# Optimal density forecast combinations

Gergely Gánics*

*Banco de España*

September 11, 2017

**Abstract**

How should researchers combine predictive densities to improve their forecasts? I propose consistent estimators of weights which deliver density forecast combinations approximating the true predictive density, conditional on the researcher's information set. Monte Carlo simulations confirm that the proposed methods work well for sample sizes of practical interest. In an empirical example of forecasting monthly US industrial production, I demonstrate that the estimator delivers density forecasts which are superior to well-known benchmarks, such as the equal weights scheme. Specifically, I show that housing permits had valuable predictive power before and after the Great Recession. Furthermore, stock returns and corporate bond spreads proved to be useful predictors during the recent crisis, suggesting that financial variables help with density forecasting in a highly leveraged economy.

***Keywords:*** Density forecasts, Forecast combinations, Probability Integral Transform, Kolmogorov–Smirnov, Cramer–von Mises, Anderson–Darling, Kullback–Leibler Information Criterion

***JEL codes:*** C13, C22, C53

*Researcher at the Banco de España. Email: gergely.ganics@bde.es. I am greatly indebted to my advisor Barbara Rossi for all the invaluable support, advice and encouragement I received. I would also like to thank Majid Al-Sadoon for his great insights that led to major improvements of the present paper, and Christian Brownlees, Eleonora Granziera, Frank Kleibergen, Malte Knüppel, Juri Marcucci, Geert Mesters, James Morley, Denis Nekipelov, Elena Pesavento, Tatevik Sekhposyan, participants of the VI$_t$ Workshop in Time Series Econometrics (2016, Zaragoza, Spain), the 2016 Barcelona GSE Summer Forum's Workshop on Time Series Econometrics and Applications in Macroeconomics and Finance (Barcelona, Spain), the 4th SIdE-IEA Workshop for PhD students in Econometrics and Empirical Economics (2016, Perugia, Italy), the CREi Macroeconomics Breakfast Seminar, the Banco de España, the Bank of Canada, the Banque de France and the Università di Bologna for numerous useful comments that considerably improved this study. All remaining errors are mine. The author gratefully acknowledges financial support from the Spanish Ministerio de Economía y Competitividad under FPI grant BES-2013-065352, the 3rd Economics Job Market Best Paper Award (Unicredit & Universities Foundation), and the 2nd Marcelo Reyes Prize in Time Series Econometrics (Universidad de Zaragoza). The views and opinions expressed herein are those of the author and do not necessarily reflect the views and opinions of the Banco de España or the Eurosystem.

# 1 Introduction

Density or distribution forecasts have become increasingly popular both in the academic literature and among professional forecasters. This success is due to their ability to provide a summary of uncertainty surrounding point forecasts, which facilitates communication between researchers, decision makers and the wider public. As Alan Greenspan stated, "a central bank needs to consider not only the most likely future path for the economy, but also the distribution of possible outcomes about that path" (Greenspan, 2004, p. 37). Well-known examples of forecasts produced in this spirit include the fan charts of the Bank of England and the Surveys of Professional Forecasters (SPF) of the Federal Reserve Bank of Philadelphia and the European Central Bank.[1]

Just as combinations of individual point forecasts have been found to be superior against a single point forecast in many settings, density combinations have been shown to outperform the density forecast of individual models (Elliott and Timmermann, 2016; Timmermann, 2006). The reasons for both are largely the same: model misspecification, structural breaks and parameter estimation uncertainty complicate the task of producing reliable forecasts. Practitioners often combine point forecasts based on simple rules or expert judgment. Convex combinations of densities can take shapes that are dissimilar to their individual components, resulting in considerably different predictions. This makes density forecast combination a more challenging task than the combination of point forecasts. While assigning equal weights to predictive densities often results in improvements (Rossi and Sekhposyan, 2014), this scheme does not offer insights into the individual models' performance, hence researchers cannot exploit information on models' predictive ability. However, the data-driven weighting scheme proposed in this study can help researchers understand and improve their forecasting methods.

In the present paper, I focus on estimators of density combination weights based on the Probability Integral Transform or PIT (Rosenblatt, 1952; Diebold et al., 1998), which is defined as the researcher's predictive cumulative distribution function (CDF) evaluated at the actual realization. The underlying idea of the PIT is remarkably simple yet powerful: the PIT is uniformly distributed if and only if the predictive density used by the researcher coincides with the true predictive density conditional on the researcher's information set, which is the notion of optimality in this paper. Discrepancies between the true, unknown predictive distribution and the researcher's density forecast show up in the distribution of the PIT, which can be used to design tests. The present paper builds on this idea, but instead of using it for testing purposes, I invert the problem and estimate the combination weights by minimizing the distance between the uniform distribution and the empirical distribution of the convex combination of PITs using either

[1] Elder et al. (2005) provide an assessment of the Bank of England's fan charts. For a recent overview of the ECB's SPF, see European Central Bank (2014). A list of papers using the Philadelphia Fed's SPF can be found at https://www.phil.frb.org/research-and-data/real-time-center/survey-of-professional-forecasters/academic-bibliography.

the Kolmogorov–Smirnov, the Cramer–von Mises or the Anderson–Darling statistic. I show that this method leads to consistent weight estimators that generate either an optimal forecast density combination or one closest to it.

This paper's contributions are summarized as follows. First, building on the PIT, I develop consistent weight estimators delivering density forecasts which either correspond to the true predictive density conditional on the researcher's information set, or are closest to it when measured in the Kolmogorov–Smirnov, Cramer–von Mises or Anderson–Darling sense. This result holds even if the true predictive density is not included in the pool of models used by the researcher. "Model" is understood in a wide sense, including survey and judgmental forecasts, and no knowledge of the underlying model generating the density forecast is required. Second, I provide a formal theory to estimate density forecast combination weights using the Kullback–Leibler Information Criterion (KLIC) and I compare the PIT-based and KLIC-based estimators in Monte Carlo simulations covering a wide range of DGPs and sample sizes, providing valuable assistance to researchers. The simulation results suggest that the PIT-based estimator using the Anderson–Darling distance and the KLIC-based estimator yield precise weight estimates even for moderate sample sizes. Third, I demonstrate that the novel PIT-based forecast combination method delivers one-month-ahead forecasts of US industrial production growth which are superior to the widely used equal weights benchmark. The weight estimates show that housing permits were a useful predictor in the years preceding and following the Great Recession. Furthermore, financial variables, especially corporate bond spreads received considerable weight during and after the recent financial crisis.

The literature on combining point forecasts according to an optimality criterion, such as minimizing the expected mean squared forecast error, started with the celebrated paper by Bates and Granger (1969) and includes numerous contributions, both empirical, such as Stock and Watson (2004), and theoretical, for example Cheng and Hansen (2015) and Claeskens et al. (2016).[2] While density forecast *evaluation* has been widely studied (Diebold et al., 1998; Corradi and Swanson, 2006a,c; Rossi and Sekhposyan, 2014, 2016), the *estimation* of density combination weights with respect to an optimality criterion has received less attention.

My theoretical contribution is related to several strands of the literature on density forecast combinations. Using logarithmic predictive scores, Hall and Mitchell (2007) propose optimal weights with respect to the KLIC. In contrast, I focus on estimators based on the PIT, although for completeness I also discuss their KLIC-based estimator and provide theoretical results for it, complementing the empirical analysis in Hall and Mitchell (2007). In a related paper, Geweke and Amisano (2011) provide theoretical results on linear prediction pools based on the KLIC. In the present study I show strong consistency of the PIT-based estimators and also provide an alternative proof of the

[2]For a comprehensive overview on the combination of point forecasts, see Elliott and Timmermann (2016) and Timmermann (2006).

consistency of the KLIC-based estimator. Pauwels and Vasnev (2016) deal with the practical implementation of estimating combination weights and provide a comparison of alternative weighting schemes through a number of Monte Carlo simulations, with a specific focus on small samples. In contrast, my simulations cover a wide range of Data Generating Processes (DGPs) and investigate both the PIT- and the KLIC-based estimators' properties in small and large samples, thereby I can offer advice to practitioners. The estimators proposed in the present paper are justified on frequentist grounds. For a recent treatment of Bayesian estimation of predictive density combination weights, see Billio et al. (2013) and Del Negro et al. (2016). While those papers use computationally intensive non-linear filtering methods, the estimators proposed in this study can be implemented using a standard optimization algorithm and do not rely on priors. Furthermore, my approach does not require knowledge of the model that generated the density forecast, therefore it can be applied to survey or judgmental forecasts as well.

From an empirical perspective, since the onset of the Great Recession, several papers have focused on exploiting non-Gaussian features of macroeconomic data, along with time-varying volatility. Cúrdia et al. (2014), using a Dynamic Stochastic General Equilibrium (DSGE) model, show that incorporating stochastic volatility and using a fat-tailed shock distribution substantially improves the model's fit. In contrast, my empirical application uses an ensemble of simple, non-structural univariate Autoregressive Distributed Lag (ARDL) models, and combines their predictive densities to achieve calibrated one-month-ahead density forecasts of US industrial production. In a recent paper, Rossi and Sekhposyan (2014) demonstrated that convex combinations of ARDL models' predictive densities deliver well-calibrated density forecasts. In terms of point forecasts, Gürkaynak et al. (2013) showed that univariate autoregressive models often outperform multivariate DSGE and Vector Autoregressive (VAR) models. Clark and Ravazzolo (2015) provide an extensive comparison of both point and density forecasts generated by univariate and multivariate Bayesian (Vector) Autoregressive (BVAR) models with a number of volatility specifications, using quarterly real-time US data. They conclude that stochastic volatility materially improves density forecasts of output growth, especially in the short-run. In the present study, I let a rolling window estimation scheme account for possible time-variation in volatility.

In their recent study, Chiu et al. (2015), using BVAR models demonstrate that in an out-of-sample forecasting exercise, it is mainly fat tailed shocks and not stochastic volatility that considerably improves density forecasts of industrial production. In a related paper, Chiu et al. (2016) investigate the mixture of normal distributions as predictive density, using a regime switching model, where the parameters of the normal distributions depend on the current, hidden state of the economy. The authors show that such a flexible specification delivers sizable gains in terms of density forecasts of industrial production relative to a Gaussian BVAR. Waggoner and Zha (2012) demonstrate how a DSGE and a BVAR model can be integrated into a common framework, using a

Markov-switching structure that drives the weights associated with the models. However, their paper focuses on improving the models' in-sample fit rather than their forecasting performance. Related to the previous papers, I also allow for non-Gaussian predictive distributions, but instead of specifying a regime switching model, I estimate the weights generating non-normal predictive distributions either through the KLIC or the PIT. This procedure allows me to focus on fine-tuning the forecasts without having to posit an underlying model for the regimes. Moreover, by taking the predictive densities as given, I can avoid the pitfalls associated with the joint estimation of the predictive densities and the mixture weights.[3] As I will demonstrate, the estimated weights are informative of the state of the US economy. Specifically, I show that data on housing permits was the best predictor of US industrial production growth in the years leading to the Great Recession. Furthermore, financial variables (corporate bond spreads and stock returns) proved to be useful predictors during the recent financial crisis. While Ng and Wright (2013) presented similar results about financial variables for point forecasts, to my best knowledge, this is the first paper that demonstrates these findings for density forecasts.

The remainder of the paper is organized as follows. Section 2 introduces the notation and the definitions used throughout the paper. Section 3 describes the forecasting environment and the proposed density forecast combination method, while Section 4 provides the results of Monte Carlo exercises. An empirical application of forecasting US industrial production is presented in Section 5, then Section 6 concludes. The proofs are collected in Appendix A, while additional technical details and results can be found in Appendices B to F.

## 2 Notation and definitions

In this section, I introduce the notation and definitions used in the present paper and discuss the assumptions of the estimation procedure.

Consider the stochastic process $\left\{Z_t : \Omega \to \mathbb{R}^{k+1}\right\}_{t=1}^{T+h}$ defined on a complete probability space $(\Omega, \mathcal{F}, P)$. The observed vector $Z_t$ is partitioned as $Z_t = (y_t, X_t')'$, where $y_t : \Omega \to \mathbb{R}$ is the variable of interest and $X_t : \Omega \to \mathbb{R}^k$ is a vector of predictors. Let $\mathcal{F}_t$ denote the filtration associated with the stochastic process $\{Z_t\}$ and let $\mathcal{I}_t \subset \mathcal{F}_t$ denote the information at time $t$ that is relevant to the determination of the outcome $y_{t+h}$. Furthermore, let $\phi^*_{t+h}(y|\mathcal{I}_t)$ be the corresponding true conditional density.[4] In what follows, the abbreviation *iid.* stands for independent and identically distributed, and $\mathcal{N}(\mu, \mathbb{V})$ is the normal distribution with mean vector $\mu$ and covariance matrix $\mathbb{V}$.

[3] For an overview of this problem, see Chapter 1 of Rossi (2014).

[4] Throughout the present paper, $\phi(\cdot|\cdot)$ and $\Phi(\cdot|\cdot)$ stand for any conditional probability density function and cumulative distribution function, respectively, not necessarily those of the normal distribution. I also assume that all random variables possess probability density functions. With a slight abuse of notation, I do not make a distinction between the random variable and its realization, as it should be clear from the context which is meant.

Convergence in probability and almost sure convergence are denoted by $\xrightarrow{p}$ and $\xrightarrow{a.s.}$, respectively.

The available sample of size $T+h$ is utilized as follows. At forecast origin $f$, the researcher has $\mathcal{M}$ models at hand, which are indexed by $m = 1, \ldots, \mathcal{M}$.[5] These models are estimated in rolling windows of size $R$, where each estimation is based on the truncated information set $\mathcal{I}_{t-R+1}^{t}$, containing information between $t-R+1$ and $t$. The time index $t$ runs from $t = f - G - h + 1$ to $t = f - h$, where $G$ is the total number of rolling windows, as it will be explained later. At each $t$, each of the models imply an $h$-step-ahead density forecast of $y_{t+h}$, with typical element $\phi_{t+h}^{m}(y|\mathcal{I}_{t-R+1}^{t})$. The forecaster uses the convex combination of the $\mathcal{M}$ predictive densities (highlighted by the $C$ superscript), denoted by

$$\phi_{t+h}^{C}(y|\mathcal{I}_{t-R+1}^{t}) \equiv \sum_{m=1}^{\mathcal{M}} w_m \phi_{t+h}^{m}(y|\mathcal{I}_{t-R+1}^{t}), \tag{1}$$

where the $m$ superscript indexes the densities. The corresponding cumulative predictive distributions are then given by

$$\Phi_{t+h}^{C}(\overline{y}|\mathcal{I}_{t-R+1}^{t}) = \int_{-\infty}^{\overline{y}} \sum_{m=1}^{\mathcal{M}} w_m \phi_{t+h}^{m}(y|\mathcal{I}_{t-R+1}^{t})\, \mathrm{d}y = \sum_{m=1}^{\mathcal{M}} w_m \Phi_{t+h}^{m}(\overline{y}|\mathcal{I}_{t-R+1}^{t}). \tag{2}$$

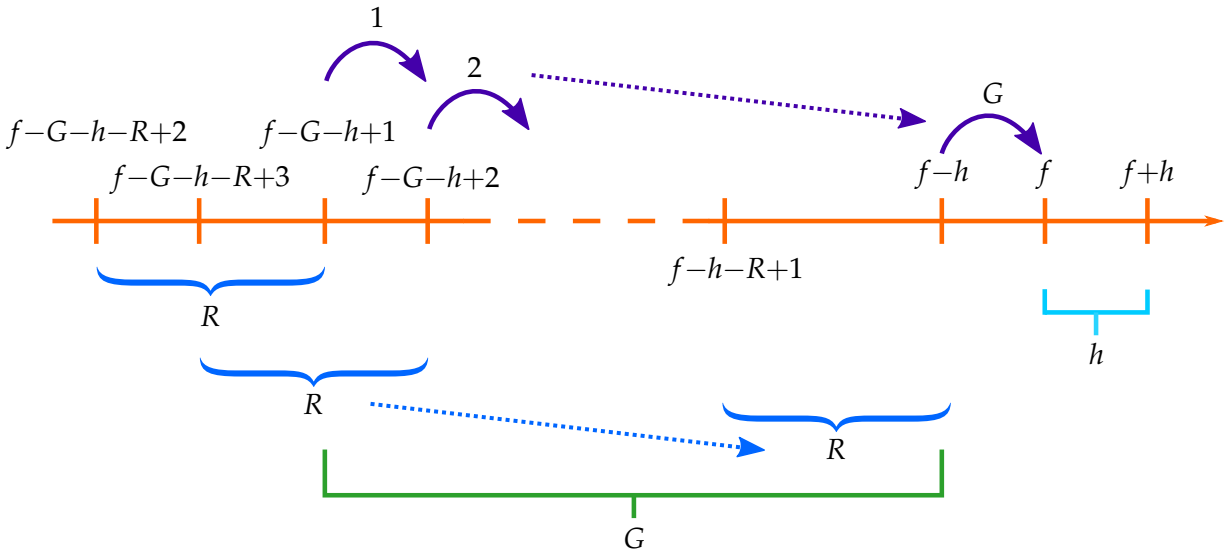
By requiring that the weights $w_m$ satisfy $w_m \geq 0$ for all $m = 1, \ldots, \mathcal{M}$ and $\sum_{m=1}^{\mathcal{M}} w_m = 1$, it is guaranteed that the combination of the individual densities (respectively, CDFs) is a density (respectively, CDF) itself. The weights are collected in a vector $w \equiv (w_1, \ldots, w_{\mathcal{M}})'$. Equivalently, $w \in \Delta^{\mathcal{M}-1}$, where $\Delta^{\mathcal{M}-1}$ is the $\mathcal{M}-1$ unit simplex.

The estimation procedure is repeated in a similar way for all forecast origins $f = G + h + R - 1, \ldots, T$. This scheme yields a total number of $P = T - G - h - R$ out-of-sample density forecasts with the corresponding realizations, which could be used to assess the performance of the forecast combinations. Figure 1 provides a graphical illustration of the proposed estimation scheme. By using a rolling window scheme, researchers can potentially alleviate problems related to structural instabilities. Furthermore, for reasons explained later, it is necessary to keep the density estimation window size $R$ finite (ie. "small") and the combination window size $G$ "large".

The true distribution of $y_{t+h}$ *conditional* on $\mathcal{I}_{t-R+1}^{t}$ is denoted by $\Phi_{t+h}^{*}(\overline{y}|\mathcal{I}_{t-R+1}^{t})$. If for a given $w$, $\sum_{m=1}^{\mathcal{M}} w_m \Phi_{t+h}^{m}(\overline{y}|\mathcal{I}_{t-R+1}^{t})$ coincides with $\Phi_{t+h}^{*}(\overline{y}|\mathcal{I}_{t-R+1}^{t})$, then the forecast is said to satisfy *probabilistic calibration*. If, in addition, for a given $w$ the conditional distribution used by the researcher is the same as the true predictive distribution of $y_{t+h}$ given $\mathcal{I}_t$, that is $\sum_{m=1}^{\mathcal{M}} w_m \Phi_{t+h}^{m}(\overline{y}|\mathcal{I}_{t-R+1}^{t}) = \Phi_{t+h}^{*}(\overline{y}|\mathcal{I}_t)$, then the forecast is said to satisfy

[5] The model set $\mathcal{M}$ is allowed to vary across forecast origins ($\mathcal{M}_f$ in notation), thereby allowing researchers to tailor the pool of forecasting models according to their past performance. However, evaluating the gains from this extension is left for future research.

Figure 1: Proposed estimation scheme




*Note*: $f$ and $f+h$ denote the forecast origin and the target date, respectively. The researcher estimates each model in rolling windows of size $R$, which are indicated by curly (blue) braces and collects the $h$-period-ahead predictive distributions and the corresponding realizations, indicated by curved (purple) arrows, forming a sequence of size $G$, which is used to estimate combination weights.

*complete calibration*.[6] It is important to note that neither notion of calibration requires that the true predictive density $\phi^*_{t+h}(y|\mathcal{I}_t)$ belong to the set of $\mathcal{M}$ densities. In practice, researchers often do not know the true predictive density of $y_{t+h}$, and the most they can aspire to is producing the best forecast conditional on the specific information set – that is, producing a probabilistically calibrated forecast.

The following stylized example, inspired by Corradi and Swanson (2006b,c), illustrates the difference between probabilistic and complete calibration and features dynamic misspecification. For simplicity, I abstract from parameter estimation error.

**Example 1.** Let us assume that the true DGP for $y_{t+1}$ is a stationary normal AR(2) process, given by $y_{t+1} = \alpha_1 y_t + \alpha_2 y_{t-1} + \varepsilon_{t+1}$ where $\varepsilon_{t+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0, \sigma^2)$; that is, the density of $y_{t+1}$ conditional on $\mathcal{I}_t = \{y_t, y_{t-1}\}$ is $\phi^*_{t+1}(y_{t+1}|\mathcal{I}_t) = \mathcal{N}(\alpha_1 y_t + \alpha_2 y_{t-1}, \sigma^2)$. Therefore the joint distribution of $(y_{t+1}, y_t, y_{t-1})'$ is a multivariate normal with covariance matrix $\Sigma$. Furthermore, by properties of the normal distribution, the distribution of $y_{t+1}$ conditional on $y_t$ alone is also normal, formally $\phi^*_{t+1}(y_{t+1}|y_t) = \mathcal{N}(\widetilde{\alpha} y_t, \widetilde{\sigma}^2)$, where $\widetilde{\alpha}$ and $\widetilde{\sigma}^2$ can be computed from $\Sigma$.

Suppose that the researcher conditions his or her one-step-ahead forecast on only one lag of the dependent variable, $(R = 1, \mathcal{I}^t_{t-R+1} = y_t)$ but maintains the normality assumption, which amounts to using the predictive density $\phi_{t+1}(y_{t+1}|\mathcal{I}^t_{t-R+1}) = \mathcal{N}(\widetilde{\alpha} y_t, \widetilde{\sigma}^2)$, corresponding to a dynamically misspecified AR(1) model. In this case, it is easy to see that while the forecast is not completely calibrated due to the omission of $y_{t-1}$, it is still probabilistically calibrated, as given the researcher's information set (now consisting of

[6] For an overview of different modes of calibration, see Gneiting et al. (2007) or Mitchell and Wallis (2011).

$y_t$), the predictive density is correct, $\phi_{t+1}(y_{t+1}|\mathcal{I}^t_{t-R+1}) = \phi^*_{t+1}(y_{t+1}|\mathcal{I}^t_{t-R+1})$. For more details on this example, see Appendix B. ▲

It is important to emphasize that the researcher does not need to know the true DGP in order to produce probabilistically calibrated forecasts, as Example 1 illustrates. Therefore this is a weak notion of calibration, making it attractive for practitioners.

## 2.1 The Probability Integral Transform

The Probability Integral Transform (PIT) is defined as

$$z_{t+h} \equiv \int_{-\infty}^{y_{t+h}} \phi^C_{t+h}(y|\mathcal{I}^t_{t-R+1})\,\mathrm{d}y = \Phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})\,, \tag{3}$$

where $\Phi^C_{t+h}(\cdot|\cdot)$ denotes the conditional CDF corresponding to the conditional predictive density $\phi^C_{t+h}(\cdot|\cdot)$. It is easy to see that if and only if the forecast is probabilistically calibrated, then $z_{t+h} \sim \mathcal{U}(0,1)$, that is $z_{t+h}$ has the standard uniform distribution. For a proof of this well-known result, see Corradi and Swanson (2006a, pp. 784–785).[7]

The following example shows how the lack of probabilistic calibration can be detected through the investigation of the PITs. It also demonstrates how the PDFs (probability density functions) and the CDFs of the PITs can provide useful information on which region of the true predictive distribution the researcher's forecast is unable to match.

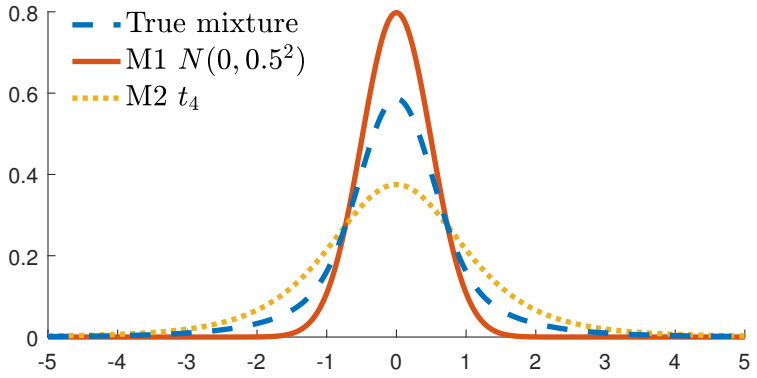
**Example 2.** Let us assume that the true forecast density of $y_{t+1}$ is a mixture of a normal density with mean zero and variance $0.5^2$ and a Student's t-density with 4 degrees of freedom (denoted by $t_4$) with mixture weights $(w_1, w_2)' = (0.5, 0.5)'$. That is, we have $\phi^*_{t+1}(y_{t+1}|\mathcal{I}_t) = 0.5\mathcal{N}(0, 0.5^2) + 0.5t_4$. The forecaster uses three predictive densities. Assume that the first incorrect predictive density is the normal component of the mixture density, $\phi^1_{t+1}(y_{t+1}|\mathcal{I}^t_{t-R+1}) = \mathcal{N}(0, 0.5^2)$ and the second one is the Student's t component, $\phi^2_{t+1}(y_{t+1}|\mathcal{I}^t_{t-R+1}) = t_4$. Furthermore, the third density is the correct mixture density.

Figure 2 displays the three PDFs. We can see that while the means of the incorrectly calibrated densities are the same as the true forecast density's mean, their tails are markedly different, with the normal density featuring thinner and Student's t-density displaying thicker tails than the true mixture density.

I calculated the PITs using each of the three models above. The PDFs of each of the PITs in Figure 3 immediately reveal that using the true density delivers uniformly distributed PITs, while the $t$ (normal) density would imply many more (much less) extreme observations in both tails, therefore the densities of the PITs show a typical hump (regular U) shape. In Figure 4, we can see that the CDF of the PITs obtained

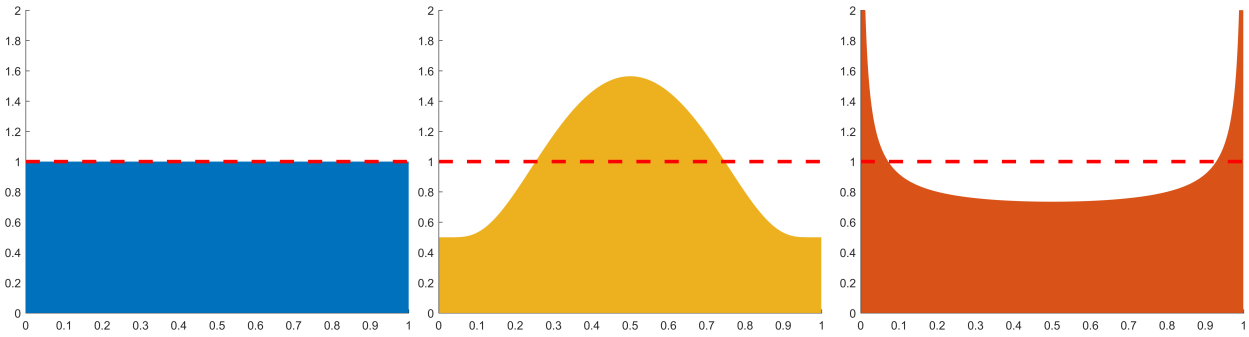
[7] The original result is usually attributed to Rosenblatt (1952), while in the econometrics literature it was introduced by Diebold et al. (1998). The discussion in Corradi and Swanson (2006a) and Gneiting et al. (2007) is the closest to the framework of the present study.

Figure 2: Probability density functions of candidate forecast densities




by using the true mixture density coincides with the 45 degree line corresponding to the CDF of the uniform distribution. On the other hand, the incorrect densities deliver PITs whose CDFs display S-shaped and inverted S-shaped patterns, which are typical in situations when the tail behaviors of the assumed and the true distributions differ. ▲

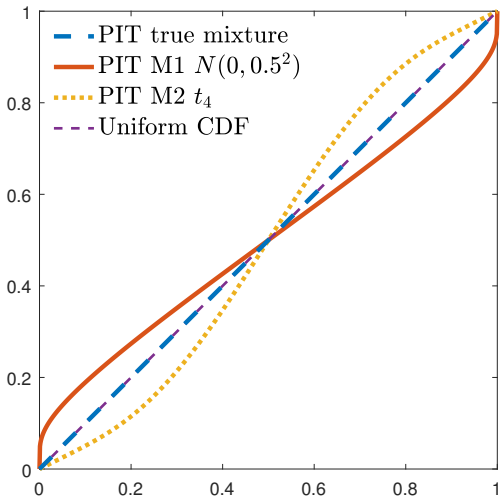
Figure 3: Probability density functions of PITs



(a) True mixture density (b) Incorrect $t_4$ density (c) Incorrect $\mathcal{N}(0, 0.5^2)$ density

*Note*: Horizontal dashed (red) line corresponds to uniform density.

Figure 4: Cumulative distribution functions of PITs of candidate densities



If the forecast is completely calibrated, then as Diebold et al. (1998) showed, the PITs are at most $h-1$ dependent. In practice, it is rather unreasonable to assume that the researcher has completely calibrated forecasts at hand (e.g. because of omitted variables, such as in Example 1) and instead I investigate how to ensure that the combined forecast is going to be as close as possible to being probabilistically calibrated *given* the information available at the forecast origin. That is, this paper takes the estimated predictive densities as given. This leads to the question of estimating the weight vector $w$.

Let us define

$$\xi_{t+h}(r,w) \equiv 1\left[\Phi^{C}_{t+h}(y_{t+h}|\mathcal{I}^{t}_{t-R+1}) \leq r\right] - r = 1\left[z_{t+h} \leq r\right] - r \tag{4}$$

at a given quantile denoted by $r \in [0,1]$ where $1[\cdot]$ stands for the indicator function. Consider $\Psi(r,w) \equiv P(z_{t+h} \leq r) - r$ and its sample counterpart:

$$\Psi_G(r,w) \equiv G^{-1} \sum_{t=f-G-h+1}^{f-h} \xi_{t+h}(r,w)\,, \tag{5}$$

which measures the vertical distance between the empirical CDF of the PIT and the CDF of the uniform distribution (the 45 degree line) at quantile $r$, where $G$ is the number of observations used to evaluate the PITs up to and including the forecast origin $f$. Recall that over the full sample, the forecast origin $f$ ranges from $G+R+h-1$ to $T$.

Three widely known test statistics that measure the discrepancy between CDFs are the Kolmogorov–Smirnov, the Cramer–von Mises and the Anderson–Darling statistics (Anderson and Darling, 1952), which have been used in recent studies to test the uniformity of PITs (see, for example Corradi and Swanson (2006c); Rossi and Sekhposyan (2013, 2014, 2016)). Let $\rho \subset [0,1]$ denote a finite union of neither empty nor singleton, closed intervals on the unit interval, which depends on the researcher's interests. The choice of $\rho$ is discussed below.

I use the Kolmogorov–Smirnov, the Cramer–von Mises and the Anderson–Darling statistics as objective functions[8] in the following forms:

$$K_G(w) \equiv \sup_{r\in\rho} \left|\Psi_G(r,w)\right|, \tag{6}$$

$$C_G(w) \equiv \int_{\rho} \Psi^2_G(r,w)\,\mathrm{d}r\,, \tag{7}$$

$$A_G(w) \equiv \int_{\rho} \frac{\Psi^2_G(r,w)}{r(1-r)}\,\mathrm{d}r\,. \tag{8}$$

The Kolmogorov–Smirnov statistic measures the largest absolute deviation of the

[8]Sometimes I refer to the Kolmogorov–Smirnov-, the Cramer–von Mises- and the Anderson–Darling-type objective functions using the abbreviations KS, CvM and AD, respectively.

empirical CDF from the 45 degree line. On the other hand, the Cramer–von Mises statistic takes into account all the deviations from the 45 degree line by measuring the total deviation. Furthermore, the Anderson–Darling statistic weighs the deviations by the inverse of the variance of the CDF, making it more sensitive to deviations in the tails than in the central region. These features of the CvM and the AD objective functions potentially lead to more precise estimators, as the Monte Carlo simulations will demonstrate.

In some situations, practitioners may be interested in obtaining probabilistically calibrated forecasts focusing only on specific parts of the predictive distribution. For example, finance researchers often forecast one-day-ahead Value at Risk (VaR) at the 5% level, that is, they want to obtain the threshold loss value $\bar{l}_{t+1}$ such that the ex-ante probability that their loss $l_{t+1}$ will exceed the threshold is 5%. As they are interested in forecasting the 5% quantile of the distribution of $l_{t+1}$, they might want to focus on the left tail of the predictive distribution, corresponding to $\rho = [0, 0.05]$. On the other hand, if a researcher is interested in the full predictive distribution, then $\rho = [0, 1]$, while if he or she wants to focus attention on the lower and upper 5 percentiles, then $\rho = [0, 0.05] \cup [0.95, 1]$ is appropriate.

## 2.2 The Kullback–Leibler Information Criterion

While the Kolmogorov–Smirnov, the Cramer–von Mises and the Anderson–Darling distances (collectively, PIT-based measures) provide one way to measure discrepancies between distributions, they are not the only ones. Another example is the Kullback–Leibler Information Criterion (KLIC), which was proposed as an objective function for density forecast combinations by Hall and Mitchell (2007).[9]

Similarly to the PIT-based objective functions, let $\varrho$ denote a finite union of closed, non-empty, non-singleton intervals on the support of the true conditional distribution $\Phi^*_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})$. As before, the researcher can set $\varrho$, for example focusing on discrepancies in the $[-3\%, 0\%]$ range when forecasting recessions. If the whole distribution is of interest, then $\varrho$ can be set as the whole real line. The KLIC between the distributions $\Phi^*_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})$ and $\Phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})$ with corresponding densities $\phi^*_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})$ and $\phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})$, over the region of interest $\varrho$ is defined as

$$\text{KLIC}_\varrho(\Phi^*_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}), \Phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})) \tag{9}$$

$$\equiv \int_{-\infty}^{\infty} \phi^*_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}) \log \frac{\phi^*_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})}{\phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})} 1[y_{t+h} \in \varrho] \, dy_{t+h} \tag{10}$$

$$= E_{\phi^*}\left\{ \left( \log \phi^*_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}) - \log \phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}) \right) 1[y_{t+h} \in \varrho] \right\} \tag{11}$$

[9] The KLIC has been used extensively in the econometrics literature, see for example the seminal paper by White (1982) on Quasi Maximum Likelihood Estimators (QMLE).

$$= E_{\phi^*}\left\{\log \phi^*_{t+h}(y_{t+h}|\mathfrak{I}^t_{t-R+1})1[y_{t+h} \in \varrho]\right\} - E_{\phi^*}\left\{\log \phi^C_{t+h}(y_{t+h}|\mathfrak{I}^t_{t-R+1})1[y_{t+h} \in \varrho]\right\}, \tag{12}$$

where the subscripts in Equations (11) and (12) remind us that the expectations are taken with respect to the *true* predictive density. It is well known that KLIC $\geq 0$, and KLIC $= 0$ if and only if $\Phi^*_{t+h}(y_{t+h}|\mathfrak{I}^t_{t-R+1}) = \Phi^C_{t+h}(y_{t+h}|\mathfrak{I}^t_{t-R+1})$ almost surely, and larger values of the KLIC correspond to larger discrepancy between the true and the combined densities. The KLIC can be interpreted as the surprise experienced on average when we believe that $\phi^C_{t+h}(y_{t+h}|\mathfrak{I}^t_{t-R+1})$ is the true predictive density but then we are informed that it is $\phi^*_{t+h}(y_{t+h}|\mathfrak{I}^t_{t-R+1})$ instead (White, 1994, Chapter 2, p.9). The first term in Equation (12) does not depend on the weights, hence the minimizer of the KLIC with respect to the weights is the minimizer of the second term alone and therefore the first term can be treated as a constant. Based on the above definition of the KLIC, the average KLIC (leaving out the constant term) is given by

$$\overline{\text{KLIC}}_0 \equiv G^{-1} \sum_{t=f-G-h+1}^{f-h} -E_{\phi^*}\left\{\log \phi^C_{t+h}(y_{t+h}|\mathfrak{I}^t_{t-R+1})1[y_{t+h} \in \varrho]\right\}, \tag{13}$$

where the average is taken over the $G$ time periods preceding the forecast origin $f$. Hall and Mitchell (2007) proposed the sample counterpart of the KLIC as objective function to estimate the combination weights:

$$\text{KLIC}_G(w) = G^{-1} \sum_{t=f-G-h+1}^{f-h} \left\{-\log \phi^C_{t+h}(y_{t+h}|\mathfrak{I}^t_{t-R+1})1[y_{t+h} \in \varrho]\right\}. \tag{14}$$

As we can see, the KLIC is fully operational without specifying the true predictive distribution, which is clearly a desirable property, also enjoyed by the PIT-based measures. Similarly to the PIT-based estimators, the KLIC-type estimator can also target specific regions of the predictive density.

Some remarks are in order. Imagine a forecaster who wants to answer the question: what is the range of values that will contain next month's inflation with, say 90% probability? Clearly, if the researcher matches the whole predictive distribution, then he or she is going to be able to answer this question. Restricting $\rho$ or $\varrho$ can potentially lead to more precise density forecasts, as Diks et al. (2011) demonstrated for the KLIC-type estimator. However, there is a trade-off. Focusing on a specific part of the distribution means that the sample size must be considerably larger than when using an unrestricted estimator. Alternatively, the estimator should be able to minimize the discrepancy between the true and the combined distributions much "better" in the subset of interest than over the whole distribution. The evaluation of potential gains resulting from such restrictions is outside the scope of the present paper.

# 3 Estimators and assumptions

In this section I will discuss how the aforementioned statistics defined in Equations (6) to (8) and (14) can be used as objective functions to estimate the weights and I outline the assumptions that render the estimators consistent.

As discussed in Section 2, obtaining probabilistically calibrated combined forecasts amounts to using a forecast density combination that delivers uniform PITs. We can invert this problem and say the following: let us estimate the combination weights by minimizing the distance between the empirical CDF of the PITs and the CDF of the uniform distribution. Formally, the "optimal" estimated weights are defined as

$$\widehat{w} = \underset{w \in \Delta^{\mathcal{M}-1}}{\operatorname{argmin}} \, T_G(w) \,, \tag{15}$$

where $T_G(w)$ is either $K_G(w), C_G(w)$ or $A_G(w)$.[10] Similarly, the estimated KLIC weights are defined as

$$\widehat{w} = \underset{w \in \Delta^{\mathcal{M}-1}}{\operatorname{argmin}} \, \mathrm{KLIC}_G(w) \,. \tag{16}$$

Before stating and discussing the assumptions that guarantee consistency of the estimators defined in Equations (15) and (16), it is worth understanding why consistency has a direct appeal to forecasters in this framework. Suppose that a researcher wants to combine models' point forecasts. Based on the past performance of the respective models and possibly some expert information, the researcher might be able to discard a number of models whose forecasts are considered implausible and then weigh the remaining models' point forecasts using either some data-driven procedure or expert judgment. On the other hand, when combining density forecasts, the forecaster is in a more difficult situation, as density forecasts are high-dimensional objects, and depending on the weights, the shape of the combined density could differ largely from the shape of its components, as the Monte Carlo simulations of Section 4 will demonstrate. Therefore it is of both theoretical and practical importance that the estimator proposed in this paper is consistent for the weight vector that in population either delivers probabilistically calibrated forecasts or minimizes the discrepancy between the combined density and the true predictive density (or their PITs).

## 3.1 PIT-based estimators

In what follows, I state and discuss the assumptions that render the PIT-based estimators consistent. Statements involving "for all $t$" are understood as $t$ ranges from $t = f - G -$

[10] The definition reflects that weights are re-estimated at forecast origins $f = G + R + h - 1, \ldots, T$, allowing for time-variation over different forecast origins. This also applies to the KLIC-based estimator.

$h+1$ to $f-h$, which is the sample period used to estimate the combination weights.

**Assumption 1** (Dependence). *$\{Z_t\}$ is $\phi$-mixing of size $-k/(2k-1), k \geq 1$ or $\alpha$-mixing of size $-k/(k-1), k > 1$.*

**Assumption 2** (Region of interest). *$\rho \subset [0,1]$ is a finite union of neither empty nor singleton, closed intervals on the unit interval, which depends on the researcher's interests.*

**Assumption 3** (Continuity). *The combined CDF is continuously distributed, formally $P\left[\Phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}) = r\right] = 0$ for all $(w,r) \in \Delta^{\mathcal{M}-1} \times \rho$ and for all $t$.*

**Assumption 4** (Estimation scheme). *$R < \infty$ as $G, T \to \infty$, $1 \leq h < \infty$ and fixed. The number of models $\mathcal{M}$ is finite.*

**Assumption 5** (Identification). *There exists a unique $w^* \in \Delta^{\mathcal{M}-1}$ such that $w^* \in \Delta^{\mathcal{M}-1}$ minimizes $K_0(w) \equiv \sup_{r\in\rho} |\Psi_0(r,w)|$, $C_0(w) \equiv \int_\rho \Psi_0^2(r,w)\,\mathrm{d}r$ or $A_0(w) \equiv \int_\rho \frac{\Psi_0^2(r,w)}{r(1-r)}\,\mathrm{d}r$, which are the population counterparts of $K_G(w)$, $C_G(w)$ and $A_G(w)$, respectively, and where $\Psi_0(w,r) \equiv G^{-1}\sum_{t=f-G-h+1}^{f-h} E[\xi_{t+h}(w,r)]$ is the population counterpart of $\Psi_G(w,r)$.*

**Assumption 6** (Anderson–Darling assumption). *There exists $0 < \delta < 0.5$ such that $\sup_{w\in\Delta^{\mathcal{M}-1}} \left| \int_0^\delta \frac{\Psi_G^2(w,r)-\Psi_0^2(w,r)}{r(1-r)}\,\mathrm{d}r \right| \xrightarrow{a.s.} 0$ and $\sup_{w\in\Delta^{\mathcal{M}-1}} \left| \int_{1-\delta}^1 \frac{\Psi_G^2(w,r)-\Psi_0^2(w,r)}{r(1-r)}\,\mathrm{d}r \right| \xrightarrow{a.s.} 0$.*

Assumption 1 is a dependence assumption frequently used in the forecasting literature (Giacomini and White, 2006; Corradi and Swanson, 2006a; Rossi and Sekhposyan, 2013). It allows the DGP to be fairly heterogeneous, but limits its memory and rules out unit-root processes, for example. This assumption is not restrictive in the sense that it is possible to replace it by an alternative one, provided that also leads to a strong or weak law of large numbers. In the latter case, consistency weakens to convergence in probability.

Assumption 2 lets the researcher focus on a specific part of the predictive distribution. For example, $\rho = [0, 0.05]$ is appropriate when performing VaR analysis at the 5% level. Assumption 3 is a mild assumption on the continuity of the combined CDF, which is satisfied in most applications in macroeconometrics and finance. Assumption 4 sets the estimation scheme, using finite (rolling) windows to estimate the parameters of the predictive densities and a "large" sample period used to estimate the combination weights. The former is necessary as the mixing property of the observables is only guaranteed to carry over to functions – in this case the predictive densities – of a finite number of observables. The latter part ($G \to \infty$) is required to invoke a law of large numbers. Assumption 5 is an identification condition. It covers the case of correct specification, that is, if the true predictive distribution can be expressed as the convex combination of the individual predictive distributions, corresponding to $\sum_{m=1}^{\mathcal{M}} w_m^* \Phi^m_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}) = \Phi^*_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})$ for all $t$. It also allows for misspecification, provided there is a unique minimizer of the population objective function.[11] In the former case, the population

[11] For an overview of the estimation of misspecified models, see White (1994).

objective function is zero at the true weight vector $w^*$, that is $K_0(w^*) = C_0(w^*) = A_0(w^*) = 0$, as the population CDF of the PIT is the 45 degree line. In the case of misspecification, the different population objective functions might yield different minimizers, therefore the pseudo-true weight vector $w^*$ might differ across estimators.[12]

Assumption 6 is a technical condition, which is only required for the Anderson–Darling-type objective function $A_G(w)$ and only if $\rho$ contains 0 or 1. This assumption ensures that the discrepancy between the objective function and its population counterpart remains asymptotically negligible uniformly in $w$ in a neighborhood of the endpoints of $[0,1]$. This difficulty arises in the case of the Anderson–Darling objective function because the weighting function $[r(1-r)]^{-1}$ is not integrable over $[0,1]$, with singularities occurring at the endpoints. To avoid introducing additional technical details, Assumption 6 is stated directly, rather than as a result that follows from low-level assumptions. In a wide range of Monte Carlo exercises (see Section 4) I never encountered a situation when the Anderson–Darling-type estimator failed to converge.

**Theorem 1** (Consistency). *Under Assumptions 1 to 6, the estimator defined in Equation (15) is strongly consistent, that is* $\widehat{w} \xrightarrow{a.s.} w^*$, *where* $w^*$ *is the weight vector that minimizes the population objective function* $K_0(w), C_0(w)$ *or* $A_0(w)$.

*Proof.* See Appendix A. ■

## 3.2 KLIC-based estimator

In this subsection I state and discuss some additional assumptions guaranteeing that the KLIC-based estimator defined in Equation (16) is strongly consistent. Assumptions involving "for all $t$" are understood as $t$ ranges from $t = f - G - h + 1$ to $f - h$.

**Assumption 7** (Region of interest). *$\varrho$ is the finite union of closed, non-empty, non-singleton intervals on the support of the true conditional distribution* $\Phi^*_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})$.

**Assumption 8** (Existence). $E_{\phi^*}\left\{\log \phi^*_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}) 1[y_{t+h} \in \varrho]\right\}$ *exists for all* $t$.

**Assumption 9** (Continuity). *Over* $\varrho$, $\log \phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})$ *is continuous in* $w$ *for all* $t$.

**Assumption 10** (Dominance). *Over* $\varrho$, $|\log \phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})| \leq b(y_{t+h})$ *for all* $w \in \Delta^{\mathcal{M}-1}$, *and* $b(y_{t+h})$ *is integrable with respect to the distribution of* $y_{t+h}$ *for all* $t$.

**Assumption 11** (Moment condition). *Over* $\varrho$, $E|(\log \phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}))|^{k+\tau} < \Delta < \infty$ *for some* $\tau > 0$ *for all* $t$ *and for all* $w \in \Delta^{\mathcal{M}-1}$.

**Assumption 12** (Identification). *There exists a unique* $w^* \in \Delta^{\mathcal{M}-1}$ *such that* $w^* \in \Delta^{\mathcal{M}-1}$ *minimizes* $\overline{KLIC}_0$ *defined in Equation (13).*

[12] As a side-note, I mention that in some cases the identification assumption does not hold, as we saw in Example 1, where $w = (0,1)' \neq (1,0)' = \tilde{w}$ both deliver uniform PITs.

Assumption 7 lets the researcher focus on a specific part of the predictive distribution. Assumption 8 allows separation of the terms in the expectation operator and proceed from Equation (11) to Equation (12). Assumption 9 is a continuity assumption which is satisfied in most relevant applications. Assumption 10 is required to convert a pointwise strong law of large numbers into a uniform one. The moment condition imposed by Assumption 11 is necessary to invoke the same strong law of large numbers for mixing processes as in the case of the PIT-based estimators, but while in that case $|\xi_{t+h}(w,r)| \leq 1$ implies that all of its moments are uniformly bounded, in the case of the KLIC estimator this assumption needs to be stated. Assumption 12 is an identification condition, either assuming correct specification, corresponding to $\sum_{m=1}^{\mathcal{M}} w_m^* \Phi_{t+h}^m(y_{t+h}|\mathcal{I}_{t-R+1}^t) = \Phi_{t+h}^*(y_{t+h}|\mathcal{I}_{t-R+1}^t)$ for all $t$, and also allowing for misspecification, similarly to Assumption 5.

**Theorem 2** (Consistency)**.** *Under Assumptions 1, 4 and 7 to 12, the estimator defined in Equation (16) is strongly consistent, that is* $\widehat{w} \xrightarrow{a.s.} w^*$*, where* $w^*$ *is the weight vector that minimizes the population objective function* $\overline{KLIC}_0$*.*

*Proof.* See Appendix A. ■

*Remark.* Theorems 1 and 2 show consistency of the respective estimators but do not establish their asymptotic distribution. Asymptotic normality can be proved following Newey and McFadden (1994) if the all the entries of $w^*$ are in the interior of the parameter space. However, from an empirical perspective this seems to be a rather demanding condition. Alternatively, the results of Andrews (1999) suggest that the asymptotic distribution of the PIT- and KLIC-based estimators are more complicated if some elements of $w^*$ are on the boundary of the parameter space. The investigation of this topic is left for future research. ▲

# 4 Monte Carlo study

To investigate the finite sample behavior of the proposed forecast density combination estimator, I performed a number of Monte Carlo simulations using a variety of DGPs.

Before presenting the results, a few remarks are in order. All simulations were repeated 2000 times. Without loss of generality I used the true parameters of the individual predictive densities. Clearly, if the models' parameters entering the predictive densities were estimated, then the true combined density would likely be a different convex combination of the densities. However, Appendix D contains results for a DGP where the parameters of the predictive densities were estimated. The sample sizes used to estimate the weight vector $w$ vary as $G = \{80, 200, 500, 1000, 2000\}$, offering guidance to practitioners using long time series (in finance, for example) and relatively smaller samples (in macroeconomics, for example).

To preserve space, this section shows the distribution of the estimators for $G = \{80, 500, 2000\}$, while the remaining cases of $G = \{200, 1000\}$ can be found in Appendix D. The likelihood functions of the models are listed in Appendix F. In what follows, I first describe each DGP in the Monte Carlo exercise, then I discuss the simulation results.

## 4.1 Monte Carlo set-up – DGP 1

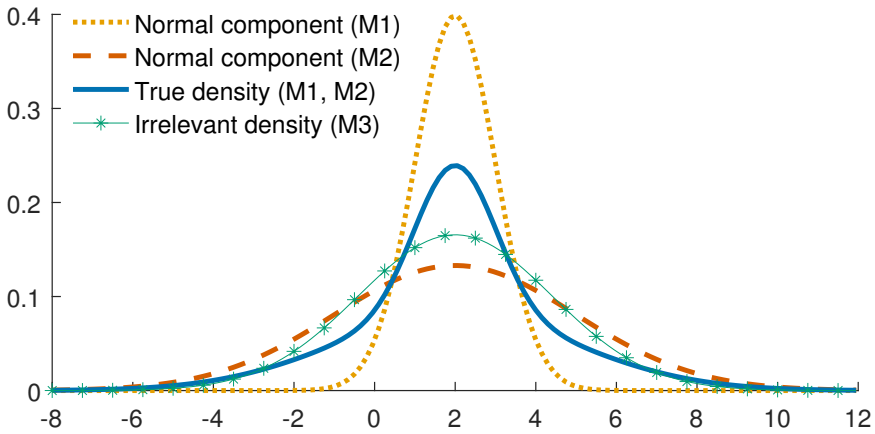
Both DGP 1a and DGP 1b feature three AR(1) models with *iid*. normal error terms. The models labeled as M1, M2 and M3 are given by

$$y_{t+h} = c^{(j)} + \rho_1^{(j)} y_t + \varepsilon_{t+h} \qquad \varepsilon_{t+h} \overset{\text{iid.}}{\sim} \mathcal{N}(0, \sigma_j^2)\,, \tag{17}$$

where the superscript $j \in \{1, 2, 3\}$ corresponds to models M1, M2 and M3, respectively. DGP 1a demonstrates the estimators' performance in a one-step-ahead forecasting scenario ($h = 1$), while DGP 1b mimics a two-step-ahead forecasting exercise ($h = 2$). I consider direct and not iterated density forecasts as the former offer the advantage of closed-form expressions of predictive densities, which implies no additional simulation burden.[13] However, this paper's framework allows for both direct and iterated forecasts.

In both cases, the true DGP is the mixture of models M1 and M2, with weights $(w_1, w_2)' = (0.4, 0.6)'$. M3 is added to demonstrate how the different estimators compare in eliminating this irrelevant density ($w_3 = 0$). Furthermore, M3 is specified such that its predictive density's first three moments match those of the true mixture density. The parameters are shown in Table 1 and Figure 5 displays the predictive densities.

Figure 5: DGPs 1a and 1b – Comparison of predictive densities



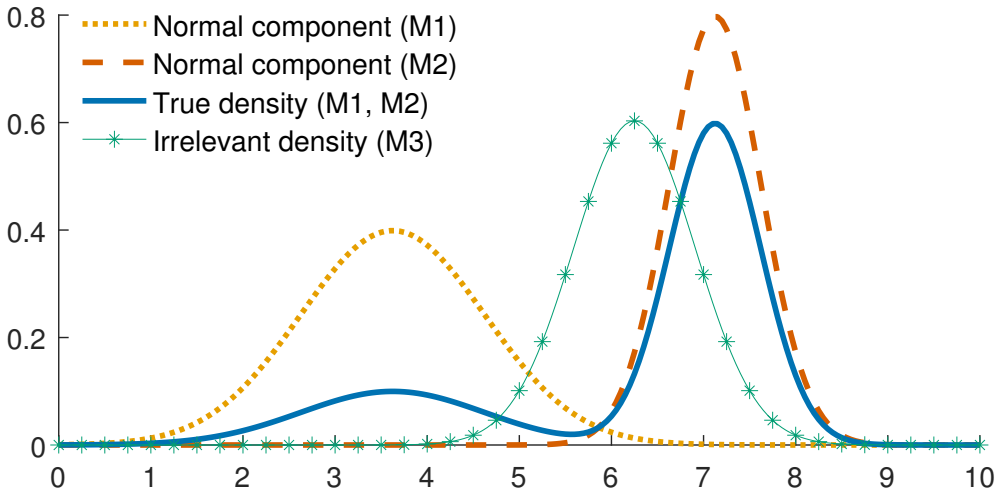

*Note*: The figure shows the predictive density of $y_{t+1}$ (that of $y_{t+2}$ in the case of DGP 1b), according to each model (M1, M2, M3) in the model set, and according to the true, mixture density. The values of $y_t$ are set to the unconditional expected value of $y_t$.

[13] Based on a wide range of models estimated using 170 US macroeconomic time series, Marcellino et al. (2006) suggested that iterated point forecasts often outperform their direct counterparts in the mean squared forecast error sense. Whether this holds in the case of density forecasts is certainly an interesting question but it is outside of the scope of the present study.

## 4.2 Monte Carlo set-up – DGP 2

In this experiment, I investigate the estimators' performance when the true DGP implies a bimodal predictive density. This could be relevant in a number of empirical applications, such as when forecasting output. In this case, the probability mass around the lower mode corresponds to periods of weak economic activity, while the majority of the mass is around a higher mode, corresponding to normal times. All three models M1, M2 and M3 share the common autoregressive structure as in the case of DGP 1 with $h = 1$, specified in Equation (17). The mixture weights are $(w_1, w_2, w_3)' = (0.25, 0.75, 0)'$. Table 1 contains the models' parameters, while Figure 6 shows the corresponding predictive densities.

Figure 6: DGP 2 – Comparison of predictive densities




*Note*: The figure shows the predictive density of $y_{t+1}$, according to each model (M1, M2, M3) in the model set, and according to the true, mixture density. The value of $y_t$ is set to the unconditional expected value of $y_t$.

## 4.3 Monte Carlo set-up – DGP 3

In order to demonstrate that the estimators perform well in a real-world scenario and to anticipate the empirical application, the parameters of DGP 3 are based on estimates of US industrial production.[14] Using monthly data on US industrial production growth between January 2008 and February 2016, I estimated two AR(2) models, specified as

$$M1 : y_{t+1} = c_1 + \rho_1^{(1)} y_t + \rho_2^{(1)} y_{t-1} + \sigma_1 \nu_{t+1} \qquad \nu_{t+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0,1)\,, \tag{18}$$

$$M2 : y_{t+1} = c_2 + \rho_1^{(2)} y_t + \rho_2^{(2)} y_{t-1} + \sigma_2 \varepsilon_{t+1} \qquad \varepsilon_{t+1} \overset{\text{iid.}}{\sim} t_\nu^{\text{s}}\,, \tag{19}$$

where $t_\nu^{\text{s}}$ stands for the standardized Student's t-distribution, with $\nu > 2$ degrees of freedom. The mixture weights are $(w_1, w_2)' = (0.4, 0.6)'$, and I added a normal AR(2)

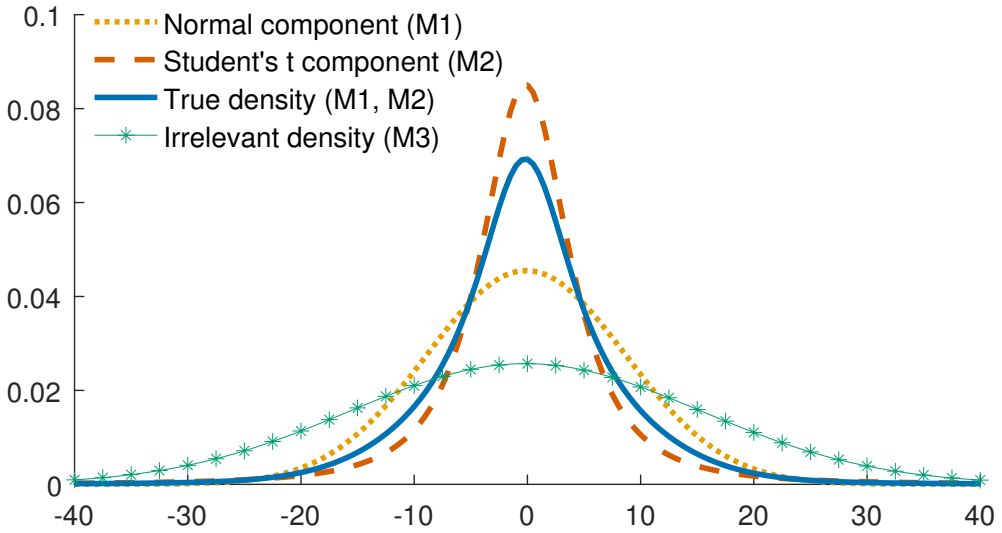
[14] More details on the data can be found in Section 5.

process to the model set, specified as

$$M3 : y_{t+1} = c_3 + \rho_1^{(3)} y_t + \rho_2^{(3)} y_{t-1} + \sigma_3 \eta_{t+1} \qquad \eta_{t+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0,1)\,, \tag{20}$$

where the parameterization $c_3 = c_1 w_1 + c_2 w_2$, $\rho_1^{(3)} = w_1 \rho_1^{(1)} + w_2 \rho_1^{(2)}$, $\rho_2^{(3)} = w_1 \rho_2^{(1)} + w_2 \rho_2^{(2)}$ and $\sigma_3^2 = w_1 \sigma_1^2 + w_2 \sigma_2^2$ guarantees that the first two moments of the predictive distribution of $y_{t+1}$ are the same for the mixture and the irrelevant models. Table 1 contains the parameters of the models and Figure 7 presents the predictive densities.

Figure 7: DGP 3 – Comparison of predictive densities




*Note*: The figure shows the predictive density of $y_{t+1}$, according to each model (M1, M2, M3) in the model set, and according to the true, mixture density. The values of $y_t$ and $y_{t-1}$ are set to the unconditional expected value of $y_t$.

Table 1: Simulation design

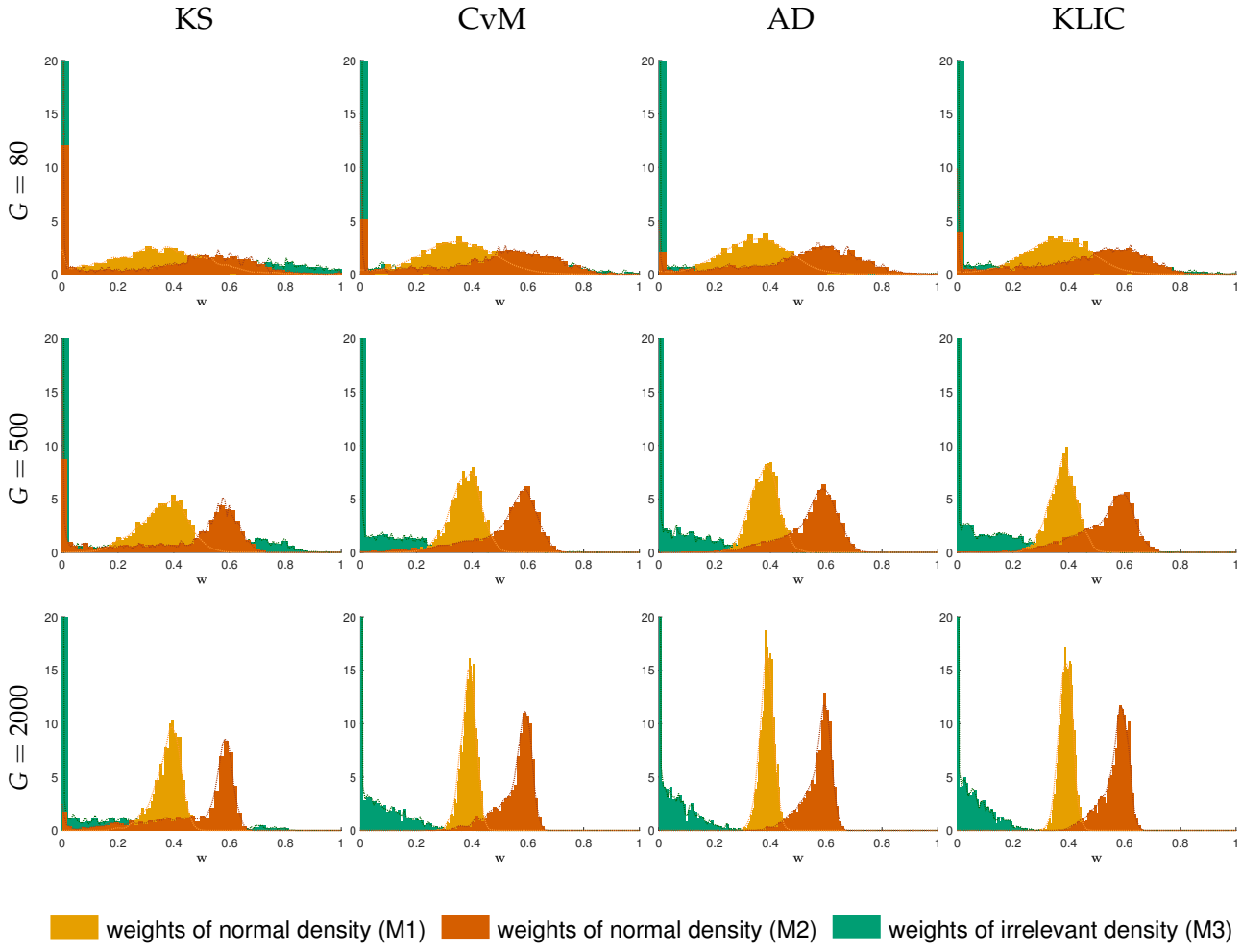
| | Model | $c$ | $\rho_1$ | $\rho_2$ | $\sigma^2$ | $\nu$ | $w_j$ |
|---|---|---|---|---|---|---|---|
| DGP1 | M1 | 1 | 0.5 | 0 | 1 | — | 0.4 |
| | M2 | 1 | 0.5 | 0 | 9 | — | 0.6 |
| | M3 | 1 | 0.5 | 0 | 5.8 | — | 0 |
| DGP2 | M1 | −2 | 0.9 | 0 | 1 | — | 0.25 |
| | M2 | 1.5 | 0.9 | 0 | 0.25 | — | 0.75 |
| | M3 | 0.63 | 0.9 | 0 | 0.44 | — | 0 |
| DGP3 | M1 | −0.02 | 0.31 | 0.21 | 76.87 | — | 0.4 |
| | M2 | −0.11 | 0.24 | 0.32 | 350.32 | 2.10 | 0.6 |
| | M3 | −0.07 | 0.27 | 0.27 | 240.94 | — | 0 |

*Note*: For each DGP and each forecasting model (M1 – M3) the table lists the constant ($c$), the autoregressive parameters ($\rho_1, \rho_2$), and the variance parameter ($\sigma^2$) of the predictive distribution. M2 in DGP 3 is specified using a Student's t predictive distribution, with degrees of freedom parameter $\nu$. For each DGP, the predictive distributions of M1 and M2 are weighted using the weights in the last column, $w_j$.

## 4.4 Monte Carlo results

Considering DGPs 1a and 1b first, in Figures 8 and 9 we can see that as the sample size increases from $G = 80$ to $G = 2000$, all the estimators deliver more precise estimates of the true parameter vector $w = (0.4, 0.6, 0)'$, demonstrating consistency. However, it is also apparent that the Anderson–Darling- and the KLIC-based estimators dominate the other two, both in terms of location and dispersion, at all sample sizes considered. This ranking holds in all the Monte Carlo experiments.

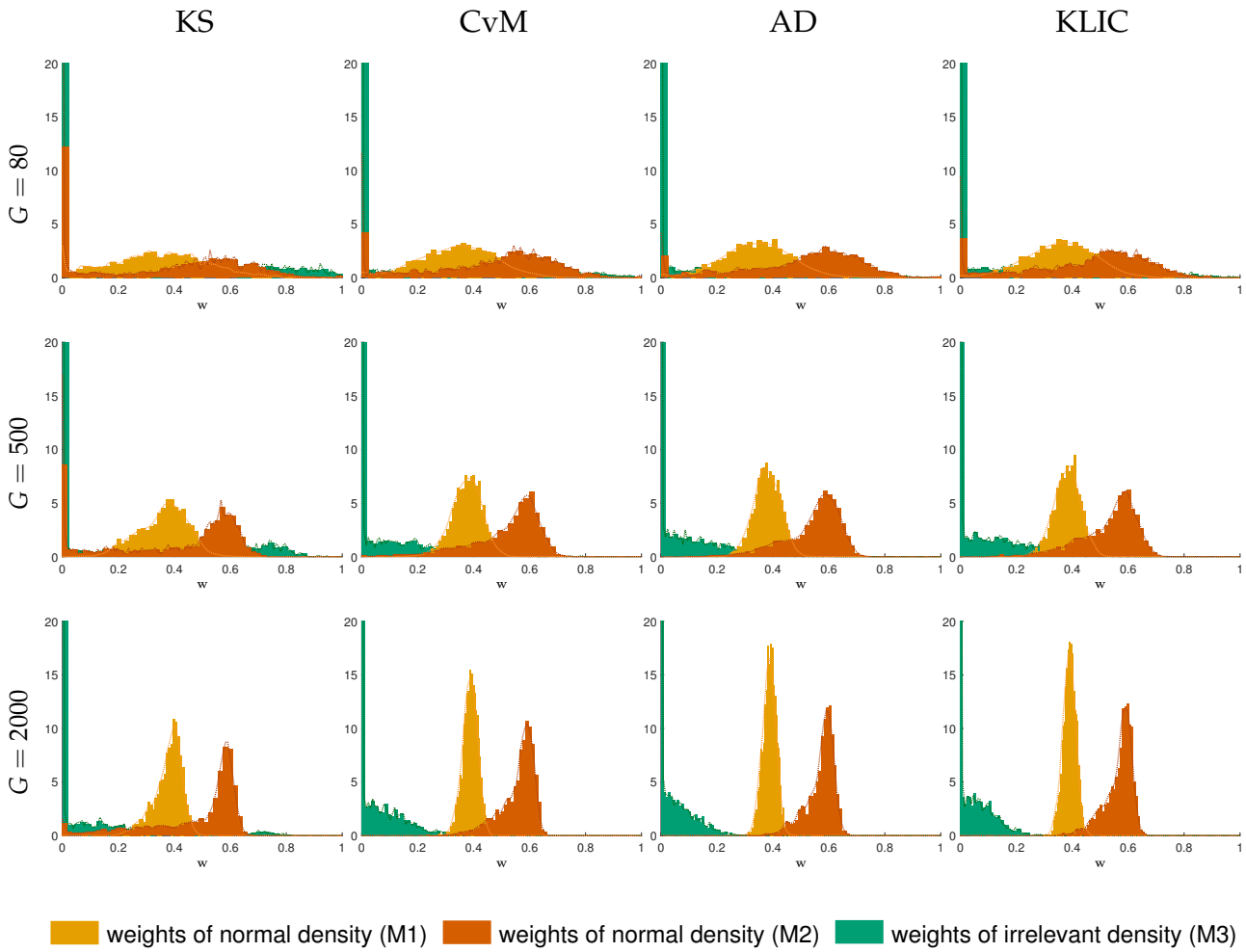
Figure 8: Monte Carlo results for DGP 1a, true parameter vector $w = (0.4, 0.6, 0)'$




*Note*: $G$ denotes the sample size. KS, CvM, AD and KLIC stand for the Kolmogorov–Smirnov-, the Cramer–von Mises-, the Anderson–Darling- and the KLIC-based estimators, respectively. Histograms and kernel density estimates are based on 2000 Monte Carlo replications.

Also, it is worth mentioning that while the AD and the KLIC estimators perform well at eliminating the irrelevant density (M3) even at sample size $G = 80$, the KS estimator still gives considerable weight to this model with large probability, and this improves rather slowly as $G$ increases. Moreover, we can see that increasing the forecast horizon from $h = 1$ to $h = 2$ has no impact on the estimators' performance.

Figure 9: Monte Carlo results for DGP 1b, true parameter vector $w = (0.4, 0.6, 0)'$




*Note*: $G$ denotes the sample size. KS, CvM, AD and KLIC stand for the Kolmogorov–Smirnov-, the Cramer–von Mises-, the Anderson–Darling- and the KLIC-based estimators, respectively. Histograms and kernel density estimates are based on 2000 Monte Carlo replications.

Tables 2 and 3 display the bias, variance and mean squared error for all sample sizes and objective functions. The figures support that the Kolmogorov–Smirnov objective function performs considerably worse than its competitors. As the KS-estimator is based on the largest deviation of the PIT from the 45 degree line, this estimator is unable to distinguish between the densities in such a nuanced way as the rest of the estimators.

Table 2: DGP 1a, Monte Carlo summary statistics for different sample sizes $G$ and objective functions $K_G(w), C_G(w), A_G(w)$ and $\text{KLIC}_G(w)$

| Sample size | Statistic | KS | | | CvM | | | AD | | | KLIC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $G = 80$ | Bias | −0.05 | −0.26 | 0.31 | −0.06 | −0.16 | 0.21 | −0.06 | −0.10 | 0.15 | −0.04 | −0.15 | 0.20 |
| | Var | 0.03 | 0.08 | 0.13 | 0.00 | 0.00 | 0.00 | 0.02 | 0.06 | 0.08 | 0.01 | 0.05 | 0.07 |
| | MSE | 0.03 | 0.14 | 0.23 | 0.02 | 0.08 | 0.13 | 0.02 | 0.06 | 0.08 | 0.02 | 0.07 | 0.11 |
| $G = 200$ | Bias | −0.05 | −0.22 | 0.27 | −0.04 | −0.12 | 0.16 | −0.03 | −0.08 | 0.11 | −0.03 | −0.11 | 0.13 |
| | Var | 0.02 | 0.07 | 0.11 | 0.00 | 0.00 | 0.00 | 0.01 | 0.03 | 0.05 | 0.01 | 0.03 | 0.04 |
| | MSE | 0.02 | 0.12 | 0.19 | 0.01 | 0.05 | 0.08 | 0.01 | 0.03 | 0.04 | 0.01 | 0.04 | 0.05 |
| $G = 500$ | Bias | −0.04 | −0.20 | 0.24 | −0.02 | −0.08 | 0.10 | −0.02 | −0.05 | 0.07 | −0.02 | −0.07 | 0.09 |
| | Var | 0.01 | 0.06 | 0.09 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.02 | 0.00 | 0.01 | 0.01 |
| | MSE | 0.01 | 0.10 | 0.15 | 0.00 | 0.02 | 0.03 | 0.00 | 0.01 | 0.02 | 0.00 | 0.02 | 0.02 |
| $G = 1000$ | Bias | −0.03 | −0.15 | 0.18 | −0.02 | −0.06 | 0.07 | −0.02 | −0.04 | 0.06 | −0.01 | −0.05 | 0.06 |
| | Var | 0.00 | 0.04 | 0.06 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 |
| | MSE | 0.01 | 0.06 | 0.09 | 0.00 | 0.01 | 0.02 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 |
| $G = 2000$ | Bias | −0.02 | −0.12 | 0.15 | −0.01 | −0.04 | 0.05 | −0.01 | −0.03 | 0.04 | −0.01 | −0.03 | 0.04 |
| | Var | 0.00 | 0.03 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| | MSE | 0.00 | 0.04 | 0.07 | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |

*Note*: In the four main columns with headers KS, CvM, AD and KLIC, the table shows the estimates of the bias, variance (Var) and mean squared error (MSE) for each of the components of the weight vector $w$. True weights: $w = (0.4, 0.6, 0)'$. Statistics are based on 2000 Monte Carlo replications.

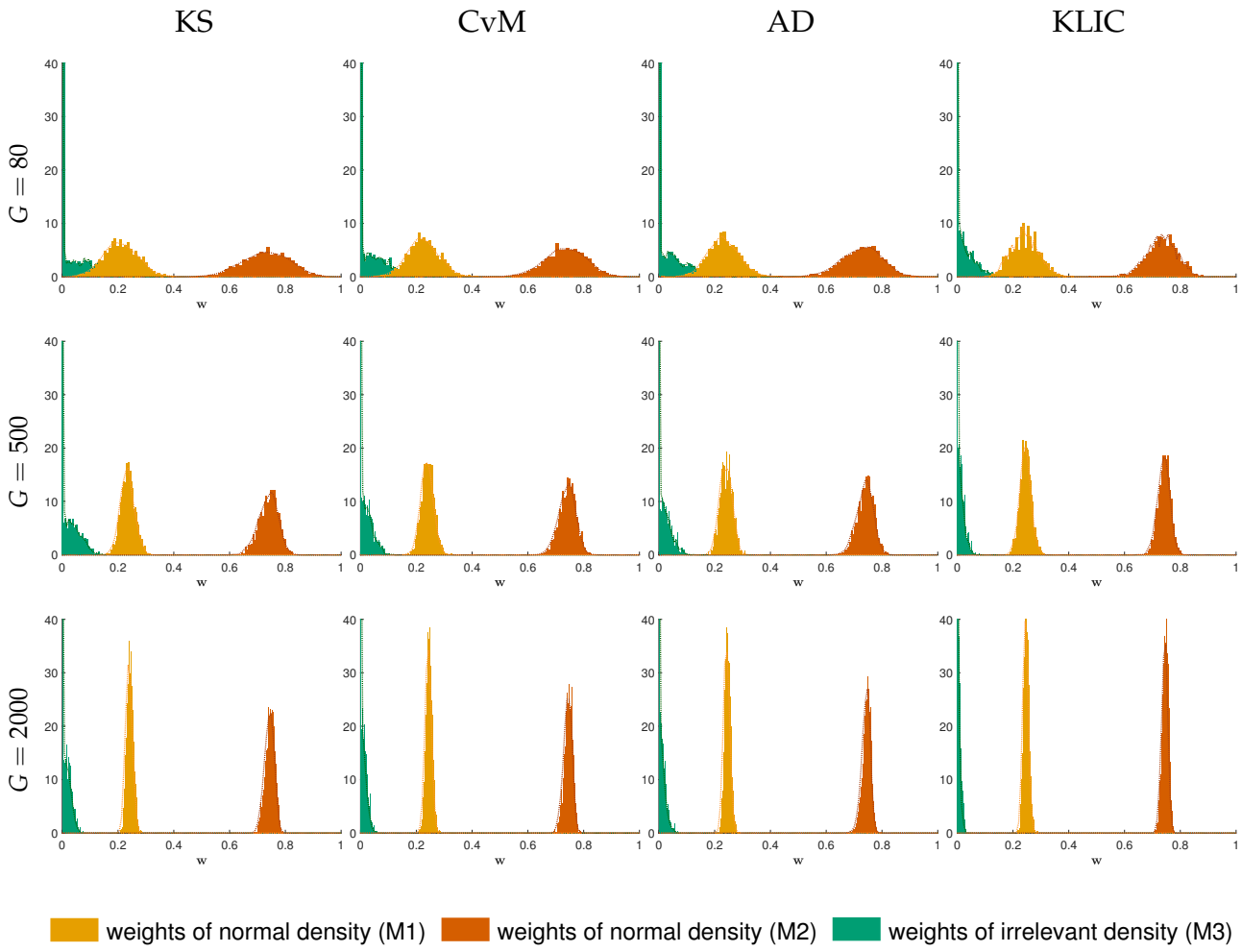
Table 3: DGP 1b, Monte Carlo summary statistics for different sample sizes $G$ and objective functions $K_G(w), C_G(w), A_G(w)$ and $\text{KLIC}_G(w)$

| Sample size | Statistic | KS | | | CvM | | | AD | | | KLIC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $G = 80$ | Bias | −0.06 | −0.25 | 0.31 | −0.06 | −0.15 | 0.21 | −0.06 | −0.10 | 0.15 | −0.04 | −0.15 | 0.19 |
| | Var | 0.03 | 0.08 | 0.13 | 0.02 | 0.06 | 0.08 | 0.01 | 0.05 | 0.06 | 0.01 | 0.05 | 0.07 |
| | MSE | 0.03 | 0.14 | 0.23 | 0.02 | 0.08 | 0.12 | 0.02 | 0.06 | 0.08 | 0.02 | 0.07 | 0.10 |
| $G = 200$ | Bias | −0.05 | −0.24 | 0.29 | −0.04 | −0.12 | 0.16 | −0.03 | −0.07 | 0.11 | −0.03 | −0.10 | 0.13 |
| | Var | 0.01 | 0.07 | 0.12 | 0.01 | 0.04 | 0.05 | 0.01 | 0.02 | 0.03 | 0.01 | 0.03 | 0.03 |
| | MSE | 0.02 | 0.12 | 0.20 | 0.01 | 0.05 | 0.08 | 0.01 | 0.03 | 0.04 | 0.01 | 0.04 | 0.05 |
| $G = 500$ | Bias | −0.04 | −0.20 | 0.24 | −0.02 | −0.08 | 0.10 | −0.02 | −0.05 | 0.07 | −0.02 | −0.07 | 0.08 |
| | Var | 0.01 | 0.05 | 0.09 | 0.00 | 0.02 | 0.02 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 |
| | MSE | 0.01 | 0.09 | 0.14 | 0.00 | 0.02 | 0.03 | 0.00 | 0.01 | 0.02 | 0.00 | 0.01 | 0.02 |
| $G = 1000$ | Bias | −0.03 | −0.16 | 0.19 | −0.02 | −0.05 | 0.07 | −0.02 | −0.04 | 0.06 | −0.01 | −0.04 | 0.05 |
| | Var | 0.00 | 0.04 | 0.07 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 |
| | MSE | 0.01 | 0.07 | 0.10 | 0.00 | 0.01 | 0.02 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 |
| $G = 2000$ | Bias | −0.02 | −0.12 | 0.14 | −0.01 | −0.04 | 0.05 | −0.01 | −0.03 | 0.04 | −0.01 | −0.03 | 0.04 |
| | Var | 0.00 | 0.03 | 0.04 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | MSE | 0.00 | 0.04 | 0.06 | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

*Note*: In the four main columns with headers KS, CvM, AD and KLIC, the table shows the estimates of the bias, variance (Var) and mean squared error (MSE) for each of the components of the weight vector $w$. True weights: $w = (0.4, 0.6, 0)'$. Statistics are based on 2000 Monte Carlo replications.

Next, in the case of DGP 2, Figure 10 clearly demonstrates that in an empirically potentially relevant scenario, even the Kolmogorov–Smirnov estimator delivers excellent results, on par with the CvM, AD and KLIC estimators, even for such small samples as $G = 80$. It is also worth noting that in this case, the difference between the estimators is visually indistinguishable both in terms of location and dispersion of the estimates. The individual forecasting models M1 and M2 concentrate mass in different areas of the real line, which considerably improves the performance of all estimators.

Figure 10: Monte Carlo results for DGP 2, true parameter vector $w = (0.25, 0.75, 0)'$




*Note*: $G$ denotes the sample size. KS, CvM, AD and KLIC stand for the Kolmogorov–Smirnov-, the Cramer–von Mises-, the Anderson–Darling- and the KLIC-based estimators, respectively. Histograms and kernel density estimates are based on 2000 Monte Carlo replications.

As Table 4 shows, all the estimators perform excellently when the individual models assign most of the probability mass to fairly remote regions. Compared to the previous DGPs, the Kolmogorov–Smirnov estimator's performance is remarkable, as the column labeled KS reveals.

Table 4: DGP 2, Monte Carlo summary statistics for different sample sizes $G$ and objective functions $K_G(w), C_G(w), A_G(w)$ and $\text{KLIC}_G(w)$

| Sample size | Statistic | KS | | | CvM | | | AD | | | KLIC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $G = 80$ | Bias | −0.04 | −0.02 | 0.05 | −0.02 | −0.01 | 0.04 | −0.02 | −0.02 | 0.04 | −0.00 | −0.02 | 0.02 |
| | Var | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| | MSE | 0.01 | 0.01 | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| $G = 200$ | Bias | −0.02 | −0.01 | 0.04 | −0.01 | −0.01 | 0.02 | −0.01 | −0.01 | 0.02 | −0.00 | −0.01 | 0.01 |
| | Var | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | MSE | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $G = 500$ | Bias | −0.01 | −0.01 | 0.02 | −0.01 | −0.01 | 0.01 | −0.01 | −0.01 | 0.01 | −0.00 | −0.01 | 0.01 |
| | Var | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | MSE | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $G = 1000$ | Bias | −0.01 | −0.01 | 0.02 | −0.00 | −0.01 | 0.01 | −0.00 | −0.01 | 0.01 | −0.00 | −0.00 | 0.01 |
| | Var | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | MSE | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $G = 2000$ | Bias | −0.01 | −0.00 | 0.01 | −0.00 | −0.00 | 0.01 | −0.00 | −0.00 | 0.01 | −0.00 | −0.00 | 0.00 |
| | Var | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | MSE | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

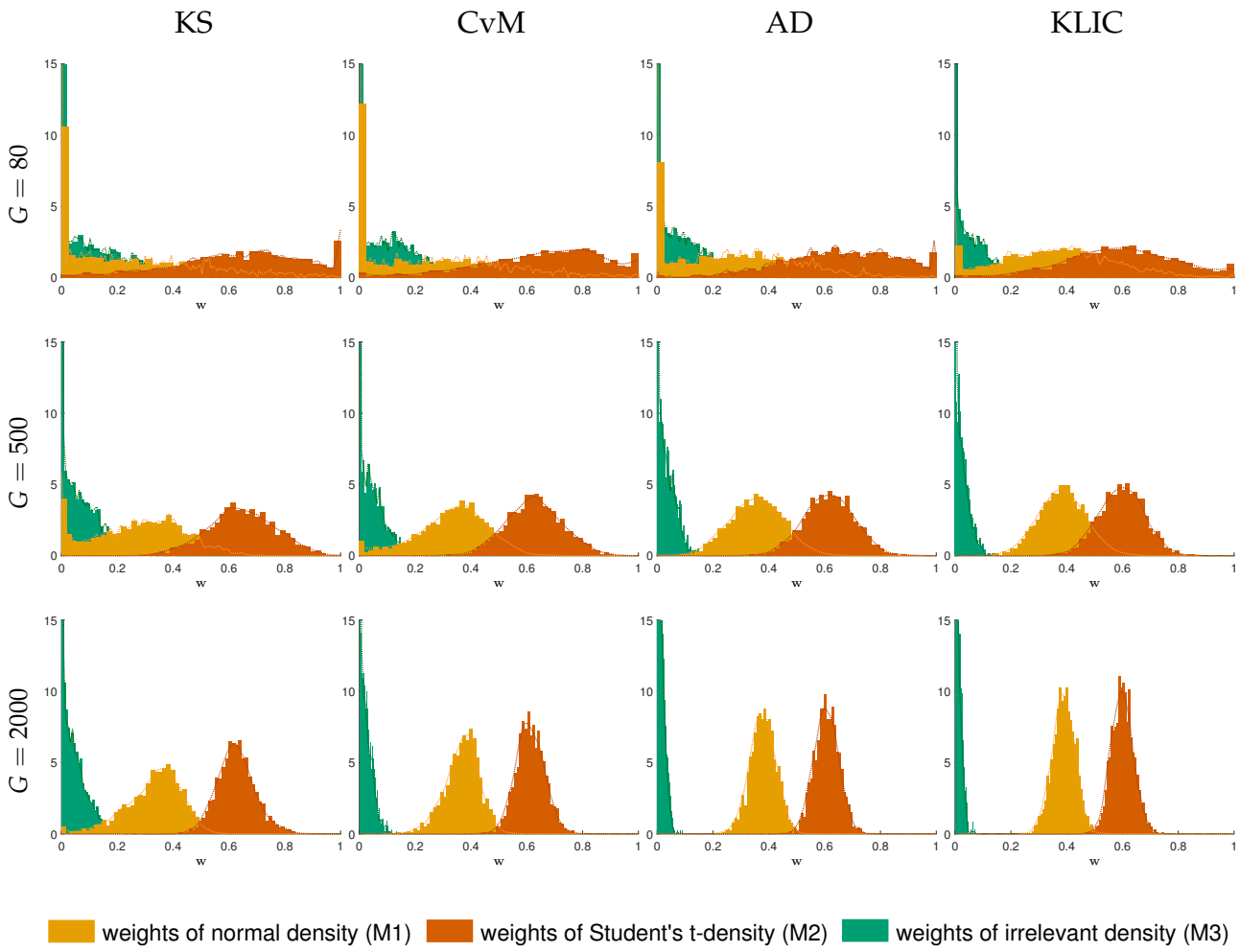
*Note*: In the four main columns with headers KS, CvM, AD and KLIC, the table shows the estimates of the bias, variance (Var) and mean squared error (MSE) for each of the components of the weight vector $w$. True weights: $w = (0.25, 0.75, 0)'$. Statistics are based on 2000 Monte Carlo replications.

In the case of DGP 3, which is based on empirically relevant models, we can see again in Figure 11 that the AD and KLIC estimators dominate the other two, with the latter delivering slightly less dispersed estimates. Table 5 shows that the relative ranking of the estimators is similar to the case of DGPs 1a and 1b, with the KLIC and the Anderson–Darling estimators clearly delivering more precise estimates in the mean squared error sense. Intuitively, this result is due to the similar means implied by the individual models, in which case the Kolmogorov–Smirnov estimator performs poorly.

In addition to these four DGPs, Appendix D reports additional simulation results, covering: (*i*) more persistent time series, (*ii*) the mixture of three predictive densities, resulting in a trimodal true density, (*iii*) the mixture of autoregressive conditionally heteroskedastic and AR(1) models, and (*iv*) predictive densities with estimated parameters. All the additional simulations confirm the conclusions, which are as follows.

The estimators based on the Anderson–Darling statistic and the KLIC typically outperform the Kolmogorov–Smirnov and Cramer–von Mises estimators in the mean squared error sense. Furthermore, a sample size as low as $G = 200$ observations is often sufficient for fairly precise weight estimates, with no economically meaningful differences

Figure 11: Monte Carlo results for DGP 3, true parameter vector $w = (0.4, 0.6, 0)'$




*Note*: $G$ denotes the sample size. KS, CvM, AD and KLIC stand for the Kolmogorov–Smirnov-, the Cramer–von Mises-, the Anderson–Darling- and the KLIC-based estimators, respectively. Histograms and kernel density estimates are based on 2000 Monte Carlo replications.

between the CvM, AD and KLIC-based estimators. These numerical results confirm the consistency of the proposed estimators and suggest that in empirical applications, the Anderson–Darling- or the KLIC-type estimator should be preferred.

Table 5: DGP 3, Monte Carlo summary statistics for different sample sizes $G$ and objective functions $K_G(w), C_G(w), A_G(w)$ and $\text{KLIC}_G(w)$

| Sample size | Statistic | KS | | | CvM | | | AD | | | KLIC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $G = 80$ | Bias | −0.15 | 0.04 | 0.11 | −0.14 | 0.05 | 0.09 | −0.13 | 0.06 | 0.07 | −0.03 | −0.01 | 0.04 |
| | Var | 0.06 | 0.05 | 0.02 | 0.06 | 0.05 | 0.01 | 0.05 | 0.04 | 0.01 | 0.04 | 0.04 | 0.00 |
| | MSE | 0.08 | 0.05 | 0.03 | 0.08 | 0.05 | 0.02 | 0.06 | 0.04 | 0.01 | 0.04 | 0.04 | 0.00 |
| $G = 200$ | Bias | −0.14 | 0.05 | 0.09 | −0.11 | 0.05 | 0.06 | −0.08 | 0.03 | 0.04 | −0.02 | −0.00 | 0.02 |
| | Var | 0.04 | 0.03 | 0.01 | 0.04 | 0.02 | 0.01 | 0.02 | 0.02 | 0.00 | 0.02 | 0.02 | 0.00 |
| | MSE | 0.06 | 0.03 | 0.02 | 0.05 | 0.03 | 0.01 | 0.03 | 0.02 | 0.00 | 0.02 | 0.02 | 0.00 |
| $G = 500$ | Bias | −0.12 | 0.05 | 0.07 | −0.06 | 0.03 | 0.03 | −0.04 | 0.02 | 0.02 | −0.01 | −0.00 | 0.02 |
| | Var | 0.03 | 0.01 | 0.00 | 0.02 | 0.01 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 | 0.00 |
| | MSE | 0.04 | 0.02 | 0.01 | 0.02 | 0.01 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 | 0.00 |
| $G = 1000$ | Bias | −0.09 | 0.04 | 0.05 | −0.04 | 0.02 | 0.02 | −0.03 | 0.02 | 0.01 | −0.01 | 0.00 | 0.01 |
| | Var | 0.02 | 0.01 | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | MSE | 0.03 | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $G = 2000$ | Bias | −0.07 | 0.03 | 0.04 | −0.03 | 0.01 | 0.02 | −0.02 | 0.01 | 0.01 | −0.01 | 0.00 | 0.01 |
| | Var | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | MSE | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

*Note*: In the four main columns with headers KS, CvM, AD and KLIC, the table shows the estimates of the bias, variance (Var) and mean squared error (MSE) for each of the components of the weight vector $w$. True weights: $w = (0.4, 0.6, 0)'$. Statistics are based on 2000 Monte Carlo replications.

# 5 Empirical application

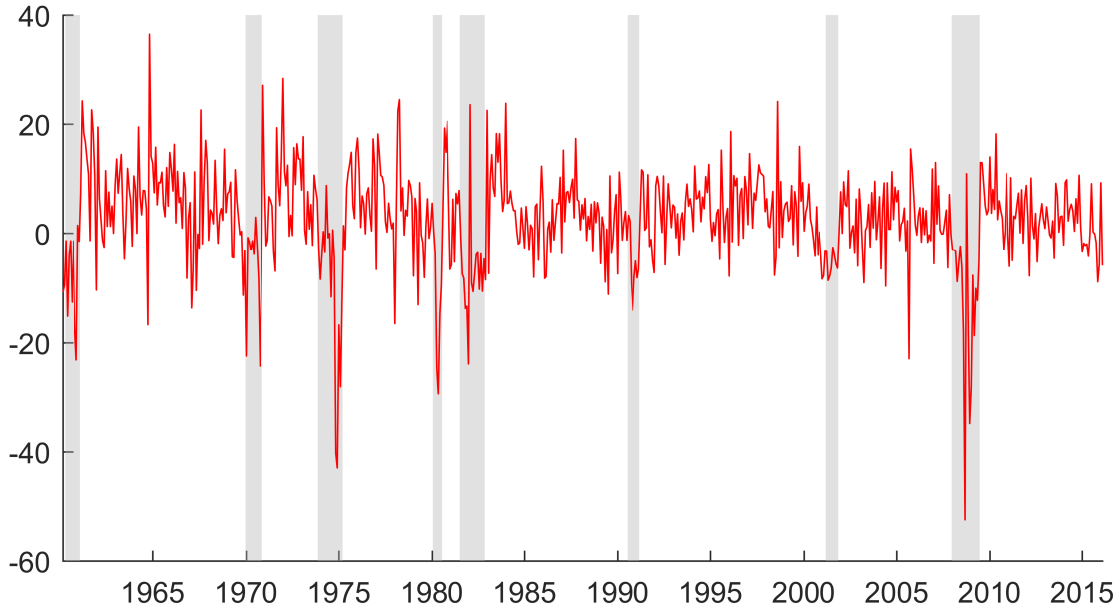
In this section I apply the proposed methodology to obtain one-month-ahead ($h = 1$) density forecast combinations of annualized US industrial production (IP) growth. Consider the time series and the unconditional distribution of annualized US IP growth between March 1960 and February 2016, shown in Figures 12 and 13, respectively. As we can see in Figure 13, the unconditional distribution shows more kurtosis ($\kappa = 7.47$) and is more negatively skewed ($s = -0.93$) than the normal distribution with the same mean ($\mu = 2.60$) and standard deviation ($\sigma = 9.03$), whose PDF is also plotted for ease of comparison, along with the kernel density estimate of IP growth.

While the non-Gaussian *un*conditional distribution does not necessarily imply non-Gaussian conditional distribution, it is worth investigating how the proposed data-dependent density forecast combination procedures — which are capable of generating a variety of forecast densities — perform in an empirical exercise.

## 5.1 Models and data

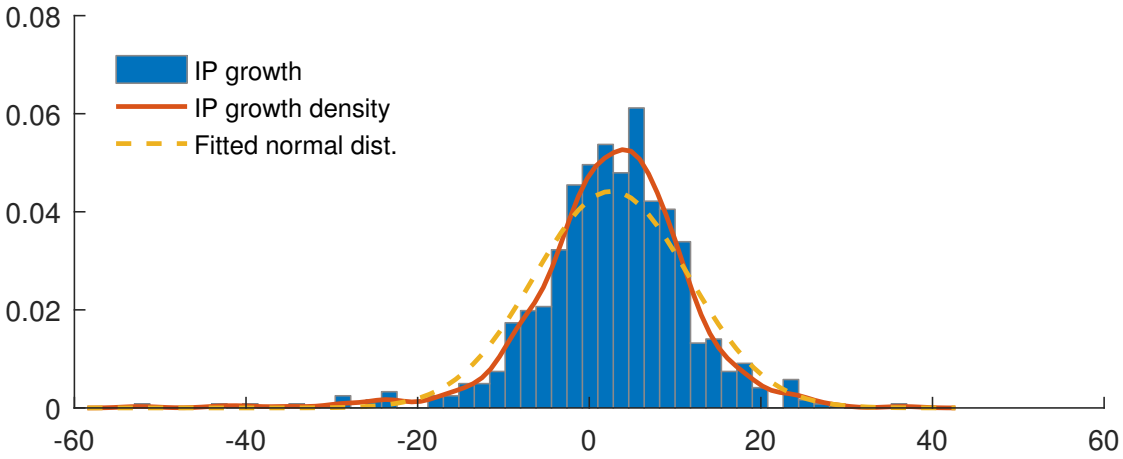
Based on their empirical success documented by Stock and Watson (2003), Granger and Jeon (2004), and more recently by Rossi and Sekhposyan (2014), I consider linear

Figure 12: Annualized US IP growth between March 1960 and February 2016




*Note*: Shaded areas are NBER recession periods.

Figure 13: Normalized histogram of annualized US IP growth between March 1960 and February 2016




Autoregressive Distributed Lag (ARDL) models of the following form:

$$y_{\tau+1} = c + \sum_{j=0}^{1} \beta_j y_{\tau-j} + \sum_{j=0}^{1} \gamma_j x_{\tau-j} + \sqrt{\sigma^2}\varepsilon_{\tau+1} \qquad \varepsilon_{\tau+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0,1)\,, \tag{21}$$

where $y_\tau$ is annualized US IP growth in month $\tau$, that is $y_\tau \equiv 1200\Delta \log(\text{IP}_\tau)$ where $\Delta$ is the first difference operator, $c$ is a constant term, $\beta_j$s are coefficients of the autoregressive terms while $\gamma_j$s are coefficients of the additional explanatory variables and $\sqrt{\sigma^2}$ scales the error term $\varepsilon_{\tau+1}$.[15] The lag length was specified following Granger and Jeon (2004), who demonstrated that on average, approximately two lags provide the best (in terms of

[15] Appendix F contains a detailed description of the models.

Root Mean Squared Error) forecasts for output series. All the data were obtained from the March 2016 vintage of the FRED-MD database (McCracken and Ng, 2016).

Some explanation regarding the $y_\tau$ and $x_\tau$ variables is in order. First, the chosen measure of industrial production is the INDPRO series (ID: 3), which measures total industrial production. Second, the possible elements of $x_\tau$ are the following variables, with the identifiers in the original database in parentheses: New Private Housing Permits SAAR (ID: 55), ISM : New Orders Index (ID: 61), S&P's Common Stock Price Index: Composite (ID: 80) and Moody's Seasoned Baa Corporate Bond Yield minus FEDFUNDS (ID: 100). Out of these four variables, I included them one by one, obtaining four different specifications. Furthermore, I estimated the pure AR(2) model, without additional regressors. The error term $\varepsilon_{\tau+1}$ is specified as *iid*. standard normal. In total, the model set $\mathcal{M}$ contains five models. To obtain stationary series, I took the log difference of the S&P index (and multiplied it by 100 to convert it into percents) and the log of the housing permits series, while the other variables were left untransformed, following McCracken and Ng (2016) and Carriero et al. (2015).[16] The resulting series are shown in Figure 14.

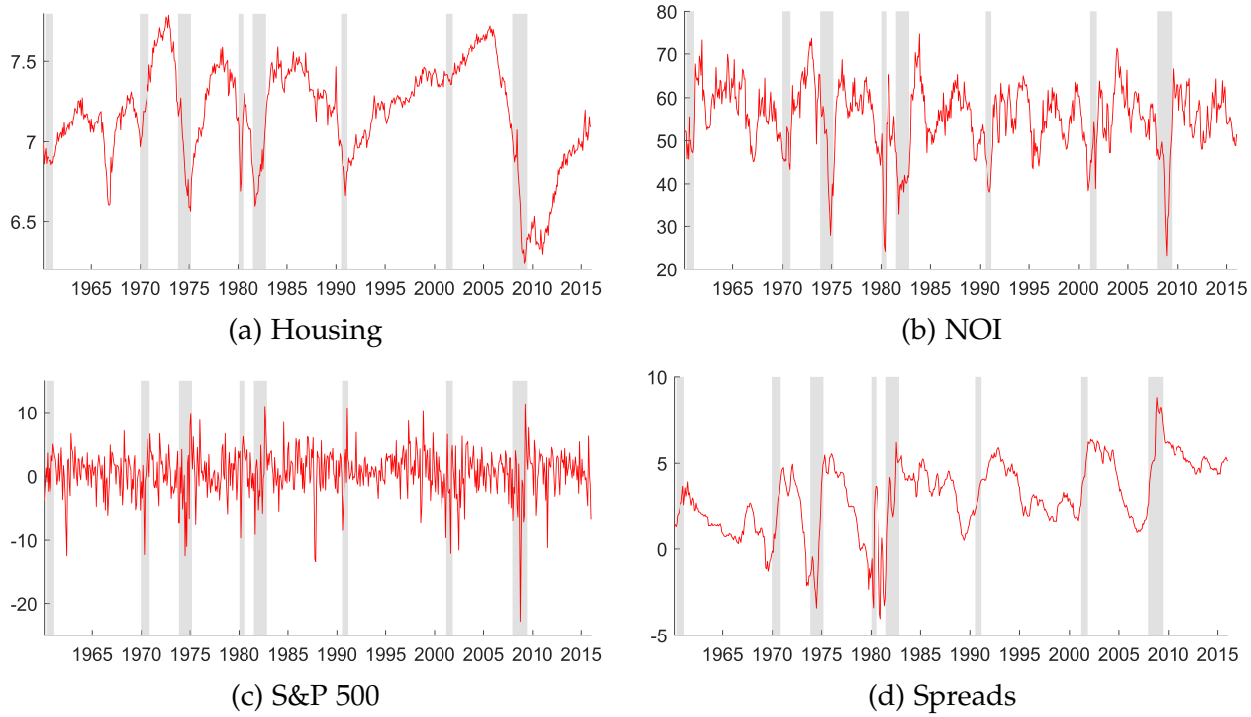
A salient feature of the housing data series is the almost uninterrupted increase since the early 1990s, which went into free fall during the recent financial crisis and recovered after the Great Recession, as Figure 14a shows. It is also remarkable that unlike in earlier recessions, housing permits did not plummet during the 2001 recession. Figure 14d reveals the sudden surge in corporate bond spreads at the onset of the financial crisis, which will turn out to be of great importance in this forecasting exercise.

All models are estimated using Maximum Likelihood in rolling windows of $R = 120$ months, with forecast origins $f$ and target dates $f + h$ ranging from February 1985 to January 2016 and March 1985 to February 2016, respectively.

To illustrate the estimation procedure, consider the first forecast origin $f$, corresponding to February 1985. The first window to estimate the models of Equation (21) contains data indexed by $\tau = \{\text{February 1960}, \ldots, \text{January 1970}\}$, which delivers out-of-sample (with respect to *this* estimation sample) predictive distributions for March 1970, by plugging in the observed values of the explanatory variables corresponding to February 1970. These predictive distributions are evaluated at the realized value of industrial production growth in March 1970, yielding the corresponding PITs. Then the window is moved one month forward. Given the results of the Monte Carlo experiments in Section 4, this procedure is repeated $G = 180$ times, until the last model estimation window reaches $\tau = \{\text{January 1975}, \ldots, \text{December 1984}\}$ and the last out-of-sample predictive distributions and PITs correspond to February 1985. This sequence of PITs form the input of the Anderson–Darling-type objective function $A_G(w)$ and the KLIC objective function $\text{KLIC}_G(w)$, resulting in weight estimates $\widehat{w}^{\text{AD}}_{\text{1985:M2}}$ and $\widehat{w}^{\text{KLIC}}_{\text{1985:M2}}$, respectively. Then, the actual realized values of the right hand side variables corresponding to $\tau =$ February 1985

[16] For each series, the Augmented Dickey–Fuller test (Dickey and Fuller, 1979) with drift and 12 lags indicates rejection of the null hypothesis of unit root at the 5% level.

Figure 14: Time series of all predictors between February 1960 and January 2016




(a) Housing (b) NOI

(c) S&P 500 (d) Spreads

*Note*: Housing stands for New Private Housing Permits, New Order Index stands for ISM: New Orders Index, S&P 500 is the S&P 500 stock index returns while Spread is Moody's Baa Corporate Bond Yield minus Fed funds rate. The series were transformed as described in the main text.

are substituted in the estimated last regressions and the previously obtained weights are used to construct either the Anderson–Darling- or the KLIC-based density forecasts corresponding to March 1985 and the corresponding out-of-sample value of the PIT is recorded. The above procedure is repeated for the remaining forecast origins, until $f$ reaches January 2016. As a result, we will have $P = 372$ observations of truly out-of-sample PITs, whose values were obtained using only preceding observations, both for model and weight estimation. This sequence of PITs spans March 1985 and February 2016, which is the out-of-sample evaluation period used to evaluate different combination schemes, as explained later.

To compare the PIT- and KLIC-based estimators to existing methods, the forecasting exercise was also performed using $(i)$ equal weights, $(ii)$ the AR(2), $(iii)$ a single model selected by the Bayesian Information Criterion (BIC) (Schwarz, 1978), and $(iv)$ Bayesian Model Averaging (BMA). All of these benchmarks have been demonstrated to perform well in empirical exercises.

Kascha and Ravazzolo (2010) and Rossi and Sekhposyan (2014) found that the equal weights combination scheme performs well when forecasting inflation with a large number of simple models. The AR(2) model with normal error terms, denoted by AR(2)-N, was shown to be a tough benchmark in point forecasting exercises, see for example Del Negro and Schorfheide (2013). Note that this benchmark could be interpreted as assigning a weight of 1 to the AR(2) model and a weight of 0 to all the other models.

The BIC of model $m$ at forecast origin $f$ is defined as

$$\text{BIC}_m \equiv -2 \sum_{t=f-R}^{f-1} \log \ell_m(y_{t+1}|z_t^m; \widehat{\theta}_m) + k_m \log(R)\,, \tag{22}$$

where $\ell_m(\cdot|\cdot)$ is the conditional likelihood function, $z_t^m$ is the vector of explanatory variables, and $\widehat{\theta}_m = (\widehat{c}, \widehat{\beta}_0, \widehat{\beta}_1, \widehat{\gamma}_0, \widehat{\gamma}_1, \widehat{\sigma}^2)'$ is the $k_m \times 1$ vector of parameter estimates (the index $m$ emphasizes that all these objects depend on the actual model). In words, at each forecast origin and for each model $m \in \{1, \ldots, 5\}$, I evaluate the likelihood function at the estimated parameters and compute the BIC. According to Kass and Raftery (1995) and Hoeting et al. (1999), model selection based on the BIC is a reliable approximation to model selection based on the highest posterior model probability. Granger and Jeon (2004) found the BIC to perform well in a forecast comparison including a large number of US macroeconomic series. In a recent empirical study on point forecasts, Gürkaynak et al. (2013) showed that simple, univariate autoregressive models, whose lag length is selected using the BIC, often outperform VAR and DSGE models when forecasting output growth at short horizons and inflation at long horizons.[17]

Kass and Raftery (1995) and Hoeting et al. (1999) demonstrated that the Bayesian Model Averaging approach can be approximated by combining the BIC values, where model $m$'s weight is given by

$$w_m = \frac{\exp(-0.5BIC_m)}{\sum_{i=1}^{5} \exp(-0.5BIC_i)}\,. \tag{23}$$

Rossi and Sekhposyan (2014) reported that in a density forecasting framework, BMA (BMA-OLS in their terminology) delivered mixed results when forecasting US GDP growth and inflation. More precisely, equal weights dominated BMA when forecasting output growth one quarter ahead or predicting inflation one and four quarters ahead. However, they both delivered well-calibrated predictive densities for GDP growth four quarters ahead.

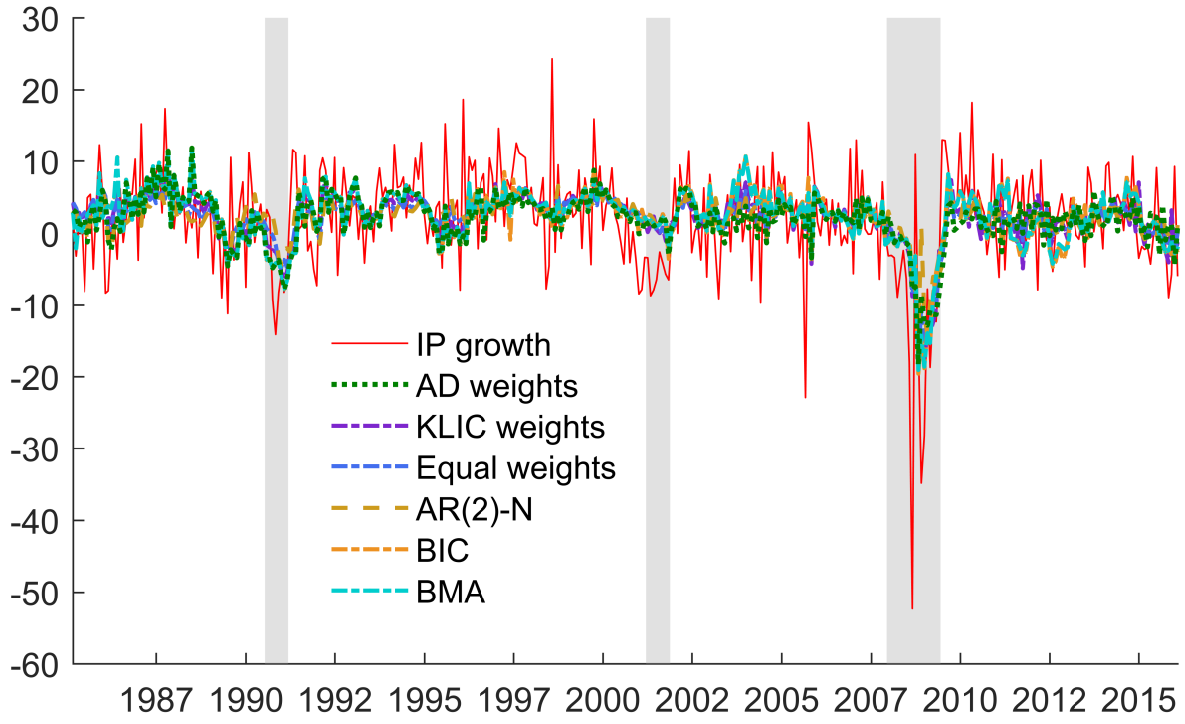
## 5.2 Results: point forecasts

Figure 15 shows the point forecasts (conditional means) of all the forecast combination schemes between March 1985 and February 2016.

We can see that while all models seem to capture the "slow-moving" component of the conditional mean of IP growth, high-frequency movements in the data remain largely unexplained. A formal comparison of Mean Squared Forecast Errors (MSFEs) can be found in Table 6, using the Diebold-Mariano test (Diebold and Mariano, 1995)

[17] For theoretical and simulation results demonstrating the virtues of the BIC in a time series forecasting framework, I refer to Inoue and Kilian (2006) and the studies cited therein.

Figure 15: Point forecasts of US industrial production growth




*Note*: Shaded areas are NBER recession periods.

and following the methodology of Giacomini and White (2006). Specifically, the null hypothesis is that the conditional forecasting performance of each alternative model (Anderson–Darling weights, KLIC weights, equal weights, BIC and BMA) measured by their respective squared forecast error is the same as the benchmark AR(2)-N model, while the alternative hypothesis is that the given alternative model has lower expected squared forecast error. Therefore the MSFE loss difference series were calculated as the squared forecast errors of the AR(2)-N model minus the given competitor's squared forecast errors. The critical values were obtained using the standard normal approximation of the distribution of the test statistic under the null, with rejection region in the right tail. This setting corresponds to the view that it is interesting to investigate whether model combinations deliver significantly superior point forecasting performance compared to the simplest benchmark.

As Table 6 shows, the KLIC weights combination significantly outperforms the benchmark AR(2)-N model at the usual significance levels, while the equal weights scheme delivers a $p$-value of 0.09. This is somewhat surprising, as the superior point forecasting performance of the equal weights model combination has been demonstrated in the literature in a variety of settings, see for example Granger and Jeon (2004), Timmermann (2006) or Elliott and Timmermann (2016). While the Anderson–Darling weight combination scheme fails to deliver significantly better point forecasts than the benchmark, it is remarkable that it performs on par with such a tough benchmark. Recall that the PIT-based weighting scheme is designed to deliver probabilistically calibrated

Table 6: Mean Squared Forecast Errors and Diebold–Mariano tests

| Model | MSFE | DM statistic | $p$-value |
|---|---|---|---|
| AR(2)-N | 3.64 | — | — |
| AD weights | 1.00 | −0.10 | 0.54 |
| KLIC weights | 0.93 | 2.86 | 0.00 |
| Equal weights | 0.96 | 1.36 | 0.09 |
| BIC | 0.97 | 0.75 | 0.23 |
| BMA | 0.96 | 1.17 | 0.12 |

*Note*: The rows correspond to the six forecasting methods, while the columns correspond to the Mean Squared Forecast Error (actual, non-annualized value in the first row, MSFE ratios as fractions of the AR(2)-N benchmark in the remaining rows), the Diebold–Mariano test statistic and its $p$-value. The DM statistic was calculated using the HAC estimator by Newey and West (1987), using a bandwidth of $\lfloor 0.75P^{1/3} \rfloor = 5$.

*density* forecasts. Whether it lives up to this expectation is investigated in the next section.

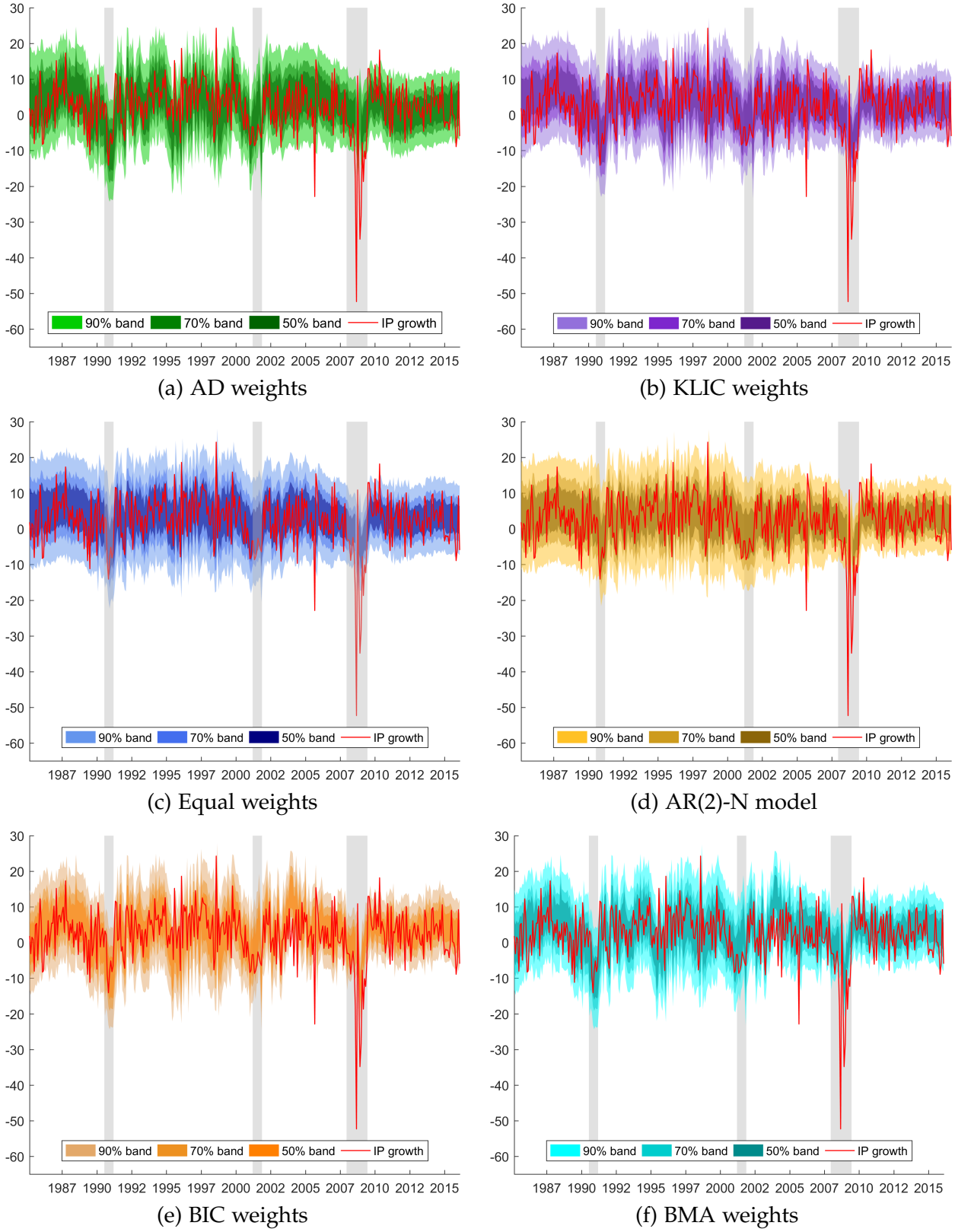
## 5.3 Results: density forecasts

Next, let us consider the density forecasts obtained by the six competing methods. First, in Figure 16 we can see central, equal tailed 90%, 70% and 50% bands of the one-step-ahead combined predictive densities at each forecast target date, ranging from March 1985 to February 2016. Visual inspection suggests that it is not easy to discriminate between the density forecasting schemes. On average, they seem to perform similarly, and not surprisingly they all miss the lowest point of the Great Recession, when in September 2008, US industrial production decreased by 4.36% compared to the previous month (the annualized figure is a striking 52.3%).

In Figure 17 we can see the histograms of the PITs associated with the six forecasting methods. By comparing Figure 17a and Figure 17b, we can see that the Anderson–Darling weight combination slightly misses periods of low growth or even contractions and puts somewhat more mass in the central part of the density than ideal, while the KLIC-based combination fails to capture extreme events in both tails. As Figure 17c and Figure 17d show, the equal weights scheme and the AR(2)-N model display this behavior in a more pronounced way. Figure 17e and Figure 17f suggest that BIC-based model selection and BMA weights provide better density forecasts than the previous two competitors.

Figure 18 shows the empirical CDFs of the PITs and the ideal, uniform CDF corresponding to the 45 degree line. As we can see, Figure 18 confirms the earlier assertions, as the empirical CDF of the AR(2)-N model and the equal weights combination are below the 45 degree line until approximately 0.5 and then run well above the diagonal. On the other hand, the Anderson–Darling and KLIC weights deliver more uniformly distributed

Figure 16: Equal-tailed forecast bands of one-month-ahead US IP growth




(a) AD weights

(b) KLIC weights

(c) Equal weights

(d) AR(2)-N model

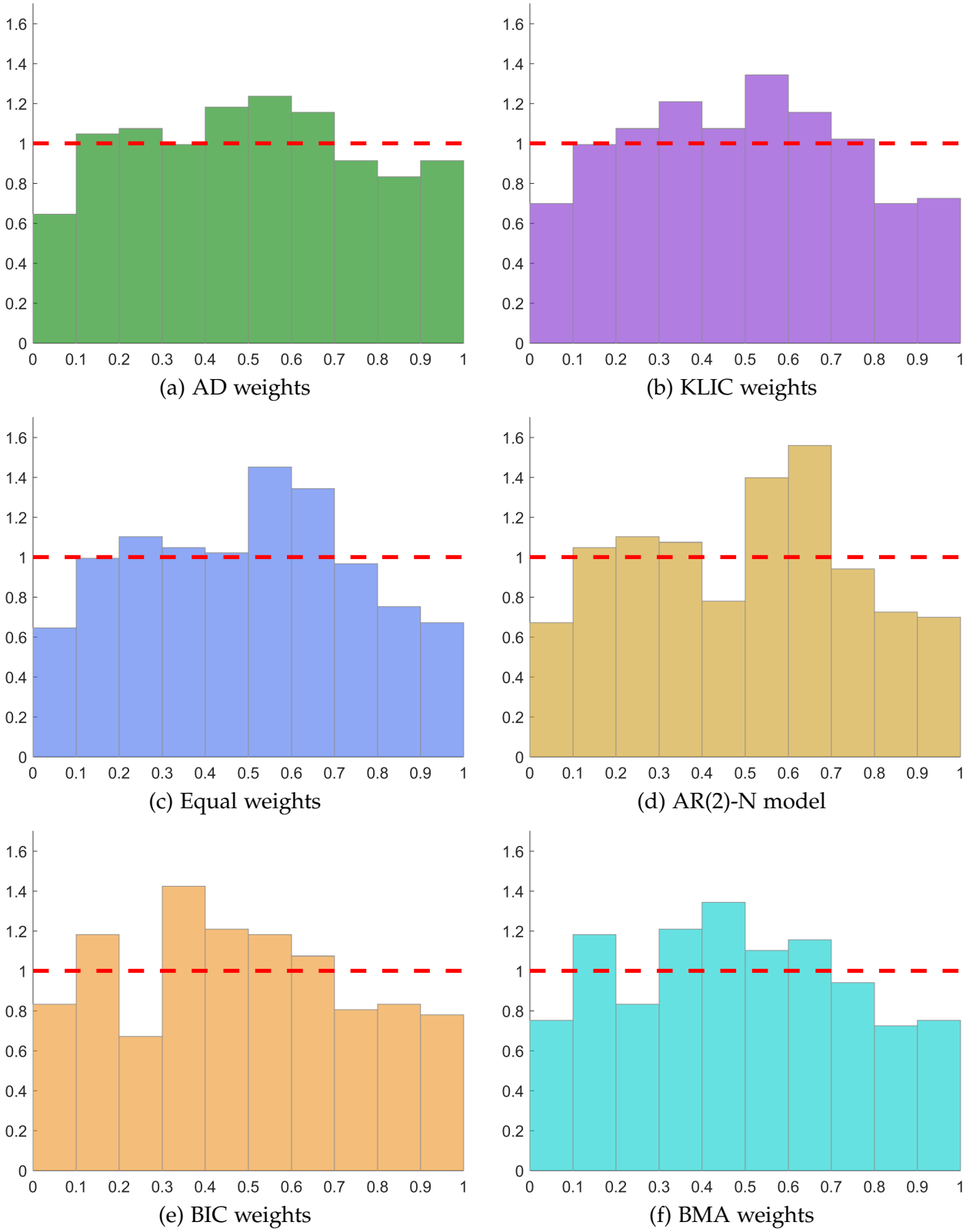
(e) BIC weights

(f) BMA weights

*Note*: Shaded areas are NBER recession periods.

PITs. It is also clear that the empirical CDF of the AD weighting scheme runs closest to the uniform CDF, and the BIC slightly outperforms BMA weights.

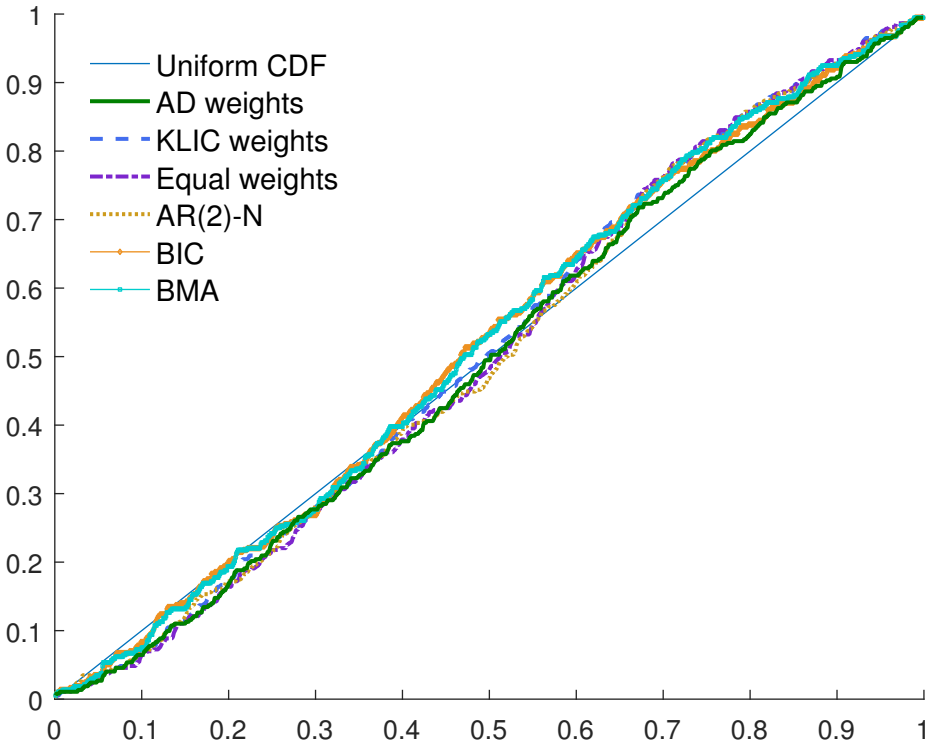
Figure 17: Normalized histograms of PITs




*Note*: Horizontal dashed (red) line corresponds to uniform density.

To formally evaluate whether each density forecasting scheme delivers probabilistically calibrated forecasts, I test the uniformity of the PITs using the test developed by Rossi and Sekhposyan (2016). Under the null hypothesis of uniformity, their test

Figure 18: Empirical CDF of PITs




allows for dynamic misspecification and maintains parameter estimation uncertainty, in line with this paper's framework, as the proposed optimal weighting scheme allows for both as well. Table 7 shows the results of the test of correct specification of each density combination method. As we can see, the Anderson–Darling weights, the BIC, and BMA deliver probabilistically calibrated forecasts of industrial production according to the Kolmogorov–Smirnov and the Cramer–von Mises-type test statistics, by not being able to reject the null even at the 10% level. Furthermore, the KLIC and the AR(2)-N also generate calibrated forecasts at the 5% level. It is reassuring that the proposed optimal weighting scheme is able to produce probabilistically calibrated forecasts in a setting where equal weighting surprisingly fails. Therefore we can conclude that the Anderson–Darling-based estimator, and to a lesser extent, the KLIC-based estimator are capable of delivering well-calibrated density forecasts.

This discussion has so far focused on evaluating the various density forecasts of US industrial production. However, it is also interesting how the combination weights of each model evolved over the out-of-sample period (March 1985 to February 2016), which is shown in Figures 19 to 21.

In Figure 19a, we can see that using the Anderson–Darling weights, apart from the beginning of the sample period, until the early 2000s, the model with the New Orders Index dominated the model pool. From the early 2000s, new housing permits proved to be by far the best predictor of industrial production, which highlights the importance of the housing sector as one of the drivers of the bubble leading to the financial crisis. During and after the Great Recession, the models featuring the corporate bond yield spread and

Table 7: Rossi and Sekhposyan (2016) test on correct specification of conditional predictive densities

| Models | Kolmogorov–Smirnov | Cramer–von Mises |
|---|---|---|
| AD weights | 0.90 (0.38) | 0.24 (0.22) |
| KLIC weights | 1.28 (0.08) | 0.42 (0.06) |
| Equal weights | 1.39 (0.05) | 0.50 (0.04) |
| AR(2)-N | 1.31 (0.08) | 0.40 (0.09) |
| BIC | 1.16 (0.17) | 0.32 (0.16) |
| BMA | 1.28 (0.10) | 0.38 (0.11) |

*Note*: The rows correspond to the six forecasting methods, while the columns correspond to the two test statistics. In each cell, the first entry is the test statistic, the second one, in parentheses is the p-value. The $p$-values were calculated using the HAC estimator by Newey and West (1987) using a bandwidth of $\lfloor 0.75P^{1/3} \rfloor = 5$. The number of Monte Carlo simulations to obtain asymptotic critical values was 200,000.

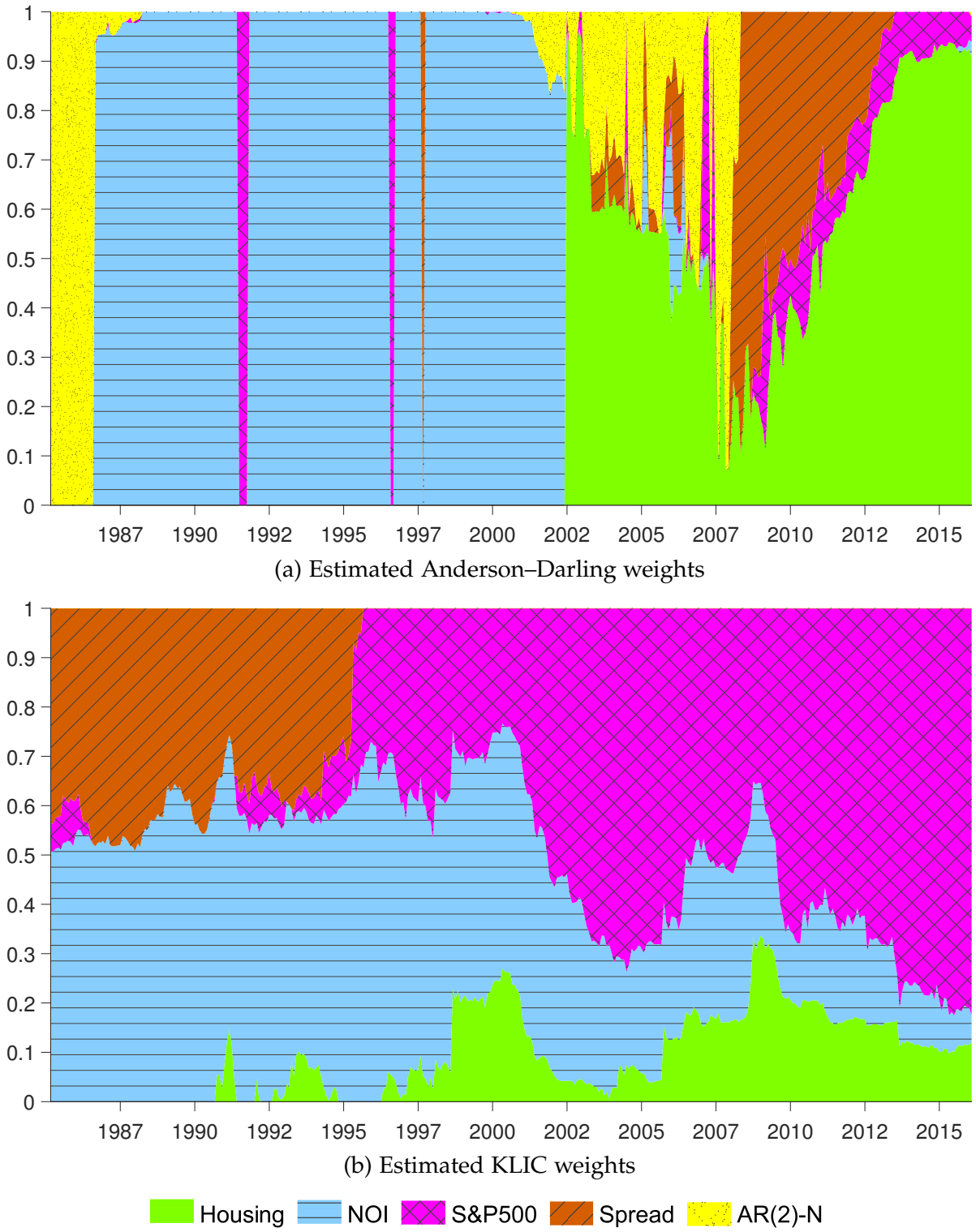
the S&P 500 received large weight. It is remarkable that the optimal combination scheme using Anderson–Darling weights was able to capture the predictive power of the spread variable at the beginning of the financial crisis, as highlighted in the "Spread" panel of Figure 21. These findings are similar to the conclusions of Ng and Wright (2013), who suggest that the predictive content of individual variables displays rather large variations over time and financial data proved to be useful predictors of output in the wake of the Great Recession. As they explain, in a more leveraged economy, interest rate spreads have stronger effect on output through channels affecting firms' finances. However, to my knowledge, the present paper is the first showing in an out-of-sample forecasting exercise that during and after the Great Recession, density forecasts of models that feature a spread variable also perform better in predicting industrial production. Interestingly, since around 2009, housing permits have again emerged as a powerful predictor.

Figure 19b shows that the weights based on the KLIC do not show such pronounced patterns as the AD weights, although we can see that new housing permits appear to contain predictive power sporadically, and spread data received considerable weight only until 1995. KLIC weights also suggest that the New Orders Index has gradually lost its predictive power. However, this weighting scheme increasingly favors the S&P 500 index since 1995, which is in contrast to the earlier results using Anderson–Darling weights.[18]

An explanation of this difference is that at each forecast origin, the individual models' Anderson–Darling statistics displayed more dispersion than their KLIC values, and the PIT-based estimator was able to exploit this variation across models. For a more detailed

[18] Figure E.1 in Appendix E displays the ratio of the inverse in-sample residual variances of each model relative to the sum of the inverse residual variances. Bates and Granger (1969) recommended this ratio as an estimator of the optimal weights, minimizing the expected Root Mean Squared Forecast Error. The figure displays very stable weights, all around 1/5, corresponding to equal weights. This confirms that the PIT- and KLIC-based weight estimates are not driven by the models' in-sample fit.

Figure 19: Time-variation of estimated AD and KLIC weights, area plots




(a) Estimated Anderson–Darling weights

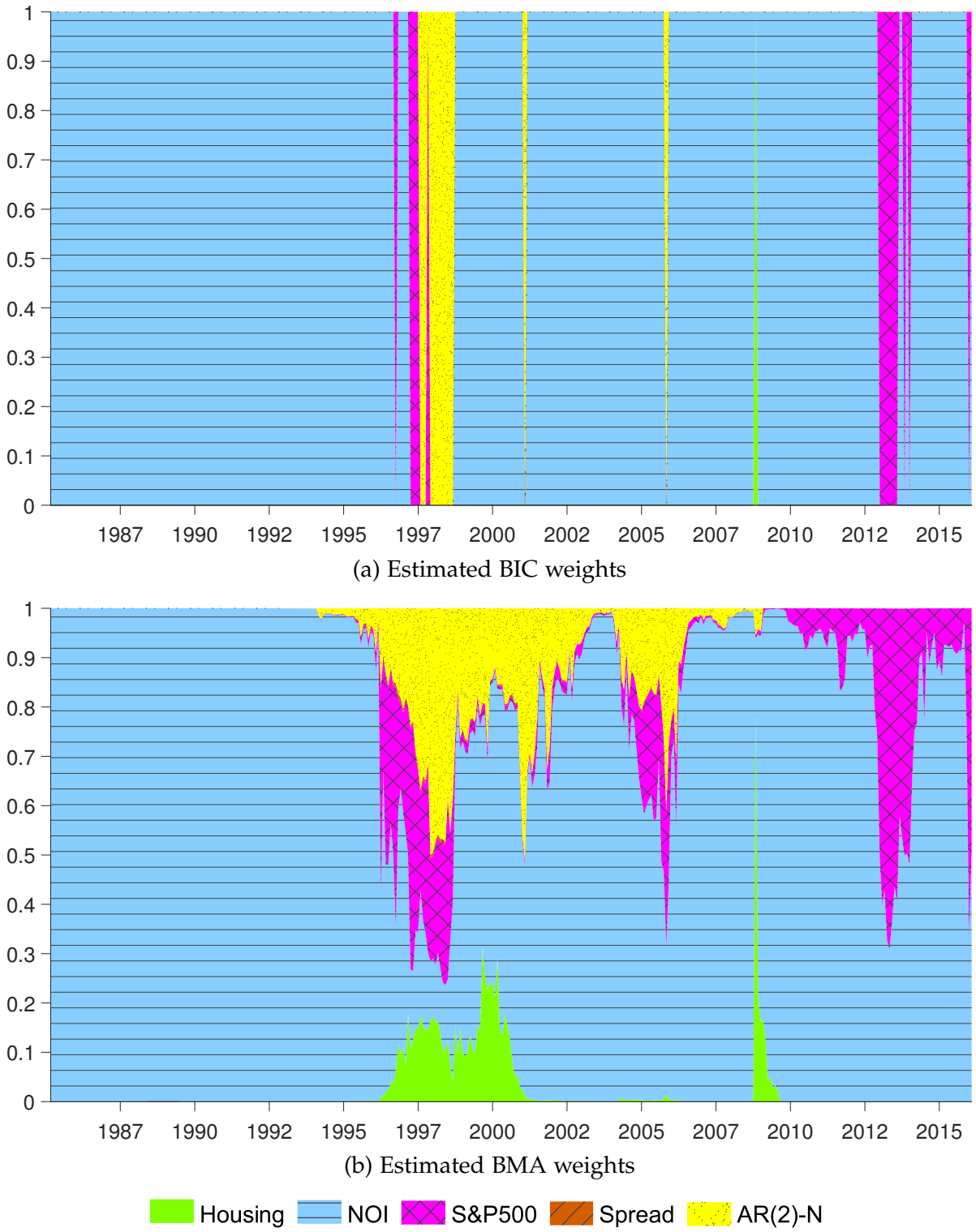
(b) Estimated KLIC weights

*Note*: The out-of-sample period starts in March 1985 and ends in February 2016, with a total number of $P = 372$ months. Housing stands for Housing Permits, NOI stands for ISM: New Orders Index, S&P 500 is the S&P 500 stock index returns while Spread is Moody's Baa Corporate Bond Yield minus Fed funds rate.

analysis and supporting evidence, see Appendix E.

Figure 20a and Figure 20b show that both the BIC and BMA overwhelmingly favored the model featuring the New Orders Index variable, and other models received some weight only sporadically, without a clear and interpretable pattern.

Figure 20: Time-variation of estimated BIC and BMA weights, area plots




(a) Estimated BIC weights

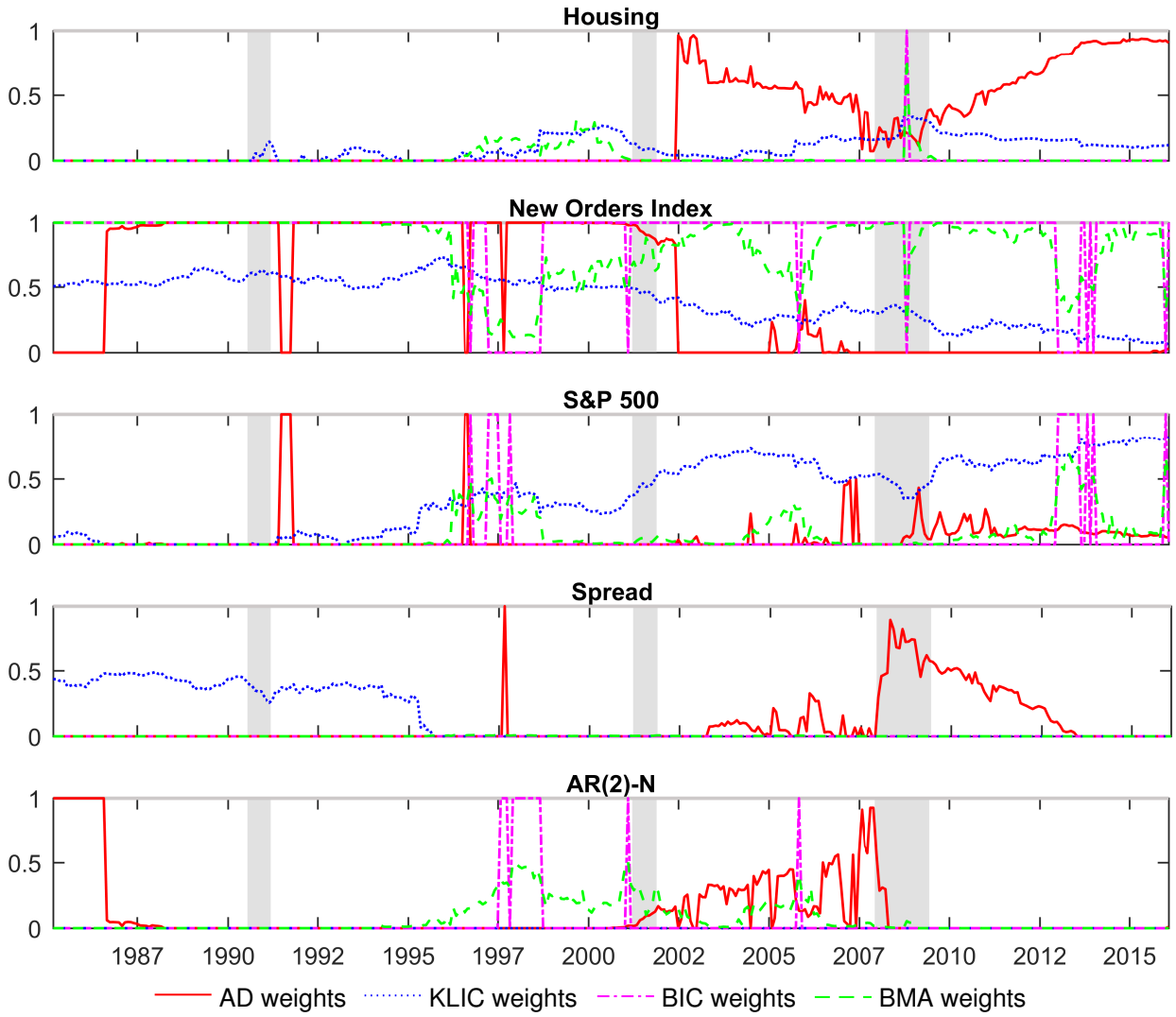
(b) Estimated BMA weights

*Note*: The out-of-sample period starts in March 1985 and ends in February 2016, with a total number of $P = 372$ months. Housing stands for Housing Permits, NOI stands for ISM: New Orders Index, S&P 500 is the S&P 500 stock index returns while Spread is Moody's Baa Corporate Bond Yield minus Fed funds rate.

Figure 21 displays the same information as discussed above, partitioning by forecasting model rather than weight estimation method.

Based on the empirical results, several conclusions arise. First, model combinations can help density forecasting if the weights are carefully estimated, using either

Figure 21: Time-variation of estimated density forecast weights, line plots




*Note*: The plots display the time-variation in the estimated weights for each model (row-wise), using the Anderson–Darling objective function (red solid line), the KLIC objective function (blue dotted line), the BIC (magenta dash-dot line), and BMA (green dashed line). The out-of-sample period starts in March 1985 and ends in February 2016, with a total number of $P = 372$ months. Housing stands for Housing Permits, NOI stands for ISM: New Orders Index, S&P 500 is the S&P 500 stock index returns while Spread is Moody's Baa Corporate Bond Yield minus Fed funds rate. Shaded areas are NBER recession periods.

the Anderson–Darling-type objective function, or to a lesser extent, the KLIC objective function. Second, the variables with most information content change over time and the PIT-based optimal weights provide valuable insights into what was driving industrial production. Specifically, housing permits and financial variables stand out as economically meaningful explanatory variables, the former since the early 2000s and the latter since the recent financial crisis and the recession that followed. Related to the previous points, non-Gaussian density forecasts perform considerably better than Gaussian ones.

# 6 Conclusion

This paper's contributions are summarized as follows. First, I proposed consistent estimators of convex combination weights to approximate the true predictive density. The framework of this study uses a weak notion of forecast calibration that takes into account the information set (the models) that the researcher uses in a given forecasting scenario. Most of the existing literature discusses *testing* whether density forecasts are correctly calibrated, but *estimating* the combination weights has received considerably less attention, which is the topic of the present paper.

Second, Monte Carlo experiments confirmed that the proposed asymptotic theory performs well for sample sizes which are relevant in macroeconometrics and finance.

Third, an empirical exercise demonstrated that this paper's methodology improves on individual models' density forecasts of US industrial production and delivers probabilistically calibrated forecast densities. Furthermore, the estimated weights highlight the importance of non-Gaussian predictive densities, and they are also intuitively interpretable. They demonstrate that the housing market was one of the drivers of output growth before and after the recent financial crisis. Moreover, corporate bond yield spreads contain considerable predictive content, especially during the Great Recession. To my best knowledge, these findings are novel in the literature on density forecasts.

The present paper offers several avenues for further research. The empirical exercise suggests that weight estimates display persistence. Therefore, a potential theoretical extension would be incorporating the information contained in past weights to improve the estimators. Furthermore, the time-variation of the weight estimates implies that structural breaks might be present in the data. Hence, another direction for further study would be to develop a testing procedure to detect breaks. This would allow researchers to make statistically well-founded statements about break dates, which could improve their forecasting strategies. Another possibility is the inclusion of a penalty term to shrink the weights towards zero, focusing on the most relevant models. This would allow forecasters to considerably extend the model set and control the estimators' mean squared error at the same time through a bias-variance trade-off. From an empirical perspective, it would be interesting to see how the proposed weight estimation method compares to recent, Bayesian approaches, suggested by Waggoner and Zha (2012), Billio et al. (2013), and Del Negro et al. (2016). Moreover, this paper's framework is general enough to include structural DSGE models or survey forecasts in the model set. This could enhance our understanding of the relative merits of these approaches in terms of density forecasts. Practitioners in the fields of finance and risk management could also take advantage of the estimators proposed in this paper by constructing more precise Value at Risk estimates using combinations of density forecasts, and focusing on a specific part of the predictive distribution.

# References


Anderson, T. W. and Darling, A. D. (1952). Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23(2):193–212.

Andrews, D. W. K. (1999). Estimation When a Parameter is on a Boundary. *Econometrica*, 67(6):1341–1383.

Bates, J. M. and Granger, C. W. J. (1969). The Combination of Forecasts. *OR*, 20(4):451–468.

Billingsley, P. (1995). *Probability and Measure*. Wiley series in probability and mathematical statistics. John Wiley & Sons, Inc., New York, 3rd edition.

Billio, M., Casarin, R., Ravazzolo, F., and van Dijk, H. K. (2013). Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics*, 177(2):213–232.

Carriero, A., Clark, T. E., and Marcellino, M. (2015). Bayesian VARs: Specification Choices and Forecast Accuracy. *Journal of Applied Econometrics*, 30(1):46–73.

Cheng, X. and Hansen, B. E. (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics*, 186(2):280–293.

Chiu, C.-W. J., Mumtaz, H., and Pinter, G. (2015). Forecasting with VAR models: Fat tails and stochastic volatility. Working Paper No. 528, Bank of England.

Chiu, C.-W. J., Mumtaz, H., and Pinter, G. (2016). VAR Models with Non-Gaussian Shocks. Discussion Paper No. 1609, Centre for Macroeconomics (CFM).

Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762.

Clark, T. E. and Ravazzolo, F. (2015). Macroeconomic Forecasting Performance under Alternative Specifications of Time-Varying Volatility. *Journal of Applied Econometrics*, 30(4):551–575.

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

Corradi, V. and Swanson, N. R. (2006b). Chapter 5 Predictive Density Evaluation. In Elliott, G., Granger, C. W. J., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, pages 197–284. Elsevier.

Corradi, V. and Swanson, N. R. (2006c). Predictive density and conditional confidence interval accuracy tests. *Journal of Econometrics*, 135(1-2):187–228.

Cúrdia, V., del Negro, M., and Greenwald, D. L. (2014). Rare shocks, great recessions. *Journal of Applied Econometrics*, 29(7):1031–1052.

Del Negro, M., Hasegawa, R. B., and Schorfheide, F. (2016). Dynamic prediction pools: An investigation of financial frictions and forecasting performance. *Journal of Econometrics*, 192(2):391–405.

Del Negro, M. and Schorfheide, F. (2013). DSGE Model-Based Forecasting. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2-A, pages 57 – 140. Elsevier, Amsterdam.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association*, 74(366):427–431.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts. *International Economic Review*, 39(4):863–883.

Diebold, F. X. and Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13(3):253.

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

Elder, R., Kapetanios, G., Taylor, T., and Yates, T. (2005). Assessing the MPC's fan charts. *Bank of England Quarterly Bulletin*, (Autumn 2005):326–348.

Elliott, G. and Timmermann, A. (2016). *Economic Forecasting*. Princeton University Press, Princeton, New Jersey, first edition.

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, 50(4):987–1007.

European Central Bank (2014). Fifteen years of the ECB Survey of Professional Forecasters. *European Central Bank Monthly Bulletin*, (January 2014):55–67.

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

Giacomini, R. and White, H. (2006). Tests of Conditional Predictive Ability. *Econometrica*, 74(6):1545–1578.

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

Granger, C. and Jeon, Y. (2004). Forecasting Performance of Information Criteria with Many Macro Series. *Journal of Applied Statistics*, 31(10):1227–1240.

Greenspan, A. (2004). Risk and uncertainty in monetary policy. *American Economic Review*, 94(2):33–40.

Gürkaynak, R. S., Kisacikoglu, B., and Rossi, B. (2013). Do DSGE Models Forecast More Accurately Out-of-Sample than VAR Models? volume 32: VAR Models in Macroeconomics - New Developments and Applications: Essays in Honor of Christopher A. Sims of *Advances in Econometrics*, pages 27–79. Emerald Group Publishing Limited.

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

Hoeting, J. A., Madigan, D. A., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14(4):382–417.

Inoue, A. and Kilian, L. (2006). On the selection of forecasting models. *Journal of Econometrics*, 130(2):273–306.

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

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795.

Marcellino, M., Stock, J. H., and Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135(1-2):499–526.

McCracken, M. W. and Ng, S. (2016). FRED-MD: A Monthly Database for Macroeconomic Research. *Journal of Business & Economic Statistics*, 34(4):574–589.

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

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In McFadden, D. and Engle, R., editors, *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier, Amsterdam.

Newey, W. K. and West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3):703.

Ng, S. and Wright, J. H. (2013). Facts and Challenges from the Great Recession for Forecasting and Macroeconomic Modeling. *Journal of Economic Literature*, 51(4):1120–1154.

Pauwels, L. L. and Vasnev, A. L. (2016). A note on the estimation of optimal weights for density forecast combinations. *International Journal of Forecasting*, 32(2):391–397.

Rosenblatt, M. (1952). Remarks on a Multivariate Transformation. *Ann. Math. Statist.*, 23(3):470–472.

Rossi, B. and Sekhposyan, T. (2013). Conditional predictive density evaluation in the presence of instabilities. *Journal of Econometrics*, 177(2):199–212.

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

Rossi, B. and Sekhposyan, T. (2016). Alternative Tests for Correct Specification of Conditional Predictive Densities. Working Paper No. 758, Barcelona GSE.

Rossi, P. E. (2014). *Bayesian Non- and Semi-Parametric Methods and Applications*. Princeton University Press, Princeton, New Jersey.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

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

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

Tauchen, G. (1985). Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics*, 30(1):415–443.

Timmermann, A. (2006). Chapter 4 Forecast Combinations. In Elliott, G., Granger, C. W. J., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, pages 135–196. Elsevier.

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

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50(1):1.

White, H. (1994). *Estimation, Inference and Specification Analysis*. Number 22 in Econometric Society Monographs. Cambridge University Press, Cambridge.

White, H. (2001). *Asymptotic Theory for Econometricians*. Economic theory, econometrics, and mathematical economics. Academic Press, New York, revised edition.

# Appendices

## A Proofs

*Proof of Theorem 1.* In the first part of the proof, I show almost sure uniform convergence of the sample average of $\xi_{t+h}(w,r)$ to its expected value, following Lemma 1 presented in Tauchen (1985). In the second part, I tailor the remainder of the proof by considering the objective functions $K_G(w)$, $C_G(w)$ and $A_G(w)$ separately. To save on notation and avoid clutter, the time index of the variable of interest runs from 1 to $G$ in the proof. The extension to the general rolling window case is straightforward by replacing the time indices by $t = f - G - h + 1, \ldots, f - h$ where $f = G + R + h - 1, \ldots, T$.

Let us fix $\varepsilon > 0$ for a given $(w,r)$. As $|\xi_{t+h}(w,r)| \leq 1$, it follows that $\lambda_{t+h}(w,r) \equiv E[\xi_{t+h}(w,r)]$ is finite. Note that $\Delta^{\mathcal{M}-1}$ is compact with the Euclidean metric $d_E^{\Delta^{\mathcal{M}-1}}$ on $\mathbb{R}^{\mathcal{M}}$ for example, and so is $\rho \subset [0,1]$, again with the Euclidean metric $d_E^{\rho}$ on $\mathbb{R}$, for instance (the latter is ensured by Assumption 2). Therefore, it follows that the Cartesian product of these sets, $\Delta^{\mathcal{M}-1} \times \rho$ is also compact with the metric $d_C \equiv \max(d_E^{\Delta^{\mathcal{M}-1}}, d_E^{\rho})$ on $\mathbb{R}^{\mathcal{M}+1}$, for example. By definition, $\xi_{t+h}(\cdot,\cdot)$ is almost surely continuous at $(w,r)$, discontinuity occurring when $\Phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}) = r$, which happens only on a set of probability zero by Assumption 3. Therefore, by the dominated convergence theorem, we have that $\lambda_{t+h}(w,r)$ is continuous at $(w,r)$, for all $(w,r)$. Next, let us define

$$u_{t+h}(w,r,d) \equiv \sup_{d_C((\tilde{w},\tilde{r}),(w,r)) \leq d} |\xi_{t+h}(\tilde{w},\tilde{r}) - \xi_{t+h}(w,r)|. \tag{A.1}$$

Recall that $\xi_{t+h}(w,r)$ is almost surely continuous at $(w,r)$, where the null set depends on $(w,r)$, by Assumption 3. Note that $u_{t+h}(w,r,d)$ is measurable, as the separability of $\xi_{t+h}(w,r)$ can be shown along the lines of Section 38 of Billingsley (1995) and therefore we can equivalently take the supremum over $(\tilde{w},\tilde{r}) \in \Delta^{\mathcal{M}-1} \times \rho \cap \mathbb{Q}^{\mathcal{M}+1}$, that is $d_C((\tilde{w},\tilde{r}),(w,r)) \leq d$, as the rationals constitute a countable, dense subset of $\Delta^{\mathcal{M}-1} \times \rho$. Therefore, $\lim_{d\to 0} u_{t+h}(w,r,d) = 0$, almost surely. Then by the dominated convergence theorem, there exists a $\bar{d}_C(w,r)$ such that if $d \leq \bar{d}_C(w,r)$, then we have that $E[u_t(w,r,d)] \leq \varepsilon$. Let $B((w,r),\bar{d}_C(w,r))$ denote an open ball of $\Delta^{\mathcal{M}-1} \times \rho$ of radius $\bar{d}_C(w,r)$ centered at $(w,r)$. Clearly, $\cup_{(w,r)\in\Delta^{\mathcal{M}-1}\times\rho} B((w,r),\bar{d}_C(w,r))$ cover $\Delta^{\mathcal{M}-1} \times \rho$ and by the compactness of $\Delta^{\mathcal{M}-1} \times \rho$, there is a finite cover such that $\Delta^{\mathcal{M}-1} \times \rho \subset \cup_{k=1}^{K} B((w_k,r_k),\bar{d}_C(w_k,r_k))$. For notational convenience, let us define $\mu_{t+h,k} \equiv E[u_{t+h}(w_k,r_k,\bar{d}_C(w_k,r_k))]$. Note that if $(w,r) \in B((w_k,r_k),\bar{d}_C(w_k,r_k))$, then $\mu_{t+j,k} \leq \varepsilon$ and $|\lambda_{t+h}(w,r) - \lambda_{t+h}(w_k,r_k)| \leq \varepsilon$. Let $(w,r) \in B((w_k,r_k),\bar{d}_C(w_k,r_k))$ and consider

$$\left| \frac{1}{G} \sum_{t=1}^{G} \xi_{t+h}(w,r) - \frac{1}{G} \sum_{t=1}^{G} \lambda_{t+h}(w,r) \right| \tag{A.2}$$

$$
\begin{aligned}
&\leq \left| \frac{1}{G}\sum_{t=1}^{G} \xi_{t+h}(w,r) - \frac{1}{G}\sum_{t=1}^{G} \xi_{t+h}(w_k,r_k) \right| + \\
&\quad \left| \frac{1}{G}\sum_{t=1}^{G} \xi_{t+h}(w_k,r_k) - \frac{1}{G}\sum_{t=1}^{G} \lambda_{t+h}(w_k,r_k) \right| + \qquad \text{(A.3)} \\
&\quad \left| \frac{1}{G}\sum_{t=1}^{G} \lambda_{t+h}(w_k,r_k) - \frac{1}{G}\sum_{t=1}^{G} \lambda_{t+h}(w,r) \right| \\
&\leq \frac{1}{G}\sum_{t=1}^{G} |\xi_{t+h}(w,r) - \xi_{t+h}(w_k,r_k)| + \frac{1}{G}\sum_{t=1}^{G} |\xi_{t+h}(w_k,r_k) - \lambda_{t+h}(w_k,r_k)| + \\
&\quad \frac{1}{G}\sum_{t=1}^{G} |\lambda_{t+h}(w_k,r_k) - \lambda_{t+h}(w,r)| \qquad \text{(A.4)} \\
&\leq \left[ \frac{1}{G}\sum_{t=1}^{G} u_{t+h}(w_k,r_k,\bar{d}_C(w_k,r_k)) - \mu_{t+h,k} \right] + \\
&\quad \frac{1}{G}\sum_{t=1}^{G} \mu_{t+h,k} + \frac{1}{G}\sum_{t=1}^{G} |\xi_{t+h}(w_k,r_k) - \lambda_{t+h}(w_k,r_k)| + \qquad \text{(A.5)} \\
&\quad \frac{1}{G}\sum_{t=1}^{G} |\lambda_{t+h}(w_k,r_k) - \lambda_{t+h}(w,r)| \ ,
\end{aligned}
$$

where Equation (A.3) follows from adding and subtracting the four terms in the middle and then I took absolute values by pairs. In Equation (A.4), I used the triangle inequality. In Equation (A.5) I used Equation (A.1) and added and subtracted $G^{-1}\sum_{t=1}^{G}\mu_{t+h,k}$. Note that by Assumption 4, $R$ is finite, therefore $\xi_{t+h}(w,r)$ is mixing of the same size as $Z_t$ by Theorem 3.49 of White (2001), thus we can apply a strong law of large numbers (Corollary 3.48 of White (2001)) on the first and the third terms of the above expression. That is, there is a $G_k(\varepsilon)$ such that if $G > G_k(\varepsilon)$, then these terms are less than or equal to $\varepsilon$ almost surely, thus the whole expression is less than or equal to $4\varepsilon$ almost surely (the second and the fourth terms each are less than or equal to $\varepsilon$ by construction).[19] Furthermore, if $G > \max_{k=1,\ldots,K} G_k(\varepsilon)$, then we have

$$
\sup_{(w,r)\in\Delta^{\mathcal{M}-1}\times\rho} \left| \frac{1}{G}\sum_{t=1}^{G} \xi_{t+h}(w,r) - \frac{1}{G}\sum_{t=1}^{G} \lambda_{t+h}(w,r) \right| \leq 4\varepsilon \qquad \text{(A.6)}
$$

almost surely, therefore as $G \to \infty$, we have

$$
\sup_{(w,r)\in\Delta^{\mathcal{M}-1}\times\rho} \left| \frac{1}{G}\sum_{t=1}^{G} \xi_{t+h}(w,r) - \frac{1}{G}\sum_{t=1}^{G} \lambda_{t+h}(w,r) \right| \xrightarrow{a.s.} 0 \, . \qquad \text{(A.7)}
$$

Let us define $\Psi_0(w,r) \equiv G^{-1}\sum_{t=1}^{G}\lambda_{t+h}(w,r)$, which is the population counterpart of

[19] Note that no additional moment assumption concerning $\xi_{t+h}(w,r)$ is necessary, as $|\xi_{t+h}(w,r)| \leq 1$, thus the moment condition of the cited law of large numbers is satisfied.

$\Psi_G(w,r) \equiv G^{-1}\sum_{t=1}^{G}\xi_{t+h}(w,r)$. Therefore, we have that:

$$\sup_{(w,r)\in\Delta^{\mathcal{M}-1}\times\rho}|\Psi_G(w,r)-\Psi_0(w,r)|\xrightarrow{a.s.}0\,. \tag{A.8}$$

Next, we tailor the remainder of the proof considering each objective function separately.

▶ **Case 1:** Kolmogorov–Smirnov objective function $K_G(w)$. I want to show that

$$\sup_{w\in\Delta^{\mathcal{M}-1}}\left|\sup_{r\in\rho}|\Psi_G(w,r)|-\sup_{r\in\rho}|\Psi_0(w,r)|\right|\xrightarrow{a.s.}0\,. \tag{A.9}$$

Consider the following inequalities:

$$\begin{aligned}
&\sup_{w\in\Delta^{\mathcal{M}-1}}\left|\sup_{r\in\rho}|\Psi_G(w,r)|-\sup_{r\in\rho}|\Psi_0(w,r)|\right| \\
&\leq \sup_{w\in\Delta^{\mathcal{M}-1}}\sup_{r\in\rho}\big||\Psi_G(w,r)|-|\Psi_0(w,r)|\big| \\
&\leq \sup_{w\in\Delta^{\mathcal{M}-1}}\sup_{r\in\rho}|\Psi_G(w,r)-\Psi_0(w,r)| \\
&\leq \sup_{(w,r)\in\Delta^{\mathcal{M}-1}\times\rho}|\Psi_G(w,r)-\Psi_0(w,r)|\,,
\end{aligned}$$

where I applied basic properties of the supremum and the reverse triangle inequality. Therefore we have

$$\sup_{w\in\Delta^{\mathcal{M}-1}}\left|\sup_{r\in\rho}|\Psi_G(w,r)|-\sup_{r\in\rho}|\Psi_0(w,r)|\right|\xrightarrow{a.s.}0\,. \tag{A.10}$$

▶ **Case 2:** Cramer–von Mises objective function $C_G(w)$. I want to show that

$$\sup_{w\in\Delta^{\mathcal{M}-1}}\left|\int_{r\in\rho}\Psi_G^2(w,r)\,\mathrm{d}r-\int_{r\in\rho}\Psi_0^2(w,r)\,\mathrm{d}r\right|\xrightarrow{a.s.}0\,. \tag{A.11}$$

Consider the following inequalities:

$$\begin{aligned}
&\left|\int_{r\in\rho}\Psi_G^2(w,r)\,\mathrm{d}r-\int_{r\in\rho}\Psi_0^2(w,r)\,\mathrm{d}r\right| \\
&=\left|\int_{r\in\rho}\Psi_G^2(w,r)-\Psi_0^2(w,r)\,\mathrm{d}r\right| \\
&\leq\int_{r\in\rho}\left|\Psi_G^2(w,r)-\Psi_0^2(w,r)\right|\,\mathrm{d}r
\end{aligned}$$

$$
\begin{aligned}
&\leq \sup_{r\in\rho} \left|\Psi_G^2(w,r) - \Psi_0^2(w,r)\right| \\
&= \sup_{r\in\rho} \left[|\Psi_G(w,r) - \Psi_0(w,r)| \cdot |\Psi_G(w,r) + \Psi_0(w,r)|\right] \\
&\leq \sup_{r\in\rho} |\Psi_G(w,r) - \Psi_0(w,r)| \cdot 2\,.
\end{aligned}
$$

Therefore, given that $\varepsilon > 0$ was arbitrary, it follows that

$$
\sup_{w\in\Delta^{\mathcal{M}-1}} \left| \int_{r\in\rho} \Psi_G^2(w,r)\,\mathrm{d}r - \int_{r\in\rho} \Psi_0^2(w,r)\,\mathrm{d}r \right| \xrightarrow{a.s.} 0\,. \tag{A.12}
$$

▶ **Case 3:** Anderson–Darling objective function $A_G(w)$. I want to show that

$$
\sup_{w\in\Delta^{\mathcal{M}-1}} \left| \int_{r\in\rho} \frac{\Psi_G^2(w,r)}{r(1-r)}\,\mathrm{d}r - \int_{r\in\rho} \frac{\Psi_0^2(w,r)}{r(1-r)}\,\mathrm{d}r \right| \xrightarrow{a.s.} 0\,. \tag{A.13}
$$

For clarity of exposition, I only discuss the case when $\rho = [0,1]$, given that the proof can be easily tailored to other cases, as it is shown below. Consider the following inequality:

$$
\begin{aligned}
&\left| \int_0^1 \frac{\Psi_G^2(w,r)}{r(1-r)}\,\mathrm{d}r - \int_0^1 \frac{\Psi_0^2(w,r)}{r(1-r)}\,\mathrm{d}r \right| \\
\leq &\left| \int_0^\delta \frac{\Psi_G^2(w,r) - \Psi_0^2(w,r)}{r(1-r)}\,\mathrm{d}r \right| + \left| \int_{1-\delta}^1 \frac{\Psi_G^2(w,r) - \Psi_0^2(w,r)}{r(1-r)}\,\mathrm{d}r \right| \\
+ &\left| \int_\delta^{1-\delta} \frac{\Psi_G^2(w,r) - \Psi_0^2(w,r)}{r(1-r)}\,\mathrm{d}r \right|\,.
\end{aligned}
$$

Next, consider the following inequalities related to the last term in the previous inequality:

$$
\begin{aligned}
&\left| \int_\delta^{1-\delta} \frac{\Psi_G^2(w,r) - \Psi_0^2(w,r)}{r(1-r)}\,\mathrm{d}r \right| \\
=&\left| \int_\delta^{1-\delta} \frac{[\Psi_G(w,r) + \Psi_0(w,r)]\,[\Psi_G(w,r) - \Psi_0(w,r)]}{r(1-r)}\,\mathrm{d}r \right| \\
\leq &\int_\delta^{1-\delta} \frac{|\Psi_G(w,r) + \Psi_0(w,r)|\,|\Psi_G(w,r) - \Psi_0(w,r)|}{r(1-r)}\,\mathrm{d}r \\
\leq &2\int_\delta^{1-\delta} \frac{|\Psi_G(w,r) - \Psi_0(w,r)|}{r(1-r)}\,\mathrm{d}r \\
\leq &2\int_\delta^{1-\delta} \frac{\sup_{r\in[0,1]} |\Psi_G(w,r) - \Psi_0(w,r)|}{r(1-r)}\,\mathrm{d}r
\end{aligned}
$$

$$
\begin{aligned}
&\leq 2 \sup_{r\in[0,1]} |\Psi_G(w,r) - \Psi_0(w,r)| \int_{\delta}^{1-\delta} \frac{1}{r(1-r)}\,\mathrm{d}r \\
&= 2 \sup_{r\in[0,1]} |\Psi_G(w,r) - \Psi_0(w,r)| \left[\log(r) - \log(1-r)\right]_{\delta}^{1-\delta} .
\end{aligned}
$$

Using Assumption 6, we have that

$$
\sup_{w\in\Delta^{\mathcal{M}-1}} \left| \int_{r\in\rho} \frac{\Psi_G^2(w,r)}{r(1-r)}\,\mathrm{d}r - \int_{r\in\rho} \frac{\Psi_0^2(w,r)}{r(1-r)}\,\mathrm{d}r \right| \xrightarrow{a.s.} 0 . \tag{A.14}
$$

The results obtained above, coupled with Assumption 5 allow us to invoke Theorem 2.1 in Newey and McFadden (1994), therefore we conclude that $\widehat{w} \xrightarrow{a.s.} w^*$.

Remark: we can also define our extremum estimator as

$$
\widehat{w} \in \Delta^{\mathcal{M}-1} \text{ s.t. } T_G(\widehat{w}) \leq \inf_{w\in\Delta^{\mathcal{M}-1}} T_G(w) + h , \tag{A.15}
$$

where $h$ is either $o_{\text{a.s.}}(1)$ or $o_p(1)$ which would deliver exactly the same consistency result as above, using the definition in Equation (15), as (Newey and McFadden, 1994, Section 2.1, pp. 2121-2122) noted (clearly, if $h$ is only $o_p(1)$ but not $o_{\text{a.s.}}(1)$, then our estimator would be weakly but not strongly consistent). Informally, the difference lies in the fact that unlike Equation (15), Equation (A.15) allows for an asymptotically vanishing discrepancy between the true minimizer of $T_G(w)$ and the actual estimator that the researcher uses. ■

*Proof of Theorem 2.* The proof is analogous to the first part of the proof of Theorem 1, hence for the sake of brevity I only highlight the differences. First, note that Assumptions 8 and 10 let us separate the terms in Equation (12). Let us define $\zeta_{t+h}(w) \equiv -\log \phi^{C}_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}) 1[y_{t+h} \in \varrho]$ and $\lambda_{t+h}(w) \equiv E_{\phi^*}\zeta_{t+h}(w)$ where the finiteness of $\lambda_{t+h}(w)$ follows from Assumption 10. Then using Assumption 9, we have that $\lambda_{t+h}(w)$ is continuous in $w$ by the dominated convergence theorem. $u_{t+h}(w,d)$ is defined similarly as in Equation (A.1) and its measurability follows from the continuity of $\zeta_{t+h}(w)$. The remainder of the proof follows the same logic as in the first part of the proof of Theorem 1 and is therefore omitted. However, note that in this case we require the moment condition of Assumption 11 to invoke the strong law of large numbers (Corollary 3.48 of White (2001)). Having arrived at

$$
\sup_{w\in\Delta^{\mathcal{M}-1}} \left| \frac{1}{G}\sum_{t=1}^{G} \zeta_{t+h}(w) - \frac{1}{G}\sum_{t=1}^{G} \lambda_{t+h}(w) \right| \xrightarrow{a.s.} 0 , \tag{A.16}
$$

by using Assumption 12, we can invoke Theorem 2.1 in Newey and McFadden (1994), therefore we conclude that $\widehat{w} \xrightarrow{a.s.} w^*$.

The same remark applies as in the proof of Theorem 1. ■

## B Differences between probabilistic and complete calibration

To illustrate the difference between probabilistic and complete calibration, consider the following stylized example, inspired by Corradi and Swanson (2006b,c). For simplicity I abstract from parameter estimation error. Let us assume that the true DGP for $y_{t+1}$ is a stationary normal AR(2) process, given by

$$y_{t+1} = \alpha_1 y_t + \alpha_2 y_{t-1} + \varepsilon_{t+1} \qquad \varepsilon_{t+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0, \sigma^2)\,, \tag{B.1}$$

that is the density of $y_{t+1}$ conditional on $\mathcal{I}_t = \{y_t, y_{t-1}\}$ is

$$\phi^*_{t+1}(y_{t+1}|\mathcal{I}_t) = \mathcal{N}(\alpha_1 y_t + \alpha_2 y_{t-1}, \sigma^2)\,. \tag{B.2}$$

It can be shown either by recursive backward substitution or using the Wold decomposition theorem that the joint distribution of $(y_{t+1}, y_t, y_{t-1})'$ is a multivariate normal, formally

$$(y_{t+1}, y_t, y_{t-1})' \sim \mathcal{N}(\mu, \Sigma)\,, \tag{B.3}$$

where the mean vector $\mu$ is a $3 \times 1$ vector of zeros and the $(i,j)$th element of the covariance matrix $\Sigma$ is given by $\Sigma_{i,j} = \gamma_{|i-j|}$, where $\gamma_{|i-j|}$ is the $|i-j|$th order autocovariance of the process. Furthermore, by properties of the normal distribution, it is true that the distribution of $y_{t+1}$ conditional on $y_t$ alone is also normal, formally

$$\phi^*_{t+1}(y_{t+1}|y_t) = \mathcal{N}(\widetilde{\alpha} y_t, \widetilde{\sigma}^2)\,, \tag{B.4}$$

where $\widetilde{\alpha}$ and $\widetilde{\sigma}^2$ can be found from $\Sigma$, specifically $\widetilde{\alpha} = \gamma_1/\gamma_0$ and $\widetilde{\sigma}^2 = \left(1 - \widetilde{\alpha}^2\right)\gamma_0$.

Suppose that the researcher conditions his or her forecast on only one lag of the dependent variable, $(R = 1, \mathcal{I}^t_{t-R+1} = y_t)$ but still maintains the normality assumption, implying the predictive density

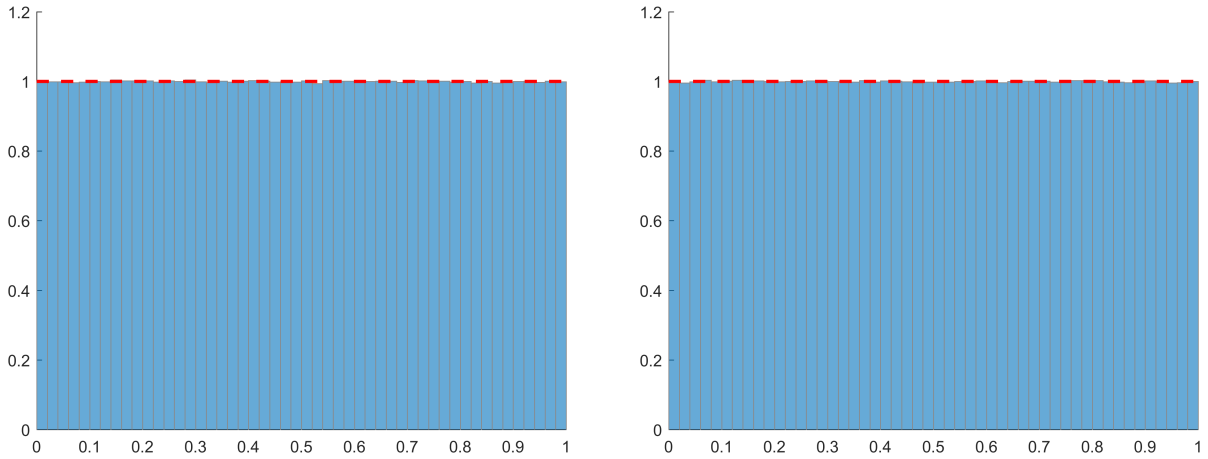
$$\phi_{t+1}(y_{t+1}|\mathcal{I}^t_{t-R+1}) = \mathcal{N}(\widetilde{\alpha} y_t, \widetilde{\sigma}^2)\,. \tag{B.5}$$

In this case, it is easy to see that while this forecast is not completely calibrated, as it misses $y_{t-1}$, it is still probabilistically calibrated, as given the researcher's information set (now consisting of $y_t$), the predictive density is correct, $\phi_{t+1}(y_{t+1}|\mathcal{I}^t_{t-R+1}) = \phi^*_{t+1}(y_{t+1}|\mathcal{I}^t_{t-R+1})$.

I repeated the exercise outlined in Example 2 using the models in Example 1, setting $\alpha_1 = 0.4, \alpha_2 = 0.3, \sigma^2 = 1$. As the histograms in Figure B.1 show, the resulting CDFs of both the correctly specified AR(2) and the dynamically misspecified AR(1) are uniformly distributed. In Figure B.2 we see the CDFs of the PITs of both models, which are

indistinguishable from the 45 degree line, corresponding to the uniform distribution, confirming the earlier theoretical result.

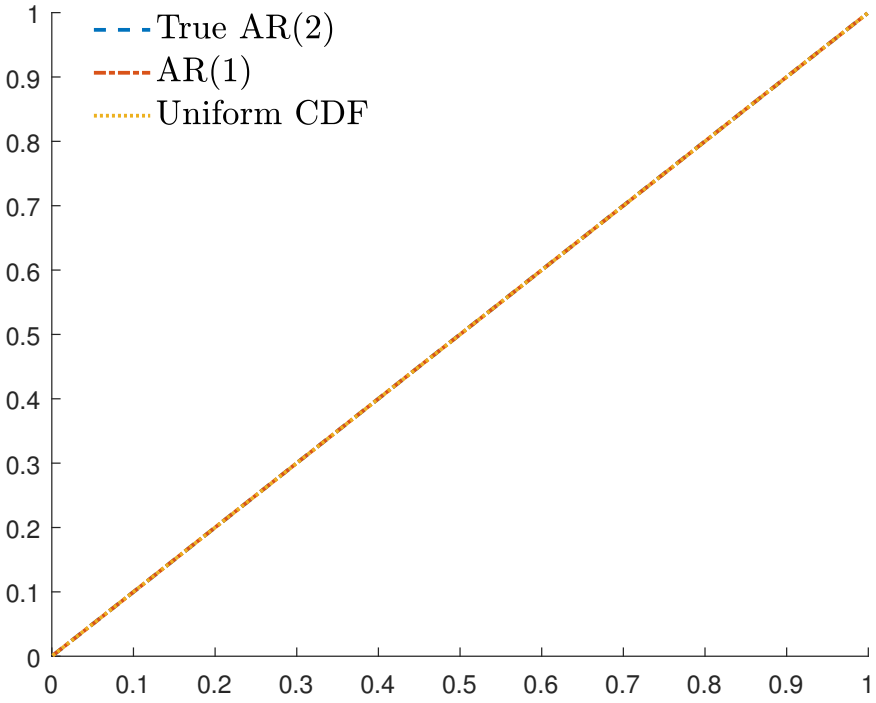
Figure B.1: Normalized histograms of PITs




(a) True AR(2) forecast density (b) AR(1) forecast density

*Note*: Horizontal (red) dashed line corresponds to uniform density.

Figure B.2: Cumulative distribution functions of PITs of candidate densities




## C Optimization algorithm

Given that the non-linear extremum estimators proposed in the present paper do not have closed form solutions, I need to use a numerical optimizer. The optimizer that operates on the unit simplex is MATLAB's built-in `fminsearch` algorithm. This is

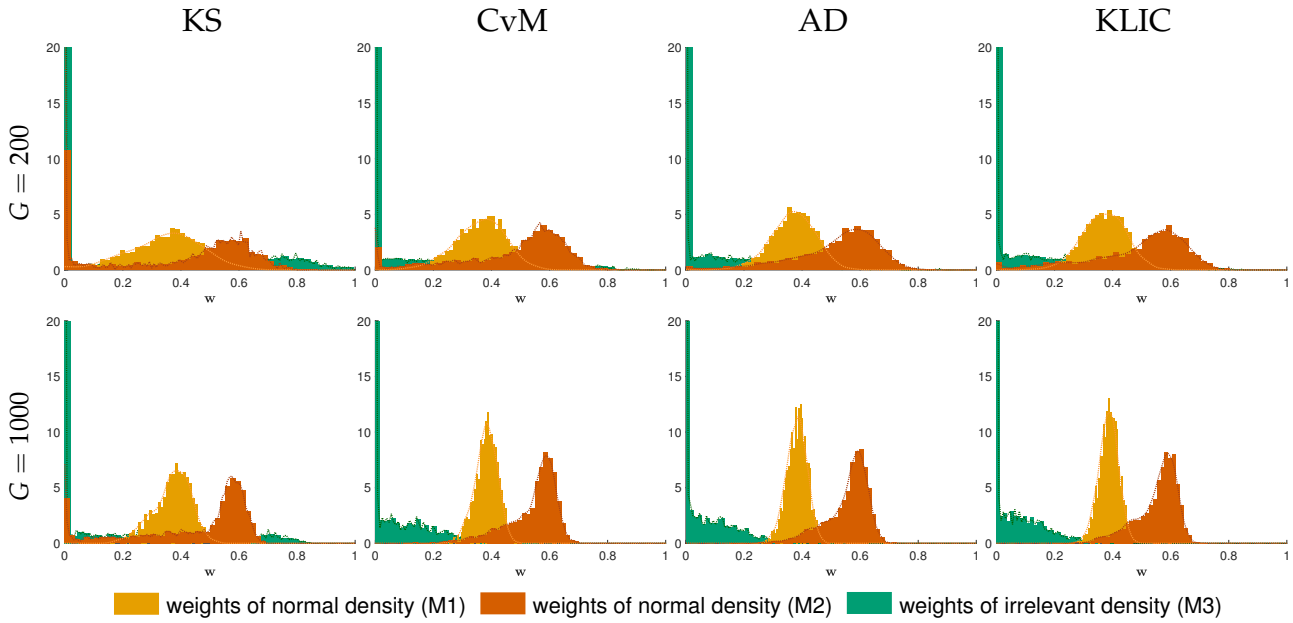
an unconstrained derivative-free optimizer, and I transformed each element of the unconstrained weight vector using the hyperbolic tangent function. The reason why I could not use derivative-based optimizers is that the empirical CDFs are step functions. Also, in practical applications, even with a moderate (5-10) number of models, grid search methods are computationally infeasible for any reasonably fine grid (100-200 points along each dimension). As the `fminsearch` algorithm is not a global optimizer, I used multiple starting points, uniformly distributed on the unit simplex (25 and 50 points in the Monte Carlo simulations and the empirical exercise, respectively) and chose the parameter vector that resulted in the smallest value of the objective function.

# D Monte Carlo – additional figures and DGPs

Figures D.1 to D.4 display the histograms and kernel density estimates for all DGPs and objective functions, for $G = \{200, 1000\}$, which were omitted from Section 4.4 to preserve space. Furthermore, a number of additional DGPs are used to illustrate the estimators' performance.

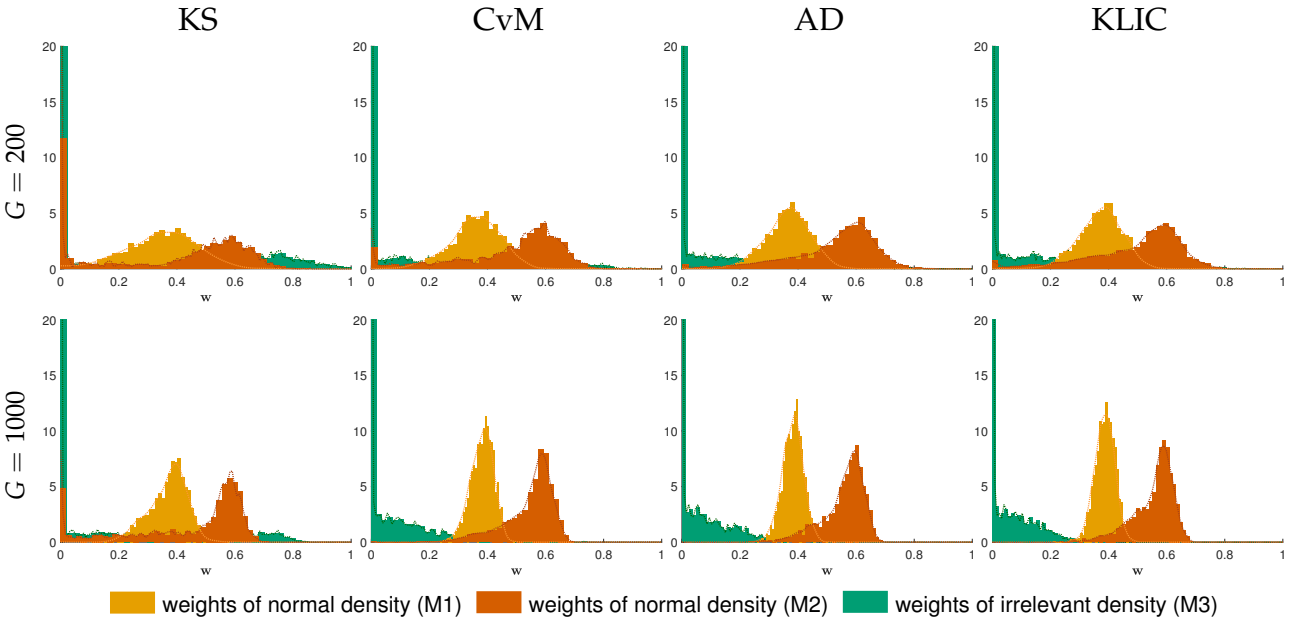
## D.1 Additional figures – DGPs 1a, 1b, 2 and 3

Figure D.1: Additional Monte Carlo results for DGP 1a, true parameter vector $w = (0.4, 0.6, 0)'$



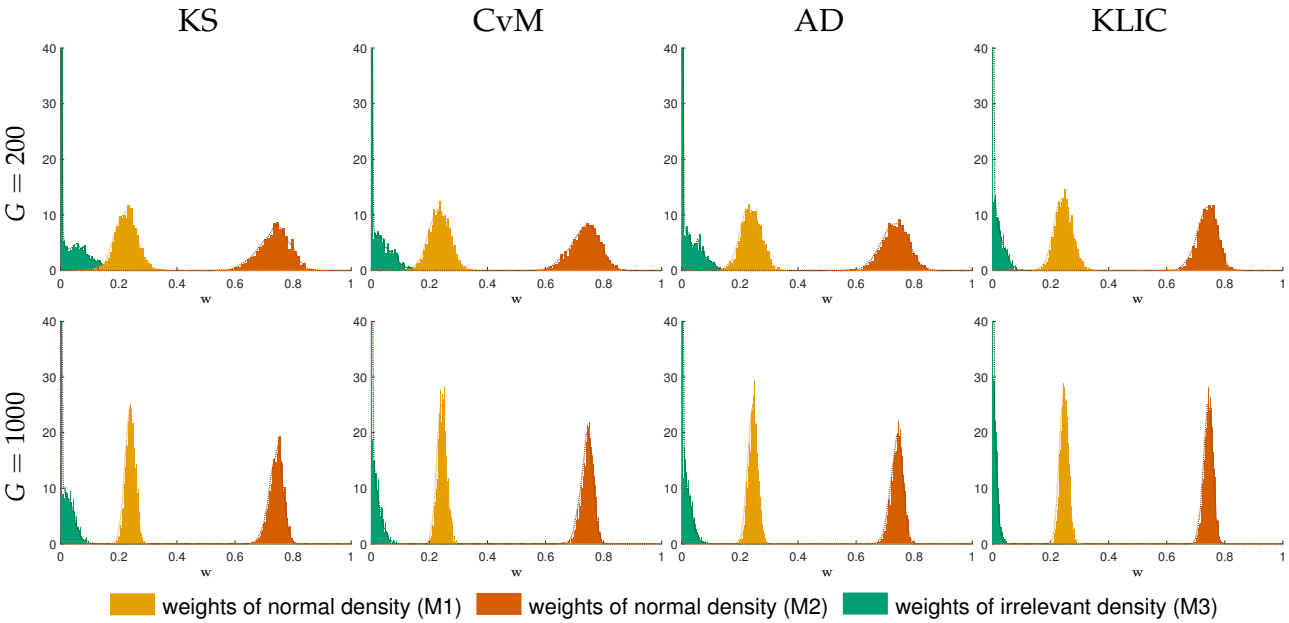

*Note*: $G$ denotes the sample size. KS, CvM, AD and KLIC stand for the Kolmogorov–Smirnov-, the Cramer–von Mises-, the Anderson–Darling- and the KLIC-based estimators, respectively. Histograms and kernel density estimates based on 2000 Monte Carlo replications.

Figure D.2: Additional Monte Carlo results for DGP 1b, true parameter vector $w = (0.4, 0.6, 0)'$



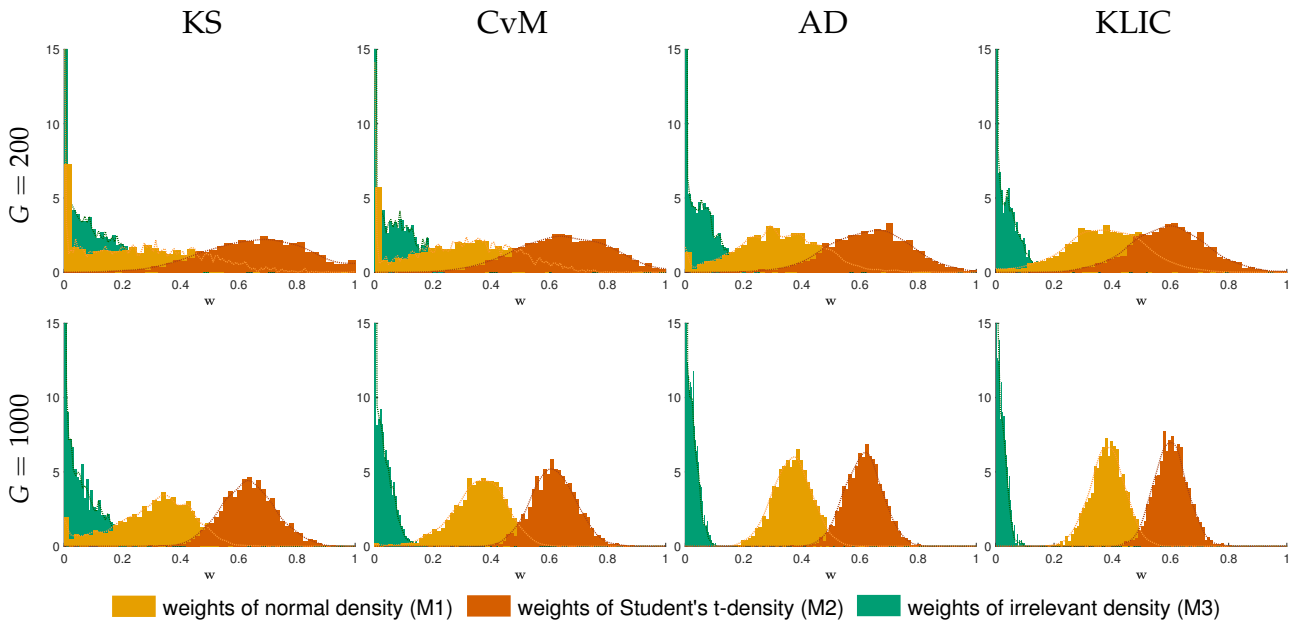

*Note*: $G$ denotes the sample size. KS, CvM, AD and KLIC stand for the Kolmogorov–Smirnov-, the Cramer–von Mises-, the Anderson–Darling- and the KLIC-based estimators, respectively. Histograms and kernel density estimates are based on 2000 Monte Carlo replications.

Figure D.3: Additional Monte Carlo results for DGP 2, true parameter vector $w = (0.25, 0.75, 0)'$




*Note*: $G$ denotes the sample size. KS, CvM, AD and KLIC stand for the Kolmogorov–Smirnov-, the Cramer–von Mises-, the Anderson–Darling- and the KLIC-based estimators, respectively. Histograms and kernel density estimates are based on 2000 Monte Carlo replications.

Figure D.4: Additional Monte Carlo results for DGP 3, true parameter vector $w = (0.4, 0.6, 0)'$

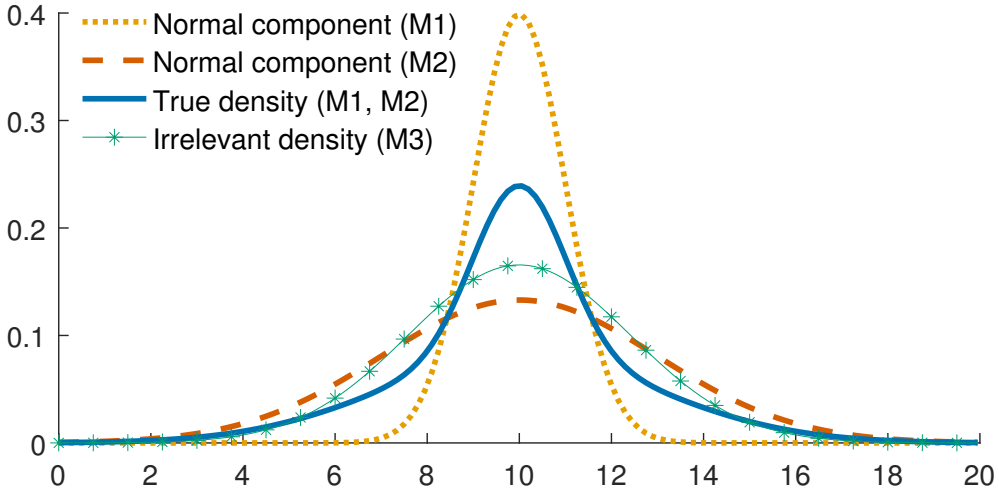



*Note*: $G$ denotes the sample size. KS, CvM, AD and KLIC stand for the Kolmogorov–Smirnov-, the Cramer–von Mises-, the Anderson–Darling- and the KLIC-based estimators, respectively. Histograms and kernel density estimates are based on 2000 Monte Carlo replications.

### D.2 Monte Carlo set-up – DGP 1c

This Monte Carlo experiment builds on DGP 1a. The only modification is that the autoregressive coefficient is increased from $\rho = 0.5$ to $\rho = 0.9$ to see if it affects the estimators' performance when the time series are more persistent. Figure D.5 displays the predictive densities.

Figure D.5: DGP 1c – Comparison of densities




*Note*: Models M1 – M3 are defined as in Section 4.1, with the difference of a higher autoregressive parameter of $\rho = 0.9$. The value of $y_t$ is set to the unconditional expected value of $y_t$.

### D.3 Monte Carlo set-up – DGP 4

In this experiment, I investigate the estimators' performance when the true DGP implies a trimodal predictive density, which has a rather "unusual" shape. This example demonstrates that the proposed estimators perform well even in such complicated cases. The DGP is specified as a mixture of the following models:

$$M1 : y_{t+1} = c_1 + 0.9y_t + \nu_{t+1} \qquad \nu_{t+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0, \sigma_1^2)\,, \tag{D.1}$$

$$M2 : y_{t+1} = c_2 + 0.9y_t + \varepsilon_{t+1} \qquad \varepsilon_{t+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0, \sigma_2^2)\,, \tag{D.2}$$

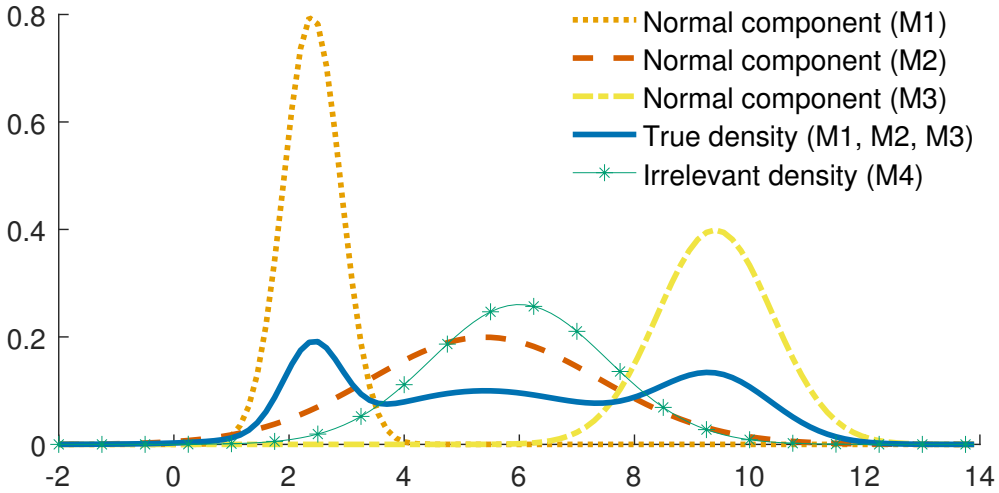
$$M3 : y_{t+1} = c_3 + 0.9y_t + \lambda_{t+1} \qquad \lambda_{t+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0, \sigma_3^2)\,, \tag{D.3}$$

with intercepts $c_1 = -3, c_2 = 0, c_3 = 4$, variances $\sigma_1^2 = 0.5^2, \sigma_2^2 = 2^2, \sigma_3^2 = 1^2$ and mixture weights $(w_1, w_2, w_3)' = (0.2, 0.5, 0.3)'$. A fourth model was added to the pool, specified as

$$M4 : y_{t+1} = c_4 + 0.9y_t + \eta_{t+1} \qquad \eta_{t+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0, \sigma_4^2)\,, \tag{D.4}$$

where the parameterization $c_4 = w_1c_1 + w_2c_2 + w_3c_3$ and $\sigma_4^2 = w_1\sigma_1^2 + w_2\sigma_2^2 + w_3\sigma_3^2$ guarantees that the first two moments of the predictive distribution of $y_{t+1}$ are the same for the mixture and the irrelevant models. Figure D.6 displays the predictive densities.

Figure D.6: DGP 4– Comparison of densities




*Note*: Normal components (M1), (M2) and (M3) refer to the predictive density of $y_{t+1}$ according to models M1, M2 and M3, respectively. True density (M1, M2, M3) is the mixture of the above densities with the correct weights $(w_1, w_2, w_3)' = (0.2, 0.5, 0.3)'$. Irrelevant density (M4) specified as a normal density with the same mean and variance as the true density. The value of $y_t$ is set to the unconditional expected value of $y_t$.

### D.4 Monte Carlo set-up – DGP 5

In this experiment, the true DGP is the mixture of an AR(1) process with *iid.* innovations (M1) and an AR(1) process where the innovations follow an autoregressive conditionally

heteroskedastic (ARCH, Engle (1982)) process (M2). The DGP is specified as the mixture of the following models:

$$M1: y_{t+1} = c_1 + \rho_1 y_t + \nu_{t+1} \qquad \nu_{t+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0, \sigma_1^2), \tag{D.5}$$

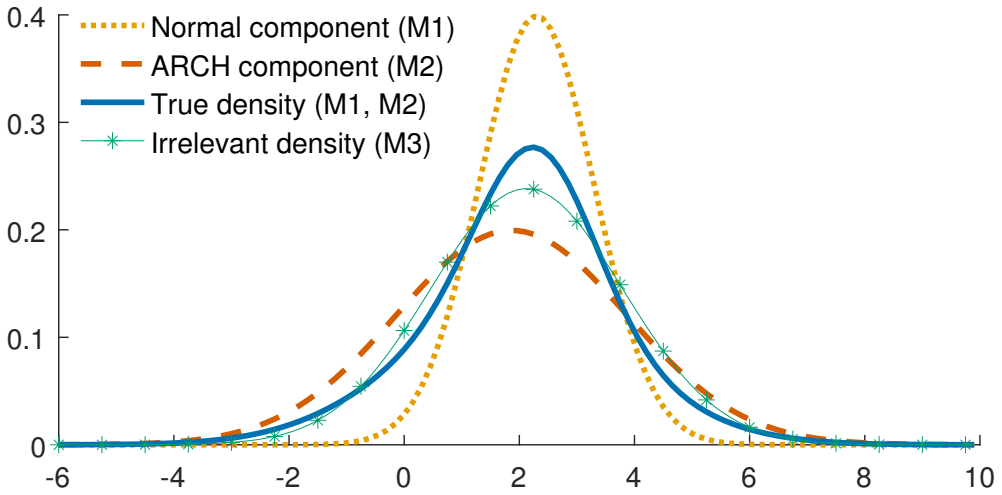
$$M2: y_{t+1} = c_2 + \rho_2 y_t + \sqrt{\sigma_{2,t+1}^2}\varepsilon_{t+1}, \quad \sigma_{2,t+1}^2 = \alpha_0 + \alpha_1 \varepsilon_t^2 \qquad \varepsilon_{t+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0,1), \tag{D.6}$$

with intercepts $c_1 = c_2 = 1$, autoregressive coefficients $\rho_1 = 0.4, \rho_2 = 0.6$ variance $\sigma_1^2 = 1$, ARCH coefficients $\alpha_0 = 2, \alpha_1 = 0.5$ and mixture weights $(w_1, w_2)' = (0.4, 0.6)'$. In the case of M2, the ARCH specification implies that the expected value of $\sigma_{2,t}^2$ is $\kappa \equiv E(\sigma_{2,t}^2) = \alpha_0/(1-\alpha_1)$. Once again, a third model was added to the pool, specified as

$$M3: y_{t+1} = c_3 + \rho_3 y_t + \eta_{t+1} \qquad \eta_{t+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0, \sigma_3^2), \tag{D.7}$$

where the parameterization $c_3 = w_1 c_1 + w_2 c_2$, $\rho_3 = w_1 \rho_1 + w_2 \rho_2$ and $\sigma_3^2 = w_1 \sigma_1^2 + w_2 \kappa$ guarantees that the first two moments of the predictive distribution of $y_{t+1}$ are the same for the mixture and the irrelevant models. Figure D.7 displays the predictive densities.

Figure D.7: DGP 5 – Comparison of densities




*Note*: Normal component (M1) and ARCH component (M2) refer to the predictive density of $y_{t+1}$, according to models M1 and M2, respectively. True density (M1, M2) is the mixture of the above densities with the correct weights $(w_1, w_2)' = (0.4, 0.6)'$. Irrelevant density (M3) specified as a normal density with the same mean and variance as the true density. The value of $y_t$ is set to the unconditional expected value of $y_t$.

### D.5 Monte Carlo set-up – DGP 6

This Monte Carlo set-up demonstrates the estimators' performance when the parameters of the predictive densities are estimated. The DGP is specified as the mixture of the following models:

$$M1: y_{t+1} = c_1 + \nu_{t+1} \qquad \nu_{t+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0, \sigma_1^2), \tag{D.8}$$

$$M2 : y_{t+1} = c_2 + \sqrt{\sigma^2_{2,t+1}}\varepsilon_{t+1}, \quad \sigma^2_{2,t+1} = \alpha_0 + \alpha_1 \varepsilon_t^2 \qquad \varepsilon_{t+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0,1)\,, \tag{D.9}$$

with intercepts $c_1 = c_2 = 1$, variance $\sigma_1^2 = 0.3$, ARCH coefficients $\alpha_0 = 0.2, \alpha_1 = 0.2$, and weights $(w_1, w_1)' = (0.4, 0.6)'$. In order to keep the problem tractable, the observations are generated sequentially (after an initial sample of size $R = 100$), based on the rolling window parameter estimates with window size $R = 100$, therefore the parameters listed above only correspond to the initial sample period. Once again, a third, irrelevant model was added to the pool, specified as

$$M3 : y_{t+1} = c_3 + \eta_{t+1} \qquad \eta_{t+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0, \sigma_3^2)\,, \tag{D.10}$$

where the parameterization $c_3 = w_1\widehat{c}_1 + w_2\widehat{c}_2$ and $\sigma_3^2 = w_1\widehat{\sigma}_1^2 + w_2\widehat{\sigma}^2_{2,t+1}$ guarantees that the first two moments of the predictive distribution of $y_{t+1}$ are the same for the mixture and the irrelevant models (note the "hats", emphasizing the estimated nature of the parameters). The Monte Carlo simulations were performed with $G = \{200, 500, 1000, 2000\}$, to keep $G > R$.

### D.6 Monte Carlo results – DGPs 1c, 4, 5 and 6

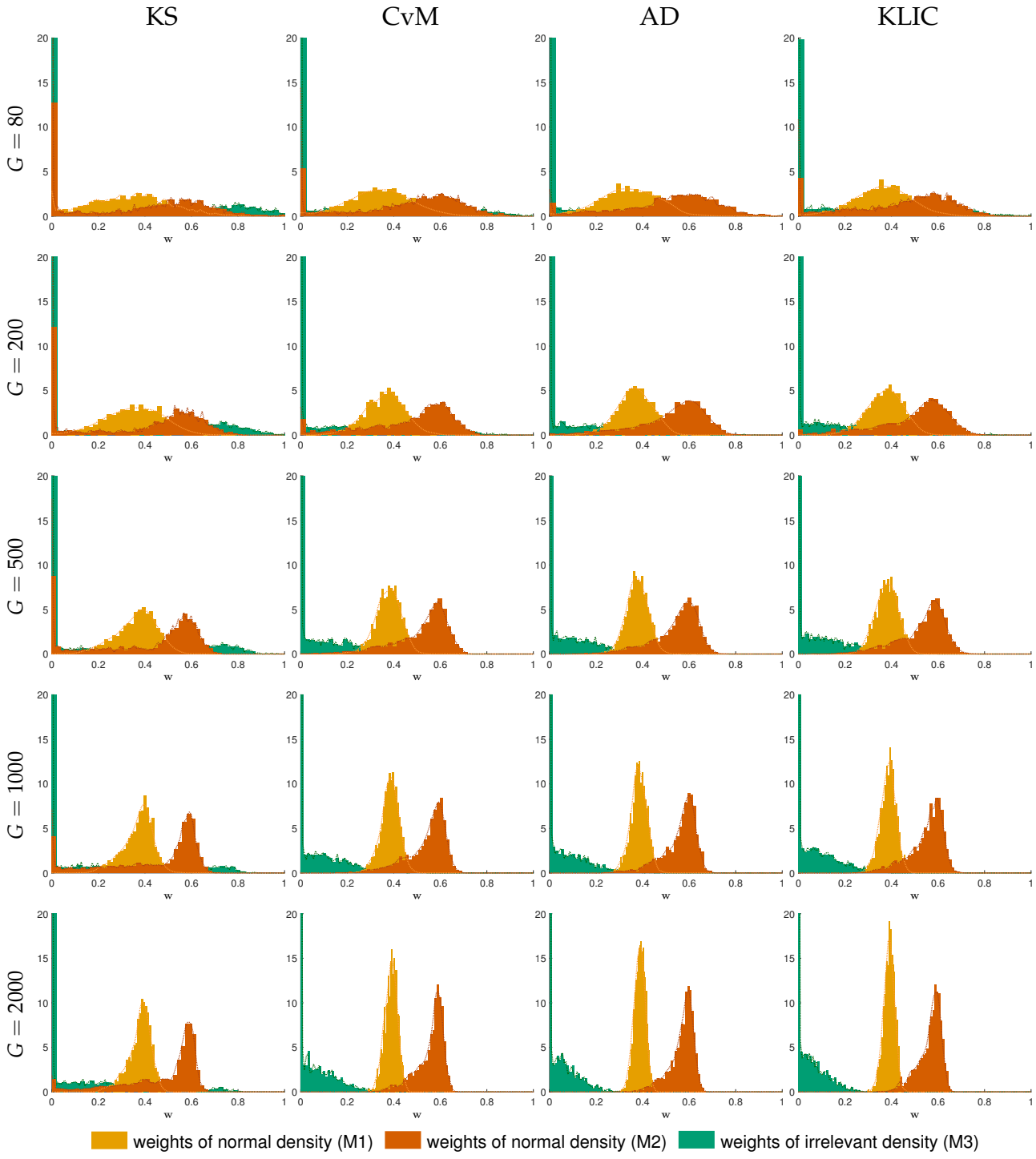
As Table D.1 and Figure D.8 show, increasing the autoregressive coefficient from $\rho = 0.5$ to $\rho = 0.9$ in DGP 1c does not affect the performance of any of the estimators.

Table D.1: DGP 1c, Monte Carlo summary statistics for different sample sizes $G$ and objective functions $K_G(w), C_G(w), A_G(w)$ and $\text{KLIC}_G(w)$

| Sample size | Statistic | KS | | | CvM | | | AD | | | KLIC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bias | −0.06 | −0.26 | 0.32 | −0.06 | −0.15 | 0.21 | −0.06 | −0.09 | 0.15 | −0.04 | −0.16 | 0.20 |
| $G = 80$ | Var | 0.03 | 0.08 | 0.13 | 0.02 | 0.06 | 0.08 | 0.02 | 0.04 | 0.05 | 0.01 | 0.05 | 0.07 |
| | MSE | 0.03 | 0.15 | 0.24 | 0.02 | 0.08 | 0.13 | 0.02 | 0.05 | 0.07 | 0.02 | 0.08 | 0.11 |
| | Bias | −0.04 | −0.23 | 0.27 | −0.04 | −0.13 | 0.16 | −0.03 | −0.07 | 0.10 | −0.02 | −0.10 | 0.13 |
| $G = 200$ | Var | 0.02 | 0.07 | 0.11 | 0.01 | 0.03 | 0.05 | 0.01 | 0.02 | 0.03 | 0.01 | 0.02 | 0.03 |
| | MSE | 0.02 | 0.12 | 0.19 | 0.01 | 0.05 | 0.08 | 0.01 | 0.03 | 0.04 | 0.01 | 0.03 | 0.05 |
| | Bias | −0.03 | −0.20 | 0.23 | −0.02 | −0.08 | 0.10 | −0.02 | −0.05 | 0.07 | −0.02 | −0.07 | 0.08 |
| $G = 500$ | Var | 0.01 | 0.05 | 0.09 | 0.00 | 0.01 | 0.02 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 |
| | MSE | 0.01 | 0.09 | 0.14 | 0.00 | 0.02 | 0.03 | 0.00 | 0.01 | 0.02 | 0.00 | 0.01 | 0.02 |
| | Bias | −0.03 | −0.16 | 0.19 | −0.01 | −0.06 | 0.07 | −0.01 | −0.04 | 0.05 | −0.01 | −0.05 | 0.06 |
| $G = 1000$ | Var | 0.00 | 0.04 | 0.06 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 |
| | MSE | 0.00 | 0.07 | 0.10 | 0.00 | 0.01 | 0.02 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 |
| | Bias | −0.02 | −0.12 | 0.14 | −0.01 | −0.04 | 0.05 | −0.01 | −0.03 | 0.04 | −0.01 | −0.03 | 0.04 |
| $G = 2000$ | Var | 0.00 | 0.03 | 0.04 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | MSE | 0.00 | 0.04 | 0.06 | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 |

*Note*: In the four main columns with headers KS, CvM, AD and KLIC, the table shows the estimates of the bias, variance (Var) and mean squared error (MSE) for each of the components of the weight vector $w$. True weights: $w = (0.4, 0.6, 0)'$. Statistics are based on 2000 Monte Carlo replications.

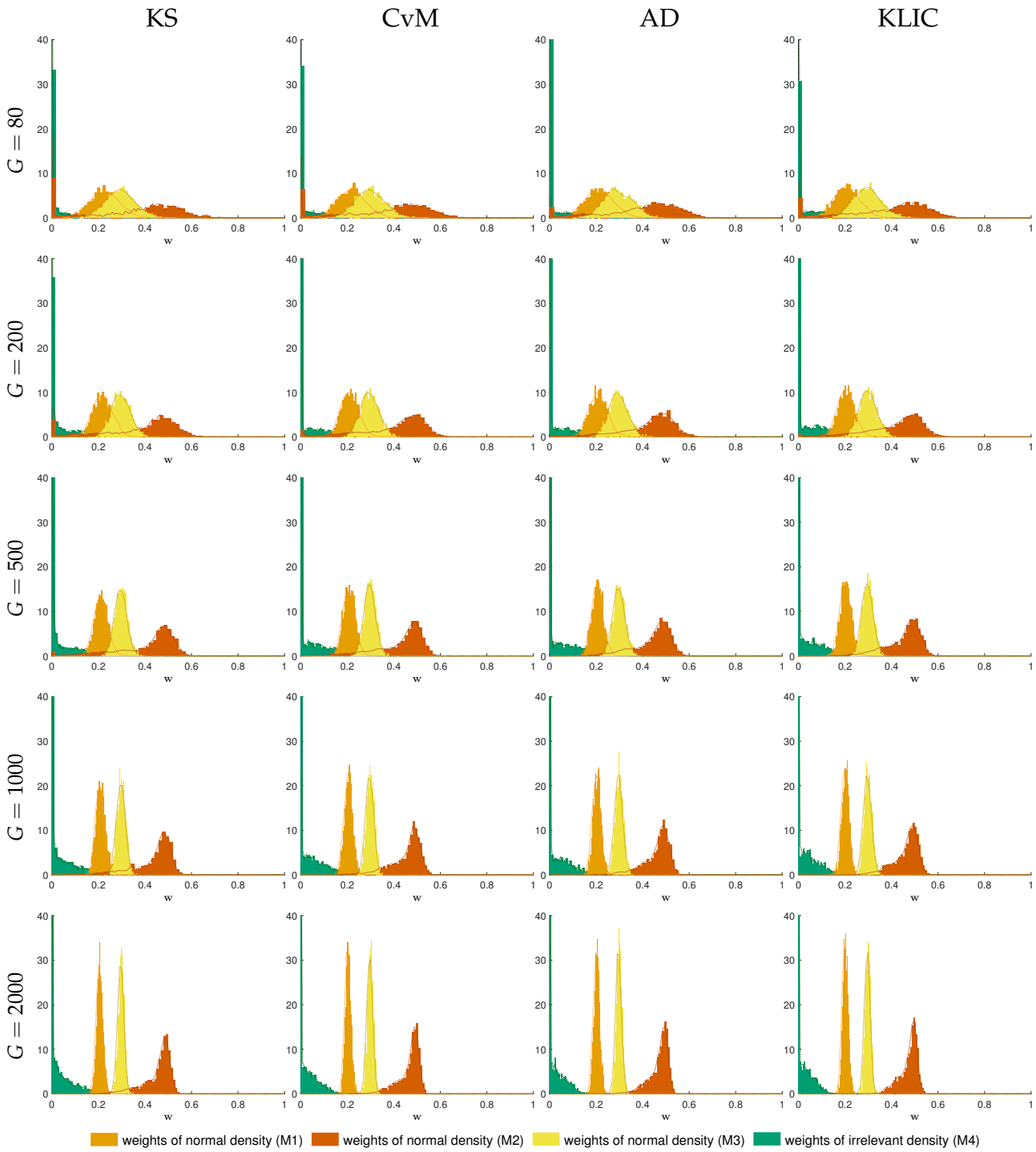
Figure D.8: Monte Carlo results for DGP 1c, true parameter vector $w = (0.4, 0.6, 0)'$




*Note*: $G$ denotes the sample size. KS, CvM, AD and KLIC stand for the Kolmogorov–Smirnov-, the Cramer–von Mises-, the Anderson–Darling- and the KLIC-based estimators, respectively. Histograms and kernel density estimates are based on 2000 Monte Carlo replications.

In the case of DGP 4, Figure D.9 and Table D.2 show that when increasing the number of potential models to four, all estimators still deliver satisfactory results and consistency is clearly demonstrated.

Figure D.9: Monte Carlo results for DGP 4, true parameter vector $w = (0.2, 0.5, 0.3, 0)'$




*Note*: $G$ denotes the sample size. KS, CvM, AD and KLIC stand for the Kolmogorov–Smirnov-, the Cramer–von Mises-, the Anderson–Darling- and the KLIC-based estimators, respectively. Histograms and kernel density estimates are based on 2000 Monte Carlo replications.

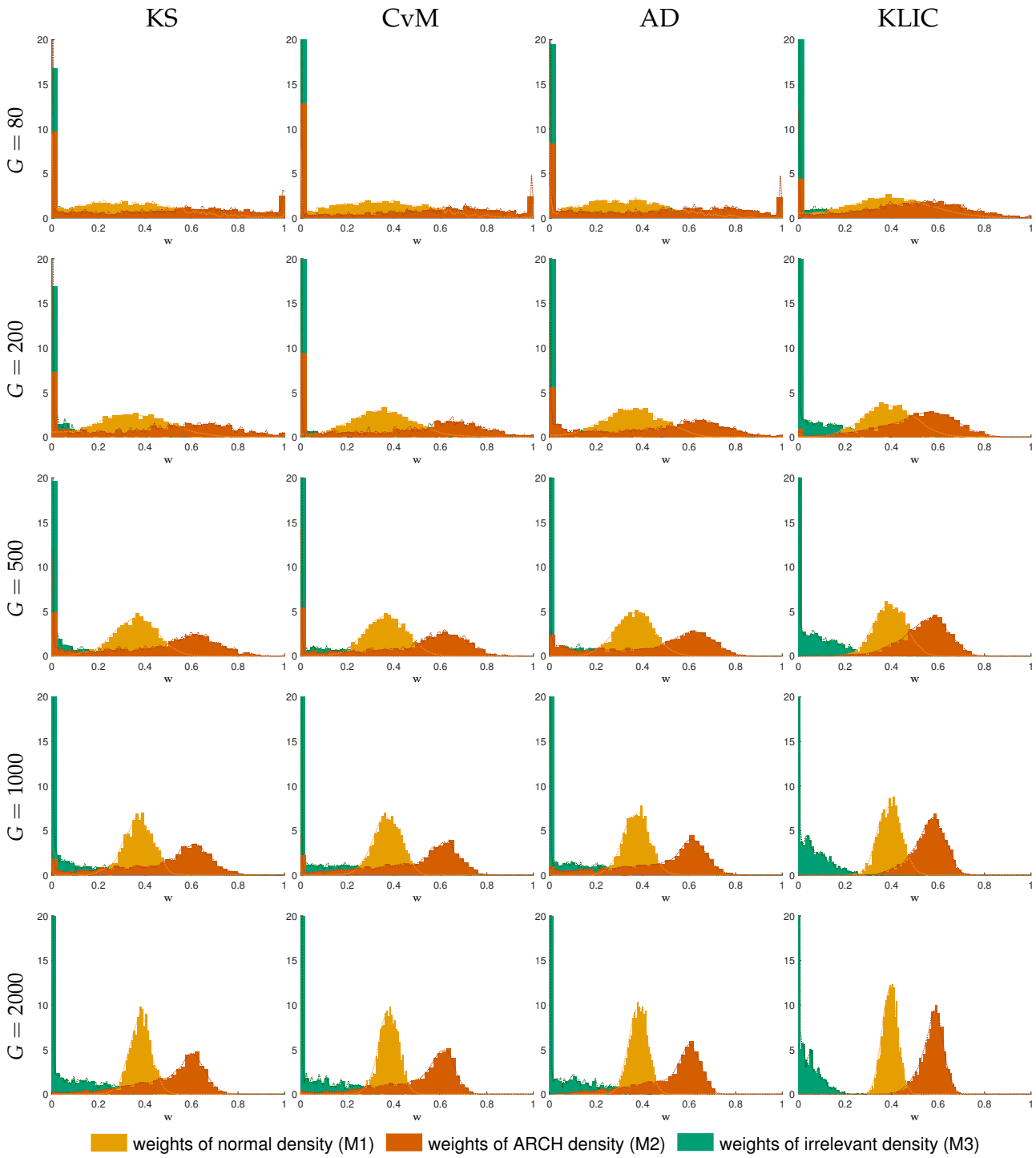
Table D.2: DGP 4, Monte Carlo summary statistics for different sample sizes $G$ and objective functions $K_G(w), C_G(w), A_G(w)$ and $\text{KLIC}_G(w)$

| Sample size | Statistic | KS | | | | CvM | | | | AD | | | | KLIC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $G = 80$ | Bias | 0.03 | −0.16 | −0.00 | 0.13 | 0.02 | −0.14 | 0.00 | 0.12 | 0.02 | −0.11 | −0.00 | 0.09 | 0.01 | −0.12 | −0.00 | 0.11 |
| | Var | 0.00 | 0.04 | 0.00 | 0.03 | 0.00 | 0.04 | 0.00 | 0.03 | 0.00 | 0.03 | 0.00 | 0.02 | 0.00 | 0.03 | 0.00 | 0.02 |
| | MSE | 0.00 | 0.07 | 0.00 | 0.05 | 0.00 | 0.06 | 0.00 | 0.04 | 0.00 | 0.04 | 0.00 | 0.03 | 0.00 | 0.05 | 0.00 | 0.03 |
| $G = 200$ | Bias | 0.02 | −0.12 | −0.00 | 0.10 | 0.01 | −0.09 | −0.00 | 0.08 | 0.01 | −0.07 | −0.00 | 0.06 | 0.01 | −0.08 | 0.00 | 0.07 |
| | Var | 0.00 | 0.03 | 0.00 | 0.02 | 0.00 | 0.02 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.02 | 0.00 | 0.01 |
| | MSE | 0.00 | 0.04 | 0.00 | 0.03 | 0.00 | 0.03 | 0.00 | 0.02 | 0.00 | 0.02 | 0.00 | 0.01 | 0.00 | 0.02 | 0.00 | 0.02 |
| $G = 500$ | Bias | 0.01 | −0.08 | −0.00 | 0.07 | 0.01 | −0.07 | −0.00 | 0.06 | 0.01 | −0.05 | 0.00 | 0.04 | 0.00 | −0.05 | −0.00 | 0.04 |
| | Var | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 |
| | MSE | 0.00 | 0.02 | 0.00 | 0.02 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 |
| $G = 1000$ | Bias | 0.01 | −0.06 | −0.00 | 0.05 | 0.01 | −0.04 | −0.00 | 0.04 | 0.01 | −0.04 | 0.00 | 0.03 | 0.00 | −0.04 | 0.00 | 0.03 |
| | Var | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | MSE | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $G = 2000$ | Bias | 0.01 | −0.04 | −0.00 | 0.04 | 0.00 | −0.03 | 0.00 | 0.03 | 0.00 | −0.03 | 0.00 | 0.02 | 0.00 | −0.02 | −0.00 | 0.02 |
| | Var | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | MSE | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

*Note*: In the four main columns with headers KS, CvM, AD and KLIC, the table shows the estimates of the bias, variance (Var) and mean squared error (MSE) for each of the components of the weight vector $w$. True weights: $w = (0.2, 0.5, 0.3, 0)'$. Statistics are based on 2000 Monte Carlo replications.

Inspecting Figure D.10 and Table D.3, we can see that in the case of DGP 5, the AD estimator seems to slightly dominate the KLIC estimator, and the KS and CvM estimators perform the worst.

Figure D.10: Monte Carlo results for DGP 5, true parameter vector $w = (0.4, 0.6, 0)'$




*Note*: $G$ denotes the sample size. KS, CvM, AD and KLIC stand for the Kolmogorov–Smirnov-, the Cramer–von Mises-, the Anderson–Darling- and the KLIC-based estimators, respectively. Histograms and kernel density estimates are based on 2000 Monte Carlo replications.

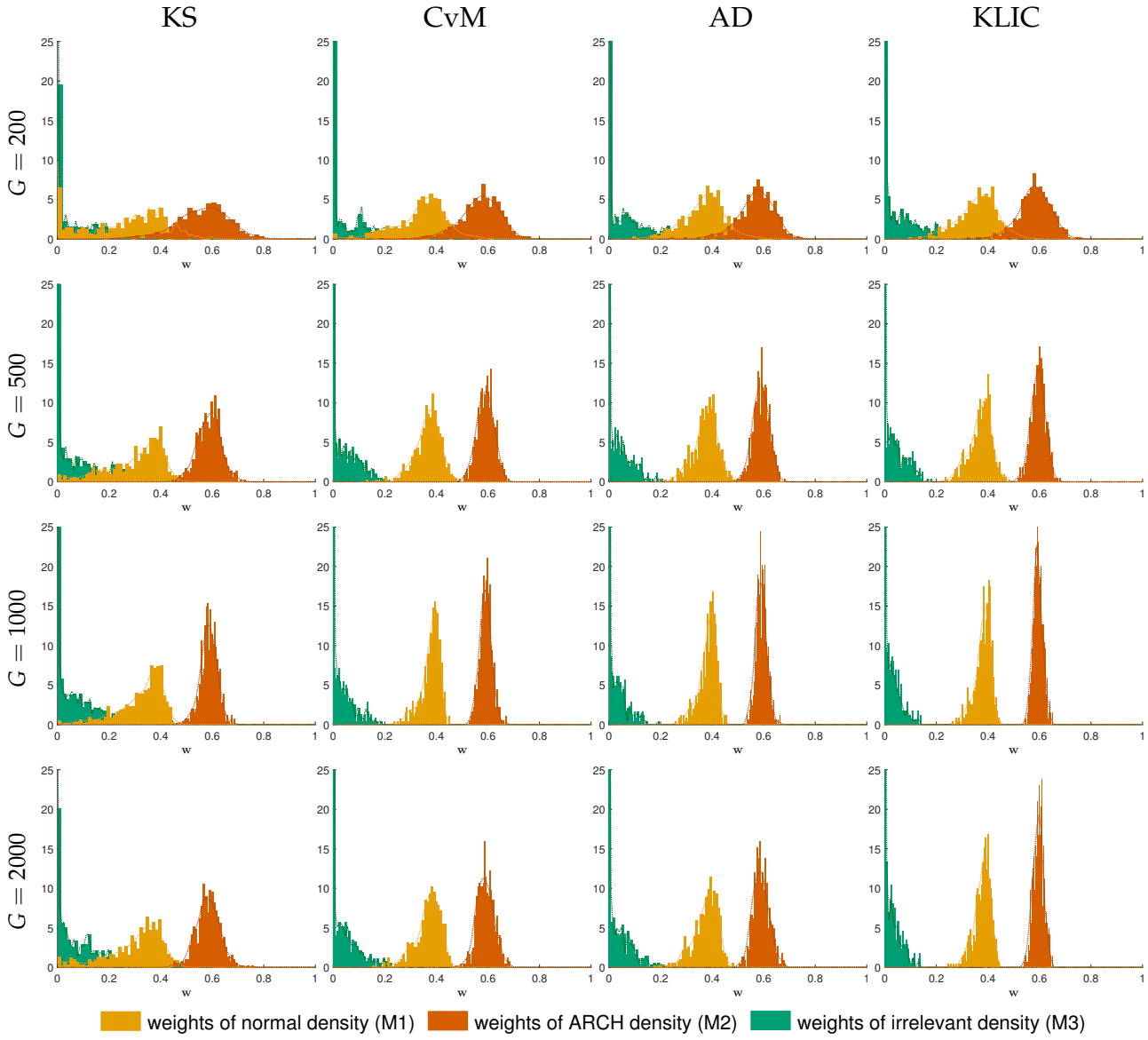
Table D.3: DGP 5, Monte Carlo summary statistics for different sample sizes $G$ and objective functions $K_G(w), C_G(w), A_G(w)$ and $\text{KLIC}_G(w)$

| Sample size | Statistic | KS | | | CvM | | | AD | | | KLIC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $G = 80$ | Bias | −0.05 | −0.26 | 0.31 | −0.06 | −0.16 | 0.21 | −0.06 | −0.10 | 0.15 | −0.04 | −0.15 | 0.20 |
| | Var | 0.00 | 0.00 | 0.00 | 0.03 | 0.08 | 0.13 | 0.01 | 0.05 | 0.06 | 0.01 | 0.05 | 0.07 |
| | MSE | 0.03 | 0.14 | 0.23 | 0.02 | 0.08 | 0.13 | 0.02 | 0.06 | 0.08 | 0.02 | 0.07 | 0.11 |
| $G = 200$ | Bias | −0.05 | −0.22 | 0.27 | −0.04 | −0.12 | 0.16 | −0.03 | −0.08 | 0.11 | −0.03 | −0.11 | 0.13 |
| | Var | 0.00 | 0.00 | 0.00 | 0.02 | 0.07 | 0.11 | 0.01 | 0.02 | 0.03 | 0.01 | 0.03 | 0.04 |
| | MSE | 0.02 | 0.12 | 0.19 | 0.01 | 0.05 | 0.08 | 0.01 | 0.03 | 0.04 | 0.01 | 0.04 | 0.05 |
| $G = 500$ | Bias | −0.04 | −0.20 | 0.24 | −0.02 | −0.08 | 0.10 | −0.02 | −0.05 | 0.07 | −0.02 | −0.07 | 0.09 |
| | Var | 0.00 | 0.00 | 0.00 | 0.01 | 0.06 | 0.09 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 |
| | MSE | 0.01 | 0.10 | 0.15 | 0.00 | 0.02 | 0.03 | 0.00 | 0.01 | 0.02 | 0.00 | 0.02 | 0.02 |
| $G = 1000$ | Bias | 0.16 | −0.31 | 0.16 | 0.18 | −0.24 | 0.06 | 0.18 | −0.22 | 0.04 | 0.19 | −0.22 | 0.04 |
| | Var | 0.00 | 0.00 | 0.00 | 0.01 | 0.02 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | MSE | 0.03 | 0.12 | 0.07 | 0.03 | 0.06 | 0.01 | 0.04 | 0.05 | 0.00 | 0.04 | 0.05 | 0.00 |
| $G = 2000$ | Bias | 0.17 | −0.29 | 0.12 | 0.19 | −0.23 | 0.04 | 0.19 | −0.22 | 0.03 | 0.19 | −0.22 | 0.03 |
| | Var | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | MSE | 0.03 | 0.09 | 0.04 | 0.04 | 0.05 | 0.00 | 0.04 | 0.05 | 0.00 | 0.04 | 0.05 | 0.00 |

*Note*: In the four main columns with headers KS, CvM, AD and KLIC, the table shows the estimates of the bias, variance (Var) and mean squared error (MSE) for each of the components of the weight vector $w$. True weights: $w = (0.4, 0.6, 0)'$. Statistics are based on 2000 Monte Carlo replications.

Figure D.11 and Table D.4 show that, in line with the theoretical results of the paper, all estimators are consistent for the true weight vector. These results confirm that the Anderson–Darling and the KLIC estimators are slightly better than the Cramer–von Mises-type estimator, which in turn outperforms the Kolmogorov–Smirnov-type estimator.

Figure D.11: Monte Carlo results for DGP 6, true parameter vector $w = (0.4, 0.6, 0)'$




*Note*: $G$ denotes the sample size. KS, CvM, AD and KLIC stand for the Kolmogorov–Smirnov-, the Cramer–von Mises-, the Anderson–Darling- and the KLIC-based estimators, respectively. Histograms and kernel density estimates are based on 2000 Monte Carlo replications.

Table D.4: DGP 6, Monte Carlo summary statistics for different sample sizes $G$ and objective functions $K_G(w), C_G(w), A_G(w)$ and $\text{KLIC}_G(w)$

| Sample size | Statistic | KS | | | CvM | | | AD | | | KLIC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $G = 200$ | Bias | −0.13 | −0.04 | 0.17 | −0.06 | −0.03 | 0.09 | −0.03 | −0.03 | 0.06 | −0.04 | −0.02 | 0.05 |
| | Var | 0.02 | 0.01 | 0.04 | 0.01 | 0.01 | 0.02 | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| | MSE | 0.04 | 0.01 | 0.06 | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.01 |
| $G = 500$ | Bias | −0.10 | −0.01 | 0.11 | −0.03 | −0.01 | 0.04 | −0.02 | −0.01 | 0.03 | −0.03 | −0.00 | 0.03 |
| | Var | 0.01 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | MSE | 0.02 | 0.00 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $G = 1000$ | Bias | −0.07 | −0.01 | 0.08 | −0.02 | −0.01 | 0.03 | −0.01 | −0.01 | 0.02 | −0.02 | −0.00 | 0.02 |
| | Var | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | MSE | 0.01 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $G = 2000$ | Bias | −0.10 | −0.01 | 0.11 | −0.03 | −0.01 | 0.04 | −0.02 | −0.01 | 0.03 | −0.02 | −0.00 | 0.02 |
| | Var | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | MSE | 0.02 | 0.00 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

*Note*: In the four main columns with headers KS, CvM, AD and KLIC, the table shows the estimates of the bias, variance (Var) and mean squared error (MSE) for each of the components of the weight vector $w$. True weights: $w = (0.4, 0.6, 0)'$. Statistics are based on 2000 Monte Carlo replications.

# E Empirical exercise – additional results

Figure E.1 shows the ratio of the inverse of the in-sample residual variances of each model, relative to the sum of the inverses, calculated in the last rolling window at each forecast origin. Bates and Granger (1969) recommended this ratio as an estimator of the optimal weights, minimizing the expected Root Mean Squared Forecast Error. The figure displays very stable weights, all around 1/5, corresponding to equal weights.

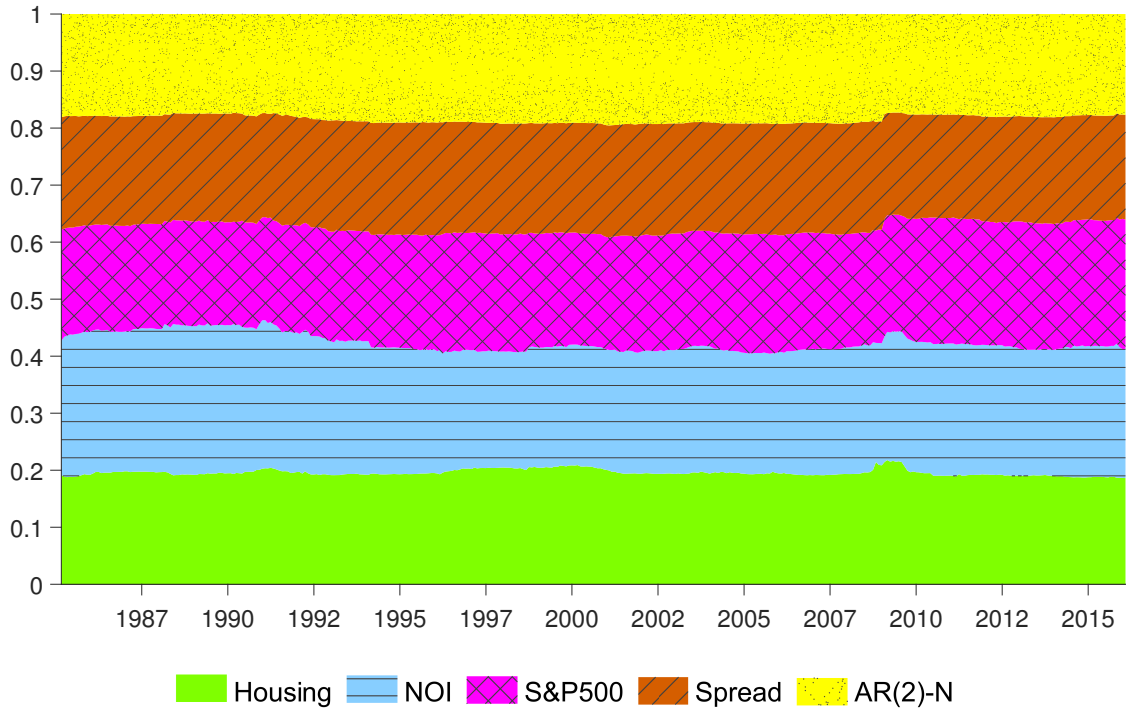
Figure E.2 shows the values of the Anderson–Darling and the KLIC objective functions for each model at each forecast origin.

As Figure E.2a confirms, the model including the New Orders Index produced the best *in-sample* density forecasts until around 2002. From about 2002 to 2009, the values of the Anderson–Darling objective function corresponding to all the other models were lower than those of the model with the New Orders Index. Furthermore, they moved closely together until around 2010, when corporate bond spreads gained considerable predictive power. Moreover, housing permits have delivered the best density forecasts since 2013. When considering the KLIC estimator, Figure E.2b shows that corporate bond spreads featured prominently until around 1996, along with the New Orders Index.

The individual models' KLIC values do not show such dispersion as in the case of the Anderson–Darling estimator. This suggests that the AD estimator was able to exploit the differences between the individual models' predictive densities more successfully than the KLIC estimator. As Table 7 showed, this gain resulted in superior *out-of-sample* density forecasts.

A visual comparison of Figure E.2a and Figure E.2b reveals that both the Anderson–

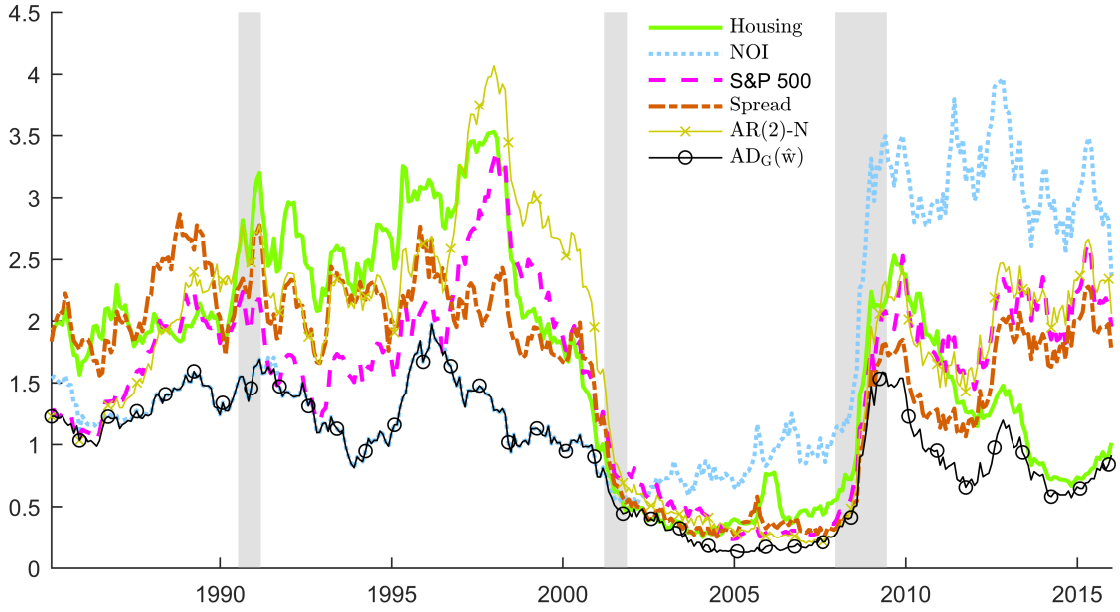
Figure E.1: Ratios of inverse in-sample residual variances



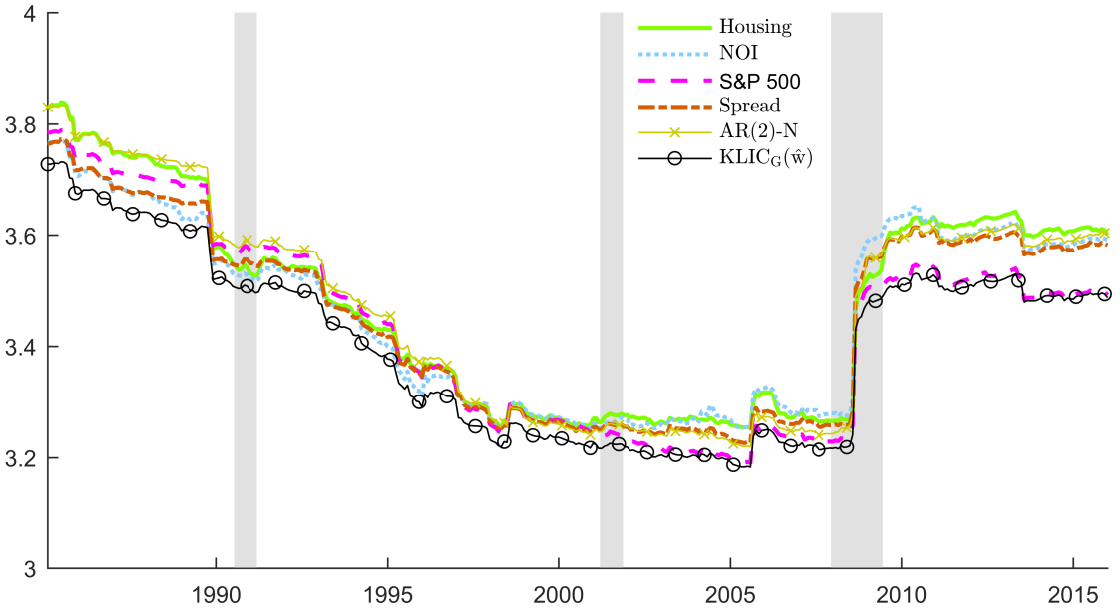

*Note*: The sample period (end of the last rolling window of size $R = 120$) starts in February 1985 and ends in January 2016, with a total number of $P = 372$ months. Housing stands for Housing Permits, NOI stands for ISM: New Orders Index, S&P 500 is the S&P 500 stock index returns while Spread is Moody's Baa Corporate Bond Yield minus Fed funds rate.

Darling and the KLIC statistics imply that US industrial production growth was the most predictable from around 1999 until shortly before the Great Recession. However, while the individual models' Anderson–Darling statistics in Figure E.2a show an upward trend (corresponding to less predictive power) until approximately 1998, the KLIC displays an uninterrupted downward trend (corresponding to more predictive power) in Figure E.2b. The Great Recession reversed this improvement in predictability.

Figure E.2: Time-variation of the values of the Anderson–Darling and the KLIC objective functions




(a) Anderson–Darling objective function




(b) KLIC objective function

*Note*: The forecast origins range from February 1985 to January 2016, with a total number of $P = 372$ months. Housing stands for Housing Permits, NOI stands for ISM: New Orders Index, S&P 500 is the S&P 500 stock index returns while Spread is Moody's Baa Corporate Bond Yield minus Fed funds rate. Shaded areas are NBER recession periods. $\text{AD}_G(\widehat{w})$ and $\text{KLIC}_G(\widehat{w})$ are the values of the AD and KLIC objective functions using the model combinations, respectively, evaluated at the corresponding weight estimates.

## F Likelihoods

This section lists the likelihoods used in the Monte Carlo simulations (Section 4 and Appendix D) and the empirical exercise (Section 5). To simplify notation, consider the model $y_{t+1} = z_t'\beta + \sqrt{\sigma^2}\varepsilon_{t+1}$, where $\varepsilon_{t+1}$ is either *iid*. standard normal, *iid*. standardized Student's $t$, or its variance follows an ARCH(1) process (Engle, 1982) with *iid*. standard normal innovations.

The conditional likelihoods are denoted by $\ell(y_{t+1}|z_t;\beta,\sigma^2)$, $\ell(y_{t+1}|z_t;\beta,\sigma^2,\nu)$ and $\ell(y_{t+1}|z_t;\beta,\alpha_0,\alpha_1)$, respectively.

1. Standard normal:

$$\ell(y_{t+1}|x_t;\beta,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{-0.5}}\exp\left(-\frac{1}{2}\frac{(y_{t+1}-z_t'\beta)^2}{\sigma^2}\right). \tag{F.1}$$

2. Standardized Student's $t$:

$$\ell(y_{t+1}|x_t;\beta,\sigma^2,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\sigma^2(\nu-2)\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{(y_{t+1}-z_t'\beta)^2}{(\nu-2)\sigma^2}\right)^{-\frac{\nu+1}{2}}, \tag{F.2}$$

   where $\nu$ is the degrees of freedom parameter, restricted to be greater than 2 so that the variance is finite, and $\Gamma(\cdot)$ is the gamma function.

3. ARCH(1) model with normal innovations: similar to the standard normal case above, replacing $\sigma^2$ by

$$\sigma^2_{t+1} = \alpha_0 + \alpha_1\varepsilon_t^2, \tag{F.3}$$

   where $(\alpha_0,\alpha_1)$ are additional parameters entering the likelihood function.

The sample log-likelihoods and the scores follow in a straightforward way.