DEUTSCHE
BUNDESBANK
EUROSYSTEM

Discussion Paper

Deutsche Bundesbank
No 07/2018

How far can we forecast? Statistical tests of the predictive content

Jörg Breitung
(University of Cologne)

Malte Knüppel
(Deutsche Bundesbank)

Deutsche Bundesbank, Wilhelm-Epstein-Straße 14, 60431 Frankfurt am Main,
Postfach 10 06 02, 60006 Frankfurt am Main

Tel +49 69 9566-0

Please address all orders in writing to: Deutsche Bundesbank,
Press and Public Relations Division, at the above address or via fax +49 69 9566-3077

Internet http://www.bundesbank.de

ISBN 978–3–95729–436–4 (Printversion)
ISBN 978–3–95729–437–1 (Internetversion)

Non-technical summary

Research Question

Forecasts are often made for several consecutive periods ahead. For example, Consensus Economics collects quarterly forecasts for up to six quarters into the future from research institutes and other professional forecasters. Yet, especially longer-term forecasts possibly do not provide any information beyond that contained in the long-run mean of the target variable. Such forecasts are deemed to be uninformative. Therefore, it is desirable to be able to determine the largest horizon for which informative forecasts can be made. Up to now, only descriptive methods have been available for this purpose.

Contribution

We develop two statistical tests designed to identify the largest forecast horizon for which forecasts are still informative. One of the tests compares the mean-squared prediction error to the variance of the target variable. If the mean-squared prediction error does not sufficiently exceed the variance, the forecast is classified as being informative. The other test focusses on the correlation of the forecasts with the target variable. According to this test, the forecast is classified as informative if the correlation is positive and sufficiently different from zero.

Results

The correlation-based test tends be more reliable in small samples according to our simulation results. We apply both tests to the macroeconomic forecasts provided by Consensus Economics for the G7-countries and the euro area, namely to the mean across the individual forecasts for the respective country. This mean forecast is commonly considered to be very accurate. It turns out that, for instance, concerning the growth rate of the real gross domestic product on average, across countries, the forecasts are informative for up to two quarters ahead only.

Nichttechnische Zusammenfassung

Fragestellung

Prognosen werden oft für mehrere aufeinanderfolgene Perioden erstellt. So reichen beispielsweise die durch Consensus Economics gesammelten Quartalsprognosen von Forschungsinstituten und anderen professionellen Prognostikern bis zu sechs Quartale in die Zukunft. Gerade bei längerfristigen Prognosen besteht aber die Möglichkeit, dass sie keine Informationen enthalten, die über den langfristigen Mittelwert der prognostizierten Variablen hinausgehen. Eine solche Prognose besitzt keine Aussagekraft mehr. Es ist daher wünschenswert, den größten Zeithorizont bestimmen zu können, an dem Prognosen noch aussagekräftig sind. Dafür stehen bisher lediglich deskriptive Methoden zur Verfügung.

Beitrag

Wir entwickeln zwei statistische Tests, die den größten noch aussagekräftigen Zeithorizont einer Prognose ermitteln können. Einer der Tests vergleicht den mittleren quadratischen Prognosefehler mit der Varianz der Zielgröße. Falls der mittlere quadratische Prognosefehler nicht ausreichend größer als die Varianz ausfällt, gilt die Prognose als aussagekräftig. Der andere Test untersucht die Korrelation der Prognosen mit den entsprechenden Beobachtungen. Nach diesem Test gilt die Prognose als aussagekräftig, wenn die Korrelation positiv und stark genug von null verschieden ist.

Ergebnisse

In Simulationen zeigt sich, dass der auf Korrelation beruhende Test in kleinen Stichproben etwas verlässlichere Ergebnisse liefert. Wir wenden beide Tests auf die von Consensus Economics gesammelten Quartalsprognosen für die G7-Länder und den Euroraum an, und zwar auf die Durchschnittsprognose, die sich aus dem Mittelwert der Einzelprognosen für das jeweilige Land ergibt. Diese Durchschnittsprognose gilt im Allgemeinen als sehr treffgenau. Dabei stellt sich zum Beispiel heraus, dass die Wachstumsrate des realen Bruttoinlandsprodukts im Mittel über alle betrachteten Länder nur für bis zu zwei Quartale im Voraus in aussagekräftiger Weise prognostiziert werden kann.

Deutsche Bundesbank Discussion Paper No 07/2018

How far can we forecast? Statistical tests of the predictive content*

Jörg Breitung
University of Cologne

Malte Knüppel
Deutsche Bundesbank

Abstract

Forecasts are useless whenever the forecast error variance fails to be smaller than the unconditional variance of the target variable. This paper develops tests for the null hypothesis that forecasts become uninformative beyond some limiting forecast horizon h^*. Following Diebold and Mariano (DM, 1995) we propose a test based on the comparison of the mean-squared error of the forecast and the sample variance. We show that the resulting test does not possess a limiting normal distribution and suggest two simple modifications of the DM-type test with different limiting null distributions. Furthermore, a forecast encompassing test is developed that tends to better control the size of the test. In our empirical analysis, we apply our tests to macroeconomic forecasts from the survey of Consensus Economics. Our results suggest that forecasts of macroeconomic key variables are barely informative beyond 2–4 quarters ahead.

Keywords: Hypothesis Testing, Predictive Accuracy, Informativeness

JEL classification: C12, C32, C53.

*Contact address: Jörg Breitung, University of Cologne, Institute of Econometrics, 50923 Cologne, Germany. Email: breitung@statistik.uni-koeln.de, malte.knueppel@bundesbank.de. We would like to thank Todd Clark, Matei Demetrescu, and seminar participants at the Workshop on Forecasting 2017 at the Deutsche Bundesbank, the Central Bank Forecasting Conference 2017 at the Federal Reserve Bank of St. Louis, and the 1st Vienna Workshop on Economic Forecasting 2018 for helpful comments and suggestions. The views expressed in this paper do not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem.

1 Introduction

The choice of the largest forecast horizon appears to be an important issue for decision-makers. For example, in recent years, several central banks, including the Federal Reserve and the European Central Bank, decided to increase the horizon of their macroeconomic forecasts.[1] Yet, it is unclear whether the additional forecasts for these larger horizons provide valuable information, since their forecast error variance might be as large as the unconditional variance of the target variable. While statistical tools for such an assessment, based on the approach of Parzen (1981), have been proposed in the literature, formal statistical tests have not been available. The purpose of this paper is to develop such tests, thereby determining the largest informative forecast horizon.

The empirical literature reports few and differing results concerning the largest informative forecast horizon. The differences are at least partly due to different transformations, as pointed out by Galbraith and Tkacz (2007). For example, concerning quarterly GDP, they find that forecasts of quarter-on-quarter growth are barely informative beyond a forecast horizon of one quarter, but that for year-on-year forecasts this horizon increases to about 4 quarters. Concerning annual GDP growth, Isiklar and Lahiri (2007) find that forecasts are informative for horizons up to 6 quarters. Diebold and Kilian (2001) report even larger horizons for HP-filtered or linearly detrended GDP.

The tests provided in this paper are directly related to the predictability measures used in the studies mentioned. Diebold and Kilian (2001) develop a measure for predictability by comparing the loss function (say mean-squared error) of the short-run and long-run forecasts. Since our focus is on forecasting stationary time series subject to a quadratic loss function, i.e. on conditional mean forecasts, our benchmark is the unconditional mean of the time series, as proposed in Nelson (1976), Parzen (1981) and Clements and Hendry (1998).[2] We also discuss, however, how our approach can be applied to nonstationary time series.

The predictability measure suggested by Nelson (1976) and others is asymptotically equivalent to the R^2 from a regression of the realized values on their h-step-ahead forecasts and a constant. Accordingly, the null hypothesis of no predictive power is equivalent to the null hypothesis that the forecasts are not correlated with the actual values. Indeed, this is the null hypothesis underlying the encompassing version of our predictability test. In contrast, our Diebold-Mariano (1995) type test statistic directly compares the loss associated with the model-based forecast and the unconditional mean, where the unconditional mean is estimated by the mean of the evaluation sample. Therefore, the only data required for both tests are the forecasts and the actual values within the evaluation sample. This setup makes the tests applicable to forecasts from unknown models like survey forecasts.

[1] See Knüppel (2018) for a survey.

[2] Diebold and Kilian (2001) attribute this predictability measure to Granger and Newbold (1986).

This feature of the tests is important because such forecasts are often considered to be more accurate than other common forecasting approaches, as documented for inflation survey forecasts by Ang, Bekaert, and Wei (2007).

It is important to notice that, in general, the model-based forecast function nests a constant as a special case. Accordingly, comparing the model-based forecast and the unconditional mean should be treated as a nested forecast comparison in the spirit of Clark and McCracken (2001) and West (1996). It is therefore not surprising that the Diebold-Mariano type statistic has a nonstandard limiting distribution. To sidestep this difficulty we suggest a simple modification of the test statistic that results in an asymptotically χ^2 distributed random variable provided the null hypothesis is true. On the other hand, our encompassing variant of the test is equivalent to the (HAC) t-statistic from a regression of the actual values on the forecasts and a constant. We provide conditions for the standard limiting null distribution which provide a reasonable approximation in empirical practice.

The rest of this paper is organized as follows. In Section 2 we introduce our testing framework and alternative concepts of predictability are discussed in Section 3. The Diebold-Mariano-type test and the encompassing test are analyzed in Sections 4 and 5. In Section 6 the local power of the tests is studied. Section 7 investigates the small sample properties by means of Monte Carlo experiments and in Section 8, the proposed tests are applied to forecasts of key macroeconomic variables as reported by Consensus Economics. Section 9 concludes.

2 Model framework

Let $\{y_{1+h}, \ldots, y_{n+h}\}$ denote the set of n observed actual values corresponding to the model forecasts $\widehat{y}_{t+h|t}$, $t = 1, \ldots, n$, based on the relevant information set $\mathcal{I}_t$ associated with time period t. We assume that y_t is generated by a stationary and ergodic stochastic process $\{Y_t\}$ and the model forecasts are realizations of the forecast generating process $\{Y^{\theta}_{t+h|t}\}$, where θ is the parameter vector of the forecasting model. As a simple example, assume that the target variable is generated by the univariate AR(1) process $Y_t = \alpha Y_{t-1} + u_t$ with $|\alpha| < 1$. In this example $Y^{\theta}_{t+h|t} = \alpha^h Y_t$ with $\theta = \alpha$. The actual forecast realization is denoted by $\widehat{y}_{t+h|t} = y^{\widehat{\theta}_t}_{t+h|t} = \widehat{\alpha}^h_t y_t$, where $\widehat{\theta}_t = \widehat{\alpha}_t$ denotes some consistent estimate of θ based on the observations up to period t. Following West (1996) we distinguish two different estimation schemes. The recursive scheme fixes the starting point of the estimation sample at $t = -T + 1$ and adapts an increasing end point such that $\mathcal{S}_{T:t} = \{-T+1, -T+2, \ldots, t\}$ and, hence, $\widehat{\theta}_t$ indicates an estimate based on $t+T$ observations. The rolling-window estimation scheme fixes the size of the estimation samples to T time periods, that is, $\mathcal{S}_{t|T} = \{t-T+1, t-T+2, \ldots, t\}$. It is important to note that we do not assume that the parameter estimates of the model and the sample size T are known. For our analysis we only need to observe the actual values $\{y_{1+h}, \ldots, y_{n+h}\}$ and their h-step

ahead forecasts $\{\widehat{y}_{1+h}, \ldots, \widehat{y}_{n+h}\}$.

The assumptions that characterize the process $\{Y_t\}$ are summarized in

Assumption 1 *Let $Y_t = \mu + u_t$ with $u_t = \phi(L)\varepsilon_t$, $\phi(L) = 1 + \phi_1 L + \phi_2 L^2 + \cdots$ is a lag polynomial with all roots outside the unit circle, $\sum_{i=1}^{\infty} |\phi_i| < \infty$ and ε_t is an i.i.d. white noise process with $\mathbb{E}(\varepsilon_t) = 0$ and $\mathbb{E}(\varepsilon_t^2) = \sigma_\varepsilon^2$. Furthermore $\mathbb{E}|\varepsilon_t|^{2+\delta} < \infty$ for some $\delta > 0$.*

Note that our null hypothesis of an uninformative forecast implies restrictions on the polynomial $\phi(L)$. For a univariate forecast, where $Y^{\theta}_{t+h|t}$ is a function of $Y_t, Y_{t-1}, \ldots$, the null hypothesis requires $\phi_s = 0$ for $s \geq h$.

In the next section it is argued that if Y_t is integrated of order one (that is ΔY_t is stationary) the tests are applied to the differenced series $\Delta Y_t = Y_t - Y_{t-1}$. The assumptions of a linear process and constant variances are not essential and may be relaxed at the cost of a more demanding asymptotic framework. We are interested in testing the null hypothesis that the forecast function $Y^{\theta}_{t+h|t}$ is not informative for Y_{t+h} in the sense that

$$H_0 : \quad \mathbb{E}(e^2_{t+h|t}) \geq \mathbb{E}(Y_{t+h} - \mu)^2, \tag{1}$$

where $e_{t+h|t} = Y_{t+h} - Y^{\theta}_{t+h|t}$ is the "theoretical" forecast error where θ is assumed to be known. We define the *maximum forecast horizon* h^* as $h^* = h_{\min} - 1$, where $h_{\min}$ is the smallest value of h for which condition (1) is fulfilled.

Let us consider a simple dynamic regression model of the form

$$Y_t = \alpha + \beta X_{t-1} + e_t \ , \tag{2}$$

where X_t and e_t are independent white noise processes with expectation zero and $\beta \neq 0$. If the forecast is correctly specified $Y^{\theta}_{t+1|t} = \alpha + \beta X_t$ and $Y^{\theta}_{t+h|t} = \alpha$ for $h > 1$. For $h = 1$ the condition (1) is violated but for $h = 2, 3, \ldots$ the forecast becomes uninformative. Accordingly the maximum forecast horizon is $h^* = 1$.

More insight can be gained by rewriting the mean-squared prediction error (MSPE) as

$$\mathbb{E}(e^2_{t+h|t}) = \mathbb{E}(Y_{t+h} - \mu)^2 - 2\mathbb{E}(Y_{t+h} - \mu)(Y_{t+h|t} - \mu) + \mathbb{E}(Y_{t+h|t} - \mu)^2. \tag{3}$$

Obviously, a sufficient condition for the forecast being uninformative is $Y_{t+h|t} = \mu$. Another sufficient condition is that $\mathbb{E}(Y_{t+h} - \mu)(Y_{t+h|t} - \mu) = 0$ as $\mathbb{E}(Y_{t+h|t} - \mu)^2 \geq 0$. Furthermore for a rational forecast with $\mathbb{E}(e_{t+h|t} Y^{\theta}_{t+h|t}) = 0$ it follows that

$$\begin{aligned} \mathbb{E}(Y_{t+h} - \mu)(Y^{\theta}_{t+h|t} - \mu) &= \mathbb{E}(e_{t+h} + Y^{\theta}_{t+h|t} - \mu)(Y^{\theta}_{t+h|t} - \mu) && (4) \\ &= \mathbb{E}(Y^{\theta}_{t+h|t} - \mu)^2 && (5) \end{aligned}$$

Combining (3) and (5) yields

$$\mathbb{E}(e^2_{t+h|t}) = \mathbb{E}(Y_{t+h} - \mu)^2 - \mathbb{E}(Y^{\theta}_{t+h|t} - \mu)^2. \tag{6}$$

Thus, for rational forecasts the conditions (i) $Y^{\theta}_{t+h|t} = \mu$ and (ii) $\text{cov}(Y_{t+h}, Y^{\theta}_{t+h|t}) = 0$ are equivalent to the null hypothesis (1).

It is important to note that the maximum forecast horizon h^* can be identified by sequentially applying a consistent test for (1). The null hypothesis (1) is tested for $h = 1, 2, ...$ until it is not rejected for the first time. Then, h^* is identified as the penultimate horizon tested. Provided that the tests are consistent, this identification is correct with probability approaching $1-\alpha$ as $n \to \infty$ with α denoting the significance level of the test. Therefore α must tend to zero to achieve a consistent selection rule for h^* (see Remark 5 below). It should be noted that the forecast error variances of rational forecasts are monotonously increasing with respect to the forecast horizon. Thus, if some forecast is informative at some horizon h it must also be uninformative for any higher horizon and we therefore can stop the testing sequence whenever the test does not reject for the first time.

3 Measuring predictability

For assessing the predictive content, Theil (1958) proposed (among other measures) the following inequality coefficient:

$$U2(h) \;=\; \sqrt{\frac{\sum_{t=1}^{n}(Y_{t+h} - \widehat{Y}_{t+h|t})^2}{\sum_{t=1}^{n}(Y_{t+h} - Y^0_{t+h})^2}}$$

where Y^0_{t+h} denotes some "naive forecast" (typically the no-change forecast). The model based forecast is uninformative whenever $U2(h)$ is close to unity. For a stationary variable the unconditional mean is a natural "naive" (resp. uninformative) forecast, whereas the no-change forecast is better suited for nonstationary (integrated) target variables; see e.g. Isiklar and Lahiri (2007) for an application to the Survey of Professional Forecasters.

If the unconditional mean is employed as the benchmark, the inequality coefficient U2(h) is related to the $R^2(h)$ measure proposed by Nelson (1976) and Diebold and Kilian (2001) given by

$$R^2(h) \;=\; 1 - \frac{var(\widehat{e}_{t+h|t})}{var(Y_{t+h})}\;,$$

where $\widehat{e}_{t+h|t} = Y_{t+h} - \widehat{Y}_{t+h|t}$ denotes the model-based forecast error. In practice the variance is estimated by the sample variance such that $R^2(h) = 1 - U2(h)^2$ with $Y^0_{t+h} =$

$\overline{Y}_h = n^{-1}\sum_{t=1}^{n} Y_{t+h}$ yielding the sample analog

$$\widehat{R}^2(h) \;=\; 1 - \frac{\sum_{t=1}^{n} \widehat{e}^2_{t+h|t}}{\sum_{t=1}^{n} \left(Y_{t+h} - \overline{Y}_h\right)^2}. \tag{7}$$

Note however that this measure may become negative. An alternative measure with the usual properties is obtained as the R^2 from a (Mincer-Zarnowitz type) regression of Y_{t+h} on the forecast $\widehat{Y}_{t+h|t}$ yielding the square of the sample correlation between the actual values and the forecasts (see Section 5).

Diebold and Kilian (2001) proposed a generalized measure of predictability

$$Q(\mathcal{L}, h, k) \;=\; 1 - \frac{\mathbb{E}[\mathcal{L}(e_{t+h|t})]}{\mathbb{E}[\mathcal{L}(e_{t+k|t})]} \quad \text{for } k > h,$$

where $\mathcal{L}(\cdot)$ indicates the loss function and $e_{t+k|t}$ denotes the long-run prediction error. If (i) Y_t is stationary, (ii) the loss function is quadratic and (iii) $k \to \infty$, the Diebold-Kilian measure and the Nelson measure coincide. In what follows we focus on the Nelson measure and propose a test for the hypothesis $R^2(h) = 0$.

As argued by Diebold and Kilian (2001) their predictability measure is also valid for nonstationary variables, whereas in this case the Nelson measure tends to unity as $n \to \infty$, no matter of the predictive content.[3] The latter approach remains valid, however, if it is applied to the differenced series. Since

$$\begin{aligned} e_{t+h|t} &= Y_{t+h} - Y^{\theta}_{t+h|t} \\ &= (Y_{t+h} - Y_t) - (Y^{\theta}_{t+h|t} - Y_t) \\ &= \left(\sum_{s=1}^{h} \Delta Y_{t+s}\right) - \left(\sum_{s=1}^{h} \Delta Y^{\theta}_{t+s|t}\right) \\ &= \sum_{s=1}^{h} \Delta e_{t+s|t} \ , \end{aligned}$$

with $\Delta e_{t+s|t} = \Delta Y_{t+s} - \Delta Y^{\theta}_{t+s|t}$ and $\Delta Y^{\theta}_{t+s|t} = Y^{\theta}_{t+s|t} - Y^{\theta}_{t+s-1|t}$ is the forecast of the differenced series, it follows that Y_{t+h} is not predictable whenever $\{\Delta Y_{t+1}, \cdots, \Delta Y_{t+h}\}$ are jointly unpredictable. Hence, for nonstationary (integrated) time series the predictability tests are applied to the differenced series for $s \in \{1, \ldots, h\}$.

[3] More precisely, if Y_t is $I(1)$ and the forecast error $e_{t+h|t}$ is $I(0)$ for fixed h, then the sample analog of the ratio $var(e_{t+h|t})/var(Y_{t+h})$ is $O_p(T^{-1})$ such that $\widehat{R}^2$ tends to unity.

4 Diebold-Mariano type test statistics

A natural test statistic for the hypothesis (1) is the statistic proposed by Diebold and Mariano (1995), which compares the sample MSPE of two competitive forecasts. In our case we are interested in analyzing the loss differential of the model-based forecast $\widehat{Y}_{t+h|t}$ and the uninformative forecast $Y_t - \widehat{\mu}$, where $\widehat{\mu}$ denotes some suitable estimator of the unconditional mean. It is important to notice that for any reasonable model forecast that allows for a constant mean as a special case, the forecasts are nested in the sense of West (1996). Therefore, under the null hypothesis the loss differential is driven by the estimation errors $\widehat{\theta} - \theta$ and $\widehat{\mu} - \mu$. It is well known that in such cases the limiting distribution of the Diebold-Mariano statistic is nonstandard and depends on unknown nuisance parameters (cf. West (1996) and Clark and McCracken (2001)).

In what follows we sidestep this problem by using estimators of θ and μ from different samples. The model parameters θ are estimated recursively[4] using the observations $\mathcal{S}_{-T:t} = \{-T+1, -T+2, \ldots, t\}$ whereas the evaluation sample is $\mathcal{F}^h_{1:n} = \{1+h, 2+h, \ldots, n+h\}$. Notice that these two samples overlap for all $t \geq 1+h$. If the estimation sample $\mathcal{S}_{-T:t}$ is large relative to the evaluation sample, this overlap will be asymptotically negligible. Specifically, we decompose the recursive estimator as $\widehat{\theta}_t = \widehat{\theta}_0 + O_p(n^{1/2}/T)$, where θ_0 (the estimator up to $t = 0$, that is, without overlapping observations) is such that the distribution of $\widehat{\theta}_t$ is asymptotically independent of $\widehat{\mu}_h = n^{-1}(Y_{1+h}, Y_{2+h} + \cdots + Y_{n+h}) = \overline{Y}_h$ under the null hypothesis. This escapes the problem of a nested forecast comparison.[5]

Since our test focuses on information of the evaluation sample we use $\overline{Y}_h = n^{-1}\sum_{t=h+1}^{n+h} Y_t$ as an estimator for $\mu = \mathbb{E}(Y_t)$. For the forecast functions $Y^{\theta}_{t+h|t}$ and $\widehat{Y}_{t+h|t} = Y^{\widehat{\theta}_t}_{t+h|t}$ we make the following assumption:

Assumption 2 *(i) Under the null hypothesis there exists some h^* such that $Y^{\theta}_{t+h|t} = \mu$ for all $h > h^*$. (ii) Under the null hypothesis, $u_{t+h} = Y_{t+h} - \mu$ is independent of the estimation error: $\mathbb{E}(u_{t+h}|\widehat{\theta}_t, \widehat{\theta}_{t-1}, \ldots) = 0$. (iii) The parameters are estimated consistently with*

$$a) \quad \widehat{\theta}_0 - \theta = O_p(T^{-1/2})$$

$$b) \quad \sup_{t\in\{1,\ldots,n\}} ||\widehat{\theta}_t - \widehat{\theta}_0|| = O_p\left(\frac{\sqrt{n}}{T}\right)$$

(iv) Let $D_{t+h}(\theta) = \partial Y^{\theta}_{t+h|t}/\partial\theta$ and $\overline{D}_h(\theta) = n^{-1}\sum_{t=1}^{n} D_{t+h}(\theta)$. For all $\theta^* \in [\theta - \epsilon, \theta + \epsilon]$

[4] Our analysis carries over to a rolling window estimation scheme if we assume that the window size T gets large relative to the size of the evaluation period.

[5] A related but fundamentally different approach is suggested by Calhoun (2016), where a fixed-length rolling window and a recursive estimation scheme are used to compute the forecast errors of the two competing forecasting methods.

with some $\epsilon > 0$

$$\frac{1}{n}\sum_{t=1}^{n}[D_{t+h}(\theta^*) - \overline{D}_h(\theta^*)]^2 \xrightarrow{p} \overline{D}^2 \quad \textit{with } 0 < \overline{D}^2 < \infty$$
$$\mathbb{E}|D_{t+h}(\theta^*)u_{t+h}|^{2+\delta} < \infty \quad \textit{for some } \delta > 0 \textit{ and all } t.$$

Part *(i)* is the null hypothesis of the test. Part *(ii)* is an implication of the null hypothesis which claims that the time series is not predictable given the information set $\mathcal{I}_t$, which includes the estimation error $\widehat{\theta}_t - \theta$. Part *(iii) a)* supposes the usual convergence rate of the estimation error in the estimated parameter vector $\widehat{\theta}_0$ based on the pre-evaluation sample $\{-T+1, \ldots, 0\}$, whereas *(iii) b)* limits the variation of estimators in the recursive estimation scheme within the evaluation sample.

To illustrate this assumption, consider the forecast based on the regression model with $\widehat{Y}_{t+h|t} = \widehat{\alpha}_t + \widehat{\beta}_t x_t$, where $\widehat{\beta}_t$ is the least squares estimator based on the $T + t$ time periods $\{-T+1, \ldots, t\}$. If x_t is stationary and, without loss of generality, $\mathbb{E}(x_t) = 0$, then $\mathbb{E}(\widehat{\beta}_0 - \beta)^2 = \sigma_u^2/(T\sigma_x^2)$, where $\sigma_x^2 = \mathbb{E}(x_t^2)$. Obviously, Assumption 2 *(iii) a)* is fulfilled. To analyze $\widehat{\beta}_t - \widehat{\beta}_0$ we write

$$\widehat{\beta}_0 = \frac{\sum_{s=-T+1}^{0} x_s u_s}{\sum_{s=-T+1}^{0} x_s^2} = (1+\kappa_{T,t})\frac{\sum_{s=-T+1}^{0} x_s u_s}{\sum_{s=-T+1}^{t} x_s^2} \quad \text{where } \kappa_{T,t} = \frac{\sum_{s=1}^{t} x_s^2}{\sum_{s=-T+1}^{0} x_s^2} = O_p\left(\frac{t}{T}\right).$$

It follows for $t \leq n$ that

$$\begin{aligned}\widehat{\beta}_t - \widehat{\beta}_0 &= \frac{\sum_{s=1}^{t} x_s u_s}{\sum_{s=-T+1}^{t} x_s^2} + O_p\left(\frac{t}{T^{3/2}}\right) \\ &= O_p\left(\frac{\sqrt{t}}{T}\right) + O_p\left(\frac{t}{T}\cdot\frac{1}{\sqrt{T}}\right) = O_p\left(\frac{\sqrt{t}}{T}\right).\end{aligned}$$

Hence, also part *(iii) b)* is satisfied. In our simple example $\overline{D}^2 = \sigma_x^2$ and, thus, part *(iv)* is fulfilled as well.

Let us first consider a test statistic constructed in the spirit of Diebold and Mariano (1995):

$$d_h = \frac{1}{\widehat{\omega}_\delta \sqrt{n}} \sum_{t=1}^{n} \delta_t^h \,, \tag{8}$$

where $\delta_t^h = \widehat{e}_{t+h|t}^2 - (Y_{t+h} - \overline{Y}_h)^2$ is the loss differential, $\overline{Y}_h = n^{-1}\sum_{t=1}^{n} Y_{t+h}$ denotes the evaluation sample mean, and $\widehat{\omega}_\delta^2$ denotes a consistent long-run variance estimator applied to δ_t^h. The following theorem presents the asymptotic distribution for the case that the number of observations T for estimating the parameters is large relative to the number of forecasts n:

Theorem 1 *Assume that Assumptions 1–2 hold. If* $T \to \infty$, $n \to \infty$, $n/T \to 0$ *we have*

$$d_{h^*} = \sqrt{n}\frac{|\overline{u}|}{2\widehat{\omega}_u} + O_p\left(\frac{n}{T}\right) \xrightarrow{d} \frac{|z|}{2},$$

where z *is a standard normally distributed random variable and* $\widehat{\omega}_u^2$ *denotes the analogous estimator for the long-run variance* $\omega_u^2 = \lim_{n\to\infty} \mathbb{E}\left(n^{-1}\sum_{t=1}^{n} u_{t+h}\right)^2$ *of* $u_{t+h} = Y_{t+h} - \mu$.

This finding gives rise to two variants of an adjusted Diebold-Mariano statistic:

Corollary 2 *Under Assumptions 1–2,* $h > h^*$, $T \to \infty$, $n \to \infty$, $n/T \to 0$ *it follows that*

$$2d_h \xrightarrow{d} |\mathcal{N}(0,1)|$$
$$\widetilde{d}_h = \frac{1}{\widehat{\omega}_u^2}\sum_{t=1}^{n} \delta_t^h \xrightarrow{d} \chi_1^2$$

where $\widehat{\omega}_u^2$ *is a consistent estimator for the long-run variance of* $u_t = Y_t - \mu$.

Both tests reject for *small* values of the test statistics. For example, for a significance level of 0.05, the critical value for $\widetilde{d}_h$ is 0.0039 and the corresponding value for $2d_h$ is 0.0627. It might be interesting to note that, for this reason, a test based on $(2d_h)^2$ and the test based on $\widetilde{d}_h$ are not asymptotically equivalent although the limiting distribution under the null hypothesis is the same. To be more precise, under the alternative, the test statistics $2d_h$ and $\widetilde{d}_h$ tend to be negative, such that squaring $2d_h$ can yield *large positive* values.

REMARK 1: Under the alternative with $\mathbb{E}(Y_{t+h} - \widehat{Y}_{t+h|t})^2 < \mathbb{E}(u_{t+h}^2)$ it follows that $d_h = O_p(\sqrt{n})$, whereas $\widetilde{d}_h = O_p(n)$. This is due to the fact that $\widehat{\omega}_\delta^2 = 4\overline{u}_h^2\widehat{\omega}_u^2 + O_p(T^{-1/2}) + O_p(n^{-2})$. Since under a fixed alternative $\overline{u}_h^2 = O_p(1)$ (instead of $O_p(n^{-1})$ under the null hypothesis) the denominator of $2d_h$ changes the order of magnitude while the order of magnitude of the denominator of $\widetilde{d}_h$ remains the same. This does not imply, however, that $2d_h$ is more powerful against alternatives in the vicinity of the null hypothesis. In fact under local alternatives both test statistics possess the same asymptotic power.

REMARK 2: The term $O_p(n/T)$ is driven by the estimation error $\widehat{\theta} - \theta$. Following West (1996) it is possible to work out the limiting distribution for the case that n/T converges to some constant. We do not think however that such limiting results are useful in practice as the asymptotic distribution involves the derivative $D_{t+h}(\theta)$ and the covariance matrix of $\widehat{\theta} - \theta$. Such information is typically not available or difficult to obtain (for example if the forecasts are based on a factor model).

In order to compare the small sample properties of the two test statistics of Corollary 2 we perform a Monte Carlo experiment, where the data are generated as $Y_{t+1} = \mu + u_{t+1}$.

Table 1: Actual sizes for various n/T combinations

	$n=25$		$n=50$		$n=100$		$n=200$	
T	$2d_1$	$\widetilde{d}_1$	$2d_1$	$\widetilde{d}_1$	$2d_1$	$\widetilde{d}_1$	$2d_1$	$\widetilde{d}_1$
50	0.087	0.092	0.070	0.073	0.044	0.047	0.024	0.027
100	0.100	0.105	0.088	0.093	0.066	0.069	0.041	0.043
200	0.113	0.119	0.109	0.115	0.083	0.088	0.062	0.065
500	0.114	0.120	0.110	0.118	0.110	0.116	0.096	0.102
1,000	0.111	0.119	0.110	0.119	0.120	0.127	0.118	0.126
10,000	0.081	0.088	0.090	0.098	0.102	0.109	0.108	0.115
50,000	0.059	0.062	0.063	0.066	0.073	0.077	0.077	0.082
500,000	0.050	0.050	0.054	0.056	0.056	0.059	0.064	0.068
∞	0.049	0.049	0.049	0.049	0.050	0.050	0.050	0.050

Note: The nominal size of the tests is 0.05. The limiting distributions of the DM-type test statistics $2d_1$ and $\widetilde{d}_1$ are presented in Corollary 2. For $T=\infty$ the test statistics are computed using the true parameter values. Results are based on 10,000 simulations. Tests statistics are based on OLS standard errors without degrees-of-freedom correction.

Since the test statistic is not affected by the value of μ and the variance σ_u^2, we set $\mu=0$ and $\sigma_u^2=1$. The forecast is based on the model $\widehat{Y}_{t+1|t}=\widehat{a}_t+\widehat{b}_tX_t$, where $\widehat{\theta}_t=(\widehat{a}_t,\widehat{\beta}_t)'$ denotes the vector of OLS estimates from a simple regression of Y_{t+1} on a constant and X_t based on the sample $\{-T+1,\ldots,t\}$. X_t is standard normally distributed. The nominal size of all tests is 0.05, and 10,000 replications are used to compute the rejection rates.

From the empirical sizes for various combinations of n and T presented in Table 1, it is evident that for realistic sample sizes such as $n=50$ and $T=200$, say, the tests suffer from a substantial size bias. This is not surprising as the critical value for the χ_1^2 distribution is very close to zero, and, thus, a large amount of probability mass is located in the vicinity of the critical value. Accordingly, even small, asymptotically negligible terms may have large effects on the actual size in finite samples. Very large estimation samples are needed in order to obtain actual sizes being close to their nominal counterpart.

Remark 3: It is interesting to analyze the properties of the test for situations where part *(ii)* of Assumption 2 is violated. Assume that the model forecast is uninformative and biased with $\mathbb{E}(Y_{t+h|t})=m\neq\mu$. It follows from the proof of Theorem 1 that

$$\frac{1}{n}\sum_{t=1+h}^{n+h}\delta_t^h=\overline{u}_h^2+(m-\mu)^2+O_p(T^{-1})\ .$$

Therefore if the forecast is biased, the test tends to be conservative in the sense that the rejection probability converges to zero as $n\to\infty$.

Next, assume that the uninformative forecast $Y^{\theta}_{t+h|t}$ is unbiased but different from a

constant: $Y^{\theta}_{t+h|t} = \mu + v^h_t$, where $var(v^h_t) = \tau^2_h > 0$. For example, consider the unbiased forecast function $Y^{\theta}_{t+h|t} = \beta X_t$ with and $X_t \overset{iid}{\sim} \mathcal{N}(\theta_1, \theta_2)$ and $\beta = \mu/\theta_1$. If X_t and Y_{t+h} are uncorrelated we have $Y^{\theta}_{t+h|t} = \mu + v^h_t$, where $v^h_t = \beta(X_t - \theta_1)$. It is not difficult to see that in this case $n^{-1}\sum_{t=1+h}^{n+h} \delta^h_t = \overline{u}^2_h + \tau^2_h + O_p(T^{-1})$ resulting again in a conservative test.

5 Encompassing tests

To overcome the small sample problems of the DM-type test statistic we consider a variant of the test based on the encompassing principle. As shown in Section 2, a test of the null hypothesis (1) is equivalent to the null hypothesis

$$H_0': \quad \mathbb{E}(Y_{t+h} - \mu)(Y^{\theta}_{t+h|t} - \mu) \;=\; 0$$

whenever the forecast is rational with $\mathbb{E}(Y_{t+h} - Y^{\theta}_{t+h|t}|Y^{\theta}_{t+h|t}) = 0$. Accordingly, the null hypothesis can be rejected if the correlation between Y_{t+h} and $\widehat{Y}_{t+h|t}$ is positive. Another view on the relationship between this approach and the DM-type statistic emerges from the decomposition

$$\begin{aligned}\sum_{t=1}^{n} \delta^h_t &= \sum_{t=1}^{n} \Big[Y_{t+h} - \overline{Y}_h - (\widehat{Y}_{t+h|t} - \overline{Y}_h)\Big]^2 - \big(Y_{t+h} - \overline{Y}_h\big)^2 \\ &= \sum_{t=1}^{n} (\widehat{Y}_{t+h|t} - \overline{Y}_h)^2 - 2\sum_{t=1}^{n} (Y_{t+h} - \overline{Y}_h)(\widehat{Y}_{t+h|t} - \overline{Y}_h).\end{aligned}$$

Since the first term on the right hand side is non-negative, it does not improve the power of the test and can be neglected. Thus, the DM-type test statistic is mainly driven by the covariance between Y_{t+h} and $\widehat{Y}_{t+h|t}$.

This gives rise to a one-sided t-test of $\beta_{1,h} = 0$ vs. $\beta_{1,h} > 0$ in the Mincer-Zarnowitz regression

$$Y_{t+h} = \beta_{0,h} + \beta_{1,h}\widehat{Y}_{t+h|t} + \epsilon_{t+h}\ , \tag{9}$$

where the error is typically autocorrelated up to lag $h-1$ due to the overlapping forecast horizon. Note that the Mincer-Zarnowitz test for rational (or efficient) forecasts focusses on the restrictions $\beta_{0,h} = 0$ and $\beta_{1,h} = 1$. Indeed, if the forecast is *informative* and rational, the forecast error $Y_{t+h} - \widehat{Y}_{t+h|t}$ should be uncorrelated with $\widehat{Y}_{t+h|t}$, which implies the restrictions considered by Mincer and Zarnowitz (1969). Our null hypothesis is that the forecast is *uninformative* which results in testing the null hypothesis $\beta_{1,h} = 0$ against the alternative $\beta_{1,h} > 0$, whereas the constant $\beta_{0,h} = \mu$ is unrestricted.

This test admits an interpretation as a forecast encompassing test (cf. Chong and Hendry, 1986, and Clements and Hendry, 1993) that is based on a convex combination of

the model-based forecast $\widehat{Y}_{t+h|t}$ and the unconditional mean $\overline{Y}_h$:

$$Y_{t+h} = \lambda_h \widehat{Y}_{t+h|t} + (1-\lambda)\overline{Y}_h + \epsilon_{t+h}$$
$$Y_{t+h} - \overline{Y}_h = \lambda_h (\widehat{Y}_{t+h|t} - \overline{Y}_h) + \epsilon_{t+h}$$

and, therefore, a test of $\beta_{1,h} = 0$ is equivalent to a test of $\lambda_h = 0$.

In our asymptotic analysis we focus on the LM-type test statistic:

$$\widehat{\varrho}_h = \frac{1}{\widehat{\omega}_\xi \sqrt{n}} \sum_{t=1}^{n} \xi_t^h \tag{10}$$

where

$$\xi_t^h = (Y_{t+h} - \overline{Y}_h)(\widehat{Y}_{t+h|t} - \overline{\widehat{Y}}_h)$$

with $\overline{\widehat{Y}}_h = n^{-1} \sum_{t=1}^{n} \widehat{Y}_{t+h|t}$ and $\widehat{\omega}_\xi^2$ denoting the long-run variance

$$\widehat{\omega}_\xi^2 = \widehat{\gamma}_0^\xi + 2 \sum_{j=1}^{k} w_j^k \widehat{\gamma}_j^\xi$$
$$\widehat{\gamma}_j^\xi = \frac{1}{n} \sum_{t=j+1}^{n} \xi_t^h \xi_{t-j}^h \ .$$

In practice any other asymptotically equivalent test such as the usual t-test of $\beta_{1,h} = 0$ in the regression (9) with appropriate HAC standard errors can be used.

A technical problem with a regression like (9) is that under the null hypothesis the regressor $\widehat{Y}_{t+h|t}$ and the constant are asymptotically collinear. To sidestep this problem we show in the proof of Theorem 3 that if $n/T \to 0$

$$\widehat{Y}_{t+h|t} - \overline{Y}_h \ \approx \ (\widehat{\theta}_0 - \theta)(D_{t+h}(\theta) - \overline{D}_h(\theta))$$

and, thus, the test of $\beta_{1,h} = 0$ in regression (9) is asymptotically equivalent to the test of $\beta_{1,h}^* = 0$ in the regressions

$$Y_{t+h} = \beta_0^* + \beta_{1,h}^* D_{t+h}(\theta) + \eta_{t+h}$$
$$\text{or} \quad Y_{t+h} = \beta_{0,h}^* + \beta_{1,h}^* D_{t+h}(\widehat{\theta}) + \widetilde{\eta}_{t+h} \ ,$$

where $\beta_{1,h}^* = (\widehat{\theta}_0 - \theta)\beta_{1,h}$. The details are provided in the proof of

Theorem 3 *Under Assumptions 1–2, a recursive forecasting scheme with $h > h^*$, $T \to \infty$, $n \to \infty$ and $n/T \to 0$ we have $\widehat{\varrho}_h \xrightarrow{d} \mathcal{N}(0,1)$, where $\widehat{\varrho}_h$ is defined in (10).*

REMARK 4: It is interesting to note that the DM-type test statistic can be interpreted as the likelihood ratio test of the null hypothesis $\beta_{1,h} = 0$ against the joint alternative

Table 2: Actual sizes for various n/T combinations

	$n=25$		$n=50$		$n=100$		$n=200$	
T	$\widehat{\beta}_{1,1}$	$\widehat{\varrho}_1$	$\widehat{\beta}_{1,1}$	$\widehat{\varrho}_1$	$\widehat{\beta}_{1,1}$	$\widehat{\varrho}_1$	$\widehat{\beta}_{1,1}$	$\widehat{\varrho}_1$
50	0.025	0.022	0.017	0.018	0.014	0.016	0.012	0.012
100	0.027	0.024	0.022	0.021	0.014	0.014	0.015	0.015
200	0.029	0.028	0.024	0.024	0.017	0.018	0.015	0.016
500	0.040	0.033	0.028	0.026	0.021	0.021	0.019	0.020
1,000	0.041	0.037	0.033	0.032	0.025	0.025	0.023	0.023
10,000	0.050	0.044	0.041	0.039	0.037	0.035	0.032	0.032
50,000	0.062	0.052	0.052	0.049	0.048	0.046	0.040	0.038
500,000	0.056	0.048	0.052	0.046	0.048	0.045	0.047	0.045

Note: The nominal size of the tests is 0.05. $\widehat{\beta}_{1,1}$ denotes the regression-based encompassing test using (9), $\widehat{\varrho}_1$ denotes the LM-type encompassing test based on (10). Results are based on 10,000 replications. Tests statistics are based on OLS standard errors without degrees-of-freedom correction.

$\beta_{0,h} = 0$ and $\beta_{1,h} = 1$ in the regression model (9), where the alternative is equivalent to the null hypothesis of the Mincer-Zarnovitz test for an informative and rational (efficient) forecast. Under the null hypothesis the log-likelihood function is a function of $s_0^2 = \sum_{t=1}^{n}(Y_{t+h} - \overline{Y}_n)^2$, whereas under the alternative the log-likelihood depends on $s_1^2 = \sum_{t=1}^{n}(Y_{t+h} - \widehat{Y}_{t+h|h})^2$. Thus, the logarithm of the likelihood ratio is a function of

$$s_0^2 - s_1^2 = \sum_{t=1}^{n} \delta_t$$

used in the numerator of the DM-type statistics $2d_h$ and $\widetilde{d}_h$.

Remark 5: In contrast to the DM-type test, the encompassing test turns out to be slightly conservative for most combinations of n and T presented in Table 2. For empirically relevant sample sizes, i.e. for $T \ll 10,000$, the tests are always conservative. While the empirical sizes for a given value of n vary depending on T, these variations are smaller than those of the DM-type tests.

Remark 6: As mentioned above, a consistent selection rule for the maximum forecast horizon h^* requires that the size of the test tends to zero as $n \to \infty$. One possibility is to apply the critical value $\kappa \log(n)$ with some $\kappa > 0$. It is not difficult to see that under the alternative $\rho_h = O_p(n^{1/2})$ such that for $h \leq h^*$ we obtain $\lim_{n\to\infty} P(\rho_h < -\kappa \log(n)) = 1$, whereas for $h > h^*$ we have $\lim_{n\to\infty} P(\rho_h < -\kappa \log(n)) = 0$. Thus the decision rule that selects the last rejection in the sequence of tests with $h = 1, 2, \ldots$ is weakly consistent. Note that for $n = 27$ the critical value $-\log(27)/2 = -1.65$ is similar to the one-sided 0.05 critical value of a standard normal distribution. This suggests to set $\kappa = 1/2$ in order

to generate selection rules roughly equivalent to usual hypothesis testing.

6 Local power

In order to gain some insight into the relative power of the two different types of tests, i.e. the power of the DM-type tests versus the power of the encompassing tests, we analyse local power against a suitable sequence of local alternatives. Consider the alternative of an informative forecast with

$$Y_{t+1} = \mu + \beta X_t + u_{t+1}$$

where X_t is an i.i.d. regressor with $\mathbb{E}(X_t) = 0$ and $\mathbb{E}(X_t^2) = \sigma_x^2 > 0$, u_t is white noise with $\mathbb{E}(u_t) = 0$, $\mathbb{E}(u_t^2) = \sigma_u^2$ and u_{t+1} is independent of $\{X_t, X_{t-1}, \ldots\}$. As $T \to \infty$ we have $\widehat{\theta}_t \xrightarrow{p} \theta$, $\theta = (\mu, \beta)'$, and $\mathbb{E}(Y_{t+1} - Y_{t+1|t}^{\theta})^2 = \sigma_u^2$, where $Y_{t+1|t}^{\theta} = \mu + \beta X_t$ and $\mathbb{E}(Y_{t+1} - \mu)^2 = \sigma_u^2 + \beta^2 \sigma_x^2$. If $\beta \neq 0$ the forecast is informative and $\widetilde{d}_1$ and $\widehat{\varrho}_1$ are $O_p(\sqrt{n})$. Accordingly, both tests are consistent against fixed alternatives $\beta \neq 0$. The asymptotic power of the tests can be studied by considering a local alternative of the form $\beta = c/\sqrt{n}$. The asymptotic distributions of the DM-type test $\widetilde{d}_1$ and the encompassing test $\widehat{\varrho}_1$ are presented in

Theorem 4 *Under the sequence of alternatives $\beta = c/\sqrt{n}$, $X_t \sim iid(0, \sigma_x^2)$, Assumptions 1 – 2 and $n/\sqrt{T} \to 0$ it follows that*

$$\widetilde{d}_1 \xrightarrow{d} z_1^2 - 2\lambda z_2 - \lambda^2 \tag{11}$$

$$\widehat{\varrho}_1 \xrightarrow{d} sign(c) z_2 + \lambda, \tag{12}$$

where $\lambda^2 = c^2 \sigma_x^2 / \sigma_u^2$ denotes the signal-to-noise ratio and z_1 and z_2 represent two independent standard normally distributed random variables.

Accordingly, the DM-type test and the encompassing test are not asymptotically equivalent. Figure 1 compares the resulting local power curves using the significance level 0.05. The DM-type test is more powerful in the vicinity of the null hypothesis, whereas the relative power of the encompassing test increases when c or the variance ratio σ_x^2/σ_u^2 gets large. The ratio λ^2 can also be represented as

$$\lambda^2 = n \frac{R(1)^2}{1 - R(1)^2},$$

using Nelsons predictability measure presented in (7). If, for instance, $n\, R(1)^2 \geq 10$, then $\lambda^2 > 10$, and both tests have a local power of at least 93% for a significance level of 0.05. It should also be noted that the power curves are symmetric with respect to the parameter c as the distribution remains the same if z_2 is replaced by $-z_2$ in (11) and (12).

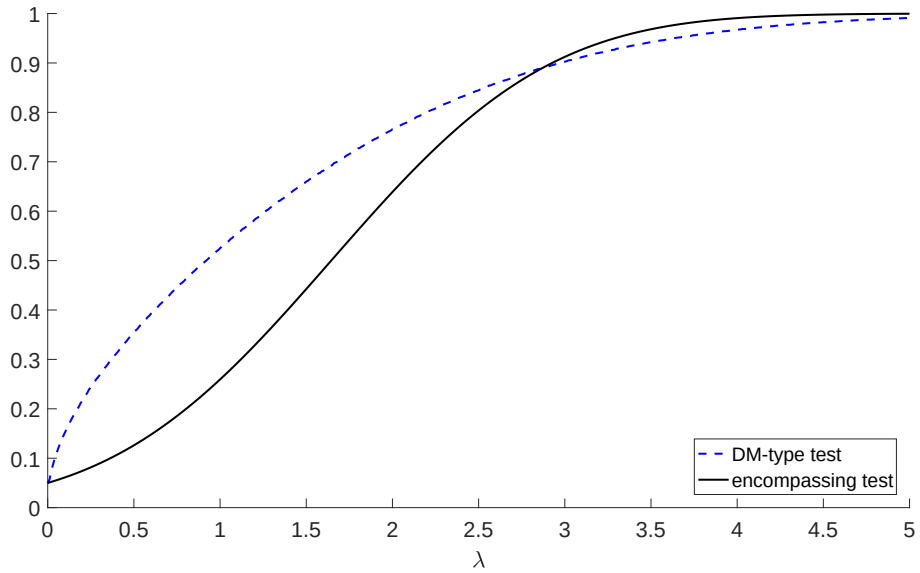

Figure 1: Local power curves

7 Monte Carlo experiments

In order to gain insight into the small sample behaviour of the tests, we conduct Monte Carlo experiments based on the cases displayed in Table 3. In the first four cases, univariate models are considered, whereas the last case refers to a simple multivariate model. The first three forecast models refer to moving-average models considered in Stock and Watson (2007) for the first difference of quarterly US inflation. The first process is based on their MA(1)-model estimated for the post-1984 period, whereas the second process refers to their pre-1984 estimation results. The third process is based on the quarterly version of Nelson and Schwert's (1997) model reported in Stock and Watson (2007). For all three cases, the forecast models are misspecified, because the data-generating processes (DGP) are MA(1) and MA(2) processes, but the forecast models assume an AR(1) process. Moreover, a constant is estimated. Note that the respective null hypotheses $h^* = 1$ and $h^* = 2$ are nevertheless correct. The fourth process uses the estimation result for an AR(1)-process of US GDP growth from 1996q3 to 2016q1 which corresponds to the sample used in the empirical application below. In this case, h^* does not exist. The last process mimics a forecasting equation for financial returns and implies an $R(1)^2$ of about 0.04 for $h = 1$. This would be considered a "large" value in forecasting stock price returns given the usual empirical results as reported, for example, in Fama and French (1988). In this case, the maximum forecast horizon is $h^* = 1$, since $R(h)^2 = 0$ for $h > 1$.

Table 3: Cases considered for Monte Carlo simulations

cases	DGP	h^*	forecast model	$R^2(h^*)$
$MA(1)_a$-$AR(1)$	$Y_t = \varepsilon_t - 0.28\varepsilon_{t-1}$	1	$\widehat{Y}_{t+h} = \widehat{\theta}_1^h + \widehat{\theta}_2^h Y_t$	0.07
$MA(1)_b$-$AR(1)$	$Y_t = \varepsilon_t - 0.66\varepsilon_{t-1}$	1	$\widehat{Y}_{t+h} = \widehat{\theta}_1^h + \widehat{\theta}_2^h Y_t$	0.21
$MA(2)$-$AR(1)$	$Y_t = \varepsilon_t - 0.49\varepsilon_{t-1} - 0.16\varepsilon_{t-2}$	2	$\widehat{Y}_{t+h} = \widehat{\theta}_1^h + \widehat{\theta}_2^h Y_t$	0.02
$AR(1)$-$AR(1)$	$Y_t = 0.33 + 0.42Y_{t-1} + \varepsilon_t$	–	$\widehat{Y}_{t+h} = \widehat{\theta}_1^h + \widehat{\theta}_2^h Y_t$	–
multivar.	$Y_t = 0.2X_{t-1} + \varepsilon_t$	1	$\widehat{Y}_{t+h} = \widehat{\theta}_1^h + \widehat{\theta}_2^h X_t$	0.04

Note: ε_t and X_t are $iid\ N(0,1)$. h^* is the maximum forecast horizon. $R^2(h^*)$ is the asymptotic R^2 of the forecast model at horizon h^*.

Forecasts are made in a direct manner, i.e. for each forecast horizon the target variable Y_{t+h} is regressed on the explanatory variables known at time t. We calculate the standard errors according to Newey and West (1987) using the automatic lag length selection procedure proposed by Andrews (1991) and a significance level of 0.05.

Table 4 displays the results for AR(1) forecasts, when the data is generated by the $\text{MA}_a(1)$ model, hence $h^* = 1$. The evaluation sample includes $n = 50$ or $n = 100$ forecasts, the initial estimation samples are based on $T = 100$ observations, and a recursive estimation scheme is employed. The tests are conducted sequentially for the forecast horizons $h = 1, 2, 3, 4$. The last forecast horizon where the test rejects is identified as horizon $\hat{h}^*$. If the test does not reject for any horizon, $\hat{h}^* \geq 4$. In addition to the tests presented above, the classical DM test is considered by using a standard normal distribution for the test statistic specified in equation (8). Given the values of $R^2(h^*)$ and n considered, the local power results of Section 6, and the fact that T is not very large, one can expect the tests to encounter certain difficulties in correctly detecting h^*, except for the case $MA(1)_b$-$AR(1)$.

With $n = 50$, the DM-type and encompassing tests have a power of at least 0.44 at $h = 1$, whereas the classical DM test attains 0.12 only. With $n = 100$, the power of the DM-type and encompassing tests reaches about 0.8. The classical DM test achieves 0.20 and its size is close to zero. Both DM-type tests are over-sized, whereas both encompassing tests are conservative. In terms of power, the two DM-type tests are similar and outperform the encompassing tests. This is likely to be partly due to their too large size, but also partly to their higher local power in the vicinity of the null hypothesis. Among the encompassing tests, the regression-based test using (9) clearly is more powerful than the LM-type test

based on (10).

For $n = 50$ the DM-type tests and the regression-based encompassing test identify the maximum forecast horizon correctly in about 60% of the replications, and this number rises to roughly 80% with $n = 100$. The LM-type encompassing test is slightly less successful. $\hat{h}^*$ based on the classical DM test has a strong downward bias due to its lack of power. With a rolling estimation window instead of a recursive estimation scheme, the results turn out to be very similar (see Table 9 in Appendix C).

For the remaining Monte Carlo experiments, we do not report results for the DM-type test based on $2d_h$, because it performs almost identically to $\widetilde{d}_h$. Moreover, we focus on the regression-based encompassing test, because it has good size properties and more power than the LM-type encompassing test. Results for all cases except MA$_a$(1)-AR(1) are reported in Table 5. We do not consider the classical DM test, because of its problems in detecting h^* documented above.

Concerning the case MA$_b$(1)-AR(1), the absolute value of the MA-coefficient is much larger than in the case MA$_a$(1)-AR(1), making it considerably easier for the tests to detect $h^* = 1$. With $n = 50$ as well as with $n = 100$, both tests reject in almost all replications at $h = 1$. The DM-type test rejects far too often for $h \geq 2$, whereas the encompassing test has almost the correct size.[6] Mainly due to the size distortion, the DM-type test detects the correct h^* in 80% to 85% of the replications only, whereas the encompassing test attains about 95%. Moreover, $\hat{h}^*$ has an upward bias which is more pronounced for the DM-type test because of its size distortion.

In the case MA(2)-AR(1), the second MA-coefficient is very close to zero, making it difficult to identify $h^* = 2$. Note that the ratio of the mean-squared prediction error (MSPE) to the evaluation-sample variance is virtually equal to 1.[7] While for $h = 1$, both tests reject in about 85% of the replications with $n = 50$ and in 97% of the replications with $n = 100$, these numbers are considerably lower for $h = 2$. The DM-type test yields a rejection probability of 40% to 45%, whereas the encompassing tests attains about 20%. The higher numbers of the former test are again at least partly related to its size distortion. The maximum forecast horizon h^* is detected correctly in about 10% to 20% of the replications by the encompassing test and in about 25% to 35% by the DM-type test. The most frequent value of $\hat{h}^*$ equals 1 in all cases considered. Thus, with the MA(2)-specification chosen here, larger evaluation samples are needed in order to reliably determine h^*.

The rejection probabilities for $h = 1$ and $h = 2$ in the AR(1)-AR(1) case are not too different from the MA(2)-AR(1) case. Accordingly, the most frequent value of $\hat{h}^*$ equals

[6] In simulations not reported here, it turns out that the size distortions of the DM-type test are far less pronounced if the MA-coefficient is positive. With a value of 0.66 instead of -0.66, the size equals about 0.10.

[7] The variance of the evaluation sample is calculated dividing by n, so that it equals the MSPE of the evaluation-sample mean.

Table 4: Results for case '$MA(1)_a$-$AR(1)$'

forecast horizon h	0	1	2	3	4	0	1	2	3	4
		$T = 100, n = 50$					$T = 100, n = 100$			
MSPE / Variance		0.96	1.03	1.03	1.03		0.95	1.02	1.02	1.02
rejections										
DM-type tests										
$\widetilde{d}_h$		0.71	0.13	0.12	0.12		0.83	0.09	0.09	0.09
$2d_h$		0.72	0.13	0.13	0.13		0.84	0.10	0.10	0.10
encompassing tests										
$\widehat{\beta}_{1,h}$		0.61	0.04	0.04	0.03		0.81	0.03	0.02	0.02
$\widehat{\varrho}_h$		0.44	0.02	0.02	0.02		0.75	0.02	0.02	0.02
classical DM test		0.12	0.01	0.01	0.01		0.20	0.00	0.00	0.00
$\hat{h}^*$										
DM-type tests										
$\widetilde{d}_h$	0.29	**0.62**	0.07	0.02	0.01	0.17	**0.76**	0.06	0.01	0.00
$2d_h$	0.28	**0.63**	0.07	0.02	0.01	0.16	**0.76**	0.06	0.02	0.00
encompassing tests										
$\widehat{\beta}_{1,h}$	0.39	**0.59**	0.02	0.00	0.00	0.19	**0.79**	0.02	0.00	0.00
$\widehat{\varrho}_h$	0.56	**0.44**	0.01	0.00	0.00	0.25	**0.74**	0.01	0.00	0.00
classical DM test	0.88	**0.11**	0.00	0.00	0.00	0.80	**0.20**	0.00	0.00	0.00

Note: The values displayed in the category 'rejections' denote the percentage of rejections for each horizon h. The values displayed in the category '$\hat{h}^*$' denote the percentage of cases in which h is identified as the maximum forecast horizon. The estimation is carried out recursively, and T denotes the number of observations used for the first parameter estimation. n is the number of observations for evaluation. The significance level is set to 0.05. 'classical DM test' refers to the test statistic proposed by Diebold and Mariano (1995). 'MSPE' is the mean-squared prediction error. The in-sample variance in the MSPE-variance-ratio is calculated dividing by n. Bold entries refer to the true h^*. If a test rejects for all horizons, $\hat{h}^*$ is set equal to the largest horizon $h = 4$. $2d_h$ denotes the test statistic distributed as $|\mathcal{N}(0,1)|$, $\widetilde{d}_h$ the test statistic distributed as χ^2_1 under the null. $\beta_{1,h}$ denotes the regression-based test using (9), $\widehat{\varrho}_h$ denotes the LM-type test based on (10). Results are based on 10,000 simulations.

Table 5: Results of the remaining cases

forecast horizon h	0	1	2	3	4	0	1	2	3	4
		$T = 100, n = 50$					$T = 100, n = 100$			
					$MA(1)_b$-$AR(1)$					
MSPE / Variance		0.80	1.02	1.02	1.02		0.80	1.02	1.02	1.01
rejections										
DM-type test		0.97	0.19	0.18	0.19		1.00	0.14	0.15	0.14
encompassing test		0.99	0.05	0.05	0.04		1.00	0.04	0.03	0.03
$\hat{h}^*$										
DM-type test	0.03	**0.80**	0.11	0.05	0.02	0.00	**0.85**	0.09	0.04	0.01
encompassing test	0.01	**0.94**	0.04	0.01	0.00	0.00	**0.96**	0.03	0.01	0.00
					$MA(2)$-$AR(1)$					
MSPE / Variance		0.91	1.00	1.02	1.02		0.91	1.00	1.02	1.01
rejections										
DM-type test		0.88	0.40	0.17	0.16		0.97	0.45	0.13	0.13
encompassing test		0.84	0.19	0.04	0.04		0.97	0.23	0.03	0.03
$\hat{h}^*$										
DM-type test	0.12	0.57	**0.25**	0.05	0.01	0.03	0.55	**0.36**	0.05	0.01
encompassing test	0.16	0.71	**0.11**	0.01	0.00	0.03	0.76	**0.21**	0.01	0.00
					$AR(1)$-$AR(1)$					
MSPE / Variance		0.88	1.03	1.06	1.07		0.86	1.01	1.03	1.04
rejections										
DM-type test		0.84	0.28	0.10	0.08		0.96	0.40	0.12	0.07
encompassing test		0.88	0.20	0.06	0.05		0.99	0.33	0.06	0.04
$\hat{h}^*$										
DM-type test	0.16	0.57	0.20	0.05	**0.02**	0.04	0.56	0.30	0.07	**0.03**
encompassing test	0.12	0.68	0.16	0.02	**0.01**	0.01	0.66	0.28	0.04	**0.01**
					multivar.					
MSPE / Variance		1.00	1.04	1.04	1.04		0.99	1.02	1.02	1.02
rejections										
DM-type test		0.50	0.08	0.08	0.09		0.62	0.06	0.06	0.07
encompassing test		0.33	0.03	0.03	0.03		0.51	0.02	0.02	0.02
$\hat{h}^*$										
DM-type test	0.50	**0.45**	0.04	0.00	0.00	0.38	**0.58**	0.04	0.00	0.00
encompassing test	0.67	**0.33**	0.01	0.00	0.00	0.49	**0.50**	0.01	0.00	0.00

Note: The DM-type test uses $\widetilde{d}_h$, the encompassing test employs $\beta_{1,h}$. For further information, see Tables 3 and 4.

1 for both tests and both values of n considered. The correct identification of $h^* \geq 4$ only happens in 1% to 3% of the replications.

Finally, in the multivariate case, the DM-type test rejects at $h = 1$ in 50% to about 60% of the replications, whereas the encompassing test does so in about 35% to 50%. However, the latter test is conservative, whereas the former rejects a little too often under the null, i.e. for $h > 1$. With $n = 50$, both tests tend to reject predictability, i.e. they yield $\hat{h}^* = 0$ in at least 50% of the replications. With $n = 100$, $\hat{h}^* = 1$ in about 60% of the replications with the DM-type test, whereas the encompassing test reaches 50%.

As the multivariate case is conformable with the local power analysis of Section 6, it is especially interesting to compare the results with the theoretical findings presented in Theorem 4. For the corresponding value $c = 0.2\sqrt{n}$ and $\lambda = c \cdot 1$, one gets $\lambda = \sqrt{2}$ for $n = 50$ and $\lambda = 2$ for $n = 100$. The corresponding power for the DM-type test is 0.64 with $n = 50$ and 0.77 with $n = 100$, and for the encompassing test 0.41 with $n = 50$ and 0.64 with $n = 100$. These values are moderately larger than the respective rejection probabilities for the multivariate case at $h = 1$ reported in Table 5. These differences are driven by the parameter estimation error. If one sets, for example, $T = 1,000$ instead of $T = 100$, the rejection probabilities obtained via simulations become very similar to those following from Theorem 4.

8 Empirical results

For the empirical application of the tests, we employ quarterly survey forecasts collected by Consensus Economics. The mean of the forecasts across all panelists is known to be a very accurate forecast, as documented, for example, by Ang, Bekaert, and Wei (2007) for inflation forecasts. We consider survey forecasts as being generated by some empirical model. One may argue, however, that survey forecasts do not involve any parameters to be estimated. This would be a comfortable situation for our analysis as in this case the $O_p(n/T)$ terms due to estimated parameters drop out and the asymptotic results should be more reliable.

Another view on survey forecasts is the notion that survey forecasters derive their predictions of Y_{t+h} as some function of the past, say $f(Y_t, Y_{t-1}, \ldots)$ for a univariate forecast. Such a function is similar to a model-based forecast where we first specify some (linear) function on past values and plug in some estimated parameter values that are again a function of past values. For a univariate zero-mean AR(1) forecast this approach result in the forecast function

$$\widehat{Y}_{t+h,t} = \frac{\sum_{s=t-T+h}^{t} Y_s Y_{s-h}}{\sum_{s=t-T+h}^{t} Y_{s-h}}$$
$$Y_t = f(Y_t, Y_{t-1}, \ldots, Y_{t-T}).$$

Note that this forecast function is a (nonlinear) filter of the past values that does not depend on parameters anymore. In our theoretical framework we assume that the (infeasible) forecast is based on the optimal filter represented by $Y^{\theta}_{t+h|t}$ (typically a conditional mean function), whereas the feasible forecast is given by $\widehat{Y}_{t+h|t} = Y^{\widehat{\theta}}_{t+h|t} = f(Y_t, Y_{t-1}, \ldots, Y_{t-T})$. No matter how the filter $f(Y_t, Y_{t-1}, \ldots, Y_{t-T})$ is derived (maybe some "guess" based on past observations, maybe using some statistical plug-in estimators for the parameters) the relevant issue is whether the difference $\widehat{Y}_{t+h|t} - Y^{\theta}_{t+h|t}$ is sufficiently small such that it does not affect the test decision. In other words, we only assume that survey forecasts are just another way to estimate the conditional mean function (i.e. by combining expert knowledge and human intelligence). For our analysis it is sufficient to assume that the order of magnitude of $f(Y_t, Y_{t-1}, \ldots, Y_{t-T}) - Y^{\theta}_{t+h|t}$ from the survey forecasts is similar to $Y^{\widehat{\theta}}_{t+h|t} - Y^{\theta}_{t+h|t}$ resulting from the parametric model framework. Based on the existing empirical literature, we have no reason to suppose that survey forecasts perform systematically worse than forecasts derived from some statistical model.

In our empirical analysis we consider forecasts of real GDP growth and real private consumption growth, because these are the only quarter-on-quarter (q-o-q) growth rates in the survey. For the indices of consumer prices (CPI), only forecasts for quarterly year-on-year (y-o-y) rates are available. Given the importance of inflation forecasts, we also include these forecasts in our analysis. However, given the y-o-y definition, and denoting the forecast horizon for the current quarter, i.e. for the nowcast by $h = 0$, we can expect to find $h^* \geq 2$. This is because knowledge about past values of the price index enables the forecasters to mechanically produce forecasts which have lower mean-squared prediction errors than the unconditional mean up to $h = 2$.[8] In addition to these variables, we also investigate the forecasts of the end-quarter values of the 3-month interest rate. Since interest rates show signs of non-stationarity in the sample under study, we use the first differences of this variable. The countries under study are the United States, the euro area (labeled 'Eurozone' by Consensus Economics), Japan, Germany, the United Kingdom, Italy, Canada, and France.

Since, in each quarter, Consensus Economics also provides data for recent quarters, we can employ this real-time data for the evaluation of the forecasts. We use the second vintage of all variables mentioned.

Considering forecasts for up to $h = 6$ quarters ahead, the balanced sample of forecasts and realizations starts in the third quarter of 1996 and ends in the first quarter of 2016, yielding a sample size of $n = 79$. However, the sample sizes for individual variables can

[8]The year-on-year rate for $h = 2$ equals the sum of the quarter-on-quarter rates for $h = -1, 0, 1, 2$. Using the observed quarter-on-quarter rate for $h = -1$ and the unconditional mean as the forecast of the quarter-on-quarter rates for the latter three horizons yields an MSPE for the year-on-year rate forecast for $h = 2$ which is lower than the variance of the year-on-year rates by construction. If information on the current quarter is available, the maximum forecast horizon must be equal to or larger than 3.

be smaller, mainly due to changes in the survey. For example, in the beginning of the sample, the survey switched from asking for West German variables to variables for the reunified Germany, and we only consider the latter forecasts. The 3-month interest rate to be forecasted for Germany, Italy and France changed in the first quarter of 1999 from country-specific rates to the Euribor. Only the Euribor forecasts enter our analysis. While they can be expected to be similar across the three countries, differences might emerge, for instance, due to the smaller number of forecasters for Italy. For Japan, the target variable of the interest rate forecasts changed from the 3-month Yen certificate of deposit to the TIBOR in the second quarter of 2010, and we only use the former forecasts.

In some countries, several large changes of the value-added tax rate (VAT) occurred. Since these changes are commonly announced well in advance, their occurrence can have a major impact on the predictability of inflation. In addition, the growth rate of real private consumption tends to be strongly negatively correlated with the VAT rate change contemporaneously, and in addition, there is a strong positive correlation in the quarter prior to the VAT rate change. We use the following rule in order to limit the impact of these changes on our analysis: For countries with at least two VAT rate changes of at least 2 percentage points, we shorten the sample such that the effects of these changes are excluded. This rule leads to sample modifications for Japan and the UK. We are going to report results obtained without this modification in the text, but not in the following figures and tables. More details on the variables and samples entering our analysis are given in Appendix B.

The quarterly forecasts are usually gathered in the first half of the last month of a quarter. Therefore, the forecasters can be expected to have information about the variable of interest in the current quarter, i.e. for the forecast (resp. nowcast) horizon $h = 0$.[9] Concerning inflation, at least the inflation rate for the first month of the current quarter should be known when the forecast is made. This implies that one can expect $\hat{h}^* \geq 3$ for the y-o-y inflation rates.

As an example forecast consider the inflation forecasts for the United States provided in 2016 as presented in Figure 2. What is striking about the forecasts for longer horizons is that they tend to settle at a value of about 2.3 which is almost identical to the mean of inflation in the evaluation period, being equal to 2.27 percent.

The empirical maximum forecast horizons $\hat{h}^*$ determined by the tests are shown in Table 6. The sequential p-values of the tests giving rise to these values of $\hat{h}^*$ are displayed in Figures 3 to 6. Notably, $\hat{h}^*$ is virtually always smaller than the largest forecast horizon of $h = 6$. The encompassing test implies larger values of $\hat{h}^*$ than the DM-type test in several cases. This may be due to potential biases of the forecasts. The larger the bias is at $\hat{h}^* + 1$, where $\hat{h}^*$ is determined by the DM-type test, the more likely it is that the

[9] However, for the q-o-q growth rate of consumption in Japan and in the UK, the nowcast will turn out to be uninformative and hence $\hat{h}^* = -1$.

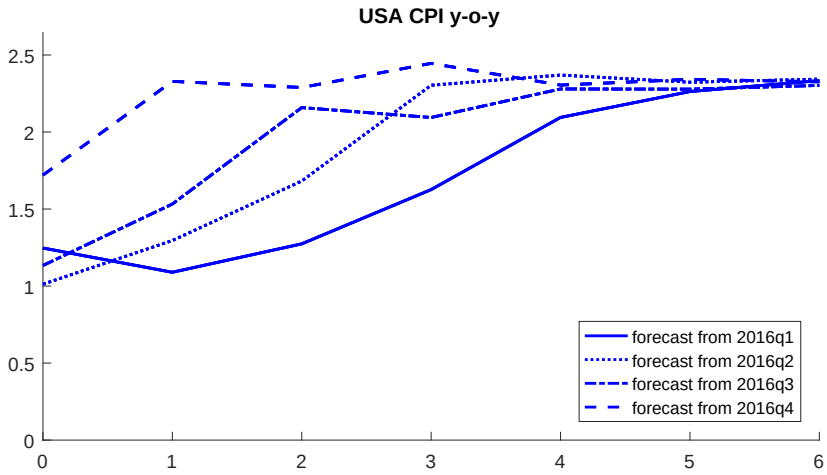

Figure 2: Forecasts for year-on-year US CPI inflation rates. The number on the x-axis denotes the forecast horizon in quarters with 0 being the nowcast.

encompassing test still rejects at this horizon.

For the q-o-q growth rates of real GDP growth, $\hat{h}^*$ tends to range from 1 to 3 quarters. Only for Italy and France, the encompassing test leads to larger values. The median of $\hat{h}^*$ across countries is 1.5 quarters according to the DM-type test, and 2 quarters according to the encompassing test.

Concerning the y-o-y growth rates of the CPI, $\hat{h}^*$ is mostly equal to 3 quarters, which is also the median across countries according to both tests. Only for Canada, both tests indicate a higher value of $\hat{h}^* = 4$. For the Euro area and Japan, at least one test indicates a larger value. Given the considerations with respect to y-o-y rates and nowcasts made at the end of $h = 0$, these results indicate substantial difficulties in making informative inflation forecasts. Using the full sample for Japan and the UK leaves the results for the UK unchanged, but leads to $\hat{h}^* \geq 6$ for Japan according to both tests. Thus, as to be expected, strong changes in the VAT rate which are announced well in advance can render inflation forecasts informative even at larger horizons.

The results for the q-o-q growth rates of real private consumption growth vary strongly across tests and countries. The encompassing test often implies pronouncedly larger values of $\hat{h}^*$ than the DM-type test. Moreover, $\hat{h}^*$ according to the encompassing test is mostly larger than in the case of real GDP growth. These results might be due to the facts that even small announced changes in the VAT rate are relatively important for private consumption, and that consumption forecasts often tend to biased. Indeed, using the full sample for Japan and the UK leads to $\hat{h}^* \geq 6$ for both countries according to both tests.

Table 6: Maximum forecast horizons in quarters determined by DM-type and encompassing tests

	US	EA	JP	DE	UK	IT	CA	FR	median
GDP q-o-q									
DM-type test	2	2	1	1	3	1	1	2	1.5
encompassing test	2	2	1	2	3	5	1	4	2
CPI y-o-y									
DM-type test	3	5	3	2	2	3	4	3	3
encompassing test	3	3	4	3	3	3	4	3	3
PrivCons q-o-q									
DM-type test	3	1	-1	0	-1	1	0	2	0.5
encompassing test	3	3	0	3	3	3	1	5	3
d(3m rate)									
DM-type test	1		0	3	2	2	1	2	2
encompassing test	2		0	2	6	3	1	2	2

Note: The DM-type test uses $\widetilde{d}_h$, the encompassing test employs $\beta_{1,h}$. 'GDP q-o-q' denotes quarter-on-quarter growth rates of real GDP, 'CPI y-o-y' year-on-year growth rates of consumer prices, 'PrivCons q-o-q' quarter-on-quarter growth rates of real private consumption, and 'd(3m rate)' quarter-on-quarter changes of the end-quarter 3-month interest rate. The abbreviations used for the countries are 'US' for the United States, 'EA' for the euro area, 'JP' for Japan, 'DE' for Germany, 'UK' for the United Kingdom, 'IT' for Italy, 'CA' for Canada, and 'FR' for France. For further information on the variables, see the text and Appendix B.

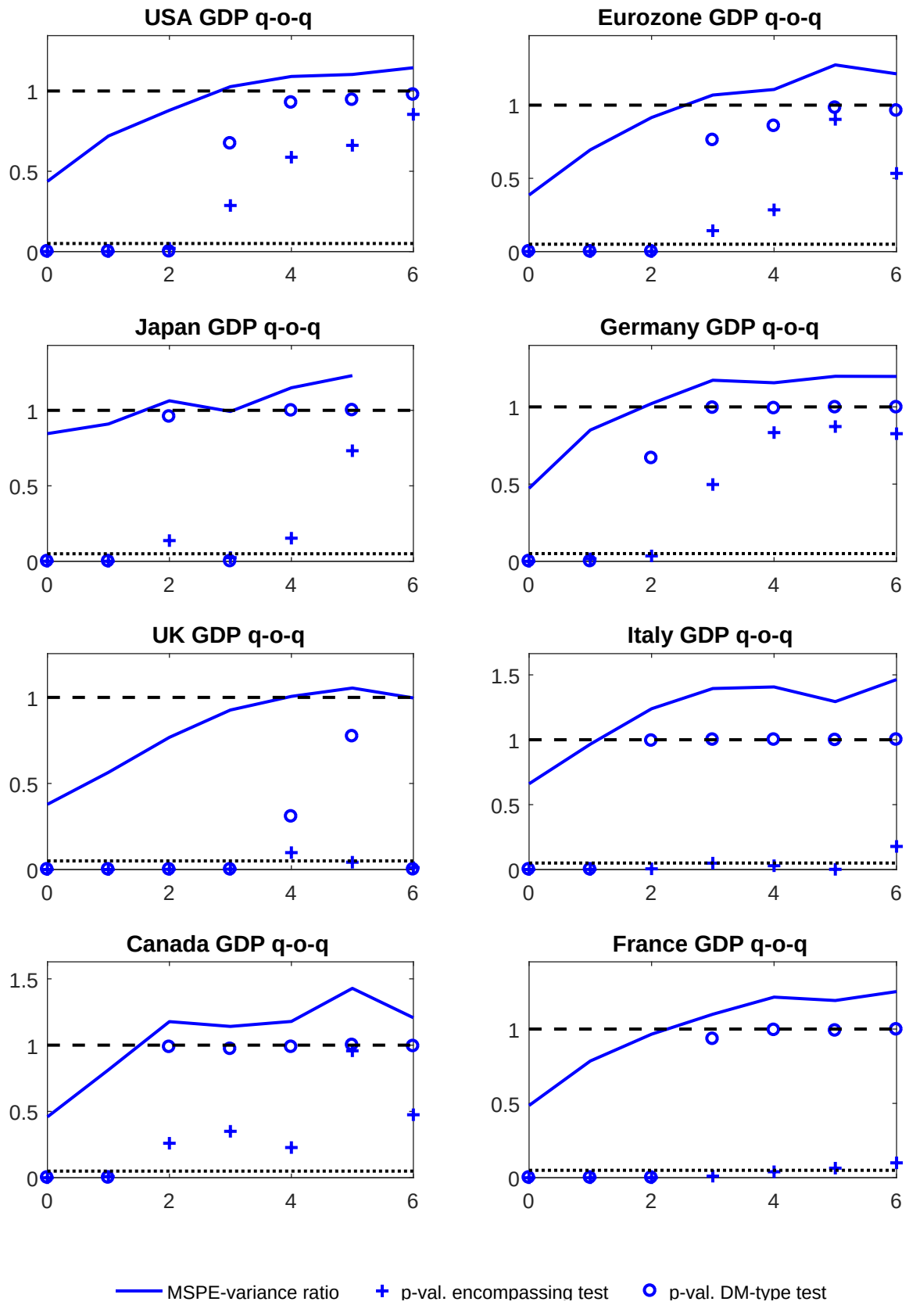

Figure 3: Test results for quarter-on-quarter growth rates of real GDP. The number on the x-axis denotes the forecast horizon in quarters with 0 being the nowcast. The dotted line is at 0.05, corresponding to the significance level of the tests. The dashed line is at 1. The solid line indicates the MSPE-variance ratio. The DM-type test uses $\widetilde{d}_h$, the encompassing test employs $\beta_{1,h}$. The maximum forecast horizon $\hat{h}^*$ identified by a test equals the horizon before the smallest horizon for which the p-value of the test exceeds 0.05.

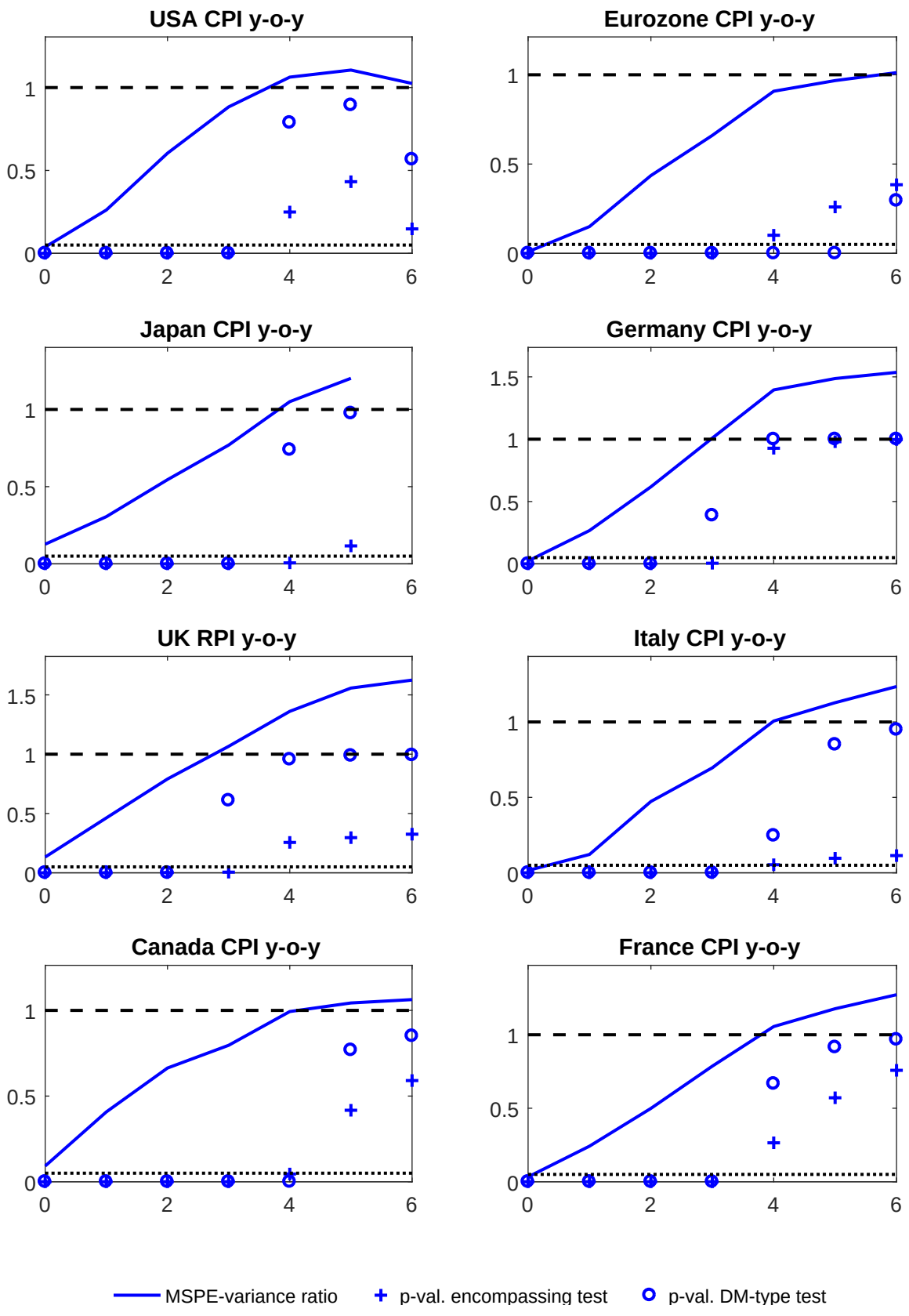

Figure 4: Test results for year-on-year growth rates of the CPI (the RPI in the case of the UK). For further explanations, see Figure 3.

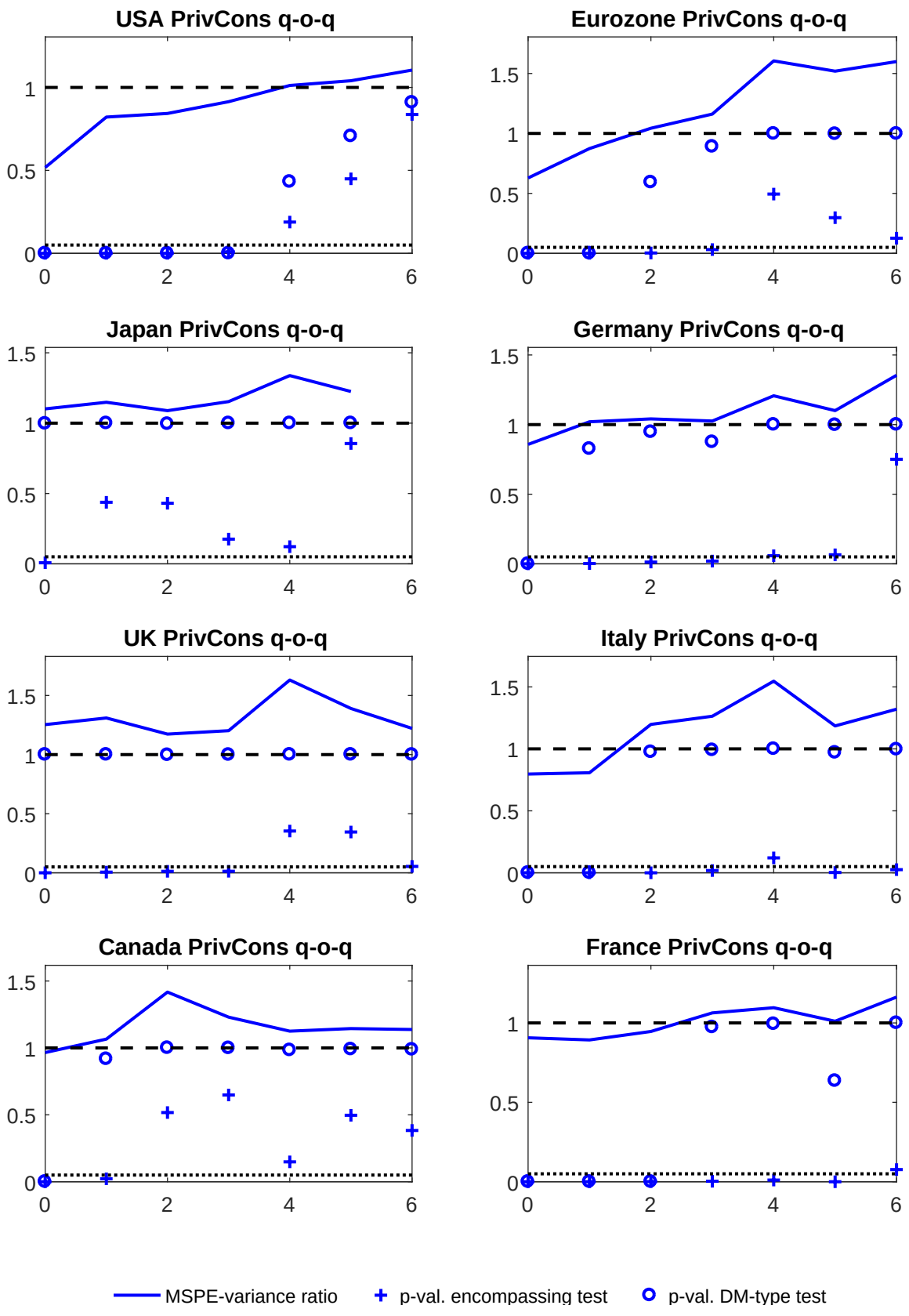

Figure 5: Test results for quarter-on-quarter growth rates of real private consumption. For further explanations, see Figure 3.

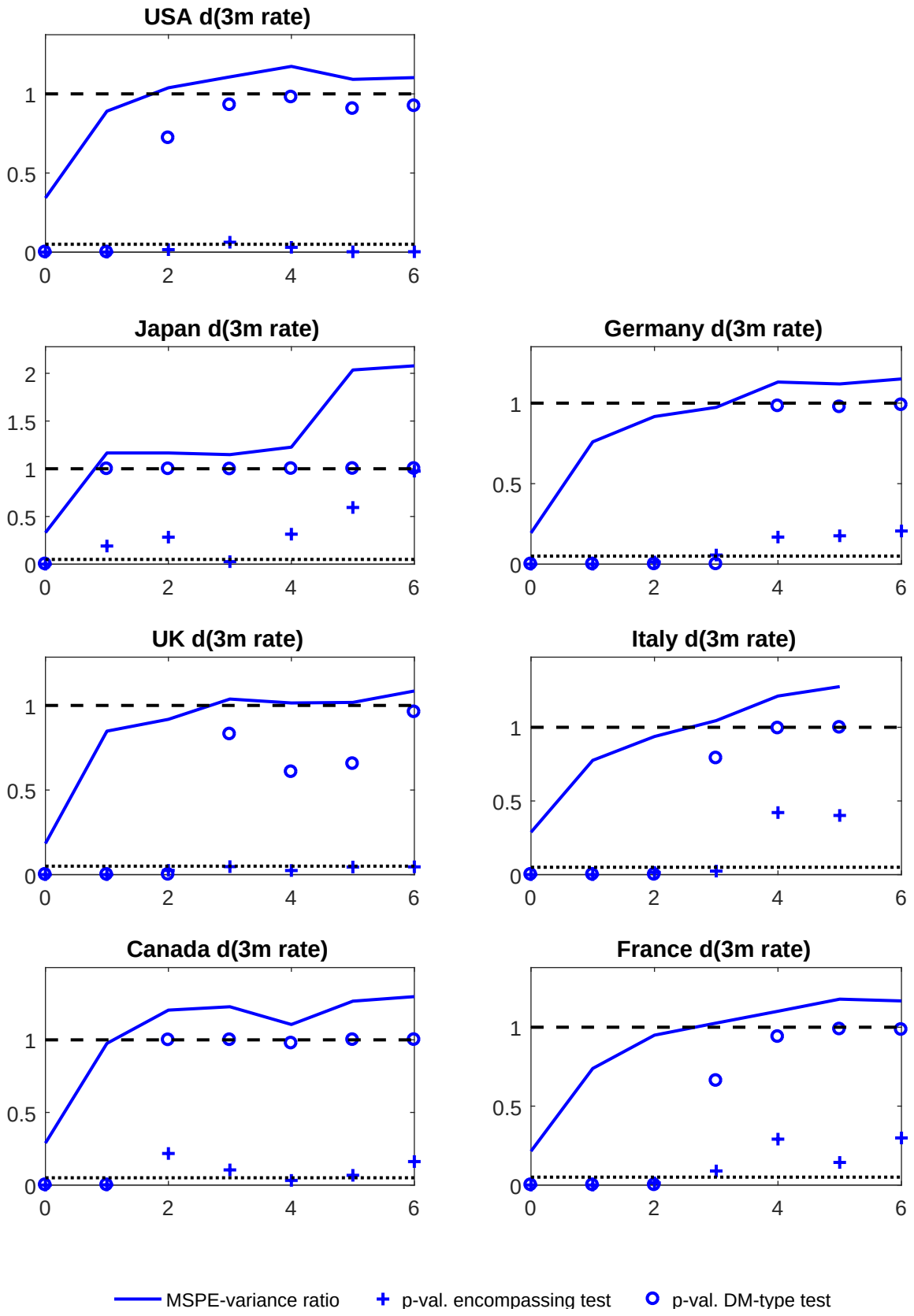

Figure 6: Test results for the change in the end-quarter 3-month interest rate. For further explanations, see Figure 3.

With the restricted sample, for the UK, the DM-type test indicates that not even the nowcast is informative.

The maximum forecast horizon for the change in the 3-month interest rate mostly varies between 1 quarter and 3 quarters. For the UK, the encompassing test finds informative forecasts at least up to $h = 6$. Except for the latter case, the two $\hat{h}^*$ found by the tests for a given country do not differ by more than 1 quarter. For Japan, only the nowcasts turn out to be informative.

9 Concluding remarks

This paper develops a forecast evaluation framework for testing the null hypothesis that the forecast at some pre-specified horizon h is uninformative. The tests are constructed such that they can be used even if the forecasts and the corresponding realizations are the only data available to the evaluator. The proposed tests can be applied sequentially to identify the maximum forecast horizon of the predictions. We show that due to the nested nature of the forecast comparison, the standard Diebold-Mariano (DM) type test statistic has a nonstandard limiting distribution and suffers from a severe loss of power. To overcome this problem, we adjust the test statistic and derive alternative tests from the encompassing principle that result in a coefficient test for a Mincer-Zarnowitz regression. Our analysis of the local power reveals that the DM-type test statistic is more powerful in the vicinity of the null hypothesis, whereas it performs similar to the encompassing test if the forecasts are more informative. In our Monte Carlo simulations we find that the DM-type test suffers from considerable size distortions in reasonable sample sizes, whereas the regression variant of the encompassing test exhibits reliable sizes.

In the empirical analysis, we apply our tests to macroeconomic forecasts from the survey of Consensus Economics. Our results suggest that forecasts of macroeconomic key variables are hardly informative beyond 2–4 quarters ahead. Our results confirm earlier (anecdotal) findings from macroeconomic forecasting. The main contribution of our work is to provide statistical tests that allow the forecaster to assess the maximum forecast horizon of the forecast of interest.

It is worth mentioning that our testing approach (as any other empirical methodology) has two major limitations. First, the estimated maximum forecast horizon may be biased downwards if the predictive power is weak but not negligible. A similar caveat applies if the number of forecasts in the evaluation sample is small. Second, the estimated maximum forecast horizon depends on the approach that generates the forecasts. If the approach fails to exploit important information it may produce uninformative forecasts, while a richer forecasting procedure may result in informative forecasts. Accordingly, any qualification of the informative content is conditional on the forecasting approach.

References

Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59(3), 817–858.

Ang, A., G. Bekaert, and M. Wei (2007): "Do macro variables, asset markets, or surveys forecast inflation better?," *Journal of Monetary Economics*, 54(4), 1163–1212.

Calhoun, G. (2016): "An asymptotically normal out-of-sample test based on mixed estimation windows," *Iowa State University, mimeo*.

Chong, Y. Y., and D. F. Hendry (1986): "Econometric Evaluation of Linear Macro-Economic Models," *Review of Economic Studies*, 53(4), 671–690.

Clark, T. E., and M. W. McCracken (2001): "Tests of equal forecast accuracy and encompassing for nested models," *Journal of Econometrics*, 105(1), 85–110.

Clements, M., and D. Hendry (1998): *Forecasting Economic Time Series*, no. 9780521634809 in Cambridge Books. Cambridge University Press.

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

Diebold, F. X., and L. Kilian (2001): "Measuring predictability: theory and macroeconomic applications," *Journal of Applied Econometrics*, 16(6), 657–669.

Diebold, F. X., and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13(3), 253–63.

Fama, E. F., and K. R. French (1988): "Dividend yields and expected stock returns," *Journal of Financial Economics*, 22(1), 3 – 25.

Galbraith, J. W., and G. Tkacz (2007): "Forecast content and content horizons for some important macroeconomic time series," *Canadian Journal of Economics*, 40(3), 935–953.

Granger, C. W. J., and P. Newbold (1986): *Forecasting Economic Time Series*, no. 9780122951831 in Elsevier Monographs. Elsevier.

Isiklar, G., and K. Lahiri (2007): "How far ahead can we forecast? Evidence from cross-country surveys," *International Journal of Forecasting*, 23(2), 167–187.

Knüppel, M. (2018): "Forecast-Error-Based Estimation of Forecast Uncertainty When the Horizon Is Increased," *International Journal of Forecasting*, 34(1), 105–116.

Mincer, J. A., and V. Zarnowitz (1969): "The Evaluation of Economic Forecasts," in *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, ed. by J. A. Mincer, chap. 1, pp. 1–46. NBER.

Nelson, C. R. (1976): "The Interpretation of R^2 in Autoregressive-Moving Average Time Series Models," *The American Statistician*, 30(4), 175–180.

Newey, W. K., and K. D. West (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55(3), 703–708.

Parzen, E. (1981): "Time Series Model Identification and Prediction Variance Horizon," in *Applied Time Series Analysis II*, ed. by D. F. Findley, pp. 415 – 447. Academic Press.

Stock, J. H., and M. W. Watson (2007): "Why Has U.S. Inflation Become Harder to Forecast?," *Journal of Money, Credit and Banking*, 39(s1), 3–33.

West, K. D. (1996): "Asymptotic Inference about Predictive Ability," *Econometrica*, 64(5), 1067–1084.

Appendix A: Proofs

Proof of Theorem 1:

Let $h > h^*$. Applying a mean-value expansion of the form $Y^{\widehat{\theta}_t}_{t+h|t} = Y^{\theta}_{t+h|t} + D_{t+h}(\bar{\theta}_t)(\widehat{\theta}_t - \theta)$, where $\bar{\theta}_t$ denotes a value between $\widehat{\theta}_t$ and θ yields

$$\begin{aligned}
\delta^h_t &= [u_{t+h} - D_{t+h}(\bar{\theta}_t)(\widehat{\theta}_t - \theta)]^2 - (u_{t+h} - \overline{u}_h)^2 \\
\delta^h_t &= u^2_{t+h} - (u_{t+h} - \overline{u}_h)^2 - 2u_{t+h}D_{t+h}(\bar{\theta}_t)(\widehat{\theta}_t - \theta) + D_{t+h}(\bar{\theta}_t)^2(\widehat{\theta}_t - \theta)^2 \\
&= \overline{u}_h(2u_{t+h} - \overline{u}_h) - 2u_{t+h}D_{t+h}(\bar{\theta}_t)(\widehat{\theta}_t - \theta) + D_{t+h}(\bar{\theta}_t)^2(\widehat{\theta}_t - \theta)^2
\end{aligned}$$

where $\overline{u}_h = n^{-1}\sum_{t=1+h}^{n+h} u_t$ and

$$\begin{aligned}
\widehat{\theta}_t - \theta &= (\widehat{\theta}_t - \widehat{\theta}_0) + (\widehat{\theta}_0 - \theta) \\
&= O_p\left(\sqrt{\frac{n}{T}} \cdot T^{-1/2}\right) + O_p(T^{-1/2}) \;=\; O_p(T^{-1/2})
\end{aligned}$$

due to Assumption 2 *(iii)*. Furthermore, *(ii)* and *(iv)* imply

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n}\delta^h_t &= \overline{u}^2_h - 2\frac{1}{n}\sum_{t=1}^{n} u_{t+h}D_{t+h}(\bar{\theta}_t)(\widehat{\theta}_t - \theta) + \frac{1}{n}\sum_{t=1}^{n} D_{t+h}(\bar{\theta}_t)^2(\widehat{\theta}_t - \theta)^2 \\
&= \overline{u}^2_h + O_p(T^{-1})
\end{aligned}$$

and

$$\begin{aligned}
\widehat{\gamma}_\delta(j) &= \frac{1}{n}\sum_{t=j+1}^{n}\delta^h_t\delta^h_{t-j} \\
&= \frac{1}{n}\overline{u}^2_h\sum_{t=j+1}^{n}(2u_{t+h} - \overline{u}_h)(2u_{t+h-j} - \overline{u}_h) \;+\; O_p(T^{-1}) \\
&= \frac{1}{n}\overline{u}^2_h\left[\left(\sum_{t=j+1}^{n} 4u_{t+h}u_{t+h-j}\right) - 3n\,\overline{u}^2_h\right] + O_p(T^{-1}) \\
&= \overline{u}^2_h\left(\frac{1}{n}\sum_{t=j+1}^{n} 4u_{t+h}u_{t+h-j}\right) + O_p(n^{-2}) + O_p(T^{-1}).
\end{aligned}$$

Hence

$$\begin{aligned} f\widehat{\omega}_\delta^2 &= \widehat{\gamma}_\delta(0) + 2\sum_{j=1}^{h-1}\widehat{\gamma}_\delta(j) \\ &= 4\overline{u}_h^2\left(\widehat{\gamma}_u(0) + 2\sum_{j=1}^{h-1}\widehat{\gamma}_u(j)\right) + O_p(T^{-1}) + O_p(n^{-2}) \\ &= 4\overline{u}_h^2\widehat{\omega}_u^2 + O_p(T^{-1}) + O_p(n^{-2}). \end{aligned}$$

Thus,

$$\begin{aligned} d_h &= \frac{1}{\widehat{\omega}_\delta\sqrt{n}}\sum_{t=1}^{h}\delta_t^h \\ &= \frac{(\sqrt{n}\,\overline{u}_h)^2 + O_p(n/T)}{\sqrt{4n\,\overline{u}_h^2\widehat{\omega}_u^2 + O_p(n/T) + O_p(n^{-1})}} \\ &= \frac{\sqrt{n}|\overline{u}_h|}{2\widehat{\omega}_u} + O_p\left(\frac{n}{T}\right) \xrightarrow{d} \frac{|z|}{2} \end{aligned}$$

where z is a standard normally distributed random variable.

Proof of Corollary 2:

The distribution of $2d_h$ follows directly from Theorem 1. As shown in Theorem 1 we have

$$\sum_{t=1}^{n}\delta_t^h = \left(\sqrt{n}\,\overline{u}\right)^2 + O_p(n/T)\ .$$

If $n/T \to 0$ and $\widehat{\omega}_u^2$ is a consistent estimator of the long-run variance of $u_{t+h} = Y_{t+h} - \mu$ then $\widetilde{d}_h$ possesses a χ^2 limiting distribution with one degree of freedom.

Proof of Theorem 3:

Consider some $h > h^*$. We first analyze

$$\sum_{t=1}^{n}(\widehat{Y}_{t+h|t} - \overline{\widehat{Y}}_h)(Y_{t+h} - \overline{Y}_h) = \sum_{t=1}^{n}\widehat{Y}_{t+h|t}(u_{t+h} - \overline{u}_h).$$

An important problem with analysing this expression is that the estimation error in $\widehat{Y}_{t+h|t} = Y_{t+h|t}^{\widehat{\theta}_t}$ is correlated with $\overline{u}_h$. To sidestep this difficulty we decompose the forecast into one component $Y_{t+h|t}^{\widehat{\theta}_0}$ that is independent of $\{u_{1+h}, \ldots, u_{n+h}\}$ and show that the remaining component is asymptotically negligible. Applying a mean value expansion

yields

$$\widehat{Y}_{t+h|t} = Y^{\widehat{\theta}_t}_{t+h|t} = Y^{\widehat{\theta}_0}_{t+h|t} + D_{t+h}(\bar{\theta}_t)(\widehat{\theta}_t - \widehat{\theta}_0)$$

where $D_{t+h}(\theta) = \partial Y^{\theta}_{t+h|t}/\partial\theta$ and $\bar{\theta}_t$ denotes some value between θ_0 and $\widehat{\theta}_t$. Note that by Assumption 2 $Y^{\widehat{\theta}_0}_{t+h|t}$ is uncorrelated with all $u_{1+h}, u_{2+h}, \ldots, u_{n+h}$. Accordingly,

$$\sum_{t=1}^{n}\left[Y^{\widehat{\theta}_0}_{t+h|t} + D_{t+h}(\bar{\theta}_t)(\widehat{\theta}_t - \widehat{\theta}_0)\right](u_{t+h} - \overline{u}_h) = A_{T,n} + B^1_{T,n} + B^2_{T,n}$$

where

$$\begin{aligned} A_{T,n} &= \sum_{t=1}^{n} Y^{\widehat{\theta}_0}_{t+h|t}(u_{t+h} - \overline{u}_h) \\ B^1_{T,n} &= \sum_{t=1}^{n} D_{t+h}(\bar{\theta}_t)(\widehat{\theta}_t - \widehat{\theta}_0)u_{t+h} \\ B^2_{T,n} &= \overline{u}_h \sum_{t=1}^{n} D_{t+h}(\bar{\theta}_t)(\widehat{\theta}_t - \widehat{\theta}_0). \end{aligned}$$

Another mean value expansion around the true value θ with $Y^{\theta}_{t+h|t} = \mu$ yields

$$\begin{aligned} A_{T,n} &= (\widehat{\theta}_0 - \theta)\sum_{t=1}^{n} D_{t+h}(\bar{\theta}_0)u_{t+h} - (\widehat{\theta}_0 - \theta)\overline{u}_h \sum_{t=1}^{n} D_{t+h}(\bar{\theta}_0) \\ &= A^1_{T,n} + A^2_{T,n} \end{aligned}$$

where $\bar{\theta}_0$ is some value between $\widehat{\theta}_0$ and θ. Since $\widehat{\theta}_0$ and $D_{t+h}(\bar{\theta}_0)$ are uncorrelated with u_{t+h} it follows that $A^1_{T,n} = O_p(T^{-1/2})O_p(n^{1/2})$, whereas $A^2_{T,n} = O_p(T^{-1/2})O_p(n^{-1/2})O_p(n)$. Thus, $A_{T,n}$ is $O_p(\sqrt{n/T})$. Under the null hypothesis $\widehat{\theta}_t - \widehat{\theta}_0$ and $D_{t+h}(\bar{\theta}_t)$ are uncorrelated with u_{t+h}. Furthermore Assumption 2 (iii) and (iv) imply

$$\sum_{t=1}^{n}(\widehat{\theta}_t - \widehat{\theta}_0)^2 D_{t+h}(\bar{\theta}_t)^2 u^2_{t+h} = O_p\left(\frac{n}{T^2}\right)\sum_{t=1}^{n} D_{t+h}(\bar{\theta}_t)^2 u^2_{t+h} \;=\; O_p\left(\frac{n^2}{T^2}\right).$$

and, therefore,

$$B^1_{T,n} = \sum_{t=1}^{n}(\widehat{\theta}_t - \widehat{\theta}_0)D_{t+h}(\bar{\theta}_t)u_{t+h} = O_p\left(\frac{n}{T}\right).$$

Since

$$\sum_{t=1}^{n}(\widehat{\theta}_t - \widehat{\theta}_0)D_{t+h}(\bar{\theta}_t) = O_p\left(\frac{\sqrt{n}}{T}\right)\sum_{t=1}^{n} D_{t+h}(\bar{\theta}_t)$$
$$= O_p\left(\frac{n^{3/2}}{T}\right)$$

it follows that $B_{T,n}^2 = O_p(n^{-1/2})O_p(n^{3/2}/T) = O_p(n/T)$. As $n/T \to 0$ It follows that

$$\sqrt{\frac{T}{n}}\sum_{t=1}^{n}(\widehat{Y}_{t+h|t} - \overline{Y}_h)(Y_{t+h} - \overline{Y}_h) = \sqrt{\frac{T}{n}}(A_{T,n} + B_{T,n}^1 + B_{T,n}^2)$$
$$= \sqrt{\frac{T}{n}}A_{T,n} + O_p\left(\sqrt{\frac{n}{T}}\right)$$
$$= \sqrt{T}(\widehat{\theta}_0 - \theta)\frac{1}{\sqrt{n}}\sum_{t=1}^{n} D_{t+h}(\bar{\theta}_0)(u_{t+h} - \overline{u}_h) + O_p\left(\sqrt{\frac{n}{T}}\right)$$

Next we analyze

$$\sum_{t=1}^{n}(\widehat{Y}_{t+h|t} - \overline{\widehat{Y}}_h)^2(Y_{t+h} - \overline{Y}_h)^2 \;=\; \sum_{t=1}^{n}(\widehat{Y}_{t+h|t} - \overline{\widehat{Y}}_h)^2(u_{t+h} - \overline{u}_h)^2.$$

Using the above mean value expansions we obtain

$$\widehat{Y}_{t+h|t} = Y_{t+h|t}^{\widehat{\theta}_0} + D_{t+h}(\bar{\theta}_t)(\widehat{\theta}_t - \widehat{\theta}_0)$$
$$= \mu + D_{t+h}(\bar{\theta}_0)(\widehat{\theta}_0 - \theta) + D_{t+h}(\bar{\theta}_t)(\widehat{\theta}_t - \widehat{\theta}_0)$$
$$\widehat{Y}_{t+h|t} - \overline{\widehat{Y}}_h = \widetilde{D}_{t+h}(\bar{\theta}_0)(\widehat{\theta}_0 - \theta) + \widetilde{\Psi}_{t+h}(\widehat{\theta}_t, \widehat{\theta}_0)$$

where

$$\widetilde{D}_{t+h}(\bar{\theta}_0) = D_{t+h}(\bar{\theta}_0) - n^{-1}\sum_{s=1}^{n} D_{s+h}(\bar{\theta}_0)$$

$$\widetilde{\Psi}_{t+h}(\widehat{\theta}_t, \widehat{\theta}_0) = D_{t+h}(\bar{\theta}_0)(\widehat{\theta}_t - \widehat{\theta}_0) - \frac{1}{n}\sum_{s=1}^{n} D_{s+h}(\bar{\theta}_0)(\widehat{\theta}_s - \widehat{\theta}_0).$$

It follows that

$$\begin{aligned}\sum_{t=1}^{n}(\widehat{Y}_{t+h|t}-\overline{\widehat{Y}}_h)^2(Y_{t+h}-\overline{Y}_h)^2 &= (\widehat{\theta}_0-\theta)^2\sum_{t=1}^{n}\widetilde{D}_{t+h}(\bar{\theta}_0)^2(u_{t+h}-\overline{u}_h)^2\\ &+\sum_{t=1}^{n}\widetilde{\Psi}_{t+h}(\widehat{\theta}_t,\widehat{\theta}_0)^2(u_{t+h}-\overline{u}_h)^2\\ &+2(\widehat{\theta}_0-\theta)\sum_{t=1}^{n}\widetilde{D}_{t+h}(\bar{\theta}_0)\widetilde{\Psi}_{t+h}(\widehat{\theta}_t,\widehat{\theta}_0)(u_{t+j}-\overline{u}_h)^2\\ &= C^0_{T,n}+C^1_{T,n}+C^2_{T,n}\ .\end{aligned}$$

For the leading term we obtain

$$C^0_{T,n}=O_p(T^{-1})O_p(n)\ =\ O_p(n/T)$$

For the second term we note that

$$\sum_{t=1}^{n}(\widehat{\theta}_t-\widehat{\theta}_0)^2D_{t+h}(\bar{\theta}_t)^2(u_{t+h}-\overline{u}_h)^2\ =\ O_p(n^2/T^2).$$

Since the mean adjustment does not affect the order of magnitude we conclude that

$$C^1_{T,n}=O_p(n^2/T^2)$$

For the last term we obtain

$$\sum_{t=1}^{n}(\widehat{\theta}_t-\widehat{\theta}_0)D_{t+h}(\bar{\theta}_0)D_{t+h}(\bar{\theta}_t)\ =\ O_p(n^{3/2}/T)$$

and, since the mean-adjustment does not affect the order of magnitude,

$$C^2_{T,n}=O_p(n^{3/2}/T^{3/2}).$$

Combining these results yields

$$\frac{T}{n}\sum_{t=1}^{n}(\widehat{Y}_{t+h|t}-\overline{\widehat{Y}}_h)^2(Y_{t+h}-\overline{Y}_h)^2=T(\widehat{\theta}_0-\theta)^2\frac{1}{n}\sum_{t=1}^{n}\widetilde{D}_{t+h}(\bar{\theta}_0)^2(u_t-\overline{u}_h)^2+O_p\left(\sqrt{\frac{n}{T}}\right).$$

In the same manner it can be shown that for $j=1,2,\ldots,h$

$$\begin{aligned}&\frac{T}{n}\sum_{t=1+j}^{n}(\widehat{Y}_{t+h|t}-\overline{\widehat{Y}}_h)(\widehat{Y}_{t+h-j|t}-\overline{\widehat{Y}}_h)(Y_{t+h}-\overline{Y}_h)(Y_{t+h-j}-\overline{Y}_h)\\ &=T(\widehat{\theta}_0-\theta)^2\frac{1}{n}\sum_{t=1}^{n}\widetilde{D}_{t+h}(\bar{\theta}_0)\widetilde{D}_{t+h-j}(\bar{\theta}_0)(u_{t+h}-\overline{u}_h)(u_{t+h-j}-\overline{u}_h)+O_p\left(\sqrt{\frac{n}{T}}\right).\end{aligned}$$

Define $\widehat{V}_{n,T} = \widehat{\Gamma}_{0,n,T} + 2\sum_{j=1}^{h-1} \widehat{\Gamma}_{j,n,T}$ where

$$\widehat{\Gamma}_{j,n,T} \;=\; \frac{1}{n}\sum_{t=1+j}^{n} (\widehat{Y}_{t+h|t} - \overline{\widehat{Y}}_h)(\widehat{Y}_{t+h-j|t} - \overline{\widehat{Y}}_h)(Y_{t+h} - \overline{Y}_h)(Y_{t+h-j} - \overline{Y}_h).$$

It follows that

$$\frac{T}{n}\widehat{V}_{n,T} = T(\widehat{\theta}_0 - \theta)^2 \mathbb{E}\left[\frac{1}{n}\left(\sum_{t=1}^{n} \widetilde{D}_{t+h}(\theta)(u_{t+h} - \overline{u}_h)\right)^2\right] + o_p(1).$$

Applying a suitable version of the central limit theorem it follows that

$$\frac{1}{\sqrt{nV_{n,T}}}\sum_{t=1}^{n}(\widehat{Y}_{t+h|t} - \overline{Y}_h)(u_{t+h} - \overline{u}_n) = \frac{\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \widetilde{D}_{t+h}(\theta)(u_{t+h} - \overline{u}_h)}{\sqrt{\mathbb{E}\left[\frac{1}{n}\left(\sum_{t=1}^{n} \widetilde{D}_{t+h}(\theta)(u_{t+h} - \overline{u}_h)\right)^2\right]}} + o_p(1)$$

$$\xrightarrow{d} \mathcal{N}(0,1)$$

Proof of Theorem 4:

Under the local alternative, we have for $h = 1$

$$Y_{t+1} - \overline{Y}_1 = u_{t+1} - \overline{u}_1 + (c/\sqrt{n})(X_t - \overline{X})$$

and the model prediction error is given by $\widehat{e}_{t+h|t} = u_{t+1} + O_p(T^{-1/2})$. Following the proof of Theorem 1 we obtain for $n/T \to 0$

$$\begin{aligned}\sum_{t=1}^{n} \delta_t^1 &= (\sqrt{n}\,\overline{u}_1)^2 - \frac{2c}{\sqrt{n}}\sum_{t=1}^{n}(X_t - \overline{X})u_{t+1} - \frac{c^2}{n}\sum_{t=1}^{n}(X_t - \overline{X})^2 + O_p(n/T) \\ &= (\sqrt{n}\,\overline{u}_1)^2 - 2c\sigma_u\sigma_x R_n - c^2\sigma_x^2 + o_p(1) \\ &\xrightarrow{d} \sigma_u^2 z_1^2 - 2\sigma_u\sigma_x c z_2 - c^2\sigma_x^2\ ,\end{aligned}$$

where

$$\begin{aligned}\frac{\sqrt{n}}{\sigma_u}\overline{u} \;&\xrightarrow{d}\; z_1 \;\overset{d}{=}\; \mathcal{N}(0,1) \\ R_n \;&=\; \frac{1}{\sigma_u\sigma_x\sqrt{n}}\sum_{t=1}^{n}(X_t - \overline{X})u_{t+1} \;\xrightarrow{d}\; z_2 \;\overset{d}{=}\; \mathcal{N}(0,1)\end{aligned}$$

Accordingly, for the modified DM statistic we obtain

$$\widetilde{d}_1 = \frac{1}{\widehat{\sigma}_u^2}\sum_{t=1}^{n}\delta_t^1 \;\xrightarrow{d}\; z_1^2 - 2c\frac{\sigma_x}{\sigma_u}z_2 - c^2\frac{\sigma_x^2}{\sigma_u^2}$$

where $\widehat{\sigma}_u^2 = n^{-1}\sum_{t=1}^{n}(u_{t+1} - \overline{u}_1 + (c/\sqrt{n})X_t)^2 = \sigma_u^2 + O_p(n^{-1/2})$.

Using

$$\widehat{Y}_{t+1|t} = \frac{c}{\sqrt{n}}X_t + O_p(T^{-1/2})$$

we have for $n/\sqrt{T} \to 0$

$$\begin{aligned}\sum_{t=1}^{n}\xi_t^1 &= \sum_{t=1}^{n}\left[u_{t+1} - \overline{u}_1 + \frac{c}{\sqrt{n}}(X_t - \overline{X})\right]\widehat{Y}_{t+1|t} \\ &= \frac{c}{\sqrt{n}}\sum_{t=1}^{n}u_{t+1}(X_t - \overline{X}) + \frac{c^2}{n}\sum_{t=1}^{n}(X_t - \overline{X})^2 + O_p(n/\sqrt{T}) \\ &\xrightarrow{d} c\,\sigma_u\sigma_x z_2 + c^2\sigma_x^2\ .\end{aligned}$$

Furthermore

$$\begin{aligned}n\widehat{\omega}_\xi^2 \;&=\; \sum_{t=1}^{n}(\xi_t^1)^2 = \frac{c^2}{n}\sum_{t=1}^{n}(u_t - \overline{u}_1)^2(X_t - \overline{X})^2 + o_p(1) \\ &\xrightarrow{p} c^2\sigma_u^2\sigma_x^2\end{aligned}$$

and, thus,

$$\widehat{\varrho}_1 = \frac{\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\xi_t^1}{\sqrt{\widehat{\omega}_\xi^2}} \;\xrightarrow{d}\; \mathrm{sign}(c)z_2 + |c|\frac{\sigma_x}{\sigma_u}$$

with $\mathrm{sign}(a) = 1$ if $a \geq 0$ and $\mathrm{sign}(a) = -1$ for $a < 0$.

Appendix B - Data Descriptions

Concerning the quarter-on-quarter growth rates of real GDP, the only change that occurred in the sample is from West German GDP to the GDP of the reunified Germany in the fourth quarter of 1995. Forecasts for the euro area started being collected in the last quarter of 2002 for all variables except the 3-month interest rate.

The inflation measure used is the year-on-year growth rate of the index of consumer prices (CPI) for all countries except for the UK, where the retail price index (RPI) is used, because the sample of forecasts for the CPI does not start until 2004. Inflation forecasts for the reunified Germany started in the fourth quarter of 1996.

Private consumption is measured by the personal consumption expenditures in the US and Canada, by private consumption in Japan, Germany, and the Euro area, and by household consumption in France, the UK, and Italy. Private consumption forecasts for the reunified Germany started in the fourth quarter of 1995.

The 3-month interest rate is measured at the last day of the quarter. The interest rate used in the analysis is the 3-month treasury bill rate for the US and Canada, the 3-month Yen certificate of deposit rate for Japan, with the sample ending in the first quarter in 2010, the 3-month Euribor in Germany, Italy, and France, with the sample starting in the first quarter of 1999, and the 3-month interbank rate for the UK.

The changes in the VAT rates are listed in Table 7. Since Japan and the UK experienced two VAT rate changes of at least 2 percentage points, we adapt their samples of inflation and private consumption. For Japan, both samples start in the second quarter of 1997 and end in the fourth quarter of 2013, because in the first quarter of 2014, real private consumption already increased substantially due to the following VAT rate increase. For the UK, the samples continue to start in the first quarter of 1995, but end in the second quarter of 2008.

All resulting sample sizes can be found in Table 8.

Table 7: Changes in the value-added tax rates in percentage points

		VAT rate in pp		
Country	Date	from	to	change
Japan	Apr 97	3	5	2
Japan	Apr 14	5	8	3
Germany	Apr 98	15	16	1
Germany	Jan 07	16	19	3
France	Apr 00	20.6	19.6	-1
Italy	Oct 97	19	20	1
Italy	Sep 11	20	21	1
Italy	Oct 13	21	22	1
UK	Dec 08	17.5	15	-2.5
UK	Jan 10	15	17.5	2.5
UK	Jan 11	17.5	20	2.5
Canada	Jul 06	7	6	-1
Canada	Jan 08	6	5	-1

Table 8: Numbers of observations n

	US	EA	JP	DE	UK	IT	CA	FR
GDP q-o-q	79	48	79	76	79	79	79	79
CPI y-o-y	79	48	63	72	48	79	79	79
PrivCons q-o-q	79	48	63	76	48	79	79	79
d(3m rate)	79		53	63	79	63	79	63

Note: 'GDP q-o-q' denotes quarter-on-quarter growth rates of real GDP, 'CPI y-o-y' year-on-year growth rates of consumer prices, 'PrivCons q-o-q' quarter-on-quarter growth rates of real private consumption, and 'd(3m rate)' quarter-on-quarter changes of the end-quarter 3-month interest rate. The abbreviations used for the countries are 'US' for the United States, 'EA' for the euro area, 'JP' for Japan, 'DE' for Germany, 'UK' for the United Kingdom, 'IT' for Italy, 'CA' for Canada, and 'FR' for France.

Appendix C

Table 9: Results for case '$MA(1)_a$-$AR(1)$'

forecast horizon h	0	1	2	3	4	0	1	2	3	4
		$T = 100, n = 50$					$T = 100, n = 100$			
MSPE / Variance		0.96	1.03	1.03	1.03		0.96	1.03	1.02	1.02
rejections										
DM-type tests										
$\widetilde{d}_h$		0.68	0.12	0.12	0.12		0.80	0.07	0.07	0.08
d_h		0.69	0.12	0.12	0.12		0.81	0.08	0.08	0.08
encompassing tests										
$\widehat{\beta}_{1,h}$		0.57	0.04	0.03	0.03		0.75	0.02	0.03	0.03
$\widehat{\varrho}_h$		0.42	0.02	0.02	0.02		0.68	0.02	0.02	0.02
classical DM test		0.10	0.00	0.00	0.00		0.15	0.00	0.00	0.00
$\hat{h}^*$										
DM-type tests										
$\widetilde{d}_h$	0.32	**0.60**	0.06	0.02	0.01	0.20	**0.74**	0.04	0.01	0.00
d_h	0.31	**0.61**	0.06	0.02	0.01	0.19	**0.75**	0.05	0.01	0.00
encompassing tests										
$\widehat{\beta}_{1,h}$	0.43	**0.56**	0.02	0.00	0.00	0.25	**0.73**	0.02	0.00	0.00
$\widehat{\varrho}_h$	0.58	**0.41**	0.00	0.00	0.00	0.32	**0.67**	0.01	0.00	0.00
classical DM test	0.90	**0.10**	0.00	0.00	0.00	0.85	**0.15**	0.00	0.00	0.00

Note: For explanations, see Tables 3 and 4. In contrast to Table 4, the estimation is carried out using a rolling window of length T.