7-2012

# Statistical tests for multiple forecast comparison

Roberto MARIANO
*Singapore Management University*, rsmariano@smu.edu.sg

Daniel P. A. PREVE
*Singapore Management University*, dpreve@smu.edu.sg

## Citation

MARIANO, Roberto and PREVE, Daniel P. A.. Statistical tests for multiple forecast comparison. (2012). *Journal of Econometrics*. 169, (1), 123-130.
**Available at:** https://ink.library.smu.edu.sg/soe_research/2331

# STATISTICAL TESTS FOR MULTIPLE FORECAST COMPARISON

ROBERTO S. MARIANO[†] AND DANIEL PREVE[‡]

ABSTRACT. We consider a multivariate version of the Diebold-Mariano test for equal predictive ability of three or more forecasting models. The Wald-type test, $S$, which has a null distribution that is asymptotically chi-squared, is shown to be generally invariant with respect to the ordering of the models being compared. Finite-sample corrections for the test are also developed. Monte Carlo simulations indicate that $S$ has reasonable size properties in large samples but tends to be oversized in moderate samples. The finite-sample correction succeeds in correcting for size, but only partially. For the size-adjusted tests, power increases with sample size, as expected. It is speculated that further finite-sample improvements can be achieved using Hotelling's $T^2$ or bootstrap critical values.

*JEL classification:* C12, C52

*Keywords:* Forecast comparison; Multivariate tests of equal predictive ability; Diebold-Mariano test; Finite-sample correction

[†] University of Pennsylvania and Singapore Management University. [‡] Uppsala University. Address correspondence to Daniel Preve, Department of Statistics, Uppsala University, Box 513, 751 20 Uppsala, Sweden. Tel.: +46 18-471 11 39; fax: +46 18-471 78 67; E-mail address: daniel.preve@statistics.uu.se.
The authors thank two anonymous referees, Rolf Larsson, Bent Nielsen and conference participants at the Society for Financial Econometrics Inaugural Conference (2008, New York), the Conference in Honor of Peter C. B. Phillips (2008, Singapore), the Far Eastern and South Asian Meeting of the Econometric Society (2008, Singapore) and seminar participants at the Nanyang Technological University for their comments and suggestions. The authors are responsible for remaining errors. Both authors gratefully acknowledge partial support from the Sim Kee Boon Institute for Financial Economics at Singapore Management University. The second author gratefully acknowledges research support from the Jan Wallander and Tom Hedelius Research Foundation (grant P 2006-0166:1).

## 1. Introduction

In empirical applications it is often the case that two or more time series models are available for forecasting a particular variable of interest. For example, in econometrics different assumptions regarding the nature of the data generating process of an economic variable can result in a number of different forecasting models. With forecasts from a number of alternative models it is inevitable that the sample will show differences in predictive ability between the different models. Consequently it is of importance to investigate how likely it is that this outcome is due to pure chance, that is, whether the observed difference is statistically significant or not. Various tests have been proposed to test whether one or more of a number of alternative models stand out in terms of predictive ability. See, for example, White (2000), Hansen (2005), Romano and Wolf (2005) and Hansen et al. (2011). Diebold and Mariano (1995, DM) proposed an asymptotically standard normally distributed test statistic for equal predictive ability (EPA) between two alternative models. West (1996) took into account that the actual forecasts that appear in a test statistic may depend on estimated parameters. A sizeable literature has also developed concerning tests for nested models. See, among others, Giacomini and White (2006), Clark and West (2006), Clark and West (2007) and McCracken (2007).

In this paper we consider a multivariate version of the DM test for EPA of two or more non-nested forecasting models. Our framework differs from that of White (2000), Hansen (2005) and Romano and Wolf (2005) in that we test for EPA, whereas they test for superior predictive ability (SPA). Tests for EPA form a natural basis for more elaborate procedures, such as the model confidence set (MCS) of Hansen et al. (2011). As explained in Hansen et al. (2011), the MCS has a number of advantages over tests for SPA.

A number of 'model-free' tests for EPA (in the sense that the models that generated the forecasts need not be at one's disposal), that compare the forecasts of two alternative time series models, are available, for example, see Mariano (2002). Some commonly used head-to-head tests are the Morgan (1939) and Granger and Newbold (1977), and the Meese and Rogoff (1988) tests for equal mean squared errors, and Christiano's (1989) test for equal root mean square errors. DM used standard results to derive a test statistic in a more general setting. In their approach, they consider two sequences of forecasts ($\hat{y}_{11}, ..., \hat{y}_{1P}$ and $\hat{y}_{21}, ..., \hat{y}_{2P}$ say) of a scalar time series $\{y_t\}_{t=1}^{P}$ and propose a conceptually simple test to assess the expected loss associated with each of the forecast sequences. The quality of each forecast is evaluated by some real-valued loss function $g(\cdot)$ of the forecast error. Important examples of $g$ include $g(x) = x^2$ (squared loss) and $g(x) = |x|$ (absolute loss). In this setting, the null hypothesis of EPA is $E\, d_t = 0$ where

$$d_t = g(\hat{y}_{1t} - y_t) - g(\hat{y}_{2t} - y_t), \tag{1.1}$$

is the 'loss differential' at time $t$. Under the assumption that the time series $\{d_t\}$ is covariance stationary, they conclude that the statistic

$$\frac{\bar{d}}{\sqrt{\hat{\omega}/P}}, \tag{1.2}$$

is asymptotically standard normally distributed under the null, where $\bar{d}$ is the sample mean of the loss differential series and $\hat{\omega}$ is a consistent estimator of the asymptotic variance of $\sqrt{P}\bar{d}$, sometimes referred to as the 'long run variance' of $\bar{d}$.

Harvey et al. (1997, HLN) addressed the finite-sample properties of the DM statistic. Under the additional assumption that all autocovariances of $\{d_t\}$ beyond some lag length $q$ are zero such that

$$\omega = \gamma(0) + 2\sum_{h=1}^{q} \gamma(h),$$

where $\gamma(h)$ is the $h$th autocovariance, they propose a finite-sample correction of the DM test based on an approximately unbiased estimator of the variance of $\bar{d}$. HLN argue that their assumption for the autocovariances can be motivated by the fact that for optimal $n$-step-ahead forecasts the sequence of forecast errors follows a moving average (MA) process of order $(n-1)$, and that this result can be expected to hold at least approximately for many sets of forecasts.

The rest of the paper is organized as follows. In Section 2 we consider a multivariate version of the DM test for EPA and show that it is invariant with respect to the ordering of the alternative forecasting models for a wide range of covariance matrix estimators. In Section 3 we show that the finite-sample correction of HLN for the DM test extends to our multivariate setting. Section 4 reports various simulation results and Section 5 concludes. In the later section we briefly discuss issues of parameter estimation uncertainty and how to proceed once the null hypothesis of EPA has been rejected. Mathematical proofs are collected in the Appendix. An extended Appendix available on request from the authors contains some results mentioned in the text but omitted from the paper to save space.

## 2. A Multivariate Version of the DM Test

In this paper, we consider the setting where we have a small number of non-nested forecasting models, say, in the single digits, that are to be compared in terms of a general loss function. Like the DM test, the test we consider is model-free. That is, in contrast to West (1996), we only assume that the information at the disposal of an analyst consists of time series of forecasts and actual values of the predictand. The models that generated the forecasts, and their associated estimators, are *potentially* unknown. For example, this situation arises when financial institutions provide forecasts without disclosing the models that generated the forecasts or in judgmental forecasting where the forecaster may use non-time series information to improve the forecasts (see Lawrence et al. 2006 for a review of judgmental forecasting techniques).

We are interested to know whether all of the models perform equally well in terms of a specific loss function. Let

$$\{e_{it}\} = \{\hat{y}_{it} - y_t\}, \quad i = 1, ..., k+1$$

be $k+1$ time series of forecast errors from $k+1$ alternative models, and let $g : \mathbf{R} \to \mathbf{R}$ denote some specified loss function. For example, the $y_{it}$ could be moving average, or filter, rules. We wish to test the hypothesis that

$$E\, g(e_{1t}) = E\, g(e_{2t}) = ... = E\, g(e_{k+1,t}), \tag{2.1}$$

in other words, that all alternative models have EPA under the loss function $g$. An equivalent way of stating hypothesis (2.1) is $E\,g(e_{jt}) = E\,g(e_{j+1,t})$ for all $j = 1, ..., k$. Define

$$d_{jt} = g(e_{jt}) - g(e_{j+1,t}), \quad j = 1, ..., k \tag{2.2}$$

and consider the $k$ loss differential series $\{d_{jt}\}$ and vector

$$\mathbf{d}_t = (d_{1t}, ..., d_{kt})', \quad t = 0, \pm 1, \pm 2, ... \tag{2.3}$$

Under hypothesis (2.1) $E\,\mathbf{d}_t = \mathbf{0}$. Hence, it is natural to base a test for EPA on the vector of observed sample means,

$$\bar{\mathbf{d}} = \frac{1}{P} \sum_{t=1}^{P} \mathbf{d}_t. \tag{2.4}$$

Suppose that $\{\mathbf{d}_t\}$ is covariance stationary with Wold representation

$$\mathbf{d}_t = \boldsymbol{\mu} + \boldsymbol{\epsilon}_t + \boldsymbol{\Psi}_1 \boldsymbol{\epsilon}_{t-1} + \boldsymbol{\Psi}_2 \boldsymbol{\epsilon}_{t-2} + ... \tag{2.5}$$

In this setting

$$\sqrt{P}(\bar{\mathbf{d}} - \boldsymbol{\mu}) \xrightarrow{d} N_k(\mathbf{0}, \boldsymbol{\Omega}),$$

as $P \to \infty$ under well-known conditions, where

$$\boldsymbol{\Omega} = \boldsymbol{\Gamma}(0) + \sum_{h=1}^{\infty} \big[\boldsymbol{\Gamma}(h) + \boldsymbol{\Gamma}'(h)\big], \tag{2.6}$$

is the long run variance and

$$\boldsymbol{\Gamma}(h) = E\,(\mathbf{d}_t - \boldsymbol{\mu})(\mathbf{d}_{t-h} - \boldsymbol{\mu})', \quad h = 0, 1, 2, ...$$

are the (auto)covariance matrices of $\{\mathbf{d}_t\}$. Thus, the following proposition holds, providing a multivariate version of the DM test.

**Proposition 1.** *Suppose that $\boldsymbol{\Omega}$ is nonsingular, then*

$$P\,(\bar{\mathbf{d}} - \boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\bar{\mathbf{d}} - \boldsymbol{\mu}) \xrightarrow{d} \chi^2_k,$$

*as $P \to \infty$ and the Wald statistic*

$$S = P\,\bar{\mathbf{d}}'\hat{\boldsymbol{\Omega}}^{-1}\bar{\mathbf{d}},$$

*where $\hat{\boldsymbol{\Omega}}$ is a consistent estimator of $\boldsymbol{\Omega}$, has a limiting chi-square distribution with k degrees of freedom under the null hypothesis of EPA ($\boldsymbol{\mu} = \mathbf{0}$).*

The situation we have in mind is one where the number of alternative models, $k+1$, is small relative to the sample size, $P$, so that the $k \times k$ covariance matrix $\boldsymbol{\Omega}$ can be reliably estimated (cf. Hansen 2005 and Hansen et al. 2011), and where the models are non-nested to avoid a potential non-chi-square limiting distribution (cf. McCracken 2007). Typically $\boldsymbol{\Omega}$ would be consistently estimated using a heteroskedasticity and autocorrelation consistent (HAC) covariance matrix estimator. The $S$ statistic of Proposition 1 is along the lines of that considered by West et al. (1993) and West and Cho (1995), among others.

By construction, any reordering of the alternative models, and hence of the $k+1$ time series of forecast errors, alters the dynamics of $\mathbf{d}_t$ in (2.3). The next proposition gives conditions under which the limiting null distribution of $S$, and the test value, is unaffected by reordering.

**Proposition 2.** *Suppose that $\hat{\mathbf{\Omega}}$ is of the form*

$$\hat{\mathbf{\Omega}} = \hat{\mathbf{\Gamma}}(0) + \sum_{h=1}^{m} \kappa(h,m)\big[\hat{\mathbf{\Gamma}}(h) + \hat{\mathbf{\Gamma}}'(h)\big],$$

*where $m(P)$ is an integer-valued truncation point, $\kappa(\cdot,\cdot)$ is a real-valued kernel weight and $\hat{\mathbf{\Gamma}}(h)$ is the sample covariance matrix at lag $h$. Then the $S$ statistic of Proposition 1 is invariant to any permutation (reordering) of the alternative models.*

That is, for each permutation of the models, we get the same limiting $\chi^2_k$ distribution under the null of EPA. Moreover, when computing $S$, we get the same test value for all permutations of the models (regardless of whether the null is true or not).

In view of Proposition 2 invariance holds quite generally. For example, invariance holds for test statistics $S$ using the Newey and West (1987) estimator, Andrews (1991) estimators and the truncated estimator in (3.1). Invariance also holds for nonsingular transformations of any ordering, such as the chi-square test considered in Hubrich and West (2010) which looks at the vector of loss differentials relative to one of the forecasting procedures.

## 3. A Finite-Sample Correction

In this section we provide a modified test with potentially better finite-sample properties than the $S$ statistic presented in the previous section. In so doing, we show that the finite-sample correction of HLN for the DM test extends to our multivariate setting.

Following HLN, we now assume further that $\{\mathbf{d}_t\}$ can be represented by a finite $q$th order vector moving average process. In this setting $\mathbf{\Gamma}(h)$ in (2.6) equals $\mathbf{0}$ for every $h > q$ and $\mathbf{\Omega}$ can be consistently estimated by the truncated estimator

$$\hat{\mathbf{\Omega}} = \hat{\mathbf{\Gamma}}(0) + \sum_{h=1}^{q} \big[\hat{\mathbf{\Gamma}}(h) + \hat{\mathbf{\Gamma}}'(h)\big], \tag{3.1}$$

where

$$\hat{\mathbf{\Gamma}}(h) = \frac{1}{P} \sum_{t=h+1}^{P} (\mathbf{d}_t - \bar{\mathbf{d}})(\mathbf{d}_{t-h} - \bar{\mathbf{d}})', \quad h = 0, 1, 2, ...$$

is the sample covariance matrix at lag $h$. The modified test $S_c$ ($S$ corrected) relies on the use of an approximately unbiased estimator of the variance of $\bar{\mathbf{d}}$. Here are the assumptions under which the result holds.

**Assumption 1.** $\{\mathbf{d}_t\}$ *is a vector MA(q) process,*

$$\mathbf{d}_t = \boldsymbol{\mu} + \boldsymbol{\epsilon}_t + \sum_{i=1}^{q} \mathbf{\Psi}_i \boldsymbol{\epsilon}_{t-i},$$

*where $\{\boldsymbol{\epsilon}_t\}$ is a (iid) vector white noise process with mean zero, positive definite covariance matrix and bounded fourth moments.*

It can be shown that, under Assumption 1, the *exact* variance of $\bar{\mathbf{d}}$ is

$$P \operatorname{Var} \bar{\mathbf{d}} = \mathbf{\Gamma}(0) + \sum_{h=1}^{q} \Big(\frac{P-h}{P}\Big) \big[\mathbf{\Gamma}(h) + \mathbf{\Gamma}'(h)\big]. \tag{3.2}$$

A large sample approximation for (3.2) is given by $\mathbf{\Omega}$. As the sample size $P$ goes to infinity, the variance of $\sqrt{P}\,\bar{\mathbf{d}}$ tends to $\mathbf{\Omega}$. In our multivariate version of the DM test we estimate $\operatorname{Var}\bar{\mathbf{d}}$ by $P^{-1}\hat{\mathbf{\Omega}}$. In finite samples, however, $P^{-1}\hat{\mathbf{\Omega}}$ is biased when the truncated estimator is used. This happens since $E\,\hat{\mathbf{\Gamma}}(h)$ is different from $\big(\frac{P-h}{P}\big)\mathbf{\Gamma}(h)$. For illustration, consider the simple setting when $q = 0$. Then

$$E\big(P^{-1}\hat{\mathbf{\Omega}}\big) = P^{-2}(P-1)\mathbf{\Gamma}(0),$$

which is different from $\operatorname{Var}\bar{\mathbf{d}} = P^{-1}\mathbf{\Gamma}(0)$. It is because of this reason that HLN derive an approximate bias correction of $P^{-1}\hat{\mathbf{\Omega}}$ for the univariate case, $k = 1$, when $\{\mathbf{d}_t\}$ is a scalar MA($q$) process. The next proposition shows that the finite-sample correction of HLN for the DM test extends also to our multivariate setting, $k \geq 1$.

**Proposition 3.** *Under Assumption 1, if $P > q$, an approximately unbiased estimator of* $\operatorname{Var}\bar{\mathbf{d}}$ *is given by*

$$(cP)^{-1}\hat{\mathbf{\Omega}}, \tag{3.3}$$

*where $\hat{\mathbf{\Omega}}$ is the truncated estimator and*

$$c = \frac{P - 1 - 2q + P^{-1}q(q+1)}{P}. \tag{3.4}$$

*The error in the approximation is of order $O(P^{-3})$. The finite-sample corrected version (3.3) of $P^{-1}\hat{\mathbf{\Omega}}$ leads to the modified test statistic*

$$S_c = cS,$$

*where $S$ is the test statistic of Proposition 1.*

Even if the observed loss-differential series is not generated by a vector MA($q$) process, the simple truncated model can be expected to provide a reasonably good approximation to the true underlying process if it is of the stationary VARMA type. In this case $q$ in (3.4) is a lag length beyond which we are willing to assume that the correlation between $\mathbf{d}_t$ and $\mathbf{d}_{t-h}$ is essentially zero. In practical applications the value of $q$ may be assessed empirically, for example, as described in Tiao and Box (1981). A test at significance level $\alpha$ can be conducted by rejecting the null hypothesis of EPA whenever $S_c > \chi^2_{k,1-\alpha}$, where $\chi^2_{k,1-\alpha}$ is the $(1-\alpha)$ quantile of a $\chi^2_k$ distribution.

In general $c$ will tend to be less than 1, implying a downward correction of $S$ to obtain $S_c$. Thus the test $S$ will tend to be oversized, rejecting a true null hypothesis more often than the nominal size of the test. In deriving $c$ in the Appendix, terms of order $P^{-1}$ were retained in the approximation of $E\,\tilde{\mathbf{\Gamma}}(h)$ as written out in (A.4). We can use the last two terms in (A.3) to include terms of order $P^{-2}$ to get an even finer finite-sample correction. Further Monte Carlo studies, extending the one in Section 4 below, can shed light on potential additional gains when the finite-sample correction is pushed to higher orders of magnitude.

## 4. Monte Carlo Results

To explore whether the behavior of $S$ can be improved by the suggested finite-sample correction we also perform a Monte Carlo study. The study investigates the size and power properties of the test statistics $S$ and $S_c$ using the truncated estimator. All the reported experiments share a common initial state of the generator for pseudo random number generation and are carried out using MATLAB.[1]

For ease of exposition, we consider the simple case when the process generating the vector loss differential series $\{\mathbf{d}_t\}_{t=1}^P$ is a $k$-dimensional MA($q$) with Gaussian noise, $\mathbf{d}_t = \boldsymbol{\mu} + \boldsymbol{\epsilon}_t + \sum_{i=1}^q \boldsymbol{\Psi}_i \boldsymbol{\epsilon}_{t-i}$. Contemporaneously correlated realizations $\boldsymbol{\epsilon}_{1-q}, ..., \boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_P$ of iid (pseudo) random vectors are drawn from a multivariate normal distribution, $\boldsymbol{\epsilon}_t \sim N_k(\mathbf{0}, \boldsymbol{\Sigma})$. $\boldsymbol{\Sigma}$ is the contemporaneous correlation matrix given by $\boldsymbol{\Sigma} = \rho\mathbf{1} - (\rho - 1)\mathbf{I}$, where $\mathbf{1}$ and $\mathbf{I}$ are $k \times k$ unity and identity matrices, respectively, and $0 \leq \rho < 1$. $\boldsymbol{\Psi}_i = \psi^i \mathbf{A}$, where $\mathbf{A}$ is a $k \times k$ diagonal matrix with nonzero entries $a_{jj} = 1/\sqrt{j}$, implying that $\operatorname{Var} d_{jt} \geq \operatorname{Var} d_{j+1,t}$ with equality if and only if $q = 0$. The parameter $\boldsymbol{\mu}$ is $\mathbf{0}$ and different from $\mathbf{0}$ in the size and power experiments, respectively. We consider sample sizes of $P = 100, 500$ and 1000.

**Size Results.** Table 1 reports empirical sizes of the test statistics $S$ and $S_c$ at the asymptotic significance level $\alpha = 0.1$. Each table entry is based on 100 000 Monte Carlo replications and rounded to three decimal places. For $k = 1$ $S_c$ is the HLN test statistic. For $q = 0$ the bias correction of $S_c$ is *exact*. It is seen that the proposed test $S$ can be quite seriously oversized in moderate samples and that this problem becomes more acute as $k$ and/or $q$ increase. It seems clear that the modified test adjusts for this problem. In all experiments, the modified test $S_c$ performs better than $S$, although it also tends to be oversized. For example, when $k = q = 2$, $\rho = \psi = 0.9$ and $P = 100$ the empirical sizes of $S$ and $S_c$ are 0.142 and 0.130, respectively. Even for sample sizes as large as $P = 1000$ the $S$ statistic benefits noticeably from the finite-sample correction, as reflected by the results for $S_c$. Thus, the finite-sample modification for the test $S$ provides important (although not complete) size corrections.

**Power Results.** Since there are significant size-distortions in finite-samples, power comparison is carried out adjusting for size. To this end, empirical critical values are calculated using the simulated samples in the size study. More specifically, let $S_{(1)}, ..., S_{(90\,000)}, ..., S_{(100\,000)}$ be the ordered sample from one of the Monte Carlo experiments in the size study. Then a $\alpha = 0.1$ empirical critical value for $S$ is given by $S_{(90\,000)}$. Similarly, if $c > 0$, the corresponding critical value for the modified test $S_c = cS$ is given by $cS_{(90\,000)}$.[2] This shows that $S$ and $S_c$ have the same size-adjusted power. That is, using empirical critical values, the probabilities that $S$ and $S_c$ will correctly lead to the rejection of a *false* null hypothesis are the same.

In the power experiments we let $\mu_j = r - 1$ if $j = 1$ and zero if $j > 1$, which is consistent with that $E\,g(e_{jt}) = rE\,g(e_{j+1,t})$ if $j = 1$ and $E\,g(e_{jt}) = E\,g(e_{j+1,t}) = 1$ otherwise. Hence, for $r > 1$ this is consistent with that the predictive ability of the first

[1] MATLAB code for generating Tables 1–2 can be downloaded from http://www.mysmu.edu/staff/danielpreve.

[2] By (3.4), a sufficient condition for $c$ to be greater than zero is that $P > 2q + 1$.

TABLE 1. Empirical sizes: Each table entry (based on 100 000 Monte Carlo replications) reports the frequency of the simulations in which the *true* null of EPA is rejected at the asymptotic significance level $\alpha = 0.1$. The process generating a loss differential series of length $P$ is a $k$-dimensional MA($q$) with Gaussian noise. The parameter $\rho$ controls the contemporaneous correlation of the noise. The parameter $\psi$ controls the strength of the serial correlation. $S_c$ is the finite-sample corrected version of the original test $S$.

| | | | $\rho = \psi = 0.5$ | | | $\rho = 0.5, \psi = 0.9$ | | | $\rho = 0.9, \psi = 0.5$ | | | $\rho = \psi = 0.9$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $k$ | $q$ | Test | $P = 100$ | $P = 500$ | $P = 1000$ | $P = 100$ | $P = 500$ | $P = 1000$ | $P = 100$ | $P = 500$ | $P = 1000$ | $P = 100$ | $P = 500$ | $P = 1000$ |
| 1 | 0 | $S$ | 0.105 | 0.101 | 0.101 | 0.105 | 0.101 | 0.101 | 0.105 | 0.101 | 0.101 | 0.105 | 0.101 | 0.101 |
| | | $S_c$ | 0.103 | 0.101 | 0.100 | 0.103 | 0.101 | 0.100 | 0.103 | 0.101 | 0.100 | 0.103 | 0.101 | 0.100 |
| | 1 | $S$ | 0.112 | 0.103 | 0.101 | 0.111 | 0.103 | 0.101 | 0.112 | 0.103 | 0.101 | 0.111 | 0.103 | 0.101 |
| | | $S_c$ | 0.107 | 0.102 | 0.101 | 0.106 | 0.102 | 0.101 | 0.107 | 0.102 | 0.101 | 0.106 | 0.102 | 0.101 |
| | 2 | $S$ | 0.120 | 0.105 | 0.103 | 0.120 | 0.105 | 0.102 | 0.120 | 0.105 | 0.103 | 0.120 | 0.105 | 0.102 |
| | | $S_c$ | 0.111 | 0.103 | 0.102 | 0.111 | 0.103 | 0.101 | 0.111 | 0.103 | 0.102 | 0.111 | 0.103 | 0.101 |
| | 3 | $S$ | 0.131 | 0.107 | 0.103 | 0.129 | 0.107 | 0.103 | 0.131 | 0.107 | 0.103 | 0.129 | 0.107 | 0.103 |
| | | $S_c$ | 0.118 | 0.105 | 0.102 | 0.116 | 0.104 | 0.102 | 0.118 | 0.105 | 0.102 | 0.116 | 0.104 | 0.102 |
| | 4 | $S$ | 0.142 | 0.109 | 0.104 | 0.139 | 0.108 | 0.104 | 0.142 | 0.109 | 0.104 | 0.139 | 0.108 | 0.104 |
| | | $S_c$ | 0.125 | 0.105 | 0.103 | 0.122 | 0.105 | 0.102 | 0.125 | 0.105 | 0.103 | 0.122 | 0.105 | 0.102 |
| 2 | 0 | $S$ | 0.110 | 0.101 | 0.102 | 0.110 | 0.101 | 0.102 | 0.110 | 0.101 | 0.102 | 0.110 | 0.101 | 0.102 |
| | | $S_c$ | 0.108 | 0.101 | 0.102 | 0.108 | 0.101 | 0.102 | 0.108 | 0.101 | 0.102 | 0.108 | 0.101 | 0.102 |
| | 1 | $S$ | 0.126 | 0.104 | 0.103 | 0.125 | 0.104 | 0.103 | 0.126 | 0.104 | 0.103 | 0.125 | 0.104 | 0.103 |
| | | $S_c$ | 0.118 | 0.103 | 0.102 | 0.118 | 0.103 | 0.102 | 0.118 | 0.103 | 0.102 | 0.117 | 0.103 | 0.102 |
| | 2 | $S$ | 0.144 | 0.108 | 0.105 | 0.142 | 0.108 | 0.105 | 0.144 | 0.108 | 0.105 | 0.142 | 0.108 | 0.105 |
| | | $S_c$ | 0.132 | 0.105 | 0.104 | 0.130 | 0.105 | 0.104 | 0.132 | 0.105 | 0.104 | 0.130 | 0.105 | 0.104 |
| | 3 | $S$ | 0.166 | 0.112 | 0.107 | 0.161 | 0.111 | 0.107 | 0.167 | 0.112 | 0.107 | 0.162 | 0.111 | 0.107 |
| | | $S_c$ | 0.148 | 0.108 | 0.105 | 0.143 | 0.108 | 0.105 | 0.148 | 0.109 | 0.105 | 0.143 | 0.108 | 0.105 |
| | 4 | $S$ | 0.188 | 0.116 | 0.109 | 0.181 | 0.114 | 0.108 | 0.189 | 0.115 | 0.109 | 0.181 | 0.114 | 0.108 |
| | | $S_c$ | 0.165 | 0.111 | 0.107 | 0.156 | 0.110 | 0.106 | 0.165 | 0.111 | 0.107 | 0.157 | 0.110 | 0.106 |
| 3 | 0 | $S$ | 0.116 | 0.103 | 0.102 | 0.116 | 0.103 | 0.102 | 0.116 | 0.103 | 0.102 | 0.116 | 0.103 | 0.102 |
| | | $S_c$ | 0.113 | 0.102 | 0.101 | 0.113 | 0.102 | 0.101 | 0.113 | 0.102 | 0.101 | 0.113 | 0.102 | 0.101 |
| | 1 | $S$ | 0.141 | 0.109 | 0.104 | 0.140 | 0.108 | 0.104 | 0.142 | 0.108 | 0.104 | 0.141 | 0.108 | 0.104 |
| | | $S_c$ | 0.133 | 0.107 | 0.103 | 0.131 | 0.107 | 0.103 | 0.133 | 0.107 | 0.103 | 0.132 | 0.107 | 0.103 |
| | 2 | $S$ | 0.172 | 0.113 | 0.107 | 0.168 | 0.113 | 0.107 | 0.174 | 0.113 | 0.107 | 0.169 | 0.113 | 0.107 |
| | | $S_c$ | 0.156 | 0.111 | 0.106 | 0.152 | 0.110 | 0.105 | 0.157 | 0.110 | 0.106 | 0.153 | 0.110 | 0.105 |
| | 3 | $S$ | 0.209 | 0.120 | 0.109 | 0.198 | 0.118 | 0.109 | 0.210 | 0.120 | 0.110 | 0.199 | 0.118 | 0.109 |
| | | $S_c$ | 0.185 | 0.115 | 0.107 | 0.174 | 0.114 | 0.107 | 0.187 | 0.115 | 0.108 | 0.175 | 0.114 | 0.107 |
| | 4 | $S$ | 0.251 | 0.126 | 0.113 | 0.232 | 0.124 | 0.112 | 0.252 | 0.126 | 0.113 | 0.233 | 0.124 | 0.112 |
| | | $S_c$ | 0.219 | 0.121 | 0.110 | 0.200 | 0.118 | 0.110 | 0.221 | 0.121 | 0.110 | 0.201 | 0.119 | 0.109 |
| 4 | 0 | $S$ | 0.123 | 0.104 | 0.103 | 0.123 | 0.104 | 0.103 | 0.123 | 0.104 | 0.103 | 0.123 | 0.104 | 0.103 |
| | | $S_c$ | 0.120 | 0.104 | 0.103 | 0.120 | 0.104 | 0.103 | 0.120 | 0.104 | 0.103 | 0.120 | 0.104 | 0.103 |
| | 1 | $S$ | 0.159 | 0.111 | 0.107 | 0.156 | 0.111 | 0.107 | 0.161 | 0.112 | 0.107 | 0.158 | 0.111 | 0.107 |
| | | $S_c$ | 0.148 | 0.109 | 0.106 | 0.146 | 0.109 | 0.106 | 0.150 | 0.110 | 0.106 | 0.147 | 0.109 | 0.106 |
| | 2 | $S$ | 0.205 | 0.119 | 0.110 | 0.197 | 0.118 | 0.109 | 0.208 | 0.119 | 0.110 | 0.199 | 0.118 | 0.109 |
| | | $S_c$ | 0.185 | 0.116 | 0.108 | 0.177 | 0.115 | 0.108 | 0.188 | 0.116 | 0.108 | 0.180 | 0.115 | 0.108 |
| | 3 | $S$ | 0.263 | 0.126 | 0.114 | 0.244 | 0.125 | 0.113 | 0.266 | 0.126 | 0.114 | 0.246 | 0.125 | 0.113 |
| | | $S_c$ | 0.233 | 0.121 | 0.112 | 0.213 | 0.120 | 0.110 | 0.236 | 0.122 | 0.112 | 0.216 | 0.120 | 0.111 |
| | 4 | $S$ | 0.327 | 0.136 | 0.118 | 0.292 | 0.133 | 0.116 | 0.329 | 0.137 | 0.118 | 0.296 | 0.133 | 0.117 |
| | | $S_c$ | 0.290 | 0.130 | 0.115 | 0.254 | 0.127 | 0.113 | 0.292 | 0.130 | 0.115 | 0.258 | 0.127 | 0.114 |

Table 2. Empirical power: Each table entry (based on 100 000 Monte Carlo replications) reports the frequency of the simulations in which the *false* null of EPA is rejected using the empirical critical value, calculated as the $(1-\alpha)$ quantile of the distribution of $S$ computed in the corresponding size experiment of Table 1.

| | | | $\rho=\psi=0.5$ | | | $\rho=0.5,\psi=0.9$ | | | $\rho=0.9,\psi=0.5$ | | | $\rho=\psi=0.9$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $k$ | $q$ | Test | $P=100$ | $P=500$ | $P=1000$ | $P=100$ | $P=500$ | $P=1000$ | $P=100$ | $P=500$ | $P=1000$ | $P=100$ | $P=500$ | $P=1000$ |
| 1 | 0 | $S$ | 0.799 | 1.000 | 1.000 | 0.799 | 1.000 | 1.000 | 0.799 | 1.000 | 1.000 | 0.799 | 1.000 | 1.000 |
| | 1 | $S$ | 0.501 | 0.981 | 1.000 | 0.368 | 0.901 | 0.994 | 0.501 | 0.981 | 1.000 | 0.368 | 0.901 | 0.994 |
| | 2 | $S$ | 0.406 | 0.937 | 0.998 | 0.237 | 0.659 | 0.897 | 0.406 | 0.937 | 0.998 | 0.237 | 0.659 | 0.897 |
| | 3 | $S$ | 0.364 | 0.906 | 0.995 | 0.184 | 0.487 | 0.740 | 0.364 | 0.906 | 0.995 | 0.184 | 0.487 | 0.740 |
| | 4 | $S$ | 0.340 | 0.887 | 0.992 | 0.159 | 0.385 | 0.609 | 0.340 | 0.887 | 0.992 | 0.159 | 0.385 | 0.609 |
| 2 | 0 | $S$ | 0.819 | 1.000 | 1.000 | 0.819 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | 1 | $S$ | 0.499 | 0.990 | 1.000 | 0.360 | 0.924 | 0.997 | 0.960 | 1.000 | 1.000 | 0.842 | 1.000 | 1.000 |
| | 2 | $S$ | 0.389 | 0.955 | 0.999 | 0.224 | 0.674 | 0.918 | 0.881 | 1.000 | 1.000 | 0.556 | 0.997 | 1.000 |
| | 3 | $S$ | 0.342 | 0.926 | 0.998 | 0.175 | 0.490 | 0.761 | 0.819 | 1.000 | 1.000 | 0.388 | 0.958 | 1.000 |
| | 4 | $S$ | 0.312 | 0.908 | 0.996 | 0.150 | 0.381 | 0.618 | 0.771 | 1.000 | 1.000 | 0.298 | 0.872 | 0.992 |
| 3 | 0 | $S$ | 0.813 | 1.000 | 1.000 | 0.813 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | 1 | $S$ | 0.474 | 0.990 | 1.000 | 0.338 | 0.922 | 0.998 | 0.976 | 1.000 | 1.000 | 0.877 | 1.000 | 1.000 |
| | 2 | $S$ | 0.365 | 0.955 | 0.999 | 0.210 | 0.658 | 0.919 | 0.908 | 1.000 | 1.000 | 0.580 | 0.999 | 1.000 |
| | 3 | $S$ | 0.316 | 0.924 | 0.998 | 0.164 | 0.470 | 0.753 | 0.840 | 1.000 | 1.000 | 0.396 | 0.976 | 1.000 |
| | 4 | $S$ | 0.283 | 0.903 | 0.997 | 0.143 | 0.361 | 0.605 | 0.777 | 1.000 | 1.000 | 0.298 | 0.907 | 0.997 |
| 4 | 0 | $S$ | 0.797 | 1.000 | 1.000 | 0.797 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | 1 | $S$ | 0.450 | 0.989 | 1.000 | 0.317 | 0.916 | 0.998 | 0.977 | 1.000 | 1.000 | 0.876 | 1.000 | 1.000 |
| | 2 | $S$ | 0.338 | 0.950 | 0.999 | 0.197 | 0.639 | 0.913 | 0.901 | 1.000 | 1.000 | 0.564 | 0.999 | 1.000 |
| | 3 | $S$ | 0.286 | 0.916 | 0.998 | 0.155 | 0.450 | 0.737 | 0.814 | 1.000 | 1.000 | 0.378 | 0.979 | 1.000 |
| | 4 | $S$ | 0.244 | 0.892 | 0.997 | 0.137 | 0.343 | 0.583 | 0.717 | 1.000 | 1.000 | 0.282 | 0.910 | 0.998 |

model is lower than the others under $g(\cdot)$, ensuring that the null hypothesis of EPA is false. Table 2 reports size-adjusted rejection frequencies of the test $S$ (and hence also $S_c$) for $r = 1.25$. In general, power decreases in $q$ and $\psi$ and increases in $P$.

**Remarks.** In their paper, HLN argue that a further intuitively reasonable modification of the DM test is to compare (1.2) and their modified statistic, respectively, with critical values from the Student's $t$ distribution with $(P-1)$ degrees of freedom, rather than from a standard normal distribution. In an extensive simulation study, including heavy-tailed forecast error distributions, the authors conclude that both of their proposed modifications of the DM test are worth making in finite samples. Along the lines of HLN, a further modification of the tests $S$ and $S_c$ is to compare the statistics $\frac{P-1}{P}S$ and $\frac{P-1}{P}S_c$, respectively, with critical values from Hotelling's $T^2$ distribution with parameters $k$ and $(P-1)$, rather than from the $\chi^2_k$ distribution. Similar to HLN, such an approach can be motivated by the observation that the exact finite-sample distribution of $\frac{P-1}{P}S$ is $T^2(k, P-1)$ in the case when $\mathbf{d}_1, ..., \mathbf{d}_P$ are independent multivariate normal random vectors with common distribution $N_k(\mathbf{0}, \mathbf{\Sigma})$ and $\hat{\mathbf{\Omega}} = \hat{\mathbf{\Gamma}}(0)$. See, for example, Corollary 3.5.1.1 in Mardia et al. (2000). The potential gains of such an approach is the subject of further studies.

Finally, it is important to acknowledge a limitation of the tests that we employ in this section. While the truncated estimator, (3.1), used when computing $S$ and $S_c$ is HAC, it need not be positive semidefinite (PSD) in finite samples if $q > 0$ (Newey and West, 1987). This property may interfere with hypothesis testing. In our Monte Carlo study this became apparent in small samples. For example, in the most extreme experiment ($k = q = 4$, $\rho = 0.9$, $\psi = 0.5$ and $P = 100$) the number of *negative* $S$ statistics were 633 (0.6%). In recognition of this, it would be interesting to compare results for the two tests studied with results for a third test using one of the PSD estimators discussed in the Summary and Concluding Remarks section below, in a more extensive Monte Carlo study which also includes non-Gaussian loss differentials.

## 5. Summary and Concluding Remarks

In this paper, we considered a multivariate version of the DM test for EPA and showed that it is invariant with respect to the ordering of the alternative forecasting models for a wide range of covariance matrix estimators. The test statistic $S$ has an asymptotic $\chi^2_k$ distribution under the null hypothesis of EPA. Additionally, we showed that the finite-sample correction of HLN for the DM test extends to our multivariate setting, resulting in the modified test $S_c$. Simulations indicated that the $S$ statistic using the truncated estimator, (3.1), has reasonable size properties for a small number of alternative models and large samples, and that improved finite-sample properties are obtained by correcting the test for finite-sample bias in the estimated variance of $\bar{\mathbf{d}}$. This suggests that a natural environment for $S$ (and $S_c$) is when a large number of forecasts from a small number of non-nested alternative models are available for comparison. It was speculated that further finite-sample improvements can be achieved using Hotelling's $T^2$ critical values.

It was remarked that although (3.1), used when computing the tests in the Monte Carlo study, is a HAC estimator of $\mathbf{\Omega}$ it has the drawback that it need not be PSD in finite samples. If a PSD estimator of $\mathbf{\Omega}$ is required and it is known that the $\mathbf{\Psi}_i$ in

(2.5) are zero after lag $q$, one can use the consistent and by construction PSD long run variance estimators discussed in Section 2 of West (2008). In the more general case in which the order $q$ is taken to be unknown, or is infinite, one can instead use a HAC estimator where the truncation point tends to infinity at a suitable rate. In all cases one can also use bootstrap critical values, along the lines of Corradi and Swanson (2006) and Corradi and Swanson (2007). We conjecture that bootstrapping can further improve the finite sample behavior of $S$. The potential gains of such bootstrap schemes is the subject of further studies.

**Parameter Estimation Uncertainty.** In order to compare the out-of-sample predictive ability when one or more of the alternative models are parametric, the time series $\{y_t\}_{t=1}^{R+P}$ is split into two subsamples. The first $R$ observations are used for model estimation and the last $P$ observations is the hold-back sample used for forecast evaluation. Typically a recursive scheme is used, where the size of the sample used for estimation successively increases as new forecasts are made.[3]

West (1996) observed that if any of the models are parametric (1.1) depend on estimated parameters such that the limiting distribution of (1.2) may not be standard normal. West showed how to adjust for the effects of parameter estimation error when conducting inference.[4] West also showed that asymptotic irrelevance holds quite generally whenever $\lim_{P,R\to\infty} P/R = 0$.[5] For related results, see also McCracken (2000) and Corradi et al. (2001). In this case asymptotic inference does not require adjusting for parameter estimation error and the limiting null distribution of (1.2) is standard normal.

If one or more of the alternative models is parametric, it is our reading of West (1996) and McCracken (2000) that asymptotic irrelevance holds quite generally for the tests $S$ and $S_c$ whenever $P/R \to 0$ as $P$ and $R$ tend to infinity. In this case asymptotic inference does not require adjusting for parameter estimation error and the limiting null distribution of $S$ (and $S_c$) is $\chi^2_k$.

As pointed out by, for example, McCracken (2007) the results in DM and West (1996) may not apply if the alternative models are nested. In this case the numerator and denominator of (1.2) may vanish asymptotically in a way such that the limiting distribution is non-normal.[6]

Giacomini and White (2006) considered an asymptotic framework, where $R$ is bounded and $P \to \infty$, that justifies the use of the DM test in the case of nested or non-nested parametric models. As noted by a referee, it appears that this framework can be used to ensure that the limiting null distribution of $S$ (and $S_c$) is $\chi^2_k$ in the case where the alternative models are parametric and potentially nested.

**Rejection of the Null of EPA.** In practice, a rejection of the null hypothesis using $S$ or $S_c$ only suggests that one or more of the alternative models stand out in terms

[3] Some other commonly used schemes are the rolling and fixed schemes.

[4] The adjustment requires that the models and their associated estimators are known. In addition, its calculations may be quite involved.

[5] In other words, the contribution of parameter estimation error vanishes asymptotically if $P$ grows at a slower rate than $R$.

[6] If one model nests the other, smaller and correctly specified, model and parameters are estimated consistently as $R \to \infty$.

of predictive ability. It is usually of interest to identify these models. The following multistep procedure, which is along the lines of Hansen et al. (2011), can be used to eliminate models with poor sample performance. In step one, test the null of EPA using $S$ or $S_c$ at level $\alpha$. In step two, stop if the null is accepted. If not, use an elimination rule to remove one of the alternative models and repeat the procedure in step one for the remaining models.

As suggested by a second referee, an elimination rule can be based on the DM statistic in (1.2). More specifically, define $s_i = \bar{d}_i/\sqrt{\hat{\omega}_{ii}/P}$ $(i = 1, ..., k)$, where $\hat{\omega}_{ii}$ is the $i$th diagonal element of $\hat{\mathbf{\Omega}}$, and let $j = \arg\max_i |s_i|$. Suppose that $g : \mathbf{R} \to \mathbf{R}^+$. Then, in view of (2.2), eliminate model $j$ if the value of $s_j$ is positive. Otherwise, eliminate model $j + 1$.

## Appendix A. Proofs

**Lemma 1.** *Suppose that* $\mathbf{B}$ *is a nonsingular* $k \times k$ *matrix and let*

$$\hat{\mathbf{\Omega}}_* = \hat{\mathbf{\Gamma}}_*(0) + \sum_{h=1}^{m} \kappa(h, m)\big[\hat{\mathbf{\Gamma}}_*(h) + \hat{\mathbf{\Gamma}}_*'(h)\big],$$

*where*

$$\hat{\mathbf{\Gamma}}_*(h) = \frac{1}{P}\sum_{t=h+1}^{P}(\mathbf{B}\mathbf{d}_t - \bar{\mathbf{d}}_*)(\mathbf{B}\mathbf{d}_{t-h} - \bar{\mathbf{d}}_*)' \text{ and } \bar{\mathbf{d}}_* = \frac{1}{P}\sum_{t=1}^{P}\mathbf{B}\mathbf{d}_t.$$

*If* $\hat{\mathbf{\Omega}}$ *is nonsingular, then*

$$P(\bar{\mathbf{d}}_* - \boldsymbol{\mu}_*)'\hat{\mathbf{\Omega}}_*^{-1}(\bar{\mathbf{d}}_* - \boldsymbol{\mu}_*) = P(\bar{\mathbf{d}} - \boldsymbol{\mu})'\hat{\mathbf{\Omega}}^{-1}(\bar{\mathbf{d}} - \boldsymbol{\mu}),$$

*where* $\boldsymbol{\mu}_* = \mathbf{B}\boldsymbol{\mu}$.

*Proof.* Since $\bar{\mathbf{d}}_* = \mathbf{B}\bar{\mathbf{d}}$ and

$$\hat{\mathbf{\Gamma}}_*(h) = \frac{1}{P}\sum_{t=h+1}^{P}\mathbf{B}(\mathbf{d}_t - \bar{\mathbf{d}})(\mathbf{d}_{t-h} - \bar{\mathbf{d}})'\mathbf{B}' = \mathbf{B}\hat{\mathbf{\Gamma}}(h)\mathbf{B}',$$

it follows that

$$\hat{\mathbf{\Omega}}_* = \mathbf{B}\Big\{\hat{\mathbf{\Gamma}}(0) + \sum_{h=1}^{m}\kappa(h, m)\big[\hat{\mathbf{\Gamma}}(h) + \hat{\mathbf{\Gamma}}'(h)\big]\Big\}\mathbf{B}' = \mathbf{B}\hat{\mathbf{\Omega}}\mathbf{B}'.$$

Finally, because both $\mathbf{B}$ and $\hat{\mathbf{\Omega}}$ are nonsingular,

$$\begin{aligned}&(\bar{\mathbf{d}}_* - \boldsymbol{\mu}_*)'\hat{\mathbf{\Omega}}_*^{-1}(\bar{\mathbf{d}}_* - \boldsymbol{\mu}_*)\\ &= (\bar{\mathbf{d}} - \boldsymbol{\mu})'\mathbf{B}'(\mathbf{B}\hat{\mathbf{\Omega}}\mathbf{B}')^{-1}\mathbf{B}(\bar{\mathbf{d}} - \boldsymbol{\mu}) = (\bar{\mathbf{d}} - \boldsymbol{\mu})'\hat{\mathbf{\Omega}}^{-1}(\bar{\mathbf{d}} - \boldsymbol{\mu}).\end{aligned}$$

□

**Proof of Proposition 2.** The total number of possible permutations of the forecast errors at time $t$ that use all $k+1$ errors is equal to $(k+1)!$. Any such permutation $\mathbf{e}_t^*$ is given by the relation $\mathbf{e}_t^* = \mathbf{P}\mathbf{e}_t$, where $\mathbf{P}$ is a $(k+1) \times (k+1)$ permutation matrix and $\mathbf{e}_t$ is the (arbitrary) ordering $\mathbf{e}_t = (e_{1t}, ..., e_{k+1,t})'$. Denote by $\mathbf{D}$ the $k \times (k+1)$ matrix

$$\mathbf{D} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{pmatrix},$$

and let the multi-variable mapping $\mathbf{g} : \mathbf{R}^{k+1} \to \mathbf{R}^{k+1}$ be given by

$$\mathbf{g}(\mathbf{x}) = \begin{pmatrix} g(x_1) \\ \vdots \\ g(x_{k+1}) \end{pmatrix}.$$

Now, any possible loss differential vector $\mathbf{d}_t^*$ at time $t$ is given by the relation

$$\mathbf{d}_t^* = \mathbf{D}\mathbf{g}(\mathbf{e}_t^*) = \mathbf{D}\mathbf{g}(\mathbf{P}\mathbf{e}_t),$$

for some permutation matrix $\mathbf{P}$. For example, for $\mathbf{P} = \mathbf{I}$, $\mathbf{d}_t^*$ is equal to $\mathbf{d}_t$. First we will show that there always exists a transformation matrix $\mathbf{B}$ such that $\mathbf{B}\mathbf{d}_t = \mathbf{d}_t^*$. Next, we will show that $\mathbf{B}$ is nonsingular.

It follows that, if $\mathbf{B}$ exists, we must have that

$$\mathbf{B}\mathbf{D}\mathbf{g}(\mathbf{e}_t) = \mathbf{D}\mathbf{g}(\mathbf{P}\mathbf{e}_t) = \mathbf{D}\mathbf{P}\mathbf{g}(\mathbf{e}_t), \tag{A.1}$$

where the second equality follows since $\mathbf{g}(\mathbf{P}\mathbf{x}) = \mathbf{P}\mathbf{g}(\mathbf{x})$. A right inverse of $\mathbf{D}$ is given by the $(k+1) \times k$ matrix

$$\mathbf{D}^- = \begin{pmatrix} 1 & \cdots & 1 & 0 \\ 0 & \ddots & \vdots & \vdots \\ \vdots & \ddots & 1 & 0 \\ 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & -1 \end{pmatrix}.$$

Consequently, the $k \times k$ transformation matrix $\mathbf{B} = \mathbf{D}\mathbf{P}\mathbf{D}^-$ satisfies (A.1) and $\mathbf{B}\mathbf{d}_t = \mathbf{d}_t^*$. In view of Lemma 1, it only remains to show that $\mathbf{B}$ is nonsingular.

By Theorem A.6.2 in Mardia et al. (2000), the non-zero eigenvalues of $\mathbf{B}$ and the $(k+1) \times (k+1)$ matrix $\mathbf{P}\mathbf{D}^-\mathbf{D}$ are the same. Moreover, since $\mathbf{P}$ is a permutation matrix, the number of non-zero eigenvalues of $\mathbf{P}\mathbf{D}^-\mathbf{D}$ and $\mathbf{D}^-\mathbf{D}$ are the same. Using

Laplace's formula on the last column, it is readily seen that

$$\mathbf{D}^{-}\mathbf{D} = \begin{pmatrix} 1 & 0 & \cdots & 0 & -1 & 0 \\ 0 & 1 & \ddots & \vdots & -1 & 0 \\ \vdots & \ddots & \ddots & 0 & \vdots & \vdots \\ 0 & \cdots & 0 & 1 & -1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & -1 & 1 \end{pmatrix},$$

has $k$ non-zero eigenvalues. Thus, in view of the theorem, $\mathbf{B}$ has $k$ non-zero eigenvalues. Hence $\mathbf{B}$ is nonsingular. □

**Proof of Proposition 3.** Following the outline of the proof in HLN, the estimator in (3.1) can be written as

$$\hat{\mathbf{\Omega}} = \tilde{\mathbf{\Gamma}}(0) + \sum_{h=1}^{q} \Big(\frac{P-h}{P}\Big) \big[\tilde{\mathbf{\Gamma}}(h) + \tilde{\mathbf{\Gamma}}'(h)\big], \tag{A.2}$$

where

$$\tilde{\mathbf{\Gamma}}(h) = \Big(\frac{P}{P-h}\Big)\hat{\mathbf{\Gamma}}(h) = \frac{1}{P-h} \sum_{t=h+1}^{P} (\mathbf{d}_t - \bar{\mathbf{d}})(\mathbf{d}_{t-h} - \bar{\mathbf{d}})',$$

and

$$\begin{aligned} & \sum_{t=h+1}^{P} (\mathbf{d}_t - \bar{\mathbf{d}})(\mathbf{d}_{t-h} - \bar{\mathbf{d}})' \\ &= \sum_{t=h+1}^{P} \mathbf{d}_t \mathbf{d}_{t-h}' - \sum_{t=h+1}^{P} \mathbf{d}_t \bar{\mathbf{d}}' - \sum_{t=h+1}^{P} \bar{\mathbf{d}} \mathbf{d}_{t-h}' + (P-h)\bar{\mathbf{d}}\,\bar{\mathbf{d}}' \\ &= \sum_{t=h+1}^{P} (\mathbf{d}_t \mathbf{d}_{t-h}' - \boldsymbol{\mu}\boldsymbol{\mu}') - (P+h)(\bar{\mathbf{d}}\,\bar{\mathbf{d}}' - \boldsymbol{\mu}\boldsymbol{\mu}') \\ & + \sum_{t=1}^{h} (\mathbf{d}_t \bar{\mathbf{d}}' - \boldsymbol{\mu}\boldsymbol{\mu}') + \sum_{t=P-h+1}^{P} (\bar{\mathbf{d}} \mathbf{d}_t' - \boldsymbol{\mu}\boldsymbol{\mu}'). \end{aligned}$$

To arrive at the second equality, note that

$$\sum_{t=h+1}^{P} \mathbf{d}_t \bar{\mathbf{d}}' = P\bar{\mathbf{d}}\bar{\mathbf{d}}' - \sum_{t=1}^{h} \mathbf{d}_t \bar{\mathbf{d}}',$$

and

$$\sum_{t=h+1}^{P} \bar{\mathbf{d}} \mathbf{d}_{t-h}' = \sum_{t=1}^{P-h} \bar{\mathbf{d}} \mathbf{d}_t' = P\bar{\mathbf{d}}\bar{\mathbf{d}}' - \sum_{t=P-h+1}^{P} \bar{\mathbf{d}} \mathbf{d}_t'.$$

By induction in $h$, it is readily verified that for $P > h$

$$\begin{aligned}
P\,E\Big[\sum_{t=1}^{h}(\mathbf{d}_t\bar{\mathbf{d}}'-\boldsymbol{\mu}\boldsymbol{\mu}')\Big] &= P\,E\Big[\sum_{t=P-h+1}^{P}(\bar{\mathbf{d}}\mathbf{d}_t'-\boldsymbol{\mu}\boldsymbol{\mu}')\Big] \\
&= \sum_{i=1}^{h-1}(h-i)\boldsymbol{\Gamma}(i)+\sum_{i=0}^{P-h}h\boldsymbol{\Gamma}'(i)+\sum_{i=1}^{h-1}(h-i)\boldsymbol{\Gamma}'(P-h+i).
\end{aligned}$$

Hence, for $h = 0, 1, ..., q$

$$\begin{aligned}
&E\,\tilde{\boldsymbol{\Gamma}}(h) \\
&= \frac{1}{P-h}E\Big[\sum_{t=h+1}^{P}(\mathbf{d}_t-\bar{\mathbf{d}})(\mathbf{d}_{t-h}-\bar{\mathbf{d}})'\Big] \\
&= \frac{1}{P-h}E\Big[\sum_{t=h+1}^{P}(\mathbf{d}_t\mathbf{d}_{t-h}'-\boldsymbol{\mu}\boldsymbol{\mu}')-(P+h)(\bar{\mathbf{d}}\,\bar{\mathbf{d}}'-\boldsymbol{\mu}\boldsymbol{\mu}') \\
&+\sum_{t=1}^{h}(\mathbf{d}_t\bar{\mathbf{d}}'-\boldsymbol{\mu}\boldsymbol{\mu}')+\sum_{t=P-h+1}^{P}(\bar{\mathbf{d}}\mathbf{d}_t'-\boldsymbol{\mu}\boldsymbol{\mu}')\Big] = \boldsymbol{\Gamma}(h)-\Big(\frac{P+h}{P-h}\Big)\mathrm{Var}\,\bar{\mathbf{d}} \\
&+\frac{2}{P(P-h)}\Big[\sum_{i=1}^{h-1}(h-i)\boldsymbol{\Gamma}(i)+\sum_{i=0}^{P-h}h\boldsymbol{\Gamma}'(i)+\sum_{i=1}^{h-1}(h-i)\boldsymbol{\Gamma}'(P-h+i)\Big].
\end{aligned} \tag{A.3}$$

The final term in (A.3) is zero if $h = 0$ and of order $P^{-2}$ for $h > 0$. Thus, for $h = 0, 1, ..., q$ we have that

$$E\,\tilde{\boldsymbol{\Gamma}}(h) = \boldsymbol{\Gamma}(h)-\Big(\frac{P+h}{P-h}\Big)\mathrm{Var}\,\bar{\mathbf{d}}+O(P^{-2}) = \boldsymbol{\Gamma}(h)-\mathrm{Var}\,\bar{\mathbf{d}}+O(P^{-2}). \tag{A.4}$$

In view of (A.2), (A.4) and (3.2) the expected value of $P^{-1}\hat{\boldsymbol{\Omega}}$ is

$$\begin{aligned}
&E\big(P^{-1}\hat{\boldsymbol{\Omega}}\big) \\
&= \frac{1}{P}E\Big\{\tilde{\boldsymbol{\Gamma}}(0)+\sum_{h=1}^{q}\Big(\frac{P-h}{P}\Big)\big[\tilde{\boldsymbol{\Gamma}}(h)+\tilde{\boldsymbol{\Gamma}}'(h)\big]\Big\} \\
&= \frac{1}{P}\Big\{\boldsymbol{\Gamma}(0)-\mathrm{Var}\,\bar{\mathbf{d}}+\sum_{h=1}^{q}\Big(\frac{P-h}{P}\Big)\big[\boldsymbol{\Gamma}(h)+\boldsymbol{\Gamma}'(h)\big] \\
&-2\sum_{h=1}^{q}\Big(\frac{P-h}{P}\Big)\mathrm{Var}\,\bar{\mathbf{d}}+O(P^{-2})\Big\} \\
&= \frac{1}{P}\Big[P-1-\frac{2}{P}\sum_{h=1}^{q}(P-h)\Big]\mathrm{Var}\,\bar{\mathbf{d}}+O(P^{-3}) \\
&= \frac{P-1-2q+P^{-1}q(q+1)}{P}\mathrm{Var}\,\bar{\mathbf{d}}+O(P^{-3}).
\end{aligned} \tag{A.5}$$

□

## References

Andrews, D.W.K., 1991, Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817–858.

Christiano, L.J., 1989, $p^*$: Not the inflation forecasters holy grail. Federal Reserve Bank of Minneapolis Quarterly Review 13, 3–18.

Clark, T.E. and K.D. West, 2006, Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. Journal of Econometrics 135, 155–186.

Clark, T.E. and K.D. West, 2007, Approximately normal tests for equal predictive accuracy in nested models. Journal of Econometrics 138, 291–311.

Corradi, V. and N.R. Swanson, 2006, Predictive density and conditional confidence interval accuracy tests. Journal of Econometrics 135, 187–228.

Corradi, V. and N.R. Swanson, 2007, Nonparametric bootstrap procedures for predictive inference based on recursive estimation schemes. International Economic Review 48, 67–109.

Corradi, V., Swanson N.R. and C. Olivetti, 2001, Predictive ability with cointegrated variables. Journal of Econometrics 104, 315–358.

Diebold, F.X. and R.S. Mariano, 1995, Comparing predictive accuracy. Journal of Business & Economic Statistics 13, 134–145.

Giacomini, R. and H. White, 2006, Tests of conditional predictive ability. Econometrica 74, 1545–1578.

Granger, C. and P. Newbold, 1977, Forecasting economic time series, Academic Press.

Hansen, P.R., 2005, A test for superior predictive ability. Journal of Business & Economic Statistics 23, 365–380.

Hansen, P.R., Lunde A. and J.M. Nason, 2011, The model confidence set. Econometrica 79, 453–497.

Harvey, D., Leybourne, S. and P. Newbold, 1997, Testing the equality of prediction mean squared errors. International Journal of Forecasting 13, 281–291.

Hubrich, K. and K.D. West, 2010, Forecast evaluation of small nested model sets. Journal of Applied Econometrics 25, 574–594.

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

Mardia, K., Kent, J. and J. Bibby, 2000, Multivariate analysis, Academic Press.

Mariano, R.S., 2002, Testing forecast accuracy, in: M.P. Clements and D.F. Hendry, (Eds.), A companion to economic forecasting, Wiley-Blackwell, pp. 284–298.

McCracken, M.W., 2000, Robust out-of-sample inference. Journal of Econometrics 99, 195–223.

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

Meese, R. and K. Rogoff, 1988, Was it real? the exchange rate-interest rate differential relation over the modern floating-rate period. Journal of Finance 43, 933–948.

Morgan, W.A., 1939, A test for significance of the difference between the two variances in a sample from a normal bivariate population. Biometrika 31, 13–19.

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

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

Tiao, G.C. and G.E.P. Box, 1981, Modeling multiple time series with applications. Journal of the American Statistical Association 76, 802–816.

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

West, K.D., 2008, Heteroskedasticity and autocorrelation corrections, in: S.N. Durlauf and L.E. Blume, (Eds.), The new Palgrave dictionary of economics, Palgrave Macmillan, pp. 6–12.

West, K.D. and D. Cho, 1995, The predictive ability of several models of exchange rate volatility. Journal of Econometrics 69, 367–391.

West, K.D., Edison, H.J. and D. Cho, 1993, A utility-based comparison of some models of exchange rate volatility. Journal of International Economics 35, 23–45.

White, H., 2000, A reality check for data snooping. Econometrica 68, 1097–1126.