bottom panel shows the SR weight before and after Lehman brothers collapse. SR has most of the mass in the left tail for the forecast made in 2008M6. The SR density forecast results not very accurate in 2008M12 (as Figure 11 shows). Our methodology increases WN weights in the left tail when the new forecast is made. All the four graphs reveal that weights have highly nonlinear multimodal posterior distributions, in particular during crisis periods, and therefore just selecting one of the two models based on the mode or the median might be not optimal.

The results for the asset allocation exercise strengthen previous statistical accuracy evidence. Panel B in Table 2 reports results for three different risk aversion coefficients, $\gamma = (4, 6, 8)$. The survey forecasts give the highest mean portfolio returns in all three cases. But they also provide the highest portfolio standard deviations. Our combination scheme gives marginally lower returns, but the standard deviation is substantially lower, resulting in higher Sharpe Ratios and higher utility. In eight cases of nine it outperforms passive benchmark strategies, giving positive r fees. The other forecast strategies outperform the passive strategy of investing 100% of the portfolio in the stock market, but not the mixed strategy and investing 100% of the portfolio in the risk free asset. Therefore, our nonlinear distributional state-space predictive density gives the highest gain when the utility function is also highly nonlinear, as those of portfolio investors. Results are robust to reasonable transaction costs.

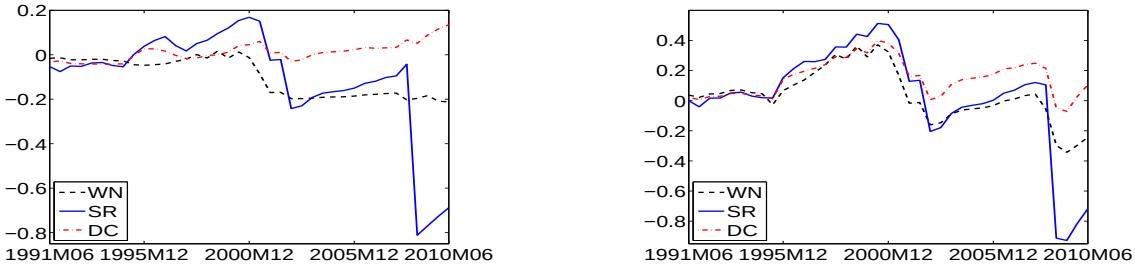
Figure 13: SR weight contours



Note: The plots show the contours for the survey forecast (SR) weight in our density combination scheme (DC) for four different dates when the forecasts were made.

Finally, Fig. 14 plots the differential between the utility values given by the three active investment strategies, $u(R_{B,t+1})$ B=SR, WN, DC, versus that of the passive strategies which invest all in the risk free asset or 50% in the risky asset and the remaining in the free risk asset. Results confirm intuitions given by the statistical evaluation: the economic gains from our combination strategies are larger during turbulent periods such as the 2001 and 2008 recessions. Relying on the SR individual models, which perform more accurately during normal times, can reduce substantially investors' economic wealth.

Figure 14: Utility value evolution



Note: Left: Power utility differentials of the three active investment strategies based on the predictive densities versus a passive strategy to invest 50% on the risky asset and 50% on the risk free asset. Right: Power utility differentials of the three active investment strategies based on the predictive densities versus a passive strategy to invest 100% on the risk free asset. The risk aversion coefficient γ is set to 6.

7 Conclusion

This paper proposes a general combination approach with time-varying weights for a set of predictive densities of models that are commonly used in macroeconomics and finance. The proposed method is based on a distributional state space representation of the weights in the combination scheme and on Bayesian filtering of the optimal weights. The distributional state-space form and the use of Sequential Monte Carlo allow us to extend the combination strategies to a nonlinear and non-Gaussian context and generalize the existing optimal weighting procedures based on Kalman and Hamilton filters. Our methodology can cope with incomplete models and different choices of the weight dynamics. The operational use of the method is assessed first in simulation exercises and then using US GDP and inflation forecast densities generated by some well known forecasting models and, also, through densities of returns of the S&P500 generated by a survey and a white noise model. In the application to macroeconomics, nonlinear density combination schemes with learning outperform, in terms of root mean square prediction error; Kullback Leibler information criterion; and cumulative rank probability score, BMA and BMA with optimal log score weights. Specifically, for the macro series we find that incompleteness of the models is relatively large in the 70's, the beginning of the 80's and during the recent financial crisis; while it is lower during the Great Moderation. The predicted probabilities of recession accurately compare with the NBER business cycle dating. Model weights have substantial uncertainty attached. The application to the financial forecasts shows that the proposed method allows one to combine forecast densities of different nature, model-based and survey-based, and that it gives the best predictive performance in terms of utility-based measures. Specifically, with respect to the returns of the S&P 500 series we find that an investment strategy using a combination of predictions from professional forecasters and from a white noise model put more weight on the white noise model in the beginning of the 90's and switches to giving more weight to the left tail of the professional forecasts during the start of the financial crisis around 2008. Information on the complete predictive distribution and not just basic moments turns out to be important in all cases investigated.

We end this paper by listing some topics for further research. The approach can be extended by using a richer set of models. Then challenges are the computational burden and the use of approximate methods, such as forgetting factor in the Kalman filter, see, e.g. Raftery et al. [2010], Koop and Korobillis [2012] and Koop and Korobilis [2012]. Parallelization techniques using, for instance,

Graphical Processing Units, are promising avenues for research. We intend to report on this in the near future.

References

G. Amisano and R. Giacomini. Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics*, 25(2):177–190, 2007.

A. Ang, G. Bekaert, and M. Wei. Do macro variables, asset markets, or surveys forecast inflation better? *Journal of Monetary Economics*, 54(4):1163–1212, 2007.

N. Barberis. Investing for the Long Run When Returns are Predictable. *Journal of Finance*, 55: 225–264, 2000.

G. A. Barnard. New methods of quality control. *Journal of the Royal Statistical Society, Series A*, 126:255–259, 1963.

J. M. Bates and C. W. J. Granger. Combination of Forecasts. *Operational Research Quarterly*, 20: 451–468, 1969.

J. Berkowitz. Testing Density Forecasts, with Applications to Risk Management. *Journal of Business & Economic Statistics*, 19(4):465–74, 2001.

M. Billio and R. Casarin. Identifying Business Cycle Turning Points with Sequential Monte Carlo: An Online and Real-Time Application to the Euro Area. *Journal of Forecasting*, 29:145–167, 2010.

G. Boero, J. Smith, and K. F. Wallis. Uncertainty and disagreement in economic prediction: the bank of england survey of external forecasters. *Economic Journal*, 118:1107–1127, 2008.

W.A. Branch. The theory of rational heterogeneous expectations: Evidence from survey data on inflation expectations. *Economic Journal*, 114:592–621, 2004.

R. Casarin and J.-M. Marin. Online data processing: Comparison of bayesian regularized particle filters. *Electronic Journal of Statistics*, 3:239–258, 2009.

T. Clark and M. W. McCracken. Advances in forecast evaluation. In A. Timmermann and G. Elliott, editors, *Handbook of Economic Forecasting*. Elsevier, Amsterdam, 2012.

T. Clark and F. Ravazzolo. The macroeconomic forecasting performance of autoregressive models with alternative specifications of time-varying volatility. Technical report, FRB of Cleveland Working Paper 12-18, 2012.

T. Clark and K. West. Approximately Normal Tests for Equal Predictive Accuracy in Nested Models. *Journal of Econometrics*, 138(1):291–311, 2007.

M. Clements and D. Hendry. *Forecasting Economic Time Series*. Cambridge University Press, 1998.

T. Cogley and T.J. Sargent. Drifts and volatilities: Monetary policies and outcomes in the post-world war ii u.s. *Review of Economic Dynamics*, 8:262–302, 2005.

D. Creal. A survey of sequential monte carlo methods for economics and finance. *Econometric Reviews*, 31(3):245–296, 2009.

A. D'Agostino, L. Gambetti, and D. Giannone. Macroeconomic forecasting and structural change. *Journal of Applied Econometrics*, forthcoming, 2011.

T. Dangl and M. Halling. Predictive regressions with time-varying coefficients. *Journal of Financial Economics*, 106(1):157–181, 2012.

P. Del Moral. *Feynman-Kac Formulae*. Springer Verlag, New York, 2004.

F. X. Diebold and P. Pauly. Structural change and the combination of forecasts. *Journal of Forecasting*, 6:21–40, 1987.

F.X. Diebold and R.S. Mariano. Comparing predictive accuracy. *Journal of Business and Economic Stastistics*, 13:253–263, 1995.

A. Doucet, J. G. Freitas, and J. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer Verlag, New York, 2001.

E. F. Fama and M. R. Gibbons. A comparison of inflation forecasts. *Journal of Monetary Economics*, 13(3):327–348, 1984.

J. Fleming, C. Kirby, and B. Ostdiek. The Economic Value of Volatility Timing. *Journal of Finance*, 56:329–352, 2001.

E. I. George. Dilution priors: Compensating for model space redundancy. In J. O. Berger, T. T. Cai, and I. M. Johnstone, editors, *Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown*, pages 158–165. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2010.

J. Geweke. *Complete and Incomplete Econometric Models*. Princeton: Priceton University Press, 2010.

J. Geweke and G. Amisano. Comparing and evaluating Bayesian predictive distributions of asset returns. *International Journal of Forecasting*, 26(2):216–230, 2010a.

J. Geweke and G. Amisano. Optimal prediction pools. *Journal of Econometrics*, 164(2):130–141, 2010b.

J. Geweke and C. Whiteman. Bayesian forecasting. In G. Elliot, C.W.J. Granger, and A.G. Timmermann, editors, *Handbook of Economic Forecasting*. North-Holland, 2006.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

T. Gneiting and R. Ranjan. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business and Economic Stastistics*, 29:411–422, 2011.

C. W. J. Granger. Invited review combining forecasts - twenty years later. *Journal of Forecasting*, 8: 167–173, 2006.

C. W. J. Granger and R. Ramanathan. Improved Methods of Combining Forecasts. *Journal of Forecasting*, 3:197–204, 1984.

J.J.J Groen, R. Paap, and F. Ravazzolo. Real-time inflation forecasting in a changing world. *Journal of Business and Economic Stastistics*, forthcoming, 2012.

G. K. Grunwald, A. E. Raftery, and P. Guttorp. Time series of continuous proportions. *Journal of the Royal Statistical Society, Series B*, 55:103–116, 1993.

M. Guidolin and A. Timmermann. Forecasts of US Short-term Interest Rates: A Flexible Forecast Combination Approach. *Journal of Econometrics*, 150:297–311, 2009.

S.G. Hall and J. Mitchell. Combining density forecasts. *International Journal of Forecasting*, 23:1–13, 2007.

B. Hansen. Least squares model averaging. *Econometrica*, 75:1175–1189, 2007.

B. Hansen. Least squares forecast averaging. *Journal of Econometrics*, 146:342–350, 2008.

J. Harrison and M. West. *Bayesian Forecasting and Dynamic Models, 2nd Ed.* Springer Verlag, New York, 1997.

D. Harvey, S. Leybourne, and P. Newbold. Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13:281–291, 1997.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14:382–417, 1999.

L. Hoogerheide, R. Kleijn, R. Ravazzolo, H. K. van Dijk, and M. Verbeek. Forecast Accuracy and Economic Gains from Bayesian Model Averaging using Time Varying Weights. *Journal of Forecasting*, 29(1-2):251–269, 2010.

L. Hoogerheide, F. Ravazzolo, and H.K. van Dijk. Backtesting value-at-risk using forecasts for multiple horizons - a comment on the forecast rationality tests based on multi-horizon bounds. *Journal of Business and Economic Statistics*, 30(1):30–33, 2012.

G. Huerta, W. Jiang, and M. Tanner. Time series modeling via hierarchical mixtures. *Statistica Sinica*, 13:1097–1118, 2003.

M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6:181–214, 1994.

A. S. Jore, J. Mitchell, and S. P. Vahey. Combining forecast densities from VARs with uncertain instabilities. *Journal of Applied Econometrics*, 25(4):621–634, 2010.

C. Kascha and F. Ravazzolo. Combining Inflation Density Forecasts. *Journal of Forecasting*, 29(1-2): 231–250, 2010.

Y. Kitamura. Econometric Comparisons of Conditional Models. Discussion paper,, University of Pennsylvania, 2002.

G. Koop. *Bayesian Econometrics*. John Wiley and Sons, 2003.

G. Koop and D. Korobilis. Large time-varying parameter vars. Technical report, University of Glasgow, 2012.

G. Koop and D. Korobillis. Forecasting inflation using dynamic model averaging. *International Economic Review*, 53(3):867–886, 2012.

D. Korobilis. Var forecasting using bayesian variable selection. *Journal of Applied Econometrics*, Forthcoming, 2011.

K. Lahiri and X. Sheng. Measuring forecast uncertainty by disagreement: the missing link. *Journal of Applied Econometrics*, 128:137–164, 2010.

M. Lanne. Properties of Market-Based and Survey Macroeconomic Forecasts for Different Data Releases. *Economics Bulletin*, 29(3):2231–2240, 2009.

F. Legland and N. Oudjane. Stability and uniform approximation of nonlinear filters using the Hilbert metric and application to particle filters. *Annals of Applied Probability*, 14(1):144–187, 2004.

E. Ley and M. F. J. Steel. On the effect of prior assumptions in bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, 24(4):651–674, 2009.

F. Li, R. Kohn, and M. Villani. Flexible modelling of conditional distributions using smooth mixtures of asymmetric student t densities. *Journal of Statistical Planning and Inference*, 140:3638–3654, 2010.

H. Liang, G. Zou, A.T.K. Wan, and X. Zhang. Optimal weight choice for frequentist model averaging estimator. *Journal of American Statistical Association*, 106:1053–1066, 2011.

J. S. Liu and M. West. Combined parameter and state estimation in simulation based filtering. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

Y. P. Mehra. Survey measures of expected inflation : revisiting the issues of predictive content and rationality. *Economic Quarterly*, 3:17–36, 2002.

J. Mitchell and S. G. Hall. Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESER "fan" charts of inflation. *Oxford Bulletin of Economics and Statistics*, 67:995–1033, 2005.

C. Musso, N. Oudjane, and F. Legland. Improving regularised particle filters. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

A.J. Patton and A. Timmermann. Forecast rationality tests based on multi-horizon bounds. *Journal of Business & Economic Statistics*, 30(1):1–17, 2012.

A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, 133:1155–1174, 2005.

A.E. Raftery, M. Karny, and P. Ettler. Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, 52:52–66, 2010.

F. Ravazzolo and S.V. Vahey. Forecast densities for economic aggregates from disaggregate ensembles. Technical Report 2010/02, Norges Bank, 2012.

H. V. Roberts. Probabilistic prediction. *Journal of American Statistical Association*, 60:50–62, 1965.

B. W. Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, London, 1986.

J.M. Sloughter, T. Gneiting, and A. E. Raftery. Probabilistic Wind Speed Forecasting Using Ensembles and Bayesian Model Averaging. *Journal of the American Statistical Association*, 105:25–35, 2010.

N. Terui and H. K. van Dijk. Combined forecasts from linear and nonlinear time series models. *International Journal of Forecasting*, 18:421–438, 2002.

L. B. Thomas. Survey Measures of Expected U.S. Inflation. *Journal of Economic Perspectives*, 13(4): 125–144, 1999.

A. Timmermann. Forecast combinations. In G. Elliot, C.W.J. Granger, and A.G. Timmermann, editors, *North-Holland*, volume 1 of *Handbook of Economic Forecasting*, chapter 4, pages 135–196. Elsevier, 2006.

I. Welch and A. Goyal. A Comprehensive Look at the Empirical Performance of Equity Premium prediction. *Review of Financial Studies*, 21(4):253–303, 2008.

K. D. West, H. J. Edison, and D. Cho. A utility-based comparison of some models of exchange rate volatility. *Journal of International Economics*, 35(1-2):23–45, 1993.

K.D. West. Asymptotic inference about predictive ability. *Econometrica*, 64:1067–1084, 1996.

R. L. Winkler. Combining probability distributions from dependent information sources. *Management Science*, 27:479–488, 1981.

V. Zarnowitz. Consensus and uncertainty in economic prediction. *National Bureau of Economic Research*, 17:492–518, 1992.

Appendix A - Data

Gross Domestic Product and Inflation

The first data set focuses on US real GDP and US inflation. We collect quarterly seasonally adjusted US GDP from 1960:Q1 to 2009:Q4 available from the US Department of Commerce, Bureau of Economic Analysis (BEA). In a pseudo-real-time out-of-sample forecasting exercise, we model and forecast the 1-step ahead quarterly growth rate, $100(\log(\text{GDP}_t) - \log(\text{GDP}_{t-1}))$[10]. For inflation we consider the quarterly growth rate of the seasonally adjusted PCE deflator, $100(\log(\text{PCE}_t) - \log(\text{PCE}_{t-1}))$, from 1960:Q1 to 2009:Q4, also collected from the BEA website.

In forecasting we use an initial in-sample period from 1960:Q1 to 1969:Q4 to obtain initial parameter estimates and we forecast GDP and PCE growth figures for 1970:Q1. We then extend the estimation sample with the value in 1970:Q1, re-estimating parameters, and forecast the next value for 1970:Q2. By iterating this procedure up to the last value in the sample we end up with a total of 160 forecasts.

We consider $K = 6$ time series models which are widely applied to forecast macroeconomic variables. Two models are linear specifications: an univariate autoregressive model of order one (AR) and a bivariate vector autoregressive model for GDP and PCE, of order one (VAR). We also apply four time-varying parameter specifications: a two-state Markov-switching autoregressive model of order one (ARMS) and a two-state Markov-switching vector autoregressive model of order one for GDP and inflation (VARMS); a time-varying autoregressive model with stochastic volatility (TVPARSV) and a time-varying vector autoregressive model with stochastic volatility (TVPVARSV). Therefore, our model set includes constant parameter univariate and multivariate specification; univariate and multivariate models with discrete breaks (Markov-Switiching specifications); and univariate and multivariate models with continuous breaks.

We estimate models using Bayesian inference with weak-informative conjugate priors and produce 1-step ahead predictive density via direct simulations for AR and VAR, see, e.g. Koop [2003] for details; we use Gibbs sampling algorithm for ARMS and VARMS, see, e.g. Geweke and Amisano [2010a] and TVPARSV and TVPVARSV, see e.g., D'Agostino et al. [2011] and Cogley and Sargent [2005] for details. For both classes of models we simulate $M = 1,000$ (independent) draws to approximate the

[10] We do not consider data revisions and use data from the 2010:Q1 vintage.

predictive likelihood of the GDP and inflation. Forecast combination practice usually considers point forecasts, e.g. the median of the predictive densities. The uncertainty around the point forecasts is, however, very large and should be carefully estimated due to its key role in decision making, see discussions in ,e.g., Geweke [2010]. The aim of our paper is to propose a general combination method of the predictive densities which can cope with the uncertainty and increase the accuracy of both density and point forecasts.

Survey Forecasts on Standard and Poor's 500

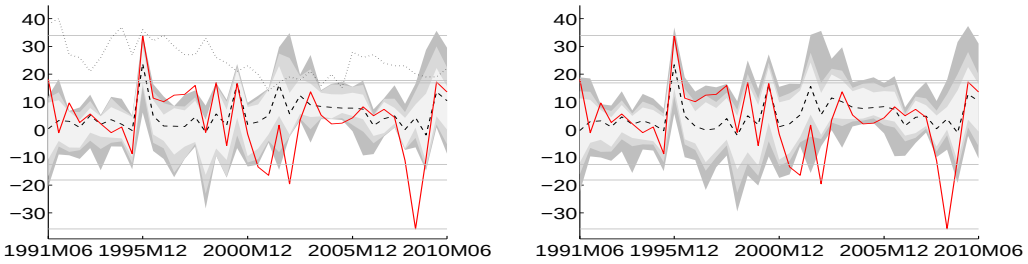
Several papers have documented that survey expectations have substantial forecasting power for macroeconomic variables. For example, Thomas [1999] and Mehra [2002] show that surveys outperform simple time-series benchmarks for forecasting inflation. Ang et al. [2007] make a comprehensive comparison of several survey measures of inflation for the US with a wide set of econometric models: time series ARIMA models, regressions using real activity measures motivated by the Phillips curve, and term structure models. Results indicate that surveys outperform these methods in point forecasting inflation.

The demand for forecasts for accurate financial variables has grown fast in recent years due to several reasons, such as changing regulations, increased sophistication of instruments, technological advances and recent global recessions. But compared to macroeconomic applications, financial surveys are still rare and difficult to access. Moreover, research on the properties of these databases such as their forecasting power is almost absent. The exceptions are few and relate mainly to interest rates. For example Fama and Gibbons [1984] compare term structure forecasts with the Livingston survey and to particular derivative products; Lanne [2009] focuses on economic binary options on the change in US non-farm payrolls.

We collect six month ahead forecasts for the Standard & Poor's 500 (S&P 500) stock price index from the Livingston survey.[11] The Livingston Survey was started in 1946 by the late columnist Joseph Livingston and it is the oldest continuous survey of economists' expectations. The Federal Reserve Bank of Philadelphia took responsibility for the survey in 1990. The survey is conducted twice a year, in June and December, and participants are asked different questions depending on the variable of interest. Questions about future movements of stock prices were proposed to participants from

[11] See for data and documentation www.philadelphiafed.org/research-and-data/real-time-center/livingston-survey/

Figure 15: Livingston survey fan charts for the S&P 500. Left: survey data empirical densities. Right: nonparametric density estimates



Note: The shadowed areas (from dark to light gray level) and the horizontal lines represent the 1%, 5%, 10%, 50%, 90%, 95% and 99% percentiles of the corresponding density forecast and of the sample distribution respectively, the black dashed line the point forecast and the red solid line shows the realized values for S&P 500 percent log returns, for each out-of-sample observation. The dotted black line shows the number of not-missing responses of the survey available at each date.

the first investigation made by Livingston in 1946, but the definition of the variable and the base years have changed several times. Since the responsibility passed to the Federal Reserve Bank of Philadelphia, questionnaires refer only to the S&P500. So the first six month ahead forecast we have, with a small but reasonable number of answers and a coherent index, is from December 1990 for June 1991.[12] The last one is made in December 2009 for June 2010, for a total of 39 observations. The surveys provide individual forecasts for the index value, we transform them in percent log-returns using realized index values contained in the survey database, that is $\tilde{y}_{t+1,i} = 100(\log(\tilde{p}_{t+1,i}) - \log(p_t))$ with $\tilde{p}_{t+1,i}$ the forecast for the index value at time $t+1$ of individual i made at time t and p_t the value of the index at time t as reported in the database and given to participants at the time that the forecast is made. Left chart in Figure 15 shows fan charts from the Livingston survey. The forecast density is constructed by grouping all the responses at each period. The number of survey forecasts can vary over time (black dotted line on the left chart); the survey participants (units) may not respond and the unit identity can vary. A problem of missing data can arise from both these situations. We do not deal with the imputation problem because we are not interested in the single agent forecast process. On the contrary, we consider the survey as an unbalanced panel and estimate over time an aggregate predictive density. We account for the uncertainty in the empirical density by using a nonparametric

[12]The survey also contains twelve month ahead forecasts and from June 1992 one month ahead forecasts (just twice at year). We focus on six month ahead forecasts, which is the database with more observations.

kernel density estimator:

$$p(\tilde{y}_t|y_{1:t-1}) = \frac{1}{hN_t}\sum_{k=1}^{N_t} K(h^{-1}(y_t - \tilde{y}_{k,t})) \tag{54}$$

on the survey forecasts $\tilde{y}_{k,t}$, with $k = 1, \ldots, N_t$, where N_t denotes that the time-varying number of available forecasts. For the kernel K we consider a Gaussian probability density function with an optimal bandwidth h (see for example Silverman [1986]). Our nonparametric density estimator can be interpreted as density forecast combination with equal weights. For optimal weights in the case of constant number of forecast, see Sloughter et al. [2010]. Zarnowitz [1992] derives combined density by aggregating point and interval forecasts for each density moment individually. Then, we simulate $M = 1,000$ draws from the estimated density. The right chart in Figure 15 shows the nonparametric simulated forecast densities. Left and right charts in Figure 15 look similar, but the nonparametric estimated forecasts span wider intervals as further uncertainties are considered in their construction. Both parametric and nonparametric estimates tend to understate the predictive uncertainty as reported in Boero et al. [2008] and Lahiri and Sheng [2010].

The survey forecasts predict accurately some sharp upward movements as in the second semester of 1995 or in the late 90's, but miss substantial drops during recession periods. The figure also shows that the forecast densities have time-varying volatility and fat-tails.

Appendix B - Combination schemes

Combining Prediction Density

A more parsimonious model than the one presented in Section 2 is given by

$$p(\mathbf{y}_t|W_t, \tilde{\mathbf{y}}_t) \propto \exp\left\{-\frac{1}{2}\left(\mathbf{y}_t - \sum_{k=1}^{K} w_{k,t}\tilde{\mathbf{y}}_{k,t}\right)' \Sigma^{-1}\left(\mathbf{y}_t - \sum_{k=1}^{K} w_{k,t}\tilde{\mathbf{y}}_{k,t}\right)\right\} \tag{55}$$

where $\mathbf{w}_t = (w_{1,t}, \ldots, w_{K,t}) \in \Delta_{[0,1]^K}$. In this model all the elements of the prediction $\mathbf{y}_{k,t}$ given by the k-th model have the same weight, while the weights may vary across the models.

Moreover, as an alternative to the Gaussian distribution, heavy-tailed distributions could be used to account for extreme values which are not captured by the pool of predictive densities.

Example 1 - (*Student-t combination scheme*)

In this scheme the conditional density of the observable is

$$p(\mathbf{y}_t|W_t, \tilde{\mathbf{y}}_t) \propto \left(1 + \frac{1}{\nu}\left(\mathbf{y}_t - W_t\tilde{\mathbf{y}}_t\right)' \Sigma^{-1} \left(\mathbf{y}_t - W_t\tilde{\mathbf{y}}_t\right)\right)^{-\frac{\nu+L}{2}} \tag{56}$$

where Σ is the precision matrix and $\nu > 2$ is the degrees-of-freedom parameter. The scheme could be extended to asymmetric Student-t as in Li et al. [2010].

■

Example 2 - (*Mixture of experts*)

Similarly to Jordan and Jacobs [1994] and Huerta et al. [2003], the density of the observable is

$$p(\mathbf{y}_t|\tilde{\mathbf{y}}_t) = \sum_{k=1}^{K} p\left(W_{k,t}|\mathbf{y}_{1:t-1}, \tilde{\mathbf{y}}_{1:t-1}\right) p\left(\tilde{\mathbf{y}}_{k,t}\right) \tag{57}$$

where $p\left(W_t|\mathbf{y}_{1:t-1}, \tilde{\mathbf{y}}_{1:t-1}\right)$ is the mixture weight associated to model k, which might be specified similarly to forms in section 3.

Such expression does not allow for the the assumption that all models are false and in the limit one of the weight will tend to one as discussed in Geweke and Amisano [2010b].

■

Weights

We present two alternatives to the continuous weights we have discussed in 3.

Example 3 - (*Dirichlet Weights*)

The weight model based on the multivariate logistic transform does not lead to an easy analytical evaluation of the dependence structure between the weights. An alternative specification of the weight dynamics makes use of the Dirichlet distribution $\mathcal{D}_K\left(\alpha_1, \ldots, \alpha_K\right)$ in order to define a Dirichlet autoregressive model.

$$\mathbf{x}_t^l \sim \mathcal{D}_{KL}\left(\eta_{1,t}^l\phi, \ldots, \eta_{KL-1,t}^l\phi, \eta_{KL,t}^l\phi\right) \tag{58}$$

where $\phi > 0$ is the precision parameter and $\boldsymbol{\eta}_t^l = \mathbf{g}(\mathbf{w}_{t-1}^l)$ with $\mathbf{w}_t^l \perp \boldsymbol{\varepsilon}_s^l$, $\forall\, s, t$. Due to the property of the Dirichlet random variable, the multivariate transform of the latent process is the identity function

and it possible to set $\mathbf{w}_t^l = \mathbf{x}_t^l$.

An advantage of using the Dirichlet model is that it is naturally defined on the standard K-dimensional simplex and that the conditional mean and variance and the covariance can be easily calculated. See for example the seminal paper of Grunwald et al. [1993] for a nonlinear time series model for data defined on the standard simplex.

The main drawback in the use of this weighting distribution is that, conditional on the past, the correlation between the weights is negative. Moreover it is not easy to model dependence between the observable and the weights. A possible way would be to introduce dependence through a common latent factor. We leave these issues as topics for future research.

■

Moreover, we consider weights with discontinuous dynamics. In fact, in many applied contexts the discontinuity (e.g. due to structural breaks) in the data generating process (DGP) calls for a sudden change of the current combination of the prediction densities.

Example 4 - (*Markov-switching Weighting Schemes*) We suggest the use of Markov-switching processes to account for the discontinuous dynamics of the weights. In fact, in many applied contexts the discontinuity (e.g. due to structural breaks) in the data generating process calls for a sudden variation of the current combination of the predictive densities.

We focus on Gaussian combination schemes with the learning mechanism presented in the section 2. The weight specification strategies, presented in the following, can, however, be easily extended to more general models to account for a more complex dependence structure between the weights of different components for the various predictors $\mathbf{y}_{k,t}$.

Consider the following Markov-switching scheme.

$$p(\mathbf{y}_t|W_t, \Sigma_t, \tilde{\mathbf{y}}_t) \propto \exp\left\{-\frac{1}{2}\left(\mathbf{y}_t - W_t\tilde{\mathbf{y}}_t\right)' \Sigma_t^{-1}\left(\mathbf{y}_t - W_t\tilde{\mathbf{y}}_t\right)\right\} \tag{59}$$

$$\Sigma_t = \sum_{r=0}^{R-1} D_r \mathbb{I}_{\{r\}}(s_t) \tag{60}$$

$$s_t \sim P(s_t = i|s_{t-1} = j) = p_{ij}, \quad \forall i, j \in \{0, \ldots, R-1\} \tag{61}$$

where D_r are positive define matrices. The l-th row of W_t is $\mathbf{w}_t^l = \mathbf{g}(\mathbf{x}_t^l)$ and is a function of the latent factors $\mathbf{x}_t^l$ and $\boldsymbol{\xi}_t = (\xi_{1,t}, \ldots, \xi_{L,t})$ with the following dynamics

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \boldsymbol{\mu}_t, \mathbf{y}_{1:t-1}, \tilde{\mathbf{y}}_{1:t-1}) \propto \exp\left\{-\frac{1}{2}\left(\Delta\mathbf{x}_t - \boldsymbol{\mu}_t + \Delta\mathbf{e}_t\right)' \Lambda^{-1} \left(\Delta\mathbf{x}_t - \boldsymbol{\mu}_t + \Delta\mathbf{e}_t\right)\right\} \tag{62}$$

$$\boldsymbol{\mu}_t = (\mu_{1,t}, \ldots, \mu_{KL^2,t}) \tag{63}$$

$$\mu_{l,t} = \sum_{r=0}^{Q-1} d_{l,r} \mathbb{I}_{\{r\}}(\xi_{l,t}) \tag{64}$$

$$\xi_{l,t} \sim P(\xi_{l,t} = i|\xi_{l,t-1} = j) = p_{ij}, \tag{65}$$

$\forall i, j \in \{0, \ldots, Q-1\}$, with $l = 1, \ldots, KL^2$. We assume $\xi_{l,t} \perp s_u \ \forall t, u$ and $\xi_{l,t} \perp \xi_{j,u} \ \forall l \neq j$ and $\forall s, t$.

It is possible to reduce the number of parameters to be estimated by considering the following Markov-switching weighting structure

$$p(\mathbf{y}_t|W_t, s_t, \tilde{\mathbf{y}}_t) \propto \exp\left\{-\frac{1}{2}\left(\mathbf{y}_t - \sum_{k=1}^{K} \mathbf{w}_{k,t} \odot \tilde{\mathbf{y}}_{k,t}\right)' \Sigma_{s_t}^{-1} \left(\mathbf{y}_t - \sum_{k=1}^{K} \mathbf{w}_{k,t} \odot \tilde{\mathbf{y}}_{k,t}\right)\right\} \tag{66}$$

$$\Sigma_{s_t} = \Sigma\psi(s_t) + (1 - \psi(s_t))I_L \tag{67}$$

$$s_t \sim P(s_t = i|s_{t-1} = j) = p_{ij}, \quad \forall i, j \in \{0, 1\} \tag{68}$$

with $\mathbf{w}_{k,t} = (w_{k,t}^1, \ldots, w_{k,t}^L)$ and $\psi(s_t) : \{0, 1\} \mapsto [0, 1]$. We let $\psi(0) = 1$ and $\psi(0) > \psi(1)$ as identifiability constraint.

The dynamics of $\mathbf{w}_t^l = (w_{1,t}^l, \ldots, w_{K,t}^l)' = \mathbf{g}(\mathbf{x}_t^l)$ is driven by the latent factors

$$p(\mathbf{x}_t^l|\mathbf{x}_t^l, \boldsymbol{\mu}_t^l, \mathbf{y}_{1:t-1}, \tilde{\mathbf{y}}_{1:t-1}) \propto \exp\left\{-\frac{1}{2}\left(\Delta\mathbf{x}_t^l - \boldsymbol{\mu}_t^l + \Delta\mathbf{e}_t^l\right)' \Lambda^{-1} \left(\Delta\mathbf{x}_t^l - \boldsymbol{\mu}_t^l + \Delta\mathbf{e}_t^l\right)\right\} \tag{69}$$

$$\boldsymbol{\mu}_t^l = \boldsymbol{\mu}_0 + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\xi_{l,t} \tag{70}$$

$$\xi_{l,t} \sim P(\xi_{l,t} = i|\xi_{l,t-1} = j) = p_{ij}, \quad \forall i, j \in \{0, 1\} \tag{71}$$

with $l = 1, \ldots, L$. We assume $\mu_{k,0} < \mu_{k,1}$ for identifiability purposes and $\xi_{l,t} \perp s_u \ \forall t, u$ and $\xi_{l,t} \perp \xi_{j,u}$ $\forall l \neq j$ and $\forall s, t$.

■

Appendix C - Sequential Monte Carlo

As an example of the filtering procedure applied in our analysis, we give in the following the pseudo-code of a simple sequential Monte Carlo procedure adapted to the basic TVW model. Let $\mathbf{x}_t$ be the vector of transformed weights and assume, to simplify the exposition, that the parameters are known. Then at time t with $t = 1, \ldots, \bar{t}$, the SMC algorithm performs the following steps:

– Given $\{\Xi_t^j\}_{j=1}^M$, with $\Xi_t^j = \{\mathbf{x}_t^{i,j}, \omega_t^{i,j}\}_{i=1}^N$ and for $j = 1, \ldots, M$

- Generate $\tilde{\mathbf{y}}_{t+1}^j$ from $p(\tilde{\mathbf{y}}_{t+1}^j | \mathbf{y}_{1:t})$
- For $i = 1, \ldots, N$
  1. Generate $\mathbf{x}_{t+1}^{i,j}$ from $\mathcal{N}_K(\mathbf{x}_t^{i,j}, \Lambda)$
  2. Generate $\mathbf{y}_{t+1}^{i,j}$ from $p(\mathbf{y}_{t+1} | \mathbf{x}_{t+1}^{i,j}, \tilde{\mathbf{y}}_{t+1}^1, \ldots, \tilde{\mathbf{y}}_{t+1}^M)$
  3. Update the weights

$$\tilde{\omega}_{t+1}^{i,j} \propto \omega_t^{i,j} \exp\left\{ -0.5 \left(\mathbf{y}_{t+1} - \sum_{k=1}^K w_{k,t}^{i,j} \tilde{\mathbf{y}}_{k,t}^j \right)' \Sigma^{-1} \left(\mathbf{y}_{t+1} - \sum_{k=1}^K w_{k,t}^{i,j} \tilde{\mathbf{y}}_{k,t}^j \right) \right\}$$

  where $w_{k,t}^{i,j} = \exp(x_{k,t}^{i,j}) / \sum_{k=1}^K \exp\{x_{k,t}^{i,j}\}$

- Evaluate the Effective Sample Size (ESS_t^j)
- Normalize the weights $\omega_{t+1}^{i,j} = \tilde{\omega}_{t+1}^{i,j} / \sum_{i=1}^N \tilde{\omega}_{t+1}^{i,j}$ for $i = 1, \ldots, N$
- If $ESS_t^j \leq \kappa$ then resample from Ξ_t^j