Hungnes, Håvard

Working Paper

Encompassing tests for evaluating multi-step system forecasts invariant to linear transformations

Discussion Papers, No. 871

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

Suggested Citation: Hungnes, Håvard (2018) : Encompassing tests for evaluating multi-step system forecasts invariant to linear transformations, Discussion Papers, No. 871, Statistics Norway, Research Department, Oslo

This Version is available at:
https://hdl.handle.net/10419/192853

Discussion Papers
Statistics Norway
Research department

No. 871
February 2018

Statistics Norway

Statistisk sentralbyrå

Håvard Hungnes

Encompassing tests for evaluating multi-step system forecasts invariant to linear transformations

Discussion Papers No. 871, February 2018
Statistics Norway, Research Department

Håvard Hungnes

Encompassing tests for evaluating multi-step system forecasts invariant to linear transformations

Abstract:
The paper suggests two encompassing tests for evaluating multi-step system forecasts invariant to linear transformations. An invariant measure for forecast accuracy is necessary as the conclusions otherwise can depend on how the forecasts are reported (e.g., as in level or growth rates). Therefore, a measure based on the prediction likelihood of the forecast for all variables at all horizons is used. Both tests are based on a generalization of the encompassing test for univariate forecasts where potential heteroscedasticity and autocorrelation in the forecasts are considered. The tests are used in evaluating quarterly multi-step system forecasts made by Statistics Norway.

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

JEL classification: C32; C53.

Acknowledgements: Thanks to Pål Boug, Jennifer Castle, John Muellbauer, and Terje Skjerpen 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

I 1993 påviste to forskere, Michael P. Clements og David F. Hendry, at evaluering av prognoser for enkeltvariabler for hver prognosehorisont ikke er invariant for lineære transformasjoner av prognosene. 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 flere perioder framover, mens prognosene for den andre modellen er best når man vurderer oljeprisveksten. Forskerne foreslo derfor et mål for hele systemet av prognoser som er invariant for slike lineære transformasjoner. Men 25 år senere finnes det relativt lite forskning som evaluere nøyaktigheten av hele systemet på tvers av alle prognosehorisontene.

I denne artikkelen utleder jeg såkalte omslutnings-tester (encompassing-tester) for å sammenligne to sett med prognoser. Et sett med prognoser omslutter et annet sett med prognoser hvis sistnevnte ikke inneholder ytterligere informasjon, det vil si at det første settet av prognoser ikke kan forbedres ved å utnytte informasjonen i sistnevnte sett av prognoser.

De utledede testene brukes til å undersøke om SSBs prognoser omslutter prognoser basert på en tilfeldig gang (random walk) prosess, noe resultatene viser at de gjør. Testene viser også at prognoser offentliggjort i et bestemt kvartal omslutter prognosene publisert i det foregående kvartalet. I artikkelen undersøker jeg prognosene for BNP, KPI og arbeidsledighetsraten for inneværende og neste år.

1 Introduction

Clements and Hendry (1993) showed 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 forecasts based on one of the models are best when considering the oil price in levels, but that the forecasts of another model are best when considering the oil price growth. Clements and Hendry (1993) suggested a measure 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 on evaluating the accuracy of the whole system across all forecasting horizons."

The following two examples illustrate the importance of considering forecasts of "the whole system" (Example 1) and forecasts "across all forecasting horizons" (Example 2):

Example 1 ("the whole system") *Suppose you forecasted private consumption and income to grow by 2 percent in period t. When period t is finished, the National Accounts numbers show that both private consumption and income increased by 3 percent. You missed both private consumption and private income by 1 percentage point. However, your implied forecast on the savings ratio was spot on!*

Example 2 ("across all forecasting horizons") *Suppose you forecast the consumer price index (CPI) to increase by 2 percent in both years t and $t+1$. Then it turns out that CPI grew by 1 percent in year t and 3 percent in year $t+1$. Measured by the annual CPI-growth, you missed by 1 percentage point each year. However, your implied prediction of the CPI-level in year $t+1$ was correct!*

Usually, measures for forecast accuracy only consider forecast for one variable at one forecasting horizon. Measures such as mean absolute forecast errors and mean square forecast errors (or variants of these) are usually applied. For measuring the accuracy of a system of forecasts variants of these individual measures can be applied. One example is the mean (or sum) of the mean square forecast errors. Kolsrud (2015) suggests applying a prediction box covering a pre-given fraction of the forecast errors as a measure of the forecast accuracy. Unfortunately, none of these measures are invariant to the transformations in Example 1 and Example 2. However, as suggested by Clements and Hendry (1993), using the determinant

of the covariance matrix of the whole system of forecasts yields a measure that is invariant to scale-preserving linear transformations of the forecasts. This measure is equivalent to the predictive likelihood, see Bjørnstad (1990).

In this paper, I consider encompassing tests for comparing two sets of forecasts. One set of forecasts encompasses another set of forecasts if the latter does not include any additional information, i.e., the former set of forecasts cannot be improved by knowing the latter set of forecasts. Granger and Newbold (1973) defined the preferred forecasts as "computationally efficient" with respect to the latter. Chong and Hendry (1986) and Clements and Hendry (1993) apply the formulation that the preferred forecasts "encompass" the competing forecasts.

Harvey et al. (1998) consider a test for forecast encompassing when there exist two forecasts of the same variable and develop a test with small size distortion. In the present paper, this test is modified such that it can be used to test if forecasts of one vector of variables over a range of forecasting horizons encompass another vector of forecasts.

The tests are used to investigate if the forecasts made by Statistics Norway encompass forecasts based on a random walk model, which the results show they do. The tests also show that forecasts made in one particular quarter of the year encompass the forecasts made in the previous quarter. In the analysis, I investigate the forecasts of GDP, CPI and the unemployment rate for the current and the next year jointly.

The rest of the paper is organized as follows: In Section 2 the theoretical background for the encompassing tests as well as the proposed tests are presented. In Section 3 the power and size of the tests are investigated. In Section 4 the proposed encompassing tests are applied to examine the forecasts made by Statistics Norway. Section 5 concludes.

2 Theory

2.1 Measures of forecast accuracy and ranking of forecasts

Let $y^i_{t+h|t}$ be the forecast of variable i in period $t+h$ made in period t. In the present paper I assume that the value of y in period t is not known in period t; hence forecasts for the current period (i.e., nowcasting) can be made and is denoted $y^i_{t|t}$. The prediction error of the forecast of variable i in period $t+h$ made in period t is defined as $e^i_{t+h|t} \equiv y^i_{t+h} - y^i_{t+h|t}$, where y^i_{t+h} is

the outcome of variable i in period $t+h$.

The observed Mean Square Forecast Error (MSFE) is given by

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

which expresses the mean square forecast error of variable i forecasted h periods for forecasts made in T consecutive periods. The MSFE (or the root of MSFE) is a widely used measure for the accuracy of forecast and ranking of forecasts 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 ranking forecasts when $h > 0$ depends on how the forecasts are measured, see Clements and Hendry (1993). Only in the case of $h = 0$, i.e., the forecasts for the current period, rankings based on the observed (univariate) MSFE are invariant of linear transformations of the forecasts, see Clements and Hendry (1993, 1998).

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

A matrix version of the observable MSFE would then be

$$V^{i}_{H} = T^{-1}\sum_{t=1}^{T}\mathbf{e}^{i}_{t,H|t}\mathbf{e}^{i\prime}_{t,H|t}, \tag{2}$$

which is here denoted the MSFE Matrix (or MSFEM). This matrix is of dimension $(H+1)\times(H+1)$ and it is not obvious how to rank forecasts based on this measure.

One approach for ranking forecasts could be the trace of MSFEM, which is the sum of the mean square forecast errors. Ranking based on this criterion is an often used approach, as also noted in Christoffersen and Diebold (1998) and Hendry and Martinez (2017). A recent example is Bjørnland et al. (2017), who apply the square roots of the mean of individual squared forecast errors, which is just a simple transformation of the trace of MSFEM that does not alter

the ranking. However, as shown by Clements and Hendry (1993), this measure is not invariant to linear transformations of the forecasts and, then, linear transformations of the forecast errors. Let $(V_H^i)^A$ and $V_H^i)^B$ be the MSFEM of two different forecasting models. Furthermore, let $\mathcal{M}(V_H^i)^A\mathcal{M}'$ and $\mathcal{M}(V_H^i)^B\mathcal{M}'$ be the MSFEM of the linear transformed versions of the two forecasts where $\mathcal{M}$ is an $(H+1)\times(H+1)$ full rank matrix expressing the linear transformation.[1] If $trace\left((V_H^i)^A\right) < trace\left((V_H^i)^B\right)$, it does not follow that $trace\left(\mathcal{M}(V_H^i)^A\mathcal{M}'\right) < trace\left(\mathcal{M}(V_H^i)^B\mathcal{M}'\right)$. This implies that a linear transformation, such as considering differences instead of levels, can alter the ranking of two models.

Clements and Hendry (1993) suggest using the determinant of MSFEM. The ranking can then be based on this measure. If $|(V_H^i)^A| < |(V_H^i)^B|$, then the forecast based on model A with MSFE given by $(V_H^i)^A$ is ranked over (i.e., preferred over) the forecast of model B with MSFE given by $(V_H^i)^B$. This measure is invariant to linear transformations of the forecasts, i.e. $|(V_H^i)^A| < |(V_H^i)^B|$ implies $|\mathcal{M}(V_H^i)^A\mathcal{M}'| < |\mathcal{M}(V_H^i)^B\mathcal{M}'|$.[2] Furthermore, if $(V_H^i)^B - (V_H^i)^A \succ 0$, (i.e., that the difference between the two MSFEMs is positive definite, see, e.g. Dhrymes, 1984, prop. 66) then it follows that $|(V_H^i)^A| < |(V_H^i)^B|$.

The lack of invariance for MSFE is also present across different variables. To present a measure of forecast errors that is also invariant to linear transformations of different variables, let $\mathbf{Y}_{t,H|t}$ be a vector of $\mathbf{y}_{t,H|t}^i$ for all variables, $i = 1, 2, \ldots, N$, such that $\mathbf{Y}_{t,H|t} = \left(\mathbf{y}_{t,H|t}^{1\prime}, \mathbf{y}_{t,H|t}^{2\prime}, \ldots, \mathbf{y}_{t,H|t}^{N\prime}\right)'$ is a vector with $M \equiv N(H+1)$ elements. The prediction error of $\mathbf{Y}_{t,H|t}$ is then given by $\mathbf{E}_{t,H|t} \equiv \mathbf{Y}_{t,H} - \mathbf{Y}_{t,H|t}$, where $\mathbf{Y}_{t,H} = \left(\mathbf{y}_{t,H}^{1\prime}, \mathbf{y}_{t,H}^{2\prime}, \ldots, \mathbf{y}_{t,H}^{N\prime}\right)'$, such that $\mathbf{E}_{t,H|t} = \left(\mathbf{e}_{t,H|t}^{1\prime}, \mathbf{e}_{t,H|t}^{2\prime}, \ldots, \mathbf{e}_{t,H|t}^{N\prime}\right)'$. The observable MSFEM for this forecast system is

$$V_H = T^{-1}\sum_{t=1}^{T}\mathbf{E}_{t,H|t}\mathbf{E}_{t,H|t}', \tag{3}$$

which is an $M \times M$ matrix. As above, the determinant of this matrix is an invariant measure for ranking of forecasts.

[1] Clements and Hendry (1993) also assume $|\mathcal{M}| = 1$, which can be interpreted as a scale-preserving transformation. Then $trace(\mathcal{M}(V_H^i)^A\mathcal{M}') \neq trace((V_H^i)^A)$ shows the lack of invariance. However, since two forecasts are compared, we do not need M to have this property; see also Schmidt (1993).

[2] The implication follows from $|\mathcal{M}\ (V_H^i)^j\mathcal{M}'| = |\mathcal{M}|^2|(\hat{V}_H^i)^j|$ for $j = A, B$.

2.2 Encompassing

Bates and Granger (1969) and Chong and Hendry (1986) suggest an encompassing test that can be used to test if one forecast is inferior to another, i.e., it contains no additional information. Hence, this is a stronger criterion than that one forecast is ranked better than another forecast based on the determinant of their MSFEM.

Consider two different forecasts of variable i in period $t+h$ made in period t; denoted $\left(y^i_{t+h|t}\right)^A$ and $\left(y^i_{t+h|t}\right)^B$. Then, consider the "composite artificial model"

$$y^i_{t+h} = (1-\alpha)\left(y^i_{t+h|t}\right)^A + \alpha\left(y^i_{t+h|t}\right)^B + u^i_{t+h|t}, \tag{4}$$

which is a weighted average of the two forecasts with the weight α and error term $u^i_{t+h|t}$.[3] The forecast-encompassing test of the hypothesis $\alpha = 0$ investigates whether forecast A contains all information (i.e., there is no additional information in forecast B). Likewise, the hypothesis $\alpha = 1$ implies that forecast B contains all information. Any other outcome implies that neither model encompasses the other. Unfortunately, in the same manner as MSFE, this test is not invariant to linear transformations of the forecasted variable, as the illustration in Ericsson (2008) shows. Hence, the forecasts based on one model (say A) could be considered better than the forecasts of another model (B) when the forecasts are measured in levels, but forecasts from B could be preferred over forecasts from A when measured in differences.

A generalization of this test could be to consider the vector version where all forecast horizons up to H for all variables are considered jointly, i.e.

$$\mathbf{Y}_{t,H} = (I_M - \Gamma)\left(\mathbf{Y}_{t,H|t}\right)^A + \Gamma\left(\mathbf{Y}_{t,H|t}\right)^B + \mathbf{U}_{t+H|t}, \tag{5}$$

where the error term is $\mathbf{U}_{t+H|t} = \left(\mathbf{u}^{1\prime}_{t,H|t}, \mathbf{u}^{2\prime}_{t,H|t}, \ldots, \mathbf{u}^{N\prime}_{t,H|t}\right)'$ with $\mathbf{u}^i_{t,H|t} = \left(u^i_{t|t}, u^i_{t+1|t}, \ldots, u^i_{t+H|t}\right)'$.

The test for forecast vector A encompassing the composite model is then $\Gamma = \mathbf{0}$ and the test for forecast vector B encompassing the composite model is $\Gamma = I_M$. Otherwise, neither forecast encompasses the other.

A simplified (i.e., restricted) version of this test would be to consider the "composite arti-

[3] See Ericsson (1993) for a discussion of why this formulation for the encompassing test is preferred to more general formulations where the weights are not restricted to sum to unity and possibly an intercept is included.

ficial model" as a weighted average of the two forecasts. Then, $\Gamma = \alpha I_M$ where α is a scalar, i.e., Γ is a matrix where all the diagonal elements are equal to α and all other elements are zero. Then the "composite artificial model" becomes

$$\mathbf{Y}_{t,H} = (1-\alpha)\left(\mathbf{Y}_{t,H|t}\right)^A + \alpha\left(\mathbf{Y}_{t,H|t}\right)^B + \mathbf{U}_{t,H|t}. \tag{6}$$

Both (5) and (6) are used in Section 4 to test if one set of forecasts is encompassing another. Note that if we subtract $\left(\mathbf{Y}_{t,H}\right)^A$ on both sides of (6) we get the formulation

$$\mathbf{E}^A_{t,H|t} = \alpha\left[\left(\mathbf{E}_{t,H|t}\right)^A - \left(\mathbf{E}_{t,H|t}\right)^B\right] + \mathbf{U}_{t,H|t}, \tag{7}$$

which is usually applied for univariate encompassing tests, see e.g. Harvey et al. (1998). Furthermore, define $\mathbf{D}_{t,H|t} \equiv \left(\mathbf{E}_{t,H|t}\right)^A - \left(\mathbf{E}_{t,H|t}\right)^B$, which is the difference in the forecast errors between the two forecasts. It follows that $\mathbf{D}_{t,H|t}$ is also the difference between the two forecasts, i.e., $\mathbf{D}_{t,H|t} = \left(\mathbf{Y}_{t,H|t}\right)^B - \left(\mathbf{Y}_{t,H|t}\right)^A$.

The parameter α in (4) and (6) is usually estimated with OLS with the formulation in (7). The distribution of this estimator is non-standard, see e.g., Harvey et al. (1998) and Harvey and Newbold (2000) in the case of univariate forecasts. There are two important reasons why the distribution is non-standard. First, the distribution of the forecast errors can be non-normal. For example, Harvey et al. (1998) show that if two univariate forecasts errors (i.e., when $n = 1$ and $H = 0$) are generated by the bivariate Student's t-distribution — see Dunnett and Sobel (1954) — with 4 degrees of freedom, then for the nominal 5%-level of the t-test the true asymptotic size is 12.2%. Second, the forecast errors will be autocorrelated since the forecasting horizons overlap (i.e., with $H > 0$) each other.

2.3 The simplified encompassing test statistics

By defining the vectors $\mathbf{y} = \left(\mathbf{E}^{A\prime}_{1,H|1}, \mathbf{E}^{A\prime}_{2,H|2}, \ldots, \mathbf{E}^{A\prime}_{T,H|T}\right)'$ and $\mathbf{x} = \left(\mathbf{D}'_{1,H|1}, \mathbf{D}'_{2,H|2}, \ldots, \mathbf{D}'_{T,H|T}\right)'$, (7) can be formulated as

$$\mathbf{y} = \alpha\mathbf{x} + \varepsilon, \tag{8}$$

with $\varepsilon = \left(\mathbf{U}'_{1,H|1}, \mathbf{U}'_{2,H|2}, \ldots, \mathbf{U}'_{T,H|T}\right)'$.

When ignoring that the forecast errors given at different time periods can be correlated the conditional estimators for α and the covariance matrix for ε are given by (when ignoring possible degrees of freedom adjustments for the covariance matrix)

$$\hat{\alpha}_{(\hat{\Omega})} = \left(\frac{1}{TM} \mathbf{x}' \left(I_T \otimes \hat{\Omega}^{-1}_{(\hat{\alpha})} \right) \mathbf{x} \right)^{-1} \left(\frac{1}{TM} \mathbf{x}' \left(I_T \otimes \hat{\Omega}^{-1}_{(\hat{\alpha})} \right) \mathbf{y} \right), \tag{9}$$

$$\hat{\Omega}_{(\hat{\alpha})} = \frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{E}^{A}_{t,H|t} - \hat{\alpha}_{(\hat{\Omega})} \mathbf{D}_{t,H|t} \right) \left(\mathbf{E}^{A}_{t,H|t} - \hat{\alpha}_{(\hat{\Omega})} \mathbf{D}_{t,H|t} \right)', \tag{10}$$

where $\otimes$ indicates the Kronecker product and the subscript in parenthesis indicates that the estimates are a function of another estimate. The estimates in (9) and (10) can be obtained by an iterative procedure until convergence, and the final estimates will equal the ones obtained with full information maximum likelihood, see Oberhofer and Kmenta (1974).

The estimator of the variance of the estimator in (9), when considering that forecasts made at different time periods can be correlated is given by

$$\begin{aligned} &\widehat{Var\left(\hat{\alpha}_{(\hat{\Omega})}\right)}_{(a)} \\ =&\frac{1}{TM} \left[\frac{1}{TM} \mathbf{x}' \left(I_T \otimes \hat{\Omega}^{-1}_{(\hat{\alpha})} \right) \mathbf{x} \right]^{-2} \left[\frac{1}{TM} \mathbf{x}' \left(I_T \otimes \hat{\Omega}^{-1}_{(\hat{\alpha})} \right) \hat{\Sigma}_{(a)} \left(I_T \otimes \hat{\Omega}^{-1}_{(\hat{\alpha})} \right) \mathbf{x} \right], \end{aligned} \tag{11}$$

where the $TM \times TM$ matrix $\hat{\Sigma}_{(a)}$ is the covariance matrix of ε. The parameter a usually is set equal to the estimated α, but I include it as a separate parameter to allow it to take other values. It can be shown that with H steps ahead forecasts (including nowcasting), there will be autocorrelation up to order H and no autocorrelations above order H. The reason is that a forecast made in period $t+H$ will overlap with a forecast made in period t, as both sets of forecasts will involve forecasts of variables for period $t+H$. However, a forecast made in period $t+H+1$ will not overlap with a forecast made in period t, and hence the prediction 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 Diebold and Mariano (1995), Harvey et al. (1997, 1998), and Harvey and Newbold (2000), among others. Hence, the block element $(t, t+l)$ in

$\hat{\Sigma}_{(a)}$ is given by

$$\{\hat{\Sigma}_{(a)}\}_{t,t+l} = \begin{cases} \hat{\mathbf{U}}_{(a),t,H|t}\hat{\mathbf{U}}'_{(a),t+l,H|t+l} & \text{for } l = 0, 1, \ldots, H \\ \mathbf{0}_{M\times M} & \text{otherwise,} \end{cases} \tag{12}$$

where

$$\hat{\mathbf{U}}_{(a),t,H|t} = \mathbf{E}^{A}_{t,H|t} - a\mathbf{D}_{t,H|t}. \tag{13}$$

Hence, $\hat{\mathbf{U}}_{(0),t,H|t} = \mathbf{E}^{A}_{t,H|t}$ and $\hat{\mathbf{U}}_{(1),t,H|t} = \mathbf{E}^{B}_{t,H|t}$.

The last term in (11) can then be written as

$$\begin{aligned} \mathbf{Q}_{(a)} =& \frac{1}{TM}\mathbf{x}'\left(I_T \otimes \hat{\Omega}^{-1}_{(\hat{\alpha})}\right)\hat{\Sigma}_{(a)}\left(I_T \otimes \hat{\Omega}^{-1}_{(\alpha)}\right)\mathbf{x} \\ =& \frac{1}{T}\left[\sum_{t=1}^{T} d^2_{(a),t} + 2\sum_{l=1}^{H}\sum_{t=1}^{T-l} w_l d_{(a),t} d_{(a),t+l}\right], \end{aligned} \tag{14}$$

where

$$d_{(a),t} = \frac{1}{M^{1/2}}\mathbf{D}_{t,H|t}\hat{\Omega}^{-1}_{(\hat{\alpha})}\hat{\mathbf{U}}_{(a),t,H|t} \tag{15}$$

can be interpreted as a generalized measure of the difference between the two forecasts, and where w_l must be equal to unity for the equality in (14) to hold. However, to secure that the estimated variance in (11) is positive, Newey and West (1987) suggest using $w_l = 1 - \frac{l}{H+1}$ ($l = 1, \ldots, H$), as will be used here.

Based on this, two alternative t-tests can be formulated for testing the null hypothesis of $\alpha = 0$ or $\alpha = 1$: one where the estimated value of α is used in the expression for the variance in (11) and another where the value under the null hypothesis is used. Both t-statistics can be formulated as

$$\frac{\hat{\alpha} - \alpha_0}{\sqrt{\widehat{Var}\left(\hat{\alpha}\right)_{(a)}}} = (TM)^{1/2} w_0^{1/2} \bar{d}_{(\alpha_0)} \mathbf{Q}^{-1/2}_{(a)}, \tag{16}$$

where α_0 is the value of α under the null hypothesis, and $\bar{d}_{(a)}$ is the sample mean of (15). When the test statistic is computed for the estimated α in the expression of the variance, we have $a = \hat{\alpha}$; and when it is computed under the null, $a = 0$ or $a = 1$, depending on the null hypothesis. For the encompassing test of a forecast of only one variable, Harvey et al. (1997)

derive the correction factor $w_0 = T^{-1}\left[T - 1 - 2H + T^{-1}H(H+1)\right]$. This correction factor is because one in (14) divide by the sample size T instead of the number of autocovariances $T - l$ (where $l = 1, 2, \ldots, H$).

In the univariate case, i.e., when only forecast of one variable and not a vector is made, Harvey et al. (1998) and Harvey and Newbold (2000) show that both variants of the test in (16) have severe size distortions; the version with $a = \alpha$ over-rejects and the other under-rejects. Therefore Harvey et al. (1998) and Harvey and Newbold (2000) suggest a modification of the test where $\mathbf{Q}_{(a)}$ in (16) is replaced with

$$\mathbf{Q}^*_{(a)} = \frac{1}{T}\left[\sum_{t=1}^{T}\left(d_{(a),t} - \bar{d}_{(a)}\right)^2 + 2\sum_{l=1}^{H}\sum_{t=1}^{T-l} w_l \left(d_{(a),t} - \bar{d}_{(a)}\right)\left(d_{(a),t+l} - \bar{d}_{(a)}\right)\right]. \tag{17}$$

They show that this modified expression with a equal to α under the null hypothesis, the corresponding t-test has only small size distortions in the univariate case.

Henningsen and Hamann (2007) discuss the degrees of freedom for t-tests in systems. They distinguish between system-based t-tests and equation-based t-tests and write that in the literature sometimes the degrees of freedom of the entire system (total number of observations in all equations minus the total number of estimated coefficients) is applied. In other cases the degrees of freedom of the single equation (the number of observations in the equation minus the number of estimated coefficients in the equation) is used. In OxMetrics the results of t-tests are reported using the equation-based degrees of freedom, see (Doornik and Hendry, 2013, sec. 5.3). Here, the equation-based t-test is applied, as it has much smaller size distortions than the system-based t-test. An additional question concerns what we mean by the number of estimated coefficients (that we need to adjust for in the t-test). The way (8) is formulated, only one coefficient is estimated under the alternative and, hence, only one coefficient is included in each equation. This indicates that we should use $T - 1$ degrees of freedom in the test. However, to estimate this coefficient, we start out with a system with M regressors in each equation, but with restrictions both within each equation and across equations. Then $T - M$ seems to be the appropriate degrees of freedom in the test. Simulation results, see Section 3, indicate that the 95 percent quantile in the statistic in (16) increases with M, and that the adjustment based

on $T - M$ performs better than using $T - 1$.[4]

An additional argument for using the smaller number of degrees of freedom is that t-tests based on HAC covariance estimates generally exhibit substantial size distortions, see, e.g., Andrews (1991), Andrews and Monahan (1992), and den Haan and Levin (1997). Usually, the distortions lead to over-rejecting. Applying $T - M$ degrees of freedom instead of $TM - M$ degrees of freedom in the test can therefore also reduce distortion bias.

2.4 The full encompassing test statistics

Now, consider the encompassing test based on the formulation in (5), i.e., with the more general weight matrix Γ. Then (8) is replaced by

$$\mathbf{y} = (\mathbf{D} \otimes I_M)\, vec\Gamma + \varepsilon, \tag{18}$$

where $\mathbf{D} = \left(\left(\mathbf{D}_{1,H|1}\right), \left(\mathbf{D}_{2,H|2}\right), \ldots, \left(\mathbf{D}_{T,H|T}\right)\right)'$ being a $T \times M$ matrix and vec is the vector operator. The estimator of $vec\Gamma$ becomes

$$vec\hat{\Gamma} = \left(\frac{1}{T}\mathbf{D}'\mathbf{D} \otimes I_M\right)^{-1} \left(\frac{1}{T}\left(\mathbf{D} \otimes I_M\right)' \mathbf{y}\right) \tag{19}$$

with variance

$$\begin{aligned} &\widehat{Var\left(vec\left(\hat{\Gamma}\right)\right)} \\ =&\frac{1}{T}\left(\frac{1}{T}\mathbf{D}'\mathbf{D} \otimes I_{M)}\right)^{-1} \left[\frac{1}{T}\left(\mathbf{D} \otimes I_M\right)' \hat{\Sigma}_{(G)} \left(\mathbf{D} \otimes I_M\right)\right] \left(\frac{1}{T}\mathbf{D}'\mathbf{D} \otimes I_{M)}\right)^{-1}, \end{aligned} \tag{20}$$

where $\hat{\Sigma}_{(G)}$ has the properties in (12) with $\hat{\mathbf{U}}_{(G),t,H|t}$ given by (13) where the scalar a is replaced with the arbitrary $M \times M$ matrix G. The term in the square brackets, which we define as $\mathbf{Q}_{(G)}$,

[4] In addition, the simulation results support that applying the "total number of observations in all equations" — i.e., TM, as the system-based t-test calls for — is not appropriate here.

becomes:

$$\begin{aligned}\mathbf{Q}_{(G)} =& \frac{1}{T}\left(\mathbf{D}\otimes I_M\right)'\hat{\Sigma}_{(G)}\left(\mathbf{D}\otimes I_M\right)\\ =& \frac{1}{T}\sum_{t=1}^{T}\left[\bigtriangledown_{(G),t}\bigtriangledown'_{(G),t}\right]+\frac{1}{T}\sum_{l=1}^{H}\sum_{t=1}^{T-l}w_l\left[\bigtriangledown_{(G),t}\bigtriangledown'_{(G),t+l}+\bigtriangledown_{(G),t+l}\bigtriangledown'_{(G),t}\right],\end{aligned}$$

where $\bigtriangledown_{(G),t} = \left(\left(\mathbf{D}_{t,H|t}\right)\otimes I_M\right)\hat{\mathbf{U}}_{(G),t,H|t}$. Hence, $\bigtriangledown_{(G),t}$ has dimension $M^2\times 1$ and corresponds to the variable $d_{(a),t}$ in the simple case. The equality requires $w_l = 1$, but I follow Newey and West (1987) and use $w_l = 1-\frac{l}{H+l}$ to secure the $M^2\times M^2$ matrix $\mathbf{Q}_{(G)}$ to be positive definite.

Let Γ_0 be the Γ-matrix under the null hypothesis. Then, provided that $\mathbf{Q}_{(G)}$ is nonsingular, the F-test statistic is given by

$$F = \frac{T}{M^2}w_0\bar{\bigtriangledown}'_{(\Gamma_0)}\mathbf{Q}^{-1}_{(G)}\bar{\bigtriangledown}_{(\Gamma_0)}, \tag{21}$$

where $\bar{\bigtriangledown}_{(G)}$ is the sample mean of $\bigtriangledown_{(G),t}$ with G set equal to Γ_0, and w_0 is as given for the t-test.

A similar modification to the one suggested in Harvey et al. (1998) and Harvey and Newbold (2000), see (17), would be

$$\begin{aligned}\mathbf{Q}^{*}_{(\Gamma_0)} =& \frac{1}{T}\sum_{t=1}^{T}\left[\left(\bigtriangledown_{(\Gamma_0),t}-\bar{\bigtriangledown}_{(\Gamma_0)}\right)\left(\bigtriangledown_{(\Gamma_0),t}-\bar{\bigtriangledown}_{(\Gamma_0)}\right)'\right]\\ &+\frac{1}{T}\sum_{l=1}^{H}\sum_{t=1}^{T-l}w_l\left[\left(\bigtriangledown_{(\Gamma_0),t}-\bar{\bigtriangledown}_{(\Gamma_0)}\right)\left(\bigtriangledown_{(\Gamma_0),t+l}-\bar{\bigtriangledown}_{(a)}\right)'\right.\\ &\qquad\left.+\left(\bigtriangledown_{(\Gamma_0),t+l}-\bar{\bigtriangledown}_{(\Gamma_0)}\right)\left(\bigtriangledown_{(\Gamma_0),t}-\bar{\bigtriangledown}_{(\Gamma_0)}\right)'\right].\end{aligned}$$

The matrix $\mathbf{Q}^{*}_{(\Gamma_0)}$ is not the conventional covariance matrix as it allows for heteroscedasticity and autocorrelations. However, if $\mathbf{Q}^{*}_{(\Gamma_0)}$ were the conventional covariance matrix of $\bigtriangledown_{(\Gamma_0),t}$, then the statistics $T\bar{\bigtriangledown}'_{(\Gamma_0)}\mathbf{Q}^{*-1}_{(\Gamma_0)}\bar{\bigtriangledown}_{(\Gamma_0)}$ would take the form of Hotelling's T^2-statistic, see Hotelling (1931). Following Harvey and Newbold (2000), I assume that this expression still can be approximated with Hotellings T^2-statistic. Under the null hypothesis, this statistic has the distribution $\frac{M^2(T-1)}{T-M^2}F_{(M^2,T-M^2)}$ where $F_{(d_1,d_2)}$ is the F-distribution with d_1 degrees of freedom in the numerator and d_2 degrees of freedom in the denominator. Hence, when applying $\mathbf{Q}^{*}_{(\Gamma_0)}$, we

use the test statistic

$$F^* = \frac{T}{M^2}\frac{T-M^2}{T-1} w_0 \bar{\nabla}'_{(\Gamma_0)} \mathbf{Q}^{*-1}_{(\Gamma_0)} \bar{\nabla}_{(\Gamma_0)}. \tag{22}$$

Note that when $M = 1$ the F-statistic is identical to the t-statistic.

3 Size and power of the tests

3.1 Size of the tests

To test the empirical size of the two tests, we consider (5) or (6) where $\alpha = 0$ or $\Gamma = 0$, respectively, under the null hypothesis. Hence, we generate the forecast error (vector) for forecast A — the forecast with the 'correct' forecast — as a "white noice" process; $\mathbf{E}^A_{t,H|t} \sim N(0, I_M)$. The forecasts from the other forecast B have no additional information on the variables forecasted, and we implicitly generate these forecasts through $\mathbf{D}_{t,H|t} = \mathbf{E}^A_{t,H|t} - \mathbf{E}^B_{t,H|t} \sim N(0, v^2 I_M)$ with the scalar $v \neq 0$. The joint forecasts errors are then generated by

$$\begin{pmatrix} \mathbf{E}^A_{t,H|t} \\ \mathbf{E}^B_{t,H|t} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} I_M & I_M \\ I_M & (1+v^2) I_M \end{pmatrix} \right), \tag{23}$$

which is similar to the simulation design in Harvey and Newbold (2000) when they only consider the forecast of one variable at one horizon (but with more than two forecasts). Furthermore, as also noted by Harvey and Newbold (2000), the null distributions will be invariant to the choice of v as long as $v \neq 0$. Hence, we will not consider different values for v in the simulations.

In Table 1 the actual size of the t-test and F-test based on the estimated parameters ($\mathbf{Q}_{(\hat{\alpha})}$ or $\mathbf{Q}_{(\hat{\Gamma})}$) and the parameters under the null ($\mathbf{Q}^*_{(0)}$) for different values of T, M, H at both the 1 and 5 percent significance level is reported. The F-test is not applicable when T is small and M is large (i.e., when $T \leq M^2$), both because $\mathbf{Q}$ is singular and the degrees of freedom in the test becomes negative.

Based on the simulations reported in Table 1 we can draw the following conclusions: First, there are large distortions for the t-test in small samples for $M \geq 4$, though applying $\mathbf{Q}^*$ reduces these distortions. Second, the actual size is higher for $H = 1$ than for $H = 0$. Third,

Table 1: Size of tests of forecast encompassing

			$\mathbf{Q}_{(\hat{\alpha})}$ or $\mathbf{Q}_{(\hat{\Gamma})}$				$\mathbf{Q}^*_{(0)}$			
			$H=0$		$H=1$		$H=0$		$H=1$	
T	M	Test statistic	1 pct.	5 pct.	1 pct.	5 pct.	1 pct.	5 pct.	1 pct.	5 pct.
12	1	t and F	0.0381	0.1066			0.0051	0.0396		
	2	t	0.0530	0.1350	0.0660	0.1611	0.0080	0.0613	0.0102	0.0591
		F	0.0921	0.2244	0.1575	0.3200	0.0019	0.0223	0.0113	0.0552
	3	t	0.0723	0.1804			0.0151	0.0914		
		F	0.0699	0.2405			0.0062	0.0298		
	6	t	0.1779	0.3431	0.2015	0.3721	0.0328	0.2082	0.0372	0.1939
		F	—	—	—	—	—	—	—	—
25	1	t and F	0.0225	0.0787			0.0066	0.0418		
	2	t	0.0304	0.0936	0.0366	0.1059	0.0107	0.0649	0.0129	0.0657
		F	0.0572	0.1495	0.0854	0.2007	0.0049	0.0332	0.0092	0.0548
	3	t	0.0302	0.0978			0.0140	0.0724		
		F	0.0905	0.2317			0.0038	0.0301		
	6	t	0.0614	0.1579	0.0718	0.1714	0.0383	0.1328	0.0395	0.1298
		F	—	—	—	—	—	—	—	—
100	1	t and F	0.0144	0.0602			0.0096	0.0511		
	2	t	0.0143	0.0567	0.0151	0.0608	0.0111	0.0516	0.0120	0.0526
		F	0.0220	0.0804	0.0276	0.0896	0.0078	0.0479	0.0093	0.0513
	3	t	0.0151	0.0645			0.0124	0.0596		
		F	0.0314	0.1048			0.0058	0.0442		
	6	t	0.0170	0.0687	0.0176	0.0715	0.0150	0.0661	0.0154	0.0652
		F	0.0884	0.2299	0.3094	0.5326	0.0062	0.0386	0.0741	0.2121
1000	1	t and F	0.0108	0.0507			0.0105	0.0500		
	2	t	0.0103	0.0521	0.0102	0.0528	0.0101	0.0516	0.0101	0.0514
		F	0.0103	0.0524	0.0113	0.0529	0.0089	0.0493	0.0096	0.0499
	3	t	0.0118	0.0528			0.0114	0.0524		
		F	0.0100	0.0542			0.0084	0.0472		
	6	t	0.0113	0.0520	0.0116	0.0523	0.0112	0.0519	0.0114	0.0513
		F	0.0144	0.0624	0.0186	0.0768	0.0101	0.0460	0.0117	0.0552

Note: table shows actual size of test for both the simple encompassing test (t-test) and the full encompassing test (F-test) for different values of T, M, H and the nominal size (1 pct. and 5 pct.), all derived both under the estimated parameters ($\mathbf{Q}_{(\hat{\alpha})}$ or $\mathbf{Q}_{(\hat{\Gamma})}$) and under the null hypothesis ($\mathbf{Q}^*_{(0)}$). The figures are based on 10 000 replications.

Table 2: Simulated quantiles

			$\mathbf{Q}_{(\hat{\alpha})}$ or $\mathbf{Q}_{(\hat{\Gamma})}$				$\mathbf{Q}^*_{(0)}$			
			$H=0$		$H=1$		$H=0$		$H=1$	
T	M	Test statistic	99 pct.	95 pct.	99 pct.	95 pct.	99 pct.	95 pct.	99 pct.	95 pct.
12	1	t and F	20.5075	8.1793			7.9988	4.4452		
	2	t	23.7051	10.4784	30.0717	12.2992	9.3708	5.3911	10.1162	5.3868
		F	22.3048	10.1090	35.0118	14.9118	4.7309	3.0139	7.3413	3.9532
	3	t	32.2494	13.9104			11.8542	6.8649		
		F	127.8313	37.0102			19.6239	6.4436		
	6	t	92.3865	37.5830	121.4018	44.3172	19.4085	12.1573	21.3042	12.2590
		F	—	—	—	—	—	—	—	—
25	1	t and F	11.5567	5.4355			6.7849	3.9612		
	2	t	11.7374	6.2409	12.9193	6.7828	7.9774	4.8257	8.4338	4.8142
		F	7.5258	4.6052	9.6957	5.4812	3.7157	2.5677	4.2437	2.9059
	3	t	12.3028	6.3915			8.9193	5.1731		
		F	7.8456	4.7543			3.1525	2.2572		
	6	t	16.9092	9.2091	18.6325	10.0863	12.7018	7.2961	13.0938	7.4885
		F	—	—	—	—	—	—	—	—
100	1	t and F	7.6941	4.3455			6.7822	3.9668		
	2	t	7.7824	4.1940	8.2824	4.3090	7.3065	3.9806	7.5822	4.0124
		F	4.0801	2.8478	4.3245	3.0040	3.3884	2.4368	3.4800	2.4907
	3	t	7.8439	4.4102			7.3644	4.2384		
		F	3.1845	2.3810			2.4570	1.9497		
	6	t	7.9349	4.6424	8.1253	4.7991	7.6570	4.4768	7.6415	4.5598
		F	2.6039	2.1209	3.6557	2.8621	1.8674	1.5420	2.5104	2.0561
1000	1	t and F	6.8227	3.8723			6.7088	3.8501		
	2	t	6.7621	3.9307	6.7381	3.9255	6.6770	3.9151	6.6683	3.8919
		F	3.3627	2.4115	3.3889	2.4215	3.2911	2.3736	3.3099	2.3784
	3	t	6.9007	3.9561			6.8270	3.9398		
		F	2.4208	1.9141			2.3436	1.8702		
	6	t	6.8263	3.9281	6.8737	3.9164	6.8025	3.9166	6.8830	3.9083
		F	1.7213	1.4684	1.7446	1.5013	1.6512	1.4161	1.6719	1.4448

Note: table shows simulated quantiles (99 pct. and 95 pct.) of test for both the simple encompassing test (t-test) and the full encompassing test (F-test) for different values of T, M, and H, all derived both under the estimated parameters ($\mathbf{Q}_{(\hat{\alpha})}$ or $\mathbf{Q}_{(\hat{\Gamma})}$) and under the null hypothesis ($\mathbf{Q}^*_{(0)}$). The figures are based on 10 000 replications.

Table 3: Power of tests

			$\mathbf{Q}_{(\hat{\alpha})}$ or $\mathbf{Q}_{(\hat{\Gamma})}$				$\mathbf{Q}^*_{(0)}$			
T&			$H=0$		$H=1$		$H=0$		$H=1$	
$(av)^2$	M	Test statistic	1 pct.	5 pct.	1 pct.	5 pct.	1 pct.	5 pct.	1 pct.	5 pct.
12	1	t and F	0.0860	0.2880			0.0548	0.2281		
0.5^2	2	t	0.1744	0.4533	0.1570	0.4390	0.1542	0.4227	0.1260	0.3936
		F	0.0567	0.2086	0.0526	0.1958	0.0444	0.1653	0.0297	0.1436
	3	t	0.2335	0.5472			0.2032	0.5176		
		F	0.0271	0.1170			0.0116	0.0699		
	6	t	0.2521	0.6143	0.2148	0.5870	0.2440	0.5719	0.1886	0.5061
		F	—	—	—	—	—	—	—	—
25	1	t and F	0.0919	0.2795			0.0873	0.2666		
0.3^2	2	t	0.2109	0.4509	0.2036	0.4413	0.2068	0.4436	0.1751	0.4303
		F	0.0779	0.2324	0.0641	0.2228	0.0698	0.2217	0.0643	0.1929
	3	t	0.3365	0.6171			0.3323	0.6146		
		F	0.0530	0.1965			0.0442	0.1631		
	6	t	0.5863	0.8219	0.5595	0.8052	0.5748	0.8367	0.5253	0.8115
		F	—	—	—	—	—	—	—	—
100	1	t and F	0.2486	0.4888			0.2424	0.4895		
0.2^2	2	t	0.5302	0.7897	0.5063	0.7854	0.5185	0.7885	0.4940	0.7834
		F	0.2971	0.5282	0.2870	0.5156	0.2895	0.5246	0.2781	0.5193
	3	t	0.7613	0.9115			0.7619	0.9121		
		F	0.3046	0.5576			0.3095	0.5499		
	6	t	0.9818	0.9964	0.9806	0.9964	0.9818	0.9964	0.9803	0.9964
		F	0.2530	0.4973	0.1944	0.4479	0.1920	0.4559	0.1618	0.3800
1000 &	1	t and F	0.7083	0.8778			0.7088	0.8772		
0.1^2	2	t	0.9708	0.9918	0.9711	0.9920	0.9713	0.9918	0.9713	0.9920
		F	0.8859	0.9640	0.8833	0.9632	0.8856	0.9637	0.8844	0.9633
	3	t	0.9976	0.9997			0.9976	0.9997		
		F	0.9473	0.9867			0.9500	0.9871		
	6	t	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		F	0.9828	0.9979	0.9837	0.9978	0.9849	0.9978	0.9837	0.9976

Note: table shows power of test for both the simple encompassing test (t-test) and the full encompassing test (F-test) for different values of T, M, H, $(av)^2$ and the nominal size (1 pct. and 5 pct.), all derived both under the estimated parameters ($\mathbf{Q}_{(\hat{\alpha})}$ or $\mathbf{Q}_{(\hat{\Gamma})}$) and under the null hypothesis ($Q^*_{(0)}$). The figures are based on 10 000 replications.

the change in size is much more profound for the F-tests than for the t-tests when applying $\mathbf{Q}^*$ instead of $\mathbf{Q}$. Fourth, for $H = 0$ using $\mathbf{Q}^*$, the F-tests have generally an actual size smaller than the nominal size. Using $\mathbf{Q}$, the F-test has an actual size bigger than the nominal size. Fifth, the distortions increase in the dimension M for the t-test. Sixth and finally, and as expected, the size distortions decrease when the sample increases in T.

Since there can be large size distortions Table 2 reports the simulated 99 percent and 95 percent quantiles. Hence, these quantiles can be applied instead of threshold values from the t- and F-distributions, especially when there are large size distortions.

3.2 Power of the tests of forecasting encompassing

To indicate the power of the tests the following processes are simulated. First, as for the test of the size, $D_{t,h|t} = E^B_{t,h|t} - E^A_{t,H|t} \sim N\left(0, v^2 I_M\right)$ with $v \neq 0$ is generated. Then $E^A_{t,h|t} - aD_{t,h|t} \sim N\left(0, I_M\right)$ is generated for the scalar a. For $a = 0$ this is the same process as was used for simulating the size of the test. Hence, for $a = 0$ the null hypothesis that forecasts A encompass forecasts B is correct. Therefore, when $a \neq 0$ the null hypothesis is not correct.

In Table 3 I have used $v = 1$ and different values of a, depending on the sample size (T). However, it can be shown — see, e.g., Harvey and Newbold (2000) — that the simulation only depends on $(av)^2$ (for $v \neq 0$), so to derive the power for different values of a and v only the different values of their squared product need to be examined. The table reports the fractions the simulations exceed the simulated quantiles in Table 2. Hence, the tests do not gain power if their actual size is greater than the nominal size.

From the simulations the following conclusions can be drawn: First, the t-test has much more power than the F-test. Second, the power is approximately the same when using $\mathbf{Q}^*$ as when using $\mathbf{Q}$. Third, the power seems to increase with the dimension M for the t-tests and to decrease for the F-tests. Forth, the power is smaller for $H = 1$ than for $H = 0$.

4 Encompassing tests for forecasts made by Statistics Norway

Statistics Norway has, with a few exceptions, published forecasts every quarter for many variables for the year the forecast was made as well as the following year since the 1st quarter in

1990. Among these variables are Mainland GDP[5], CPI and the unemployment rate (UR).

Even though a quarterly model is used in generating the forecasts, the forecasts are only published in annual terms. Therefore, in testing the forecasts from Statistics Norway, I only consider forecasts of annual values.

Statistics Norway did not publish forecasts in the 2nd quarter of 1990 and 1991. In the analysis, I have set those forecasts equal to the forecasts made in the 1st quarter. Also in the 3rd quarter in 2013 Statistics Norway did not publish a forecast. Here, also, the forecast for the 3rd quarter is set equal to the previously published forecast, i.e., the forecast from the 2nd quarter that year.[6] The latest forecast considered in the examination is the forecast made in the 4th quarter of 2014, which includes forecasts of variables for the year 2015. Hence, the sample spans 25 years of forecasts.

The published numbers for CPI and the unemployment rate are never revised. The published numbers for variables from the National Accounts, such as the Mainland GDP, can be revised in many quarters until they are fixed. Though also these 'fixed' numbers can be revised due to revisions in the System of National Accounts. In the analysis undertaken here, the first published number of Mainland GDP-growth is used.

4.1 Comparison with random walk

Table 4 compares the forecasts by Statistics Norway with the forecasts generated by a random walk. The forecasts based on a random walk are equal to the last observed value of the variables, i.e. the last observed GDP-growth (for Mainland-Norway), the last observed year-on-year CPI-growth, and the last observed level of unemployment over a calendar year. Hence, the forecasted values for the growth provided by a random walk given at time t, q (i.e., quarter q in year t) for Mainland GDP in both year t and year $t+1$ is equal to the Mainland GDP growth in year $t-1$. Similarly, the random walk forecast implies that the CPI-growth in both year t and year $t+1$ is equal to the CPI-growth in year $t-1$, and the unemployment rate in t and $t+1$ is equal to the unemployment rate in year $t-1$.

[5]Mainland Norway consists of all domestic production activity except exploration of crude oil and natural gas, transport via pipelines and ocean transport. The term was revised as a part of the main revision of the national accounts in 2014. Before this, service activities incidental to oil and gas were also excluded from Mainland Norway.

[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.

The forecasts by Statistics Norway, given at time t, q for GDP, CPI, and UR, for horizons 0 and 1, are given by the vector

$$\mathbf{Y}_{t,1|(t,q)} = \left(\Delta GDP_{t|(t,q)}, \Delta GDP_{t+1|(t,q)}, \Delta CPI_{t|(t,q)}, \Delta CPI_{t+1|(t,q)}, UR_{t|(t,q)}, UR_{t+1|(t,q)}\right)', \quad (24)$$

where Δ indicates that we are considering the growth of that particular variable. The ordering of the elements in the vectors in (24) is not important, only that they are in the same order in the two forecasts we are comparing when applying the simplified encompassing test. The distinction between level and growth in (24) is only important for the simplified encompassing test when we are comparing with forecasts generated by a random walk, as it makes it easier to formulate the corresponding forecast from the random walk model. The corresponding forecasts of the random walk model are then the last observation of the corresponding variable in (24), i.e., $\mathbf{Y}^{RW}_{t,1|(t,q)} = (\Delta GDP_{t-1}, \Delta GDP_{t-1}, \Delta CPI_{t-1}, \Delta CPI_{t-1}, UR_{t-1}, UR_{t-1})'$.[7]

For the full encompassing test of the full system the regression considered is therefore

$$\mathbf{Y}_{t,1} = (1-\Gamma)\,\mathbf{Y}_{t,1|(t,q)} + \Gamma \mathbf{Y}^{RW}_{t,1|(t,q)} + u_{t,q}, \quad (25)$$

where $\mathbf{Y}_{t,1} = (\Delta GDP_t, \Delta GDP_{t+1}, \Delta CPI_t, \Delta CPI_{t+1}, UR_t, UR_{t+1})$, and $N = 3$, $H = 1$, and — therefore — $M = 6$. For the simple encompassing test the coefficient matrix is restricted by $\Gamma = \alpha I_M$ where α is a scalar.

In the first block of Table 4, one is testing if the forecasts of the full system made in one quarter encompass the forecasts based on a random walk, as in (25). In the first line, two numbers are reported for each quarter; the estimated α and its standard error.[8] For forecasts made in both the 2nd and 3rd quarter the weights on the forecasts based on the random walk are less than 0.01 (i.e., less than one percent). Based on the standard error (derived by the estimated α) these estimates are clearly not significantly different from zero. Also for the 4th quarter, the weight on the forecasts based on the random walk is small (less than 3 percent), but

[7] If all variables in $\mathbf{Y}_{t|(t,q)}$ were in levels, we could formulate the corresponding forecasts generated by a random walk as $(GDP_{t-1} + \Delta GDP_{t-1}, GDP_{t-1} + 2\Delta GDP_{t-1}, CPI_{t-1} + \Delta CPI_{t-1}, CPI_{t-1} + 2\Delta CPI_{t-1}, UR_{t-1}, UR_{t-1})$. The results in Table 4 would then be unchanged, also for the simplified encompassing test provided that also $\mathbf{Y}_{t,1}$ were measured in levels.

[8] The standard error is implicitly derived such that the t-value is given in (16) with $a = \alpha$.

Table 4: Comparison with Random Walk

	Q1	Q2	Q3	Q4
Both horizons				
$\hat{\alpha}_{(\hat{\Omega})}$	0.1148 (0.0386)	0.0002 (0.0240)	0.0042 (0.0178)	0.0287 (0.0102)
$H_0 : \alpha = 0$	(0.0433) 7.0309 [0.0162]	(0.0256) 0.0001 [0.9941]	(0.0195) 0.0455 [0.8335]	(0.0171) 2.8276 [0.1099]
$H_0 : \alpha = 1$	(0.2172) 16.6073 [0.0007]**	(0.2756) 13.1653 [0.0019]**	(0.3016) 10.8992 [0.0040]*	(0.2979) 10.6329 [0.0043]*
Nowcasting				
$\hat{\alpha}_{(\hat{\Omega})}$	0.1128 (0.0552)	-0.0138 (0.0275)	0.0168 (0.0199)	0.0205 (0.0103)
$H_0 : \alpha = 0$	(0.0576) 3.8318 [0.0637]	(0.0282) 0.2413 [0.6284]	(0.0229) 0.5425 [0.4695]	(0.0128) 2.5715 [0.1237]
$H_0 : \alpha = 1$	(0.1850) 22.9940 [0.0001]**	(0.2609) 15.0972 [0.0009]**	(0.2593) 14.3790 [0.0011]**	(0.2635) 13.8142 [0.0013]**
$H_0 : \Gamma = \mathbf{0}$	0.4941 [0.8397]	0.5126 [0.8272]	0.3169 [0.9440]	0.5932 [0.7716]
$H_0 : \Gamma = \mathbf{I}_M$	3.0052 [0.0804]*	2.4934 [0.1208]*	3.1232 [0.0737]*	3.7472 [0.0477]**
GDP $H = 1$				
$\hat{\alpha}_{(\hat{\Omega})}$	0.1572 (0.1127)	0.0358 (0.0842)	-0.0006 (0.0801)	0.0054 (0.0396)
$H_0 : \alpha = 0$	(0.0902) 3.0353 [0.0954]	(0.0807) 0.1968 [0.6616]	(0.0853) 0.0000 [0.9946]	(0.0414) 0.0168 [0.8981]
$H_0 : \alpha = 1$	(0.3840) 4.8167 [0.0390]*	(0.4109) 5.5055 [0.0284]*	(0.3911) 6.5445 [0.0179]*	(0.3867) 6.6152 [0.0174]*
$H_0 : \Gamma = \mathbf{0}$	1.0511 [0.4104]	0.3654 [0.8298]	0.9699 [0.4494]	0.3569 [0.8357]
$H_0 : \Gamma = \mathbf{I}_M$	1.2162 [0.3404]	1.3728 [0.2848]	1.5459 [0.2339]	1.8802 [0.1603]
CPI $H = 1$				
$\hat{\alpha}_{(\hat{\Omega})}$	0.0414 (0.0634)	-0.0320 (0.0434)	0.0028 (0.0242)	0.0205 (0.0117)
$H_0 : \alpha = 0$	(0.0684) 0.3654 [0.5517]	(0.0466) 0.4727 [0.4989]	(0.0263) 0.0116 [0.9151]	(0.0162) 1.6060 [0.2183]
$H_0 : \alpha = 1$	(0.3378) 8.0552 [0.0096]*	(0.3527) 8.5631 [0.0078]**	(0.3362) 8.7980 [0.0071]**	(0.3148) 9.6780 [0.0051]**
$H_0 : \Gamma = \mathbf{0}$	1.0939 [0.3911]	1.9145 [0.1543]	2.2744 [0.1037]	0.7677 [0.5609]
$H_0 : \Gamma = \mathbf{I}_M$	4.6855 [0.0099]**	3.1912 [0.0398]*	5.0329 [0.0073]**	4.6209 [0.0104]**
UR $H = 1$				
$\hat{\alpha}_{(\hat{\Omega})}$	0.2295 (0.1492)	0.1213 (0.1432)	0.0488 (0.0708)	0.0260 (0.0601)
$H_0 : \alpha = 0$	(0.1715) 1.7918 [0.1944]	(0.1631) 0.5533 [0.4648]	(0.0724) 0.4533 [0.5078]	(0.0603) 0.1863 [0.6702]
$H_0 : \alpha = 1$	(0.2570) 8.9876 [0.0066]**	(0.2637) 11.1066 [0.0030]**	(0.3257) 8.5292 [0.0079]**	(0.3021) 10.3954 [0.0039]**
$H_0 : \Gamma = \mathbf{0}$	2.3031 [0.1005]	0.7745 [0.5568]	0.8700 [0.5020]	0.9816 [0.4436]
$H_0 : \Gamma = \mathbf{I}_M$	2.1431 [0.1197]	3.1341 [0.0421]*	3.7003 [0.0242]*	3.1425 [0.0417]*

Note: Simple and full encompassing test for forecasts made by Statistics Norway encompass a random walk. Q1 – Q4 indicates the quarter (of the year) the forecast is made. For $\hat{\alpha}_{(\hat{\Omega})}$ the estimated value and its standard errors (implicitly derived such that the t-value is given in (16) with $a = \alpha$) is reported. For the simple tests standard errors (derived under the null with $\mathbf{Q}^*$), F-value (squared t-value), and the corresponding p-value to the F-value are reported, in addition to asterisks, where one asterisks denote that the test statistics exceeds the 95 percent critical value and two asterisks denote that the test statistics exceeds the 99 percent critical value. For the full test the F-value with corresponding p-value is reported together with asterisks.

with an even smaller standard error. A standard t-test would lead to rejection of the hypothesis $\alpha = 0$. The same is the case for the forecasts made in the 1st quarter, where both the estimated α and its standard error are larger.

In the line "$H_0 : \alpha = 0$" the null hypothesis of $\alpha = 0$ is tested when the standard error is derived under the null (and $\mathbf{Q}^*$ is applied). In the parenthesis, the corresponding standard error to the t-test is reported. Also, the F-value (i.e., the square of the t-test) is reported, and the p-value based on the t-test is reported in the square brackets. Finally, two asterisks succeeding the p-value denote that the F-value exceeds the simulated 99 percent quantile reported in Table 2; similarly, one asterisk shows that the F-value of the test exceeds the simulated 95 percent quantile. The standard errors are somewhat larger when derived under the null hypothesis

(and $\mathbf{Q}^*$) than when derived based on the estimated α.

The results of the test $\alpha = 0$ show that the forecasts made in the 2nd, 3rd, and 4th quarter encompass the forecasts based on a random walk. For the forecasts made in the 1st quarter, the encompassing test is rejected at the 5 percent level when applying the t-test. However, when compared to the simulated 95 percent quantile, also the forecasts made in the 1st quarter encompass the forecasts based on a random walk.

In the line "$H_0 : \alpha = 1$" the null hypothesis $\alpha = 1$ is rejected for the forecasts made in all four quarters. Hence, forecasts made by Statistics Norway do significantly improve the forecasts based on a random walk. From the estimated α and its standard error, this is not surprising. Though, when applying the standard error under the null hypothesis, the standard error increases dramatically: for the forecasts made in the 4th quarter the standard error is almost 30 times higher when derived under $\alpha = 1$ than for the estimated α. Even with these large increases in the standard error, the hypothesis that $\alpha = 1$ is clearly rejected.

The F-test for the full system forecast ($M = 6$) is not applicable since $T < M^2$.

In the remaining tests, only subsets of the system are applied. The results are relatively similar to the ones with the full system when considering the t-test. The F-test supports the t-test when considering nowcasting and the forecasts of CPI and UR, where $\Gamma = 0$ is not rejected, but (with one exception) $\Gamma = I$ is rejected (at the 5 percent significance level).

When only considering the forecasts of GDP (for mainland Norway) the hypothesis $\Gamma = I$ cannot be rejected despite that the hypothesis $\alpha = 1$ is rejected in the t-tests for forecasts made in all quarters. These diverging results can be due to the low power of the F-test.

4.2 Improved forecast through the year?

The encompassing tests can also be used to test if a forecast made at a point in time outperforms a forecast of the same variables for the same periods made at an earlier point in time. If the latest forecast does not encompass the previous forecast, one could make a better forecast by using a weighted average of the two forecasts. In this case, the latest published forecast is not efficient.

In the upper part of Table 5 I report test results related to whether forecasts of Mainland GDP, CPI, and UR for the current and next year improves throughout the calendar year. To

test if the forecast improves throughout the year, we consider the following regression

$$\mathbf{Y}_{t,1} = (1-\Gamma)\,\mathbf{Y}_{t,1|(t,q_1)} + \Gamma \mathbf{Y}_{t,1|(t,q_2)} + \mathbf{u}_{t,1}, \tag{26}$$

with $\Gamma = \alpha I_m$ for the simplified encompassing test. Suppose $q_1 > q_2$. Then we hope the forecast made in quarter q_1 ($q_1 = 2,3,4$) should include all information that was available at quarter q_2 ($q_2 = 1,2,3$) plus the extra information that has come available between the two times the forecasts were made. Hence, we should not expect a combined forecast of $\mathbf{Y}_{t+1|(t,q_2)}$ and $\mathbf{Y}_{t+1|(t,q_1)}$ to outperform the latter (i.e., the latest forecast). Therefore, we hope that the hypothesis $\alpha = 0$ is not rejected. Finally, we would like the hypothesis $\alpha = 1$ to be rejected.

The upper part of Table 5 reports results of the encompassing tests for the system with all three variables ($N = 3$) and forecasts for both the current and next year ($H = 1$), hence $M = 6$, as in (24). In the first column, I test if the forecasts made in the 2nd quarter of the year encompass the forecasts made in the 1st quarter. In the remaining two columns I also compare the forecasts with the forecasts made in the previous quarter; hence I always use $q_2 = q_1 - 1$.

The estimate of α is close to zero (-0.0631 when comparing the 2nd quarter forecasts to the 1st quarter forecasts; 0.0805 when comparing the 3nd quarter forecasts and the 2nd quarter forecasts; and 0.0769 when comparing the two latest forecast of the year), indicating that the best combined forecast puts full weight on the latest forecasts and no weight on the previous forecasts. The standard errors (reported in parenthesis and based on the estimated α) are relatively large, so based on the t-test derived for the estimated α the hypothesis that $\alpha = 0$ is not rejected for any of the compared quarters. The exception is for the test of whether the forecasts made in the 4th quarter encompass the forecasts of the 3rd quarter; here the t-value is about 2.3, whereas the critical value (based on the t-distribution) is about 2.1 with $T - M = 25 - 6 = 19$ degrees of freedom at the 5 percent significance level in a two-sided test.

The next line in the table reports the t-test for $\alpha = 0$ when $\mathbf{Q}^*$ is applied. For this test the null hypothesis $\alpha = 0$ is not rejected for any of the three cases.

The results of the test of the hypothesis $\alpha = 1$ indicate the power of the encompassing test. Note that the standard error is between three and four times higher than for the test of the hypothesis $\alpha = 0$. If I had used the estimated standard error, the hypothesis of $\alpha = 1$

Table 5: Improved forecasts throughout the year?

	Q2 vs. Q1	Q3 vs. Q2	Q4 vs. Q3
Both horizons			
$\hat{\alpha}_{(\hat{\Omega})}$	-0.0631 (0.0815)	0.0805 (0.0702)	0.0769 (0.0332)
$H_0 : \alpha = 0$	(0.0844) 0.5583 [0.4646]	(0.0947) 0.7213 [0.4069]	(0.0384) 4.0135 [0.0604]
$H_0 : \alpha = 1$	(0.2709) 15.4001 [0.0010]**	(0.3802) 5.8483 [0.0264]	(0.1302) 50.2976 [0.0000]**
Nowcasting			
$\hat{\alpha}_{(\hat{\Omega})}$	-0.2073 (0.0818)	-0.0499 (0.0762)	0.0616 (0.0652)
$H_0 : \alpha = 0$	(0.0825) 6.3095 [0.0203]*	(0.0910) 0.3007 [0.5892]	(0.0697) 0.7828 [0.3863]
$H_0 : \alpha = 1$	(0.2449) 24.2974 [0.0001]**	(0.5220) 4.0461 [0.0573]	(0.1672) 31.4824 [0.0000]**
$H_0 : \Gamma = \mathbf{0}$	2.3479 [0.1367]*	2.9062 [0.0867]*	0.7609 [0.6566]
$H_0 : \Gamma = \mathbf{I}_M$	3.2286 [0.0682]**	1.5538 [0.2872]	5.2640 [0.0198]**
GDP $H = 1$			
$\hat{\alpha}_{(\hat{\Omega})}$	-0.3904 (0.2880)	-0.1669 (0.3035)	0.1271 (0.1429)
$H_0 : \alpha = 0$	(0.3231) 1.4596 [0.2398]	(0.3487) 0.2290 [0.6370]	(0.1680) 0.5724 [0.4573]
$H_0 : \alpha = 1$	(0.4922) 7.9809 [0.0099]*	(0.5663) 4.2461 [0.0514]	(0.2003) 18.9874 [0.0003]**
$H_0 : \Gamma = \mathbf{0}$	1.0657 [0.4037]	0.3473 [0.8422]	0.3892 [0.8134]
$H_0 : \Gamma = \mathbf{I}_M$	2.7394 [0.0631]	1.7255 [0.1908]	6.2784 [0.0027]**
CPI $H = 1$			
$\hat{\alpha}_{(\hat{\Omega})}$	0.0769 (0.0648)	-0.0397 (0.0855)	0.0366 (0.0466)
$H_0 : \alpha = 0$	(0.0768) 1.0007 [0.3280]	(0.0775) 0.2624 [0.6135]	(0.0487) 0.5647 [0.4603]
$H_0 : \alpha = 1$	(0.3590) 6.6126 [0.0174]*	(0.5987) 3.0158 [0.0964]	(0.2133) 20.4055 [0.0002]**
$H_0 : \Gamma = \mathbf{0}$	0.8458 [0.5154]	0.9376 [0.4659]	0.6097 [0.6612]
$H_0 : \Gamma = \mathbf{I}_M$	4.2448 [0.0146]**	1.1145 [0.3821]	6.1579 [0.0030]**
UR $H = 1$			
$\hat{\alpha}_{(\hat{\Omega})}$	-0.1647 (0.1822)	0.1494 (0.1582)	0.1597 (0.1389)
$H_0 : \alpha = 0$	(0.2127) 0.5999 [0.4469]	(0.1474) 1.0272 [0.3218]	(0.1715) 0.8662 [0.3621]
$H_0 : \alpha = 1$	(0.3999) 8.4842 [0.0081]**	(0.3665) 5.3866 [0.0300]*	(0.1769) 22.5628 [0.0001]**
$H_0 : \Gamma = \mathbf{0}$	0.3976 [0.8076]	0.8560 [0.5097]	1.0807 [0.3969]
$H_0 : \Gamma = \mathbf{I}_M$	3.0439 [0.0461]*	1.3921 [0.2786]	5.1950 [0.0064]**

Note: Column "Q2 vs. Q1" compares the forecasts made by Statistics Norway in the second quarter (of the year) with the forecasts made in the first quarter. Similarly for the columns "Q3 vs. Q2" and "Q4 vs. Q3". For further explanation, see Table 4.

would clearly have been rejected. Here, however, using the standard error under the null (and $\mathbf{Q}^*$), makes it more difficult to reject the null hypothesis. Still, I reject the null hypothesis that the forecasts made in the 1st quarter encompass the forecasts of the 2nd quarter at the 1 percent level, both when judged by the t-distribution and when compared to the simulated quantile. Also, I reject that the forecasts made in the 3rd quarter encompass the forecasts from the 4th quarter. However, I cannot reject that forecasts made at the 2nd quarter encompass the forecasts from the 3rd quarter when applying the simulated quantile. When applying the t-distribution the p-value shows that the null hypothesis (that $\alpha = 1$) is rejected when applying a 5 percent significance level.

In the remainder of the table, similar tests are conducted of a subsystem of the variables in the full system. The test results are similar to the results for the full system. In the simplified encompassing test for comparing nowcasting of the three variables made in the first two quarters of the year, the estimated weight on the forecasts made in the 1st quarter is -0.2073.[9] With a standard error based on the estimated α of about 0.08, the null hypothesis of $\alpha = 0$ is rejected. This hypothesis is also rejected using the standard error based on α under the null hypothesis (and for the test of the hypothesis $\Gamma = 0$ for the full encompassing test using the variance matrix under the null hypothesis). This results could indicate that Statistics Norway has not put enough weight on new information available between the publications of the 1st and the 2nd quarter.

5 Conclusions

This paper presents two alternative encompassing tests for system forecasts. Both tests involve a correct specification of the system under the null hypothesis that one set of forecasts encompasses another set of forecasts. However, under the alternative, one of the tests (the simplified test, i.e., the t-test) involves a more restrictive formulation than the other test (the general test, i.e., the F-test).

For both of the tests, two different assumptions with respect to the variance are considered; one employs the estimated parameter(s), the other the value under the null hypothesis.

Simulation results show that for both the simplified and the general test, the version where the value of the variance under the null is used has the smallest size distortions. Furthermore, the simplified test shows the highest power.

The two versions of the two tests are used to examine forecasts published quarterly over 25 years from Statistics Norway. I find that the forecasts made by Statistics Norway encompass forecasts based on a random walk. Furthermore, forecasts made in one quarter generally encompass forecasts made in the previous quarter.

[9] Often, the restriction $0 \leq \alpha \leq 1$ is imposed. Even with two unbiased forecasts, the optimal weight on one of them can be negative. As pointed out by Granger and Newbold (1986, ch. 9) for the univariate forecast: "an inferior forecast may still be worth including with negative weight on the grounds that its relatively high error variance is outweighed by a large ρ value — that is to say, the part of the variable of interest left unexplained by it is sufficiently strongly related to the part left unexplained by the better forecast" where "ρ is the correlation between the two forecast errors".

References

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

Andrews, D. W. K. and Monahan, J. C. (1992). An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator. *Econometrica*, 60(4):953–966.

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

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.

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

Christoffersen, P. F. and Diebold, F. X. (1998). Cointegration and Long-Horizon Forecasting. *Journal of Business & Economic Statistics*, 16(4):450–456.

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.

den Haan, W. J. and Levin, A. T. (1997). A practitioner's guide to robust covariance matrix estimation. In *Handbook of Statistics*, volume 15, chapter 12, pages 299–342. Elsevier.

Dhrymes, P. J. (1984). *Mathematics for Econometrics*. Springer-Verlag, New York, 2nd edition.

Diebold, F. X. and Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business and Economic Statistics*, 13(3):253–265.

Doornik, J. A. and Hendry, D. F. (2013). *Modelling Dynamic Systems - PcGive 14: Volume II*, volume II.

Dunnett, C. W. and Sobel, M. (1954). A Bivariate Generalization of Student's t-Distribution, with Tables for Certain Special Cases. *Biometrika*, 41(1-2):153–169.

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.

Ericsson, N. R. (1993). On the limitations of comparing mean square forecast errors: Clarifications and extensions. *Journal of Forecasting*, 12(8):644–651.

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.

Granger, C. W. J. and Newbold, P. (1973). Some comments on the evaluation of economic forecasts. *Applied Economics*, 5(1):35–47.

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

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

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.

Henningsen, A. and Hamann, J. D. (2007). systemfit : A Package for Estimating Systems of Simultaneous Equations in R. *Journal of Statistical Software*, 23(4).

Hotelling, H. (1931). The Generalization of Student ' s Ratio. *The Annals of Mathematical Statistics*, 2(3):360–378.

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

Newey, W. K. and West, K. D. (1987). A simple positive semi-definite heteroskedasticity and autocorrelation-consistent covariance matrix. *Econometrica*, 55:703–708.

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

Schmidt, P. (1993). On the limitations of comparing mean square forecast errors: Comment. *Journal of Forecasting*, 12(8):660–662.

Statistics Norway

Postal address:
PO Box 8131 Dept
NO-0033 Oslo

Office address:
Akersveien 26, Oslo
Oterveien 23, Kongsvinger

E-mail: ssb@ssb.no
Internet: www.ssb.no
Telephone: + 47 62 88 50 00

ISSN: 1892-753X

Statistisk sentralbyrå
Statistics Norway

Design: Siri Boquist