Combining Predictive Densities using Bayesian Filtering with Applications to US Economics Data*

Monica Billio [†] Roberto Casarin[†**]

Francesco Ravazzolo[‡] Herman K. van Dijk[§‖]

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

[‡]Norges Bank

[§]Econometric Institute, Erasmus University Rotterdam

[‖]Department of Econometrics VU University Amsterdam and Tinbergen Institute

Revised, September 30, 2011

Abstract

Using a Bayesian framework this paper provides a multivariate combination approach to prediction based on a distributional state space representation of predictive densities from alternative models. In the proposed approach the model set can be incomplete, meaning that all the models can be false. Several multivariate time-varying combination strategies are introduced. In particular, a weight dynamics driven by the past performance of the predictive densities is considered and the use of learning mechanisms. The approach is assessed using statistical and utility-based performance measures for evaluating density forecasts of US macroeconomic time series and of surveys of stock market prices.

*We thank for constructive comments conference and seminar participants at: the 1st meeting of the European Seminar on Bayesian Econometrics, the 4th CSDA International Conference on Computational and Financial Econometrics, the 6th Eurostat Colloquium and Norges Bank, the 4th Italian Congress of Econometrics and Empirical Economics, the NBER Summer Institute 2011 Forecasting and Empirical Methods in Macroeconomics and Finance Workshop, the 2011 European Economic Association and Econometric Society. The views expressed in this paper are our own and do not necessarily reflect those of Norges Bank.

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

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 available information on the variable to be predicted and, as a consequence, to possibly produce better forecasts. Most of the literature on forecast combinations in economics and finance focus on point forecasts. However the value of the forecasts can be increased by supplementing point forecasts with some measures of uncertainty. For example, interval and density forecasts are considered important parts of the communication from (central) banks to the public and also for the decision-making process on financial asset allocation.

In the literature there is growing focus on and many different approaches to model combination. One of the first-mentioned papers on forecasting with model combinations is Barnard [1963], who studied air passenger data, 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 started with Bates and Granger [1969] about combining predictions from different forecasting models. See Granger [2006] for an updated review on forecast combination. Granger and Ramanathan [1984] extend Bates and Granger [1969] and propose combining the forecasts with

unrestricted regression coefficients as weights. Liang et al. [2011] derive optimal weights in a similar framework. Hansen [2007] and Hansen [2008] compute optimal weights by maximizing a Mallow criterium. Terui and van Dijk [2002] generalize the least squares model weights by representing the dynamic forecast combination as a state space. In their work the weights are assumed to follow a random walk process. This approach has been successfully extended by Guidolin and Timmermann [2009], who introduced Markov-switching weights, and by Hoogerheide et al. [2010] and Groen et al. [2009], who proposed robust time-varying weights and accounted for both model and parameter uncertainty in model averaging. In these papers the model space is possibly incomplete, extending standard BMA where the correct model is supposed to exist (in the limit).

In the following, we assume that the weights associated with the predictive densities are time-varying and propose a general distributional state-space representation of the predictive densities and of the combination schemes. Our system allows for all the models to be false and therefore the model set is misspecified as discussed in Geweke [2010] and Amisano and Geweke [2010]. In this sense we extend the state-space representation of Terui and van Dijk [2002] and Hoogerheide et al. [2010]. For a review on distributional state-space representation in the Bayesian literature, see Harrison and West [1997]. We also extend 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 the models are false. Our approach is general enough to include multivariate linear and Gaussian models, dynamic mixtures and Markov-switching models, 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. It also implies that we can estimate (optimal) model weights that maximize general utility functions 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 a discrete probability distribution 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 prediction densities. A learning mechanism is introduced to allow the dynamics of each weight to be driven by the past and current performances of the predictive densities in the combination scheme.

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 the 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]. For example 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.

To show the practical and operational implications of the proposed approach, this paper focuses on the problem of combining density forecasts from two relevant economic datasets. The first one is given by density forecasts on two economic time series: the quarterly series of US Gross Domestic Product (GDP) and US inflation as measured by the Personal Consumption Expenditures (PCE) deflator. The 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. To the best of our knowledge, there are no other papers applying this general density combination method. The second dataset considers density forecasts on the 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, 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.

The structure of the paper is as follows. Section 2 describes the datasets and introduces combinations of prediction densities in a multivariate context. Section 3 presents different models for the weights dynamics and introduces the learning mechanism. In the Appendix alternative combination schemes and the relationships with some existing schemes in the literature are briefly discussed. Section 4 describes the nonlinear filtering problem and shows how Sequential Monte Carlo methods could be used to combine prediction densities. Section 5 provides the results of the application of the proposed combination method to the macroeconomic and financial datasets. Section 6 concludes.

2 Data and Method

In order to motivate the operational implications of our approach to combining predictive densities, we start with an exploratory data analysis and subsequently discuss our methodology.

2.1 Gross Domestic Product and Inflation

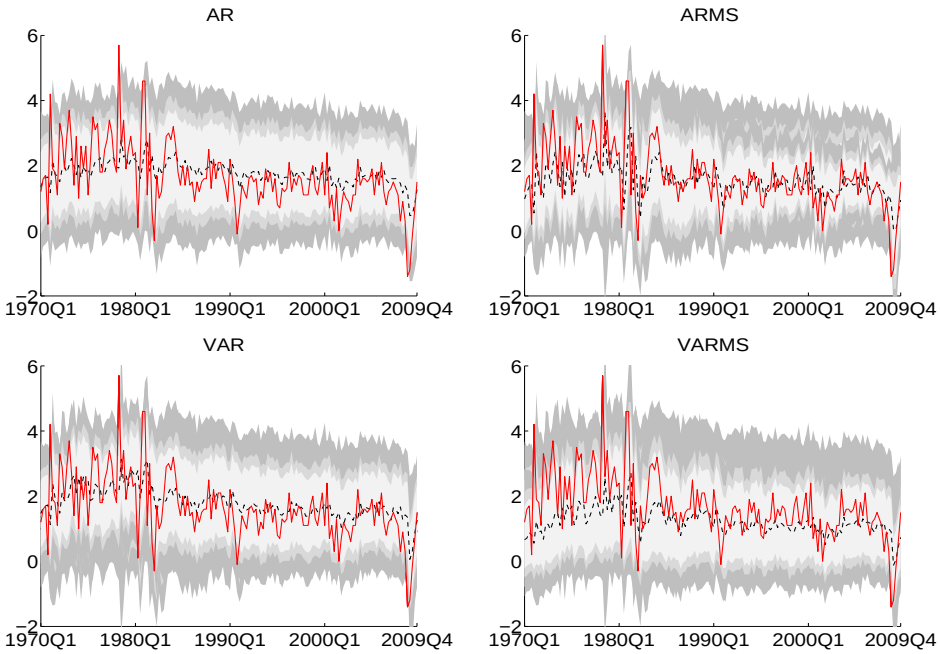
The first data set focuses on US GDP and US inflation. We collect quarterly seasonally adjusted US GDP from 1960:Q1 to 2009:Q4 available from the US Department of Commerce, Bureau of Economic Analysis (BEA). In a pseudo-real-time out-of-sample forecasting exercise, we model and forecast the 1-step ahead quarterly growth rate, $100(\log(\text{GDP}_t) - \log(\text{GDP}_{t-1}))$[1]. 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 $K = 4$ 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 two 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). 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 a Gibbs sampling algorithm for ARMS and VARMS, see, e.g. Geweke and Amisano [2010] for details. For both classes of models we simulate $M = 1,000$ (independent) draws to approximate the predictive likelihood of the GDP. Forecast combination practice usually considers point forecasts, e.g. the median of the predictive densities (black

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

Figure 1: GDP density forecast generated by different models



Note: Fan charts for empirical forecast density. In each chart the shadowed areas (from dark to light gray level) represent the 1%, 5%, 10%, 50%, 90%, 95% and 99% percentiles of the corresponding density forecast, the black dashed line the point forecasts and the red solid line shows the realized values for the US GDP percent growth.

dashed lines in Fig. 1). The uncertainty around the point forecasts is, however, very large (see percentiles in Fig. 1) and should be carefully estimated due to its key role in decision making. The aim of our paper is to propose a general combination method of the predictive densities which can reduce the uncertainty and increase the accuracy of both density and point forecasts.

2.2 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.

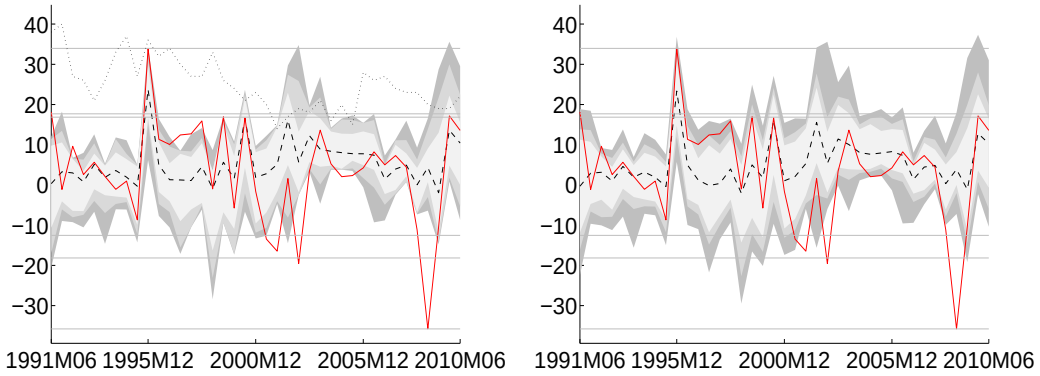
The demand for forecasts for accurate financial variables has grown fast in recent years due to several reasons, such as changing regulations, increased sophistication of instruments, technological advances and recent global recessions. But compared to macroeconomic applications, financial surveys are 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.[2] 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 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.[3] The last one is made in December 2009 for June 2010, for a total of 39 observations.

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

[3]The survey also contains twelve month ahead forecasts and from June 1992 one month ahead forecasts. We focus on six month ahead forecasts, which is the database with more observations.

Figure 2: 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 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 2 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 density. We account for the uncertainty in the empirical density

by using a nonparametric kernel density estimator:

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

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 2 shows the nonparametric simulated forecast densities. Left and right charts in Figure 2 look similar, but the nonparametric estimated forecasts span wider intervals as further uncertainties are considered in their construction.

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.

2.3 Combining Multivariate Prediction Densities

Let t be the time index, with $t = 1, \ldots, \bar{t}$, then given a sequence of vectors $\mathbf{x}_u$ with $u = s, \ldots, t$ and $s \leq t$ we denote with $\mathbf{x}_{s:t} = (\mathbf{x}_s, \ldots, \mathbf{x}_t)$ the collection of these vectors. We denote with $\mathbf{y}_t \in \mathcal{Y} \subset \mathbb{R}^L$ the vector of observable variables, $\tilde{\mathbf{y}}_{k,t} \in \mathcal{Y} \subset \mathbb{R}^L$ the typical k-th one-step ahead predictor for $\mathbf{y}_t$, where $k = 1, \ldots, K$. For the sake of simplicity we present the new combination method for the one-step ahead forecasting horizon. The methodology easily extends to multi-step ahead forecasting horizons.

We assume that the observable vector 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 $p(\tilde{\mathbf{y}}_{k,t}|\mathbf{y}_{1:t-1})$. In order to simplify the exposition, 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 matrix with the predictors in the columns and vec is an operator that stacks the columns of a matrix into a 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.

A combination scheme of a set of predictive densities is a probabilistic relation between the density of the observable variable and a set of predictive densities. We assume that the 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

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

where the dependence structure between the observable and the predictive is not defined yet. This relation might be misspecified because all the models are false or the true DGP is a combination of unknown and unobserved models that statistical and econometric tools can only partially approximate. In the following, in order to model the possibly misspecified dependence between forecasting models, we consider a parametric latent variable model. We also assume that the model is dynamic to capture the time variations in the dependence structure.

In order to define the latent variable model and the combination scheme we introduce first 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.

Secondly we introduce the latent model that is a matrix-valued stochastic process, $W_t \in \mathcal{W} \subset \mathbb{R}^L \times \mathbb{R}^{KL}$, which represents the time-varying weights of the combination scheme. Denote with $w^l_{k,t}$ the k-th column and l-th row elements of W_t, then we assume that the vectors $\mathbf{w}^l_t = (w^l_{1,t}, \ldots, w^l_{KL,t})'$ in the rows of W_t satisfy $\mathbf{w}^l_t \in \Delta_{[0,1]^{KL}}$.

The definition of the latent space as the standard simplex and the consequent restrictions on the dynamics of the weight process allow us to estimate a time series of $[0, 1]$ weights at time $t - 1$ when a forecast is made for $\mathbf{y}_t$. This 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 $\Delta_{[0,1]^{LK}}$.[4] In such a way one can interpret the weights as a discrete probability density over the set of predictors.

We assume that at time t, the time-varying weight process 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. (2) as

$$p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) = \int_{\mathcal{Y}^{Kt}} \left(\int_{\mathcal{W}} p(\mathbf{y}_t|W_t, \tilde{\mathbf{y}}_t) p(W_t|\mathbf{y}_{1:t-1}, \tilde{\mathbf{y}}_{1:t-1}) dW_t \right) p(\tilde{\mathbf{y}}_{1:t}|\mathbf{y}_{1:t-1}) d\tilde{\mathbf{y}}_{1:t} \qquad (3)$$

In the following, we assume that the time-varying weights have a first-order Markovian dynamics and that they may depend on the past values $\tilde{\mathbf{y}}_{1:t-1}$ of the predictors. Thus the weights at time t have $p(W_t|W_{t-1}, \tilde{\mathbf{y}}_{1:t-1})$ as conditional transition density. We usually assume that the weight dynamics depend on the recent values of the predictors, i.e.

$$p(W_t|W_{t-1}, \tilde{\mathbf{y}}_{1:t-1}) = p(W_t|W_{t-1}, \tilde{\mathbf{y}}_{t-\tau:t-1}) \qquad (4)$$

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

with $\tau > 0$.

Under these assumptions, the first integral in Eq. (3) is now defined on the set $\mathcal{Y}^{K(\tau+1)}$ and is taken with respect to a probability measure that has $p(\tilde{\mathbf{y}}_{t-\tau:t}|\mathbf{y}_{1:t-1})$ as joint predictive density. Moreover the conditional predictive density of W_t in Eq. (3) can be further decomposed as follows

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

The above assumptions do not alter the general validity of the proposed approach for the combination of the predictive densities. In fact, the proposed combination method extends previous model pooling by assuming possibly non-Gaussian predictive densities as well as nonlinear weights dynamics that maximize general utility functions.

As a conclusion of this section we present a possible specification of the conditional predictive density $p(\mathbf{y}_t|W_t, \tilde{\mathbf{y}}_t)$. In the appendix we present two further examples which allow for heavy-tailed conditional distributions. In the next section we will consider a specification for the weights transition density $p(W_t|W_{t-1}, \tilde{\mathbf{y}}_{1:t-1})$.

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

$$p(\mathbf{y}_t|W_t, \tilde{\mathbf{y}}_t) \propto \exp\left\{-\frac{1}{2}\left(\mathbf{y}_t - W_t\tilde{\mathbf{y}}_t\right)' \Sigma^{-1} \left(\mathbf{y}_t - W_t\tilde{\mathbf{y}}_t\right)\right\} \tag{5}$$

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

$$p(\mathbf{y}_t|W_t,\tilde{\mathbf{y}}_t)\propto\exp\left\{-\frac{1}{2}\left(\mathbf{y}_t-\sum_{k=1}^{K}\mathbf{w}_{k,t}\odot\tilde{\mathbf{y}}_{k,t}\right)'\Sigma^{-1}\left(\mathbf{y}_t-\sum_{k=1}^{K}\mathbf{w}_{k,t}\odot\tilde{\mathbf{y}}_{k,t}\right)\right\} \quad (6)$$

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 $\mathbf{y}_t$.

A more parsimonious model than the previous one is given by

$$p(\mathbf{y}_t|W_t,\tilde{\mathbf{y}}_t)\propto\exp\left\{-\frac{1}{2}\left(\mathbf{y}_t-\sum_{k=1}^{K}w_{k,t}\tilde{\mathbf{y}}_{k,t}\right)'\Sigma^{-1}\left(\mathbf{y}_t-\sum_{k=1}^{K}w_{k,t}\tilde{\mathbf{y}}_{k,t}\right)\right\} \quad (7)$$

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.

3 Weight Dynamics

In this section we present some existing and new specifications of the conditional density of the weights given in Eq. (4). In order to write the density of the combination models in a more general and compact form, 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 $\mathbf{x}_t^l=(x_{1,t}^l,\ldots,x_{KL,t}^l)'\in\mathcal{X}\subset\mathbb{R}^{KL}$. Then, 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 $\Delta_{[0,1]^K}$, we introduce the

multivariate transform $\mathbf{g} = (g_1, \ldots, g_{KL})'$

$$\mathbf{g} : \left[\begin{array}{ll} \mathbb{R}^{KL} & \to \Delta_{[0,1]^{KL}} \\ \mathbf{x}_t^l & \mapsto \mathbf{w}_t = (g_1(\mathbf{x}_t^l), \ldots, g_{KL}(\mathbf{x}_t^l))' \end{array} \right. \tag{8}$$

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

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

$$g_k(\mathbf{x}^l) = \frac{\exp\{x_k^l\}}{\sum_{j=1}^{KL} \exp\{x_j^l\}}, \quad \text{with } k = 1, \ldots, KL \tag{9}$$

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 then these estimates might vary along the recursion. Thus following the same idea, which is underlying the recursive least squares regression model, it is possible to replace the parameters x_k^l with a stochastic process $x_{k,t}^l$ which accounts for the time variation of the weight estimates and assume the trivial dynamics $x_{k,t}^l = x_{k,t-1}^l$, $\forall t$ and $l = 1, \ldots, L$.

We generalize this simple time-varying weight scheme. In our first specification of W_t, we assume that the weights have their own fluctuations generated by the latent process

$$\mathbf{x}_t \sim p(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{10}$$

with a non-degenerate distribution and then apply the transform g defined in Eq. (8)

$$\mathbf{w}_t^l = \mathbf{g}(\mathbf{x}_t^l), \quad l = 1, \ldots, L \tag{11}$$

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.

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

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) \propto \exp\left\{-\frac{1}{2}\left(\mathbf{x}_t - \mathbf{x}_{t-1}\right)' \Lambda^{-1} \left(\mathbf{x}_t - \mathbf{x}_{t-1}\right)\right\} \tag{12}$$

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

$$\mathbf{w}_t^l = \frac{\exp\{x_k^l\}}{\sum_{j=1}^{KL} \exp\{x_j^l\}}, \quad \text{with}\, k = 1, \ldots, KL$$

with $l = 1, \ldots, L$.

■

3.2 Learning Mechanism

We consider learning strategies based on the distribution of the forecast errors. More precisely, we evaluate the past performance of each prediction model and compare it with the performances of the other models.

The contribution of this section is to generalize the weight structures given in the previous sections and related literature (see for example Hoogerheide et al. [2010]) by including a learning strategy in the weight dynamics and by estimating, with nonlinear

filtering, the weight posterior probability. Therefore the weights are explicitly driven by the past and current forecast errors and capture the residual evolution of the combination scheme by the dynamic structure. In this sense our approach generalizes the existing literature on adaptive estimation schemes (see the seminal work of Bates and Granger [1969]). 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 last prediction error. We consider a moving window of τ observations and define the distance matrix $E_t^l = (\mathbf{e}_t^{l,1}, \ldots, \mathbf{e}_t^{l,L})$, where $\mathbf{e}_t^{l,d} = (e_{1,t}^{l,d}, \ldots, e_{K,t}^{l,d})'$, with $d = 1, \ldots, L$, is a vector of exponentially weighted average errors

$$e_{k,t}^{l,d} = (1-\lambda) \sum_{i=1}^{\tau} \lambda^{i-1} (y_{t-i}^l - \widehat{y}_{k,t-i}^{l,d})^2 \tag{13}$$

with $\lambda \in (0,1)$ a smoothing parameter and $\widehat{y}_{k,t-i}^{l,d}$ is the point forecast at time t given by model k for the variable y_{t-i}^l. Define $\mathbf{e}_t = \text{vec}(E_t)$, where $E_t = (E_t^1, \ldots, E_t^L)$, then we introduce the following weight model

$$\begin{aligned} \mathbf{w}_t^l &= \mathbf{g}(\mathbf{x}_t^l), \quad l = 1, \ldots, L & (14) \\ \mathbf{x}_t &= \mathbf{z}_t - \mathbf{e}_t & (15) \\ \mathbf{z}_t &= \mathbf{z}_{t-1} & (16) \end{aligned}$$

where $\mathbf{z}_t = \text{vec}(z_t^1, \ldots, z_t^L)$ and $\mathbf{z}_t^l \in \mathbb{R}^{KL}$. The model can be rewritten as follows

$$\begin{aligned} \mathbf{w}_t^l &= \mathbf{g}(\mathbf{x}_t^l), \quad l = 1, \ldots, L & (17) \\ \mathbf{x}_t &= \mathbf{x}_{t-1} - \Delta \mathbf{e}_t & (18) \end{aligned}$$

where $\Delta \mathbf{e}_t = \mathbf{e}_t - \mathbf{e}_{t-1}$. For the l-th variable in the model, with $l = 1, \ldots, L$, an increase at time t of the average forecasting error, i.e. $(e^{l,d}_{k,t} - e^{l,d}_{k,t-1}) > 0$, implies a reduction in the value of the weight associated to the $d-$th variable of the k-th predictor in the prediction density for the l-th variables in $\mathbf{y}_t$.

We notice that for $\tau = 1$ the model reduces to

$$x^l_{r,t} = x^l_{r,t-1} - (1-\lambda)\left[(y^l_{t-1} - \tilde{y}^{l,d}_{k,t-1})^2 - (y^l_{t-2} - \tilde{y}^{l,d}_{k,t-2})^2\right]$$

where $r = K(d-1) + k$.

We include the exponentially weighted learning strategy into the weight dynamics and estimate the posterior distribution of $\mathbf{x}_t$ accounting for the density of the conditional errors $p_\lambda(e^{l,d}_{k,t} | \widetilde{y}^{l,d}_{k,t-1:t-\tau}, y^l_{1:t-1})$ induced by Eq. (13).

It should also be noted that this specification strategy allows us to compute weights associated with very general utility functions and dynamics. Moreover we extend the previous section by introducing an error term in the weight dynamics in order to account for irregular variations in the weights and consider the following conditional densities.

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

$$p(\mathbf{x}_t | \mathbf{x}_{t-1}, \tilde{\mathbf{y}}_{1:t-1}) \propto \exp\left\{-\frac{1}{2}\left(\mathbf{x}_t - \mathbf{x}_{t-1} + \Delta \mathbf{e}_t\right)' \Lambda^{-1} \left(\mathbf{x}_t - \mathbf{x}_{t-1} + \Delta \mathbf{e}_t\right)\right\} \qquad (19)$$

■

Summary of the applied combination scheme

In the following 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

$$\mathbf{w}_t^l = \frac{\exp\{x_k^l\}}{\sum_{j=1}^{KL}\exp\{x_j^l\}}, \quad \text{with } k = 1, \ldots, KL$$

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

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:

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

4 Non-linear Filtering and Prediction

The density of the observable variables conditional on the combination scheme and on the predictions and the density of the weights of the scheme conditional on the prediction errors represent 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 $\tilde{\mathbf{y}}_t = (\tilde{\mathbf{y}}'_{1,t}, \ldots, \tilde{\mathbf{y}}'_{K,t})' \in \mathcal{Y} \subset \mathbb{R}^{KL}$ stand from a $\mathcal{F}_{t-1}$-measurable stochastic process associated with the predictive densities of the K

different models in the pool. Let $\mathbf{w}_t = (\mathbf{w}'_{1,t}, \dots \mathbf{w}'_{K,t})' \in \mathcal{X} \subset \mathbb{R}^{KL}$ 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 optimal predictive model. Include the parameter vector into the state vector and thus define the augmented state vector $\mathbf{z}_t = (\mathbf{w}_t, \boldsymbol{\theta}) \in \mathcal{Y} \times \Theta$. The distributional state space form of the optimal forecast model is

$$\begin{aligned} \mathbf{y}_t|\mathbf{z}_t, \tilde{\mathbf{y}}_t &\sim p(\mathbf{y}_t|\mathbf{z}_t, \tilde{\mathbf{y}}_t) && (20) \\ \mathbf{z}_t|\mathbf{z}_{t-1} &\sim p(\mathbf{z}_t|\mathbf{z}_{t-1}, \tilde{\mathbf{y}}_{1:t-1}) && (21) \\ \mathbf{z}_0 &\sim p(\mathbf{z}_0) && (22) \end{aligned}$$

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

$$\begin{aligned} p(\mathbf{z}_{t+1}|\mathbf{y}_{1:t}, \tilde{\mathbf{y}}_{1:t}) &= \int_{\mathcal{X}} p(\mathbf{z}_{t+1}|\mathbf{z}_t, \tilde{\mathbf{y}}_{1:t}) p(\mathbf{z}_t|\mathbf{y}_{1:t}, \tilde{\mathbf{y}}_{1:t}) d\mathbf{z}_t && (23) \\ p(\mathbf{z}_{t+1}|\mathbf{y}_{1:t+1}, \tilde{\mathbf{y}}_{1:t+1}) &\propto p(\mathbf{y}_{t+1}|\mathbf{z}_{t+1}, \tilde{\mathbf{y}}_{t+1}) p(\mathbf{z}_{t+1}|\mathbf{y}_{1:t}, \tilde{\mathbf{y}}_{1:t}) && (24) \end{aligned}$$

A major element of interest is the marginal predictive density of the observable variables

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

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.

An analytical solution of the previous filtering and prediction problems is not

known for the non-linear models presented in the previous sections, thus we apply a numerical approximation method. More specifically we consider a sequential Monte Carlo (SMC) approach to filtering. 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

$$p_N(\mathbf{z}_{t+1}|\mathbf{y}_{1:t}, \tilde{\mathbf{y}}_{1:t}) = \sum_{i=1}^{N} p(\mathbf{z}_{t+1}|\mathbf{z}_t, \tilde{\mathbf{y}}_{1:t})\omega_t^i \delta_{\mathbf{z}_t^i}(\mathbf{z}_t) \tag{25}$$

$$p_N(\mathbf{z}_{t+1}|\mathbf{y}_{1:t+1}, \tilde{\mathbf{y}}_{1:t+1}) = \sum_{i=1}^{N} \omega_{t+1}^i \delta_{\mathbf{z}_{t+1}^i}(\mathbf{z}_{t+1}) \tag{26}$$

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

$$p_N(\mathbf{y}_{t+1}|\mathbf{y}_{1:t}, \tilde{\mathbf{y}}_{1:t+1}) = \sum_{i=1}^{N} \omega_t^i \delta_{\mathbf{y}_{t+1}^i}(\mathbf{y}_{t+1}) \tag{27}$$

where $\mathbf{y}_{t+1}^i$ has been simulated from the measurement density $p(\mathbf{y}_{t+1}|\mathbf{z}_{t+1}^i,\ \tilde{\mathbf{y}}_{t+1}, \boldsymbol{\theta})$. 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

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

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 j independent values $\mathbf{z}^j_{1:t+1}$, 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 $\tilde{\mathbf{y}}^j_{1:t+1}$, in order to generate the particle set $\Xi^{i,j}_t = \{\mathbf{z}^{i,j}_{1:t}, \omega^{i,j}_t\}^N_{i=1}$, with $j = 1, \ldots, M$. At the last step, simulate $\mathbf{y}^{i,j}_{t+1}$ from $p(\mathbf{y}_{t+1}|\mathbf{z}^{i,j}_{t+1}, \tilde{\mathbf{y}}^j_{t+1})$ and obtain the following empirical predictive density

$$p_{N,M}(\mathbf{y}_{t+1}|\mathbf{y}_{1:t}) = \frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N}\omega^{i,j}_t \delta_{\mathbf{y}^{i,j}_{t+1}}(\mathbf{y}_{t+1}) \tag{28}$$

5 Empirical Applications

5.1 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 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]' statics (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.

Following Welch and Goyal [2008] we investigate how square prediction varies over time by a graphical inspection of the Cumulative Squared Prediction Error Difference

(CSPED):

$$CSPED_{k,t+1} = \sum_{s=\underline{t}}^{t} \widehat{f}_{k,s+1},$$

where $\widehat{f}_{k,t+1} = e_{AR,t+1} - e_{k,t+1}$ with k =VAR, ARMS, VARMS. Increases in $CSPED_{k,t+1}$ indicate that the alternative to the benchmark (AR model) predicts better at out-of-sample observation $t+1$.

We evaluate the predictive densities using a test of absolute forecast accuracy. Like Diebold et al. [1998], we utilize the Probability Integral Transforms (PITS), of the realization of the variable with respect to the forecast densities. A forecast density is preferred if the density is correctly calibrated, regardless of the forecasters loss function. The PITS at time $t+1$ are:

$$PITS_{k,t+1} = \int_{-\infty}^{y_{t+1}} p(\tilde{u}_{k,t+1}|y_{1:t}) d\tilde{u}_{k,t+1}.$$

and should be uniformly, independently and identically distributed if the forecast densities $p(\tilde{y}_{k,t+1}|y_{1:t})$, for $t = \underline{t}, \ldots, \bar{t}$, are correctly calibrated. Hence, calibration evaluation requires the application of tests for goodness of fit. We apply the Berkowitz [2001] test for zero mean, unit variance and independence of the PITS. The null of the test is no calibration failure.

Turning to our analysis of relative predictive accuracy, we consider a Kullback Leibler Information Criterion (KLIC) based test, utilizing the expected difference in the Logarithmic Scores of the candidate forecast densities; see for example Kitamura [2002], Mitchell and Hall [2005], Amisano and Giacomini [2007], Kascha and Ravazzolo [2010] and Caporin and Pres [2010]. Geweke and Amisano [2010] and Mitchell and Wallis [2010] discuss the value of information-based methods for evaluating forecast densities that are well calibrated on the basis of PITS tests. 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

$$\begin{aligned} \text{KLIC}_{k,t+1} &= \int p(y_{t+1}|y_{1:t}) \ln \frac{p(y_{t+1}|y_{1:t})}{p(\tilde{y}_{k,t+1}|y_{1:t})} dy_{t+1}, \\ &= \mathbb{E}_t[\ln p(y_{t+1}|y_{1:t}) - \ln p(\tilde{y}_{k,t+1}|y_{1:t}))]. \end{aligned} \tag{29}$$

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 $p(\tilde{y}_{k,t+1}|y_{1:t})$:

$$\overline{KLIC}_k = \frac{1}{t^*} \sum_{t=\underline{t}}^{\bar{t}} [\ln p(y_{t+1}|y_{1:t}) - \ln p(\tilde{y}_{k,t+1}|y_{1:t})]. \tag{30}$$

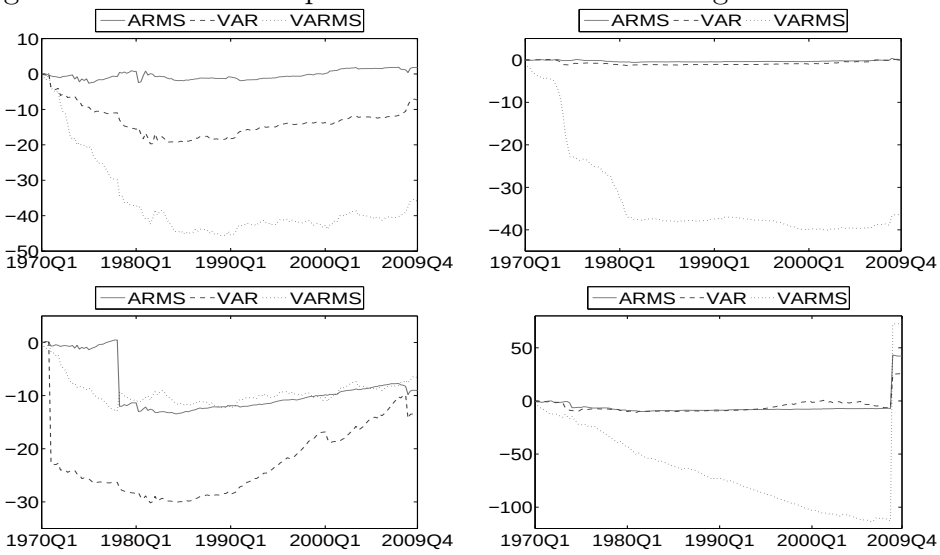
Even though we do not know the true density, we can still compare multiple densities, $p(\tilde{y}_{k,t+1}|y_{1:t})$. For the comparison of two competing models, it is sufficient to consider the Logarithmic Score (LS), which corresponds to the latter term in the above sum,

$$LS_k = -\frac{1}{t^*} \sum_{t=\underline{t}}^{\bar{t}} \ln p(\tilde{y}_{k,t+1}|y_{1:t}), \tag{31}$$

for all k and to choose the model for which the expression in (31) is minimal, or as we report in our tables, the opposite of the expression in (31) is maximal. Differences in KLIC can be statistically tested. We apply a test of equal accuracy of two density forecasts for nested models similar to Mitchell and Hall [2005], Giacomini and White [2006] and Amisano and Giacomini [2007]. For the two 1-step ahead density forecasts, $p(\tilde{y}_{AR,t+1}|y_{1:t})$ and $p(\tilde{y}_{k,t+1}|y_{1:t})$ we consider the loss differential

$$d_{k,t+1} = \ln p(\tilde{y}_{AR,t+1}|y_{1:t}) - \ln p(\tilde{y}_{k,t+1}|y_{1:t}).$$

Figure 3: Cumulative Square Prediction Error and Log Score Differences



Note: Cumulative Square Prediction Error Difference (first line) and the Cumulative Log Score Difference (second line), relative to the benchmark AR model, for the alternative models for forecasting US GDP growth (left column) and US PCE growth (right column) over the forecasting samples 1970-2009.

and apply the following Wald test:

$$GW_k = t^* \left(\frac{1}{t^*} \sum_{t=\underline{t}}^{\bar{t}} h_{k,t} d_{k,t+1} \right)' \widehat{\Sigma}_{k,t+1} \left(\frac{1}{t^*} \sum_{t=\underline{t}}^{\bar{t}} h_{k,t} d_{k,t+1} \right), \tag{32}$$

where $h_{k,t} = (1, d_{k,t})'$, and $\widehat{\Sigma}_{k,t+1}$ is the HAC estimator for the variance of $(h_{k,t} d_{k,t+1})$. The null is of the test is equal predictability.

Analogous to our use of the CSPED for graphically examining relative MSPEs over time, and following Kascha and Ravazzolo [2010], we define the Cumulative Log Score Difference ($CLSD$):

$$CLSD_{k,t+1} = -\sum_{s=\underline{t}}^{t} d_{k,s+1}, \tag{33}$$

If $CLSD_{k,t+1}$ increases at observation $t+1$, this indicates that the alternative to the AR benchmark has a higher log score.

Table 1: Forecast accuracy for the univariate case.

	AR	VAR	ARMS	VARMS	BMA	TVW	TVW(λ, τ)
RMSPE	0.882	0.875	0.907	1.000	0.885	0.799	0.691
CW		1.625	1.274	1.587	-0.103	7.185	7.984
LS	-1.323	-1.381	-1.403	-1.361	-2.791	-1.146	-1.151
GW		0.337	0.003	0.008	0.001	0.016	0.020
PITS	0.038	0.098	0.164	0.000	0.316	0.468	0.851

Note: AR, VAR, $ARMS$ and $VARMS$: individual models defined in Section 2. BMA: constant weights Bayesian Model Averaging. 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: Clark and West's test statistics. LS: average Logarithmic Score over the evaluation period. GW: p-value of the Wald statistics for the LS. PITS: p-value of the test of zero mean, unit variance and independence of the inverse normal cumulative distribution function transformed PIT, with a maintained assumption of normality for transformed PITS.

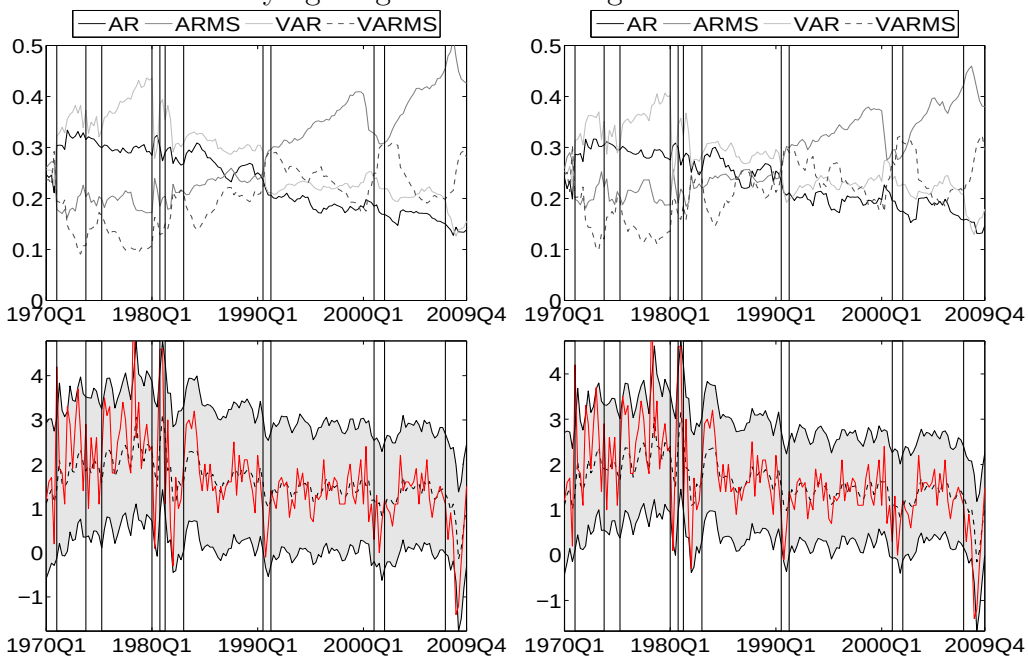
5.2 Application to GDP

First we evaluate the performance of the individual models for forecasting US GDP growth. The results in Table 1 indicate that the linear models produce the most accurate point and density forecasts. The left column of figure 3 shows that the predictive accuracy of the AR model is high in the initial 15 years of the sample and deteriorates after the structural break due to the Great Moderation. Time-varying models capture the break and their accuracy increases in the second part of the forecasting sample.

Secondly, we apply three combination schemes. The first one is a Bayesian model averaging (BMA) approach similar to Jore et al. [2010] and Hoogerheide et al. [2010]. The weights are computed as in (9) where x_k^l is equal to the cumulative log score in (31). See, e.g., Hoogerheide et al. [2010] for further details.

The other two methods are derived from our contribution in equations from (2) to (4). We only combine the i-th predictive densities of each predictor $\tilde{\mathbf{y}}_{k,t}$ of $\mathbf{y}_t$ in order to have a prediction of the i-th element of $\mathbf{y}_t$ as in equation (6). First we consider time-varying weights (TVW) with logistic-Gaussian dynamics and without learning (see equation (12)). The third scheme computes weights with learning (TVW(λ, τ))

Figure 4: Combination forecasts. Left column: time-varying weights without learning. Right column: time-varying weights with learning.



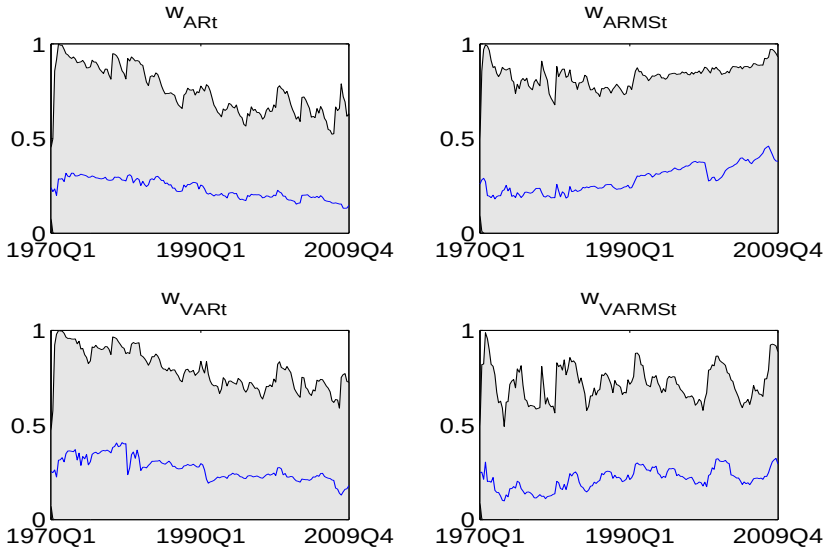
Note: Top: Average filtered weights for the GDP forecasts with models AR, ARMS, VAR e VARMS. Bottom: estimated mean (solid line) and 2.5% and 97.5% quantiles (gray area) of the marginal prediction density for $\mathbf{y}_t$. Vertical lines: NBER business cycle expansion and contraction dates.

as in (19). 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.

The results of the comparison are given in Table 1. We observe that the time-varying weights model and the TVW model with learning both outperform the standard BMA and the single models. In particular the TVW(λ, τ), with smoothing factor $\lambda = 0.95$ and window size $\tau = 9$, sensibly outperforms the TVW model in terms of RMSPE and LS. For this reason, in the multivariate setup, we consider weight updating schemes with a learning mechanism. The values of λ and τ have been chosen on the basis of the optimal RMSPE as discussed below. All the densities are correctly specified following the Berkowitz [2001] test on PITs.

The weight for the AR model in BMA is dominant, as one could expect from the results in the left column of Fig. 3. The average over the different draws of the filtered

Figure 5: 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 quantiles are obtained using the different draws from the predictive densities.

time-varying weights and the resulting approximated predictive density are, on the contrary, given for the TVW and TVW(λ, τ) schemes in the left and right columns respectively of Fig. 4. All the average weights are positive and larger than 0.1, none is above to 0.5. The average weight for the AR model is never the biggest one as in BMA and decreases over time. There are several variations in the average weights, in particular for the VARMS model. It starts low and it increases substantially in the last 10 observations of our sample, during the recent financial crisis. The weights for the TVW(λ, τ) schemes are more volatile than for the TVW scheme, but differences are very marginal. Fig. 5 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 which accounts for both model and parameter uncertainty. The figures show that the weight uncertainty is enormous and neglecting it can be very misleading.

To study the behavior of the RMSPE of the TVW(λ, τ) density combining

Table 2: Forecast accuracy for combination schemes with learning.

TVW(λ, τ)									
	$\tau = 1$			$\tau = 9$			$\tau = 20$		
λ	0.95	0.5	0.1	0.95	0.5	0.1	0.95	0.5	0.1
RMSPE	0.716	0.720	0.738	0.691	0.710	0.714	0.729	0.736	0.743
CW	7.907	7.914	8.026	7.984	8.007	7.878	8.010	8.191	8.144
LS	-1.193	-1.019	-1.024	-1.151	-1.222	-1.112	-1.177	-1.136	-1.001
GW	0.032	0.038	0.051	0.020	0.046	0.057	0.021	0.004	0.030
PITS	0.905	0.724	0.706	0.851	0.664	0.539	0.865	0.705	0.694

Note: see Tab. 1 for a detailed description.

strategy, we consider different parameter setting. Table 5.2 gives a comparison of the optimal TVW(λ, τ) prediction scheme with the TVW(λ, τ) predictions corresponding to different parameter settings.

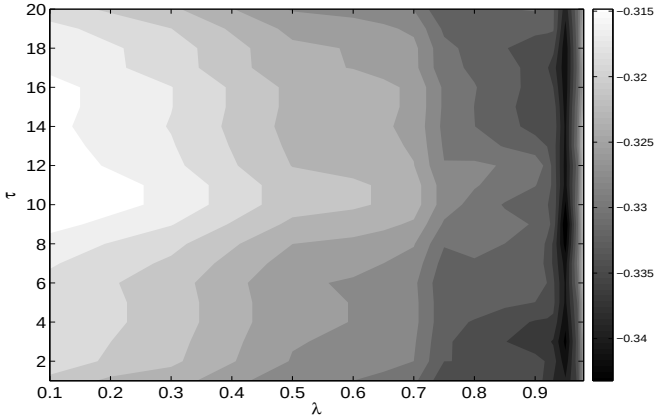
We also estimate optimal values for the smoothing parameters and the window size via a grid search. We set the grid $\lambda \in [0.1, 1]$ with step size 0.01 and $\tau \in \{1, 2, \ldots, 20\}$ with step size 1 and on the GDP dataset, for each point of the grid we iterate 10 times the SMC estimation procedure and evaluate the RMSPE. The level sets of the resulting approximated RMSPE surface are given in Fig. 6.

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 $\tau = 9$.

5.3 Multivariate Application to GDP and PCE

We extend the previous combination strategy to the multivariate prediction density of US GDP and PCE inflation. We still use $K = 4$ models, and we produce forecasts for the AR and ARMS for PCE. We use the joint predictive densities for the VAR and the VARMS. We consider the first and the third combination schemes. BMA averages models separately for GDP and PCE; our combination method is multivariate by construction and can combine forecasts for a vector of variables. We apply previous evaluation statistics and present results individually for each series of interest.

Figure 6: 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.

Results in Table 3 are very encouraging. Multivariate combination results in marginally less accurate point forecasts for GDP, but improve density forecasting in terms of LS. The TVW(λ, τ) gives the most accurate point and density forecast, and it is the only approach that suggests correct calibrated density at 5% level of significance.

Figure 7 shows that PCE average weights (or model average probability) are more volatile than GDP average probability, ARMS has an higher probability and VARMS a lower probability. VARMS seems the less adequate model even if it has the highest average LS, although we observe a reversal in this phenomenon in the last part of the sample with an increase (from 0.04 to 0.2) in the VARMS probability and a decrease (from 0.7 to 0.3) in the ARMS probability. A similar pattern for the model probabilities can be observed for GDP.

5.4 Application to Finance

We use stock returns collected from the Livingston survey and consider the nonparametric estimated density forecasts as one possible way to predict future stock

Table 3: Results for the multivariate case.

GDP						
	AR	VAR	ARMS	VARMS	BMA	TVW(λ, τ)
RMSPE	0.882	0.875	0.907	1.000	0.885	0.718
CW		1.625	1.274	1.587	-0.103	8.554
LS	-1.323	-1.381	-1.403	-1.361	-2.791	-1.012
GW		0.337	0.003	0.008	0.001	0.015
PITS	0.038	0.098	0.164	0.000	0.316	0.958
PCE						
	AR	VAR	ARMS	VARMS	BMA	TVW(λ, τ)
RMSPE	0.385	0.384	0.384	0.612	0.382	0.307
CW		1.036	1.902	1.476	1.234	6.715
LS	-1.538	-1.267	-1.373	-1.090	-1.759	-0.538
GW		0.008	0.024	0.007	0.020	0.024
PITS	0.001	0.000	0.000	0.000	0.000	0.095

Note: see Tab. 1 for a detailed description.

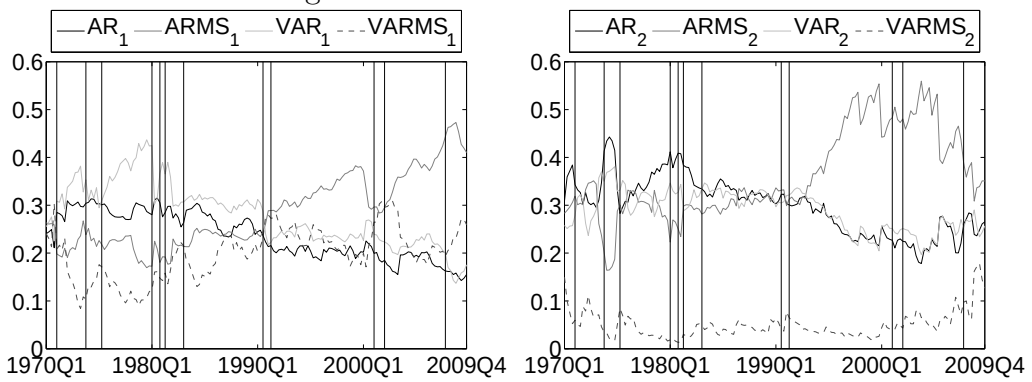
returns as discussed in Section 2. We call these survey forecasts (SR). The second alternative is a white noise model (WN).[5] This model assumes and thus forecasts that log returns are normally distributed with mean and standard deviation equal to the unconditional (up to time t for forecasting at time $t+1$) 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]). Finally, we apply our combination scheme from (2) to (4) with time-varying weights (TVW) with logistic-Gaussian dynamics and learning (see equation (12)).

We evaluate the statistical accuracy of point forecasts, the survey forecasts and the combination schemes in terms of the root mean square error (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 Kullback Leibler Information Criterion (KLIC) as in the previous section.

Moreover, as an investor is more interested in the economic value of a forecasting

[5]In the interest of brevity, we restrict this exercise to two individual models. Extensions to larger sets of individual models is straightforward.

Figure 7: Multivariate combination



Note: Time-varying weights for AR, ARMS, VAR e VARMS models for the GDP (left chart) and the PCE prediction (right chart). Vertical lines: NBER business cycle expansion and contraction dates.

model than its precision, we test our conclusions in 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.[6]

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 upon a forecast of the stock index return. We do not allow for short-sales or leveraging, constraining α_{t+1} to be in the $[0, 1]$ interval (see Barberis [2000]). The investor is assumed to maximize a power utility function with coefficient γ of relative risk aversion:

$$u(R_{t+1}) = \frac{R_{t+1}^{1-\gamma}}{1-\gamma}, \qquad \gamma > 1, \tag{34}$$

where R_{t+1} is the wealth at time $t+1$, which is equal to

$$R_{t+1} = R_t \left((1-\alpha_{t+1})\exp(y_{f,t+1}) + \alpha_{t+1}\exp(y_{f,t+1} + \tilde{y}_{t+1})\right), \tag{35}$$

where R_t denotes initial wealth, $y_{f,t+1}$ the 1-step ahead risk free rate and $\tilde{y}_{t+1}$ the

[6]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.

1-step ahead forecast of the stock index return in excess of the risk free made at time t.

Without loss of generality we set initial wealth equal to one, i.e. $R_0 = 1$, such that the investor's optimization problem is given by

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

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

$$\max_{\alpha_{t+1}\in[0,1]} \int u(R_{t+1}) p(\tilde{y}_{t+1}|y_{1:t}) d\tilde{y}_{t+1}. \tag{36}$$

We approximate the integral in (36) by generating with the SMC procedure MN equally weighted independent draws $\{y^g_{t+1}, w^g_{t+1}\}_{g=1}^{MN}$ from the predictive density $p(\tilde{y}_{t+1}|y_{1:t})$, and then use a numerical optimization method to find:

$$\max_{\alpha_{t+1}\in[0,1]} \frac{1}{MN} \sum_{g=1}^{MN} \left(\frac{((1-\alpha_{t+1})\exp(y_{f,t+1}) + \alpha_{t+1}\exp(y_{f,t+1}+\tilde{y}^g_{t+1}))^{1-\gamma}}{1-\gamma} \right) \tag{37}$$

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 (34). 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 $\underline{t}$. To compare alternative strategies we compute the multiplication factor of wealth that would equate their average utilities. For example, suppose we compare two strategies A and B. The wealth provided at time $t+1$ by the two resulting portfolios is denoted as $R_{A,t+1}$ and $R_{B,t+1}$, respectively. We then determine the value of Δ such that

$$\sum_{t=\underline{t}}^{\bar{t}} u(R_{A,t+1}) = \sum_{t=\underline{t}}^{\bar{t}} u(R_{B,t+1}/\exp(r)). \tag{38}$$

Following West et al. [1993], we interpret r as the maximum performance fee the investor would be willing to pay to switch from strategy A to strategy B.[7] For comparison of multiple investment strategies, it is useful to note that – under a power utility specification – the performance fee an investor is willing to pay to switch from strategy A to strategy B can also be computed as the difference between the performance fees of these strategies with respect to a third strategy C.[8] We use this property to infer the added value of strategies based on individual models and combination schemes by computing r 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, the portfolio weights in the active investment strategies change every month, and the portfolio must be rebalanced accordingly. Hence, transaction costs play a non-trivial role and should be taken into account when evaluating the relative performance of different strategies. Rebalancing 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

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

[8] This follows from the fact that combining (38) for the comparisons of strategies A and B with C, $\sum_t u(R_{C,t+1}) = \sum_t u(R_{A,t+1}/\exp(r_A))$ and $\sum_t u(R_{C,t+1}) = \sum_t u(R_{B,t+1}/\exp(r_B))$, gives $\sum_t u(R_{A,t+1}/\exp(r_A)) = \sum_t u(R_{B,t+1}/\exp(r_B))$. Using the power utility specification in (34), this can be rewritten as $\sum_t u(R_{A,t+1}) = \sum_t u(R_{B,t+1}/\exp(r_B - r_A))$.

that transaction costs amount to a fixed percentage c on each traded dollar. Setting the initial wealth R_t equal to 1 for simplicity, transaction costs at time $t+1$ are equal to

$$c_{t+1} = 2c|\alpha_{t+1} - \alpha_t| \tag{39}$$

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 4 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, different in terms of density forecasts: the highest log score is for our combination scheme. Figure 8 plots density forecasts given by the three approaches. The density forecasts of the survey are too narrow and therefore highly penalized 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 the realization never exceeds the 2.5% and 97.5% percentiles.

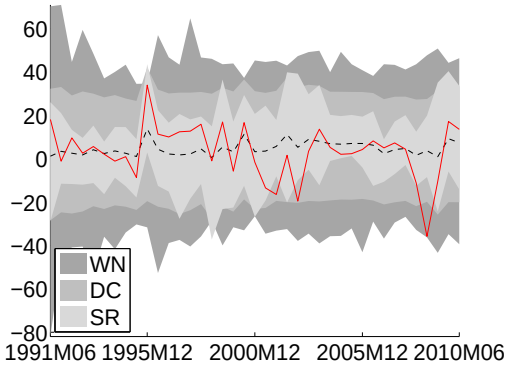
Figure 9 shows the combination weights with learning for the individual forecasts. The weights seem to converge to a $\{0, 1\}$ optimal solution, where the survey has all the weight towards the end of the period even if the uncertainty is still substantial. Changing regulations, increased sophistication 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. As our distributional state-space representation of the predictive density assumes that the model space is possible incomplete, it appears to infer

Table 4: 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	-	-	-	-	-	-
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

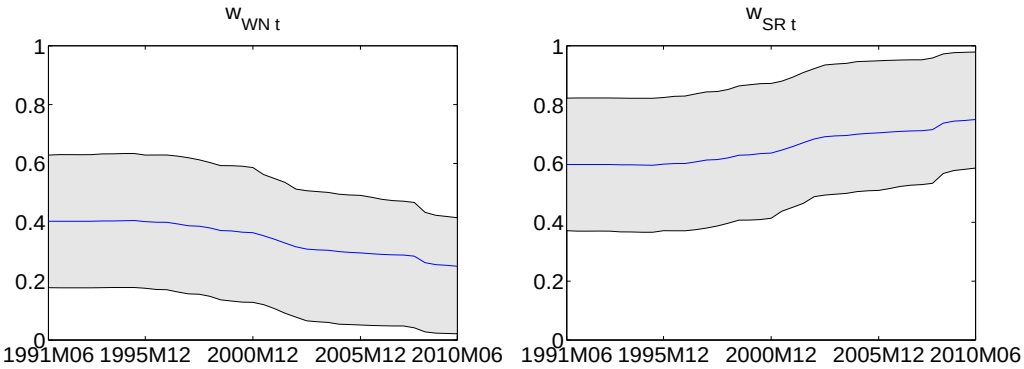
Note: In Panel A the root mean square prediction error (RMSPE), the correctly predicted sign ratio (SIGN) and the Logarithmic Score (LS) for the individual models and combination schemes in forecasting the six month ahead S&P500 index over the sample December 1990 - June 2010. WN, SR and DC denote strategies based on excess return forecasts from the White Noise model, the Livingston-based forecasts and our density combination scheme in equation (2)-(4) and (12). 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 $\gamma = (4, 6, 8)$.

Figure 8: 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.

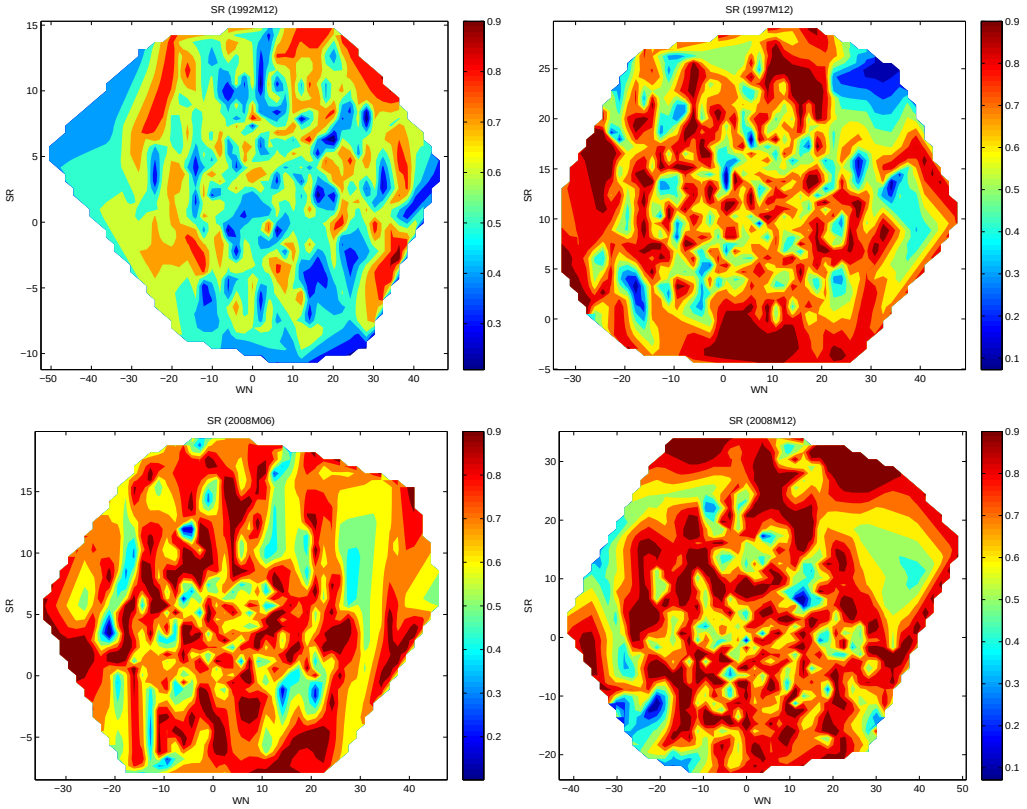
Figure 9: Combination weights for the S&P 500 forecasts



properly forecast uncertainties.

Figure 10 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 8 shows). Our methodology increases WN weights in the left

Figure 10: 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.

tail when the new forecast is made. All the four graph reveal that weights have highly nonlinear multimodal posterior distributions.

The results for the asset allocation exercise strengthen previous statistical accuracy evidence. Panel B in Table 4 reports results for three different risk aversion coefficients, $\gamma = (4, 5, 8)$. 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. Finally, results are robust to reasonable transaction costs.

6 Conclusion

This paper proposes a general combination approach with several predictive densities that are commonly used in macroeconomics and finance. The proposed method is based on a distributional state-space representation of the prediction model and of the combination scheme and on a 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 model spaces and different choices of the weight dynamics. The operational use of the method is assessed through a comparison with standard BMA on U.S. GDP and inflation forecast densities generated by some well known forecasting models and with the Standard & Poor's 500 forecast densities generated by a survey. The paper analyzes the effectiveness of the methodology in both the univariate and multivariate setup and finds that, in the application to macroeconomics, nonlinear density combination schemes with learning outperform, in terms of root mean square prediction error and the Kullback Leibler information criterion, both the BMA and the time-varying combination without learning. 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 prediction performance in terms of utility-based measures.

Appendix A - Combination schemes

Combining Prediction Density

As an alternative to the Gaussian distribution used in section 2.3, 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

$$p(\mathbf{y}_t|W_t, \tilde{\mathbf{y}}_t) \propto \left(1 + \frac{1}{\nu}\left(\mathbf{y}_t - W_t\tilde{\mathbf{y}}_t\right)'\Sigma^{-1}\left(\mathbf{y}_t - W_t\tilde{\mathbf{y}}_t\right)\right)^{-\frac{\nu+L}{2}} \tag{40}$$

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

$$p(\mathbf{y}_t|\tilde{\mathbf{y}}_t) = \sum_{k=1}^{K} p\left(W_{k,t}|\mathbf{y}_{1:t-1}, \tilde{\mathbf{y}}_{1:t-1}\right) p\left(\tilde{\mathbf{y}}_{k,t}\right) \tag{41}$$

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 Amisano and Geweke [2010].

■

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 $\mathcal{D}_K(\alpha_1, \ldots, \alpha_K)$ in order to define a Dirichlet autoregressive model.

$$\mathbf{x}_t^l \sim \mathcal{D}_{KL}\left(\eta_{1,t}^l \phi, \ldots, \eta_{KL-1,t}^l \phi, \eta_{KL,t}^l \phi\right) \tag{42}$$

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 2.3. 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.

$$
\begin{aligned}
p(\mathbf{y}_t|W_t, \Sigma_t, \tilde{\mathbf{y}}_t) &\propto \exp\left\{-\frac{1}{2}\left(\mathbf{y}_t - W_t\tilde{\mathbf{y}}_t\right)'\Sigma_t^{-1}\left(\mathbf{y}_t - W_t\tilde{\mathbf{y}}_t\right)\right\} && (43)\\
\Sigma_t &= \sum_{r=0}^{R-1} D_r \mathbb{I}_{\{r\}}(s_t) && (44)\\
s_t &\sim P(s_t = i|s_{t-1} = j) = p_{ij}, \quad \forall i, j \in \{0, \ldots, R-1\} && (45)
\end{aligned}
$$

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

$$
\begin{aligned}
p(\mathbf{x}_t|\mathbf{x}_{t-1}, \boldsymbol{\mu}_t, \tilde{\mathbf{y}}_{1:t-1}) &\propto \exp\left\{-\frac{1}{2}\left(\Delta\mathbf{x}_t - \boldsymbol{\mu}_t + \Delta\mathbf{e}_t\right)'\Lambda^{-1}\left(\Delta\mathbf{x}_t - \boldsymbol{\mu}_t + \Delta\mathbf{e}_t\right)\right\} && (46)\\
\boldsymbol{\mu}_t &= (\mu_{1,t}, \ldots, \mu_{KL^2,t})'\} && (47)\\
\mu_{l,t} &= \sum_{r=0}^{Q-1} d_{l,r}\mathbb{I}_{\{r\}}(\xi_{l,t}) && (48)\\
\xi_{l,t} &\sim P(\xi_{l,t} = i|\xi_{l,t-1} = j) = p_{ij}, && (49)
\end{aligned}
$$

$\forall i, j \in \{0, \ldots, Q-1\}$, 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

$$p(\mathbf{y}_t|W_t, s_t, \tilde{\mathbf{y}}_t) \propto \exp\left\{-\frac{1}{2}\left(\mathbf{y}_t - \sum_{k=1}^{K} \mathbf{w}_{k,t} \odot \tilde{\mathbf{y}}_{k,t}\right)' \Sigma_{s_t}^{-1} \left(\mathbf{y}_t - \sum_{k=1}^{K} \mathbf{w}_{k,t} \odot \tilde{\mathbf{y}}_{k,t}\right)\right\} \quad (50)$$

$$\Sigma_{s_t} = \Sigma\psi(s_t) + (1 - \psi(s_t))I_L \quad (51)$$

$$s_t \sim P(s_t = i|s_{t-1} = j) = p_{ij}, \quad \forall i, j \in \{0, 1\} \quad (52)$$

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

$$p(\mathbf{x}_t^l|\mathbf{x}_t^l, \boldsymbol{\mu}_t^l, \tilde{\mathbf{y}}_{1:t-1}) \propto \exp\left\{-\frac{1}{2}\left(\Delta\mathbf{x}_t^l - \boldsymbol{\mu}_t^l + \Delta\mathbf{e}_t^l\right)' \Lambda^{-1} \left(\Delta\mathbf{x}_t^l - \boldsymbol{\mu}_t^l + \Delta\mathbf{e}_t^l\right)\right\} \quad (53)$$

$$\boldsymbol{\mu}_t^l = \boldsymbol{\mu}_0 + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\xi_{l,t} \quad (54)$$

$$\xi_{l,t} \sim P(\xi_{l,t} = i|\xi_{l,t-1} = j) = p_{ij}, \quad \forall i, j \in \{0, 1\} \quad (55)$$

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 B - 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. See Creal [2009] for a recent survey on the field of sequential Monte Carlo methods and their applications to economics. 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 $\mathcal{N}_K(\mathbf{x}_t^{i,j}, \sigma_\eta I_K)$
  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

$$\tilde{\omega}_{t+1}^{i,j} \propto \omega_t^{i,j} \exp\left\{-0.5\sigma^{-2}\left(\mathbf{y}_{t+1} - \sum_{k=1}^K w_{k,t}^{i,j}\mathbf{y}_{k,t}^j\right)^2\right\}$$

  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

References

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

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.

M. Caporin and J. Pres. Modelling and forecasting wind speed intensity for weather risk management. *Computational Statistics and Data Analysis*, forthcoming, 2010.

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

D. Creal. A survey of sequential monte carlo methods for economics and finance. *Econometric Reviews*, Forthcoming(0018), 2009.

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, T. Gunther, and A. S. Tay. Evaluating Density Forecasts with Applications to Finance and Management. *International Economic Review*, 39: 863–883, 1998.

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.

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, 2010.

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.

R. Giacomini and H. White. Tests of conditional predictive ability. *Econometrica*, 74 (6):1545–1578, 2006.

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. Groen, R. Paap, and F. Ravazzolo. Real-Time Inflation Forecasting in a Changing World. Technical Report, 2009.

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.

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.

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.

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.

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

J. Mitchell and K. F. Wallis. Evaluating Density Forecasts: Forecast Combinations, Model Mixtures, Calibration and Sharpness. *Journal of Applied Econometrics*, forthcoming, 2010.

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

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.

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.