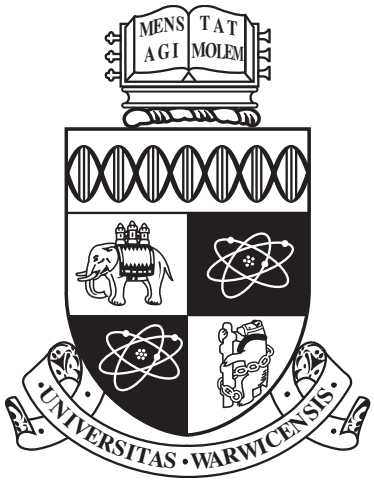


Combining, Evaluating and Constraining Predictive Distributions

by

Giulia Mantoan

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Warwick Business School

November 2021

THE UNIVERSITY OF
WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

I wish to express all my gratitude to the people who put time and thoughts into the development of this work. First and foremost, I am extremely grateful to my supervisors Ana Galvao and James Mitchell, for their invaluable advice, continuous support, and patience during my Ph.D. studies. Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. I would like to offer my special thanks to the members of Ph.D. review committees, Philippe Mueller and Anthony Garratt, for their thoughtful insights on preliminary versions of this thesis. I would like to express my sincere gratitude to Knut Are Aastveit and Saskia ter Ellen for offering me the opportunity to work with them in Norges Bank and for their help and advice.

My Ph.D. journey would not have been the same without the incredible people I have met in these years at Warwick. I would like to thank Anastasia, Danilo, Ida, Danae and Ila, whom I admire greatly and I am proud to call my friends. I am deeply grateful to Fosca, Caterina and Ana; despite being spread around the world, they were always present and shared with me all the ups and downs.

This thesis is not the product of a single person but the expression of a community. It wouldn't be possible without the support of the Macroeconomics Policy and Forecasting Research Network of the Finance Group in Warwick Business School. Lastly would like to thank the ESRC and WBS Finance Group scholarships for the financial support that allowed me to pursue my Ph.D.

Dedication

To my parents Agnese and Antonello,
who always allowed me to bring books to the table.

Declarations

I declare that, Chapter (2) and Chapter (4) are the result of my own work. Chapter (3) is co-authored with my supervisors Ana Galvao and James Mitchell. Their contribution to this chapter mainly regards the conceptualisation and the methodology, while I contributed on writing, data curation, coding, derivations and estimations.
Moreover, I declare that this thesis has not been submitted for any degree at the University of Warwick or any other institution.

Abstract

This thesis consists in three essays on predictive distributions, in particular their combination, calibration and constraint.

Chapter (2), entitled "Combination of Probabilistic Forecasts: a comparison between inference approaches", aims to compare two inference approaches for combinations found in the literature. One is arguably more common in macroeconomics and finance literature, the other is more used in statistics. Both inference approaches have pros and cons, but no analysis has been found in literature about which approach is the most accurate. The paper's results find clear evidence favouring one approach or the other based on the problem and data at hand.

Chapter (3) is entitled "Are Central Banks' Fan charts Reliable? On Calibration of Density Path Forecasts". Central Banks regularly publish fan charts of macroeconomic variables, communicating forecasts for several horizons. Although fan charts contains three types of information: point forecasts (the path), the probability around each point forecast (bands around it), and variable's dynamics across horizons (i.e. the path), existing absolute evaluation approaches neglect the latter. Practitioners evaluate the calibration of fan charts testing the forecast accuracy horizon by horizon, not considering any joint calibration of the path. This paper describes fan charts as density path forecasts, discusses the impact of horizon-dependence in the evaluation and proposes calibration tests to assess whether Central Banks publish reliable forecasts. We proposed several calibration tests, analysing their size and power, demonstrating that, according to our test, the Bank published on average non-calibrated path density forecasts.

Chapter (4), entitled "Generalised Constraints for Predictive Distributions: a

Bayesian Approach" investigates the concept of constraining density forecasts. Often policymakers wish to impose a feature to predictive distributions (such as moments constraint, tails behaviour, shifts in support). Although moment constraining is well discussed in the literature (i.e. by exponential tilting), little study has been done on constraining specific parts of the density's support. This forecast constraining shifts individual predictive densities using Bayesian Importance Sampling. This approach is applied to forecast US GDP under the Covid-19 pandemic: density forecasts from statical models are constrained to the survey of professional forecasters (SPF) in the left tail.

Chapter 1

Introduction

In this thesis, I investigate combining, calibrating, and constraining predictive distributions for macroeconomic time-series, relevant topics for both academic and policy-making purposes. The use of probabilistic forecasts over the past three decades has surged over point forecasts Tay and Wallis [2000], Gneiting and Katzfuss [2014], Clark [2011]. The motivation comes from the probabilistic nature of the future state of the world as discussed by (Dawid [1984]), and with it, the need for inclusion of some degree of uncertainty. Model uncertainty is crucial in forecasting since if the analyst uses an inappropriate model, then forecasts will be less accurate. A popular approach to address the model uncertainty is the combination. For example, suppose a set of forecasts (from different models) is available. Then it has been established empirically that a weighted linear combination of these forecasts will often be more accurate on average than any of the individual forecasts (Clemen and Winkler [1986], Wallis [2005], Chatfield [1996] Hendry and Clements [2004]).

Chapter (2) aims to contribute to probabilistic forecast combination proposing a comparison between two inference approaches. The first, called "two-step" is arguably the most popular in applied econometrics and finance, and it takes individual probability forecasts as given and then combines them. The second, called "one-step" is studied extensively by the statistical literature, and it estimates forecasts' parameters and combination weights simultaneously. First, I propose an empirical exercise to analyse the forecast accuracy of the two approaches. The application consists of forecasting US real output growth and inflation, combining a set of 31 individual models. The empirical exercise leaves us with no clear indication over which combination approach is the most accurate. Then, I tried to shed light on the different performances in controlled environments. Several types of DGPs and misspecifications for forecasting models have been

studied. The main takeaway is that the trade-off between parameter estimation noise and forecast accuracy typical of the one-step approach is crucial.

Concurrently with the increasing of interest in probabilistic forecasts, it becomes more and more important to assess their forecast accuracy such as in Diebold et al. [1997], Clements [2004], Corradi and Swanson [2006b], Boero et al. [2011], Wolters [2015] and correct specification of uncertainty around forecasting models. **Chapter (3)** focuses on forecasting evaluation and, in particular, propose an absolute evaluation criterion for path density forecasts. One example of the employment of path density forecast is the Central Bank's fan charts. This graph joins the realisations with the most likely "path" for GDP growth, inflation and unemployment for 1 to H periods in the future. In this way, fan charts are informative about the point forecast at a specific horizon $h = 1, \ldots, H$ and about the path the economy will follow up to period H. In addition, fan charts present confidence bands around the prediction that gives the reader an understanding of its likelihood. As predictions become increasingly uncertain the further into the future one goes, the forecast ranges out, creating distinctive "fan" shapes, hence the name. First, the chapter defines the path density forecast and discusses a series of testing strategies. The existing absolute evaluation approaches assess the forecast accuracy separately across horizons. However, in doing so, they implicitly assume independence across horizons of path forecast. We discuss the importance of horizon dependence in path density forecast and identify two main strategies for calibration tests. Whether to adopt one or another depends on the information the research has about the horizon dependence. If the researcher has information about it, they can use tests on marginal and conditional distributions of PITs; If the researcher does not have any information, they can use test statistics on marginal distributions and some "sup tests". We contrast and compare the statistics in simulation exercises to investigate their size and power characteristics and apply them to evaluate the calibration of Bank of England fan charts.

In recent years, policymakers have shifted their focus from forecasts uncertainty, in general, to be particularly interested in quantifying macroeconomic downside tail risk, often referred to as GDP-at-risk. More recently, it has become crucial to correctly predict adverse events since the impact of the COVID-19 pandemic on growth challenged forecasting with usual predictors (especially the second quarter of 2020). Motivated by this issue, in **Chapter (4)**, I focus on constraining predictive distributions for US growth to Survey of Professional Forecasters (SPF) probability of negative growth. Besides the COVID-19 case, policymakers and practitioners often wish to impose a desirable feature on predictive distributions (such as moments constraint, tails behaviour, shifts in support,...). Although constraining moments' distributions is well discussed in the literature

(i.e. by exponential tilting), it is often unclear which moment to use as a constraint. This paper aims to generalise the constraints to any desirable feature of the distribution. The constraints are imposed by approximation of the target (constrained) distribution using a mixture of Student-t distributions. The resulting density distributions can give a probability different from zero to the actual realisation.

In addition to my thesis, I have worked on an external project with two researchers at Norges Bank: Knut Are Aastveit and Saskia ter Ellen. This paper develops a forecast combination scheme that assigns weights to the individual predictive density forecasts based on quantile scores. Compared to standard combination schemes, our approach has the advantage of assigning a different set of combination weights to the various quantiles of the predictive distribution. The results show that density forecasts from our approach outperform forecasts from commonly used combination approaches.

Chapter 2

Combination of Probabilistic Forecasts: a comparison between inference approaches

2.1 Introduction

Forecast a future state of the world is by nature probabilistic and it includes some degree of uncertainty ([Dawid, 1984]). Although prominent forecasting models reduced it, uncertainty cannot be totally eliminated. For this reason, over the past three decades the use of probabilistic forecasts over point forecasts has surged (Tay and Wallis [2000], Gneiting and Katzfuss [2014], Clark [2011]). Concurrently with the increasing of interest in probabilistic forecasts, it becomes more and more important to assess their forecast accuracy such as in Diebold et al. [1997], Clements [2004], Corradi and Swanson [2006b], Boero et al. [2011], Wolters [2015] and to correct specification of uncertainty around forecasting models.

Model uncertainty is crucial in forecasting since if the analyst uses an inappropriate model, then forecasts will be less accurate. A popular approach to address the model uncertainty is the combination. For example, suppose you have produced forecasts by several methods. Then it has been established empirically that a weighted linear combination of these forecasts will often be more accurate on average than any of the individual forecasts (Clemen and Winkler [1986], Wallis [2005], Chatfield [1996] Hendry and Clements [2004]).

Arguably, "two-step" combination procedure is the most popular in applied macroeconomics and finance (Genest et al. [1984], Diebold and Lopez [1996], Clemen

[1989], Hall and Mitchell [2007], Mitchell and Wallis [2011], Kascha and Ravazzolo [2010], Kapetanios et al. [2015], and Geweke and Amisano [2011] among others). In this paper it is called "two-step" procedure the combination approach that takes individual probability forecasts as given and then combines them.

However, in literature (especially in statistics), another combination approach is also studied, here called the "one-step" procedure. This paper calls the "one-step" procedure the combination approach that estimates individual forecasts' parameters and combination weight simultaneously. The most popular one-step approach uses the finite mixture distribution (Everitt [2014]) as combined predictive density of individual models (Elliott and Timmermann [2005], Waggoner and Zha [2012], Ravazzolo and Vahey [2014], Gupta and Dhingra [2012], Raftery et al. [2005]). Given the inference complexity, the one-step approach has been less used by practitioners and in applied works. However, the algorithm employed here by Diebolt and Robert [1994] and Frühwirth-Schnatter [2006], overcomes the main drawbacks of finite mixture distributions as a combination approach. The main takeaway is that the trade-off between parameter estimation noise and forecast accuracy typical of the one-step approach is crucial. However, the trade-off is overcome by the one-step ability to account for the dependence between the mixture's components.

This paper proposes a comparison between these two approaches to the combination of density forecasts. The comparison is sustained by the empirical evidence that neither of the two approaches is more accurate than the other in a systematic way; in the absence of theoretical background, this paper aims to understand the discriminants that affect the different performances in simulation exercises.

The rest of the paper is organised as follows: Section (2.2) shows empirical evidence of the impact the combination choice has on forecast accuracy; Section (2.3) studies the different performances through simulation exercises; Section (2.5) contains concluding remarks.

2.2 Empirical Evidence

This section provides empirical evidence of the impact the combination choice has on forecast accuracy. The exercise set up, dataset and forecasting models follow Rossi and Sekhposyan [2014]; however, for the purpose of this paper, a broader set of combinations is employed.

Following Rossi and Sekhposyan [2014], the probability forecasts for US real output growth and inflation are achieved by some of the most commonly used macroe-

conomic predictors. The dataset consists of $K = 31$ variables selected from Stock and Watson [2003] dataset, among which asset prices, real economic activity, wage, price and money variables. Data are collected at a quarterly frequency, updated up to 2018Q1 and adequately transformed. Table (2.3) presents a detailed description of the variables and their transformations. 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 individual forecasting models (to be combined) are Autoregressive Distributed Lag (ADL) models, where the K predictors are used one-at-a-time. The forecasting model is:

$$y_{t+1,k} = \beta_0 + \beta_1(L)X_{t,k} + \beta_2(L)y_t + \varepsilon_{t+1} \tag{2.1}$$

for $t = 1, \ldots, T$. The variable of interest is either $y_t = 400 \ln (GDP_t/GDP_{t-1})$ or $y_t = 400 \ln (P_t/P_{t-1}) - 400 \ln (P_{t-1}/P_{t-2})$ where GDP and P are the real output growth and GDP deflator, respectively. $X_{t,k}$ denotes the k-th variable for $k = 1, \ldots, K$ as in Rossi and Sekhposyan [2014]. The dataset used to predict 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 in equation (2.1) are linearly combined according to one and two-step approaches.

2.2.1 Two-step combination procedures

Two-step combination procedures are so called since they consist in two phases: first, the K individual forecasting models are estimated in Equation (2.1); subsequently, the K probabilistic forecasts obtained in step one are combined using the combination weights adequately estimated. The linear combination scheme assumes the from:

$$y_{t+1,c} = \sum_{k=1}^{K} \eta_k(y_{t+1,k}) \tag{2.2}$$

where $y_{t+1,c}$ denotes the combined probability forecast, $y_{t+1,k}$ are the density forecasts obtained in Equation (2.1) and η_k are the combination weights. In this paper the first step consists in an OLS estimation, while three different approaches to estimation of the

weights are compared:

- Equal weights: this widely used combination method consists in attaching the same combination weight to every forecast, regardless their accuracy.

$$\eta_k = 1/K \tag{2.3}$$

  This simple two-step approach has been proved empirically to outperform more sophisticated combination methods. This phenomenon is known as the "forecast combination puzzle" and documented by Genre et al. [2013] for expert forecasts dataset.

- Bayesian Model Averaging (BMA) pooling model: the combination weights are proportional to models' posterior probability:

$$\eta_k := P_t(y_{t+1,k}|y_t) = \frac{p(y_t|y_{t+1,k})p(y_{t+1,k})}{\sum_{k=1}^{K} p(y_t|y_{t+1,k})p(y_{t+1,k})} \quad s.t. \quad \eta_k > 0, \quad \sum_{k=1}^{K} \eta_k = 1 \tag{2.4}$$

  where $y_{t+1,k}$ is the probability forecast obtained in equation (2.1) and y_t is the data at time t. Bayesian inference follows Rossi and Sekhposyan [2014] and Wright [2009].

- Optimal weight (CGM): proposed by Hall and Mitchell [2007], held from the idea of determining combination weights based on some objective criterion or cost function, such as the logarithmic score. Combination weights are obtained by maximizing a logarithmic score function:

$$\eta_k = \frac{1}{T-1} \sum_{t=1}^{T-1} ln(y_{t+1,k}) \quad s.t. \quad \eta_k > 0, \quad \sum_{k=1}^{K} \eta_k = 1 \tag{2.5}$$

  which is known as the log predictive score. Given the size of K, the inference algorithm for η_k in Conflitti et al. [2015] is used.

The last two methods are conceptually different: the BMA is born as a model selection method, with the purpose to elicit, among a bundle of alternatives, the model that better fits the data. For this reason, the BMA assumes the "true" model is in the set of models considered, the assumption that the optimal pooling does not make. The BMA is then an "improper" combination approach since this assumption contradicts the principle of forecast combination: none of the alternatives at hand is the correct model.

2.2.2 One-step combination procedure

One-step procedure is so called since it estimates forecast models and combination weights simultaneously. This procedure addresses the combination issue in Equation (2.2) as a finite mixture of univariate Gaussian components. Its predictive density takes the form:

$$p(y_{t+1}|\boldsymbol{y}_t, \boldsymbol{\theta}_k) = \sum_{k=1}^{K} p(y_{t+1,k}|\boldsymbol{y}_t, \boldsymbol{\theta}_k)\eta_k. \tag{2.6}$$

Each finite mixture is defined by three parameters: the number of components K, the components' parameters vector $\boldsymbol{\theta}_k$, and the mixing proportions $\eta_k(y_t, \boldsymbol{\theta}_k)$. In this case: the number of components K is known and it corresponds to the number of individual forecasts to combine. The components' parameters $\boldsymbol{\theta}_k$ is the vector if size $D = 3K$ defined by the type of forecast models. More explicitly, consider a mixture model of K normal component densities: $p(y_{t+1,k}|y_t, \boldsymbol{\theta}_k) = f_N(y_{t+1,k}|\mu_{k,t+1}, \sigma^2_{k,t+1})$ with $\mu_{k,t+1} = \mathbb{E}(y_{t+1}|\boldsymbol{y}_t, \boldsymbol{\theta}_k)$ and $\sigma^2_{k,t+1} = Var(y_{t+1}|\boldsymbol{y}_t, \boldsymbol{\theta}_k)$ being the conditional mean and variance. In this paper each component's mean follows a ADL process of order p_k: $\mu_{k,t+1} = \boldsymbol{\beta}'_k \boldsymbol{z}_k$ where $\boldsymbol{\beta}_k = [\beta_0, \beta_1, \beta_2]$, $\boldsymbol{z}_k = [1\ X_{t:t-p_k+1}\ y_{t:t-p_k+1}]$ and $y_{t:t-p_k+1} = \{y_t, y_{t-1}, \ldots, y_{t-p_k+1}\}$. Finally, the mixing proportions $\eta_k(y_t, \boldsymbol{\theta}_k)$ correspond to combination weights and they follows a multinomial distribution:

$$\eta_k \sim M\left(1, \left[\frac{p_1 f_N(y_t; \mu_{1,t}, \sigma^2_1)}{\sum_{k=1}^{K} p_k f_N(y_t; \mu_{k,t}, \sigma^2_k)}, \ldots, \frac{p_K f_N(y_t; \mu_{K,t}, \sigma^2_K)}{\sum_{k=1}^{K} p_k f_N(y_t; \mu_{k,t}, \sigma^2_k)}\right]\right) \tag{2.7}$$

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

Inference

Given the high parametrisation, the estimation approach employed here is the Bayesian inference technique of MCMC estimation using two-block Gibbs sampling. To allow for a fair comparison between combinations, the weakness of the finite mixture model in combining a significant number of components is addressed by imposing dependent priors for parameters following Frühwirth-Schnatter [2006].

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, $\boldsymbol{\theta}_k$ and σ_k) is:

$$p(\eta_k|\boldsymbol{S}) \sim \mathcal{D}(e_1(\boldsymbol{S}), \ldots, e_K(\boldsymbol{S})) \tag{2.8}$$

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 $\boldsymbol{\theta}_k$ and σ_k^2 given the weights and all observations assigned to group k are normally distributed:

$$p(\boldsymbol{\theta}_k|\boldsymbol{\sigma}_k^2, \boldsymbol{y}_t, \boldsymbol{S}) \sim \mathcal{N}(a_k, A_k) \tag{2.9}$$

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

$$p(\sigma_k^2|\boldsymbol{\theta}_k, \boldsymbol{y}_t, \boldsymbol{S}^{m-1}) \sim \mathcal{G}^{-1}(c_N, C_N) \tag{2.10}$$

where:

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

and where $\varepsilon_k = \boldsymbol{y}_t - Z_k\boldsymbol{\theta}_k$. A prior dependence among the component parameters is introduced. Following Richardson and Green [1997], the parameter C_0 is treated as an unknown hyper-parameter with a prior of its own.

$$p(C_0|\boldsymbol{S}^{m-1}, \theta_k, \sigma_k^2, \boldsymbol{y}_t) \propto \prod_{k=1}^{K} p(\sigma_k^2|C_0)p(C_0) \propto \prod_{k=1}^{K}\left(C_0^{c_0}\exp\{-\frac{C_0}{\sigma_k^2}\}\right)C_0^{g_0-1}\exp\{-G_0C_0\} \tag{2.11}$$

which is the Kernel of a $\mathcal{G}(g_N, G_N)$-density with $g_N = G_0 + KC_0$ and $G_N = G_0 + \sum_{k=1}^{K}\frac{1}{\sigma_{\varepsilon,k}^2}$. The joint prior takes the form of a hierarchical independent prior:

$$p(\theta_k, \sigma_k^2, C_0) = \prod_{k=1}^{K} p(\theta_k)\prod_{k=1}^{K} p(\sigma_k^2|C_0)p(C_0) \tag{2.12}$$

where θ_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:

$$p(\boldsymbol{\theta}_k) = \mathcal{D}(e_0, e_0|\boldsymbol{S})\prod_{k=1}^{2}\mathcal{N}(\boldsymbol{\phi}_k|a_0, A_0)\mathcal{IG}(\sigma_k^2|c_N, C_N). \tag{2.13}$$

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(1)). The algorithm corresponds to algorithm 8.1 in Frühwirth-Schnatter [2006] for the case of univariate normal mixture regression model (references to algorithm 6.1).

2.2.3 Results

Combination approaches are evaluated with respect to the benchmark AR(1) according to several criteria: the logarithmic score, the continuously ranked probability 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). The logarithmic scoring rule gives a higher score to a forecast that provides a high probability to the realisation. The forecaster aims to maximise the log score and, for elicitation purposes, to select the forecasting model that obtains a higher log score. The CRPS and TW-CRPS are positively valued such that a forecast with a lower score indicates that it performs better than the alternative. According to the PITs, the density forecast is called "calibrated" if its PIT values are iid uniform. One way to check the uniformity is to plot the empirical CDF of the PIT values against the 45 degrees line (CDF of uniform distribution). Please refer to Rosenblatt [1952] and Diebold et al. [1997] for discussion.

Combined density forecasts for output growth and inflation are evaluated in Table (2.1) and Figure (1). According to the scores, BMA and CGM two-step procedures are more accurate than the one-step procedure in combining probability forecasts for output growth. Average log scores and CRPS are in favour of the two-step BMA combination model. From PITs cumulative distribution functions in Figure (1), it is clear that none of the models is well-calibrated; however, the one-step procedure is much closer to the 45 degrees line than the other alternatives displaying the fact that even though the one-step approach is the less accurate according to relative scores, it is the more consistent to the data. For inflation, the result is less clear: log scores favour one-step procedure while CRPS favours the benchmark AR(1) model. From the inspection of the PITs, it is evident that the one-step procedure is well-calibrated, at least in the central part of the distribution.

In summary, the empirical exercise performed in this section leaves us with no clear indication over which combination approach is the most accurate. Whether the conflict between evaluation criteria has been treated in the literature, it is still unclear

why different combination approaches perform better (worse) for output growth and fail (succeed) for inflation. The results also depend on other factors such as the number of forecasts K to combine and the time window. The following section will try to shed light on what affects the different performances in controlled environments.

Table 2.1: AR(1) benchmark vs. One-step and Two-step alternatives, in-sample forecasting results, forecast horizon $h = 1$, 1985Q1-2010Q1

	Output growth					Inflation rate				
	AR(1)	Two-step CGM	Two-step BMA	Two-step EQ	One-step	AR(1)	Two-step CGM	Two-step BMA	Two-step EQ	One-step
Log Score	**-9.1549**	**-8.9942**	**-8.8205**	**-9.1231**	**-9.0074**	**-1.8479**	**-1.9259**	**-1.8984**	**-1.8941**	**-1.8075**
		(0.0417)	(0.0001)	(0.1354)	(0)		(0.0009)	(0.0002)	(0.0001)	(0.565)
			(0.002)	(0.0342)	(0.416)			(0.03)	(0.0314)	(0.0001)
				(0)	(0.0017)				(0.0961)	(0)
					(0.3435)					(0)
CRPS	**3.0744**	**3.0429**	**3.0182**	**3.0758**	**3.1733**	**0.8293**	**0.8757**	**0.8573**	**0.8556**	**0.9034**
		(0.0033)	(0)	(0.1649)	(0)		(0)	(0)	(0)	(0)
			(0.0006)	(0.0044)	(0)			(0.001)	(0.0008)	(0.0001)
				(0)	(0)				(0.1027)	(0)
					(0)					(0)
TW-CRPS	**3.0745**	**3.0429**	**3.0182**	**3.0759**	**3.1733**	**0.8293**	**0.8758**	**0.8576**	**0.8556**	**0.9034**
		(0.0033)	(0)	(0.1657)	(0)		(0)	(0)	(0)	(0)
			(0.0006)	(0.0044)	(0)			(0.001)	(0.0008)	(0.0001)
				(0)	(0)				(0.0859)	(0)
					(0)					(0)

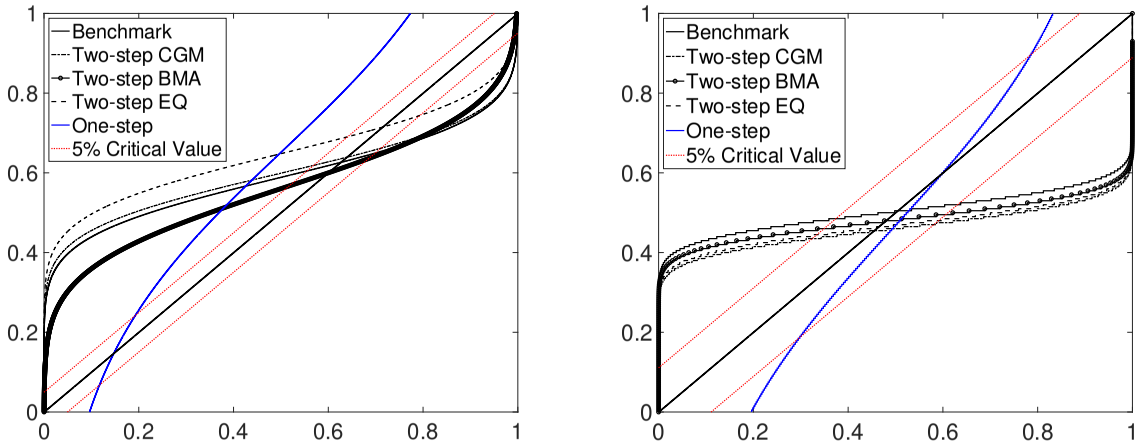


Figure 1: Empirical CDF of Probability Integral Transform (PITs) of combined densities forecasts for output growth (on the left) and inflation (on the right). In red the CDF of the PITs under the null hypothesis of calibration (45 degrees line) and the 5% critical value bands.

2.3 Simulation Exercises

This section moves from the aim of understanding the mechanisms behind the results seen in the previous section. Neither of the combination approaches used in this paper seems to be superior to the others in a consistent matter. From the empirical evidence, it is not clear what causes the higher accuracy of one approach over the others. In the absence of a theoretical discussion in the literature, this paper explains it through simulation exercises.

In particular, a series of simulation exercises investigates how different sources of misspecifications affect the accuracy of combination procedures. Each experiment is characterised by a Data Generating Process (DGP), the definitions of forecasting models and their pooling schemes. Combination approaches are the same as before: the mixtures of normals distributions as one step approach and equal weighted, CGM and BMA as two-step procedures.
Let us consider y_t as a vector of values generated by a function g:

$$y_t = g(V_1, V_2, V_3) + u_{t+1} \tag{2.14}$$

where $u \sim \mathcal{N}(0, \sigma^2)$ and V_1, V_2, V_3 are exogenous variables. In order to forecast one-step ahead values for y_t, the researcher has two forecasting models f_1 and f_2:

$$\begin{aligned} y_{t+1,f_1} &= f_1(\theta_{1,0} + \theta_{1,1}V_1 + \theta_{1,2}V_2 + \theta_{1,3}V_3) + u_{t+1} \\ y_{t+1,f_2} &= f_2(\theta_{2,0} + \theta_{2,1}V_1 + \theta_{2,2}V_2 + \theta_{2,3}V_3) + u_{t+1} \end{aligned} \tag{2.15}$$

which are combined according the one and two step procedures seen in the previous section, i.e.:

$$y_{t+1,c} = \eta_1 y_{t+1,f_1} + \eta_2 y_{t+1,f_2}. \tag{2.16}$$

Each simulation setup differs by different choices for parameters both for individual forecasting models and for explanatory variables $\{V_1, V_2, V_3\}$ (both for DGPs and forecasts).

Three families of DGPs are investigated: unimodal, bimodal and nonlinear, aiming to exploit which approach to combination helps to detect the unimodality, multimodality, and nonlinearity of data. For each family of DGP, several misspecifications are introduced to the forecasting models f_1 and f_2 to identify which combination approach is more likely to overcome the individual model's misspecification through combination.

Results are displayed in Table (2.4) and Figures (2.1-2.9) in the form of the

accuracy loss a researcher will incur choosing a combination of forecasts instead of the DGP. The accuracy loss is then a rate of scores such as:

$$\text{Accuracy Loss} = \frac{(SCORE_{combination} - SCORE_{DGP})}{SCORE_{DGP}} \tag{2.17}$$

For elicitation purposes, the researcher should select the combination approach with smaller accuracy loss. In order to investigate if and how the two approaches work in different environments, three samples of different sizes ($T = 50$, $T = 200$, $T = 1000$) are drawn from the following DGPs.

2.3.1 Unimodal Data Generating Processes

The issue of unimodality and multimodality of the data is well studied in Everitt [1985]: albeit a combination of several unimodal distributions is a good proposal to fit a multimodal DGP, it may well fit a unimodal DGP under certain conditions that regard their moments. Therefore, this section aims to study how well different combination procedures fit unimodal DGPs. The first DGP is obtain imposing $V_1 = y_{t-1}$, and $V_2 = V_3 = 0$ to equation (2.14) such that:

DGP A: $y_t = \beta_0 + \beta_1 y_{t-1} + u_t \qquad u \sim \mathcal{N}(0, \sigma^2)$

It consists then in a AR(1) process and parameter values are assumed to be $\beta_0 = 0.5, \beta_1 = 0.8, \sigma^2 = 0.6$. Two types of misspecification in the forecasting models are considered:

A1 : f_1 is the correct model, f_2 is noise.

$$\begin{aligned} f_1 &: y_{t+1} \sim \mathcal{N}(\theta_{0,1} + \theta_{1,1} y_t, \sigma_1^2) \\ f_2 &: y_{t+1} \sim \mathcal{N}(\theta_{0,2} + \theta_{1,2} x_t, \sigma_2^2) \end{aligned} \tag{2.18}$$

The first forecast model f_1 follows a AR(1) process as the DGP, so it is correctly specified. The second forecast model f_2 is a function of an exogenous variable $x_t \sim \mathcal{N}(-1, 0.5)$ that is assumed to be independent from y_t.

This exercise is employed to study the effectiveness of the combination approach. It is desirable that a combination approach rules out the noisy model (f_2) collapsing the weight η_2 or parameters $\theta_{0,2}, \theta_{1,2}$ and σ_2 to zero. Both combination approaches should detect the correct model; however, the difference in inference approach may favour the two-step approach at least in small samples.

From Panel (A1) of Table (2.4) we can see that one-step approach delivers the most accurate combined density forecast when the sample size is sufficiently large. For small sample size (i.e. $T = 50$), the two-step CGM approach bets one-step according to log-score and the tail-weighted version of CRPS. An explanation for this evidence can be found in the fact that estimating parameters and weights at the same time (using one-step apporach) may increase the parameter estimation error and raise the forecast error variance.

A2 : f_1 is the correct model, f_2 nests f_1.

$$\begin{aligned} f_1 &: y_{t+1} \sim \mathcal{N}(\theta_0 + \theta_1 y_t, \sigma^2_{\varepsilon_1}) \\ f_2 &: y_{t+1} \sim \mathcal{N}(\theta_0 + \theta_1 y_t + \theta_2 x_t, \sigma^2_{\varepsilon_1}) \end{aligned} \tag{2.19}$$

As in the previous set up, f_1 follows a AR(1) process as the DGP, so it is correctly specified. However now f_2 is a function of the independent exogenous variable $x_t \sim \mathcal{N}(-1, 0.5)$ and y_t. f_2 then nests model for f_1, and it is correctly specified when $\theta_2 = 0$. This exercise is employed to study the effectiveness of combining forecasts from nested models. Clark and McCracken [2009] provides a theoretical analysis of combining forecasts from nested models, finding that, under model uncertainty, it will improve the forecast accuracy. The nature of this exercise wants to investigate whether both approaches improve the forecast accuracy to combination in the same manner.

From Panel (A2) of Table (2.4) we can see that introducing a nested model in the combination affects the performance favours the one-step approach more than the two-step alternatives. Both CGM and BMA approach lose accuracy with respect to the previous experiment (A1), such that the one-step approach is now the most accurate combination approach even in a small sample ($T = 50$). An explanation for this evidence can found in the inference of the one-step procedure: estimating all the parameters jointly enable the combination to detect the irrelevance of the noise term x_t, irrelevance that is "hidden" in the first step of the two-step procedures.

2.3.2 Bimodal DGP

Following Everitt [1985], the multimodal structure of the combination of distribution well suits multimodal data. The following experiments aim to check this theoretical proof and detect how model misspecification affects different approaches to combination in this environment.

The second DGP is obtain imposing to equation (2.14): $V_1 : f_N(\theta_{0,1} + \theta_{1,1}y_t, \sigma_1^2)$, $V_2 : f_N(\theta_{0,2} + \theta_{1,2}y_{t-1}, \sigma_2^2)$, $V_3 = 0$ and g being a concave function: $\eta_1 + \eta_2 = 1$ such that:

$$\textbf{DGP B1:} \quad y_{t+1} = \eta_1 f_N(\beta_{0,1} + \beta_{1,1}y_t, \sigma_1^2) + \eta_2 f_N(\beta_{0,2} + \beta_{1,2}y_{t-1}, \sigma_2^2)$$

The variable of interest is distributed as a mixture of two AR models with different orders: one AR(1) and one AR(2) where, $\eta_1 + \eta_2 = 1$, $\eta_k > 0$. This model setup has a long tradition in mixture autoregressive models for nonlinear time series. In particular, the parameters chosen to simulate data are taken from the results in Wong and Li [2000]. The parameters of the first component are assumed being equal to: $\mu_1 = -1; \beta_{0,1} = -0.5; \beta_{1,1} = 0.5, \sigma_1^2 = 0.6$, while for the second component of the mixture: $\mu_2 = 1; \beta_{0,2} = 0.7; \beta_{1,2} = 0.2, \sigma_2^2 = 0.3$. Mixing probabilities are set at $\eta_1 = 0.25, \eta_2 = 0.75$. The probabilistic forecasts to be combined are:

$$\begin{aligned} f_1 &: y_{t+1} \sim f_N(\theta_{0,1} + \theta_{1,1}y_t, \sigma_1^2) \\ f_2 &: y_{t+1} \sim f_N(\theta_{0,2} + \theta_{1,2}y_{t-1}, \sigma_2^2) \end{aligned} \tag{2.20}$$

In this set up both individual models are misspecified, but their linear combination mimics the DGP. This exercise aims to study the trade-off between parameter estimation noise and forecast accuracy in a bivariate environment. Both combination approaches are supposed to combine the two components correctly; however, the highly parametrised inference in the one-step approach may cost forecast accuracy. It would be interesting to exploit how big T has to be to let the one-step approach correctly estimate the DGP. From Panel (B1) of Table (2.4), we can see that, according to log-score and for all three sample sizes studied here, two-step approaches CGM and BMA outperform the one-step procedure. However, it is interesting to notice that the two versions of CRPS indicate that the one-step procedure is more accurate than the two-step alternatives at $T = 50$ and $T = 200$. Considering that CRPS is a measure for forecast accuracy more robust to outliers than log-score, we can prefer a one-step approach instead of two-step approach. However, the result for $T = 1000$ is puzzling: despite the increasing sample size, the trade-off between parameter estimation noise and forecast accuracy is still substantial and affects one-step approach forecast accuracy so much to let us move in favour BMA two-step approach.

The third DGP is obtain imposing to equation (2.14): $\beta_3 = 0$, $V_1 : f_N(\theta_{0,1} +$

$\theta_{1,1}y_t, \sigma_1^2)$, $V_2 : f_N(\theta_{0,2} + \theta_{1,2}x_t, \sigma_2^2)$ and g being a concave function such as $\eta_1 + \eta_2 = 1$ such that:

$$\textbf{DGP B2:} \quad y_{t+1} = \eta_1 f_N(\beta_{0,1} + \beta_{1,1}y_t, \sigma_1^2) + \eta_2 f_N(\beta_{0,2} + \beta_{1,2}x_t, \sigma_2^2)$$

The variable of interest y_{t+1} is distributed as a mixture of an AR(1) model and a linear regression model with one explanatory variable x_t. This time x_t is not a noise term but is a variable correlated with y_t with correlation $\rho = 0.2$. It has then some explanatory power for y_t. Parameters are set in the same way as the previous exercise.

As before, the probabilistic forecasts to be combined are:

$$\begin{aligned} f_1 &: y_{t+1} \sim f_N(\theta_{0,1} + \theta_{1,1}y_t, \sigma_1^2) \\ f_2 &: y_{t+1} \sim f_N(\theta_{0,2} + \theta_{1,2}x_t, \sigma_2^2) \end{aligned} \tag{2.21}$$

In this set up both individual models are misspecified, but their linear combination mimics the DGP. This exercise aims to study whether the effect of the trade-off between parameter estimation noise and forecast accuracy is still present when an explanatory variable is considered.

From Panel (B2) of Table (2.4), we can see that at, according to log-score, CRPS and TW-CRPS and for all the sample sizes considered one-step approach delivers the most accurate combined probability forecast. The estimation drawback of the one-step approach is then overcome by the ability to account for dependence between the mixture's components.

2.3.3 Nonlinear DGP

The simulation exercises in this section aim to investigate whether combination procedures can fit nonlinear data. The nonlinearity analysed is structural breaks. The motivation behind this is twofold. First, Stock and Watson [1996], and subsequent papers find that a wide variety of economic time series is subject to structural breaks; this comparison between combination approaches cannot be complete without testing their predictive ability under breaks. Second, the evidence of instability of parameters of autoregressive models fitted to economic time series subject to structural breaks. For this reason, this section applies the comparison between the combination procedures to the framework subject to different breaks. The exercise aims to test the hypothesis that a combination of forecasts can overcome the lack of predictability of component models under breaks.

Regardless of the majority of literature on breaks, this paper does not aim to detect breaks in the time series. Indeed this information is treated as unknown. Instead, the exercise wants to assess how well combination approaches that do not consider breaks perform. The third type of DGP is obtained imposing to equation (2.14) $V_1 = y_{t-1}$, $V_2 = x_{t-1}$ and $V_3 = h(t)$ is a function of time that defines the type of structural break in the sample. The variable of interest y_t is then simulated from a AR(1) model and it exhibits a break at point $t = T_b$.

DGP C: $y_t = \mathbf{B}\boldsymbol{X}_t + \varepsilon_t \quad \epsilon_t \overset{\text{iid}}{\sim} (0, \boldsymbol{\sigma}_\varepsilon^2)$

$$\boldsymbol{X}_t = \begin{bmatrix} 1 \\ y_{t-1} \\ x_{t-1} \end{bmatrix} \tag{2.22}$$

where x_{t-1} is an exogenous variable assumed to be independent from y_t. V_3 imposes the break to parameters $\boldsymbol{B}$ and $\boldsymbol{\sigma}_\varepsilon^2$ such that:

$$\begin{aligned} \mathbf{B} &= \begin{cases} [\beta_0 \quad \beta_1 \quad \beta_2] \ , & \text{if } t < T_b \\ [\beta_0 + d_0 \quad \beta_1 + d_1 \quad \beta_2 + d_2] \ , & \text{if } T_b \leq t \leq T. \end{cases} \\ \boldsymbol{\sigma}_\varepsilon^2 &= \begin{cases} \sigma_\varepsilon^2 \ , & \text{if } t < T_b \\ \sigma_\varepsilon^2 + s \ , & \text{if } T_b \leq t \leq T. \end{cases} \end{aligned} \tag{2.23}$$

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, and to be independent of y_t: a noisy variable.

Let us impose that $\beta_2 = 0$, such that the first part of the sample is a simple AR(1) model, and then a break is imposed on 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_ε^2 (experiments 8-9). All the types of breaks are presented in Table (2.2). 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: $T = \{50, 200, 1000\}$.

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 2.2: Simulation set-up under a single break scenario. Parameters' values are assumed to be $\beta_0 = 0.5$, $\beta_1 = 0.8$, $\beta_2 = 0$. 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.

Concerning the individual models, two cases are considered: a complete and an incomplete model set. They are the same as in the unimodal exercise i.e. simulation set up (**A1** and **A2**).

C1 : Incomplete model set combines two misspecified models:

$$\begin{aligned} f_1 &: y_{t+1} \sim \mathcal{N}(\theta_{0,1} + \theta_{1,1} y_t, \sigma_1^2) \\ f_2 &: y_{t+1} \sim \mathcal{N}(\theta_{0,2} + \theta_{1,2} x_t, \sigma_2^2) \end{aligned} \tag{2.24}$$

In this setup, f_1 is correctly specified for the sample period before the break, but both are misspecified for the remaining part of the sample.

C2 : Complete model set combines a misspecified model f_1 and a second model f_2 that mimic the GDP (in the no breaks scenario):

$$\begin{aligned} f_1 &: y_{t+1} \sim \mathcal{N}(\theta_{0,1} + \theta_{1,1} y_t, \sigma_1^2) \\ f_2 &: y_{t+1} \sim \mathcal{N}(\theta_{0,2} + \theta_{1,2} y_t + \theta_{2,2} x_t, \sigma_2^2) \end{aligned} \tag{2.25}$$

Since two-step tends to attach extreme values of weights to the components, it is more accurate in experiments $(1 - 4, 8, 9)$ (where f_1 mimics the structure of the DGP). On the contrary, it is supposed to be less accurate in the experiments $(5 - 7)$ where neither component mimics the DGP. The size of the break matters as well; large

breaks 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, the two-step becomes more accurate in the case of incomplete model set, less accurate in the complete-model-set case. The timing of the break matters as well; it seems reasonable that performances worsen when the break is located at the end of the sample because parameters and weights estimates are biased. However, in the presence of a large sample size T, the post-break subsample can be large enough to correct this bias.

Figures (2.1) and (2.2) regard DGPs which are AR(1) processes with a break in the intercept. As expected, in the case of incomplete model set, 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, the one-step approach does not become more accurate with a large break.

In the case of complete model set, the 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) in Figure (2.3) and (4) in Figure (2.4) and experiments (8) in Figure (2.8) and (9) in Figure (2.9) . Figure (2.5) and (2.6) 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 procedures must be preferred in the incomplete (complete)-model-set case. When the break is large (exp. 6, figure 2.6), the one-step is more accurate in the complete-model-set case and in the incomplete-model-set case for small samples (i.e. $T = 50$ and $T = 200$) (apart from the case in which the break is close to the end of the sample). Finally, let us consider experiment (7) in Figure (2.7), where the break regards both the AR dynamics and the exogenous variable parameter. The one-step procedure is more accurate than the two-step in all cases besides the large sample, incomplete-model-set case.

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

2.4 Results from the simulation exercises

Arguably, the two-step procedure is the most popular in applied macroeconomics; however, the one-step procedure accounts for dependency among forecasts. Moreover, since using the one-step combination approach is computationally more elaborated, endowing the decision-maker with a tool to discriminate when it is worth is crucial.

In this section a series of simulation exercises investigates the different performances of one-step and two-step procedures in controlled environments. Three families of data generating processes are considered: unimodal, multimodal and non-linearity given by breaks. To forecast these DGP, the researcher is endowed with two predictive distributions, whose features vary across simulation setups. The literature inspires the designs of simulations in the combination of density forecasts in macroeconomics, such as the difference between the combination of complete and incomplete model sets, dependence among forecast models and presence of breaks in the time series object of the forecast. From the simulation exercises presented in the section, one can infer that: the one-step approach delivers more accurate combined forecasts when one forecast model nests the other; when the time series is subject to breaks and the sample size is sufficiently large (i.e. greater or equal than 50 observations). Conversely, the two-step procedure must be preferred when the sample size is small, and the components are nonnested.

The results obtained in this paper are subject to the design of the simulation exercise, (in particular the characteristics of the data generating processes) and the decision on using a one-step or two-step procedure is based on these circumstances. Further works would expand the simulation set to different pooling schemes (such as logarithmic and beta-transformed pools) and to a higher number of forecasting models (hence linking the highly parametrised application exercise).

2.5 Conclusions

This paper proposes a comparison between two inference approaches to the combination of density forecasts. The empirical exercise performed in this Section (2.2) leaves us with no clear indication over which combination approach is the most accurate. Whether the conflict between evaluation criteria has been treated in the literature, it is still unclear why different combination approaches perform better (worse) for output growth and fail (succeed) for inflation. The results also depend on other factors such as the number of forecasts K to combine and the time window. Section (2.3) tried to shed light on understanding what affects the different performances in controlled environments. Several types of DGPs and misspecifications for forecasting models have been studied. The trade-off between parameter estimation noise and forecast accuracy typical of the one-step approach is crucial to identify which combination approach to use given the forecasting problem at hand: the one-step approach delivers more accurate combined forecasts when one forecast model nests the other; when the time series is subject to

breaks and the sample size is sufficiently large (i.e. greater or equal than 50 observations). Conversely, the two-step procedure must be preferred when the sample size is small, and the components are nonnested.

2.6 Appendices

2.6.1 Conflitti et al. [2015] Algorithm for Two-step Optimal Weight Maximisation

As described in section (2.2), the optimality problem reduces to the maximisation of the concave cost function:

$$\Phi(\omega_j^*) = \frac{1}{T}\sum_{t=1}^{T} \ln\, g(Y_t) \tag{2.26}$$

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 2.26 can be rewritten as:

$$\Phi(\omega_j^*) = \frac{1}{T}\sum_{t=1}^{T} \ln\, (\hat{G}\omega_j). \tag{2.27}$$

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

$$\Phi(\omega_j^*) = \frac{1}{T}\sum_{t=1}^{T} \ln\, (\hat{G}\omega_j) - \lambda\sum_{j=1}^{m} \omega_j. \tag{2.28}$$

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

$$\Psi_\lambda(\omega_j, a_j) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{m} \frac{\hat{G}_{jt}a_j}{\sum_{j=1}^{m}\hat{G}_{jt}a_j} \ln\left(\frac{\omega_j}{a_j}\sum_{j=1}^{m}\hat{G}_{jt}a_j\right) - \lambda\sum_{j=1}^{m}\omega_j. \tag{2.29}$$

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

$$\omega_{j,\lambda}^{k+1} = \underset{\omega}{\text{argmax}}\ \ \Psi_\lambda(\omega_j, \omega_{j,\lambda}^{k}) \tag{2.30}$$

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

$$\omega_j^{k+1} = \omega_j^k \frac{1}{T} \sum_{t=1}^{T} \frac{\hat{G}_{jt}}{\sum_{j=1}^{m} \hat{G}_{jt} \omega_j^k}. \tag{2.31}$$

The nonnegative constrain is satisfied by 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 (2.29) and the constraints. The algorithm also has a stop criterion based on a negligible difference between two successive iterates.

2.6.2 The One-step approach Bayesian Inference

Algorithm 1 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**

A. 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 $C_k = C_0 + \frac{1}{2}\epsilon_k'\epsilon$

B. 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:

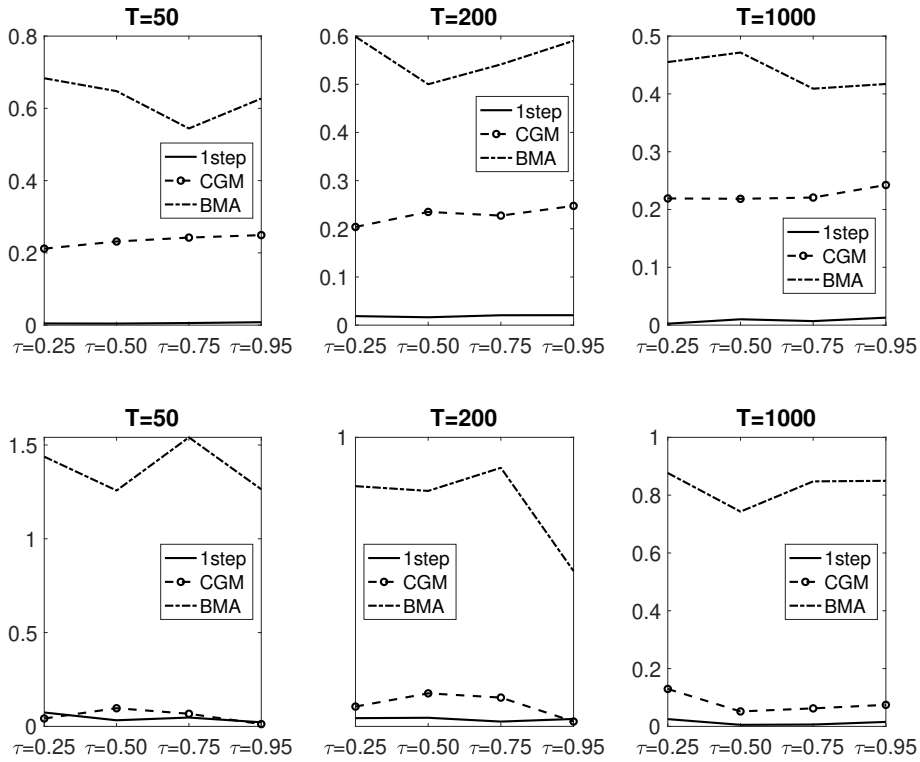
$$p(y_{t+1}|y_t^o, \boldsymbol{\theta}_k) = \frac{1}{M + M_0} \sum_{m=1}^{M} \left(\sum_{k=1}^{K} \eta_k^m p(y_{t+1}|\theta_{k,t+1}^m) \right)$$

Table 2.3: 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
rrovnght@us	level	59:M1-18:M6	FEDFUNDS	Int. Rate: Fed Funds (Effective)	F
rrtbill@us	level	59:M1-18:M6	TB3MS	Int. Rate: 3-Mn Tr. Bill, Sec Mkt Rate	F
rrbnds@us	level	59:M1-18:M6	GS1	Int. Rate: US Tr. Const. Mat., 1-Yr	F
rrbndm@us	level	59:M1-18:M6	GS5	Int. Rate: US Tr. Const. Mat., 5-Yr	F
rrbndl@us	level	59:M1-18:M6	GS10	Int. Rate: US Tr. Const. Mat., 10-Yr	F
rstockp@us	Δln	59:M1-18:M6	SP500	US Share Prices: S&P 500	F
rexrate@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
rmon0@us	Δln	59:M1-18:M6	AMBSL	Monetary Base, sa	I
rmon1@us	Δln	59:M1-18:M6	M1SL	Money: M1, sa	I
rmon2@us	Δln	59:M1-18:M6	M2SL	Money: M2, sa	I
rmon3@us	Δln	59:M1-06:M2	M3SL	Money: M3, sa	I

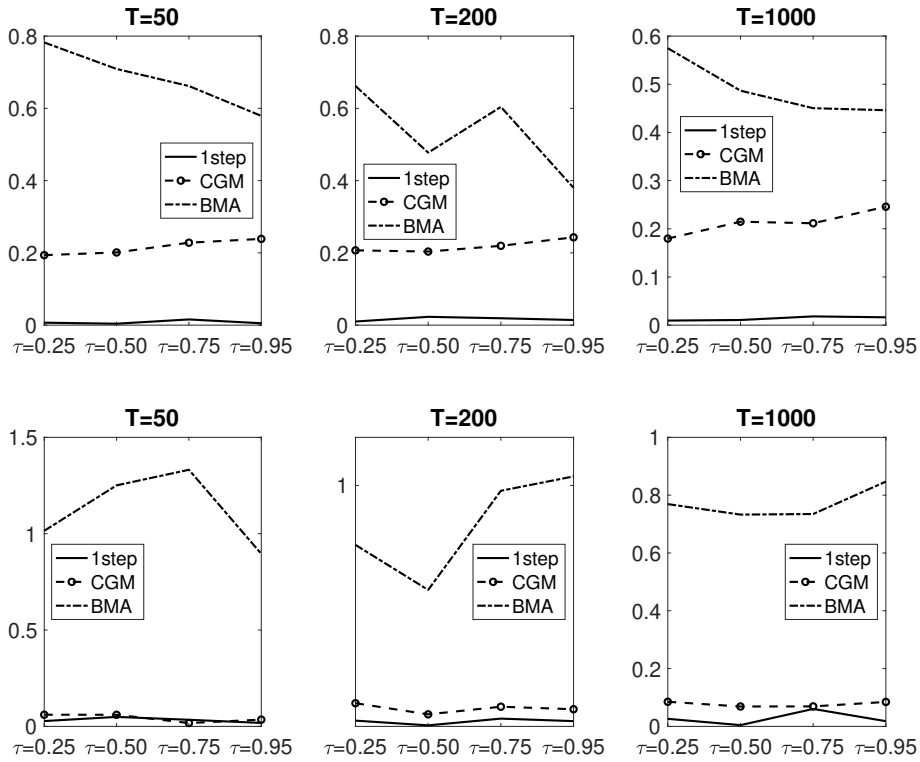
Notes: Sources abbreviated as "F" denotes Federal Reserve Economic Data (FRED) and "I" IMF International Financial Statistics. The "r" in front of the variable name denotes the transformation in real terms.

Figure 2.1: Accuracy Loss of one-step and two-step in presence of a small break in the intercept (exp.#1). Nested case in top three graphs, nonnested case in the bottom three.



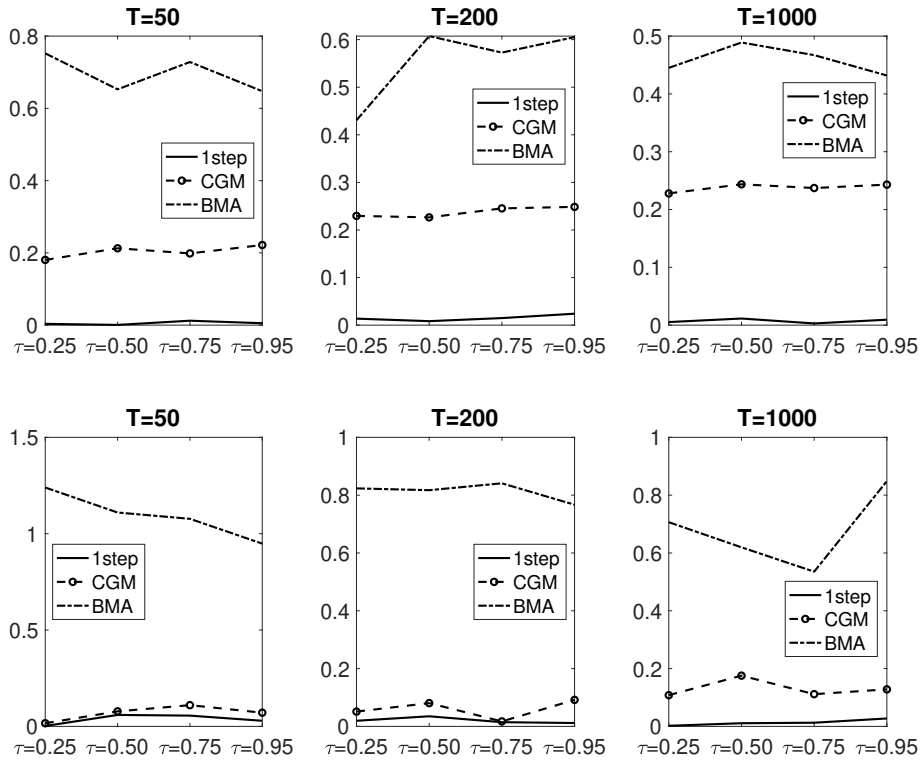
CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the DGP. Each point identifies the CRPS score obtained from different simulation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

Figure 2.2: Accuracy Loss of one-step and two-step in presence of a large break in the intercept (exp.#2). Nested case in top three graphs, nonnested case in the bottom three.



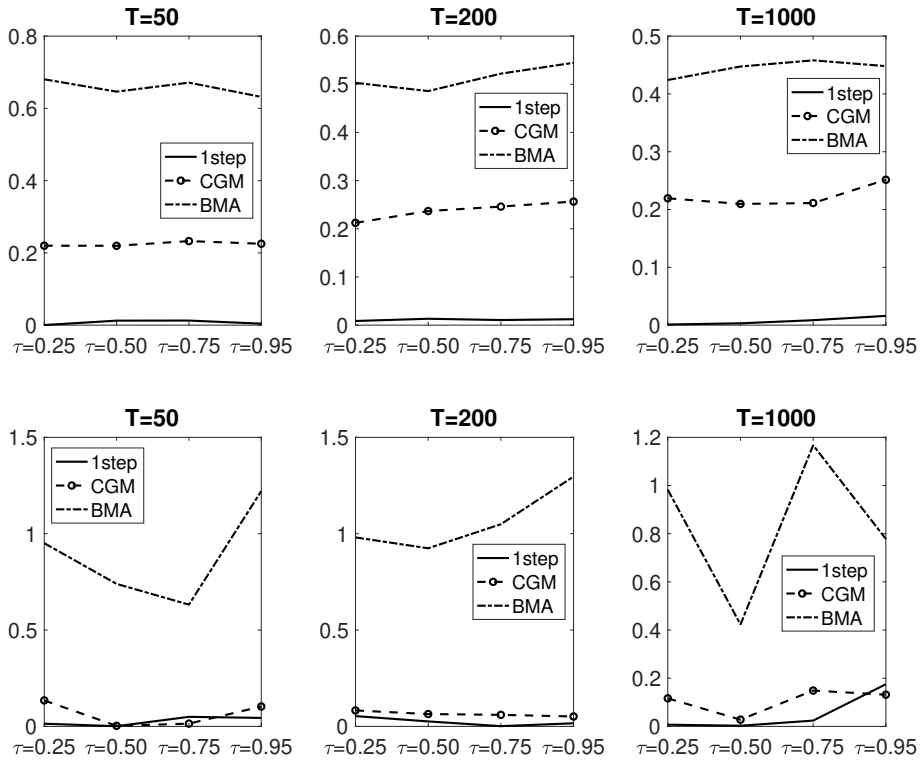
CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the DGP. Each point identifies the CRPS score obtained from different simulation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

Figure 2.3: Accuracy Loss of one-step and two-step in presence of a small break in AR(1) dynamics (exp.#3). Nested case in top three graphs, nonnested case in the bottom three.



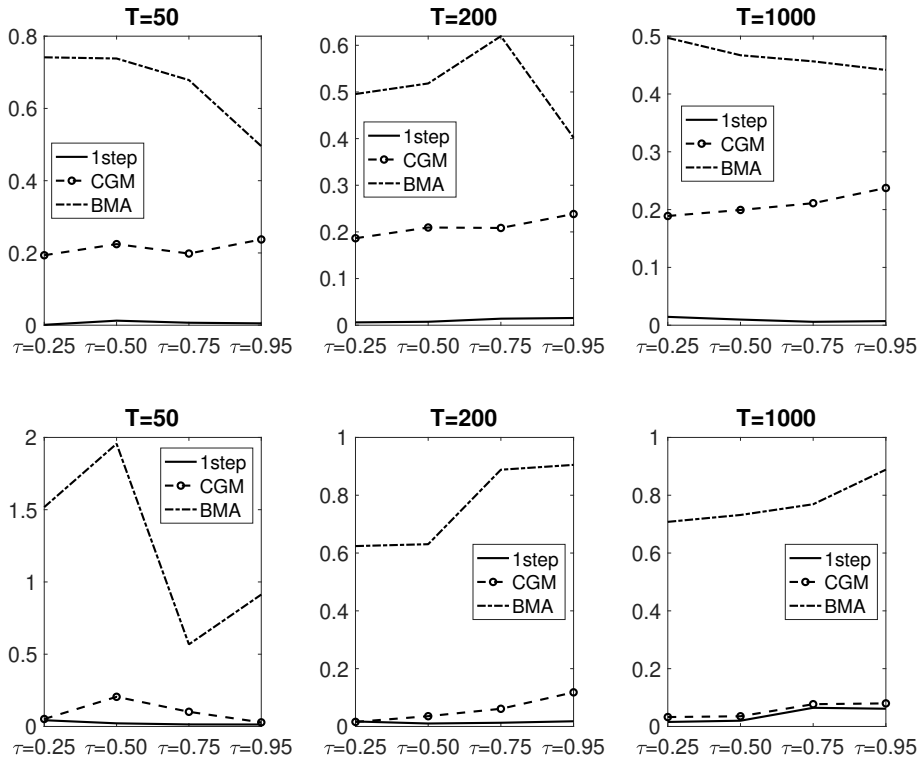
CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the DGP. Each point identifies the CRPS score obtained from different simulation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

Figure 2.4: Accuracy Loss of one-step and two-step in presence of a large break in AR(1) dynamics (exp.#4). Nested case in top three graphs, nonnested case in the bottom three.



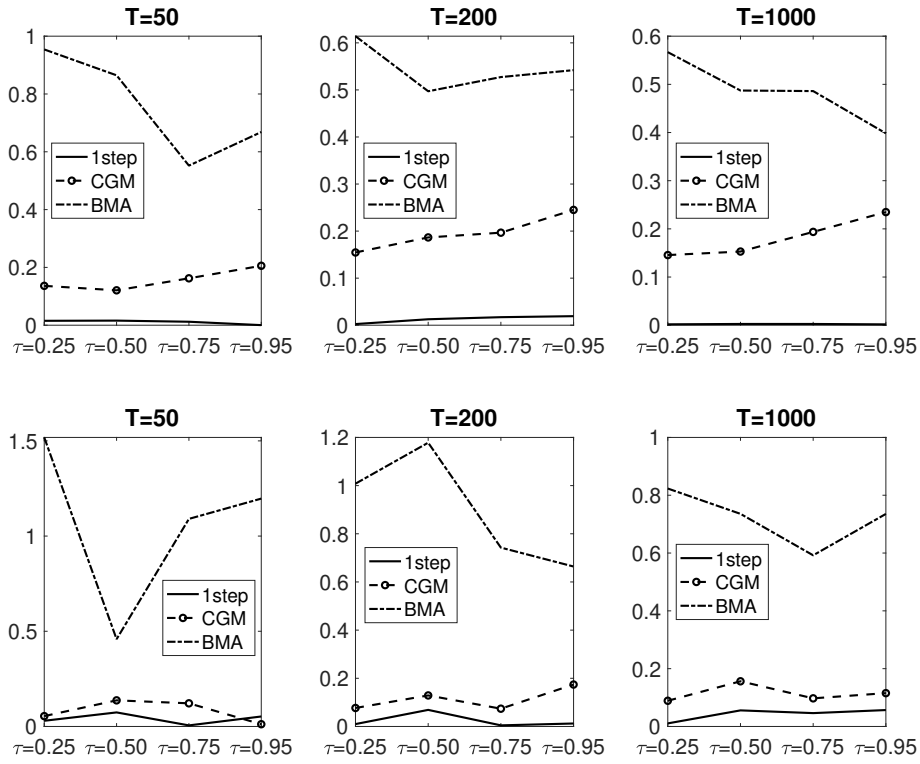
CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the DGP. Each point identifies the CRPS score obtained from different simulation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

Figure 2.5: Accuracy Loss of one-step and two-step in presence of a small break in exogenous variable coefficient (exp.#5). Nested case in top three graphs, nonnested case in the bottom three.



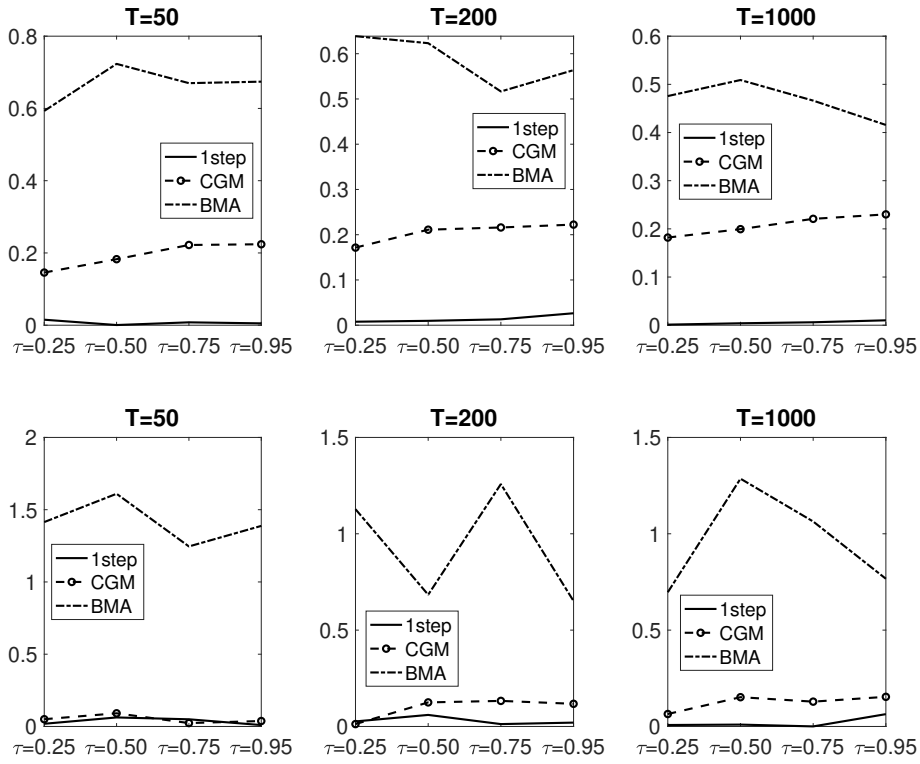
CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the DGP. Each point identifies the CRPS score obtained from different simulation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

Figure 2.6: Accuracy Loss of one-step and two-step in presence of a large break in exogenous variable coefficient (exp.#6). Nested case in top three graphs, nonnested case in the bottom three.



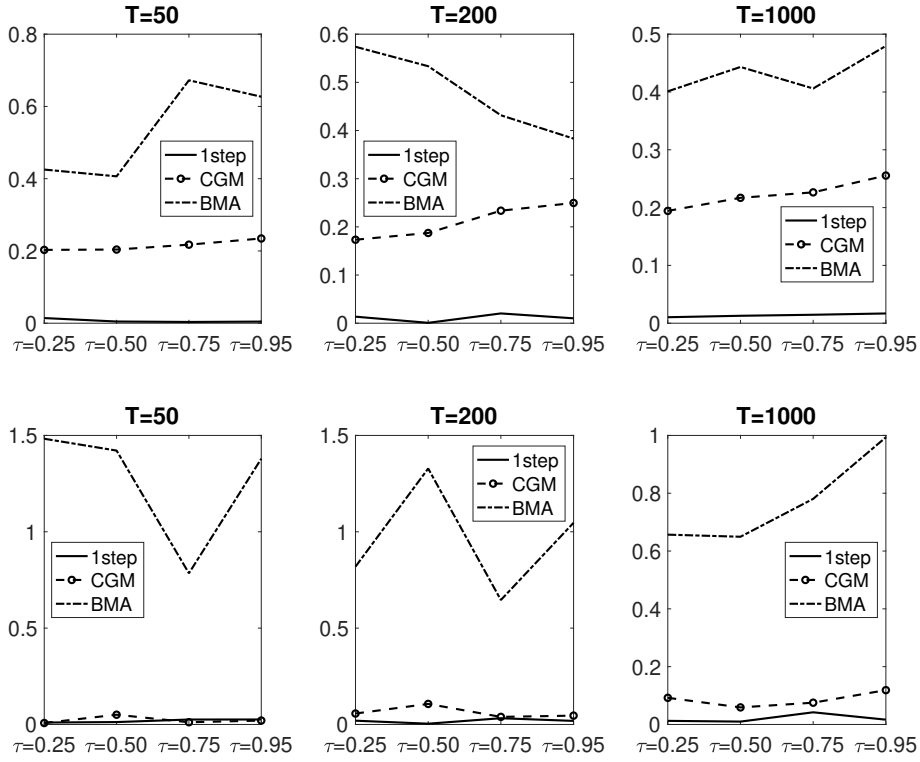
CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the DGP. Each point identifies the CRPS score obtained from different simulation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

Figure 2.7: Accuracy Loss of one-step and two-step in presence of a break in both AR(1) dynamics and exogenous variable coefficient (exp.#7). Nested case in top three graphs, nonnested case in the bottom three.



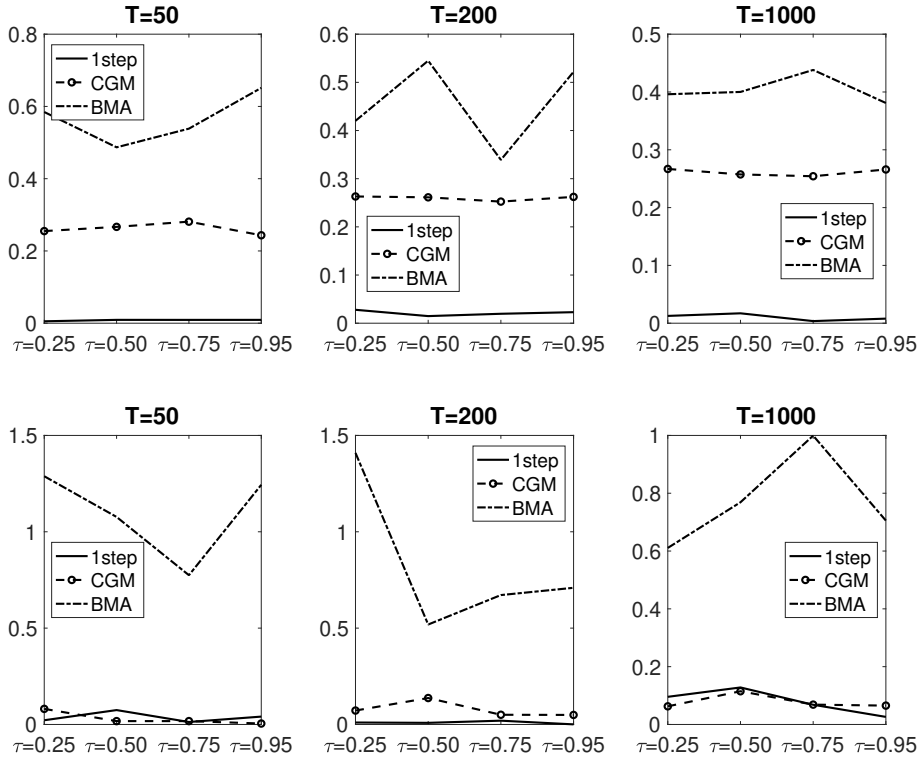
CRPS rates calculated against the DGP i.e. $(CRPS_{combination} - CRPS_{DGP})/CRPS_{DGP}$. Y axis report the loss in accuracy in percentage points. An higher CRPS rate indicates a higher loss in accuracy with respect to the DGP. Each point identifies the CRPS score obtained from different simulation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

Figure 2.8: Accuracy Loss of one-step and two-step in presence of an increase in post-break variance (exp.#8). Nested case in top three graphs, nonnested 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 DGP. 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 2.9: Accuracy Loss of one-step and two-step in presence of a decrease in post-break variance (exp.#9). Nested case in top three graphs, nonnested 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 DGP. Each point identifies the CRPS score obtained from different simulation setups, i.e. break timing (in X axis) and sample size ($T = 50, 200, 1000$).

Table 2.4: Simulated accuracy Loss of combined forecasts against the DGP: forecasts are combined according to one and two-step procedures.

		Panel A: Unimodal DGPs					
		DGP A1			DGP A2		
		One-step	Two-step CGM	Two-step BMA	One-step	Two-step CGM	Two-step BMA
$T = 50$	Log Score	0.0013	0.0005	0.1508	0.0009	0.0342	0.1522
	CRPS	0.0472	0.0500	0.5575	0.0154	0.2864	0.5690
	TW-CRPS	0.0624	0.0426	0.5592	0.0064	0.2804	0.5676
		One-step	Two-step CGM	Two-step BMA	One-step	Two-step CGM	Two-step BMA
$T = 200$	Log Score	0.0103	0.0125	0.0928	0.0010	0.0294	0.1490
	CRPS	0.0480	0.2468	0.3740	0.0004	0.2310	0.5295
	TW-CRPS	0.0458	0.2430	0.3680	0.0039	0.2412	0.5305
		One-step	Two-step CGM	Two-step BMA	One-step	Two-step CGM	Two-step BMA
$T = 1000$	Log Score	0.0093	0.0306	0.1442	0.0060	0.0317	0.1414
	CRPS	0.0328	0.2476	0.4156	0.0220	0.2608	0.4473
	TW-CRPS	0.0298	0.2491	0.4137	0.0149	0.2579	0.4461
		Panel B: Bimodal DGPs					
		DGP B1			DGP B2		
		One-step	Two-step CGM	Two-step BMA	One-step	Two-step CGM	Two-step BMA
$T = 50$	Log Score	0.0728	0.0293	0.0227	0.0018	0.0220	0.2040
	CRPS	0.3470	1.0758	0.8026	0.0068	0.1707	0.9443
	TW-CRPS	0.5274	1.1159	0.8364	0.0043	0.1688	0.9416
		One-step	Two-step CGM	Two-step BMA	One-step	Two-step CGM	Two-step BMA
$T = 200$	Log Score	0.0206	0.0240	0.0152	0.0023	0.0237	0.2081
	CRPS	0.3626	1.1586	0.6924	0.0033	0.1806	0.6382
	TW-CRPS	0.2560	1.1338	0.6750	0.0044	0.1808	0.6394
		One-step	Two-step CGM	Two-step BMA	One-step	Two-step CGM	Two-step BMA
$T = 1000$	Log Score	0.0339	0.0256	0.0038	0.0065	0.0211	0.2017
	CRPS	0.6060	0.9400	0.0770	0.0194	0.1700	0.6650
	TW-CRPS	0.2849	0.9189	0.0406	0.0157	0.1621	0.6660

Notes: Log Scores and CRPS rates are calculated against the DGP i.e. $(SCORE_{combination} - SCORE_{DGP})/SCORE_{DGP}$. An higher score rate indicates a higher loss in accuracy with respect to the DGP. T indicates the size of the simulated sample.

Chapter 3

Are Central Banks' Fan charts Reliable? On Calibration of Density Path Forecasts

3.1 Introduction

Forecasting a future state of the world is by nature probabilistic: forecasts include some degree of uncertainty that cannot be eliminated ([Dawid, 1982]). A way to overcome this issue is to communicate the degree of uncertainty together with the point forecast: in the last three decades, the employment of probabilistic forecast over point forecast has surged (Tay and Wallis [2000] and Gneiting and Katzfuss [2014] among the others). The characteristic feature of probabilist forecasts is to consider, through density or cumulative distribution, a range of candidate values for point forecasts and the probability associated with them. Central Banks communicate the density forecasts by publishing the so-called "fan chart". This graph joins the realisations with the most likely "path" for GDP growth, inflation and unemployment for 1 to H periods in the future. In this way, fan charts are informative about the point forecast at a specific horizon $h = 1, \ldots, H$ and about the path the economy will follow up to period H. In addition, fan charts present confidence bands around the prediction that gives the reader an understanding of its likelihood. As predictions become increasingly uncertain the further into the future one goes, the forecast ranges out, creating distinctive "fan" shapes, hence the name. Bank of England was the first central bank to publish fan charts in its "Inflation Report" since 1997, coining the name, but fan charts are also popular in other central banks' reports. According to Hammond et al. [2012], in 2012, 18 central banks regularly

publish fan charts: Armenia, Brazil, Colombia, Hungary, Perù, Poland, South Africa, Thailand, Turkey and UK publish fan charts for inflation and GDP; in addition to these, Israel, Czech Republic and Norway also publish fan charts for the key policy rate; Canada, Czech Republic and Ghana distinguish from the headline and core inflation; the Philippines publishes fan chart only for inflation; finally Sweden publishes fan chart for the repo rate besides CPI and GDP.

In addition to point and probabilistic forecasts, the fan charts also display the dynamics across horizons. For example, if the economy is forecasted to go through a recession, from the fan chart, one can infer how severe the recession will be (point forecast), how likely this forecast is (bands around the point) and how long the economy will take to exit the recession. This "third dimension" information of fan charts is called here "dynamics across horizons" or "horizon-dependency". Researchers often neglect horizon-dependency when they evaluate fan charts' forecast ability. A first attempt has been made by Martinez [2017], proposing a relative evaluation criterion for point path forecasts, i.e. an elicitation criterion to select, among a set of alternative path forecasts, the most accurate one. However, this paper does not consider path density forecasts.

This paper aims to propose an absolute evaluation criterion for path density forecasts. First, we are interested in absolute instead of relative criteria because we would like to have a tool to judge the forecast ability of the central bank's fan charts. Second, we decide to consider all the information in fan charts and not just the point forecast, as it is common to the literature on density forecasts.

Existing absolute evaluation approaches test the forecast accuracy separately across horizons. In doing so, the researcher implicitly assumes independence across horizons of path forecast. Denote for example a path forecast of variable of interest y_t for horizons $h = 1, \dots, H$ with $f_t(y_{t+1}, y_{t+2}, \dots, y_{t+H})$. This is a joint distribution of the horizon-specific predictive distributions. Using Sklar's theorem, we can represent any joint distribution as:

$$f_t(y_{t+1}, y_{t+2}, \dots, y_{t+H}) = f_t(y_{t+1}) \cdot f_t(y_{t+2}) \dots f_t(y_{t+H}) \cdot C(\cdot) \tag{3.1}$$

where $C(\cdot)$ is the copula function that identifies the dependence structure among forecasts. Evaluating forecast accuracy, as done in literature, equals assuming $C(\cdot) = 1$. In our application to fan charts, the family of $C(\cdot)$ is unknown since Bank of England publishes partial information on joint density. This paper provides a measure of loss of accuracy in absolute evaluation given by the missing information, allowing for dependence across horizons (called here horizon dependence) and let $C(\cdot)$ be different from

1. This paper will answer the following questions: are Bank of England fan charts calibrated? How can we test the path density calibration? Do we have information about the horizon-dependency? Which test statistic should we use?

The results show that the fan chart for GDP growth and unemployment are not calibrated while inflation is. The choice of the test depends on the availability of information about the time dependence. If they are available, we propose to evaluate PITs of marginal and conditional distributions; if no, we propose a set of alternative "sup tests" and a technique to approximate the horizon dependence. Finally, we explore the properties of a set of the most used test statistics by Monte Carlo exercises.

The rest of the paper is organised as follows: Section (3.2) presents the notion of path density forecast; and defines its calibration; Section (3.3) discusses the issue of time-dependence and proposes several approaches as evaluation criteria; Section (3.4) shows, through Monte Carlo simulations, the size and power properties of the calibration test statistics; Section (3.5) applies the tests to Bank of England fan charts; finally, Section (3.6) contains concluding remarks.

3.2 Fan chart as Path Density Forecast and its Calibration

A path density forecast is a sequence of forecasts 1 to H periods in the future. It is informative about the prediction at a specific horizon and the path the variable will follow up to H. However, examining the most likely path is only half of the story of understanding the distribution of possible outcomes about that path. Path forecast is much more than a simple collection of predictions: it is informative of the dynamics of the variable of interest. A rich literature addresses the construction of pathwise confidence bands from the H marginal predictive densities. Hall and Titterington [1988] proposes confidence bands in a nonparametric density estimation; Jordà and Marcellino [2010] introduces a method for the construction of bands based on the joint asymptotic distribution of forecast error. However, as asymptotic methods rely on a large number of observations, they may provide poor results for small samples. In finite samples, the paths are often obtained by bootstrapping such as in Wolf and Wunderli [2015] or other simulation-based methods (for example in Vidoni [2017], Garratt et al. [2003], Schüssler and Trede [2016].

One example of the employment of path density forecasts is fan charts. Central banks, as other organisations, communicate their monetary policies through graphs that show the future state of the economy for 1 to H periods in the future. Fan charts are then informative about the prediction at a specific horizon and about the economy's

path to period H. In addition to the most likely path for an economy, fan charts present confidence bands that give the reader an understanding of its likelihood. However, this information is difficult to extract by the subjective inspection of the fan charts. Thus, it is difficult to perceive the change in the dynamics from one publication to its revision. For this reason, it is crucial to evaluate the accuracy of paths density forecasts with respect to the actual realisations.

The type of evaluation that targets the "consistency" of forecasts to the data is better known as probabilistic calibration. Following Dawid [1982] prequential principle, the predictive distributions need to be assessed only based on the forecast-observation pairs. Diebold et al. [1997] proposed the use of the probability integral transform (PIT) value firstly introduced by Rosenblatt [1952] for this purpose. If the density forecast for a specific horizon h, $f_t(y_{t+h})$ is probabilistically calibrated then its PITs z_{t+h} have a uniform distribution.

$$z_{t+h} = \int_{-\infty}^{y_{t+h}} f_t(y_{t+h})dy \sim \text{ i.i.d. } U(0,1) \tag{3.2}$$

for $t = 1, \ldots, T$ and y_{t+h} being the realization. Checking the calibration can be done by visual inspection methods involving histograms and correlograms of probability integral transforms or using testing procedures developed by many studies. In this paper we will use Kolmogorov, Cramer-von Mises, Berkowitz [2001] and Knüppel [2015] statistics.

3.2.1 Calibration of path density

We are interested in assessing whether $F_t(y_{t+h})$ is correctly specified, i.e.:

$$H_0 : F_t(y_{t+1}, \ldots, y_{t+H}) = G_t(y_{t+1}, \ldots, y_{t+H}) \tag{3.3}$$

where G_t is the true data generating process for path the variable of interest y_t and $F_t(y_{t+1}, \ldots, y_{t+H})$ denotes the associated probabilistic path forecasts made at time t. We call a forecast that satisfy Equation (3.3), a calibrated forecast. In this paper, for calibration we refer to probabilistic calibration defined by Gneiting et al. [2007] as the "statistical consistency between the distributional forecasts and the observation". The sequence $F_t(y_{t+1}, \ldots, y_{t+H})$ is probabilistically calibrated relative to the sequence $G_t(y_{t+1}, \ldots, y_{t+H})$ if:

$$\sum_{t=1}^{T} G_t \circ F_t^{-1}(p) \to p \quad \text{for all} \quad p \in (0,1). \tag{3.4}$$

In a uni-dimensional (univariate and/or single horizon) framework the probabilistic calibration is equivalent to the iid uniformity of probability integral transforms (PIT) values. Defined by Rosenblatt [1952], the PIT is the value p_t that predictive CDF $F_t(y_{t+h})$ attains at the observation y_{t+h}. The connection between PITs values and calibration can be achieved by substituting the empirical distribution function $\mathbb{1}\{y_{t+h} \leq y\}$ for the data generating distribution $G_t(y)$, $y \in \mathbf{R}$ and noting that $y_{t+h} \leq F_t^{-1}(p)$ if and only if $p_t \leq p$. Equation (3.4) becomes:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{p_t < p\} \to p \quad \text{almost surely} \quad \text{for all} \quad p. \tag{3.5}$$

The empirical sequence of PIT values p_t in terms of probabilistic calibration is characterised by asymptotic uniformity. Following Dawid [1982] and Diebold et al. [1997], the probability integral transform values has become a cornerstone of forecast evaluation. Uniformity can be assessed in several ways; one is to plot the empirical CDF of the PIT values and comparing to the identity function. Although visual inspection can be informative, for example, detecting the reasons for forecast deficiency, sometimes the employment of formal tests is preferable since they are not affected by subjective interpretability. The assessment of probabilistic calibration throughout formal testing tracked back to Pearson [1933]. Statistical tests for uniformity are compared in Marhuenda et al. [2005].

Unfortunately, the definition of calibration for path forecasts is more complicated since the path has a multi-dimensional joint distribution.
Let $F_t(y_{t+1}), F_t(y_{t+2}), \ldots, F(y_{t+H})$ be a sequence of probabilistic forecasts for the variable of interest y_t at horizons $h = 1, 2, \ldots, H$. Let $\Psi = (y_{t+1}, y_{t+2}, \ldots, y_{t+H})$ be a continuous random variable whose outcomes consist of ordered H-tuples of realisations, with the h-coordinate lying in the set R^h. The sample space Ω of Ψ is the Cartesian product of the R_h's:

$$\Omega = \mathbf{R}_1 \times \mathbf{R}_2 \times \ldots \times \mathbf{R}_H. \tag{3.6}$$

The path density forecast is defined as the joint distribution of Φ that can be expressed by the joint cumulative distribution function:

$$F_t(y_{t+1}, \ldots, y_{t+H}) = P(F_t(y_{t+1}) < y_{t+1}, F_t(y_{t+2}) < y_{t+2} \ldots F_t(y_{t+H}) < y_{t+H}). \tag{3.7}$$

or the joint pdf f of Ψ is defined by:

$$f_t(y_{t+1}, \ldots, y_{t+H}) = \int_{-\infty}^{y_{t+1}} \ldots \int_{-\infty}^{y_{t+H}} f_t(y_{t+1}) \ldots f_t(y_{t+H}) \quad dy_H \ldots dy_1 \tag{3.8}$$

$$f_t(y_{t+1}, \ldots, y_{t+H}) = \frac{\partial^H F_t(y_{t+1}, \ldots, y_{t+H})}{\partial y_{t+1}, \partial y_{t+2}, \ldots, \partial y_{t+H}} \tag{3.9}$$

The probabilistic forecasts $F_t(y_{t+h})$ can be obtained by a parametric model, in this paper we treat the set of parameters as known.

The literature about the evaluation of path forecasts is still embryonic. Some procedures propose to evaluate the single horizons independently and the conclusion about the calibration of the stream of forecasts. In doing so, the horizon-dependency is neglected: each prediction depends on the likelihood and moments of the previous one. To evaluate the calibration of the paths instead of the individual forecasts for the different horizons, one has to consider the horizon dependence. Martinez [2017] has made some steps in this direction, which accounts for the covariance between horizons using the general forecast-error second-moment matrix proposed by Clements and Hendry [1993] as a metric of path forecast accuracy. However, it treats the path forecast as a mere stream of points and not its density. This paper accounts for the horizon-dependence adapting the evaluation criteria of the multivariate literature. Since the works of Diebold et al. [1999], multivariate density evaluation has become popular in the fields of time series forecasting and risk evaluation. From it, Clements and Smith [2002], Ko and Park [2013] and other works take a stand. Multivariate evaluation criteria are natural candidates for our multi horizon evaluation task: each horizon is treated as a variable, depending on the previous horizon and the information set.

Multihorizons evaluation tests are optimal approaches to assess path density calibration. However, they assume some knowledge about the dynamics underlying the horizons. For Bank of England fan charts, the dynamics between horizons is not disclosed: the only information published regards density forecasts for each horizon. Our paper proposes two approaches based on availability of information about horizon-dependence.

3.3 Econometric Framework

This section defines and contrasts two calibration tests: one based on marginal and conditional distributions and one based on marginal distributions only. The choice of the test depends on the availability of information about the horizon dependence. The

lack of knowledge about horizon dependence in path density forecast can have various nature. It can come from a problem of disclosing (such as the Bank of England fan chart) or because the path is built from horizon-specific models. Building paths from indirect forecasts gives us total information about the joint distribution, while paths from direct forecasts are a clear example of the latter case.

Regardless of the reason, if the information is available, we can decompose the joint distribution in marginals and conditionals; in this case, we propose to test calibration on the pits of marginal and conditional distribution as in Section (3.3.1); otherwise, we propose a set of alternative test on marginal distributions only in Section (3.3.2).

3.3.1 Tests on PITs of Marginal and Conditional distributions

This section proposes and contrasts different approaches to path evaluation found in literature of multivariate evaluation.

Decomposition of Joint Distribution

In the absence of a multi-dimensional notion of calibration, we can adapt the strategy of decomposing the multi-dimensional density forecast into uni-dimensional densities and apply the definition just mentioned before. However, one open question is how to decompose the path CDF (3.7) or PDF (3.9) adequately. A first, natural strategy could be to use the chain rule of conditional distributions, which allows to write any joint distribution as the product of conditional and marginals, i.e.:

$$F_t(y_{t+1}, \ldots, y_{t+H}) = F_t(y_{t+H}|y_{t+H-1}, \ldots, y_{t+1}) \ldots F_t(y_{t+2}|y_{t+1})F_t(y_{t+1}) \tag{3.10}$$

This decomposition has been used in the literature of evaluation of multivariate forecasts from the work of Diebold et al. [1999]. However, in our framework, the equation (3.10) can be simplified according to the law of iterated projections in:

$$F_t(y_{t+1}, \ldots, y_{t+H}) = F_t(y_{t+H}|y_{t+H-1}) \ldots F_t(y_{t+2}|y_{t+1})F_t(y_{t+1}) \tag{3.11}$$

For an example, please refer to the Appendix (3.7.1) for the decomposition of path using a AR(p) forecasting model. Thanks to this decomposition of path density forecast into uni-dimensional distributions, the calibration can be assessed easily with standard tests. The task consists of taking the probability integral transform of each element of equation (3.11), i.e.:

$$\{F_t(y_{t+H}|y_{t+H-1}), \ldots, F_t(y_{t+2}|y_{t+1}), F_t(y_{t+1})\} \tag{3.12}$$

then testing their uniformity. However, assessing the calibration of each distribution in (3.12) is not equivalent to evaluate the calibration of the overall path, which we are interested in. In other words, the PITs must be "aggregated" back in order to perform joint calibration tests. In the following section, a set of vectors of PITs will be presented and discussed.

A. **The Diebold et al. [1999] approach**

Following Diebold et al. [1999], the literature has proposed several approaches we can apply to test whether a set of "multi-horizon" densities coincides with the true path density. Diebold et al. [1999] proposes a factorisation of the joint distribution as the product of its conditional and marginal distributions as in (3.11). Consider a three-horizon path, the joint distribution is factorised in:

$$f_t(y_{t+1}, y_{t+2}, y_{t+3}) = f_t(y_{t+3}|y_{t+2}) f_t(y_{t+2}|y_{t+1}) f_t(y_{t+1}) \tag{3.13}$$

for $t = 1 \dots T$. This procedure produces a set of H pits (Z) series (labelled $Z_{h=3|t+2}, Z_{h=2|t+1}, Z_{h=1|t}$). Where $z_{h|t}$ is by definition:

$$z_{h|t} = \int_{-\infty}^{y_{t+h}} f_{t+h}(y_{t+h}) dy_{t+h} \sim \mathrm{U}(0,1) \tag{3.14}$$

for $t = 1, \dots, T$ and $Z = [z_{t=1}, z_{t=2}, \dots, z_{t=T}]$. Where $f_{t+h}(y_{t+h})$ is defined as the density forecast. The vector to test for uniformity is then:

$$Z_{DHT} = \begin{bmatrix} Z_{h=3|h=2,t} \\ Z_{h=2|h=1,t} \\ Z_{h=1|t} \end{bmatrix} \tag{3.15}$$

which has dimension $TH \times 1$. The approach of stacking conditional and marginal components has been criticized by Clements and Smith [2000] and subsequent literature of multivariate. Its main drawback is that stacking PITs values is subject to ordering of PITs and it neglects any dependence among the t value of $Z_{h=3|h=2,t}$ with the t value of $Z_{h=2|h=1,t}$ and $Z_{h=1|t}$, which are all dependent to starting value t. The effect of this dependence on uniformity tests is discussed in Appendix (3.7.3).

B. **The Clements and Smith [2000] Approach**

Clements and Smith [2000] proposed a new factorization of joint density forecasts

that consists in the product of the PITs vectors in Diebold et al. [1999]. This joint evaluation has been proposed to allow the correlation among variables to be non-linear. For each t, the $N = T$ dimensional vector of PITs to test for uniformity is:

$$Z_{CSp} = \left[Z_{h=3|h=2,t} \times Z_{h=2|h=1,t} \times Z_{h=1|t} \right] \tag{3.16}$$

Assessing the calibration of the path forecasts consists in testing the uniformity of the Z_{CSp} vector of PITs. The product is shown to be i.i.d. $U(0,1)$ sequences under the null hypothesis. The testing set up proposed by Clements and Smith [2002], crucial in the multi-variable setting, is not as important in the multi-horizon case. The reason has to be found in the lack of contemporaneous correlation of variables when forecasting multi-step-ahead. To see this, consider the multi-horizon setting with iterative forecasts using an AR(1) model:

$$y_t = \beta y_{t-1} + u_t \tag{3.17}$$

where $u_t \sim N(0, \sigma^2)$.

Then the $h = 1$ and $h = 2$ period ahead marginal density indirect forecasts are:

$$f(y_{t+1|t}) = N(\beta y_t, \sigma^2) \tag{3.18}$$

$$f(y_{t+2|t}) = N(\beta^2 y_t, (1+\beta^2)\sigma^2) \tag{3.19}$$

and the conditional forecast is:

$$f(y_{t+2|t+1}) = N(\beta y_{t+1}, \sigma^2) \tag{3.20}$$

So the inverse normal CDF, Φ, transformed PITS (i.e. the standardized forecast errors) are:

$$z^*_{t+1|t} = \frac{u_{t+1}}{\sigma} \tag{3.21}$$

$$z^*_{t+2|t} = \frac{\beta u_{t+1} + u_{t+2}}{\sqrt{(1+\beta^2)}\sigma} \tag{3.22}$$

$$z^*_{t+2|t+1} = \frac{u_{t+2}}{\sigma} \tag{3.23}$$

There is no issue (unlike below) of contemporaneous correlation affecting the power of the tests. Although, there may be a question of whether tests of calibration are more powerful than the Diebold et al. [1999] case.

In the multivariate case, Diebold et al. [1999] is misspecified. To see this, consider the VAR(0) DGP:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N\left(\begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right) \tag{3.24}$$

Then the marginal density forecasts are

$$f_t(y_1) = N(m_1, \sigma_1^2) \tag{3.25}$$

$$f_t(y_2) = N(m_2, \sigma_2^2) \tag{3.26}$$

and the conditional forecast is

$$f_t(y_{2|1}) = N(\rho\frac{\sigma_2}{\sigma_1}m_1, (1-\rho^2)\sigma_2^2) \tag{3.27}$$

So the inverse normal CDF, Φ, transformed PITS are

$$z_1^* = \frac{(y_1 - m_1)}{\sigma_1} \tag{3.28}$$

$$z_2^* = \frac{(y_2 - m_2)}{\sigma_2} \tag{3.29}$$

$$z_{2|1}^* = \frac{\left(y_2 - \rho\frac{\sigma_2}{\sigma_1}m_1\right)}{\sqrt{(1-\rho^2)}\sigma_2} \tag{3.30}$$

So there is now the problem, when $m_1 = m_2 = 0$, that if we (incorrectly) assume independence $\rho = 0$, both z_1^* and $z_{2|1}^*$ are still distributed standard normal - so calibration failure (in terms of ρ) is undetected. But z_1^* and $z_{2|1}^*$ are not independent (when it is incorrectly assumed that $\rho = 0$), hence the tests of Clements and Smith, and Ko and Park.

Multi-horizon: direct forecasting

Now consider forming the forecasts via direct forecasting

$$y_t = \beta_1 y_{t-1} + u_{1t} \tag{3.31}$$

$$y_t = \beta_2 y_{t-2} + u_{2t} \tag{3.32}$$

where:

$$\begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right) \tag{3.33}$$

Recall, direct forecasts have been shown to be more robust to misspecifications (Bhansali, Marcellino et al, Pesaran et al.).

Then the standardized PITS are

$$z^*_{t+1|t} = \frac{u_{1t+1}}{\sigma_1} \tag{3.34}$$

$$z^*_{t+2|t} = \frac{u_{2t+2}}{\sigma_2} \tag{3.35}$$

$$z^*_{2|1} = \frac{\left(y_{t+2} - \rho\frac{\sigma_2}{\sigma_1}\beta_1 y_t\right)}{\sqrt{(1-\rho^2)}\sigma_2} \tag{3.36}$$

In this framework, we can see that parameter ρ is unknown. While in a indirect path forecast, ρ is a transformation of model parameters (depending on the model), in this case we do not know ρ. In parallel, for the case of Bank of England Fancharts, all we observe is u_{1t+1} and u_{2t+1}.

Please refer to Appendix (3.7.2) for the transformation for $H > 2$ case following Dovern and Manner [2016] online appendix.

C. **The Ko and Park [2013]Approach**

In the multivariate setup, Ko and Park [2013] identifies an asymmetric behaviour in the power of Clements and Smith [2000]. To overcome this lack of power, Ko and Park [2013] proposed a location-adjusted transformation. Such as $\{Z_{i|j,t} \times Z_{j,t}\}$ of Clements and Smith [2002] becomes $\{(Z_{i|j,t} - \mathbb{E}(Z_{i|j,t})) \times (Z_{j,t} - \mathbb{E}(Z_{i,t})\}$. Hence, the vector of PITs to test for uniformity becomes:

$$Z_{KP} = \begin{bmatrix} (Z_{h=1|t} - \mathbb{E}(Z_{h=1|t})) \times (Z_{h=2|h=1,t} - \mathbb{E}(Z_{h=2|h=1,t}) \times \dots \\ \times (Z_{h=3|h=2,t}) - \mathbb{E}(Z_{h=3|h=2,t})) \end{bmatrix} \tag{3.37}$$

for any $t = 1, \dots, T$. Testing this sequence of PITs to be uniformly distributed in $(0, 1)$ results to have better empirical power than the previous tests and to be free from asymmetries in the multivariate case, and we consider it here as a variation of Clements and Smith [2000] approach. Please refer to Appendix (3.7.2) for the transformation for $H > 2$ case following Dovern and Manner [2016] online appendix.

A discussion on the horizon-dependence issue and how it affects these four approaches differently can be found in Appendix (3.7.3).

The next section will discuss in details the types of uniformity test performed in this paper.

Uniformity Tests for PITs

Once the statistics has been defined assessing the well-calibration of the path forecasts consists in testing the uniformity of the PITs. A series of uniformity test is available in literature.

- **Uniformity Test of Probability integral Transform.** The tests that belong to this category tests the null hypothesis:

$$H_0 : \boldsymbol{z} \sim U(0,1). \tag{3.38}$$

Two uniformity of the PITs tests are the Kolmogorov-Smirnov (κ) and Cramer-von Mises (C) tests. That have test statistics:

$$\kappa = \underset{r\in[0,1]}{sup} |\Psi(r)| \tag{3.39}$$

$$C = \int_0^1 \Psi(r)^2 dr \tag{3.40}$$

where $\Psi(r)$ is the empirical process:

$$\Psi(r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} I\{\boldsymbol{z} \leq r\} - r \tag{3.41}$$

and $r \in [0,1]$ is the vector of PITs under null hypothesis of calibration (i.e. uniformly distributed). The test reject H_0 at the $\alpha * 100\%$ significance level if $\kappa > \kappa_\alpha$ and $C > C_\alpha$. Usually critical values are derived by analytical calculation (see Durbin [1973] and Smirnov [1948]). However, this statistic ignores the serially correlated PITs discussed as the first type of dependence in Appendix (3.7.3). As noted by Rossi and Sekhposyan [2019], the critical values depend on the nuisance parameters that appear in the covariance matrix of the PITs. Rossi and Sekhposyan [2019] continues recommending to use critical values from a block version of the weighting bootstrap by Inoue

[2001] described in Rossi and Sekhposyan [2019]. We apply the evaluation using both sets of critical values.

- **Normality tests of Inverse Normal transformation (INT).** The tests that belong to this category tests the null hypothesis:

$$H_0 : \Phi^{-1}(\boldsymbol{z}) \sim \mathcal{N}(0,1). \tag{3.42}$$

The test is a likelihood ratio (LR) test proposed by Berkowitz [2001] can be applied to INT to test simultaneously for zero mean, unit variance and zero autocorrelation. Since, we are in presence of serial correlation, we will use a version of the test where only the first two hypotheses are tested. The LR joint test that pits values have mean and variance equal to $(0, 1)$ is:

$$LR_B = -2(L(0,1) - L(\hat{\mu}, \hat{\sigma}^2)) \tag{3.43}$$

Under the null hypothesis, the test statistic is distributed $\chi^2(2)$, chi-squares with 2 degrees of freedom. The exact log-likelihood function associated is:

$$L(\mu, \sigma^2) = -\frac{T}{2} log(2\pi) - \frac{T}{2} log(\sigma^2) - \sum_{t=1}^{T} \frac{(z_t - \mu)^2}{2\sigma^2} \tag{3.44}$$

According to Berkowitz [2001], the advantage of tests based on the inverse normal transformation of the PITs is that they are more powerful than tests of uniformity applied directly to the PITs, at least in small samples; the limitation is that they detect violations of normality only through the first two, and not higher, moments, whereas PIT-based tests can detect any departure from uniformity. This variant of the test has been used in several application among which Mitchell and Hall [2005], Clements [2004] and Jore et al. [2010b]. A variation of Berkowitz [2001] that are valid also for serial correlation is proposed by Bai and Ng [2005].

- **Raw-Moments Tests** A raw-moments test for the calibration of multi-step ahead density forecasts are proposed by Knüppel [2015]. The raw-moments tests are based on the standardised PITs (S-PITs):

$$\text{S-PITs} = \sqrt{12}\left(z_t - \frac{1}{2}\right) \tag{3.45}$$

where Z_t is the vector of PITs. Under the null of probabilistic calibration:

$$H_0 = \text{S-PITs} \sim (-\sqrt{3}, \sqrt{3}) \tag{3.46}$$

Let the n^{th} raw moment of S-PITs be:

$$m_n = \mathbb{E}[\text{S-PIT}^r] \tag{3.47}$$

with $n \in \mathcal{N}^+$. Let us denote that the vector of N empirical raw-moments of interest as $[\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_N]$ and the vector of moments under null hypothesis as $[m_1, m_2, \ldots, m_N]$. Than the vector $\hat{\boldsymbol{D}}_{1,2,\ldots,N}$ denoting the difference between both vectors, is given by:

$$\hat{D}_{1,2,\ldots,N} = \begin{bmatrix} \hat{m}_1 - m_1 \\ \hat{m}_2 - m_2 \\ \vdots \\ \hat{m}_N - m_N \end{bmatrix} \tag{3.48}$$

which converges to a multivariate normal distribution:

$$\sqrt{T}\hat{D}_{1,2,\ldots,N} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_{1,2,\ldots,N}) \tag{3.49}$$

where T is the sample size and $\boldsymbol{\Omega}_{1,2,\ldots,N}$ the long-run covariance matrix of the vector $\hat{D}_{1,2,\ldots,N}$. A test statistic for the vector of S-PITs can then be written as:

$$\hat{\alpha}_{1,2,\ldots,N} = T\hat{\boldsymbol{D}}'_{1,2,\ldots,N}\hat{\boldsymbol{\Omega}}^{-1}_{1,2,\ldots,R}\hat{\boldsymbol{D}}_{1,2,\ldots,N} \tag{3.50}$$

which under the null $\hat{\alpha}_{1,2,\ldots,N} \to \chi^2(N)$.

Knüppel [2015] test has the advantages of having a higher power when the misspecification affects moments included in the test, however, it's not clear how big the misspecification has to be in other to be preferred to κ or C tests. Some previous works justify the use of κ and C since they tests the entire distribution and not just some moments. In the multi-step-ahead case, Knüppel [2015] has a HAC estimator which is consistently estimated, while in κ or C the covariance matrix is large and a block bootstrap procedure to estimate critical values is suggested. In this paper we specified Knüppel [2015] test on $n = 4$ moments.

3.3.2 Tests on PITs of Marginal Distributions

Often, the decomposition strategy proposed in the previous section is not available. One reason is that the information regarding the conditional distribution is not disclosed (as in the case of Bank of England's fan charts). Another reason is that different models are used to predict specific horizon, and the conditioning losses its meaning. This section proposes an alternative strategy to evaluate path density calibration when only horizon-specific density distributions are available.

Stacked Pits of Marginal Distributions

Intuitively, the first example of calibration PITs on marginal distribution consists in evaluating stacked PITs of forecasts horizon-by-horizon. In this case, the joint distribution is:

$$F_t(y_{t+1}, y_{t+2}, y_{t+3}) = F_t(y_{t+1})F_t(y_{t+2})F_t(y_{t+3}) \tag{3.51}$$

and the feasible vector to test for uniformity is then:

$$Z_M = \left[Z_{h=1|t}; Z_{h=2|t}; Z_{h=3|t}\right] \tag{3.52}$$

The beauty of this approach is that not knowing the nature of $C(\cdot)$ in Equation (3.1), we ignore it. However, we must remember that Equation (3.51) has an approximation error.

It is critical then questioning how much of the power and size we lose if one does not acknowledge some degree of dependency, as our practice to evaluate paths does. This section proposes several simulation exercises to investigate how much the missing time-dependence matters for the size and power of calibration tests.

"Sup-tests"

The calibration test proposed here is a direct translation from the relative evaluation criterion of superior predictive ability (SPA), which entitles a long tradition in the literature. Among the most recent works, Quaedvlieg [2021] generalises the Diebold and Mariano [2002] test to a multi-horizon framework. The accuracy of two or more multi-horizon forecasts is compared at each horizon, leading to incoherent results when one forecast do not perform better than the others in all the horizons. For this reason,

they propose several tests to help the researcher to elicit the best model for the multi-horizon forecast. Quaedvlieg [2021] introduces two definitions for superior predictive ability: uniform and average. The concepts of uniform and average SPA link to first- and second-order forecast dominance, respectively, and stochastic dominance (e.g., Linton et al. [2005]; Linton et al. [2010]). More generally, the tests are closely related to multivariate inequality tests (e.g., Bartholomew [1961]; Wolak [1987]). Patton and Timmermann [2010] proposed a solution similar in the context of monotonicity in asset pricing relationships.

This paper contributes to this literature by applying the superior predictive ability to absolute evaluation. Following Rossi and Sekhposyan [2019], we assume that the researcher has divided the available sample of size $T + h$ into an in-sample portion of size R and an out-of-sample portion of size P. The null hypothesis of path calibration is:

$$\begin{aligned} &H_0 : F_t(y_{t+1}, \ldots, y_{t+H} \leq u|\mathcal{I}_t) = F_0(u|\mathcal{I}_t) \\ &H_1 : \text{the negation of } H_0 \end{aligned} \tag{3.53}$$

where $F_t(y_{t+h}|\mathcal{I}_t) = \int_{-\infty}^{y_{t+h}} f_t(u|\mathcal{I}_t)$, $\mathcal{I}_t$ is the information set available at time t, y_{t+h} is the realisation at $t + h$ of the random variable u.

For each horizon, we can state the empirical PITs as:

$$\begin{aligned} &\Pr(F_t(y|\mathcal{I}_t) \leq r) \\ &= \Pr\left(\int_{-\infty}^{y_{t+h}} f_t(y|\mathcal{I}_t)dy \leq r \right) \\ \Psi_{P,h}(y_{t+h}) &\equiv P^{-1/2} \sum_{t=R}^{T} (1\{F_{t+h}(y_{t+h}) \leq r\} - r) \end{aligned} \tag{3.54}$$

and $r \in [0, 1]$.

Two statistics are employed to test the null: a "sup" statistic on the maximum $\Psi_{P,h}(y_{t+h})$ over h and a "sup" statistic on the average of $\Psi_{P,h}(y_{t+h})$ over h. Let us define the statistic:

- **Max Sup statistic:** A path density forecast is strictly calibrated if each horizon is calibrated. The test statistics are the following:

$$\kappa_{P,s} = \max_{h \in \{1,\ldots,H\}} \left(\sup_{r \in [0,1]} |\Psi_{P,h}(y_{t+h})| \right) \quad C_{P,s} = \max_{h \in \{1,\ldots,H\}} \left(\int_0^1 \Psi_{P,h}(y_{t+h})^2 dy_{t+h} \right) \tag{3.55}$$

The test reject H_0 at the $\alpha * 100\%$ significance level if $\kappa_{P,h} > \kappa_\alpha$ and $C_{P,h} > C_\alpha$.

According to this test, the null hypothesis of path calibration is rejected when at least the pits of one horizon are not uniform.

- **Weighted Sup statistic**. This second definition is more lenient than the previous; the path density forecast is allowed to be not calibrated in some horizons, although it is calibrated "on average". According to this test, the null hypothesis of path calibration is rejected when the averaged pits over horizons are not uniform. W denotes the vector of weights that multiplies PITs for each horizon. The weights are assigned based on a specific preference of the researcher. For example, one can assume the same calibration level at each horizon (i.e. $W = H$) or allow some degree of misspecification on a longer horizon without making the test reject the path calibration (i.e. descending weights). The test statistics available for this test are:

$$\begin{aligned} \kappa_{P,a} &= \max_{h\in\{1,\dots,H\}} \left(\sup_{r\in[0,1]} |W_{h\in\{1,\dots,H\}} \times \Psi_{P,h}(y_{t+h})| \right) \\ C_{P,a} &= \max_{h\in\{1,\dots,H\}} \left(\int_0^1 W_{h\in\{1,\dots,H\}} \times \Psi_{P,h}(y_{t+h})^2 dy_{t+h} \right) \end{aligned} \tag{3.56}$$

The test reject H_0 at the $\alpha * 100\%$ significance level if $\kappa_{P,h} > \kappa_\alpha$ and $C_{P,h} > C_\alpha$.

Bootstrap Critical Values

Critical values for Equations (3.55, 3.56) can be based on the following bootstrap procedure. Let l be the block length and η_t be a continuous random variable that is used for random weighting in the block weighted bootstrap. $\{\eta_t^j\}_{t+R}^{T-l+1}$ are independent random variables, independent of z_t, with zero mean, variance $1/l$ and $\mathbb{E}(\eta_i^4) = O(1/l^2)$, where $l \to \infty$ as $T \to \infty$ and $l = o(P^{1/2})$. The bootstrap can be implemented in practice using the following step-by-step procedure:

A. Construct the test statistics κ_P and C_P for each horizon following (3.55, 3.56).

B. For each horizon let J be the maximum number of bootstrap replications. For $j = 1, 2, \dots, J$, generate $\{\kappa^*_{P,j,h}\}_{j=1}^J$ and $\{C^*_{P,j,h}\}_{j=1}^J$ where κ^* and C^* are based on draws$\{\eta_t^j\}_{t+R}^{T-l+1}$;

C. For each horizon estimate the level-α critical values $\hat{c}^J_{\kappa,\alpha,h}$ and $\hat{c}^J_{C,\alpha,h}$ by choosing α-100 percentiles from $\{\kappa^*_{P,j,h}\}_{j=1}^J\}$ and $\{C^*_{P,j,h}\}_{j=1}^J\}$, respectively.

D. Compute the path critical values according to:

- $\hat{c}^J_{\kappa,\alpha} = \max_{h\in\{1,\dots,H\}} \hat{c}^J_{\kappa,\alpha,h}$ and $\hat{c}^J_{C,\alpha,h} = \max_{h\in\{1,\dots,H\}} \hat{c}^J_{C,\alpha,h}$ (for 3.55);
- $\hat{c}^J_{\kappa,\alpha} = \frac{1}{W}\hat{c}^J_{\kappa,\alpha,h}$ and $\hat{c}^J_{C,\alpha,h} = \frac{1}{W}\hat{c}^J_{C,\alpha,h}$ for(3.56 weighting each horizon according to a vector of weight W arbitrary chosen.)

E. Reject H_0 at the $\alpha \cdot 100\%$ significance level if $\{\kappa_P > \hat{c}^J\}_{\kappa,\alpha}$ and $C_P > \hat{c}^J_{C,\alpha}$

In the following section we will present some Monte Carlo Simulations where the different strategies of evaluation will be compared.

3.4 Monte Carlo Simulations

This section shows some Monte Carlo simulations to study the size and power of test statistics to evaluate the path density forecast presented previously. Throughout this simulation, we will investigate the size and power of test statistics of both approaches to evaluation: the first using marginal and conditional distributions in Sections (3.4.1) and (3.4.2) and the second using marginal distributions only in Sections (3.4.3) and (3.4.4).

3.4.1 Size Experiments for tests on PITs of Marginal and Conditional distributions

In this series of simulations we want to investigate the impact of sample size, degree of temporal dependence on the multiplicity of calibration tests statistics.

We generate the path density forecasts for $H = 4$ horizons from an $AR(4)$ model:

$$DGP : y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \phi_4 y_{t-4} + \varepsilon_t \tag{3.57}$$

for $T = [25, 50, 100, 500, 1000]$ where ε_t are i.i.d. normal $(0, \sigma^2)$. Several sets of autoregressive coefficients are used to consider different ways of persistence $\boldsymbol{\phi} = \{0.3, 0.2, 0.1, 0.1\}$, (Tables 3.4, 3.5, 3.6) $\boldsymbol{\phi} = \{0.1, 0.1, 0.1, 0.1\}$ (Tables 3.7, 3.8, 3.9). We also reduce the model to a AR(1) imposing the autoregressive coefficient to be $\boldsymbol{\phi} = \{0, 0, 0, 0\}$, equal to uniform distribution, and $\boldsymbol{\phi} = \{0.1, 0, 0, 0\}$,$\boldsymbol{\phi} = \{0.5, 0, 0, 0\}$ and $\boldsymbol{\phi} = \{0.9, 0, 0, 0\}$ to test the effect of increasing in autocorrelation on size of the tests (Tables 3.1, 3.2, 3.3). In this experiment there is no parameter uncertainty. Density forecasts for horizons $h = 1, \dots, 4$ are jointly distributed according to:

$$\begin{bmatrix} y_{t+1} \\ y_{t+2} \\ \vdots \\ y_{t+4} \end{bmatrix} \sim N \begin{bmatrix} \boldsymbol{\phi} \boldsymbol{y}_t \\ \boldsymbol{\phi}^2 \boldsymbol{y}_t \\ \vdots \\ \boldsymbol{\phi}^4 \boldsymbol{y}_t \end{bmatrix}, \begin{bmatrix} \sigma^2 & \phi_1 \sigma^2 & \dots & \phi_1^3 \sigma^2 \\ \phi_1 \sigma^2 & (1+\phi_1^2)\sigma^2 & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1^3 \sigma^2 & \dots & \dots & \sigma^2 \sum_{j=0}^{3} \phi_j^2 \end{bmatrix} \tag{3.58}$$

where $\boldsymbol{\phi} = [\phi_1 \phi_2 \phi_3 \phi_4]$ and $\boldsymbol{y}_t = [y_t, y_{t-1}, y_{t-2}, \dots, y_1]$.

We calculate Kolmogorov-Smirnov, Cramer-von Mises, Berkowitz [2001] and Knüppel [2015] statistics for the vectors of PITs z_{DHT}, z_{CS}, z_{KP} in equations (3.15, 3.16, and 3.37)) and z_M which contains marginal distributions only. We repeat the simulation 1000 times and report the sizes of tests in Tables (3.1-3.9).
Tables (3.1, 3.4 and 3.7) show empirical rejection frequencies for each marginal distribution $\{z_{h=1|t}, z_{h=2|t}, z_{h=3|t}, z_{h=4|t}\}$, while Tables (3.2, 3.5 and 3.8) show empirical rejection frequencies for each conditional distribution $\{z_{h=2|t+1}, z_{h=3|t+2}, z_{h=4|t+3}\}$. Results for path PITs $z_M, z_{DHT}, z_{CS}, z_{KP}$ are reported in tables (3.3, 3.6 and 3.9).

From the results, we can draw a series of comments on the size properties of the test statistics:

A. We see that, as expected, the size of the tests improves with the increase in sample size.

B. Comparing the size properties of the statistics at the variation of ϕ, we can answer whether the level of horizon dependence affects the evaluation since, in indirect forecasts, the time dependence is given by autoregressive parameters. We note that ϕ matters for small samples, while in larger samples, the simulation provides equivalent size for the statistics.

C. We can discuss the performance of the different statistics: bootstrap version of Kolmogoroz-Smirnov and Cramer-von Mises are correctly sized, Berkowitz does not work well, or it is undersized, Knuppel test needs a large sample size (i.e. at least $T = 100$) to be correctly sized.

D. The performance of vectors of pits changes across tests and T considered, but z_M and z_{DHT} performs overall better than the alternatives.

3.4.2 Power Experiments for tests on PITs of Marginal and Conditional distributions

We consider three types of misspecifications to evaluate the power properties of the tests of calibration of path density forecasts.

A. First, the DGP follows an iid normal distribution, and the path density forecast is built under the hypothesis that DGP follows an AR(4) model. Forecast model parameters are estimated through OLS. Results of tests for each horizon forecast (both marginal and conditional distributions) and entire path are displayed in Table (3.10). All the tests display a good power even at a lower sample size (i.e. $T = 25$).

B. Second, the DGP follows an AR(4) process, while the path density forecast follows an AR(1) model. The DGP has richer dynamics than the model used to forecast the $H = 4$ steps. The AR(4) model used to simulate the DGP has parameters $\beta = [0.3, 0.2, 0.1, 0.1]$ while forecast model parameters are estimated through OLS. Please refer to Table (3.11) for result of tests. From this, we can conclude that the power of the tests increases with sample size and that the Knuppel test is the most powerful of the tests considered in this paper.

C. Third, following Knüppel [2015], the DGP follows a MA(1) process while the forecast model is an AR(1) process. The persistence of AR(1) model is fixed at $\beta = 0$, $\beta = 0.5$ or $\beta = 0.9$. Please refer to the result about calibration of the single-horizon forecasts (marginals) in Table (3.12), conditionals in Table (3.13) and the entire path in Table (3.14). Note that for $\beta = 0$, the forecasts are correctly specified, and it should be considered as the size of the tests. In this case, all tests (excluding Berkowitz) have good power at a large sample size (i.e. $T = 500$ or higher). In addition, the higher the β coefficient, the higher the probability of rejecting the null hypothesis of calibration. We find these patterns also in the calibration of the vector of pits, although z_M and z_{DHL} also have a rejection probability closer to nominal level at $\beta = 0$ than alternative methods.

In conclusion we can state that the test statistics of z_M and z_{DHT} are, across simulations, the ones with better size and power.

3.4.3 Size Experiments of tests on PITs of Marginal distributions

In this series of simulation we want to investigate the impact of sample size, on the multiplicity of calibration tests statistics. To investigate the size properties of our "sup

tests" we consider forecasts are based on model parameters estimated in rolling windows for $t = R, ..., T$. We consider several values for in-sample estimation window of $R = [25, 50, 100, 200, 500, 1000]$ and out-of-sample evaluation period $P = [25, 50, 100, 200, 500, 1000]$ to evaluate the performance of the proposed procedure. While our assumptions require R to be finite, we consider both small and large values of R to investigate the robustness of our methodology when R is large. Following Rossi and Sekhposyan [2019], the DGP is a integrated moving average (IMA) model:

DGP IMA: $\Delta t_t = \mu_t + \epsilon_t - \rho\epsilon_{t-1}$, $\epsilon_t \sim i.i.d.N(0, 1.261)$, $\rho = 0.275$ and μ_t is defined as:

$$\mu_t = R^{-1} \sum_{j=t_R}^{t-1} \Delta y_j. \tag{3.59}$$

The parameters for the monte carlo design are from Stock and Watson [2007] 1960:I-1983:IV sample period (i.e. great inflation period). Forecast model: $\Delta y_t = \beta + e_t$, $e_t \sim i.i.d.N(0, 1 + \rho^2)$ with multi-step ahead forecasting model equal to:

$$\Delta y_t = \mu_t + \epsilon_t - \sum_{j=1}^{h} \rho^j \epsilon_{t-1} \tag{3.60}$$

where $h = 1, \ldots, H$ are the forecasting horizons. We will analyse the size of the tests for $H = 4$ and $H = 12$. Tables (3.15 and 3.16) refer to the evaluation of path density forecasts with $H = 4$ horizons. The empirical p-values of the test using bootstrap κ and C have reasonable size. Table (3.16) displays the empirical rejection probabilities for the sup tests. The tests for path calibration using max sup statistic in Equation (3.55) are correctly sized. We adopt two versions of weighted sup statistic: one that simply applies a equal weighting scheme on the statistics for each horizon (i.e. $[0.25, 0.25, 0.25, 0.25]$) and a descending weighting scheme, that applies a higher weight on shorter horizon than longer horizons (i.e. $[0.40, 0.30, 0.20, 0.10]$). The tests for both definition are slightly undersized. Tables (3.20 and 3.21) refer to the evaluation of path density forecasts with $H = 12$ horizons. The "sup tests" with $H = 12$ horizons displayed a better size, both for strict calibration and averaged calibration, stating that they are more accurate with increasing the number of horizons.

3.4.4 Power Experiments of tests on PITs of Marginal distributions

Three misspecifications are considered in power exercises: a misspecification in mean and error variance:

A. Power exercise 1: $\mu_t + 2$. This first power exercise is simple. It assumes that the forecasters has a biased mean of the path density forecast. We can see from Table (3.17) that the calibration tests have a good power, already at in-sample period of $R = 100$. Here as well, the tests for weighted sup statistic have lower power than the max sup but only for short in-sample periods.

B. Power Exercise 2 assumes the variance to be lower than the DGP, $\epsilon_t \sim i.i.d.N(0, 1)$. Interestingly from Tables (3.18), we can see that the tests need a larger sample size to reject significantly the calibration.

C. Power Exercise 3: assumes the variance to be higher than the DGP, $\epsilon_t \sim i.i.d.N(0, 1.261 * 2))$. Compared to the previous case, here the tests have higher power (results are displayed in Table 3.19).

In conclusion, this section wanted to highlight the properties of the tests by simulated data. To summarise, the sample size matters for the performance of the tests; the degree of horizon dependence does not affect the properties of the tests; among the vectors of pits, the better sized and more powerful vectors are z_M and z_{DHT}; among the sup tests, the max sup statistic displays the best properties. The choice of statistics depends on several aspects: Kolmogorov-Smirnov and Cramer-von Mise have better size and power when employing the bootstrapped version; choosing these two tests will depend on the computational issues that the estimation entitles. Berkowitz test is undersized, and the Knuppel test works only with a large enough (i.e. T=100) sample size, but it is the most powerful.

3.5 Empirical Applications

In this section the path density evaluation examined in theory and through simulation will be apply to Bank of England fan charts. As already mentioned, Bank of England publishes, at a quarterly frequency, one-year inflation forecasts for $h = 1, \ldots, H = 13$ steps ahead. The dataset for fan charts spans from 2004Q1 up to 2020Q1. The charts are based on a number of conditioning assumption that reflects the deliberations of the Mon-

etary Policy Committee (MPC) that are collected the document called "Conditioning assumption, MPC key judgements and indicative projection" document.

Analytically, fan charts have a two-piece normal distribution, for details see Wallis [2004]. The Bank of England publishes a set of parameters that describes the distribution: mode (mo), mean (m), median (me), uncertainty (un) and skewness (sk). The uncertainty is a parameter of the two-piece normal distribution. The skew statistic is defined as mean minus mode. Two-piece normal probability density distribution takes the form:

$$f_t(y_t) = \begin{cases} \frac{2}{\sqrt{2\pi}(\sigma_1+\sigma_2)} \exp\Big(- \frac{(y_t - mo)^2}{2\sigma_1^2} \Big), & \text{if } y_t \leq mo \\ \frac{2}{\sqrt{2\pi}(\sigma_1+\sigma_2)} \exp\Big(- \frac{(y_t - mo)^2}{2\sigma_2^2} \Big), & \text{if } y_t > mo \end{cases} \tag{3.61}$$

with mo being the mode parameter and σ_1 and σ_2 are obtained by:

$$\sigma_1 = \sqrt{un^2/(1+\gamma)} \quad \sigma_2 = \sqrt{un^2/(1-\gamma)} \tag{3.62}$$

γ is defined by:

$$\begin{cases} \gamma = \sqrt{\gamma_2} & \text{if } s \geq 0 \\ \gamma = -\sqrt{\gamma_2} & \text{if } s < 0 \end{cases} \tag{3.63}$$

where $s = \frac{m - mo}{un}$ and $\gamma_2 = 1 - 4(\frac{\sqrt{(1+\pi s^2)} - 1}{\pi(\frac{m-mo}{un})^2})^2$. Then, according to Clements [2004], we can compute the pit values as follows:

$$P(Y \leq y) = \begin{cases} \frac{2}{\sqrt{2\pi}(\sigma_1+\sigma_2)} \exp\Big(- \frac{(y_t - m)^2}{2\sigma_1^2} \Big), & \text{if } y_t \leq m \\ \frac{\sigma_1 - \sigma_2}{\sigma_1+\sigma_2} + \frac{2}{\sqrt{2\pi}(\sigma_1+\sigma_2)} \exp\Big(- \frac{(y_t - m)^2}{2\sigma_2^2} \Big), & \text{if } y_t > m \end{cases} \tag{3.64}$$

PIT values for $t + h$ for $h = 1, \ldots, H = 13$ steps called $Z_{BOE,t}$ ahead are tested for uniformity. The vector of PIT values is obtained evaluating the calibration at each forecast origin $t = 1, \ldots, T = 52$. The forecast horizons span from 2004Q1 to 2020Q1. The dataset is reduced since we need realization for inflation in order to compute the PITs; the dataset from $T = 65$ is then reduce to $T = 52$.

With the information published by Bank of England, we can assess the joint calibration only using the statistics based on PITs of marginal distributions discussed in Section (3.3.2). The other joint tests are not feasible since the information about horizon dependence is not disclosed, and the estimation of conditional distributions is impossible.

However we can approximate the covariance between horizons δ_h using a GMM approach. This approach consists in estimating the correlation among horizons as the increment in the variance of density forecast at horizon $h > 1$ compared to previous horizon. For each forecast origin and horizon, the variance of a two-piece normal distribution can be derived from parameters σ_1 and σ_2 in equation (3.62):

$$\sigma_{BOE,t,h} = (1 - \frac{2}{\pi})(\sigma_{2,t,h} - \sigma_{1,t,h})^2 + \sigma_{1,t,h}\sigma_{2,t,h} \tag{3.65}$$

and for each horizon, the variance can be estimated according to its sample mean:

$$\sigma^2_{BOE,h=1} = \frac{1}{T}\sum_{t=1}^{T} \hat{\sigma}_{BOE,t} \tag{3.66}$$

Hence, the covariances among horizons can be approximates as:

$$\begin{aligned} cov(y_{t+1}, y_{t+1}) &= \sigma^2_{BOE,h=1} \\ cov(y_{t+2}, y_{t+1}) &= \sqrt{\delta_2}\sigma^2_{BOE,h=1} \\ cov(y_{t+3}, y_{t+2}) &= \sqrt{\delta_3}\sigma^2_{BOE,h=1} \\ &\vdots \\ cov(y_{t+H}, y_{t+H-1}) &= \sqrt{\delta_H}\sigma^2_{BOE,h=1} \end{aligned} \tag{3.67}$$

Again, this is an approximation of the correlation among horizons because we do not know which process defines the temporal dependence of fan charts when these are built by the Bank. However, with this approach we allow the dependence to change at each horizon. The estimation of δ_h is carried out as follows:

$$\begin{aligned} \delta_2 &= \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\hat{\sigma}_{t,h=2} - \hat{\sigma}_{t,h=1}}{\hat{\sigma}_{t,h=1}}\right] \\ \delta_3 &= \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\hat{\sigma}_{t,h=3} - \hat{\sigma}_{t,h=2}}{\hat{\sigma}_{t,h=1}}\right] \\ &\vdots \\ \delta_H &= \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\hat{\sigma}_{t,h=H} - \hat{\sigma}_{t,h=H-1}}{\hat{\sigma}_{t,h=1}}\right] \end{aligned} \tag{3.68}$$

Using these correlation between horizons, we can then approximate the conditional dis-

tributions for horizons $h = 2, \ldots, H$ given the previous one, by:

$$\begin{aligned} f(y_{h=2|h=1}) &= \mathcal{N}(y_{t+1}, \sqrt{\delta_2}\sigma_{t,h=1}) \\ f(y_{h=3|h=2}) &= \mathcal{N}(y_{t+2}, \sqrt{\delta_3}\sigma_{t,h=1}) \\ &\vdots \\ f(y_{h=H|h=H-1}) &= \mathcal{N}(y_{t+H-1}, \sqrt{\delta_H}\sigma_{t,h=1}) \end{aligned} \tag{3.69}$$

This paper evaluates three fan charts published by the Bank of England: inflation rate, GDP growth rate and unemployment rate. For each variable, we organised the results for calibration tests as follow: tests for calibration of marginal distributions (respectively Table 3.22, 3.26 and 3.30), their PITs histograms in Figures (3.1, 3.2, 3.3), the correlation coefficients we estimated between horizons in tables (3.23, 3.27 and 3.31), tests for calibration of conditional distributions (respectively Table 3.24, 3.28 and 3.32), and finally path density tests (both vectorised and sup max) in Table (3.25, 3.29 and 3.33).

For Kolmogorov-Smirnov, Cramer-von-Mises tests, Berkowitz and Knuppel, the results display the value of statistic in the first column and the critical values at 5%. According to all tests, inflation rate fan charts are calibrated at each horizon, except for Berkowitz (evidence also found in the previous section). The calibration of path density forecasts is verified according to the unfeasible approach z_{DHT} and all the feasible ones (z_M, and all the version of sup tests).

The results are different for the GDP growth fan charts. All test rejects the null hypothesis of calibration at least for horizons $h = \{1, \ldots, 7\}$, while conditional distributions are overall calibrated. Consequently, the joint test for calibration is rejected by all "sup" tests, z_{CS} and z_{KP}. The only tests that cannot reject the null hypothesis are z_M and z_{DHT} using Kolmogorov and Cramer-von-Mises statistics.

Overall, the Bank of England path density forecast for the unemployment rate is not calibrated. The null hypothesis of calibration is rejected for both single-horizon and the path density forecast.

Given the size of the time series, this application exercise is comparable with the sizes exercises for T=50 or (P=25, R=25). We do not consider the Knuppel test results since it needs a larger sample size (at least double), as shown in the Monte Carlo exercises.

3.6 Conclusions

This paper proposes an absolute evaluation criterion for path density forecasts. After the path density forecast has been defined, a series of testing strategies have been discussed. We identified two main tests that depend on information about horizon dependence. If the researcher has information about the dependence and can build conditional forecast distributions, then they can use vectors of PITs, among which the better sized and more powerful vectors are z_M and z_{DHT}; If the research does not have any information, they can use a vector of marginal distributions and some "sup tests". Among the "sup tests", the max sup statistic displays the best properties.

The choice of test statistics depends on several aspects: Kolmogorov-Smirnov and Cramer-von Mise have better size and power when employing the bootstrapped version; choosing these two tests will depend on the computational issues that the estimation entitles. Berkowitz test is undersized, and the Knuppel test works only with a large enough (i.e. T=100) sample size, but it is the most powerful.

To answer whether the Bank of England fan chart is calibrated, we applied the tests to Bank of England fan charts published from 2004Q1 up to 2020Q1. From our analysis, we can say that the calibration of path density forecast for inflation rate is not rejected by the majority of our tests, either horizon-by-horizon or jointly; for GDP growth rate and unemployment is rejected.

3.7 Appendices

3.7.1 Decomposition Example: AR(p)-generated Path Density Forecast

Let us simplify the set up considering a path density forecast of three horizons (i.e. $H = 3$) obtain form an AR(p) process:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \tag{3.70}$$

where ε_t follows a White Noise distribution $(0, 1)$. Equivalently, it can be written in companion form:

$$\begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p+1} \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 & \dots & \phi_{p-1} & \phi_p \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ y_{t-3} \\ \vdots \\ y_{t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{3.71}$$

or:

$$\boldsymbol{y}_t = \boldsymbol{B}\boldsymbol{y}_{t-1} + \boldsymbol{\eta}_t \tag{3.72}$$

Starting at t and iterating forward h periods, it gives:

$$\boldsymbol{y}_{t+h|t} = \boldsymbol{B}^h \boldsymbol{y}_t + \boldsymbol{\eta}_{t+h} + \boldsymbol{B}\boldsymbol{\eta}_{t+h-1} + \cdots + \boldsymbol{B}^{h-1}\boldsymbol{\eta}_{t+1} \tag{3.73}$$

The first equation of the system (3.73) characterises the value of y_{t+j} for $j = 1, \dots, H$. Let us denote b_{11}^j as the element $(1, 1)$ of the coefficient matrix $\boldsymbol{B}^j$. For j-step ahead, the first equation then becomes:

$$\begin{aligned} y_{t+j|t} =& b_{11}^j y_t + b_{12}^j y_{t-1} + \cdots + b_{1p}^j y_{t-p+1} + \\ & \varepsilon_{t+j} + \psi_1 \varepsilon_{t+j-1} + \psi_2 \varepsilon_{t+j-2} + \cdots + \psi_{j-1} \varepsilon_{t+1} \end{aligned} \tag{3.74}$$

where $\psi_1 = b_{11}^1 = \phi_1$, $\psi_2 = b_{11}^2 = \phi_1^2 + \phi_2$, and $\psi_{j-1} = b_{11}^j = \phi_1 \psi_{j-1} + \phi_2 \psi_{j-2} + \cdots + \phi_p \psi_{j-p}$.

Up to this point, we specified the forecasts obtained by a AR(p) model. Using

the decomposition in equation (3.10), the path density forecasts for $H = 3$, becomes:

$$f_t(y_{t+1}, y_{t+2}, y_{t+2}) = f_t(y_{t+H}|y_{t+H-1}, y_{t+H-2}) f_t(y_{t+H-1}|y_{t+H-2}) f_t(y_{t+H-2}) \tag{3.75}$$

However, under the law of iterated projections, $f_t(y_{t+H}|y_{t+H-1})$ contains the same information as $f_t(y_{t+H}|y_{t+H-1}, y_{t+H-2})$.

$$f_t(y_{t+h|t+h-1}) \sim \mathcal{N}\left(\boldsymbol{B}^h \boldsymbol{y}_t + \sum_{j=1}^{h-1} \boldsymbol{B}^{h-j} \boldsymbol{y}_{t+j}, \Sigma_\eta + \sum_{j=1}^{h-1} \boldsymbol{B}^j \boldsymbol{\eta}_{t+j}^2 \boldsymbol{B}^{j'}\right) \tag{3.76}$$

$$f_t(y_{t+h|t+h-2}) \sim \mathcal{N}\left(\boldsymbol{F}^h \boldsymbol{y}_t + \sum_{j=1}^{h-2} \boldsymbol{B}^{h-j} \boldsymbol{y}_{t+j}, \Sigma_\eta + \sum_{j=1}^{h-2} \boldsymbol{B}^j \boldsymbol{\eta}_{t+j}^2 \boldsymbol{B}^{j'}\right) \tag{3.77}$$

From this example it is easy to see that all the information in $f_t(y_{t+h|t+h-2})$ are contained in $f_t(y_{t+h|t+h-1})$ and then the joint distribution of path forecast is totally described by equation (3.11).

3.7.2 Derivation of transformations of the distributions of z_{CS} and z_{KP} for $h > 2$

Clements and Smith [2002]

In the paper, the number of variables considered is 2. The authors then call z^* the product of two PITs vectors. Assuming the independence of the two vectors, the distribution function of their product is:

$$F_{z^*} = z^* - z^* \ln Z^* \quad 0 < z^* < 1 \tag{3.78}$$

For more details of the derivation, please refer to Clements and Smith [2002]. However, we need to generalise the transformation to d number of elements in the vector to be multiplied. Consider the following change of variable for our case of $d = 4$:

$$\begin{aligned} z_1^* &= z_1 z_2 z_3 z_4 \\ z_2^* &= z_2 z_3 z_4 \\ z_3^* &= z_3 z_4 \\ z_4^* &= z_4 \end{aligned} \tag{3.79}$$

The determinant of the Jacobian for the inverse transformation is: $|J| = \frac{1}{z_2^* z_3^* z_4^*}$, and therefore the joint distribution of $z_1^* z_2^* z_3^* z_4^*$ is:

$$\begin{aligned} f_{z_1^* z_2^* z_3^* z_4^*} &= |J| f(z_1^* z_2^* z_3^* z_4^*) = \frac{1}{z_2^* z_3^* z_4^*} f\left(\frac{z_1^*}{z_2^* z_3^* z_4^*}\right) \\ f_{z_1^* z_2^* z_3^* z_4^*} &= \frac{(-1)^{d-1}}{d-1!} log^{d-1}(\frac{z_1^*}{z_2^* z_3^* z_4^*}) \frac{1}{z_2^* z_3^* z_4^*} \end{aligned} \tag{3.80}$$

Since we are interested in the distribution of z_1^*, let us integrate out of equation z_2^*, z_3^*, z_4^*. The marginal probability distribution of z_1^* is:

$$f_d(z_1^*) = \frac{(-1)^d}{d!} log^d(z_1^*) \tag{3.81}$$

For our case of $d = 4$, the transformation of cdf of z_1^* is equal to:

$$F_{d=4}(z_1^*) = \frac{z_1^*}{24}\left(log^4(z_1^*) - 4log^3(z_1^*) + 12log^2(z_1^*) - 24log(z_1^*) + 24\right) \tag{3.82}$$

Applying this transformation, the vector z_1^* has a uniform $U(0,1)$ distribution.

Ko and Park [2013]

Moving from a critique to Clements and Smith [2002], Ko and Park [2013] uses the case of 2 variables as well. Let us call z^* the location-adjusted transformation of two PITs vectors. Its distribution function is:

$$F(z^*) = \begin{cases} -2z^* ln2 + 2z^* - 2z^* ln(2z^*) + 1/2 & z^* > 0, \\ -2z^* ln2 + 2z^* - 2z^* ln(-2z^*) + 1/2 & z^* \leq 0 \end{cases} \tag{3.83}$$

For more details of the derivation, please refer to Ko and Park [2013]. However, we need to generalise the transformation to d number of vectors. Consider the following change of variable for our case of $d = 4$:

$$\begin{aligned} z_1^* &= (z_1 - \mathbb{E}(z_1)) \times (z_2 - \mathbb{E}(z_2)) \times (z_3 - \mathbb{E}(z_3)) \times (z_4 - \mathbb{E}(z_4)) \\ z_2^* &= (z_2 - \mathbb{E}(z_2)) \times (z_3 - \mathbb{E}(z_3)) \times (z_4 - \mathbb{E}(z_4)) \\ z_3^* &= (z_3 - \mathbb{E}(z_3)) \times (z_4 - \mathbb{E}(z_4)) \\ z_4^* &= z_4 - \mathbb{E}(z_4) \end{aligned} \tag{3.84}$$

The determinant of the Jacobian for the inverse transformation is: $|J| = \frac{1}{z_2^* z_3^* z_4^*}$, and therefore the joint distribution of $z_1^* z_2^* z_3^* z_4^*$ is:

$$\begin{aligned} f_{z_1^* z_2^* z_3^* z_4^*} &= |J| f(z_1^* z_2^* z_3^* z_4^*) = \left| \frac{1}{z_2^* z_3^* z_4^*} \right| f\left(\frac{z_1^*}{z_2^* z_3^* z_4^*} \right) \\ f_{z_1^* z_2^* z_3^* z_4^*} &= \frac{(2)^{d-1}}{d-1!} log^{d-1}\left(\frac{z_2^* z_3^* z_4^*}{2^d |z_1^*|} \right) \frac{1}{z_2^* z_3^* z_4^*} \end{aligned} \tag{3.85}$$

Since we are interested in the distribution of z_1^*, let us integrate out of equation z_2^*, z_3^*, z_4^*. The marginal probability distribution of z_1^* is:

$$f_d(z_1^*) = \frac{2^{d-1}}{(d-1)!} log^{d-1} \left| \frac{1}{(2^d z_1^*)} \right| \tag{3.86}$$

For our case of $d = 4$, the transformation of cdf of z_1^* is equal to:

$$F_{d=4}(z_1^*) = \frac{|z_1^*|}{3} \left(log^3 \left| \frac{1}{16 z_1^*} \right| + 3 log^2 \left| \frac{1}{16 z_1^*} \right| + 6 log \left| \frac{1}{16 z_1^*} \right| + 6 \right) \tag{3.87}$$

Applying this transformation, the vector z_1^* is now distributed according to a uniform $U(0,1)$ distribution.

3.7.3 The effect of temporal dependence on calibration's tests

The previous section concludes with a set of alternative vectors of uni-dimensional PITs to test for calibration. Before moving to analyse uniformity tests, one crucial aspect has to be discussed. Following Rosenblatt [1952], the PITs Z_t are uniform (under probabilistic calibration) subject to independence of PIT values z_t. If PITs are independent, they can be used directly for testing the calibration of density forecasts, employing for example the Kolmogorov-Smirnov test. Consider that:

$$F_t(y_{t+h}) = \{F_{t=1}(y_{t+h}), F_{t=2}(y_{t+h}), \ldots, F_{t=T}(y_{t+h})\} \tag{3.88}$$

and its probability integral transform is:

$$Z_{t+h|t} = \{z_{t+h|t=1}, z_{t+h|t=1}, \ldots, z_{t+h|t=1}\} \tag{3.89}$$

and so any conditional or marginal distribution in equations (3.11-3.51).

Hamill [2001] states how the use of formal tests is often hindered by complex dependence structures, particularly in the case of PIT values are spatially or temporarily

aggregated. The time-dependence that characterises path forecasts interferes with the evaluation procedure making the PITs dependent since for $h > 1$, h-step ahead forecasts have serially correlated errors making their PITs values serially correlated. Since individual PIT values are serially correlated, the vector of PITs are serially correlated.

This effect is well-discussed in the literature of multi-horizon forecasts Knüppel [2015] and regards the serial dependence among $\{z_{t+h|t=1}, z_{t+h|t=1}, \ldots, z_{t+h|t=1}\}$. $z_{t+h|t}$ will be serially dependent if $F_t(y_{t+h})$ are serially dependent when $h > 1$. To show this, consider, for example, $H = 2$ forecasts obtained from an AR(1) forecasting model:

$$y_t = \beta y_{t-1} + \varepsilon_t \tag{3.90}$$

where $t = 1, \ldots, T$ and ε_t are iid normal $(0, \sigma_\varepsilon)$ distributed. To prove that $y_{t+2|t}$ is serially dependent let us prove that:

$$cov(y_{t+2|t=1}, y_{t+2|t=2}) \neq 0 \tag{3.91}$$

where:

$$\begin{aligned} y_{t+2|t=1} &= \beta^2 y_{t=1} + \beta\varepsilon_{t+1|t=1} + \varepsilon_{t+2|t=1} \\ &= \beta^2 y_{t=1} + \beta\varepsilon_{t=2} + \varepsilon_{t=3} \\ y_{t+2|t=2} &= \beta^2 y_{t=2} + \beta\varepsilon_{t+1|t=2} + \varepsilon_{t+2|t=2} \\ &= \beta^2 y_{t=2} + \beta\varepsilon_{t=3} + \varepsilon_{t=4} \end{aligned} \tag{3.92}$$

It is easy to see that this two density forecasts are not independent to each other, and this is true for any other $t = 1, \ldots, T$. The covariance $cov(\varepsilon_{t+2|t=1}, \beta\varepsilon_{t+1|t=2}) \neq 0$ and $cov(y_{t+2|t=1}, y_{t+2|t=2}) = \beta\sigma_\varepsilon^2$. This is due to the fact that $\varepsilon_{t+2|t=1}$ and $\varepsilon_{t+1|t=2}$ are two notations for error terms that regards the same $t = 3$ horizon, they differ only by the point in time they are calculated i.e. $t = 2$ or $t = 1$. The serial dependence in the error term regards only marginals $y_{t+h|t}$ and not conditional distributions $(y_{t+h|t+h-1,t})$. To prove that $y_{t+2|t+1,t}$ is NOT serially dependent let us prove that:

$$cov(y_{t+2|t+1,t=1}, y_{t+2|t+1|t=2}) = 0 \tag{3.93}$$

where:

$$\begin{aligned} y_{t+2|t+1,t=1} &= \beta\hat{y}_{t+1|t=1} + \varepsilon_{t+2|t=1} \\ y_{t+2|t+1,t=2} &= \beta\hat{y}_{t+1|t=2} + \varepsilon_{t+2|t=2} \end{aligned} \tag{3.94}$$

where $\beta\hat{y}_{t+1|t=1}$ is the forecast made at $t = 1$ for horizon $h = 1$.

$$Cov(y_{t+2|t+1,t=1}, y_{t+2|t+1,t=2}) = 0 \tag{3.95}$$

Given that the serial dependence in z_t affects density forecasts for $h > 2$, $Z_{t+2|t}$ and $Z_{t+3|t}$ (but not $Z_{t+1|t}$), the components of vector Z_M are serially dependent and then Z_M itself. Defining the path forecast as a sequence of conditional forecasts help us avoiding the serial correlation in the error terms that affects the h-step ahead marginal distribution. Indeed it does not affect any of the "optimal but unfeasible" vector of PITs since they are function of one-step ahead forecasts $Z_{h=3|h=2,t}; Z_{h=2|h=1,t}; Z_{h=1|t}$. This is the first reason for which Z_{DHT}, Z_{CS}, Z_{KP} are superior to Z_M. However, serial correlation of errors is not an invalidating issue, since several solutions to test uniformity of serially correlated PITs are available in literature. A first approach consists in changing the associated critical values: Corradi and Swanson [2006a] and Rossi and Sekhposyan [2019] propose Kolmogorov-Smirnov and Cramér-von-Mises types of tests that account for serially correlation using empirical critical values. A second approach refers to the inverse normal transformation (INT) of PITs proposed by Smith [1985] and Berkowitz [2001]. Under probabilistic transformation, the INT of PIT values are normally distributed. These tests (including skewness and kurtosis normality tests Mitchell and Wallis [2011]) are valid also in presence of serially correlation. A third approach proposed by Knüppel [2015], is based of raw moments. More details about the tests, their strengths and limits will be given in section (C).

Table 3.1: Size Test: Empirical Rejection Frequencies for tests of marginal distributions for sample size T and $\boldsymbol{\phi} = \{\phi, 0, 0, 0\}$, $\sigma^2 = 1$.Nominal size: $\alpha = 0.05$.

		Kolmogorov-Smirnov (κ)								Cramer-von Mises (C)								Berkowitz Test				Knuppel Test			
T	ϕ	Traditional c.v.=1.36				Bootstrapped c.v.				Traditional c.v.=0.46				Bootstrapped c.v.											
		z_1	z_2	z_3	z_4	z_1	z_2	z_3	z_4	z_1	z_2	z_3	z_4	z_1	z_2	z_3	z_4	z_1	z_2	z_3	z_4	z_1	z_2	z_3	z_4
$T = 25$	0	0.038	0.036	0.032	0.038	0.120	0.120	0.120	0.120	0.070	0.110	0.110	0.080	0.100	0.100	0.100	0.100	0.051	0.02	0.021	0.023	0.025	0.000	0.001	0.001
	0.1	0.038	0.050	0.052	0.052	0.140	0.140	0.140	0.140	0.060	0.050	0.07	0.070	0.120	0.120	0.120	0.120	0.047	0.065	0.021	0.023	0.025	0.000	0.000	0.001
	0.5	0.038	0.099	0.150	0.173	0.320	0.320	0.320	0.320	0.160	0.190	0.190	0.230	0.320	0.320	0.320	0.320	0.047	0.139	0.021	0.023	0.025	0.000	0.000	0.000
	0.9	0.038	0.123	0.222	0.294	0.100	0.100	0.100	0.100	0.140	0.190	0.250	0.280	0.050	0.050	0.050	0.050	0.047	0.171	0.021	0.023	0.025	0.000	0.000	0.003
$T = 50$	0	0.050	0.090	0.020	0.050	0.110	0.110	0.110	0.110	0.090	0.150	0.080	0.110	0.070	0.070	0.070	0.070	0.050	0.090	0.050	0.070	0.015	0.010	0.009	0.013
	0.1	0.050	0.040	0.050	0.050	0.140	0.140	0.140	0.140	0.150	0.120	0.080	0.080	0.130	0.130	0.130	0.130	0.060	0.060	0.050	0.050	0.015	0.012	0.011	0.018
	0.5	0.040	0.090	0.150	0.170	0.02	0.02	0.02	0.02	0.150	0.090	0.120	0.150	0.1	0.1	0.1	0.1	0.030	0.100	0.180	0.200	0.015	0.018	0.022	0.028
	0.9	0.030	0.160	0.250	0.340	0.03	0.03	0.03	0.03	0.140	0.170	0.200	0.190	0.05	0.05	0.05	0.05	0.080	0.180	0.280	0.350	0.015	0.022	0.035	0.041
$T = 100$	0	0.041	0.048	0.048	0.046	0.06	0.06	0.06	0.06	0.132	0.122	0.112	0.13	0.07	0.07	0.07	0.07	0.058	0.057	0.057	0.058	0.036	0.034	0.034	0.030
	0.1	0.041	0.063	0.074	0.071	0.06	0.06	0.06	0.06	0.132	0.124	0.124	0.128	0.07	0.07	0.07	0.07	0.058	0.082	0.083	0.079	0.036	0.035	0.041	0.037
	0.5	0.041	0.121	0.153	0.185	0.04	0.04	0.04	0.04	0.070	0.130	0.100	0.110	0.06	0.06	0.06	0.06	0.058	0.142	0.194	0.227	0.036	0.032	0.037	0.044
	0.9	0.041	0.14	0.223	0.284	0.05	0.05	0.05	0.130	0.180	0.180	0.210	0.1	0.1	0.1	0.1	0.1	0.058	0.161	0.263	0.333	0.036	0.042	0.051	0.079
$T = 500$	0	0.049	0.049	0.051	0.053	0.1	0.1	0.1	0.1	0.096	0.098	0.104	0.098	0.06	0.06	0.06	0.06	0.056	0.054	0.056	0.061	0.036	0.034	0.034	0.030
	0.1	0.049	0.072	0.071	0.073	0.06	0.06	0.06	0.06	0.096	0.104	0.102	0.106	0.08	0.08	0.08	0.08	0.056	0.073	0.077	0.082	0.046	0.049	0.044	0.046
	0.5	0.049	0.118	0.161	0.192	0.05	0.05	0.05	0.05	0.070	0.060	0.008	0.050	0.1	0.1	0.1	0.1	0.056	0.137	0.205	0.234	0.046	0.045	0.054	0.056
	0.9	0.049	0.134	0.236	0.323	0.05	0.05	0.05	0.05	0.090	0.130	0.140	0.150	0.1	0.1	0.1	0.1	0.056	0.169	0.278	0.347	0.046	0.050	0.066	0.081
$T = 1000$	0	0.041	0.041	0.044	0.042	0.05	0.05	0.05	0.05	0.106	0.11	0.098	0.106	0.05	0.05	0.05	0.05	0.045	0.044	0.045	0.046	0.037	0.038	0.040	0.037
	0.1	0.041	0.053	0.056	0.057	0.05	0.05	0.05	0.05	0.106	0.102	0.100	0.108	0.05	0.05	0.05	0.05	0.045	0.065	0.064	0.064	0.037	0.039	0.038	0.038
	0.5	0.041	0.106	0.143	0.162	0.05	0.05	0.05	0.05	0.060	0.040	0.040	0.030	0.05	0.05	0.05	0.05	0.045	0.124	0.179	0.203	0.037	0.050	0.059	0.058
	0.9	0.041	0.125	0.230	0.286	0.05	0.05	0.05	0.05	0.090	0.090	0.110	0.080	0.04	0.04	0.04	0.04	0.045	0.153	0.247	0.325	0.037	0.049	0.064	0.079

Note: The table reports empirical p-values for the several tests for various. Bootstrapped critical values are obtained by block bootstrap following Rossi and Sekhposyan [2019]. The number of Monte Carlo replications is 1000.

Table 3.2: Size Test: Empirical Rejection Frequencies for tests of Conditional distributions for sample size T and $\boldsymbol{\phi} = \{\phi, 0, 0, 0\}$, $\sigma^2 = 1$.Nominal size: $\alpha = 0.05$.

		Kolmogorov-Smirnov (κ)						Cramer-von Mises (C)						Berkowitz Test			Knuppel Test		
T	ϕ	Traditional c.v.=1.36			Bootstrapped c.v.			Traditional c.v.=0.46			Bootstrapped c.v.								
		$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$
$T = 25$	0	0.036	0.032	0.038	0.05	0.36	0.07	0.043	0.043	0.051	0.05	0.05	0.05	0.021	0.023	0.025	0.000	0.001	0.001
	0.1	0.038	0.036	0.032	0.07	0.07	0.05	0.047	0.043	0.043	0.05	0.05	0.05	0.021	0.023	0.025	0.000	0.001	0.001
	0.5	0.036	0.032	0.038	0.07	0.05	0.05	0.047	0.043	0.043	0.05	0.05	0.05	0.021	0.023	0.025	0.000	0.001	0.001
	0.9	0.036	0.032	0.038	0.048	0.050	0.070	0.047	0.043	0.043	0.05	0.05	0.05	0.021	0.023	0.025	0.000	0.001	0.001
$T = 50$	0	0.049	0.051	0.053	0.04	0.08	0.057	0.057	0.054	0.056	0.04	0.04	0.039	0.019	0.018	0.017	0.01	0.009	0.013
	0.1	0.048	0.048	0.046	0.04	0.08	0.057	0.057	0.058	0.060	0.051	0.041	0.13	0.019	0.018	0.017	0.01	0.009	0.013
	0.5	0.049	0.051	0.053	0.050	0.050	0.050	0.054	0.056	0.061	0.050	0.050	0.050	0.019	0.018	0.017	0.01	0.009	0.013
	0.9	0.049	0.051	0.053	0.054	0.050	0.050	0.054	0.056	0.056	0.061	0.050	0.050	0.019	0.018	0.017	0.01	0.009	0.013
$T = 100$	0	0.049	0.051	0.053	0.050	0.050	0.050	0.054	0.056	0.061	0.050	0.050	0.050	0.01	0.013	0.011	0.034	0.034	0.030
	0.1	0.049	0.051	0.053	0.050	0.050	0.050	0.054	0.056	0.061	0.050	0.050	0.050	0.01	0.013	0.011	0.034	0.034	0.030
	0.5	0.051	0.053	0.056	0.050	0.050	0.050	0.054	0.056	0.061	0.050	0.050	0.050	0.01	0.013	0.011	0.034	0.034	0.030
	0.9	0.041	0.044	0.042	0.050	0.050	0.050	0.044	0.045	0.046	0.050	0.050	0.050	0.01	0.013	0.011	0.034	0.034	0.030
$T = 500$	0	0.041	0.044	0.042	0.050	0.050	0.050	0.044	0.045	0.046	0.05	0.05	0.05	0.013	0.012	0.011	0.046	0.045	0.046
	0.1	0.041	0.044	0.042	0.050	0.050	0.050	0.044	0.045	0.046	0.050	0.050	0.050	0.013	0.012	0.011	0.046	0.045	0.046
	0.5	0.041	0.044	0.042	0.050	0.050	0.050	0.044	0.045	0.046	0.050	0.050	0.050	0.013	0.012	0.011	0.046	0.045	0.046
	0.9	0.029	0.031	0.032	0.050	0.050	0.050	0.035	0.037	0.037	0.050	0.050	0.050	0.013	0.012	0.011	0.046	0.045	0.046
$T = 1000$	0	0.029	0.031	0.032	0.050	0.050	0.050	0.035	0.037	0.037	0.050	0.050	0.050	0.014	0.012	0.013	0.038	0.040	0.037
	0.1	0.029	0.031	0.032	0.050	0.050	0.050	0.035	0.037	0.037	0.050	0.050	0.050	0.014	0.012	0.013	0.038	0.040	0.037
	0.5	0.029	0.031	0.032	0.050	0.050	0.050	0.035	0.037	0.037	0.050	0.050	0.050	0.014	0.012	0.013	0.038	0.040	0.037
	0.9	0.029	0.031	0.032	0.050	0.050	0.050	0.035	0.037	0.037	0.050	0.050	0.050	0.014	0.012	0.013	0.038	0.040	0.037

Note: The table reports empirical p-values for the several tests for various. Bootstrapped critical values are obtained by block bootstrap following Rossi and Sekhposyan [2019]. The number of Monte Carlo replications is 1000.

Table 3.3: Size Test: Empirical Rejection Frequencies for tests of joint distributions for sample size T and $\boldsymbol{\phi} = \{\phi, 0, 0, 0\}$, $\sigma^2 = 1$. Nominal size: $\alpha = 0.05$.

		Kolmogorov-Smirnov (κ)								Cramer-von Mises (C)								Berkowitz Test				Knuppel Test			
T	ϕ	Traditional c.v.=1.36				Bootstrapped c.v.				Traditional c.v.=0.46				Bootstrapped c.v.											
		z_M	z_{DHT}	z_{CS}	z_{KP}	z_M	z_{DHT}	z_{CS}	z_{KP}	z_M	z_{DHT}	z_{CS}	z_{KP}	z_M	z_{DHT}	z_{CS}	z_{KP}	z_M	z_{DHT}	z_{CS}	z_{KP}	z_M	z_{DHT}	z_{CS}	z_{KP}
$T = 25$	0	0.027	0.027	0.33	0.112	0.05	0.05	0.36	0.07	0.043	0.043	0.357	0.098	0.048	0.049	0.360	0.085	0.084	0.084	0.221	0.051	0.000	0.000	0.002	0.000
	0.1	0.029	0.027	0.33	0.112	0.05	0.07	0.36	0.07	0.054	0.043	0.357	0.098	0.048	0.048	0.360	0.083	0.094	0.084	0.221	0.051	0.000	0.000	0.002	0.000
	0.5	0.059	0.027	0.33	0.112	0.14	0.05	0.36	0.07	0.115	0.043	0.357	0.098	0.048	0.090	0.366	0.85	0.225	0.084	0.221	0.051	0.000	0.000	0.002	0.000
	0.9	0.069	0.027	0.33	0.112	0.049	0.050	0.360	0.070	0.177	0.043	0.357	0.098	0.049	0.290	0.064	0.84	0.351	0.084	0.221	0.051	0.000	0.000	0.002	0.000
$T = 50$	0	0.038	0.038	0.342	0.141	0.040	0.040	0.390	0.080	0.056	0.056	0.374	0.14	0.0.070	0.050	0.260	0.150	0.111	0.111	0.319	0.064	0.007	0.007	0.063	0.059
	0.1	0.043	0.038	0.342	0.141	0.040	0.040	0.390	0.050	0.073	0.056	0.374	0.14	0.080	0.070	0.250	0.110	0.112	0.111	0.319	0.064	0.007	0.007	0.063	0.059
	0.5	0.062	0.038	0.342	0.141	0.110	0.040	0.139	0.080	0.132	0.056	0.374	0.14	0.090	0.050	0.250	0.120	0.345	0.111	0.319	0.064	0.007	0.007	0.063	0.059
	0.9	0.075	0.038	0.342	0.141	0.140	0.040	0.139	0.080	0.17	0.056	0.374	0.14	0.080	0.070	0.220	0.140	0.453	0.111	0.319	0.064	0.007	0.007	0.063	0.059
$T = 100$	0	0.049	0.049	0.374	0.153	0.030	0.030	0.243	0.100	0.054	0.054	0.373	0.129	0.070	0.080	0.240	0.110	0.085	0.085	0.393	0.058	0.032	0.032	0.087	0.079
	0.1	0.049	0.049	0.374	0.153	0.030	0.030	0.143	0.100	0.068	0.054	0.373	0.129	0.070	0.080	0.200	0.120	0.117	0.085	0.393	0.058	0.030	0.032	0.087	0.079
	0.5	0.06	0.049	0.374	0.153	0.060	0.030	0.343	0.100	0.134	0.054	0.373	0.129	0.100	0.090	0.020	0.130	0.418	0.085	0.393	0.058	0.032	0.032	0.087	0.079
	0.9	0.076	0.049	0.374	0.153	0.021	0.030	0.343	0.100	0.185	0.054	0.373	0.129	0.110	0.060	0.211	0.090	0.462	0.085	0.393	0.058	0.032	0.032	0.087	0.079
$T = 500$	0	0.041	0.041	0.341	0.138	0.070	0.070	0.041	0.080	0.043	0.043	0.358	0.121	0.110	0.130	0.170	0.070	0.087	0.087	0.445	0.047	0.046	0.046	0.095	0.079
	0.1	0.037	0.041	0.341	0.138	0.080	0.070	0.041	0.080	0.059	0.043	0.358	0.121	0.100	0.080	0.150	0.070	0.184	0.087	0.445	0.047	0.045	0.046	0.095	0.079
	0.5	0.047	0.041	0.341	0.138	0.018	0.070	0.041	0.080	0.122	0.043	0.358	0.121	0.060	0.080	0.150	0.070	0.451	0.087	0.445	0.047	0.028	0.046	0.095	0.079
	0.9	0.076	0.041	0.341	0.138	0.022	0.070	0.041	0.080	0.162	0.043	0.358	0.121	0.091	0.110	0.130	0.080	0.471	0.087	0.445	0.047	0.015	0.046	0.095	0.079
$T = 1000$	0	0.031	0.031	0.309	0.111	0.022	0.070	0.041	0.080	0.036	0.036	0.345	0.117	0.090	0.110	0.130	0.080	0.075	0.075	0.483	0.048	0.036	0.036	0.091	0.069
	0.1	0.031	0.031	0.309	0.111	0.040	0.040	0.036	0.110	0.049	0.036	0.345	0.117	0.060	0.030	0.070	0.040	0.297	0.075	0.483	0.048	0.040	0.036	0.091	0.069
	0.5	0.043	0.031	0.309	0.111	0.050	0.040	0.036	0.110	0.112	0.036	0.345	0.117	0.020	0.020	0.040	0.070	0.499	0.075	0.483	0.048	0.033	0.036	0.091	0.069
	0.9	0.055	0.031	0.309	0.111	0.050	0.040	0.036	0.110	0.151	0.036	0.345	0.117	0.020	0.040	0.080	0.060	0.51	0.075	0.483	0.048	0.036	0.036	0.091	0.069

Note: The table reports empirical p-values for the several tests for various. Bootstrapped critical values are obtained by block bootstrap following Rossi and Sekhposyan [2019]. The number of Monte Carlo replications is 1000.

Table 3.4: Size Test: Empirical Rejection Frequencies for uniformity tests of marginal distributions. $\boldsymbol{\phi} = \{0.3, 0.2, 0.1, 0.1\}$, $\sigma^2 = 1$

	Kolmogorov-Smirnov (κ)							
	Traditional c.v.=1.36				Bootstrapped c.v.			
	z_1	z_2	z_3	z_4	z_1	z_2	z_3	z_4
T=25	0.038	0.08	0.131	0.171	0.026	0.026	0.026	0.029
T=50	0.041	0.094	0.135	0.182	0.020	0.020	0.020	0.016
T=100	0.049	0.085	0.15	0.197	0.028	0.028	0.028	0.028
T=500	0.041	0.100	0.125	0.181	0.027	0.027	0.027	0.027
T=1000	0.029	0.117	0.138	0.159	0.029	0.029	0.029	0.029
	Cramer-von Mises (C)							
	Traditional c.v.=0.46				Bootstrapped c.v.			
	z_1	z_2	z_3	z_4	z_1	z_2	z_3	z_4
T=25	0.047	0.09	0.162	0.212	0.260	0.260	0.260	0.260
T=50	0.058	0.104	0.165	0.229	0.140	0.140	0.140	0.140
T=100	0.056	0.104	0.177	0.239	0.170	0.170	0.170	0.170
T=500	0.045	0.118	0.162	0.209	0.130	0.130	0.130	0.130
T=1000	0.036	0.124	0.168	0.199	0.090	0.090	0.090	0.090
	Berkowitz Test				Knuppel Test			
	z_1	z_2	z_3	z_4	z_1	z_2	z_3	z_4
T=25	0.02	0.134	0.267	0.338	0.000	0.000	0.000	0.001
T=50	0.019	0.326	0.487	0.591	0.015	0.019	0.015	0.016
T=100	0.009	0.64	0.793	0.874	0.036	0.056	0.05	0.065
T=500	0.209	0.013	1	1	0.046	0.176	0.105	0.076
T=1000	0.199	0.013	1	1	0.037	0.291	0.152	0.084

Note: The table reports empirical rejection frequencies for the test statistic KS at the 5% nominal size for various sample sizes T. Bootstrapped critical values obtained by block bootstrap following Rossi and Sekhposyan [2019]. The number of Monte Carlo replications is 1000.

Table 3.5: Size Test: Empirical Rejection Frequencies for uniformity tests of conditional distributions. $\boldsymbol{\phi} = \{0.3, 0.2, 0.1, 0.1\}$, $\sigma^2 = 1$

	Kolmogorov-Smirnov (κ)							
	Traditional c.v.=1.36				Bootstrapped c.v.			
	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$
T=25	0.038	0.036	0.032	0.038	0.030	0.030	0.030	0.030
T=50	0.041	0.048	0.048	0.046	0.030	0.030	0.030	0.030
T=100	0.049	0.049	0.051	0.053	0.040	0.040	0.040	0.040
T=500	0.041	0.041	0.044	0.042	0.070	0.070	0.070	0.070
T=1000	0.029	0.029	0.031	0.032	0.050	0.050	0.050	0.050
	Cramer-von Mises (C)							
	Traditional c.v.=0.46				Bootstrapped c.v.			
	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$
T=25	0.047	0.043	0.043	0.051	0.08	0.080	0.080	0.080
T=50	0.058	0.057	0.057	0.058	0.08	0.080	0.080	0.080
T=100	0.056	0.054	0.056	0.061	0.07	0.070	0.070	0.070
T=500	0.045	0.044	0.045	0.046	0.11	0.110	0.110	0.110
T=1000	0.036	0.035	0.037	0.037	0.04	0.040	0.040	0.040
	Berkowitz Test				Knuppel Test			
	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$
T=25	0.020	0.021	0.023	0.025	0.000	0.000	0.001	0.001
T=50	0.019	0.019	0.018	0.017	0.015	0.010	0.009	0.013
T=100	0.009	0.010	0.013	0.011	0.035	0.036	0.034	0.030
T=500	0.013	0.013	0.012	0.011	0.046	0.046	0.045	0.046
T=1000	0.013	0.014	0.012	0.013	0.037	0.038	0.040	0.037

Note: The table reports empirical rejection frequencies for the test statistic KS at the 5% nominal size for various sample sizes T. Critical values obtained by block bootstrap following Rossi and Sekhposyan [2019]. The number of Monte Carlo replications is 1,000.

Table 3.6: Size Test: Empirical Rejection Frequencies for uniformity tests of vectors. $\boldsymbol{\phi} = \{0.3, 0.2, 0.1, 0.1\}$, $\sigma^2 = 1$

	Kolmogorov-Smirnov (κ)							
	Traditional c.v.=1.36				Bootstrapped c.v.			
	z_M	z_{DHT}	z_{CS}	z_{KP}	z_M	z_{DHT}	z_{CS}	z_{KP}
T=25	0.04	0.027	0.330	0.112	0.140	0.050	0.360	0.070
T=50	0.054	0.038	0.342	0.141	0.090	0.040	0.390	0.080
T=100	0.049	0.049	0.374	0.153	0.140	0.030	0.430	0.100
T=500	0.05	0.041	0.341	0.138	0.180	0.070	0.410	0.080
T=1000	0.043	0.031	0.309	0.111	0.100	0.040	0.360	0.110
	Cramer-von Mises (C)							
	Traditional c.v.=0.46				Bootstrapped c.v.			
	z_M	z_{DHT}	z_{CS}	z_{KP}	z_M	z_{DHT}	z_{CS}	z_{KP}
T=25	0.102	0.043	0.357	0.098	0.160	0.010	0.350	0.170
T=50	0.117	0.056	0.374	0.14	0.060	0.080	0.230	0.120
T=100	0.125	0.054	0.373	0.129	0.080	0.070	0.210	170
T=500	0.118	0.043	0.358	0.121	0.100	0.110	0.150	0.080
T=1000	0.106	0.036	0.345	0.117	0.040	0.040	0.070	0.060
	Berkowitz Test				Knuppel Test			
	z_M	z_{DHT}	z_{CS}	z_{KP}	z_M	z_{DHT}	z_{CS}	z_{KP}
T=25	0.175	0.084	0.221	0.051	0.051	0.000	0.000	0.002
T=50	0.237	0.111	0.319	0.064	0.006	0.007	0.063	0.059
T=100	0.334	0.085	0.393	0.058	0.022	0.032	0.087	0.079
T=500	0.444	0.087	0.445	0.047	0.050	0.046	0.095	0.079
T=1000	0.506	0.075	0.483	0.048	0.066	0.036	0.091	0.069

Note: The table reports empirical rejection frequencies for the test statistic KS at the 5% nominal size for various sample sizes T. Critical values obtained by block bootstrap following Rossi and Sekhposyan [2019]. The number of Monte Carlo replications is 1,000.

Table 3.7: Size Test: Empirical Rejection Frequencies for uniformity tests of marginal distributions. $\boldsymbol{\phi} = \{0.1, 0.1, 0.1, 0.1\}$, $\sigma^2 = 1$

	Kolmogorov-Smirnov (κ)							
	Traditional c.v.=1.36				Bootstrapped c.v.			
	z_1	z_2	z_3	z_4	z_1	z_2	z_3	z_4
T=25	0.038	0.049	0.065	0.083	0.29	0.29	0.29	0.29
T=50	0.041	0.062	0.083	0.105	0.16	0.16	0.16	0.14
T=100	0.049	0.069	0.081	0.117	0.14	0.14	0.14	0.14
T=500	0.041	0.051	0.069	0.092	0.1	0.1	0.1	0.1
T=1000	0.029	0.043	0.058	0.084	0.09	0.09	0.09	0.09
	Cramer-von Mises (C)							
	Traditional c.v.=0.46				Bootstrapped c.v.			
	z_1	z_2	z_3	z_4	z_1	z_2	z_3	z_4
T=25	0.047	0.064	0.086	0.111	0.260	0.260	0.260	0.260
T=50	0.058	0.08	0.104	0.129	0.140	0.140	0.140	0.140
T=100	0.056	0.073	0.101	0.137	0.170	0.170	0.170	0.170
T=500	0.045	0.066	0.085	0.117	0.130	0.130	0.130	0.130
T=1000	0.036	0.056	0.076	0.106	0.090	0.090	0.09	0.09
	Berkowitz Test				Knuppel Test			
	z_1	z_2	z_3	z_4	z_1	z_2	z_3	z_4
T=25	0.02	0.028	0.046	0.062	0.000	0.000	0.000	0.001
T=50	0.019	0.050	0.061	0.080	0.015	0.012	0.010	0.015
T=100	0.009	0.071	0.117	0.144	0.036	0.034	0.038	0.041
T=500	0.013	0.384	0.463	0.543	0.046	0.047	0.049	0.060
T=1000	0.013	0.773	0.832	0.899	0.037	0.041	0.045	0.056

Note: The table reports empirical rejection frequencies for the test statistic KS at the 5% nominal size for various sample sizes T. Bootstrapped critical values obtained by block bootstrap following Rossi and Sekhposyan [2019]. The number of Monte Carlo replications is 1000.

Table 3.8: Size Test: Empirical Rejection Frequencies for uniformity tests of conditional distributions. $\boldsymbol{\phi} = \{0.1, 0.1, 0.1, 0.1\}$, $\sigma^2 = 1$

	Kolmogorov-Smirnov (κ)							
	Traditional c.v.=1.36				Bootstrapped c.v.			
	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$
T=25	0.038	0.036	0.032	0.038	0.140	0.140	0.140	0.140
T=50	0.041	0.048	0.048	0.046	0.100	0.100	0.100	0.100
T=100	0.049	0.049	0.051	0.053	0.100	0.100	0.100	0.100
T=500	0.041	0.041	0.044	0.042	0.070	0.070	0.070	0.070
T=1000	0.029	0.029	0.031	0.032	0.050	0.050	0.050	0.050
	Cramer-von Mises (C)							
	Traditional c.v.=0.46				Bootstrapped c.v.			
	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$
T=25	0.047	0.043	0.043	0.051	0.080	0.080	0.080	0.080
T=50	0.058	0.057	0.057	0.058	0.080	0.080	0.080	0.080
T=100	0.056	0.054	0.056	0.061	0.070	0.070	0.070	0.070
T=500	0.045	0.044	0.045	0.046	0.110	0.110	0.110	0.110
T=1000	0.036	0.035	0.037	0.037	0.040	0.040	0.040	0.040
	Berkowitz Test				Knuppel Test			
	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$
T=25	0.02	0.021	0.023	0.025	0.000	0.000	0.001	0.001
T=50	0.019	0.019	0.018	0.017	0.010	0.009	0.009	0.013
T=100	0.009	0.01	0.013	0.011	0.036	0.034	0.034	0.030
T=500	0.013	0.013	0.012	0.011	0.046	0.046	0.045	0.046
T=1000	0.013	0.014	0.012	0.013	0.037	0.038	0.040	0.037

Note: The table reports empirical rejection frequencies for the test statistic KS at the 5% nominal size for various sample sizes T. Critical values obtained by block bootstrap following Rossi and Sekhposyan [2019]. The number of Monte Carlo replications is 1,000.

Table 3.9: Size Test: Empirical Rejection Frequencies for uniformity tests of vectors. $\boldsymbol{\phi} = \{0.1, 0.1, 0.1, 0.1\}$, $\sigma^2 = 1$

	Kolmogorov-Smirnov (κ)							
	Traditional c.v.=1.36				Bootstrapped c.v.			
	z_M	z_{DHT}	z_{CS}	z_{KP}	z_M	z_{DHT}	z_{CS}	z_{KP}
T=25	0.030	0.027	0.330	0.112	0.110	0.050	0.360	0.070
T=50	0.036	0.038	0.342	0.141	0.050	0.040	0.390	0.080
T=100	0.044	0.049	0.374	0.153	0.050	0.030	0.343	0.100
T=500	0.033	0.041	0.341	0.138	0.110	0.070	0.341	0.080
T=1000	0.023	0.031	0.309	0.111	0.060	0.040	0.360	0.110
	Cramer-von Mises (C)							
	Traditional c.v.=0.46				Bootstrapped c.v.			
	z_M	z_{DHT}	z_{CS}	z_{KP}	z_M	z_{DHT}	z_{CS}	z_{KP}
T=25	0.067	0.043	0.357	0.098	0.160	0.100	0.350	0.17
T=50	0.083	0.056	0.374	0.140	0.080	0.050	0.230	0.150
T=10	0.081	0.054	0.373	0.129	0.090	0.060	0.220	0.130
T=500	0.068	0.043	0.358	0.121	0.080	0.110	0.180	0.090
T=1000	0.062	0.036	0.345	0.117	0.050	0.040	0.060	0.060
	Berkowitz Test				Knuppel Test			
	z_M	z_{DHT}	z_{CS}	z_{KP}	z_M	z_{DHT}	z_{CS}	z_{KP}
T=25	0.100	0.084	0.221	0.051	0.005	0.000	0.000	0.002
T=50	0.116	0.111	0.319	0.064	0.006	0.007	0.063	0.059
T=100	0.132	0.085	0.393	0.058	0.027	0.032	0.087	0.079
T=500	0.121	0.202	0.087	0.445	0.047	0.046	0.095	0.079
T=1000	0.327	0.075	0.483	0.048	0.043	0.036	0.091	0.069

Note: The table reports empirical rejection frequencies for the test statistic KS at the 5% nominal size for various sample sizes T. Critical values obtained by block bootstrap following Rossi and Sekhposyan [2019]. The number of Monte Carlo replications is 1,000.

Table 3.10: Power Simulation 1: Rejection Probabilities when DGP is iid and path has a AR(4) process.

Panel A: Rejection probabilities Marginal distributions

Size	z_1	z_2	z_3	z_4	z_1	z_2	z_3	z_4
	Kolmogorov-Smirnov (κ)				Cramer-von Mises (C)			
T=25	0.971	0.976	0.989	0.995	1	1	1	1
T=50	1	1	1	1	1	1	1	1
T=100	1	1	1	1	1	1	1	1
T=500	1	1	1	1	1	1	1	1
T=1000	1	1	1	1	1	1	1	1
	Berkowitz Test				Knuppel Test			
T=25	0	0	0	0.001	0.002	0.003	0.005	0.007
T=50	0	0	0	0.001	1	1	1	1
T=100	0	0	0	0.001	1	1	1	1
T=500	0	0	0	0.001	1	1	1	1
T=1000	0	0	0	0.001	1	1	1	1

Panel B: Rejection probabilities Conditional distributions

	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$
	Kolmogorov-Smirnov (κ)				Cramer-von Mises (C)			
T=25	0.971	0.977	0.975	0.978	1	1	1	1
T=50	1	1	1	1	1	1	1	1
T=100	1	1	1	1	1	1	1	1
T=500	1	1	1	1	1	1	1	1
T=1000	1	1	1	1	1	1	1	1
	Berkowitz Test				Knuppel Test			
T=25	0	0	0	0.001	0.002	0.001	0.002	0
T=50	0	0	0	0.001	1	1	1	1
T=100	0	0	0	0.001	1	1	1	1
T=500	0	0	0	0.001	1	1	1	1
T=1000	0	0	0	0.001	1	1	1	1

Panel C: Rejection probabilities joint distribution

	z_M	z_{DHL}	z_{CS}	z_{KP}	z_S	z_{DHL}	z_{CS}	z_{KP}
	Kolmogorov-Smirnov (κ)				Cramer-von Mises (C)			
T=25	0.942	0.969	1	1	1	1	1	1
T=50	1	1	1	1	1	1	1	1
T=100	1	1	1	1	1	1	1	1
T=500	1	1	1	1	1	1	1	1
T=1000	1	1	1	1	1	1	1	1
	Berkowitz Test				Knuppel Test			
T=25	1	1	0.994	1	1	1	0.963	0.003
T=50	1	1	0.999	1	1	1	1	1
T=100	1	1	1	1	1	1	1	1
T=500	1	1	1	1	1	1	1	1
T=1000	1	1	1	1	1	1	1	1

Note: The table reports empirical p-values for the several tests for various. The number of Monte Carlo replications is 1000.

Table 3.11: Power Simulation 2: Rejection Probabilities when DGP is AR(4) and path has a AR(1) process.

Panel A: Rejection probabilities Marginal distributions

Size	z_1	z_2	z_3	z_4	z_1	z_2	z_3	z_4
	Kolmogorov-Smirnov (κ)				Cramer-von Mises (C)			
T=25	0.275	0.364	0.402	0.392	0.352	0.429	0.465	0.457
T=50	0.348	0.42	0.444	0.458	0.401	0.475	0.493	0.508
T=100	0.384	0.441	0.466	0.473	0.426	0.492	0.515	0.527
T=500	0.567	0.577	0.618	0.627	0.586	0.593	0.637	0.648
T=1000	0.757	0.733	0.778	0.79	0.842	0.788	0.847	0.859
	Berkowitz Test				Knuppel Test			
T=25	0.056	0.12	0.17	0.189	0.001	0.002	0.002	0.005
T=50	0.074	0.218	0.254	0.265	0.099	0.109	0.125	0.129
T=100	0.100	0.356	0.386	0.401	0.246	0.229	0.246	0.260
T=500	0.336	0.829	0.842	0.851	0.728	0.596	0.663	0.672
T=1000	0.509	0.983	0.984	0.985	0.972	0.884	0.942	0.951

Panel B: Rejection probabilities Conditional distributions

	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$
	Kolmogorov-Smirnov (κ)				Cramer-von Mises (C)			
T=25	0.275	0.289	0.3	0.283	0.352	0.362	0.357	0.362
T=50	0.348	0.347	0.349	0.35	0.401	0.406	0.405	0.412
T=100	0.384	0.38	0.38	0.387	0.426	0.42	0.427	0.436
T=500	0.567	0.566	0.564	0.568	0.586	0.591	0.594	0.595
T=1000	0.757	0.756	0.758	0.756	0.842	0.84	0.836	0.840
	Berkowitz Test				Knuppel Test			
T=25	0.056	0.052	0.052	0.052	0.001	0.001	0.001	0.001
T=50	0.074	0.078	0.084	0.084	0.099	0.099	0.099	0.098
T=100	0.100	0.093	0.095	0.095	0.246	0.248	0.248	0.254
T=500	0.336	0.331	0.325	0.331	0.728	0.726	0.721	0.729
T=1000	0.509	0.514	0.509	0.506	0.972	0.973	0.97	0.969

Panel C: Rejection probabilities joint distribution

	z_M	z_{DHL}	z_{CS}	z_{KP}	z_S	z_{DHL}	z_{CS}	z_{KP}
	Kolmogorov-Smirnov (κ)				Cramer-von Mises (C)			
T=25	0.301	0.27	0.683	0.175	0.418	0.351	0.703	0.165
T=50	0.361	0.344	0.701	0.3	0.472	0.404	0.719	0.287
T=100	0.394	0.375	0.739	0.367	0.483	0.42	0.746	0.35
T=500	0.536	0.563	0.933	0.747	0.615	0.586	0.944	0.77
T=1000	0.725	0.762	0.997	0.942	0.829	0.84	0.999	0.954
	Berkowitz Test				Knuppel Test			
T=25	0.100	0.096	0.36	0.091	0.835	0.842	0.025	0.000
T=50	0.188	0.168	0.442	0.185	0.85	0.859	0.172	0.051
T=100	0.266	0.257	0.454	0.333	0.887	0.893	0.319	0.122
T=500	0.618	0.72	0.481	0.882	0.981	0.987	0.859	0.632
T=1000	0.785	0.867	0.544	0.992	1	1	0.989	0.929

Note: The table reports empirical p-values for the several tests for various. The number of Monte Carlo replications is 1000.

Table 3.12: Power Simulation 3: Rejection Probabilities for calibration of single horizon forecasts (marginals) when DGP is MA(1) and path has a AR(1) process.

	AR(1)	Kolmogorov-Smirnov (κ)				Cramer-von Mises (C)				Berkowitz Test				Knuppel Test			
	coeff.	z_1	z_2	z_3	z_4	z_1	z_2	z_3	z_4	z_1	z_2	z_3	z_4	z_1	z_2	z_3	z_4
$T = 25$	$\beta = 0$	0.015	0.043	0.065	0.096	0.013	0.054	0.078	0.116	0.004	0.145	0.372	0.363	0	0	0.001	0
	$\beta = 0.5$	0.002	0.001	0.001	0.005	0.007	0.003	0.001	0.012	0.034	0.026	0.356	0.369	0	0	0.001	0.001
	$\beta = 0.9$	0.062	0.068	0.064	0.067	0.069	0.089	0.072	0.088	0.011	0.051	0.362	0.369	0	0	0	0
$T = 50$	$\beta = 0$	0.014	0.058	0.079	0.094	0.017	0.064	0.103	0.119	0.007	0.413	0.73	0.717	0.006	0.008	0.017	0.026
	$\beta = 0.5$	0.01	0.002	0.006	0.02	0.017	0.003	0.005	0.029	0.08	0.085	0.754	0.688	0.016	0.009	0.033	0.158
	$\beta = 0.9$	0.169	0.172	0.095	0.082	0.278	0.21	0.1	0.094	0.029	0.16	0.752	0.71	0.112	0.089	0.025	0.013
$T = 100$	$\beta = 0$	0.019	0.051	0.077	0.1	0.016	0.054	0.093	0.115	0.001	0.817	0.983	0.979	0.019	0.018	0.033	0.04
	$\beta = 0.5$	0.026	0.01	0.004	0.057	0.054	0.01	0.004	0.121	0.188	0.294	0.985	0.952	0.147	0.043	0.118	0.552
	$\beta = 0.9$	0.514	0.408	0.158	0.114	0.677	0.521	0.197	0.124	0.06	0.49	0.989	0.969	0.802	0.593	0.191	0.084
$T = 500$	$\beta = 0$	0.015	0.049	0.078	0.095	0.012	0.059	0.082	0.105	0.003	1	1	1	0.056	0.063	0.066	0.099
	$\beta = 0.5$	0.621	0.068	0.041	0.828	0.805	0.141	0.083	0.93	0.962	0.999	1	1	0.984	0.444	0.421	0.998
	$\beta = 0.9$	1	1	0.832	0.494	1	1	0.929	0.576	0.62	1	1	1	1	1	0.993	0.849
$T = 1000$	$\beta = 0$	0.014	0.043	0.058	0.083	0.012	0.049	0.082	0.106	0	1	1	1	0.08	0.091	0.065	0.123
	$\beta = 0.5$	0.979	0.265	0.117	0.997	0.998	0.44	0.233	1	1	1	1	1	1	0.837	0.698	1
	$\beta = 0.9$	1	1	0.993	0.865	1	1	1	0.946	0.962	1	1	1	1	1	1	1

Note: The table reports empirical p-values for the several tests for various. The number of Monte Carlo replications is 1000.

Table 3.13: Power Simulation 3: Rejection Probabilities for calibration of single horizon forecasts (conditionals) when DGP is MA(1) and path has a AR(1) process.

	AR(1)	Kolmogorov-Smirnov (κ)				Cramer-von Mises (C)				Berkowitz Test				Knuppel Test			
	coeff.	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$	z_1	$z_{2\|1}$	$z_{3\|2}$	$z_{4\|3}$
$T = 25$	$\beta = 0$	0.015	0.016	0.013	0.013	0.013	0.013	0.01	0.016	0.002	0.001	0.002	0.001	0	0	0	0
	$\beta = 0.5$	0.002	0.002	0.001	0.004	0.007	0.007	0.005	0.004	0.036	0.034	0.04	0.041	0	0	0	0
	$\beta = 0.9$	0.062	0.061	0.064	0.065	0.069	0.082	0.081	0.073	0.366	0.369	0.355	0.36	0	0.001	0.001	0.001
$T = 50$	$\beta = 0$	0.014	0.017	0.013	0.018	0.017	0.017	0.015	0.015	0.001	0.001	0	0.001	0.006	0.007	0.006	0.005
	$\beta = 0.5$	0.01	0.008	0.01	0.007	0.017	0.016	0.015	0.013	0.127	0.132	0.138	0.135	0.016	0.016	0.018	0.016
	$\beta = 0.9$	0.169	0.174	0.165	0.188	0.278	0.278	0.281	0.276	0.773	0.771	0.785	0.769	0.112	0.119	0.116	0.113
$T = 100$	$\beta = 0$	0.019	0.02	0.019	0.022	0.016	0.018	0.018	0.016	0.001	0	0	0	0.019	0.018	0.021	0.02
	$\beta = 0.5$	0.026	0.024	0.023	0.022	0.054	0.059	0.05	0.051	0.307	0.31	0.308	0.31	0.147	0.137	0.139	0.139
	$\beta = 0.9$	0.514	0.514	0.512	0.511	0.677	0.673	0.673	0.664	0.984	0.984	0.981	0.982	0.802	0.798	0.794	0.803
$T = 500$	$\beta = 0$	0.015	0.018	0.017	0.017	0.012	0.012	0.011	0.013	0.007	0.007	0.006	0.005	0.056	0.06	0.059	0.06
	$\beta = 0.5$	0.621	0.632	0.624	0.626	0.805	0.802	0.804	0.809	0.997	0.998	0.997	0.996	0.984	0.985	0.986	0.986
	$\beta = 0.9$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
$T = 100$	$\beta = 0$	0.014	0.012	0.01	0.011	0.012	0.012	0.011	0.012	0.017	0.017	0.019	0.02	0.08	0.079	0.079	0.083
	$\beta = 0.5$	0.979	0.979	0.981	0.982	0.998	0.998	0.998	0.999	1	1	1	1	1	1	1	1
	$\beta = 0.9$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Note: The table reports empirical p-values for the several tests for various. The number of Monte Carlo replications is 1000.

Table 3.14: Power Simulation 3: Rejection Probabilities for calibration of path forecast when DGP is MA(1) and path has a AR(1) process.

	AR(1)	Kolmogorov-Smirnov (κ)				Cramer-von Mises (C)				Berkowitz Test				Knuppel Test			
	coeff.	z_M	z_{DHT}	z_{CS}	z_{KP}	z_S	z_{DHT}	z_{CS}	z_{KP}	z_M	z_{DHT}	z_{CS}	z_{KP}	z_S	z_{DHT}	z_{CS}	z_{KP}
$T = 25$	$\beta = 0$	0.014	0.009	0.262	0.133	0.044	0.006	0.299	0.124	0.539	0.118	0.936	0.127	0	0	0.001	0.001
	$\beta = 0.5$	0	0.001	0.322	0.264	0	0.004	0.38	0.271	0.217	0.31	0.831	0.138	0	0	0.007	0.001
	$\beta = 0.9$	0.005	0.047	0.661	0.43	0.03	0.054	0.704	0.457	0.347	0.212	0.943	0.124	0	0	0.019	0
$T = 50$	$\beta = 0$	0.018	0.012	0.285	0.172	0.058	0.014	0.318	0.149	0.776	0.131	0.999	0.131	0.002	0.002	0.087	0.045
	$\beta = 0.5$	0	0.007	0.604	0.369	0	0.013	0.662	0.373	0.356	0.501	0.997	0.133	0	0.014	0.158	0.078
	$\beta = 0.9$	0.016	0.159	0.886	0.652	0.072	0.27	0.913	0.695	0.529	0.328	0.999	0.122	0.001	0.09	0.343	0.195
$T = 100$	$\beta = 0$	0.022	0.017	0.301	0.18	0.051	0.018	0.325	0.176	0.969	0.126	1	0.141	0.007	0.017	0.149	0.078
	$\beta = 0.5$	0	0.022	0.828	0.53	0	0.056	0.894	0.551	0.577	0.726	1	0.145	0	0.132	0.362	0.241
	$\beta = 0.9$	0.064	0.51	0.983	0.864	0.234	0.672	0.988	0.905	0.806	0.486	1	0.128	0.222	0.792	0.825	0.628
$T = 500$	$\beta = 0$	0.016	0.014	0.498	0.264	0.051	0.011	0.521	0.26	1	0.108	1	0.183	0.021	0.059	0.346	0.153
	$\beta = 0.5$	0	0.619	1	0.984	0	0.804	1	0.983	0.988	0.999	1	0.225	0.003	0.986	1	0.982
	$\beta = 0.9$	0.971	1	1	1	0.999	1	1	1	1	0.973	1	0.179	1	1	1	1
$T = 100$	$\beta = 0$	0.01	0.013	0.611	0.321	0.031	0.011	0.662	0.346	1	0.122	1	0.2	0.021	0.08	0.511	0.253
	$\beta = 0.5$	0	0.982	1	1	0	0.998	1	1	1	1	1	0.28	0.002	1	1	1
	$\beta = 0.9$	1	1	1	1	1	1	1	1	1	1	1	0.215	1	1	1	1

Note: The table reports empirical p-values for the several tests for various. The number of Monte Carlo replications is 1000.

Table 3.15: "Sup test" Size Properties for the first $H = 4$ forecast horizons in DGP IMA.

		κ_P						C_P					
	P/R	25	50	100	200	500	1000	25	50	100	200	500	1000
$h=1$	5	0.114	0.110	0.078	0.065	0.064	0.066	0.085	0.086	0.063	0.067	0.060	0.052
	50	0.116	0.104	0.091	0.077	0.061	0.055	0.097	0.080	0.076	0.066	0.049	0.045
	100	0.121	0.094	0.088	0.085	0.055	0.053	0.084	0.091	0.074	0.073	0.058	0.043
	200	0.107	0.098	0.082	0.079	0.064	0.048	0.086	0.079	0.075	0.069	0.06	0.056
	500	0.097	0.094	0.077	0.080	0.049	0.049	0.088	0.074	0.054	0.060	0.05	0.056
	1000	0.124	0.099	0.085	0.045	0.050	0.045	0.097	0.085	0.078	0.047	0.055	0.044
$h=2$	25	0.067	0.074	0.069	0.057	0.053	0.048	0.047	0.050	0.046	0.046	0.046	0.045
	50	0.084	0.081	0.066	0.074	0.049	0.053	0.055	0.054	0.045	0.054	0.047	0.048
	100	0.077	0.085	0.069	0.066	0.060	0.041	0.055	0.066	0.055	0.043	0.045	0.033
	200	0.068	0.064	0.055	0.068	0.040	0.048	0.048	0.040	0.043	0.045	0.035	0.051
	500	0.066	0.060	0.057	0.056	0.055	0.050	0.048	0.045	0.037	0.049	0.044	0.044
	1000	0.090	0.071	0.058	0.044	0.044	0.049	0.049	0.051	0.049	0.030	0.049	0.034
$h=3$	25	0.058	0.056	0.063	0.047	0.045	0.047	0.035	0.039	0.038	0.035	0.037	0.033
	50	0.074	0.066	0.058	0.052	0.046	0.048	0.037	0.039	0.042	0.041	0.038	0.034
	100	0.058	0.066	0.061	0.057	0.054	0.040	0.037	0.048	0.045	0.035	0.047	0.033
	200	0.040	0.042	0.040	0.058	0.041	0.048	0.039	0.032	0.032	0.036	0.031	0.047
	500	0.049	0.045	0.056	0.051	0.047	0.047	0.032	0.038	0.039	0.033	0.034	0.051
	1000	0.066	0.063	0.051	0.039	0.052	0.033	0.037	0.040	0.039	0.027	0.039	0.027
$h=4$	25	0.054	0.040	0.056	0.050	0.034	0.044	0.030	0.028	0.036	0.038	0.028	0.038
	50	0.046	0.066	0.038	0.046	0.048	0.044	0.020	0.040	0.038	0.046	0.038	0.026
	100	0.070	0.068	0.068	0.038	0.056	0.040	0.030	0.056	0.048	0.024	0.038	0.028
	200	0.040	0.046	0.038	0.074	0.034	0.050	0.040	0.038	0.038	0.036	0.036	0.044
	500	0.058	0.044	0.048	0.044	0.042	0.050	0.032	0.026	0.026	0.030	0.044	0.046
	1000	0.058	0.060	0.042	0.032	0.042	0.042	0.022	0.034	0.020	0.022	0.040	0.032

Table 3.16: Sizes of Path Calibration Tests in equations (3.55, and 3.56) in case of DGP IMA for $H = 4$.

		κ_P						C_P					
	P/R	25	50	100	200	500	1000	25	50	100	200	500	1000
Max Sup	25	0.127	0.123	0.090	0.078	0.071	0.070	0.084	0.089	0.064	0.067	0.061	0.053
	50	0.138	0.115	0.104	0.086	0.070	0.062	0.099	0.084	0.077	0.067	0.049	0.046
	100	0.135	0.110	0.103	0.093	0.063	0.057	0.082	0.094	0.076	0.073	0.058	0.043
	200	0.123	0.113	0.091	0.089	0.066	0.050	0.084	0.079	0.076	0.071	0.060	0.056
	500	0.114	0.107	0.084	0.085	0.059	0.060	0.087	0.073	0.055	0.060	0.051	0.057
	1000	0.138	0.116	0.096	0.053	0.055	0.049	0.090	0.088	0.081	0.047	0.055	0.044
W_{equal}	25	0.044	0.054	0.057	0.044	0.041	0.039	0.033	0.044	0.041	0.045	0.043	0.040
	50	0.068	0.053	0.054	0.046	0.041	0.038	0.047	0.045	0.045	0.051	0.040	0.039
	100	0.058	0.064	0.059	0.053	0.046	0.028	0.049	0.056	0.053	0.042	0.039	0.031
	200	0.052	0.044	0.040	0.056	0.032	0.045	0.038	0.036	0.042	0.046	0.040	0.048
	500	0.044	0.045	0.043	0.047	0.041	0.043	0.040	0.042	0.038	0.046	0.039	0.045
	1000	0.063	0.056	0.050	0.033	0.034	0.031	0.044	0.045	0.048	0.023	0.042	0.028
$W_{descending}$	25	0.061	0.060	0.061	0.046	0.045	0.040	0.044	0.063	0.048	0.050	0.050	0.043
	50	0.074	0.058	0.062	0.052	0.045	0.039	0.056	0.058	0.054	0.056	0.045	0.042
	100	0.071	0.071	0.063	0.056	0.045	0.033	0.059	0.067	0.063	0.051	0.040	0.036
	200	0.061	0.057	0.050	0.057	0.042	0.045	0.051	0.050	0.051	0.052	0.040	0.048
	500	0.049	0.055	0.044	0.054	0.047	0.040	0.049	0.049	0.038	0.052	0.043	0.047
	1000	0.076	0.065	0.052	0.041	0.038	0.033	0.057	0.056	0.056	0.029	0.044	0.031

Table 3.17: Power exercise 1 for Path Calibration Tests in equations (3.55, and 3.56): $\mu_t + 2$ and $H = 4$

		κ_P						C_P					
	P/R	25	50	100	200	500	1000	25	50	100	200	500	1000
Max Sup	25	0.660	0.862	0.990	1	1	1	0.646	0.868	0.992	1	1	1
	50	0.710	0.882	0.990	1	1	1	0.678	0.890	0.992	1	1	1
	100	0.724	0.878	0.984	1	1	1	0.676	0.886	0.992	1	1	1
	200	0.704	0.898	0.994	1	1	1	0.672	0.908	0.996	1	1	1
	500	0.708	0.892	0.990	1	1	1	0.678	0.898	0.990	1	1	1
	1000	0.668	0.880	0.992	1	1	1	0.634	0.882	0.998	1	1	1
W_{equal}	25	0.580	0.858	0.994	1	1	1	0.596	0.880	0.998	1	1	1
	50	0.604	0.858	0.988	1	1	1	0.600	0.886	0.996	1	1	1
	100	0.612	0.858	0.980	1	1	1	0.618	0.896	0.998	1	1	1
	200	0.622	0.884	0.994	1	1	1	0.622	0.908	0.996	1	1	1
	500	0.632	0.870	0.988	1	1	1	0.632	0.894	0.994	1	1	1
	1000	0.600	0.846	0.994	1	1	1	0.586	0.882	1	1	1	1
$W_{descending}$	25	0.566	0.830	0.980	1	1	1	0.570	0.854	0.992	1	1	1
	50	0.594	0.840	0.980	1	1	1	0.602	0.872	0.994	1	1	1
	100	0.590	0.848	0.976	1	1	1	0.608	0.882	0.992	1	1	1
	200	0.612	0.862	0.992	1	1	1	0.608	0.882	0.994	1	1	1
	500	0.618	0.858	0.986	1	1	1	0.610	0.886	0.988	1	1	1
	1000	0.592	0.838	0.992	1	1	1	0.578	0.862	0.992	1	1	1

Table 3.18: Power exercise 2 of Path Calibration Tests in equations (3.55, and 3.56): $\epsilon_t \sim i.i.d.N(0,1)$ and $H = 4$

		κ_P						C_P					
	P/R	25	50	100	200	500	1000	25	50	100	200	500	1000
Max Sup	25	0.270	0.298	0.396	0.632	0.954	1	0.140	0.176	0.266	0.526	0.950	1
	50	0.272	0.352	0.446	0.628	0.954	1	0.114	0.174	0.268	0.510	0.950	1
	100	0.268	0.324	0.448	0.592	0.950	1	0.132	0.178	0.284	0.516	0.936	1
	200	0.254	0.254	0.376	0.574	0.954	1	0.110	0.140	0.200	0.460	0.942	1
	500	0.236	0.276	0.412	0.624	0.964	1	0.112	0.152	0.250	0.508	0.966	1
	1000	0.262	0.310	0.376	0.612	0.966	1	0.140	0.162	0.248	0.478	0.956	1
W_{equal}	25	0.152	0.212	0.344	0.62	0.968	1	0.076	0.128	0.274	0.602	0.974	1
	50	0.160	0.266	0.380	0.602	0.968	1	0.068	0.142	0.288	0.588	0.980	1
	100	0.144	0.236	0.364	0.568	0.962	1	0.082	0.144	0.302	0.584	0.972	1
	200	0.122	0.174	0.304	0.536	0.958	1	0.048	0.096	0.232	0.528	0.974	1
	500	0.138	0.196	0.342	0.608	0.972	1	0.044	0.124	0.270	0.556	0.982	1
	1000	0.146	0.208	0.306	0.576	0.964	1	0.070	0.116	0.242	0.554	0.978	1
$W_{descending}$	25	0.166	0.226	0.358	0.614	0.960	1	0.098	0.140	0.270	0.576	0.966	1
	50	0.166	0.258	0.382	0.610	0.964	1	0.076	0.150	0.274	0.568	0.976	1
	100	0.164	0.256	0.370	0.568	0.960	1	0.086	0.148	0.286	0.542	0.962	1
	200	0.140	0.180	0.314	0.536	0.952	1	0.064	0.116	0.218	0.494	0.958	1
	500	0.140	0.200	0.344	0.600	0.966	1	0.062	0.120	0.256	0.538	0.980	1
	1000	0.178	0.218	0.306	0.562	0.960	1	0.098	0.128	0.234	0.524	0.976	1

Table 3.19: Power exercise 3 of Path Calibration Tests in equations (3.55, and 3.56): $\epsilon_t \sim i.i.d.N(0, 1.261 * 2)$ and $H = 4$

		κ_P						C_P					
	P/R	25	50	100	200	500	1000	25	50	100	200	500	1000
Max Sup	25	0.684	0.956	1	1	1	1	0.906	0.998	1	1	1	1
	50	0.644	0.968	1	1	1	1	0.916	0.996	1	1	1	1
	100	0.678	0.978	1	1	1	1	0.906	1	1	1	1	1
	200	0.728	0.970	1	1	1	1	0.934	0.998	1	1	1	1
	500	0.672	0.978	1	1	1	1	0.904	1	1	1	1	1
	1000	0.634	0.958	1	1	1	1	0.876	1	1	1	1	1
W_{equal}	25	0.634	0.972	1	1	1	1	0.886	1	1	1	1	1
	50	0.560	0.972	1	1	1	1	0.914	1	1	1	1	1
	100	0.616	0.984	1	1	1	1	0.892	1	1	1	1	1
	200	0.674	0.982	1	1	1	1	0.918	1	1	1	1	1
	500	0.618	0.978	1	1	1	1	0.894	1	1	1	1	1
	1000	0.582	0.972	1	1	1	1	0.876	1	1	1	1	1
$W_{descending}$	25	0.628	0.966	1	1	1	1	0.886	1	1	1	1	1
	50	0.574	0.968	1	1	1	1	0.914	1	1	1	1	1
	100	0.618	0.984	1	1	1	1	0.894	1	1	1	1	1
	200	0.688	0.974	1	1	1	1	0.920	1	1	1	1	1
	500	0.626	0.972	1	1	1	1	0.894	1	1	1	1	1
	1000	0.570	0.970	1	1	1	1	0.872	1	1	1	1	1

Table 3.20: Sizes of Path Calibration Tests in equations (3.55, and 3.56) in case of DGP IMA for H=12.

		κ_P						C_P					
	P/R	25	50	100	200	500	1000	25	50	100	200	500	1000
$h=4$	25	0.054	0.040	0.056	0.050	0.034	0.044	0.030	0.028	0.036	0.038	0.028	0.038
	50	0.046	0.066	0.038	0.046	0.048	0.044	0.020	0.040	0.038	0.046	0.038	0.026
	100	0.070	0.068	0.068	0.038	0.056	0.040	0.030	0.056	0.048	0.024	0.038	0.028
	200	0.040	0.046	0.038	0.074	0.034	0.050	0.040	0.038	0.038	0.036	0.036	0.044
	500	0.058	0.044	0.048	0.044	0.042	0.050	0.032	0.026	0.026	0.030	0.044	0.046
	1000	0.058	0.060	0.042	0.032	0.042	0.042	0.022	0.034	0.020	0.022	0.040	0.032
$h=8$	25	0.042	0.054	0.062	0.048	0.034	0.040	0.026	0.034	0.032	0.026	0.030	0.030
	50	0.046	0.052	0.040	0.048	0.054	0.052	0.026	0.034	0.028	0.038	0.042	0.038
	100	0.086	0.066	0.07	0.044	0.058	0.042	0.044	0.054	0.052	0.020	0.044	0.034
	200	0.056	0.040	0.050	0.070	0.032	0.052	0.036	0.02	0.038	0.036	0.026	0.048
	500	0.048	0.056	0.044	0.052	0.044	0.048	0.028	0.034	0.026	0.038	0.040	0.042
	1000	0.054	0.066	0.038	0.034	0.042	0.046	0.024	0.028	0.026	0.022	0.034	0.030
$h=12$	25	0.034	0.052	0.070	0.052	0.044	0.040	0.028	0.036	0.040	0.038	0.030	0.030
	50	0.044	0.052	0.040	0.048	0.048	0.046	0.028	0.026	0.032	0.028	0.030	0.034
	100	0.064	0.070	0.064	0.038	0.056	0.048	0.036	0.050	0.038	0.020	0.038	0.030
	200	0.060	0.054	0.040	0.058	0.032	0.046	0.028	0.032	0.032	0.030	0.026	0.054
	500	0.046	0.058	0.040	0.054	0.034	0.044	0.028	0.032	0.030	0.034	0.040	0.040
	1000	0.040	0.054	0.032	0.038	0.044	0.036	0.026	0.026	0.018	0.020	0.038	0.032

Table 3.21: Sizes of Path Calibration Tests in equations (3.55, and 3.56) in case of DGP IMA and H=12.

	P/R	25	50	100	200	500	1000	25	50	100	200	500	1000
Strict	25	0.156	0.140	0.118	0.094	0.076	0.070	0.100	0.110	0.076	0.068	0.066	0.052
	50	0.150	0.148	0.106	0.090	0.070	0.062	0.090	0.090	0.080	0.082	0.050	0.050
	100	0.140	0.128	0.146	0.094	0.070	0.064	0.086	0.118	0.104	0.070	0.054	0.040
	200	0.132	0.128	0.072	0.092	0.056	0.054	0.098	0.084	0.066	0.062	0.068	0.068
	500	0.122	0.134	0.096	0.088	0.058	0.050	0.078	0.074	0.066	0.070	0.052	0.046
	1000	0.166	0.120	0.090	0.048	0.064	0.060	0.090	0.074	0.074	0.042	0.064	0.054
W_{equal}	25	0.024	0.046	0.062	0.042	0.042	0.038	0.018	0.028	0.030	0.032	0.032	0.036
	50	0.024	0.038	0.032	0.044	0.042	0.046	0.020	0.022	0.038	0.046	0.034	0.038
	100	0.044	0.040	0.062	0.038	0.056	0.034	0.024	0.044	0.040	0.024	0.032	0.030
	200	0.032	0.032	0.036	0.062	0.028	0.046	0.030	0.020	0.038	0.034	0.022	0.052
	500	0.024	0.032	0.036	0.036	0.036	0.036	0.018	0.026	0.020	0.032	0.042	0.042
	1000	0.036	0.038	0.026	0.026	0.034	0.040	0.016	0.020	0.024	0.020	0.038	0.030
$W_{descending}$	25	0.032	0.046	0.058	0.044	0.042	0.038	0.022	0.038	0.036	0.036	0.040	0.036
	50	0.030	0.036	0.034	0.048	0.042	0.046	0.026	0.022	0.040	0.056	0.040	0.038
	100	0.046	0.060	0.062	0.034	0.054	0.030	0.040	0.048	0.048	0.028	0.040	0.032
	200	0.028	0.032	0.032	0.064	0.028	0.046	0.036	0.036	0.038	0.038	0.028	0.050
	500	0.034	0.030	0.040	0.034	0.036	0.038	0.020	0.026	0.024	0.038	0.046	0.038
	1000	0.042	0.044	0.030	0.028	0.034	0.040	0.020	0.034	0.032	0.020	0.040	0.034

Table 3.22: Uniformity tests of Bank of England Fan Charts Inflation at each horizon h.

	KvStat	KvCV	CvMStat	CvMCV	BerkowitzStat	BerkowitzCV	KnuppelStat	KnuppelCV
h=1	0.98295	1.4768	0.17808	0.70887	15.3584	5.9915	3.8164	9.4877
h=2	0.68392	1.4909	0.068743	0.80843	17.532	5.9915	2.3827	9.4877
h=3	0.99504	1.5307	0.19167	0.95518	14.8258	5.9915	2.2369	9.4877
h=4	1.1563	1.6121	0.43974	0.99042	19.1634	5.9915	4.0122	9.4877
h=5	1.4858	1.6907	0.64212	1.0182	23.2371	5.9915	5.1442	9.4877
h=6	1.4057	1.6213	0.74386	1.0774	26.2473	5.9915	4.1375	9.4877
h=7	1.484	1.6389	0.80311	1.0117	25.439	5.9915	3.2113	9.4877
h=8	1.6068	1.6365	0.706	0.9161	28.005	5.9915	3.5694	9.4877
h=9	1.4195	1.6771	0.61893	0.95207	33.5271	5.9915	3.7477	9.4877
h=10	1.2612	1.6183	0.56509	0.83102	26.8302	5.9915	4.1393	9.4877
h=11	1.2637	1.706	0.41506	0.85167	27.1024	5.9915	3.347	9.4877
h=12	1.0206	1.7031	0.35538	0.86637	24.876	5.9915	4.0671	9.4877
h=13	1.1406	1.7236	0.35273	0.87032	23.5249	5.9915	3.9432	9.4877

The table shows test statistics ("Stat") and critical values ("CV") obtained evaluating each forecast horizon spanning from 2004Q1 to 2020Q1. Sample size: 52 observations. KvCV and CvMCV refer to bootstrap critical values for Kolmogorov-Smirnov and Cramer-von Mises respectively. Traditional critical values are equal to 1.36 for Kolmogorov-Smirnov and 0.46 for Cramer-von Mises. The null hypothesis of calibration is rejected if the test static is greater than the respective critical value.

Table 3.23: Empirical Correlations between horizons for Bank of England Fan Charts for inflation at horizon $h = 1 : H$ and the previous horizon forecast.

h=1	h=2	h=3	h=4	h=5	h=6	h=7	h=8	h=9	h=10	h=11	h=12	h=13
0.2480	1.0389	1.1346	0.9229	0.4502	0.3623	0.4534	0.5409	0.5217	0.1788	0.1290	0.1759	0.0312

Figure 3.1: Histogram of pits values for marginal distributions. Inflation forecasts by Bank of England Fan charts at horizons $h = 1, \ldots, 13$.

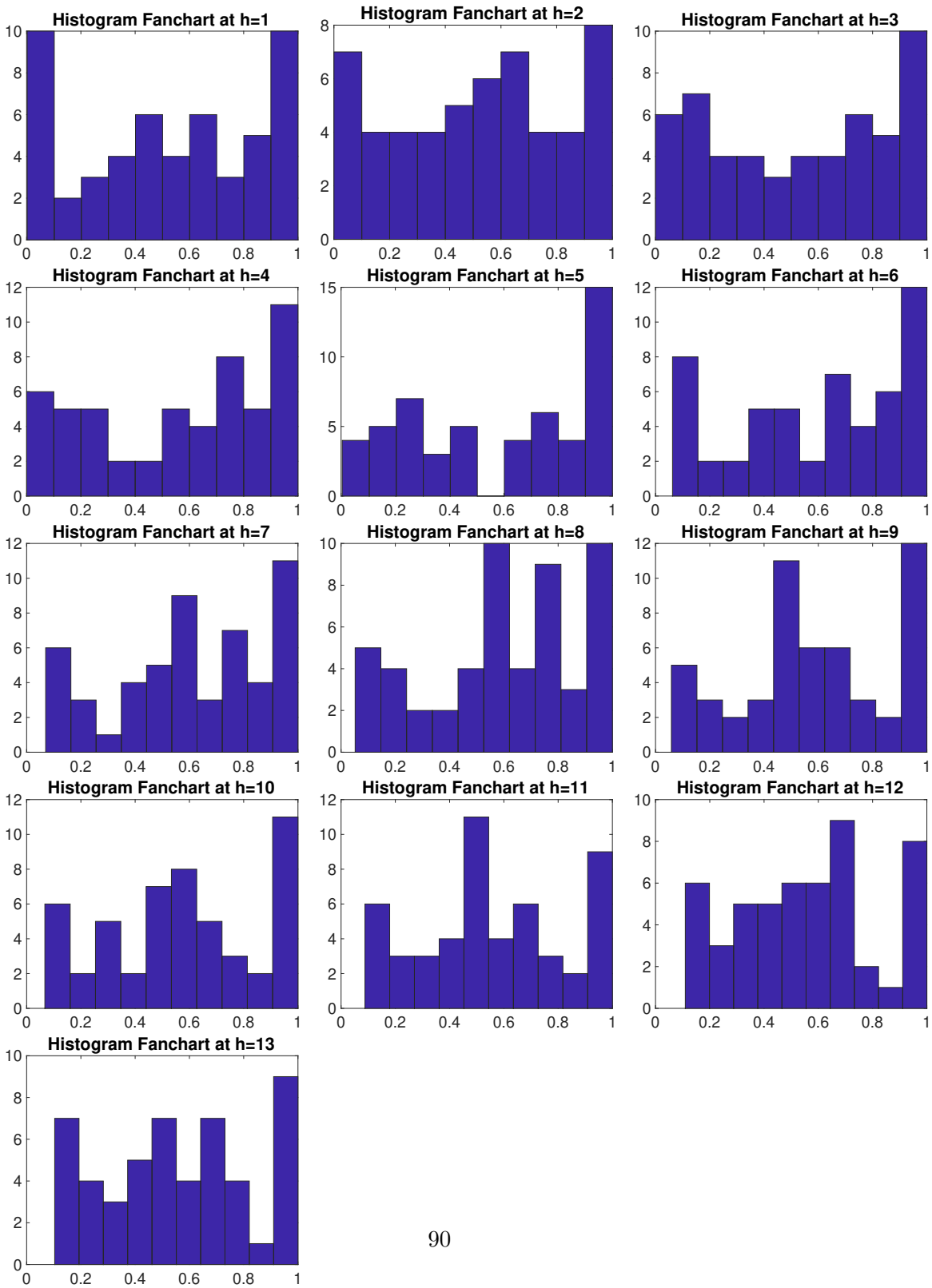


Table 3.24: Uniformity tests of Conditional distribution of Bank of England Fan Charts for Inflation at horizon h give the previous horizon forecast.

	KvStat	KvCV	CvMStat	CvMCV	BerkowitzStat	BerkowitzCV	KnuppelStat	KnuppelCV
$z_{2\|1}$	1.0700	1.4835	0.18462	0.61695	14.6137	5.9915	3.1695	9.4877
$z_{3\|2}$	0.7554	1.5461	0.10078	0.86384	19.377	5.9915	0.34295	9.4877
$z_{4\|3}$	1.1132	1.6462	0.2646	1.1102	24.5342	5.9915	2.9974	9.4877
$z_{5\|4}$	1.1369	1.6296	0.38032	1.218	35.3487	5.9915	3.7828	9.4877
$z_{6\|5}$	1.1699	1.8001	0.4316	1.4178	46.6568	5.9915	3.1515	9.4877
$z_{7\|6}$	1.2786	1.7338	0.38936	1.3355	48.9546	5.9915	2.3015	9.4877
$z_{8\|7}$	1.3567	1.6773	0.50139	1.2276	47.2792	5.9915	3.2228	9.4877
$z_{9\|8}$	1.7052	1.6196	0.74475	1.2591	51.1992	5.9915	4.2439	9.4877
$z_{10\|9}$	1.7498	1.7412	0.8532	1.5075	59.9285	5.9915	4.2314	9.4877
$z_{11\|10}$	1.7125	1.7511	0.9964	1.5966	65.2507	5.9915	6.002	9.4877
$z_{12\|11}$	1.7732	1.7537	1.0372	1.4383	69.1626	5.9915	5.3245	9.4877
$z_{13\|12}$	1.5893	1.6345	0.99193	1.3308	66.3947	5.9915	5.2549	9.4877

The table shows test statistics ("Stat") and critical values ("CV") obtained evaluating each forecast horizon spanning from 2004Q1 to 2020Q1. Sample size: 52 observations. KvCV and CvMCV refer to bootstrap critical values for Kolmogorov-Smirnov and Cramer-von Mises respectively. Traditional critical values are equal to 1.36 for Kolmogorov-Smirnov and 0.46 for Cramer-von Mises. The null hypothesis of calibration is rejected if the test static is greater than the respective critical value.

Table 3.25: Uniformity tests of Vectors of PITs of Bank of England Fan Charts for Inflation rate at horizon $h = 1, \ldots, 13$.

	Unfeasible				Feasible		
	z_{DHT}	z_{CS}	z_{KP}	z_M	SupMax	SupAverage	SupDes
KvStat	1.039	3.143	3.34	0.835	1.607	1.224	1.179
KvCV Bootstp	1.755	1.763	1.569	1.693	1.724	1.625	1.589
KvCV Trad	1.360	1.360	1.360	1.360	1.360	1.360	1.360
CvMStat	0.375	3.204	5.286	0.312	0.803	0.468	0.427
CvMCV Bootstp	1.206	1.872	0.644	0.961	1.077	0.912	0.911
CvMCVtrad	0.460	0.460	0.460	0.460	0.460	0.460	0.460
BerkowitzStat	524.63	141.652	65.988	277.158			
BerkowitzCV	5.9915	5.9915	5.9915	5.9915			
KnuppelStat	21.494	5.843	7.179	29.604			
KnuppelCV	9.4877	9.4877	9.4877	9.4877			

The table shows test statistics ("Stat") and critical values ("CV") obtained evaluating each forecast horizon spanning from 2004Q1 to 2020Q1. KvCV and CvMCV refer to bootstrap critical values for Kolmogorov-Smirnov and Cramer-von Mises respectively. Sample size: 52 observations. The null hypothesis of path calibration is rejected if the test static is greater than the respective critical value.

Table 3.26: Uniformity tests of Bank of England Fan Charts GDP growth at each horizon h.

	KvStat	KvCV	CvMStat	CvMCV	BerkowitzStat	BerkowitzCV	KnuppelStat	KnuppelCV
h=1	2.4078	1.6459	2.4864	0.67167	41.5815	5.9915	4.6573	9.4877
h=2	2.4349	1.4892	2.6863	0.5444	36.0102	5.9915	4.7524	9.4877
h=3	2.5431	1.5714	2.9219	0.57347	33.1247	5.9915	4.7991	9.4877
h=4	2.2354	1.4979	2.5546	0.53151	31.9758	5.9915	6.9382	9.4877
h=5	2.0994	1.538	2.2386	0.78829	28.7306	5.9915	6.8713	9.4877
h=6	1.9286	1.6918	1.5777	0.83001	20.4326	5.9915	5.8849	9.4877
h=7	1.5652	1.5638	0.86212	0.70421	18.9003	5.9915	6.5274	9.4877
h=8	1.6388	1.6718	0.6234	0.74802	18.0925	5.9915	6.5573	9.4877
h=9	1.4875	1.4849	0.57123	0.64563	16.6564	5.9915	6.3992	9.4877
h=10	1.2534	1.5009	0.3864	0.6818	13.1712	5.9915	6.0832	9.4877
h=11	0.8553	1.6213	0.21681	0.82781	12.9874	5.9915	5.1528	9.4877
h=12	0.71077	1.6107	0.11554	0.81182	12.1361	5.9915	3.8847	9.4877
h=13	0.70655	1.6652	0.094451	0.80077	10.5692	5.9915	2.1282	9.4877

The table shows test statistics ("Stat") and critical values ("CV") obtained evaluating each forecast horizon spanning from 2004Q1 to 2020Q1. Sample size: 52 observations. KvCV and CvMCV refer to bootstrap critical values for Kolmogorov-Smirnov and Cramer-von Mises respectively. The null hypothesis of calibration is rejected if the test static is greater than the respective critical value.

Table 3.27: Empirical Correlations between horizons for Bank of England Fan Charts for GDP at horizon $h = 1 : H$ and the previous horizon forecast.

h=1	h=2	h=3	h=4	h=5	h=6	h=7	h=8	h=9	h=10	h=11	h=12	h=13
1.2432	0.2266	0.3421	0.3964	0.2234	0.1131	0.1357	0.1372	0.0759	0.1155	0.0851	0.1008	0.0506

Figure 3.2: Histogram of pits values for marginal distributions. GDP growth rate forecasts by Bank of England Fan charts at horizons $h = 1, \ldots, 13$.

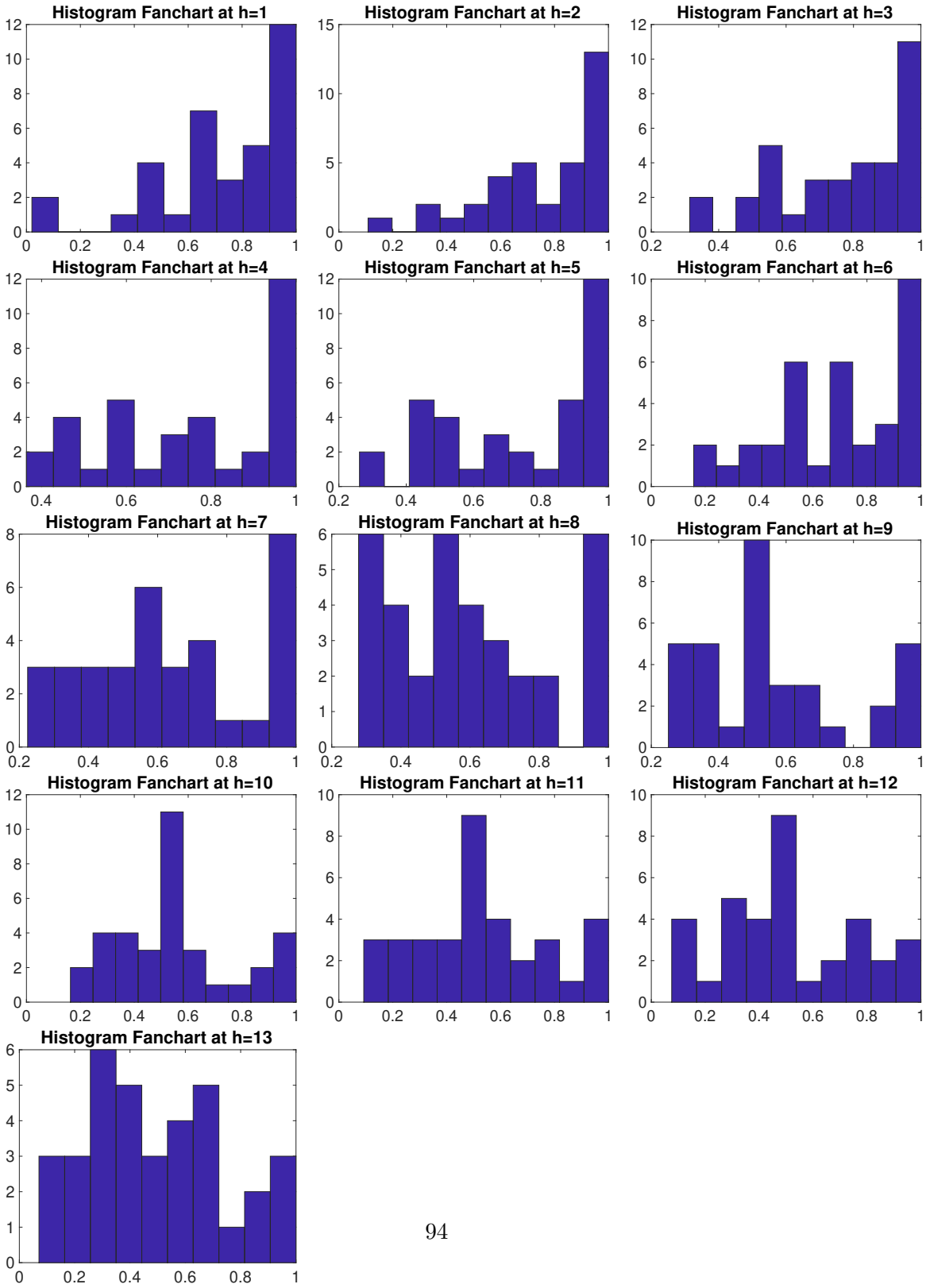


Table 3.28: Uniformity tests of Conditional distribution of Bank of England Fan Charts for GDP growth at horizon h give the previous horizon forecast.

	KvStat	KvCV	CvMStat	CvMCV	BerkowitzStat	BerkowitzCV	KnuppelStat	KnuppelCV
$z_{2\|1}$	1.0548	1.3717	0.17784	0.57447	7.7259	5.9915	4.342	9.4877
$z_{3\|2}$	0.88572	1.4972	0.15759	0.78186	10.2459	5.9915	2.3382	9.4877
$z_{4\|3}$	1.1688	1.583	0.26364	0.92049	11.6051	5.9915	2.2783	9.4877
$z_{5\|4}$	1.1114	1.5937	0.27211	0.94337	15.4367	5.9915	2.0366	9.4877
$z_{6\|5}$	1.01	1.7249	0.17894	1.201	20.8864	5.9915	1.7744	9.4877
$z_{7\|6}$	0.73275	1.6783	0.083718	1.0208	20.6604	5.9915	0.661	9.4877
$z_{8\|7}$	0.82149	1.665	0.11948	1.1001	18.5047	5.9915	1.6557	9.4877
$z_{9\|8}$	1.2365	1.5008	0.30182	1.1167	21.5017	5.9915	3.9456	9.4877
$z_{10\|9}$	1.1883	1.6719	0.29625	1.1579	24.5123	5.9915	3.6421	9.4877
$z_{11\|10}$	1.2018	1.678	0.3875	1.4043	31.4416	5.9915	6.1914	9.4877
$z_{12\|11}$	1.3125	1.6696	0.41042	1.2368	38.2726	5.9915	5.0099	9.4877
$z_{13\|12}$	1.1004	1.6718	0.31434	1.2485	40.7789	5.9915	4.1864	9.4877

The table shows test statistics ("Stat") and critical values ("CV") obtained evaluating each forecast horizon spanning from 2004Q1 to 2020Q1. KvCV and CvMCV refer to bootstrap critical values for Kolmogorov-Smirnov and Cramer-von Mises respectively. Sample size: 52 observations. The null hypothesis of calibration is rejected if the test static is greater than the respective critical value.

Table 3.29: Uniformity tests of Vectors of PITs of Bank of England Fan Charts for GDP growth at horizon $h = 1, \ldots, 13$.

	Unfeasible				Feasible		
	z_{DHT}	z_{CS}	z_{KP}	z_M	SupMax	SupAverage	SupDes
KvStat	0.805	2.939	1.875	1.295	2.543	1.682	2.012
KvCV	1.684	1.686	1.413	1.769	1.692	1.581	1.572
KvCV Trad	1.360	1.360	1.360	1.360	1.360	1.360	1.360
CvMStat	0.208	2.728	1.357	0.812	2.922	1.333	1.87
CvMCV	1.138	1.506	0.724	0.851	0.83	0.705	0.669
CvMCVtrad	0.460	0.460	0.460	0.460	0.460	0.460	0.460
BerkowitzStat	264.361	85.813	20.928	279.86			
BerkowitzCV	5.9915	5.9915	5.9915	5.9915			
KnuppelStat	19.783	5.154	4.657	38.841			
KnuppelCV	9.4877	9.4877	9.4877	9.4877			

The table shows test statistics ("Stat") and critical values ("CV") obtained evaluating each forecast horizon spanning from 2004Q1 to 2020Q1. KvCV and CvMCV refer to bootstrap critical values for Kolmogorov-Smirnov and Cramer-von Mises respectively. Sample size: 52 observations. The null hypothesis of path calibration is rejected if the test static is greater than the respective critical value.

Table 3.30: Uniformity tests of Bank of England Fan Charts Unemployment at each horizon h.

	KvStat	KvCV	CvMStat	CvMCV	BerkowitzStat	BerkowitzCV	KnuppelStat	KnuppelCV
h=1	1.771	1.260	0.803	0.261	2.646	5.9915	5.800	9.4877
h=2	1.528	1.234	0.884	0.281	3.090	5.9915	3.807	9.4877
h=3	1.766	1.166	1.138	0.226	3.064	5.9915	2.936	9.4877
h=4	2.188	1.164	1.549	0.226	4.501	5.9915	2.893	9.4877
h=5	2.164	1.215	1.586	0.249	5.076	5.9915	2.777	9.4877
h=6	2.153	1.168	1.657	0.212	5.174	5.9915	3.507	9.4877
h=7	2.235	1.222	1.783	0.216	5.190	5.9915	4.024	9.4877
h=8	2.379	1.232	1.916	0.214	6.047	5.9915	4.048	9.4877
h=9	2.472	1.246	2.000	0.221	6.173	5.9915	4.633	9.4877
h=10	2.406	1.229	1.921	0.197	5.845	5.9915	5.186	9.4877
h=11	2.357	1.321	1.989	0.213	5.702	5.9915	5.050	9.4877
h=12	2.426	1.130	2.051	0.143	5.901	5.9915	5.521	9.4877
h=13	2.422	1.211	2.146	0.168	6.404	5.9915	5.445	9.4877

The table shows test statistics ("Stat") and critical values ("CV") obtained evaluating each forecast horizon spanning from 2004Q1 to 2020Q1. Sample size: 52 observations. KvCV and CvMCV refer to bootstrap critical values for Kolmogorov-Smirnov and Cramer-von Mises respectively. The null hypothesis of calibration is rejected if the test static is greater than the respective critical value.

Table 3.31: Empirical Correlations between horizons for Bank of England Fan Charts for Unemployment at horizon $h = 1 : H$ and the previous horizon forecast.

h=1	h=2	h=3	h=4	h=5	h=6	h=7	h=8	h=9	h=10	h=11	h=12	h=13
3.5081	0.2453	0.5355	0.4824	0.8780	0.3893	0.3572	0.4639	0.5607	0.3405	1.0255	0.3965	0.2922

Figure 3.3: Histogram of pits values for marginal distributions. Unemployment rate forecasts by Bank of England Fan charts at horizons $h = 1, \ldots, 13$.

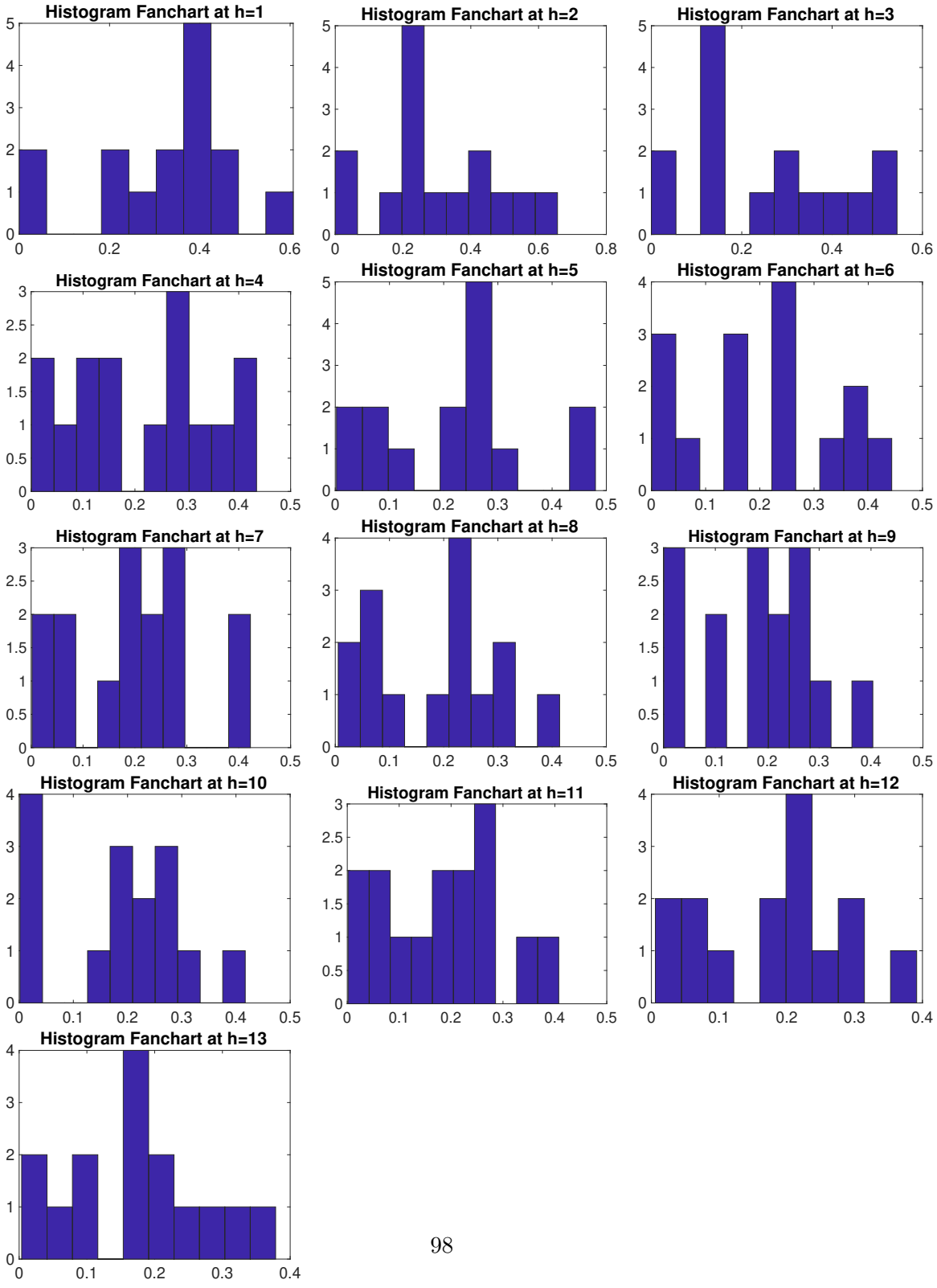


Table 3.32: Uniformity tests of Conditional distribution of Bank of England Fan Charts for Unemployment at horizon h give the previous horizon forecast.

	KvStat	KvCV	CvMStat	CvMCV	BerkowitzStat	BerkowitzCV	KnuppelStat	KnuppelCV
$z_{2\|1}$	1.8397	1.184	0.78556	0.18045	8.6239	5.9915	4.6138	9.4877
$z_{3\|2}$	2.0604	1.26	0.99098	0.18608	6.1253	5.9915	3.92	9.4877
$z_{4\|3}$	2.3509	1.4294	1.3604	0.16855	6.8338	5.9915	3.2413	9.4877
$z_{5\|4}$	2.2618	1.1685	1.4446	0.16629	5.2003	5.9915	2.5752	9.4877
$z_{6\|5}$	2.3741	1.252	1.6641	0.16525	5.1068	5.9915	3.9888	9.4877
$z_{7\|6}$	2.4012	1.2523	1.8638	0.15383	5.7323	5.9915	4.2178	9.4877
$z_{8\|7}$	2.5136	1.268	2.1916	0.13086	6.733	5.9915	5.1697	9.4877
$z_{9\|8}$	2.5303	1.2483	2.3216	0.1378	7.0781	5.9915	5.3121	9.4877
$z_{10\|9}$	2.5794	1.2503	2.4554	0.15202	8.3655	5.9915	5.2343	9.4877
$z_{11\|10}$	2.5536	1.2492	2.319	0.15358	7.2016	5.9915	5.1276	9.4877
$z_{12\|11}$	2.6917	1.2897	2.76	0.1257	10.0579	5.9915	4.7592	9.4877
$z_{13\|12}$	2.8699	1.2851	2.9601	0.12282	11.7947	5.9915	4.8369	9.4877

The table shows test statistics ("Stat") and critical values ("CV") obtained evaluating each forecast horizon spanning from 2004Q1 to 2020Q1. KvCV and CvMCV refer to bootstrap critical values for Kolmogorov-Smirnov and Cramer-von Mises respectively. Sample size: 52 observations. The null hypothesis of calibration is rejected if the test static is greater than the respective critical value.

Table 3.33: Uniformity tests of Vectors of PITs of Bank of England Fan Charts for GDP growth at horizon $h = 1, \ldots, 13$.

	Unfeasible				Feasible		
	z_{DHT}	z_{CS}	z_{KP}	z_M	SupMax	SupAverage	SupDes
KvStat	2.043	2.461	2.912	1.966	2.472	2.175	2.04
KvCV	1.412	1.269	1.176	1.413	1.321	1.215	1.214
KvCV Trad	1.360	1.360	1.360	1.360	1.360	1.360	1.360
CvMStat	1.561	2.629	3.135	1.559	2.145	1.648	1.434
CvMCV	0.273	0.189	0.409	0.262	0.282	0.218	0.233
CvMCVtrad	0.460	0.460	0.460	0.460	0.460	0.460	0.460
BerkowitzStat	94.282	15.41	47.665	91.515			
BerkowitzCV	5.9915	5.9915	5.9915	5.9915			
KnuppelStat	11.971	4.904	3.879	16.038			
KnuppelCV	9.4877	9.4877	9.4877	9.4877			

The table shows test statistics ("Stat") and critical values ("CV") obtained evaluating each forecast horizon spanning from 2004Q1 to 2020Q1. KvCV and CvMCV refer to bootstrap critical values for Kolmogorov-Smirnov and Cramer-von Mises respectively. Sample size: 52 observations. The null hypothesis of path calibration is rejected if the test static is greater than the respective critical value.

Chapter 4

Generalised Constraint for Predictive Distributions: a Bayesian Approach

4.1 Introduction

Policymakers and practitioners often wish to impose a desirable feature on predictive distributions (such as moments constraint, tails behaviour, shifts in support, etc...). Although constraining moments' distributions are well discussed in the literature (i.e. by exponential tilting Robertson et al. [2005], Krüger et al. [2017], Giacomini and Ragusa [2014]), it is often unclear which moment one should constrain. This paper aims to generalise the constraints to any desirable feature of the distribution. The constraints are imposed by approximating the target (constrained) distribution using a mixture of Student-t distributions. The mixture's parameters are estimated using the technique of Importance Sampling following Ardia et al. [2009]. The advantage of using this technique instead of exponential tilting is that the constraint can include any feature of the distribution, not just its moments.

This approach is applied to forecast US GDP during the Covid-19 pandemic: a combination of statical models is constrained to the higher probability of a negative growth event, recorded by the survey of professional forecasters (SPF). The predictive distribution is estimated via a Bayesian Inference of Quantile Regression model, following Kozumi and Kobayashi [2011] and it has been used to allow for non-linearity in the GDP of GDP growth. A set of six predictive distributions are obtained respectively using the National Financial Condition Index (NFCI) (following [Adrian et al., 2019]),

the University of Michigan Consumer Sentiment Index (ICS), the credit spread that measures the difference between BAA corporate bond yield and the ten-year treasury yield (following [Liu and Moench, 2016]), residential investments [Aastveit et al., 2019], and the unemployment rate [Marcellino, 2006]; in addition to lagged values of real GDP growth.

The resulting predictive distributions are then constrained to external information of the probability of negative growth. Such source is a judgmental forecast from a Central Bank or a survey of professional forecasters, motivated by evidence that such forecasts often provide useful information beyond that contained in econometric models (e.g. Ang et al. [2007], Faust and Wright [2013]). The data comes from the SPF variable "recess", which gives the mean responses for the probability of a decline in the level of chain-weighted real GDP in the current and following quarters. These are declines in the level of chain-weighted real GDP from one quarter to the next, beginning with a decline in the current quarter (the quarter in which the survey was conducted) compared with the quarter prior.

The SPF probability of negative growth is, on average, lower than the probability from forecasts; however, in rare events, the SPF probability of negative growth is higher and more accurate. One example is the negative growth in 2020:Q2. The impact of the COVID-19 pandemic on growth at 2020:Q2 was difficult to forecast with the usual predictors, while SPF displays a spike in the probability of this adverse event. This paper imposes the distribution to have a probability of being negative equal to the one express by SPF, and this constraint improves the forecast-ability of the quantile forecasts.

This paper uses Importance Sampling (IS) methodology to include external information about the probability of a specific support region. Importance sampling was first proposed by Hammersley et al. [1965] and lately introduced in econometrics by Kloek and Van Dijk [1978]. Importance sampling is an inference technique for estimating a target distribution using an instrumental or "candidate" distribution. IS uses weights to correct the fact that we sample from a candidate distribution $q(x)$ instead of the target distribution $p(x)$. Similar to the entropic tilting optimisation, this paper modifies the distribution by reweighing draws; however, instead of drawing them from the predictive distribution (the target distribution in IS framework), they are drawn from a candidate distribution that approximate the target satisfying the desired constraints. The methodology adopted here is a modified version of Ardia et al. [2009], where the candidate distribution is a mixture of Student-t. Once applied the constraint through importance sampling, the density forecasts are able to give a probability different from zero to the actual realisation.

The rest of the paper is organised as follows: Section (4.2) presents the dataset used; Section (4.3) introduces the quantile regression model used to obtain the predictive distributions; Section (4.4) discusses the importance sampling estimation technique and the inclusion of external information as constraint; finally, Section (4.5) concludes.

4.2 Data

This paper considers $K = 5$ different predictors. These are leading indicators that cover a broad range of the macroeconomy, and that earlier studies have found to be useful for predicting GDP growth and recessions. A vast amount of research has shown that various economic and financial variables contain predictive information about future economic recessions and downturns. A recent study by Adrian et al. [2019] argues that financial conditions are particularly informative above future downside macroeconomic risk. In addition to these studies, several other variables have also been regarded as leading recession indicators for GDP growth and recessions, including stock prices (Estrella and Mishkin [1998] and Stock and Watson [2003]), the index of leading economic indicators (Berge and Jordà [2011] and Stock and Watson [1989]), oil prices (Hamilton [1983, 1996] and Ravazzolo and Rothman [2013, 2016]), survey data (Hansson et al. [2005], Claveria et al. [2007]); and residential investments (Aastveit et al. [2019]).

The following five variables are predictors in quantile regression: the National Financial Condition Index (NFCI), the University of Michigan Consumer Sentiment Index (ICS), the credit spread that measures the difference between BAA corporate bond yield and the 10-year treasury yield, residential investments, the unemployment rate.

The list of the $K = 5$ predictors with their data source and the data transformation is presented in Table (4.1). All our data series covers the period 1973Q1-2021Q1. Thus, the full out-of-sample forecasting evaluation period runs from 1993Q1-2021Q1. Predictive densities are updated recursively for forecast horizons $H = 1$ (one quarter ahead) based on models that are estimated using an expanding window.

As mentioned before, additional external information is added to this predictive distribution: a measure of the probability of negative growth. The data comes from the SPF variable "recess", which gives the mean responses for the probability of a decline in the level of chain-weighted real GDP in the current and following quarters. These are declines in the level of chain-weighted real GDP from one quarter to the next, beginning with a decline in the current quarter (the quarter in which the survey was conducted) compared with the quarter prior. The Federal Reserve Bank of Philadelphia publishes

this variable four quarters ahead of forecasts plus the current quarter. This paper will forecast only a one-quarter-ahead forecast.

4.3 Quantile regression models to forecast US GDP growth

This section presents the quantile regression model used to estimate quantile predictors for GDP growth. Section (4.3.1) introduces the quantile regression framework; Section (4.3.2) derives the predictive distributions from the parameters' estimates and displays the prior and posterior distributions used in Bayesian inference. Finally, Section (4.3.3) discusses the features of quantile regression distributions, compares their probability of negative growth to SPF and presents the failure of these models to forecast the plunge of GDP growth following the COVID-19 pandemic out-spring.

4.3.1 Quantile regression models

Quantile regression is a statistical procedure that explores the non-linear relationship between quantiles of the response distribution and available covariates. Quantile regression has a long tradition in the econometric and statistic literature since Koenker and Bassett Jr [1978] seminar paper. The idea behind quantile regression rises from the doubt that mean or conditional expectation (typical of linear regression) would adequately characterize statistical relationships among variables. The quantile regression is then a way to handle non-linearity in the data. In principle, one would like to know the entire conditional distribution function that relates the dependent variable with the predictors. However, in practice, quantile regression is based on minimizing sums of asymmetrically weighted absolute residuals.

Quantile regression generalizes traditional least-squares regression by fitting a distinct regression line for each quantile of the variable of interest distribution. However, least squares regression only produces coefficients that allow us to fit the mean of the dependent variable conditional on some explanatory variables. In that respect, quantile regression may be more appropriate for making inferences about predictive distributions and assessing the forecast uncertainty.

Consider the quantile regression model given by

$$y_t = x_t \beta_\tau + \varepsilon_t \tag{4.1}$$

for $t = 1, \ldots, T$. Here $\tau = [1, \ldots, 10]$ identifies the quantile, ε_t is the error term whose

distribution (with density, $f(\tau(\cdot))$ is restricted to have the τ^{th} quantile equal to zero, that is, $\int_{-\infty}^{0} f(\tau(t))dt = \tau$. Traditionally, quantile regression estimation for β_τ proceeds by minimizing:

$$\sum_{t=1}^{T} \rho_\tau(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_\tau), \tag{4.2}$$

where $\rho_\tau(\cdot)$ is the check (or loss) function defined by

$$\rho_\tau(u) = (\tau - I(u < 0)\}u \tag{4.3}$$

and $I(\cdot)$ denotes the usual indicator function. Since a set of quantiles often provides more complete description of the response distribution than the mean, quantile regression offers a practically important alternative to classical mean regression.

Since, however, the check function is not differentiable at zero, we cannot derive explicit solutions to the minimization problem. To solve this issue, we follow Kozumi and Kobayashi [2011] approach to Bayesian quantile regression models using the asymmetric Laplace distribution for the error term.

4.3.2 Predictive quantile function for GDP

For each variable $\{k = 1, \ldots, K\}$, a predictive distribution for GDP growth is obtained using a ARDL model:

$$y_{t+h,\tau,k} = \boldsymbol{x}_{t,k}'\boldsymbol{\beta}_\tau + \sigma\theta z_{t+h} + \sigma\tau\sqrt{z_{t+h}}u_{t+h} \tag{4.4}$$

where $\boldsymbol{x}_{t,k}'$ is the vector of lagged values of y_t (with maximum lag r) and of one of the K predictors (with maximum lag p). In the empirical application, the number of lags p and r are selected using BIC selection criterion with a maximum of four lags. The error term takes the form $\varepsilon_{t+1} = \sigma\theta z_{t+h} + \sigma\tau\sqrt{z_{t+h}}u_{t+h}$ as in Kozumi and Kobayashi [2011], where $z_{t+h} \sim Exponential(1)$, u_{t+h} has a standard normal distribution and $\sigma \sim IG(n_0/2, s_0/2)$. Finally, we will focus on forecast horizons $h = \{1\}$ in our empirical application.

Bayesian Inference

We consider the linear model given by:

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_\tau + \varepsilon_t \tag{4.5}$$

where τ denotes the quantile and assume that ε_t has the asymmetric Laplace distribution with density:

$$f_\tau(\varepsilon_t) = \tau(1-\tau)\exp\{-\rho_\tau(\varepsilon_t)\} \tag{4.6}$$

where $\rho_\tau(\varepsilon_t) = \varepsilon_t\{\tau - I(\varepsilon_t < 0)\}$. The mean and variance of the asymmetric Laplace distribution are given by:

$$\mathbb{E}(\varepsilon_t) = \frac{1-2\tau}{\tau(1-\tau)} \quad Var(\varepsilon_t) = \frac{1-2\tau+2\tau^2}{\tau^2(1-\tau)^2} \tag{4.7}$$

To develop a Gibbs sampling algorithm for the quantile regression model, we use a mixture representation based on exponential and normal distribution by Kotz et al. [2012]. Following Kozumi and Kobayashi [2011] the error term ε_t as:

$$\varepsilon_t = \sigma\theta z_t + \sigma\delta\sqrt{z_t}u_t \tag{4.8}$$

where σ is the scale parameter, $z_t \sim Exponential(1)$ and $u_t \sim N(0,1)$ are mutually independent, and:

$$\theta = \frac{1-2\tau}{\tau(1-\tau)} \qquad \delta^2 = \frac{2}{\tau(1-\tau)} \tag{4.9}$$

From this we can rewrite y_t as:

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_\tau + \sigma\theta z_t + \sigma\delta\sqrt{z_t}u_t \tag{4.10}$$

To facilitate the inference, we adopt the reparametrization by Kozumi and Kobayashi [2011]:

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_\tau + \theta\upsilon_t + \delta\sqrt{\upsilon\sigma}u_t \tag{4.11}$$

where $\upsilon_t = \sigma z_t$. We assume that $\boldsymbol{\beta}_\tau \sim \mathcal{N}(\boldsymbol{\beta}_{\tau 0}, \boldsymbol{B}_{\tau 0}$ and $\sigma \sim \mathcal{IG}(n_0 72, s_0/2)$, where $\mathcal{IG}(a,b)$ denotes an inverse Gamma distribution with parameters a and b. The conditional distribution of y_t given z_t is normal with mean $\boldsymbol{x}_t'\boldsymbol{\beta}_\tau + \theta\upsilon_t$ and variance $\delta^2\upsilon_t$. The joint density of $\boldsymbol{y} = (y_1, \ldots, y_T)'$ is given by:

$$f(\boldsymbol{y}|\boldsymbol{\beta}_\tau, \boldsymbol{z}, \sigma) \propto \left(\prod_{t=1}^{T} \upsilon_t^{-1/2}\right) exp\left\{-\sum_{t=1}^{T} \frac{(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_\tau - \theta\upsilon_t)^2}{2\delta^2\upsilon_t}\right\}, \tag{4.12}$$

We need to sample $\boldsymbol{\beta}_\tau$, $\boldsymbol{v} = (\upsilon_1, \ldots, \upsilon_t)'$ and σ from their conditionals distributions: The full conditional density of $\boldsymbol{\beta}_\tau$ is given by:

$$\boldsymbol{\beta}_\tau|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{v}, \sigma \sim \mathcal{N}(\bar{\beta}_\tau, \bar{V}_\beta), \tag{4.13}$$

where:

$$\bar{V}_{\beta}^{-1} = \left(\sum_{t=1}^{T} \frac{x_t' x_t}{\delta^2 \sigma \upsilon_t} + \boldsymbol{B}_{\tau_0}^{-1} \right) \quad \bar{\beta}_{\tau} = \bar{V}_{\beta} \left[\sum_{t=1}^{T} \frac{x_t (y_t - \theta \upsilon_t)}{\delta^2 \sigma \upsilon_t} + \boldsymbol{B}_{\tau_0}^{-1} \boldsymbol{\beta}_{\tau_0} \right] \tag{4.14}$$

and assuming a normal prior

$$\boldsymbol{\beta}_{\tau} \sim \mathcal{N}(\boldsymbol{\beta}_{\tau_0}, \boldsymbol{B}_{\tau_0}) \tag{4.15}$$

where $\boldsymbol{\beta}_{\tau_0}$ and $\boldsymbol{B}_{\tau_0}$ are the prior mean and variance covariance matrix of $\boldsymbol{\beta}_{\tau}$. Priors for the quantile betas have been chosen to have mean zero and variance 1000.

The full conditional distribution of υ_t is proportional to:

$$\upsilon_t | \boldsymbol{y}_t, \boldsymbol{x}_t \boldsymbol{\beta}_{\tau} \sigma \sim GIG(1/2, \xi_t, \gamma_t) \tag{4.16}$$

where:

$$\xi_t = (y_t - \boldsymbol{x}_t' \boldsymbol{\beta})^2 / \delta^2 \sigma \quad \gamma_t^2 = 2/\sigma + \theta^2 / \delta^2 \sigma \tag{4.17}$$

and where GIG denotes the Generalized Inverse Gaussian distribution which pdf for the general case $GIG(v, a, b)$ is:

$$f(x|v, a, b) = \frac{(b/a)^v}{2K_v(ab)} x^{v-1} exp\left\{ -1/2(a^2 x^{-1} + b^2 x) \right\}, \quad x > 0, \quad -\infty < v < \infty, \quad a, b \leq 0 \tag{4.18}$$

and K_v is a modified Bessel function of the third kind. By noting that $\upsilon_t \sim \mathcal{E}(\sigma)$, the full conditional density of σ is proportional to:

$$\sigma | \boldsymbol{y}_t, \boldsymbol{x}_t \boldsymbol{\beta}_{\tau} \boldsymbol{v} \sim IG(n/2, s/2) \tag{4.19}$$

where $n = n_0 + 3n$ and $s = s_0 + 2 \sum_{t=1}^{T} \upsilon_t + (y_t + \boldsymbol{x}_t' \boldsymbol{\beta}_{\tau} + \theta \upsilon_t)^{`2} / \tau^2 \upsilon_t$

The posterior distribution is calculated using 8000 replications after a burn-in period of 4000 replications. For each replication, a quantile predictive function is calculated for the following:

$$p(y_{t+h} | \boldsymbol{y}_t, \boldsymbol{\vartheta}) = \int p(y_{t+h} | \boldsymbol{\vartheta}, \boldsymbol{y}_t) p(\boldsymbol{\vartheta} | \boldsymbol{y}_t) d\boldsymbol{\vartheta} \tag{4.20}$$

where $p(\boldsymbol{\vartheta} | \boldsymbol{y}_t)$ corresponds to the posterior distributions of the parameters' set $\boldsymbol{\vartheta} = \{\boldsymbol{\beta}_{\tau}, \theta, \delta, \upsilon_t\}$, and $p(\boldsymbol{y}_{t+h} | \boldsymbol{\vartheta})$ corresponds to:

$$p(\boldsymbol{y}_{\tau, t+h} | \boldsymbol{\vartheta}) = \int \mathcal{N}(y_{t+1} | x_T' \boldsymbol{\beta}_{\tau} + \theta z_T \upsilon_T, \delta^2 z_T \upsilon_T) d\boldsymbol{\beta}_{\tau} d\upsilon_t \tag{4.21}$$

4.3.3 Empirical Results - Quantile regression

Six predictive distributions are obtained from the previous section by fitting a kernel function on the quantile points. Figure (4.1) shows the probability of negative growth using each of those densities compared to the same information disclosed by the Survey of Professional Forecasters. All six models forecast, on average gives a higher probability of an adverse event than the experts. For some events, the experts predict a higher likelihood of negative GDP growth. We can observe three spikes: during the financial crisis (last quarter of 2008), the 2001 recession and the COVID-19 pandemic (second quarter of 2020). It seems that, for these particular events, accounting for the SPF probability of negative growth would be beneficial for the forecast accuracy. Figure (4.2) displays density forecasts for the first quarter of COVID-19 pandemic: 2020:Q2. All the predictive distributions available fail to cover the support of the actual realisation for GDP growth.

Including the SPF probability of negative growth in the forecasting model appears not only to better forecast extreme events (such as recessions) but also in the rest of the dataset to reduce the variance of density forecasts. The following section will present the methodology used to incorporate SPF probability of negative growth, constraining the predictive distributions.

4.4 Inclusion of external information in predictive density using Importance Sampling

This section presents the Importance Sampling methodology used to include external information about the probability of a specific support part. After a short excursus on importance sampling in Section (4.4.1), the section presents the candidate distribution proposed, called adaptive mixture of Student-t distributions in Section (4.4.2), a slightly changed version from the one proposed by Ardia et al. [2009]. In this version of the algorithm, **Step 0 and 1** allows external information to impact the IS weight distribution. Steps (2a, 2b, 2c) rest invariant to Ardia et al. [2009].

4.4.1 Background on importance sampling and motivation

Importance sampling was first proposed by Hammersley et al. [1965] and lately introduced in econometrics by Kloek and Van Dijk [1978]. Importance sampling is an inference technique for estimating a target distribution using an instrumental or "candidate"

distribution. Then, IS uses weights to correct the fact that we sample from the instrumental distribution $q(y)$ instead of the target distribution $p(y)$. It is often employed to approximate (target) distributions that are not easy to sample from. It is based on the following relationship:

$$\mathbb{P}(y \in \mathcal{Y}) = \int_{\mathcal{Y}} p(y)dy = \int_{\mathcal{Y}} q(y)\frac{p(y)}{q(y)}dy = \int_{\mathcal{Y}} q(y)w(y)dy \tag{4.22}$$

for all $q(\cdot)$, such that $q(y) > 0$ for (almost) all y with $\phi(y) > 0$,w(y)=$\frac{p(y)}{q(y)}$. We can generalise this identity by considering the expectation $\mathbb{E}_p(g(\mathcal{Y}))$ of a measurable function g:

$$\mathbb{E}_p(g(\mathcal{Y})) = \int p(y)g(y)dy = \int q(y)\frac{p(y)}{q(y)}g(y)dy = \int \mathbb{E}_q(w(y) \cdot g(y)) \tag{4.23}$$

where $\mathbb{E}_p$ denotes the expectation with respect to the target density $p(y)$ and $\mathbb{E}_q$ denotes the expectation with respect to the importance approximation $q(y)$. The importance sampling estimator of $\mathbb{E}_p(g(\mathcal{Y}))$ is obtained as the sample counter-part of the right-hand side of Equation (4.23):

$$\hat{g} = \frac{\sum_{i=1}^{N} g(y_i)w(y_i)}{\sum_{i=1}^{N} w(y_i)} \tag{4.24}$$

where $\{y_i|1, \ldots, N\}$ is a sample of draws from the importance density $q(y)$.

4.4.2 Adaptive mixture of Student-*t* distributions

In this paper, the target distribution $p(y_{t+h})$ in Equation (4.20) has a Gaussian kernel smoother of quantile predictions:

$$\phi(y_{t+h}|\hat{\boldsymbol{\vartheta}}) \propto \exp\{-1/2\frac{(y_t - x_t'\hat{\beta}_q - \theta\hat{z}_t)^2}{(\tau\sqrt{\hat{z}_t})^2}\} \tag{4.25}$$

where $\hat{\boldsymbol{\vartheta}} = \{\hat{\boldsymbol{\beta}}_\tau, \theta, \delta, \hat{z}_t\}$ subject to the probability of negative support being equal to the probability of negative growth published by SPF, i.e.:

$$\int_{-\infty}^{0} \phi(y_{t+h}|\hat{\boldsymbol{\vartheta}})dy_{t+h} = \text{Prob}(y_{t+h} \leq 0) \tag{4.26}$$

The problem can be solved in several ways, such as entropic tilting optimization or chance-constrained programming. The first is an optimization problem subject to a constraint that regards one of the distribution's moments; the second is a modelling tool that allows incorporating uncertainty into optimization problems using Sample Average

Approximation (SAA). This paper proposes a flexible tool that can impose any constraint on distribution and not use ad-hoc procedures based on the constraining preference at hand. Similar to the entropic tilting optimization, this paper considers modifying the distribution by reweighing draws; however, instead of drawing from $p(y_{t+h}|\hat{\boldsymbol{\vartheta}})$ (the target distribution in this framework) they are drawn from a candidate distribution that approximate the target satisfying the desired constraints.

The candidate distribution is an adaptive mixture of Student-t developed in Hoogerheide et al. [2007]. The density of a mixture of student-t distributions can be written as:

$$q(y_{t+h}) = \sum_{k=1}^{K} \eta_k t(y_{t+h}|\mu_k, \Sigma_k, \nu) \tag{4.27}$$

where η_k for $(k = 1, \ldots, K)$ are the mixing probabilities of the Student-t components, $0 \leq \eta_k \leq 1$, $\sum_{k=1}^{K} = 1$, and $t(y_{t+h}|\mu_k, \Sigma_k, \nu)$ is a Student-t density with mode μ_k, scale Σ_k, and ν degrees of freedom:

$$t(y_{t+h}|\mu_k, \Sigma_k, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})(\pi\nu)^{1/2}} \times (\Sigma_k)^{-1/2}(1 + \frac{(y_{t+h} - \mu_k)^2\Sigma_k^{-1}}{\nu})^{-(\nu+1)/2} \tag{4.28}$$

The adaptive mixture approach determines K, η_k, μ_k and Σ_k $(k = 1, \ldots K)$ based on a kernel function $\phi(y_{t+h})$ of the target density $p(y_{t+h})$. It consists of the following steps:

Step 0 - Initiation Following Ardia et al. [2009], compute the mode μ_1 and the scale Σ_1 of the first Student-t distribution in the mixture as $\mu_1 = \arg\max_{y_{t+h} \in Y} \log\phi(y_{t+h})$, the mode of the log kernel function, and Σ_1 as minus the Hessian of $\log \phi(y_{t+h})$ evaluated at its mode μ_1. Then draw a set of N_s points $\mathcal{Y}_i$ $(i = 1, \ldots, Ns)$ from the first candidate density $q(y_{t+h}) = t(\mu_1, \Sigma_1, \nu)$, with small ν to allow for fat tails. The degrees of freedom ν are chosen to be equal to one $(\nu = 1)$ since it enables the method to deal with fat-tailed target distribution and it makes it easier for the iterative procedure to detect modes that are far apart.

This paper modifies the algorithm by including a constraint in this step: all the draws from the candidate distributions have to satisfy the constraint that the probability of the ith draw being negative or equal to zero is equal to:

$$\mathbb{P}(\mathcal{Y}_i \leq 0) = \int_{-\infty}^{0} \phi(y_{t+h}|\boldsymbol{\vartheta}) dy_{t+h} \tag{4.29}$$

the algorithm adopted here is designed in such a way that it draws from the candidate

as many times as it is necessary to satisfy the condition that the number of negative draws has to be equal to $N_s/\mathbb{P}(\mathcal{Y}_i \leq 0)$ (and so the number of positive draws have to be equal to $N_s - (N_s/\mathbb{P}(\mathcal{Y}_i \leq 0))$). Any draws that do not respect these conditions are discarded.

After that, add components to the mixture, iteratively, by performing the following steps:

Step 1 - Evaluate the distribution weights Following Ardia et al. [2009], in this step the points $\mathcal{Y}_i$ $(i = 1, \ldots, Ns)$ are reweighed based on the Importance Sampling weights:

$$w(\mathcal{Y}_i) = \frac{\phi(\mathcal{Y}_i)}{q((\mathcal{Y}_i))} \quad \text{for } i = 1, \ldots, N_s \tag{4.30}$$

Where $\phi(\mathcal{Y}_i)$ denotes the Kernel of target distribution $p(y_{t+h})$ evaluated at draws $\mathcal{Y}_i$ and $q((\mathcal{Y}_i))$ denotes the candidate distribution evaluated at each draw $\mathcal{Y}_i$. $w(\mathcal{Y}_i)$ is then a measure of the distance between target and candidate kernels computed at the same points. When the two kernel are identical, the weight is equal to one and the candidate distribution has approximated the target distribution. However, in the case of this paper, the IS is not used to approximate the target distribution but to constrain it using a candidate. For this reason, here the structure of IS weights is a bit different:

$$w(\mathcal{Y}_i) = \begin{cases} \phi(\mathcal{Y}_i)/q((\mathcal{Y}_i)), & \text{if } \mathcal{Y}_i > 0 \\ 1, & \text{if } \mathcal{Y}_i \leq 0 \end{cases} \tag{4.31}$$

In the positive part of the support, IS weights are computed in the traditional way: as the ratio between kernels evaluated at (positive) points. For the negative part of the support, the weights are imposed to be equal to one to ensure that the negative draws respect conditions imposed in Step 0. To describe the procedure in a broader way, it ensures that the negative part of the support is estimated according to SPF probability of negative draws (step 0), while the positive part of the support is estimated according to the target distribution that in this case is equal to the predictive distribution obtained via quantile regression.

Step 2a - Iterate on the number of components Add another Student-t distribution with density $t(\mu_k, \Sigma_k, \nu)$ to the mixture with

$$\mu_k = \arg\max_{y_{t+h} \in Y} \log w(I(\mathcal{Y} > 0)) \tag{4.32}$$

where I denotes the indicative function; and Σ_1 as minus the Hessian of $\log w(I(\mathcal{Y} > 0))$. The idea behind this step is to choose the initial value for the maximization procedure for computing μ_k as the point $\mathcal{Y}_i$ with the highest weight $\{w(\mathcal{Y}_i), i = 1, \ldots, N_s\}$. Compared to Ardia et al. [2009], the algorithm employed here chooses the initial value of maximisation in the subsample of positive draws under the assumption that the probability of negative growth is never higher than 50%. A weighting system that accounts for it would be designed in the subsequent development of this paper.

The rest of the algorithm follows Ardia et al. [2009].

Step 2b - Optimize the mixing probabilities Choose the probabilities η_k for $k = (1, \ldots, K)$ in the mixture $q(\mathcal{Y})$ defined in Equation (4.27) by minimizing the coefficient of variation of the importance sampling weights. First draw again N_s points $\mathcal{Y}_{k,i}$ from each component $t(\mathcal{Y}_{k,i}|\mu_k, \Sigma_k, \nu)$ the minimize:

$$\mathbb{E}[w(\mathcal{Y})^2]/\mathbb{E}[w(\mathcal{Y})]^2 \tag{4.33}$$

with respect to η_k $(k = 1, \ldots, K)$, where:

$$\mathbb{E}[w(\mathcal{Y})^2] = \frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{k=1}^{K} \eta_k w(\mathcal{Y}_{k,i})^2 \tag{4.34}$$

and

$$w(\mathcal{Y}_{k,i}) = \frac{\phi(\mathcal{Y}_{k,i})}{\sum_{k=1}^{K} \eta_k t(\mathcal{Y}_{k,i}|\mu_k, \Sigma_k, \nu)} \tag{4.35}$$

Step 2c - Draw from the mixture Draw a sample of N_s points $\mathcal{Y}_i$, $(i = 1, \ldots, N_s)$ from the new mixture of Student-t distributions,

$$q(\mathcal{Y}_i) = \sum_{k=1}^{K} \eta_k t(\mathcal{Y}_i|\mu_k, \Sigma_k, \nu) \tag{4.36}$$

4.4.3 Empirical Results - Importance Sampling

Figure (4.3) displays the density forecasts for 2020:Q2 using the adaptive mixture of Student-t distributions in the previous Equation. Using external information to reweight the draws from the distributions obtained in Section (4.3.1) leads to a shift of the distribution in the negative region, and the mixture of Student-t distribution allows for multimodality. Although the probability at the realisation (in red) is still relatively low, all predictive densities now cover its support.

4.5 Conclusions and Further Developments

This paper proposes a flexible tool to impose a desirable feature to predictive distributions. The constraints are imposed by approximation of the target (constrained) distribution using a mixture of Student-t distributions. The mixture's parameters are estimated using the technique of Importance Sampling. The advantage of using this technique instead of exponential tilting is that the constraint can include any feature of the distribution, not just its moments. This paper applies the technique to constrain forecasts for US GDP growth. The predictive distribution is estimated using a set of predictors and contained a probability of negative growth equal to the one published by the Survey of Professional Forecasters. Predictive distributions are estimated using quantile regression models with Bayesian inference. The SPF probability of negative growth is, on average, lower than the probability from forecasts; however, in rare events, the SPF probability of negative growth is higher and more accurate. One example is the GDP negative growth in 2020:Q2. The impact of the COVID-19 pandemic on growth at 2020:Q2 was still difficult to forecast with the predictors used in the past to forecast GDP growth, while SPF displays a spike in the probability of negative events. This paper then imposes the distribution to have a probability of being negative equal to the one express by SPF, and this constraint improves the forecastability of the quantile forecasts. The density distributions give a probability different from zero to the realisation.

Despite the encouraging results obtained, further work is necessary to improve the approach proposed. First, the estimation of predictive distribution parameters in the IS; this would allow for variability parameters to accommodate the external information of SPF probability of negative growth; second, an improvement of the IS weighting system that currently is based on a series of assumptions that are not always verified; third, apply the technique to a combination of the predictive distributions.

Table 4.1: Description of Data Series

Label	Trans	Period	Real-Time	Description	Source
rgdp	Δln	59:Q1-20:Q2	73Q1-20Q2	Real GDP growth, sa	AL
NFCI	level	71:Q1-20:Q2	11Q2-20Q2	National Financial Conditions Index	Chicago Fed
ICS	level-100	60:Q1-20:Q2	98Q3-20Q2	Consumer Sentiment Index	AL
CrSpread	Level	53:Q2-20:Q2	none	Credit Spread: BAA corporate bond yield - 10-year treasury	F
U	Δ log	48:Q1-20:Q2	65Q4-20Q2	Unemployment rate	AL
ResInv	$\Delta\%$	47:Q2-20:Q2	65Q4-20Q2	Real Gross Private Domestic Investment: Fixed Investment: Residential	AL
Recess	level	68:Q4-21:Q2	68:Q4-21:Q2	Recess	PFed

Notes: Sources abbreviated as "F" denotes Federal Reserve Economic Data (FRED), as "AL" denotes Federal Reserve Economic Real-Data (ALFRED) dataset and as "PFed" the Federal Reserve Bank of Philadelphia.

Figure 4.1: Probability of negative growth over 1996:Q2 to 2021Q1: comparison between quantile forecasts and SPF.

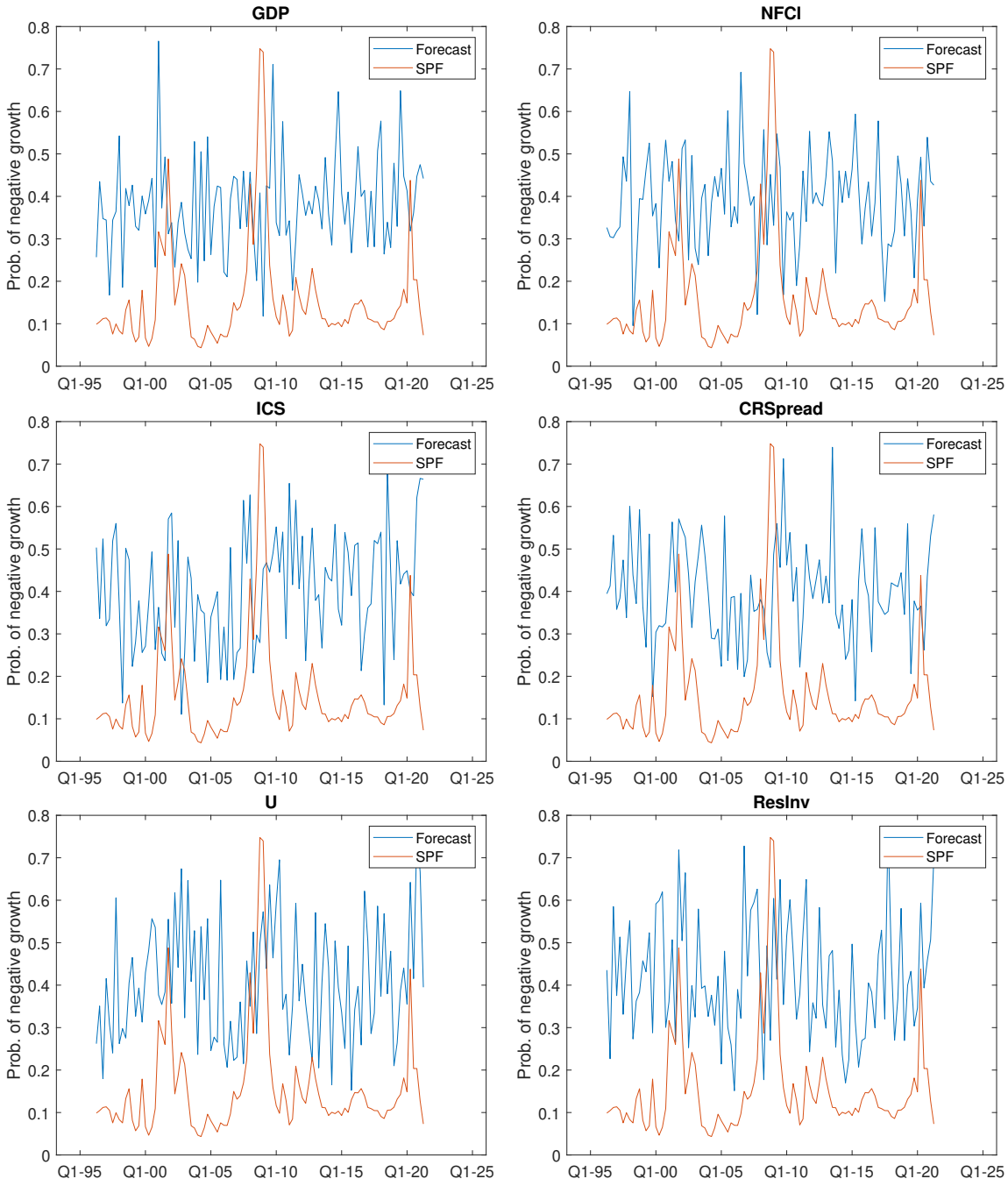


Figure 4.2: Density forecast obtained using Quantile Regression for the 6 alternative models for 2020:Q2 and the actual realisation.

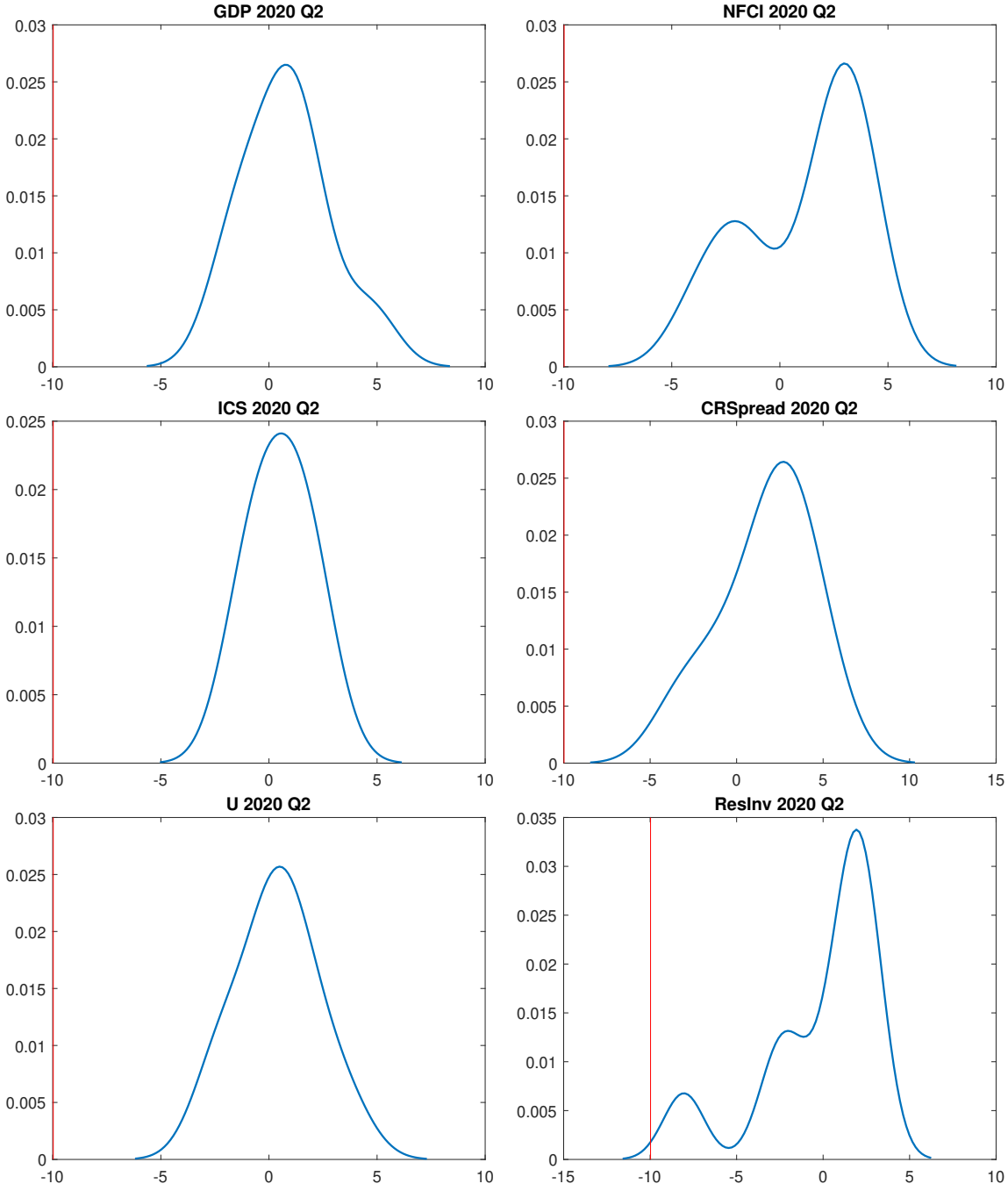
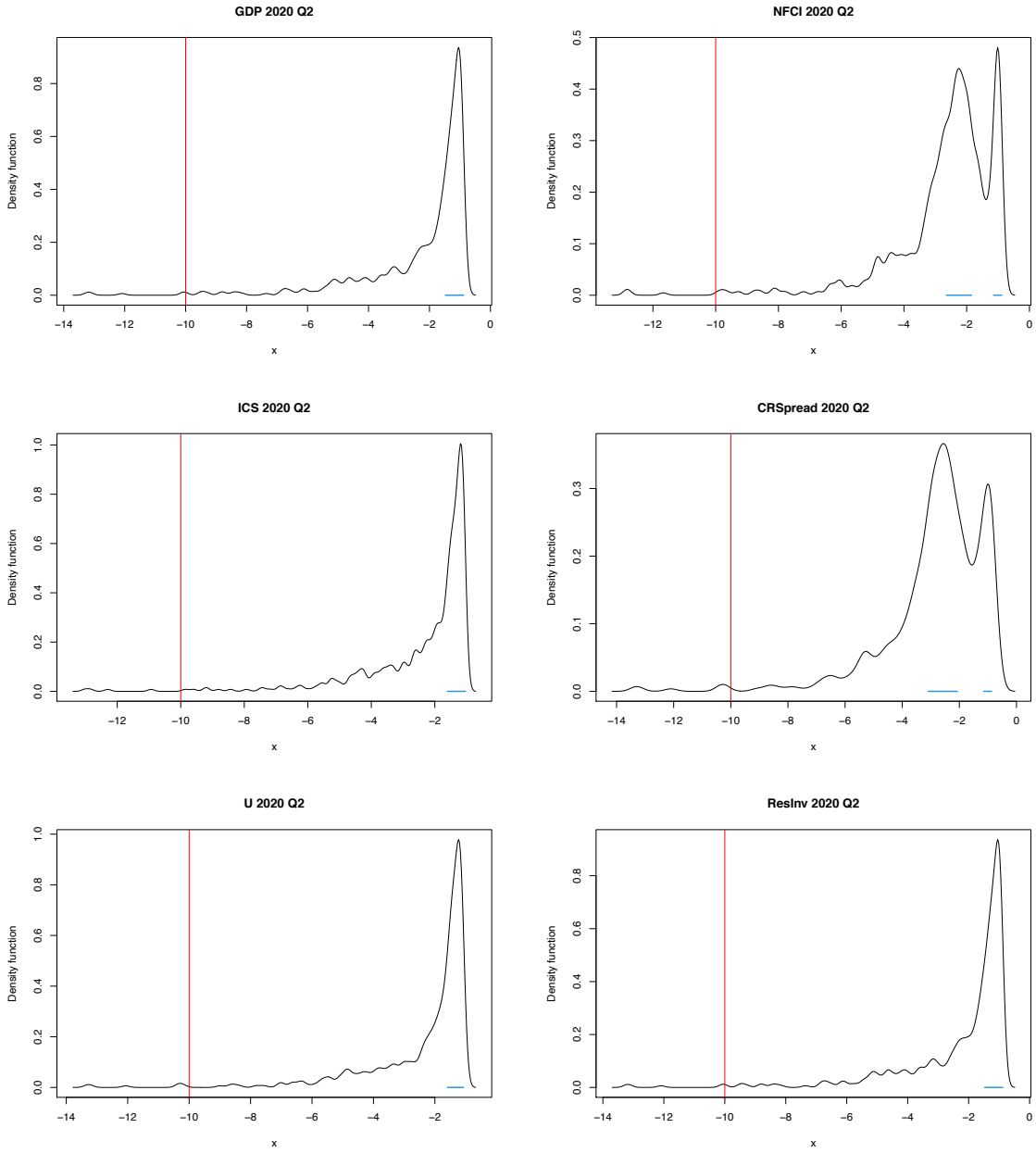


Figure 4.3: Density forecast obtained using Adaptive mixture of Student-t distribution for the 6 alternative models for 2020:Q2 and the actual realisation.



Chapter 5

Conclusions

In this thesis, I focus on combining, calibrating and constraining predictive distributions for macroeconomic time-series. This thesis is divided into three chapters. The first chapter investigates the effect of inference approaches on forecast accuracy of the density combination. The second proposes an absolute evaluation criterion for path density forecasts. The third focuses on the inclusion of external information on model-based predictive distributions by constraining it.

Chapter (2) proposes a comparison between two approaches to the combination of density forecasts. The first, called "two-step" is arguably the most popular in applied econometrics and finance, and it takes individual probability forecasts as given and then combines them. The second, called "one-step" is studied extensively by the statistical literature, and it estimates forecasts' parameters and combination weights simultaneously. First, I propose an empirical exercise to investigate the forecast accuracy of the two approaches. The application consists of forecasting US real output growth and inflation, combining 31 individual models. The empirical exercise leaves us with no clear indication over which combination approach is the most accurate. Then, I tried to shed light on understanding what affects the different performances in controlled environments. Several types of DGPs tried to shed light on understanding what affects the different performances in controlled environments. The main takeaway is that the trade-off between parameter estimation noise and forecast accuracy typical of the one-step approach is crucial. However, this trade-off is overcome by the one-step approach by its ability to account for the dependence between the mixture's components.

Chapter (3) proposes an absolute evaluation criterion for path density forecasts. One example of the employment of path density forecast is the Central Bank's fan charts. First, the chapter defines the path density forecast and discusses a series of testing

strategies. We identified two main tests that depend on information about horizon dependence. If the researcher has information about the dependence and can build conditional forecast distributions, then he can use vectors of pits, among which the better sized and more powerful vectors are z_M and z_{DHT}; If the research does not have any information, he can use a vector of marginal distributions and some "sup tests". Among the "sup tests", the strict calibration test displays the best properties. The choice of test statistics depends on several aspects: Kolmogorov-Smirnov and Cramer-von Mise have better size and power when employing the bootstrapped version; choosing these two tests will depend on the computational issues that the estimation entitles. Berkowitz test is undersized, and the Knuppel test works only with a large enough (i.e. T=100) sample size, but it is the most powerful. To answer whether the Bank of England fan chart is calibrated, we applied the tests to Bank of England fan charts published from 2004Q1 up to 2020Q1. From our analysis, we can say that the calibration of path density forecast for inflation rate is not rejected by the majority of our tests, either horizon-by-horizon or jointly; for GDP growth rate and unemployment is rejected.

Chapter (4) proposes a flexible tool to impose a desirable feature to predictive distributions. This chapter is motivated by the need for policymakers and practitioners to impose a desirable feature on predictive distributions. Examples of the feature can be moments constraint, tails behaviour, shifts in support, etc. Although constraining moments' distributions are well discussed in the literature (i.e. by exponential tilting Robertson et al. [2005], Krüger et al. [2017], Giacomini and Ragusa [2014]), it is often unclear which moment one should constrain. This chapter aims to generalise the constraints to any desirable distribution feature by approximating the target (constrained) distribution using a mixture of Student-t distributions. The mixture's parameters are estimated using the technique of Importance Sampling following Ardia et al. [2009]. The advantage of using this technique instead of exponential tilting is that the constraint can include any feature of the distribution, not just its moments, without increasing the variance. This chapter focuses on constraining the probability of a part of the distribution support to be equal to some external information. The technique is applied to constrained forecasts for US GDP growth. First, a set of density forecasts are obtained by quantile regression, then each of these is constrained to SPF probability of negative growth. Particularly interesting is the application to the COVID-19 pandemic. The impact of the COVID-19 pandemic on growth at 2020:Q2 was difficult to forecast with the usual predictors, while SPF displays a spike in the probability of this adverse event. This paper imposes the distribution to have a probability of being negative equal to the one express by SPF, and this constraint improves the forecast-ability of the quantile

predictors.

In addition to this work, I have an ongoing project with two researchers at Norges Bank: Knut Are Aastveit and Saskia ter Ellen. The paper focuses on the same topic of this thesis: the combination of density forecasts and their evaluation. We develop a forecasts combination scheme that assigns weights to the individual predictive density forecasts based on quantile scores. The paper is motivated by the limits of common combination approaches that ignore that some models may be good at forecasting the mean of the distribution but poor in the tails. In contrast, other models may provide accurate forecasts in the tail put less accurate forecasts for the mean of the distribution. The need for a coherent methodology that gives policymakers the flexibility to construct density forecasts that incorporates the heterogeneity in accuracy across regions of the forecast distribution from multiple sources cannot be understated. This paper addresses the issues above by proposing a new alternative forecast combination approach; which aims to obtain overall more accurate density forecasts by assigning a set of combination weights to the various quantities of the individual density forecasts. To achieve this goal, we first produce individual forecasts using Bayesian quantile regression models as in Kozumi and Kobayashi [2011]. They are then combined using a novel quantile combination approach. Each quantile of the combined density forecast is constructed as a weighted combination of the individual forecasts for the corresponding quantile. To account for the heterogeneity in forecast accuracy from the models across the various parts of the distribution, we allocate the quantile-specific weights from each model using the quantile score by Gneiting and Ranjan [2011]. As highlighted by Gneiting and Ranjan [2011], the quantile score is a strictly proper scoring rule, which is a weighted version (decomposition) of the continuously ranked probability score (CRPS). In an empirical application, we demonstrate the usefulness of our novel quantile combination approach to forecasting the real GDP growth rate for the Unites States for 1993Q1-2020Q2 using a real-time dataset. We combine predictive distributions from $K = 5$ quantile regression models. Each quantile regression model consists of lagged GDP growth and one additional predictor (with lags). Motivated by the recent paper by Adrian et al. [2019] and the vast literature on predicting economic recessions, we include the following predictors, the National Financial Condition Index (NFCI), the University of Michigan Consumer Sentiment Index (ICS), a credit spread that measures the difference between BAA corporate bond yield and the 10 year treasury yield, residential investments and the unemployment rate. Our novel quantile combination approach extends the findings of earlier forecast combination and GDP-at-risk literature in several ways. First, we show that density forecasts from our quantile combination approach outperforms fore-

casts from commonly used combination approaches such as Bayesian Model Averaging (BMA), the optimal combination of density forecasts (OptComb) suggested by Hall and Mitchell [2007] and Geweke and Amisano [2011], recursive logarithmic score weights as in Jore et al. [2010a] and equal weights. This holds irrespective of using the CRPS or any threshold or quantile weighted version of the CRPS that emphasise performance in either the centre, left or right tail of the distribution as a measure of forecast accuracy. The latter therefore indicates that the relative gains in terms of forecasting performance from our model are not specific to observations in a certain region of the distribution or to specific subperiods in our forecasting sample. Instead, we find a steady improvement over time and in all quantiles of the GDP distribution. Second, we show that forecasts from a quantile regression outperform forecasts from the linear regression for each model. This complements findings in Korobilis [2017] and Mazzi and Mitchell [2019] that quantile regression methods can be useful for macroeconomic forecasting. Third, while Adrian et al. [2019] argue that financial conditions are particularly informative about future downside macroeconomic risk, we show that quantile regressions that include variables such as residential investments and credit spread provide somewhat more accurate forecasts for the lower left quantile of the GDP distribution than quantile regressions that include the NFCI. This suggests that also other variables other than the NFCI are informative about future downside macroeconomic risk. Finally, our paper is also related to Opschoor et al. [2017] that assess the merits of density forecast combination schemes that assign weights to individual density forecasts based on the censored likelihood scoring rule of Diks et al. [2011] and the CRPS of Gneiting and Ranjan [2011]. While in their paper, they use this approach in the context of measuring downside risk (Value-at-Risk) in equity markets using recently developed individual volatility models, our paper differs in three important aspects. First, our combination approach differs as we assign weights to individual density forecasts based on quantile scores. Second, our goal is different as we aim to obtain density forecasts that are overall more accurate for all parts of the distribution and not only for the lower tail. Finally, we focus on forecasting GDP growth, arguably the most important macroeconomic variable, instead of measuring downside risk in equity markets.

Bibliography

Knut Are Aastveit, André K Anundsen, and Eyo I Herstad. Residential investment and recession predictability. *International Journal of Forecasting*, 35(4):1790–1799, 2019.

Tobias Adrian, Nina Boyarchenko, and Domenico Giannone. Vulnerable growth. *American Economic Review*, 109(4):1263–89, 2019.

Andrew Ang, Geert Bekaert, and Min Wei. Do macro variables, asset markets, or surveys forecast inflation better? *Journal of monetary Economics*, 54(4):1163–1212, 2007.

David Ardia, Lennart F Hoogerheide, and Herman K Van Dijk. Adaptive mixture of student-t distributions as a flexible candidate distribution for efficient simulation: The r package admit. *Journal of Statistical Software*, 29(3):1–32, 2009.

Jushan Bai and Serena Ng. Tests for skewness, kurtosis, and normality for time series data. *Journal of Business & Economic Statistics*, 23(1):49–60, 2005.

DJ Bartholomew. A test of homogeneity of means under restricted alternatives. *Journal of the Royal Statistical Society: Series B (Methodological)*, 23(2):239–272, 1961.

Travis J Berge and Òscar Jordà. Evaluating the classification of economic activity into recessions and expansions. *American Economic Journal: Macroeconomics*, 3(2):246–77, 2011.

Jeremy Berkowitz. Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, 19(4):465–474, 2001.

Gianna Boero, Jeremy Smith, and Kenneth F Wallis. Scoring rules and survey density forecasts. *International Journal of Forecasting*, 27(2):379–393, 2011.

Chris Chatfield. Model uncertainty and forecast accuracy. *Journal of Forecasting*, 15 (7):495–508, 1996.

Todd E Clark. Real-time density forecasts from bayesian vector autoregressions with stochastic volatility. *Journal of Business & Economic Statistics*, 29(3):327–341, 2011.

Todd E Clark and Michael W McCracken. Combining forecasts from nested models. *Oxford Bulletin of Economics and Statistics*, 71(3):303–329, 2009.

Oscar Claveria, Ernest Pons, and Raúl Ramos. Business and consumer expectations and macroeconomic forecasts. *International Journal of Forecasting*, 23(1):47–69, 2007.

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

Robert T Clemen and Robert L Winkler. Combining economic forecasts. *Journal of Business & Economic Statistics*, 4(1):39–46, 1986.

Michael P Clements. Evaluating the bank of england density forecasts of inflation. *The Economic Journal*, 114(498):844–866, 2004.

Michael P Clements and David F Hendry. On the limitations of comparing mean square forecast errors. *Journal of Forecasting*, 12(8):617–637, 1993.

Michael P Clements and Jeremy Smith. Evaluating the forecast densities of linear and non-linear models: applications to output growth and unemployment. *Journal of Forecasting*, 19(4):255–276, 2000.

Michael P Clements and Jeremy Smith. Evaluating multivariate forecast densities: a comparison of two approaches. *International Journal of Forecasting*, 18(3):397–407, 2002.

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

Valentina Corradi and Norman R Swanson. Bootstrap conditional distribution tests in the presence of dynamic misspecification. *Journal of Econometrics*, 133(2):779–806, 2006a.

Valentina Corradi and Norman R Swanson. Predictive density evaluation. *Handbook of economic forecasting*, 1:197–284, 2006b.

A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

A Philip Dawid. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, pages 278–292, 1984.

Francis X Diebold and Jose A Lopez. 8 forecast evaluation and combination. *Handbook of statistics*, 14:241–268, 1996.

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

Francis X Diebold, Todd A Gunther, and Anthony S Tay. Evaluating density forecasts, 1997.

Francis X Diebold, Jinyong Hahn, and Anthony S Tay. Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange. *Review of Economics and Statistics*, 81(4):661–673, 1999.

Jean Diebolt and Christian P Robert. Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 363–375, 1994.

Cees Diks, Valentyn Panchenko, and Dick van Dijk. Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163(2):215–230, 2011.

Jonas Dovern and Hans Manner. Order invariant evaluation of multivariate density forecasts. Technical report, Discussion Paper Series, 2016.

James Durbin. *Distribution theory for tests based on the sample distribution function*, volume 9. Siam, 1973.

Graham Elliott and Allan Timmermann. Optimal forecast combination under regime switching. *International Economic Review*, 46(4):1081–1102, 2005.

Arturo Estrella and Frederic S Mishkin. Predicting us recessions: Financial variables as leading indicators. *Review of Economics and Statistics*, 80(1):45–61, 1998.

Brian S Everitt. *Mixture Distributions—I.* Wiley Online Library, 1985.

Brian S Everitt. Finite mixture distributions. *Wiley StatsRef: Statistics Reference Online*, 2014.

Jon Faust and Jonathan H Wright. Forecasting inflation. In *Handbook of economic forecasting*, volume 2, pages 2–56. Elsevier, 2013.

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

Anthony Garratt, Kevin Lee, M Hashem Pesaran, and Yongcheol Shin. Forecast uncertainties in macroeconomic modeling: An application to the uk economy. *Journal of the American Statistical Association*, 98(464):829–838, 2003.

Christian Genest, Samaradasa Weerahandi, and James V Zidek. Aggregating opinions through logarithmic pooling. *Theory and decision*, 17(1):61–70, 1984.

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

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

Raffaella Giacomini and Giuseppe Ragusa. Theory-coherent forecasting. *Journal of Econometrics*, 182(1):145–155, 2014.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.

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

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

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

Aditya Gupta and Bhuwan Dhingra. Stock market prediction using hidden markov models. In *2012 Students Conference on Engineering and Systems*, pages 1–4. IEEE, 2012.

Peter Hall and DM Titterington. On confidence bands in nonparametric density estimation and regression. *Journal of Multivariate Analysis*, 27(1):228–254, 1988.

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

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

James D Hamilton. Oil and the macroeconomy since world war ii. *Journal of political economy*, 91(2):228–248, 1983.

James D Hamilton. This is what happened to the oil price-macroeconomy relationship. *Journal of Monetary Economics*, 38(2):215–220, 1996.

JM Hammersley, DC Handscomb, and George Weiss. Monte carlo methods. *Physics Today*, 18(2):55, 1965.

Gill Hammond et al. State of the art of inflation targeting. *Handbooks*, 2012.

Jesper Hansson, Per Jansson, and Mårten Löf. Business survey data: Do they help in forecasting gdp growth? *International Journal of Forecasting*, 21(2):377–389, 2005.

David F Hendry and Michael P Clements. Pooling of forecasts. *The Econometrics Journal*, 7(1):1–31, 2004.

Lennart F Hoogerheide, Johan F Kaashoek, and Herman K Van Dijk. On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: an application of flexible sampling methods using neural networks. *Journal of Econometrics*, 139(1):154–180, 2007.

Atsushi Inoue. Testing for distributional change in time series. *Econometric theory*, 17 (1):156–187, 2001.

Òscar Jordà and Massimiliano Marcellino. Path forecast evaluation. *Journal of Applied Econometrics*, 25(4):635–662, 2010.

Anne Sofie Jore, James Mitchell, and Shaun P. Vahey. Combining forecast densities from vars with uncertain instabilities. *Journal of Applied Econometrics*, 25(4):621–634, 2010a.

Anne Sofie Jore, James Mitchell, and Shaun P Vahey. Combining forecast densities from vars with uncertain instabilities. *Journal of Applied Econometrics*, 25(4):621–634, 2010b.

G Kapetanios, James Mitchell, Simon Price, and Nicholas Fawcett. Generalised density forecast combinations. *Journal of Econometrics*, 188(1):150–165, 2015.

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

Tuen Kloek and Herman K Van Dijk. Bayesian estimates of equation system parameters: an application of integration by monte carlo. *Econometrica: Journal of the Econometric Society*, pages 1–19, 1978.

Malte Knüppel. Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *Journal of Business & Economic Statistics*, 33(2):270–281, 2015.

Stanley IM Ko and Sung Y Park. Multivariate density forecast evaluation: a modified approach. *International Journal of Forecasting*, 29(3):431–441, 2013.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

Dimitris Korobilis. Quantile regression forecasts of inflation under model uncertainty. *International Journal of Forecasting*, 33(1):11–20, 2017.

Samuel Kotz, Tomasz Kozubowski, and Krzystof Podgorski. *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Springer Science & Business Media, 2012.

Hideo Kozumi and Genya Kobayashi. Gibbs sampling methods for bayesian quantile regression. *Journal of statistical computation and simulation*, 81(11):1565–1578, 2011.

Fabian Krüger, Todd E Clark, and Francesco Ravazzolo. Using entropic tilting to combine bvar forecasts with external nowcasts. *Journal of Business & Economic Statistics*, 35(3):470–485, 2017.

Oliver Linton, Esfandiar Maasoumi, and Yoon-Jae Whang. Consistent testing for stochastic dominance under general sampling schemes. *The Review of Economic Studies*, 72(3):735–765, 2005.

Oliver Linton, Kyungchul Song, and Yoon-Jae Whang. An improved bootstrap test of stochastic dominance. *Journal of Econometrics*, 154(2):186–202, 2010.

Weiling Liu and Emanuel Moench. What predicts us recessions? *International Journal of Forecasting*, 32(4):1138–1150, 2016.

Massimiliano Marcellino. Leading indicators. In Graham Elliott, Clive W. J. Granger, and Allan Timmermann, editors, *Handbook of Economic Forecasting*, volume 1, pages 879–960. Elsevier, Amsterdam, 2006.

Y Marhuenda, D Morales, and MC Pardo. A comparison of uniformity tests. *Statistics*, 39(4):315–327, 2005.

Andrew Martinez. Testing for differences in path forecast accuracy: Forecast-error dynamics matter. 2017.

G. L. Mazzi and J. Mitchell. Nowcasting Euro Area GDP Growth Using Quantile Regression. *mimeo*, 2019.

James Mitchell and Stephen G Hall. Evaluating, comparing and combining density forecasts using the klic with an application to the bank of england and niesr 'fan'charts of inflation. *Oxford bulletin of economics and statistics*, 67:995–1033, 2005.

James Mitchell and Kenneth F Wallis. Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness. *Journal of Applied Econometrics*, 26(6):1023–1040, 2011.

Anne Opschoor, Dick Van Dijk, and Michel van der Wel. Combining density forecasts using focused scoring rules. *Journal of Applied Econometrics*, 32(7):1298–1313, 2017.

Andrew J Patton and Allan Timmermann. Why do forecasters disagree? lessons from the term structure of cross-sectional dispersion. *Journal of Monetary Economics*, 57 (7):803–820, 2010.

Karl Pearson. On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, pages 379–410, 1933.

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

Rogier Quaedvlieg. Multi-horizon forecast comparison. *Journal of Business & Economic Statistics*, 39(1):40–53, 2021.

Adrian E Raftery, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174, 2005.

Francesco Ravazzolo and Philip Rothman. Oil and us gdp: A real-time out-of-sample examination. *Journal of Money, Credit and Banking*, 45(2-3):449–463, 2013.

Francesco Ravazzolo and Philip Rothman. Oil-price density forecasts of us gdp. *Studies in Nonlinear Dynamics & Econometrics*, 20(4):441–453, 2016.

Francesco Ravazzolo and Shaun P Vahey. Forecast densities for economic aggregates from disaggregate ensembles. *Studies in Nonlinear Dynamics & Econometrics*, 18(4): 367–381, 2014.

Sylvia Richardson and Peter J Green. 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, 1997.

John C Robertson, Ellis W Tallman, and Charles H Whiteman. Forecasting using relative entropy. *Journal of Money, Credit and Banking*, pages 383–401, 2005.

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

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

Barbara Rossi and Tatevik Sekhposyan. Alternative tests for correct specification of conditional predictive densities. *Journal of Econometrics*, 208(2):638–657, 2019.

Rainer Schüssler and Mark Trede. Constructing minimum-width confidence bands. *Economics Letters*, 145:182–185, 2016.

Nickolay Smirnov. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281, 1948.

JQ Smith. Diagnostic checks of non-standard time series models. *Journal of Forecasting*, 4(3):283–291, 1985.

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

James H Stock and Mark W Watson. New indexes of coincident and leading economic indicators. *NBER macroeconomics annual*, 4:351–394, 1989.

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

James H Stock and Mark W Watson. Why has us inflation become harder to forecast? *Journal of Money, Credit and banking*, 39:3–33, 2007.

Anthony S Tay and Kenneth F Wallis. Density forecasting: a survey. *Journal of forecasting*, 19(4):235–254, 2000.

Paolo Vidoni. Improved multivariate prediction regions for markov process models. *Statistical Methods & Applications*, 26(1):1–18, 2017.

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

Kenneth F Wallis. An assessment of bank of england and national institute inflation forecast uncertainties. *National Institute Economic Review*, 189(1):64–71, 2004.

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

Frank A Wolak. An exact test for multiple inequality and equality constraints in the linear regression model. *Journal of the American Statistical Association*, 82(399): 782–793, 1987.

Michael Wolf and Dan Wunderli. Bootstrap joint prediction regions. *Journal of Time Series Analysis*, 36(3):352–376, 2015.

Maik H Wolters. Evaluating point and density forecasts of dsge models. *Journal of Applied Econometrics*, 30(1):74–96, 2015.

Chun Shan Wong and Wai Keung Li. On a mixture autoregressive model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):95–115, 2000.

Jonathan H Wright. Forecasting us inflation by bayesian model averaging. *Journal of Forecasting*, 28(2):131–144, 2009.