(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

International Journal of Forecasting 33 (2017) 618–626

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Improving the power of the Diebold–Mariano–West test for least squares predictions

Walter J. Mayer [a,*], Feng Liu [a], Xin Dang [b]

[a] Department of Economics, University of Mississippi, University MS, 38677, United States

[b] Department of Mathematics, University of Mississippi, University MS, 38677, United States

ARTICLE INFO

Keywords:
Prediction
Hypothesis tests

ABSTRACT

We propose a more powerful version of the test of Diebold and Mariano (1995) and West (1996) for comparing least squares predictors based on non-nested models when the parameter being tested is the expected difference between the squared prediction errors. The proposed test improves the asymptotic power by using a more efficient estimator of the parameter being tested than that used in the literature. The estimator used by the standard version of the test depends on the individual predictions and realizations only through the observations on the prediction errors. However, the parameter being tested can also be expressed in terms of moments of the predictors and the predicted variable, some of which cannot be identified separately by the observations on the prediction errors alone. Parameterizing these moments in a GMM framework and drawing on the theory of West (1996), we devise more powerful versions of the test by exploiting a restriction that is maintained routinely under the null hypothesis by West (1996, Assumption 2b) and later studies. This restriction requires only finite second-order moments and covariance stationarity in order to ensure that the population linear projection exists. Simulation experiments show that the potential gains in power can be substantial.

1. Introduction

The test of out-of-sample predictive accuracy that was proposed by Diebold and Mariano (1995) and West (1996) is widely regarded as an important test for comparing two predictors.[1] Applications include the studies by Andreou, Ghysels, and Kourtellos (2013), Corradi, Swanson, and Olivetta (2001), Hanson and Lunde (2005), Hong and Lee (2003), Mark (1995) and Swanson and White (1997). Formal asymptotic theory was first presented by West (1996), and further developed by Clark and McCracken (2001, 2014), McCracken (2007) and others. The present paper proposes a more powerful version of the test, obtained by using a more efficient estimator of the parameter being tested than that which is used in the literature. The parameter being tested is the expected difference between functions of the prediction errors. The standard version of the test estimates it using the sample mean of the difference; as a consequence, the individual observations on the predictions and realizations generally enter the test statistic only through the prediction errors. However, the parameter being tested can also be expressed in terms of moments of the predictors and the predicted variable, some of which cannot be identified separately by the

* Corresponding author.
E-mail addresses: wmayer@olemiss.edu (W.J. Mayer), xdang@olemiss.edu (X. Dang).

[1] Unlike Diebold and Mariano (1995), West (1996) accounts for parameter estimation explicitly.

http://dx.doi.org/10.1016/j.ijforecast.2017.01.008

observations on the prediction errors alone.[2] This raises the possibility of using the individual observations on the predictors and the predicted variable to construct more efficient estimators of the tested parameter,[3] leading to tests with greater asymptotic power.

We pursue this for the commonly-tested hypothesis of equal mean squared errors in the case of least squares predictors, and propose a new version of the test based on a more efficient estimator of the expected difference between the squared prediction errors. We parameterize the moments of the predictors in a GMM framework, and derive a more efficient estimator by incorporating a restriction that is part of the weak regularity conditions that are maintained routinely under the null hypothesis by West (1996, Assumption 2b) and later studies. In the context of linear least squares prediction, this restriction requires only finite second-order moments and covariance stationarity. While this requirement is just a standard restriction on the data-generating process, it might be considered more restrictive than the setup of Diebold and Mariano (1995), in which the underlying predictors are completely unspecified. In applications that involve evaluating predictions from surveys, for example, it is clearly desirable to leave the underlying predictors unspecified and impose restrictions on the prediction errors directly. On the other hand, the proposed test should be useful given the large number of studies that evaluate predictions under the conditions of the present paper. Examples of studies that compare the mean squared errors of least squares predictors under covariance stationarity include those by Clark and McCracken (2006), Mark (1995), and Stock and Watson (2002, 2007), among others.

The proposed test and the GMM estimator on which it is based are described in Section 2. Section 3 reports evidence from simulation experiments on the efficiency of the estimator and the size and power of the tests. Consistent with the asymptotic efficiency of the GMM estimator, the simulations confirm substantial power gains for the proposed test relative to the standard version of the Diebold–Mariano–West (DMW) test. Proposition 1 in Section 2 formally establishes the asymptotic normality and efficiency of the GMM estimator relative to the estimator on which the standard DMW test is based. The proof of Proposition 1 follows from an application of West (1996, Theorem 4.1). Devising a DMW-type test based on the GMM estimator is straightforward for non-nested regression models.[4] However, as is the case for the standard DMW test, certain technical problems arise when it is applied to nested and overlapping models. For the standard version of the DMW test, these problems were studied in a series of papers by Clark and McCracken.[5] For the new GMM version proposed here, the problems take the form of duplicated moment conditions and singular covariance matrices. These problems are discussed briefly in Section 2.3, but left as directions for future research. It is well-known from GMM theory that increasing the number of moments generally improves the asymptotic efficiency. Under the assumption of covariance stationarity, the population linear projection error is uncorrelated with any linear combination of the predictor variables. Section 2.4 exploits this to expand the moment functions used to define the GMM estimator in Section 2.2. For the simulation experiments in Section 3, tests based on the expanded moment functions are found to be more powerful for both small and large samples, but to have worse sizes in small samples. Section 4 illustrates the use of the test, with a forecasting application to monthly US industrial production, while Section 5 concludes.

2. GMM version of the DMW test

2.1. Null hypothesis, predictors and restrictions

Given the data available at time t, we consider two competing predictors $\hat{y}_{1,t+\tau}$ and $\hat{y}_{2,t+\tau}$ of a variable $y_{t+\tau}$ at time $t+\tau$ with prediction errors $e_{1,t+\tau} = y_{t+\tau} - \hat{y}_{1,t+\tau}$ and $e_{2,t+\tau} = y_{t+\tau} - \hat{y}_{2,t+\tau}$. The DMW test can be used to test the null hypothesis of equal predictive accuracy for a wide range of loss functions and predictors. An important special case is equal mean squared errors:

$$H_o : \theta \equiv E(e^2_{1,t+\tau} - e^2_{2,t+\tau}) = 0. \tag{1}$$

We assume that Eq. (1) is to be tested using P of the τ-step-ahead predictions. The predictors are of the form $\hat{y}_{j,t+\tau} = X_{j,t+\tau}\hat{\beta}_{j,t}$ $(j = 1, 2)$, where $\hat{\beta}_{jt} = (\hat{\beta}_{jt,0}, \hat{\beta}_{jt,1}, \hat{\beta}_{jt,2})'$ is the least squares estimator computed from the regression of y_t on $X_{j,t} = (1, x_{j,t}, x_t)$ using a minimum of R in-sample observations, where $x_{j,t}$ denotes the vector of predictor variables that are specific to predictor j and x_t is the common set of predictor variables. Following West (1996), we assume a recursive forecasting scheme which is also used commonly in practice.[6] Therefore, $\hat{\beta}_{jt} = \left(\sum_{s=1}^{t} X'_{j,s}X_{j,s}\right)^{-1} \sum_{s=1}^{t} X'_{j,s}y_s$ for $t = R, \ldots, R+P-1$, and $j = 1$ and 2. Given the P out-of-sample predictions, the standard DMW test of Eq. (1) is based on the following estimator:

$$\hat{\theta}_e = \frac{1}{P}\sum_{t=R}^{P+R-1}(e_{1,t+\tau} - e_{2,t+\tau})(e_{1,t+\tau} + e_{2,t+\tau}), \tag{2}$$

where the subscript "e" emphasizes that Eq. (2) is computed from the observations on the prediction errors.

2 Diebold and Mariano (1995, p. 254) note that the test is not limited to testing functions of the prediction error, but can be applied to functions in which the realization and the prediction enter separately. Of course, the same argument applies if the moments of the predicted variable and the predictors cannot be identified by the observations on such functions.

3 The potential inefficiency of replacing a sample of observations (here, the individual observations on the predictors and predicted variable) with functions of the observations (here, the prediction errors) can be quantified in terms of information matrices. Specifically, if $g(X)$ is a measurable function of a random variable X with a density that is a function of a vector θ, it can be shown that difference between the information matrices with respect to θ, $J_X - J_{g(X)}$, is non-negative definite (Rao, 1973, pp. 330–331).

4 Recent studies that have used DMW tests for comparing non-nested models include those by Andreou et al. (2013) and Naes, Skjeltorp, and Ødegaard (2011).

5 See for example Clark and McCracken (2001, 2014) and McCracken (2007).

6 This assumption accommodates our application of West (1996, Theorem 4.1) in Section 2.3. West (1996) assumes a recursive forecasting scheme but considers extensions to fixed and rolling schemes in an unpublished working paper; see West (1994).

Since θ is a function of the moments of $\hat{y}_{1,t+\tau}$, $\hat{y}_{2,t+\tau}$ and $y_{t+\tau}$, the individual observations on $\hat{y}_{1,t+\tau}$, $\hat{y}_{2,t+\tau}$ and $y_{t+\tau}$ can be used to devise an asymptotically more efficient estimator than Eq. (2). The basis for our approach is the following assumption:

Assumption 1. For $j = 1, 2$, the sequence $\{X_{j,t} \; y_t\}$ is covariance-stationary and ergodic for second moments, and $E(X'_{j,t}X_{j,t})$ is nonsingular.

Assumption 1 imposes weak regularity conditions that are maintained routinely under the null hypothesis in the literature; see for example West (1996, pp. 1070–1071) and Clark and McCracken (2014, p. 418). It ensures that the moments $E(X'_{j,t}X_{j,t})$ and $E(X'_{j,t}y_t)$ are finite, that $E(X'_{j,t}X_{j,t})^{-1}$ exists, and therefore, that the linear projection $\beta_j \equiv [E(X'_{j,t}X_{j,t})]^{-1}E(X'_{j,t}y_t)$ exists for $j = 1, 2$. Under Assumption 1, the law of large numbers can also be applied to show that $\hat{\beta}_{jt}$ converges in probability (as $t \to \infty$) to β_j.[7] One well-known property of the linear projection coefficient is that $E[X'_{j,t}(y_t - X_{j,t}\beta_j)] = 0$, which is a special case of Assumption 2b of West (1996).[8] This in turn implies the following property which is used to restrict θ below:

$$E[X_{j,t}\beta_j(y_t - X_{j,t}\beta_j)] = 0. \tag{3}$$

The predictors and prediction errors obviously depend on the estimators $\hat{\beta}_{jt}$. To reflect this, we let $\hat{\beta}_j = (\hat{\beta}_{jR}, \ldots, \hat{\beta}_{j,R+P-1})$ and $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$, and rewrite $\hat{\theta}_e$ and θ as $\hat{\theta}_e(\hat{\beta})$ and $\theta(\hat{\beta})$ respectively. As was emphasized by Clark and McCracken (2013), hypotheses like that in Eq. (1) can be interpreted in terms of the finite-sample or population-level predictive accuracy. The former concerns $\theta(\hat{\beta})$ for finite R, whereas the latter concerns $\theta(\hat{\beta})$ where $\hat{\beta}$ is replaced with $\beta = (\beta_1, \beta_2)$, giving $\theta(\beta)$. Following most of the previous work, we focus on the population-level predictive accuracy, and thus interpret Eq. (1) as "$\theta(\beta) = 0$".[9]

Assumption 1 provides a basis for efficiency gains over $\hat{\theta}_e(\hat{\beta})$ through restricted GMM estimation of the tested parameter $\theta(\beta)$. Substituting $e_{1,t+\tau} - e_{2,t+\tau} = \hat{y}_{2,t+\tau} - \hat{y}_{1,t+\tau}$ and $e_{1,t+\tau} + e_{2,t+\tau} = y_{t+\tau} - \hat{y}_{2,t+\tau} + y_{t+\tau} - \hat{y}_{1,t+\tau}$ into $\hat{\theta}_e(\hat{\beta})$ and $\theta(\beta)$ yields:

$$\begin{aligned}\hat{\theta}_e(\hat{\beta}) = {} & P^{-1}\sum_{t=R}^{P+R-1} X_{2,t+\tau}\hat{\beta}_{2t}(y_{t+\tau} - X_{2,t+\tau}\hat{\beta}_{2t}) \\ & - P^{-1}\sum_{t=R}^{P+R-1} X_{1,t+\tau}\hat{\beta}_{1t}(y_{t+\tau} - X_{1,t+\tau}\hat{\beta}_{1t}) \\ & + P^{-1}\sum_{t=R}^{P+R-1} X_{2,t+\tau}\hat{\beta}_{2t}y_{t+\tau} \\ & - P^{-1}\sum_{t=R}^{P+R-1} X_{1,t+\tau}\hat{\beta}_{1t}y_{t+\tau}\end{aligned} \tag{4}$$

$$\begin{aligned}\theta(\beta) = {} & E[X_{2,t+\tau}\beta_2(y_{t+\tau} - X_{2,t+\tau}\beta_2)] \\ & - E[X_{1,t+\tau}\beta_1(y_{t+\tau} - X_{1,t+\tau}\beta_1)] \\ & + E[X_{2,t+\tau}\beta_2 y_{t+\tau}] - E[X_{1,t+\tau}\beta_1 y_{t+\tau}].\end{aligned} \tag{5}$$

Eqs. (4) and (5) reveal that $\hat{\theta}_e(\hat{\beta})$ is equivalent to an estimator that replaces the population moments in Eq. (5) with the corresponding sample moments. As such, Eq. (4) is inefficient because it ignores the restrictions on Eq. (5) that are implied by Assumption 1, namely Eq. (3), which restricts the first two terms on the right hand side of Eq. (5) to be zero.[10] A more efficient restricted estimator of $\theta(\beta)$ that incorporates Eq. (3) is developed in the next subsection.

The tests developed in the following sections maintain Eq. (3) under the null hypothesis of equal mean squared errors: $\theta(\beta) = 0$. It is important to emphasize that the only restriction required for Eq. (3) to hold is Assumption 1. Hence, the null considered in the present paper is the same as those considered in other studies that maintain finite second-order moments and covariance stationarity under the null. It should also be noted that we do not make any assumptions about the statistical properties of the predictors in finite samples, and therefore *do not* require the predictors to be unbiased or efficient. Such properties would require the imposition of assumptions that were much more restrictive than Assumption 1, which the present approach avoids. For example, unbiasedness, which requires that $E(y_t - X_{j,t}\hat{\beta}_{jt}) = 0$ holds for finite values of t, would require the addition of the assumption that $X_{j,t}\beta_j$ is the conditional mean of y_t given $X_{j,t}$. In contrast, Assumption 1, and consequently Eq. (3), only requires covariance stationarity and the existence of certain moments.[11] Assumption 1 does not require either predictor to be based on a correctly specified model of the conditional mean or any other statistical functional.

[7] Our assumption of a "covariance-stationary process, ergodic for second moments" follows the work of Hamilton (1994, p. 47). Covariance stationarity implies that $E(X'_{j,t}X_{j,t})$ and $E(X'_{j,t}y_t)$ are finite and time-invariant, while "ergodic for second moments" implies that they are estimated consistently by their sample counterparts (Hamilton, 1994, p. 76). Consequently, $\hat{\beta}_{jt}$ converges in probability to $\beta_j \equiv [E(X'_{j,t}X_{j,t})]^{-1}E(X'_{j,t}y_t)$ if $E(X'_{j,t}X_{j,t})$ is nonsingular. For a discussion of sufficient conditions for second-moment ergodicity, see Hamilton (1994, Ch. 7).

[8] $E[X'_{j,t}(y_t - X_{j,t}\beta_j)] = 0$ is easy to verify by substituting for β_j. The quantity $X_{j,t}\beta_j$ is known as the linear projection of y_t on $X_{j,t}$, and is defined formally as the best linear predictor of y_t given $X_{j,t}$ (Hamilton, 1994, p. 74). The consistent estimation of linear projection coefficients requires much weaker assumptions than the coefficients of structural or causal regression models, which require assumptions about the functional forms of conditional means and other quantities. The linear projection exists for any set of random variables with finite variances. For further discussions, see Hamilton (1994, Ch. 4), Hanson (unpublished, Ch. 2) and Wooldridge (2010, Ch. 2). Applications of linear projections include that of Chamberlain (1982).

[9] Exceptions include Clark and McCracken (2015) and Giacomini and White (2006).

[10] The sample counterparts of Eq. (3) clearly hold for the R in-sample observations, since the least squares predictors, $\hat{y}_{1t}$ and $\hat{y}_{2t}$, are computed using these observations. However, the sample counterparts of Eq. (3) do not generally hold for the P observations that are used to compute $\hat{\theta}_e(\hat{\beta})$, and thus, they are not incorporated into the estimation of $\theta(\beta)$.

[11] The weaker assumptions imposed on the predictors can also be seen as an advantage of focusing on population-level predictive accuracy rather than finite-sample predictive accuracy. Developing an analogous approach for the latter would require Eq. (3) to hold with β replaced by $\hat{\beta}$, which, in turn, would require $X_{jt}\beta_j$ to be the conditional mean.

2.2. Restricted GMM estimation of $\theta(\beta)$

Next, we devise a more efficient restricted estimator of $\theta(\beta)$ using a GMM framework in which the moments in Eq. (5) are estimated jointly, subject to Eq. (3). Note that Eqs. (5) and (3) can be written as:

$$\theta(\beta) = 2E(X_{2,t+\tau}\beta_2 y_{t+\tau}) - E(X_{2,t+\tau}\beta_2)^2 - 2E(X_{1,t+\tau}\beta_1 y_{t+\tau}) + E(X_{1,t+\tau}\beta_1)^2 \quad (6)$$

$$E(X_{j,t+\tau}\beta_j y_{t+\tau}) = E([X_{j,t+\tau}\beta_j]^2) \quad j = 1, 2. \quad (7)$$

Let μ denote the vector of the moments in Eq. (6):

$$\mu = [E(y_{t+\tau}X_{1,t+\tau}\beta_1), E(y_{t+\tau}X_{2,t+\tau}\beta_2), E([X_{1,t+\tau}\beta_1]^2), E([X_{2,t+\tau}\beta_2]^2)]'.$$

In what follows, each element of μ is treated as a parameter to be estimated. The orthogonality condition for the GMM estimator of μ is $E[g_{t+\tau}(\mu, \beta)] = 0$, where $g_{t+\tau}(\mu, \beta) = m_{t+\tau}(\beta) - \mu$ and

$$m_{t+\tau}(\beta) = [y_{t+\tau}X_{1,t+\tau}\beta_1, y_{t+\tau}X_{2,t+\tau}\beta_2, (X_{1,t+\tau}\beta_1)^2, (X_{2,t+\tau}\beta_2)^2]'.$$

Thus, the feasible sample analog of $E[g_{t+\tau}(\mu, \beta)]$ is $\bar{g}(\mu, \hat{\beta}) = P^{-1}\sum_{t=R}^{R+P-1} g_{t+\tau}(\mu, \hat{\beta}_t)$, where $\hat{\beta}_t = (\hat{\beta}_{1t}, \hat{\beta}_{2t})$. The two restrictions in Eq. (7) can be expressed as $Q\mu = 0$, where

$$Q = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}.$$

The restricted GMM estimator of μ, $\tilde{\mu}(\hat{\beta})$, then solves the problem:

$$\min_{\mu} \bar{g}(\mu, \hat{\beta})'\hat{W}\bar{g}(\mu, \hat{\beta}), \quad \text{subject to } Q\mu = 0, \quad (8)$$

where $\hat{W}$ is a weighting matrix. The solution to Eq. (8) is:

$$\tilde{\mu}(\hat{\beta}) = \hat{A}\tilde{\mu}^*(\hat{\beta}), \quad (9)$$

where $\hat{A} = I - \hat{W}^{-1}Q'[Q\hat{W}^{-1}Q']^{-1}Q$ and $\tilde{\mu}^*(\hat{\beta}) = P^{-1}\sum_{t=R}^{R+P-1} m_{t+\tau}(\hat{\beta}_t)$ is the unrestricted GMM estimator of μ. The restricted GMM estimator of $\theta(\beta)$ is $\tilde{\theta}_y(\hat{\beta}) = c\tilde{\mu}(\hat{\beta})$, where $c = \begin{pmatrix}-2 & 2 & 1 & -1\end{pmatrix}$, while $\hat{\theta}_e(\hat{\beta}) = c\tilde{\mu}^*(\hat{\beta})$ is the unrestricted GMM estimator.

Under the following assumptions, Proposition 1 below follows from an application of West (1996, Theorem 4.1), provides the limiting distributions of $\sqrt{P}\tilde{\theta}_y(\hat{\beta})$ and $\sqrt{P}\hat{\theta}_e(\hat{\beta})$, and establishes the asymptotic efficiency of $\tilde{\theta}_y(\hat{\beta})$ relative to $\hat{\theta}_e(\hat{\beta})$.

Assumption 2. $\text{plim}_{P\to\infty}\hat{W} = W$, where W is positive definite.

Assumption 3. (a) Let $\nabla m_t(\beta) = \partial m_t(\beta)/\partial\beta$, $\varepsilon_{t,j} = y_t - X_{j,t}\beta_j$, and $\|*\|$ be the Euclidean norm. Then, for $j = 1, 2$ and some $d > 1$, $\sup_t E\|[vec(\nabla m_t(\beta))', m_t(\beta)', X'_{j,t}\varepsilon_{t,j}]'\|^{4d} < \infty$.

(b) For $j = 1, 2$ and some $d > 1$, $[vec(\nabla m_t(\beta) - E\nabla m_t(\beta))', (m_t(\beta) - Em_t(\beta))', X'_{j,t}\varepsilon_{t,j}]'$ is strong mixing, with mixing coefficients of size $-3d/(d-1)$.

(c) For $j = 1, 2$, $[vec(\nabla m_t(\beta))', m_t(\beta)', X'_{j,t}\varepsilon_{t,j}]'$ is covariance stationary.

(d) Let $V_{mm} = \sum_{j=-\infty}^{\infty}\Gamma_{mm}(j)$, where $\Gamma_{mm}(j) = E[(m_t(\beta) - Em_t(\beta))(m_{t-j}(\beta) - Em_t(\beta))]$. Then, V_{mm} is positive definite.

Assumption 4. $\lim_{R,P\to\infty} P/R = \pi$, where $0 \le \pi \le \infty$.

Assumptions 3 and 4 are the same as Assumptions 3 and 4 of West (1996, p. 1073), but with different notation. Assumption 3 restricts serial correlation, while Assumption 4 specifies the asymptotics for the numbers of in-sample and out-of-sample observations.

Proposition 1. *For $j = 1, 2$: let $B_j(t) = \left(t^{-1}\sum_{s=1}^{t} X'_{j,s}X_{j,s}\right)^{-1}$, $B(t) = diag(B_1(t), B_2(t))$, $B = p\lim_{t\to\infty} B(t)$, $H_j(t) = \left(t^{-1}\sum_{s=1}^{t} X'_{j,s}\varepsilon_{j,s}\right)$, $H(t) = [H_1(t)', H_2(t)']'$, $h_t(\beta) = \left(X_{1,t}\varepsilon_{1,t} \quad X_{2,t}\varepsilon_{2,t}\right)'$, $\Gamma_{mh}(j) = E(m_t(\beta) - Em_t(\beta))h'_{t-j}$, $\Gamma_{hh}(j) = E[h_t h'_{t-j}]$, $V_{mh} = \sum_{j=-\infty}^{\infty}\Gamma_{mh}(j)$, $V_{hh} = \sum_{j=-\infty}^{\infty}\Gamma_{hh}(j)$, $V_\beta = BV_{hh}B'$, $S = \begin{pmatrix} V_{mm} & V_{mh}B' \\ BV'_{mh} & V_\beta \end{pmatrix}$, $\Pi = 1 - \pi^{-1}\ln(1+\pi)$ for $0 < \pi < \infty$, $\Pi = 0$ for $\pi = 0$, and $\Pi = 1$ for $\pi = \infty$.*

If S is positive definite, then under Assumptions 1–4:

(i) $\sqrt{P}(\mu^*(\hat{\beta}) - \mu) \xrightarrow{dist} N(0, \Omega)$, where

$$\Omega = V_{mm} + \Pi[(E\nabla m_t(\beta))BV'_{mh} + V_{mh}B'(E\nabla m_t(\beta))'] + 2\Pi(E\nabla m_t(\beta))V_\beta(E\nabla m_t(\beta))'.$$

(ii) $\sqrt{P}(\tilde{\theta}_y(\hat{\beta}) - \theta(\beta)) \xrightarrow{dist} N(0, cA\Omega A'c')$ where $A = I - W^{-1}Q'[QW^{-1}Q']^{-1}Q$.

(iii) $\sqrt{P}(\hat{\theta}_e(\hat{\beta}) - \theta(\beta)) \xrightarrow{dist} N(0, c\Omega c')$.

(iv) Let $W = \Omega^{-1}$. Then, $c\Omega c' - cA\Omega A'c' = c\Omega Q'[Q\Omega Q']^{-1}Q\Omega c'$.

Proof. See the Appendix.

It is straightforward to use Proposition 1 to devise DMW tests when the predictors are based on non-nested models. In this case, the asymptotic variance–covariance of $\mu^*(\hat{\beta})$, Ω, is generally nonsingular and the tests are asymptotically standard normal. The original DMW test statistic is:

$$DMW(\hat{\theta}_e(\hat{\beta})) = \sqrt{P}\hat{\theta}_e(\hat{\beta})/\sqrt{c\hat{\Omega}c'}, \quad (10)$$

while the analog based on $\tilde{\theta}_y(\hat{\beta})$ is:

$$DMW(\tilde{\theta}_y(\hat{\beta})) = \sqrt{P}\tilde{\theta}_y(\hat{\beta})/\sqrt{c\hat{A}\hat{\Omega}\hat{A}'c'}, \quad (11)$$

where $\hat{\Omega}$ is a consistent estimate of Ω. It follows from parts (ii) and (iii) of Proposition 1 that Eqs. (10) and (11) are each asymptotically standard normal under the null hypothesis, $\theta(\beta) = 0$. It follows from part (iv) that $\tilde{\theta}_y(\hat{\beta})$ is asymptotically efficient relative to $\hat{\theta}_e(\hat{\beta})$ when $\tilde{\theta}_y(\hat{\beta})$ is based on the optimal choice of the weight matrix, $W = \Omega^{-1}$. The asymptotic efficiency advantage suggests that Eq. (11) may be more powerful than Eq. (10). Consistent estimation of the components of Ω is discussed by West

(1996, pp. 1074–1075). A Newey and West (1987) type estimator can be used for V_{mm}, V_{mh} and V_β; π can be estimated consistently by P/R, and B and $E\nabla m_t(\beta)$ by their sample analogues. The matrix Ω is the sum of four terms. As was noted by West (1996, p. 1072), the first term, V_{mm}, reflects the uncertainty that is present in the estimation of μ conditional on the value of β, while the remaining three terms reflect the uncertainty due to the estimation of β. If $\pi = 0$, the latter terms do not contribute to the asymptotic variance of either $\hat{\theta}_e(\hat{\beta})$ or $\tilde{\theta}_y(\hat{\beta})$. However, if $\pi \neq 0$, then the uncertainty due to the estimation of β contributes to the asymptotic variance of $\tilde{\theta}_y(\hat{\beta})$ but not to the asymptotic variance of $\hat{\theta}_e(\hat{\beta})$. This can be seen by noting that

$$\nabla m_t(\beta) = \begin{pmatrix} y_t X_{1,t} & 0 \\ 0 & y_t X_{2,t} \\ 2X_{1,t}\beta_1 X_{1,t} & 0 \\ 0 & 2X_{2,t}\beta_2 X_{2,t} \end{pmatrix},$$

and therefore $cE\nabla m_t(\beta) = 0$, since β is defined as a projection coefficient. For the optimal weighting matrix $W = \Omega^{-1}$, the expressions for the asymptotic variances of $\hat{\theta}_e(\hat{\beta})$ and $\tilde{\theta}_y(\hat{\beta})$ simplify to:

$$c\Omega c' = cV_{mm}c' \tag{12}$$

$$cA\Omega A'c' = cV_{mm}c' - c\Omega Q'(Q\Omega Q')^{-1}Q\Omega c' \tag{13}$$

respectively, where $c\Omega = cV_{mm} + c\Pi V_{mh}B'(E\nabla m_t(\beta))'$.

2.3. Nested and overlapping models

The DMW tests are less straightforward for nested and overlapping models. For these models, $X_{1t}\beta_1 = X_{2t}\beta_2$ can characterize the null hypothesis, which results in duplicated moment conditions, and consequently, a singular Ω. When $X_{1t}\beta_1 = X_{2t}\beta_2$, we have $c\Omega c' = 0$, $\sqrt{P}\hat{\theta}_e(\hat{\beta}) = o_p(1)$, and therefore the asymptotic distribution of Eq. (10) is not obvious.[12] For the special case of one-step-ahead prediction and serially uncorrelated prediction errors, Clark and McCracken (2001) and McCracken (2007) show that Eq. (10) has a non-standard distribution that is a function of Brownian motion if $\pi > 0$, and is asymptotically standard normal if $\pi = 0$.[13]

Unlike $\hat{\theta}_e(\hat{\beta})$, the asymptotic distribution of the restricted estimator $\tilde{\theta}_y(\hat{\beta})$ depends on the choice of W. One problem with nested and overlapping models is that the standard optimal choice $W = \Omega^{-1}$ is obviously not available if Ω is singular. In a different context, Peñaranda and Sentana (2012, pp. 306–308) extended Hanson's (1982) optimal GMM theory to address the problem of a singular asymptotic covariance matrix.[14] They proposed a two-step GMM estimator that is asymptotically equivalent to the infeasible optimal (asymptotically efficient) GMM based on the generalized inverse of the singular covariance matrix. For the present problem, this would entail replacing Assumption 2 with the assumption that $p\lim_{P\to\infty}\hat{\Omega}^+ = \Omega^+$ for some estimator $\hat{\Omega}^+$ of the generalized inverse Ω^+. However, it is not clear how such an estimator could be constructed for the present problem. Unlike ordinary inverses, generalized inverses are discontinuous. Consequently, as noted by Andrews (1987), consistency of $\hat{\Omega}$ for Ω does not generally ensure consistency of $\hat{\Omega}^+$ for Ω^+. Andrews (1987, Theorem 2) provides an additional necessary and sufficient condition for the latter to hold. The condition is $\text{Prob}[rank(\hat{\Omega}) = rank(\Omega)] \to 1$ as $P \to \infty$. However, this condition is unlikely to be satisfied in the present case if $\hat{\Omega}$ is a standard covariance estimator, because $\hat{\Omega}$ will generally be nonsingular if $X_{1,t+1}\hat{\beta}_{1t} \neq X_{2,t+1}\hat{\beta}_{2t}$, which will hold with probability one for all finite values of P if the components of $X_{1,t+1}$ or $X_{2,t+1}$ are distributed continuously. We leave the problem of constructing consistent estimators for Ω^+ for future research.

Yet another possible approach would be a two-step procedure in which the first step consists of testing the hypothesis that the models are overlapping against the alternative that they are non-nested. Such tests have been proposed by Clark and McCracken (2014), Marcellino and Rossi (2008) and Vuong (1989). If the test fails to reject the hypothesis of overlapping models, the procedure stops. If the hypothesis of overlapping models is rejected, the second step tests the hypothesis $\theta(\beta) = 0$ with Eq. (11) using an estimate of the optimal weight matrix for $\tilde{\theta}_y(\hat{\beta})$.

2.4. Additional moments and restrictions

Under Assumption 1, $E(X'_{j,t+\tau}(y_{t+\tau} - X_{j,t+\tau}\beta_j)) = 0$, which implies that the product of $y_{t+\tau} - X_{j,t+\tau}\beta_j$ and any linear combination of the predictor variables has zero expectation. As a consequence, many different specifications of μ and $m_{t+\tau}(\beta)$ are possible. Thus far, our specification consists of the four cross-product moments in the restriction $E[X_{j,t+\tau}\beta_j(y_{t+\tau} - X_{j,t+\tau}\beta_j)] = 0, j = 1, 2$, giving:

$$\begin{aligned} \mu &= [E(y_{t+\tau}X_{1,t+\tau}\beta_1), E(y_{t+\tau}X_{2,t+\tau}\beta_2), \\ &\quad E([X_{1,t+\tau}\beta_1]^2), E([X_{2,t+\tau}\beta_2]^2)]' \\ m_{t+\tau}(\beta) &= [y_{t+\tau}X_{1,t+\tau}\beta_1, y_{t+\tau}X_{2,t+\tau}\beta_2, (X_{1,t+\tau}\beta_1)^2, \\ &\quad (X_{2,t+\tau}\beta_2)^2]'. \end{aligned} \tag{14}$$

Six additional restrictions implied by $E(X'_{j,t+\tau}(y_{t+\tau} - X_{j,t+\tau}\beta_j)) = 0$ are $E[x_{t+\tau}\beta_{j2}(y_{t+\tau} - X_{i,t+\tau}\beta_i)] = 0$ and

12 $\sqrt{P}\hat{\theta}_e(\hat{\beta}) = o_p(1)$ follows from equation (4.1) of West (1996), since $cm_t(\beta) = 0$.

13 Hanson and Timmerman (2015) demonstrate that, for nested models, Eq. (10) can be expressed as the difference between two conventional Wald statistics testing the hypothesis that the coefficient in the larger model is zero, one statistic based on the full sample and the other on a subsample. They show that this dilutes the local power over the conventional full-sample Wald test, and argue that this raises "serious questions" about the testing of the population-level accuracy for nested models using out-of-sample tests.

14 See also Diez de los Rios (2015).

$E(y_{t+\tau} - X_{j,t+\tau}\beta_j) = 0, i, j = 1, 2$. This suggests using the following expanded vectors:

$$\begin{aligned}\dot{\mu} = [&E(y_{t+\tau}X_{1,t+\tau}\beta_1), E(y_{t+\tau}X_{2,t+\tau}\beta_2),\\ &E([X_{1,t+\tau}\beta_1]^2), E([X_{2,t+\tau}\beta_2]^2), E(y_{t+\tau}),\\ &E(X_{1,t+\tau}\beta_1), E(X_{2,t+\tau}\beta_2), E(x_{t+\tau}\beta_{12}y_{t+\tau}),\\ &E(x_{t+\tau}\beta_{22}y_{t+\tau}), E(x_{t+\tau}\beta_{12}X_{1,t+\tau}\beta_1),\\ &E(x_{t+\tau}\beta_{12}X_{2,t+\tau}\beta_2), E(x_{t+\tau}\beta_{22}X_{2,t+\tau}\beta_2),\\ &E(x_{t+\tau}\beta_{22}X_{1,t+\tau}\beta_1)]'\end{aligned}$$

$$\begin{aligned}\dot{m}_{t+\tau}(\beta) = [&y_{t+\tau}X_{1,t+\tau}\beta_1, y_{t+\tau}X_{2,t+\tau}\beta_2, [X_{1,t+\tau}\beta_1]^2,\\ &[X_{2,t+\tau}\beta_2]^2, y_{t+\tau}, X_{1,t+\tau}\beta_1, X_{2,t+\tau}\beta_2,\\ &x_{t+\tau}\beta_{12}y_{t+\tau}, x_{t+\tau}\beta_{22}y_{t+\tau},\\ &x_{t+\tau}\beta_{12}X_{1,t+\tau}\beta_1, x_{t+\tau}\beta_{12}X_{2,t+\tau}\beta_2,\\ &x_{t+\tau}\beta_{22}X_{2,t+\tau}\beta_2, x_{t+\tau}\beta_{22}X_{1,t+\tau}\beta_1]'. \qquad (15)\end{aligned}$$

It is well known from standard GMM theory that incorporating additional moments and restrictions generally improves the asymptotic efficiency. Consequently, Eq. (15) offers potential efficiency gains over Eq. (14), at least asymptotically.

Extending the theory in Section 2.2 to accommodate Eq. (15) is straightforward: it is simply a matter of replacing $m_{t+\tau}(\beta)$ and μ in Proposition 1 with $\dot{m}_{t+\tau}(\beta)$ and $\dot{\mu}$. Using Eq. (15) instead of Eq. (14) does not change $\hat{\theta}_e(\hat{\beta})$, since unrestricted GMM is equivalent to equation-by-equation OLS in the present setting. However, it does change the asymptotic covariance of $\sqrt{P}\tilde{\theta}_y(\beta)$, and consequently, the optimal weighting matrix.

3. Simulation experiments

3.1. Design

Simulation experiments were conducted to examine the finite-sample properties of the various estimators and tests. Samples were drawn from the following model:

$$y_{t+1} = 0.3y_t + \delta_1 x_{1,t} + \delta_2 x_{2,t} + u_{t+1}, \qquad (16)$$

where u_t and $x_{i,t}$ are independent and serially uncorrelated, $u_t \sim N(0, 10)$ and $x_{i,t} \sim N(0, 0.5)$. The tests were evaluated for one-step-ahead ($\tau = 1$) predictions from the two models:

$$y_{t+1} = \beta_{10} + \beta_{11}x_{1,t} + \beta_{12}y_t + u_{1,t+1} \qquad (17a)$$

$$y_{t+1} = \beta_{20} + \beta_{21}x_{2,t} + \beta_{22}y_t + u_{2,t+1}, \qquad (17b)$$

which were estimated by OLS, and a recursive scheme was used to generate predictions.

In all experiments, tests were conducted at the 10% significance level and for 5000 replications. Experiments were run for $P = 20, 50, 100, 200$, with $R = jP$ and $j = 2, 3, 4, 5$. Samples were generated from Eq. (16) using a non-nested specification in which we set $\delta_1 = -2$ in all experiments. The expected value of the squared prediction error in this case equals $10 + 0.5\delta_2^2$ for Eq. (17a) 12 for Eq. (17b). We set $\delta_2 = -2$ in the size experiments, and $\delta_2 = -1$ in the power experiments. In what follows, the DMW test based on a given estimator is denoted by DMW("estimator"). The restricted GMM estimator based on the optimal weight matrix for j moments is denoted by $\tilde{\theta}_{y(j)}(\hat{\beta})$, $j = 4, 13$, and computed from the two-step GMM estimator $\tilde{\mu}(\hat{\beta})$ given by Eq. (9). The first-step estimator is OLS. The weight matrix is the inverse of a consistent estimate of the asymptotic covariance. The latter covariance estimator is serial-correlation robust, and uses a Bartlett kernel and the automatic lag-selection algorithm proposed by Newey and West (1994).

Table 1
Standard deviations under the null hypothesis.

P	R	$\hat{\theta}_e(\hat{\beta})$	$\tilde{\theta}_{y(13)}(\hat{\beta})$	$\tilde{\theta}_{y(4)}(\hat{\beta})$
20	40	3.21	2.16	2.18
20	60	3.18	1.93	1.94
20	80	3.25	1.91	1.92
20	100	3.04	1.78	1.79
50	100	1.97	1.27	1.33
50	150	1.86	1.15	1.22
50	200	1.90	1.11	1.18
50	250	1.89	1.03	1.09
100	200	1.38	0.881	0.947
100	300	1.34	0.809	0.888
100	400	1.34	0.755	0.816
100	500	1.33	0.726	0.793
200	400	0.966	0.627	0.684
200	600	0.954	0.547	0.618
200	800	0.945	0.510	0.567
200	1000	0.942	0.483	0.529

3.2. Relative efficiencies of the GMM estimators

We have argued that tests based on $\tilde{\theta}_{y(4)}(\hat{\beta})$ and $\tilde{\theta}_{y(13)}(\hat{\beta})$ may have greater power than those based on $\hat{\theta}_e(\hat{\beta})$ because of the asymptotic efficiency of restricted estimation. Before evaluating the tests, we first examine the efficiency issue by comparing the sample means and standard deviations of $\hat{\theta}_e(\hat{\beta})$, $\tilde{\theta}_{y(4)}(\hat{\beta})$ and $\tilde{\theta}_{y(13)}(\hat{\beta})$ for the samples used to evaluate the tests. Under the null hypothesis of equal mean squared prediction errors, the standard deviations and means are reported in Tables 1 and 2 respectively. In all cases, the sample means are much smaller than the standard deviations, suggesting that biasedness is not a significant source of estimation error. Consistent with the asymptotic efficiency of restricted estimation, the standard deviations of $\tilde{\theta}_{y(13)}(\hat{\beta})$ and $\tilde{\theta}_{y(4)}(\hat{\beta})$ are smaller than those of $\hat{\theta}_e(\hat{\beta})$ for all P and R. Also as expected, the standard deviations of $\tilde{\theta}_{y(4)}(\hat{\beta})$ are (slightly) larger than those of $\tilde{\theta}_{y(13)}(\hat{\beta})$. Averaging the percentages over R, the standard deviation of $\tilde{\theta}_{y(13)}(\hat{\beta})$ is about 61% of the standard deviation of $\hat{\theta}_e(\hat{\beta})$ for $P = 20$, 59% of that for $P = 50$ and $P = 100$, and 57% of that for $P = 200$. The efficiency gains generally increase as P/R decreases. The percentages range from 64% to 67% for $P/R = 1/2$, from 57% to 61% for $P/R = 1/3$, from 52% to 59% for $P/R = 1/4$, and from 52% to 55% for $P/R = 1/5$. The efficiency gains for $\tilde{\theta}_{y(4)}(\hat{\beta})$ over $\hat{\theta}_e(\hat{\beta})$ are only slightly smaller.

3.3. Size and power results

Next, we evaluate the size and power of the tests. For all tests, the estimator of the asymptotic covariance

Table 2
Means under the null hypothesis.

P	R	$\hat{\theta}_e(\hat{\beta})$	$\tilde{\theta}_{y(13)}(\hat{\beta})$	$\tilde{\theta}_{y(4)}(\hat{\beta})$
20	40	−0.019	−0.026	−0.036
20	60	−0.196	0.013	0.019
20	80	−0.097	0.046	−0.043
20	100	−0.094	−0.082	−0.066
50	100	0.103	−0.048	−0.053
50	150	0.153	0.069	−0.068
50	200	−0.044	−0.021	0.032
50	250	0.076	−0.037	0.023
100	200	−0.033	−0.041	−0.017
100	300	0.016	0.004	0.004
100	400	0.011	−0.023	−0.025
100	500	0.016	0.009	0.010
200	400	0.012	0.021	0.002
200	600	0.001	−0.007	−0.007
200	800	0.003	0.039	0.016
200	1000	−0.007	−0.020	−0.006

matrix Ω follows the discussion below Proposition 1. The estimator uses a serial-correlation robust Newey–West estimator that is specified with a Bartlett kernel and the automatic lag-selection algorithm proposed by Newey and West (1994). It is important to recognize that the choice of a truncation lag can affect the size and power of the tests in finite samples. The advantage of the lag-selection algorithm is that the truncation lag is at least optimal in an asymptotic mean square error sense.[15] As Proposition 1 shows, Ω also depends on the limit of P/R, π. When R is large relative to P, one might consider setting $\pi = 0$, which greatly simplifies the estimated asymptotic covariance. However, initial experiments indicated that $DMW(\tilde{\theta}_{y(4)}(\hat{\beta}))$ and $DMW(\tilde{\theta}_{y(13)}(\hat{\beta}))$ were greatly oversized for all P and R when $\pi = 0$ was imposed. For example, for the 10% nominal size under the null hypothesis, the rejection rates for $DMW(\tilde{\theta}_{y(4)}(\hat{\beta}))$ with $\pi = 0$ imposed ranged from an average of 30% for $P = 200$ to an average of 57% for $P = 20$. Although $\hat{\beta}$ is asymptotically irrelevant in the distributions of $\tilde{\theta}_{y(4)}(\hat{\beta})$ and $\tilde{\theta}_{y(13)}(\hat{\beta})$ if the true value of π is zero, the distributions generally depend on the distribution of $\hat{\beta}$ in finite samples. Imposing $\pi = 0$ in finite samples neglects this dependence, meaning that the estimated asymptotic variance may be a poor approximation to the finite-sample variance. Consistent with this, the sizes of $DMW(\tilde{\theta}_{y(4)}(\hat{\beta}))$ and $DMW(\tilde{\theta}_{y(13)}(\hat{\beta}))$ improved considerably when $\pi = 0$ was not imposed. As is discussed below, the rejection rates of $DMW(\tilde{\theta}_{y(4)}(\hat{\beta}))$ and $DMW(\tilde{\theta}_{y(13)}(\hat{\beta}))$ are typically within 2%–3% of the nominal size when $\pi = 0$ is not imposed. In contrast, the value of π is not an issue for $DMW(\hat{\theta}_e(\hat{\beta}))$. As Eqs. (12) and (13) reveal, unlike $\tilde{\theta}_{y(4)}(\hat{\beta})$ and $\tilde{\theta}_{y(13)}(\hat{\beta})$, $\hat{\beta}$ is asymptotically irrelevant in the distribution of $\hat{\theta}_e(\hat{\beta})$ for all values of π.

Table 3 reports the rejection rates under the null hypothesis for a nominal size of 10%. In line with previous studies, the standard form of the DM test, $DMW(\hat{\theta}_e(\hat{\beta}))$, performs well in terms of size. It is slightly oversized in most cases, but the rejection rates are within 2% of the nominal size for all cases in which $P \geq 50$ and within about 3% for $P = 20$. Clark and McCracken (2014, pp. 425–426) also found $DMW(\hat{\theta}_e(\hat{\beta}))$ to be slightly oversized for similar sample sizes. They conjecture that both P and R might have to be larger "for the asymptotics to kick in". In all but two cases, the rejection rates of $DMW(\tilde{\theta}_{y(13)}(\hat{\beta}))$ are within about 2% of the nominal size for $P \geq 50$ with $P/R \leq 1/3$. The rejection rates of $DMW(\tilde{\theta}_{y(4)}(\hat{\beta}))$ are somewhat better, as they are within 2% of the nominal size for $P \geq 20$ with $P/R \leq 1/3$ for 14 cases out of 15. Therefore, $DMW(\tilde{\theta}_{y(4)}(\hat{\beta}))$ performs about as well as $DMW(\hat{\theta}_e(\hat{\beta}))$ in these cases. However, one difference is that, whereas $DMW(\hat{\theta}_e(\hat{\beta}))$ tends to be slightly oversized, $DMW(\tilde{\theta}_{y(4)}(\hat{\beta}))$ and $DMW(\tilde{\theta}_{y(13)}(\hat{\beta}))$ tend to be undersized, and therefore are more conservative. It should also be noted that the rejection rate of $DMW(\hat{\theta}_e(\hat{\beta}))$ is generally less sensitive to P/R than that of either $DMW(\tilde{\theta}_{y(4)}(\hat{\beta}))$ or $DMW(\tilde{\theta}_{y(13)}(\hat{\beta}))$. Again, this might be due to the fact that the asymptotic variance of $\hat{\theta}_e(\hat{\beta})$ does not depend on π.

Table 4 reports the rejection rates under the alternative hypothesis. In terms of power, the rejection rate of $DMW(\hat{\theta}_e(\hat{\beta}))$ is also less sensitive to the value of P/R than the other tests are. For the case of $P \geq 50$ with $P/R \leq 1/3$, in which the tests have similar sizes, $DMW(\tilde{\theta}_{y(13)}(\hat{\beta}))$ and $DMW(\tilde{\theta}_{y(4)}(\hat{\beta}))$ are much more powerful than $DMW(\hat{\theta}_e(\hat{\beta}))$. For example, with $P/R = 1/5$, the rejection rate of $DMW(\tilde{\theta}_{y(13)}(\hat{\beta}))$ is about 2.3 times greater than that of $DMW(\hat{\theta}_e(\hat{\beta}))$ for $P = 50$, 2.1 times greater for $P = 100$ and 1.5 times greater for $P = 200$. As expected, $DMW(\tilde{\theta}_{y(4)}(\hat{\beta}))$ is less powerful than $DMW(\tilde{\theta}_{y(13)}(\hat{\beta}))$ but more powerful than $DMW(\hat{\theta}_e(\hat{\beta}))$.

4. Illustrative empirical application

Next, we illustrate the tests using a forecasting application to monthly US industrial production. The data were downloaded from the FRED website of the Federal Reserve Bank of St. Louis. One forecasting model is a regression of the growth rate of the industrial production index on a constant, one lag of the growth rate and two lags of the spread between Moody's Aaa and Baa corporate bond yields. The other model replaces the credit spread with the log of housing starts for privately owned housing. The data are monthly from 1959:03 to 2015:08. We applied the tests to one-month-ahead forecasts for the period 2011:06–2015:08. The models were estimated recursively. Using 4999 bootstrap draws, we then applied the bootstrap test of Clark and McCracken (2014, p. 421) to determine whether the models are overlapping or non-nested. The null hypothesis that the models are overlapping can be rejected at the 5% level. For the null hypothesis of equal predictive accuracy, we obtain $DMW(\hat{\theta}_e(\hat{\beta})) = 1.47$, which is insignificant at the 10% level, but $DMW(\tilde{\theta}_{y(13)}(\hat{\beta})) = 3.025$ and $DMW(\tilde{\theta}_{y(4)}(\hat{\beta})) = 2.162$, which are significant at the 1% and 5% levels, respectively.

15 For a summary of the parameters used in the lag-selection algorithm, see Table 1 of Newey and West (1994).

Table 3
Rejection rates for a nominal size of 10% under the null hypothesis.

P	R	$DMW(\hat{\theta}_e(\hat{\beta}))$	$DMW(\tilde{\theta}_{y(13)}(\hat{\beta}))$	$DMW(\tilde{\theta}_{y(4)}(\hat{\beta}))$
20	40	0.116	0.070	0.053
20	60	0.131	0.103	0.085
20	80	0.142	0.157	0.109
20	100	0.129	0.178	0.120
50	100	0.119	0.067	0.069
50	150	0.112	0.090	0.088
50	200	0.093	0.104	0.109
50	250	0.109	0.144	0.115
100	200	0.112	0.064	0.069
100	300	0.099	0.093	0.089
100	400	0.103	0.105	0.092
100	500	0.103	0.117	0.112
200	400	0.099	0.056	0.073
200	600	0.107	0.057	0.075
200	800	0.115	0.083	0.090
200	1000	0.091	0.085	0.095

Table 4
Rejection rates for a nominal size of 10% under the alternative hypothesis.

P	R	$DMW(\hat{\theta}_e(\hat{\beta}))$	$DMW(\tilde{\theta}_{y(13)}(\hat{\beta}))$	$DMW(\tilde{\theta}_{y(4)}(\hat{\beta}))$
20	40	0.189	0.160	0.087
20	60	0.192	0.218	0.157
20	80	0.181	0.323	0.257
20	100	0.194	0.380	0.295
50	100	0.281	0.348	0.339
50	150	0.272	0.533	0.452
50	200	0.278	0.589	0.516
50	250	0.282	0.653	0.577
100	200	0.433	0.652	0.618
100	300	0.420	0.778	0.723
100	400	0.439	0.874	0.811
100	500	0.424	0.910	0.842
200	400	0.655	0.898	0.889
200	600	0.626	0.976	0.945
200	800	0.643	0.986	0.973
200	1000	0.672	0.996	0.986

5. Conclusion

The test developed by Diebold and Mariano (1995) and West (1996) is regarded widely as being an important test that addresses the need to formally assess whether differences in the accuracies of two predictors are purely sampling error. Using a GMM framework, we proposed a more powerful version that can be used to compare the accuracies of regression predictors based on non-nested models. One advantage of the version proposed by Diebold and Mariano (1995) is that it circumvents the problem of imposing assumptions on the predictors by imposing assumptions on the prediction errors directly. However, as we demonstrated, this comes at a cost in terms of power. Specifically, we showed that more powerful versions of the tests can be obtained by exploiting the properties of linear projections that only require assumptions as to the existence of certain moments and covariance-stationarity. The simulation experiments illustrate that the potential gains in power can be considerable. Possible directions for future research include extensions to the hypothesis of finite-sample (as opposed to population-level) predictive accuracy, and the estimation of the optimal weight matrix when the covariance matrix is singular for tests of overlapping and nested models.

Acknowledgments

The authors would like to thank the editor Michael McCracken, the associate editor, and an anonymous referee for helpful comments.

Appendix

Part (i) of Proposition 1 can be proved by showing that Assumptions 1–4 of West (1996) hold for $m_t(\beta)$. The result then follows directly from Theorem 4.1 of West (1996), which also assumes a recursive forecasting scheme. As was noted above, our Assumptions 3 and 4 are Assumptions 3 and 4 of West (1996) with different notation. Clearly, $m_t(\beta)$ is measurable and twice continuously differentiable, with second-order derivatives that do not depend on β. Under our Assumption 1, the second-order derivatives have finite expectations. Thus, Assumption 1 of West (1996) holds. Next, note that under our Assumption 1 $\hat{\beta}_t - \beta = B(t)H(t)$, where $B(t) = diag(B_1(t), B_2(t))$, $H(t) =$

$[H_1(t)', H_2(t)']' = t^{-1}\sum_{s=1}^{t} h_s(\beta)$, $B = p\lim_{t\to\infty} B(t)$, B has full column rank, and $Eh_t(\beta) = 0$. Therefore, Assumption 2 of West (1996) holds. Consequently, part (i) follows from West (1996, Theorem 4.1).

Part (ii) follows from part (i) by noting that $\hat{A}$ converges in probability to A under Assumption 2 and $\sqrt{P}(\tilde{\theta}_y(\hat{\beta}) - \theta(\beta)) = c\hat{A}\sqrt{P}(\mu^*(\hat{\beta}) - \mu)$, since $Q\mu = 0$ under Assumption 1.

Part (iii) follows from part (i) by noting that $\sqrt{P}(\hat{\theta}_e(\hat{\beta}) - \theta(\beta)) = c\sqrt{P}(\mu^*(\hat{\beta}) - \mu)$. Part (iv) then follows from substituting $W = \Omega^{-1}$ into A and multiplying out $cA\Omega A'c'$.

References

Andreou, E., Ghysels, E., & Kourtellos, A. (2013). Should macroeconomic forecasters use daily financial data and how? *Journal of Business and Economic Statistics*, *31*(1), 240–251.

Andrews, D. W. K. (1987). Asymptotic results for generalized Wald tests. *Econometric Theory*, *3*, 348–358.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, *18*, 4–46.

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

Clark, T. E., & McCracken, M. W. (2006). The predictive content of the output gap for inflation: resolving in-sample and out-of-sample evidence. *Journal of Money, Credit and Banking*, *38*(5), 1127–1148.

Clark, T. E., & McCracken, M. W. (2013). Advances in forecast evaluation. In *Handbook of economic forecasting, Vol. 2* (pp. 1107–1201).

Clark, T. E., & McCracken, M. W. (2014). Tests of equal forecast accuracy for overlapping models. *Journal of Applied Econometrics*, *29*, 415–430.

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

Corradi, V., Swanson, N. R., & Olivetta, C. (2001). Predictive ability with cointegrated variables. *Journal of Econometrics*, *104*, 315–358.

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

Diez de los Rios, A. (2015). Optimal asymptotic least squares estimation in a singular set-up. *Economics Letters*, *128*, 83–86.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*(6), 1545–1578.

Hamilton, J. D. (1994). *Time series analysis*. Princeton: Princeton University Press.

Hanson, L. P. (1982). Large sample properties of generalized method of moment estimators. *Econometrica*, *50*, 1029–1054.

Hanson, B.E. 2016. Econometrics, (unpublished textbook manuscript), University of Wisconsin.

Hanson, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a GAR(1,1) model? *Journal of Applied Econometrics*, *20*, 873–889.

Hanson, P. R., & Timmerman, A. (2015). Equivalence between out-of-sample forecast comparisons and Wald statistics. *Econometrica*, *83*(6), 2485–2505.

Hong, Y., & Lee, T. H. (2003). Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. *Review of Economics and Statistics*, *85*(4), 1048–1062.

Marcellino, M., & Rossi, B. (2008). Model selection for nested and overlapping nonlinear, dynamic and possibly misspecified models. *Oxford Bulletin of Economics and Statistics*, *70*, 867–893.

Mark, N. C. (1995). Exchange rates and fundamentals: evidence on long-horizon predictability. *American Economic Review*, *85*, 201–218.

McCracken, M. W. (2007). Asymptotics for out of sample tests of Granger causality. *Journal of Econometrics*, *140*(2), 719–752.

Naes, R., Skjeltorp, J. A., & Ødegaard, B. A. (2011). Stock market liquidity and business cycle. *Journal of Finance*, *66*, 139–176.

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

Newey, W. K., & West, K. D. (1994). Automatic lag selection in covariance estimation. *Review of Economic Studies*, *61*, 631–653.

Peñaranda, F., & Sentana, E. (2012). Spanning tests and stochastic discount mean–variance Frontiers: a unifying approach. *Journal of Econometrics*, *170*, 303–324.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: John Wiley & Sons.

Stock, J. H., & Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics*, *20*(2), 147–162.

Stock, J. H., & Watson, M. W. (2007). Why has US inflation become harder to forecast? *Journal of Money, Credit and Banking*, *39*(1), 3–33.

Swanson, N. R., & White, H. (1997). A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. *Review of Economics and Statistics*, *79*(4), 540–550.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*, 307–333.

West, K.D. 1994. Asymptotic inference about predictive ability, (Manuscript), University of Wisconsin.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, *64*, 1067–1084.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: MIT Press.

Walter J. Mayer is Professor of Economics and Graduate Program Coordinator at the University of Mississippi, where he teaches undergraduate and graduate econometrics. He has published articles in various journals, including the *Journal of Econometrics, Journal of Business and Economic Statistics, International Journal Forecasting, Economics Letters, Computational Economics, and Journal of Housing Economics*. He received a B.A. in Economics from the University of Missouri-St. Louis and a Ph.D. in Economics from the University of Florida.

Feng Liu is a Ph.D. Candidate in Economics at the University of Mississippi. Feng obtained her master's degree in Economics at Shanghai University of Finance and Economics in 2011.

Xin Dang received a Ph.D. in Statistics from the University of Texas at Dallas in 2005. Currently she is an associate professor in the Department of Mathematics at the University of Mississippi. Her research interests include robust and nonparametric statistics, statistical computing, multivariate data analysis, and econometrics. In particular, she has focused on statistical learning, data depth and application, and robust procedures. Dr. Dang is a member of the Institute of Mathematical Statistics, the International Institute of Forecasters and the International Chinese Statistical Association.