Essays in Macroeconometrics

Gergely Ákos Gánics

TESI DOCTORAL UPF / 2017

DIRECTOR DE LA TESI
Professor Barbara Rossi
Departament d'Economia i Empresa

Édesapám emlékének.

Acknowledgments

This thesis represents not only the summary of the research I have carried out at UPF in the past years, it is also a piece of work which was made possible thanks to a large number of people.

First and foremost I offer my heartfelt thanks to my advisor Barbara Rossi, whose constant encouragement, enthusiasm and dedication helped me overcome the obstacles and reach this stage. Her brilliant insights and guidance proved to be invaluable throughout the years. *Grazie di cuore per tutto*, Barbara!

I owe special thanks to Majid Al-Sadoon, whose patience and commitment to precision greatly improved my papers. Furthermore, Christian Brownlees and Geert Mesters provided valuable advice countless times, for which I am very thankful.

Among the professors who helped me at various points during my studies, I would like to thank Jordi Galí, Christian Matthes, Dávid Krisztián Nagy, Kristoffer Nimark, Balázs Kotosz, Erzsébet Kovács and Kolos Csaba Ágoston.

I received very helpful comments on my first chapter from Isaac Baley, Davide Debortoli, Eleonora Granziera, Frank Kleibergen, Malte Knüppel, Juri Marcucci, James Morley, Denis Nekipelov, Elena Pesavento, Tatevik Sekhposyan and participants at the VI_t Workshop in Time Series Econometrics, the 2016 Barcelona GSE Summer Forum's Workshop on Time Series Econometrics and Applications in Macroeconomics and Finance, the 4th SIdE-IEA Workshop for PhD students in Econometrics and Empirical Economics, the CREi Macroeconomics Breakfast Seminar, the Banco de España, the Bank of Canada, the Banque de France and the Università di Bologna. In addition, Kristoffer Nimark, and participants of the V_t Workshop in Time Series Econometrics and the 2015 Belgrade Young Economists Conference provided useful advice on the second paper of the present thesis. Collaborating with Atsushi Inoue on the third chapter was a great pleasure.

The financial stability thanks to the Obra Social "la Caixa" – Barcelona GSE Scholarship, the Spanish Ministerio de Economía y Competitividad (FPI grant BES–2013–065352) and UPF helped me focus on my studies and enjoy the fascinating city of Barcelona. *¡Muchas gracias España! Moltes gràcies Barcelona!*

I would also like to thank Marta Araque, Laura Agustí and Mariona Novoa for their fantastic help with administrative duties and for being so kind and supportive throughout my time at UPF.

During my journey as a PhD student, I had the privilege to meet great people and forge long-lasting friendships. Carlos, César, David, Frank, Lore, Matías and

Stef: I will always be grateful for your advice, for cheering me up in the hardest of times, and being my friends. Derrick, Donghai, Flo, Jagdish, Jenny, Karolis, Luca, Stefan and Christian: you made my time in Barcelona and particularly on campus so much fun. Sanyi, Viktor, Raj, Szandra, Ivett, Lilla, Csaba, Szepi and Denisa: you were always there (and sometimes here) for me, and I always cherished the time we spent together.

Finally, I want to thank my wonderful family for all their caring support throughout my six years in Barcelona. In particular, my sister, my mother and my grandmother deserve all my gratitude for their unconditional love, help and encouragement. Without you, I would not have been able to complete this journey. I wish I could ever return all you have done for me. *Szívből köszönök szépen mindent!*

Abstract

This thesis consists of three chapters on topics in macroeconometrics. Chapter 1 provides a novel estimator of combination weights which delivers well-calibrated density forecasts. In an empirical example of forecasting US industrial production, I show that my proposed methodology outperforms several benchmark combination schemes, and the weights indicate that financial variables proved to be useful predictors during the Great Recession. Chapter 2 investigates time-variation in the forecasting performance of structural Dynamic Stochastic General Equilibrium models and reduced-form statistical models. I show that the models' in-sample forecasting ability was strongly related to their out-of-sample performance before the recent financial crisis, but this link considerably weakened at the onset of the crisis. In Chapter 3 we propose a methodology to construct confidence intervals for the strength of identification in both instrumental variable models and Structural Vector Autoregressive models identified with an external instrument. We illustrate the proposed method using three leading empirical examples: the New Keynesian Phillips Curve, a linearized Euler equation, and a Structural Vector Autoregressive model describing the dynamic effects of oil shocks.

Resum

La present tesi es composa de tres capítols sobre temes de macroeconometria. El capítol 1 introdueix un nou estimador de combinacions de pesos que dóna prediccions de densitat ben calibrades. En un exemple empíric de predicció de la producció industrial dels EUA, demostro que l'aplicació d'aquesta metodologia millora molts dels esquemes de combinació de referència i els pesos indiquen que les variables financeres són predictors útils de la Gran Recessió. El capítol 2 investiga la variació temporal en la capacitat de predicció dels models dinàmics estocàstics d'equilibri general i dels models estadístics de forma reduïda. Demostro que la capacitat de predicció del model dins de la mostra estava fortament relacionada amb el seu rendiment fora de la mostra abans de la recent crisi financera, però aquest vincle es fa feble amb l'inici de la crisi. En el capítol 3 proposem una metodologia per construir intervals de confiança per la força d'identificació tan en models de variables instrumentals com en models estructurals de vectors autoregressius identificats amb un instrument extern. Il·lustrem la metodologia proposada utilitzant tres exemples empírics importants: La Corba de Phillips Neokeynesiana, una equació d'Euler linealitzada i un model estructural de vectors autoregressius que descriu les dinàmiques dels efectes dels xocs del petroli.

Preface

This thesis consists of three self-contained chapters on topics in macroeconometrics, both theoretical and empirical. In particular, density forecast combination, point forecast evaluation, and weak identification are the main themes of the papers.

Chapter 1, "Optimal Density Forecast Combinations", studies how researchers should 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.

In Chapter 2, "Forecasting with DSGE versus Reduced-Form Models: A Time-Variation Perspective", the out-of-sample forecasting performance of a leading Dynamic Stochastic General Equilibrium (DSGE) model is investigated. First, I demonstrate that, while the model delivers competitive forecasts against a number of statistical models, its predictive ability displays time-variation. Generally, in turbulent times, such as the recent financial crisis, simpler statistical models forecast better. Second, I show that swings in the model's absolute and relative out-of-sample performance are strongly related to its in-sample performance. Specifically, I find that the DSGE model's in-sample fit is highly informative in the early 2000s but the financial crisis deteriorated this link. Third, I find that extending a DSGE model with financial frictions results in better forecasting performance in times of financial distress but not in other times.

In Chapter 3, "Confidence Intervals for the Strength of Identification" (joint with Atsushi Inoue and Barbara Rossi), we propose a novel methodology to construct confidence intervals for the strength of identification in both instrumental variable models as well as Structural Vector Autoregressive models identified with an external instrument. Unlike tests for weak instruments, whose distributions are non-standard and depend on nuisance parameters that cannot be consistently

estimated, the confidence intervals are straightforward and computationally easy to calculate, as they are obtained from inverting chi-squared distributions. Another appealing feature of our methodology is that it is valid in the presence of heteroskedasticity and serial correlation. Monte Carlo simulations show that the confidence intervals have good small sample coverage. We illustrate the proposed method to measure the strength of identification in three leading empirical situations: the New Keynesian Phillips Curve, a linearized Euler equation, and a Structural Vector Autoregressive model describing the dynamic effects of oil shocks.

Contents

List of Figures

List of Tables

Chapter 1

Optimal Density Forecast Combinations

1.1 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

[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`.

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 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 esti-

mator 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 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).

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

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 1.2 introduces the notation and the definitions used throughout the paper. Section 1.3 describes the forecasting environment and the proposed density forecast combination method, while Section 1.4 provides the results of Monte Carlo exercises. An empirical application of forecasting US industrial production is presented in Section 1.5, then Section 1.6 concludes. The proofs are collected in Appendix A, while additional technical details and results can be found in Appendices B to F.

1.2 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

[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

and identically distributed, and $\mathcal{N}(\mu, \mathbb{V})$ is the normal distribution with mean vector μ and covariance matrix $\mathbb{V}$. 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

$$\phi_{t+h}^{C}(y|\mathcal{I}_{t-R+1}^{t}) \equiv \sum_{m=1}^{\mathcal{M}} w_m \phi_{t+h}^{m}(y|\mathcal{I}_{t-R+1}^{t}), \tag{1.1}$$

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

$$\Phi_{t+h}^{C}(\overline{y}|\mathcal{I}_{t-R+1}^{t}) = \int_{-\infty}^{\overline{y}} \sum_{m=1}^{\mathcal{M}} w_m \phi_{t+h}^{m}(y|\mathcal{I}_{t-R+1}^{t})\, \mathrm{d}y \tag{1.2}$$

$$= \sum_{m=1}^{\mathcal{M}} w_m \Phi_{t+h}^{m}(\overline{y}|\mathcal{I}_{t-R+1}^{t}). \tag{1.3}$$

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.

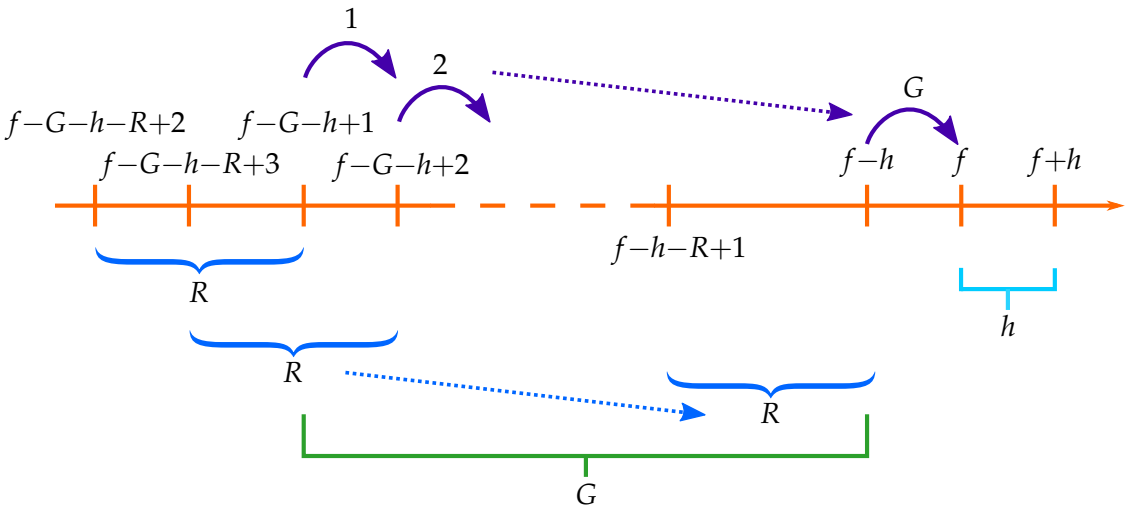
The estimation procedure is repeated in a similar way for all forecast origins $f = G + h + R - 1, \ldots, T$. This scheme yields a total number of $P = T - G - h - R$ out-of-sample density forecasts with the corresponding realizations, which could be used to assess the performance of the forecast combinations. Figure 1.1 provides a graphical illustration of the proposed estimation scheme. By using a rolling window scheme, researchers can potentially alleviate problems

realization, as it should be clear from the context which is meant.

[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.

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".

Figure 1.1: 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.

The true distribution of y_{t+h} *conditional* on $\mathcal{I}^t_{t-R+1}$ is denoted by $\Phi^*_{t+h}(\overline{y}|\mathcal{I}^t_{t-R+1})$. If for a given w, $\sum_{m=1}^{\mathcal{M}} w_m \Phi^m_{t+h}(\overline{y}|\mathcal{I}^t_{t-R+1})$ coincides with $\Phi^*_{t+h}(\overline{y}|\mathcal{I}^t_{t-R+1})$, 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 $\sum_{m=1}^{\mathcal{M}} w_m \Phi^m_{t+h}(\overline{y}|\mathcal{I}^t_{t-R+1}) = \Phi^*_{t+h}(\overline{y}|\mathcal{I}_t)$, then the forecast is said to satisfy *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.

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

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 Σ. 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 Σ.

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 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.

1.2.1 The Probability Integral Transform

The Probability Integral Transform (PIT) is defined as

$$z_{t+h} \equiv \int_{-\infty}^{y_{t+h}} \phi^C_{t+h}(y|\mathcal{I}^t_{t-R+1})\, \mathrm{d}y = \Phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})\,, \tag{1.4}$$

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

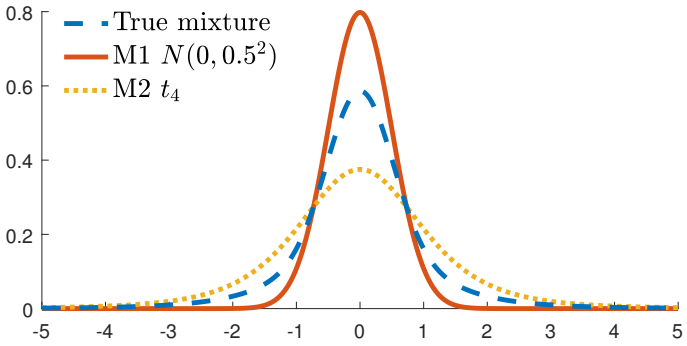
[7] The original result is usually attributed to Rosenblatt (1952), while in the econometrics literature it was introduced by Diebold et al. (1998). The discussion in Corradi and Swanson (2006a) and Gneiting et al. (2007) is the closest to the framework of the present study.

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, $\phi^1_{t+1}(y_{t+1}|\mathcal{J}^t_{t-R+1}) = \mathcal{N}(0, 0.5^2)$ and the second one is the Student's t component, $\phi^2_{t+1}(y_{t+1}|\mathcal{J}^t_{t-R+1}) = t_4$. Furthermore, the third density is the correct mixture density.

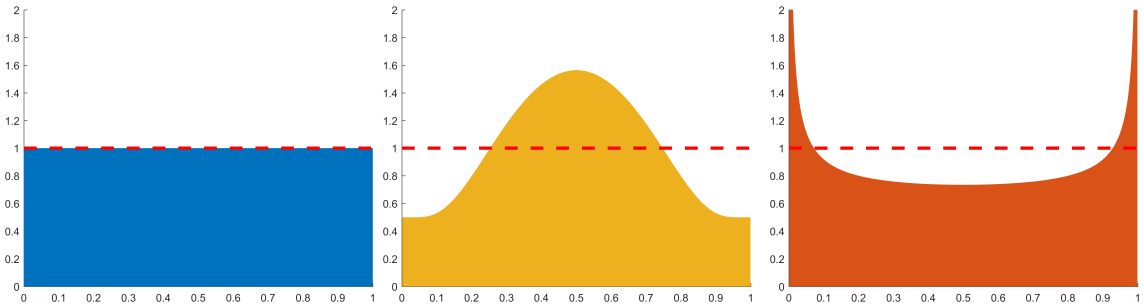
Figure 1.2 displays the three PDFs. We can see that while the means of the incorrectly calibrated densities are the same as the true forecast density's mean, their tails are markedly different, with the normal density featuring thinner and Student's t-density displaying thicker tails than the true mixture density.

Figure 1.2: Probability density functions of candidate forecast densities



I calculated the PITs using each of the three models above. The PDFs of each of the PITs in Figure 1.3 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 1.4, we can see that the CDF of the PITs obtained 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. ▲

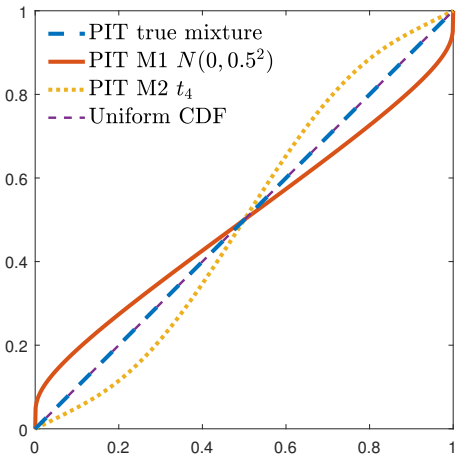
Figure 1.3: Probability density functions of PITs



(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 1.4: 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

$$\xi_{t+h}(r, w) \equiv 1\left[\Phi^{C}_{t+h}(y_{t+h}|\mathcal{I}^{t}_{t-R+1}) \leq r\right] - r = 1\left[z_{t+h} \leq r\right] - r \tag{1.5}$$

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:

$$\Psi_G(r, w) \equiv G^{-1} \sum_{t=f-G-h+1}^{f-h} \xi_{t+h}(r, w)\,, \tag{1.6}$$

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 ρ 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:

$$K_G(w) \equiv \sup_{r \in \rho} |\Psi_G(r, w)|\,, \tag{1.7}$$

$$C_G(w) \equiv \int_{\rho} \Psi_G^2(r, w)\, \mathrm{d}r\,, \tag{1.8}$$

$$A_G(w) \equiv \int_{\rho} \frac{\Psi_G^2(r, w)}{r(1-r)}\, \mathrm{d}r\,. \tag{1.9}$$

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

[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.

listically 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.

1.2.2 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 ϱ 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 ϱ, for example focusing on discrepancies in the $[-3\%, 0\%]$ range when forecasting recessions. If the whole distribution is of interest, then ϱ 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 ϱ is defined as

$$\text{KLIC}_\varrho(\Phi^*_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}), \Phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})) \tag{1.10}$$

$$\equiv \int_{-\infty}^{\infty} \phi^*_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}) \log \frac{\phi^*_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})}{\phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})} 1[y_{t+h} \in \varrho] \, \mathrm{d}y_{t+h} \tag{1.11}$$

$$= E_{\phi^*}\left\{ \left(\log \phi^*_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}) - \log \phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}) \right) 1[y_{t+h} \in \varrho] \right\} \tag{1.12}$$

$$\begin{aligned} &= E_{\phi^*}\left\{ \log \phi^*_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}) 1[y_{t+h} \in \varrho] \right\} - \\ &\quad E_{\phi^*}\left\{ \log \phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}) 1[y_{t+h} \in \varrho] \right\}, \end{aligned} \tag{1.13}$$

where the subscripts in Equations (1.12) and (1.13) remind us that the expectations

[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).

are taken with respect to the *true* predictive density. It is well known that KLIC ≥ 0, and KLIC $= 0$ if and only if $\Phi^*_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}) = \Phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})$ 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 $\phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})$ is the true predictive density but then we are informed that it is $\phi^*_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})$ instead (White, 1994, Chapter 2, p.9). The first term in Equation (1.13) 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

$$\overline{\text{KLIC}}_0 \equiv G^{-1} \sum_{t=f-G-h+1}^{f-h} -E_{\phi^*}\left\{\log \phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}) 1[y_{t+h} \in \varrho]\right\}, \qquad (1.14)$$

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:

$$\text{KLIC}_G(w) = G^{-1} \sum_{t=f-G-h+1}^{f-h} \left\{-\log \phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}) 1[y_{t+h} \in \varrho]\right\}. \qquad (1.15)$$

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 ρ or ϱ 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.

1.3 Estimators and assumptions

In this section I will discuss how the aforementioned statistics defined in Equations (1.7) to (1.9) and (1.15) can be used as objective functions to estimate the weights and I outline the assumptions that render the estimators consistent.

As discussed in Section 1.2, 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

$$\widehat{w} = \underset{w \in \Delta^{\mathcal{M}-1}}{\operatorname{argmin}} \, T_G(w) \,, \tag{1.16}$$

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

$$\widehat{w} = \underset{w \in \Delta^{\mathcal{M}-1}}{\operatorname{argmin}} \, \mathrm{KLIC}_G(w) \,. \tag{1.17}$$

Before stating and discussing the assumptions that guarantee consistency of the estimators defined in Equations (1.16) and (1.17), 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 1.4 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).

[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.

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

Assumption 1 (Dependence). *$\{Z_t\}$ is ϕ-mixing of size $-k/(2k-1), k \geq 1$ or α-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 $\sum_{m=1}^{\mathcal{M}} w_m^* \Phi_{t+h}^m(y_{t+h}|\mathcal{I}_{t-R+1}^t) = \Phi_{t+h}^*(y_{t+h}|\mathcal{I}_{t-R+1}^t)$ for all t. It also allows for misspecification, provided there is a unique minimizer of the population objective function.[11] In the former case, the population 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 ρ 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 1.4) 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* (1.16) *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. ■

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

[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.

1.3.2 KLIC-based estimator

In this subsection I state and discuss some additional assumptions guaranteeing that the KLIC-based estimator defined in Equation (1.17) 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). *ϱ 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 ϱ, $\log \phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})$ is continuous in w for all t.*

Assumption 10 (Dominance). *Over ϱ, $|\log \phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})| \leq b(y_{t+h})$ 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 ϱ, $E|(\log \phi^C_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}))|^{k+\tau} < \Delta < \infty$ 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* (1.14).

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 (1.12) to Equation (1.13). 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 $\sum_{m=1}^{\mathcal{M}} w^*_m \Phi^m_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1}) = \Phi^*_{t+h}(y_{t+h}|\mathcal{I}^t_{t-R+1})$ for all t, 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* (1.17) *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. Both 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 strictly greater than zero. 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. ▲

1.4 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.

1.4.1 Monte Carlo set-up – DGP 1

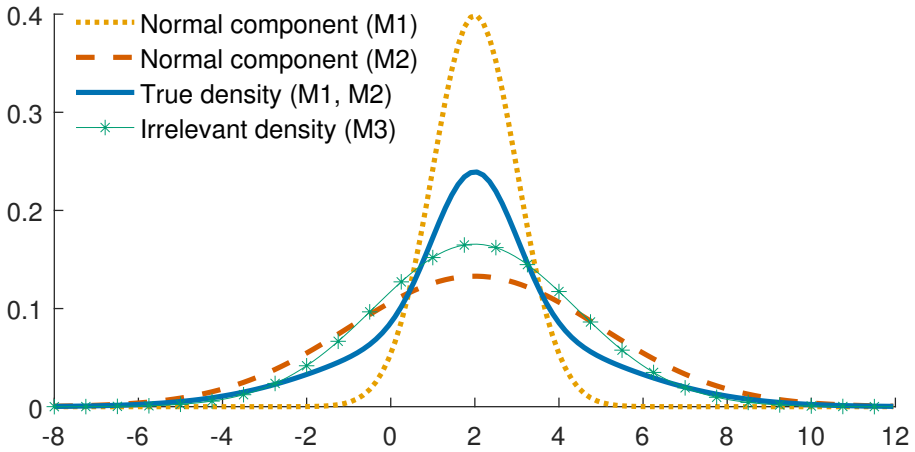
Both DGP 1a and DGP 1b feature three AR(1) models with *iid*. normal error terms. The models labeled as M1, M2 and M3 are given by

$$y_{t+h} = c^{(j)} + \rho_1^{(j)} y_t + \varepsilon_{t+h} \qquad \varepsilon_{t+h} \overset{\text{iid.}}{\sim} \mathcal{N}(0, \sigma_j^2)\,, \tag{1.18}$$

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 1.1 and Figure 1.5 displays the predictive densities.

Figure 1.5: 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.

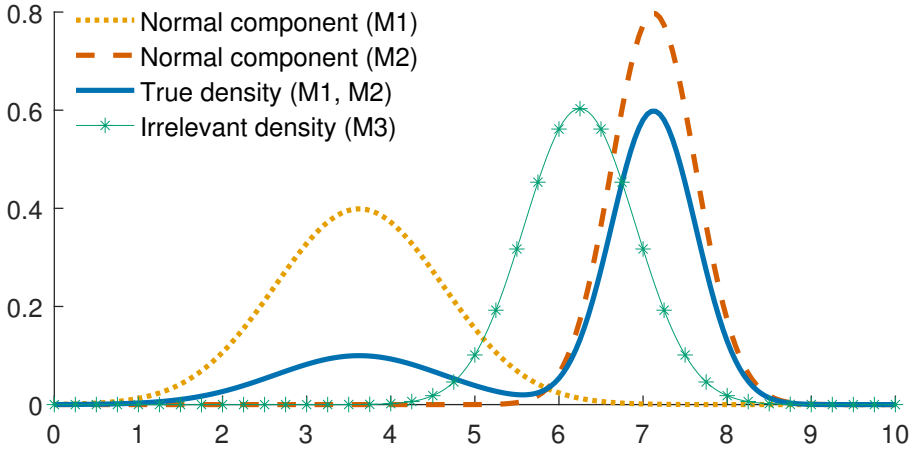
1.4.2 Monte Carlo set-up – DGP 2

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

[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.

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 (1.18). The mixture weights are $(w_1, w_2, w_3)' = (0.25, 0.75, 0)'$. Table 1.1 contains the models' parameters, while Figure 1.6 shows the corresponding predictive densities.

Figure 1.6: 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.

1.4.3 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

$$M1 : y_{t+1} = c_1 + \rho_1^{(1)} y_t + \rho_2^{(1)} y_{t-1} + \sigma_1 \nu_{t+1} \qquad \nu_{t+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0,1)\,, \tag{1.19}$$

$$M2 : y_{t+1} = c_2 + \rho_1^{(2)} y_t + \rho_2^{(2)} y_{t-1} + \sigma_2 \varepsilon_{t+1} \qquad \varepsilon_{t+1} \overset{\text{iid.}}{\sim} t_\nu^{\text{s}}\,, \tag{1.20}$$

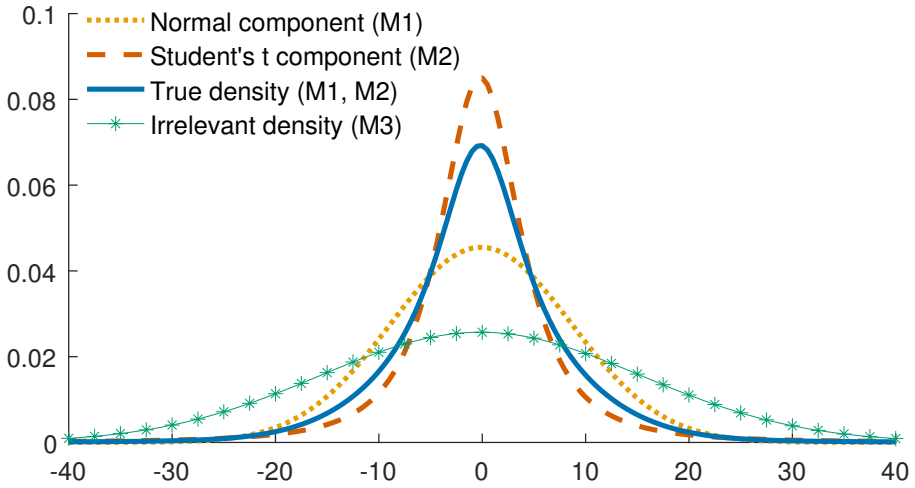
where t_ν^{s} stands for the standardized Student's t-distribution, with $\nu > 2$ degrees of freedom. The mixture weights are $(w_1, w_2)' = (0.4, 0.6)'$, and I added a normal AR(2) process to the model set, specified as

$$M3 : y_{t+1} = c_3 + \rho_1^{(3)} y_t + \rho_2^{(3)} y_{t-1} + \sigma_3 \eta_{t+1} \qquad \eta_{t+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0,1)\,, \tag{1.21}$$

[14] More details on the data can be found in Section 1.5.

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 1.1 contains the parameters of the models and Figure 1.7 presents the predictive densities.

Figure 1.7: 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 1.1: Simulation design

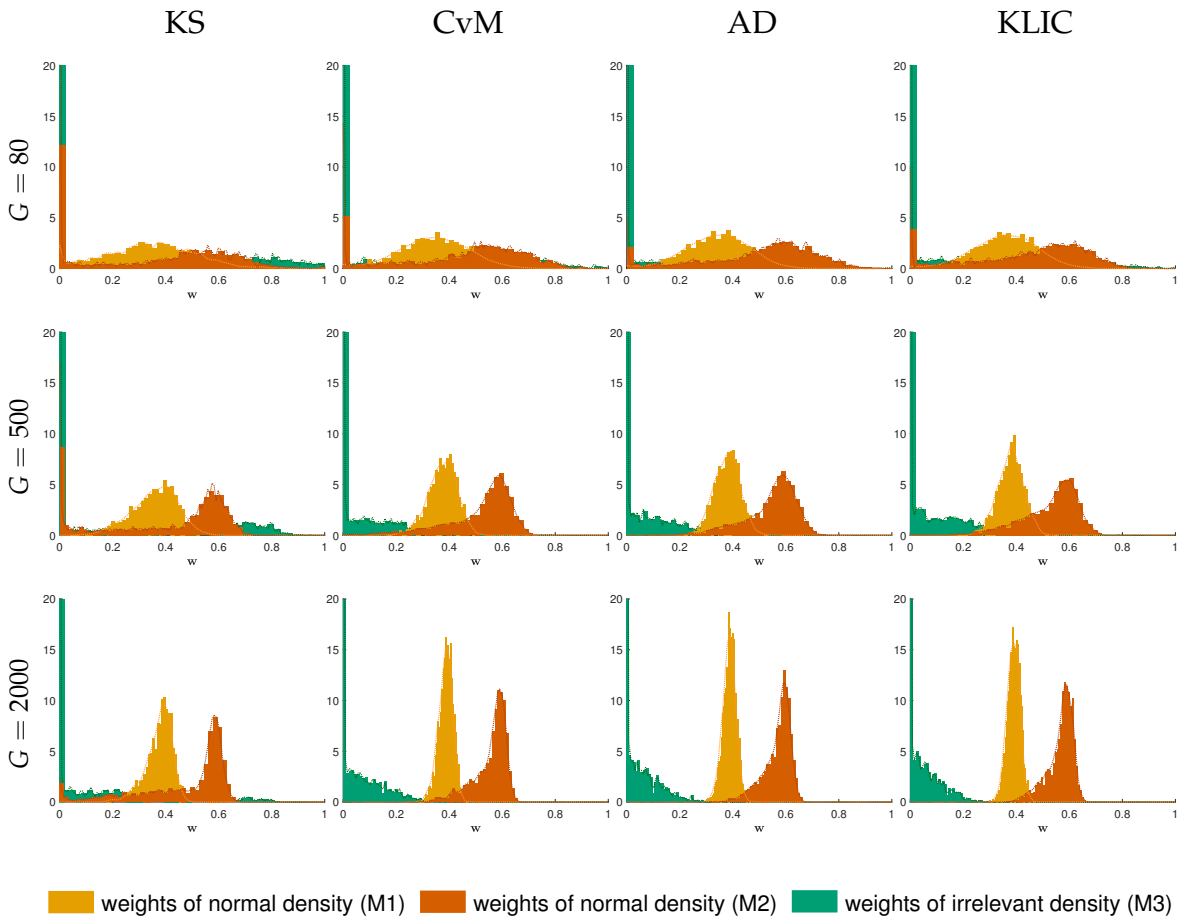
	Model	c	ρ_1	ρ_2	σ^2	ν	w_j
DGP1	M1	1	0.5	0	1	—	0.4
	M2	1	0.5	0	9	—	0.6
	M3	1	0.5	0	5.8	—	0
DGP2	M1	−2	0.9	0	1	—	0.25
	M2	1.5	0.9	0	0.25	—	0.75
	M3	0.63	0.9	0	0.44	—	0
DGP3	M1	−0.02	0.31	0.21	76.87	—	0.4
	M2	−0.11	0.24	0.32	350.32	2.10	0.6
	M3	−0.07	0.27	0.27	240.94	—	0

Note: For each DGP and each forecasting model (M1 – M3) the table lists the constant (c), the autoregressive parameters (ρ_1, ρ_2), and the variance parameter (σ^2) of the predictive distribution. M2 in DGP 3 is specified using a Student's t predictive distribution, with degrees of freedom parameter ν. For each DGP, the predictive distributions of M1 and M2 are weighted using the weights in the last column, w_j.

1.4.4 Monte Carlo results

Considering DGPs 1a and 1b first, in Figures 1.8 and 1.9 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 1.8: Monte Carlo results for DGP 1a, true parameter vector $w = (0.4, 0.6, 0)'$

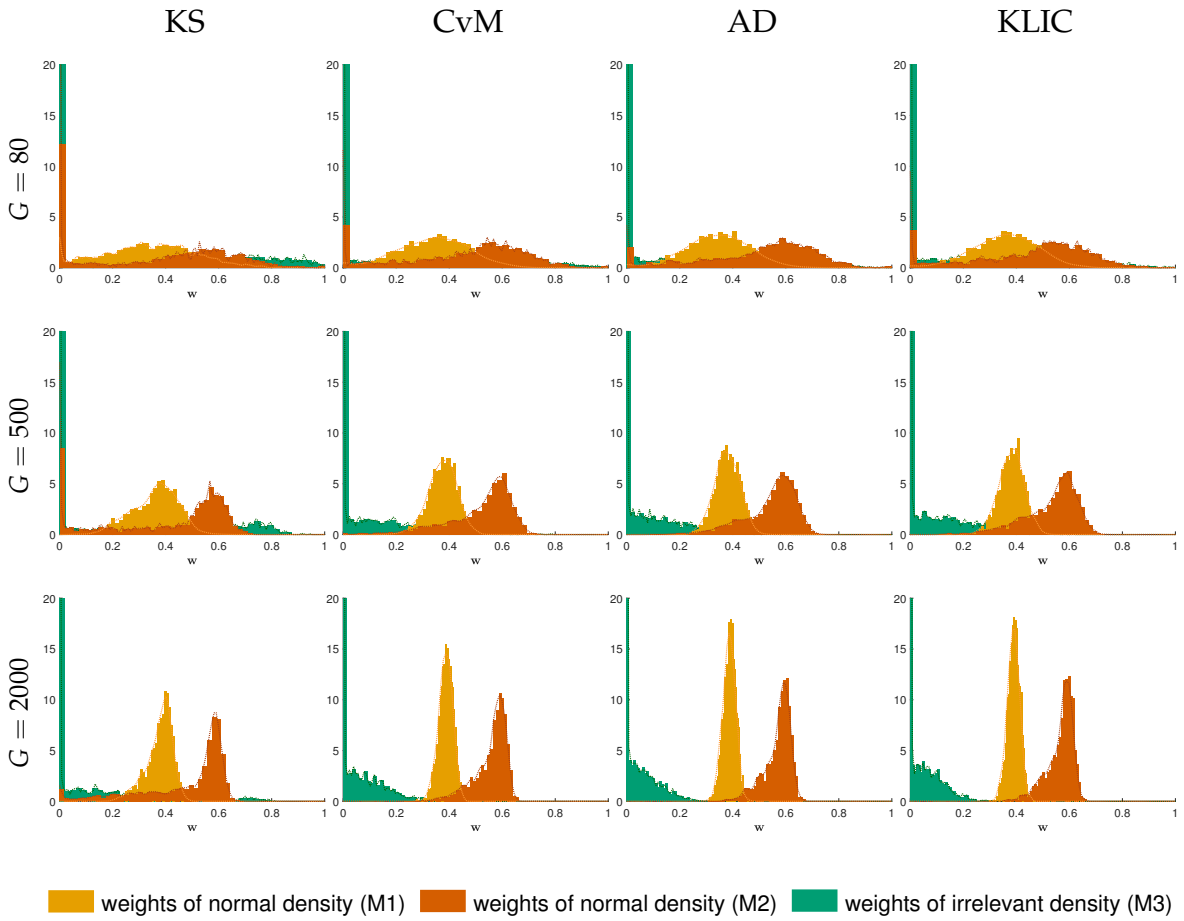


Note: G denotes the sample size. KS, CvM, AD and KLIC stand for the Kolmogorov–Smirnov-, the Cramer–von Mises-, the Anderson–Darling- and the KLIC-based estimators, respectively. Histograms and kernel density estimates are based on 2000 Monte Carlo replications.

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 1.9: 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 1.2 and 1.3 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 1.2: 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.

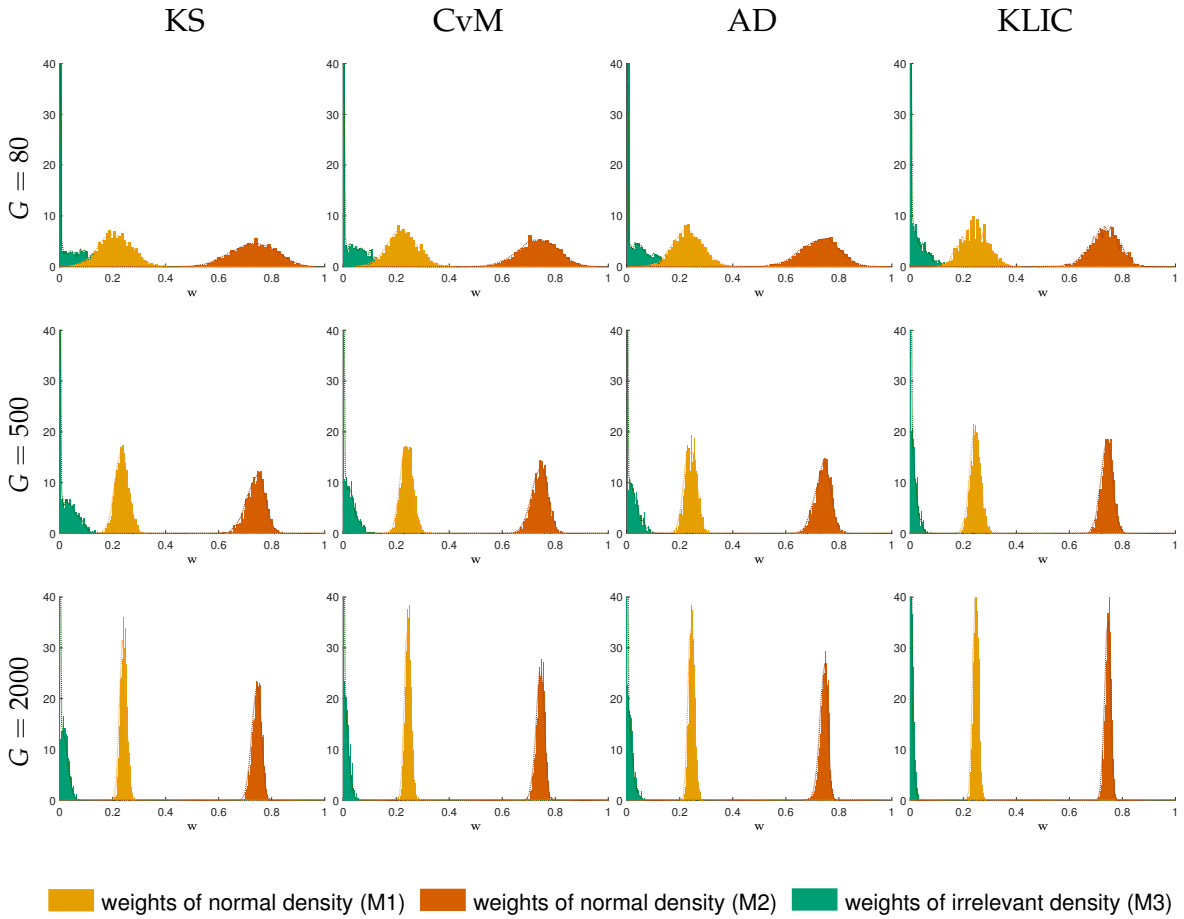
Table 1.3: DGP 1b, Monte Carlo summary statistics for different sample sizes G and objective functions $K_G(w), C_G(w), A_G(w)$ and $\text{KLIC}_G(w)$

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 1.10 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 1.10: 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 1.4 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 1.4: DGP 2, Monte Carlo summary statistics for different sample sizes G and objective functions $K_G(w)$, $C_G(w)$, $A_G(w)$ and $\text{KLIC}_G(w)$

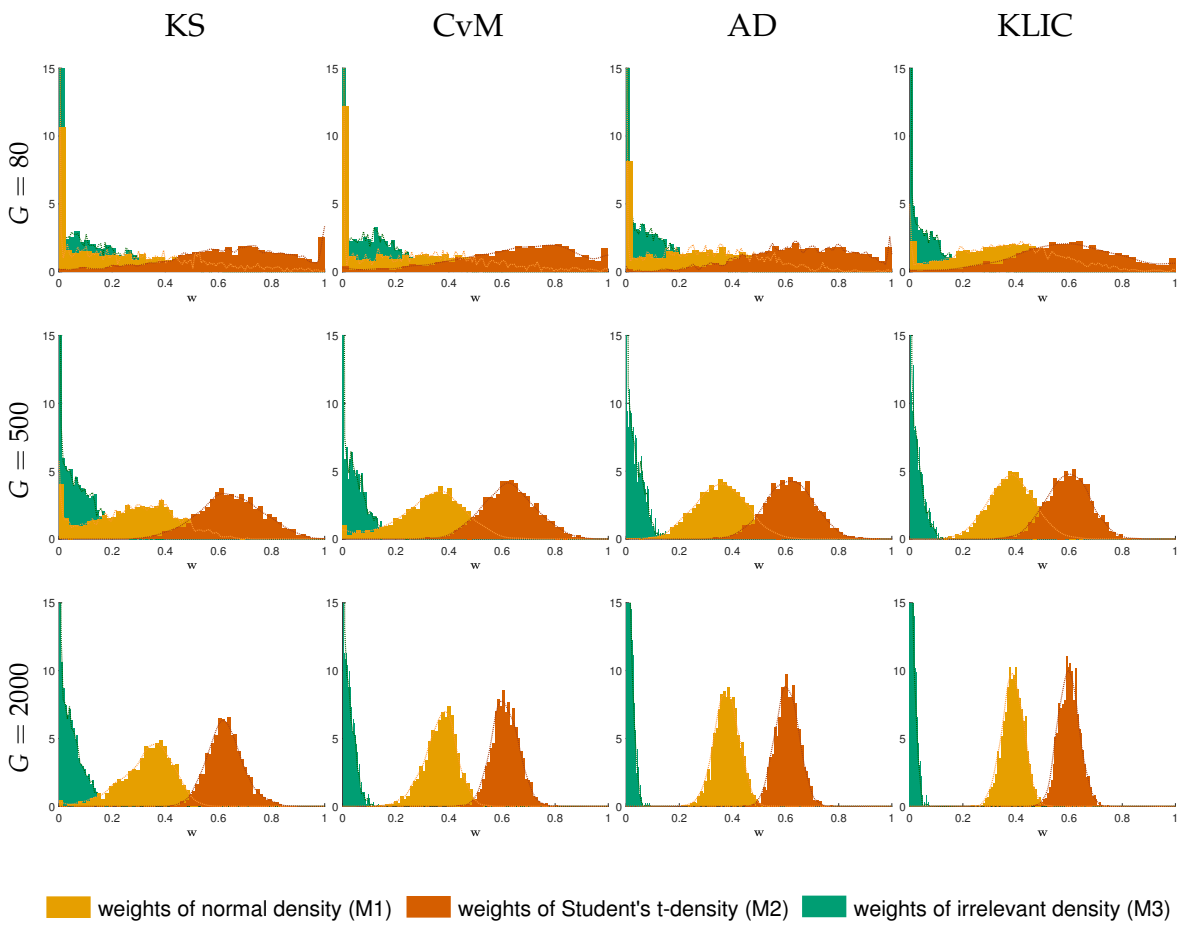
Sample size	Statistic	KS			CvM			AD			KLIC		
$G = 80$	Bias	−0.04	−0.02	0.05	−0.02	−0.01	0.04	−0.02	−0.02	0.04	−0.00	−0.02	0.02
	Var	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00
	MSE	0.01	0.01	0.01	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00
$G = 200$	Bias	−0.02	−0.01	0.04	−0.01	−0.01	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 1.11 that the AD and KLIC estimators dominate the other two, with the latter delivering slightly less dispersed estimates. Table 1.5 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.

Figure 1.11: 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.

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 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 1.5: 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

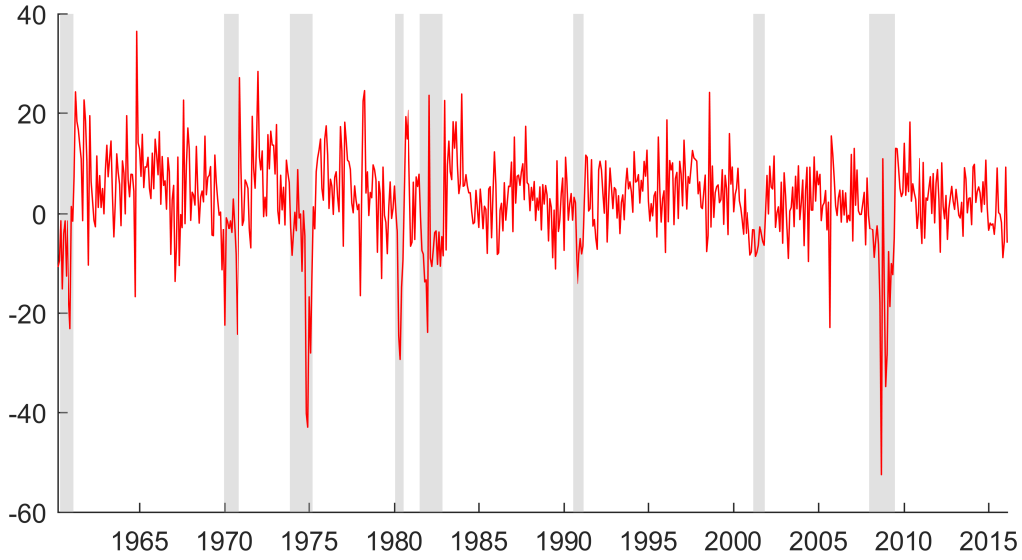
Note: In the four main columns with headers KS, CvM, AD and KLIC, the table shows the estimates of the bias, variance (Var) and mean squared error (MSE) for each of the components of the weight vector w. True weights: $w = (0.4, 0.6, 0)'$. Statistics are based on 2000 Monte Carlo replications.

1.5 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 1.12 and 1.13, respectively. As we can see in Figure 1.13, 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.

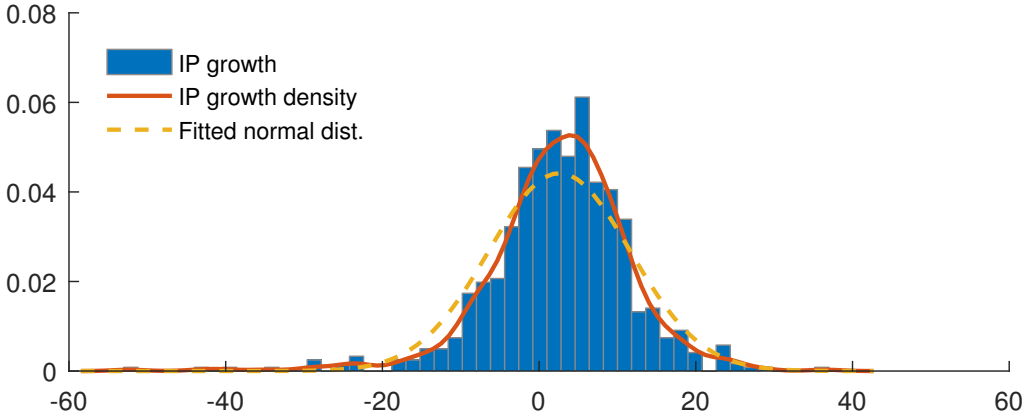
While the non-Gaussian *un*conditional distribution does not necessarily imply non-Gaussian conditional distribution, it is worth investigating how the proposed data-dependent density forecast combination procedures — which are capable of generating a variety of forecast densities — perform in an empirical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

Figure 1.12: Annualized US IP growth between March 1960 and February 2016



Note: Shaded areas are NBER recession periods.

Figure 1.13: Normalized histogram of annualized US IP growth between March 1960 and February 2016



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

$$y_{\tau+1} = c + \sum_{j=0}^{1} \beta_j y_{\tau-j} + \sum_{j=0}^{1} \gamma_j x_{\tau-j} + \sqrt{\sigma^2}\varepsilon_{\tau+1} \qquad \varepsilon_{\tau+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0,1)\,, \tag{1.22}$$

where y_τ is annualized US IP growth in month τ, that is $y_\tau \equiv 1200\Delta\log(\text{IP}_\tau)$ where Δ is the first difference operator, c is a constant term, β_js are coefficients of

the autoregressive terms while γ_js 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 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_τ and x_τ 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_τ 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 1.14.

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 1.14a shows. It is also remarkable that unlike in earlier recessions, housing permits did not plummet during the 2001 recession. Figure 1.14d 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.

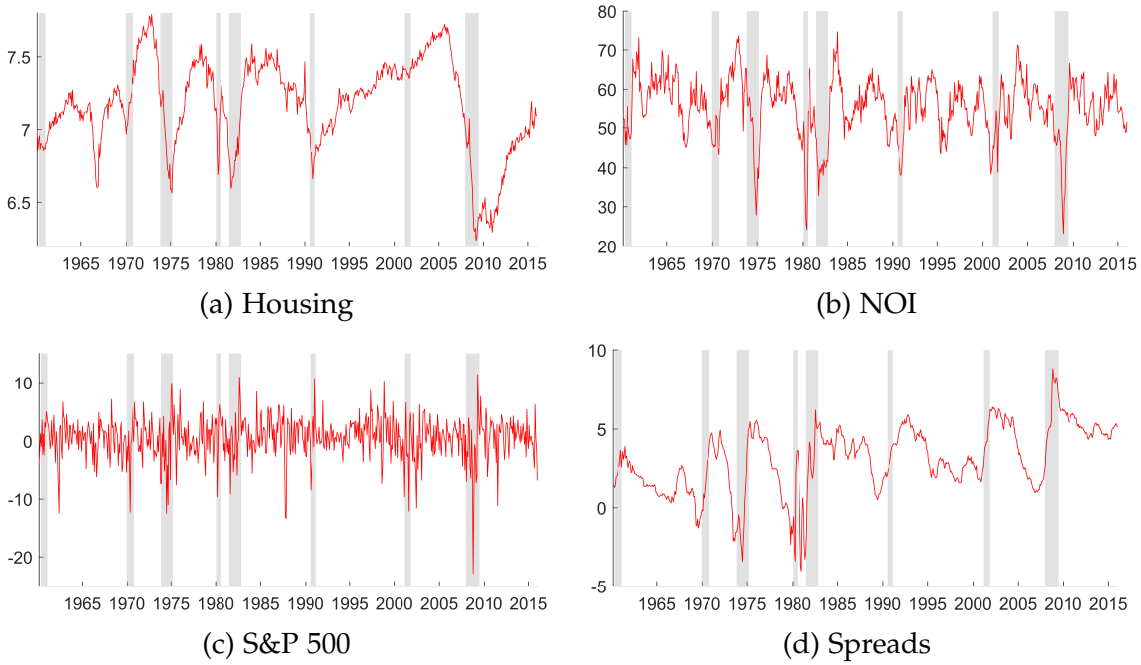
All models are estimated using Maximum Likelihood in rolling windows of $R = 120$ months, with forecast origins f and target dates $f + h$ ranging from February 1985 to January 2016 and March 1985 to February 2016, respectively.

To illustrate the estimation procedure, consider the first forecast origin f, corresponding to February 1985. The first window to estimate the models of Equation (1.22) contains data indexed by $\tau = \{\text{February 1960}, \ldots, \text{January 1970}\}$, which delivers out-of-sample (with respect to *this* estimation sample) predictive

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

[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 1.14: 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.

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 1.4, 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 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

$$\text{BIC}_m \equiv -2 \sum_{t=f-R}^{f-1} \log \ell_m(y_{t+1}|z_t^m; \widehat{\theta}_m) + k_m \log(R)\,, \tag{1.23}$$

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

[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.

values, where model m's weight is given by

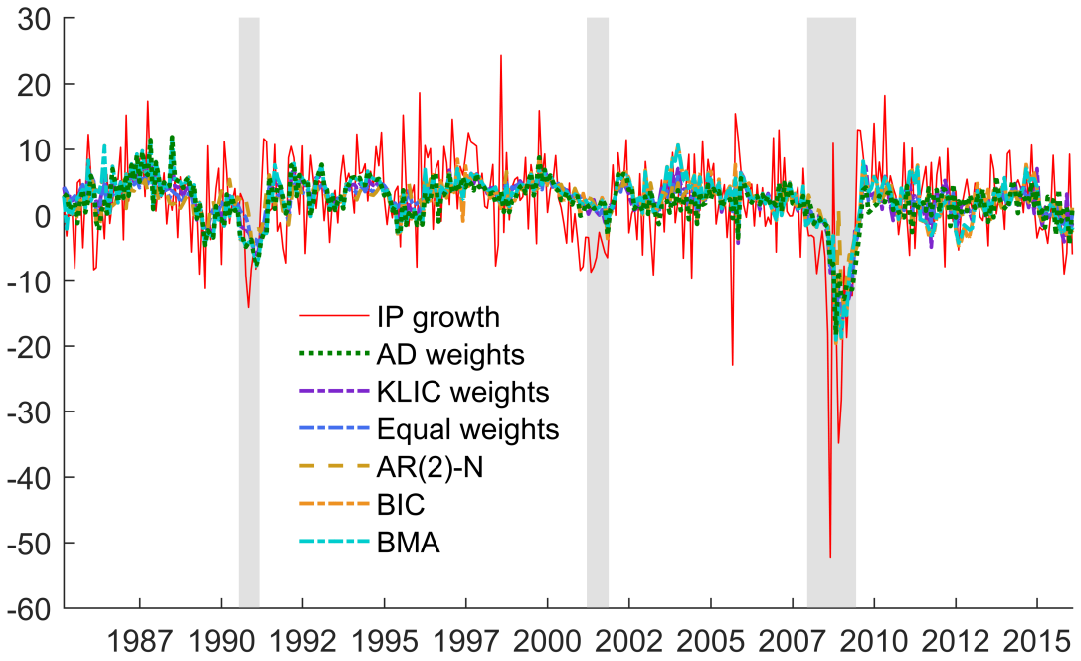
$$w_m = \frac{\exp(-0.5BIC_m)}{\sum_{i=1}^{5}\exp(-0.5BIC_i)}. \tag{1.24}$$

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.

1.5.2 Results: point forecasts

Figure 1.15 shows the point forecasts (conditional means) of all the forecast combination schemes between March 1985 and February 2016.

Figure 1.15: Point forecasts of US industrial production growth



Note: Shaded areas are NBER recession periods.

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 1.6, using the Diebold-Mariano test (Diebold and Mariano, 1995) 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.

Table 1.6: 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 $\lfloor 0.75P^{1/3} \rfloor = 5$.

As Table 1.6 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 *density* forecasts. Whether it lives up to this expectation is investigated in the next section.

1.5.3 Results: density forecasts

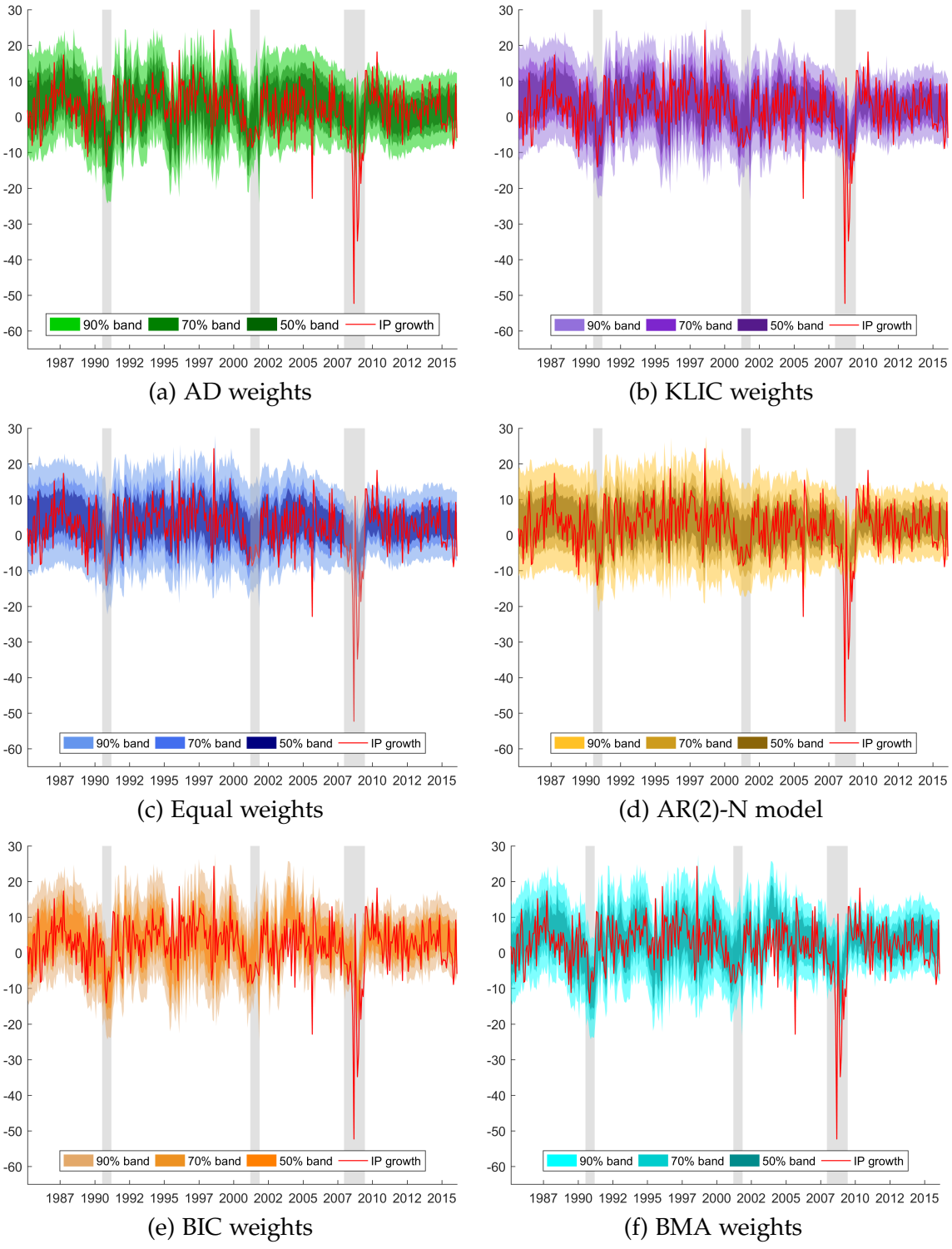
Next, let us consider the density forecasts obtained by the six competing methods. First, in Figure 1.16 we can see central, equal tailed 90%, 70% and 50% bands of the one-step-ahead combined predictive densities at each forecast target date, ranging from March 1985 to February 2016. Visual inspection 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 1.17 we can see the histograms of the PITs associated with the six forecasting methods. By comparing Figure 1.17a and Figure 1.17b, 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 1.17c and Figure 1.17d show, the equal weights scheme and the AR(2)-N model display this behavior in a more pronounced way. Figure 1.17e and Figure 1.17f suggest that BIC-based model selection and BMA weights provide better density forecasts than the previous two competitors.

Figure 1.18 shows the empirical CDFs of the PITs and the ideal, uniform CDF corresponding to the 45 degree line. As we can see, Figure 1.18 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 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.

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 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 1.7 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

Figure 1.16: Equal-tailed forecast bands of one-month-ahead US IP growth



(a) AD weights (b) KLIC weights

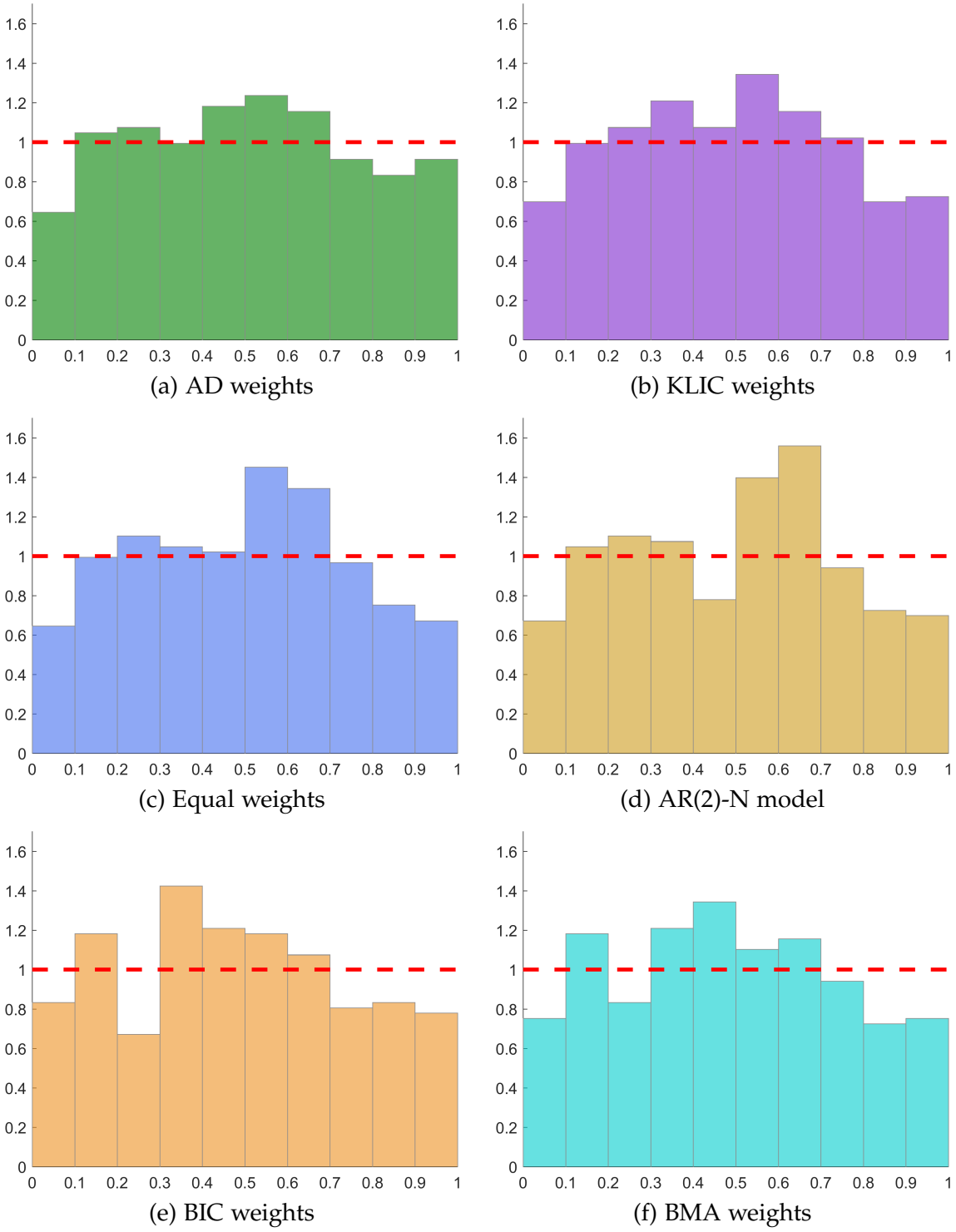
(c) Equal weights (d) AR(2)-N model

(e) BIC weights (f) BMA weights

Note: Shaded areas are NBER recession periods.

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

Figure 1.17: Normalized histograms of PITs



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

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.

Figure 1.18: Empirical CDF of PITs

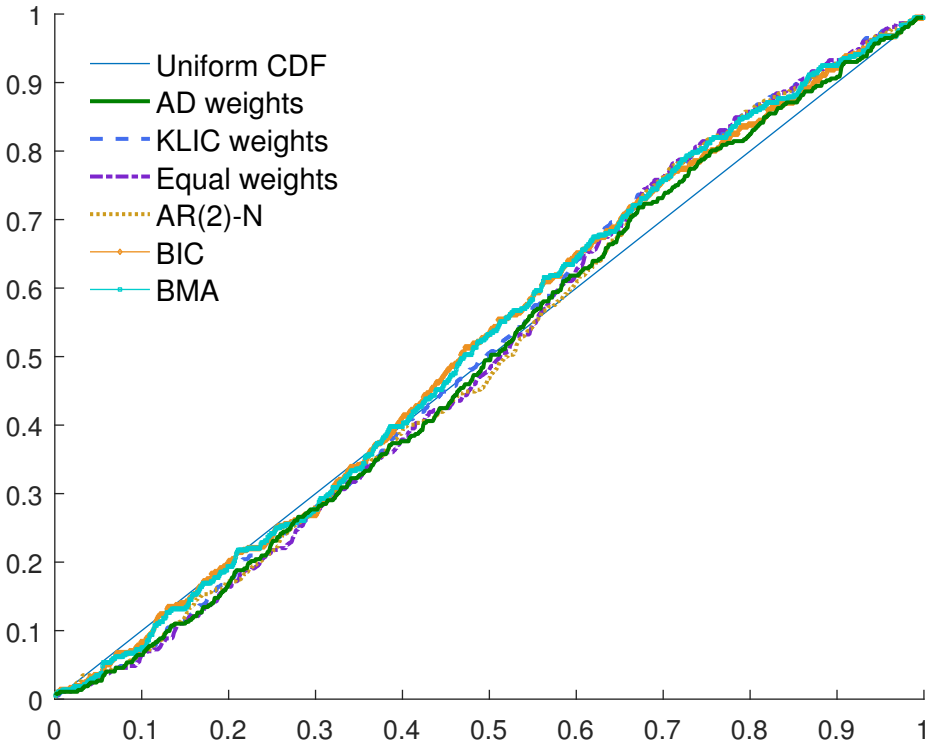


Table 1.7: Rossi and Sekhposyan (2016) test on correct specification of conditional predictive densities

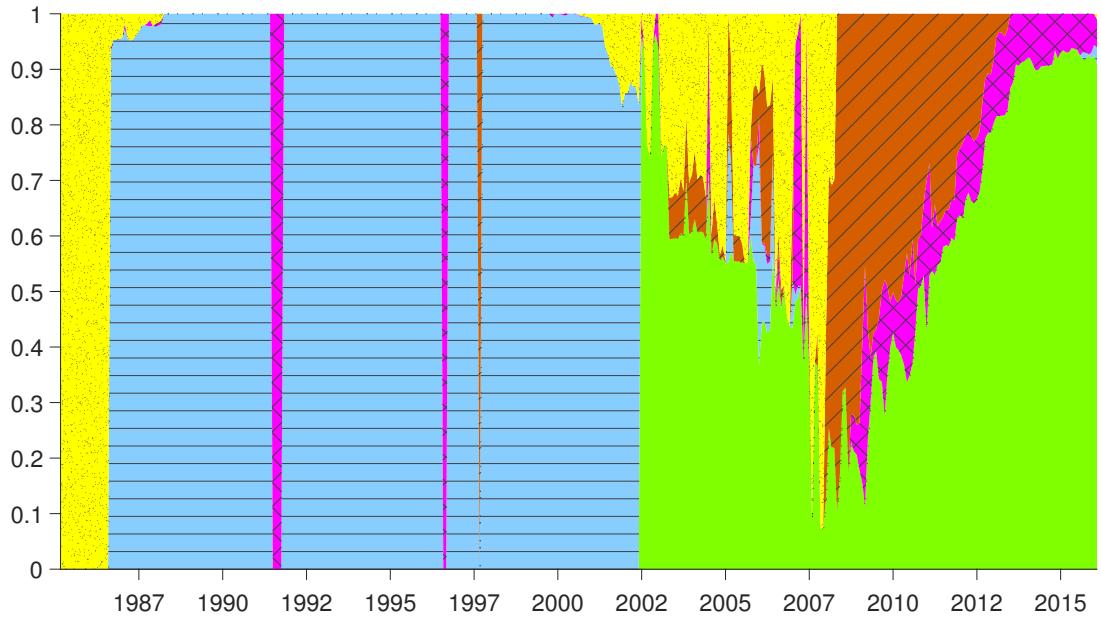
Models	Kolmogorov–Smirnov	Cramer–von Mises
AD weights	0.90 (0.38)	0.24 (0.22)
KLIC weights	1.28 (0.08)	0.42 (0.06)
Equal weights	1.39 (0.05)	0.50 (0.04)
AR(2)-N	1.31 (0.08)	0.40 (0.09)
BIC	1.16 (0.17)	0.32 (0.16)
BMA	1.28 (0.10)	0.38 (0.11)

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.

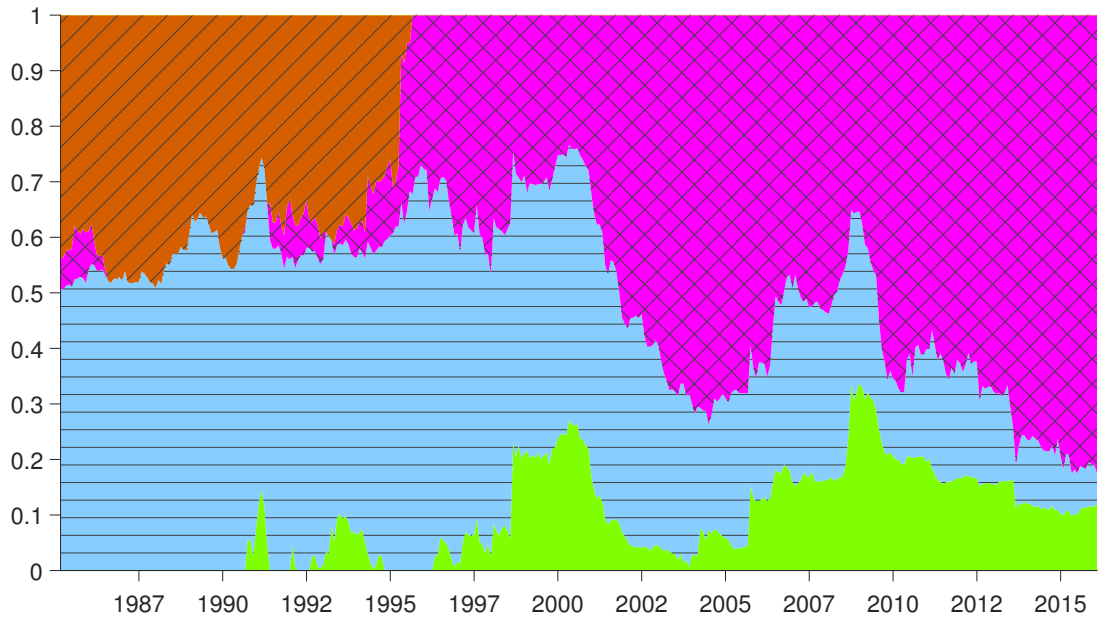
This discussion has so far focused on evaluating the various density forecasts of US industrial production. However, it is also interesting how the combination weights of each model evolved over the out-of-sample period (March 1985 to February 2016), which is shown in Figures 1.19 to 1.21.

In Figure 1.19a, we can see that using the Anderson–Darling weights, apart

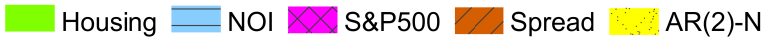
Figure 1.19: Time-variation of estimated AD and KLIC weights, area plots



(a) Estimated Anderson–Darling weights



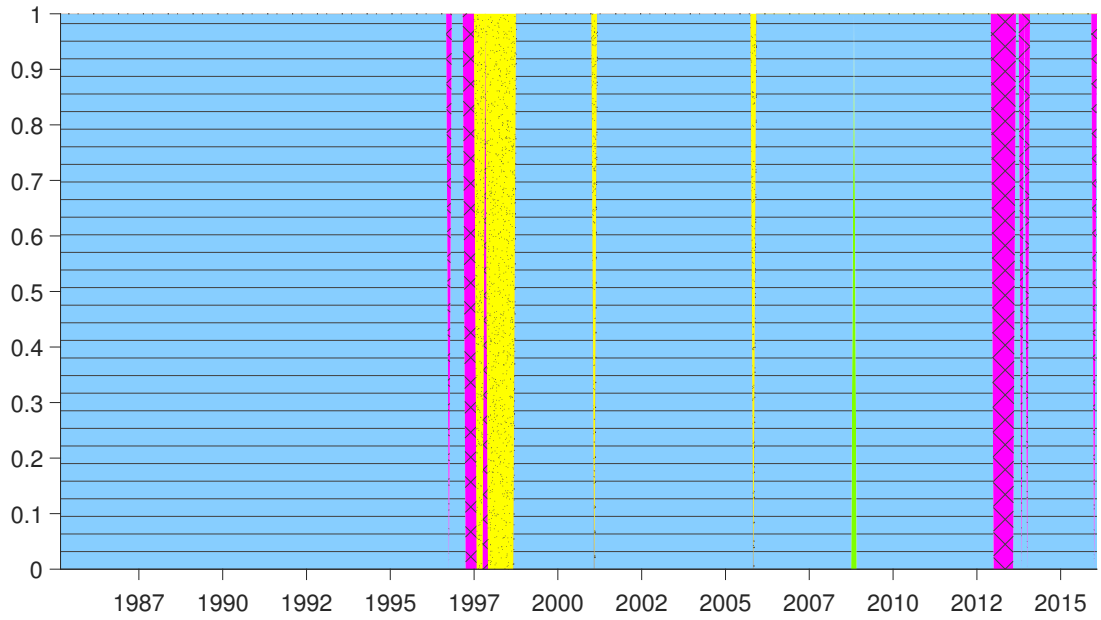
(b) Estimated KLIC weights



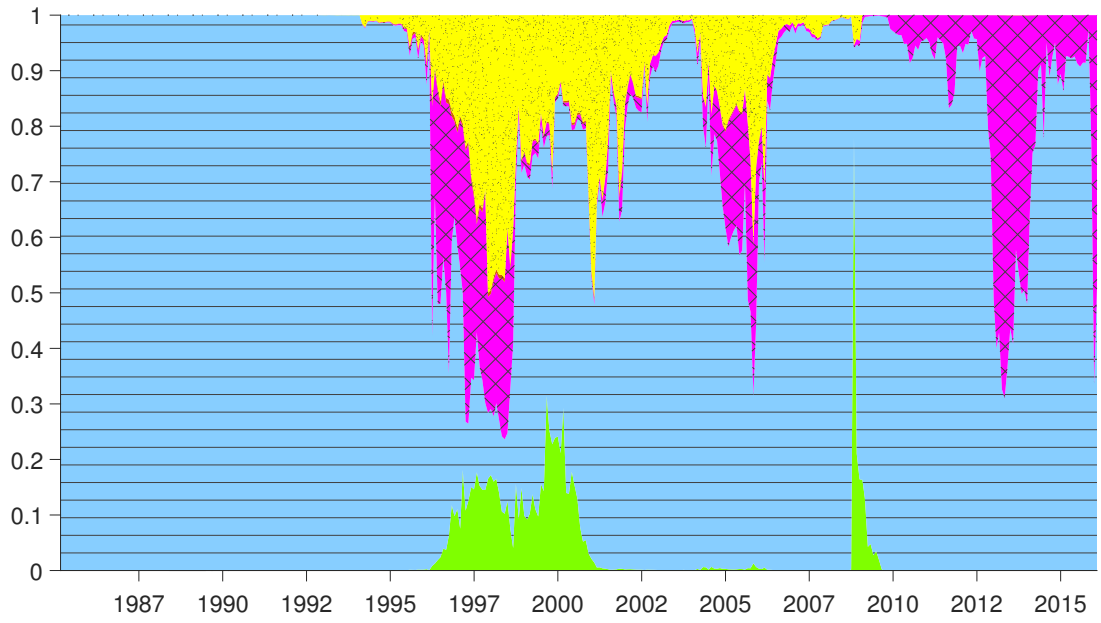
Note: The out-of-sample period starts in March 1985 and ends in February 2016, with a total number of $P = 372$ months. Housing stands for Housing Permits, NOI stands for ISM: New Orders Index, S&P 500 is the S&P 500 stock index returns while Spread is Moody's Baa Corporate Bond Yield minus Fed funds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

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,

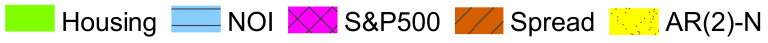
Figure 1.20: Time-variation of estimated BIC and BMA weights, area plots



(a) Estimated BIC weights



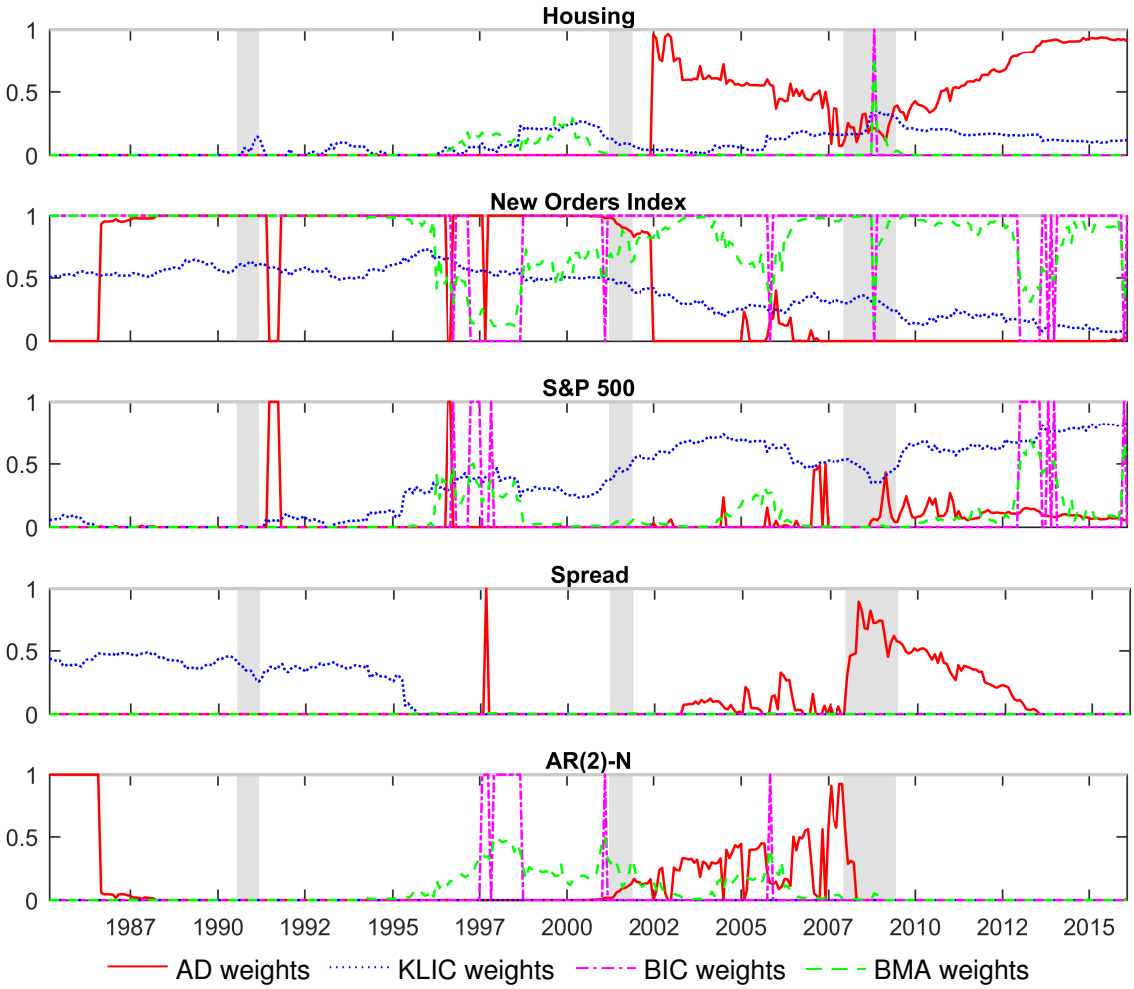
(b) Estimated BMA weights



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.

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 the S&P 500 received large

Figure 1.21: 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.

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 1.21. 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 1.19b 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 analysis and supporting evidence, see Appendix E.

Figure 1.20a and Figure 1.20b 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 1.21 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 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.

[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.

1.6 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.

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^{ρ} 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 $\Phi_{t+h}^{C}(y_{t+h}|\mathcal{I}_{t-R+1}^{t}) = r$, 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

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

$$\leq \left| \frac{1}{G} \sum_{t=1}^{G} \xi_{t+h}(w, r) - \frac{1}{G} \sum_{t=1}^{G} \xi_{t+h}(w_k, r_k) \right| +$$

$$\left| \frac{1}{G} \sum_{t=1}^{G} \xi_{t+h}(w_k, r_k) - \frac{1}{G} \sum_{t=1}^{G} \lambda_{t+h}(w_k, r_k) \right| + \tag{A.3}$$

$$\left| \frac{1}{G} \sum_{t=1}^{G} \lambda_{t+h}(w_k, r_k) - \frac{1}{G} \sum_{t=1}^{G} \lambda_{t+h}(w, r) \right|$$

$$\leq \frac{1}{G} \sum_{t=1}^{G} |\xi_{t+h}(w, r) - \xi_{t+h}(w_k, r_k)| + \frac{1}{G} \sum_{t=1}^{G} |\xi_{t+h}(w_k, r_k) - \lambda_{t+h}(w_k, r_k)| + \tag{A.4}$$

$$\frac{1}{G} \sum_{t=1}^{G} |\lambda_{t+h}(w_k, r_k) - \lambda_{t+h}(w, r)|$$

$$\leq \left[\frac{1}{G} \sum_{t=1}^{G} u_{t+h}(w_k, r_k, \bar{d}_C(w_k, r_k)) - \mu_{t+h,k} \right] +$$

$$\frac{1}{G} \sum_{t=1}^{G} \mu_{t+h,k} + \frac{1}{G} \sum_{t=1}^{G} |\xi_{t+h}(w_k, r_k) - \lambda_{t+h}(w_k, r_k)| + \tag{A.5}$$

$$\frac{1}{G} \sum_{t=1}^{G} |\lambda_{t+h}(w_k, r_k) - \lambda_{t+h}(w, r)| \ ,$$

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 ε almost surely, thus the whole expression is less than or equal to 4ε almost surely (the second and the fourth terms each are less than or equal to ε by construction).[19] Furthermore, if $G > \max_{k=1,\ldots,K} G_k(\varepsilon)$, then we have

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

[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.

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

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

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

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

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:

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

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 (1.16), 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 (1.16), 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 (1.13). 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 μ is a 3×1 vector of zeros and the (i,j)th element of the covariance matrix Σ 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 Σ, specifically $\widetilde{\alpha} = \gamma_1/\gamma_0$ and $\widetilde{\sigma}^2 = \left(1 - \widetilde{\alpha}^2\right)\gamma_0$.

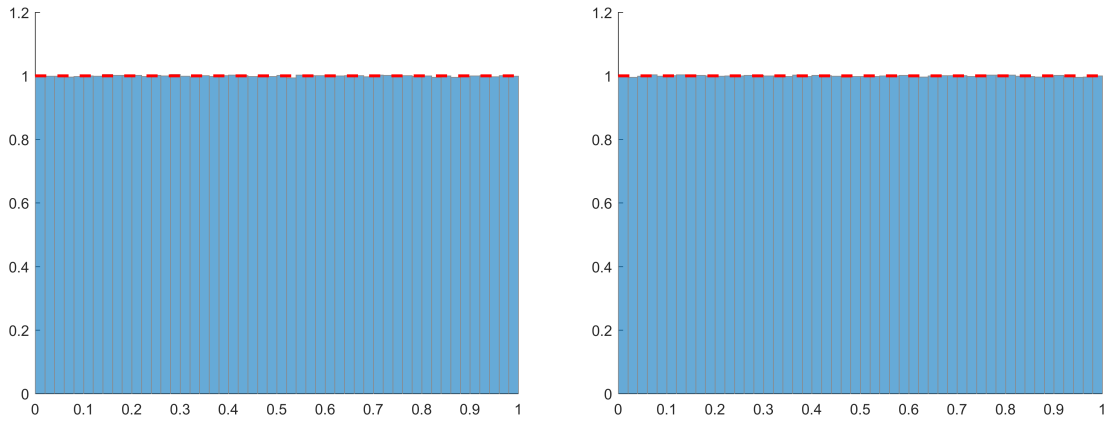
Suppose that the researcher conditions his or her forecast on only one lag of the dependent variable, $(R = 1, \mathcal{I}^t_{t-R+1} = y_t)$ but still maintains the normality assumption, implying the predictive density

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

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.

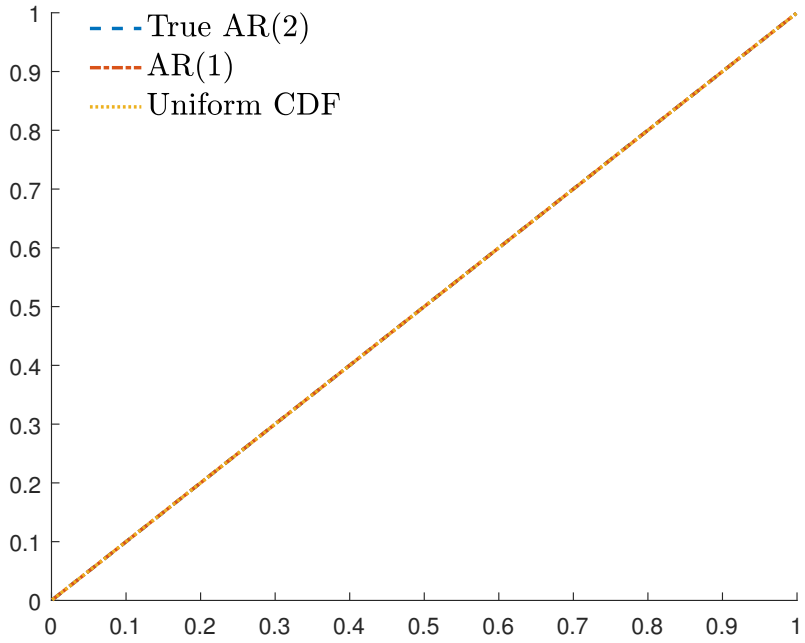
Figure B.1: Normalized histograms of PITs



(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

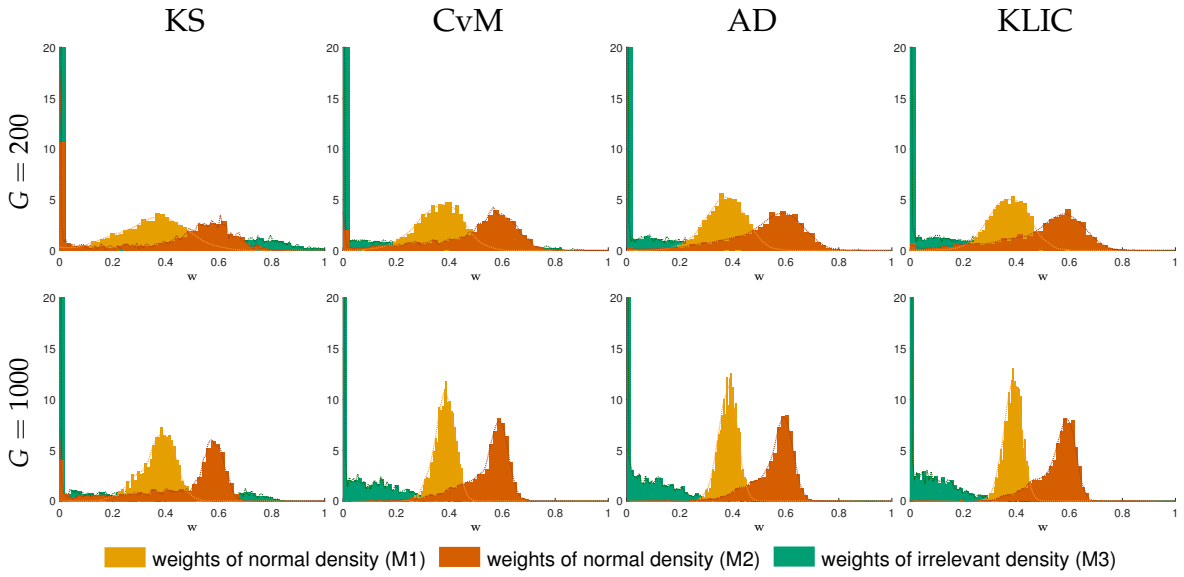
Given that the non-linear extremum estimators proposed in the present paper do not have closed form solutions, I need to use a numerical optimizer. The optimizer that operates on the unit simplex is MATLAB's built-in `fminsearch` algorithm. This is an unconstrained derivative-free optimizer, and I transformed 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 1.4.4 to preserve space. Furthermore, a number of additional DGPs are used to illustrate the estimators' performance.

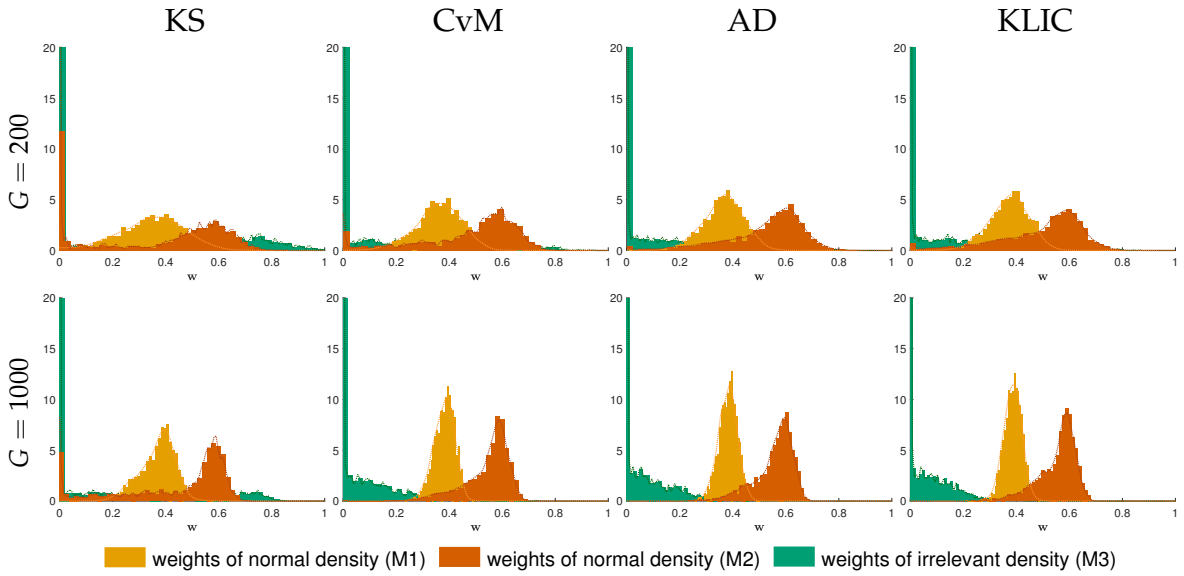
Additional figures – DGPs 1a, 1b, 2 and 3

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



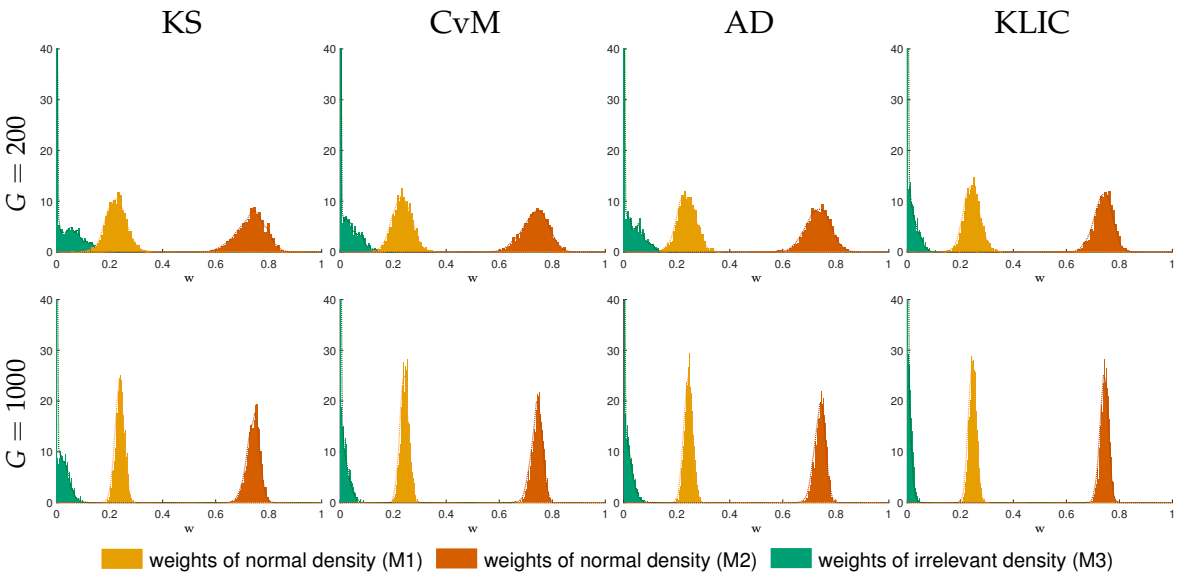
Note: G denotes the sample size. KS, CvM, AD and KLIC stand for the Kolmogorov–Smirnov-, the Cramer–von Mises-, the Anderson–Darling- and the KLIC-based estimators, respectively. Histograms and kernel density estimates based on 2000 Monte Carlo replications.

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



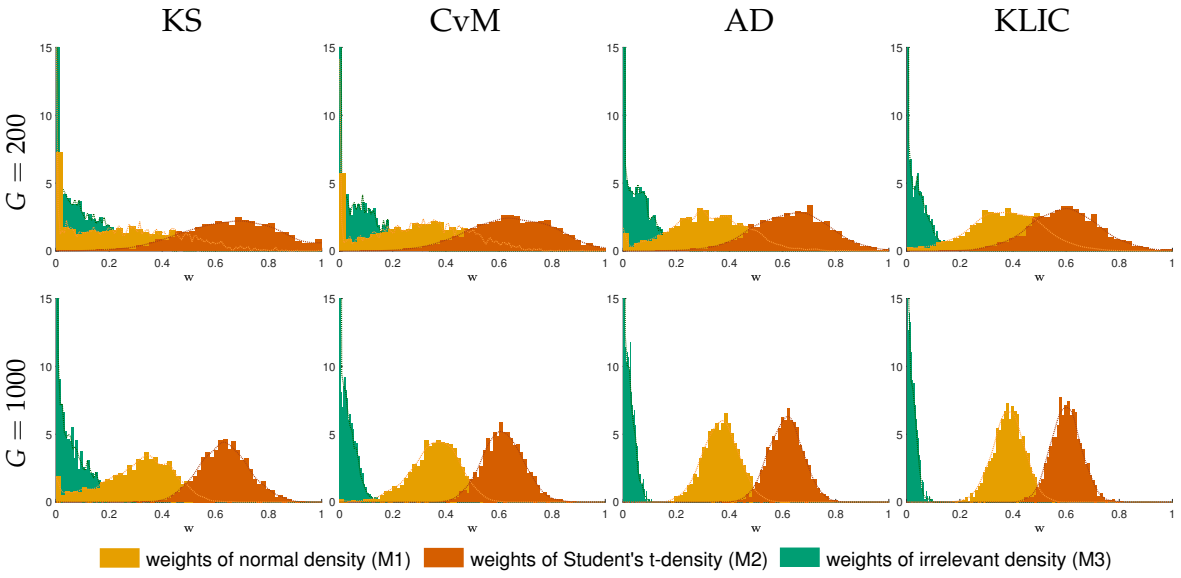
Note: G denotes the sample size. KS, CvM, AD and KLIC stand for the Kolmogorov–Smirnov-, the Cramer–von Mises-, the Anderson–Darling- and the KLIC-based estimators, respectively. Histograms and kernel density estimates are based on 2000 Monte Carlo replications.

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 $w = (0.4, 0.6, 0)'$

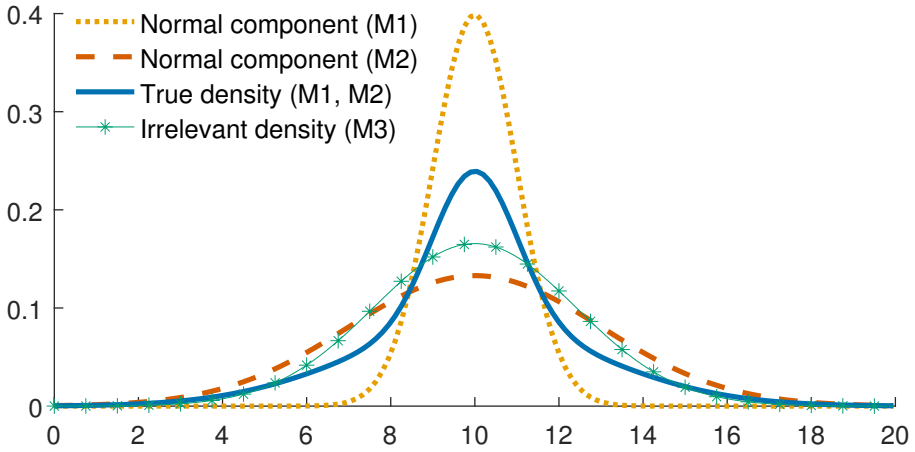


Note: G denotes the sample size. KS, CvM, AD and KLIC stand for the Kolmogorov–Smirnov-, the Cramer–von Mises-, the Anderson–Darling- and the KLIC-based estimators, respectively. Histograms and kernel density estimates are based on 2000 Monte Carlo replications.

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 1.4.1, 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.

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:

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

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

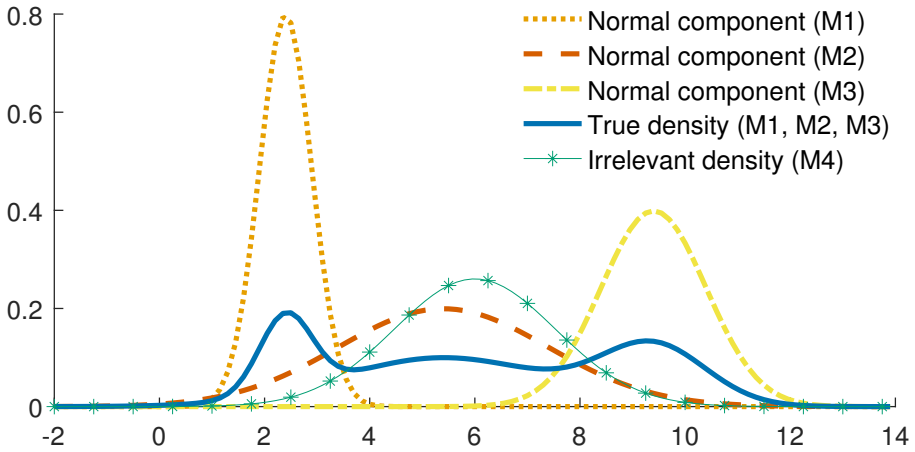
$$M3 : y_{t+1} = c_3 + 0.9y_t + \lambda_{t+1} \qquad \lambda_{t+1} \overset{\text{iid.}}{\sim} \mathcal{N}(0, \sigma_3^2)\,, \tag{D.3}$$

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.

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:

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

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

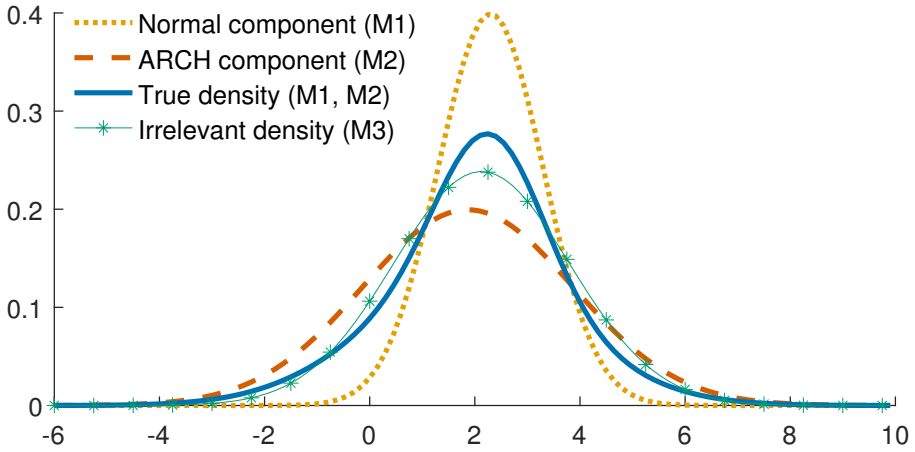
with intercepts $c_1 = c_2 = 1$, autoregressive coefficients $\rho_1 = 0.4, \rho_2 = 0.6$ variance $\sigma_1^2 = 1$, ARCH coefficients $\alpha_0 = 2, \alpha_1 = 0.5$ and mixture weights $(w_1, w_2)' = (0.4, 0.6)'$. In the case of M2, the ARCH specification implies that the expected value of $\sigma_{2,t}^2$ 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.

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:

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

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

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

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

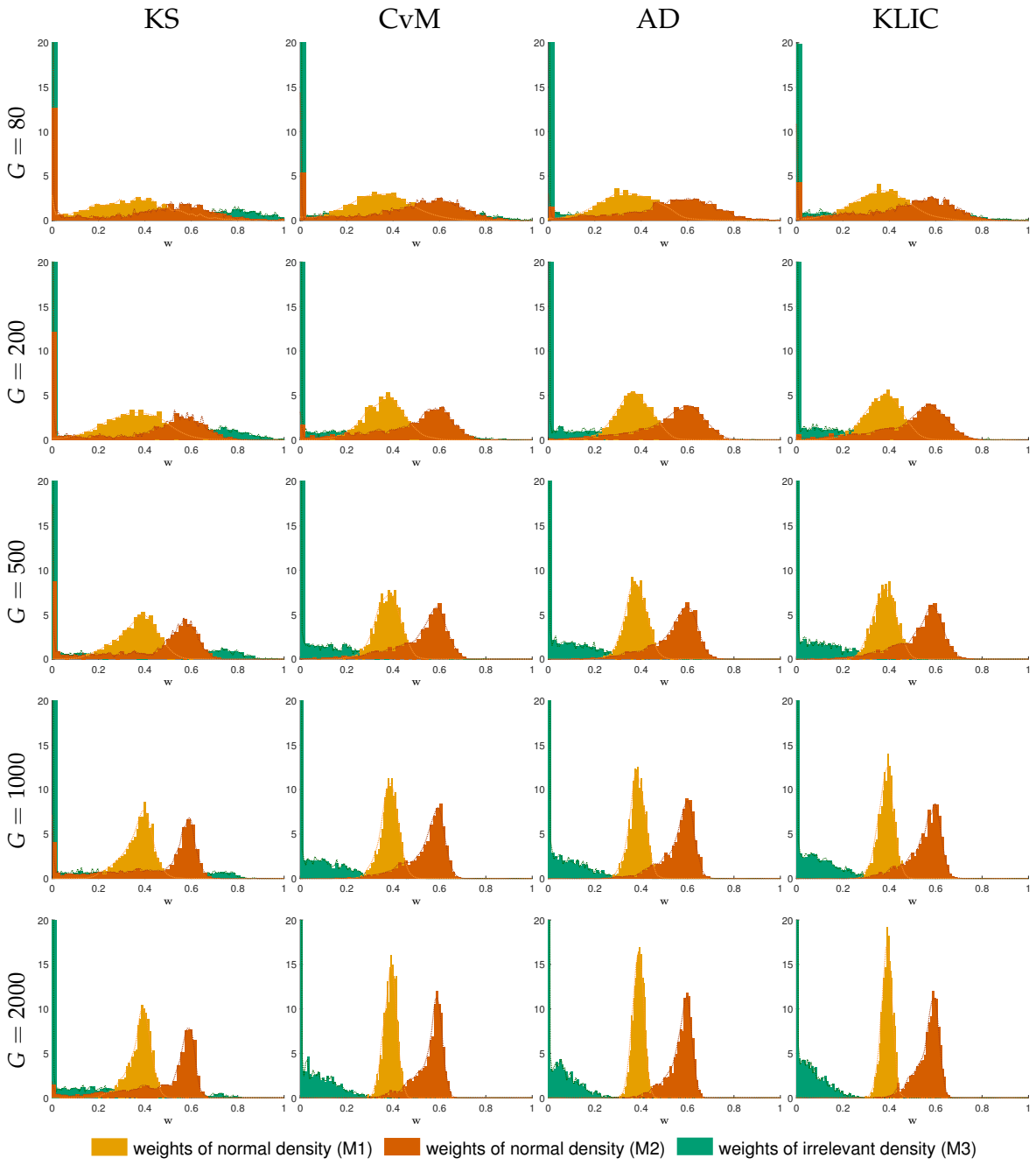
As Table D.1 and Figure D.8 show, increasing the autoregressive coefficient from $\rho = 0.5$ to $\rho = 0.9$ in DGP 1c does not affect the performance of any of the estimators.

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		
$G = 80$	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
	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
$G = 200$	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
	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
$G = 500$	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
	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
$G = 1000$	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
	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
$G = 2000$	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	0.01	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.

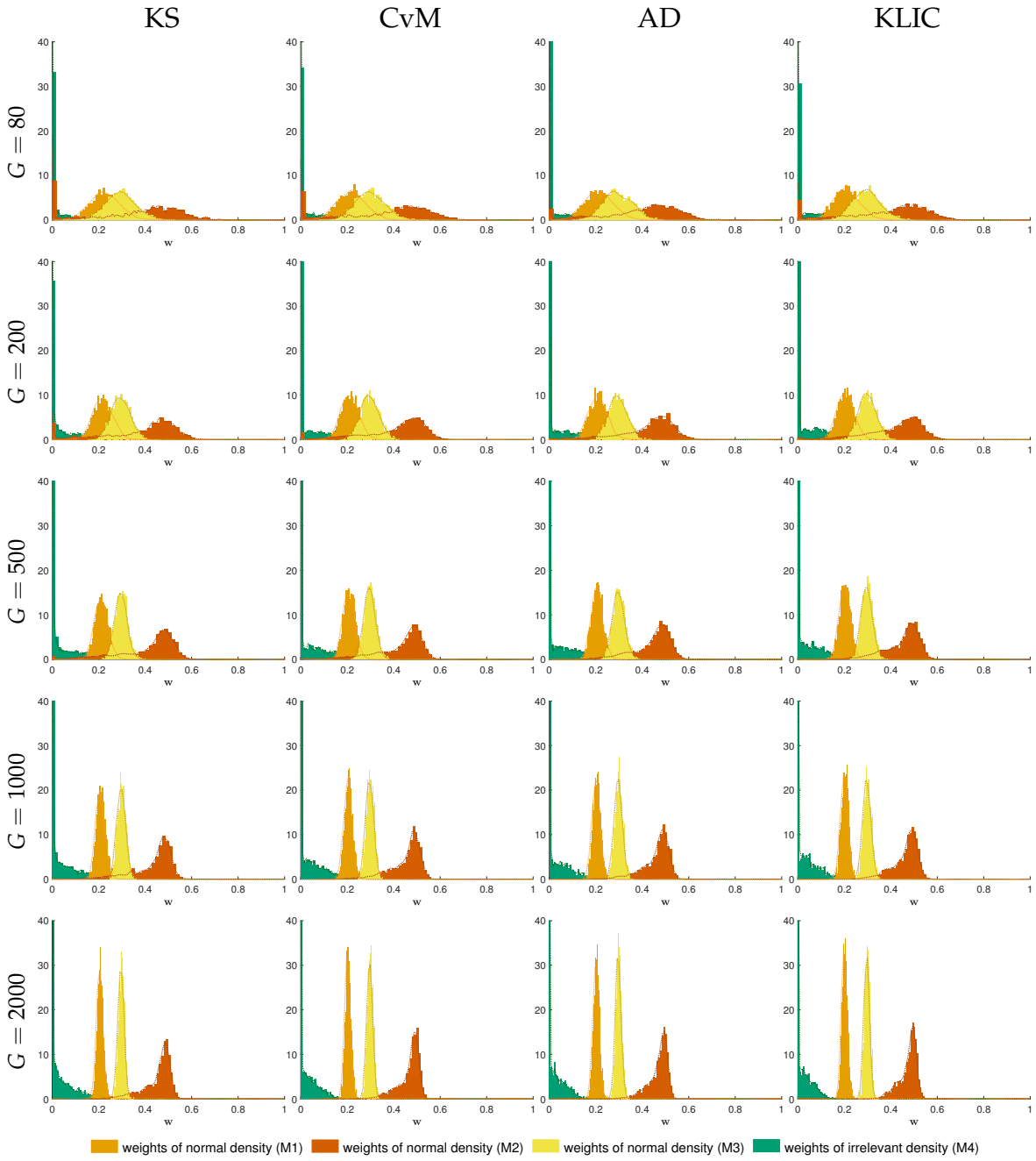
Figure D.8: Monte Carlo results for DGP 1c, true parameter vector $w = (0.4, 0.6, 0)'$



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.

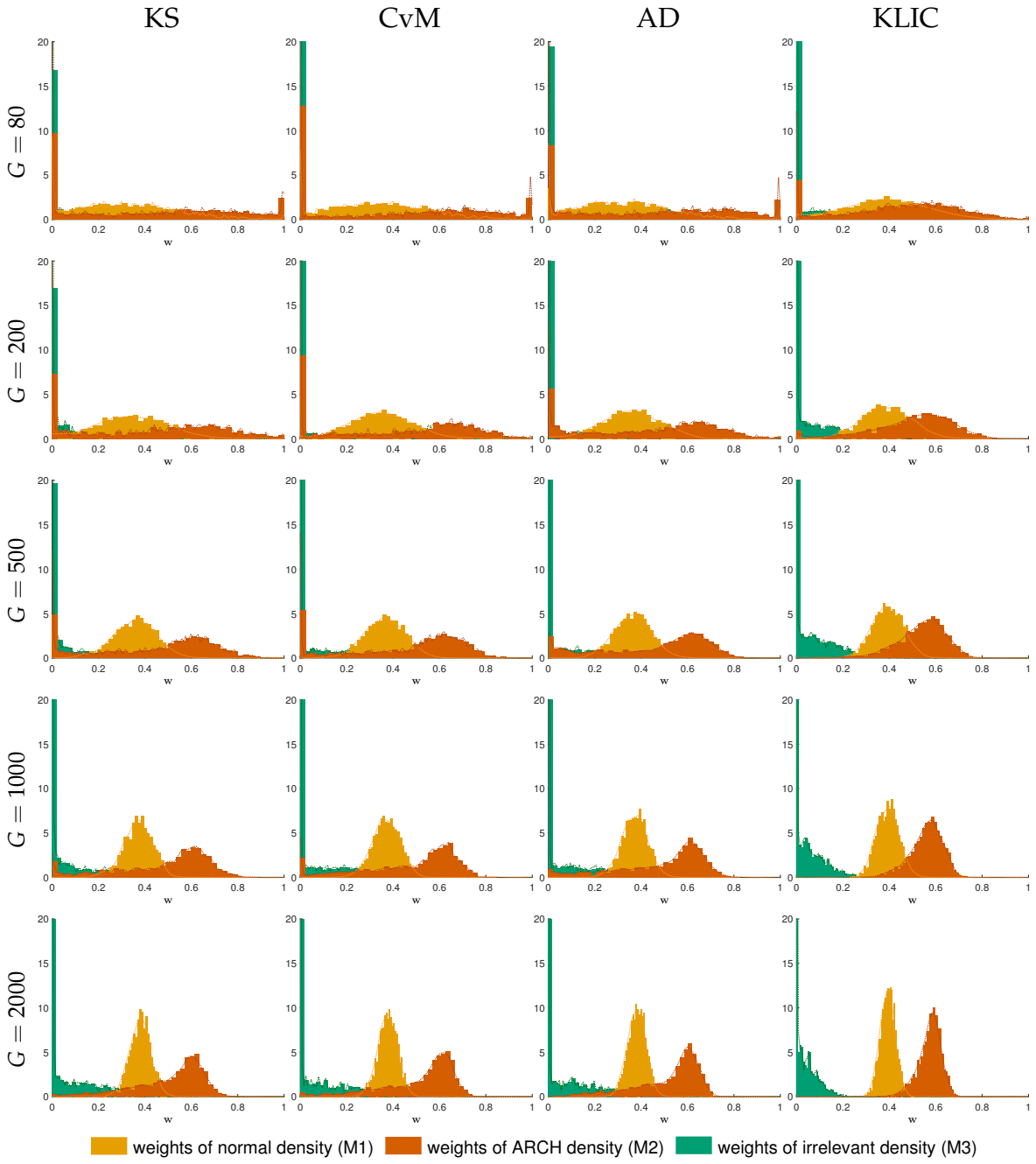
Table D.2: DGP 4, Monte Carlo summary statistics for different sample sizes G and objective functions $K_G(w), C_G(w), A_G(w)$ and $\text{KLIC}_G(w)$

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.

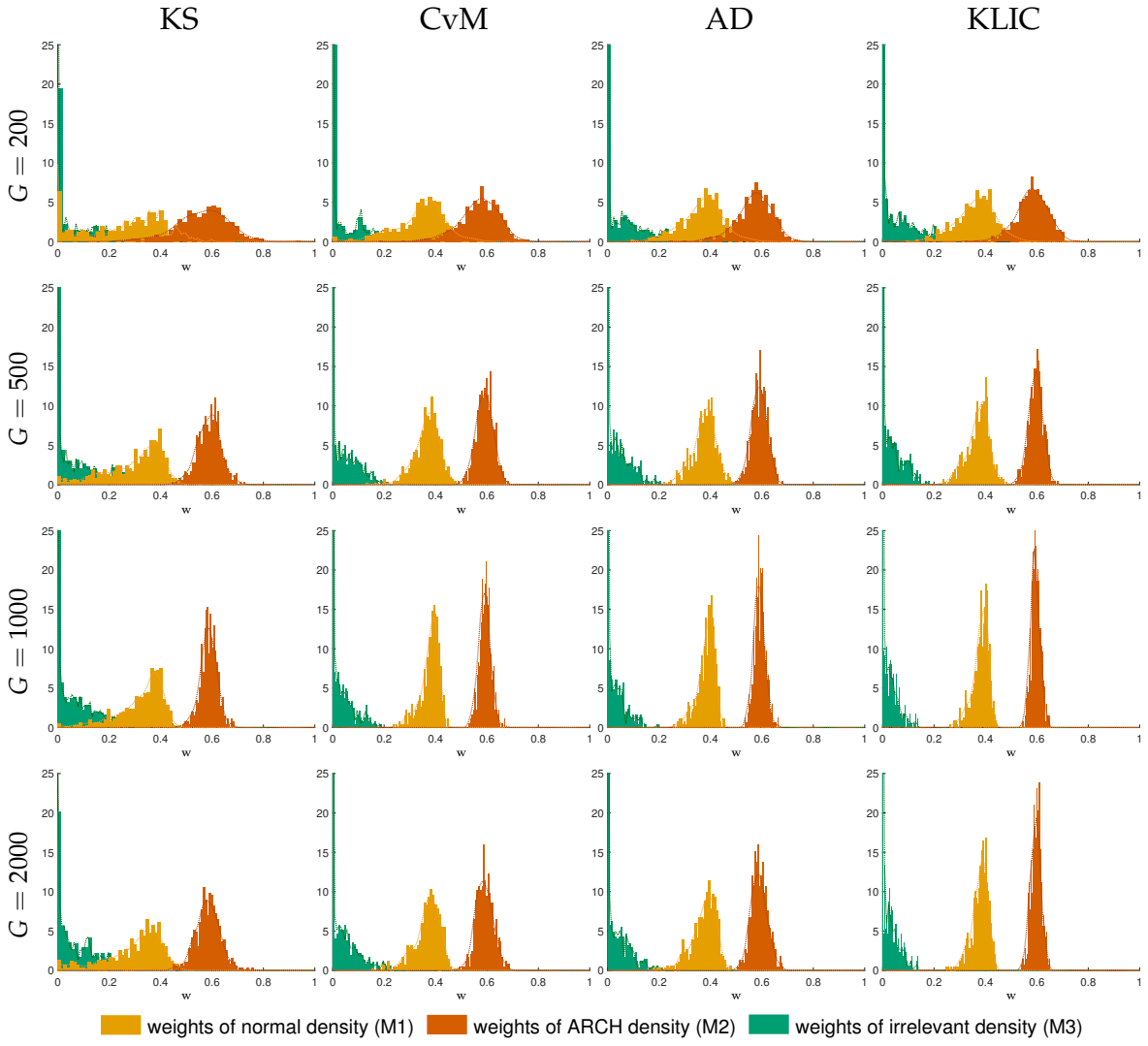
Table D.3: DGP 5, Monte Carlo summary statistics for different sample sizes G and objective functions $K_G(w), C_G(w), A_G(w)$ and $\text{KLIC}_G(w)$

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

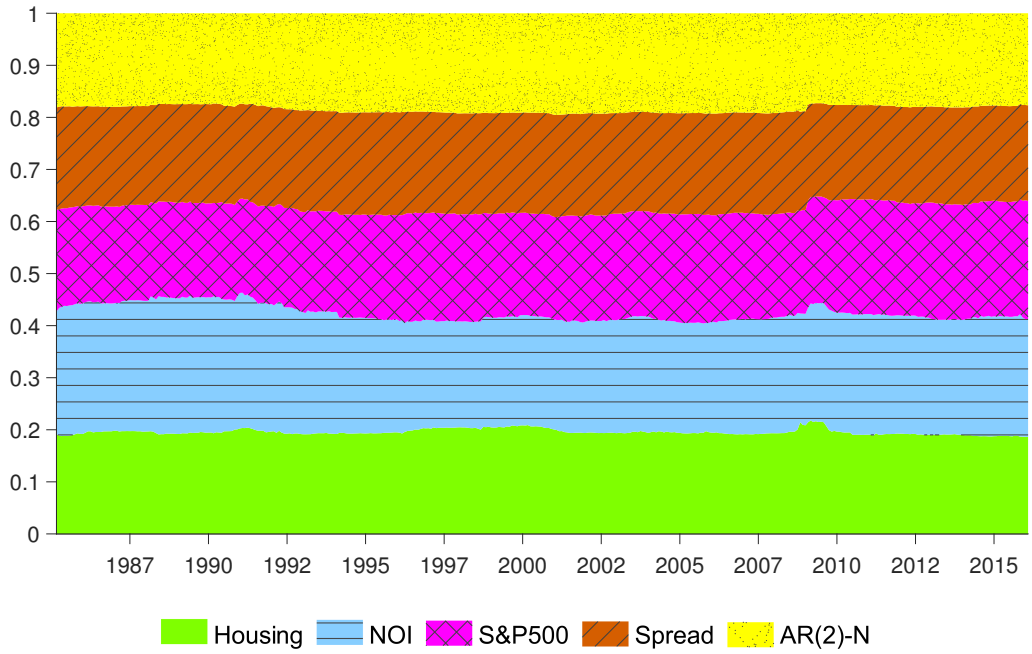
Figure E.1 shows the ratio of the inverse of the in-sample residual variances of each model, relative to the sum of the inverses, calculated in the last rolling window at each forecast origin. Bates and Granger (1969) recommended this ratio as an estimator of the optimal weights, minimizing the expected 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

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

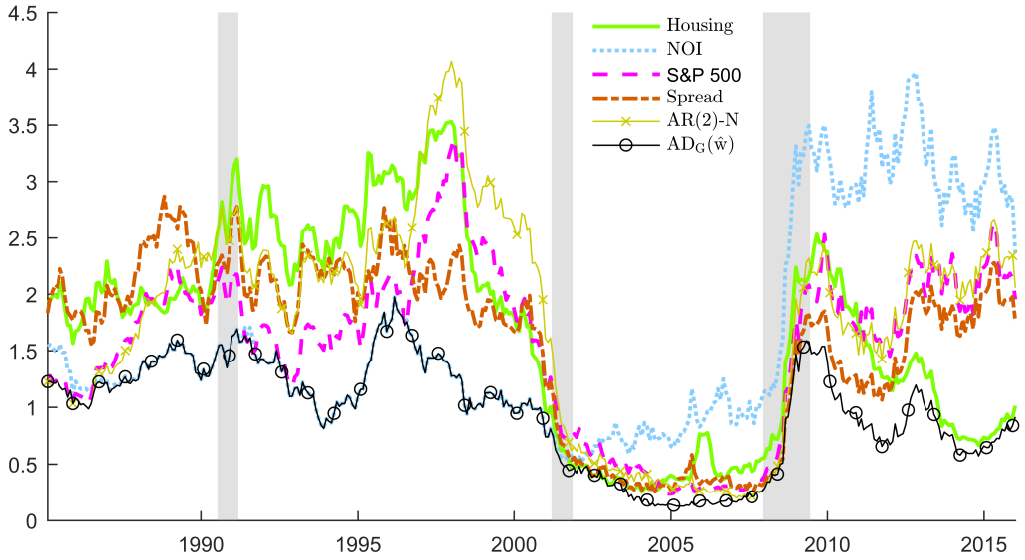


Note: The sample period (end of the last rolling window of size $R = 120$) starts in February 1985 and ends in January 2016, with a total number of $P = 372$ months. Housing stands for Housing Permits, NOI stands for ISM: New Orders Index, S&P 500 is the S&P 500 stock index returns while Spread is Moody's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

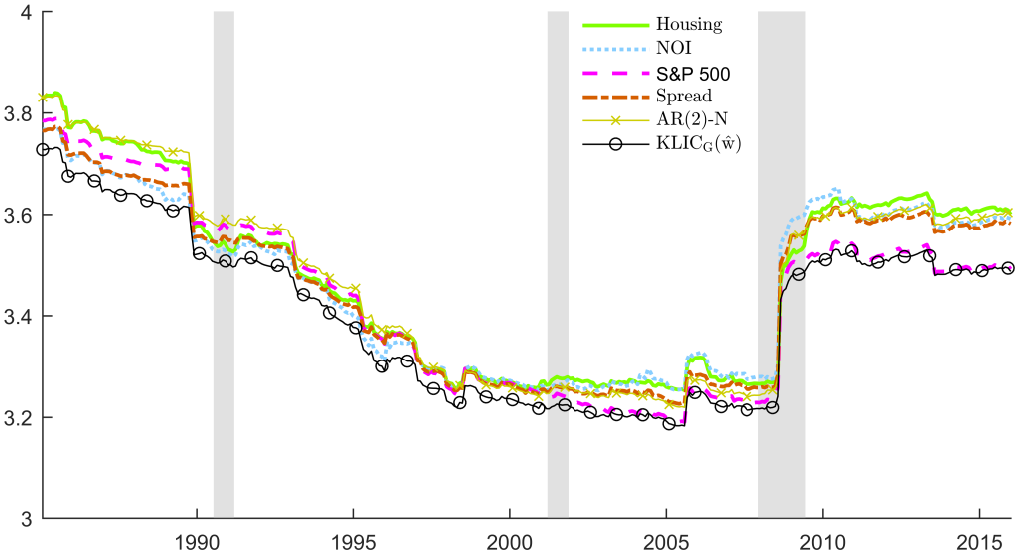
was able to exploit the differences between the individual models' predictive densities more successfully than the KLIC estimator. As Table 1.7 showed, this gain resulted in superior *out-of-sample* density forecasts.

A visual comparison of Figure E.2a and Figure E.2b reveals that both the Anderson–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 1.4 and Appendix D) and the empirical exercise (Section 1.5). To simplify notation, consider the model $y_{t+1} = z_t'\beta + \sqrt{\sigma^2}\varepsilon_{t+1}$, where ε_{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 ν 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 σ^2 by

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

   where (α_0,α_1) are additional parameters entering the likelihood function.

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

Chapter 2

Forecasting with DSGE versus Reduced-Form Models: A Time-Variation Perspective

2.1 Introduction

Nowadays a considerable fraction of economic forecasts are produced using Dynamic Stochastic General Equilibrium (DSGE) models, both in academia and at other organizations. In this paper I focus on evaluating the forecasting performance of the widely known DSGE model by Smets and Wouters (2007), henceforth SW model, and its version extended with financial frictions (SW-FF, following Del Negro and Schorfheide (2013)) against a number of statistical models. The SW model's in-sample and out-of-sample forecasting ability has made it an attractive benchmark in the forecasting literature. However, the literature has primarily focused on evaluating this model's forecasting performance (in both absolute and relative terms) and documenting its time-varying nature, rather than understanding the driving forces behind this behavior.

This paper's main contribution is investigating when DSGE models forecast better than their reduced-form competitors. Moreover, I explain how the time-variation in the models' out-of-sample forecasting performance can be linked to their in-sample performance. My results demonstrate that in-sample relative performance is informative about out-of-sample performance, but this relationship varies over time. In particular, before the recent Great Recession, this relationship strengthened, but the crisis considerably weakened this correlation. That is, in periods of economic stability, researchers can rely on models' relative in-sample performance when selecting forecasting models.

My paper is related to several contributions in the literature. First of all, in their original paper, Smets and Wouters (2007) briefly discussed their model's forecasting performance against a VAR(1) and a BVAR(4) model (VAR stands for Vector Autoregressive, BVAR stands for Bayesian VAR), concluding that at short horizons (1 quarter ahead) their model and the BVAR(4) are both competitive against the VAR(1), while at longer horizons (4 to 12 quarters ahead), the DSGE model delivers more precise forecasts than its competitors (more or less uniformly for all observable variables, including GDP growth, inflation, and interest rate). However, they did not perform a rigorous statistical analysis of the SW model's out-of-sample forecasting performance, which is the main goal of the present study. Del Negro and Schorfheide (2013) perform an analysis using variants of the SW model and compare them to a Bayesian AR(2) and judgmental forecasts (Greenbook, Blue Chip). They conclude that the SW model is competitive in forecasting GDP growth, inflation, and interest rate, especially in the medium run (4 quarters ahead) but they do not investigate whether the differences in forecasting performance are statistically significant. They consider an interesting extension of the baseline SW model by incorporating financial frictions, which I also use in this paper. The authors show that in the recent crisis, this extension considerably improves the DSGE model's forecasting ability. In addition, they discuss this model's performance from the early 1990s but not earlier – my paper also contributes on that front. In a related paper, Kolasa and Rubaszek (2015) show extending a baseline DSGE model with financial frictions in an explicitly modeled housing market significantly improves both point and density forecasts during episodes of financial turmoil, but not in tranquil times.

Edge et al. (2010) showed that while the SW model produces forecasts for GDP growth and inflation that are at least as good as Greenbook and BVAR forecasts — except for 1–3-quarter-ahead inflation forecasts, where the Greenbook dominates —, all the methods in their exercise delivered very poor quality forecasts, according to Mincer and Zarnowitz (1969) type regressions. However, they did not publish the details of a formal statistical comparison of the competing models, which is an objective of my study.

In a recent paper, Gürkaynak et al. (2013) revisit the findings in the literature on the forecasting performance of the SW model against reduced form models (AR, VAR, BVAR, random walk) and reach a number of conclusions. First, they confirm that no forecasting method is efficient, in line with the findings of Edge et al. (2010). Second, they demonstrate that relative forecasting performance shows much variability in terms of variables and forecast horizons: GDP growth

is better forecast by a simple AR model at short horizons, while the SW model performs better at longer horizons, whereas the findings are reversed in the case of inflation. Third, they suggest that the Bayesian VAR should not be used as the benchmark reduced form forecasting model, as simpler AR or VAR models are often superior. However, to my best knowledge, the literature has not tried to explain *why* in some periods the Smets and Wouters model performs better than its reduced form competitors. Furthermore, it is of both theoretical and practical interest to understand where the changes in the out-of-sample forecasting performance originate from. Can we say something about the expected (relative or absolute) performance of the models based on their in-sample fit? Answering these questions is in the focus of this paper.

To summarize, the forecasting performance of the Smets and Wouters model has been intensively investigated, and the literature seems to agree on a number of "stylized" facts. First, while the model does not deliver optimal forecasts, neither do its competitors. Second, its relative forecasting performance depends on the variable of interest as well as on the forecast horizon. Third, statistical models often outperform the structural DSGE model. However, there is no clear answer to the question if DSGE models should be preferred to reduced-form models purely on forecasting grounds. A possible explanation of the conflicting results in the literature is that all the papers mentioned earlier differ along a number of dimensions: the estimation periods and schemes, the evaluation periods and whether they use real-time or fully revised data. Table 2.1 provides a brief summary of these differences.

In this paper, I demonstrate the time-variation of the relative forecasting performance of the Smets and Wouters model against statistical models. Furthermore, I show that swings in the model's absolute and relative out-of-sample performance are strongly related to its in-sample performance. I find that the DSGE model's in-sample fit was highly informative in the early 2000s until the recent financial crisis. Moreover, I demonstrate that extending a DSGE model with financial frictions results in better forecasting performance in times of financial distress but not in other times, similarly to the findings of Kolasa and Rubaszek (2015).

The remainder of the paper is organized as follows. Section 2.2 introduces the models and data used in the paper, while Section 2.3 describes the forecasting framework. In Section 2.4 I discuss my main findings, and Section 2.5 concludes. Additional empirical results can be found in Appendices A and B, and Appendix C contains a detailed description of the data used in this paper.

Table 2.1: Overview of forecasting exercises with the SW model

Paper	Forecast period	Estimation	Real-Time
Smets and Wouters (2007)	90:Q1–04:Q4	Rec., 66:Q1–89:Q4	No
Edge et al. (2010)	92:Q1–04:Q4	Rec., 65:Q1–91:Q3	Yes
Del Negro and Schorfheide (2013)	92:Q1–11:Q1	Rec., 64:Q1–91:Q3	Yes
Gürkaynak et al. (2013)	92:Q1–05:Q4	Roll., 72:Q1–91:Q4	Yes

Note: Forecast periods always refer to one-quarter-ahead forecasts. Real-Time is whether the authors used real-time or revised data. "Rec." and "Roll." stand for recursive and rolling window estimation, respectively. Gürkaynak et al. (2013) used $w = 80$ quarters of data.

2.2 Models, data and estimation

In this section I briefly summarize the main features of the models used in the forecasting exercise, along with the estimation procedures.

The Smets and Wouters (2007) model is a benchmark New-Keynesian rational expectations DSGE model. It features a representative household maximizing its expected utility (over consumption with external habit and labor) on an infinite horizon, an exogenously modeled government, and a monetary authority that follows a Taylor rule (reacting to inflation and the output gap). The labor and goods markets are differentiated, guaranteeing some monopoly power over wages/prices. Price and wage setting are staggered, following Calvo (1983), and they feature indexation. In addition, capital adjustment costs and variable capital utilization are included. The seven exogenous disturbances that drive the stochastic behavior of the SW model are: total factor productivity, investment-specific technology, risk premium, exogenous spending, price mark-up, wage mark-up, and monetary policy shocks. The model is log-linearized around its balanced growth path.

In addition, I also used an extension of the original SW model, following Del Negro and Schorfheide (2013), who add financial frictions to the baseline model along the lines of Bernanke et al. (1999). They document that after the onset of the recent financial crisis, the DSGE model that explicitly accounts for financial factors, forecasts considerably better than the baseline model. In this paper I will refer to that model as SW-FF.

The SW model utilizes data on seven key US macroeconomic variables: the log difference of per capita real GDP, real consumption, real investment and the real wage, log hours worked, the log difference of the GDP deflator, and the Fed funds rate. In addition, the SW-FF model uses an interest rate spread. Quarterly data on observables are obtained from the St. Louis Fed's FRED database and the

U.S. Bureau of Labor Statistics. The full sample spans 1966:Q1 – 2013:Q3. Further details on the data can be found in Appendix C.

In what follows, *iid*. means independent and identically distributed, and $\mathcal{N}(\mu, \mathbb{V})$ is the normal distribution with mean vector μ and covariance matrix $\mathbb{V}$. The statistical models are the following:

1. Vector Autoregressive model of order p, VAR(p) given by

$$\mathbf{y}_t = \mathbf{c} + \sum_{q=1}^{p} \mathbf{A}_q \mathbf{y}_{t-q} + \mathbf{u}_t \,, \mathbf{u}_t \overset{\text{iid.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\mathbf{u}) \,, \tag{2.1}$$

   where $\mathbf{y}_t = (y_{1,t}, \ldots, y_{7,t})'$ is the vector of the seven observables (or eight in the case of the SW-FF model by appending $y_{8,t}$),

2. Bayesian VAR(p), which has the same structure as VAR(p) but assuming a prior distribution on the parameters, which shrinks the parameters towards univariate random walks,[1]

3. Autoregressive model of order p, AR(p), which follows from VAR(p) by including only the variable of interest in $\mathbf{y}_t$ and $\mathbf{y}_{t-1}$,

4. Random Walk and Random Walk with Drift, RW and RWD, respectively, given by $y_{j,t} = c + y_{j,t-1} + u_t$, where the j subscript refer to the variable of interest, and the restriction $c = 0$ corresponds to the Random Walk.

The AR, VAR and BVAR models are estimated using both fixed and variable lag length. The fixed lag length equals 1 in the case of the AR and VAR models, and 4 in the case of the BVAR, in line with Smets and Wouters (2007). The variable lag length is selected at each estimation window based on the Bayesian Information Criterion (Schwarz, 1978) with maximum lag length $p_{\max} = 3$.[2]

The statistical models (except the BVARs) are estimated by OLS. The BVAR models are estimated using the MATLAB code on Christopher Sims's website.[3] Bayesian estimation of the SW model is performed using Dynare (Adjemian et al., 2011), implementing the Random Walk Metropolis-Hastings (RWMH) method

[1] Precisely for this reason, four variables (real GDP, consumption, investment and the real wage) enter the model in log levels. However, forecasts are still produced for the stationary variables. The parameters characterizing the prior distribution and the likelihood are the same as used by Smets and Wouters (2007).

[2] In the present paper, I do not consider models with time-varying parameters or stochastic volatility. First, the sample size is rather small to incorporate these features. Second, the models are re-estimated in rolling windows, which at least partly accounts for time-variation in the parameters.

[3] MATLAB code available at `http://www.princeton.edu/~sims/`.

(see An and Schorfheide (2007), for example). The resulting draws from the posterior distribution of the parameters are used to generate the h-period in-sample and out-of-sample posterior distribution of forecasts of real GDP growth, inflation and the Federal funds rate. Then, according to the squared forecast error loss function, the sample mean (at the given quarter, for each horizon) is used to estimate the conditional expected value of the forecasts.

2.3 Forecasting

This section outlines how the forecasts are generated. My main objective is to identify the factors that determine the DSGE and the statistical models' forecasting performance over time. For this purpose, I perform a pseudo out-of-sample forecasting exercise. First, following the literature (Smets and Wouters (2007) and Gürkaynak et al. (2013), among others), I choose the quadratic loss function, and focus on the squared forecast error for variable of interest j at time t, $y_{j,t}$, which corresponds to $(y_{j,t} - \hat{y}_{j,t}^{(m)})^2$, where $\hat{y}_{j,t}^{(m)}$ is the h-period-ahead forecast of model m for time t. The forecast horizon is h, where $h = \{1, 4, 8\}$.[4] Each model is estimated in a rolling window fashion with window size $R = 60$, explained in what follows for the $h = 1$ case.[5,6] The first estimation sample consists of $R = 60$ observations, from 1966:Q1 to 1980:Q4, resulting in a sequence of in-sample and out-of-sample point forecasts for the three key variables at the given horizon.[7] For variable of interest j, let $\widehat{\mathcal{L}}_{j,t}^{(m)}$ denote the last in-sample loss of model m estimated up to period t and $\widehat{L}_{j,t+h}^{(m)}$ the corresponding h-period-ahead out-of-sample loss, where the "hat" notation is used to emphasize that the loss functions are evaluated at the parameter estimates. For each model and for each variable of interest I calculate the last in-sample forecast error loss corresponding to 1980:Q4, that is $\widehat{\mathcal{L}}_{j,1980:Q4}^{(m)} = (y_{j,1980:Q4} - \hat{y}_{j,1980:Q4}^{(m)})^2$ and the $h = 1$ period ahead out-of-sample loss, that is $\widehat{L}_{j,1981:Q1}^{(m)} = (y_{j,1981:Q1} - \hat{y}_{j,1981:Q1}^{(m)})^2$. The next estimation sample again consists of $R = 60$ observations, from 1966:Q2 to 1981:Q1, which are used to generate forecasts of the variables of interest and the corresponding forecast

[4] Throughout the paper, the terms "horizon", "time" and "period" mean quarters, as all the models are estimated on quarterly data. Forecasts are obtained in an iterated manner.

[5] I selected the window size to balance between producing precise forecasts and to obtain a relatively long out-of-sample period. According to Castelnuovo (2012), my choice is around the average window size used in the literature.

[6] The $h = \{4, 8\}$ cases are handled analogously, keeping in mind that the full sample spans 1966:Q1 – 2013:Q3.

[7] Note that all the models considered in this paper produce predictive densities. Corresponding to the squared forecast error loss function, I use the mean of each predictive density.

losses. I continue this procedure until reaching $R + P - 1 = R + 130 = 190 = T$, corresponding to in-sample data up to 1998:Q3 – 2013:Q2 and last out-of-sample observation in 2013:Q3. This results in a total number of $P = 131$ in-sample and out-of-sample forecast losses for each variable of interest and for each model.

2.4 Results

2.4.1 A preliminary comparison

A preliminary comparison of the predictive ability of the models is displayed in Table 2.2. First, I calculated the root mean squared forecast error (RMSFE) of each model m over the full out-of-sample period, where RMSFE is defined as

$$\text{RMSFE}^m \equiv \sqrt{P^{-1} \sum_{t=R}^{T} \widehat{L}_{t+h}^{(m)}} \,. \tag{2.2}$$

In the table, for each forecast horizon $h = \{1, 4, 8\}$ and each variable, the RMSFE column displays the following: the first row (labeled SW) is the RMSFE of the SW model, while the other rows contain the gains(+)/losses(-) *relative* to the SW model, that is $100 \times \left(1 - \frac{\text{RMSFE}^m}{\text{RMSFE}^{\text{SW}}}\right)$, and dagger shows the model with lowest RMSFE (that is, the best forecast method). In addition, column SWP tells us in what fraction of periods the SW model produced more precise forecasts than a given competitor.

Some interesting patterns emerge. The SW and SW-FF models dominate the statistical models when forecasting inflation and interest rate 4 and 8 quarters ahead. Note the particularly sizable differences in the inflation column for $h = 4$ and in the inflation and interest rate columns for $h = 8$. The SW and SW-FF models forecast GDP the most precisely in the short run, while the AR(p) model performs slightly better in the long run.

We can therefore conclude that there is no consistently "best" forecasting model, as the forecasting methods' relative performance depends on both the variable of interest and the forecast horizon. My findings are not completely in line with those of Gürkaynak et al. (2013), but they do not contradict them either. The reasons behind the different ranking of the forecasting approaches might originate from several sources. In terms of data, we used different sample periods, for both estimation and forecast evaluation. Furthermore, they used real-time data, while my data are fully revised, available as of November, 2013

(see Table 2.1). Moreover, their paper uses the direct forecasting method, whereas I used the iterated approach. These differences highlight that care must be taken when making general statements about these models' forecasting performance.

Table 2.2: Root mean squared forecast errors and comparisons relative to the SW model

		GDP growth		inflation		interest rate	
		RMSFE	SWP	RMSFE	SWP	RMSFE	SWP
$h=1$	SW	0.664	—	0.221	—	0.178	—
	SW-FF	2.2†	5.3	7.3	8.4	0.0	56.5
	VAR(4)	−56.5	95.4	−18.1	50.4	−1.1	52.7
	VAR(p)	−56.8	95.4	−17.6	50.4	−1.1	52.7
	BVAR(1)	−0.6	58.0	−4.1	48.9	3.4†	17.6
	BVAR(p)	−3.6	58.8	−4.1	48.9	1.1	37.4
	AR(1)	−4.2	89.3	3.2†	34.4	−2.8	64.9
	AR(p)	−3.9	80.2	2.7	31.3	−18.5	57.3
	RW	−18.5	97.7	−0.5	40.5	0.0	43.5
	RWD	−19.7	97.7	−2.3	44.3	−3.4	63.4
$h=4$	SW	0.739	—	0.264	—	0.435†	—
	SW-FF	0.0	57.8	8.7†	25.0	−1.6	55.5
	VAR(4)	−8.4	69.5	−74.2	65.6	−2.3	49.2
	VAR(p)	−8.3	69.5	−75.0	66.4	−2.3	49.2
	BVAR(1)	6.0†	6.3	−46.2	85.2	−5.5	64.8
	BVAR(p)	5.5	15.6	−49.2	90.6	−2.8	54.7
	AR(1)	2.2	35.9	−26.9	92.2	−6.0	65.6
	AR(p)	1.5	44.5	−23.5	85.9	−1.6	39.8
	RW	−31.8	100.0	−12.9	52.3	−5.7	50.8
	RWD	−34.4	100.0	−16.3	56.3	−10.8	57.0
$h=8$	SW	0.680	—	0.272†	—	0.559†	—
	SW-FF	−3.7	71.0	−5.2	62.1	−3.2	94.4
	VAR(4)	−16.0	78.2	−130.1	80.6	−24.9	77.4
	VAR(p)	−16.0	78.2	−131.3	80.6	−25.6	77.4
	BVAR(1)	−4.3	75.8	−109.9	99.2	−32.6	100.0
	BVAR(p)	−3.2	71.0	−114.3	99.2	−31.3	100.0
	AR(1)	0.0	46.8	−61.0	99.2	−25.6	76.6
	AR(p)	0.3†	39.5	−55.9	97.6	−29.0	58.9
	RW	−50.1	91.1	−48.5	76.6	−27.4	68.5
	RWD	−54.6	92.7	−52.9	75.0	−33.3	70.2

Note: root mean squared forecast errors of the SW model (row SW) and gains(+)/losses(-) in percentage terms relative to the SW model (all other rows). Dagger (†) shows model with lowest RMSFE. SWP is the proportion of out-of-sample periods when the SW model had lower squared forecast error than a given competitor.

Next, after confirming that there is no uniformly "best" model measured by relative RMSFEs, let us see if we observe *statistically* significant differences over the full out-of-sample period.

2.4.2 Tests of predictive ability

In this section, I seek to answer the following question: which model forecast *significantly*[8] better throughout the full out-of-sample period both unconditionally and conditionally? I use the methodology developed by Giacomini and White (2006), which is a generalization and extension of the well-known papers by Diebold and Mariano (1995) and West (1996). The main idea is to compare two forecasting methods, and see if their performance differed significantly over time, conditional on some information available to the forecaster. If so, then this information can be exploited by a researcher when selecting a forecasting method. To formalize the idea, consider the following null hypothesis:

$$\begin{aligned} H_0: \quad & \mathrm{E}\left[L_{t+h}(y_{t+h}, f_t(\widehat{\beta}_{1,t})) - L_{t+h}(y_{t+h}, g_t(\widehat{\beta}_{2,t}))|\mathcal{G}_t\right] = 0 \\ & \text{almost surely, } t = 1, 2, \dots, \end{aligned} \tag{2.3}$$

where $L_{t+h}(y_{t+h}, f_t(\widehat{\beta}_{1,t}))$ denotes the forecast losses of the first forecasting method (now always the SW model), characterized by the t-measurable function $f_t(.)$ and parameter estimates $\widehat{\beta}_{1,t}$ (and the second term is analogous for the second forecasting method), and $\mathcal{G}_t$ is some information set. In words, we are interested in testing whether the losses incurred by using the two forecasting methods differ in expected value. When comparing the unconditional performance of the competing forecasting methods, $\mathcal{G}_t$ is the trivial σ-field. However, when we are interested in the conditional forecasting performance, we want to find information not contained in past average forecasting performance that can be used to predict which model is going to perform better at a given future date. In this case we have $\mathcal{G}_t = \mathcal{F}_t$, that is the time t-information set. Giacomini and White (2006) propose the following Wald-type test statistic:

$$T_{P,h}^{k} \equiv P\left(P^{-1}\sum_{t=R}^{T} k_t \Delta L_{t+h}\right)' \widetilde{\Omega}_P^{-1}\left(P^{-1}\sum_{t=R}^{T} k_t \Delta L_{t+h}\right), \tag{2.4}$$

where $\Delta L_{t+h} \equiv L_{t+h}^{(1)}(\hat{\beta}_{1,t}) - L_{t+h}^{(2)}(\hat{\beta}_{2,t})$, that is the difference of the forecast loss series evaluated at the estimated parameters[9] of each model, k_t is any $(q \times 1)$

[8] Throughout the paper, I always consider significance at the 5% level.

[9] Note that this is a favorable feature of the Giacomini-White approach for two reasons. First, the rolling window estimation scheme naturally entails parameter estimation uncertainty, even as the sample size $T \to \infty$. Second, the Bayesian estimation of the BVAR and SW models inherently implies parameter uncertainty, as in the Bayesian framework the sample of a given size is treated as fixed, and the parameters are random variables characterized by their distributions, implied by the prior and the likelihood.

$\mathcal{F}_t$-measurable function, and $\widetilde{\Omega}_P$ is a suitable (possibly heteroskedasticity and autocorrelation consistent, HAC) $q \times q$ estimator of the asymptotic variance of $P^{-1/2} \sum_{t=R}^{T} k_t \Delta L_{t+h}$. As Giacomini and White (2006) showed, under the null hypothesis the test statistic converges in distribution to a χ^2 distributed random variable with q degrees of freedom, formally $T_{P,h}^{k} \xrightarrow{d} \chi_q^2$. It is easy to see that the evaluation of *un*conditional[10] predictive ability is a special case of the conditional one, by setting $k_t = 1$.

Table 2.3 summarizes the findings. In terms of unconditional predictive ability, the DMW columns show that the SW and SW-FF models did not forecast significantly worse than any of their competitors for any variables at any horizons, but actually the improvements are statistically significant in a number of cases, particularly when forecasting GDP growth and inflation 8 quarters ahead. I consider this as a quite strong argument in favor of the SW and SW-FF models, along the lines of Del Negro and Schorfheide (2013): DSGE modeling offers an all-round "package" to perform policy experiments and evaluate counterfactual scenarios in addition to forecasting, and all these along with a natural way to quantify uncertainty surrounding estimates and forecasts. Given that all the other modeling approaches are rather limited in such ways, it is reassuring that when it comes to forecasting, the SW model *on average* never does a worse job but in fact dominates its competitors, especially at long horizons. It is also remarkable that the SW-FF model significantly outperformed the SW model when forecasting GDP growth and inflation one quarter ahead but not in the case of interest rate or at longer horizons, which suggests that incorporating financial frictions mainly improves short-horizon forecasting ability.

Second, the GW column shows the Giacomini–White test statistics, when the conditioning set for forecasting variable $y_{j,t+h}$ is $k_t = (1, y_{j,t-1}, y_{j,t-2})'$. As we can see, it can not be rejected that conditionally the SW model and any of its competitors forecast equally well. Of course, this result depends on the conditioning set k_t. However, including the first two lags of all three variables of interest does not change the conclusions. To see if we can gain insight into how the models performed over the business cycle, I included recession probabilities[11] in the conditioning set k_t. However, this did not help forecasting which method

[10] When calculating the test statistic in the unconditional case, I use the form advocated by Diebold-Mariano-West, which is informative about which model performs better – the test statistic suggested by Giacomini and White is the square of the former.

[11] The recession probabilities were calculated by Jeremy Piger. Data are available on Piger's website at `http://pages.uoregon.edu/jpiger/us_recession_probs.htm/`. However, it must be noted that these are smoothed and not filtered recession probabilities, that is they use the full sample of data available.

produces more precise forecasts. For each variable, the third column labeled CP contains the fraction of cases (in percents), when the regression of ΔL_{t+h} on k_t correctly predicted the better forecasting model – we can see that unfortunately the numbers are not much different from 50% (which is the benchmark case, corresponding to a coin flip).

To sum up, while unconditionally there seems to be significant evidence in favor of the SW model, especially when forecasting at the long horizon, we apparently cannot (yet) exploit this difference by conditioning on information available at time t. A valid concern that arises when analyzing the results of Table 2.3 is the problem of multiple comparisons. In order to address these complications, Table 2.3 also contains the results using Hochberg's (1988) Bonferroni-type correction method, as suggested by Giacomini and White (2006) — significant differences are in bold.

A relevant question is whether the SW model is outperformed by any of the alternatives, at any horizon and for any variable. A positive answer would be a very strong warning sign against the baseline DSGE model's forecasting ability. The Superior Predictive Ability (SPA) test by Hansen (2005) seeks to answer precisely this question. The null hypothesis considered by the SPA test is that the benchmark model — in our case, the SW model — is not inferior to *any* of the alternatives. The results shown in Table 2.4 confirm the previous finding, namely that the SW model has the best performance when forecasting inflation or interest rate at longer horizons (4 and 8 quarters ahead), and while its performance is somewhat worse at the 1 quarter ahead horizon and when forecasting GDP, it is not significantly worse than any of the statistical models or the DSGE model with financial frictions.[12]

In Figures 2.1 to 2.5, we can see a number of selected forecast comparisons, plotting the realizations and predictions, along with forecast error losses and marking the periods when the SW model produced lower loss. After inspecting the figures, it becomes apparent that even in cases when one model dominates the other *on average* over the full out-of-sample period (the SW model has almost 50% lower RMSFE than the RW for inflation 8 quarters ahead, and the DSGE model dominates the VAR(4) by 25% when forecasting interest rate 8 quarters ahead!), there is time variation in the relative predictive ability of even the best

[12]In order to identify the "best" forecasting model, I also calculated Model Confidence Sets for each horizon and for each variable of interest, following Hansen et al. (2011). The results in Table A.1 in Appendix A indicate that the SW model is always inside the Model Confidence Set, but so are many other statistical models. This confirms that the data are not informative enough to shrink the pool of candidate models to a *much* smaller set.

Table 2.3: Tests of predictive ability

		GDP growth			inflation			interest rate		
		DMW	GW	CP	DMW	GW	CP	DMW	GW	CP
$h=1$	SW-FF	2.16*	0.05	58.8	2.61*	0.07	58.8	−0.09	0.06	55.7
	VAR(4)	**−2.73***	0.13	62.6	−1.32	0.05	59.5	−0.09	0.01	51.9
	VAR(p)	**−2.75***	0.14	63.4	−1.29	0.05	59.5	−0.08	0.01	49.6
	BVAR(1)	−0.07	0.00	52.7	−0.71	0.06	56.5	0.35	0.01	52.7
	BVAR(p)	−0.36	0.01	51.9	−0.72	0.06	56.5	0.14	0.01	50.4
	AR(1)	−0.61	0.01	42.0	0.68	0.06	60.3	−0.27	0.00	60.3
	AR(p)	−0.59	0.01	45.8	0.57	0.03	64.1	−1.01	0.02	53.4
	RW	−1.65	0.03	55.7	−0.15	0.08	63.4	0.02	0.00	61.8
	RWD	−1.72	0.03	55.0	−0.45	0.07	62.6	−0.28	0.00	60.3
$h=4$	SW-FF	0.00	0.36	52.3	1.57	6.06	55.5	−0.26	1.51	59.4
	VAR(4)	−0.96	6.17	60.2	−1.71	4.34	53.1	−0.21	2.00	44.5
	VAR(p)	−0.95	6.18	60.2	−1.73	4.56	53.9	−0.22	2.04	44.5
	BVAR(1)	0.78	1.25	57.0	−1.88	3.27	57.8	−0.44	2.44	50.8
	BVAR(p)	0.72	1.48	50.8	−1.95	3.49	57.0	−0.24	1.96	48.4
	AR(1)	0.29	0.91	45.3	−1.58	4.60	52.3	−0.45	2.07	43.0
	AR(p)	0.20	1.04	45.3	−1.38	4.28	54.7	−0.13	1.75	59.4
	RW	−1.85	3.48	53.1	−0.88	2.30	57.8	−0.40	1.77	51.6
	RWD	−1.95	3.78	53.1	−1.04	2.14	57.0	−0.67	1.55	52.3
$h=8$	SW-FF	−0.89	1.53	51.6	−0.53	4.66	66.9	−0.85	0.73	45.2
	VAR(4)	−1.97*	4.52	68.5	−2.12*	7.44	61.3	−1.74	5.21	56.5
	VAR(p)	−1.97*	4.52	68.5	−2.13*	7.70	61.3	−1.81	5.63	58.1
	BVAR(1)	−0.64	0.79	51.6	**−2.65***	4.92	71.0	−2.11*	5.37	64.5
	BVAR(p)	−0.52	0.73	47.6	**−2.65***	4.88	71.0	−2.08*	5.62	62.1
	AR(1)	0.02	0.10	49.2	**−2.53***	3.38	60.5	−1.44	2.13	54.8
	AR(p)	0.05	0.09	51.6	**−2.30***	3.00	54.0	−1.44	2.81	54.0
	RW	**−2.48***	5.26	59.7	−1.93	3.41	65.3	−1.53	2.02	55.6
	RWD	**−2.55***	5.61	60.5	−1.89	3.32	60.5	−1.65	2.32	58.9

Note: DMW is the Diebold-Mariano-West test statistic of equal unconditional predictive ability (negative values showing lower average losses of the SW model), and GW is the Giacomini-White test statistic for testing equal conditional predictive ability. CP is the proportion of cases (in %) when the Giacomini-White procedure correctly predicted the better forecasting model. Asymptotic variances estimated by the Newey and West (1987) HAC estimator: truncation lag is $h-1$ for the conditional GW test, while $\lfloor P^{1/4} \rfloor = 3$ for the unconditional DMW test. Asterisks $(^*)$ show significance at the 5% level. Statistics in bold mean significant differences after the Hochberg (1988) correction was applied.

(measured by their RMSFEs) models. It is remarkable that the Smets and Wouters model extended by financial frictions dominates the benchmark SW model after the onset of the financial crisis (note the spike in Figure 2.5), but the baseline SW model performed better during earlier recessions, similarly to the results of Kolasa and Rubaszek (2015). Furthermore, this is in line with the findings of Ng

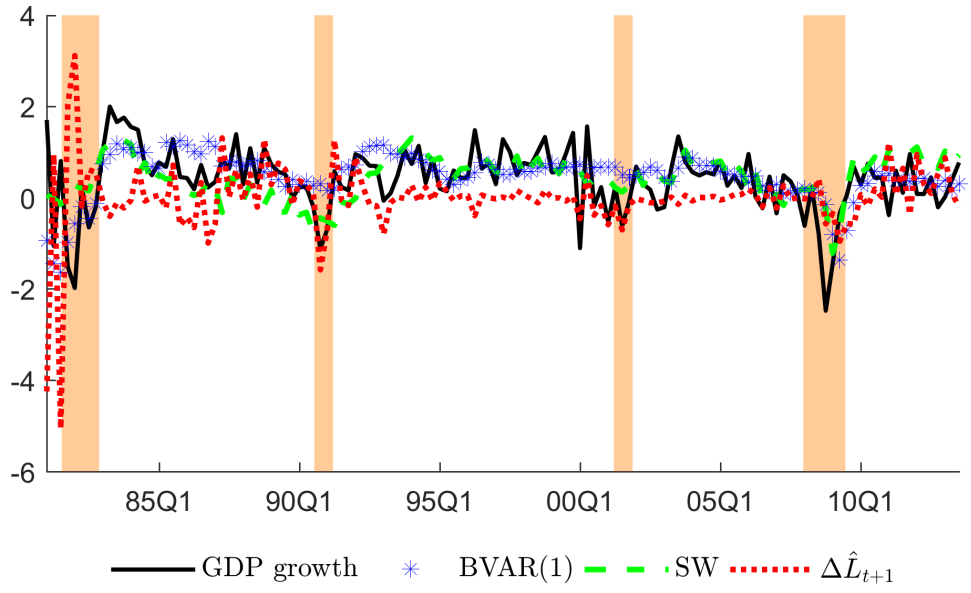
Table 2.4: Test of Superior Predictive Ability, p-values

	GDP growth	inflation	interest rate
$h = 1$	0.42	0.44	0.72
$h = 4$	0.54	1.00	1.00
$h = 8$	0.62	1.00	1.00

Note: p-values of test of Superior Predictive Ability of the SW model as benchmark, following Hansen (2005). Results based on 10,000 bootstrap replications, with block bootstrap length of 20 quarters of data. The results are robust to block length (15 and 30) and to using the stationary bootstrap.

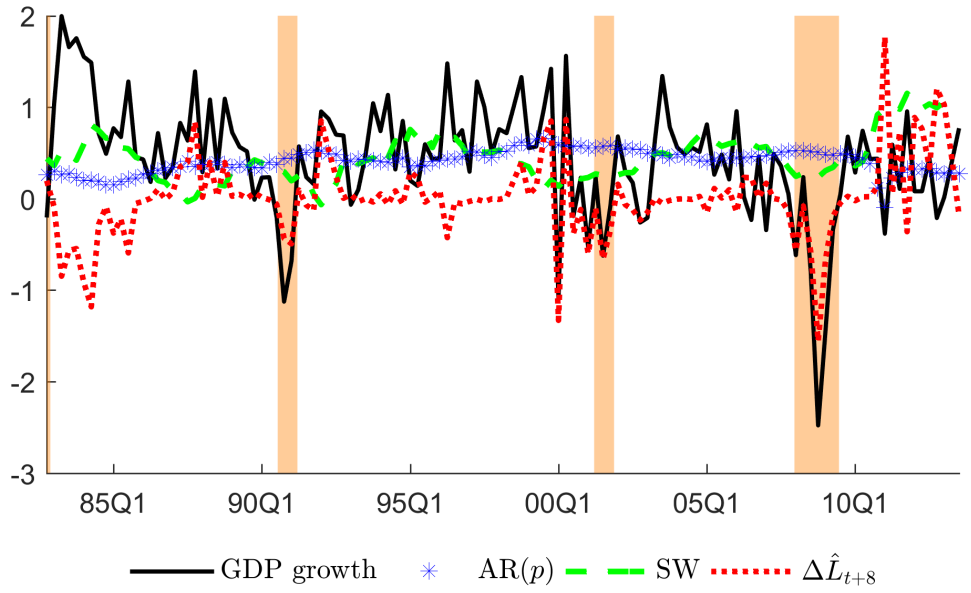
and Wright (2013), who note that the recent Great Recession was different from other post—World War II economic downturns. They offer explanations as to why forecasting during the crisis was different than before. One is that in a highly leveraged economy, credit spreads are more informative about the future paths of macroeconomic variables. Another possible explanation is that the origins of the crisis were not monetary/fiscal policy shocks or other exogenous demand or supply shifters but rather a housing/credit bubble.

Figure 2.1: SW vs. BVAR(1) 1-quarter-ahead GDP growth forecasts



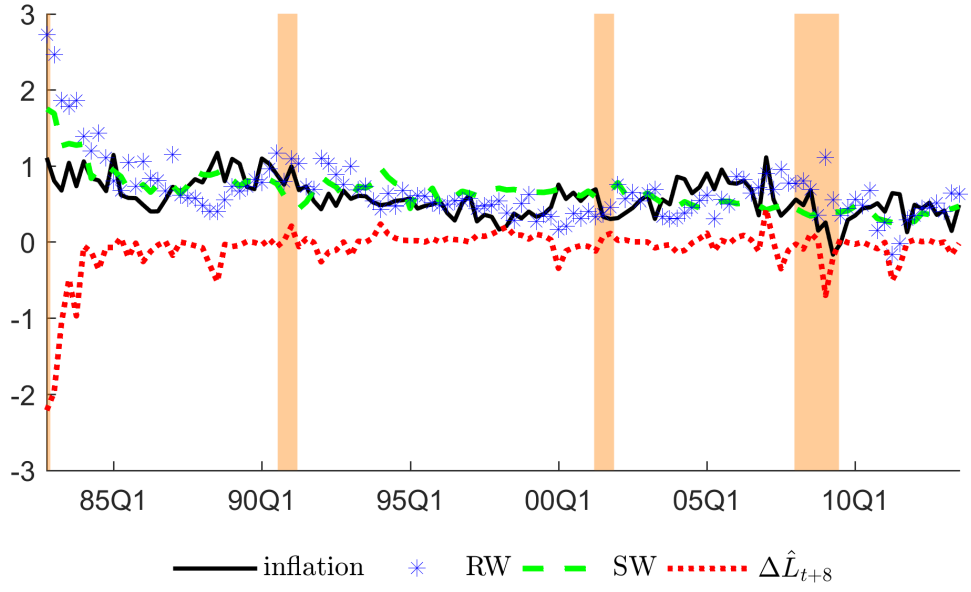
Note: Solid black line: GDP growth data. Blue asterisks: BVAR(1) forecasts. Dashed green line: SW model's forecasts. Dotted red line: forecast loss (squared forecast error) differences (SW–BVAR), positive (negative) values indicate SW (BVAR(1)) model forecasts worse. Shaded areas are NBER recession dates.

Figure 2.2: SW vs. AR(p) 8-quarter-ahead GDP growth forecasts



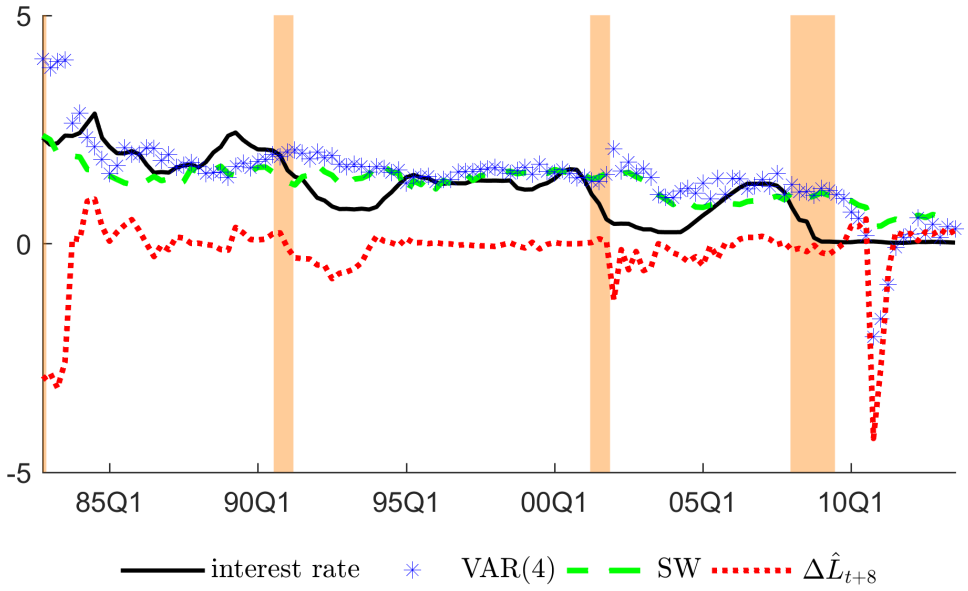
Note: Solid black line: GDP growth data. Blue asterisks: AR(p) forecasts. Dashed green line: SW model's forecasts. Dotted red line: forecast loss (squared forecast error) differences (SW – AR(p)), positive (negative) values indicate SW (AR(p)) model forecasts worse. Shaded areas are NBER recession dates.

Figure 2.3: SW vs. RW, 8-quarter-ahead inflation forecasts



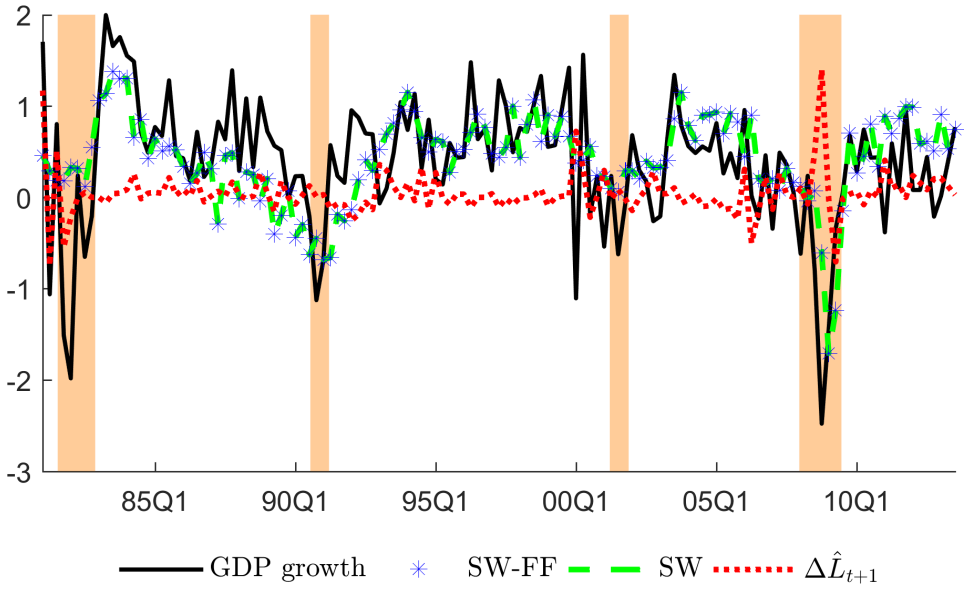
Note: Solid black line: Inflation data. Blue asterisks: RW forecasts. Dashed green line: SW model's forecasts. Dotted red line: forecast loss (squared forecast error) differences (SW – RW), positive (negative) values indicate SW (RW) model forecasts worse. Shaded areas are NBER recession dates.

Figure 2.4: SW vs. VAR(4), 8-quarter-ahead interest rate forecasts



Note: Solid black line: Interest rate data. Blue asterisks: VAR(4) forecasts. Dashed green line: SW model's forecasts. Dotted red line: forecast loss (squared forecast error) differences (SW – VAR(4)), positive (negative) values indicate SW (VAR(4)) model forecasts worse. Shaded areas are NBER recession dates.

Figure 2.5: SW vs. SW-FF, 1-quarter-ahead GDP-growth forecasts



Note: Solid black line: GDP growth data. Blue asterisks: SW-FF forecasts. Dashed green line: SW model's forecasts. Dotted red line: forecast loss (squared forecast error) differences (SW – SW-FF), positive (negative) values indicate SW (SW-FF) model forecasts worse. Shaded areas are NBER recession dates.

The figures suggest that there is considerable time-variation in the relative performance of the models considered, and investigating the sources of this instability is the topic of the next section.

2.4.3 A decomposition approach

As we have concluded that there seems to be time-variation in the models' relative forecasting performance but it cannot be easily exploited, now I turn to a different approach to shed light on understanding why in some periods the DSGE models forecast better than in other periods.

Rossi and Sekhposyan (2011) propose decomposing the difference between models' forecast losses and the average expected loss into three components, measuring (1): time-variation relative to the average full out-of-sample loss, (2): average expected out-of-sample loss based on the in-sample performance, and (3): a term capturing average unexpected forecast error loss.

To fix ideas, let us see their decomposition:[13]

$$\frac{1}{m}\sum_{t=R+\tau-m}^{R+\tau-1}\left[\widehat{L}_{t+h}-\mathrm{E}\left(L_{t+h}\right)\right]=\left(A_{\tau,P}-\overline{A}_{\tau,P}\right)+\left(B_P-\overline{B}_P\right)+\left(U_P-\overline{U}_P\right)\ , \quad (2.5)$$

where

$$A_{\tau,P}\equiv m^{-1}\sum_{t=R+\tau-m}^{R+\tau-1}\widehat{L}_{t+h}-P^{-1}\sum_{t=R}^{T}\widehat{L}_{t+h} \quad (2.6)$$

measures instabilities in the relative forecasting performance over a rolling window of size m, τ is a running index from m to P, and $\overline{A}_{\tau,P}=\mathrm{E}(A_{\tau,P})$,

$$B_P\equiv\left(P^{-1}\sum_{t=R}^{T}\widehat{\mathcal{L}}_{t+h}\right)\widehat{\beta} \quad (2.7)$$

with $\widehat{\beta}$ being the OLS estimate of β in $\widehat{L}_{t+h}=\beta\widehat{\mathcal{L}_t}+\widehat{u}_{t+h}$, $(t=R,\ldots,T)$, reflecting how much of the average out-of-sample forecasting ability was predictable based on the in-sample fit, $\overline{B}_P=\beta\,\mathrm{E}(\widehat{\mathcal{L}}_t)$, and finally

$$U_P\equiv P^{-1}\sum_{t=R}^{T}\widehat{u}_{t+h} \quad (2.8)$$

measures the average unexpected loss, with $\overline{U}_P=\mathrm{E}(\widehat{L}_{t+h}-\beta\,\mathrm{E}(\widehat{\mathcal{L}}_t))$.

[13] Their framework can accommodate both relative and absolute measures of forecast losses. In what follows, with a slight abuse of notation, I will not use the Δ symbol as before, but it should be used whenever necessary. Furthermore, while the decomposition is applied to each variable of interest j and model m, I omitted these indices to simplify notation.

I am interested in testing the following hypotheses: first, was there time-variation in the models' relative forecasting performance, where the null hypothesis is $H_{0,A} : \overline{A}_{\tau,P} = 0, \forall\, \tau = m, \ldots, P$. Second, is out-of-sample performance predictable on the basis of in-sample performance, where the null is $H_{0,B} : \overline{B}_P = 0$. Third, is the unexplained component significantly different from zero, which is an indication of over-fitting, where we test the null $H_{0,U} : \overline{U}_P = 0$. The respective test statistics are denoted by Γ_A, Γ_B, and Γ_U.

Table 2.5 shows the results of the decomposition of the forecast error losses (window size $m = 25$), both in absolute terms (row labeled Null) and in relative terms (SW model versus SW-FF or a statistical model).[14] The general patterns are the following: time-variation in the average performance (Γ_A) does not seem to explain much of the relative forecasting performance in most of the cases. However, the correlation between the in-sample and out-of-sample performance (Γ_B column) is statistically significant in an overwhelming majority of cases, and in most of those cases the model which was better in-sample was also better out-of-sample ($+$ signs in the B_P column). Furthermore, the over-fitting component is significant when analyzing the *absolute* forecasting performance of the SW model, suggesting that the SW model is over-parameterized.

To sum it up, we can see that there is predictive content in the models' in-sample performance over the full out-of-sample period. However, instabilities relative to the average full out-of-sample performance do not contribute significantly to the fluctuation of average forecast losses around their expected value. This raises the question whether this time-variation comes from a time-varying β, that is whether the correlation between the in- and out-of-sample losses is not stable over time. Figures 2.6 to 2.8 present evidence supporting this conjecture. It is interesting to see that the β linking the out-of-sample and in-sample relative performances of the models is higher in the early 2000s until the onset of the crisis, than in other time periods. It is particularly striking how the in-sample and out-of-sample forecasting link broke down around the Great Recession, as shown in Figure 2.7. This confirms that in tranquil times, models' (relative) in-sample performance is more positively related to their (relative) out-of-sample performance. In other words, a researcher facing the decision which model to use can select a forecasting model based on its relative in-sample fit, carefully taking into account the time-varying nature of this relationship.

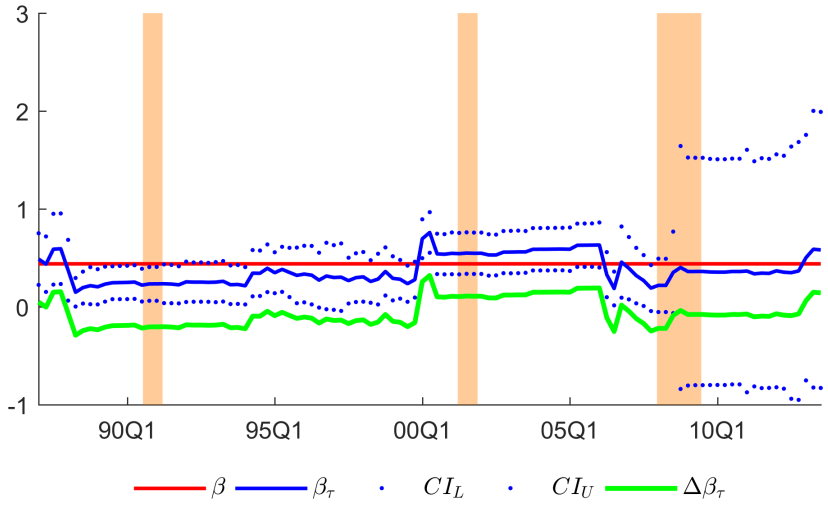
[14] As a robustness check, I also performed the tests using a window size $m = 40$. The results in Appendix B indicate that the main conclusions are largely unchanged.

Table 2.5: Decomposition of out-of-sample forecast error losses

		GDP growth				inflation				interest rate			
		Γ_A	Γ_B	B_P	Γ_U	Γ_A	Γ_B	B_P	Γ_U	Γ_A	Γ_B	B_P	Γ_U
$h=1$	Null	3.57	6.09*	+	3.12*	4.31	14.33*	+	1.33	8.16*	20.87*	+	−0.27
	SW-FF	4.28	2.42*	+	1.91	4.24	2.70*	+	2.43*	5.81	2.70*	−	−0.30
	VAR(4)	6.40	−2.65*	+	−2.88*	6.39	−13.11*	+	−0.31	5.27	−20.78*	+	0.10
	VAR(p)	6.40	−2.65*	+	−2.92*	6.41	−13.13*	+	−0.30	5.26	−20.79*	+	0.11
	BVAR(1)	3.28	2.77*	−	−0.15	5.91	−13.58*	+	0.75	3.25	−20.87*	−	0.60
	BVAR(p)	3.44	−2.68*	+	−0.34	6.10	−13.65*	+	0.91	3.23	−20.83*	−	0.35
	AR(1)	4.07	1.30	−	−0.84	3.51	−13.60*	−	1.84	3.27	−20.94*	+	0.66
	AR(p)	4.18	1.59	−	−0.69	3.41	−13.76*	−	1.72	4.05	−21.11*	+	−0.25
	RW	4.00	0.71	−	−1.71	3.64	−13.40*	+	0.64	2.95	−20.97*	−	0.68
	RWD	3.92	0.64	−	−1.78	3.53	−13.40*	+	0.43	3.44	−21.02*	+	0.71
$h=4$	Null	4.84	6.51*	+	2.95*	3.68	10.02*	+	2.51*	4.02	9.58*	+	0.69
	SW-FF	6.48	0.59	+	−0.07	6.31	1.83	+	1.03	3.46	2.91*	−	−0.54
	VAR(4)	6.62	−1.48	+	−0.87	6.28	−6.45*	+	−0.93	4.04	9.62*	−	−1.05
	VAR(p)	6.62	−1.47	+	−0.86	6.27	−6.44*	+	−0.94	4.03	9.61*	−	−1.06
	BVAR(1)	3.98	−2.80*	−	1.01	8.30*	−8.03*	+	−0.44	3.30	−9.66*	+	0.10
	BVAR(p)	3.91	−2.56*	−	1.00	8.49*	−7.80*	+	−0.66	3.24	−9.71*	+	−0.01
	AR(1)	4.05	1.41	+	0.14	8.43*	−7.42*	+	−0.85	3.86	−9.15*	+	0.34
	AR(p)	3.61	1.77	+	−0.03	8.69*	−7.54*	+	−0.79	3.53	−9.36*	+	0.52
	RW	5.22	−1.70	+	−1.64	5.32	−7.17*	+	−0.21	3.15	−9.07*	+	0.56
	RWD	5.19	−1.62	+	−1.80	5.43	−7.13*	+	−0.34	3.07	−9.02*	+	0.56
$h=8$	Null	4.29	6.74*	+	3.08*	2.79	7.56*	+	1.06	4.90	6.59*	+	2.54*
	SW-FF	7.22*	−1.72	+	−0.68	4.77	0.99	−	−0.55	5.81	2.21*	−	−1.11
	VAR(4)	6.70*	−1.48	+	−1.91	5.53	−3.83*	+	−0.51	3.56	−5.97*	+	0.42
	VAR(p)	6.71*	−1.48	+	−1.92	5.52	−3.83*	+	−0.53	3.64	−5.97*	+	0.36
	BVAR(1)	4.62	−3.49*	+	−0.41	8.36*	−5.39*	+	−1.62	4.41	−6.30*	+	0.21
	BVAR(p)	4.76	−3.47*	+	−0.28	8.37*	−5.05*	+	−2.04*	3.97	−6.31*	+	0.08
	AR(1)	2.38	−1.88	−	0.08	8.81*	−4.82*	+	−2.17*	2.83	−5.73*	+	1.87
	AR(p)	2.43	−2.19*	−	0.08	9.04*	−4.99*	+	−2.04*	3.60	−5.86*	+	1.77
	RW	4.30	−2.41*	+	−2.51*	6.29	−3.31*	+	−0.59	4.01	−5.09*	+	1.91
	RWD	4.34	−2.23*	+	−2.68*	6.22	−3.37*	+	−0.50	4.18	−5.08*	+	1.82

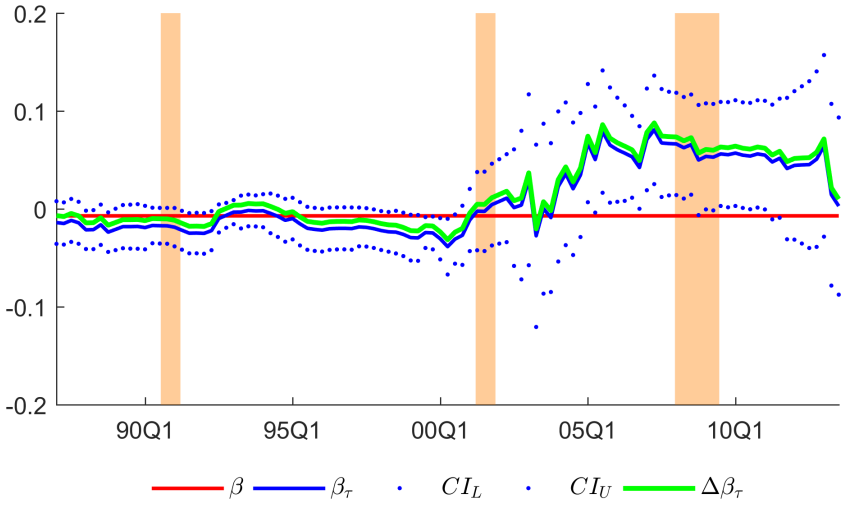
Note: Γ_A is the time variation component, Γ_B measures predictability of out-of-sample losses based on in-sample losses, and Γ_U is the unpredictable (over-fitting) component. The $+/-$ signs in the B_P column show whether the model that performed better in sample was also better/worse out of sample. "Null" means absolute forecast error losses of the SW model, while the rest of the rows display the given model's forecasting ability compared to the SW model. Asterisks $(^*)$ show significance at the 5% level.

Figure 2.6: SW model, GDP growth forecast 1 quarter ahead, time variation in β



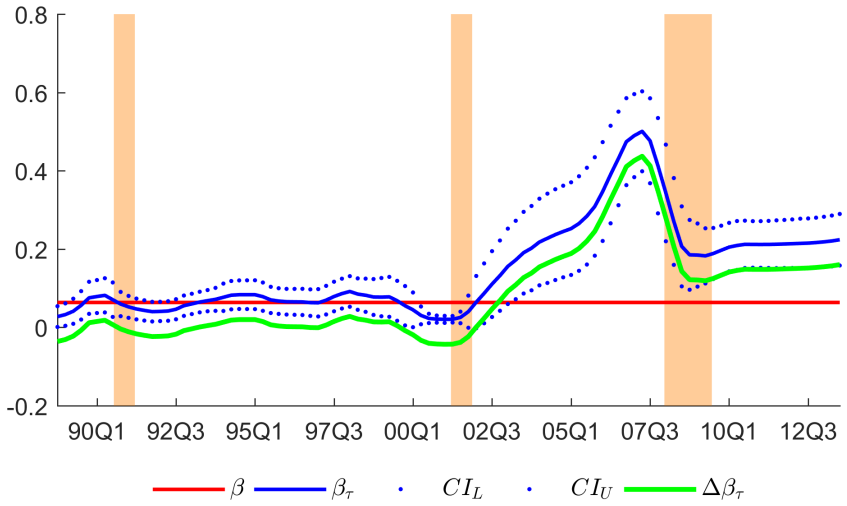
Note: Horizontal (red) line: regression of out-of-sample losses on in-sample losses over the full out-of-sample period. Dark (blue) line: regression β_τ estimated in rolling windows of 25 observations, where the timing is according to the end of the window. Light (green) line: time-variation in β, that is $\beta_\tau - \beta$. Dots are the lower and upper bounds of 95% confidence intervals for β_τ , calculated in each window. Shaded areas are NBER recession dates.

Figure 2.7: SW vs. SW-FF model, inflation forecast 1 quarter ahead, time variation in β



Note: Horizontal (red) line: regression of out-of-sample losses on in-sample losses over the full out-of-sample period. Dark (blue) line: regression β_τ estimated in rolling windows of 25 observations, where the timing is according to the end of the window. Light (green) line: time-variation in β, that is $\beta_\tau - \beta$. Dots are the lower and upper bounds of 95% confidence intervals for β_τ , calculated in each window. Shaded areas are NBER recession dates.

Figure 2.8: SW model, interest rate forecast 8 quarters ahead, time variation in β



Note: Horizontal (red) line: regression of out-of-sample losses on in-sample losses over the full out-of-sample period. Dark (blue) line: regression β_τ estimated in rolling windows of 25 observations, where the timing is according to the end of the window. Light (green) line: time-variation in β, that is $\beta_\tau - \beta$. Dots are the lower and upper bounds of 95% confidence intervals for β_τ, calculated in each window. Shaded areas are NBER recession dates.

2.5 Conclusion

The existing literature has found that the Smets and Wouters model delivers competitive forecasts for three main variables of interest (GDP growth, inflation, and interest rate) when compared against a number of benchmark reduced-form models. However, while the previous statement is valid when considering the full out-of-sample forecast period (1981Q1:2013Q3), I present evidence of time-variation in the model's relative performance. I find that there is no consistently "best" forecasting method, rather it depends on the variable of interest and the forecast horizon. Furthermore, this paper also shows that models' out-of-sample forecasting performance is significantly correlated with their in-sample performance and this information can potentially be exploited by a researcher to select a forecasting method. However, this relationship is time-varying, which makes it difficult to use the information content of in-sample performance in order to improve forecasts.

This study could be extended in several directions. One common concern regarding studies that use rolling windows is whether the results are robust against different window sizes. Another possibility is using real-time vintages of data to more closely mimic the actual forecasting scenario a researcher faces in

practice. Re-estimating the models using the data collected by Edge et al. (2010) (also used by (Gürkaynak et al., 2013)) would certainly be interesting, but as Gürkaynak et al. (2013) and Edge et al. (2010) emphasize, the prior distributions of the DGSE models (and the Bayesian VARs) inherently carry present-day beliefs about the parameter values and the very structure of the model.

As we could see, time-variation is present between models' in-sample and out-of-sample performance. This suggests that decomposing the B_P component further could provide relevant insights into how models' in-sample and out-of-sample forecasts are related over time – if the relationship between the losses or the changes in losses themselves contribute more to the time-variation.

One could estimate the DSGE models in their non-linear form, as there is considerable evidence that taking into account the nonlinearities both in the solution (non-linear approximation of policy rules as opposed to log-linearization) and the estimation (particle filter against the Kalman filter) considerably increases the in-sample fit of this class of models (Fernández-Villaverde and Rubio-Ramírez, 2005). Moreover, in light of the recent crisis, taking into account the zero lower bound might improve a model's fit. Whether this translates into more precise forecasts is an open question, outside the scope of the present paper.

Appendices

A Model Confidence Sets

Table A.1: Model Confidence Sets

	GDP growth
$h = 1$	AR(1), AR(p), BVAR(1), BVAR(p), RW, **SW**, SW-FF
$h = 4$	AR(1), AR(p), BVAR(1), BVAR(p), RW, VAR(4), VAR(p), **SW**, SW-FF
$h = 8$	AR(1), AR(p), BVAR(1), BVAR(p), VAR(4), VAR(p), **SW**, SW-FF
	inflation
$h = 1$	AR(1), AR(p), BVAR(1), BVAR(p), RW, RWD, VAR(4), VAR(p), **SW**, SW-FF
$h = 4$	AR(1), AR(p), BVAR(1), BVAR(p), RW, RWD, VAR(4), VAR(p), **SW**
$h = 8$	AR(1), AR(p), BVAR(1), BVAR(p), RW, RWD, VAR(4), VAR(p), **SW**, SW-FF
	interest rate
$h = 1$	AR(1), AR(p), BVAR(1), BVAR(p), RW, RWD, VAR(4), VAR(p), **SW**
$h = 4$	AR(1), AR(p), BVAR(1), BVAR(p), RW, RWD, VAR(4), VAR(p), **SW**
$h = 8$	AR(1), AR(p), BVAR(1), BVAR(p), RW, RWD, VAR(4), VAR(p), **SW**, SW-FF

Note: Model Confidence Sets, following Hansen et al. (2011). Results based on threshold p-value=0.05, 10000 bootstrap replications, with block bootstrap length of 20 quarters of data. The results are robust to block length (15 and 30) and to using the stationary bootstrap.

B Robustness checks

Table B.1 evaluates the robustness of the decomposition in the main text (cfr. Table 2.5) using a window size $m = 40$ and discarding the first 2 years of data. As we can see, the main conclusions are unchanged.

Table B.1: Decomposition of out-of-sample forecast error losses

		Γ_A	Γ_B	B_P	Γ_U	Γ_A	Γ_B	B_P	Γ_U	Γ_A	Γ_B	B_P	Γ_U
$h=1$	Null	2.47	6.09*	+	3.12*	3.96	14.33*	+	1.33	4.39	20.87*	+	−0.27
	SW-FF	1.93	2.42*	+	1.91	2.80	2.70*	+	2.43*	2.23	2.70*	−	−0.30
	VAR(4)	3.80	−2.65*	+	−2.88*	3.43	−13.11*	+	−0.31	2.61	−20.78*	+	0.10
	VAR(p)	3.79	−2.65*	+	−2.92*	3.46	−13.13*	+	−0.30	2.59	−20.79*	+	0.11
	BVAR(1)	3.09	2.77*	−	−0.15	3.15	−13.58*	+	0.75	2.22	−20.87*	−	0.60
	BVAR(p)	2.56	−2.68*	+	−0.34	3.25	−13.65*	+	0.91	2.54	−20.83*	−	0.35
	AR(1)	2.97	1.30	−	−0.84	2.43	−13.60*	−	1.84	2.22	−20.94*	+	0.66
	AR(p)	3.04	1.59	−	−0.69	2.74	−13.76*	−	1.72	2.18	−21.11*	+	−0.25
	RW	2.71	0.71	−	−1.71	3.44	−13.40*	+	0.64	2.31	−20.97*	−	0.68
	RWD	2.67	0.64	−	−1.78	3.44	−13.40*	+	0.43	2.54	−21.02*	+	0.71
$h=4$	Null	2.85	6.51*	+	2.95*	2.95	10.02*	+	2.51*	2.29	9.58*	+	0.69
	SW-FF	3.45	0.59	+	−0.07	1.98	1.83	+	1.03	2.41	2.91*	−	−0.54
	VAR(4)	3.70	−1.48	+	−0.87	3.26	−6.45*	+	−0.93	3.20	9.62*	−	−1.05
	VAR(p)	3.71	−1.47	+	−0.86	3.24	−6.44*	+	−0.94	3.19	9.61*	−	−1.06
	BVAR(1)	2.99	−2.80*	−	1.01	4.21	−8.03*	+	−0.44	1.74	−9.66*	+	0.10
	BVAR(p)	2.93	−2.56*	−	1.00	4.37	−7.80*	+	−0.66	1.62	−9.71*	+	−0.01
	AR(1)	2.37	1.41	+	0.14	4.53	−7.42*	+	−0.85	1.43	−9.15*	+	0.34
	AR(p)	2.29	1.77	+	−0.03	4.50	−7.54*	+	−0.79	1.75	−9.36*	+	0.52
	RW	2.56	−1.70	+	−1.64	2.93	−7.17*	+	−0.21	2.21	−9.07*	+	0.56
	RWD	2.57	−1.62	+	−1.80	2.93	−7.13*	+	−0.34	2.22	−9.02*	+	0.56
$h=8$	Null	2.07	6.74*	+	3.08*	2.39	7.56*	+	1.06	3.27	6.59*	+	2.54*
	SW-FF	4.27	−1.72	+	−0.68	2.28	0.99	−	−0.55	4.04	2.21*	−	−1.11
	VAR(4)	3.92	−1.48	+	−1.91	3.20	−3.83*	+	−0.51	1.78	−5.97*	+	0.42
	VAR(p)	3.91	−1.48	+	−1.92	3.18	−3.83*	+	−0.53	1.84	−5.97*	+	0.36
	BVAR(1)	2.76	−3.49*	+	−0.41	4.87*	−5.39*	+	−1.62	2.55	−6.30*	+	0.21
	BVAR(p)	2.78	−3.47*	+	−0.28	4.94*	−5.05*	+	−2.04*	2.40	−6.31*	+	0.08
	AR(1)	2.89	−1.88	−	0.08	5.67*	−4.82*	+	−2.17*	2.19	−5.73*	+	1.87
	AR(p)	2.90	−2.19*	−	0.08	5.50*	−4.99*	+	−2.04*	2.36	−5.86*	+	1.77
	RW	2.50	−2.41*	+	−2.51*	3.44	−3.31*	+	−0.59	2.77	−5.09*	+	1.91
	RWD	2.62	−2.23*	+	−2.68*	3.31	−3.37*	+	−0.50	2.74	−5.08*	+	1.82

Note: Γ_A is the time variation component, Γ_B measures predictability of out-of-sample losses based on in-sample losses, and Γ_U is the unpredictable (over-fitting) component. The $+/-$ signs in the B_P column show whether the model that performed better in sample was also better/worse out of sample. "Null" means absolute forecast error losses of the SW model, while the rest of the rows display the given model's forecasting ability compared to the SW model. Asterisks $(^*)$ show significance at the 5% level.

C Data appendix

This section contains a description of the data used for the estimation exercises, following Smets and Wouters (2007) and Del Negro and Schorfheide (2013). The full sample spans 1966:Q1 – 2013:Q3. The eight observables with the required transformations of the DSGE models are:

- $\text{consumption}_t = \log((\text{PCEC}_t/\text{GDPDEF}_t)/\text{LNSindex}_t) \times 100$,
- $\text{investment}_t = \log((\text{FPI}_t/\text{GDPDEF}_t)/\text{LNSindex}_t) \times 100$,
- $\text{output}_t = \log(\text{GDPC96}_t/\text{LNSindex}_t) \times 100$,
- $\text{hours}_t = \log((\text{PRS85006023}_t \times \text{CE16OV}_t/100)/\text{LNSindex}_t) \times 100$,
- $\text{inflation}_t = \log(\text{GDPDEF}_t/\text{GDPDEF}_{t-1}) \times 100$,
- $\text{real wage}_t = \log(\text{PRS85006103}_t/\text{GDPDEF}_t) \times 100$,
- $\text{interest rate}_t = \text{Federal Funds Rate}_t/4$,
- $\text{spread}_t = \text{BAA10Y}_t$ (only used by the SW-FF model).

The original data are the following:

- GDPC96 : Real Gross Domestic Product - Billions of Chained 1996 Dollars, Seasonally Adjusted Annual Rate.
- GDPDEF : Gross Domestic Product - Implicit Price Deflator - 1996=100, Seasonally Adjusted.
- PCEC : Personal Consumption Expenditures - Billions of Dollars, Seasonally Adjusted Annual Rate.
- FPI : Fixed Private Investment - Billions of Dollars, Seasonally Adjusted Annual Rate.
- CE16OV : Civilian Employment: Sixteen Years & Over, Thousands, Seasonally Adjusted. (CE16OV index: CE16OV normalized such that (1992:Q3)=1.)
- Federal Funds Rate : Averages of Daily Figures - Percent. Before 1954: 3-Month Treasury Bill Rate, Secondary Market Averages of Business Days, Discount Basis.
- LFU800000000 : Population level - 16 Years and Older - Not Seasonally Adjusted.

- LNS10000000 : Labor Force Status : Civilian noninstitutional population - Age : 16 years and over - Seasonally Adjusted - Number in thousands. Before 1976: LFU800000000 : Population level - 16 Years and Older. LNSindex: LNS10000000 normalized such that (1992:Q3)=1.
- PRS85006023 - Nonfarm Business, All Persons, Average Weekly Hours Duration: index, 1992 = 100, Seasonally Adjusted.
- PRS85006103 - Nonfarm Business, All Persons, Hourly Compensation Duration: index, 1992 = 100, Seasonally Adjusted.
- BAA10Y - Moody's Seasoned Baa Corporate Bond Yield Relative to Yield on 10-Year Treasury Constant Maturity.

.

Except for the population and labor force data (series LFU800000000 and LNS10000000), all series were downloaded from the St. Louis Fed's FRED database (`http://research.stlouisfed.org/fred2/`). The two aforementioned series were downloaded from the website of the U.S. Bureau of Labor Statistics (`http://www.bls.gov/`). All data series are the latest, fully revised vintages available as of November, 2013.

Chapter 3

Confidence Intervals for the Strength of Identification

(joint with Atsushi Inoue and Barbara Rossi)

3.1 Introduction

In this paper, we propose a novel methodology to construct confidence intervals for the strength of identification in linear instrumental variables (IV) models. The methodology has several advantages. A first advantage is that it is robust to the presence of weak instruments. It is well-known that the presence of weak instruments invalidates standard inference (Stock et al., 2002). Our methodology provides guidance on the strength of instruments to applied researchers.

A second advantage is that the confidence intervals are straightforward and computationally easy to calculate, as they are obtained from inverting asymptotic chi-squared distributions. The simplicity of our confidence intervals distinguishes our methodology from weak instrument tests, whose distributions are typically non-standard and depend on nuisance parameters that cannot be consistently estimated. For example, Stock and Yogo (2005) suggested an approach to evaluate the severity of the weak instrument problem in specific empirical applications based on the first-stage F-statistic. The first-stage F-statistic is the F-statistic on the strength of the instrument identification. Our complementary approach is instead based on constructing a confidence interval for the strength of identification. It might be surprising that the confidence intervals can be obtained by inverting limiting chi-squared distributions while the test statistics have non-standard limiting distributions. The intuition is that the test statistics are based on the

difference between the estimate of the strength of identification and its value under the null hypothesis, where the null hypothesis is that of weak identification. Hence, the difference between the two contains information on the true strength of identification and how close to zero that is, which cannot be consistently estimated and results in a limiting distribution that is non-standard. Confidence intervals, instead, are based on the difference between the estimate and the true strength of identification, rather than its value under the null hypothesis, whose limiting distribution does not depend on how close to zero the strength of identification is. Interestingly, this is a rather peculiar feature of the weak instrument problem, which cannot be applied to other non-standard situations resulting from the fact that the parameter is local to the null hypothesis, such as confidence intervals for highly persistent (local-to-unity) autoregressive processes. The reason is that, in the local-to-unity framework, the difference between the estimated largest root and its true value is still a function of the local-to-unity parameter in the Ornstein–Uhlenbeck process that approximates the autoregressive process itself. In our weak instrument case, instead, the local-to-zero parameter does not affect the limiting distribution of the variables themselves.

A third advantage of our methodology is that it is general enough to be applied to both IV as well as Structural Vector Autoregressive (SVAR) models with external instruments.[1] It can also be applied in the presence of heteroskedasticity and serial correlation. In fact, with the exception of Montiel Olea and Pflueger (2013), tests for weak instruments in IV regressions assume homoskedasticity and no serial correlation. Since the construction of confidence intervals for the strength of identification is based on inverting a limiting chi-squared distribution, the methodology can be easily applied no matter whether the disturbances are homoskedastic and serially uncorrelated or not – in the latter case, one will simply use a Heteroskedasticity and Autocorrelation Consistent (HAC) estimator to take into account heteroskedasticity and/or serial correlation.

We show in Monte Carlo simulations that our method results in good coverage in finite samples in the homoskedastic IV model, the heteroskedastic and serially correlated IV model and the external instrument SVAR model.

We illustrate the usefulness of our methodology in several empirical applications. The first empirical analysis involves the New Keynesian Phillips Curve. We find that the identification of the parameters is somewhat weak, consistently with several results in the literature (see e.g. Mavroeidis et al. (2014), and Klei-

[1] External instruments are also called proxy variables. In this paper we use the terminology "external instrument".

bergen and Mavroeidis (2009)), although, interestingly, it changed over time. In particular, it has become weaker over time. In the second exercise, we estimate the elasticity of intertemporal substitution using linearized Euler equations. Our confidence intervals confirm that weak identification is indeed a serious problem in this case as well, preventing reliable estimation of the elasticity of intertemporal substitution. In the third empirical example, we analyze the identification of a SVAR model with external instruments, where the instruments are oil shocks. We show that using Hamilton's (2003) oil shocks leads to more precise estimates of the dynamic effects of oil shocks than Kilian's (2008) oil shock series.

Our paper is related to the literature on testing for the strength of instruments in linear IV models. Stock and Yogo (2005) provided critical values for the first-stage F-statistics to test whether instruments are weak in the homoskedastic and serially uncorrelated IV model, while Montiel Olea and Pflueger (2013), derived the limiting distribution of an appropriate first-stage F-statistic under heteroskedasticity and serial correlation when there is only one included endogenous variable. Our paper is also related to the literature on weak identification in SVAR models, in particular Montiel Olea et al. (2016), who construct confidence sets for impulse-response functions which are robust to weak identification, and Lunsford (2016), who formalizes the problem of a weak instrumental variable in an external instrument SVAR model and provides an F-statistic to test if the external instrument is weak. Differently from the papers above, we propose asymptotic confidence intervals for the strength of instruments, which can be used in each of the models discussed earlier.

The remainder of this paper is organized as follows. Section 3.2 describes the econometric frameworks and models we consider, while Section 3.3 provides the results of Monte Carlo simulations. Section 3.4 presents empirical applications illustrating our proposed methodology, and Section 3.5 concludes. The proofs are collected in Appendix A, Appendices B and C contain additional results, and Appendix D provides the description of the data we use in the empirical exercises.

3.2 Econometric frameworks

In this section, we describe the three econometric frameworks we consider, and the corresponding confidence intervals that we propose. Throughout the paper, T denotes the sample size, $\xrightarrow{p}$ and $\xrightarrow{d}$ stand for convergence in probability and in distribution, respectively. The Euclidean norm of vector a is denoted by $\|a\|$, tr $(\cdot)$

is the trace operator, while vec $(\cdot)$ is the vectorization operator. The abbreviation *iid*. stands for independent and identically distributed, $\mathcal{N}(\mu, \mathbb{V})$ is the normal distribution with mean vector μ and covariance matrix $\mathbb{V}$, and for any matrix A, $P_A \equiv A(A'A)^{-1}A'$ and $M_A \equiv I - P_A$.

3.2.1 The homoskedastic IV model

Consider the model of Staiger and Stock (1997) and Stock and Yogo (2005) (henceforth SSY), whose notation we follow:

$$y = Y\beta + X\gamma + u, \tag{3.1}$$

$$Y = Z\Pi + X\Phi + V, \tag{3.2}$$

where y is a $(T \times 1)$ vector and Y is a $(T \times n)$ matrix of included endogenous variables. X is a $(T \times K_1)$ matrix of included exogenous variables (including a column of ones if there is a constant in Equation (3.1)), and Z is a $(T \times K_2)$ matrix of excluded exogenous variables. β is an $(n \times 1)$, while γ is a $(K_1 \times 1)$ vector of coefficients. Π is a matrix of coefficients of dimension $(K_2 \times n)$, and Φ is a $(K_1 \times n)$ matrix of coefficients. Furthermore, u is a $(T \times 1)$ vector of errors, and V is a $(T \times n)$ matrix of errors. Equation (3.1) is the structural equation of interest to the researcher and Equation (3.2) is the first stage equation relating the matrix of endogenous regressor(s) Y to the matrix of instrument(s) Z.[2]

We also define $X_t = (X_{1t}, \ldots, X_{K_1 t})'$, $Z_t = (Z_{1t}, \ldots, Z_{K_2 t})'$, $V_t = (V_{1t}, \ldots, V_{nt})'$, $\underline{Z}_t = (X_t', Z_t')'$ as the vectors of the tth observations of the respective variables. For $t = 1, \ldots, T$, the population second moment matrices Σ and Q are as follows :

$$\Sigma = \mathrm{E}\left[\begin{pmatrix} u_t \\ V_t \end{pmatrix} \begin{pmatrix} u_t & V_t' \end{pmatrix} \right] = \begin{bmatrix} \sigma_{uu} & \Sigma_{uV} \\ \Sigma_{Vu} & \Sigma_{VV} \end{bmatrix}, \tag{3.3}$$

$$Q = \mathrm{E}\left(\underline{Z}_t \underline{Z}_t'\right) = \begin{bmatrix} Q_{XX} & Q_{XZ} \\ Q_{ZX} & Q_{ZZ} \end{bmatrix}, \tag{3.4}$$

which are assumed to be positive definite.

In this section we make the same assumptions as SSY.

Assumption $\mathbf{L_\Pi}$. $\Pi = \Pi_T = C/\sqrt{T}$, *where* C *is a fixed* $K_2 \times n$ *matrix.*

[2]Note that if one uses the lagged dependent variables as exogenous variables in Equation (3.2), then one needs to include those in Equation (3.1) as well.

Assumption M. *The following limits hold jointly for fixed K_2 as $T \to \infty$:*

(a) $(T^{-1}u'u, T^{-1}V'u, T^{-1}V'V) \xrightarrow{p} (\sigma_{uu}, \Sigma_{Vu}, \Sigma_{VV})$,

(b) $T^{-1}\underline{Z}'\underline{Z} \xrightarrow{p} Q$,

(c) $(T^{-1/2}X'u, T^{-1/2}Z'u, T^{-1/2}X'V, T^{-1/2}Z'V) \xrightarrow{d} (\Psi_{Xu}, \Psi_{Zu}, \Psi_{XV}, \Psi_{ZV})$, *where* $\Psi \equiv [\Psi_{Xu}', \Psi_{Zu}', vec(\Psi_{XV})', vec(\Psi_{ZV})']' \sim \mathcal{N}(0, \Sigma \otimes Q)$.

Assumption L_Π models Π as local to zero, leading to weak instruments, while Assumption M ensures that the appropriately scaled moments of the errors and the variables obey a weak law of large numbers and a central limit theorem. Part (c) of Assumption M corresponds most naturally to serially uncorrelated and conditionally homoskedastic errors, which is restrictive in a number of applications. This assumption will be substantially relaxed in Section 3.2.2.

When there is only one endogenous regressor in Equation (3.1), that is, $n = 1$, then Σ_{VV} is a scalar σ_v^2, and in the absence of included exogenous regressors X, the concentration parameter is defined as:

$$\mu^2 \equiv \Pi' Z' Z \Pi / \sigma_v^2 , \tag{3.5}$$

which plays a role similar to that of the sample size when deriving the asymptotic distribution of the two-stage least squares (TSLS) estimator with fixed instruments and *iid*. normal errors, as Rothenberg (1984) demonstrated.

As Stock and Yogo (2005) showed, the asymptotic maximum bias (relative to the OLS estimator) of a number of k-class instrumental variables estimators (TSLS, limited information maximum likelihood, Fuller-k and bias-adjusted TSLS) and the asymptotic maximum size distortion of Wald-tests on β can be characterized in terms of the minimum eigenvalue of the matrix $\Lambda \equiv \lambda'\lambda / K_2$, where $\lambda \equiv \Omega^{1/2} C \Sigma_{VV}^{-1/2}$ and $\Omega \equiv Q_{ZZ} - Q_{ZX} Q_{XX}^{-1} Q_{XZ}$. They also note that this matrix is the weak instrument limit of the concentration matrix, the multivariate analog of μ^2 defined in Equation (3.5):

$$\Xi = \frac{1}{K_2} \Sigma_{VV}^{-1/2\prime} \Pi Z' Z \Pi \Sigma_{VV}^{-1/2} \xrightarrow{p} \Lambda . \tag{3.6}$$

When $n \geq 1$, the matrix analog of the first-stage F-statistic testing the null hypothesis $\Pi = 0$ is

$$G_T = \frac{1}{K_2} \widehat{\Sigma}_{VV}^{-\frac{1}{2}\prime} Y^{\perp\prime} P_{Z^\perp} Y^\perp \widehat{\Sigma}_{VV}^{-\frac{1}{2}} ,$$

where $\widehat{\Sigma}_{VV} = Y'M_{\underline{Z}}Y/(T - K_1 - K_2)$, $\underline{Z} = [X\,Z]$, $Y^{\perp} \equiv M_X Y$ and $Z^{\perp} \equiv M_X Z$. The Cragg and Donald (1993) and Stock and Yogo (2005) test statistic is the minimum eigenvalue of G_T, $g_{\min} = \text{mineval}(G_T)$. As Stock and Yogo (2005) showed, the asymptotic distributions of G_T and $g_{\min}$ are

$$G_T \xrightarrow{d} \nu_1/K_2\,, \tag{3.7}$$

$$g_{\min} \xrightarrow{d} \text{mineval}(\nu_1/K_2)\,, \tag{3.8}$$

respectively, where ν_1 has a non-central Wishart distribution with noncentrality matrix $\lambda'\lambda = K_2\Lambda$. While formally G_T is the usual F-statistic testing $\Pi = 0$, its asymptotic distribution is derived under Assumption L_{Π}, yielding this non-standard result. In contrast, our procedure circumvents this problem by building on an appropriate distance between the OLS estimate of Π_T and its true value.

Building on the asymptotic distribution of $g_{\min}$, Stock and Yogo's (2005) procedure tests whether the instruments are strong enough either in terms of being less biased than a pre-specified tolerance, or if the Wald-test on β does not display higher size distortion than a threshold. While their method cannot provide a confidence set for $\text{mineval}(\Lambda)$, our method is specifically designed to do so, offering guidance on *how* weak or strong the instruments are.

For the asymptotic theory to be developed, it is convenient to project out the exogenous regressors X. That is, let us define $V^{\perp} \equiv M_X V$, $y^{\perp} \equiv M_X y$ and $u^{\perp} \equiv M_X u$ in addition to $Y^{\perp}$ and $Z^{\perp}$ defined earlier. Using this notation, Equation (3.2) reads as

$$Y^{\perp} = Z^{\perp}\Pi + V^{\perp}\,. \tag{3.9}$$

Let $Z_t^{\perp\prime}$ and $V_t^{\perp\prime}$ be the tth row of $Z^{\perp}$ and $V^{\perp}$, respectively. By the exogeneity of X, $\mathrm{E}(X_t V_t') = 0$, hence $\Sigma_{V^{\perp}V^{\perp}} \equiv \mathrm{E}\left(V_t^{\perp} V_t^{\perp\prime}\right) = \Sigma_{VV}$. Furthermore, simple algebra shows that $\Omega \equiv Q_{ZZ} - Q_{ZX}Q_{XX}^{-1}Q_{XZ} = Q_{Z^{\perp}Z^{\perp}}$, where $Q_{Z^{\perp}Z^{\perp}} \equiv \mathrm{E}\left(Z_t^{\perp} Z_t^{\perp\prime}\right)$.

Our proposed confidence interval builds on the asymptotic distribution of the OLS estimator of Π_T in Equation (3.9), denoted by $\widehat{\Pi}_T$:

$$\sqrt{T}\left(\widehat{\Pi}_T - \Pi_T\right) = \left(T^{-1}Z^{\perp\prime}Z^{\perp}\right)^{-1} T^{-1/2}Z^{\perp\prime}V^{\perp} \tag{3.10}$$

$$\sqrt{T}\,\text{vec}\left(\widehat{\Pi}_T - \Pi_T\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{V^{\perp}V^{\perp}} \otimes Q_{Z^{\perp}Z^{\perp}}^{-1}\right)\,, \tag{3.11}$$

$$\text{vec}\left(\widehat{C} - C\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{V^{\perp}V^{\perp}} \otimes Q_{Z^{\perp}Z^{\perp}}^{-1}\right)\,, \tag{3.12}$$

where Equation (3.11) follows from Assumption M, and in Equation (3.12) we

used that $\Pi_T = C/\sqrt{T}$, and $\widehat{C} = \widehat{\Pi}_T\sqrt{T}$.[3] Next, consider the Wald statistic

$$\left[\text{vec}\left(\widehat{C} - C\right)\right]' \left[\Sigma_{V^\perp V^\perp} \otimes Q^{-1}_{Z^\perp Z^\perp}\right]^{-1} \left[\text{vec}\left(\widehat{C} - C\right)\right] \xrightarrow{d} \chi^2_{nK_2}, \tag{3.13}$$

where $\chi^2_{nK_2}$ stands for a chi-squared random variable with nK_2 degrees of freedom. Using $\widehat{\Sigma}_{V^\perp V^\perp} = \widehat{V}^{\perp\prime}\widehat{V}^\perp/T \xrightarrow{p} \Sigma_{V^\perp V^\perp}$, where $\widehat{V}^\perp$ is the matrix of OLS residuals, and $\widehat{Q}_{Z^\perp Z^\perp} = Z^{\perp\prime}Z^\perp/T \xrightarrow{p} Q_{Z^\perp Z^\perp}$ (both follow from Assumption M), Slutsky's theorem and the continuous mapping theorem imply

$$\mathcal{W}(C) \equiv \left[\text{vec}\left(\widehat{C} - C\right)\right]' \left[\widehat{\Sigma}_{V^\perp V^\perp} \otimes \widehat{Q}^{-1}_{Z^\perp Z^\perp}\right]^{-1} \left[\text{vec}\left(\widehat{C} - C\right)\right] \xrightarrow{d} \chi^2_{nK_2}. \tag{3.14}$$

By taking the $(1-\alpha)$ quantile of the $\chi^2_{nK_2}$ distribution (denoted by $\chi^2_{nK_2,1-\alpha}$), the Wald statistic $\mathcal{W}(C)$ can be inverted to obtain an asymptotically valid $(1-\alpha)$ level confidence set for C, which is formally defined as

$$\text{CI}^C_{1-\alpha} \equiv \left\{\forall\, \widetilde{C} \in \mathbb{R}^{K_2\times n} : \mathcal{W}\left(\widetilde{C}\right) \leq \chi^2_{nK_2,1-\alpha}\right\}. \tag{3.15}$$

Recall that $\Lambda = \Sigma^{-1/2'}_{V^\perp V^\perp} C' Q_{Z^\perp Z^\perp} C \Sigma^{-1/2}_{V^\perp V^\perp}/K_2$. Let us define

$$\widetilde{\Lambda}(\widetilde{C}) \equiv \widehat{\Sigma}^{-1/2'}_{V^\perp V^\perp} \widetilde{C}' \widehat{Q}_{Z^\perp Z^\perp} \widetilde{C} \widehat{\Sigma}^{-1/2}_{V^\perp V^\perp}/K_2, \tag{3.16}$$

which is a continuous function of $\widetilde{C}$, and the consistent estimates of $\Sigma_{V^\perp V^\perp}$ and $Q_{Z^\perp Z^\perp}$ have replaced their population counterparts. Our proposed $(1-\alpha)$ level asymptotic confidence interval for mineval(Λ) is

$$\text{CI}^\Lambda_{1-\alpha} \equiv \left[\min_{\widetilde{C}\in \text{CI}^C_{1-\alpha}} \text{mineval}(\widetilde{\Lambda}(\widetilde{C})), \max_{\widetilde{C}\in \text{CI}^C_{1-\alpha}} \text{mineval}(\widetilde{\Lambda}(\widetilde{C}))\right]. \tag{3.17}$$

We summarize our results in the following proposition.

Proposition 1 (Confidence Interval Validity Under Homoskedasticity and Uncorrelatedness)**.** *Under Assumptions L_Π and M, $CI^\Lambda_{1-\alpha}$ is an asymptotically valid confidence interval for mineval (Λ), that is* $\lim_{T\to\infty} P\left(\textit{mineval}\,(\Lambda) \in CI^\Lambda_{1-\alpha}\right) \geq 1-\alpha$.

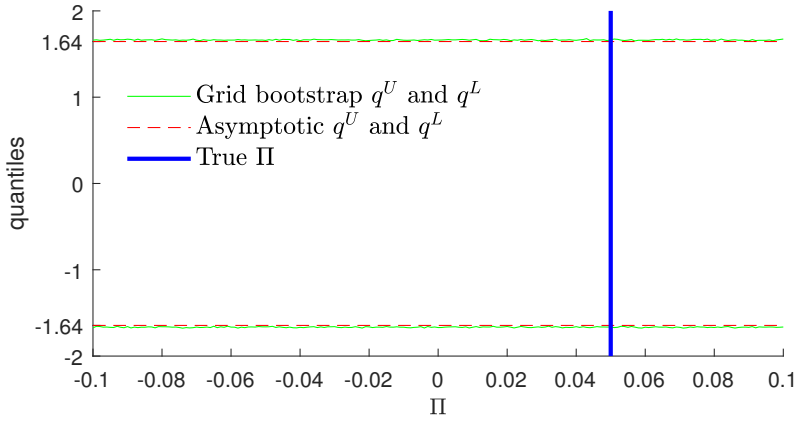
Proof. See Appendix A. ■

We note that as $\widetilde{\Lambda}\left(\widetilde{C}\right)$ is not a one-to-one function of $\widetilde{C}$ in general, our proposed confidence interval is conservative.

[3] Note that $\widehat{C}$ is an inconsistent estimator of C. However, for our purposes the asymptotic normality result of Equation (3.12) is sufficient.

Example 1. To illustrate the validity and accuracy of the asymptotic normal approximation, consider a small Monte Carlo study. Let us specify $Y = Z\Pi_T + X\Phi + V$ such that $(Z_t, V_t)' \overset{\text{iid.}}{\sim} \mathcal{N}(0, I_2)$, $\Pi_T = C/\sqrt{T}$ with $C = 0.5$ and $T = 100$, and $X_t = 1$ with $\Phi = 1$. Hansen's (1999) grid bootstrap is asymptotically valid in the presence of a weak instrument. Using the grid bootstrap, we can simulate the distribution of the usual t-statistic testing the null hypothesis of $\Pi = \Pi_0$ at each point on a fine grid A_G, which we specify as ranging from -0.1 to 0.1, with increments of 0.01. At each point on A_G, we simulate the distribution of the t-statistic $B = 999$ times (by resampling the estimated residuals with replacement), and estimate the 5th and 95th percentiles (q^L and q^U) of the simulated distribution. We repeated the above exercise 200 times, and calculated the means of q^L and q^U at each point on A_G across the 200 replications. The results shown in Figure 3.1 confirm that the simulated quantiles of the t-statistic are constant and virtually indistinguishable from their asymptotic counterparts (± 1.64). ▲

Figure 3.1: The grid bootstrap and asymptotic quantiles of the t-statistic

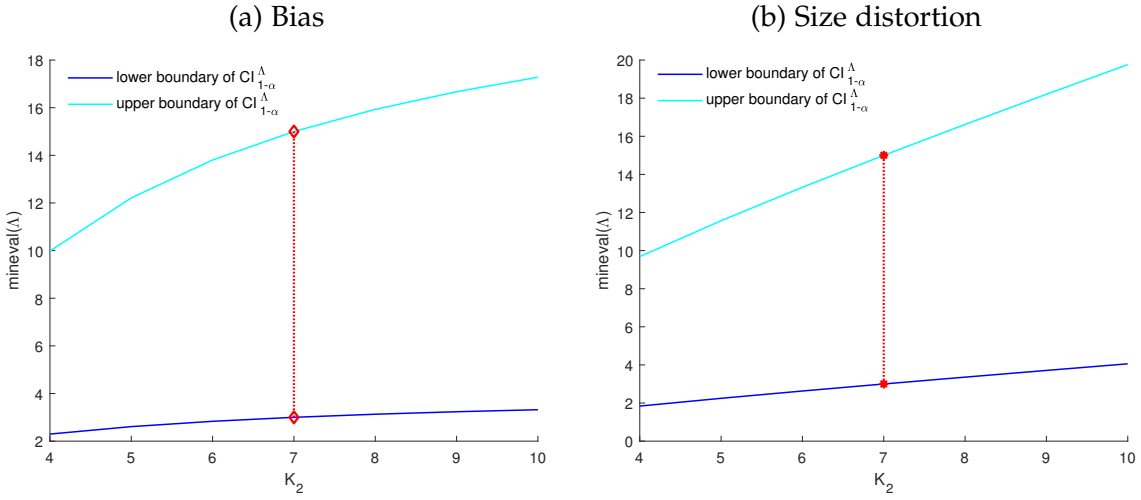


Having obtained a $(1-\alpha)$ level confidence interval for mineval (Λ), researchers can summarize the strength of identification by calculating the values of lower and upper maximum bias and size distortion corresponding to the upper and lower endpoints of the confidence interval. In Appendix B we provide the boundary values of mineval (Λ) for $n = \{1, 2, 3\}$ endogenous variables and $K_2 = n+2, \ldots, 20$ (bias) and $K_2 = n, \ldots, 20$ (size distortion) for a fine grid of maximum bias and size distortion for the TSLS estimator, extending the simulation results of Stock and Yogo (2005).[4]

[4]Following Stock and Yogo (2005), we calculated the size distortion assuming the Wald test on β has a nominal level 5%.

Example 2. Figure 3.2 illustrates the proposed procedure. Let us assume that at the $(1-\alpha) = 0.90$ level we obtain the confidence interval $\text{CI}^{\Lambda}_{0.90} = [3, 15]$ when using $n = 2$ endogenous regressors and $K_2 = 7$ instruments. Based on the simulation results, we looked up the corresponding bias and size distortion values, and plotted the resulting intervals (dashed vertical lines), along with the boundary values of mineval(Λ) that would lead to the same amount of bias and size distortion. We obtained the interval $[0.04, 0.19]$ for maximum bias, and $[0.11, 0.34]$ for size distortion. ▲

Figure 3.2: An example of using $\text{CI}^{\Lambda}_{1-\alpha}$



Note: In the left (right) panel, we calculated the maximum bias (size distortion) corresponding to $\text{CI}^{\Lambda}_{1-\alpha} = [3, 15]$ (dashed vertical lines) for $n = 2$ endogenous regressors and $K_2 = 7$ instruments. The diamond and the asterisk markers show the resulting intervals, while the curves serve illustrative purpose only, showing the confidence interval for mineval(Λ) while keeping the bias and size distortion at the same value.

3.2.2 The heteroskedastic/autocorrelated linear IV model

The assumption of homoskedastic errors used in the previous section is restrictive in a number of applications. In those cases, following the Stock and Yogo (2005) testing method or applying our proposed confidence interval could lead to incorrect inference on the strength of instruments. As a solution to this problem, Montiel Olea and Pflueger (2013) proposed a measure of the strength of instruments which applies to general (heteroskedastic, autocorrelated or clustered) errors, but assumes there is only one endogenous regressor ($n = 1$).

Following Montiel Olea and Pflueger (2013), consider the linear IV model in

its reduced form, where the exogenous regressors X have been projected out:

$$y^{\perp} = Z^{\perp}\Pi\beta + v_1, \tag{3.18}$$

$$Y^{\perp} = Z^{\perp}\Pi + v_2\,, \tag{3.19}$$

where Equation (3.18) is the structural equation of interest in reduced form, while Equation (3.19) is the first stage equation linking the endogenous regressor $Y^{\perp}$ with the instruments $Z^{\perp}$. Both $y^{\perp}$ and $Y^{\perp}$ are $(T \times 1)$ vectors, $Z^{\perp}$ is a $(T \times K_2)$ matrix of instruments, β is a scalar coefficient, Π is a $(K_2 \times 1)$ vector of coefficients, while $v_1 \equiv V^{\perp}\beta + u^{\perp}$ and $v_2 \equiv V^{\perp}$ are $(T \times 1)$ vectors of errors. Furthermore, $Z^{\perp}$ is orthogonalized such that $Z^{\perp\prime}Z^{\perp}/T = I_{K_2}$.

Montiel Olea and Pflueger (2013) adopt Assumption L_{Π} of SSY to model weak instruments, but considerably weaken their moment assumptions as follows:

Assumption HL. *The following limits hold as $T \to \infty$:*

(a) $\begin{pmatrix} T^{-1/2}Z^{\perp\prime}v_1 \\ T^{-1/2}Z^{\perp\prime}v_2 \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, W)$ *for some positive definite* $W = \begin{pmatrix} W_1 & W_{12} \\ W_{12}' & W_2 \end{pmatrix}$,

(b) $[v_1\, v_2]'\,[v_1\, v_2]\,/T \xrightarrow{p} K$ *for some positive definite* $K = \begin{pmatrix} \kappa_1^2 & \kappa_{12} \\ \kappa_{12} & \kappa_2^2 \end{pmatrix}$,

(c) There exists a sequence of positive definite estimates $\widehat{W}$, measurable with respect to $\{y_t^{\perp}, Y_t^{\perp}, Z_t^{\perp}\}_{t=1}^{T}$, such that $\widehat{W} \xrightarrow{p} W$.

As we can see, unlike Assumption M of SSY, these high level assumptions do not restrict W to take the form of $K \otimes I_{K_2}$, and therefore they can encompass a wide range of error structures, including heteroskedastic, autocorrelated or clustered (in panel data settings) error terms.

Montiel Olea and Pflueger (2013) focus on the TSLS and the Limited Information Maximum Likelihood (LIML) estimators, and their testing procedure is designed to decide if the instruments $Z^{\perp}$ are such that the Nagar (1959) bias of the TSLS or the LIML estimators exceeds a given fraction τ relative to a worst-case benchmark. Formally, their Theorem 1 shows that the Nagar (1959) bias of the estimators $e \in \{\text{TSLS, LIML}\}$ is given by

$$N_e(\beta, C, W, K) = \mu^{-2} n_e(\beta, \overline{C}, W, K)\,, \tag{3.20}$$

where

$$n_{\text{TSLS}}(\beta, \overline{C}, W, K) = \frac{\text{tr}(S_{12})}{S_2}\left(1 - 2\frac{\overline{C}' S_{12}\overline{C}}{\text{tr}(S_{12})}\right), \tag{3.21}$$

$$n_{\text{LIML}}(\beta, \overline{C}, W, K) = \frac{1}{\text{tr}(S_2)}\left(\text{tr}(S_{12}) - \frac{\sigma_{12}}{\sigma_1^2}\text{tr}(S_1) - \overline{C}'\left(2S_{12} - \frac{\sigma_{12}}{\sigma_1^2}S_1\right)\overline{C}\right), \tag{3.22}$$

and where C is written as $C = \|C\|\overline{C}$,

$$\mu^2 \equiv \|C\|^2 / \text{tr}(W_2), \tag{3.23}$$

$S_1 = W_1 - 2\beta W_{12} + \beta^2 W_2$, $S_{12} = W_{12} - \beta W_2$, $S_2 = W_2$, $\sigma_1^2 = \kappa_1^2 - 2\beta\kappa_{12} + \beta^2\kappa_2^2$, $\sigma_{12} = \kappa_{12} - \beta\kappa_2^2$, and $\sigma_2^2 = \kappa_2^2$. The benchmark bias is defined as $\text{BM}(\beta, W) \equiv \sqrt{\text{tr}(S_1)/\text{tr}(S_2)}$.

Their null and alternative hypotheses are formulated for a given threshold $\tau \in [0, 1]$, long-run covariance matrix W, covariance matrix K, and estimator $e \in \{\text{TSLS, LIML}\}$ as

$$H_e^0 : \mu^2 \in \mathcal{H}_e(W, K) \text{ versus } H_e^1 : \mu^2 \notin \mathcal{H}_e(W, K), \tag{3.24}$$

where $\mathcal{H}_e(W, K)$ is the set of μ^2 for which the absolute value of the relative Nagar (1959) bias exceeds the threshold τ set by the researcher, formally:

$$\mathcal{H}_e(W, K) \equiv \left\{\mu^2 \in \mathbb{R}_+ : \sup_{\beta \in \mathbb{R}, \overline{C} \in \mathcal{S}^{K_2 - 1}} \frac{|N_e(\beta, \mu\sqrt{\text{tr}(W_2)}\,\overline{C}, W, K)|}{\text{BM}(\beta, W)} > \tau\right\}, \tag{3.25}$$

where $\mathcal{S}^{K_2 - 1}$ is the $K_2 - 1$ dimensional unit sphere.

Montiel Olea and Pflueger (2013) propose the so-called *effective first-stage F-statistic* defined as

$$\widehat{F}_{\text{eff}} \equiv \frac{Y^{\perp\prime}(Z^\perp Z^{\perp\prime}/T)Y^\perp}{\text{tr}(\widehat{W}_2)}, \tag{3.26}$$

and provide a method to test the null hypothesis of weak instruments. However, their procedure cannot guide researchers on *how* weak or strong their instruments are. On the other hand, our proposed methodology allows researchers to go beyond hypothesis testing by providing an asymptotic confidence interval for the parameter μ^2, which determines the strength of the instruments.

Our proposed confidence interval for μ^2 is constructed similarly to that of Section 3.2.1. In particular, consider the asymptotic distribution of the OLS

estimator of Π_T in Equation (3.19):

$$\sqrt{T}\left(\widehat{\Pi}_T - \Pi_T\right) = T^{-1/2} Z^{\perp\prime} v_2 \,, \tag{3.27}$$

$$\sqrt{T}\left(\widehat{\Pi}_T - \Pi_T\right) \xrightarrow{d} \mathcal{N}\left(0, W_2\right) \,, \tag{3.28}$$

$$\widehat{C} - C \xrightarrow{d} \mathcal{N}\left(0, W_2\right) \,, \tag{3.29}$$

where we used the normalization $Z^{\perp\prime} Z^{\perp}/T = I_{K_2}$, the central limit theorem of Assumption HL, and Assumption L_Π. Moreover, by Slutsky's theorem, the continuous mapping theorem and part (c) of Assumption HL, the Wald statistic is asymptotically chi-squared distributed with K_2 degrees of freedom, formally

$$\mathcal{W}(C) \equiv \left(\widehat{C} - C\right)' \widehat{W}_2^{-1} \left(\widehat{C} - C\right) \xrightarrow{d} \chi^2_{K_2} \,, \tag{3.30}$$

where $\widehat{W}_2$ is the lower right $(K_2 \times K_2)$ block of the consistent estimator $\widehat{W}$. Analogously to the procedure in Section 3.2.1, by taking the $(1-\alpha)$ quantile of the $\chi^2_{K_2}$ distribution (denoted by $\chi^2_{K_2,1-\alpha}$), inverting the Wald statistic $\mathcal{W}(C)$ yields an asymptotically valid $(1-\alpha)$ level confidence set for C, which is defined as

$$\mathrm{CI}^C_{1-\alpha} \equiv \left\{ \forall\, \widetilde{C} \in \mathbb{R}^{K_2} : \mathcal{W}(\widetilde{C}) \leq \chi^2_{K_2,1-\alpha} \right\} . \tag{3.31}$$

Recall that $\mu^2 \equiv \|C\|^2 / \operatorname{tr}(W_2)$. Let us define

$$\widetilde{\mu}^2(\widetilde{C}) \equiv \|\widetilde{C}\|^2 / \operatorname{tr}(\widehat{W}_2) \,, \tag{3.32}$$

which is a function of $\widetilde{C}$, and the consistent estimator $\widehat{W}_2$ has replaced its population counterpart W_2. Our proposed $(1-\alpha)$ level asymptotic confidence interval for μ^2 is

$$\mathrm{CI}^{\mu^2}_{1-\alpha} \equiv \left[\min_{\widetilde{C} \in \mathrm{CI}^C_{1-\alpha}} \widetilde{\mu}^2\left(\widetilde{C}\right), \max_{\widetilde{C} \in \mathrm{CI}^C_{1-\alpha}} \widetilde{\mu}^2\left(\widetilde{C}\right) \right] . \tag{3.33}$$

We summarize our results in the following proposition.

Proposition 2 (Confidence Interval Validity Under Heteroskedasticity and Autocorrelation)**.** *Under Assumptions* L_Π *and HL,* $CI^{\mu^2}_{1-\alpha}$ *is an asymptotically valid confidence interval for* μ^2*, that is* $\lim_{T\to\infty} P\left(\mu^2 \in CI^{\mu^2}_{1-\alpha}\right) \geq 1-\alpha$*.*

Proof. See Appendix A. ■

We note that as $\widetilde{\mu}^2\left(\widetilde{C}\right)$ is not a one-to-one function of $\widetilde{C}$ in general, our proposed confidence interval is conservative.

Montiel Olea and Pflueger (2013) note that the weak instrument set for μ^2 takes the form $\mathcal{H}_e(W,K) = [0, B_e(W,K)/\tau)$, where

$$B_e(W,K) \equiv \sup_{\beta\in\mathbb{R},\bar{C}\in\mathcal{S}^{K_2-1}} \frac{|n_e(\beta,\bar{C},W,K)|}{\mathrm{BM}(\beta,W)} < \infty. \tag{3.34}$$

Using $\widehat{W}$ and $\widehat{K} \equiv [\hat{v}_1\,\hat{v}_2]'[\hat{v}_1\,\hat{v}_2]/T$, where $[\hat{v}_1\,\hat{v}_2]$ is the matrix of OLS residuals calculated from Equations (3.18) and (3.19), we propose that researchers could summarize the strength of instruments by calculating the endpoints of the interval

$$\widehat{\tau}_L^U \equiv \left[\widehat{\tau}_L, \widehat{\tau}^U\right], \tag{3.35}$$

where[5]

$$\widehat{\tau}_L \equiv \min\left(1, \left[\max_{\widetilde{C}\in \mathrm{CI}^C_{1-\alpha}} \widetilde{\mu}^2\left(\widetilde{C}\right)\right]^{-1} B_e\left(\widehat{W},\widehat{K}\right)\right), \tag{3.36}$$

$$\widehat{\tau}_U \equiv \min\left(1, \left[\min_{\widetilde{C}\in \mathrm{CI}^C_{1-\alpha}} \widetilde{\mu}^2\left(\widetilde{C}\right)\right]^{-1} B_e\left(\widehat{W},\widehat{K}\right)\right). \tag{3.37}$$

3.2.3 The external instrument SVAR model

Since the papers of Stock and Watson (2012) and Mertens and Ravn (2013), the macroeconomics literature has frequently used the external instrument identification approach to estimate dynamic effects of various macroeconomic shocks of interest (tax, monetary, oil price shocks, etc.) in SVARs as an alternative to more traditional identification schemes, such as Cholesky or sign restrictions. In a simplified way, this identification approach relies on finding an observable variable z_t not contained in the VAR which satisfies two conditions: it is correlated with the (unobserved) shock of interest (relevance), and it is uncorrelated with all the other structural shocks (exogeneity). This choice of vocabulary parallels that of the IV literature, already suggesting that the estimators proposed in the aforementioned papers can be written as IV estimators.

Recently Lunsford (2016) proposed a test of instrument strength in this framework, although using a slightly different estimator than Montiel Olea et al. (2012). However, similarly to the tests of Stock and Yogo (2005) and Montiel Olea and Pflueger (2013), his methodology is not suited either to inform researchers on *how* weak or strong their external instruments are. In this section, we first des-

[5]Clearly, the proposed interval $\widehat{\tau}_L^U$ is meaningful only when the ratios are well-defined.

cribe Montiel Olea et al.'s (2012) approach, and then propose a valid confidence interval for the parameter that determines the strength of identification in their framework.

In this section we use a notation that is specific to the SVAR literature with external instruments. Let $U_j = \left(u_{j,1}, \ldots, u_{j,T}\right)'$, $E_j = \left(e_{j,1}, \ldots, e_{j,T}\right)'$, $N_j = \left(\nu_{j,1}, \ldots, \nu_{j,T}\right)'$, $E = (\epsilon_1, \ldots, \epsilon_T)'$. Table 3.1 summarizes how the notation in this section differs from the notation in Sections 3.2.1 and 3.2.2.

Table 3.1: Correspondence between the variables used in the external instrument SVAR, the homoskedastic and the heteroskedastic/autocorrelated IV models

External instrument SVAR	Homoskedastic IV	Heteroskedastic/autocorrelated IV
U_j	$y^\perp$	$y^\perp$
U_1	$Y^\perp$	$Y^\perp$
b_{j1}	β	β
E_j	$u^\perp$	$u^\perp$
π_T	Π_T	Π_T
Z	$Z^\perp$	$Z^\perp$
N_j	$V^\perp\beta + u^\perp$	v_1
E	$V^\perp$	v_2

Note: The objects with subscript j in the first column correspond to the objects in the second and third columns for a given $j = 2, \ldots k$.

Consider the VAR(p)

$$Y_t = A_0 + \sum_{l=1}^{p} A_l Y_{t-l} + u_t\,, \tag{3.38}$$

where Y_t is an $(k \times 1)$ vector of variables and u_t is an $(k \times 1)$ vector of VAR innovations. The coefficients are collected in $\mathrm{Y} \equiv \mathrm{vec}\left(A_0, A_1, \ldots, A_p\right)$. The $(k \times 1)$ vector of structural shocks v_t is given by

$$u_t = B v_t\,, \tag{3.39}$$

where B is an invertible $(k \times k)$ matrix.

The external instrument SVAR literature focuses on estimating the column of B corresponding to the structural shock, and without loss of generality we assume this is the first column, denoted by B_1.

Before introducing the SVAR assumptions, it is helpful to rewrite Equation (3.39) as

$$\begin{bmatrix} u_{1,t} \\ (1\times 1) \\ u_{2,t} \\ (k-1)\times 1 \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ (1\times 1) & 1\times(k-1) \\ b_{21} & b_{22} \\ (k-1)\times 1 & (k-1)\times(k-1) \end{bmatrix} \begin{bmatrix} v_{1,t} \\ (1\times 1) \\ v_{2,t} \\ (k-1)\times 1 \end{bmatrix}, \tag{3.40}$$

where $v_{1,t}$ is the structural shock of interest, $v_{2,t}$ contains all the other $k-1$ structural shocks, and $B_1 = (b_{11}, b_{21}')'$ is the first column of B. Following Montiel Olea et al. (2012), we normalize $b_{11} = 1$. The external instrument SVAR identifies B_1 by finding a variable z_t not in Y_t such that the relevance $\mathrm{E}(v_{1,t}z_t) = \pi \neq 0$ and the exogeneity conditions $\mathrm{E}(v_{2,t}z_t) = 0$ hold. Let us define $Z \equiv (z_1, \ldots, z_T)'$, $\mathcal{Y}_t \equiv \left(1, Y_{t-1}', \ldots, Y_{t-p}'\right)'$, and $\mathcal{Y} \equiv (\mathcal{Y}_1, \ldots, \mathcal{Y}_T)'$.

We make the following set of assumptions regarding the SVAR in Equations (3.38) and (3.39).

Assumption SVAR.

(a) The VAR(p) given in Equation (3.38) *is stationary, and there exists a consistent and asymptotically normal estimator of the coefficients, denoted by* $\widehat{\mathrm{Y}}$*, that is* $\widehat{\mathrm{Y}} \xrightarrow{p} \mathrm{Y}$*, and* $\sqrt{T}\left(\widehat{\mathrm{Y}} - \mathrm{Y}\right) \xrightarrow{d} \mathcal{N}(0, \Theta)$*, where* Θ *is positive definite,*

(b) B *is a fixed, invertible* $k \times k$ *matrix,*

(c) $\mathrm{E}(v_t) = 0, \mathrm{E}(v_t v_t') = \Sigma_v$*, where* Σ_v *is positive definite, and the structural shock of interest* $v_{1,t}$ *is uncorrelated with the remaining structural shocks,* $\mathrm{E}\left(v_{1,t}v_{j,t}\right) = 0$ *for* $j = 2, \ldots, k$*,*

(d) $T^{-1}(\mathcal{Y}'Z) \xrightarrow{p} 0$*,*

(e) $\mathrm{E}\left(z_t^2\right) = 1$*.*

Part (a) is a standard assumption in the VAR literature, for a set of primitive conditions we refer to Hamilton (1994, Chapter 11, pp. 298-299). The matrix B in part (b) links the structural shocks v_t to the VAR innovations u_t. Part (c) contains moment conditions which shocks are expected to fulfill to label them as structural. Part (d) (along with part (a)) allows using consistent estimates $\widehat{u}_t$ in place of the unobserved VAR innovations u_t without changing the asymptotic properties of the estimator described below, while part (e) serves as a convenient normalization in the asymptotic theory to be developed.

Assumption E-SVAR. *z_t is exogenous with respect to $v_{2,t}$, that is* $\mathrm{E}(v_{2,t}z_t) = 0$*.*

Assumption R-SVAR. *z_t is relevant for $v_{1,t}$, that is* $\mathrm{E}(v_{1,t}z_t) = \pi \neq 0$.

Assumptions E-SVAR and R-SVAR are crucial exogeneity and relevance conditions for the estimator introduced below.

Under similar assumptions, Montiel Olea et al. (2012) and Lunsford (2016) provided consistent estimators of b_{21} and B_1, respectively.[6] In particular, Montiel Olea et al.'s (2012) estimator is

$$\widehat{b}_{21} = \frac{T^{-1}\sum_{t=1}^{T}\hat{u}_{2,t}z_t}{T^{-1}\sum_{t=1}^{T}\hat{u}_{1,t}z_t}, \tag{3.41}$$

where $\hat{u}_{1,t}$ and $\hat{u}_{2,t}$ are the estimates of the VAR residuals in Equation (3.38). Montiel Olea et al. (2012) showed that if a consistent and asymptotically normal estimator is used to estimate the VAR innovations, and part (d) of Assumption SVAR holds, then using the estimated residuals instead of the unobserved VAR innovations does not change the asymptotic properties of the estimator in Equation (3.41). This condition can be ensured by regressing z_t on $\mathcal{Y}_t$, and using the residuals $\widehat{z}_t$ in place of z_t. Furthermore, part (e) of Assumption SVAR can be ensured by standardizing the residuals, which does not affect the asymptotic properties of $\widehat{b}_{21}$. For simplicity of notation, we will continue using z_t to denote the standardized residuals, and use u_t in place of $\widehat{u}_t$.

Note that Equation (3.41) has the form of an IV estimator, where the structural and the first stage equations are given by

$$u_{j,t} = b_{j1}u_{1,t} + e_{j,t}\,, \text{ for all } j = 2,\ldots,k \tag{3.42}$$

$$u_{1,t} = \pi z_t + \nu_t\,, \tag{3.43}$$

where the same first stage equation is used for each structural equation. It is important to realize that using the estimator in Equation (3.41) is equivalent to using the $k-1$ TSLS estimators given by Equations (3.42) and (3.43).

When the correlation between z_t and $v_{1,t}$ is modeled as local to zero (similarly to the approach taken by Staiger and Stock (1997) and Stock and Yogo (2005)), both Montiel Olea et al.'s (2012) and Lunsford's (2016) estimators become inconsistent. Following the method of Stock and Yogo (2005), Lunsford (2016) provides a characterization of the weak instrument set in terms of the bias of the estimator of B_1 and proposes a test for a weak external instrument. In this paper, we model a weak external instrument as Montiel Olea et al. (2012) and Lunsford (2016), but by using the proposed estimator of the former authors, we are able to link the

[6] Lunsford (2016) did not use the normalization $b_{11} = 1$.

external instrument SVAR framework with the familiar linear IV frameworks of Sections 3.2.1 and 3.2.2, and therefore show how our previous results carry over to this case.

A weakly relevant external instrument is modeled by replacing Assumption R-SVAR by the following assumption:

Assumption π_T. $\mathrm{E}\left(v_{1,t}z_t\right) = \pi_T = C/\sqrt{T}$, *where C is a fixed, nonzero scalar.*

Note that this assumption, together with the exogeneity condition Assumption E-SVAR imply that

$$\begin{aligned}\mathrm{E}\left(u_{1,t}z_t\right) &= \mathrm{E}\left[\left(b_{11}v_{1,t} + b_{12}v_{2,t}\right)z_t\right] && (3.44)\\ &= b_{11}\,\mathrm{E}\left(v_{1,t}z_t\right) + b_{12}\,\mathrm{E}\left(v_{2,t}z_t\right) && (3.45)\\ &= \pi_T\,. && (3.46)\end{aligned}$$

Consider the coefficient in the population regression of $u_{1,t}$ on z_t

$$\left[\mathrm{E}\left(z_t^2\right)\right]^{-1}\mathrm{E}\left(z_t u_{1,t}\right) = \pi_T\,, \tag{3.47}$$

which implies that using z_t as the best linear predictor ($\widehat{\mathrm{E}}(\cdot|\cdot)$) of $u_{1,t}$, we have that

$$\widehat{\mathrm{E}}\left(u_{1,t}|z_t\right) = \pi_T z_t\,, \tag{3.48}$$

which yields

$$u_{1,t} = \pi_T z_t + \epsilon_t \tag{3.49}$$

which is exactly the first stage in Equation (3.43).

Therefore, the estimator of b_{21} in the weak external instrument SVAR can be obtained as the solution to the IV problem given by the structural, reduced form, and first stage equations, respectively:

$$\begin{aligned}u_{j,t} &= b_{j1}u_{1,t} + e_{j,t} && \text{for all } j = 2,\ldots,k && (3.50)\\ u_{j,t} &= b_{j1}\left(\pi_T z_t + \epsilon_t\right) + e_{j,t} = b_{j1}\left(\pi_T z_t\right) + \nu_{j,t} && \text{for all } j = 2,\ldots,k && (3.51)\\ u_{1,t} &= \pi_T z_t + \epsilon_t\,. && && (3.52)\end{aligned}$$

This also implies that our results on the confidence intervals for the strength of identification in Sections 3.2.1 and 3.2.2 carry over to the weak external instrument

SVAR model. In practice, it depends on the particular problem whether the homoskedastic or the heteroskedastic/autocorrelated IV model's assumptions should be applied in addition to Assumptions SVAR, E-SVAR and π_T. In either case, it is important to note that when constructing the confidence sets, the number of endogenous regressors is $n = 1$, and the number of instruments is $K_2 = 1$. Furthermore, given that our discussion builds on applying the TSLS (IV) estimator $k-1$ times to estimate the $k-1$ elements of b_{21}, our confidence sets should not be interpreted as joint ones applying to the "joint" estimator in Equation (3.41).

3.3 Monte Carlo results

In this section, we investigate the performance of the methodologies that we proposed in three leading examples. The first is the homoskedastic IV model that is a typical reference in the weak instrument literature; the second is the heteroskedastic/autocorrelated IV model. The final example is the SVAR with external instruments. The simulations are inspired by the empirical analyses that we undertake in Section 3.4. In the present section we focus on the empirical coverage rates of our proposed confidence intervals, while Appendix C contains the mean and median lengths of the confidence intervals. Furthermore, as without loss of generality we do not include exogenous regressors (X) in the Data Generating Processes (DGPs), $Y = Y^{\perp}$, $Z = Z^{\perp}$ and $V = V^{\perp}$. The numerical optimization was performed using MATLAB's built-in `fmincon` algorithm.

3.3.1 Homoskedastic IV model

Recall that the first stage equation (in the absence of X) is given by

$$Y = Z\Pi + V\,, \tag{3.53}$$

where Y is the $(T \times n)$ matrix of endogenous variables, Z is the $(T \times K_2)$ matrix of instruments, and V is the $(T \times n)$ matrix of errors. We specified $V_t \overset{\text{iid.}}{\sim} \mathcal{N}(0, I_n)$ and $Z_t \overset{\text{iid.}}{\sim} \mathcal{N}(0, I_{K_2})$, and considered $n = \{1, 2\}$, with $K_2 = \{n, \ldots, n+3\}$. For each pair (n, K_2), we considered four values of mineval(Λ) as follows: mineval$(\Lambda) = 0$ corresponds to irrelevant instruments, mineval$(\Lambda) = \{1, 10\}$ correspond to weak instruments, while strong instruments are modeled by setting mineval$(\Lambda) = 25$. To investigate the performance of our proposed confidence interval, we considered sample sizes of $T = \{100, 250, 500, 1000\}$. The number of Monte Carlo

replications was 2000. The results presented in Tables 3.2 and 3.3 confirm that our proposed confidence interval performs well across different specifications, even for relatively small samples. The simulations also show that our method is conservative when C is not a scalar.

3.3.2 Heteroskedastic/autocorrelated IV model

We considered two DGPs, labeled as DGP 1 and DGP 2, and constructed confidence intervals for $\mu^2 = \|C\| / \operatorname{tr}(W_2)$ with nominal coverage of 90%. DGP 1 is based on the DGP suggested by Montiel Olea and Pflueger (2013). In particular, let $Z_t = \left(Z_{1,t}, \ldots, Z_{K_2,t}\right)' \overset{\text{iid.}}{\sim} \mathcal{N}\left(0, I_{K_2}\right)$ and $\widetilde{v}_t = \left(\widetilde{v}_{1,t}, \widetilde{v}_{2,t}\right) \overset{\text{iid.}}{\sim} \mathcal{N}\left(0, K\right)$, where $K = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, and Z_t is independent of $\widetilde{v}_t$. Then constructing $v_{j,t} = \prod_{k=1}^{K_2} Z_{k,t} \widetilde{v}_{j,t}$ for $j = \{1, 2\}$ results in $W = 3(K \otimes I_{K_2})$. We performed Monte Carlo simulations for sample sizes of $T = \{100, 250, 500, 1000\}$, with $K_2 = \{1, 2, 3, 4\}$ instruments, and for various levels of the strength of identification $\mu^2 = \{0, 1, 10, 25\}$. W_2 was estimated using White's (1980) heteroskedasticity consistent estimator. The number of Monte Carlo replications was 2000. The results in Table 3.4 confirm that our proposed methodology delivers confidence intervals of correct coverage across a wide range of specifications.

DGP 2 is specified as follows. Let $\widetilde{Z}_t = \left(\widetilde{Z}_{1,t}, \ldots, \widetilde{Z}_{K_2,t}\right)'$, $\epsilon_t \overset{\text{iid.}}{\sim} \mathcal{N}(0, I_{K_2})$ and $\widetilde{Z}_t' = \widetilde{Z}_{t-1}' \rho I_{K_2} + \epsilon_t$, where ρ controls the persistence of the independent autoregressive processes in $\widetilde{Z}_t$. We set $\rho = 0.5$. Then we orthogonalized $\widetilde{Z}_t$ such that $Z'Z/T = I_{K_2}$. Next, we specified a moving average process $u_t = q_t + \theta q_{t-1}$, where $q_t \overset{\text{iid.}}{\sim} \mathcal{N}(0, 1)$, and $\theta = 0.4$. Finally, conditional heteroskedasticity is introduced by $v_{2,t} = Z_t \gamma u_t$, where $v_{2,t}$ is the tth element of v_2, and the $K_2 \times 1$ coefficient vector γ is specified as $\gamma = (0.5, 0, \ldots, 0)'$. This specification introduces both heteroskedasticity and autocorrelation in the process $Z_t' v_{2,t}$. As the long-run variance estimator by Newey and West (1987) delivered rather imprecise estimates of W_2, we used the moving blocks bootstrap, following Gonçalves and White (2005), with block size $b = \lfloor T^{1/3} \rfloor$ (where $\lfloor m \rfloor$ is the integer part of m), as suggested by Hall et al. (1995). As before, the number of Monte Carlo replications was 2000. The results in Table 3.5 confirm that our proposed confidence interval has correct coverage at the nominal level $(1 - \alpha) = 0.90$ across different sample sizes T, strength of identification μ^2, and number of instruments K_2.

Table 3.2: Homoskedastic IV model, coverage rates for mineval(Λ), $n = 1$ endogenous regressor, nominal level $(1 - \alpha) = 0.90$

	$K_2 = 1$				$K_2 = 2$				$K_2 = 3$				$K_2 = 4$			
mineval(Λ) =	0	1	10	25	0	1	10	25	0	1	10	25	0	1	10	25
$T = 100$	0.89	0.95	0.89	0.89	0.90	0.97	0.94	0.92	0.88	0.98	0.96	0.94	0.89	0.98	0.97	0.94
$T = 250$	0.91	0.94	0.89	0.88	0.88	0.96	0.96	0.94	0.90	0.98	0.97	0.98	0.89	0.99	0.99	0.98
$T = 500$	0.90	0.95	0.89	0.89	0.90	0.97	0.97	0.96	0.89	0.98	0.97	0.98	0.90	0.98	0.99	0.99
$T = 1000$	0.89	0.96	0.90	0.90	0.90	0.97	0.97	0.97	0.90	0.99	0.99	0.99	0.90	0.98	0.99	0.99

Note: The table shows the empirical coverage rates of the proposed confidence interval for mineval(Λ) for different sample sizes T, instrument strength mineval(Λ), and number of instruments K_2. The number of Monte Carlo simulations was 2000.

Table 3.3: Homoskedastic IV model, coverage rates for mineval(Λ), $n = 2$ endogenous regressors, nominal level $(1 - \alpha) = 0.90$

	$K_2 = 2$				$K_2 = 3$				$K_2 = 4$				$K_2 = 5$			
mineval(Λ) =	0	1	10	25	0	1	10	25	0	1	10	25	0	1	10	25
$T = 100$	1.00	1.00	0.98	0.96	1.00	1.00	0.99	0.99	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.99
$T = 250$	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$T = 500$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$T = 1000$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note: The table shows the empirical coverage rates of the proposed confidence interval for mineval(Λ) for different sample sizes T, instrument strength mineval(Λ), and number of instruments K_2. The number of Monte Carlo simulations was 2000.

Table 3.4: Heteroskedastic IV model (DGP 1), coverage rates for μ^2, nominal level $(1-\alpha) = 0.90$

	$K_2 = 1$				$K_2 = 2$				$K_2 = 3$				$K_2 = 4$			
$\mu^2 =$	0	1	10	25	0	1	10	25	0	1	10	25	0	1	10	25
$T = 100$	0.88	0.94	0.93	0.91	0.90	0.96	0.97	0.86	0.91	0.96	0.97	0.83	0.94	0.88	0.99	0.79
$T = 250$	0.89	0.94	0.93	0.94	0.90	0.96	0.94	0.89	0.89	0.96	0.94	0.91	0.95	0.94	0.90	0.91
$T = 500$	0.91	0.94	0.92	0.93	0.89	0.99	0.96	0.94	0.94	0.99	0.92	0.94	0.93	0.97	0.87	0.87
$T = 1000$	0.92	0.95	0.89	0.92	0.89	0.98	0.94	0.94	0.94	0.98	0.97	0.85	0.89	0.98	0.87	0.82

Note: The table shows the empirical coverage rates of the proposed confidence interval for μ^2 for different sample sizes T, instrument strength μ^2, and number of instruments K_2. Asymptotic variance W_2 estimated using White's (1980) heteroskedasticity consistent estimator. The number of Monte Carlo simulations was 2000.

Table 3.5: Heteroskedastic and autocorrelated IV model (DGP 2), coverage rates for μ^2, nominal level $(1-\alpha) = 0.90$

	$K_2 = 1$				$K_2 = 2$				$K_2 = 3$				$K_2 = 4$			
$\mu^2 =$	0	1	10	25	0	1	10	25	0	1	10	25	0	1	10	25
$T = 100$	0.84	0.89	0.89	0.95	0.87	0.94	0.99	0.96	0.87	0.97	1.00	0.91	0.82	0.99	1.00	0.76
$T = 250$	0.88	0.91	0.85	0.90	0.87	0.95	0.97	0.99	0.89	0.98	0.99	0.99	0.86	0.98	1.00	0.99
$T = 500$	0.87	0.91	0.83	0.80	0.88	0.95	0.94	0.98	0.85	0.98	0.98	1.00	0.87	0.98	1.00	1.00
$T = 1000$	0.88	0.90	0.82	0.76	0.89	0.95	0.92	0.92	0.90	0.98	0.97	0.98	0.87	0.99	0.98	0.99

Note: The table shows the empirical coverage rates of the proposed confidence interval for μ^2 for different sample sizes T, instrument strength μ^2, and number of instruments K_2. Asymptotic variance W_2 estimated with the moving blocks bootstrap, following Gonçalves and White (2005), with block size $b = \lfloor T^{1/3} \rfloor$, as suggested by Hall et al. (1995). The number of Monte Carlo simulations was 2000.

3.3.3 External instrument SVAR model

We specified three DGPs. In the first two, labeled as DGP 1 and DGP 2, the proxy variable is modeled as an *iid*. process: $z_t \overset{\text{iid.}}{\sim} \mathcal{N}(0,1)$. We considered a homoskedastic and a heteroskedastic process for ϵ_t. In the former (DGP 1), $\epsilon_t \overset{\text{iid.}}{\sim} \mathcal{N}(0,1)$. In the latter (DGP 2), $\epsilon_t = |\tau z_t|\tilde{\epsilon}_t$, $\tilde{\epsilon}_t \overset{\text{iid.}}{\sim} \mathcal{N}(0,1)$ and we set $\tau = 0.5$, which implies $W_2 = 3\tau^2 = 0.75$. In DGP 3, we introduced serial correlation by specifying the autoregressive process $z_t = \rho z_{t-1} + \eta_t$ for the instrument, with $\eta_t \overset{\text{iid.}}{\sim} \mathcal{N}(0, 1-\rho^2)$, $\rho = 0.2$, and specifying a moving average process for ϵ_t such that $\epsilon_t = \zeta_t + \theta\zeta_{t-1}$, $\zeta_t \overset{\text{iid.}}{\sim} \mathcal{N}(0,1)$, $\theta = 0.3$, implying $W_2 = 1.21$.

Naturally, for the homoskedastic DGP we constructed the confidence intervals based on our results in Section 3.2.1, while in the heteroskedastic and autocorrelated DGPs we used the confidence interval proposed in Section 3.2.2.

In order to investigate the coverage rate of our proposed confidence interval at different strengths of the external instrument, we performed simulations using different values of the local-to-zero parameter C. In the homoskedastic case, note that $\Lambda = C^2$. The four values of Λ we used are $\Lambda = \{0.1, 1, 10, 25\}$, with corresponding values of $C = \left\{\sqrt{0.1}, 1, \sqrt{10}, 5\right\}$. In the heteroskedastic and autocorrelated cases, we considered the same values for $\mu^2 = C^2/W_2$ as for Λ, resulting in $C = \sqrt{W_2} \times \left\{\sqrt{0.1}, 1, \sqrt{10}, 5\right\}$. The sample sizes are $T = \{100, 150, 200, 500\}$, which are typical in the macroeconomics literature using quarterly or monthly data. The simulations were performed at the nominal $(1-\alpha) = 90\%$ confidence level, and we conducted 2000 Monte Carlo replications.

The results reported in Tables 3.6 and 3.7 confirm that our proposed confidence interval delivers coverage rates close to the nominal level, with minor coverage distortions in most cases.

Table 3.6: Homoskedastic external instrument SVAR (DGP 1), coverage rates for mineval(Λ), nominal level $(1-\alpha) = 0.90$

mineval(Λ) =	0.01	1	10	25
$T = 100$	0.93	0.95	0.89	0.90
$T = 150$	0.94	0.95	0.90	0.91
$T = 200$	0.93	0.95	0.91	0.90
$T = 500$	0.94	0.96	0.90	0.89

Note: The table shows the empirical coverage rates of the proposed confidence interval for mineval(Λ) for different sample sizes T, and external instrument strength mineval(Λ). The number of Monte Carlo simulations was 2000.

Table 3.7: Heteroskedastic and autocorrelated external instrument SVAR coverage rates for μ^2, nominal level $(1-\alpha) = 0.90$

	DGP 2, Heteroskedastic				DGP 3, Autocorrelated			
$\mu^2 =$	0.01	1	10	25	0.01	1	10	25
$T = 100$	0.92	0.93	0.85	0.83	0.92	0.92	0.87	0.84
$T = 150$	0.92	0.93	0.85	0.84	0.92	0.93	0.86	0.85
$T = 200$	0.94	0.94	0.88	0.85	0.92	0.93	0.87	0.86
$T = 500$	0.94	0.95	0.89	0.88	0.93	0.94	0.89	0.88

Note: The table shows the empirical coverage rates of the proposed confidence interval for μ^2 for different sample sizes T, and external instrument strength μ^2. Asymptotic variance W_2 estimated by Gonçalves and White's (2005) bootstrap with 2999 bootstrap samples, with block length equal to one in the case of DGP 2, and block length equal to $\lfloor T^{1/3} \rfloor$ in the case of DGP 3. The number of Monte Carlo simulations was 2000.

3.4 Empirical Analysis

3.4.1 Linear homoskedastic IV model

New Keynesian macroeconomic models predict that inflation dynamics can be described by some form of the New Keynesian Phillips Curve (NKPC). In this paper, we follow Galí and Gertler's (1999) "hybrid" specification:

$$\pi_t = c + \lambda s_t + \gamma_f E_t(\pi_{t+1}) + \gamma_b \pi_{t-1} + \varepsilon_t, \tag{3.54}$$

where π_t denotes inflation in period t, s_t is the natural log of the labor income share, and E_t is the conditional expectation operator. After replacing $E_t(\pi_{t+1})$ by π_{t+1}, the model can be written as $y_t \equiv \pi_t$, $Y_t \equiv (s_t, \pi_{t+1})'$, and $X_t \equiv (1, \pi_{t-1})'$ in the notation of Section 3.2.1.

To handle the endogeneity of s_t and π_{t+1}, we used the TSLS estimator. We specified the first stage equation as:

$$Y_t = Z_t \Pi + X_t \Phi + V_t\,, \tag{3.55}$$

where Z_t is a vector of instruments. More specifically, Z_t, contains the first three lags of labor share, the Baxter–King (Baxter and King, 1999) filtered output gap (retaining cyclical fluctuations of real GDP between 6 and 32 quarters), wage inflation, interest rate spread (defined as the difference between the 10-year Treasury at constant maturity rate and the 3-month Treasury bill rate) and commodity price inflation, and the second to third lags of inflation. The quarterly

US data were downloaded from the St. Louis Fed's FRED Database. The full sample ranges from 1960:Q1 to 2017:Q1. For more information on the series and the transformations, see Appendix D. In our analysis, we focus on constructing a 90% confidence interval for the mineval(Λ) parameter.

We consider two samples. The first is 1960:Q1 to 1997:Q4, which is the same sample used in Galí and Gertler (1999). The second is the full sample, 1960:Q1 to 2017:Q1. The results are shown in Table 3.8. Thus, while considering the full sample period, the instruments are somewhat weak based on the size distortion, in the first subsample the NKPC is stronger identified.

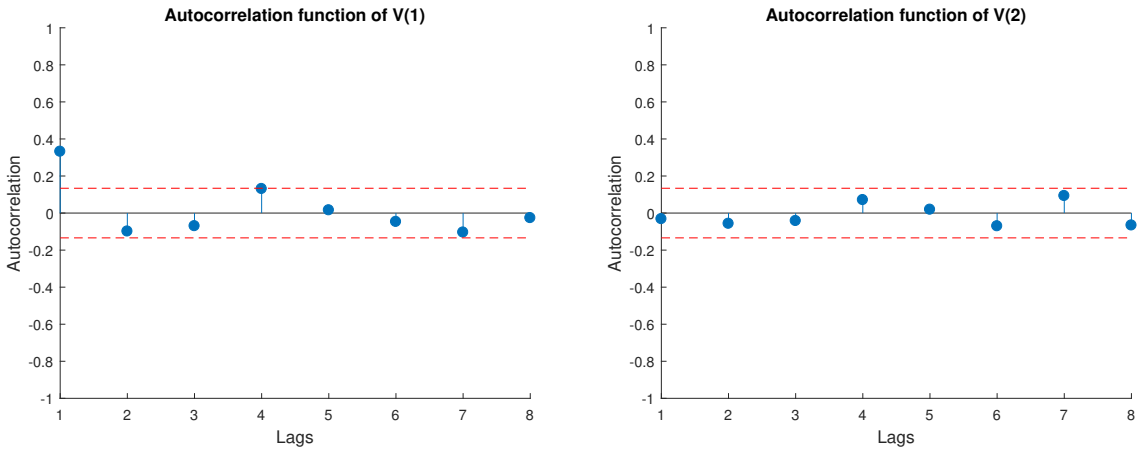
Table 3.8: Confidence intervals for the strength of identification of the NKPC

Sample period	$\text{CI}^{\Lambda}_{0.90}$	max. bias	max. size distortion
1960:Q1 – 1997:Q4	$[45.32, 64.78]$	$[0.01, 0.02]$	$[0.08, 0.09]$
1960:Q1 – 2017:Q1	$[31.84, 78.97]$	$[0.01, 0.03]$	$[0.07, 0.11]$

Note: The table shows the proposed $(1-\alpha) = 0.90$ level confidence interval for mineval(Λ), and the corresponding maximum bias and size distortion (assuming 5% nominal level for the Wald-test). The results in the first row are based on the same subsample as Galí and Gertler (1999), while the second row presents the results based on the full sample.

In the presence of heteroskedasticity or serial correlation, the concentration parameter does not have the same meaning as in the uncorrelated, homoskedastic case. However, as shown in Figure 3.3, the serial correlation of the residuals is very mild at most. White's (1980) test for conditional heteroskedasticity yielded p-values of 0.01 and 0.28 for the first and the second first-stage regressions, respectively, which might warrant some caution about our results.

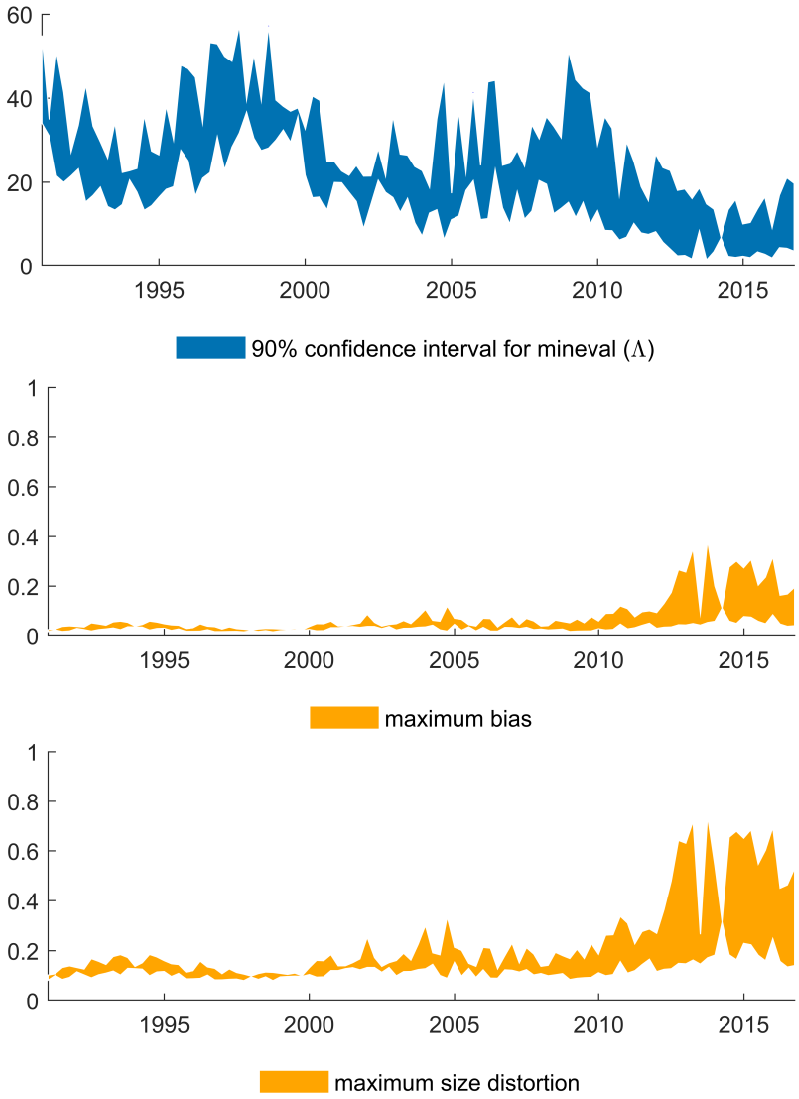
Figure 3.3: Autocorrelation of residuals in the NKPC



Note: The left panel shows the correlogram of the residuals in the first column of $\widehat{V}$ along with the 95% confidence bands, while the right panel displays the same for the second column of $\widehat{V}$.

Figure 3.4 reports empirical evidence on the strength of identification based on a rolling window size with 120 observations, corresponding to 30 years of data. The pictures show some time variation, in particular they point to weaker identification around the end of the sample, suggesting that the NKPC has become flatter during the Great Moderation, echoing the results of Kleibergen and Mavroeidis (2009). Overall, our analysis confirms the findings of Mavroeidis et al. (2014), who similarly found empirical evidence of weak identification in the New Keynesian Phillips Curve.

Figure 3.4: Evolution of the strength of identification of the NKPC



Note: The upper subfigure shows the 90% confidence interval for mineval(Λ) over time, calculated in rolling windows with 120 observations (corresponding to 30 years). The middle and bottom subfigures show the corresponding worst case maximum bias and size distortion, respectively. The timing corresponds to the end of the rolling windows.

3.4.2 Heteroskedastic/autocorrelated IV model

The elasticity of intertemporal substitution (EIS) is often estimated using a linearized Euler equation, which is commonly derived as an optimality condition of the household's problem in modern macroeconomic models. We illustrate our proposed methodology by using the same specifications of the consumption Euler equation as Yogo (2004) and Montiel Olea and Pflueger (2013). In particular, the model specification is the following:

$$\Delta c_{t+1} = \nu + \psi r_{t+1} + u_{t+1}\,, \tag{3.56}$$

$$r_{t+1} = \xi + \psi^{-1}\Delta c_{t+1} + \eta_{t+1}\,, \tag{3.57}$$

where Δc_{t+1} is consumption growth, and r_{t+1} is a real asset return, ψ is the EIS parameter, ν and ξ are constants, while u_{t+1} and η_{t+1} are stochastic disturbances. Note that Equation (3.57) (EIS ψ^{-1}) expresses the same relationship between consumption growth and returns as Equation (3.56) (EIS ψ), but often the estimates of ψ are vastly different between these two specifications. Yogo (2004) argued that weak identification can explain these contradicting results.

In the empirical analysis we construct the data set following Yogo (2004) and Montiel Olea and Pflueger (2013) for the sample period 1960:Q1 to 2017:Q1, using US data. We used real per capita consumption growth for Δc_{t+1}, and the real return on the 3-month T-bill for r_{t+1}. As Yogo (2004) noted, by using instruments dated $t-1$, ψ or its reciprocal ψ^{-1} can be still identified even if asset returns or consumption are conditionally heteroskedastic. We used the same set of instruments as Montiel Olea and Pflueger (2013), with $Z_{t-1} = \big(\text{3-month T-bill rate}_{t-1}, \Delta \log \text{CPI}_{t-1}, \Delta c_{t-1}, \log\left(\text{Div}/\text{Price}\right)_{t-1}\big)'$, where CPI is the consumer price index, and Div/Price is a dividend over price ratio, calculated over a large number of assets. Appendix D contains further details of the data.

Table 3.9 summarizes the results. First, the TSLS point estimates suggest contradicting values for ψ, which was also found by Yogo (2004) and Montiel Olea and Pflueger (2013). Moreover, comparing the results in the top and the bottom of the table shows that the robust confidence intervals clearly signal weak identification with very wide confidence intervals for μ^2 and correspondingly large intervals for the bias τ, while the confidence intervals based on the homoskedastic IV model show no evidence of identification deficiency. This confirms Montiel Olea and Pflueger's (2013) finding that the asymptotic theory developed for the homoskedastic IV model can be highly misleading when judging the strength of identification.

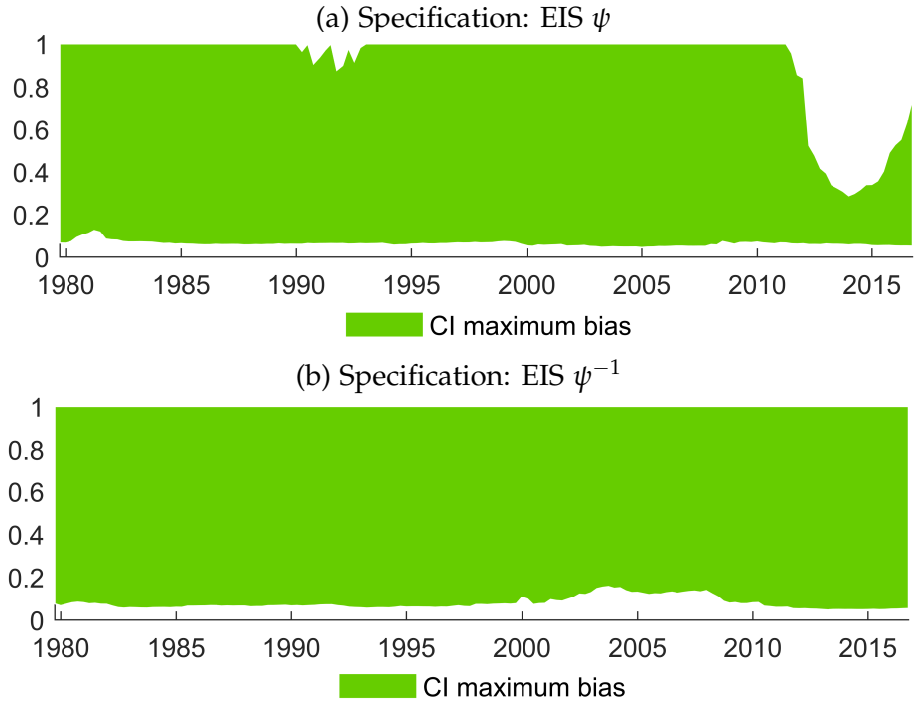
Table 3.9: EIS: Full sample results for the strength of identification

Specification	EIS ψ	EIS ψ^{-1}
point estimate	0.24	0.93
$\text{CI}^{\mu^2}_{0.90}$	$[0.83, 17.74]$	$[0.66, 14.20]$
maximum bias $\widehat{\tau}^U_L$	$[0.04, 0.92]$	$[0.05, 1.00]$
$\text{CI}^{\Lambda}_{0.90}$	$[112.20, 179.06]$	$[122.99, 192.64]$
maximum bias	$[0.00, 0.00]$	$[0.00, 0.00]$
max. size distortion	$[0.00, 0.02]$	$[0.00, 0.01]$

Note: The table displays the estimation results of the consumption Euler equations with Δc_{t+1} regressed on r_{t+1} (specification EIS ψ of Equation (3.56)), and r_{t+1} regressed on Δc_{t+1} (specification EIS ψ^{-1} of Equation (3.57)) using the TSLS estimator. The upper panel shows the $(1-\alpha) = 0.90$ level confidence interval for μ^2 (heteroskedastic/autocorrelated IV model) and the corresponding maximum bias. Asymptotic variance W_2 estimated by Gonçalves and White's (2005) bootstrap with 2999 bootstrap samples, with block length equal to $\lfloor T^{1/3} \rfloor = 6$. The lower panel displays the $(1-\alpha) = 0.90$ level confidence interval for mineval(Λ) (homoskedastic IV model) and the corresponding maximum bias and size distortion (assuming a nominal 5% Wald-test).

We also estimated Equations (3.56) and (3.57) in rolling windows of 80 quarters (corresponding to 20 years), and as Figure 3.5 shows, the conclusions are largely unchanged.

Figure 3.5: Maximum bias $\widehat{\tau}^U_L$ over time, rolling windows of $w = 80$ quarters



3.4.3 External instrument SVAR model

We illustrate the usefulness of our confidence interval with an empirical example investigating the dynamic effects of oil price shocks. Inspired by Montiel Olea et al. (2016), we specify a VAR(1) in the following $k = 4$ quarterly US variables: log difference of oil price, first difference of the Fed funds rate, log difference of CPI, and log difference of real GDP. The lag length was selected using the Bayesian Information Criterion (Schwarz, 1978), with possible lag length between one and six. In the literature, two prominent oil shock series are often used: Hamilton's (2003) twelve-month maximum deviation series ($z_t^{\mathrm{H},12}$) and Kilian's (2008) shortfall in OPEC's oil production (z_t^{K}). In the analysis, we use these shocks one at a time and provide evidence on the strength of these external instruments. Due to data availability, the sample period starts in 1971:Q1 and ends in 2004:Q3. For a detailed description of the data, see Appendix D.

Given that we cannot exclude the possibility of heteroskedasticity or autocorrelation, we used the methodology of both Sections 3.2.1 and 3.2.2 to construct confidence intervals for the strength of identification.

In Table 3.10 we can see that our confidence interval for the heteroskedastic/autocorrelated IV model signals that Hamilton's (2003) oil shock series is a strong external instrument, while Kilian's (2008) series is a rather weak one. Furthermore, as the confidence interval based on the homoskedastic IV model indicates the same, we conjecture that heteroskedasticity/autocorrelation is not of primary concern in this application. Lunsford's (2016) F-statistic also confirms our findings about the relative strength of these instruments, which is in line with Montiel Olea et al.'s (2016) results.

Table 3.10: Oil shocks: confidence intervals for the strength of identification

Oil shock	$\mathrm{CI}_{0.90}^{\mu^2}$	maximum bias $\widehat{\tau}_L^U$	$\mathrm{CI}_{0.90}^{\Lambda}$	max. size distortion	F-statistic
$z_t^{\mathrm{H},12}$	$[15.40, 52.04]$	$[0.02, 0.07]$	$[75.16, 143.30]$	$[0.00, 0.00]$	27.11
z_t^{K}	$[0.00, 4.48]$	$[0.22, 1.00]$	$[0.00, 4.77]$	$[0.11, 0.62]$	2.49

Note: The table shows the estimation results using either Hamilton's (2003) or Kilian's (2008) oil shock series as external instruments ($z_t^{\mathrm{H},12}$ and z_t^{K}, respectively). The first two main columns display the $(1-\alpha) = 0.90$ level confidence interval for μ^2 and the corresponding maximum bias (heteroskedastic/autocorrelated IV model). Asymptotic variance W_2 estimated by Gonçalves and White's (2005) bootstrap with 2999 bootstrap samples, with block length equal to $\lfloor T^{1/3} \rfloor = 5$. The third and fourth main columns contain the $(1-\alpha) = 0.90$ level confidence interval for mineval(Λ) and the corresponding maximum size distortion (assuming a 5% Wald-test) in the homoskedastic IV model. The last column contains Lunsford's (2016) F-statistic testing for a weak external instrument, whose critical values at the 10% level, leading to a maximum bias of $\{0.01, 0.05, 0.1, 0.2\}$ are $\{46.42, 12.03, 7.24, 4.58\}$, respectively.

3.5 Conclusion

In this paper we proposed confidence intervals for the strength of identification in the most well-known homoskedastic and heteroskedastic/autocorrelated linear IV models, and in Structural VARs identified with an external instrument. Therefore our proposed methodology can inform researchers working with either microeconomic or macroeconomic data on *how* strong their instruments are. Monte Carlo simulations demonstrated that the proposed confidence intervals have correct coverage even for moderate sample sizes. Furthermore, the practical implementation is easy and computationally not intensive.

In an empirical application, we showed that the New Keynesian Phillips Curve has become more weakly identified after the Great Moderation. Furthermore, our analysis of consumption Euler equations confirmed that weak identification of the model poses a serious challenge to estimate the elasticity of intertemporal substitution parameter. Finally, in a Structural VAR framework we demonstrated that Hamilton's (2003) oil shock series can be used as a strong instrument when analyzing the dynamic effects of oil shocks.

Our results suggest that the methods could be applied to construct confidence intervals for the structural parameters as well, although we do not specifically investigate it given the variety of methods to construct robust confidence intervals available in the literature (see, for example Kleibergen and Mavroeidis (2009) or Montiel Olea et al. (2016)) .

The present study could be extended in a number of directions. First, developing the asymptotic theory for the strength of identification with general error structures and multiple endogenous regressors would be of great practical use. Furthermore, it would be useful to extend our methodology to IV models in which not all the instruments are weak. Another possible avenue for future work is the generalization of our confidence interval for the strength of identification in the Structural VAR framework with multiple instruments and multiple identified shocks. These extensions are left for future research.

Appendices

A Proofs

Proof of Proposition 1. First, let us introduce some additional notation. Let us define the lower and upper endpoints of the confidence interval $\text{CI}^{\Lambda}_{1-\alpha}$ as

$$A_T \equiv \min_{\widetilde{C}\in \text{CI}^{C}_{1-\alpha}} \text{mineval}(\widehat{\Sigma}^{-1/2'}_{V^\perp V^\perp} \widetilde{C}' \widehat{Q}_{Z^\perp Z^\perp} \widetilde{C} \widehat{\Sigma}^{-1/2}_{V^\perp V^\perp} / K_2)\,, \tag{A.1}$$

$$B_T \equiv \max_{\widetilde{C}\in \text{CI}^{C}_{1-\alpha}} \text{mineval}(\widehat{\Sigma}^{-1/2'}_{V^\perp V^\perp} \widetilde{C}' \widehat{Q}_{Z^\perp Z^\perp} \widetilde{C} \widehat{\Sigma}^{-1/2}_{V^\perp V^\perp} / K_2)\,, \tag{A.2}$$

and consider their semi-population counterparts given by

$$A \equiv \min_{\widetilde{C}\in \text{CI}^{C}_{1-\alpha}} \text{mineval}(\Sigma^{-1/2'}_{V^\perp V^\perp} \widetilde{C}' Q_{Z^\perp Z^\perp} \widetilde{C} \Sigma^{-1/2}_{V^\perp V^\perp} / K_2)\,, \tag{A.3}$$

$$B \equiv \max_{\widetilde{C}\in \text{CI}^{C}_{1-\alpha}} \text{mineval}(\Sigma^{-1/2'}_{V^\perp V^\perp} \widetilde{C}' Q_{Z^\perp Z^\perp} \widetilde{C} \Sigma^{-1/2}_{V^\perp V^\perp} / K_2)\,. \tag{A.4}$$

Using this new notation, we need to prove that

$$\lim_{T\to\infty} P\left(\text{mineval}(\Lambda) \in [A_T, B_T]\right) \geq 1-\alpha\,. \tag{A.5}$$

Note that by construction, $\lim_{T\to\infty} P\left(C \in \text{CI}^{C}_{1-\alpha}\right) = 1-\alpha$, therefore it follows that $\lim_{T\to\infty} P\left(\text{mineval}(\Lambda) \in [A, B]\right) \geq 1-\alpha$.[7] Consequently, we need to prove that

$$\lim_{T\to\infty} \left[P(\text{mineval}(\Lambda) \in [A_T, B_T]) - P(\text{mineval}(\Lambda) \in [A, B])\right] = 0\,. \tag{A.6}$$

To show this, it suffices to prove that for any $\epsilon > 0$, $\lim_{T\to\infty} P(|B_T - B| > \epsilon) = 0$ (the argument for A_T and A is analogous, and therefore omitted).

Let us define $\Theta \equiv \mathbb{P}^{n\times n}_{+} \times \mathbb{P}^{K_2\times K_2}_{+} \times \mathbb{R}^{K_2\times n}$, where $\mathbb{P}^{k\times l}_{+}$ is the set of $(k\times l)$ positive definite matrices. Observe that $\theta \equiv \left(\widehat{\Sigma}_{V^\perp V^\perp}, \widehat{Q}_{Z^\perp Z^\perp}, \widehat{C}\right) \in \Theta$. Note that $f : \mathbb{R}^{K_2\times n} \times \Theta \to \mathbb{R}$ given by $f(\widetilde{C}, \theta) = \text{mineval}(\widehat{\Sigma}^{-1/2'}_{V^\perp V^\perp} \widetilde{C}' \widehat{Q}_{Z^\perp Z^\perp} \widetilde{C} \widehat{\Sigma}^{-1/2}_{V^\perp V^\perp} / K_2)$ is continuous on $\left(\mathbb{R}^{K_2\times n} \times \Theta\right)$, and $\mathcal{D} : \Theta \longrightarrow \text{CI}^{C}_{1-\alpha}$ is a compact-valued, continuous correspondence. Therefore by the maximum theorem (Sundaram, 1996, page 235), $\max_{\widetilde{C}\in \text{CI}^{C}_{1-\alpha}} \text{mineval}(\widehat{\Sigma}^{-1/2'}_{V^\perp V^\perp} \widetilde{C}' \widehat{Q}_{Z^\perp Z^\perp} \widetilde{C} \widehat{\Sigma}^{-1/2}_{V^\perp V^\perp} / K_2)$ is continuous on

[7] We used the compactness of $\text{CI}^{C}_{1-\alpha}$ and the continuity of the mineval$(\cdot)$ function. The latter follows from the fact that the eigenvalues of a matrix are continuous functions of the entries of the matrix (Theorem 2.11 on page 68 in Zhang (2011)).

Θ. Hence, using that $\widehat{\Sigma}_{V^\perp V^\perp} \xrightarrow{p} \Sigma_{V^\perp V^\perp}$ and $\widehat{Q}_{Z^\perp Z^\perp} \xrightarrow{p} Q_{Z^\perp Z^\perp}$ (both follow from Assumption M) and applying the continuous mapping theorem, we have established that for any $\epsilon > 0$, $\lim_{T\to\infty} P(|B_T - B| > \epsilon) = 0$, which concludes the proof. ■

Proof of Proposition 2. The proof is analogous to that of Proposition 1, and therefore omitted. ■

B Boundary values of mineval(Λ)

This appendix contains the simulated boundary values of mineval(Λ) for $n = \{1, 2, 3\}$ endogenous variables and $K_2 = n + 2, \ldots, 20$ (maximum bias, Tables B.1 to B.3) and $K_2 = n, \ldots, 20$ (maximum size distortion, Tables B.4 to B.6) instruments for a fine grid of maximum bias and maximum size distortion for the TSLS estimator. The simulation procedure follows Stock and Yogo (2005). Following Stock and Yogo (2005), we calculated the maximum size distortion assuming the Wald test on β in the structural equation has a nominal level 5%.

Table B.1: Simulated boundary values of mineval(Λ) for $n = 1$ endogenous regressor, for different values of maximum bias (in columns) and number of instruments K_2 (in rows)

K_2	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.2	0.21	0.22	0.23	0.24	0.25	0.26	0.27	0.28	0.29	0.3	0.31	0.32	0.33	0.34	0.35	0.36	0.37	0.38	0.39	0.4	0.41	0.42	0.43	0.44	0.45	0.46	0.47	0.48	0.49	0.5
3	33.23	17.21	11.42	8.62	6.95	5.97	5.16	4.56	4.09	3.70	3.42	3.18	2.96	2.79	2.60	2.47	2.35	2.23	2.13	2.04	1.95	1.87	1.79	1.72	1.65	1.60	1.55	1.49	1.44	1.39	1.34	1.29	1.25	1.22	1.18	1.14	1.11	1.07	1.04	1.01	0.98	0.95	0.92	0.89	0.86	0.84	0.81	0.79	0.76	0.74
4	49.69	25.18	16.66	12.48	9.97	8.37	7.18	6.28	5.59	5.02	4.58	4.20	3.88	3.61	3.36	3.15	2.97	2.80	2.65	2.52	2.39	2.28	2.17	2.07	1.98	1.90	1.82	1.75	1.68	1.61	1.55	1.49	1.44	1.39	1.34	1.29	1.25	1.20	1.16	1.12	1.08	1.05	1.01	0.98	0.95	0.92	0.89	0.86	0.83	0.80
5	59.56	29.96	19.80	14.79	11.77	9.82	8.40	7.32	6.49	5.81	5.27	4.82	4.43	4.11	3.81	3.56	3.34	3.14	2.96	2.80	2.65	2.52	2.39	2.28	2.17	2.08	1.99	1.90	1.83	1.75	1.68	1.61	1.55	1.49	1.44	1.38	1.33	1.28	1.24	1.19	1.15	1.11	1.07	1.03	1.00	0.96	0.93	0.90	0.87	0.84
6	66.15	33.15	21.89	16.33	12.98	10.78	9.21	8.01	7.09	6.34	5.73	5.23	4.80	4.44	4.11	3.84	3.59	3.37	3.17	2.99	2.83	2.68	2.54	2.42	2.31	2.20	2.10	2.01	1.92	1.84	1.76	1.69	1.62	1.56	1.50	1.44	1.39	1.34	1.29	1.24	1.19	1.15	1.11	1.07	1.03	1.00	0.96	0.93	0.89	0.86
7	70.85	35.43	23.39	17.43	13.84	11.46	9.78	8.50	7.52	6.71	6.07	5.53	5.07	4.67	4.33	4.03	3.77	3.53	3.32	3.13	2.95	2.80	2.65	2.52	2.40	2.28	2.18	2.08	1.99	1.90	1.82	1.75	1.68	1.61	1.55	1.49	1.43	1.37	1.32	1.27	1.23	1.18	1.14	1.10	1.06	1.02	0.98	0.95	0.91	0.88
8	74.38	37.14	24.51	18.26	14.48	11.98	10.22	8.87	7.84	6.99	6.31	5.75	5.26	4.85	4.49	4.18	3.90	3.65	3.43	3.23	3.05	2.89	2.73	2.60	2.47	2.35	2.24	2.14	2.04	1.95	1.87	1.79	1.72	1.65	1.58	1.52	1.46	1.40	1.35	1.30	1.25	1.20	1.16	1.11	1.07	1.03	1.00	0.96	0.93	0.89
9	77.12	38.47	25.39	18.90	14.99	12.38	10.56	9.16	8.09	7.21	6.51	5.92	5.42	4.98	4.61	4.29	4.00	3.75	3.52	3.31	3.12	2.95	2.80	2.65	2.52	2.40	2.29	2.18	2.08	1.99	1.90	1.82	1.75	1.68	1.61	1.54	1.48	1.42	1.37	1.32	1.27	1.22	1.17	1.13	1.09	1.05	1.01	0.97	0.94	0.90
10	79.32	39.53	26.09	19.42	15.39	12.70	10.83	9.39	8.29	7.39	6.66	6.06	5.54	5.09	4.71	4.38	4.09	3.82	3.59	3.37	3.18	3.01	2.85	2.70	2.57	2.44	2.33	2.21	2.11	2.02	1.93	1.85	1.77	1.70	1.63	1.56	1.50	1.44	1.39	1.33	1.28	1.23	1.19	1.14	1.10	1.06	1.02	0.98	0.94	0.91
11	81.11	40.40	26.66	19.84	15.72	12.96	11.05	9.58	8.45	7.53	6.79	6.17	5.64	5.18	4.80	4.46	4.15	3.88	3.64	3.43	3.23	3.05	2.89	2.74	2.60	2.47	2.36	2.24	2.14	2.05	1.96	1.87	1.79	1.72	1.65	1.58	1.52	1.46	1.40	1.34	1.29	1.24	1.20	1.15	1.11	1.07	1.03	0.99	0.95	0.92
12	82.61	41.12	27.13	20.19	15.99	13.18	11.23	9.73	8.58	7.65	6.89	6.26	5.72	5.26	4.86	4.52	4.21	3.94	3.69	3.47	3.27	3.09	2.92	2.77	2.63	2.50	2.38	2.27	2.16	2.07	1.97	1.89	1.81	1.73	1.66	1.59	1.53	1.47	1.41	1.36	1.30	1.25	1.20	1.16	1.12	1.07	1.03	0.99	0.96	0.92
13	83.88	41.74	27.54	20.48	16.22	13.36	11.39	9.87	8.70	7.75	6.98	6.34	5.80	5.32	4.92	4.57	4.26	3.98	3.73	3.50	3.30	3.12	2.95	2.80	2.66	2.52	2.40	2.29	2.18	2.08	1.99	1.90	1.82	1.75	1.67	1.60	1.54	1.48	1.42	1.36	1.31	1.26	1.21	1.17	1.12	1.08	1.04	1.00	0.96	0.93
14	84.96	42.26	27.88	20.74	16.42	13.52	11.52	9.98	8.80	7.84	7.06	6.41	5.86	5.38	4.97	4.62	4.30	4.02	3.76	3.54	3.33	3.15	2.98	2.82	2.68	2.54	2.42	2.30	2.20	2.10	2.01	1.92	1.84	1.76	1.68	1.61	1.55	1.49	1.43	1.37	1.32	1.27	1.22	1.17	1.13	1.08	1.04	1.00	0.97	0.93
15	85.90	42.72	28.18	20.96	16.59	13.66	11.64	10.08	8.88	7.92	7.12	6.47	5.91	5.42	5.02	4.66	4.33	4.05	3.79	3.56	3.36	3.17	3.00	2.84	2.70	2.56	2.44	2.32	2.21	2.11	2.02	1.93	1.85	1.77	1.69	1.62	1.56	1.50	1.44	1.38	1.33	1.27	1.22	1.18	1.13	1.09	1.05	1.01	0.97	0.93
16	86.72	43.12	28.44	21.15	16.74	13.78	11.74	10.17	8.96	7.98	7.18	6.52	5.96	5.47	5.05	4.69	4.36	4.08	3.82	3.59	3.38	3.19	3.02	2.86	2.71	2.57	2.45	2.33	2.22	2.12	2.03	1.94	1.86	1.78	1.70	1.63	1.56	1.50	1.44	1.38	1.33	1.28	1.23	1.18	1.14	1.09	1.05	1.01	0.97	0.94
17	87.45	43.47	28.67	21.32	16.88	13.88	11.83	10.24	9.03	8.04	7.23	6.56	6.00	5.50	5.09	4.72	4.39	4.10	3.84	3.61	3.40	3.21	3.03	2.87	2.73	2.59	2.46	2.34	2.23	2.13	2.04	1.95	1.86	1.78	1.71	1.64	1.57	1.51	1.45	1.39	1.34	1.28	1.23	1.19	1.14	1.10	1.06	1.02	0.98	0.94
18	88.10	43.78	28.88	21.47	16.99	13.98	11.91	10.31	9.08	8.09	7.28	6.60	6.03	5.53	5.12	4.75	4.42	4.13	3.86	3.63	3.42	3.23	3.05	2.89	2.74	2.60	2.47	2.35	2.24	2.14	2.05	1.95	1.87	1.79	1.71	1.64	1.58	1.51	1.45	1.39	1.34	1.29	1.24	1.19	1.14	1.10	1.06	1.02	0.98	0.94
19	88.67	44.06	29.06	21.61	17.10	14.06	11.98	10.37	9.14	8.14	7.32	6.64	6.06	5.56	5.14	4.77	4.44	4.15	3.88	3.64	3.43	3.24	3.06	2.90	2.75	2.61	2.48	2.36	2.25	2.15	2.05	1.96	1.88	1.80	1.72	1.65	1.58	1.52	1.46	1.40	1.34	1.29	1.24	1.19	1.15	1.10	1.06	1.02	0.98	0.94
20	89.19	44.31	29.23	21.73	17.19	14.14	12.04	10.42	9.18	8.18	7.36	6.67	6.09	5.59	5.17	4.79	4.46	4.16	3.90	3.66	3.45	3.25	3.07	2.91	2.76	2.62	2.49	2.37	2.26	2.16	2.06	1.97	1.88	1.80	1.72	1.65	1.59	1.52	1.46	1.40	1.35	1.29	1.24	1.20	1.15	1.10	1.06	1.02	0.98	0.95

Note: The table shows the simulated boundary values of mineval(Λ) for different values of maximum bias (in columns) and number of instruments K_2 (in rows). The simulations are based on 100,000 Monte Carlo replications, and follow Stock and Yogo (2005).

Table B.2: Simulated boundary values of mineval(Λ) for $n = 2$ endogenous regressors, for different values of maximum bias (in columns) and number of instruments K_2 (in rows)

K_2	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**	**0.1**	**0.11**	**0.12**	**0.13**	**0.14**	**0.15**	**0.16**	**0.17**	**0.18**	**0.19**	**0.2**	**0.21**	**0.22**	**0.23**	**0.24**	**0.25**	**0.26**	**0.27**	**0.28**	**0.29**	**0.3**	**0.31**	**0.32**	**0.33**	**0.34**	**0.35**	**0.36**	**0.37**	**0.38**	**0.39**	**0.4**	**0.41**	**0.42**	**0.43**	**0.44**	**0.45**	**0.46**	**0.47**	**0.48**	**0.49**	**0.5**
4	26.20	13.27	8.86	6.80	5.65	4.78	4.14	3.70	3.34	3.06	2.84	2.64	2.48	2.33	2.22	2.12	2.01	1.91	1.84	1.75	1.69	1.63	1.57	1.51	1.45	1.41	1.36	1.32	1.28	1.24	1.20	1.16	1.13	1.09	1.06	1.03	1.00	0.97	0.94	0.92	0.89	0.87	0.84	0.82	0.80	0.77	0.75	0.73	0.71	0.69
5	40.94	20.47	13.61	10.25	8.32	6.96	5.97	5.25	4.69	4.24	3.88	3.57	3.31	3.08	2.90	2.73	2.57	2.43	2.31	2.19	2.09	2.00	1.91	1.83	1.75	1.68	1.62	1.56	1.50	1.45	1.40	1.35	1.30	1.26	1.21	1.17	1.13	1.10	1.06	1.03	1.00	0.97	0.93	0.90	0.88	0.85	0.82	0.80	0.77	0.75
6	50.78	25.27	16.77	12.55	10.10	8.41	7.19	6.29	5.59	5.03	4.58	4.19	3.87	3.59	3.35	3.14	2.94	2.78	2.62	2.48	2.36	2.25	2.14	2.04	1.95	1.87	1.79	1.72	1.65	1.59	1.53	1.47	1.42	1.36	1.31	1.27	1.22	1.18	1.14	1.10	1.07	1.03	1.00	0.96	0.93	0.90	0.87	0.84	0.81	0.79
7	57.80	28.69	19.03	14.20	11.37	9.44	8.06	7.03	6.23	5.60	5.07	4.64	4.26	3.94	3.67	3.43	3.21	3.02	2.85	2.69	2.55	2.43	2.31	2.20	2.10	2.00	1.91	1.83	1.76	1.69	1.62	1.56	1.50	1.44	1.39	1.34	1.29	1.24	1.20	1.15	1.11	1.08	1.04	1.00	0.97	0.94	0.90	0.87	0.84	0.81
8	63.07	31.26	20.72	15.43	12.32	10.22	8.72	7.59	6.72	6.02	5.44	4.97	4.56	4.21	3.92	3.65	3.41	3.21	3.02	2.85	2.70	2.56	2.43	2.31	2.20	2.10	2.01	1.92	1.84	1.76	1.69	1.62	1.56	1.50	1.44	1.39	1.33	1.29	1.24	1.19	1.15	1.11	1.07	1.03	1.00	0.96	0.93	0.90	0.86	0.83
9	67.16	33.26	22.04	16.39	13.06	10.83	9.23	8.02	7.09	6.35	5.73	5.23	4.79	4.42	4.10	3.82	3.57	3.35	3.15	2.97	2.81	2.66	2.53	2.40	2.29	2.18	2.08	1.99	1.90	1.82	1.75	1.67	1.61	1.54	1.48	1.43	1.37	1.32	1.27	1.23	1.18	1.14	1.10	1.06	1.02	0.98	0.95	0.91	0.88	0.85
10	70.44	34.86	23.09	17.16	13.65	11.31	9.63	8.36	7.39	6.61	5.96	5.43	4.98	4.59	4.25	3.96	3.70	3.46	3.25	3.07	2.90	2.74	2.60	2.47	2.35	2.24	2.14	2.04	1.95	1.87	1.79	1.71	1.65	1.58	1.52	1.46	1.40	1.35	1.30	1.25	1.20	1.16	1.12	1.08	1.04	1.00	0.96	0.93	0.90	0.86
11	73.12	36.17	23.95	17.79	14.14	11.70	9.97	8.65	7.63	6.82	6.15	5.60	5.13	4.72	4.38	4.07	3.80	3.56	3.34	3.14	2.97	2.81	2.66	2.53	2.41	2.29	2.18	2.09	1.99	1.91	1.82	1.75	1.68	1.61	1.54	1.48	1.43	1.37	1.32	1.27	1.22	1.18	1.13	1.09	1.05	1.01	0.98	0.94	0.91	0.87
12	75.36	37.26	24.67	18.31	14.54	12.03	10.24	8.88	7.84	7.00	6.31	5.74	5.25	4.84	4.48	4.16	3.88	3.64	3.41	3.21	3.03	2.87	2.72	2.58	2.45	2.33	2.22	2.12	2.03	1.94	1.85	1.78	1.70	1.63	1.57	1.51	1.45	1.39	1.34	1.29	1.24	1.19	1.15	1.10	1.06	1.03	0.99	0.95	0.92	0.88
13	77.25	38.18	25.28	18.75	14.89	12.31	10.48	9.08	8.01	7.15	6.44	5.86	5.36	4.93	4.57	4.24	3.96	3.70	3.47	3.27	3.08	2.92	2.76	2.62	2.49	2.37	2.26	2.15	2.05	1.96	1.88	1.80	1.72	1.65	1.59	1.52	1.46	1.41	1.35	1.30	1.25	1.20	1.16	1.12	1.07	1.04	1.00	0.96	0.93	0.89
14	78.87	38.97	25.80	19.13	15.18	12.55	10.68	9.25	8.16	7.28	6.56	5.96	5.45	5.02	4.64	4.31	4.02	3.76	3.52	3.32	3.13	2.96	2.80	2.66	2.52	2.40	2.29	2.18	2.08	1.99	1.90	1.82	1.74	1.67	1.60	1.54	1.48	1.42	1.36	1.31	1.26	1.21	1.17	1.12	1.08	1.04	1.00	0.97	0.93	0.90
15	80.27	39.66	26.25	19.46	15.43	12.76	10.86	9.40	8.29	7.40	6.66	6.05	5.53	5.09	4.71	4.37	4.07	3.81	3.57	3.36	3.17	2.99	2.83	2.69	2.55	2.43	2.31	2.20	2.10	2.01	1.92	1.84	1.76	1.69	1.62	1.55	1.49	1.43	1.38	1.32	1.27	1.22	1.18	1.13	1.09	1.05	1.01	0.97	0.94	0.90
16	81.50	40.26	26.65	19.75	15.66	12.94	11.01	9.53	8.40	7.50	6.74	6.13	5.60	5.15	4.76	4.42	4.12	3.85	3.61	3.39	3.20	3.02	2.86	2.71	2.58	2.45	2.33	2.22	2.12	2.02	1.94	1.85	1.77	1.70	1.63	1.56	1.50	1.44	1.39	1.33	1.28	1.23	1.18	1.14	1.10	1.06	1.02	0.98	0.94	0.91
17	82.58	40.79	27.00	20.00	15.85	13.10	11.14	9.64	8.50	7.58	6.82	6.20	5.66	5.21	4.81	4.46	4.16	3.89	3.64	3.43	3.23	3.05	2.89	2.74	2.60	2.47	2.35	2.24	2.14	2.04	1.95	1.87	1.79	1.71	1.64	1.58	1.51	1.45	1.39	1.34	1.29	1.24	1.19	1.15	1.10	1.06	1.02	0.98	0.95	0.91
18	83.55	41.26	27.31	20.23	16.03	13.24	11.26	9.74	8.59	7.66	6.89	6.26	5.72	5.26	4.86	4.50	4.20	3.92	3.67	3.46	3.26	3.08	2.91	2.76	2.62	2.49	2.37	2.26	2.15	2.05	1.96	1.88	1.80	1.72	1.65	1.58	1.52	1.46	1.40	1.35	1.30	1.25	1.20	1.15	1.11	1.07	1.03	0.99	0.95	0.92
19	84.41	41.68	27.58	20.43	16.18	13.37	11.37	9.84	8.67	7.73	6.95	6.31	5.77	5.30	4.90	4.54	4.23	3.95	3.70	3.48	3.28	3.10	2.93	2.78	2.64	2.50	2.38	2.27	2.16	2.07	1.98	1.89	1.81	1.73	1.66	1.59	1.53	1.47	1.41	1.35	1.30	1.25	1.20	1.16	1.11	1.07	1.03	0.99	0.96	0.92
20	85.19	42.06	27.83	20.61	16.32	13.48	11.47	9.92	8.74	7.79	7.00	6.36	5.81	5.34	4.93	4.57	4.26	3.98	3.73	3.50	3.30	3.12	2.95	2.79	2.65	2.52	2.40	2.28	2.18	2.08	1.99	1.90	1.82	1.74	1.67	1.60	1.54	1.47	1.42	1.36	1.31	1.26	1.21	1.16	1.12	1.08	1.04	1.00	0.96	0.92

Note: The table shows the simulated boundary values of mineval(Λ) for different values of maximum bias (in columns) and number of instruments K_2 (in rows). The simulations are based on 100,000 Monte Carlo replications, and follow Stock and Yogo (2005).

Table B.3: Simulated boundary values of mineval(Λ) for $n = 3$ endogenous regressors, for different values of maximum bias (in columns) and number of instruments K_2 (in rows)

K_2	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.2	0.21	0.22	0.23	0.24	0.25	0.26	0.27	0.28	0.29	0.3	0.31	0.32	0.33	0.34	0.35	0.36	0.37	0.38	0.39	0.4	0.41	0.42	0.43	0.44	0.45	0.46	0.47	0.48	0.49	0.5
5	21.41	10.94	7.44	5.80	4.76	4.02	3.54	3.18	2.90	2.65	2.47	2.31	2.17	2.06	1.97	1.87	1.79	1.71	1.64	1.58	1.53	1.47	1.42	1.37	1.33	1.29	1.25	1.21	1.18	1.14	1.11	1.08	1.05	1.02	0.99	0.96	0.94	0.91	0.88	0.86	0.84	0.82	0.79	0.77	0.75	0.73	0.71	0.69	0.67	0.66
6	34.55	17.33	11.63	8.84	7.14	5.97	5.16	4.56	4.10	3.71	3.40	3.14	2.92	2.73	2.57	2.43	2.30	2.18	2.07	1.97	1.89	1.81	1.73	1.66	1.60	1.54	1.48	1.43	1.38	1.33	1.29	1.25	1.20	1.17	1.13	1.09	1.06	1.02	0.99	0.96	0.93	0.90	0.88	0.85	0.83	0.80	0.78	0.75	0.73	0.71
7	43.93	21.90	14.62	11.01	8.84	7.36	6.32	5.55	4.95	4.46	4.06	3.73	3.45	3.21	3.01	2.82	2.66	2.51	2.37	2.26	2.15	2.05	1.96	1.87	1.79	1.72	1.65	1.59	1.53	1.47	1.42	1.37	1.32	1.27	1.23	1.18	1.14	1.11	1.07	1.03	1.00	0.97	0.94	0.91	0.88	0.85	0.82	0.80	0.77	0.75
8	50.97	25.32	16.87	12.64	10.11	8.40	7.19	6.29	5.60	5.02	4.56	4.18	3.85	3.57	3.33	3.12	2.93	2.76	2.60	2.47	2.34	2.23	2.12	2.02	1.94	1.85	1.78	1.70	1.63	1.57	1.51	1.45	1.40	1.35	1.30	1.25	1.21	1.17	1.13	1.09	1.05	1.02	0.98	0.95	0.92	0.89	0.86	0.83	0.80	0.78
9	56.44	27.98	18.61	13.90	11.10	9.21	7.87	6.87	6.10	5.46	4.95	4.52	4.16	3.85	3.58	3.35	3.14	2.95	2.78	2.63	2.49	2.37	2.25	2.15	2.05	1.96	1.87	1.79	1.72	1.65	1.58	1.52	1.47	1.41	1.36	1.31	1.26	1.22	1.17	1.13	1.09	1.05	1.02	0.98	0.95	0.92	0.89	0.86	0.83	0.80
10	60.82	30.11	20.01	14.91	11.89	9.86	8.41	7.33	6.49	5.81	5.26	4.80	4.41	4.08	3.79	3.53	3.31	3.10	2.92	2.76	2.61	2.48	2.36	2.24	2.14	2.04	1.95	1.87	1.79	1.71	1.64	1.58	1.52	1.46	1.40	1.35	1.30	1.25	1.21	1.16	1.12	1.08	1.05	1.01	0.98	0.94	0.91	0.88	0.85	0.82
11	64.40	31.86	21.15	15.74	12.54	10.39	8.85	7.71	6.82	6.10	5.51	5.02	4.61	4.26	3.95	3.68	3.45	3.23	3.04	2.87	2.71	2.57	2.44	2.32	2.21	2.11	2.01	1.93	1.84	1.77	1.69	1.62	1.56	1.50	1.44	1.39	1.33	1.28	1.24	1.19	1.15	1.11	1.07	1.03	1.00	0.96	0.93	0.89	0.86	0.83
12	67.39	33.31	22.11	16.43	13.08	10.83	9.22	8.02	7.09	6.34	5.72	5.21	4.78	4.41	4.09	3.81	3.56	3.33	3.14	2.96	2.80	2.65	2.51	2.39	2.27	2.17	2.07	1.98	1.89	1.81	1.73	1.66	1.60	1.53	1.47	1.42	1.36	1.31	1.26	1.22	1.17	1.13	1.09	1.05	1.01	0.98	0.94	0.91	0.88	0.85
13	69.92	34.54	22.91	17.02	13.54	11.20	9.53	8.29	7.33	6.54	5.90	5.37	4.92	4.54	4.21	3.91	3.66	3.42	3.22	3.03	2.87	2.71	2.57	2.44	2.32	2.21	2.11	2.02	1.93	1.85	1.77	1.69	1.63	1.56	1.50	1.44	1.39	1.33	1.28	1.23	1.19	1.15	1.10	1.06	1.03	0.99	0.95	0.92	0.89	0.86
14	72.08	35.59	23.60	17.52	13.93	11.53	9.80	8.51	7.52	6.72	6.05	5.51	5.05	4.65	4.31	4.00	3.74	3.50	3.29	3.10	2.92	2.77	2.62	2.49	2.37	2.25	2.15	2.05	1.96	1.88	1.80	1.72	1.65	1.58	1.52	1.46	1.41	1.35	1.30	1.25	1.21	1.16	1.12	1.08	1.04	1.00	0.97	0.93	0.90	0.86
15	73.96	36.50	24.20	17.95	14.27	11.80	10.03	8.71	7.69	6.87	6.18	5.63	5.15	4.75	4.39	4.08	3.81	3.57	3.35	3.15	2.98	2.82	2.67	2.53	2.41	2.29	2.18	2.08	1.99	1.90	1.82	1.75	1.67	1.61	1.54	1.48	1.42	1.37	1.32	1.27	1.22	1.17	1.13	1.09	1.05	1.01	0.97	0.94	0.91	0.87
16	75.60	37.30	24.73	18.33	14.57	12.05	10.23	8.88	7.84	7.00	6.30	5.73	5.25	4.83	4.47	4.15	3.88	3.62	3.40	3.20	3.02	2.86	2.71	2.57	2.44	2.32	2.21	2.11	2.02	1.93	1.84	1.77	1.69	1.62	1.56	1.50	1.44	1.38	1.33	1.28	1.23	1.18	1.14	1.10	1.06	1.02	0.98	0.95	0.91	0.88
17	77.05	38.01	25.19	18.67	14.83	12.26	10.41	9.04	7.98	7.12	6.40	5.82	5.33	4.91	4.54	4.21	3.93	3.68	3.45	3.25	3.06	2.89	2.74	2.60	2.47	2.35	2.24	2.14	2.04	1.95	1.86	1.79	1.71	1.64	1.57	1.51	1.45	1.40	1.34	1.29	1.24	1.19	1.15	1.11	1.07	1.03	0.99	0.95	0.92	0.89
18	78.34	38.64	25.60	18.97	15.06	12.45	10.57	9.17	8.09	7.22	6.49	5.90	5.40	4.97	4.60	4.27	3.98	3.72	3.49	3.28	3.10	2.93	2.77	2.63	2.50	2.37	2.26	2.16	2.06	1.97	1.88	1.80	1.73	1.65	1.59	1.52	1.46	1.41	1.35	1.30	1.25	1.20	1.16	1.12	1.07	1.03	1.00	0.96	0.92	0.89
19	79.49	39.20	25.97	19.23	15.27	12.62	10.71	9.29	8.20	7.31	6.58	5.97	5.47	5.03	4.65	4.31	4.03	3.76	3.53	3.32	3.13	2.96	2.80	2.65	2.52	2.40	2.28	2.18	2.08	1.98	1.90	1.82	1.74	1.67	1.60	1.54	1.47	1.42	1.36	1.31	1.26	1.21	1.17	1.12	1.08	1.04	1.00	0.97	0.93	0.90
20	80.53	39.70	26.30	19.47	15.46	12.78	10.84	9.40	8.29	7.39	6.65	6.04	5.53	5.09	4.70	4.36	4.07	3.80	3.56	3.35	3.16	2.98	2.82	2.68	2.54	2.41	2.30	2.19	2.09	2.00	1.91	1.83	1.75	1.68	1.61	1.55	1.48	1.43	1.37	1.32	1.27	1.22	1.17	1.13	1.09	1.05	1.01	0.97	0.93	0.90

Note: The table shows the simulated boundary values of mineval(Λ) for different values of maximum bias (in columns) and number of instruments K_2 (in rows). The simulations are based on 100,000 Monte Carlo replications, and follow Stock and Yogo (2005).

Table B.4: Simulated boundary values of mineval(Λ) for $n = 1$ endogenous regressor, for different values of maximum size distortion (in columns) and number of instruments K_2 (in rows)

K_2	0.07	0.08	0.09	0.1	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.2	0.21	0.22	0.23	0.24	0.25	0.26	0.27	0.28	0.29	0.3	0.31	0.32	0.33	0.34	0.35	0.36	0.37	0.38	0.39	0.4	0.41	0.42	0.43	0.44	0.45	0.46	0.47	0.48	0.49	0.5
1	19.28	11.52	8.96	5.87	4.42	3.45	2.79	2.31	1.89	1.65	1.36	1.10	0.96	0.89	0.82	0.75	0.67	0.59	0.51	0.47	0.44	0.41	0.37	0.34	0.30	0.27	0.24	0.23	0.21	0.19	0.18	0.16	0.15	0.13	0.12	0.11	0.10	0.10	0.09	0.08	0.07	0.06	0.06	0.05
2	26.85	18.18	13.67	10.61	8.38	7.06	6.02	5.23	4.66	4.02	3.71	3.38	3.03	2.77	2.57	2.38	2.21	2.07	1.94	1.83	1.73	1.65	1.57	1.49	1.41	1.34	1.28	1.22	1.17	1.12	1.06	1.02	0.97	0.94	0.90	0.86	0.82	0.79	0.75	0.72	0.68	0.65	0.62	0.60
3	33.84	23.14	17.29	13.81	11.06	9.45	8.11	7.13	6.38	5.56	5.16	4.76	4.30	3.94	3.67	3.41	3.19	2.99	2.83	2.67	2.53	2.42	2.30	2.20	2.09	2.00	1.91	1.83	1.76	1.68	1.61	1.54	1.48	1.43	1.38	1.32	1.27	1.22	1.17	1.12	1.07	1.03	0.99	0.95
4	40.68	27.67	20.66	16.63	13.43	11.53	9.91	8.77	7.84	6.90	6.39	5.92	5.37	4.94	4.60	4.29	4.01	3.78	3.58	3.39	3.22	3.07	2.93	2.79	2.66	2.55	2.44	2.34	2.25	2.16	2.07	1.98	1.91	1.85	1.78	1.71	1.65	1.59	1.53	1.47	1.41	1.36	1.31	1.26
5	47.45	32.03	23.91	19.29	15.66	13.49	11.59	10.31	9.20	8.15	7.54	6.99	6.36	5.87	5.47	5.11	4.78	4.51	4.27	4.05	3.85	3.67	3.50	3.35	3.20	3.06	2.93	2.81	2.70	2.59	2.49	2.39	2.31	2.23	2.15	2.07	2.00	1.93	1.86	1.79	1.73	1.67	1.61	1.56
6	54.20	36.31	27.11	21.88	17.83	15.38	13.22	11.79	10.51	9.36	8.64	8.01	7.30	6.76	6.31	5.90	5.53	5.21	4.94	4.69	4.46	4.25	4.06	3.88	3.71	3.54	3.40	3.26	3.14	3.01	2.90	2.79	2.69	2.60	2.51	2.42	2.34	2.26	2.18	2.11	2.03	1.97	1.90	1.84
7	60.93	40.53	30.28	24.42	19.96	17.24	14.81	13.25	11.78	10.55	9.71	9.00	8.23	7.64	7.13	6.67	6.25	5.89	5.59	5.32	5.06	4.81	4.60	4.40	4.20	4.02	3.86	3.70	3.56	3.42	3.29	3.17	3.06	2.96	2.86	2.76	2.67	2.58	2.50	2.41	2.33	2.26	2.19	2.12
8	67.65	44.73	33.43	26.94	22.07	19.08	16.38	14.69	13.04	11.72	10.77	9.98	9.14	8.50	7.94	7.43	6.97	6.57	6.24	5.94	5.65	5.37	5.13	4.91	4.69	4.49	4.31	4.14	3.98	3.83	3.68	3.55	3.43	3.31	3.20	3.09	2.99	2.90	2.81	2.71	2.63	2.55	2.47	2.39
9	74.37	48.91	36.57	29.43	24.17	20.91	17.94	16.11	14.28	12.89	11.82	10.95	10.04	9.36	8.75	8.19	7.68	7.24	6.87	6.55	6.23	5.92	5.66	5.42	5.18	4.95	4.76	4.57	4.39	4.22	4.07	3.92	3.79	3.66	3.54	3.42	3.32	3.21	3.11	3.01	2.92	2.83	2.75	2.66
10	81.08	53.07	39.70	31.92	26.25	22.72	19.49	17.53	15.52	14.04	12.86	11.91	10.94	10.21	9.54	8.94	8.38	7.90	7.51	7.15	6.81	6.47	6.19	5.92	5.66	5.42	5.20	4.99	4.80	4.62	4.45	4.30	4.15	4.01	3.88	3.75	3.64	3.52	3.42	3.31	3.21	3.12	3.02	2.94
11	87.78	57.22	42.82	34.39	28.33	24.53	21.03	18.94	16.75	15.19	13.90	12.87	11.83	11.05	10.33	9.68	9.08	8.56	8.14	7.76	7.39	7.02	6.71	6.42	6.14	5.87	5.64	5.42	5.21	5.01	4.83	4.67	4.51	4.36	4.22	4.08	3.95	3.83	3.72	3.61	3.50	3.40	3.30	3.21
12	94.49	61.37	45.94	36.86	30.40	26.33	22.57	20.35	17.97	16.34	14.93	13.82	12.72	11.89	11.12	10.43	9.78	9.22	8.76	8.36	7.96	7.56	7.23	6.92	6.62	6.33	6.08	5.84	5.61	5.41	5.21	5.03	4.87	4.70	4.55	4.40	4.27	4.14	4.02	3.90	3.79	3.68	3.58	3.47
13	101.19	65.51	49.05	39.32	32.47	28.13	24.11	21.75	19.19	17.49	15.96	14.77	13.60	12.73	11.91	11.17	10.48	9.87	9.39	8.96	8.53	8.10	7.74	7.42	7.09	6.78	6.51	6.26	6.02	5.80	5.59	5.40	5.22	5.05	4.88	4.73	4.59	4.45	4.32	4.20	4.08	3.96	3.85	3.74
14	107.89	69.64	52.16	41.78	34.53	29.93	25.64	23.15	20.41	18.63	16.99	15.72	14.49	13.57	12.70	11.91	11.17	10.53	10.01	9.56	9.11	8.64	8.26	7.91	7.57	7.24	6.95	6.68	6.42	6.19	5.97	5.77	5.58	5.39	5.22	5.05	4.90	4.76	4.62	4.49	4.36	4.24	4.13	4.01
15	114.59	73.77	55.27	44.23	36.59	31.72	27.16	24.55	21.63	19.77	18.01	16.66	15.37	14.40	13.48	12.65	11.86	11.18	10.63	10.15	9.68	9.18	8.78	8.41	8.04	7.69	7.38	7.09	6.82	6.58	6.35	6.13	5.93	5.73	5.55	5.37	5.22	5.06	4.92	4.78	4.65	4.52	4.40	4.28
16	121.28	77.90	58.38	46.68	38.65	33.51	28.69	25.94	22.84	20.91	19.03	17.61	16.25	15.24	14.26	13.39	12.55	11.83	11.25	10.75	10.25	9.72	9.29	8.90	8.51	8.14	7.82	7.51	7.23	6.97	6.72	6.50	6.28	6.08	5.88	5.70	5.53	5.37	5.22	5.07	4.94	4.80	4.67	4.55
17	127.98	82.02	61.48	49.13	40.70	35.30	30.21	27.34	24.05	22.05	20.06	18.55	17.12	16.07	15.04	14.12	13.25	12.48	11.87	11.34	10.81	10.25	9.80	9.40	8.98	8.59	8.25	7.93	7.63	7.35	7.10	6.86	6.63	6.42	6.21	6.02	5.85	5.68	5.52	5.37	5.22	5.08	4.95	4.81
18	134.68	86.15	64.58	51.58	42.76	37.08	31.74	28.73	25.27	23.18	21.08	19.49	18.00	16.90	15.82	14.86	13.94	13.13	12.49	11.94	11.38	10.79	10.32	9.89	9.46	9.04	8.68	8.34	8.03	7.74	7.48	7.23	6.99	6.76	6.54	6.34	6.16	5.98	5.81	5.66	5.51	5.36	5.22	5.08
19	141.37	90.27	67.69	54.02	44.81	38.87	33.26	30.12	26.48	24.32	22.09	20.43	18.88	17.73	16.60	15.59	14.63	13.78	13.11	12.53	11.95	11.33	10.83	10.38	9.93	9.49	9.12	8.76	8.43	8.13	7.85	7.59	7.34	7.10	6.87	6.66	6.47	6.29	6.11	5.95	5.79	5.64	5.49	5.34
20	148.07	94.39	70.79	56.47	46.86	40.65	34.78	31.51	27.69	25.46	23.11	21.37	19.75	18.57	17.38	16.33	15.32	14.43	13.73	13.13	12.52	11.86	11.34	10.87	10.40	9.94	9.55	9.18	8.83	8.51	8.23	7.95	7.69	7.44	7.20	6.98	6.79	6.59	6.41	6.24	6.08	5.92	5.76	5.61

Note: The table shows the simulated boundary values of mineval(Λ) for different values of maximum size distortion (in columns) and number of instruments K_2 (in rows). The nominal size of the Wald-test on the structural parameter β is 5%. The simulations are based on 100,000 Monte Carlo replications, and follow Stock and Yogo (2005).

Table B.5: Simulated boundary values of mineval(Λ) for $n = 2$ endogenous regressors, for different values of maximum size distortion (in columns) and number of instruments K_2 (in rows)

K_2	**0.06**	**0.07**	**0.08**	**0.09**	**0.1**	**0.11**	**0.12**	**0.13**	**0.14**	**0.15**	**0.16**	**0.17**	**0.18**	**0.19**	**0.2**	**0.21**	**0.22**	**0.23**	**0.24**	**0.25**	**0.26**	**0.27**	**0.28**	**0.29**	**0.3**	**0.31**	**0.32**	**0.33**	**0.34**	**0.35**	**0.36**	**0.37**	**0.38**	**0.39**	**0.4**	**0.41**	**0.42**	**0.43**	**0.44**	**0.45**	**0.46**	**0.47**	**0.48**	**0.49**	**0.5**
2	14.76	9.46	4.71	2.72	1.89	1.43	0.97	0.86	0.77	0.66	0.56	0.49	0.45	0.41	0.38	0.36	0.33	0.29	0.26	0.25	0.23	0.22	0.20	0.18	0.17	0.16	0.15	0.14	0.12	0.12	0.11	0.10	0.09	0.09	0.08	0.07	0.07	0.06	0.06	0.05	0.05	0.05	0.04	0.04	0.04
3	38.91	19.31	11.91	8.49	6.47	5.17	4.50	3.81	3.26	2.89	2.59	2.37	2.19	2.02	1.89	1.76	1.65	1.55	1.44	1.35	1.27	1.21	1.16	1.10	1.04	0.99	0.95	0.91	0.87	0.83	0.79	0.75	0.72	0.68	0.65	0.62	0.60	0.58	0.56	0.54	0.52	0.50	0.48	0.46	0.44
4	55.17	26.37	16.90	12.45	9.65	7.79	6.87	5.83	5.01	4.46	4.03	3.70	3.41	3.15	2.96	2.76	2.58	2.43	2.27	2.14	2.01	1.92	1.84	1.75	1.67	1.59	1.52	1.46	1.40	1.34	1.28	1.23	1.18	1.13	1.07	1.03	1.00	0.97	0.94	0.90	0.87	0.84	0.81	0.78	0.75
5	68.28	32.31	21.01	15.68	12.26	9.96	8.78	7.49	6.46	5.78	5.24	4.80	4.42	4.09	3.84	3.58	3.36	3.16	2.97	2.80	2.64	2.51	2.41	2.29	2.19	2.09	2.00	1.93	1.85	1.78	1.70	1.63	1.57	1.51	1.44	1.38	1.34	1.30	1.26	1.22	1.18	1.14	1.10	1.07	1.03
6	79.82	37.69	24.68	18.56	14.58	11.91	10.45	8.95	7.77	6.96	6.32	5.80	5.34	4.94	4.63	4.33	4.06	3.83	3.60	3.40	3.21	3.05	2.92	2.79	2.67	2.55	2.44	2.35	2.26	2.17	2.08	2.00	1.93	1.86	1.78	1.71	1.66	1.61	1.56	1.52	1.47	1.42	1.37	1.33	1.28
7	90.46	42.74	28.10	21.22	16.75	13.73	11.99	10.31	8.99	8.06	7.35	6.73	6.19	5.73	5.37	5.03	4.72	4.45	4.19	3.96	3.75	3.56	3.40	3.25	3.11	2.98	2.86	2.75	2.65	2.54	2.44	2.35	2.27	2.19	2.10	2.03	1.97	1.91	1.85	1.79	1.74	1.69	1.63	1.58	1.53
8	100.54	47.60	31.37	23.76	18.81	15.47	13.44	11.61	10.16	9.12	8.32	7.62	7.01	6.48	6.08	5.70	5.35	5.04	4.76	4.50	4.26	4.05	3.87	3.69	3.54	3.39	3.26	3.13	3.02	2.90	2.79	2.69	2.60	2.51	2.41	2.33	2.26	2.19	2.13	2.07	2.00	1.94	1.88	1.83	1.77
9	110.24	52.33	34.52	26.21	20.81	17.16	14.84	12.86	11.30	10.15	9.27	8.48	7.80	7.21	6.77	6.36	5.96	5.62	5.31	5.02	4.76	4.52	4.32	4.12	3.96	3.80	3.65	3.51	3.38	3.25	3.13	3.02	2.92	2.82	2.72	2.63	2.55	2.48	2.40	2.33	2.26	2.20	2.13	2.07	2.00
10	119.67	56.96	37.61	28.59	22.76	18.81	16.20	14.08	12.41	11.15	10.20	9.32	8.57	7.93	7.44	6.99	6.56	6.18	5.85	5.54	5.25	4.99	4.76	4.55	4.36	4.19	4.03	3.88	3.73	3.60	3.47	3.35	3.23	3.13	3.02	2.92	2.83	2.75	2.67	2.59	2.52	2.44	2.37	2.30	2.24
11	128.92	61.52	40.64	30.94	24.67	20.43	17.53	15.27	13.50	12.14	11.12	10.15	9.33	8.63	8.11	7.62	7.15	6.74	6.37	6.04	5.74	5.45	5.19	4.96	4.77	4.58	4.40	4.24	4.08	3.93	3.79	3.67	3.55	3.43	3.32	3.21	3.12	3.02	2.94	2.85	2.77	2.69	2.61	2.54	2.47
12	138.02	66.04	43.63	33.25	26.56	22.04	18.84	16.45	14.58	13.12	12.02	10.98	10.09	9.33	8.76	8.24	7.73	7.28	6.90	6.54	6.21	5.90	5.62	5.38	5.16	4.97	4.77	4.59	4.42	4.27	4.12	3.98	3.85	3.73	3.61	3.50	3.39	3.29	3.20	3.11	3.02	2.93	2.85	2.77	2.69
13	147.02	70.51	46.59	35.54	28.43	23.63	20.14	17.62	15.65	14.09	12.92	11.79	10.83	10.02	9.41	8.86	8.31	7.83	7.42	7.04	6.69	6.35	6.05	5.78	5.56	5.35	5.14	4.95	4.76	4.60	4.44	4.30	4.16	4.03	3.90	3.78	3.67	3.56	3.46	3.36	3.27	3.18	3.09	3.00	2.92
14	155.93	74.95	49.52	37.80	30.29	25.20	21.42	18.77	16.71	15.05	13.81	12.60	11.57	10.70	10.05	9.47	8.88	8.37	7.93	7.53	7.16	6.80	6.47	6.19	5.95	5.73	5.51	5.30	5.10	4.93	4.76	4.61	4.46	4.33	4.19	4.07	3.95	3.83	3.72	3.61	3.51	3.42	3.32	3.23	3.15
15	164.76	79.37	52.44	40.05	32.13	26.77	22.69	19.92	17.76	16.01	14.70	13.40	12.31	11.39	10.69	10.07	9.45	8.90	8.44	8.02	7.63	7.24	6.89	6.59	6.34	6.10	5.87	5.65	5.44	5.26	5.08	4.92	4.77	4.62	4.48	4.35	4.22	4.10	3.98	3.87	3.76	3.66	3.56	3.46	3.37
16	173.55	83.77	55.34	42.29	33.96	28.33	23.95	21.06	18.81	16.96	15.58	14.20	13.04	12.06	11.33	10.68	10.02	9.44	8.95	8.51	8.10	7.68	7.31	6.99	6.72	6.48	6.23	6.00	5.78	5.59	5.40	5.23	5.07	4.92	4.77	4.63	4.49	4.36	4.24	4.12	4.00	3.90	3.79	3.69	3.60
17	182.28	88.15	58.22	44.52	35.79	29.88	25.21	22.20	19.86	17.91	16.45	14.99	13.77	12.74	11.96	11.28	10.58	9.97	9.46	8.99	8.56	8.12	7.72	7.39	7.11	6.85	6.60	6.35	6.11	5.91	5.72	5.54	5.37	5.21	5.06	4.91	4.77	4.63	4.49	4.37	4.25	4.14	4.03	3.92	3.82
18	190.98	92.52	61.10	46.74	37.60	31.43	26.46	23.33	20.90	18.85	17.33	15.78	14.50	13.41	12.59	11.88	11.14	10.50	9.97	9.48	9.02	8.56	8.14	7.79	7.49	7.23	6.96	6.70	6.45	6.24	6.03	5.85	5.67	5.50	5.34	5.19	5.04	4.89	4.75	4.62	4.49	4.38	4.26	4.15	4.04
19	199.64	96.88	63.97	48.95	39.41	32.97	27.70	24.46	21.94	19.79	18.20	16.57	15.22	14.08	13.22	12.48	11.70	11.02	10.47	9.96	9.49	9.00	8.55	8.18	7.88	7.60	7.31	7.04	6.78	6.56	6.35	6.15	5.97	5.79	5.63	5.47	5.31	5.15	5.01	4.87	4.74	4.61	4.49	4.38	4.27
20	208.28	101.22	66.83	51.15	41.22	34.50	28.95	25.58	22.97	20.73	19.07	17.36	15.94	14.75	13.85	13.07	12.26	11.55	10.98	10.44	9.95	9.43	8.97	8.58	8.26	7.97	7.67	7.39	7.11	6.88	6.66	6.46	6.26	6.09	5.92	5.75	5.58	5.42	5.26	5.11	4.98	4.85	4.72	4.60	4.49

Note: The table shows the simulated boundary values of mineval(Λ) for different values of maximum size distortion (in columns) and number of instruments K_2 (in rows). The nominal size of the Wald-test on the structural parameter β is 5%. The simulations are based on 100,000 Monte Carlo replications, and follow Stock and Yogo (2005).

Table B.6: Simulated boundary values of mineval(Λ) for $n = 3$ endogenous regressors, for different values of maximum size distortion (in columns) and number of instruments K_2 (in rows)

K_2	**0.06**	**0.07**	**0.08**	**0.09**	**0.1**	**0.11**	**0.12**	**0.13**	**0.14**	**0.15**	**0.16**	**0.17**	**0.18**	**0.19**	**0.2**	**0.21**	**0.22**	**0.23**	**0.24**	**0.25**	**0.26**	**0.27**	**0.28**	**0.29**	**0.3**	**0.31**	**0.32**	**0.33**	**0.34**	**0.35**	**0.36**	**0.37**	**0.38**	**0.39**	**0.4**	**0.41**	**0.42**	**0.43**	**0.44**	**0.45**	**0.46**	**0.47**	**0.48**	**0.49**	**0.5**
3	11.62	2.91	1.37	0.96	0.77	0.63	0.47	0.45	0.42	0.38	0.35	0.32	0.29	0.26	0.24	0.22	0.20	0.19	0.18	0.17	0.16	0.15	0.14	0.13	0.11	0.11	0.10	0.09	0.09	0.08	0.08	0.07	0.07	0.06	0.06	0.05	0.05	0.05	0.04	0.04	0.04	0.03	0.03	0.03	0.03
4	32.12	12.43	8.07	5.98	4.61	3.68	3.14	2.72	2.39	2.17	1.95	1.79	1.66	1.52	1.39	1.30	1.21	1.14	1.08	1.03	0.98	0.92	0.88	0.83	0.79	0.75	0.71	0.68	0.65	0.62	0.60	0.58	0.56	0.54	0.52	0.50	0.47	0.46	0.44	0.42	0.40	0.38	0.37	0.36	0.34
5	46.51	19.61	12.98	9.67	7.47	6.01	5.17	4.45	3.92	3.55	3.18	2.93	2.72	2.51	2.30	2.15	2.01	1.89	1.79	1.71	1.63	1.54	1.47	1.40	1.33	1.27	1.21	1.15	1.11	1.06	1.02	0.99	0.96	0.92	0.89	0.86	0.83	0.80	0.77	0.74	0.71	0.68	0.66	0.64	0.61
6	57.84	25.62	17.00	12.70	9.85	7.97	6.87	5.92	5.23	4.72	4.25	3.91	3.63	3.35	3.09	2.88	2.70	2.55	2.42	2.30	2.19	2.08	1.98	1.89	1.81	1.73	1.65	1.57	1.51	1.45	1.40	1.36	1.31	1.27	1.23	1.19	1.14	1.10	1.06	1.03	0.99	0.95	0.92	0.89	0.86
7	67.42	30.96	20.50	15.34	11.94	9.73	8.39	7.23	6.41	5.77	5.21	4.80	4.46	4.12	3.81	3.56	3.34	3.15	2.99	2.85	2.71	2.57	2.46	2.35	2.24	2.15	2.05	1.96	1.89	1.82	1.76	1.70	1.64	1.59	1.54	1.49	1.44	1.39	1.34	1.30	1.25	1.21	1.17	1.14	1.10
8	75.92	35.88	23.69	17.75	13.87	11.36	9.80	8.46	7.52	6.75	6.11	5.63	5.23	4.85	4.48	4.19	3.93	3.72	3.52	3.36	3.19	3.03	2.90	2.78	2.66	2.54	2.44	2.33	2.25	2.17	2.09	2.02	1.96	1.90	1.84	1.78	1.72	1.66	1.61	1.56	1.50	1.46	1.41	1.37	1.33
9	83.69	40.52	26.66	20.00	15.68	12.90	11.12	9.61	8.57	7.67	6.96	6.42	5.97	5.53	5.13	4.80	4.51	4.26	4.04	3.85	3.66	3.48	3.33	3.19	3.05	2.93	2.81	2.69	2.59	2.50	2.42	2.34	2.26	2.19	2.13	2.06	1.99	1.93	1.87	1.81	1.75	1.70	1.65	1.60	1.55
10	90.94	44.97	29.48	22.14	17.41	14.38	12.40	10.73	9.59	8.57	7.79	7.18	6.68	6.20	5.76	5.39	5.07	4.79	4.54	4.32	4.11	3.91	3.75	3.59	3.44	3.30	3.17	3.04	2.93	2.83	2.73	2.65	2.56	2.48	2.41	2.33	2.26	2.19	2.12	2.06	1.99	1.93	1.88	1.83	1.77
11	97.83	49.27	32.20	24.20	19.08	15.81	13.64	11.81	10.58	9.43	8.60	7.93	7.37	6.85	6.37	5.96	5.61	5.30	5.03	4.78	4.55	4.34	4.15	3.98	3.81	3.66	3.52	3.38	3.26	3.15	3.05	2.95	2.85	2.77	2.68	2.60	2.52	2.45	2.37	2.30	2.23	2.17	2.11	2.05	1.99
12	104.44	53.47	34.83	26.20	20.71	17.22	14.84	12.87	11.55	10.28	9.39	8.65	8.05	7.48	6.97	6.53	6.14	5.81	5.51	5.24	4.99	4.75	4.55	4.36	4.18	4.02	3.87	3.72	3.59	3.47	3.35	3.24	3.14	3.05	2.96	2.87	2.78	2.70	2.62	2.54	2.47	2.40	2.33	2.27	2.21
13	110.84	57.59	37.40	28.15	22.30	18.60	16.03	13.91	12.50	11.12	10.16	9.37	8.72	8.11	7.56	7.08	6.67	6.30	5.98	5.69	5.42	5.16	4.95	4.74	4.55	4.37	4.21	4.05	3.91	3.78	3.66	3.54	3.43	3.32	3.23	3.13	3.04	2.95	2.87	2.78	2.70	2.63	2.56	2.49	2.42
14	117.07	61.64	39.93	30.07	23.86	19.96	17.19	14.93	13.44	11.94	10.93	10.08	9.38	8.73	8.15	7.63	7.19	6.80	6.45	6.13	5.84	5.57	5.34	5.12	4.91	4.72	4.55	4.38	4.23	4.09	3.96	3.83	3.71	3.60	3.49	3.39	3.29	3.20	3.11	3.02	2.93	2.86	2.78	2.71	2.64
15	123.16	65.64	42.41	31.96	25.41	21.30	18.34	15.94	14.37	12.75	11.69	10.78	10.03	9.34	8.73	8.18	7.71	7.29	6.92	6.57	6.26	5.97	5.72	5.49	5.27	5.07	4.89	4.71	4.54	4.40	4.26	4.12	3.99	3.87	3.76	3.65	3.55	3.45	3.35	3.26	3.17	3.08	3.00	2.93	2.85
16	129.14	69.60	44.86	33.82	26.93	22.63	19.48	16.95	15.30	13.56	12.44	11.48	10.68	9.95	9.31	8.72	8.22	7.77	7.38	7.01	6.68	6.37	6.11	5.86	5.63	5.41	5.22	5.03	4.86	4.70	4.55	4.41	4.27	4.14	4.02	3.91	3.80	3.69	3.59	3.49	3.40	3.31	3.22	3.14	3.06
17	135.04	73.53	47.29	35.66	28.45	23.95	20.62	17.94	16.22	14.35	13.19	12.17	11.33	10.55	9.88	9.26	8.73	8.26	7.84	7.44	7.09	6.77	6.49	6.23	5.98	5.76	5.55	5.36	5.17	5.01	4.85	4.69	4.55	4.41	4.29	4.17	4.05	3.94	3.83	3.73	3.63	3.53	3.45	3.36	3.28
18	140.86	77.42	49.69	37.49	29.95	25.26	21.74	18.93	17.13	15.15	13.94	12.86	11.97	11.15	10.45	9.79	9.24	8.74	8.29	7.88	7.51	7.17	6.87	6.59	6.33	6.10	5.89	5.68	5.49	5.31	5.14	4.98	4.83	4.68	4.55	4.42	4.30	4.18	4.07	3.96	3.86	3.76	3.66	3.58	3.49
19	146.62	81.29	52.07	39.30	31.44	26.56	22.85	19.92	18.04	15.94	14.68	13.54	12.60	11.75	11.02	10.33	9.75	9.22	8.75	8.31	7.92	7.56	7.25	6.96	6.69	6.44	6.22	6.00	5.80	5.62	5.44	5.26	5.10	4.95	4.81	4.68	4.55	4.43	4.31	4.19	4.08	3.98	3.88	3.79	3.70
20	152.32	85.14	54.44	41.11	32.92	27.86	23.97	20.90	18.94	16.72	15.41	14.22	13.24	12.34	11.58	10.86	10.25	9.69	9.20	8.74	8.33	7.95	7.63	7.32	7.04	6.78	6.55	6.32	6.11	5.92	5.73	5.55	5.38	5.22	5.07	4.93	4.80	4.67	4.54	4.43	4.31	4.21	4.10	4.00	3.91

Note: The table shows the simulated boundary values of mineval(Λ) for different values of maximum size distortion (in columns) and number of instruments K_2 (in rows). The nominal size of the Wald-test on the structural parameter β is 5%. The simulations are based on 100,000 Monte Carlo replications, and follow Stock and Yogo (2005).

C Additional Monte Carlo results

This appendix contains the mean and median lengths of the proposed confidence intervals for each of the Monte Carlo DGPs in Section 3.3. The number of Monte Carlo replications was 2000 in each case.

Table C.1: Homoskedastic IV model, mean lengths of confidence intervals for mineval(Λ), $n = 1$ endogenous regressor, nominal level $(1 - \alpha) = 0.90$

	$K_2 = 1$				$K_2 = 2$				$K_2 = 3$				$K_2 = 4$			
mineval(Λ) =	0	1	10	25	0	1	10	25	0	1	10	25	0	1	10	25
$T = 100$	6.37	8.59	20.98	33.14	6.07	8.10	19.80	30.75	5.90	7.90	19.01	29.65	5.59	7.76	18.27	28.70
$T = 250$	6.31	8.69	20.66	33.07	5.97	8.17	19.54	30.79	5.73	7.84	19.04	29.33	5.64	7.70	18.44	28.17
$T = 500$	6.38	8.27	20.58	32.72	5.79	8.22	19.81	30.84	5.77	7.74	18.92	29.29	5.53	7.65	18.16	28.33
$T = 1000$	6.39	8.51	20.64	32.86	6.07	7.92	19.66	30.66	5.72	7.74	18.99	29.24	5.54	7.67	18.26	28.37

Note: The table shows the mean lengths of the proposed confidence interval for mineval(Λ) across 2000 Monte Carlo simulations for different sample sizes T, instrument strength mineval(Λ), and number of instruments K_2.

Table C.2: Homoskedastic IV model, median lengths of confidence intervals for mineval(Λ), $n = 1$ endogenous regressor, nominal level $(1 - \alpha) = 0.90$

	$K_2 = 1$				$K_2 = 2$				$K_2 = 3$				$K_2 = 4$			
mineval(Λ) =	0	1	10	25	0	1	10	25	0	1	10	25	0	1	10	25
$T = 100$	5.43	7.39	20.91	32.81	5.65	7.70	19.48	30.47	5.62	7.58	18.61	29.50	5.32	7.53	18.03	28.58
$T = 250$	5.28	7.50	20.73	32.96	5.38	7.58	19.49	30.64	5.34	7.45	18.92	29.09	5.37	7.48	18.47	28.19
$T = 500$	5.50	7.07	20.55	32.85	5.33	7.79	19.76	30.75	5.44	7.40	18.93	29.29	5.32	7.43	18.18	28.31
$T = 1000$	5.39	7.40	20.96	32.72	5.59	7.46	19.58	30.59	5.43	7.46	19.01	29.27	5.32	7.35	18.19	28.40

Note: The table shows the median lengths of the proposed confidence interval for mineval(Λ) across 2000 Monte Carlo simulations for different sample sizes T, instrument strength mineval(Λ), and number of instruments K_2.

Table C.3: Homoskedastic IV model, mean lengths of confidence intervals for mineval(Λ), $n = 2$ endogenous regressors, nominal level $(1 - \alpha) = 0.90$

	$K_2 = 2$				$K_2 = 3$				$K_2 = 4$				$K_2 = 5$			
mineval(Λ) =	0	1	10	25	0	1	10	25	0	1	10	25	0	1	10	25
$T = 100$	4.44	7.46	24.68	39.44	4.71	9.93	24.24	37.82	4.64	9.68	23.82	36.61	4.57	9.36	23.10	35.99
$T = 250$	4.43	7.68	24.89	39.41	4.58	9.70	24.33	37.77	4.58	9.61	23.78	36.68	4.64	9.36	23.32	36.16
$T = 500$	4.36	7.54	24.90	39.79	4.55	9.90	24.14	37.83	4.55	9.60	23.62	36.70	4.57	9.57	23.23	36.01
$T = 1000$	4.37	7.55	24.61	39.78	4.55	9.60	24.49	37.91	4.53	9.58	23.64	36.79	4.60	9.48	23.07	36.12

Note: The table shows the mean lengths of the proposed confidence interval formineval(Λ) across 2000 Monte Carlo simulations for different sample sizes T, instrument strength mineval(Λ), and number of instruments K_2.

Table C.4: Homoskedastic IV model, median lengths of confidence intervals for mineval(Λ), $n = 2$ endogenous regressors, nominal level $(1 - \alpha) = 0.90$

	$K_2 = 2$				$K_2 = 3$				$K_2 = 4$				$K_2 = 5$			
mineval(Λ) =	0	1	10	25	0	1	10	25	0	1	10	25	0	1	10	25
$T = 100$	4.19	6.89	24.70	39.41	4.53	9.54	23.86	37.49	4.54	9.49	23.72	36.26	4.46	9.15	22.97	35.69
$T = 250$	4.22	7.32	24.87	39.59	4.45	9.26	24.21	37.65	4.43	9.33	23.83	36.46	4.55	9.10	23.24	36.10
$T = 500$	4.13	7.06	25.01	39.88	4.30	9.36	24.02	37.94	4.43	9.34	23.75	36.66	4.50	9.16	23.30	36.04
$T = 1000$	4.13	7.14	24.57	39.83	4.38	9.10	24.36	37.94	4.43	9.32	23.68	36.90	4.48	9.34	23.02	36.02

Note: The table shows the median lengths of the proposed confidence interval for mineval(Λ) across 2000 Monte Carlo simulations for different sample sizes T, instrument strength mineval(Λ), and number of instruments K_2.

Table C.5: Heteroskedastic IV model (DGP 1), mean lengths of confidence intervals for μ^2, nominal level $(1-\alpha) = 0.90$

	$K_2 = 1$				$K_2 = 2$				$K_2 = 3$				$K_2 = 4$			
$\mu^2 =$	0	1	10	25	0	1	10	25	0	1	10	25	0	1	10	25
$T = 100$	6.19	8.58	20.52	28.97	6.81	9.78	20.08	27.17	8.15	10.62	20.10	25.81	9.12	12.32	20.95	24.99
$T = 250$	6.50	8.99	20.85	30.72	6.67	9.05	20.02	29.48	7.60	9.97	20.76	29.26	8.53	11.33	22.01	29.76
$T = 500$	6.28	8.63	21.23	32.34	6.23	8.29	20.81	29.80	6.71	9.60	20.62	29.76	7.78	10.44	21.54	30.75
$T = 1000$	6.53	8.55	19.80	32.04	6.22	8.21	19.65	31.37	6.54	8.75	20.57	28.79	7.29	9.60	20.68	29.46

Note: The table shows the mean lengths of confidence intervals for μ^2 across 2000 Monte Carlo simulations for different sample sizes T, instrument strength μ^2, and number of instruments K_2.

Table C.6: Heteroskedastic IV model (DGP 1), median lengths of confidence intervals for μ^2, nominal level $(1-\alpha) = 0.90$

	$K_2 = 1$				$K_2 = 2$				$K_2 = 3$				$K_2 = 4$			
$\mu^2 =$	0	1	10	25	0	1	10	25	0	1	10	25	0	1	10	25
$T = 100$	5.20	7.22	20.45	29.22	6.31	9.05	20.53	27.62	8.03	10.25	20.09	26.41	8.71	11.81	21.35	25.55
$T = 250$	5.58	7.69	20.90	30.37	6.31	8.55	19.57	29.96	7.26	9.78	20.70	29.99	8.09	11.02	22.13	30.70
$T = 500$	5.21	7.70	20.77	32.54	5.86	7.44	20.42	30.09	6.47	9.23	20.37	29.80	7.53	10.05	21.39	30.32
$T = 1000$	5.45	7.44	19.66	32.46	5.73	7.89	19.40	31.50	6.13	8.50	20.07	29.31	7.16	9.13	20.53	29.17

Note: The table shows the median lengths of confidence intervals for μ^2 across 2000 Monte Carlo simulations for different sample sizes T, instrument strength μ^2, and number of instruments K_2.

Table C.7: Heteroskedastic and autocorrelated IV model (DGP 2), mean lengths of confidence intervals for μ^2, nominal level $(1-\alpha) = 0.90$

	$K_2 = 1$				$K_2 = 2$				$K_2 = 3$				$K_2 = 4$			
$\mu^2 =$	0	1	10	25	0	1	10	25	0	1	10	25	0	1	10	25
$T = 100$	7.03	10.49	24.41	32.51	7.35	10.40	21.86	28.02	8.50	10.83	19.89	25.31	8.78	10.81	19.25	23.15
$T = 250$	6.66	9.82	24.59	35.95	7.58	10.08	22.18	32.49	7.86	10.46	20.96	29.10	8.16	10.16	20.50	27.43
$T = 500$	6.53	9.52	24.66	38.11	7.51	10.55	23.39	34.37	7.65	9.87	21.92	31.69	7.90	10.45	21.16	30.25
$T = 1000$	6.69	9.74	25.38	38.86	7.26	10.37	22.56	34.96	7.71	9.94	22.07	33.01	7.94	10.01	21.04	31.44

Note: The table shows the mean lengths of confidence intervals for μ^2 across 2000 Monte Carlo simulations for different sample sizes T, instrument strength μ^2, and number of instruments K_2.

Table C.8: Heteroskedastic and autocorrelated IV model (DGP 2), median lengths of confidence intervals for μ^2, nominal level $(1-\alpha) = 0.90$

	$K_2 = 1$				$K_2 = 2$				$K_2 = 3$				$K_2 = 4$			
$\mu^2 =$	0	1	10	25	0	1	10	25	0	1	10	25	0	1	10	25
$T = 100$	5.82	9.83	24.99	32.36	6.71	9.62	21.55	28.32	7.67	10.14	19.70	25.53	8.25	10.43	19.08	23.23
$T = 250$	5.33	8.94	24.13	35.95	6.50	8.89	21.89	32.17	7.38	9.84	20.76	29.05	7.79	9.33	20.06	27.69
$T = 500$	5.51	8.62	25.48	38.24	6.59	9.68	23.30	34.20	7.15	9.22	21.92	31.25	7.27	9.89	21.43	30.45
$T = 1000$	5.49	8.77	25.39	38.95	6.43	9.57	22.99	35.35	7.23	8.96	22.02	33.34	7.35	9.49	20.78	31.39

Note: The table shows the median lengths of confidence intervals for μ^2 across 2000 Monte Carlo simulations for different sample sizes T, instrument strength μ^2, and number of instruments K_2.

Table C.9: Homoskedastic external instrument SVAR (DGP 1), mean lengths of confidence intervals for mineval(Λ), nominal level $(1-\alpha) = 0.90$

$\Lambda =$	0.01	1	10	25
$T = 100$	6.40	8.32	20.82	32.87
$T = 150$	6.61	8.14	20.66	32.56
$T = 200$	6.69	8.40	21.10	32.79
$T = 500$	6.64	8.30	20.84	32.59

Note: The table shows the mean lengths of the proposed confidence interval for mineval(Λ) across 2000 Monte Carlo simulations for different sample sizes T, and external instrument strength mineval(Λ).

Table C.10: Homoskedastic external instrument SVAR (DGP 1), median lengths of confidence intervals for mineval(Λ), nominal level $(1-\alpha) = 0.90$

$\Lambda =$	0.01	1	10	25
$T = 100$	5.45	7.15	20.79	32.89
$T = 150$	5.69	7.07	20.64	32.55
$T = 200$	5.71	7.22	21.13	32.74
$T = 500$	5.68	7.00	20.92	32.53

Note: The table shows the median lengths of the proposed confidence interval for mineval(Λ) across 2000 Monte Carlo simulations for different sample sizes T, and external instrument strength mineval(Λ).

Table C.11: Heteroskedastic and autocorrelated external instrument SVAR, mean lengths of confidence intervals for μ^2, nominal level $(1-\alpha) = 0.90$

	DGP 2, Heteroskedastic				DGP 3, Autocorrelated			
$\mu^2 =$	0.01	1	10	25	0.01	1	10	25
$T = 100$	6.93	9.00	22.51	34.09	6.90	9.11	22.24	33.78
$T = 150$	6.84	8.77	21.74	34.04	7.04	9.02	21.97	33.26
$T = 200$	6.80	8.82	21.52	33.71	6.73	8.73	21.56	33.79
$T = 500$	6.52	8.47	21.11	33.35	6.73	8.43	21.31	33.65

Note: The table shows the mean lengths of confidence intervals for μ^2 across 2000 Monte Carlo simulations for different sample sizes T, and external instrument strength μ^2. Asymptotic variance W_2 estimated by Gonçalves and White's (2005) bootstrap with 2999 bootstrap samples, with block length equal to one in the case of DGP 2, and block length equal to $\lfloor T^{1/3} \rfloor$ in the case of DGP 3.

Table C.12: Heteroskedastic and autocorrelated external instrument SVAR, mean lengths of confidence intervals for μ^2, nominal level $(1-\alpha) = 0.90$

	DGP 2, Heteroskedastic				DGP 3, Autocorrelated			
$\mu^2 =$	0.01	1	10	25	0.01	1	10	25
$T = 100$	5.90	7.84	22.46	33.65	5.77	7.78	21.88	33.25
$T = 150$	5.77	7.58	21.45	34.05	5.88	7.76	21.81	32.93
$T = 200$	5.72	7.66	21.58	33.63	5.71	7.42	21.30	33.44
$T = 500$	5.63	7.27	20.90	33.13	5.64	7.29	21.00	33.51

Note: The table shows the median lengths of confidence intervals for μ^2 across 2000 Monte Carlo simulations for different sample sizes T, and external instrument strength μ^2. Asymptotic variance W_2 estimated by Gonçalves and White's (2005) bootstrap with 2999 bootstrap samples, with block length equal to one in the case of DGP 2, and block length equal to $\lfloor T^{1/3} \rfloor$ in the case of DGP 3.

D Data appendix

New Keynesian Phillips Curve

All US data series were downloaded from the St. Louis Fed's FRED database and cover the period 1960:Q1 to 2017:Q1. Most series are readily available at the quarterly frequency and the ones on the monthly frequency were transformed by taking quarterly averages. Our choice of the series is motivated by Galí and Gertler (1999).

The series, their mnemonics and the transformations applied as follows:

- inflation π_t: Gross Domestic Product: Implicit Price Deflator, logarithmic difference,
- labor share s_t: nonfarm business sector labor share, PRS85006173, log difference from sample average,
- wage inflation: nonfarm compensation per hour, COMPNFB, in logarithmic difference,
- output gap: real GDP, GDPC1, cyclical component (retaining fluctuations between 6 and 32 quarters) of the Baxter-King (1999) filtered logarithmic real GDP,
- interest rate spread: difference of 10-Year Treasury Constant Maturity Rate (GS10) and 3-month Treasury Bill: Secondary Market Rate (TB3MS),
- commodity price inflation: Producer Price Index for All Commodities, PPIACO, logarithmic difference.

Consumption Euler equation

The choice of the data series follows Yogo (2004) and Montiel Olea and Pflueger (2013). In particular, we used quarterly US data covering the period 1960:Q1 to 2017:Q1. We adopt the *beginning of the period* timing convention for consumption. All the data were downloaded from the St. Louis Fed's FRED database, except the data required to compute the Dividend/Price ratio, which were obtained from the Center for Research in Security Prices (CRSP) via the CRSPSift interface.

The series, their mnemonics and the transformations applied are as follows:

- consumption growth Δc_t: percentage growth rate of the sum of Real personal consumption expenditures per capita: Services (A797RX0Q048SBEA) and Real personal consumption expenditures per capita: Goods: Nondurable goods (A796RX0Q048SBEA),

- asset return r_t: 100 times the logarithm of the real return on the 3-month Treasury Bill: Secondary Market Rate (TB3MS), which was calculated as one plus the 3-month Treasury Bill rate divided by 400 minus the logarithmic difference of the Consumer Price Index for All Urban Consumers: All Items (CPIAUCSL),

- 3-month T-bill rate: 3-month Treasury Bill: Secondary Market Rate (TB3MS, divided by four)),

- inflation $\Delta \log \text{CPI}_t$: 100 times the logarithmic difference of the Consumer Price Index for All Urban Consumers: All Items (CPIAUCSL),

- $\log(\text{Div}/\text{Price})_t$: logarithm of the Dividend/Price ratio in quarter t, where the dividends are cumulated over the 11 months preceding and the last month of quarter t (12 months in total), and the resulting sum is divided by the Index Level Associated with VWRETX. The dividends in each month are calculated as the product of the Index Level Associated with VWRETX and the dividend yield. The dividend yield is calculated as (1+Value-Weighted Return-incl. dividends)/(1+Value-Weighted Return-excl. dividends)-1.

Structural VAR identified with oil shocks

The choice of the data series follows Montiel Olea et al. (2016). Due to data availability, the sample period covers 1971:Q1 to 2004:Q3. All quarterly US data were downloaded from the St. Louis Fed's FRED database, except Kilian's (2008) OPEC shortfall series, which was obtained from the author's website.

The series, their mnemonics and the transformations applied are as follows:

- inflation $\Delta \log \text{CPI}_t$: logarithmic difference of the Consumer Price Index for All Urban Consumers: All Items (CPIAUCSL),
- first difference of the Fed funds rate (FEDFUNDS),
- real GDP growth: logarithmic difference of real GDP (GDPC1),
- oil price growth: logarithmic difference of oil price (Producer Price Index by Commodity for Fuels and Related Products and Power: Crude Petroleum (Domestic Production), WPU0561),
- Hamilton's (2003) twelve-month maximum deviation series $z_t^{\text{H},12}$: if in quarter t the oil price (WPU0561 series) was higher than the highest oil price in the preceding 4 quarters $(t-1, t-2, t-3, t-4)$, then $z_t^{\text{H},12}$ is equal to the price of oil in quarter t minus the highest value in quarters $(t-1, t-2, t-3, t-4)$, otherwise $z_t^{\text{H},12}$ takes the value zero.
- Kilian's (2008) OPEC shortfall series was obtained from the author's website. For a detailed description of the series, we refer to Kilian (2008).

Bibliography

Adjemian, S., Bastani, H., Juillard, M., Karamé, F., Perendia, G., Pfeifer, J., Ratto, M., and Villemot, S. (2011). Dynare: Reference Manual, Version 4. *Dynare Working Papers*.

An, S. and Schorfheide, F. (2007). Bayesian Analysis of DSGE Models. *Econometric Reviews*, 26(2-4):113–172.

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.

Baxter, M. and King, R. G. (1999). Measuring business cycles: Approximate band-pass filters for economic time series. *The Review of Economics and Statistics*, 81(4):575–593.

Bernanke, B., Gertler, M., and Gilchrist, S. (1999). The financial accelerator in a quantitative business cycle framework. In Taylor, J. B. and Woodford, M., editors, *Handbook of Macroeconomics*, volume 1C.

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.

Calvo, G. A. (1983). Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics*, 12(3):383–398.

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

Castelnuovo, E. (2012). Fitting U.S. Trend Inflation: A Rolling-Window Approach. In Balke, N. S., Canova, F., Milani, F., and Wynne, M. A., editors, *DSGE Models in Macroeconomics: Estimation, Evaluation, and New Developments*, volume 28 of *Advances in Econometrics*, pages 201–252. Emerald Group Publishing Limited.

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.

Cragg, J. G. and Donald, S. G. (1993). Testing Identifiability and Specification in Instrumental Variable Models. *Econometric Theory*, 9(2):222–240.

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.

Edge, R. M., Gürkaynak, R. S., Reis, R., and Sims, C. A. (2010). How useful are estimated DSGE model forecasts for central bankers? [with Comments and Discussion]. *Brookings Papers on Economic Activity*, 41(2):209–259.

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.

Fernández-Villaverde, J. and Rubio-Ramírez, J. F. (2005). Estimating Dynamic Equilibrium Economies: Linear versus Nonlinear Likelihood. *Journal of Applied Econometrics*, 20(7):891–910.

Galí, J. and Gertler, M. (1999). Inflation dynamics: A structural econometric analysis. *Journal of Monetary Economics*, 44(2):195–222.

Geweke, J. and Amisano, G. (2011). Optimal prediction pools. *Journal of Econome-*

trics, 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.

Gonçalves, S. and White, H. (2005). Bootstrap Standard Error Estimates for Linear Regression. *Journal of the American Statistical Association*, 100(471):970–979.

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, P., Horowitz, J. L., and Jing, B. Y. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika*, 82(3):561–574.

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

Hamilton, J. (2003). What is an oil shock? *Journal of Econometrics*, 113(2):363–398.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton, New Jersey.

Hansen, B. E. (1999). The grid bootstrap and the autoregressive model. *The Review of Economics and Statistics*, 81(4):594–607.

Hansen, P. R. (2005). A Test for Superior Predictive Ability. *Journal of Business & Economic Statistics*, 23(4):365–380.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The Model Confidence Set. *Econometrica*, 79(2):453–497.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of signifi-

cance. *Biometrika*, 75(4):800–802.

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.

Kilian, L. (2008). Exogenous oil supply shocks: How big are they and how much do they matter for the US economy? *The Review of Economics and Statistics*, 90(2):216–240.

Kleibergen, F. and Mavroeidis, S. (2009). Weak Instrument Robust Tests in GMM and the New Keynesian Phillips Curve. *Journal of Business & Economic Statistics*, 27(3):293–311.

Kolasa, M. and Rubaszek, M. (2015). Forecasting using DSGE models with financial frictions. *International Journal of Forecasting*, 31(1):1–19.

Lunsford, K. G. (2016). Identifying Structural VARs with a Proxy Variable and a Test for a Weak Proxy. Working paper.

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.

Mavroeidis, S., Plagborg-Møller, M., and Stock, J. H. (2014). Empirical Evidence on Inflation Expectations in the New Keynesian Phillips Curve. *Journal of Economic Literature*, 52(1):124–188.

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

Mertens, K. and Ravn, M. O. (2013). The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States. *American Economic Review*, 103(4):1212–1247.

Mincer, J. A. and Zarnowitz, V. (1969). The Evaluation of Economic Forecasts. In

Mincer, J. A., editor, *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, pages 3–46. National Bureau of Economic Research, New York.

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.

Montiel Olea, J. L. and Pflueger, C. (2013). A Robust Test for Weak Instruments. *Journal of Business & Economic Statistics*, 31(3):358–369.

Montiel Olea, J. L., Stock, J., and Watson, M. W. (2012). Inference in Structural VARs with External Instruments. Presentation slides.

Montiel Olea, J. L., Stock, J. H., and Watson, M. W. (2016). Uniform Inference in SVARs Identified with External Instruments. Working paper.

Nagar, A. L. (1959). The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations. *Econometrica*, 27(4):575.

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. (2011). Understanding models' forecasting performance. *Journal of Econometrics*, 164(1):158–172.

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.

Rothenberg, T. J. (1984). Approximating the distributions of econometric estimators and test statistics. In Griliches, Z. and Intriligator, M. D., editors, *Handbook of Econometrics*, volume 2 of *Handbook of Econometrics*, pages 881–935. Elsevier.

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

Smets, F. and Wouters, R. (2007). Shocks and frictions in US business cycles: A Bayesian DSGE approach. *American Economic Review*, 97(3):586–606.

Staiger, D. and Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65(3):557–586.

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.

Stock, J. H. and Watson, M. W. (2012). Disentangling the Channels of the 2007–09 Recession. *Brookings Papers on Economic Activity*, (2012 Spring):81–135.

Stock, J. H., Wright, J. H., and Yogo, M. (2002). A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *Journal of Business & Economic Statistics*, 20(4):518–529.

Stock, J. H. and Yogo, M. (2005). Testing for Weak Instruments in Linear IV Regression. In Andrews, D. W. K., editor, *Identification and Inference for Econometric Models*, pages 80–108. Cambridge University Press, New York.

Sundaram, R. K. (1996). *A First Course in Optimization Theory*. Cambridge University Press, New York.

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.

West, K. D. (1996). Asymptotic Inference about Predictive Ability. *Econometrica*, 64(5):1067.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48(4):817–838.

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.

Yogo, M. (2004). Estimating the elasticity of intertemporal substitution when instruments are weak. *The Review of Economics and Statistics*, 86(3):797–810.

Zhang, F. (2011). *Matrix Theory*. Universitext. Springer New York, New York, NY.