Hungnes, Håvard

Working Paper

Equal predictability test for multi-step-ahead system forecasts invariant to linear transformations

Discussion Papers, No. 931

Provided in Cooperation with:
Research Department, Statistics Norway, Oslo

Suggested Citation: Hungnes, Håvard (2020) : Equal predictability test for multi-step-ahead system forecasts invariant to linear transformations, Discussion Papers, No. 931, Statistics Norway, Research Department, Oslo

This Version is available at:
https://hdl.handle.net/10419/249121

Statistisk sentralbyrå
Statistics Norway

Equal predictability test for multi-step-ahead system forecasts invariant to linear transformations

TALL SOM FORTELLER
DISCUSSION PAPERS
931

Håvard Hungnes

Discussion Papers No. 931, May 2020
Statistics Norway, Research Department

Håvard Hungnes

Equal predictability test for multi-step-ahead system forecasts invariant to linear transformations

Abstract:
The paper derives a test for equal predictability of multi-step-ahead system forecasts that is invariant to linear transformations. The test is a multivariate version of the Diebold-Mariano test. An invariant metric for multi-step-ahead system forecasts is necessary as the conclusions otherwise can depend on how the forecasts are reported (e.g., as in levels or differences; or log-levels or growth rates). The test is used in comparing quarterly multi-step-ahead system forecasts made by Statistics Norway with similar forecasts made by Norges Bank.

Keywords: Macroeconomic forecasts; Econometric models; Forecast performance; Forecast evaluation; Forecast comparison.

JEL classification: C32, C53.

Acknowledgements: Thanks to Terje Skjerpen and the referees and participants at the ITISE 2018 conference for valuable comments on an earlier version of this paper.

Address: Håvard Hungnes, Statistics Norway, Research Department. E-mail: hhu@ssb.no

Discussion Papers	comprise research papers intended for international journals or books. A preprint of a Discussion Paper may be longer and more elaborate than a standard journal article, as it may include intermediate calculations and background material etc.

Abstracts with downloadable Discussion Papers
in PDF are available on the Internet:
http://www.ssb.no/en/forskning/discussion-papers
http://ideas.repec.org/s/ssb/dispap.html

ISSN 1892-753X (electronic)

Sammendrag

Når man tester ulike prognoser for en økonomisk størrelse noen perioder fram i tid, kan resultatet fort avhenge av hvordan man måler denne størrelsen. Dette kan illustreres ved å se på to ulike prognosemodeller for oljeprisen, hvor prognoser basert på den ene modellen er best når man vurderer oljeprisen målt på nivå, mens prognosene for den andre modellen er best når man vurderer oljeprisveksten. Hvis man isteden vurderer hele prognosebanene opp mot hverandre, spiller det ingen rolle om prognosene er formulert på nivå- eller endringsform da prognosefeilene fra den ene er en lineær transformasjon av prognosefeilene fra den andre.

I artikkelen foreslås det derfor å teste hele prognosebaner opp mot hverandre. Nullhypotesen er at de to prognosebanene er like gode. Forkastes denne hypotesen, kan vi konkludere med at den ene prognosebanen er signifikant bedre enn den andre.

Testen som utvikles i denne artikkelen benyttes til å teste SSBs prognoser for Fastlands-BNP, KPI og AKU-ledighet opp mot tilsvarende prognoser fra Norges Bank. Når prognosene fra SSB og Norges Bank er avgitt på omtrent samme tid, viser resultateten her at prognosebanene er om lag like gode.

1 Introduction

Clements and Hendry (1993) show that evaluation of forecasts of individual variables at each horizon separately is not invariant to linear transformations of the forecasts. Ericsson (2008) illustrates this by considering two different models for forecasting the oil price, where the multi-step-ahead forecasts based on one of the models are considered better when the forecasts are examined in terms of levels, but where the forecasts of another model is considered better when the forecasts are evaluated in terms of growth rates. Clements and Hendry (1993) suggested a metric of the whole system of forecasts when evaluating the system forecasts. However, 25 years later, Hendry and Martinez (2017) point out that "relatively little work has been done to evaluate the accuracy of the whole system jointly." Nor are there many papers that consider the whole multi-step-ahead system forecast when comparing forecasts.

Multi-step-ahead system forecasts tell a consistent story of the economy as they describe both the path of the variables as well as the relationships between them. Multi-step-ahead forecasts of one variable identify turning points, whereas forecasts of multiple variables identify co-movements of these variables. Policy makers are also interested in system forecasts. For example, as noted by Martinez (2017), "central banks care about the future trajectory of inflation, output and unemployment."

Usually, metrics for forecast accuracy only consider forecasts for one variable at one forecasting horizon. Measures such as mean absolute forecast errors and mean squared forecast errors (or variants of these) are usually applied. Unfortunately, none of these metrics are invariant to linear transformations of the forecasts (such as measuring the forecast errors in growth rates instead of log-levels). However, Clements and Hendry (1993) suggest a metric for accuracy of a system forecast that is equivalent to the predictive likelihood; see Bjørnstad (1990), Hinkley (1979), and Mathiasen (1979). Engle (1993) suggests an alternative metric for accuracy based on a quadratic loss function.

There have been some important contributions to evaluating system forecasts. Kolsrud (2007) suggests using prediction bands for a multi-step-ahead path forecast of a univariate time series, and Kolsrud (2015) extends this to multivariate time series. Jordà and Marcellino (2010) consider the forecasts of a variable for all considered forecasting horizons jointly, and for each variable, they derive the prediction regions of this path based on the covariance matrix of the forecast errors. The prediction region will then be independent of linear transformations of the forecasts. Furthermore, considering systems of more variables, Jordà and Marcellino (2010) show how these prediction regions for the forecast path of one variable change with different assumptions about the future path of other variables. The prediction regions in Jordà and Marcellino (2010) can also be used to test for the absence of a forecasting bias. A similar test for unbiased system forecast is presented in Sinclair et al. (2012, 2015). Sinclair and Stekler (2013) apply this approach to test for biases in the revision of forecasts.

In this paper, we consider an equal predictability test for multi-step-ahead system forecasts. The test compares the predictive likelihood of the equal-weighted combination of the two system forecasts with the predictive likelihood of the optimal-weighted combination of the two forecasts. The optimal-weighted combination implies using the weights of the two system forecasts that maximizes the joint predictive likelihood. We show that the test is a multivariate version of the Diebold and Mariano (1995) test.[1] The test is also related to the encompassing test for multi-step-ahead system forecast suggested in Hungnes (2018).

Quaedvlieg (2019) and Martinez (2017) consider equal predictability tests for system forecasts. How-

[1] Pesaran and Skouras (2002) have also suggested a multivariate version of the Diebold and Mariano (1995) test where a weighting matrix must be used. The present paper motivates using the covariance matrix of the equally weighted forecast errors of the two forecast systems as this weighting matrix.

ever, by ignoring the dependency between forecast errors of different horizons and between different variables, the test in Quaedvlieg (2019) is not invariant to linear transformations. Thus, results can depend on such normalizations, for example, if the forecasts are measured in growth rates instead of log levels. The equal predictability test for system forecast in Martinez (2017) is invariant to linear transformations.

Under some conditions, the test in Diebold and Mariano (1995) is asymptotically normally distributed. However, Diebold and Mariano (1995) provide simulation experiments that show that the normal distribution can be a very poor approximation when applying the test statistic to small samples. The test will typically reject the null too often. To improve the small-sample properties, Harvey et al. (1997) suggest both a bias correction to the test statistic as well as comparing the corrected statistics with a Student-t distribution. In the present paper, we account for these two improvements.

The Diebold and Mariano (1995) test requires that the difference of the loss function based on the squared forecast errors between two forecasters (or models) is covariance stationary, see also Diebold (2015). West (1996) and Clark and McCracken (2001) consider the case where the forecasts are based on econometric models with estimated parameters and the forecast tests are conducted to compare the forecasting models (see also Clark and McCracken, 2013, Ch. 3.1).[2] West (1996) shows that the distribution of the test statistic is still asymptotically normal for non-nested forecasting models. For nested models, Clark and McCracken (2001) show that the distribution may be non-standard.

Giacomini and White (2006) consider comparisons of forecasts of models with estimated parameters. The parameters are estimated with a rolling sample, so the estimates do not converge to their true values as more observations become available. Under this estimation scheme, they show that the test of equal predictability is asymptotically normally distributed. Therefore, the approach by Giacomini and White (2006) is more in line with the Diebold and Mariano (1995) test than the tests considered by West (1996) and Clark and McCracken (2001), among others. According to Patton (2015), tests comparing forecasts from estimated models, or surveys, or judgemental forecasts, correspond to this type of Diebold and Mariano (1995) test.

The rolling sample assumption in Giacomini and White (2006) is crucially for obtaining the asymptotic normal distribution of the tests (see also Clark and McCracken, 2015; McCracken, 2019). When applying a forecasting model estimated on a recursive sample (i.e., an expanding sample with fixed starting date), the estimation bias of both a correctly specified model and an over-fitted model will vanish and the difference of the squared forecast errors will decrease with time, violating the assumption behind the Diebold and Mariano (1995) test. However, in a frequently changing economy with structural breaks in the data generating process, a rolling sample will be more effective for detecting the data generating process at the time the forecasts are made, a point also noted by Giacomini and White (2006).

The equal predictability test for multi-step-ahead system forecasts is used to compare forecasts made by Statistics Norway with forecasts made by Norges Bank. We investigate jointly the forecasts of GDP, CPI, and the unemployment rate for the same year as the forecast are made as well as for the following year.

When making forecasts of the Norwegian economy, Statistics Norway applies an econometric model. However, in the process of making the forecasts, the forecasts are also influenced by judgment where other information than what is included in the econometric model is used to improve the forecasts (see also Lawrence et al., 2006, for a review of judgmental forecasting techniques). The forecast testing we

[2] Clark and McCracken (2013) refer to this as "testing population-level predictive ability [...] that is, the accuracy of the forecasts at unknown population values of the parameters". The Diebold and Mariano (1995) test implies "testing finite-sample predictive ability [...] that is, the accuracy of the forecasts at estimated values of the parameters."

are conducting here is accordingly not a part of testing the underlying model.

The rest of the paper is organized as follows: In Section 2, the theoretical background for the test as well, as the proposed test for equal predictability, are presented. In Section 3, the proposed equal predictability test is applied to compare the forecasts made by Statistics Norway with forecasts made by Norges Bank. Section 4 concludes.

2 Theory

The necessary theory for the equal predictability test is presented in Section 2.1. The implied autocorrelation in the forecasts is derived in Section 2.2. The standard Diebold and Mariano (1995) test is presented in Section 2.3. The test statistic for equal predictability of system forecasts and its distribution is derived in Section 2.4. Small sample properties of the test are discussed in Section 2.5.

2.1 Measures of forecast accuracy

Let $y^i_{t+h|t}$ be the forecast of variable i in period $t+h$ made in period t. We assume that the value of y in period t is not known in period t; hence forecast for the current period — also referred to as nowcasting — can be made and are denoted $y^i_{t|t}$. The forecast error of variable i in period $t+h$ made in period t is defined as

$$e^i_{t+h|t} \equiv y^i_{t+h} - y^i_{t+h|t}, \tag{1}$$

where y^i_{t+h} is the outcome of variable i in period $t+h$. If the variables are measured on the logarithmic scale, the forecast error in (1) is approximately a measure of the percentage error (when disregarding the scaling factor). The use of the logarithmic scale can be appropriate for many macroeconomic variables such as GDP, where we are more interested in the forecast error measured in percent than in, say, dollars.

The Mean Squared Forecast Error (MSFE) is given by

$$T^{-1}\sum_{t=1}^{T}\left(e^i_{t+h|t}\right)^2, \tag{2}$$

which expresses the mean squared forecast error of variable i forecasted h periods for forecasts made in T consecutive periods. The MSFE (or the square root of it) is a widely used metric for the accuracy of forecast and comparison of forecasts. However, it can be problematic to use this metric when comparing a system forecast of multiple variables or multiple forecasting horizons. To see the former, consider the following example: Suppose that variable 1 is (log of) consumption and variable 2 is (log of) income. One metric of the accuracy of nowcasts of these two variables could be the sum of the MSFE of the two variables,

$$T^{-1}\sum_{t=1}^{T}\left(e^1_{t|t}\right)^2 + T^{-1}\sum_{t=1}^{T}\left(e^2_{t|t}\right)^2. \tag{3}$$

However, we could alternatively consider the sum of the MSFE errors of (log) consumption and the savings ratio (defined as the difference between the log of income and the log of consumption). This metric would then be

$$T^{-1}\sum_{t=1}^{T}\left(e^1_{t|t}\right)^2 + T^{-1}\sum_{t=1}^{T}\left(e^2_{t|t} - e^1_{t|t}\right)^2, \tag{4}$$

which is not identical to the metric in (3).

To make a metric that is invariant to linear transformations of the variables related to forecasting horizon h, we define a vector of the forecast of all N variables in period $t+h$ made in period t as $\mathbf{y}_{t+h|t} = \left(y^1_{t+h|t}, y^2_{t+h|t}, \ldots, y^N_{t+h|t}\right)'$. Similarly, the outcome of these variables in period $t+h$ is $\mathbf{y}_{t+h} = \left(y^1_{t+h}, y^2_{t+h}, \ \ldots, y^N_{t+h}\right)'$, which implies that the forecast error vector becomes $\mathbf{e}_{t+h|t} = \left(e^1_{t+h|t}, e^2_{t+h|t}, \ldots, e^N_{t+h|t}\right)'$ with elements defined as in (1). Two alternative metrics for forecast accuracy will now be considered. One of these metrics is based on a matrix version of the observable MSFE for forecasting horizon h given as

$$\hat{\mathbf{V}}_h = T^{-1} \sum_{t=1}^{T} \mathbf{e}_{t+h|t} \mathbf{e}'_{t+h|t}. \tag{5}$$

This matrix is of dimension $N \times N$, and it is not obvious how to compare forecasts based on this matrix. The metric that we considered above and which was shown not to be invariant to linear transformations, corresponds to use the trace of the matrix in (5): if $N = 2$ and $h = 0$, then the trace is given by (3).

An alternative metric for forecast accuracy is to apply a quadratic loss function, see Engle (1993). The loss of the forecast errors for the forecast made at time t is then given by

$$\mathbf{e}'_{t+h|t} \mathbf{H} \mathbf{e}_{t+h|t}, \tag{6}$$

where $\mathbf{H}$ is an $N \times N$ positive definite matrix of constants which represents the relative cost of different errors. The average loss over for the forecasts made in T periods would be $T^{-1} \sum_{t=1}^{T} \mathbf{e}_{t+h|t} \mathbf{H} \mathbf{e}'_{t+h|t}$. However, this metric will not be invariant to linear transformations of the forecasts unless the weighting matrix $\mathbf{H}$ is adjusted accordingly.

Now we will consider how to make use of these two metrics for forecast accuracy such that they are invariant to linear transformations of the forecasts. We do so by considering a linear transformation of the forecasts given by the $N \times N$ non-singular matrix $\mathbf{M}$. Assume that $\mathbf{e}_{t+h|t}$ is a vector of forecast errors for forecasts h periods ahead with one representation of the variables (which we, for simplicity, will also refer to as the original representation of the variables); and $\mathbf{e}^*_{t+h|t} = \mathbf{M} \mathbf{e}_{t+h|t}$ is the corresponding vector of forecast errors for another representation of the variables. For example, if the variables in the first representation are (logs of) consumption and (logs) of income, and in the second representation the variables are (logs of) consumption and the savings ratio. In this example the transformation matrix from the first formulation of the variables to the second formulation is given by $\mathbf{M} = \left[\begin{smallmatrix} 1 & 0 \\ -1 & 1 \end{smallmatrix}\right]$.

With the original representation of the variables the MSFE matrix is given as in (5). With the alternative representation of the variables the MSFE matrix is

$$\hat{\mathbf{V}}^*_h = T^{-1} \sum_{t=1}^{T} \mathbf{e}^*_{t+h|t} \mathbf{e}^{*\prime}_{t+h|t} = T^{-1} \mathbf{M} \sum_{t=1}^{T} \left(\mathbf{e}_{t+h|t} \mathbf{e}'_{t+h|t} \right) \mathbf{M}'. \tag{7}$$

The matrices in (5) and (7) are not equal. However, as pointed out by Clements and Hendry (1993), the determinant of (5) is equal to the determinant of (7) if $|\mathbf{M}| = 1$, a property Clements and Hendry (1993) refer to as a scale-preserving linear transformation. Taking the determinant of the MSFE matrix is equivalent to using the predictive likelihood, as suggested by Bjørnstad (1990), Hinkley (1979), and Mathiasen (1979).

It is also worth noting that the trace of $\hat{\mathbf{V}}_h$ does not equal the trace of $\hat{\mathbf{V}}^*_h$. The trace of these matrices corresponds to evaluating the accuracy of the forecasts by the sum of the individual MSFE of the

forecasted variables. Hence, the sum of the MSFE of the different forecasted variables is not invariant to linear transformations.

Now consider the metric in (6). The loss of the forecast errors made at time t is given by (6) with the original representation of the variables, and with $\mathbf{e}^{*\prime}_{t+h|t}\mathbf{H}^*\mathbf{e}^*_{t+h|t}$ with the alternative representation of the variables when we allow for another weighting matrix with this alternative representation. The two metrics of the loss of the forecast errors made at time t are equal if $\mathbf{H}^* = \mathbf{M}'^{-1}\mathbf{H}\mathbf{M}^{-1}$. Hence, if the weighting matrix is adjusted accordingly to the transformation of the forecast errors, this metric will be invariant to linear transformations.

The MSFE (or the root of it) is a widely used metric for the accuracy of forecast also for $h > 0$; see, e.g., Bjørnland et al. (2017), El-Shagi et al. (2016), Jungmittag (2016), and Kock and Teräsvirta (2016) for some recent applications. However, the MSFE for measuring the forecast accuracy when $h > 0$ depends on how the forecasts are measured, see Clements and Hendry (1993). Only in the case with one variable and $h = 0$ (univariate nowcasting) comparison based on the observed (univariate) MSFE are invariant to linear transformations of the forecasts, see Clements and Hendry (1993, 1998).

To compare forecasts generated by different models, we need to consider all forecasts up to forecast horizon H (where H denotes the longest forecast horizon). Therefore, we define $\mathbf{Y}_{t,H|t}$ to be the vector of forecasts of $\mathbf{y}_{t+h|t}$ in each period from period t to period $t+H$ made at time t, i.e., $\mathbf{Y}_{t,H|t} = \left(\mathbf{y}'_{t|t}, \mathbf{y}'_{t+1|t}, \ldots, \mathbf{y}'_{t+H|t}\right)'$. The forecast error of $\mathbf{Y}_{t,H|t}$ is given by $\mathbf{E}_{t,H|t} \equiv \mathbf{Y}_{t,H} - \mathbf{Y}_{t,H|t}$, where $\mathbf{Y}_{t,H} = \left(\mathbf{y}'_t, \mathbf{y}'_{t+1}, \ldots, \mathbf{y}'_{t+H}\right)'$ is the vector of the outcome of all the variables from period t to period $t+H$. This implies that vector of forecast errors is $\mathbf{E}_{t,H|t} = \left(\mathbf{e}'_{t|t}, \mathbf{e}'_{t+1|t}, \ldots, \mathbf{e}'_{t+H|t}\right)'$.

A matrix version of the observable MSFE for all forecasting horizons up to period H would then be

$$\mathbf{V}_H = T^{-1}\sum_{t=1}^{T}\mathbf{E}_{t,H|t}\mathbf{E}'_{t,H|t}, \tag{8}$$

which is of dimension $K \times K$ with $K = N(H+1)$. As above, the determinant of this matrix is an invariant metric for forecasts accuracy if the linear transformation is scale-preserving.

The metric based on the quadratic loss function can also be used here;

$$\mathbf{E}'_{t,H|t}\mathbf{H}\mathbf{E}_{t,H|t}$$

(with $\mathbf{H}$ now of dimension $K \times K$) which also is invariant to linear transformations if the weighting matrix is adjusted accordingly.

2.2 Autocorrelation

Multi-step-ahead forecasts lead to autocorrelation in the forecast errors. Suppose the process of the considered variables has the following Wold representation

$$\mathbf{y}_t = \sum_{i=0}^{\infty}\Gamma_i v_{t-i},$$

with $\Gamma_0 = I_N$, and where deterministic variables are ignored for simplicity. The optimal h period ahead forecast given at time t and provided that the coefficients in the infinite matrix lag polynomial $I_N + \Gamma_1 L + \Gamma_2 L^2 + \ldots$ (where L is the lag operator; $L^\ell x_t = x_{t-\ell}$) are known and that the error in period

t is not known, is

$$\mathbf{y}_{t+h|t} = \sum_{i=h+1}^{\infty} \Gamma_i v_{t+h-i},$$

and the forecast error then becomes

$$\mathbf{e}_{t+h|t} = \mathbf{y}_{t+h} - \mathbf{y}_{t+t|t} = \sum_{i=0}^{h} \Gamma_i v_{t+h-i}.$$

Therefore, we can specify the vector of forecast errors for all forecasting periods from 0 (nowcasting) to H periods ahead forecast as

$$\begin{pmatrix} \mathbf{e}_{t|t} \\ \mathbf{e}_{t+1|t} \\ \mathbf{e}_{t+2|t} \\ \vdots \\ \mathbf{e}_{t+H|t} \end{pmatrix} = \begin{pmatrix} I_N & 0 & 0 & \cdots & 0 \\ \Gamma_1 & I_N & 0 & \ddots & 0 \\ \Gamma_2 & \Gamma_1 & I_N & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \Gamma_H & \cdots & \cdots & \cdots & I_N \end{pmatrix} \begin{pmatrix} v_t \\ v_{t+1} \\ v_{t+2} \\ \vdots \\ v_{t+H} \end{pmatrix},$$

which shows that optimal forecasts up to a horizon H have autocorrelation of order H. The reason is that a forecast made in period $t+H$ will partly overlap with a forecast made in period t, as both sets of forecasts will involve forecasts of variables for period $t+H$. However, an optimal forecast made in period $t+H+1$ will not overlap with a forecast made in period t. Hence, the forecast errors in the two sets of forecasts are not expected to be correlated. This property is also shown in Hendry and Martinez (2017) and used by Harvey et al. (1997, 1998), and Harvey and Newbold (2000), among others.

2.3 The Diebold-Mariano test

Diebold and Mariano (1995) suggest a test for equal predictability. Let $e^{i,j}_{t+h|t}$ $(j = A, B)$ be the forecast error of the forecast made by forecaster j in period t of variable i in period $t+h$. Then consider a loss-difference series $d^i_{t,h} = \left(e^{i,A}_{t+h|t}\right)^2 - \left(e^{i,B}_{t+h|t}\right)^2$ for variable i made in period t with a forecasting horizon h, when applying a quadratic loss function. When the time series $\{d^i_{t,h}\}_{t=1}^{T}$ is covariance stationary with a short memory, Diebold and Mariano (1995) and Diebold (2015) show that the mean of this series is asymptotically normally distributed:

$$T^{1/2}\left(\bar{d}^i_h - \mu^i_h\right) \xrightarrow{d} N\left(0, q_{h,i}\right)$$

where $\bar{d}^i_h = T^{-1}\sum_{t=1}^{T} d^i_{t,h}$; μ^i_h is the population mean of the loss-difference for variable i at forecasting horizon h; and $q_{h,i}$ is the sum of the autocovariances of d^i_h, $q_{h,i} = \sum_{s=-\infty}^{\infty} E\left[\left(d^i_{t,h} - \mu^i_h\right)\left(d^i_{t-s,h} - \mu^i_h\right)\right]$. The test statistic Diebold and Mariano (1995) suggests for the null hypothesis $\mu^i_h = 0$ is simply

$$T^{1/2}\bar{d}^i_h \widehat{q}^{-1/2}_{h,i}, \tag{9}$$

with

$$\widehat{q}_{h,i} = \frac{1}{T}\left[\sum_{t=1}^{T}\left(d^i_{t,h} - \bar{d}^i_h\right)^2 + 2\sum_{l=1}^{\tau_H}\sum_{t=1}^{T-l}\left(d^i_{t,h} - \bar{d}^i_h\right)\left(d^i_{t+l,h} - \bar{d}^i_h\right)\right], \tag{10}$$

where τ_H is the truncation lag, with $\tau_H \geq H$ since we in Section 2.2 showed that we have autocorrelation of order H with optimal forecasts and may have a higher order of autocorrelation if the forecasts are not optimal. The null hypothesis of $\mu_h^i = 0$ implies equal predictability. If this null hypothesis is rejected and $\bar{d}_h^i < 0$, then forecast A of variable i at horizon h is significantly better than forecast B; and $\bar{d}_h^i > 0$ implies the opposite.

The test statistics in (9) only considers a univariate forecast, i.e., a forecast of one variable at one forecasting horizon. Pesaran and Skouras (2002) present the loss-difference series

$$d_t = \mathbf{E}_{t,H|t}^{A\prime} \mathbf{H} \mathbf{E}_{t,H|t}^{A} - \mathbf{E}_{t,H|t}^{B\prime} \mathbf{H} \mathbf{E}_{t,H|t}^{B} \tag{11}$$

for the multivariate quadratic model where the $K \times K$ matrix $\mathbf{H}$ depends on the parameters in the loss function. Capistrán (2006) suggests using $\mathbf{H} = I_K$ (where I_K is the identity matrix of order K), and Quaedvlieg (2019) suggests (in his weighted average loss test) using a diagonal matrix with weights along its main diagonal. However, none of these suggestions leads to a metric that is invariant to linear transformations of the forecasts.

2.4 The test statistic for equal predictability

Williams and Kloot (1953) introduce a general test for equal predictability between two models; see also Granger and Newbold (1986, Charpter 9) and Howrey (1993). Consider two different forecasts of variable i in period $t+h$ made in period t; denoted $y_{t+h|t}^{i,A}$ and $y_{t+h|t}^{i,B}$, and the following relationship:

$$y_{t+h}^{i} = (1-\alpha)\, y_{t+h|t}^{i,A} + \alpha y_{t+h|t}^{i,B} + v_{t+h|t}^{i}. \tag{12}$$

Equal predictability implies $\alpha = \frac{1}{2}$. By utilizing the definition of forecast error, the above expression can also be formulated as $e_{t+h|t}^{i,A} = \alpha \left(e_{t+h|t}^{i,A} - e_{t+h|t}^{i,B} \right) + v_{t+h|t}^{i}$. Furthermore, by applying the vectors of forecasting errors for forecasters A and B, we can write

$$\mathbf{E}_{t,H|t}^{A} = \alpha \left(\mathbf{E}_{t,H|t}^{A} - \mathbf{E}_{t,H|t}^{B} \right) + \mathbf{V}_{t,H|t}, \tag{13}$$

where the error vector is $\mathbf{V}_{t,H|t} = \left(\mathbf{v}_{t|t}', \mathbf{v}_{t+1|t}', \ldots, \mathbf{v}_{t+H|t}' \right)'$ with $\mathbf{v}_{t+h|t} = \left(v_{t+h|t}^{1}, v_{t+h|t}^{2}, \ldots, v_{t+h|t}^{N} \right)'$. Finally, by multiplying this expression with 2 and subtracting $\mathbf{E}_{t,H|t}^{A} - \mathbf{E}_{t,H|t}^{B}$ on both sides, we have

$$\mathbf{y}_t = \gamma \mathbf{x}_t + \mathbf{U}_{t,H|t}, \tag{14}$$

where $\mathbf{y}_t = \mathbf{E}_{t,H|t}^{A} + \mathbf{E}_{t,H|t}^{B}$, $\mathbf{x}_t = \mathbf{E}_{t,H|t}^{A} - \mathbf{E}_{t,H|t}^{B}$, $\gamma = 2\alpha - 1$ and $\mathbf{U}_{t,H|t} = 2\mathbf{V}_{t,H|t}$. The hypothesis of equal predictability can now be formulated as $\gamma = 0$. If, say, the estimate of γ is negative, we can test if it is significantly different from zero. If it is significantly different from zero, we say that forecast A is significantly better than forecast B.[3]

The conditional estimators for γ and the $K \times K$ covariance matrix of $\mathbf{U}_{t,H|t}$ (which we denote by Σ)

[3]Note also that the hypothesis that $\gamma = -1 \Leftrightarrow \alpha = 0$ corresponds to a test of forecast A encompassing forecast B, i.e., forecast A contains all information, so there is no additional information provided by forecast B. This test is not considered here, see Hungnes (2018) for a system version of this test.

in (14) are given by (when ignoring possible degrees of freedom adjustments for the covariance matrix)

$$\hat{\gamma}_{(\mathbf{S})} = \left(\frac{1}{T} \sum_{t=1}^{T} x_t' \mathbf{S}^{-1} x_t \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} x_t' \mathbf{S}^{-1} y_t \right), \tag{15}$$

$$\hat{\Sigma}_{(g)} = \frac{1}{T} \sum_{t=1}^{T} (y_t - g x_t) (y_t - g x_t)', \tag{16}$$

where the subscript in parenthesis indicates that the estimates are a function of another parameter or matrix of parameters, such that $\mathbf{S}$ is a $K \times K$ matrix representing some estimate of Σ and that g is a scalar representing some estimate of γ. The FIML estimates in (15) and (16) can be obtained by carrying out an iterative procedure until convergence, and the final estimates will equal the ones obtained with full information maximum likelihood estimation, see Oberhofer and Kmenta (1974). In this case $\mathbf{S} = \hat{\Sigma}_{(g)}$ and $g = \hat{\gamma}_{(\mathbf{S})}$.

An alternative to obtaining the estimates of γ and Σ is to apply some GLS estimators. In the hypothesis testing, we also consider the GLS estimators where Σ is estimated under the null hypothesis of $\gamma = 0$, denoted $\hat{\Sigma}_{(0)}$, and γ is estimated conditional on $\hat{\Sigma}_{(0)}$, i.e., $\hat{\gamma}_{(\hat{\Sigma}_{(0)})}$.

When both heteroscedaticity and autocorrelation in the forecasts made at different time periods are considered, a robust estimator of the variance of the estimate of γ is

$$\widehat{Var\left(\hat{\gamma}_{(\mathbf{S})}\right)} = \frac{1}{T} \left[\frac{1}{T} \sum_{t=1}^{T} x_t' \mathbf{S}^{-1} x_t \right]^{-2} \left[\frac{1}{T} \sum_{t=1}^{T} d^2_{(\hat{\gamma}_{(\mathbf{S})},\mathbf{S}),t} + \frac{2}{T} \sum_{l=1}^{\tau_H} \sum_{t=1}^{T-l} d_{(\hat{\gamma}_{(\mathbf{S})},\mathbf{S}),t} d_{(\hat{\gamma}_{(\mathbf{S})},\mathbf{S}),t+l} \right], \tag{17}$$

where

$$d_{(g^*,\mathbf{S}),t} = x_t' \mathbf{S}^{-1} (y_t - g^* x_t) \tag{18}$$

is a measure of how much $\hat{\gamma}$ deviates from g^*.

Section 2.2 shows that with optimal forecasts H steps ahead (including nowcasting), there will be autocorrelation up to order H and no autocorrelations above order H. Though, we do not know if the forecasts are optimal. Therefore, we allow for autocorrelation up to order τ_H, where $\tau_H \geq H$.

The term in the last square brackets in (17) expresses the variance of $d_{(\gamma_{(\mathbf{S})},\mathbf{S}),t}$. To secure that this variance is positive, we may use

$$\mathbf{Q}_{(\hat{\gamma}_{(\mathbf{S})},\mathbf{S})} = \left[\frac{1}{T} \sum_{t=1}^{T} d^2_{(\hat{\gamma}_{(\mathbf{S})},\mathbf{S}),t} + \frac{2}{T} \sum_{l=1}^{\tau_H} \sum_{t=1}^{T-l} w_l d_{(\hat{\gamma}_{(\mathbf{S})},\mathbf{S}),t} d_{(\hat{\gamma}_{(\mathbf{S})},\mathbf{S}),t+l} \right], \tag{19}$$

where $w_l = 1 - \frac{l}{H+1}$ $(l = 1, \ldots, \tau_H)$.[4]

The expression for $d_{(g^*,\mathbf{S}),t}$ in (18) with $g^* = 0$ represents a multivariate version of Diebold and Mariano (1995) loss differential. By inserting the expressions for x_t and y_t we have

$$\begin{aligned} d_{(0,\mathbf{S}),t} &= \left(\mathbf{E}^A_{t,H|t} - \mathbf{E}^B_{t,H|t} \right)' \mathbf{S}^{-1} \left(\mathbf{E}^A_{t,H|t} + \mathbf{E}^B_{t,H|t} \right) \\ &= \mathbf{E}^{A\prime}_{t,H|t} \mathbf{S}^{-1} \mathbf{E}^A_{t,H|t} - \mathbf{E}^{B\prime}_{t,H|t} \mathbf{S}^{-1} \mathbf{E}^B_{t,H|t}, \end{aligned} \tag{20}$$

which is the multivariate version of the Diebold and Mariano (1995) loss differential based on squared forecast errors. This expression is equal to (11) when we use $\mathbf{H} = \mathbf{S}^{-1}$ as the weighting matrix.

Based on the ratio of the estimate in (15) and its standard error (the square root of (17)) a test statistic

[4] Harvey et al. (2017) consider various alternatives and investigate their small sample properties in forecasting.

for testing the null hypothesis of $\gamma = 0$ can be formulated as

$$\mathbb{T}_{(\mathbf{S})} = T^{1/2} w_0^{1/2} \bar{d}_{(0,\mathbf{S})} \mathbf{Q}_{(\hat{\gamma}_{(\mathbf{S})},\mathbf{S})}^{-1/2}, \tag{21}$$

where $\bar{d}_{(0,\mathbf{S})}$ is the sample mean of (18) with $g^* = 0$, and $\mathbf{S}$ could be based on the FIML estimate or the GLS estimate. In small samples, Harvey et al. (1997) suggest that it has a t-distribution when forecasting only one variable. Furthermore, they derive the small sample correction factor $w_0 = T^{-1}\left[T - 1 - 2H + T^{-1}H(H+1)\right]$. Both the t-distribution and the correction factor will be used here.

In the univariate case, i.e., when a forecast is made for only one variable at one specific forecasting horizon and not a vector, Harvey et al. (1998) show that the test statistic that is similar to (21) over-rejects in small samples. Therefore, Harvey et al. (1997, 1998) suggest a modification of the test where the variance of d is derived relative to its sample mean. Then $\mathbf{Q}_{(\hat{\gamma},\Sigma)}$ in (21) is replaced with

$$\mathbf{Q}^*_{(0,\mathbf{S})} = \frac{1}{T}\left[\sum_{t=1}^{T}\left(d_{(0,\mathbf{S}),t} - \bar{d}_{(0,\mathbf{S})}\right)^2 + 2\sum_{l=1}^{\tau_H}\sum_{t=1}^{T-l} w_l \left(d_{(0,\mathbf{S}),t} - \bar{d}_{(0,\mathbf{S})}\right)\left(d_{(0,\mathbf{S}),t+l} - \bar{d}_{(0,\mathbf{S})}\right)\right]. \tag{22}$$

The t-statistic then becomes

$$\mathbb{T}^*_{(\mathbf{S})} = T^{1/2} w_0^{1/2} \bar{d}_{(0,\mathbf{S}_{(0)})} \mathbf{Q}^{*-1/2}_{(0,\mathbf{S})}, \tag{23}$$

where $\mathbf{S}$ could be based on the FIML or the GLS estimate.

For the test statistics in (23) to have desirable properties, we follow Giacomini and White (2006), and assume that the two forecasts are generated by measurable functions of the most recent observations of the vector $\mathbf{z}_t$, where this vector at least contains the vector of variables we are forecasting, i.e. $\mathbf{y}_t = \left(y_t^1, y_t^2, \ldots, y_t^N\right)'$. More precisely, we consider the stochastic process $\mathbf{Z} = \{\mathbf{z}_t : \Omega \to \mathbb{R}^{N+N_X}, N + N_X \in \mathbb{N}, t = 1, 2, \ldots\}$ defined on a complete probability space $(\Omega, \mathcal{F}, P)$, where the observed vector $\mathbf{z}_t$ is partitioned as $\mathbf{z}_t = (\mathbf{y}_t', \mathbf{x}_t')'$, with $\mathbf{y}_t : \Omega \to \mathbb{R}^N$ being the vector of variables being forecasted and $\mathbf{x}_t : \Omega \to \mathbb{R}^{N_X}$ being a vector of N_X predictors. Let $\mathcal{F} = \sigma\left(\mathbf{z}_1', \ldots, \mathbf{z}_t'\right)$ be the information set at time t. Suppose two alternative models are used to produce a system of path forecasts by $\mathbf{Y}^i_{t,H|t} = f_i\left(\mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \ldots, \mathbf{z}_{t-m_i}; \hat{\Pi}_{i,m_i,t}\right)$ for $i = A, B$, where f_A and f_B are measurable functions. The vector $\hat{\Pi}_{i,m_i,t}$ is estimated based on m_i most recent observations of $\mathbf{z}_t$; $\hat{\Pi}_{i,m_i,t} = \hat{\Pi}_{i,m_i,t}\left(\mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \ldots, \mathbf{z}_{t-m_i}\right)$. See also McCracken (2019) for the importance of the parameters being estimated based on a rolling estimation window.

Giacomini and White (2006) provide a theorem that here is modified to path forecasts. It applies the assumption of a mixing process, where the ϕ-mixing process is due to Ibragimov (1959, 1962) and the α-mixing process was introduced by Rosenblatt (1956).

Theorem 2.1 *Given a finite estimation window $m_i < \infty$ for $i = A, B$, suppose*

(i) $\{\mathbf{z}_t\}$ *is a mixing sequence with ϕ of size $-r/(2r-2)$, $r \geq 2$, or α of size $-r/(r-2)$, $r > 2$;*

(ii) $\mathbb{E}\left|d_{(0,\mathbf{S}),t}\right|^{2r} < \infty$ *for all t;*

(iii) $q^*_{(0,\mathbf{S})} = var\left[\sqrt{T}\bar{d}_{(0,\mathbf{S}),t}\right]$ *for all T sufficiently large.*

Then, when applying (22) with $w_l \to 1$ and $\tau_H \to \infty$ as $T \to \infty$, we have:

(a) under H_0 of $\mathbb{E}\left[\bar{d}_{(0,\mathbf{S})}\right] = 0$, $\mathbb{T}^*_{(\mathbf{S})} \xrightarrow{d} N(0,1)$ as $T \to \infty$, and

(b) under H_A *of* $\left(\mathbb{E}\left[\bar{d}_{(0,\mathbf{S})}\right]\right)^2 \geq \delta^* > 0$ *for* T *sufficiently large, for any constant* $c \in \mathbb{R}$, $P\left[\mathbb{T}^*_{(\mathbf{S})} > c\right] \to 1$ *as* $T \to \infty$.

The proof follows directly from Giacomini and White (2006, Proof of Theorem 4) where $W_t = z_t$ and $\Delta L_t = d_{(0,\mathbf{S}),t}$, see also Hungnes (2018, Appendix A).

Remark 2.1 *Theorem 2.1 is derived under the assumption of* $\mathbf{S}$ *being and* $K \times K$ *non-singular matrix such that* $\mathbf{S}^{-1}$ *and thus* $d_{(0,\mathbf{S}),t}$ *exists. Therefore, in the theorem* $\mathbf{S}$ *is not updated as* $T \to \infty$. *We may set* $\mathbf{S} = \hat{\Sigma}_{(g)}$ *with* $g = \hat{\gamma}_{(\mathbf{S})}$, *or* $S = \hat{\Sigma}_{(0)}$, *which implies that the theorem applies for both the FIML and the GLS version of the test statistic in* (23).

Remark 2.2 *By using* (14) *and* (18) *we have* $\mathbb{E}\left[\bar{d}_{(g^*,\mathbf{S})}\right] = \mathbb{E}\left[T^{-1}\sum_{t=1}^{T} x_t'\mathbf{S}^{-1}x_t\right](\gamma - g^*) + \mathbb{E}\left[x_t'\mathbf{S}^{-1}\mathbf{U}_{t,H|t}\right]$ *where* $\mathbb{E}\left[x_t'\mathbf{S}^{-1}\mathbf{U}_{t,H|t}\right] = 0$. *Let* $m_d \equiv \mathbb{E}\left[T^{-1}\sum_{t=1}^{T} x_t'\mathbf{S}^{-1}x_t\right] > 0$ *if* $\mathbf{S}$ *and thus* $\mathbf{S}^{-1}$ *is positive definite. Then we have*

- *the null hypothesis* $\mathbb{E}\left[\bar{d}_{(0,\mathbf{S})}\right] = m_d\gamma = 0$ *corresponds to the null hypothesis* $\gamma = 0$, *and*
- *the alternative hypothesis* $\left(\mathbb{E}\left[\bar{d}_{(0,\mathbf{S})}\right]\right)^2 \geq \delta^* \equiv m_d^2\delta > 0$ *corresponds to the alternative hypothesis* $\gamma^2 \geq \delta$.

Theorem 2.2 *The test statistic in* (23) *with* $\mathbf{S} = \hat{\Sigma}_{(0)}$ *from* (16) *is invariant to linear transformations of the forecasts given by the non-singular matrix* $\mathbf{M}$ *with dimension* $K \times K$.

Proof. It follows from (22) that $\mathbf{Q}^*_{(0)}$ is unaltered by such linear transformations if $d_{(0,\hat{\Sigma}_{(0)}),t}$ and $\bar{d}_{(0,\hat{\Sigma}_{(0)})}$ are unaltered by such linear transformations. Thus, it is sufficient to show that $d_{(0,\hat{\Sigma}_{(0)}),t}$ (and therefore $\bar{d}_{(0,\hat{\Sigma}_{(0)})}$) is invariant to linear transformations of the forecasts. If we consider the linear transformation given by the matrix $\mathbf{M}$, then $\mathbf{M}\mathbf{E}^j_{t,H|t}$ with $j = A, B$ is the transformed system forecasts of forecaster j. From (16) it follows that the covariance matrix corresponding to the transformed forecasts is $\mathbf{M}\hat{\Sigma}_{(0)}\mathbf{M}'$. When these terms are substituted into the definition of $d_{(0,\mathbf{S}),t}$ in (20) with $\mathbf{S} = \hat{\Sigma}_{(0)}$ for the the untransformed series, it follows that the transformation matrix $\mathbf{M}$ cancels out and, accordingly, for the transformed series, we have shown that $d_{(0,\hat{\Sigma}_{(0)}),t}$ is invariant to linear transformations of the forecasts. ■

Although the distribution of the test statistic converges to a normal distribution, the deviation between the actual distribution of the test statistic and the normal distribution might be large in small samples. Therefore, we follow Harvey et al. (1997) and compare the test statistics with a t-distribution. In addition we apply the correction factor w_0 to the test statistics. Harvey et al. (1997) showed that the test statistics in both (21) and (23) have a distribution close to a t-distribution with $T-1$ degrees of freedom in the univariate case. In the multivariate case of a system of forecasts with a vector of K forecasts, we apply the t-distribution with $TK-1$ degrees of freedom.

2.5 Size and power of the tests

Hungnes (2018) presents a Monte Carlo simulation and investigates the size and power of an encompassing test. The results from the simple encompassing test also apply to the test presented here. Hence, only the most important findings there are summed up here.

Regarding the size of the test, the most important result is that the size distortion is smaller when using the test statistic in (23) with the GLS estimates. Hence, this is the test statistic we apply in our

empirical application where we compare forecasts made by Norges Bank and Statistics Norway in the next section.

Regarding the power of the test, the Monte Carlo example in Hungnes (2018) indicates that the power increases with the dimension K in addition to the sample size T.

3 Comparison with Norges Bank

Since the 1st quarter of 1990 Statistics Norway has, with few exceptions, published forecasts every quarter for many variables for the year for in which the forecasts were made as well as the following year. Among these variables are Mainland GDP,[5] CPI and the unemployment rate. The forecasts from Norges Bank (the central bank of Norway) are available from the 4th quarter of 1992 with some exceptions mentioned below. In this period Statistics Norway has published forecasts quarterly, with the exception of the 3rd quarter in 2013 Statistics Norway. In this analysis, we have set this forecast equal to the previously published forecast, i.e., the forecast from the 2nd quarter that year.[6]

Although a quarterly model is used when making the forecasts, they are only published for the calendar year. Therefore, in testing the forecasts from Statistics Norway, we only consider forecasts of annual values.

The published numbers for CPI are never revised and the published numbers for the unemployment rate are usually never revised. However, the published numbers for variables from the National Accounts, such as Mainland GDP, can be revised in many quarters until they are fixed. These 'fixed' numbers may also be revised due to benchmark revisions in the National Accounts (where definitions are changed). In the analysis undertaken here, the first published number for Mainland GDP growth is used as the outcome of that variable. See Helliesen et al. (2020) for an analysis of the revision process in preliminary Norwegian national accounts.

In this analysis Mainland GDP and CPI are measured on the logarithmic scale. We do this for two reasons. First, the forecast errors are then approximately a measure of the percentage error (when disregarding the scaling factor of 100). For variables that are increasing over time, such as Mainland GPD and the CPI index, the assumption that the variance of the forecast error for variables measured in levels is time independent will imply that the variance of the forecast error measured in percentages will decrease over time, which we find unlikely. Second, with log-transformed data, there is a linear transformation from log-levels to growth rates, where the latter is approximately the percentage growth from one period to the next. For the unemployment rate we do not use a logarithmic transformation; this variable is already measured in percentages and we find it more intuitive to consider changes in this variable in percentage points than in percent.

Table 1 compares the forecasts made by Statistics Norway with the forecasts made by Norges Bank. Forecasts from Norges Bank made in the 1st quarter of the year is available from 1996, and forecasts from Norges Bank made in the 2nd quarter are available from 1993. These forecasts are published about the same time as the forecasts from Statistics Norway are published. Hence, these forecasts can (without problems) be compared.

From 2001 to 2012 Norges Bank only published its forecasts 3 times a year. The last forecasts in these 12 years were published at the end of October. When comparing the forecast from the third quarter

[5] Mainland GDP consists of all domestic production activity except exploration for crude oil and natural gas, pipeline transport and ocean transport. The meaning of the term was changed as a part of the benchmark revision of the national accounts in 2014. Before this, service activities incidental to oil and gas were also excluded from Mainland GDP.

[6] Due to the onset of the financial crises, Statistics Norway published an extra forecast in mid-October 2008. This extra forecast is not included in the current analysis.

from Statistics Norway and Norges Bank, we use the forecasts made in the beginning of September for Statistics Norway (except for 2013 where we use the forecasts made in the second quarter, as mentioned above). For Norges Bank, we use forecasts made about the same time in the years 1996-2000 and 2013-2014. For the years 2001-2012 we use the forecasts published in end-October, which implies that Norges Bank has about 1.5 month of information advantages in 12 out of 19 years for which we compare forecasts from the "3rd" quarter.[7]

Norges Bank has published forecasts in the 4th quarter since 1992. In the years 2001-2012 these forecasts were made in late October, while in the other years they are from December. The forecasts from Statistics Norway are from the beginning of December in all the years we compare. Hence, here Statistics Norway has an information advantage in 12 out of 23 years.

In the upper part of Table 1 (*'All variables, both horizons'*) we compare the forecast for all three variables in both the current and the next year. In the first line, two numbers are reported for the forecasts made in each quarter; the estimate of α based on FIML and its standard error. The standard error of the estimated α derived using the estimated α is based on (21) and adjusted for the relationship between α and γ;

$$\sqrt{\widehat{Var\left(\hat{\alpha}_{(\mathbf{S})}\right)}} = \frac{1}{2}\sqrt{\widehat{Var\left(\hat{\gamma}_{(\mathbf{S})}\right)}} = \frac{1}{2}\sqrt{\frac{\mathbf{Q}_{(\hat{\gamma}_{(\mathbf{S})},\mathbf{S})}}{Tw_0}}, \tag{24}$$

with $\mathbf{S} = \hat{\Sigma}_{(g)}$ where $g = \hat{\gamma}_{(\mathbf{S})}$. We report the estimated α instead of the estimated γ since the former has a more intuitive interpretation: the estimated α is the estimated optimal weight of the forecasts by Norges Bank whereas $1-\alpha$ is the corresponding optimal weight of Statistics Norway. For forecasts made in the 1st quarter (of the year) the estimate is about 0.487, indicating that the forecasts made by Statistics Norway are slightly better than the forecasts made by Norges Bank. However, seen in relation to the relatively high standard error (0.080), this indicates that the estimate is not significantly different from 0.5.

To derive the standard error of the estimated α under the null hypothesis ($\gamma = 0 \Leftrightarrow \alpha = \frac{1}{2}$), we use (24) with $\mathbf{Q}^*_{(0,\mathbf{S})}$ from (22) with $\mathbf{S} = \hat{\Sigma}_{(0)}$. This standard error, which is reported in round brackets in the next line of the cell in Table 1, is slightly higher than the standard error using the estimated α from the line above. The next two figures reported in the cell are the (absolute) t-value of the hypothesis test and the corresponding p-value in square brackets. As can be seen, we cannot reject the null hypothesis that the forecasts made by Statistics Norway and Norges Bank are equally good.

When comparing the forecasts made by Statistics Norway and Norges Bank made in the 2nd quarter of the year, we find that the estimate is not significantly different from a half (the estimate is 0.521 with a standard error given by 0.101). Hence, also for the forecasts made in the 2nd quarter by Statistics Norway and Norges Bank we cannot reject that they are equally good.

For the forecasts made in the 3rd quarter, the estimated α is 0.751 and — by both measures of the standard error — it is clearly significantly different from 0.5. This estimate implies that for forecasts made in this quarter we could compute an optimal forecast with a weight of about 75 percent for the forecast made by Norges Bank and 25 percent for the forecast made by Statistics Norway. However, here Norges Bank has an information advantage of about 1.5 month in a substantial proportion of the years the forecasts were made. Hence, we would expect Norges Bank to do better than Statistics Norway for forecasts made in the 3rd quarter of the year.

For the forecasts made in the 4th quarter, the estimated α is about 0.367, indicating that the forecasts

[7] This differs form Bjørnland et al. (2012) who use the forecasts from the 2nd quarter when they compare the forecasts from Norges Bank with a system of averaging models, SAM.

Table 1: Equal predictability tests between forecasts from Statistics Norway and Norges Bank

	Q1 (1996-2014)	Q2 (1993-2014)	Q3 (1996-2014)	Q4 (1992-2014)
All variables, both horizons				
$\hat{\alpha}$ (FIML)	0.487 (0.080)	0.521 (0.101)	0.751 (0.065)	0.367 (0.079)
$H_0 : \alpha = \frac{1}{2}$	(0.088) 0.14 [0.889]	(0.106) 0.20 [0.848]	(0.077) 3.27 [0.007]**	(0.087) 1.53 [0.145]
All variables, nowcasting				
$\hat{\alpha}$ (FIML)	0.632 (0.144)	0.522 (0.161)	0.739 (0.084)	0.219 (0.103)
$H_0 : \alpha = \frac{1}{2}$	(0.142) 0.93 [0.366]	(0.164) 0.13 [0.895]	(0.115) 2.08 [0.055]	(0.132) 2.12 [0.048]*
Mainland GDP, both horizons				
$\hat{\alpha}$ (FIML)	0.791 (0.388)	0.812 (0.263)	1.287 (0.242)	0.585 (0.125)
$H_0 : \alpha = \frac{1}{2}$	(0.423) 0.69 [0.502]	(0.287) 1.09 [0.290]	(0.364) 2.16 [0.046]*	(0.145) 0.59 [0.563]
CPI, both horizons				
$\hat{\alpha}$ (FIML)	0.561 (0.136)	0.650 (0.191)	0.646 (0.079)	0.217 (0.130)
$H_0 : \alpha = \frac{1}{2}$	(0.146) 0.41 [0.684]	(0.174) 0.86 [0.400]	(0.100) 1.45 [0.166]	(0.177) 1.60 [0.125]
Unemployment rate, both horizons				
$\hat{\alpha}$ (FIML)	0.440 (0.156)	0.466 (0.226)	0.401 (0.168)	0.232 (0.231)
$H_0 : \alpha = \frac{1}{2}$	(0.185) 0.33 [0.748]	(0.244) 0.14 [0.892]	(0.173) 0.57 [0.577]	(0.261) 1.02 [0.318]

Note: Q1 – Q4 indicate the quarter (of the year) the forecast is made. For $\hat{\alpha}$ the FIML estimate with its standard errors (implicitly derived such that the t-value is given in (21) with $g = \hat{\gamma}$) is reported. For the hypothesis tests the table reports the GLS estimate, its standard errors (derived under the null hypothesis $\gamma = 0 \Leftrightarrow \alpha = \frac{1}{2}$ with $Q^*_{(0)}$ from (22)) in round brackets, the absolute t-value, and the corresponding p-value to the t-value in the square brackets. One asterisk denotes that the p-value is less than 0.05 and two asterisks denote that the p-value is less than 0.01. For the tests, we apply $\tau_H = H$.

made by Statistics Norway in this quarter are better than the forecasts made by Norges Bank. However, the forecasts made by Statistics Norway are not significantly better than the ones made by Norges Bank, despite that Statistics Norway has had an information advantage for many of the years.

In the next part of the table ('*All variables, nowcasting*') we only consider the forecast for the current year for the three variables. The results for the forecasts made in the 1st and 2nd quarter are similar to those obtained concerning both horizons. When only considering nowcasting we do not find that the forecasts made by Norges Bank in the 3rd quarter are significantly better than the forecasts made by Statistics Norway, even though the point estimate is almost identical to the one obtained when considering both forecasting horizons. For nowcasts made in the 4th quarter we see that the forecasts made by Statistics Norway are significantly better than the ones made by Norges Bank. However, this must be contributed to the fact that Statistics Norway has an information advantage in half of the years and this information advantage is much more important when only forecasts for the current year are considered.

At the bottom of the table, the forecasts for Mainland GDP, CPI, and the unemployment rate are considered separately to see which institution does the best to forecast these variables. For Mainland GDP, Norges Bank makes the best forecasts based on the estimated α. But only for Mainland GDP are the forecasts by Norges Bank made in the 3rd quarter significantly better than the forecasts made by Statistics Norway at the 5 percent significance level.

4 Conclusions

This paper presents a test of equal predictability of multi-step-ahead system forecasts. The test, which is a multivariate version of the Diebold and Mariano (1995) test, is invariant to linear transformations of

the system forecasts. The test is used to compare the forecasts made by Statistics Norway with forecasts made by Norges Bank. We find that when the forecasts are made approximately at the same time, they are equally good in the sense that none of them are significantly better than the other. However, when one of the institutions has an information advantage, the forecast made by the one that possesses the advantage can be significantly better than the forecasts made by the other institution.

In the present paper, we consider testing for equal predictability of only two system forecasts. In the case of univariate forecasts, Mariano and Preve (2012) and West (2006, p. 105) consider an extension of the Diebold and Mariano (1995) test to test for equal predictability between three or more univariate forecasts. Mariano and Preve (2012) show that this test is invariant concerning the ordering of the forecasts being compared. The test proposed by Mariano and Preve (2012) can easily be applied to compare multiple system forecasts when each pair of system forecasts is compared as in the present paper.

The test in Mariano and Preve (2012) only considers equal predictability among all forecasts. For only two alternative forecasts this is sufficient for ranking the forecasts, as with a rejection of the null hypothesis of equal predictability it follows that one forecast is preferred over the other forecast. White (2000) provides a test for testing if there exist forecasts from at least one forecast procedure that are more accurate than a baseline forecast. This test is improved upon in Hansen (2005), and Romano and Wolf (2005) extend the procedure to identify all forecasts from forecasting procedures that are significantly better than the baseline forecast.

References

Bjørnland, H. C., Gerdrup, K. R., Jore, A. S., Smith, C., and Thorsrud, L. A. (2012). Does Forecast Combination Improve Norges Bank Inflation Forecasts? *Oxford Bulletin of Economics and Statistics*, 74(2):163–179.

Bjørnland, H. C., Ravazzolo, F., and Thorsrud, L. A. (2017). Forecasting GDP with global components: This time is different. *International Journal of Forecasting*, 33(1):153–173.

Bjørnstad, J. F. (1990). Predictive Likelihood: A Review. *Statistical Science*, 5(2):242–254.

Capistrán, C. (2006). On comparing multi-horizon forecasts. *Economics Letters*, 93(2):176–181.

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

Clark, T. E. and McCracken, M. W. (2013). Advances in Forecast Evaluation. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2B, chapter 20, pages 1107–1201. Elsevier B.V.

Clark, T. E. and McCracken, M. W. (2015). Nested forecast model comparisons: A new approach to testing equal accuracy. *Journal of Econometrics*, 186(1):160–177.

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

Clements, M. P. and Hendry, D. F. (1998). *Forecasting Economic Time Series*. Cambridge University Press.

Diebold, F. X. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests. *Journal of Business & Economic Statistics*, 33(1):1–24.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics Economic Statistics*, 13(3):253–263.

El-Shagi, M., Giesen, S., and Jung, A. (2016). Revisiting the relative forecast performances of Fed staff and private forecasters: A dynamic approach. *International Journal of Forecasting*, 32(2):313–323.

Engle, R. F. (1993). On the limitations of comparing mean square forecast errors: Comment. *Journal of Forecasting*, 12(8):642–644.

Ericsson, N. R. (2008). Comment on 'Economic Forecasting in a Changing World' (by Michael Clements and David Hendry). *Capitalism and Society*, 3(2):1–16.

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

Granger, C. W. J. and Newbold, P. (1986). *Forecasting economic time series*. Academic Press, 2nd edition.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business and Economic Statistics*, 23(4):365–380.

Harvey, D. I., Leybourne, S. J., and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291.

Harvey, D. I., Leybourne, S. J., and Newbold, P. (1998). Tests for Forecast Encompassing. *Journal of Business & Economic Statistics*, 16(2):254–259.

Harvey, D. I., Leybourne, S. J., and Whitehouse, E. J. (2017). Forecast evaluation tests and negative long-run variance estimates in small samples. *International Journal of Forecasting*, 33(4):833–847.

Harvey, D. I. and Newbold, P. (2000). Tests for multiple forecast encompassing. *Journal of Applied Econometrics*, 15(5):471–482.

Helliesen, M. K., Hungnes, H., and Skjerpen, T. (2020). Revisions in the Norwegian National Accounts accuracy , unbiasedness and efficiency in preliminary figures. *Discussion Papers xxx, Statistics Norway*.

Hendry, D. F. and Martinez, A. B. (2017). Evaluating Multi-Step System Forecasts with Relatively Few Forecast-Error Observations. *International Journal of Forecasting*, 33(784):359–372.

Hinkley, D. (1979). Predictive likelihood. *The Annals of StatisticsStatistics*, 7(4):718–728.

Howrey, P. E. (1993). On the limitations of comparing mean square forecast errors: Comment. *Journal of Forecasting*, 12(8):652–654.

Hungnes, H. (2018). Encompassing tests for evaluating multi-step system forecasts invariant to linear transformations. *Statistics Norway, Discussion Papers*, 871.

Ibragimov, I. (1959). Some limit theorems for stochastic processes stationary in the strict sense. *Dokl. Akad. Nauk SSSR 125*, pages 711–714.

Ibragimov, I. A. (1962). Some Limit Theorems for Stationary Processes. *Theory of Probability & Its Applications*, 7(4):349–382.

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

Jungmittag, A. (2016). Combination of Forecasts across Estimation Windows: An Application to Air Travel Demand. *Journal of Forecasting*, 35(4):373–380.

Kock, A. and Teräsvirta, T. (2016). Forecasting Macroeconomic Variables using Neural Network Models and Three Automated Model Selection Techniques. *Econometric Reviews*, 35(8-10):1753–1779.

Kolsrud, D. (2007). Time-simultaneous prediction band for a time series. *Journal of Forecasting*, 26(3):171–188.

Kolsrud, D. (2015). A Time-Simultaneous Prediction Box for a Multivariate Time Series. *Journal of Forecasting*, 34(8):675–693.

Lawrence, M., Goodwin, P., O'Connor, M., and Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, 22(3):493–518.

Mariano, R. S. and Preve, D. (2012). Statistical tests for multiple forecast comparison. *Journal of Econometrics*, 169(1):123–130.

Martinez, A. B. (2017). Testing for Differences in Path Forecast Accuracy: Forecast-Error Dynamics Matter. *Working Paper, Federal Reserve Bank of Cleveland*, 17-2017.

Mathiasen, P. . E. . (1979). Prediction Functions. *Scandinavian Journal of Statistics*, 6(1):1–21.

McCracken, M. W. (2019). Tests of conditional predictive ability: A comment. *Federal reserve bank of St. Louis*.

Oberhofer, W. and Kmenta, J. (1974). A general procedure for obtaining maximum likelihood estimates in generalized regression models. *Econometrica*, 42(3):579–590.

Patton, A. J. (2015). Comparing Predictive Accuracy Twenty Years Later: A Personal Perspective of the Use and Abuse of Diebold-Mariano Tests: Comment. *Journal of Business & Economic Statistics*, 33(1):22–24.

Pesaran, M. H. and Skouras, S. (2002). Decision-Based Methods for Forecast Evaluation. In Clements, M. P. and Hendry, D. F., editors, *A Companion to Economic Forecasting*, chapter 11, pages 241–267. Blackwell Publishing Ltd.

Quaedvlieg, R. (2019). Multi-Horizon Forecast Comparison. *Journal of Business & Economic Statistics*, (forthcoming):1–14.

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

Rosenblatt, M. (1956). A CENTRAL LIMIT THEOREM AND A STRONG MIXING CONDITION. *Proceedings of the National Academy of Sciences*, 42(1):43–47.

Sinclair, T. M. and Stekler, H. O. (2013). Examining the quality of early GDP component estimates. *International Journal of Forecasting*, 29(4):736–750.

Sinclair, T. M., Stekler, H. O., and Carnow, W. (2012). A new approach for evaluating economic forecasts. *Economics Bulletin*, 32(3):2332–2342.

Sinclair, T. M., Stekler, H. O., and Carnow, W. (2015). Evaluating a vector of the Fed's forecasts. *International Journal of Forecasting*, 31(1):157–164.

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

West, K. D. (2006). Forecast Evaluation. In *Handbook of Economic Forecasting*, volume 1, chapter 3, pages 99–134.

White, H. (2000). A Reality Check for Data Snooping. *Econometrica*, 68(5):1097–1126.

Williams, E. J. and Kloot, N. H. (1953). Interpolation in a series of correlated observations. *Australian Journal of Applied Science*, 4(1):1–17.