Optimal Pooling and Finite Mixture Distribution: a Comparison between Approaches to Density Forecast Combination

Giulia Mantoan
Warwick Business School

February 14, 2019

Abstract

The combination of two or more density forecasts entails a long tradition the statistics and forecasting literature. However, little attention in econometrics has been given to the finite mixture distribution as a statistical model for combining density forecasts. Combination procedures based on a mixture density distribution are able to account for parameter uncertainty in addition to weights uncertainty, which are features normally not considered in the traditional "two-step" approaches. The aim of this paper is to compare the "one-step" mixture approach with a more traditional "two step" approach for combining density forecasts. The comparison has been achieved with several Monte Carlo simulations and applications. From the comparison, the "two-step" set of procedures result to be more accurate in combining density forecasts when the sample size is small, the individual models are nonnested or when the number of forecasts to combine is high. The "one-step" is more accurate in combing density forecasts when the sample size is big enough, the individual models are nested, when data presents breaks or when the number of forecasts to combine is low.

Key words: density forecast, combination of forecasts, density forecast evaluation, macroeconomic forecasting, optimal pooling, finite mixture distribution

JEL code: C22, C53, C52, E17

1 Introduction

The problem of obtaining one, combined density forecast from multiple alternatives has been widely investigated for several decades, and empirical applications have documented the superiority of a combined forecast over its single components in terms of accuracy (Bates and Granger (1969a). Diebold and Lopez (1996) is one of the first works to highlight the need of a combination method for probability densities, they supported the thesis with economic and financial applications; the most used pooling schemes are summarized in Clemen (1989). Most of the literature in combination of density forecasts assumes the scenario in which forecasts are obtained from different sources (for example different forecasting models or opinions of several experts) and no information about how they are achieved is available [1]. To build a combined forecast that outperforms its individual components, most of the literature determines optimal weights according to a loss function. Econometric relevance of combining predictive distributions has emerged only recently from Wallis (2005) however, several fields of research have dealt with the problem of aggregating density functions [2]. Among the most recent papers, Geweke and Amisano (2011) for example, proposed a combination method based on attaching a weight on each individual forecast which minimises the loss function. Predictive densities of each model, as well as the parameters for any parametric model, are taken as given. The optimal pool concerns the estimation of model weights only.

Waggoner and Zha (2012) extended the optimal pooling approach estimating parameters and combination weights simultaneously using a Markov-switching mixture of DSGE and BVAR models, enabling the macroeconomist to address jointly both model uncertainty and parameter uncertainty. They also allow the weights to switch between two regimes to explore the possibility that the importance of a particular model may change over time. Modern analysis of Markov-switching dynamic models can be found in Hamilton

[1] For a more detailed discussion on opinion pooling schemes please refer to Genest and Zidek (1986) and subsequent papers, here I will address the issues of combining forecasts from several statistical models.

[2] Such as in weather forecasting (Diebold and Lopez (1996), aerospace programs (Barnard (1963), pharmaceutical (Davidian and Giltinan (1995), chemistry (Zhao and Truhlar (2008), nuclear energy (Becke (1988) and policy analysis (Timmermann (2000).

(1989), Chib (1996), Kim et al. (1999), and Frühwirth-Schnatter (2006).

However, Waggoner and Zha (2012) focuses on the empirical superiority of the combination over the two individual components and analyses the implication for misspecification in macroeconomics without discussing if and to what extent their approach is superior to the one proposed by Geweke and Amisano (2011) and all the subsequent literature on optimal pooling. The question is even more crucial considering that, by their own admission, the Markov-switching mixture of two medium-scale macroeconomic models is computationally demanding.

It is then legitimate to ask how would the mixture model proposed by Waggoner and Zha (2012) differs from the optimal pooling approach. The first seems more appealing since it enables the researcher to address both model uncertainty and parameter uncertainty jointly; however, this does not necessary support the thesis of achieving a more accurate combined forecast. Moreover, even when the superiority of one-step approach is assessed, it is unclear if it is always the case.

This paper tries to answer these questions proposing a comparison between the Geweke and Amisano (2011) optimal pool (called here "two-step" approach from the duality between the estimations of predictive densities and combination weights) and a simplified version of the mixture model used by Waggoner and Zha (2012) (called here "one-step" approach since models' parameters and weights are estimated simultaneously) in a time-series framework. Through the comparison between those two models, this paper aims to generalise the comparison to two different approaches to combination of density forecasts: the two-step and the one-step.

The comparison is indeed carried out in order to exploit weaknesses and strengths of both approaches. Several simulation setups and applications helped the researcher to apply the best combination approach to the type of framework she is facing.

A model that belongs to the two-step set of procedures is more accurate in combining density forecasts when the sample size is small, the individual models are nonnested and when the number of forecasts to combine is high. A model that belongs to the one-step set of procedures is more accurate in combing density forecasts when the sample size is

big enough, the individual models are nested, when data presents breaks and when the number of forecasts to combine is low.

The rest of the paper is organised as follows: Section (2) provides the theoretical background of the paper: after a brief literature review, two-step and one-step procedures are presented and the comparison among them introduced. Section (3) presents several frameworks via Monte Carlo simulations; Section (4) applies the comparison to real data, in particular it presents two empirical forecasting exercises (a small and a large set of forecasting models) in macroeconomics; Finally, section (5) contains concluding remarks.

2 How the two approaches differ? Some Theoretical aspects

Often the decision maker has several competing forecasts of the same variable of interest. For examples, forecasting models span from experts' believes to very sophisticated statistical models. Although they diverge, there is no reason to consider one more reliable than another. Moreover, since they reflect diverse information sets, eliciting one forecast may not be the best solution. Having a unique forecast which represents all the predictions available is not an easy task. Ideally, it represents all the information sets used to draw each forecast. Yet, this task is impossible to achieve in practice, since the information sets are unfeasible. That is why combination of forecasts attired important attention in the recent past. Several combination rules or methods have been proposed in literature. Let us introduce the framework following the notation in Gneiting et al. (2013). Let $\{F_1, \ldots, F_k, \ldots F_K\}$ be a bundle of K forecasts for y_t, the variable of interest. Consider the combination method as defined as a sequence of maps:

$$(F_1, ..., F_K) \mapsto H(F_1, ..., F_K|\boldsymbol{\eta}) \tag{1}$$

where H is a suitable combination formula which depends on parameter vector $\boldsymbol{\eta} = \eta_1 \ldots, \eta_K$. For explanatory purposes, consider the combination formula H introduced by Stone et al. (1961): the linear pool. It takes the form:

$$H(y_t|\boldsymbol{\eta}) = \sum_{k=1}^{K} \eta_k F_k(y_t^o). \tag{2}$$

where $F_k(y_t)$ are the individual forecasts drawn from y_t^o, the observed values of y_t and $\boldsymbol{\eta}$ the vector of combination weights. The linear pooling is simply a weighted average of the forecast available. While there is an extensive literature on combination of point forecasts (as summarised in Timmermann (2006), the combination of density forecasts has only recently gained renewed interest, due also to the advances in modelling and computing. Even if the first paper that has exploited combination for density forecast was Bates and Granger (1969b), perhaps Wallis (2005) was the first to explicitly treat the combination of predictive functions. For an extensive review of methods of combining experts' opinions see, among the others, Genest and Zidek (1986).

The existing literature in combination of density forecasts focuses on two-step procedures. It is called "two-step procedure" the approach that takes individual forecasts as given and combines them in the most appropiate way. This is the classical point of view on combination: the need to summarise information in one, unique forecast. But it is not the only possible approach to combine forecasts. Consider the case where all the individual forecasts come from statistical models. Instead of combining the forecasts that come from them, we can combine the models themselves. The approach consists in estimating individual forecasts' parameters and combination weights simultaneously. This approach is called "one-step procedure". This section reviews the main combination models in the literature, introduces the two-step and one-step approaches compared here and concludes with the evaluation criteria used to carry out the comparison.

2.1 Two-step procedure, the Optimal Prediction Pool

The concept of optimal linear pool was first derived by DeGroot and Mortera (1991) as combination scheme for opinions. Thanks to the widespread use of the linear pooling, it has been employed almost exclusively in empirical applications (for macroeconomic and financial variables Stock and Watson (1999), for output growth Stock and Watson (2004), for European macroeconomic series Marcellino (2004), for ECB Survey of Professional Forecasters, Genre et al. (2013). Among nonlinear pooling schemes, the logarithmic pool, firstly introduced by Dawid et al. (1995), has gained some popularity since it overcomes some of the difficulties associated with the linear pooling. It is indeed typically unimodal, less disperse and invariant to rescaling. Among the most recent works, it has to be mentioned Kascha and Ravazzolo (2010) and their attempt to pool individual forecasts using both logarithmic and linear combination methods. Some of the most recent literature focuses in the calibration issues, as first highlighted by Dawid et al. (1995), and how to find an optimal combination scheme that satisfies the calibration criterion at the same time. In particular, Ranjan and Gneiting (2010) and Gneiting et al. (2013) showed how beta-transformed linear pool can be a calibrated combination for density forecasts. Despite the fact that Bayesian Model Averaging (BMA) technique has not been created to combine predictive densities, the Bayesian way to estimate combination weights is also applicable to density forecasts. The BMA attaches the posterior probability to each individual model. It typically examines combinations of the form:

$$H(y_t|\boldsymbol{\eta}) = \sum_{k=1}^{K} \eta_k p(y_t|M_k) \tag{3}$$

where the weights η_t are posteriors for the individual models M_k. Let us denote the K models by $\{M_1, \ldots, M_K\}$, the prior probability that model k is the true model by $p(M_k)$, and y_t the variable of interest. Then the posterior probability of the model k

(and then the weights) are:

$$p(M_k|y_t) = \frac{p(y_t|M_k)p(M_k)}{\sum_{k=1}^{K} p(y_t|M_k)p(M_k)}. \tag{4}$$

Please refer to Hoeting et al. (1999) for a detailed discussion. When K is large, repeating the computation becomes difficult or infeasible. For this reason, several authors among which Zellner (1986) proposed to use more informative priors in order to make the estimation quicker and more accurate. Although practitioners use BMA technique to combine forecasts, the technique has been highly criticized. BMA indeed assumes that one of the models is the correct one, and in this way it tends to attach very high weight to one model or in this case forecast. This assumption is helpful to deal with misspecification, but for combination of forecasts is it quite unrealistic.

Let us present in details the optimal prediction pool. Following the notation of Geweke and Amisano (2011), consider a bundle of alternative models $\{F_1, \dots F_k, \dots F_K\}$ that provides predictive distributions for a vector of time series y_t. A prediction model could be any construction that produces a probability density for y_t denoted by $p(y_t; y_{t-1}^o, f_k)$. In this framework, the specific construction of $p(y_t; y_{t-1}^o, f_k)$ does not concern us, it is considered as given, exogenous. The K predictive densities are combined according to the linear scheme:

$$g(y_t) = \sum_{k=1}^{K} \eta_k p(y_t; y_{t-1}^o, F_k); \quad \text{where} \quad \sum_{k=1}^{K} \eta_k = 1; \quad \text{and} \quad \forall \eta_k \geq 0 \tag{5}$$

where y_{t-1}^o denotes the vector of realized observations. The restrictions on weights η_k are necessary and sufficient to assure that equation (5) is a density function for all values of weights and all arguments of any density function.

In order to obtain the optimal weight η_k^* the methodology applied here is the one presented in Hall and Mitchell (2007) and Conflitti et al. (2015), where optimal weights are obtained by minimising the difference between the true y_t and the combined probability densities $g(y_t)$. The difference is measured through the Kullback-Leiber Information

Criterion (KLIC) which is defined as:

$$KLIC = \int \ln\, y_t \ln \frac{y_t}{g(y_t)} dy_t = \mathbb{E}[\ln\, y_t - \ln\, g(y_t)]. \tag{6}$$

Thus, the optimal combined density forecast is $g^*(y_t) = \sum_{k=1}^{K} \eta_k^* p(y_t; y_{t-1}^o, F_k)$, where the optimal weight vector η_k^* minimises the KLIC distance by:

$$\eta_k^* = \underset{\eta}{\text{argmax}} \frac{1}{T} \sum_{t=1}^{T} \ln\, g(y_t) \tag{7}$$

where $\frac{1}{T}\sum_{t=1}^{T} \ln\, g(y_t)$ is the average logarithmic score of the combined density forecast over the sample $t = 1, \ldots, T$.

From now on, the combination procedure described in this section will be referred as two-step approach, since the combined forecast is achieved in two phases: a first stage in which the distinct forecasts are obtained, and the decision maker does not have any further information about how they were estimated; and a second, explicit stage during which the appropriate weight is attached to each forecast and the combined prediction is achieved.

2.2 One-step Procedure: The Markov Switching Autoregressive model

The second approach analysed here applies the flexible and parametric framework of the finite mixture distribution to the density forecast combination problem. Historically, the mixture distribution regarded multimodal variables that have indistinguishable conditional distributions. The cause of this identification issue arises from the impossibility to observe an underlying variable able to split observations in groups and only the compound distribution can be inferred. A random variable y_t is said to have a mixture distribution if the density $p(y_t)$ has the form:

$$p(y_t) = \sum_{k=1}^{K} \eta_k p(y_t|\theta_k) \tag{8}$$

where $p(y_t|\theta_k)$ are probability densities from a parametric family with the unknown parameters θ_k and where η_k denotes the nonnegative weights that sum up to one. The finite mixture distribution has a long tradition, originating in the nineteenth century with applications in physics and biology. Among them there are those of Behboodian (1975), and Redner and Walker (1984) on statistical characteristics of mixtures, Blischke (1963) on the special case of mixtures of discrete distributions and Holgersson and Jorner (1978) and Pearson (1894) among the first works on normal mixtures. In econometrics, statistical models for finite mixture distributions were employed by Quandt (1972) as switching regression models with unknown break point. However, the method of moments estimation used by Quandt (1972) is unable to identify the observations within a particular regime. The model was expanded by Goldfeld and Quandt (1973), introducing a Markov Switching model. The monograph Everitt (1985) presents a summary of the main finite mixture distributions.

Waggoner and Zha (2012) employed a one-step procedure to macroeconomics, using a Markov-switching mixture model to combine two models well known in macroeconomic theory, the DSGE and BVAR. Here, for simplicity the combination approach is described using autoregressive models.

Consider the standard AR(p) model:

$$y_t - \mu = \phi_1(y_{t-1} - \mu) + \cdots + \phi_p(y_{t-p} - \mu) + \epsilon_t \tag{9}$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$. Which is equivalent to

$$y_t = \phi_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t \tag{10}$$

with $\phi_0 = \mu(1 - \phi_1, \ldots, -\phi_p)$.

The MSAR model is obtained introducing a hidden Markov Chain S_t into equation (10).

Let S_t being an irreducible, aperiodic Markov chain starting from its ergodic distri-

bution $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ such that:

$$Pr(S_0 = k|\boldsymbol{\xi}) = \eta_k \tag{11}$$

The stochastic properties of S_t are described by the $(K \times K)$ transition matrix $\boldsymbol{\xi}$, where each element ξ_{jk} is equal to the transition probability from the state j to state k:

$$\xi_{jk} = Pr(S_t = k|S_{t-1} = j), \quad \forall j, k \in \{1, \ldots, K\} \tag{12}$$

All elements of $\xi_{j,k}$ are nonnegative and each row of the matrix $\boldsymbol{\xi}$ sum to 1.
Equation (10) becomes:

$$y_t = \phi_{S_t,0} + \phi_{S_t,1} y_{t-1} + \cdots + \phi_{S_t,p} y_{t-p} + \epsilon_t. \tag{13}$$

The notation MS(K)-AR(p) is used to denote a Markov switching model with K states and autoregressive of order p.

The one-step predictive distribution $p(y_t|y_{t-1}^o, \boldsymbol{\theta}_k)$ for y_t described by equation (13) is:

$$p(y_t|y_{t-1}^o, \boldsymbol{\theta}_k) = \sum_{k=1}^{K} p(y_t|S_t = k, y_{t-1}^o, \boldsymbol{\theta}_k) Pr(S_t = k|y_{t-1}^o, \boldsymbol{\theta}_k) \tag{14}$$

where $\boldsymbol{\theta}_k = \phi_{k,0}, \phi_{k,0}, \ldots, \phi_{k,p}, \sigma_\epsilon^2$.

2.2.1 Finite Mixture Autoregressive model as limiting case of Markov switching autoregressive model

Any standard finite mixture of distributions can be seen as a limiting case of the MSAR model in equation (13) where S_t is an i.i.d. random sequence. As said above, the

marginal distribution of y_t is given by:

$$p(y_t|y_{t-1}^0, \boldsymbol{\theta}_k) = \sum_{k=1}^{K} p(y_t|y_{t-1}^0, S_t = k, \boldsymbol{\theta}_k) Pr(S_t = k|y_{t-1}^0, \boldsymbol{\theta}_k) \tag{15}$$

Since S_t is an i.i.d. random sequence with each row of the transition matrix being equal to the weight η_k distribution, i.e.:

$$Pr(S_t = k|S_{t-1} = j) = Pr(S_t = k) = \eta_k \tag{16}$$

then equation (15) became:

$$p(y_t|y_{t-1}^0, \boldsymbol{\theta}_k) = \sum_{k=1}^{K} p(y_t|y_{t-1}^0, S_t = k, \boldsymbol{\theta}_k) \eta_k \tag{17}$$

which is identical to the general definition of combination previously defined as:

$$p(y_t|y_{t-1}^0, \boldsymbol{\theta}_k) = \sum_{k=1}^{K} \eta_k f_N(y_t; \mu_{k,t}, \sigma_k^2)$$

where $\sum_{k=1}^{K} \eta_k = 1$, $\eta_k > 0$ and where $\mu_{1,t} = \mathbb{E}(y_t|y_{t-1}^0, \boldsymbol{\theta}_1)$. Thus, a random variable y_t drawn from a standard finite mixture distribution with weight distribution $\boldsymbol{\eta}$ is observationally equivalent with a process y_t generated by a finite Markov-Switching mixture of distributions where all rows of the transition matrix of S_t are identical to $\boldsymbol{\eta}$:

$$\xi = \begin{pmatrix} \eta_1 & \dots & \eta_K \\ \vdots & & \vdots \\ \eta_1 & \dots & \eta_K. \end{pmatrix} \tag{18}$$

Weights distribute according to a a multinomial distribution:

$$\eta_k \sim M\left(1, \left[\frac{p_1 f_N(y_t; \mu_{1,t}, \sigma_1^2)}{\sum_{k=1}^{K} p_k f_N(y_t; \mu_{k,t}, \sigma_k^2)}, \dots, \frac{p_K f_N(y_t; \mu_{K,t}, \sigma_K^2)}{\sum_{k=1}^{K} p_k f_N(y_t; \mu_{k,t}, \sigma_k^2)}\right]\right) \tag{19}$$

where $\boldsymbol{p} = (p_1, ..., p_K)$, $0 \leq p_k \leq 1$ and $\sum_{k=1}^{K} p_k = 1$.

From now on, the combination procedure described in this section will be referred as "one-step approach", since the combined forecast is achieved in a single phase during which forecasts' parameters are estimated simultaneously with the combination weights.

2.3 Evaluation of Density Forecasts

Two-step and one-step are compared according to two scoring rules and one absolute criterion. Scoring rules have been developed in literature with the purpose to measure the quality of probability forecasts and to rank competing forecast models. These evaluation techniques assess the quality of probabilistic forecasts assigning a numerical score based on the observed value for the variable of interest and the forecast. Evaluation technique depends on the application of the combination and Gneiting and Raftery (2007) provides an excursus for the main scoring rules used in literature for discrete and continuous variables, and for point, quantile and interval forecasts. Here I consider three evaluation criteria for the combination of predictive densities, two scoring rules (the logarithmic score and the CRPS), and one absolute criterion: the probability integral transform (PIT).

The Logarithmic Scoring Rule Let us considerer μ as a σ-finite measure on the measurable space $(\Omega, \mathcal{A})$ and P a probability forecast which is continuous with respect to μ with μ-density p called predictive density or density forecast.
The logarithmic score is:

$$\text{LogS}\ (p, \eta) = log\ p(\eta). \tag{20}$$

This scoring rule was first proposed by Good (1952) and it is calculated as the logarithm of probability estimate for the actual outcome. The associated divergence function is the classical Kullback-Leibler distance. The logarithmic scoring rule gives higher score to a forecast that provides a high probability to the realisation of η. The aim of the

forecaster is then to maximise the log score and, for elicitation purposes, to select the forecasting model that obtains higher log score.

Continuous ranked probability score Gneiting and Raftery (2007) highlights how logarithmic score (together with others) is not sensitive to the distance, namely, the rule does not appreciate the difference between values near but not identical to the one realised. This sensitivity is crucial in the case of multimodal predictive distributions, as in our case. To address this issue, let us consider the cumulative distribution functions $f_k(y_t)$, and let $\mathbf{1}\{y \leq x\}$ denote the function that attains the value 1 if $y \leq x$ and the value 0 otherwise. The continuous ranked probability score (CRPS) proposed by Matheson and Winkler (1976) is a function of the the realised value x as well as a functional of the probability distribution function $f_k(y_t)$:

$$CRPS(f_k(y_t), x) = -\int_{-\infty}^{\infty} (f_k(y_t) - \mathbf{1}\{y \leq x\})^2 dy_t. \tag{21}$$

The use of the continuous ranked probability score has been for long time hinder by the lack of analytical expressions, until Baringhaus and Franz (2004) proposed the following:

$$CRPS(f_k(y_t), x) = 1/2\mathbb{E}|X - X'| - \mathbb{E}|X - x|, \tag{22}$$

where X and X' are independent copies of a random variable with distribution $f_k(y_t)$. Heuristically speaking, it measures the difference between the forecast $f_k(y_t)$ and a verification x. Furthermore, note that a small score indicates a good forecast.

The CRPS used here is a symmetric tail-weighted version of the CRPS proposed by Gneiting and Ranjan (2011) which rewards the forecast accuracy in the tails. The metrics is given by the score function:

$$\text{TW-CRPS } (f_k(y_t), x, (2\alpha - 1)^2) = \int_0^1 QS_\alpha(f_k(\alpha)^{-1}, x)(2\alpha - 1)^2 d\alpha \tag{23}$$

where $QS_\alpha(q, x) = 2(\boldsymbol{I}(x < q) - \alpha)(q - x)$ is the quantile score for the forecast quantile

q at the level $0 < \alpha < 1$, f_k^{-1} is the inverse distribution of the density forecast and $(2\alpha - 1)^2$ a weighting function.

Since a closed-form solution of equation (23) is not available, it can be employed large Monte Carlo samples to approximate $f_k(y_t)$ using the empirical cumulative function and to compute the integral in equation (23). Smith and Vahey (2013) proved throughout simulations that this method is more accurate than using the CRPS as derived from difference of the expectations in equation (22).

Probability Integral Transforms A popular alternative to scoring rules is the Probability Integral Transforms (PIT) of the realisation of the variable of interest with respect to its density forecasts. The PIT was first proposed by Rosenblatt (1952). Diebold et al. (1997) proposed the use of the PIT values to assess the accuracy of the predictive distribution on the basis of the forecast-observation pairs. The density forecast is called "correct" if the probability integral transforms z_t are uniform. Where:

$$z_t = \int_{-\infty}^{y_t} f_k(y_t) d(y_t). \tag{24}$$

In practice, the evaluation need a uniformity check for the PIT values, as proposed by Früiiwirth-Schnatter (1996). Even if this evaluation technique is appealing since it provides an absolute measure of accuracy of the forecasts, Hamill (2001) proved that the uniformity of PIT values is a necessary but not sufficient condition for the forecaster to be ideal.

3 Simulation Exercises

Understanding which combination procedure delivers the best combined forecast is not immediate. Since both approaches have pros and cons, a structured comparison is needed. Waggoner and Zha (2012) assessed the empirical superiority of one-step procedure combining highly parametrised models (a BVAR and DSGE) in a large dataset framework. How would the comparison changed in different framework? This section

shows how different characteristics both of data both of forecasts matter. The comparison is carried out with different simulated Data Generating Processes. They mimic different problematics: the size of dataset, the complexity of forecasting models, the number of forecasts, the level of misspecification, the nesting issue and the presence of breaks.

Section (3.1) presents the baseline case: the two approaches combine simple forecasting models; subject to different sample sizes. Section (3.2) examines how the case where one forecast model nests the other matters. Section (3.3) includes a single break in the dataset. Section (3.4) extents the framework to multiple breaks.

Section (3.1) applies the problem to a combination of two autoregressive models; more realistic simulations are included in section (3.3), where combination approaches are compared in presence of breaks.

3.1 Baseline: Combination of Two Autoregressive models

In this section a first, naive comparison is employed. One and two-step procedure combine two forecasts obtained from two autoregressive models. The aim of this exercise is to understand which combination procedure produce the most accurate combined forecast. Moreover, in order to investigate if and how the two approaches work in different environments, the sample size is allowed to vary. In a way to appreciate the effect of sample size on the performance of the two combinations. This simulation exercise assesses the predictive ability of the two combination procedures described in Section (2). For explanatory purposes, the variable of interest is distributed as a mixture of two AR models with different orders: one AR(1) and one AR(2).

$$p(y_t|y_{t-1}^o, \boldsymbol{\theta}_k) = \eta_1 f_N(y_t; \mu_{1,t}, \sigma_1^2) + \eta_2 f_N(y_t; \mu_{2,t}, \sigma_2^2) \tag{25}$$

where, $\eta_1 + \eta_2 = 1$, $\eta_k > 0$ and where $\mu_{1,t} = \mathbb{E}(y_t|y_{t-1}^o, \boldsymbol{\theta}_1) = \phi_{1,0} + \phi_{1,1} y_{t-1}$ and $\mu_{2,t} = \mathbb{E}(y_t|y_{t-1}^o, \boldsymbol{\theta}_2) = \phi_{2,0} + \phi_{2,1} y_{t-1} + \phi_{2,2} y_{t-2}$.
This model set-up has a long tradition in mixture autoregressive models for non-linear

time series. In particular, the parameters chosen to simulate data are taken from the results in Wong and Li (2000). Wong and Li (2000) uses the non-linear and bimodal Canadian lynx data to illustrate the of Mixture Autoregressive (MAR) models. They generalised the Gaussian mixture transition distribution model introduced by Le et al. (1996) to account for multimodality and heteroscedacity. They applied the model to the largely studied number of Canadian lynxes trapped in the Mackenzie River during 1821-1934. According to Wong and Li (2000), the best model to estimate this dataset is a MAR model of orders (2,2,2) i.e. a mixture of two AR, both of order two. In their setup, the mixture's components are constraint to have the same lag order. Here a DGP is simulated from the Wong and Li estimates but allowing mixture's components to have different orders.

Two-step procedure The first approach employed here to predict this DGP is a two-step procedure where two normal predictive densities have an AR process in the mean:

$$f_k = p(y_{t+1}|y_t^o, \boldsymbol{\theta}_k) = f_N(y_{t+1}; \mu_{k,t+1}, \sigma^2_{k,t+1}) \tag{26}$$

where, for $k = \{1, 2\}$, $\boldsymbol{\theta}_k = (\mu_k, \sigma_k^2)$, $\mu_{1,t+1} = \phi_{1,0} + \phi_{1,1} y_t$, and $\mu_{2,t+1} = \phi_{2,0} + \phi_{2,1} y_t + \phi_{2,2} y_{t-1}$. The first step consists in estimating through OLS the individual predictive densities. The two forecasts are then taken as given and combined in the second step according to weights η_k which are non-negative and sum up to one to ensure that the combined density

$$p(y_{t+1}|y_t^o, \boldsymbol{\theta}_k) = \eta_1 f_1 + (1 - \eta_1) f_2 \tag{27}$$

is still a probability density. As mentioned in section (2.1), optimal weights are obtained by minimising the difference between the true y_{t+1} and the combined probability densities $p(y_{t+1}|y_t^o, \boldsymbol{\theta}_k)$. The difference is measured through the Kullback-Leiber Information

Criterion (KLIC). Equation (6) becomes:

$$KLIC = \int \ln\, y_{t+1}\, \ln \frac{y_{t+1}}{p(y_{t+1}|y_t^o, \boldsymbol{\theta}_k)} dy_{t+1} = \mathbb{E}[\ln\, y_{t+1} - \ln\, p(y_{t+1}|y_t^o, \boldsymbol{\theta}_k)]. \tag{28}$$

where $p(y_{t+1}|y_t^o, \boldsymbol{\theta}_k) = \sum_{k=1}^{K} \eta_k^* f_k$ and the optimal combined density forecast in equation (7):

$$\eta_k^* = \text{argmax}\, \frac{1}{T} \sum_{t=1}^{T} \ln\, p(y_{t+1}|y_t^o, \boldsymbol{\theta}_k) \tag{29}$$

where argmax $\frac{1}{T} \sum_{t=1}^{T} \ln\, p(y_{t+1}|y_t^o, \boldsymbol{\theta}_k)$ is the average logarithmic score of the combined density forecast over the sample $t = 1, \ldots, T$. Details for the algorithm are reported in Appendix (A).

One-step procedure The second approach proposed is a finite mixture of univariate Gaussian components advocated as adequate specification for the one-step procedure. Its predictive density takes the form:

$$p(y_{t+1}|y_t^o, \boldsymbol{\theta}_k) = \sum_{k=1}^{K} p(y_{t+1}|y_t^o, \boldsymbol{\theta}_k) Pr(S_t = k|\boldsymbol{\theta}_k) \tag{30}$$

where $\boldsymbol{\theta}_k$ is the vector if size $D = 3K$ containing all the model's parameters. More explicitly, consider a mixture model of $K = 2$ normal components: $p(y_{t+1}|y_t^o, \boldsymbol{\theta}_k) = f_N(y_{t+1}; \mu_{k,t+1}, \sigma^2_{k,t+1})$ with $\mu_{k,t+1} = \mathbb{E}(y_{t+1}|y_t^o, \boldsymbol{\theta}_k)$ and $\sigma^2_{k,t+1} = Var(y_{t+1}|y_t^o, \boldsymbol{\theta}_k)$ being the conditional mean and variance. From equation (30), the predictive density $p(y_{t+1}|y_t^o, \boldsymbol{\theta}_k)$ is a mixture of two normal distributions:

$$p(y_{t+1}|y_t^o, \boldsymbol{\theta}_k) = \eta_1 f_N(y_{t+1}; \mu_{1,t+1}, \sigma^2_{1,t+1}) + (1 - \eta_1) f_N(y_{t+1}; \mu_{2,t+1}, \sigma^2_{2,t+1}). \tag{31}$$

Let us define the autoregressive process in the mean as a function of its order p_k: $\mu_{k,t+1} = \boldsymbol{\phi}_k' \boldsymbol{z}_k$ where $\boldsymbol{\phi}_k = [\phi_0, \ldots, \phi_{p_k}]$, $\boldsymbol{z}_k = [1\ y_{t:t-p_k+1}]$ and $y_{t:t-p_k+1} = \{y_t, y_{t-1}, \ldots, y_{t-p_k+1}\}$. In the case in which the mixture's components are two autoregressive models of order

$p_1 = 1$ and $p_2 = 2$: $\mu_{1,t+1} = \phi_{1,0} + \phi_{1,1} y_t$ and $\mu_{2,t+1} = \phi_{2,0} + \phi_{2,1} y_t + \phi_{2,2} y_{t-1}$.

Bayesian Inference The estimation approach employed here is the Bayesian inference technique of MCMC estimation using data augmentation and Gibbs sampling.
Assuming a Dirichlet $\mathcal{D}(e_0, \ldots, e_0)$ distribution for η_k, the posterior distribution of η_k given the indicators $\boldsymbol{S} = (S_1, S_2, \ldots, S_t, \ldots, S_T)$ (which are independent conditional on y_t and $\boldsymbol{\theta}_k$) is: $p(\eta_k|\boldsymbol{S})$:

$$p(\eta_k|\boldsymbol{S}) \sim \mathcal{D}(e_1(\boldsymbol{S}), \ldots, e_K(\boldsymbol{S}))$$

where $e_k(\boldsymbol{S}) = e_0 + N_k(\boldsymbol{S})$ and $N_k(\boldsymbol{S})$ is the number of times in which the equality $S_t = k$ is verified. The posterior conditional densities of $\boldsymbol{\phi}_k$ and σ_k^2 given the weights and all observations assigned to group k are normally distributed:

$$p(\boldsymbol{\phi}_k|\boldsymbol{\sigma}_k^2, y_t^o, \boldsymbol{S}) \sim \mathcal{N}(a_k, A_k)$$

where,

$$A_k = (A_0^{-1} + \frac{1}{\sigma_k^2}\boldsymbol{z}_k'\boldsymbol{z}_k)^{-1} \quad a_k = A_k(A_0^{-1}a_0 + \frac{1}{\sigma_k^2}\boldsymbol{z}_k' y_k)$$

and

$$p(\sigma_k|\boldsymbol{\theta}_k, y_t^o, \boldsymbol{S}^{m-1}) \sim \mathcal{G}^{-1}(c_k, C_k)$$

where:

$$c_k = c_0 + \frac{N_k}{2}, \qquad C_k = C_0 + \frac{1}{2}\epsilon_k'\epsilon_k$$

where $\epsilon_k = y_t^o - Z_k\boldsymbol{\phi}_k$

A common choice of prior takes the form:

$$p(\boldsymbol{\theta}_k) = \mathcal{D}(e_0, e_0|\boldsymbol{S}) \prod_{k=1}^{2} \mathcal{N}(\boldsymbol{\phi}_k|a_0, A_0)\mathcal{IG}(\sigma_k^2|c_0, C_0).$$

where $\mathcal{D}$ is the symmetric Dirichlet distribution and $\mathcal{IG}(\cdot|b, c)$ is the inverse Gamma

distribution with shape parameter b and scale parameter c. Details of the Algorithm can be find in Algorithm(1).

Algorithm 1 MCMC for a Normal Mixture Regression Model

Start from some initial values of $\boldsymbol{S}^0$ and repeat the following steps M times after a burn-in period long M_0.

for $m = 1, \ldots, M_0, \ldots, M + M_0$ **do**

1. parameter simulation conditional on the allocation $\boldsymbol{S}^{m-1}$:

   (a) Sample η_k from the conditional Dirichlet posterior $p(\eta_k|\boldsymbol{S})$

   (b) Sample all regression coefficients $\boldsymbol{\phi} = (\phi_{1,0}, \phi_{1,1}, \phi_{2,0}, \phi_{2,1}, \phi_{2,2})$ jointly from the posterior distribution $p(\boldsymbol{\phi}|\sigma_k^2, y_t^o, \boldsymbol{S}^{m-1}) \sim \mathcal{N}(a_k, A_k)$;

   (c) Sample each variance σ_k from the posterior distribution $\sigma_k|\boldsymbol{\phi}, y_t^o, \boldsymbol{S}^{m-1} \sim \mathcal{G}^{-1}(c_k, C_k)$

2. Classification of each observation y_t conditional on $\boldsymbol{\theta}_k$: sample each element of S_t of $\boldsymbol{S}^m$ from the conditional posterior $p(S_t|\boldsymbol{\phi}, \sigma_{\epsilon,k}^2, y_t^o)$ given by:

$$Pr(S_t = k|\boldsymbol{\phi}, \sigma_k^2, y_t^o) \propto \eta_k f_N(y_t^o; \boldsymbol{\phi}_k, \sigma_k^2)$$

end for

The posterior density estimated from the MCMC draws is:

$$p(y_{t+1}|y_t^o, \boldsymbol{\theta}_k) = \frac{1}{M + M_0} \sum_{m=1}^{M} \left(\sum_{k=1}^{K} \eta_k^m p(y_{t+1}|\theta_{k,t+1}^m) \right)$$

A starting value for the classification (i.e. $(S)^0$) is obtained thought k-means clustering of the data. The clustering partitions the data points into K clusters according to minimization of their sum over all clusters and of the within-cluster sum of point-to-cluster-centroid distances (Everitt (2004).

The term label switching has been introduced into the literature of mixture models by Redner and Walker (1984) to describe the phenomenon of invariance in the mixture likelihood under relabelling the components of a mixture model. Consider the identifi-

cation of a finite mixture distribution with K components,

$$p(y_t|\theta_k, \eta_k) = \sum_{k=1}^{K} \eta_k f_N(y_t|\boldsymbol{\theta}_k), \tag{32}$$

given a sample $y_t = (y_1, \ldots, y_T)$, the identification of equation (32) requires handling the label switching problem caused by the invariance representation with respect to reordering the components:

$$p(y_t|\boldsymbol{\theta}_k, \eta_k) = \sum_{k=1}^{K} \eta_k f_N(y_t|\boldsymbol{\theta}_k) = \sum_{k=1}^{K} \eta_{\rho(k)} f_N(y_t|\boldsymbol{\theta}_{\rho(k)}) \tag{33}$$

where ρ is an arbitrary permutation of $\{1, \ldots, K\}$. Many useful methods have been developed to force a unique labelling on draws from this posterior distribution when the number of mixture's components is known such as Celeux (1998), Celeux et al. (2000), Frühwirth-Schnatter (2001), Jasra et al. (2005), Sperrin et al. (2010). In this paper Markov Chain algorithm that takes into account the label switching is used, where the identification is achieved imposing a constraint on the components' intercepts, i.e.:

$$\phi_{0,1} < \phi_{0,2}.$$

MC Exercise Design The parameters of the first component are assumed being equal to: $\mu_1 = -1; \phi_{1,0} = -0.5; \phi_{1,1} = 0.5$, while for the second component of the mixture: $\mu_2 = 1; \phi_{2,0} = 0.7; \phi_{2,1} = 0.2 \phi_{2,2} = 0.1$.
Regarding the one-step procedure, priors have been selected to be both vague and independent. Priors for AR coefficients are $\phi_k \sim \mathcal{N}(0, 1000)$, $\sigma_k \sim \mathcal{G}^{-1}(2, 0.5)$ while equal weight is the prior for weight. A first sample of 100,000 simulated dataset has been generated from a mixture of these AR models.

In order to investigate if and how the two approaches work in different environments, three samples of different sizes (T=50, T=200, T=5000) are drawn from the same DGP expressed above.

Simulation Results Table (3) displays parameters' estimates for the two competing forecasting approaches used in this paper and DGP specification used to simulate three datasets with different sample sizes. In addition, the table shows the mean square errors between parameters' estimates and the DGP. Since the MSE is a measure of the estimation error, it has to be preferred the procedure that exhibits lower values of this criterion and it should decrease with the increasing of the sample size. Overall, both approaches perform better with the increasing of the sample size (since MSE metric drops across datasets), especially the one-step procedure which well fits the DGP with 5000 observations. Even if the superiority of the one-step approach is evident in estimating large sample sizes, it lacks in accuracy when the sample size is smaller, and using the two-step approach is a better decision. However, to properly judge which model produced the most accurate combined prediction evaluation scores are needed.

Table (4) reports three forecast evaluation criteria already presented in Section (2.3): the log score and two versions of the CRPS for the three different datasets. For the smallest sample, i.e. $T = 50$, since the two-step procedure exhibits a log score higher than the alternative one-step, it seems to provide the most accurate forecast; contradictory results can be drawn form the CRPS figures. However, with larger samples, even for $T = 200$, all three scores are in favour of the one-step procedure, since it displays higher log scores and lower CRPSs.

To stress the point of the effect of sample size in the performance of different combination procedures, I plotted the differences in log scores (the choice of this criterion is based on the fact that from the Table (4) log score shows the bigger difference between the two alternatives) varying the sample size form $T = 20$ observations to $T = 5000$. The difference between two-step and one-step procedure drops almost immediately, after the first couple of datasets (i.e. $T = 20$ and $T = 30$);

3.2 Pools of two forecasting models

Pooling schemes affect combination accuracy. This section accounts for two issues in particular: the completeness of the models sets and the nesting issue. A model set is called complete when the correct DGP is one of the competing models. Ideally, the combination procedure should identify the correct model and attach it a weight equal to one. Since two-step procedure treats individual forecasts as given, I would expect it to identify the correct model more easily then one-step approach. However, the identification would be more difficult when one model nests the other. The issue of nesting models has a long tradition in econometrics. Two-step procedure suffers from this issue, as shown in Geweke and Amisano. Given the flexibility of one-step, we can expect that it can overcome the issue and achieve the most accurate combine density forecast. The following simulation exercise aims to shed light on these issues.

The DGP for the variable of interest is a AR(1) process:

$$y_t = \theta_0 + \theta_1 y_{t-1}, + \varepsilon_t \tag{34}$$

where $\epsilon_t \overset{\text{iid}}{\sim} (0, \sigma_\varepsilon^2)$,

With regard to the individual models (to be combined by the two approaches) two complete-model-set cases are considered: a nonnested and a nested case.

- Nonnested Case: the model f_1 mimics the GDP, f_2 is a function of a variable x, independent from y_t.

$$\begin{aligned} f_1 &: y_{T+h} \sim \mathcal{N}(\hat{\theta}_0 + \hat{\theta}_1 y_{T+h-1}, \hat{\sigma}^2_{\varepsilon_1}) \\ f_2 &: y_{T+h} \sim \mathcal{N}(\hat{\beta}_0 + \hat{\beta}_1 x_{T+h-1}, \hat{\sigma}^2_{\varepsilon_2}) \end{aligned} \tag{35}$$

- Nested case: the model f_1 mimics the GDP and it is nested in the second model

f_2:

$$
\begin{aligned}
f_1 &: y_{T+h} \sim \mathcal{N}(\hat{\theta}_0 + \hat{\theta}_1 y_{T+h-1}, \hat{\sigma}^2_{\varepsilon_1}) \\
f_2 &: y_{T+h} \sim \mathcal{N}(\hat{\theta}_0 + \hat{\theta}_1 y_{T+h-1} + \hat{\beta}_1 x_{T+h-1}, \hat{\sigma}^2_{\varepsilon_2})
\end{aligned}
\tag{36}
$$

where $\hat{\theta}_0, \hat{\beta}_0, \hat{\theta}_1, \hat{\beta}_1, \hat{\sigma}^2_{\varepsilon_1}, \hat{\sigma}^2_{\varepsilon_2}$ are the parameters' estimates obtained by OLS in the two-step procedure and using a finite mixture distribution estimation for the one-step procedure. The two marginal forecasts, are then combined according to the two combination approaches already presented. The performance of nested and nonnested model sets is interesting since they are very different theoretically. The nonnested case both the marginals f_1 and f_2 will exhibits uniform PITs but f_2 will have smaller forecast error variance than f_1 and then better log scores. This fact will affect the way the two combination procedures estimates weights and parameters. Since f_2 is the true model, combining it with any other model is suboptimal. This issue can be solved attaching a weight equal to zero to the second component. In the nonnested case, the two-step procedure may be more accurate. The second step indeed will identify f_1 as the correct forecaster and attaches a weight equal to one. Parameter and weight estimation noise has a clear effect on forecast error variance. Since the one-step approach deals with them jointly; parameter uncertainty plays a crucial role, making difficult to overcome the misspecification. The flexibility of one-step procedure can, however, help in the nested case.

Figures below are examples of how nested and nonnested cases may affect the performance of combination procedure. The figures show the difference in log scores (multiply by -1) between combined density forecasts and the "best component". The "best component" has been selected taking the component of the mixture with lower Log Score. Since the Log Score multiplied by -1 has to be minimised, a positive value in the graph identify the case where the log score of the combination is higher than the one of the "best component". In this case the single model has to be preferred to its combination. Figure (2) shows the case where the DGP is an AR(1) model and the correct model

is present in both cases as first component f_1. Both combination approaches however ignore the DGP has the same specification as f_1 and combines f_1 and f_2 as specified above. We should then expect that both procedures are worse than the best component. As the right panel in figure (2) shows, the loss in accuracy using the two-step procedure instead of the best component model is less than the one-step procedure. Moreover, the larger the sample size, the smallest is the loss. Same conclusion can be drawn from the nested case: besides the case of $T = 200$, where one-step combination is better than the best component model, two-step performs better. Interesting, the two-step approach gets worse with the increasing of the sample size.

The results confirm the hypothesis that with a complete model set two-step approach is the most accurate combination procedure, expecially when the individual models are nonnested. When one model nests the other, the situation is not clear. Increasing the sample size does not increase the accuracy of combined forecasts and one-step procedure even if it performs better than in the nonnested case, it is still less accurate than the two-step procedure.

However, in presence of breaks, the performances may change. The time variation of parameters' coefficients can be capture better by the one-step procedure, especially for the nested case.

3.3 Combination in a Single-Break Scenario

This section presents a series of simulation exercises under the case of presence of structural breaks. The motivation behind this decision is twofold. First, Stock and Watson (1996) and subsequent papers find that a wide variety of economic time series are subject to structural breaks; this comparison between combination approaches can't be complete without testing their predictive ability under breaks. The second links to the evidence of instability of parameters of autoregressive models fitted to economic time series subject to structural breaks. For this reason, the following section apply the comparison between the combination procedures to framework subject to different breaks. The exercises is aim to test the hypothesis that a combination of forecasts can overcome the

lack of predictability of component models under a framework subject to breaks. The reason behind this hypothesis is that the predictability of different forecasting models varies over time and that a combination scheme can account for this. Regardless the majority of literature about breaks, this paper does not aim to detect the breaks in the time series. Indeed this information is treated as unknown. The exercise wants to assess how well combination approaches that do not consider breaks perform.

The approach is general and it allows to compares two-step and one-step procedure in combining two forecasting models. With regard to the DGP setups, it has been simulated accounting for the following factors. To ensure that the results are comparable to the existing literature, the benchmark model is the AR(1) without breaks. I introduced a breaks to this model to extend the results to the ADL specification. The variable of interest y_t is then simulated from a AR(1) model and it exhibits a break at point $t = T_b$.

$$y_t = \Phi \boldsymbol{X}_t + \varepsilon_t \tag{37}$$

where $\epsilon_t \overset{\text{iid}}{\sim} (0, \sigma_\varepsilon^2)$,

$$\boldsymbol{X}_t = \begin{bmatrix} 1 \\ y_{t-1} \\ x_{t-1} \end{bmatrix} \tag{38}$$

and x_{t-1} is an exogenous variable that has some predictive ability for y_t but it is assumed to be independent from y_t. Although this assumption is unrealistic, it is useful for the simulation at this stage and it can be relax in the future.

$$\Phi = \begin{cases} [\phi_0 \quad \phi_1 \quad \phi_2] \ , & \text{if } t < T_b \\ [\phi_0 + d_0 \quad \phi_1 + d_1 \quad \phi_2 + d_2] \ , & \text{if } T_b \leq t \leq T. \end{cases} \tag{39}$$

The variable of interest y_t is then a function of its past values and some explanatory variable x_t. x_t is assumed to be normally distributed with mean 1 and variance 0.5.

Let us impose that $\phi_2 = 0$, such that the first part of the sample is a simple AR(1)

exp. #	d_0	d_1	d_2	σ^2_ϵ	Comments
1	-0.4	0	0	0.6	small break in the intercept
2	-0.6	0	0	0.6	large break in the intercept
3	0	-0.2	0	0.6	small break in AR(1) dynamics
4	0	-0.4	0	0.6	large break in AR(1) dynamics
5	0	0	0.5	0.6	Small break in exo. var. coefficient
6	0	0	1	0.6	Large break in exo. var. coefficient
7	0	-0.2	0.5	0.6	Breaks in AR(1) and exo. var. coefficients
8	0	0	0	2	Increase in post-break variance
9	0	0	0	0.3	Decrease in post-break variance

Table 1: Simulation set-up under a single break scenario. Parameters' values are assumed to be $\phi_0 = 0.5$, $\phi_1 = 0.8$, $\phi_2 = 0$. The experiments are run under break-point at time $T_b = \tau T$, where $\tau = \{0.25, 0.50, 0.75, 0.95\}$ and $T = \{50, 200, 1000\}$ are the simple sizes.

model, and then a break is imposed to the process. The break regards: the intercept (experiments 1-2), the AR dynamics (experiments 3-4), the impact of the explanatory variable x_t on y_t (experiments 5-7), the error variance σ^2_ε (experiments 8-9). All the types of breaks are presented in Table (1). Thus DGPs in experiments number $1-5, 8, 9$ correspond to different specifications of AR(1) models.

The timing of the break has been taken into account. It has been discussed in the literature how the position of the break matters in estimating and then forecasting time series (Pesaran et al. (2006)). For this reason, a different percentage of the sample is generated by the post-break setup, i.e. $\tau = \{0.25, 0.50, 0.75, 0.95\}$. Moreover, to incorporate the framework of the previous simulation exercise, three sample sizes are examined: $T = \{50, 200, 1000\}$.

With regard to the individual models (to be combined by the two approaches) two cases are considered: a complete and an incomplete model set:

- Incomplete model set combines two misspecified models:

$$\begin{aligned} f_1 &: y_{T+h} \sim \mathcal{N}(\hat{\theta}_0 + \hat{\theta}_1 y_{T+h-1}, \hat{\sigma}^2_{\varepsilon_1}) \\ f_2 &: y_{T+h} \sim \mathcal{N}(\hat{\beta}_0 + \hat{\beta}_1 x_{T+h-1}, \hat{\sigma}^2_{\varepsilon_2}) \end{aligned} \tag{40}$$

Moreover, in absence of breaks, both models are nested in the DGP.

- Complete model set combines a misspecified model f_1 and a second model f_2 that mimic the GDP (in the no breaks scenario):

$$\begin{aligned} f_1 &: y_{T+h} \sim \mathcal{N}(\hat{\theta}_0 + \hat{\theta}_1 y_{T+h-1}, \hat{\sigma}^2_{\varepsilon_1}) \\ f_2 &: y_{T+h} \sim \mathcal{N}(\hat{\theta}_0 + \hat{\theta}_1 y_{T+h-1} + \hat{\beta}_1 x_{T+h-1}, \hat{\sigma}^2_{\varepsilon_2}) \end{aligned} \tag{41}$$

where $\hat{\theta}_0, \hat{\beta}_0, \hat{\theta}_1, \hat{\beta}_1, \hat{\sigma}^2_{\varepsilon_1}, \hat{\sigma}^2_{\varepsilon_2}$ are the parameters' estimates obtained by OLS in the two-step procedure and using a finite mixture distribution estimation for the one-step procedure.

From the previous section, I can hypothesise that two-step approach is more accurate in the nonnested case, one-step in the nested case. Since two-step tends to attach extreme values of weights to the components, it is more accurate in experiments $(1-4, 8, 9)$ (where one component mimics the structure of the DGP). On the contrary, it is supposed to be less accurate in the experiments $(5-7)$ where neither components mimics the DGP. The size of the break matters as well, large breaks make increase the parameter estimation error, especially in small samples. The one-step procedure can overcome this issue thanks to its flexibility. The sample size matters, with the increase of T two-step becomes more accurate in the nonnested case, less accurate in the nested case. The timing of the break matters as well, it seems reasonable that performances get worse when the break is located at the end of the sample because parameters and weights estimates are biased. However, in presence of a big T, the post-break subsample can be large enough to correct this biasness.

Figures (3-11) plot the relative performance of the combination procedure with respect to DGP. The relative performance consists in CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. CRPS rates then, represent the loss in accuracy the decision maker has using the combination approach instead of the correct model specification. As repeated few times in this paper, the DGP is unknown and the decision maker has to choose between the two combination approaches. The

best combination approach is the one that exploits a lower loss in accuracy.

Figures (3) and (4) regard DGPs which are AR(1) processes with a break in the intercept. As expected, in the nonnested case the two-step is better than the one-step, increasing the sample size two-step becomes more accurate, but the performance decrease when the break is large (experiment #2) especially in smaller samples. However, one-step approach does not become more accurate with large break. In the nested case one-step is more accurate than the two step and the two-step becomes worse with the increase of the sample size. The same results can be drawn for experiments (3) and (4) and experiment (8) and (9).

Figure (7) and (8) regards DGPs which are ARDL(1,1) processes with a break in the parameter of exogenous variable. When the break is small, two (one)-step procedure have to be preferred in the nonnested (nested) case. When the break is large (exp. 6, figure 8), one-step becomes much better than the two-step procedure in the nested case (as before), and in the nonnested case as well at least for small samples (i.e. $T = 50$ and $T = 200$) a part for the case in which the break is close to the end of the sample. Let us finally consider the experiment (7) where the break regards both the AR dynamics and the exogenous variable parameter. In this case the one-step procedure is more accurate than the two-step in all cases besides the large sample, nonnested case.

The results presented do not show a clear indication about the impact of break timing in our comparison.

3.4 Combination in a Multiple-BreaksScenario

To account for the possibility of multiple breaks, a double-break case is introduced considering experiments with two breaks occurring at time $t = \{T_b, T_c\}$.

$$y_t = \Phi \cdot \boldsymbol{X}_t + \eta_t \tag{42}$$

exp #	d_0	d_1	d_2	d_0^*	d_1^*	d_2^*	σ_ϵ^2	Comments
10	-0.2	0	0	0	0	0	0.6	mean reversion in intercept
11	-0.2	0	0	-0.4	0	0	0.6	decreasing trend in the intercept
12	0	-0.2	0	0	0	0	0.6	mean reversion in AR dynamics
13	0	-0.2	0	0	-0.4	0	0.6	decreasing trend in AR dynamics
14	0	0.2	0	0	0.4	0	0.6	increasing trend in AR dynamics
15	0	0	1	0	0	0	0.6	mean reversion in predictor coefficient
16	0	0	1	0	0	2	0.6	trending break in predictor coefficient
17	0	0	0	0	0	0	2	increase in post-break variance
18	0	0	0	0	0	0	0.3	decrease in post-break variance

Table 2: Simulation set-up under a double-break scenario. Parameters' values are assumed to be $\phi_0 = 0.5$, $\phi_1 = 0.8$, $\phi_2 = 0$. Experiments 10-18 are run under break-points at time $T_b = 0.33T$, and $T_c = 0.75T$ where $T = \{50, 200, 1000\}$ are the simple sizes.

where $\eta_t \overset{\text{iid}}{\sim} (0, \sigma_\eta^2)$,

$$\boldsymbol{X}_t = \begin{bmatrix} 1 \\ y_{t-1} \\ x_{t-1} \end{bmatrix} \tag{43}$$

and

$$\Phi = \begin{cases} \begin{bmatrix} \phi_0 & \phi_1 & \phi_2 \end{bmatrix}, & \text{if } t < T_b \\ \begin{bmatrix} \phi_0 + d_0 & \phi_1 + d_1 & \phi_2 + d_2 \end{bmatrix}, & \text{if } T_b \leq t \leq T_c \\ \begin{bmatrix} \phi_0 + d_0^* & \phi_1 + d_1^* & \phi_2 + d_2^* \end{bmatrix}, & \text{if } T_c \leq t \leq T \end{cases} \tag{44}$$

For simplicity, I assumed the breaks occur at one and two-third of the sample. Under this setup, there are three break segments so the AR coefficients can now either decline, increase or mean revert over the sample. To capture all these possibilities, 9 experiments are considered and presented in Table (2).

Experiment (12) for example, assumes that ϕ_1 mean reverts from 0.8 to 0.6 and back to 0.8. Conversely, experiment 13 considers ϕ_1 as result of each break which starts at 0.8, declines first to 0.6 and then to 0.4. Experiment (14) lets ϕ_1 to follow an increasing trend starting from 0.8 moving to 1 and then to 1.2 at the time of break points. Experiments (10) and (11) consider breaks in the intercept ϕ_0, experiments (15) and (16) regard breaks in the marginal coefficient ϕ_2 of x_{t-1} on y_t from zero to one and back to zero

(experiment 15) and from zero to one to two (experiment 16). Finally, experiments (17) and (18) assume an increase and decrease in the post-break variance, respectively, occurring after the second break date, T_b.

Even in this case, marginal forecasting models are allowed to be nested and not-nested:

- Non-nested case:

$$\begin{aligned} f_1 &: y_{T+h} \sim \mathcal{N}(\hat{\theta}_0 + \hat{\theta}_1 y_{T+h-1}, \hat{\sigma}^2_{\varepsilon_1}) \\ f_2 &: y_{T+h} \sim \mathcal{N}(\hat{\beta}_0 + \hat{\beta}_1 x_{T+h-1}, \hat{\sigma}^2_{\varepsilon_2}) \\ f_3 &: y_{T+h} \sim \mathcal{N}(\hat{\beta}_0 + \hat{\beta}_1 x_{T+h-1}, \hat{\sigma}^2_{\varepsilon_3}) \end{aligned} \tag{45}$$

- Nested case

$$\begin{aligned} f_1 &: y_{T+h} \sim \mathcal{N}(\hat{\theta}_0 + \hat{\theta}_1 y_{T+h-1}, \hat{\sigma}^2_{\varepsilon_1}) \\ f_2 &: y_{T+h} \sim \mathcal{N}(\hat{\theta}_0 + \hat{\theta}_1 y_{T+h-1} + \hat{\beta}_1 x_{T+h-1}, \hat{\sigma}^2_{\varepsilon_2}) \\ f_3 &: y_{T+h} \sim \mathcal{N}(\hat{\theta}_0 + \hat{\theta}_1 y_{T+h-1} + \hat{\beta}_1 x_{T+h-1}, \hat{\sigma}^2_{\varepsilon_3}) \end{aligned} \tag{46}$$

where $\hat{\theta}_0, \hat{\beta}_0, \hat{\theta}_1, \hat{\beta}_1, \hat{\sigma}^2_{\varepsilon_1}, \hat{\sigma}^2_{\varepsilon_2}$ are the parameters' estimates obtained by OLS in the two-step procedure and using a finite mixture distribution estimation for the one-step procedure. The three marginal forecasts, are then combined according to the two combination approaches already presented.

Figure (12) regards DGPs which are AR(1) processes with two break in the intercept: a mean reversion (experiment 10) and a decreasing trend. As in the single-break scenario, in the nonnested (nested) case the two-step (one-step) is better than the alternative, increasing the sample size both combination approaches become more accurate. Let us consider the three types of break imposed to the AR dynamics. Figure (13) , show how the two-step approach is better than the one-step both in the nonnested and nested cases. However, when the sample size is large, one-step has an almost identical performance than the two-step. Experiments (15) and (16) regard DGPs which are ARDL(1,1) processes with breaks in the parameter of exogenous variable. Regardless

the sign of the break imposed, one-step procedure is always better than the two-step approach, as shown in figure (14). This result supports the evidence of this type of break in the single-break scenario (experiments 5-7). Let us finally consider the experiment (17) and (18) where the break regards the error variance. Contrary of what shown in the scenario of one-break, here two-step approach performs better in all the cases considered here. All the experiments (10-18), both combination approaches become more accurate with the increasing of the sample size. The sign of changes in parameter do not affect the comparison. Thus, the conclusions drawn for the single-break scenario are verified for the multiple-break example.

3.5 Preliminary Conclusions from the simulation exercises

To conclude, in forecasting time-series with regression models that are subject to structural breaks, the choice of which combination methods to employ was not address by the main literature yet. Since the use of the one-step combination approach is computationally more elaborated, endowing the decision maker with a tool able to discriminate when the elaboration is worthy is crucial. Indeed, by the comparison done in this section several factors seem to make the difference in eliciting the combination approach: the nature of components models, the types and the numbers of break, the number of observations the variable of interest has, the position of the break(s) in the sample and the combination schemes. From the simulation exercises presented in the section, the one-step procedure delivers more accurate combined forecasts when one component nests the other, when the time-series is subject to multiple-breaks and the sample size is sufficiently large. Conversely, the two-step procedure has to be preferred when the sample size is small, when the components are nonnested.

4 Applications

This application section employs the comparison between combination approaches to real data. Generally, it exhibits breaks and medium-large sample size, two features

covered in simulations. Two set-ups are considered here: the simple case where only two competing models are combined (in section 4.1), and a more complex one, where a large set of forecasts have to be combined by two-step and one-step (in section 4.2). The number of forecasts affects the two combination procedures in a different way. One-step approach, since it estimates models' parameters and combination weights simultaneously is penalised in the second case, since it burdens the estimation procedures. While the two-step approach is not affected by this issue.

4.1 Forecasting GDP with Industrial Production Index and Employment rate

A one-step ahead in-sample density forecasting exercise is carried out to examine the predictive ability of different combination methods in forecasting real US output growth using industrial production index and non-farm payroll employment as predictors.

4.1.1 Exercise Design

One-step ahead forecasts for quarters 1985:Q1-2018:Q1 are estimated using both ex-post and real-time data; the estimation sample begins at 1964:Q1 and, for each forecast origin, models are estimated through rolling windows of the previous 80 observations.
The datasets consists in quarterly data of real output growth and monthly data of the two indicators, ex-post data are downloaded from FRED database while for the real-time exercise, the series are obtained from the Philadelphia Federal Reserve Bank's real-time database. Data for indicators were adequately transformed in quarterly data and growth rates are obtained by taking the first difference of the log of each series.

4.1.2 Predictive regressions

To estimate the GDP growth rate using the industrial production index and employment rate, the models proposed are $K = 2$ Autoregressive Distributed Lag models (ADL), one includes industrial production index and the other employment rate. Their predictive

densities are:

$$p(y_{t+1}|y_t^o, \boldsymbol{\theta}_{1,t}) = \alpha_{1,t} + \sum_{i=0}^{p} \beta_{i,t} y_{t-i}^o + \sum_{j=0}^{q} \delta_{j,t} IP_{t-j} + \varepsilon_{1,t} \tag{47}$$

where IP denotes the industrial production index, $\boldsymbol{\theta}_{1,t} = \{\alpha_{1,t}, \beta_{i,t}, \delta_{j,t}\}$ and errors are assumed to be distributed according to $\varepsilon_{1,t} \sim N(0, \sigma_{1,t}^2)$.

$$p(y_{t+1}|y_t^o, \boldsymbol{\theta}_{2,t}) = \alpha_{2,t} + \sum_{i=0}^{p} \beta_{i,t} y_{t-i}^o + \sum_{n=0}^{N} \gamma_{n,t} EMP_{t-n} + \varepsilon_{2,t} \tag{48}$$

where EMP denotes the employment growth rate, $\boldsymbol{\theta}_{2,t} = \{\alpha_{2,t}, \beta_{i,t}, \gamma_{n,t}\}$ and errors are assumed to be distributed according to $\varepsilon_{2,t} \sim N(0, \sigma_{2,t}^2)$.

The two predictive densities are then combined according to:

$$p^c(y_{t+1}|y_t^o, \boldsymbol{\theta}_{k,t}) = \sum_{k=1}^{K=2} \eta_{k,t} p(y_{t+1}|y_t^o, \boldsymbol{\theta}_{k,t}) \tag{49}$$

where $\boldsymbol{\theta}_{k,t} = \{\boldsymbol{\theta}_{1,t}, \boldsymbol{\theta}_{2,t}\}$ and the weights $\eta_{k,t}$ are non-negative and they sum up to one. Parameters $\boldsymbol{\theta}_{k,t} = \{\alpha_{1,t}, \delta_{j,t}, \alpha_{2,t}, \beta_{i,t}, \gamma_{n,t}, \eta_{k,t}\}$ are estimated according to previously refereed one-step and two-step procedures.

For the two-step procedure, parameters $\{\alpha_{1,t}, \alpha_{2,t}, \delta_{j,t}, \beta_{i,t}, \gamma_{n,t}\}$ are obtained from a OLS regression, while optimal combination weights $\eta_{k,t}$ are estimated by maximising the logarithmic score of density forecasts employing the algorithm presented in Hall and Mitchell (2007); For the one-step procedure, estimation and forecasts are obtain from a mixture ADL models, a modified version of the algorithm used for the simulation exercise.

For completeness, a benchmark model is added to the comparison exercise. A standard benchmark to forecast real GDP growth is an autoregressive model (AR) of order 1:

$$p(y_{t+1}|y_t^o, \boldsymbol{\theta}_{b,t}) = \alpha_{b,t} + \sum_{i=0}^{p=1} \beta_{b,t} y_t^o + \varepsilon_{b,t} \tag{50}$$

To estimate the density forecasts, errors are assumed to be distributed according to $\varepsilon_{b,t} \sim N(0, \sigma^2_{b,t})$.

For the benchmark model and the one-step procedure, Bayesian inference is applied with weak informative independent priors. Parameters $\{\alpha_{k,t}, \delta_{j,t}, \beta_{i,t}, \gamma_{n,t}\, \alpha_{b,t}, \beta_{b,t}$ are assumed to have a normal-inverse-gamma prior with means equal to zero and variances equal to 1000. The error variances $\sigma^2_{1,t}, \sigma^2_{2,t}, \sigma^2_{b,t}$ have an inverse-gamma distribution with degrees of freedom equal to the number of regressors including the intercept. Combination weights η_k are assumed to distribute according to a Dirichlet distribution with parameters equal to $e_{0,1} = e_{0,2} = 4$. Predictive densities are normally distributed and their medians are used as point forecasts. The posterior estimates are obtained from a sample of 10 000 replications after a burn-in phase of 1 000 simulations. Lags have been selected according to Akaike information criteria among a maximum of 5 lags to be equal to $(p = 1, q^1 = 2, q^2 = 2)$.

4.1.3 Forecast Evaluation

Using the predictive regressors presented in the previous section, I generated and evaluated both density both point forecasts. One-step and two-step forecast ability are evaluated with respect to the benchmark AR(1) model according to several criteria, already presented in Section (2.3): log score, CRPS in the version resented in Gneiting and Raftery (2007) and in its symmetric tail-weighted version proposed in Gneiting and Ranjan (2011), and the Probability Integral Transform (PIT). For completeness, the point forecast has been computed and its accuracy is measured by MSPE.

4.1.4 Results

It is interesting to stress the different change in weights among one-step and two-step procedure. As shown in Figure (16) the weights estimated through the two-step procedure are approximately equal to 0.5 for both ex-post and real-time datasets while, the one-step procedure clearly identifies periods in which the first component (ADL model with industrial production index) contributes more to achieve a more accurate forecast

and when it contribute less.

In particular, regarding the ex-post dataset, the first component of the mixture (i.e. the ADL with Industrial Production index) has a higher predictability for GDP till 1994, gaining again weight during the period 2004-2010 and in the last part of the sample; while, the second component (i.e. the ADL with Employment rate) contributes significantly more to the combined forecast in a sub-sample of roughly 9 years from 1994 to 2003 and again in 2010-2014. The same trend, but emphasised, can be tracked in the forecasting exercise using the real-time data, where the combination gives overall more weight to the first component with respect to the ex-post dataset.

To investigate the source of the difference in weights distribution between the two procedures, Figure (17) displays the probability density functions of each component of their combination. The reason for which the optimal combination for the two-step approach tends to be the equally weighted model is because the two individual forecasts display almost identical distributions. Moreover, this evidence shed light on the models' coefficients; we can indeed suppose that small values multiply exogenous variables (industrial production index and employment rate) and that the common part of the two ADLs (i.e the autoregressive) contributes more to the forecast. This evidence does not emerge in the one-step procedure: estimating all the coefficient simultaneously indeed generates much more different distributions allowing the combination approach to be more flexible.

In general, these evidences support the thesis that the two explanatory variables analysed here have a time-varying forecast ability for GDP growth rate, feature captured only by the one-step procedure.

Even if the results already presented seem to move in favour of the one-step procedure, a correct comparison among the two alternatives have to consider some proper evaluation criteria. Table (5) presents log scores, CRPSs and mean square predictive errors for one-step, two step and the benchmark model both for ex-post revised data both for real-time and in parenthesis p-values for the Diebold and Mariano (2002) t-test.

One-step procedure produces the most accurate density forecasts according to all

four criteria employed here: its average log score is higher than the one produced by two-step procedure and slightly higher than the benchmark AR(1); both CRPS measures are in favour of the one-step procedure. Figure (18) and (19) display the score difference between alternative models and the benchmark: the evidence that one-step procedure provides the most accurate density forecast is not true only for average scores but also in each point of time. From the figure (20) we can verify that one-step procedure generates the most calibrated density sample, as the empirical cumulative distribution function of its PITs is the closest one to the standard uniform cdf (black line). Despite the accuracy of density forecasts, one-step procedure does not provide the best point forecasts, its mean square predictive error indeed is the higher among the models and the most accurate forecasts are generated by the alternative two-step procedure.

4.2 Forecasting U.S. Output Growth and Inflation in a Large Macroeconomic Data Set

The previous section compares the two approaches to combination of two forecasting models. The aim of this exercise is to generate and evaluate density forecasts for real output growth and inflation using 20 predictive variables.

4.2.1 Exercise Design

Here I use Stock and Watson (2003) dataset to forecast one-step ahead output growth and inflation rates for US. The sample includes asset prices, real economic activity, wages, prices and money variables. Data are collected at quarterly frequency, updated up to 2018Q1 and adequately transformed. For example, the rates are taken in level while the rest in the natural logarithm of the difference. Stochastic or deterministic trend and seasonality have been eliminated following Stock and Watson (2003). Table (6) presents a detailed description of the variables and their transformations. The sample starts in January 1959, although some series have a later starting point. This is due to data availability constraint. One-step ahead forecasts for quarters 1985:Q1-2018:Q2 are estimated using a fixed rolling window estimation scheme with a window size of 40

observations. The posterior estimates are obtained from a sample of 10 000 replications after a burn-in phase of 7 000 simulations.

4.2.2 Predictive regressions

This exercise evaluates one-step and two-step combinations of several forecasting models. Each of those models use lags of one of K predictors in addiction to the lagged dependent variable. The forecasting model k is:

$$y_{t+1,k} = \beta_0 + \beta_1(L)X_{t,k} + \beta_2(L)y_t + \epsilon_{t+1} \tag{51}$$

where the dependent variable is either $y^k_{t+1} = 400 \ln (RGDP_{t+1}/RGDP_t)$ or $y^k_{t+1} = 400 \ln (PGDP_{t+1}/PGDP_t) - 400 \ln (PGDP_t/PGDP_{t-1})$ where RGDP and PGDP are the real GDP and the GDP deflator, respectively. $X_{t,k}$ denotes the k-th variable for $k = 1, \ldots, K$ in Stock and Watson (2003) database. The total number of variables considered in this application is $K = 20$. The dataset for output growth includes historical data for inflation, but not for output growth (and vice versa). Further, the error term ϵ_{t+1} is assumed to be Gaussian. $\beta_1(L) = \sum_{i=0}^{p} \beta_{1,i} L^i$ and $\beta_2(L) = \sum_{j=0}^{q} \beta_{2,j} L^j$, where L is the lag operator. The number of lags p and q are recursively estimated by BIC: first selecting the lag for the AR component, then the optimal lag for the additional predictor.

The K forecasting models are then combined according to a linear scheme:

$$y_{t+h,c} = \sum_{k=1}^{K} \eta_k (y_{t+1,k}) \tag{52}$$

where the weights η_k are non-negative and they sum up to one. Parameters $\Theta = \{\eta_k, \beta_0, \beta_1, \beta_2\}$ are estimated according to one-step and two-step procedures. As one-step I considered again the mixture distributions model while, as two-step, two procedures are employed. In addiction to the Conflitti et al. (2015) model, here the Bayesian Model Averaging (BMA) pooling model is considered. The reason is twofold: first it endows the comparison of a two-step Bayesian combination, then it links this work with Rossi

and Sekhposyan (2014). Indeed, it aims to evaluate predictive densities of US output growth and inflation using Stock and Watson (2003) dataset. In that case, they use ADL forecasting models and they pool densities according to BMA. The two models are conceptually different. BMA assigns weights that are proportional to models' posterior probabilities while Conflitti et al. (2015)'s weights maximise the logarithmic score of density forecasts. Moreover, the BMA assumes the "true" model is in the set of models provided, assumption that Conflitti et al. (2015) does not rely on. The BMA assumes the "true" model is in the set of models provided, assumption that Conflitti et al. (2015) does not make. For further details on BMA, please refer to section (2.1).

As benchmark model, I consider an autoregression model of order 1, where the only lagged dependent variable are used to forecast.

$$y_{t+1,b} = \beta_0 + \beta_2(L)y_t + \epsilon_{t+1} \tag{53}$$

where $\epsilon_{t+1} \sim WN(0, \sigma^2)$ and $y_{t+1,b}$ denotes the density forecast obtained by the benchmark AR(1) model.

4.2.3 Forecast Evaluation

Using the predictive regressors presented in the previous section, I generated and evaluated both density forecasts. One-step and two-step forecast ability are evaluated with respect to the benchmark AR(1) and BMA model according to several criteria, already presented in Section (2.3): log score, CRPS in the version resented in Gneiting and Raftery (2007) and in its symmetric tail-weighted version proposed in Gneiting and Ranjan (2011), the Probability Integral Transform (PIT).

4.2.4 Bayesian Inference

The estimation approach employed here is the Bayesian inference technique of MCMC estimation using two-block Gibbs sampling. In this setup, the Algorithm(2) is modified to accommodate the large number of mixture's components. Following Frühwirth-Schnatter

(2006), priors for parameters are now dependent. This modification is employed to account for the weakness of the finite mixture model (one-step procedure) and allow to have a fair comparison between combinations.

Assuming a Dirichlet $\mathcal{D}(e_0, \ldots, e_0)$ distribution for η_k, the posterior distribution of η_k given the indicators $\boldsymbol{S} = (S_1, S_2, \ldots, S_t, \ldots, S_T)$ (which are independent conditional on y_t, $\boldsymbol{\phi}_k$ and σ_k) is:

$$p(\eta_k|\boldsymbol{S}) \sim \mathcal{D}(e_1(\boldsymbol{S}), \ldots, e_K(\boldsymbol{S})) \tag{54}$$

where $e_k(\boldsymbol{S}) = e_0 + N_k(\boldsymbol{S})$ and $N_k(\boldsymbol{S})$ is the number of times in which the equality $S_t = k$ is verified. The posterior conditional densities of $\boldsymbol{\phi}_k$ and σ_k^2 given the weights and all observations assigned to group k are normally distributed:

$$p(\boldsymbol{\phi}_k|\boldsymbol{\sigma}_k^2, y_t^o, \boldsymbol{S}) \sim \mathcal{N}(a_k, A_k) \tag{55}$$

where,

$$A_k = (A_0^{-1} + \frac{1}{\sigma_k^2}\boldsymbol{z}_k'\boldsymbol{z}_k)^{-1} \quad a_k = A_k(A_0^{-1}a_0 + \frac{1}{\sigma_k^2}\boldsymbol{z}_k'y_k)$$

and

$$p(\sigma_k^2|\boldsymbol{\theta}_k, y_t^o, \boldsymbol{S}^{m-1}) \sim \mathcal{G}^{-1}(c_N, C_N) \tag{56}$$

where:

$$c_N = c_0 + \frac{N_k}{2}, \qquad C_N = C_0 + \frac{1}{2}\epsilon_k'\epsilon_k$$

and where $\epsilon_k = y_t^o - Z_k\boldsymbol{\phi}_k$. A prior dependence among the component parameters is introduced. Following Richardson and Green (1997), the parameter C_0 is treated as an unknown hyperparameter with a prior of its own.

$$p(C_0|\boldsymbol{S}^{m-1}, \phi_k, \sigma_k^2, y_t^o) \propto \prod_{k=1}^{K} p(\sigma_k^2|C_0)p(C_0) \propto \prod_{k=1}^{K}\left(C_0^{c_0}\exp\{-\frac{C_0}{\sigma_k^2}\}\right)C_0^{g_0-1}\exp\{-G_0C_0\} \tag{57}$$

which is the Kernel of a $\mathcal{G}(g_N, G_N)$-density with $g_N = G_0 + KC_0$ and $G_N = G_0 +$

$\sum_{k=1}^{K} \frac{1}{\sigma^2_{\epsilon,k}}$. The joint prior takes the form of a hierarchical independent prior:

$$p(\phi_k, \sigma_k^2, C_0) = \prod_{k=1}^{K} p(\phi_k) \prod_{k=1}^{K} p(\sigma_k^2|C_0)p(C_0) \tag{58}$$

where ϕ_k is distributed as above, and the variance has prior equal to $\sigma_k^2 \sim (c_0, C_0)$, and $C_0 \sim (g_0, G_0)$. Following Richardson and Green (1997), initial values are selected equal to $c_0 = 2$, $g_0 = 0.2$ and $G_0 = 10/R^2$ where R is the length of the observation interval.

A common choice of prior takes the form:

$$p(\boldsymbol{\theta}_k) = \mathcal{D}(e_0, e_0|\boldsymbol{S}) \prod_{k=1}^{2} \mathcal{N}(\boldsymbol{\phi}_k|a_0, A_0)\mathcal{IG}(\sigma_k^2|c_N, C_N). \tag{59}$$

where $\mathcal{D}$ is the symmetric Dirichlet distribution and $\mathcal{IG}(\cdot|b, c)$ is the inverse Gamma distribution with shape parameter b and scale parameter c. Full conditional Gibbs sampling is carried out in two steps (details in Algorithm(2). The algorthm corresponds to algorithm 8.1 in Frühwirth-Schnatter (2006) for the case of univariate normal mixture regression model (references to algorithm 6.1).

4.3 Results

Combined density forecasts for output growth are evaluated in table (7) and figures (21 -22). According to average log scores, one-step procedure is more accurate than the two-step procedure described by Conflitti et al. (2015), however two-step BMA approach beats the benchmark model, one-step approach and two-step by Conflitti et al. (2015). Average CRPS are in favour of two-step BMA combination model. Plotting scores at each point in time (figure (21), it is clear the preference for two-step BMA model against the alternatives. From PITs cumulative distribution functions in figure (22) it is clear that none of the models are well calibrated.

Same conclusions can be drawn for the combined density forecasts for inflation, evaluated in table (8) and figures (23 -24).

Algorithm 2 Uncostraint MCMC for a Normal Mixture Regression Model.

Start from some initial values of $\boldsymbol{S}^0$ and repeat the following steps M times after a burn-in period long M_0.

for $m = 1, \ldots, M_0, \ldots, M + M_0$ **do**

1. parameter simulation conditional on the allocation $\boldsymbol{S}^{m-1}$ (as in algorithm (1)):

   (a) Sample η_k from the conditional Dirichlet posterior $p(\eta_k|\boldsymbol{S})$ as in algorithm (1);

   (b) Sample each regression coefficient $\boldsymbol{\phi} = (\phi_{1,0}, \phi_{1,1}, \phi_{2,0}, \phi_{2,1}, \phi_{2,2})$ jointly from the posterior distribution $p(\boldsymbol{\phi}|\sigma_k^2, y_t^o, \boldsymbol{S}^{m-1}) \sim \mathcal{N}(a_k, A_k)$ as in algorithm (1);

   (c) Sample the random hyperparameter C_0 from $p(C_0|\boldsymbol{S}^{m-1}, \boldsymbol{x}_i\boldsymbol{\phi}_k, \sigma_k^2, y_t^o \sim \mathcal{G}(g_N, G_N)$;

   (d) Sample each variance σ_k^2 from the posterior distribution $\sigma_k|\boldsymbol{\phi}, y_t^o, \boldsymbol{S}^{m-1} \sim \mathcal{G}^{-1}(c_k, C_k)$
   Where $c_k = c_0 + \frac{N_k}{2}$ and $C_k = C_0 + \frac{1}{2}\epsilon_k'\epsilon$

2. Classification of each observation y_t conditional on $\boldsymbol{\theta}_k$: sample each element of S_i of $\boldsymbol{S}^m$ from the conditional posterior $p(S_t|\boldsymbol{\phi}, \sigma_{\epsilon,k}^2, y_t^o)$ given by:

$$Pr(S_i = k|\boldsymbol{\phi}, \sigma_k^2, y_t^o) \propto \eta_k f_N(y_t^o; \boldsymbol{x}_i\boldsymbol{\phi}_k, \sigma_k^2)$$

end for

The posterior density estimated from the MCMC draws is:

$$p(y_{t+1}|y_t^o, \boldsymbol{\theta}_k) = \frac{1}{M + M_0} \sum_{m=1}^{M} \left(\sum_{k=1}^{K} \eta_k^m p(y_{t+1}|\theta_{k,t+1}^m) \right)$$

4.4 Conclusions

In this section two-step and one-step combination models have been compared in forecasting two empirical exercises. The first exercise aims to estimate density forecast from two ADL models; one-step procedure is the most accurate combination approach in this framework. The second exercise aims to stress one weakness of the finite mixture distribution model: the failure in estimating efficiently a mixture with high number of components. A well known exercise was used to stress this issue. As expected the performance of one-step procedure drops, in this framework other combination approaches have to be preferred.

5 Conclusions and Outlook

This paper wanted to compare two branches of combination procedures, one called "two-step" and the other "one-step". Two-step and one-step are different approaches to combine density forecasts. Both well-studied in literature and used in practice, it is not clear a priori which one should be preferred. Two-step procedure has a lower complexity and the number of parameters to estimate is low. However, it ignores the way the forecasts are achieved and the data used to obtain them. On the other hand, one-step procedure can be seen as combination of information sets. The mixture of distribution assesses a certain level of flexibility, although it requires more computational effort.

The comparison has been achieved with several simulation exercises and applications. From the comparison, the two-step set of procedures result to be more accurate in combining density forecasts when the sample size is small, the individual models are nonnested or when the number of forecasts to combine is high. The one-step set of procedures is more accurate in combing density forecasts when the sample size is big enough, the individual models are nested, when data presents breaks or when the number of forecasts to combine is low.

Although the conclusions drawn from the comparison depend on the chosen combination models among the two-step and one-step bundles of models, the comparison

can be generalise to combination branches. The two models indeed are comparable in the structure, complexity and estimation approach. Both indeed are linear models and use an OLS methods for estimation; they differ only in their approach to combination. Moreover, the variety of exercises carried out in this paper covers the weaknesses and strengths of both procedures which are the same for any other two-step or one-step models. Other more sophisticated models are available in literature that can overcome some of the limits of their category, but this goes beyond the purpose of this paper.

Appendix A Conflitti et al. (2015) Algorithm for Two-step Optimal Weight Maximisation

As described in section (2), the optimality problem reduces to the maximisation of the concave cost function:

$$\Phi(\omega_j^*) = \frac{1}{T}\sum_{t=1}^{T} \ln\, g(Y_t) \tag{60}$$

where $\omega_{\hat{OPT}}$ maximises $\Phi(\omega_j^*)$ subject to the constraints $\omega_j \geq 0$ and $\sum_{j=1}^{m} \omega_j = 1$. Let us define the $T \times J$ matrix $\hat{G}$ composed by nonnegative elements $\hat{G_{tj}} = \hat{g}_t(Y_t)$. Then equation 60 can be rewritten as:

$$\Phi(\omega_j^*) = \frac{1}{T}\sum_{t=1}^{T} \ln\, (\hat{G}\omega_j). \tag{61}$$

Let as introduce the following Lagrange multiplier λ to take into account the constraints of the weights:

$$\Phi(\omega_j^*) = \frac{1}{T}\sum_{t=1}^{T} \ln\, (\hat{G}\omega_j) - \lambda\sum_{j=1}^{m} \omega_j. \tag{62}$$

Following Conflitti et al. (2015), we introduce a "surrogate" cost function depending on a vector of arbitrary weights a_j, such that:

$$\Psi_\lambda(\omega_j, a_j) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{m}\frac{\hat{G}_{jt}a_j}{\sum_{j=1}^{m}\hat{G}_{jt}a_j}\ln\left(\frac{\omega_j}{a_j}\sum_{j=1}^{m}\hat{G}_{jt}a_j\right) - \lambda\sum_{j=1}^{m}\omega_j. \tag{63}$$

Let us define the following algorithm for k numbers of iterations:

$$\omega_{j,\lambda}^{k+1} = \underset{\omega}{\text{argmax}}\ \ \Psi_\lambda(\omega_j, \omega_{j,\lambda}^{k}) \tag{64}$$

Rewriting last equation in terms of ω_j^k, the iterative algorithm becomes:

$$\omega_j^{k+1} = \omega_j^k \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{G}_{jt}}{\sum_{j=1}^{m}\hat{G}_{jt}\omega_j^k}. \tag{65}$$

The nonnegative constrain is satisfied imposing positive weights that sum to one as initial values (i.e. $\omega_j^0 = 1/m$). The iterates are expected to converge to the maximiser $\omega_{\hat{OPT}}$ due to the monotonicity of the cost function in (63) and the constraints. The algorithm has also a stop criterion based on negligible difference between two successive iterates.

References

Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *Journal of multivariate analysis*, 88(1):190–206.

Barnard, G. A. (1963). New methods of quality control. *Journal of the Royal Statistical Society. Series A (General)*, 126(2):255–258.

Bates, J. M. and Granger, C. W. (1969a). The combination of forecasts. *Journal of the Operational Research Society*, 20(4):451–468.

Bates, J. M. and Granger, C. W. (1969b). The combination of forecasts. *Journal of the Operational Research Society*, 20(4):451–468.

Becke, A. D. (1988). Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical review A*, 38(6):3098.

Behboodian, J. (1975). Structural properties and statistics of finite mixtures. *Statistical Distributions in Scientific Work*, 1:103–112.

Blischke, W. (1963). Mixtures of discrete distributions. In *Proceedings of the International Symposium on Discrete Distributions, Montreal*, pages 351–372.

Celeux, G. (1998). Bayesian inference for mixture: The label switching problem. In *Compstat*, pages 227–232. Springer.

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970.

Chib, S. (1996). Calculating posterior distributions and modal estimates in markov mixture models. *Journal of Econometrics*, 75(1):79–97.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, 5(4):559–583.

Conflitti, C., De Mol, C., and Giannone, D. (2015). Optimal combination of survey forecasts. *International Journal of Forecasting*, 31(4):1096–1103.

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data*, volume 62. CRC press.

Dawid, A., DeGroot, M., Mortera, J., Cooke, R., French, S., Genest, C., Schervish, M., Lindley, D., McConway, K., and Winkler, R. (1995). Coherent combination of experts' opinions. *Test*, 4(2):263–313.

DeGroot, M. H. and Mortera, J. (1991). Optimal linear opinion pools. *Management Science*, 37(5):546–558.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1997). Evaluating density forecasts.

Diebold, F. X. and Lopez, J. A. (1996). 8 forecast evaluation and combination. *Handbook of statistics*, 14:241–268.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144.

Everitt, B. S. (1985). *Mixture DistributionsI*. Wiley Online Library.

Everitt, B. S. (2004). Mixture distributionsi. *Encyclopedia of statistical sciences*, 7.

Frühwirth-Schnatter, S. (2001). Fully bayesian analysis of switching gaussian state space models. *Annals of the Institute of Statistical Mathematics*, 53(1):31–49.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

Früiiwirth-Schnatter, S. (1996). Recursive residuals and model diagnostics for normal and non-normal state space models. *Environmental and Ecological Statistics*, 3(4):291–309.

Genest, C. and Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, pages 114–135.

Genre, V., Kenny, G., Meyler, A., and Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, 29(1):108–121.

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

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3):411–422.

Gneiting, T., Ranjan, R., et al. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782.

Goldfeld, S. M. and Quandt, R. E. (1973). A markov model for switching regressions. *Journal of econometrics*, 1(1):3–15.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 107–114.

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

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3):550–560.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, pages 357–384.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401.

Holgersson, M. and Jorner, U. (1978). Decomposition of a mixture into normal components: a review. *International Journal of Bio-Medical Computing*, 9(5):367–392.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, pages 50–67.

Kascha, C. and Ravazzolo, F. (2010). Combining inflation density forecasts. *Journal of forecasting*, 29(1-2):231–250.

Kim, C.-J., Nelson, C. R., et al. (1999). State-space models with regime switching: classical and gibbs-sampling approaches with applications. *MIT Press Books*, 1.

Le, N. D., Martin, R. D., and Raftery, A. E. (1996). Modeling flat stretches, bursts outliers in time series using mixture transition distribution models. *Journal of the American Statistical Association*, 91(436):1504–1515.

Marcellino, M. (2004). Forecast pooling for european macroeconomic variables. *Oxford Bulletin of Economics and Statistics*, 66(1):91–112.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management science*, 22(10):1087–1096.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110.

Pesaran, M. H., Pettenuzzo, D., and Timmermann, A. (2006). Forecasting time series subject to multiple structural breaks. *The Review of Economic Studies*, 73(4):1057–1084.

Quandt, R. E. (1972). A new approach to estimating switching regressions. *Journal of the American statistical association*, 67(338):306–310.

Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):71–91.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239.

Richardson, S. and Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472.

Rossi, B. and Sekhposyan, T. (2014). Evaluating predictive densities of us output growth and inflation in a large macroeconomic data set. *International Journal of Forecasting*, 30(3):662–682.

Smith, M. and Vahey, S. (2013). Asymmetric density forecasting of us macroeconomic variables using a gaussian copula model of cross-sectional and serial dependence. *Journal of Business & Economic Statistics*, pages 1–44.

Sperrin, M., Jaki, T., and Wit, E. (2010). Probabilistic relabelling strategies for the label switching problem in bayesian mixture models. *Statistics and Computing*, 20(3):357–366.

Stock, J. H. and Watson, M. (2003). Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature*, 41(3):788–829.

Stock, J. H. and Watson, M. W. (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics*, 14(1):11–30.

Stock, J. H. and Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics*, 44(2):293–335.

Stock, J. H. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430.

Stone, M. et al. (1961). The opinion pool. *The Annals of Mathematical Statistics*, 32(4):1339–1342.

Timmermann, A. (2000). Density forecasting in economics and finance. *Journal of Forecasting*, 19(4):231–234.

Timmermann, A. (2006). Forecast combinations. *Handbook of economic forecasting*, 1:135–196.

Waggoner, D. F. and Zha, T. (2012). Confronting model misspecification in macroeconomics. *Journal of Econometrics*, 171(2):167–184.

Wallis, K. F. (2005). Combining density and interval forecasts: a modest proposal. *Oxford Bulletin of Economics and Statistics*, 67(s1):983–994.

Wong, C. S. and Li, W. K. (2000). On a mixture autoregressive model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):95–115.

Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, 81(394):446–451.

Zhao, Y. and Truhlar, D. G. (2008). Density functionals with broad applicability in chemistry. *Accounts of chemical research*, 41(2):157–167.

Table 3: Comparison between DGP of a mixture autoregressive model MAR(2,1,2) obtained by two-step and one-step procedures.

			T=50		T=200		T=5000	
		DGP	2-step	1-step	2-step	1-step	2-step	1-step
$\phi_{k,0}$	k=1	-0.5	0.8792	1.4437	0.6778	0.1001	0.5434	-0.3485
			(1.9021)	(2.8299)	(1.3871)	(1.397)	(1.0886)	(0.0281)
	k=2	0.7	0.52725	1.5288	0.48297	1.5018	0.40832	0.71812
			(0.0654)	(1.1009)	(0.0270)	(0.7606)	(0.0230)	(0.0001)
$\phi_{k,1}$	k=1	0.5	0.4443	0.0277	0.5356	0.0315	0.5483	0.2667
			(0.0007)	(0.2416)	(0.0003)	(0.2142)	(0.0084)	(0.0547)
	k=2	0.4	0.2646	-0.0057	0.3855	0.0837	0.4122	0.3617
			(0.1500)	(0.1498)	(0.3855)	(0.1070)	(0.4122)	(0.0017)
$\phi_{k,2}$	k=1	-	-	-	-	-	-	-
		-	-	-	-	-	-	-
	k=2	0.3	0.3922	0.0168	0.2818	0.0945	0.2483	0.2915
			(0.0085)	(0.0841)	(0.0003)	(0.0568)	(0.0027)	(0.0002)
σ_k	k=1	0.6	1.0304	1.7689	0.7231	0.3597	0.7892	0.5350
			(0.1852)	(0.6921)	(0.0152)	(0.1516)	(0.0358)	(0.0049)
	k=2	0.3	0.8968	0.1831	0.6662	0.5176	0.7406	0.3170
			(0.3561)	(1.1542)	(0.1341)	(0.0989)	(0.1941)	(0.0005)
η_k	k=1	0.25	0.6055	0.5670	0.4289	0.1887	0.5124	0.2003
			(0.1264)	(0.0734)	(0.0319)	(0.0571)	(0.0688)	(0.0023)
	k=2	0.75	0.3945	0.43302	0.5711	0.8113	0.4876	0.7997
			(0.1264)	(0.0734)	(0.0319)	(0.0571)	(0.0688)	(0.0023)

Notes: Table reports estimates parameters' estimates for a mixture for an AR(1) ($k = 1$) and an AR(2) ($k = 2$). Three datasets for the DGP presented in the first column are simulated, with different sample sizes: 50, 200 and 5000. In parenthesis are shown mean square errors for parameters' estimates between competing models and DGP.

Table 4: Forecast evaluation according to Log score, the Continuous Ranked Probability Score (CRPS) and its symmetric tail-weighted version (TW-CRPS) according to three simulated datasets.

	T=50			T=200			T=5000		
	DGP	2-step	1-step	DGP	2-step	1-step	DGP	2-step	1-step
Log Score	-0.9203	-0.9208	-0.9215	-0.9209	-0.9324	-0.9212	-0.9233	-0.9287	-0.9254
CRPS	0.0657	0.0690	0.0688	0.0793	0.0988	0.0786	0.0739	0.0837	0.0764
TW-CRPS	0.0657	0.0685	0.0698	0.0781	0.0977	0.0785	0.0732	0.0845	0.0769

Notes: The log score is negative: a model with higher score indicates that it performs better than the alternative; CRPS and TW˙CRPS are positive a model with lower score indicates that it performs better than the alternative.

Figure 1: Difference of Log Scores between the 2-step and the 1-step procedure for several simulated sample sizes from T=20 to T=5000.

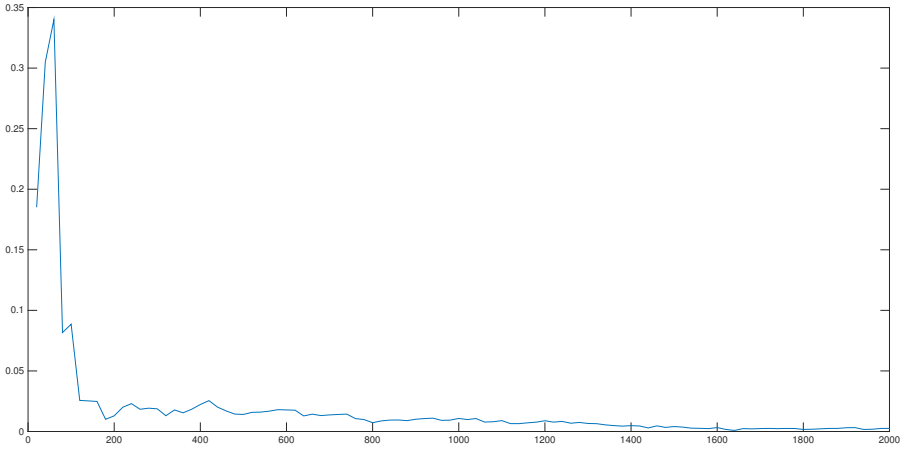
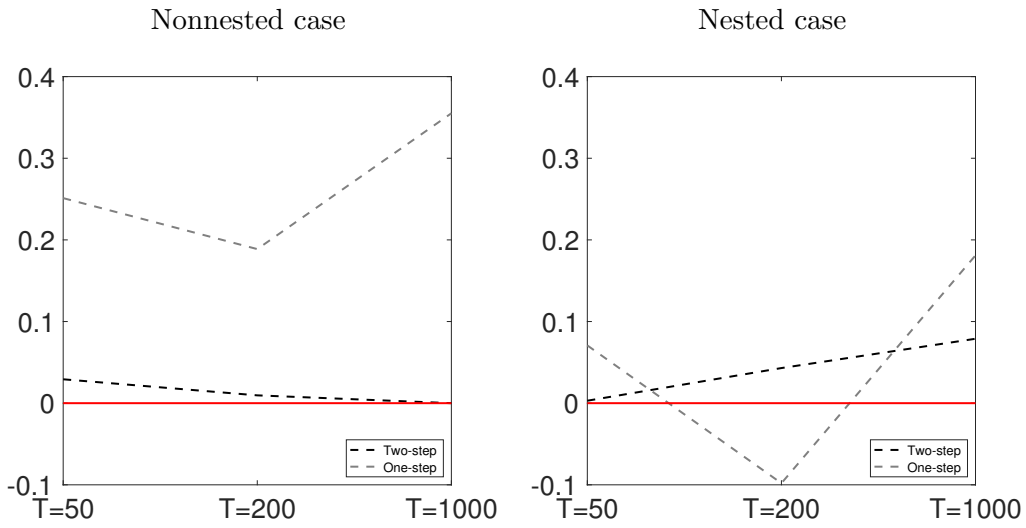
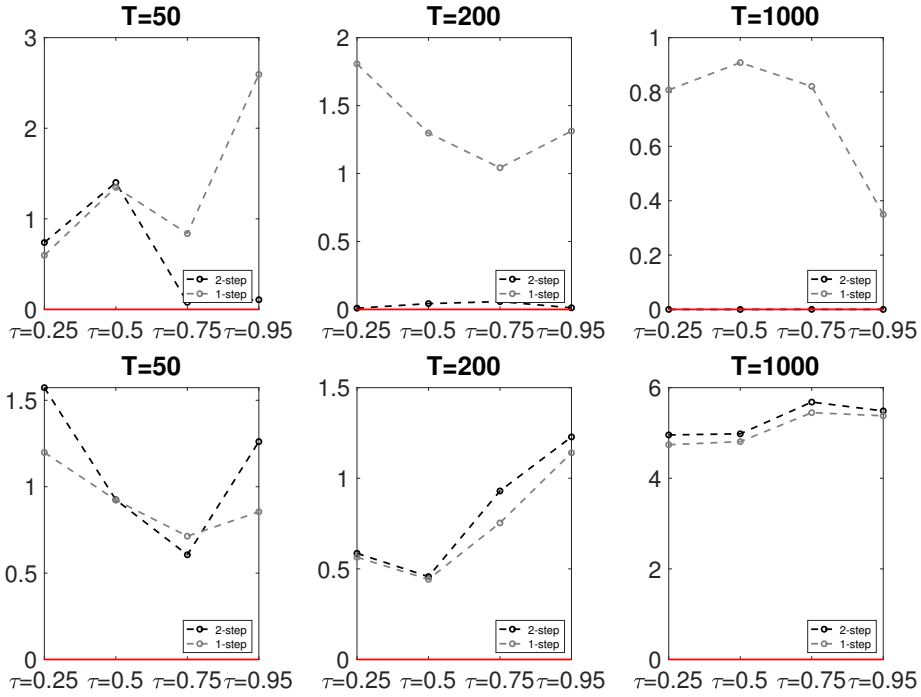


Figure 2: Absolute Difference of Log Scores between Combined Density and the Component Density with Higher Log Score in Absence of breaks (experiment 1).



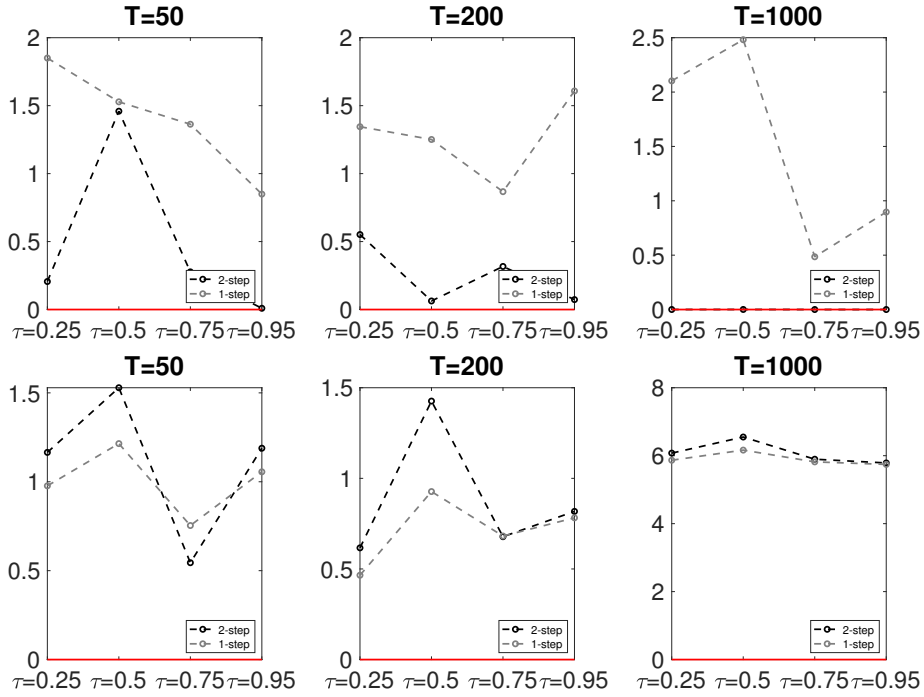
Notes: Positive numbers indicate that the single predictive performs better than the combination. Negative numbers indicate that combination performs better than the single predictive. Each point identifies the difference in scores obtained from different simulation setups, i.e. sample size ($T = 50, 200, 1000$). The left picture refers to the nonnested case, the right to the nested case.

Figure 3: Accuracy Loss of one-step and two-step in presence of a small break in the intercept (exp.#2). Nonnested case in top three graphs, nested case in the bottom three.



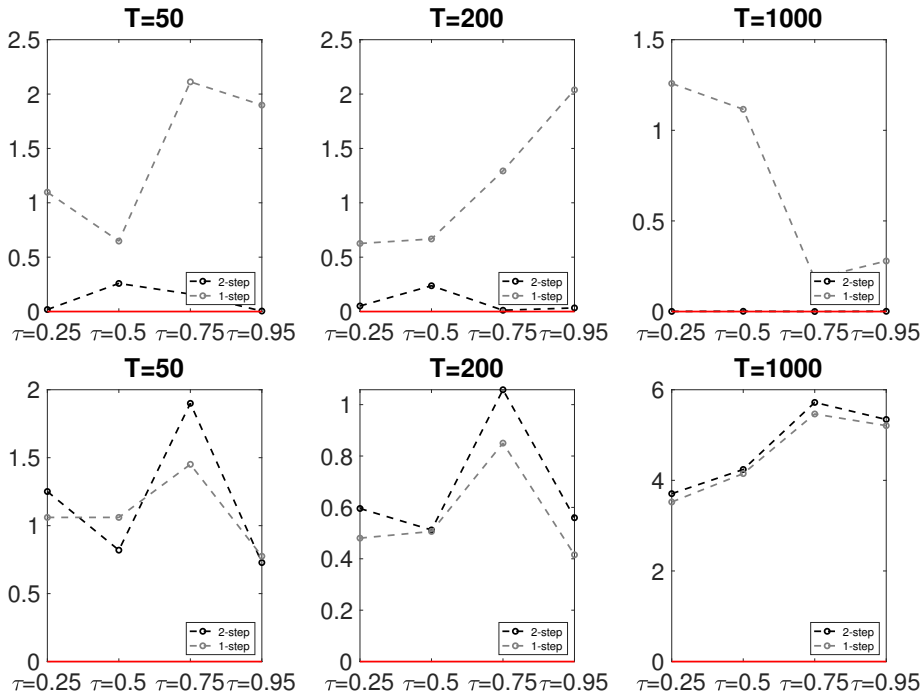
CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the GDP. Each point identifies the CRPS score obtained from different simulation setups, i.e. break timing (in X axis) and sample size ($T = 50, 200, 1000$).

Figure 4: Notes: Accuracy Loss of one-step and two-step in presence of a large break in the intercept (exp.#3). Nonnested case in top three graphs, nested case in the bottom three.



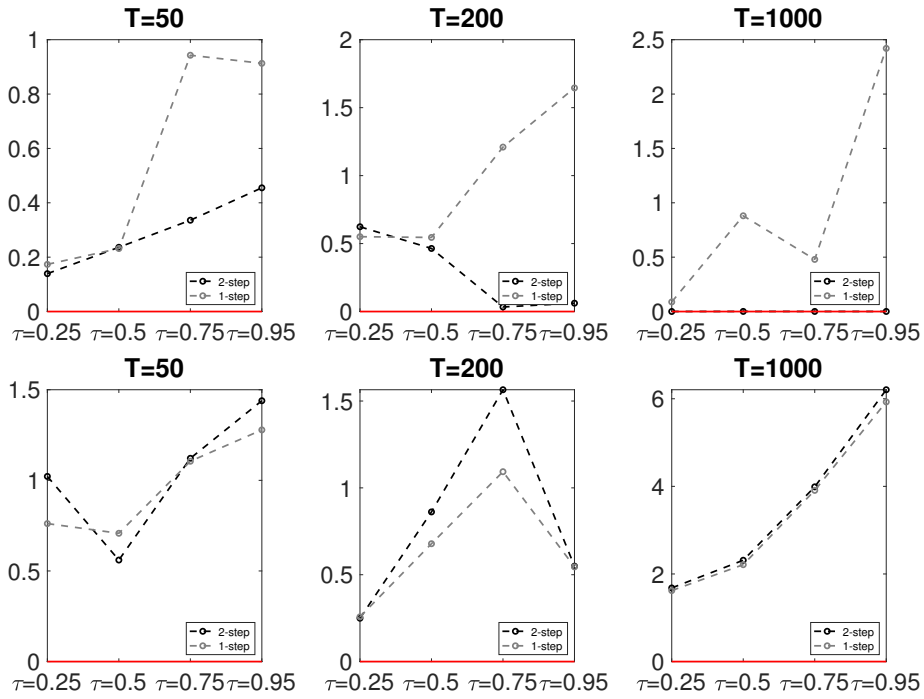
Notes: CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the GDP. Each point identifies the CRPS score obtained from different simulation setups, i.e. break timing (in X axis) and sample size ($T = 50, 200, 1000$).

Figure 5: Accuracy Loss of one-step and two-step in presence of a small break in AR(1) dynamics (exp.#4). Nonnested case in top three graphs, nested case in the bottom three.



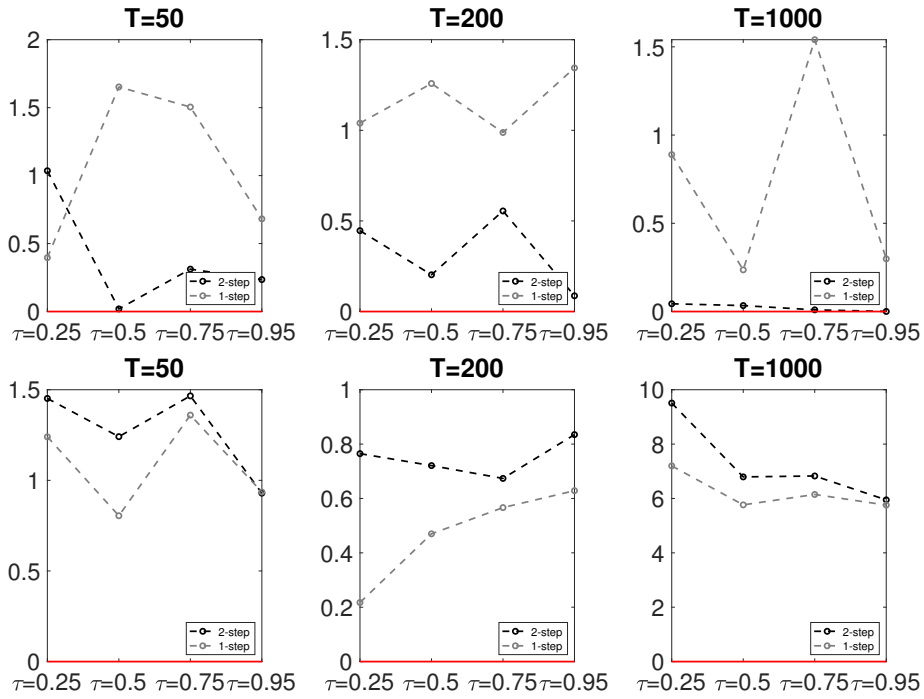
Notes: CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the GDP. Each point identifies the CRPS score obtained from different simulation setups, i.e. break timing (in X axis) and sample size ($T = 50, 200, 1000$).

Figure 6: Accuracy Loss of one-step and two-step in presence of a large break in AR(1) dynamics (exp.#5). Nonnested case in top three graphs, nested case in the bottom three.



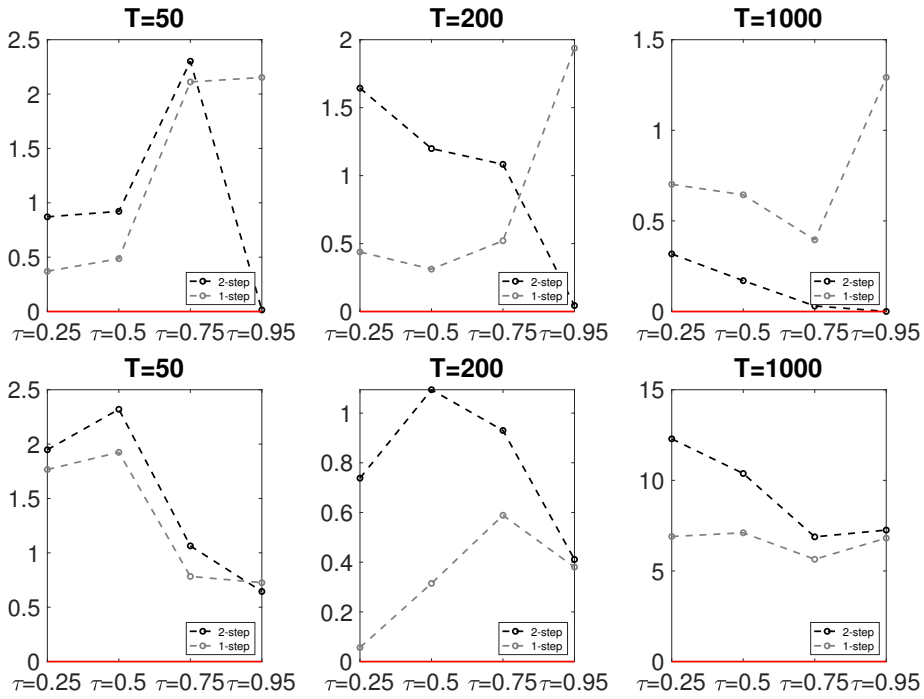
Notes: CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the GDP. Each point identifies the CRPS score obtained from different simulation setups, i.e. break timing (in X axis) and sample size ($T = 50, 200, 1000$).

Figure 7: Accuracy Loss of one-step and two-step in presence of a small break in exogenous variable coefficient (exp.#6). Nonnested case in top three graphs, nested case in the bottom three.



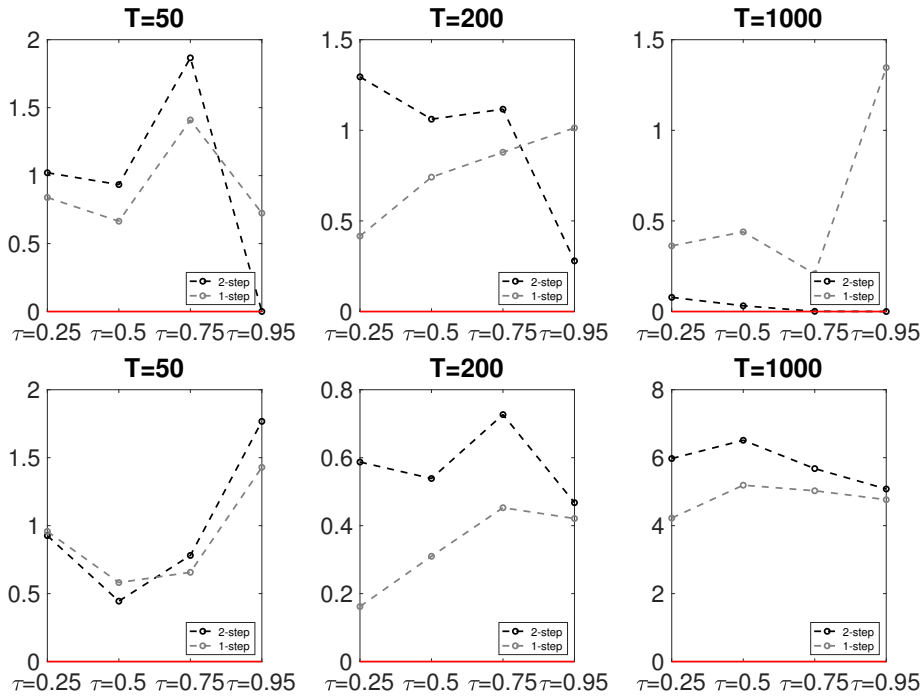
Notes: CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the GDP. Each point identifies the CRPS score obtained from different simulation setups, i.e. break timing (in X axis) and sample size ($T = 50, 200, 1000$).

Figure 8: Accuracy Loss of one-step and two-step in presence of a large break in exogenous variable coefficient (exp.#7). Nonnested case in top three graphs, nested case in the bottom three.



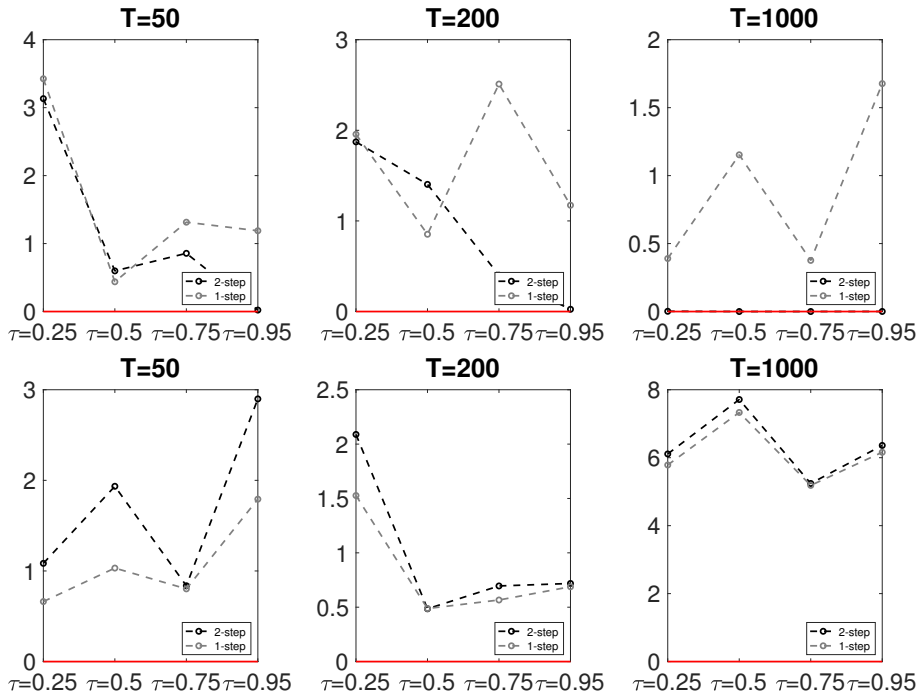
Notes: CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the GDP. Each point identifies the CRPS score obtained from different simulation setups, i.e. break timing (in X axis) and sample size ($T = 50, 200, 1000$).

Figure 9: Accuracy Loss of one-step and two-step in presence of a break in both AR(1) dynamics and exogenous variable coefficient (exp.#8). Nonnested case in top three graphs, nested case in the bottom three.



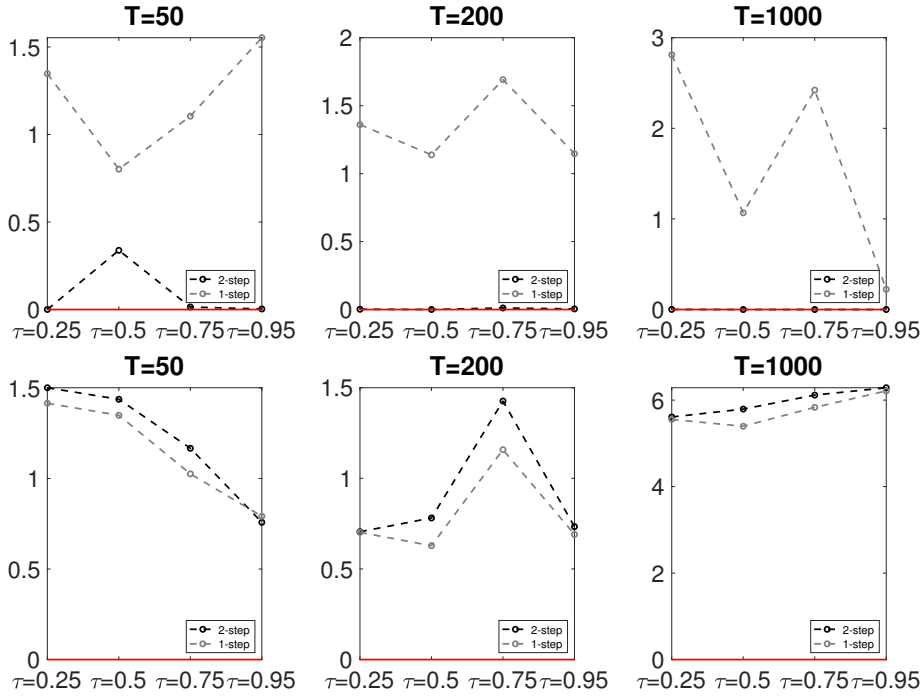
Notes: CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the GDP. Each point identifies the CRPS score obtained from different simulation setups, i.e. break timing (in X axis) and sample size ($T = 50, 200, 1000$).

Figure 10: Accuracy Loss of one-step and two-step in presence of an increase in post-break variance (exp.#7). Nonnested case in top three graphs, nested case in the bottom three.



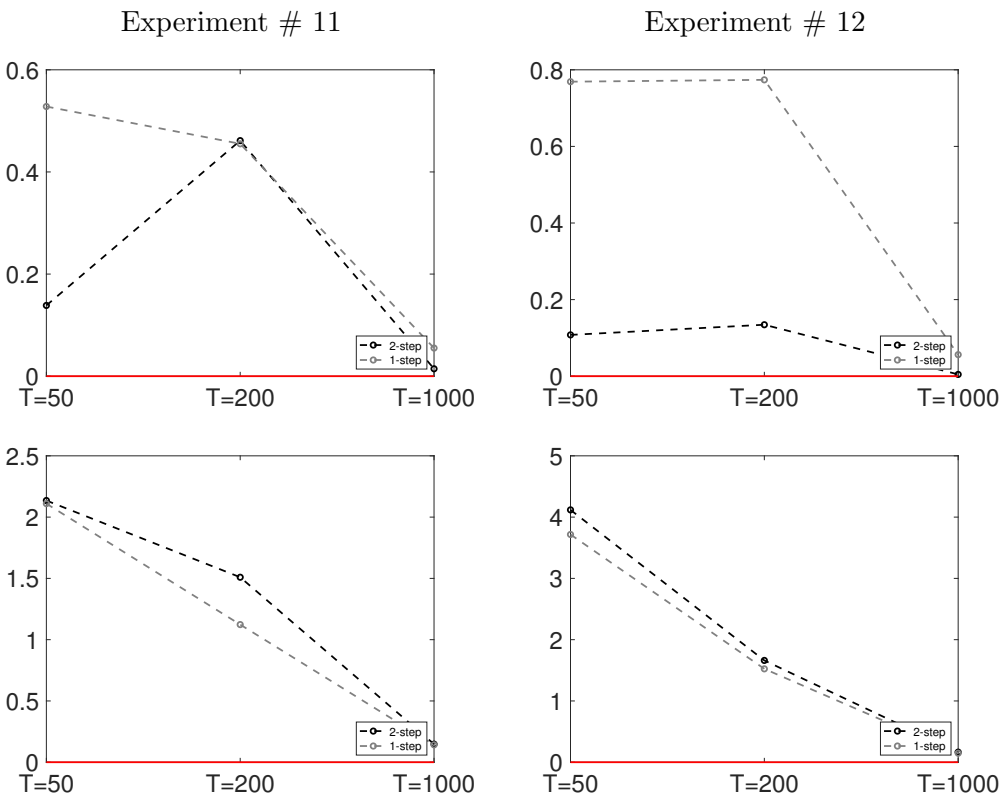
Notes: CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the GDP. Each point identifies the CRPS score obtained from different simulation setups, i.e. break timing (in X axis) and sample size ($T = 50, 200, 1000$).

Figure 11: Accuracy Loss of one-step and two-step in presence of a decrease in post-break variance (exp.#7). Nonnested case in top three graphs, nested case in the bottom three.



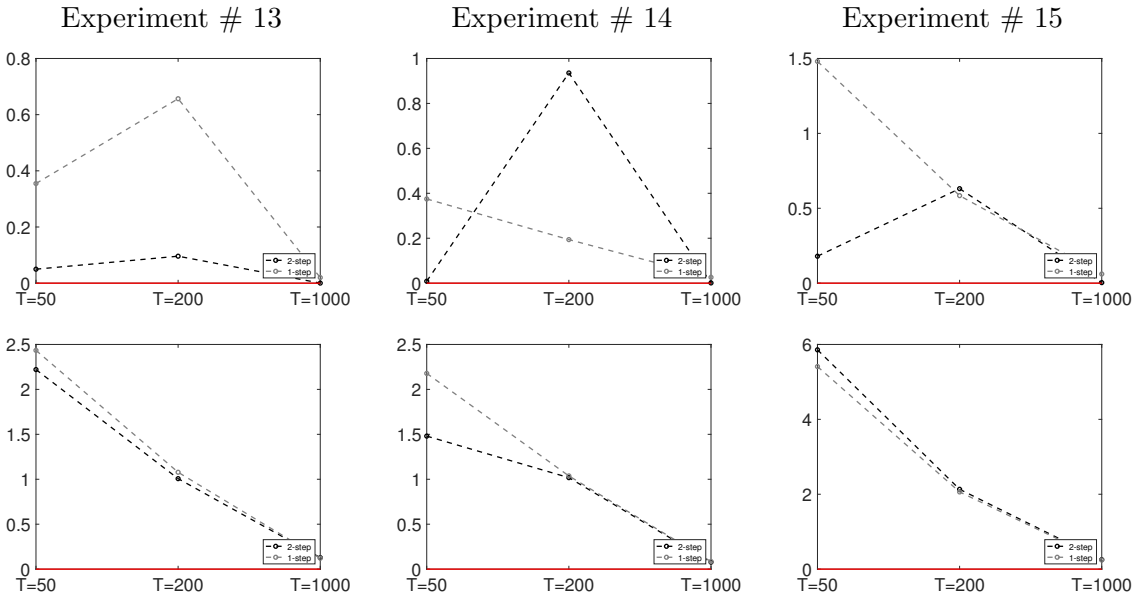
Notes: CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the GDP. Each point identifies the CRPS score obtained from different simulation setups, i.e. break timing (in X axis) and sample size ($T = 50, 200, 1000$).

Figure 12: Accuracy Loss of one-step and two-step in presence of a break in intercept. Nonnested case in top two graphs, nested case in the bottom two.



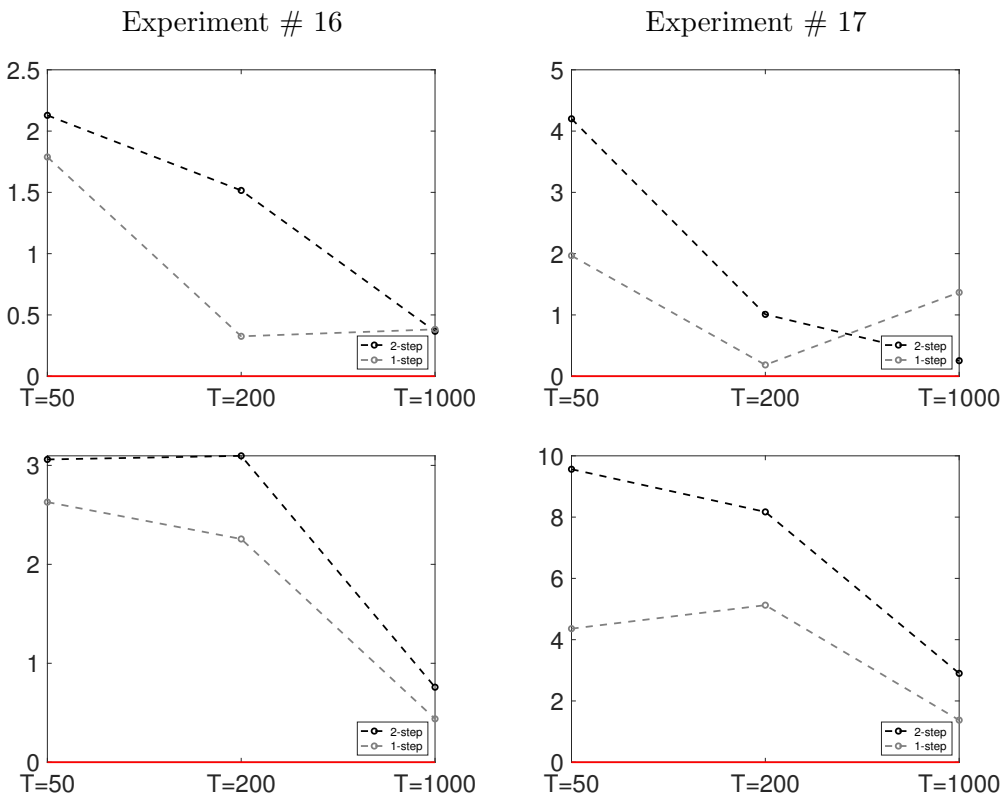
Notes: CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the GDP. Each point identifies the CRPS score obtained from different simulation setups, i.e. sample size (in X axis).

Figure 13: Accuracy Loss of one-step and two-step in presence of a break in AR dynamics. Nonnested case in top three graphs, nested case in the bottom three.



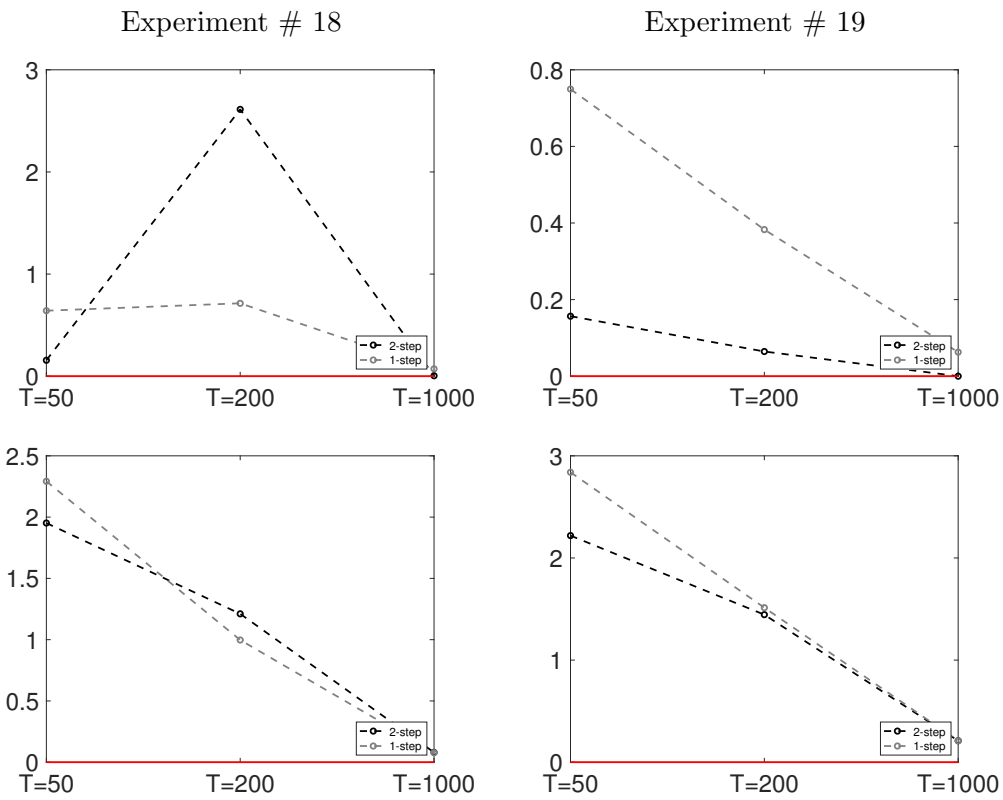
Notes: CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the GDP. Each point identifies the CRPS score obtained from different simulation setups, i.e. sample size (in X axis).

Figure 14: Accuracy Loss of one-step and two-step in presence of a break in the predictor coefficient. Nonnested case in top two graphs, nested case in the bottom two.



Notes: CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the GDP. Each point identifies the CRPS score obtained from different simulation setups, i.e. sample size (in X axis).

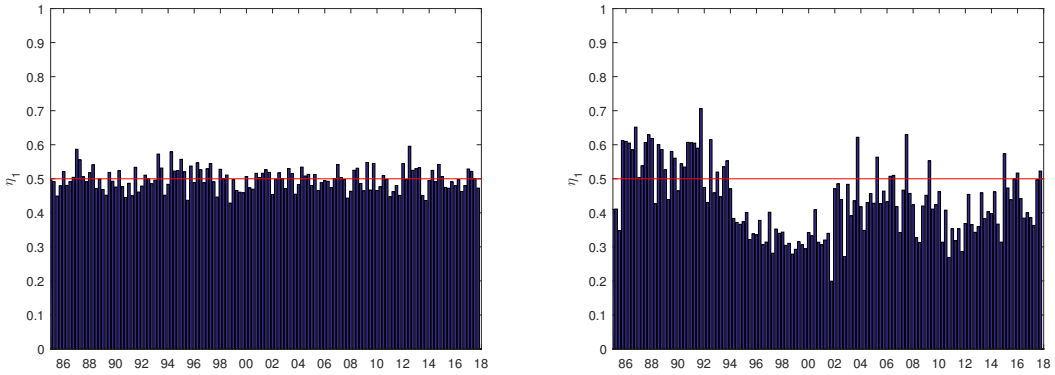
Figure 15: Accuracy Loss of one-step and two-step in presence of a break in error variance. Nonnested case in top two graphs, nested case in the bottom two.



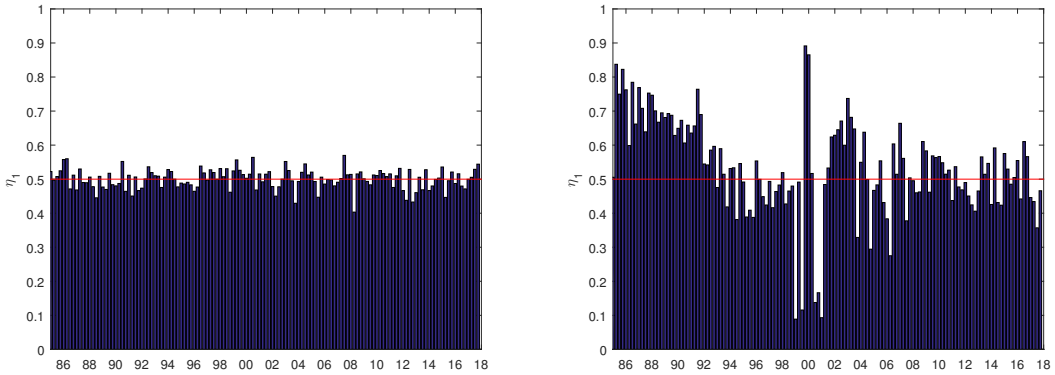
Notes: CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the GDP. Each point identifies the CRPS score obtained from different simulation setups, i.e. sample size (in X axis).

Figure 16: Combination Weights distribution over forecast sample 1985Q1-2018Q1

(a) Ex-Post Revised data



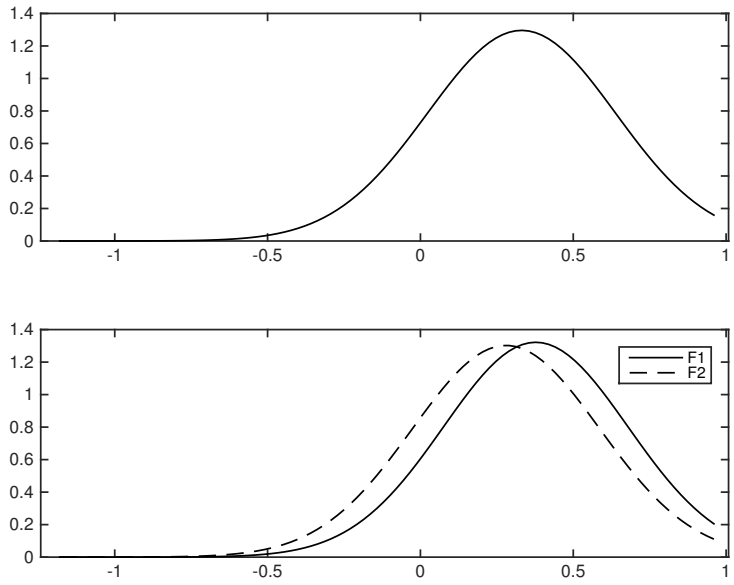
(b) Real-Time data



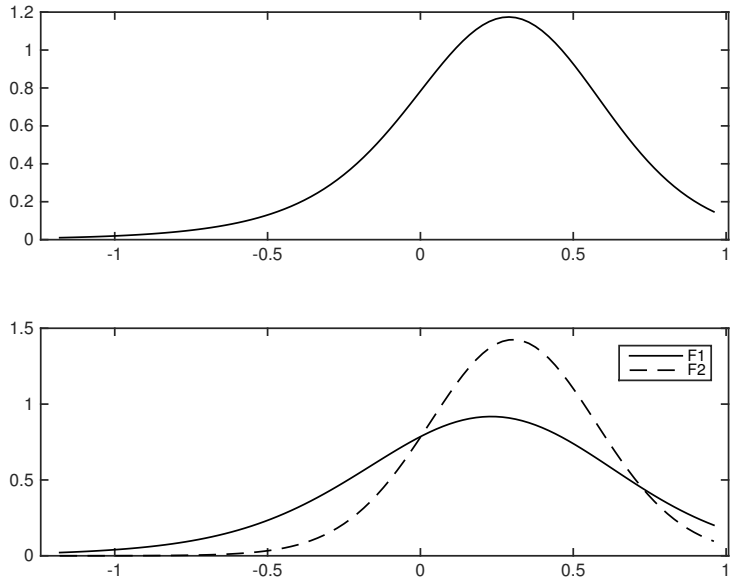
Notes: The black bars denote the weighted attached to ADL model with industrial production (η_1) as explanatory variable; estimates for any forecast origin. Since $\eta_2 = (1 - \eta_1)$, the values for the weight attached to ADL model with employment as explanatory variable are represented by the white area of the graph. Red line denoted the case of equally weighted combination, i.e. $\eta_1 = \eta_2 = 0.5$.

Figure 17: Combined and marginal density forecasts for last forecast T=2018Q.

(a) Two-step Procedure



(b) One-step Procedure



Notes: For each combination approach, the upper figure displays the density forecasts obtained combining the two components displayed in the lower figure. F_1 denotes the ADL model with industrial production, and F_2 the ADL model with empoyment rate.

Table 5: AR(1) benchmark vs. one-step and two-step alternatives, Out-of-Sample forecasting results, forecast horizon $h = 1$, 1985Q1-2018Q1; Ex-Post Revised and Real-time data.

	Ex-Post Revised data			Real-Time data		
	AR(1)	2-step	1-step	AR(1)	2-step	1-step
Log Score	-0.9818	-1.1765	-0.9134	-0.9298	-0.9974	-0.9126
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
			(0.000)			(0.000)
CRPS	0.2289	0.4879	0.1331	0.1008	0.2916	0.0756
		(0.000)	(0.000)		(0.000)	(0.000)
			(0.000)			(0.000)
TW-CRPS	0.2289	0.4880	0.1330	0.1008	0.2916	0.0760
		(0.000)	(0.000)		(0.000)	(0.000)
			(0.000)			(0.000)
MSPE	0.2209	0.2215	0.2220	0.3629	0.3573	0.3914
		(0.8236)	(0.3923)		(0.3964)	(0.0362)
			(0.8236)			(0.0521)

Notes: Table reports results for out-of-sample tests of equal predictability for models of US GDP growth at $h = 1$ step ahead. The models are estimated using rolling windows of data: the first in sample window is 1965Q1-1984Q4. The panel labelled "Ex-post Revised Data" reports the results using the latest vintage of data. The panel "Real-Time Data" reports results using vintages of real-time data, with OOS forecast errors computed using the first available real-time vintages of data. For the AR(1) benchmark and the alternative models, MSPE, average log scores, average CRPS and average tail-weighted CRPS values are reported. In parentheses are reported p-values for Diebold and Mariano (2002) t-test for equal forecast accuracy between the benchmark and the alternative model and between one-step and two-step approach.

Figure 18: Score Difference between the benchmark (AR(1) model) and the alternative models (a blue line identifies two-step procedure, a red line the one-step procedure) using Ex-Post revised data.

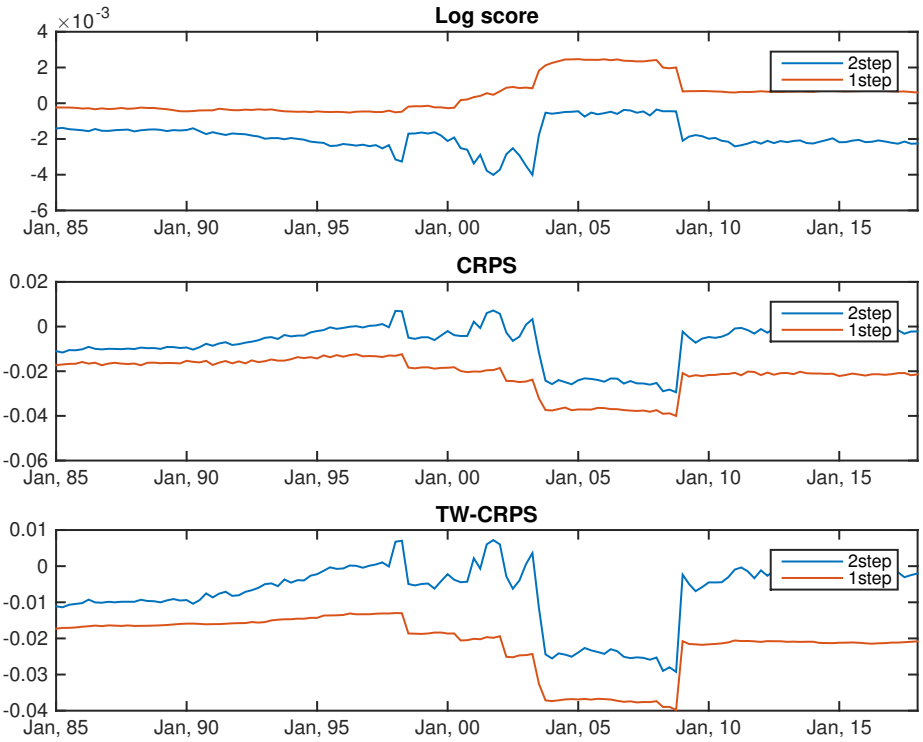


Figure 19: Score Difference between the benchmark (AR(1) model) and the alternative models (a blue line identifies two-step procedure, a red line the one-step procedure) using Real-Time data.

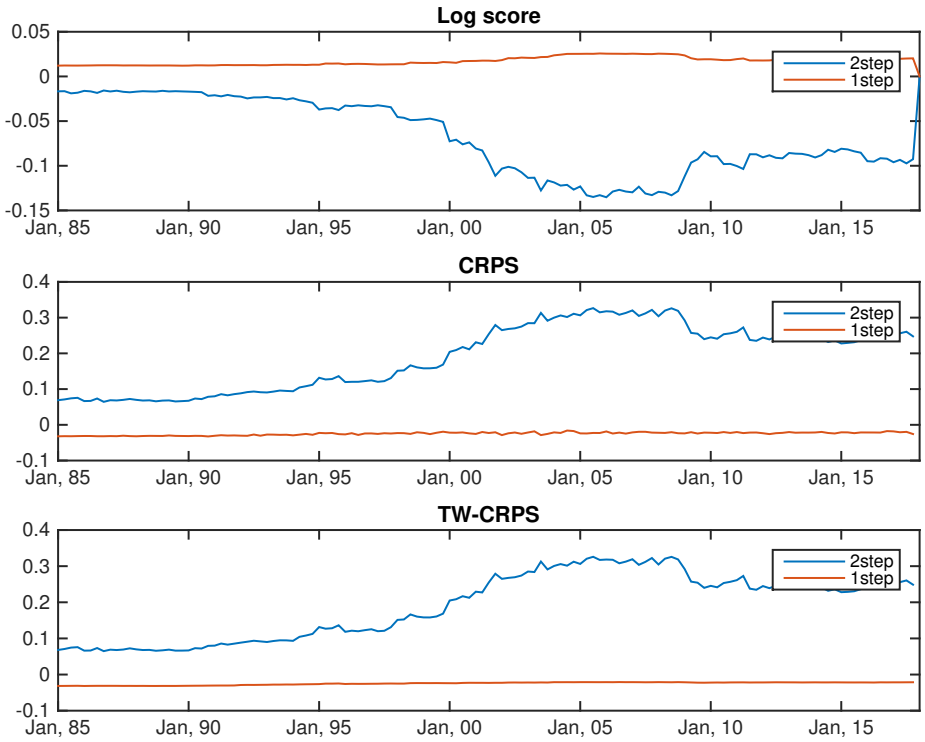
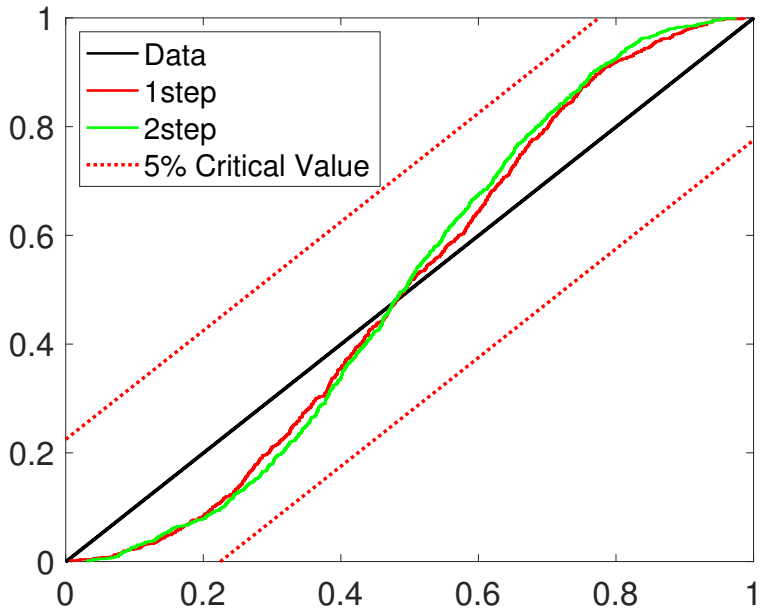


Figure 20: Probability Integral Transform (PIT) cdfs for data (black line), two-step procedure (green line) and one-step procedure (red line) using Ex-Post revised (top graph) and Real-Time (bottom graph) data. Dashed lines represent the 5% critical values bands.

(a) Ex-Post Revised data



(b) Real-Time data

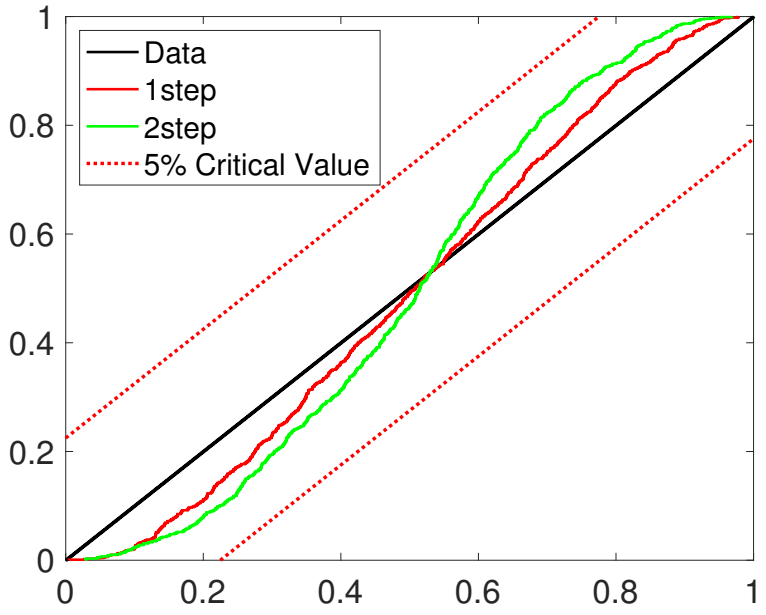


Table 6: Description of Data Series

Label	Trans	Period	Name	Description	Source
			Asset Prices		
rovnght@us	level	59:M1-18:M6	FEDFUNDS	Int. Rate: Fed Funds (Effective)	F
rtbill@us	level	59:M1-18:M6	TB3MS	Int. Rate: 3-Mn Tr. Bill, Sec Mkt Rate	F
rbnds@us	level	59:M1-18:M6	GS1	Int. Rate: US Tr. Const. Mat., 1-Yr	F
rbndm@us	level	59:M1-18:M6	GS5	Int. Rate: US Tr. Const. Mat., 5-Yr	F
rbndl@us	level	59:M1-18:M6	GS10	Int. Rate: US Tr. Const. Mat., 10-Yr	F
stockp@us	Δln	59:M1-18:M6	SP500	US Share Prices: S&P 500	F
exrate@us	Δln	73:M1-18:M6	111	NEER	I
			Real Activity		
rgdp@us	Δln	59:Q1-18:Q1	GDPC12	Real GDP, sa	F
ip@us	Δln	59:M1-18:M6	INDPRO	Industrial Production Index, sa	F
capu@us	level	59:M1-18:M6	CUMFNS	Capacity Utilization Rate: Man., sa	F
emp@	Δln	59:M1-18:M6	CE16OV	Civilian Employment: thsnds,sa	F
unemp@us	level	59:M1-18:M6	UNRATE	Civilian Unemployment,sa	F
			Wages and Prices		
pgdp@us	Δln	59:Q1-18:Q1	GDPDEF	GDP Deflator, sa	F
cpi@us	Δln	59:M1-18:M6	CPIAUCSL	CPI: Urban, All items, sa	F
ppi@us	Δln	59:M1-18:M6	PPIACO	Producer Price Index, nsa	F
earn@us	Δln	59:M1-18:M6	AHEMAN	Hourly Earnings: Man., nsa	F
			Money		
mon0@us	Δln	59:M1-18:M6	AMBSL	Monetary Base, sa	I
mon1@us	Δln	59:M1-18:M6	M1SL	Money: M1, sa	I
mon2@us	Δln	59:M1-18:M6	M2SL	Money: M2, sa	I
mon3@us	Δln	59:M1-06:M2	M3SL	Money: M3, sa	I

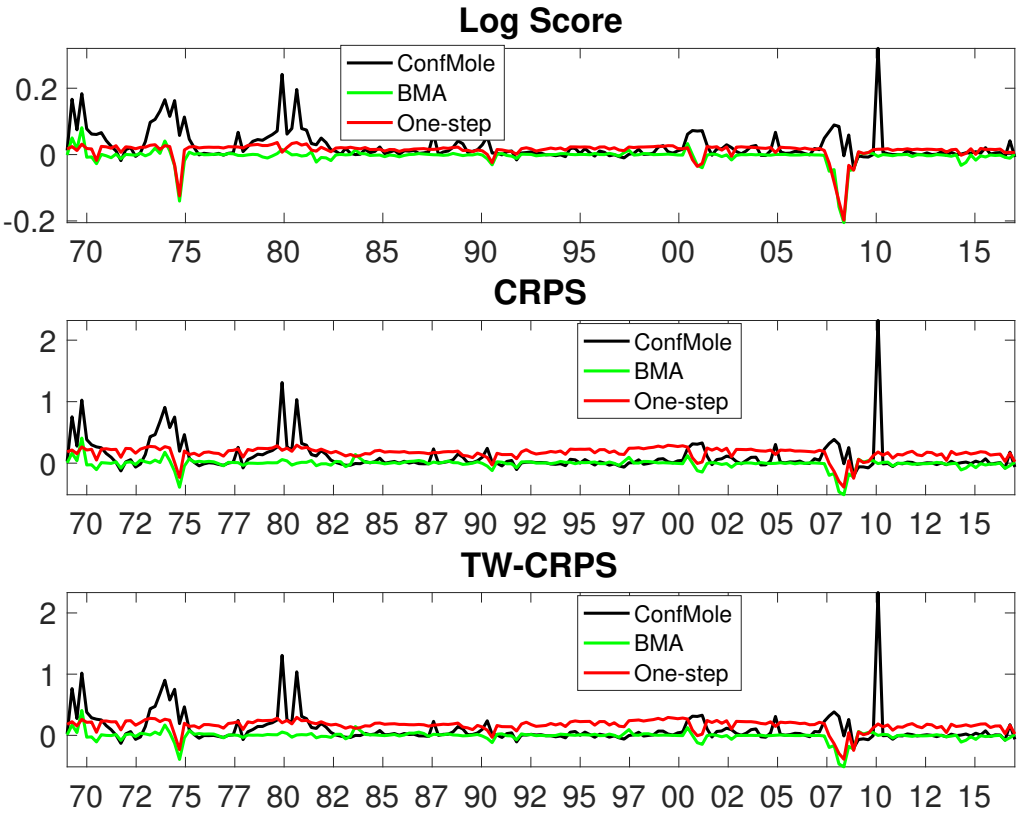
Notes: Sources abbreviated as "F" denotes Federal Reserve Economic Data (FRED) and "I" IMF International Financial Statistics.

Table 7: Average Scores of density forecasts for output growth rate obtained by AR(1) benchmark model vs. one-step and two-step alternatives, Out-of-Sample forecasting results, forecast horizon $h = 1$, 1985Q1-2018Q1

	AR(1)	Two-step CM	Two-step BMA	One-step
Log Score	-0.9378	-0.9643	-0.9333	-0.9506
	(0)	(0)	(0.0135)	(0.000)
	(0)	(0)	(0)	(0.0002)
	(0)	(0)	(0)	(0.0000)
CRPS	0.2223	0.2468	0.2178	0.2562
	(0)	(0)	(0.0731)	(0.0000)
	(0)	(0)	(0)	(0.7481)
	(0)	(0)	(0)	(0.0000)
TW-CRPS	0.2223	0.2468	0.2178	0.2562
	(0)	(0)	(0.073)	(0.0000)
	(0)	(0)	(0)	(0.7444)
	(0)	(0)	(0)	(0.0000)

Notes: Table reports results for out-of-sample tests of equal predictability for models of US GDP growth at $h = 1$ step ahead. The models are estimated using rolling windows of data: the first in sample window is 1969Q1-1978Q4. For the AR(1) benchmark and the alternative models, average log scores, average CRPS and average tail-weighted CRPS values are reported. In parentheses are reported p-values for Diebold and Mariano (2002) t-test for equal forecast accuracy between the benchmark and the alternative model and between one-step and two-step approach (Optimal weighting) and two-step approach (BMA).

Figure 21: Output Growth forecast. Accuracy loss using one-step or one of the two-step alternatives used here instead of the benchmark model (AR1). Out-of-Sample forecasting results, forecast horizon $h = 1$, 1979Q1-2018Q1



Notes: Difference in Log scores calculated against the benchmark i.e. $-1*LS_{combination} - (-1 * LS_{bench})$. CRPS rates calculated against the Benchmark i.e. $(CRPS_{combination} - CRPS_{bench})/CRPS_{bench}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the benchmark.

Figure 22: Output Growth Forecasts. Empirical CDF of the PITs for benchmark (AR1), two-step (CM), two-step (BMA) and one-step procedures, the CDF of the PITs under the null hypothesis (the 45 degree line) and the 5% critical values bands.

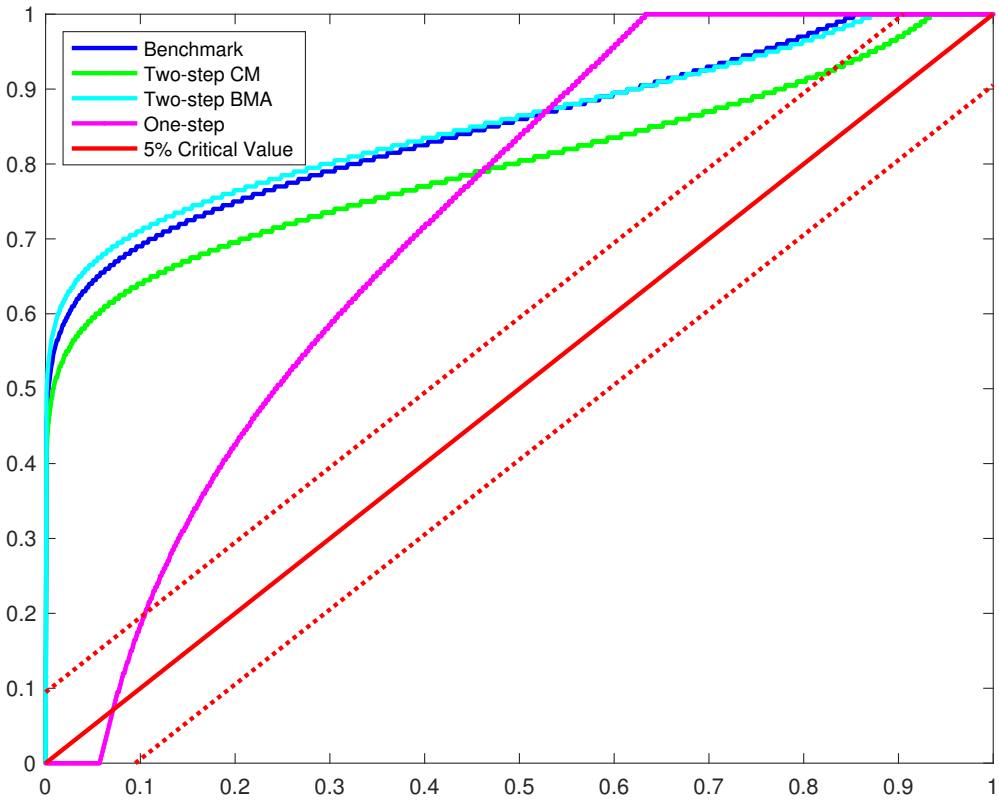
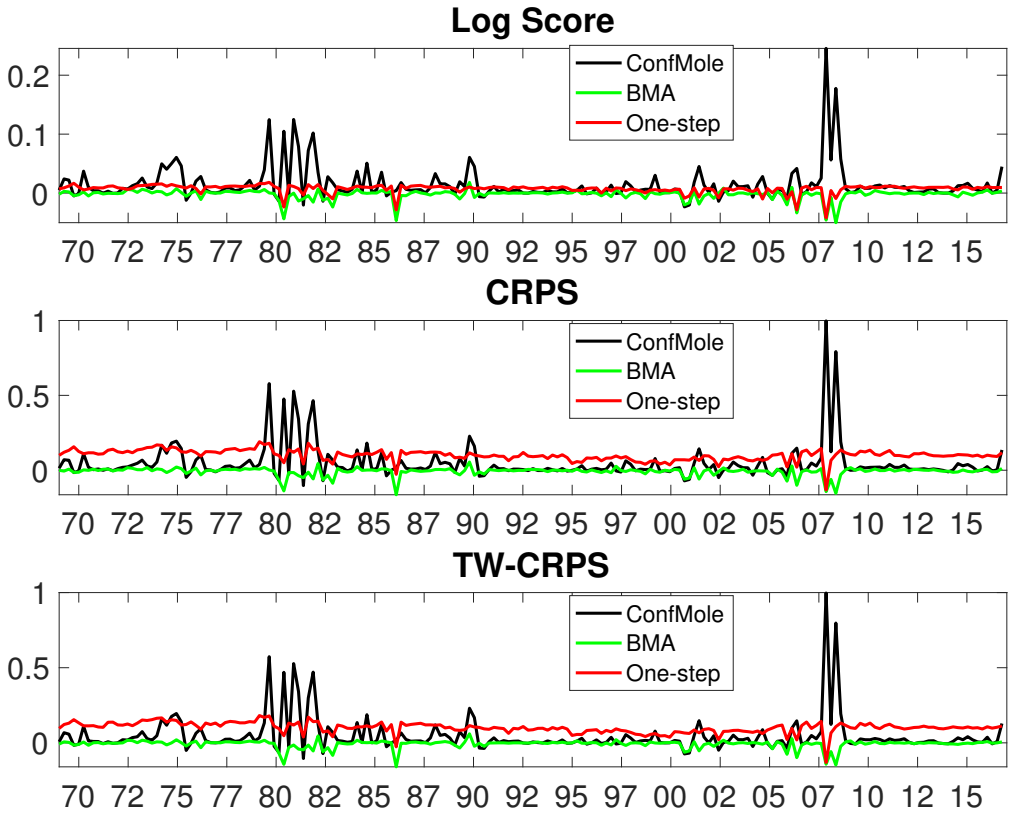


Table 8: Average Scores of density forecasts for inflation obtained by AR(1) benchmark model vs. one-step and two-step combinations, Out-of-Sample forecasting results, forecast horizon $h = 1$, 1985Q1-2018Q1

	AR(1)	Two-step CM	Two-step BMA	One-step
Log Score	-0.9269	-0.9418	-0.9248	-0.9342
	(0)	(0)	(0.0006)	(0)
	(0)	(0)	(0)	(0.002)
	(0)	(0)	(0)	(0)
CRPS	0.2356	0.2479	0.2333	0.259
	(0)	(0.0002)	(0.0002)	(0)
	(0)	(0)	(0.0001)	(0.0674)
	(0)	(0)	(0)	(0)
TW-CRPS	0.2355	0.2478	0.2334	0.259
	(0)	(0.0003)	(0.0003)	(0)
	(0)	(0)	(0.0001)	(0.0657)
	(0)	(0)	(0)	(0)

Notes: Table reports results for out-of-sample tests of equal predictability for models of inflation at $h = 1$ step ahead. The models are estimated using rolling windows of data: the first in sample window is 1969Q1-1978Q4. For the AR(1) benchmark and the alternative models, average log scores, average CRPS and average tail-weighted CRPS values are reported. In parentheses are reported p-values for Diebold and Mariano (2002) t-test for equal forecast accuracy between the benchmark and the alternative model and between one-step and two-step approach (CM) and two-step approach (BMA).

Figure 23: Inflation forecast. Accuracy loss using one-step or one of the two-step alternatives used here (CM and BMA) instead of the benchmark model (AR1). Out-of-Sample forecasting results, forecast horizon $h = 1$, 1979Q1-2018Q1



Notes: Difference in log scores calculated against the benchmark i.e. $-1 * LS_{combination} - (-1 * LS_{bench})$. CRPS rates calculated against the Benchmark i.e. $(CRPS_{combination} - CRPS_{bench})/CRPS_{bench}$. An higher CRPS rate indicates a higher loss in accuracy with respect to the benchmark.

Figure 24: Inflation forecast. Empirical CDF of the PITs for benchmark (AR1), two-step (optimal weighting), two-step (BMA) and one-step procedures, the CDF of the PITs under the null hypothesis (the 45 degree line) and the 5% critical values bands.