Journal of Statistical Planning and Inference 140 (2010) 3655–3675

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Selection between models through multi-step-ahead forecasting

Tucker S. McElroy, David F. Findley *

Statistical Research Division, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233-9100, USA

ARTICLE INFO

Available online 5 May 2010

Keywords:
ARIMA models
Diebold–Mariano tests
Incorrect models
Misspecified models
Model selection
Parameter estimation effects
Time series

ABSTRACT

We develop and show applications of two new test statistics for deciding if one ARIMA model provides significantly better h-step-ahead forecasts than another, as measured by the difference of approximations to their asymptotic mean square forecast errors. The two statistics differ in the variance estimates used for normalization. Both variance estimates are consistent even when the models considered are incorrect. Our main variance estimate is further distinguished by accounting for parameter estimation, while the simpler variance estimate treats parameters as fixed. Their broad consistency properties offer improvements to what are known as tests of Diebold and Mariano (1995) type, which are tests that treat parameters as fixed and use variance estimates that are generally not consistent in our context. We show how these statistics can be calculated for any pair of ARIMA models with the same differencing operator.

Published by Elsevier B.V.

1. Introduction

In this article, we make several contributions to the technology of testing whether two not necessarily correct time series models for an observed series have equal or differing h-step-ahead forecasting ability as assessed by estimates of mean square h-step forecast error. This work is in the tradition of Meese and Rogoff (1988), Findley (1990, 1991a), Diebold and Mariano (1995) and Rivers and Vuong (2002). Our focus is on nonstationary ARIMA models, a type of model not considered in this earlier work. Our specific approach is derived from the goodness-of-fit testing methodology of McElroy and Holan (2009) with modifications to account for the consideration of more than one model and other features of the forecast comparison setting. We account for effects of parameter estimation, which only Rivers and Vuong (2002) do among the forecasting papers cited. In contrast to Rivers and Vuong, we provide explicit formulas for the asymptotic variance of our statistic (corresponding to the σ_n^2 quantity of their Assumption 7), as well as an explicit consistent estimator of this variance. Also, our assumptions are more basic and therefore more transparent. These same advantages apply in relation to the results of West (1996), which also account for parameter estimation but are focused on out-of-sample forecasting, from a perspective more connected with regression models. Our tests, like those of the papers other than West's, are tests of in-sample forecast performance.

The approximation relation between our measure of model forecast performance (8) and the more customary average of squared forecast errors over the sample is derived in Section 2.1, after a review of some relevant aspects of ARIMA model forecasting. The central theoretical results of the paper are presented in Section 2.3, whose Theorem 1 provides the CLT and consistent estimator of its variance needed for our main test statistic (12). Section 2.4 presents results for the situation in which parameter estimation uncertainty is ignored, i.e., when estimated parameters are treated as constant. Here our consistent variance estimate simplifies, becoming reasonably straightforward to calculate for all ARIMA models, and is also

* Corresponding author.
E-mail address: david.f.findley@census.gov (D.F. Findley).

0378-3758/$ - see front matter Published by Elsevier B.V.
doi:10.1016/j.jspi.2010.04.032

applicable to the ARIMA model case of the test commonly referred to as the test of Diebold and Mariano (1995). For this test, it provides a consistent alternative to the customary variance estimate, which is consistent only in effectively correct model situations. With h=1, it also provides a consistent variance estimate, which had been lacking, for the time series generalization in Findley (1990) of the nonnested model comparison test statistic of Vuong (1989).

In Section 3, after explaining why the size study of Diebold and Mariano (1995) is invalid, we present size and power studies of our test statistics and the Diebold–Mariano statistic together with an empirical study of the application of all three statistics to competing models for series from Box et al. (1994) and Brockwell and Davis (2002). The size and power studies favor both of our new test statistics over the Diebold–Mariano statistic. All of the studies favor most our statistic that accounts for parameter estimation.

The Appendix contains proofs and the derivations of some formulas, including auxiliary formulas for algebraically computing the variance estimate that accounts for parameter uncertainty.

2. Methodology

We are interested in comparing two competing models' h-step-ahead forecasts of data from a time series Y_t which, if nonstationary, can be made stationary by application of a differencing operator, i.e., a backshift operator polynomial $\delta(B)$ whose zeroes have unit magnitude. As usual, B denotes the backshift (or lag) operator, with BX_t=X_{t-1}. To simplify the exposition, the stationary series $W_t=\delta(B)Y_t$ is assumed to be Gaussian, an assumption that can be weakened moderately. It is also assumed to be purely nondeterministic. Thus its spectral density $\tilde{f}$ is log integrable and generates its autocovariances via $\gamma_j(\tilde{f})=(2\pi)^{-1}\int_{-\pi}^{\pi}\tilde{f}(\lambda)e^{ij\lambda}\,d\lambda$, a formula that shows our convention with the constant 2π. The matrix of autocovariances is denoted $\Gamma(\tilde{f})$, i.e., $\Gamma_{jk}(\tilde{f})=\gamma_{j-k}(\tilde{f})$. The dimension of $\Gamma(\tilde{f})$ is equal to the number of $W_t=\delta(B)Y_t$ calculable from the observed Y_t.

2.1. Multi-step-ahead forecasting

We start by reviewing some basic forecasting results for nonstationary Y_t. Beyond basic formulas, the key results obtained are two concerning asymptotic properties of forecast error measures, (6) and (7).

Let $\delta(z)=1+\sum_{j=1}^{d}\delta_j z^j$ be the differencing operator such that $W_t=\delta(B)Y_t$ and let Y_t, $1-d\le t\le n$ denote the available data. Set $\tau(z)=1/\delta(z)$ expressed as a power series in $|z|<1$ with coefficients τ_j. Thus $\tau_j=1$ for j=0 and $\tau_j=-\sum_{i=0}^{j-1}\tau_i\delta_{j-i}$ for $j>0$. For any $1\le h<n$ and any $1\le t\le n-h$, we have $Y_{t+h}=[\tau]_0^{h-1}(B)W_{t+h}+\sum_{j=0}^{d-1}c_{j,h}Y_{t-j}$, where the coefficients $c_{j,h}$ depend only on the coefficients of $\delta(z)$, see Bell (1984, p. 650). The bracket notation means that the power series is truncated to powers of B between zero and $h-1$. Forecasts $\hat{Y}_{t+h|t}$ of Y_{t+h} from Y_s, $1-d\le s\le t$ are obtained from forecasts $\hat{W}_{t+h-j|t}$, $0\le j\le h-1$ of W_{t+h-j} from W_s, $1\le s\le t$ by way of

$$\hat{Y}_{t+h|t}=\sum_{j=0}^{h-1}\tau_j\hat{W}_{t+h-j|t}+\sum_{j=0}^{d-1}c_{j,h}Y_{t-j}. \tag{1}$$

Consequently, the forecast errors are given by $Y_{t+h}-\hat{Y}_{t+h|t}=\sum_{j=0}^{h-1}\tau_j(W_{t+h-j}-\hat{W}_{t+h-j|t})$.

To motivate our performance measure, we will use the forecast $\hat{W}_{t+h|t}$ obtained by truncating the filter for the forecast $W_{t+h|t}$ of W_{t+h} from the infinite past W_s, $-\infty<s\le t$. The latter forecast is given by $W_{t+h|t}=\sum_{j\ge 0}\psi_{j+h}B^j\Psi(B)^{-1}W_t$, where $\Psi(z)=\sum_{j\ge 0}\psi_j z^j$ with $\psi_0=1$ has the coefficients of the innovations (Wold, MA(∞)) representation $W_t=\sum_{j\ge 0}\psi_j\varepsilon_{t-j}$ with ε_t the error of the mean square optimal forecast of W_t from W_s, $s<t$. Since $W_{t+h}-W_{t+h|t}=[\Psi]_0^{h-1}(B)\Psi^{-1}(B)W_{t+h}=[\Psi]_0^{h-1}(B)\varepsilon_{t+h}$, this forecast error is a moving average process of order (at most) $h-1$, as is also the error process of the forecasts $Y_{t+h|t}=\sum_{j=0}^{h-1}\tau_j W_{t+h-j|t}+\sum_{j=0}^{d-1}c_{j,h}Y_{t-j}$,

$$Y_{t+h}-Y_{t+h|t}=\sum_{j=0}^{h-1}\tau_j(W_{t+h-j}-W_{t+h-j|t})=\sum_{j=0}^{h-1}\tau_j B^j[\Psi]_0^{h-1-j}(B)\Psi^{-1}(B)W_{t+h}=\sum_{j=0}^{h-1}\tau_j B^j[\Psi]_0^{h-1-j}(B)\varepsilon_{t+h}, \tag{2}$$

where the backshift operators B^j operate on the t index.

The truncated filter forecast $\hat{W}_{t+h|t}$ and its error $W_{t+h}-\hat{W}_{t+h|t}$ are obtained from the infinite past formulas given above by setting W_{t-j}=0 for $j\ge t$. Denoting the filter in (2) by

$$\eta^{(h)}(B)=\left(\sum_{j=0}^{h-1}\tau_j B^j[\Psi]_0^{h-1-j}(B)\right)\Psi^{-1}(B), \tag{3}$$

it follows that, for the associated forecast $\hat{Y}_{t|t-h}$ of Y_t, the error process $\hat{\varepsilon}_t^{(h)} = Y_t - \hat{Y}_{t|t-h}$ is given by

$$\begin{aligned}\hat{\varepsilon}_t^{(h)} &= \eta^{(h)}(B)W_t, \quad W_{t-j} = 0,\ j \geq t \\ &= \sum_{j=0}^{t-1} \eta_j^{(h)} W_{t-j}, \quad t \geq 1. \end{aligned} \tag{4}$$

Now we generalize the notation to let $\Psi(B)$ in (3) denote the innovations filter of a not necessarily correct model for W_t, the log of whose continuous spectral density is integrable. This condition guarantees the existence of a (unique) continuous $\Psi(e^{-i\lambda}) = 1 + \sum_{j=1}^{\infty} \psi_j e^{-ij\lambda}$ satisfying $\int_{-\pi}^{\pi} \log|\Psi(e^{-i\lambda})|\, d\lambda = 0$ and such that the model spectral density is equal to $\sigma^2|\Psi(e^{-i\lambda})|^2$ for some $\sigma^2 > 0$, see Pourahmadi (2001, p. 68, Theorem VII). (For an ARMA model considered for W_t, $\Psi(B)$ has the form $\Psi(B) = \Omega(B)\Xi^{-1}(B)$, where $\Xi(B)$ is the AR polynomial and $\Omega(B)$ is the MA polynomial with no zeroes of magnitude less than one.) The only further requirement on the model is $\int_{-\pi}^{\pi} |\Psi(e^{-i\lambda})|^{-2} \tilde{f}(\lambda)\, d\lambda < \infty$, to ensure that its infinite past (quasi)innovations $\varepsilon_t = \Psi(B)^{-1} W_t$ for W_t are defined. Unless the true spectral density is given by $\tilde{f}(\lambda) = \sigma^2|\Psi(e^{-i\lambda})|^2$ for some $\sigma^2 > 0$, then the series ε_t will not be white noise and $\sum_{j=0}^{h-1} \tau_j B^j \varepsilon_t$ will generally not be a moving average process of order $h-1$.

One measure commonly used to evaluate the h-step forecast performance of a model is the average of squared forecast errors $n^{-1}\sum_{t=1}^{n}[\hat{\varepsilon}_t^{(h)}]^2$, where now we let $\hat{\varepsilon}_t^{(h)}$ denote the forecast error either from the truncated predictor or from the standard finite-past predictor discussed, for example, in Findley et al. (2004, Section 3.3.1). With either predictor, for an invertible ARIMA model, Findley (1991a, Proposition 4.1) shows, under the assumption

$$\sum_{j=-\infty}^{\infty} (2^{1/2} + |j|^{1/2})|\gamma_j(\tilde{f})| < \infty \tag{5}$$

that, as $n \to \infty$, with E denoting expectation, this average converges in probability at the rate $O_p(n^{-1/2})$ to the variance of $\eta^{(h)}(B)W_t$, which is given by $E(\eta^{(h)}(B)W_t)^2 = (1/2\pi)\int_{-\pi}^{\pi} |\eta^{(h)}(e^{-i\lambda})|^2 \tilde{f}(\lambda)\, d\lambda$. The same is true for the expression on the right in

$$\frac{1}{n}\sum_{t=1}^{n}[\hat{\varepsilon}_t^{(h)}]^2 \approx \frac{1}{2\pi}\int_{-\pi}^{\pi} |\eta^{(h)}(e^{-i\lambda})|^2 \tilde{f}(\lambda)\, d\lambda \approx \frac{1}{2\pi}\int_{-\pi}^{\pi} |\eta^{(h)}(e^{-i\lambda})|^2 I(\lambda)\, d\lambda, \tag{6}$$

where $I(\lambda) = n^{-1}|\sum_{t=1}^{n} W_t e^{-it\lambda}|^2$, the continuous-frequency periodogram of W_t, $t=1,2,\ldots,n$, see Lemma 3.1.1 of Taniguchi and Kakizawa (2000). For the random variates in (6), in Section A.3 of the Appendix, we derive under (5) the stronger approximation result

$$n^{1/2}\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} |\eta^{(h)}(e^{-i\lambda})|^2 I(\lambda)\, d\lambda - \frac{1}{n}\sum_{t=1}^{n}[\hat{\varepsilon}_t^{(h)}]^2\right\} \to_p 0. \tag{7}$$

These two rate-of-convergence results assume that $\eta^{(h)}(e^{-i\lambda})$ comes from a model that does not change with n, i.e., that has fixed rather than estimated parameters. The same assumption is used by the Diebold–Mariano test statistics for the hypothesis of equal asymptotic h-step forecast accuracy, as we discuss in Section 2.4. It follows from (7) that either random variate in (6) can be used as the measure of the model's h-step-ahead forecast performance in Diebold–Mariano type tests. We will use the expression on the right in (6) because of the theoretical results that are available to derive asymptotic distributions of test statistics with this measure, including tests that account for the effects of parameter estimation.

2.2. Computing the performance measures

For computation, setting $W = (W_1, \ldots, W_n)'$ and $g(\lambda) = |\eta^{(h)}(e^{-i\lambda})|^2$, a more convenient form of the performance measure is given by

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} g(\lambda) I(\lambda)\, d\lambda = \frac{1}{n} W' \Gamma(g) W, \tag{8}$$

because, in the case of an invertible ARIMA model, standard procedures can be used to calculate $\Gamma(g)$. For in this case, $\eta^{(h)}(e^{-i\lambda}) = (\sum_{j=0}^{h-1} \tau_j e^{-ij\lambda}[\Psi]_0^{h-1-j}(e^{-i\lambda}))\Xi(e^{-i\lambda})\Omega^{-1}(e^{-i\lambda})$, and $g(\lambda)$ can be regarded as the spectral density of an ARMA model with autoregressive polynomial $\Omega(B)$ and moving average polynomial of the form $\Phi(B)\Xi(B)$. Here $\Phi(B)$ has degree $h-1$ and coefficients that can be obtained recursively from the coefficient identity implied by $\Psi(B)\Xi(B) = \Omega(B)$ and the recursion for τ_j for $0 \leq j \leq h-1$ given above. For example, in the case $h=2$ and $d \geq 1$, $\Phi(B) = 1 + (\xi_1 + \omega_1 + \tau_1)B$. As a result, after multiplying out $\Phi(B)\Xi(B)$ to obtain the AR coefficients, the entries $\gamma_j(g) = (1/2\pi)\int_{-\pi}^{\pi} g(\lambda) e^{ij\lambda}\, d\lambda$ of $\Gamma(g)$ on r.h.s. of (8) are easily calculated with a standard recursive algorithm, see Brockwell and Davis (1991, p. 95), which is implemented in *R* (see R Development Core Team, 2008) and other widely used software. Similar calculations are used to compute our consistent variance alternative for Diebold–Mariano statistics derived in Section 2.4 and the asymptotic variances used to analyze the power study results in Section 3.2.

For the finite-past forecasts, $n^{-1}\sum_{t=1}^{n}[\hat{\varepsilon}_t^{(h)}]^2$ can be computed from $\tau_j, 0 \le j \le h-1$ and the covariance matrix of the ARMA model, see (3.13)–(3.15) of Findley et al. (2004).

2.3. The test statistic

Consider a model with parameter vector θ whose spectral density f_θ is such that $\log f_\theta$ is integrable for each θ ranging over a convex compact parameter set Θ. With $\Psi_\theta(B)$ denoting the model's innovations filter, the model spectral density is assumed to have the form $f_\theta(\lambda) = \sigma^2|\Psi_\theta(e^{-i\lambda})|^2$ with σ not functionally related to how θ determines $\Psi_\theta(e^{-i\lambda})$. Set $\eta_\theta^{(h)}(B) = \sum_{j=0}^{h-1}\tau_j B^j[\Psi_\theta]_0^{h-1-j}(B)\Psi_\theta^{-1}(B)$ and $g_\theta(\lambda) = |\eta_\theta^{(h)}(e^{-i\lambda})|^2$. A quasi-maximum-likelihood estimate (QMLE) of θ is, by definition, a minimizer of $D(f_\theta, I)$ over Θ, where $D(k,h)$ is the Kullback–Leibler (KL) discrepancy:

$$D(k,h) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\log k(\lambda) + \frac{h(\lambda)}{k(\lambda)}\right)d\lambda.$$

(See Dahlhaus and Wefelmeyer, 1996 for properties of QMLEs and MLEs in incorrect model situations.) With $\tilde{f}$ denoting the true spectral density of W_t, the pseudo-true value $\tilde{\theta}$ is, by definition, the minimizer of $D(f_\theta, \tilde{f})$ over $\theta \in \Theta$, which we assume to be unique. It will be the asymptotic limit of the QMLEs (and of the MLEs). The list of assumptions we use is a simple extension of the assumptions used by McElroy and Holan (2009):

1. W_t is stationary, mean zero, Gaussian and purely nondeterministic, i.e., $\int_{-\pi}^{\pi}\log\tilde{f}(\lambda)\,d\lambda$ is finite.
2. Θ is compact and convex.
3. $\tilde{\theta}$, the pseudo-true value of the model parameter θ, exists uniquely in the interior of Θ.
4. The model spectral density $f_\theta(\lambda)$ is twice continuously differentiable in θ and is continuous in λ.
5. The weighting function $g_\theta(\lambda) = |\eta_\theta^{(h)}(e^{-i\lambda})|^2$ is twice continuously differentiable in θ, and continuous in λ.
6. The matrix $M_f(\theta)$, which is the Hessian of the KL discrepancy between f_θ and $\tilde{f}$, is nonsingular at $\theta = \tilde{\theta}$.
7. The first derivative of $f_\theta(\lambda)$ is uniformly bounded and bounded away from zero (in λ).

Apart from the Gaussian requirement, these assumptions are typical for the literature on this topic. The Gaussian assumption is needed for the theory to cover MLEs, as discussed in Dahlhaus and Wefelmeyer (1996); if only QMLEs are of interest, Gaussianity can be relaxed. If Θ specifies only invertible ARIMA models whose AR and MA polynomials have no common zeroes, then 4 and 5 hold. If, in addition, the correct model is specified by an interior point of Θ, then 3 and 6 also hold, see Brockwell and Davis (1991, Section 10.8). Further, when there is only a pseudo-true model in the interior of Θ, then 3 and 6 will continue to hold if its spectral density is sufficiently close to the true spectral density in the Kullback–Leibler sense—see Ploberger (1982). More generally, when 3 holds, it seems reasonable to expect that 6 usually will, too. Our goal is to compare the h-step-ahead forecast performance of two fitted models with the correct differencing operator for the data that have parameters $\theta^{(i)}$ in $\Theta^{(i)}$ and unique pseudo-true values $\tilde{\theta}^{(i)}$, $i=1, 2$. (The forecast lead h is the same for both models—otherwise we would not be evaluating them on the same footing.) For $i=1, 2$, we define $g_i = g_{\theta^{(i)}}$ and $\tilde{g}_i = g_{\tilde{\theta}^{(i)}}$, $i=1, 2$. In the Appendix we establish the following result.

The statistics and null hypotheses we consider can be expressed in a unified and simple way in terms of functions of the form

$$Q(f,g,\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\lambda)g_\theta(\lambda)\,d\lambda, \tag{9}$$

in which $f(\lambda)$ can be stochastic, as on the l.h.s. of (8), or nonstochastic.

Theorem 1. *Under conditions* 1–6 *for both models, with* $\hat{\theta}^{(i)}$ *denoting the QMLEs (if they are MLEs, also assume condition* 7*) over their respective parameter sets* $\Theta^{(i)}$, $i=1, 2$, *we have*

$$\sqrt{n}(Q(I,\hat{g}_1,\hat{\theta}^{(1)}) - Q(I,\hat{g}_2,\hat{\theta}^{(2)})) - \sqrt{n}(Q(\tilde{f},\tilde{g}_1,\tilde{\theta}^{(1)}) - Q(\tilde{f},\tilde{g}_2,\tilde{\theta}^{(2)})) \xrightarrow{\mathcal{L}} \mathcal{N}(0,V),$$

where V has the formula

$$V = \frac{1}{\pi}\int_{-\pi}^{\pi}\tilde{f}^2(\tilde{g}_1 + \tilde{p}_1 - \tilde{g}_2 - \tilde{p}_2)^2\,d\lambda, \tag{10}$$

with $\tilde{p}_1 = p_{\tilde{\theta}^{(1)},1}$ *and* $\tilde{p}_2 = p_{\tilde{\theta}^{(2)},2}$ *defined below. Further, V is consistently estimated by*

$$\hat{V} = \frac{1}{2\pi}\int_{-\pi}^{\pi} I^2(\hat{g}_1 + \hat{p}_1 - \hat{g}_2 - \hat{p}_2)^2\,d\lambda,$$

where $\hat{p}_1$ *and* $\hat{p}_2$ *are the result of substituting the periodogram I for* $\tilde{f}$ *and QMLEs or MLEs for pseudo-true values in the formulas defining* p_i *(for* $i=1, 2$*):*

$$p_{\theta^{(i)},i}(\lambda) = f_{\theta^{(i)}}^{-2}(\lambda)b'_{\theta^{(i)},i}M_f^{-1}(\theta^{(i)})\nabla_{\theta^{(i)}}f_{\theta^{(i)}}(\lambda),$$

$$b_{\theta^{(i)},i} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \tilde{f}(\lambda)\nabla_{\theta^{(i)}} g_{\theta^{(i)},i}(\lambda)\, d\lambda,$$

$$M_f(\theta^{(i)}) = \nabla_{\theta^{(i)}}\nabla'_{\theta^{(i)}} D(f_{\theta^{(i)}},\tilde{f}).$$

Our null hypothesis is that the pseudo-true models have equal asymptotic average squared h-step ahead forecast performance, defined as in (6),

$$H_0 : Q(\tilde{f},\tilde{g}_1,\tilde{\theta}^{(1)}) = Q(\tilde{f},\tilde{g}_2,\tilde{\theta}^{(2)}), \tag{11}$$

which is the same as $E(\eta^{(h)}_{\tilde{\theta}^{(1)}}(B)W_t)^2 = E(\eta^{(h)}_{\tilde{\theta}^{(2)}}(B)W_t)^2$. The two-sided alternative to (11), i.e. $Q(\tilde{f},\tilde{g}_1,\tilde{\theta}^{(1)}) \neq Q(\tilde{f},\tilde{g}_2,\tilde{\theta}^{(2)})$ will be the focus in our empirical study, but one-sided alternatives will be considered in the size study we present of our test statistic for (11),

$$T_{\hat{V}} = (\hat{V}/n)^{-1/2}(Q(I,\hat{g}_1,\hat{\theta}^{(1)}) - Q(I,\hat{g}_2,\hat{\theta}^{(2)})). \tag{12}$$

When (11) holds as well as the assumptions of Theorem 1, and in (10), $V > 0$, it follows that this statistic has a standard Gaussian limit distribution, $T_{\hat{V}} \stackrel{\mathcal{L}}{\Longrightarrow} \mathcal{N}(0,1)$.

Because $Q(\tilde{f},\tilde{g}_1,\tilde{\theta}^{(1)}) - Q(\tilde{f},\tilde{g}_2,\tilde{\theta}^{(2)})$ is the integral of $\tilde{f}(\tilde{g}_1 - \tilde{g}_2)$ over $[-\pi,\pi]$, the test based on $T_{\hat{V}}$ will have adequate power for distinguishing between the h-step forecasting performance of the two models when $\sqrt{n}$ times the integral of this function is adequately large in magnitude in units of $\sqrt{V}$. Recall that the pseudo-true values $\tilde{\theta}^{(1)}$ and $\tilde{\theta}^{(2)}$ are minimizers of the KL distance to the true spectrum $\tilde{f}$, and thus are associated with minimizing one-step-ahead mean square forecast error from each model. When $h > 1$, the function $\tilde{f}(\tilde{g}_1 - \tilde{g}_2)$ includes the multi-step-ahead performance of each model through the forecast error filter functions used to define $\tilde{g}_1$ and $\tilde{g}_2$.

By replacing each model's g in $Q(I,g,\theta)$ with a positively weighted linear combination of functions g over several forecast leads h, one can assess model forecast performance over all of these leads simultaneously. Future research will examine this type of diagnostic, in order to find the models that forecast well at a suite of future horizons—this is important for seasonal adjustment with X-12-ARIMA—see Findley et al. (1998)—which extends the series with one or more years of forecasts before applying seasonal adjustment filters.

2.4. The case of constant parameters

Meese and Rogoff (1988) seems to be the earliest article in which, for a stationary Gaussian series satisfying standard conditions, the limiting distribution under the null hypothesis (11) of the difference of average multi-step forecast squared errors (the l.h.s. of (6)) from two models is obtained together with an estimate of the variance of the distribution, and thereby a test statistic for the null hypothesis. The resulting test has become known as the Diebold–Mariano test through its appearance (with credit to Meese and Rogoff) in Diebold and Mariano (1995). In these references, the limiting distribution was obtained by treating the forecast errors as stationary, which is the situation of errors $\varepsilon^{(h)}_t(\theta^{(1)})$ and $\varepsilon^{(h)}_t(\theta^{(2)})$ of forecasts from the infinite past from models whose parameters $\theta^{(1)}$ and $\theta^{(2)}$ are constant rather than estimated. The assumed null hypothesis is thus

$$H_0 : Q(\tilde{f},g_1,\theta^{(1)}) = Q(\tilde{f},g_2,\theta^{(2)}). \tag{13}$$

Unaware of the work of Meese and Rogoff, for the null hypothesis (13) and for a large class of stationary time series models, Findley (1991a) obtained a limiting distribution equivalent to theirs for the errors $\hat{\varepsilon}^{(h)}_t(\theta^{(i)})$ from the standard *finite-past* predictors defined by constant parameters $\theta^{(i)}$, $i=1, 2$,

$$n^{-1/2}\left(\sum_{t=1}^{n-h}[\hat{\varepsilon}^{(h)}_t(\theta^{(1)})]^2 - \sum_{t=1}^{n-h}[\hat{\varepsilon}^{(h)}_t(\theta^{(2)})]^2\right) \stackrel{\mathcal{L}}{\Longrightarrow} \mathcal{N}(0,V_c),$$

but provided no estimator for the limiting variance, which we denote by V_c, where the subscript c indicates the treatment of the parameters as constant. Meese and Rogoff's formula for V_c will be presented below in (17) and shown to have the value

$$V_c = \frac{1}{\pi}\int_{-\pi}^{\pi} \tilde{f}^2(g_1 - g_2)^2\, d\lambda. \tag{14}$$

This is the variance that Theorem 1 yields for the constant parameter case under (13),

$$\sqrt{n}(Q(I,g_1,\theta^{(1)}) - Q(I,g_2,\theta^{(2)})) \stackrel{\mathcal{L}}{\Longrightarrow} \mathcal{N}(0,V_c),$$

because the terms in (10) involving derivatives with respect to the parameters drop out. Theorem 1 provides a consistent estimate of V_c,

$$\hat{V}_c = \frac{1}{2\pi}\int_{-\pi}^{\pi} I^2(g_1-g_2)^2\,d\lambda = \sum_{j,k=-n+1}^{n-1} \hat{\gamma}_j\hat{\gamma}_k\{\gamma_{j-k}(g_1^2)+\gamma_{j-k}(g_2^2)-2\gamma_{j-k}(g_1g_2)\}, \tag{15}$$

with $\hat{\gamma}_j = n^{-1}\sum_{t=|j|+1}^{n} W_tW_{t-|j|}$, $-n+1 \le j \le n-1$. Thus we have a test statistic

$$T_{\hat{V}_c} = (\hat{V}_c/n)^{-1/2}(Q(I,g_1,\theta^{(1)})-Q(I,g_2,\theta^{(2)})). \tag{16}$$

The simplifications of the proof of Theorem 1 that result from using constant parameters show that $T_{\hat{V}_c}$ has an $\mathcal{N}(0,1)$ limiting distribution when, along with conditions 1–3 of Theorem 1, the model spectral density and weighting functions are continuous functions of their parameters and also $g_1 \neq g_2$ holds, so that $V_c > 0$. The same is true for the time series generalization of the nonnested model comparison test statistic of Vuong (1989) in Findley (1990) if the $h=1$ instance of $\hat{V}_c$ replaces the robust estimate of asymptotic variance used in this report's applications, which was not shown to be consistent.

In the case of competing ARIMA models, the model autocovariances on the right in (15) can be calculated by identifying the coefficients of the ARMA models whose spectral densities are g_1^2, g_2^2 and g_1g_2 and from these coefficients obtaining the autocovariances, as in Section 2.2. In the numerical studies below, the parameters treated as constant are Maximum Likelihood estimates from $W_1,\ldots,W_n$. The calculation of $\hat{V}$ is much more complex because of its terms that involve derivatives.

To present Meese and Rogoff's formula for V_c and their estimate $\hat{V}_{DM}$ as described for general h by Diebold and Mariano (1995), set $v_t = \varepsilon_t^{(h)}(\theta^{(1)})+\varepsilon_t^{(h)}(\theta^{(2)})$ and $w_t = \varepsilon_t^{(h)}(\theta^{(1)})-\varepsilon_t^{(h)}(\theta^{(2)})$ and observe that $[\varepsilon_t^{(h)}(\theta^{(1)})]^2-[\varepsilon_t^{(h)}(\theta^{(2)})]^2 = v_tw_t$. Thus, the null hypothesis (11) is equivalent to $Ev_tw_t=0$, and $n^{-1/2}$ times the difference of the average squared forecast errors is equal to $n^{-1/2}\sum_{t=1}^{n} v_tw_t$, whose normal limiting distribution under the null has the well-known variance formula

$$V_{c,MR} = \sum_{r=-\infty}^{\infty}[\gamma_{vv}(r)\gamma_{ww}(r)+\gamma_{vw}(r)\gamma_{vw}(-r)], \tag{17}$$

in the Gaussian case. In Section A.4 of the Appendix, we verify that

$$V_{c,MR} = V_c. \tag{18}$$

Motivated by the fact discussed above that with correct models, the h-step-ahead forecast errors $\varepsilon_t^{(h)}$ form a moving average process of order $h-1$, Meese and Rogoff (1988) and Diebold and Mariano (1995, p. 257) propose the estimator of $V_{c,MR}=V_c$ defined by

$$\hat{V}_{DM} = \sum_{r=-h+1}^{h-1}\left(1-\frac{|r|}{n}\right)[\hat{\gamma}_{vv}(r)\hat{\gamma}_{ww}(r)+\hat{\gamma}_{vw}(r)\hat{\gamma}_{vw}(-r)], \tag{19}$$

with sample cross-covariance estimates $\hat{\gamma}_{vv}(r)$, $\hat{\gamma}_{ww}(r)$, $\hat{\gamma}_{vw}(r)\hat{\gamma}_{vw}(-r)$ defined by the observed in-sample forecast errors from the estimated models. This $\hat{V}_{DM}$ converges to

$$V_{DM} = \sum_{r=-h+1}^{h-1}[\gamma_{vv}(r)\gamma_{ww}(r)+\gamma_{vw}(r)\gamma_{vw}(-r)], \tag{20}$$

which is to be regarded, and judged, as an approximation to V_c. The equality $V_{DM}=V_c$ holds only in very special situations. For example, it holds when the series being modeled is a moving average process of order less than h, or when both models being compared contain the correct model as a special case. However, in this correct model situation, at the asymptotic (true) parameter values, $w_t=0$ and $V_{DM}=V_c=0$, a situation in which the test statistic proposed by these authors has not been shown to have a limiting distribution.[1] In the empirical results of the next section, for uniformity, the Diebold–Mariano test statistic is taken as $T_{DM} = (\hat{V}_{DM}/n)^{-1/2}(Q(I,g_1,\hat{\theta}^{(1)})-Q(I,g_2,\hat{\theta}^{(2)}))$, which differs from their actual statistic to the extent of the effect of the approximation errors in (7) for both models. However, it has the same $\mathcal{N}(0,V_c/V_{DM})$ limit distribution when $V_{DM} \neq 0$.

Remark. In the important case $h=1$, which is also relevant for likelihood-ratio-based model selection, see Findley (1990), we have $V=V_c$ because, for each model family, the vector $b(\tilde{\theta})$ of Theorem 1 is zero. This happens because in this case, $b(\theta)$ is the gradient of $(1/2\pi)\int_{-\pi}^{\pi}\tilde{f}(\lambda)|\Psi_\theta(e^{-i\lambda})|^{-2}\,d\lambda$ and the pseudo-true value $\tilde{\theta}$ minimizes this integral and lies in the interior of Θ by Assumption 3. Thus, when $h=1$, we can expect similar results from $T_{\hat{V}}$ and $T_{\hat{V}_c}$ with large enough samples.

[1] Clark and McCracken (2001, 2005) have obtained limiting distributions in this situation for related encompassing statistics comparing out-of-sample forecasting performance of competing regression models with jointly stationary variables. The distributions are determined by the limit of the ratio of the number of data held out of sample to the number of in-sample data used to forecast the withheld data.

3. Numerical studies

First we report a size study of the statistics $T_{\hat{V}}$, $T_{\hat{V}_c}$, and $T_{\hat{V}_{DM}}$ obtained from the only kind of examples known to us of pairs of incorrect models for which the null hypothesis (11) is satisfied, namely pairs of autoregressive models like those described in Findley (Section 3, 1991b) involving models with coefficient gaps. For a misspecified autoregressive model, the pseudo-true coefficient vector and its associated asymptotic mean square forecast error (AMSFE) $Q(\tilde{f},\tilde{g},\tilde{\theta}) = E(\eta_{\tilde{\theta}}^{(h)}(B)W_t)^2$ for the case $h=1$ both have simple general formulas. These facilitate finding nonnested pairs of incorrect autoregressive models with $\tilde{g}_1 \neq \tilde{g}_2$ such that for $h=1$, the null hypothesis of equal AMSFEs holds, $Q(\tilde{f},\tilde{g}_1,\tilde{\theta}^{(1)}) = Q(\tilde{f},\tilde{g}_2,\tilde{\theta}^{(2)})$, and also $V > 0$. After the size study, we present simulation-based power studies, some of which involve nonstationary series and values $h > 1$. Pseudo-true coefficients are used in the evaluation of AMSFEs and asymptotic variances V_c and V_{DM} as well as V, because estimated rather than fixed coefficients are used with each simulation and series length, and the pseudo-true coefficients are the theoretical limits of the estimates from each data generating process (DGP). Therefore, the statistics $T_{\hat{V}}$, $T_{\hat{V}_c}$, and $T_{\hat{V}_{DM}}$ differ only in their denominators, which converge to the asymptotic standard errors $\sqrt{V}$, $\sqrt{V_c}$, and $\sqrt{V_{DM}}$, respectively. These limit quantities will be shown to be good indicators of the relative power properties of the three statistics in finite samples. After the simulation studies (done with *R*), we present the results of applying $T_{\hat{V}}$, $T_{\hat{V}_c}$, and $T_{\hat{V}_{DM}}$ to recommended and alternative models for some published time series.

Remark. The simulation results presented as size studies of $T_{\hat{V}_{DM}}$ in Section 3 of Diebold and Mariano (1995) are not valid for this purpose. They assume a series W_t can exist that has two incorrect models with $|\psi_1| < 1$ whose 2-step-ahead forecast errors processes are distinct invertible MA(1) processes with the same MA coefficient. There is no such W_t: For models with $|\psi_1| < 1$, for a given MA(1) forecast error polynomial $\Omega(B) = 1+\omega_1 B$ with $|\omega_1| < 1$, the MA(1) process is unique, being given by $\Omega(B)\varepsilon_t$ where ε_t is the innovations process of W_t. Indeed, more generally, for any $h \geq 1$, if the zeroes of $[\Psi]_0^{h-1}(z)$ and $\Omega(z)$ lie in $\{|z| > 1\}$, then from $\eta^{(h)}(B)W_t = [\Psi]_0^{h-1}(B)\Psi^{-1}(B)W_t = \Omega(B)e_t$, we have $W_t = \tilde{\Psi}(B)e_t$ for $\tilde{\Psi}(B) = \Psi(B)([\Psi]_0^{h-1}(B))^{-1}\Omega(B)$. Because $\tilde{\Psi}^{-1}(B) = \Omega^{-1}(B)[\Psi]_0^{h-1}(B)\Psi^{-1}(B)$ is causal, if $e_t = \tilde{\Psi}^{-1}(B)W_t$ is white noise, then $\tilde{\Psi}(B)$ is the innovations filter of W_t and $e_t = \varepsilon_t$.

3.1. Size studies

We use the easily verified fact that when fitting a possibly incorrect AR(p) model to the time series W_t, the pseudo-true coefficient vector $\tilde{\xi} = (\tilde{\xi}_1,\ldots,\tilde{\xi}_p)'$ has the entries that minimize $E(W_t - \sum_{j=1}^p \xi_j W_{t-j})^2$. Thus $\tilde{\xi}$ is the solution to the Yule–Walker equation defined by W_t's true autocovariances, $\Gamma(\tilde{f})\tilde{\xi} = \boldsymbol{\gamma}$, where the covariance matrix is p dimensional and $\boldsymbol{\gamma} = (\gamma_1(\tilde{f}),\ldots,\gamma_p(\tilde{f}))'$. Hence, when $h=1$ the AMSFE is equal to $E(W_t - \sum_{j=1}^p \tilde{\xi}_j W_{t-j})^2 = \gamma_0(\tilde{f}) - \boldsymbol{\gamma}'\Gamma^{-1}(\tilde{f})\boldsymbol{\gamma}$.

For an AR(1) model, $\tilde{\xi}_1 = \gamma_1(\tilde{f})/\gamma_0(\tilde{f})$, the lag one autocorrelation of W_t. For the null hypothesis, we must find two different models such that the corresponding AMSFEs are equal, but without their spectra and weighting functions (evaluated at pseudo-true values) being equal—this excludes nested models, for example. Let an AR(1) be the first model, and let an AR(2) model with AR polynomial of the constrained form $\Xi(B) = 1-\xi_2 B^2$ be the second. This model's pseudo-true coefficient, the minimizer of $E(W_t - \xi_2 W_{t-2})^2$, is $\tilde{\xi}_2 = \gamma_2(\tilde{f})/\gamma_0(\tilde{f})$, the lag two autocorrelation of W_t, and the AMSFE for $h=1$ is $\gamma_0(\tilde{f}) - \gamma_2^2(\tilde{f})/\gamma_0(\tilde{f})$.

The two AMSFEs will be equal when W_t is such that $\gamma_1(\tilde{f}) = \gamma_2(\tilde{f})$. An MA(2) process of the form $W_t = (1+1/3B+1/2B^2)\varepsilon_t$ has this property and, with Gaussian white noise, will be the "Null DGP" for our size study (its innovation variance is irrelevant for our purposes). It is easy to see that $g_1 \neq g_2$ at the pseudo-true values, so the asymptotic variances of the test statistics are nonzero. We will generate data from the Null DGP in order to assess the size properties of the statistics, for $h=1$. Another choice of h, or a model with $\delta(B) \neq 1$, would have different AMSFE formulas and would thereby require other constraints on $\tilde{f}$.

For the size study,we simulated 1000 Gaussian time series from the Null DGP described above, with sample sizes n=50,100,200. The three test statistics $T_{\hat{V}}$,$T_{\hat{V}_c}$, and $T_{\hat{V}_{DM}}$ (for $h=1$) were then applied and the coverage was computed. Nominal coverage was $\alpha = 0.05, 0.10$; in Table 1 we give empirical coverage for each method,for both the left and right tails (the tests are considered as if one-sided,so the table entries should be compared against the limiting values of $\alpha = 0.05, 0.10$ for each tail). There is very similar under-coverage for $T_{\hat{V}}$ and $T_{\hat{V}_c}$ at $n=50$ with some asymmetry, which disappears as the data length increases. The similarity is expected because $h=1$,so the statistics' denominators coincide asymptotically,$\sqrt{V} = \sqrt{V_c}(= 1.239)$. (The calculation of the asymptotic variances is discussed at the end of Section 3.2.) Except at $n=50$ and $\alpha = 0.05$,$T_{\hat{V}_{DM}}$ has over-coverage throughout. The fact that $T_{\hat{V}_{DM}}$ consistently made more Type I errors is what one might expect from $\sqrt{V_{DM}} = 1.020$, and the fact that $T_{\hat{V}_{DM}}$ is asymptotically equivalent to $T_{\hat{V}_c}$ multiplied by $\sqrt{V_c}/\sqrt{V_{Dm}} = 1.215$. (Thus $\sqrt{V_{DM}}$ is not a good approximation to $\sqrt{V_c}$ in this case.) None of the tests is grossly mis-sized, but the correctly normalized test statistic $T_{\hat{V}}$ performs best.

Table 1
One-sided size coverage for $\alpha = 0.05, 0.10$ for the three $h=1$ test statistics comparing the pair of AR models, at sample sizes $n=50$, 100, 200.

Sample	$\alpha = 0.05$						$\alpha = 0.10$					
	$T_{\hat{V}}$		$T_{\hat{V}_c}$		$T_{\hat{V}_{DM}}$		$T_{\hat{V}}$		$T_{\hat{V}_c}$		$T_{\hat{V}_{DM}}$	
50	0.047	0.022	0.045	0.018	0.051	0.025	0.119	0.060	0.115	0.055	0.136	0.064
100	0.057	0.035	0.054	0.035	0.079	0.051	0.093	0.106	0.105	0.089	0.137	0.119
200	0.042	0.041	0.042	0.039	0.064	0.070	0.093	0.091	0.092	0.090	0.135	0.125

In each cell, the left hand entry is for the left tail, and the right hand entry for the right tail of the distribution from 1000 simulations of the MA(2) processes. The closeness of the results for $T_{\hat{V}}$ and $T_{\hat{V}_c}$ with larger samples is not surprising because $V_c = V$ when $h=1$.

3.2. Power studies

We now present the results of simulation experiments to determine the probabilities that one-sided tests with the three test statistics reject the null hypothesis in favor of the model with smaller AMSFE in various stationary and nonstationary situations. Consider first the fitting of an AR(1) model to data from an MA(1) process W_t with MA polynomial $\Omega(B) = 1+\omega_1 B$ ($\omega_1 \neq 0$) and unit innovation variance, $E(\varepsilon_t^2) = 1$. In this situation, the AMSFE of the AR(1) model for $h=1$ is $(1+\omega_1^2+\omega_1^4)/(1+\omega_1^2)$, which is always different from the AMSFE of the MA(1) model, $E(\varepsilon_t^2) = 1$. Hence, for comparing the $h=1$ performance of AR(1) and MA(1) models, the null hypothesis is false. From the autoregressive representation $\sum_{j=0}^{\infty}(-\omega_1)^j W_{t-j} = \varepsilon_t$, one expects it to be difficult to detect inadequacies of the AR(1) model when $|\omega_1|$ is small. Indeed, for $\omega_1 = 0.2, 0.5, 0.8$, the corresponding AR(1) AMSFE values are 1.0015, 1.05, and 1.2498, to be compared to the MA(1) AMSFE value $E[\varepsilon_t^2] = 1$ for the correct MA(1) model. We therefore proceeded with power studies using $\omega_1 = 0.5, 0.8$ and fitting the AR(1), MA(1), and MA(2) models to these processes, ignoring lower values like $\omega_1 = 0.2$. We also omitted the comparison of MA(1) to MA(2) as the pseudo-true MA(2) model coincides with the true MA(1) model, with the result that the AMSFEs are the same and $V=0$.

So we consider the comparison of AR(1) to MA(1) (in which the latter is favored, being correctly specified) as well as AR(1) to MA(2). These are similar situations, since in a sense the MA(2) is not incorrect, as it nests the true model. However, when $\omega_1 = 0.5$ the null hypothesis is close to being satisfied, so the power should be close to the α levels. But when $\omega_1 = 0.8$, the AMSFEs are sufficiently different that we can expect the power to be much higher (in favor of the moving average models against the autoregressive model). These observations are largely borne out by the results in Tables 2, 4 and 5. Of course, this discussion pertains to the $h=1$ case; different AMSFEs are involved when $h=2$ and ARIMA data and models are considered with $d=0$, 1, 2, as Table 2 details.

We also look at something slightly different: we generate data from the MA(2) model $W_t = (1+0.25B+0.5B^2)\varepsilon_t$ with unit innovation variance. Now both the AR(1) and MA(1) models are incorrect, because the MA(2) model is correct. Using the formula for AR(1) AMSFE given in Section 3.1, we obtain the value 1.205—a medium discrepancy from the optimal AMSFE of 1. Thus, in comparing the AR(1) and MA(2) fits, the latter is certainly favored, and we can expect decent power. Fitting an MA(1) to the MA(2) yields the pseudo-true value for the MA(1) model's coefficient in the form

$$\frac{(1+2\omega_2+\omega_1^2+\omega_2^2) \pm \sqrt{(1+2\omega_2+\omega_1^2+\omega_2^2)^2 - 4(\omega_1+\omega_1\omega_2)^2}}{2(\omega_1+\omega_1\omega_2)},$$

the choice of $\pm$ being made so that the coefficient's magnitude is less than 1. For the particular MA(2), the MA(1) pseudo-true value is $\tilde{\omega}_1 = \frac{1}{6}$. The AMSFE formula is

$$\frac{(1+\omega_1^2+\omega_2^2) - 2\tilde{\omega}_1(\omega_1+\omega_1\omega_2) + 2\tilde{\omega}_1^2\omega_2}{1-\tilde{\omega}_1^2},$$

times the innovation variance. In our case, this yields the AMSFE value 1.25. So if we compare the AR(1) and MA(1) (both are incorrect) fitted to this MA(2), the difference in AMSFEs is $1.205 - 1.25 = -.045$, indicating a slight preference for the AR(1). We expect the power to be low—perhaps even close to the nominal size for small samples—in this case. This is borne out by the results in Table 6 for $d=0$. Note that we could also do comparisons of MA(1) to MA(2) fits (they are not nested in this case, since their AMSFEs differ), but this is omitted for uniformity of presentation.

For the power studies, we simulate from these DGPs and fit three models, making two comparisons as discussed above. We examine the sample sizes $n=50$, 100, 200 and the α levels 0.05, 0.10 just as in the size study. We restrict the power calculation for the three statistics to the relevant tail (either left or right, depending on which model is favored), since if there is power in the positive direction there should be negligible power in the negative direction, and vice versa. When $h=1$, the differencing order is irrelevant, but for $h=2$ it is important, as the Tables 2 to 6 show.

Table 3 provides for each model comparison the pertinent AMSFE difference from Table 2, the values of $\sqrt{V}$, $\sqrt{V_c}$, and $\sqrt{V_{DM}}$, and finally the value of the AMSFE difference normalized by division by $\sqrt{V}$. It follows from Theorem 1 that the power of each of $T_{\hat{V}}$, $T_{\hat{V}_c}$, and $T_{\hat{V}_{DM}}$ statistics is governed asymptotically by the AMSFE difference divided by $\sqrt{V}$, $\sqrt{V_c}$, and

$\sqrt{V_{DM}}$, respectively. The larger such a normalized AMSFE difference is in absolute value, i.e., the smaller its denominator, the greater one can expect the power of the statistic to be in sufficiently large finite samples. In most cells of Table 3, one finds $\sqrt{V} \le \sqrt{V_c} \le \sqrt{V_{DM}}$, with $\sqrt{V} = \sqrt{V_c}$ holding only for $h=1$, and $\sqrt{V_c} = \sqrt{V_{DM}}$ holding only for $h=2$ with MA(1) processes, as dictated by results presented in Section 2.4. When the inequalities are strict in these cells, almost always in the corresponding cells of Tables 4–6, the statistic $T_{\hat{V}}$ has greater power than $T_{\hat{V}_c}$, which has greater power than $T_{\hat{V}_{DM}}$. (So in particular, accounting for parameter estimation frequently reduces the variability of the test statistic.) In general, the ordering of $\sqrt{V}$, $\sqrt{V_c}$, and $\sqrt{V_{DM}}$ from least to greatest is a good predictor of the power ranking from greatest to least of the associated tabled statistics. Further, because $\sqrt{V_{DM}}$ is evaluated by its success as an approximation to $\sqrt{V_c}$, whenever $\sqrt{V_c}/\sqrt{V_{DM}}$ differs much from one, any appreciable power advantage for $T_{\hat{V}_{DM}}$, such as is seen with the MA(2) DGP in Table 6 for $h=1$, should be viewed as an artefact of deficient approximation and disregarded. To see further evidence of the

Table 2
AMSFE values for various MA component model DGPs, leads, and models.

DGPs	MA(1), 0.5			MA(1), 0.8			MA(2)		
Models	AR(1)	MA(1)	MA(2)	AR(1)	MA(1)	MA(2)	AR(1)	MA(1)	MA(2)
$h=1$	1.05	1	1	1.250	1	1	1.205	1.250	1
$h=2$, $d=0$	1.282	1.25	1.25	1.733	1.640	1.640	1.240	1.313	1.063
$h=2$, $d=1$	3.332	3.25	3.25	4.583	4.240	4.240	2.909	3.146	2.563
$h=2$, $d=2$	7.482	7.25	7.25	9.932	8.84	8.84	6.990	7.479	6.063

The top row shows the three DGPs (MA(1) with coefficient either 0.5 or 0.8, and MA(2) with coefficients 0.25 and 0.5), and below each DGP the three ARMA component models, AR(1), MA(1), and MA(2) associated with the values of d considered, $d=0$, 1, 2. The cells of the subsequent rows present each ARMA or ARIMA model's AMSFE value for the DGP, first for $h=1$ and then, when $h=2$, according the value of d that determines the model.

Table 3
Differences between AMSFE values of Table 2 for two component model comparisons, AR(1) to MA(1) and AR(1) to MA(2) [first row], for each component model DGP [second row], MA(1) with coefficient either 0.5 or 0.8, and MA(2) with coefficients 0.25 and 0.5.

Models	AR(1) vs. MA(1)			AR(1) vs. MA(2)		
DGP	MA(1), 0.5	MA(1), 0.8	MA(2)	MA(1), 0.5	MA(1), 0.8	MA(2)
$h=1$	0.050	0.250	−0.045	0.050	0.250	0.205
	0.437	0.937	0.429	0.437	0.937	0.984
	0.437	0.937	0.429	0.437	0.937	0.984
	0.453	1.060	0.327	0.453	1.060	0.952
	0.115	0.267	−0.105	0.115	0.267	0.208
$h=2$, $d=0$	0.032	0.093	−0.073	0.032	0.093	0.177
	0.329	0.584	0.701	0.329	0.584	0.910
	0.454	0.925	0.259	0.454	0.925	0.859
	0.454	0.925	0.238	0.454	0.925	0.891
	0.097	0.159	−0.104	0.097	0.159	0.194
$h=2$, $d=1$	0.082	0.343	−0.237	0.082	0.343	0.346
	0.947	1.994	1.955	0.947	1.994	1.724
	1.127	2.321	1.209	1.127	2.321	2.112
	1.127	2.321	1.166	1.127	2.321	2.224
	0.087	0.172	−0.121	0.087	0.172	0.201
$h=2$, $d=2$	0.232	1.092	−0.489	0.232	1.092	0.927
	2.321	5.054	4.048	2.321	5.054	4.470
	2.537	5.469	3.022	2.537	5.469	4.759
	2.537	5.469	2.909	2.537	5.469	4.962
	0.100	0.216	−0.121	0.100	0.216	0.207

In the cells below a DGP specification, the AMSFE differences are shown for the model comparison, first for $h=1$ and then, when $h=2$, according the value of d that determines the model. Within each cell the actual AMSFE difference is given first, followed by $\sqrt{V}$, $\sqrt{V_c}$ and $\sqrt{V_{DM}}$, and finally the AMSFE difference normalized by division by $\sqrt{V}$. In accord with the analysis of Section 3.2, the ordering of $\sqrt{V}$, $\sqrt{V_c}$ and $\sqrt{V_{DM}}$ from least to greatest, whenever there is strict inequality, is almost always predictive of the power rankings of the statistics, from greatest to least, that are observable in Tables 4–6.

Table 4
Power for (one-sided) $\alpha = 0.05, 0.10$ for the three test statistics, at sample sizes n=50, 100, 200 and forecast leads h=1,2.

DGP: 0.5		AR(1) vs. MA(1)						AR(1) vs. MA(2)					
Lead	Sample	$T_{\hat{V}}$		$T_{\hat{V}_c}$		$T_{\hat{V}_{DM}}$		$T_{\hat{V}}$		$T_{\hat{V}_c}$		$T_{\hat{V}_{DM}}$	
h=1	50	0.060	0.185	0.051	0.152	0.025	0.098	0.087	0.239	0.075	0.208	0.048	0.174
	100	0.210	0.458	0.193	0.414	0.124	0.347	0.228	0.488	0.213	0.447	0.163	0.408
	200	0.475	0.678	0.459	0.668	0.379	0.619	0.505	0.732	0.473	0.719	0.419	0.675
h=2	50	0.072	0.208	0.109	0.221	0.106	0.200	0.073	0.211	0.060	0.161	0.048	0.144
d=0	100	0.252	0.434	0.195	0.313	0.178	0.296	0.199	0.402	0.129	0.272	0.113	0.251
	200	0.398	0.535	0.227	0.373	0.223	0.362	0.363	0.549	0.196	0.358	0.178	0.358
h=2	50	0.022	0.078	0.049	0.130	0.041	0.120	0.053	0.164	0.061	0.182	0.059	0.163
d=1	100	0.070	0.222	0.116	0.232	0.111	0.226	0.152	0.328	0.144	0.292	0.148	0.272
	200	0.215	0.402	0.199	0.345	0.197	0.327	0.290	0.473	0.221	0.372	0.213	0.365
h=2	50	0.032	0.112	0.046	0.122	0.045	0.121	0.064	0.195	0.064	0.199	0.069	0.195
d=2	100	0.109	0.302	0.133	0.293	0.127	0.284	0.180	0.402	0.172	0.375	0.180	0.360
	200	0.332	0.519	0.284	0.470	0.280	0.448	0.361	0.589	0.318	0.513	0.312	0.508

The true processes are ARIMA(0,d,1) with MA(1) coefficient equal to 0.5 for each specified d. We first compare models with the AR(1) stationary component to models with the MA(1) component, and then to models with the MA(2) component (after d differencings). When h=2, we take values d=0,1,2 so that the models compared are ARIMA(1,d,0), ARIMA(0,d,1), and ARIMA(0,d,2). In each cell, the left hand entry is power for $\alpha = 0.05$, and the right hand entry is power for $\alpha = 0.10$ (the left or right tail is taken, as appropriate in each case). The three test statistics $T_{\hat{V}}$, $T_{\hat{V}_c}$, and $T_{\hat{V}_{DM}}$ are compared. Only for n=200 do powers exceeding 0.500 occur. In these cases, $T_{\hat{V}}$ has the most power, followed by $T_{\hat{V}_c}$. The rankings of the statistics according to power in each comparison conform very well to what would be predicted from Table 3 following the discussion of Section 3.2.

Table 5
The analog of Table 4 for ARIMA(0,d,1) DGPs with MA(1) coefficient equal to 0.8.

DGP: 0.8		AR(1) vs. MA(1)						AR(1) vs. MA(2)					
Lead	Sample	$T_{\hat{V}}$		$T_{\hat{V}_c}$		$T_{\hat{V}_{DM}}$		$T_{\hat{V}}$		$T_{\hat{V}_c}$		$T_{\hat{V}_{DM}}$	
h=1	50	0.352	0.515	0.293	0.444	0.207	0.375	0.429	0.583	0.340	0.510	0.252	0.454
	100	0.823	0.900	0.763	0.865	0.689	0.829	0.859	0.929	0.795	0.891	0.739	0.867
	200	0.987	0.995	0.980	0.992	0.974	0.990	0.990	0.997	0.983	0.991	0.977	0.990
h=2	50	0.302	0.457	0.158	0.306	0.163	0.272	0.288	0.444	0.116	0.255	0.123	0.236
d=0	100	0.476	0.594	0.236	0.375	0.226	0.366	0.476	0.608	0.217	0.398	0.206	0.367
	200	0.685	0.787	0.383	0.573	0.378	0.562	0.713	0.787	0.404	0.595	0.381	0.581
h=2	50	0.110	0.236	0.112	0.207	0.120	0.191	0.200	0.365	0.138	0.276	0.142	0.262
d=1	100	0.364	0.538	0.317	0.471	0.289	0.453	0.491	0.655	0.383	0.554	0.360	0.531
	200	0.725	0.836	0.626	0.775	0.621	0.760	0.779	0.870	0.668	0.807	0.655	0.790
h=2	50	0.194	0.333	0.163	0.285	0.163	0.283	0.296	0.463	0.218	0.370	0.218	0.360
d=2	100	0.565	0.721	0.503	0.666	0.487	0.648	0.664	0.807	0.578	0.738	0.567	0.720
	200	0.906	0.952	0.874	0.928	0.852	0.925	0.927	0.970	0.896	0.951	0.886	0.947

The statistic $T_{\hat{V}}$ has the highest power and $T_{\hat{V}_c}$ the second highest in the many instances of powers exceeding 0.500, which occur more often than not with n=100 and every time but one with n=200. The rankings of the statistics according to power in each comparison conform very well to what would be predicted from Table 3 following the discussion of Section 3.2.

predictive power of the denominators, note the cells for h=2 and d=0,1,2 in Table 6 where the rare cases occur in which $T_{\hat{V}_c}$ has substantially greater power than $T_{\hat{V}}$. The corresponding cells of Table 3 show the largest values greater than 1.000 of $\sqrt{V}/\sqrt{V_c}$ in our study.

It also seems interesting that when $\sqrt{V} = \sqrt{V_c}$, then $T_{\hat{V}}$ usually has (slightly) greater power than $T_{\hat{V}_c}$ and when $\sqrt{V_c} = \sqrt{V_{DM}}$, then $T_{\hat{V}_c}$ usually has greater power than $T_{\hat{V}_{DM}}$.

Table 6
The analog of Table 4 for ARIMA(0,*d*,2) DGPs with coefficients 0.25 and 0.5.

DGP: MA(2)		AR(1) vs. MA(1)						AR(1) vs. MA(2)					
Lead	Sample	$T_{\hat{Y}}$		$T_{\hat{Y}_c}$		$T_{\hat{Y}_{DM}}$		$T_{\hat{Y}}$		$T_{\hat{Y}_c}$		$T_{\hat{Y}_{DM}}$	
$h=1$	50	0.018	0.136	0.018	0.128	0.052	0.206	0.267	0.527	0.211	0.466	0.202	0.420
	100	0.115	0.363	0.114	0.354	0.266	0.526	0.749	0.906	0.713	0.877	0.664	0.848
	200	0.351	0.622	0.341	0.613	0.626	0.852	0.991	0.999	0.988	0.999	0.979	0.997
$h=2$	50	0.004	0.088	0.729	0.868	0.689	0.831	0.209	0.444	0.204	0. 420	0.151	0.372
$d=0$	100	0.087	0.337	0.969	0.985	0.947	0.982	0.677	0.859	0.671	0.870	0.581	0.826
	200	0.337	0.630	10.00	10.00	10.00	10.00	0.979	0.994	0.990	0.999	0.979	0.999
$h=2$	50	0.004	0.114	0.230	0.598	0.230	0.553	0.178	0.372	0.102	0.264	0.061	0.213
$d=1$	100	0.126	0.452	0.810	0.942	0.788	0.913	0.601	0.784	0.406	0.672	0.343	0.591
	200	0.551	0.854	0.996	0.998	0.995	0.998	0.969	0.992	0.878	0.970	0.830	0.958
$h=2$	50	0.007	0.122	0.021	0.319	0.021	0.290	0.212	0.424	0.163	0.347	0.134	0.297
$d=2$	100	0.134	0.459	0.478	0.866	0.499	0.837	0.680	0.848	0.564	0.805	0.497	0.761
	200	0.546	0.850	0.991	0.998	0.995	0.998	0.985	0.996	0.965	0.996	0.949	0.993

Some of the cells for $h=2$ in Table 6 contain our study's rare instances in which $T_{\hat{Y}_c}$ has substantially greater power than $T_{\hat{Y}}$. The rankings of the statistics according to power in each comparison largely conform to what would be predicted from Table 3 following the discussion of Section 3.2, where it is also explained why the appreciable power advantage of $T_{\hat{Y}_{DM}}$ over $T_{\hat{Y}_c}$ and $T_{\hat{Y}}$ seen for $h=1$ with AR(1) vs. MA(1) can be discounted.

To summarize the tabled results, the sample size 50 seems insufficient to discriminate between models, but at size 100 many of the tests have greater than 50% power. The results are quite supportive of $T_{\hat{Y}}$ and also supportive of $T_{\hat{Y}_c}$ over $T_{\hat{Y}_{DM}}$.

For this power study, the variances V and V_c were calculated by expressing their integral representations (10) and (14) in terms of autocovariances of various ARMA models, in analogy with the procedure described in Section 2.2. The calculation of V_{DM} similarly used an integral re-expression of the r.h.s. of (20) based on certain algebraic simplifications. Details are omitted.

3.3. Results for published time series

To simplify our presentation of empirical results for nonstationary models, we define the ARMA component of an ARIMA(p,d,q) model for Y_t to be its ARMA(p,q) model for $W_t=(1-B)^d Y_t$.

We now consider four data examples, three of which are taken from Box et al. (1994): Chemical Process Concentration Readings (Series A); IBM Daily Common Stock Closing Prices, May 17, 1961 to November 2, 1962 (Series B); and Chemical Process Temperature Readings (Series C). We chose these series for their availability (at http://www.stat.wisc.edu/~reinsel/bjr-data/index.html) and the simplicity (lack of seasonality) of their recommended MA models. Our fourth data example is from Brockwell and Davis (2002): the Dow Jones Utilities Index, August 28 to December 18, 1972 (Series D).

For each series, we considered six candidate ARMA(p, q) component models for each order d of differencing used with the above authors' recommended models for their series. This resulted in 12 models for Series B and six for the rest. The six options for (p, q) were (0,0), (0,1), (1, 0), (1,1) (0,2) and (2,0), a suite that includes all of the recommended component models. For each series, the test statistics $T_{\hat{Y}}$, $T_{\hat{Y}_c}$, and $T_{\hat{Y}_{DM}}$ were calculated for leads $h=1,2,3$ for all 15 pairs of ARIMA models determined by a choice of d and a pair of ARMA component models. The results are reported in Table 7. The component model pairs are shown in the first column. In each row, the first model listed is the ARMA(p,q) component of Model 1 and the second is that of Model 2 in the formulas of the test statistics. The subsequent columns present the values of the test statistics for each series and recommended differencing order d. Thus, a significant positive value (at least 1.64, for this discussion) indicates that the second model has significantly better forecast performance than the first, whereas a significant negative value (at most -1.64) indicates the reverse.

We focus on the question of whether the models recommended in the cited textbooks outperform, or are outperformed, in a statistically significant way at some lead h by one of the other models considered. The values of statistics involving a recommended model are shown in boldface in the table. We use $T_{\hat{Y}}$ as the final arbiter of statistical significance when there is disagreement among $T_{\hat{Y}}$, $T_{\hat{Y}_c}$, and $T_{\hat{Y}_{DM}}$. This leads to the conclusions that no recommended model is significantly outperformed and that each series except series B has a recommended model that significantly outperforms some of the other models. Here are some details.

For Series A (197 observations), the recommended nonstationary model is an ARIMA(0,1,1). Restricting attention in the second column of Table 7 to those tests involving the ARMA(0,1) component (values in bold), one sees that all three test

Table 7
Test statistics obtained at various forecast leads (h=1,2,3) for each series, comparing the ARIMA models having the two ARMA components indicated on the left-hand column.

Pairs	Series A, d=1			Series C, d=1			Series C, d=2			Series D, d=1		
	h=1	h=2	h=3	h=1	h=2	h=3	h=1	h=2	h=3	h=1	h=2	h=3
(2,0)	−1.33	−1.79	−2.80	**0.17**	**0.18**	**0.15**	−0.61	−0.84	−0.94	−0.57	−0.64	−0.73
(1,0)	−1.29	−2.42	−3.45	**0.24**	**0.09**	**0.15**	−0.60	−1.40	−1.95	−0.58	−0.55	−0.74
	−1.26	−2.39	−3.74	**0.25**	**0.09**	**0.15**	−0.53	−1.28	−0.94	−0.71	−0.69	−0.87
(2,0)	−2.70	−2.86	−4.20	−3.51	−2.91	−2.53	−0.98	−1.17	−1.36	−1.90	−1.40	−1.18
(0,0)	−3.10	−3.51	−5.49	−3.17	−2.77	−2.44	−0.96	−1.51	−2.15	−1.50	−1.26	−1.16
	−3.26	−3.54	−5.16	−7.38	−4.34	−3.29	−0.78	−1.37	−2.31	−2.35	−1.65	−1.38
(2,0)	−4.61	−1.93	−1.17	−5.40	−4.64	−3.98	−0.65	−1.49	−1.47	−1.79	−1.13	−1.04
(1,1)	−2.99	−2.97	−3.14	−3.59	−2.58	−2.21	−3.02	−2.33	−3.23	−1.76	−1.19	−1.11
	−2.77	−3.03	−2.96	−5.96	−4.06	−2.93	−1.90	−1.89	−3.66	−2.44	−1.57	−1.32
(2,0)	**2.58**	**2.44**	**2.42**	−2.77	−2.47	−2.21	−0.48	−0.67	−0.73	−1.08	−0.99	−0.95
(0,1)	**2.58**	**2.89**	**3.13**	−2.78	−2.38	−2.15	−0.47	−1.33	−1.86	−1.07	−0.94	−0.99
	1.72	**2.59**	**3.67**	−5.04	−3.67	−2.87	−0.45	−1.24	−2.01	−1.47	−1.24	−1.18
(2,0)	2.55	2.38	2.34	−2.17	−2.03	−1.92	**0.85**	**0.98**	**1.03**	**−0.69**	**−0.81**	**−0.82**
(0,2)	2.65	2.58	2.92	−2.20	−1.96	−1.82	**0.77**	**1.16**	**1.58**	**−0.69**	**−0.81**	**−0.87**
	1.99	2.34	3.29	−3.14	−2.90	−2.42	**0.66**	**1.09**	**1.68**	**−0.79**	**−1.05**	**−1.04**
(1,0)	−2.42	−3.23	−3.40	**−3.51**	**−2.91**	**−2.53**	−0.60	−0.86	−1.05	−1.86	−1.45	−1.20
(0,0)	−2.80	−4.15	−5.06	**−3.17**	**−2.76**	**−2.44**	−0.59	−1.21	−2.05	−1.45	−1.39	−1.28
	−3.02	−4.16	−4.62	**−7.38**	**−4.33**	**−3.29**	−0.58	−1.18	−2.14	−2.24	−1.80	−1.52
(1,0)	−6.52	−1.09	−0.66	**−5.44**	**−4.65**	**−4.00**	−0.59	−0.90	−0.81	−1.89	−1.14	−1.04
(1,1)	−3.74	−3.64	−2.20	**−3.58**	**−2.58**	**−2.20**	−3.44	−3.35	−3.85	−1.98	−1.33	−1.24
	−3.03	−3.85	−1.88	**−5.95**	**−4.05**	**−2.92**	−2.04	−2.04	−5.59	−2.59	−1.73	−1.47

(1,0)	**2.57**	**2.84**	**2.69**	**−2.77**	**−2.47**	**−2.22**	0.58	0.69	0.76	−1.16	−1.02	−0.95
(0,1)	**2.56**	**3.42**	**3.30**	**−2.78**	**−2.37**	**−2.14**	0.55	1.34	2.05	−1.13	−1.07	−1.10
	1.86	**2.87**	**3.87**	**−5.03**	**−3.66**	**−2.86**	0.46	1.22	2.22	−1.59	−1.39	−1.31
(1,0)	2.72	2.84	2.58	**−2.18**	**−2.03**	**−1.92**	**0.78**	**0.99**	**1.07**	**−0.30**	**−0.80**	**−0.72**
(0,2)	2.79	3.16	3.10	**−2.19**	**−1.96**	**−1.82**	**0.73**	**1.29**	**1.75**	**−0.29**	**−1.04**	**−0.97**
	2.11	2.70	3.53	**−3.13**	**−2.88**	**−2.40**	**0.63**	**1.20**	**1.89**	**−0.42**	**−1.39**	**−1.18**
(0,0)	0.68	0.79	1.27	1.88	1.42	1.29	−0.58	−0.68	−0.55	0.18	1.69	2.35
(1,1)	0.68	4.54	2.57	2.13	2.92	2.88	−2.82	−1.31	−1.69	0.08	1.65	1.37
	0.63	4.27	2.21	4.58	4.81	3.79	−1.68	−0.89	−2.03	0.12	2.74	1.43
(0,0)	**2.99**	**3.52**	**3.68**	3.85	3.56	3.29	0.61	0.82	0.96	2.10	2.09	1.77
(0,1)	**3.41**	**4.34**	**4.67**	3.29	3.36	3.17	0.60	1.27	2.09	1.51	1.85	1.61
	3.26	**3.76**	**4.97**	7.40	5.38	4.28	0.57	1.21	2.20	2.31	2.35	1.89
(0,0)	3.48	3.70	3.73	3.72	3.25	3.02	**0.99**	**1.15**	**1.27**	**2.11**	**1.66**	**1.51**
(0,2)	4.07	4.19	4.42	3.29	3.09	2.99	**0.94**	**1.41**	**1.93**	**1.60**	**1.44**	**1.42**
	3.74	3.62	4.70	7.57	4.80	4.06	**0.77**	**1.30**	**2.08**	**2.37**	**1.81**	**1.67**
(1,1)	**4.78**	**2.49**	**2.00**	1.30	−0.02	−0.11	0.60	0.92	0.85	2.55	1.18	1.23
(0,1)	**3.60**	**3.85**	**3.57**	1.90	−0.14	−0.96	3.44	3.05	3.90	2.25	1.89	1.79
	3.15	**3.30**	**4.24**	3.75	−0.66	−0.96	2.08	2.11	5.16	2.72	2.46	1.95
(1,1)	4.34	2.38	1.93	3.74	2.22	2.09	**0.69**	**1.43**	**1.49**	**1.99**	**1.23**	**1.24**
(0,2)	3.48	3.65	3.35	3.17	3.23	3.08	**2.80**	**1.87**	**2.55**	**1.85**	**1.37**	**1.41**
	3.11	3.14	3.88	4.76	5.41	3.97	**1.88**	**1.66**	**2.79**	**2.44**	**1.73**	**1.63**
(0,1)	**0.93**	**0.99**	**0.98**	2.89	2.78	2.67	**0.74**	**0.94**	**1.02**	**1.10**	**1.06**	**1.13**
(0,2)	**1.01**	**1.29**	**1.73**	2.83	2.73	2.77	**0.70**	**1.26**	**1.70**	**1.10**	**0.99**	**1.21**
	1.11	**1.10**	**1.52**	4.09	4.15	3.75	**0.61**	**1.18**	**1.83**	**1.47**	**1.25**	**1.41**

In the subsequent column headings, the differencing order d of the ARIMA models is shown below the name of the series being forecasted. In each cell, the first, second, and third rows correspond to the three variance normalizations: $\hat{V}$, $\hat{V}_c$, and $\hat{V}_{DM}$. Values in bold highlight the comparisons with the recommended model(s).

statistics find that the ARIMA(0,1,1) model forecasts significantly better at all leads than all competitors except the ARIMA(0,1,2), whose measured forecast performance is uniformly better but never in a statistically significant way.

Series B (369 observations) has an ARIMA(0,1,0) and an ARIMA(0,1,1) as recommended models. None of the tests statistic we considered was significant: all had values between -1 and 1. These values are omitted from Table 7. The model choices of Box et al. (1994) are neither contradicted nor confirmed by our diagnostics.

For Series C (226 observations), the ARIMA(1,1,0) and ARIMA(0,2,2) models are recommended. First considering models with $d=1$, for the ARMA(1,0) component, in column 4, it is seen that the ARIMA(1,1,0) model forecasts significantly better than the ARIMA(0,1,0), ARIMA(1,1,1), ARIMA(0,1,1), and ARIMA(0,1,2) models—although the significance in this last comparison weakens at $h=3$. The tests do not detect significant forecasting differences between the ARIMA(1,1,0) and the ARIMA(2,1,0) models. For the results for the ARMA(0,2) component of the recommended model with $d=2$ in column 5, no significant differences appear between the ARIMA(0,2,2) model and its competitors. Note that use of $T_{\hat{V}_c}$ and $T_{\hat{V}_{DM}}$ would lead to incorrect conclusions of significantly better performance for the recommended ARIMA(0,2,2) over the ARIMA(1,2,1) and ARIMA(0,2,0) models at one or more leads.

Finally, for Series D (78 observations), the ARIMA(0,1,2) model is recommended. From the result of the last column of Table 7 for the ARMA(0,2) component model, the ARIMA(0,1,2) model has generally better forecast performance, but in a statistically significant way only for $h=1$ against the ARIMA(0,1,0) and ARIMA(1,1,1) models (where $T_{\hat{V}_c}$ fails to indicate significance).

Examination of the entries of Table 7 that do not involve a recommended model show many comparisons where $T_{\hat{V}_c}$ or $T_{\hat{V}_{DM}}$ indicates a significant forecast advantage whose significance is contradicted by $T_{\hat{V}}$, and one, for series D with $h=3$, where $T_{\hat{V}}$ indicates that the (1, 1, 1) model forecasts significantly better than the (0,1,0) model, but $T_{\hat{V}_c}$ and $T_{\hat{V}_{DM}}$ do not show this.

To summarize, the test with $T_{\hat{V}}$ of this paper can be used to support the recommended models for these series, often for multiple forecast leads h. (With series B, no rival model is preferred, but neither is a recommended model preferred over any rival.) The statistics that ignore coefficient estimation effects, $T_{\hat{V}_c}$ and $T_{\hat{V}_{DM}}$, lead frequently to spurious indications of statistically significant performance improvement and fail to identify some instances of significant improvement that are revealed by $T_{\hat{V}}$.

4. Concluding summary

We have introduced a new pair of test statistics for testing the null hypothesis that two competing incorrect, invertible ARMA or ARIMA-type models for a series have the same asymptotic mean square h-step forecast error. In the nonstationary case, the models are assumed to have the correct differencing operator. The numerators of both statistics are the same, being the difference of the mean square forecast error measures of the models. But they are conceptually different in that the models' parameters are treated as fixed in the simpler statistic $T_{\hat{V}_c}$ and as sample-size-dependent (Quasi-)Maximum Likelihood estimates in the statistic $T_{\hat{V}}$, which the more complex standard deviation estimate in its denominator to account for parameter estimation. The simpler statistic improves the well-known Diebold–Mariano statistic, equivalent to $T_{\hat{V}_{DM}}$ in our notation, by providing a denominator that leads to a standard normal limiting distribution when the null hypothesis of equal asymptotic mean square forecast error is satisfied, as happens with the series of our size study. No series is known to exist for which $T_{\hat{V}_{DM}}$ has this property (see the Remark of Section 3). Regarding our more comprehensive statistic, its superior finite-sample and asymptotic results in our power study clearly reveal that accepting the parameters in the numerator of these statistics as estimates, and accounting for this fact appropriately in the denominator, often yields a statistic with smaller variability both asymptotically and in samples of moderate size. R code is available from the first author (McElroy) for calculating the statistics for the general ARIMA case, including exact calculation of the first and second derivatives of functions of the model spectral densities required to account for parameter estimation as shown in formulas in the Appendix.

For the size study, we provided an example of an MA(2) series and simple pair of nonnested incorrect AR models for which the null hypothesis of equal asymptotic mean squared one-step-ahead forecast error holds in a nondegenerate way. In the study, the worst results were obtained for the Diebold–Mariano statistic, which was often quite oversized, whereas the new statistics showed a tendency to be moderately undersized.

Our empirical study of models for four published time series provided further evidence of the superior performance of $T_{\hat{V}}$ over $T_{\hat{V}_c}$ and $T_{\hat{V}_{DM}}$. Our test with $T_{\hat{V}}$ of models for four published time series formally justified the use of one or more of the models recommended by experts for these series in the application to forecasting at lead one and often at higher leads.

Acknowledgments

The communicating author (Findley) was stimulated to change fields from functional analysis to statistical time series analysis by a one-day time series workshop at the University of Cincinnati in 1973 presented by Manny Parzen and his distinguished former student Grace Wahba. He is most grateful for this influential workshop and is further grateful to Manny for subsequent acts of support and collaboration. The authors are indebted to two anonymous referees and to Michael McCracken for comments and questions that led to significant improvements. They also thank Brian Monsell for his careful reading of the manuscript.

Disclaimer. This paper is released to inform about ongoing research and to encourage discussion. The views expressed are the authors' and not necessarily those of the U.S. Census Bureau.

Appendix A

In this section we consider the asymptotic properties of statistics of the form

$$Q_n(f,g,\theta) = \frac{1}{n}\sum_{\lambda} g_\theta(\lambda) f(\lambda),$$

where g_θ is some weighting function dependent on a parameter vector θ, and the sum is over the Fourier frequencies in $(-\pi,\pi)\backslash\{0\}$. These functions g and f may or may not be random, depending on the context given below. To match the generality of McElroy and Holan (2009), in our general result, Theorem 2 below, f can be some integer power of the periodogram. In our forecasting application, where we take the first power, our results also apply (see Chen and Deo, 2000) to the approximation to $Q_n(f,g,\theta)$ defined by (9)—in other words, the integral and the sum over Fourier frequencies are asymptotically equivalent when these formulas are linear in the periodogram. Consider the situation of evaluating L models, each with its own parameter vector $\theta^{(i)}$, $i=1,2,\ldots,L$. The corresponding model spectral densities will be denoted $f_{\theta^{(i)}}$—which is an abuse of notation, since they depend on i directly in their functional form and not solely through the parameter $\theta^{(i)}$. The parameter vectors can be stacked together into one super-vector $\theta=(\theta^{(1)},\ldots,\theta^{(l)})$ with values in the Cartesian product of the L compact parameter spaces $\Theta^{(i)}$.

We assume that each model satisfies the assumptions of Section 2.1. To simplify the notation, we write g_i instead of $g_{\theta^{(i)},i}$, $\tilde{g}_i$ instead of $g_{\tilde{\theta}^{(i)},i}$, and $\hat{g}_i$ instead of $g_{\hat{\theta}^{(i)},i}$, where $\hat{\theta}^{(i)}$ denotes a parameter estimate. The result below is similar to McElroy and Holan (2009, Theorem 2), but because the statistic $Q_n(I^{j_i},\hat{g}_i,\hat{\theta}^{(i)})$ is compared to the true process quantity $j_i!Q_n(\tilde{f}^{j_i},\tilde{g}_i,\tilde{\theta}^{(i)})$ rather than to a model-based estimate (as in McElroy and Holan, 2009, Theorem 2), nontrivial modifications to the previous results are needed to establish it. The null hypotheses pertinent to each theorem are quite different; in the case of this paper two misspecified models are compared, and thus it is impossible to speak of an estimate of a correctly specified model. But in the case of McElroy and Holan (2009),the null hypothesis is that the given model is actually correct, so it makes sense to compare the data to our estimate of that model. Hence the difference in theory, which results in nontrivial differences in the proofs; this is also the reason why the resulting variances are dissimilar—the formulas for the b vector are substantially different. Here we also consider the case of L models.

Theorem 2. *Under conditions* 1–6 *with* $\hat{\theta}^{(i)}$ *the QMLEs* (*if they are MLEs, also assume condition* 7), *we have*

$$\{\sqrt{n}(Q_n(I^{j_i},\hat{g}_i,\hat{\theta}^{(i)}) - j_i!Q_n(\tilde{f}^{j_i},\tilde{g}_i,\tilde{\theta}^{(i)}))\}_{i=1}^{L} \stackrel{\mathcal{L}}{\Longrightarrow} \mathcal{N}(0,W(\tilde{\theta}))$$

as $n\to\infty$, *with* $W(\theta)$ *an* $L\times L$ *variance matrix with kl entry*

$$\begin{aligned} W_{kl}(\theta) = {} & \frac{(j_k+j_l)!-j_k!j_l!}{4\pi}\int_{-\pi}^{\pi}(g_k(\lambda)g_l(-\lambda)+g_l(\lambda)g_k(-\lambda)+2g_k(\lambda)g_l(\lambda))\tilde{f}^{j_k+j_l}(\lambda)\,d\lambda \\ & + \frac{(j_k+1)!-j_k!}{4\pi}\int_{-\pi}^{\pi}(g_k(\lambda)p_l(-\lambda)+p_l(\lambda)g_k(-\lambda)+2g_k(\lambda)p_l(\lambda))\tilde{f}^{j_k+1}(\lambda)\,d\lambda \\ & + \frac{(j_l+1)!-j_l!}{4\pi}\int_{-\pi}^{\pi}(g_l(\lambda)p_k(-\lambda)+p_k(\lambda)g_l(-\lambda)+2p_k(\lambda)g_l(\lambda))\tilde{f}^{j_l+1}(\lambda)\,d\lambda \\ & + \frac{1}{4\pi}\int_{-\pi}^{\pi}(p_k(\lambda)p_l(-\lambda)+p_l(\lambda)p_k(-\lambda)+2p_k(\lambda)p_l(\lambda))\tilde{f}^{2}(\lambda)\,d\lambda. \end{aligned}$$

These entries are defined in terms of the following quantities:

$$p_{\theta^{(i)},i}(\lambda) = f_{\theta^{(i)}}^{-2}(\lambda)b'_{\theta^{(i)},i}M_f^{-1}(\theta^{(i)})\nabla_{\theta^{(i)}}f_{\theta^{(i)}}(\lambda),$$

$$b_{\theta^{(i)},i} = \frac{j_i!}{2\pi}\int_{-\pi}^{\pi}\tilde{f}^{j_i}(\lambda)\nabla_{\theta^{(i)}}g_{\theta^{(i)},i}(\lambda)\,d\lambda,$$

$$M_f(\theta^{(i)}) = \nabla_{\theta^{(i)}}\nabla'_{\theta^{(i)}}D(f_{\theta^{(i)}},\tilde{f}).$$

Theorem 2 is stated very generally, where higher powers of the periodogram are allowed. Although this is not pursued further in this article, squares and higher powers can be used to facilitate more powerful tests (as discussed and demonstrated in an analogous situation in McElroy and Holan, 2009), and are related to the idea of using squared residual autocorrelations to test goodness-of-fit.

In this paper we are specifically interested in the case of two fitted models for the data, whose forecast performance we wish to compare. So we consider the case in which, for each $i=1$, 2, $g_i = g_{\theta^{(i)},i}$ corresponds to the weighting function g defined by (8) and (3), where the dependency on $\theta^{(i)}$ enters in through the innovations filter function $\Psi_{\theta^{(i)}}$, which is

substituted for Ψ in (3). (The forecast lead h is the same for both models—otherwise we would not be evaluating them on the same footing.) The model spectral densities are assumed to have the form $f_{\theta^{(i)}}(\lambda)=\sigma^2_{(i)}|\Psi_{\theta^{(i)}}(e^{-i\lambda})|^2$ with $\sigma_{(i)}$ not functionally related to how $\theta^{(i)}$ determines $\Psi_{\theta^{(i)}}(e^{-i\lambda})$, $i=1, 2$. Then application of Theorem 2 with $j_1=1=j_2$ shows that

$$\{\sqrt{n}(Q_n(I,\hat{g}_i,\hat{\theta}^{(i)})-Q_n(\tilde{f},\tilde{g}_i,\tilde{\theta}^{(i)}))\}_{i=1,2}\stackrel{\mathcal{L}}{\Longrightarrow}\mathcal{N}(0,W(\tilde{\theta})). \tag{A.1}$$

The entries of the asymptotic variance matrix are as follows:

$$W_{kl}(\theta)=\frac{1}{\pi}\int_{-\pi}^{\pi}\tilde{f}^2(\lambda)(g_{\theta^{(k)},k}(\lambda)+p_{\theta^{(k)},k}(\lambda))(g_{\theta^{(l)},l}(\lambda)+p_{\theta^{(l)},l}(\lambda))\,d\lambda.$$

Since $Q(I,\hat{g}_i,\hat{\theta}^{(i)})$ assesses the forecast error of each model, we construct our statistic from the difference of these quantities.

To establish Theorem 1 we then apply the vector $(1,-1)$ to the joint convergence (A.1), obtaining the asymptotic variance formula $W_{11}(\tilde{\theta})-2W_{12}(\tilde{\theta})+W_{22}(\tilde{\theta})$. This is easily shown to equal the limiting variance V of Theorem 1. The consistency of $\hat{V}$—independent of whether the null or alternative hypothesis is true—then follows from conditions 2, 3, 4, 5, and 6 (as well as condition 7 if we are considering MLEs instead of QMLs), together with Taniguchi and Kakizawa (2000, Lemma 3.1.1). This concludes the derivation.

A.1. Proof of Theorem 2

For each i we have

$$Q_n(I^{j_i},\hat{g}_i,\hat{\theta}^{(i)})-j_i!Q_n(\tilde{f}^{j_i},\tilde{g}_i,\tilde{\theta}^{(i)})=(Q_n(I^{j_i},\hat{g}_i,\hat{\theta}^{(i)})-j_i!Q_n(\tilde{f}^{j_i},\hat{g}_i,\hat{\theta}^{(i)}))+j_i!(Q_n(\tilde{f}^{j_i},\hat{g}_i,\hat{\theta}^{(i)})-Q_n(\tilde{f}^{j_i},\tilde{g}_i,\tilde{\theta}^{(i)})).$$

The first term expands to

$$\frac{1}{n}\sum_{\lambda}(I^{j_i}(\lambda)-j_i!\tilde{f}^{j_i}(\lambda))g_{\hat{\theta}^{(i)},i}(\lambda)=\frac{1}{n}\sum_{\lambda}(I^{j_i}(\lambda)-j_i!\tilde{f}^{j_i}(\lambda))(g_{\tilde{\theta}^{(i)},i}(\lambda)+O_P(n^{-1/2})).$$

Since $\sum_{\lambda}(I^{j_i}(\lambda)-j_i!\tilde{f}^{j_i}(\lambda))=O_P(n^{1/2})$ by Lemma 3.1.1 of Taniguchi and Kakizawa (2000), we have

$$\sqrt{n}(Q_n(I^{j_i},\hat{g}_i,\hat{\theta}^{(i)})-j_i!Q_n(\tilde{f}^{j_i},\hat{g}_i,\hat{\theta}^{(i)}))=o_P(1)+\frac{1}{\sqrt{n}}\sum_{\lambda}(I^{j_i}(\lambda)-j_i!\tilde{f}^{j_i}(\lambda))g_{\tilde{\theta}^{(i)},i}(\lambda).$$

For the second term we have

$$\frac{j_i!}{n}\sum_{\lambda}\tilde{f}^{j_i}(\lambda)(g_{\hat{\theta}^{(i)},i}(\lambda)-g_{\tilde{\theta}^{(i)},i}(\lambda))=\frac{j_i!}{n}\sum_{\lambda}\tilde{f}^{j_i}(\lambda)(\nabla'_{\theta^{(i)}}g_{\tilde{\theta}^{(i)},i}(\lambda)(\hat{\theta}^{(i)}-\tilde{\theta}^{(i)})+O_P(n^{-1})).$$

Now by Theorem 3.1.2 of Taniguchi and Kakizawa (2000),

$$\sqrt{n}(\hat{\theta}^{(i)}-\tilde{\theta}^{(i)})=o_P(1)+M_{\tilde{f}}^{-1}(\tilde{\theta}^{(i)})\frac{1}{\sqrt{n}}\sum_{\lambda}\nabla_{\theta^{(i)}}f_{\tilde{\theta}^{(i)},i}(\lambda)(I(\lambda)-\tilde{f}(\lambda))f^{-2}_{\tilde{\theta}^{(i)},i}(\lambda),$$

and hence

$$\sqrt{n}j_i!(Q_n(\tilde{f}^{j_i},\hat{g}_i,\hat{\theta}^{(i)})-Q_n(\tilde{f}^{j_i},\tilde{g}_i,\tilde{\theta}^{(i)}))=o_P(1)+\frac{j_i!}{n}\sum_{\lambda}\tilde{f}^{j_i}(\lambda)\nabla'_{\theta^{(i)}}g_{\tilde{\theta}^{(i)},i}(\lambda)M_{\tilde{f}}^{-1}(\tilde{\theta}^{(i)})\frac{1}{\sqrt{n}}\sum_{\omega}\nabla_{\theta^{(i)}}f_{\tilde{\theta}^{(i)},i}(\omega)(I(\omega)-\tilde{f}(\omega))f^{-2}_{\tilde{\theta}^{(i)},i}(\omega).$$

In order to prove joint convergence we use the Cramer–Wold device, so consider the dot product against the vector $\alpha=(\alpha_1,\ldots,\alpha_L)'$, which yields (up to terms tending to zero in probability)

$$\frac{1}{\sqrt{n}}\sum_{\lambda}\sum_{i=1}^{L}\alpha_i g_{\tilde{\theta}^{(i)},i}(\lambda)(I^{j_i}(\lambda)-j_i!\tilde{f}^{j_i}(\lambda))+\frac{1}{\sqrt{n}}\sum_{\lambda}\sum_{i=1}^{L}\alpha_i b'_{\tilde{\theta}^{(i)},i}M_{\tilde{f}}^{-1}(\tilde{\theta}^{(i)})\nabla_{\theta^{(i)}}f_{\tilde{\theta}^{(i)},i}(\lambda)(I(\lambda)-\tilde{f}(\lambda))f^{-2}_{\tilde{\theta}^{(i)},i}(\lambda).$$

Consider weighting functions $\phi_i(\lambda)=\alpha_i g_{\tilde{\theta}^{(i)},i}(\lambda)$ for $1\le i\le L$ and

$$\phi_0(\lambda)=\sum_{i=1}^{L}\alpha_i b'_{\tilde{\theta}^{(i)},i}M_{\tilde{f}}^{-1}(\tilde{\theta}^{(i)})\nabla_{\theta^{(i)}}f_{\tilde{\theta}^{(i)},i}(\lambda)f^{-2}_{\tilde{\theta}^{(i)},i}(\lambda).$$

Let $j_0=1$ and apply Theorem 1 of McElroy and Holan (2009):

$$\left\{\frac{1}{\sqrt{n}}\phi_i(\lambda)(I^{j_i}(\lambda)-j_i!\tilde{f}^{j_i}(\lambda))\right\}_{i=0}^{L}\stackrel{\mathcal{L}}{\Longrightarrow}\mathcal{N}(0,V(\alpha)),$$

with the variance matrix given by $V_{kl}(\alpha)$ equal to

$$\frac{(j_k+j_l)!-j_k!j_l!}{4\pi}\int_{-\pi}^{\pi}(\phi_k(\lambda)\phi_l(-\lambda)+\phi_k(-\lambda)\phi_l(\lambda)+2\phi_k(\lambda)\phi_l(\lambda))\tilde{f}^{j_k+j_l}(\lambda)\,d\lambda$$

for $0 \le k,l \le L$. Then our statistic of interest, summed against α, is asymptotically normal with variance $\sum_{k,l=0}^{L} V_{kl}(\alpha)$. This establishes joint asymptotic normality of $\sqrt{n}(Q_n(I^{j_i},\hat{g}_i,\hat{\theta}^{(i)}) - j_i! Q_n(\tilde{f}^{j_i},\tilde{g}_i,\tilde{\theta}^{(i)}))$ with variance matrix W, which has entries given as follows. If $1 \le i \ne j \le L$, then $W_{ij} = \frac{1}{2}\sum_{k,l=0}^{L}(V_{kl}(e_i+e_j) - V_{kl}(e_i) - V_{kl}(e_j))$ with e_i the ith unit vector. Also $W_{ii} = \sum_{k,l=0}^{L} V_{kl}(e_i)$, but this follows from the previous formula letting $i=j$. Simplifying these expressions (for details, see the proof of Theorem 2 in McElroy and Holan, 2009) yields the stated expressions for $W = W(\tilde{\theta})$.

A.2. Implementation details

We now discuss the implementation details for an ARMA(p,q) model. We need to compute estimates for W_{11}, $W_{12}=W_{21}$, and W_{22}, which involve the quantities $p_{\theta^{(i)},i}(\lambda)$ and $b_{\theta^{(i)},i}$ described in Theorem 1. Now in forming estimates, we replace all parameter vectors $\theta^{(i)}$ with their MLEs $\hat{\theta}^{(i)}$, and replace $\tilde{f}^2$ with $I^2/2$ (division by two is necessary for unbiased estimation, as shown in McElroy and Holan, 2009). In a similar fashion, the estimate of the Hessian matrix $M_f(\theta^{(i)})$ is constructed. We proceed first to construct $\hat{b}_{\hat{\theta}^{(i)},i}$, and then the Hessian matrix, and finally the quantities $\hat{W}_{kl}$.

We suppose that both models, after suitable differencings, yield ARMA models. As mentioned in Section 2.1, the differencing operator is assumed to be correctly specified but the stationary models may both be incorrect. So let $\theta^{(i)} = (\theta_1^{(i)},\theta_2^{(i)},\ldots,\theta_{r_i}^{(i)},\theta_{r_i+1}^{(i)})'$, where $\theta_{r_i+1}^{(i)}$ is the innovation variance, and $i=1,2$. Let the first $q_i \ge 0$ parameters be the moving average coefficients of a polynomial $\Omega^{(i)}(B)$ (written in addition convention), whereas the next $p_i \ge 0$ parameters are the autoregressive coefficients of a polynomial $\Xi^{(i)}(B)$ (written in subtraction convention). Then $r_i=p_i+q_i$. The coefficients of these polynomials are written $\omega_j^{(i)} = \theta_j^{(i)}$ for $1 \le j \le q_i$ and $\xi_j^{(i)} = \theta_{j+q_i}^{(i)}$ for $1 \le j \le p_i$. Then clearly the ARMA transfer function of the model is $\Omega^{(i)}(B)/\Xi^{(i)}(B)$, which can be written as an infinite order causal moving average $\Psi^{(i)}(B)$.

Letting $\Phi^{(i)}(B) = [\Psi^{(i)}/\delta]_0^{h-1}(B)$, we see that it is equal to $\sum_{k=0}^{h-1}\tau_k B^k \sum_{l=0}^{h-1-k}\psi_l^{(i)}B^l$ with $\phi_k^{(i)} = \sum_{l=0}^{k}\psi_l^{(i)}\tau_{k-l}$, utilizing $\tau_k = 0$ when $k<0$. From this calculation it is immediate that

$$\Phi^{(i)}(B) = [\Psi^{(i)}/\delta]_0^{h-1}(B) = \left[\frac{\Omega^{(i)}}{\delta\Xi^{(i)}}\right]_0^{h-1}(B).$$

The weighting function $g_{\theta^{(i)}}$ is then given by

$$g_{\theta^{(i)}}(\lambda) = |\Phi^{(i)}(z)|^2 |\Xi^{(i)}(z)|^2 |\Omega^{(i)}(z)|^{-2},$$

using $z = e^{-i\lambda}$ as a convenient abbreviation. Clearly the above formulation of g shows that it can be thought of as the spectral density of an ARMA process, and hence its Fourier transforms (FTs) are easily obtained; then applying (8) we at once obtain the numerator of our test statistic (12). As for the computation of $\hat{b}_{\hat{\theta}^{(i)},i}$ we must determine the gradient of g, along with its FTs. We denote the sum of a complex number ζ with its conjugate $\overline{\zeta}$ via the notation $\zeta^\sharp = \zeta + \overline{\zeta}$. Then we note that $\partial/\partial\omega_j|\Omega(z)|^2 = \{z^{-j}\Omega(z)\}^\sharp$ and $\partial/\partial\xi_j|\Xi(z)|^2 = -\{z^{-j}\Xi(z)\}^\sharp$. Furthermore,

$$\frac{\partial}{\partial\theta_j^{(i)}} g_{\theta^{(i)}}(\lambda) = \left\{\left[\frac{j}{\delta\Xi^{(i)}}\right]_0^{h-1}(z)\Phi^{(i)}(z^{-1})\right\}^\sharp |\Xi^{(i)}(z)|^2|\Omega^{(i)}(z)|^{-2} - \{z^{-j}\Omega^{(i)}(z)\}^\sharp |\Phi^{(i)}(z)|^2|\Xi^{(i)}(z)|^2|\Omega^{(i)}(z)|^{-4}$$

for $1 \le j \le q_i$, and for $1 \le j \le p_i$ we have

$$\frac{\partial}{\partial\theta_{j+q_i}^{(i)}} g_{\theta^{(i)}}(\lambda) = \left\{\left[\frac{j\Omega^{(i)}}{\delta(\Xi^{(i)})^2}\right]_0^{h-1}(z)\Phi^{(i)}(z^{-1})\right\}^\sharp |\Xi^{(i)}(z)|^2|\Omega^{(i)}(z)|^{-2} - \{z^{-j}\Xi^{(i)}(z)\}^\sharp |\Phi^{(i)}(z)|^2|\Omega^{(i)}(z)|^{-2}.$$

Of course the derivative with respect to $\theta_{r_i+1}^{(i)}$ is zero, since g is scale-invariant. Since $b_{\theta^{(i)},i}$ is estimated by integrating I against the gradient of $g_{\theta^{(i)},i}$—and evaluating all at parameter estimates—we can use the formula (8) to compute the components of b, so long as we have the FTs of the gradient of g. The following general formula will be used repeatedly: suppose that a and c are polynomials and s is some even function of λ with FTs given by the sequence $\gamma_k(s)$; then the lag t FT of $z^k a(z^{-1})z^{-l}c(z)s(\lambda)$ is given by

$$\sum_{m=0}^{deg(a)} a_m \sum_{n=0}^{deg(c)} c_n \gamma_{t+m-n-k+l}(s). \tag{A.2}$$

It can be shown that if $e(z)$ is the polynomial obtained by multiplying $c(z)$ and the reversed polynomial of $a(z)$, i.e., $z^{deg(a)}a(z^{-1})$, then the inner product of the coefficient vector $[e_0,e_1,\ldots,e_{deg(a)+deg(c)}]$ with the vector of FTs $[\gamma_{t-k+l+deg(a)}(s),\ldots,\gamma_{t-k+l-deg(c)}]$ will yield (A.2); this is useful in the encoding of these formulas.

Applying (A.2) to the gradient of $g_{\theta^{(i)},i}$ for $1 \le j \le q_i$ with $a(z) = [1/\delta\Xi^{(i)}]_0^{h-1}(z)$, k=l=0, and $c(z) = \Phi(z)$ yields $\sum_{m=0}^{h-1} a_{m-j} \sum_{n=0}^{h-1} \phi_n^{(i)}(\gamma_{t+m-n}(|\Xi^{(i)}|^2|\Omega^{(i)}|^{-2}) + \gamma_{t-m+n}(|\Xi^{(i)}|^2|\Omega^{(i)}|^{-2}))$. The coefficients of $a(z)$ are easily obtained by finding the causal moving average form of $1/\delta(z)\Xi^{(i)}(z)$, and truncating to lag $h-1$. All together, the lag t FT for $\partial g_{\theta^{(i)},i}/\partial\theta_j^{(i)}$ and $1 \le j \le q_i$ is

$$\sum_{m=0}^{h-1} a_{m-j} \sum_{n=0}^{h-1} \phi_n^{(i)}(\gamma_{t+m-n}(|\Xi^{(i)}|^2|\Omega^{(i)}|^{-2}) + \gamma_{t-m+n}(|\Xi^{(i)}|^2|\Omega^{(i)}|^{-2}))$$
$$- \sum_{m=0}^{q_i} \omega_m^{(i)}(\gamma_{t+m-j}(|\Phi^{(i)}(z)|^2|\Xi^{(i)}|^2|\Omega^{(i)}|^{-4}) + \gamma_{t-m+j}(|\Phi^{(i)}(z)|^2|\Xi^{(i)}|^2|\Omega^{(i)}|^{-4})).$$

Now turning to the AR portion, let $1 \le j \le p_i$ and set $a(z) = [\Omega^{(i)}/\delta(\Xi^{(i)})^2]_0^{h-1}(z)$ so that the FT at lag t of the $j+q_i$th derivative of $g_{\theta^{(i)},i}$ is

$$\sum_{m=0}^{h-1} a_{m-j} \sum_{n=0}^{h-1} \phi_n^{(i)}(\gamma_{t+m-n}(|\Xi^{(i)}|^2|\Omega^{(i)}|^{-2}) + \gamma_{t-m+n}(|\Xi^{(i)}|^2|\Omega^{(i)}|^{-2}))$$
$$+ \sum_{m=0}^{p_i} \xi_m^{(i)}(\gamma_{t+m-j}(|\Phi^{(i)}(z)|^2|\Omega^{(i)}|^{-2}) + \gamma_{t-m+j}(|\Phi^{(i)}(z)|^2|\Omega^{(i)}|^{-2})),$$

where by an abuse of notation we here set $\xi_0^{(i)} = -1$ (because of the subtraction convention). This completes the description of the computation of b; this will be a consistent estimate of the true b, since the parameter estimates converge to pseudo-true values and the integrated periodogram converges to $\tilde{f}$.

Next we consider the Hessian of the KB discrepancy, which under H_0 is not the same as the Fisher information matrix, unfortunately:

$$[M_f(\theta^{(i)})]_{jk} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\partial^2 f_{\theta^{(i)}}(\lambda)}{\partial\theta_j^{(i)}\partial\theta_k^{(i)}}\left(1 - \frac{\tilde{f}(\lambda)}{f_{\theta^{(i)}}(\lambda)}\right) f_{\theta^{(i)}}^{-1}(\lambda) + \frac{\partial f_{\theta^{(i)}}(\lambda)}{\partial\theta_j^{(i)}}\frac{\partial f_{\theta^{(i)}}(\lambda)}{\partial\theta_k^{(i)}}\left(2\frac{\tilde{f}(\lambda)}{f_{\theta^{(i)}}(\lambda)} - 1\right) f_{\theta^{(i)}}^{-2}(\lambda)\, d\lambda.$$

Since the model is fixed in these calculations, we can suppress the i superscript for the moment. The gradient of f is given by

$$\frac{\partial}{\partial\theta_j} f_\theta(\lambda) = \begin{cases} \{z^{-j}\Omega(z)\}^\sharp |\Xi(z)|^{-2}\theta_{r+1}, & 1 \le j \le q, \\ \{z^{-j+q}\Xi(z)\}^\sharp |\Xi(z)|^{-4}|\Omega(z)|^2\theta_{r+1}, & q+1 \le j \le r, \\ |\Omega(z)|^2|\Xi(z)|^{-2}. \end{cases}$$

The mixed partial derivatives are given by

$$\frac{\partial^2}{\partial\theta_j\partial\theta_k} f_\theta(\lambda) = \begin{cases} \{z^{k-j}\}^\sharp |\Xi(z)|^{-2}\theta_{r+1}, & 1 \le j,\ k \le q, \\ \{z^{-j}\Omega(z)\}^\sharp \{z^{-k+q}\Xi(z)\}^\sharp |\Xi(z)|^{-4}\theta_{r+1}, & 1 \le j \le q,\ q+1 \le k \le r, \\ \{z^{-j}\Omega(z)\}^\sharp |\Xi(z)|^{-2}, & 1 \le j \le q,\ k = r+1, \\ \{z^{-j+q}\Xi(z)\}^\sharp \{z^{-k}\Omega(z)\}^\sharp |\Xi(z)|^{-4}\theta_{r+1}, & q+1 \le j \le r,\ 1 \le k \le q, \\ 2\{z^{-j-k}\Xi^2(z)\}^\sharp |\Xi(z)|^{-6}|\Omega(z)|^2\theta_{r+1} & \\ \quad + \{z^{k-j}\}^\sharp |\Xi(z)|^{-4}|\Omega(z)|^2\theta_{r+1}, & q+1 \le j,\ k \le r, \\ \{z^{-j+q}\Xi(z)\}^\sharp |\Xi(z)|^{-2}|\Omega(z)|^2, & q+1 \le j \le r,\ k = r+1, \\ \{z^{-k}\Omega(z)\}^\sharp |\Xi(z)|^{-2}, & j = r+1,\ 1 \le k \le q, \\ \{z^{-k+q}\Xi(z)\}^\sharp |\Xi(z)|^{-2}|\Omega(z)|^2, & j = r+1,\ q+1 \le k \le r, \\ 0, & j = k = r+1. \end{cases}$$

First considering the terms that do not involve $\tilde{f}$, we have

$$f_\theta^{-1}(\lambda)\frac{\partial^2 f_\theta(\lambda)}{\partial\theta_j\partial\theta_k} - f_\theta^{-2}(\lambda)\frac{\partial f_\theta(\lambda)}{\partial\theta_j}\frac{\partial f_\theta(\lambda)}{\partial\theta_k} = f_\theta^{-2}(\lambda)\begin{cases} -\{z^{-j-k}\Omega^2(z)\}^\sharp |\Xi(z)|^{-4}\theta_{r+1}^2, & 1 \le j,\ k \le q, \\ \{z^{-j-k}\Xi^2(z)\}^\sharp |\Xi(z)|^{-8}|\Omega(z)|^4\theta_{r+1}^2, & q+1 \le j,\ k \le r, \\ -|\Xi(z)|^{-4}|\Omega(z)|^4, & j = k = r+1, \\ 0 & \text{else}, \end{cases}$$

which integrates to zero unless $j=k=r+1$, in which case we obtain $-\theta_{r+1}^{-2}$. Now for the terms involving $\tilde{f}$ we have

$$2f_\theta^{-3}(\lambda)\frac{\partial f_\theta(\lambda)}{\partial\theta_j}\frac{\partial f_\theta(\lambda)}{\partial\theta_k}-f_\theta^{-2}(\lambda)\frac{\partial^2 f_\theta(\lambda)}{\partial\theta_j\partial\theta_k}=\begin{cases}(\{z^{-j-k}\Omega^2(z)\}^\sharp+\{z^{-j}\Omega(z)\}^\sharp\{z^{-k}\Omega(z)\}^\sharp)\dfrac{|\Xi(z)|^2}{|\Omega(z)|^6\theta_{r+1}}, & 1\le j,\ k\le q,\\ \{z^{-j}\Omega(z)\}^\sharp\{z^{-k+q}\Xi(z)\}^\sharp|\Omega(z)|^{-4}\theta_{r+1}^{-1}, & 1\le j\le q,\ q+1\le k\le r,\\ \{z^{-j}\Omega(z)\}^\sharp|\Omega(z)|^{-4}|\Xi(z)|^2\theta_{r+1}^{-2}, & 1\le j\le q,\ k=r+1,\\ \{z^{-j+q}\Xi(z)\}^\sharp\{z^{-k}\Omega(z)\}^\sharp|\Omega(z)|^{-4}\theta_{r+1}^{-1}, & q+1\le j\le r,\ 1\le k\le q,\\ \{z^{j-k}\}^\sharp|\Omega(z)|^{-2}\theta_{r+1}^{-1}, & q+1\le j,\ k\le r,\\ \{z^{-j+q}\Xi(z)\}^\sharp|\Omega(z)|^{-2}\theta_{r+1}^{-2}, & q+1\le j\le r,\ k=r+1,\\ \{z^{-k}\Omega(z)\}^\sharp|\Xi(z)|^2|\Omega(z)|^{-4}\theta_{r+1}^{-2}, & j=r+1,\ 1\le k\le q,\\ \{z^{-k+q}\Xi(z)\}^\sharp|\Omega(z)|^{-2}\theta_{r+1}^{-2}, & j=r+1,\ q+1\le k\le r,\\ 2|\Xi(z)|^2|\Omega(z)|^{-2}\theta_{r+1}^{-3}, & j=k=r+1.\end{cases}$$

The corresponding FTs can now be easily obtained using (A.2), and the estimate of M_f is formed by utilizing (8) once again. Noting that $R'\Gamma(\nu)R=(2\pi)^{-1}\int_{-\pi}^{\pi}\nu(\lambda)I^2(\lambda)\,d\lambda$ for any bounded $\nu(\lambda)$ and R the $2n-1$ vector of sample autocovariances (i.e., $R_j=n^{-1}\sum_{t=1}^{n-|j-n|}W_tW_{t+|j-n|}$), it follows that $\hat{W}_{kl}(\theta)$ is given by

$$R'\Gamma(g_{\theta^{(k)}}g_{\theta^{(l)}})R+\hat{b}'_{\theta^{(l)},l}\hat{M}_f(\theta^{(l)})^{-1}\left\{R'\Gamma\left(g_{\theta^{(k)}}f_{\theta^{(l)}}^{-2}\frac{\partial f_{\theta^{(l)}}}{\partial\theta_j^{(l)}}\right)R\right\}_{j=1}^{r_l+1}+\hat{b}'_{\theta^{(k)},k}\hat{M}_f(\theta^{(k)})^{-1}\left\{R'\Gamma\left(g_{\theta^{(l)}}f_{\theta^{(k)}}^{-2}\frac{\partial f_{\theta^{(k)}}}{\partial\theta_j^{(k)}}\right)R\right\}_{j=1}^{r_k+1}$$
$$+\hat{b}'_{\theta^{(l)},l}\hat{M}_f(\theta^{(l)})^{-1}\left\{R'\Gamma\left(f_{\theta^{(l)}}^{-2}f_{\theta^{(k)}}^{-2}\frac{\partial f_{\theta^{(l)}}}{\partial\theta_i^{(l)}}\frac{\partial f_{\theta^{(k)}}}{\partial\theta_j^{(k)}}\right)R\right\}_{i,j=1}^{r_l+1,r_k+1}\hat{M}_f(\theta^{(k)})^{-1}\hat{b}_{\theta^{(k)},k},$$

with $k,l=1,2$. From the description of g given above, it is clear how to obtain the first term. The middle two terms require the computation of $g_{\theta^{(k)}}f_{\theta^{(l)}}^{-2}\partial f_{\theta^{(l)}}/\partial\theta_j^{(l)}$, which is

$$\begin{cases}\{z^{-j}\Omega^{(l)}(z)\}^\sharp|\Xi^{(l)}(z)|^2|\Omega^{(l)}(z)|^{-4}|\Phi^{(k)}(z)|^2|\Xi^{(k)}(z)|^2|\Omega^{(k)}(z)|^{-2}\theta_{r_l+1}^{-1}, & 1\le j\le q_l,\\ \{z^{-j+q_l}\Xi^{(l)}(z)\}^\sharp|\Omega^{(l)}(z)|^{-2}|\Phi^{(k)}(z)|^2|\Xi^{(k)}(z)|^2|\Omega^{(k)}(z)|^{-2}\theta_{r_l+1}^{-1}, & q_l+1\le j\le r_l,\\ |\Xi^{(l)}(z)|^2|\Omega^{(l)}(z)|^{-2}|\Phi^{(k)}(z)|^2|\Xi^{(k)}(z)|^2|\Omega^{(k)}(z)|^{-2}/\theta_{r_l+1}^{-2}, & j=r_l+1.\end{cases}$$

The last term requires the computation of

$$f_{\theta^{(l)}}^{-2}\frac{\partial f_{\theta^{(l)}}}{\partial\theta_i^{(l)}}f_{\theta^{(k)}}^{-2}\frac{\partial f_{\theta^{(k)}}}{\partial\theta_j^{(k)}}=\begin{cases}\dfrac{\{z^{-j}\Omega^{(k)}(z)\}^\sharp\{z^{-i}\Omega^{(l)}(z)\}^\sharp|\Xi^{(k)}(z)|^2|\Xi^{(l)}(z)|^2}{|\Omega^{(k)}(z)|^4|\Omega^{(l)}(z)|^4\theta_{r_k+1}\theta_{r_l+1}}, & 1\le j\le q_k,\ 1\le i\le q_l,\\ \dfrac{\{z^{-j}\Omega^{(k)}(z)\}^\sharp\{z^{-i+q_l}\Xi^{(l)}(z)\}^\sharp|\Xi^{(k)}(z)|^2}{|\Omega^{(k)}(z)|^4|\Omega^{(l)}(z)|^2\theta_{r_k+1}\theta_{r_l+1}}, & 1\le j\le q_k,\ q_l+1\le i\le r_l,\\ \dfrac{\{z^{-j}\Omega^{(k)}(z)\}^\sharp|\Xi^{(k)}(z)|^2|\Xi^{(l)}(z)|^2}{|\Omega^{(k)}(z)|^4|\Omega^{(l)}(z)|^2\theta_{r_k+1}\theta_{r_l+1}^2}, & 1\le j\le q_k,\ i=r_l+1,\\ \dfrac{\{z^{-j+q_k}\Xi^{(k)}(z)\}^\sharp\{z^{-i}\Omega^{(l)}(z)\}^\sharp|\Xi^{(l)}(z)|^2}{|\Omega^{(k)}(z)|^2|\Omega^{(l)}(z)|^4\theta_{r_k+1}\theta_{r_l+1}}, & q_k+1\le j\le r_k,\ 1\le i\le q_l,\\ \dfrac{\{z^{j-q_k}\Xi^{(k)}(z)\}^\sharp\{z^{i-q_l}\Xi^{(l)}(z)\}^\sharp}{|\Omega^{(k)}(z)|^2|\Omega^{(l)}(z)|^2\theta_{r_k+1}\theta_{r_l+1}}, & q_k+1\le j\le r_k,\ q_l+1\le i\le r_l,\\ \dfrac{\{z^{-j+q_k}\Xi^{(k)}(z)\}^\sharp}{|\Omega^{(k)}(z)|^2|\Omega^{(l)}(z)|^2|\Xi^{(l)}(z)|^2\theta_{r_k+1}\theta_{r_l+1}^2}, & q_k+1\le j\le r_k,\ i=r_l+1,\\ \dfrac{\{z^{-i+q_l}\Omega^{(l)}(z)\}^\sharp|\Xi^{(k)}(z)|^2|\Xi^{(l)}(z)|^2}{|\Omega^{(k)}(z)|^2|\Omega^{(l)}(z)|^4\theta_{r_k+1}^2\theta_{r_l+1}}, & j=r_k+1,\ 1\le i\le q_l,\\ \dfrac{\{z^{-i+q_l}\Xi^{(l)}(z)\}^\sharp|\Xi^{(k)}(z)|^2}{|\Omega^{(k)}(z)|^2|\Omega^{(l)}(z)|^2\theta_{r_k+1}^2\theta_{r_l+1}}, & j=r_k+1,\ q_l+1\le i\le r_l,\\ \dfrac{|\Xi^{(k)}(z)|^2|\Xi^{(l)}(z)|^2}{|\Omega^{(k)}(z)|^2|\Omega^{(l)}(z)|^2\theta_{r_k+1}^2\theta_{r_l+1}^2}, & j=r_k+1,\ i=r_l+1.\end{cases}$$

From these expressions the FTs can be obtained, and $\hat{W}_{kl}(\hat{\theta})$ can be computed; from the previous discussion this converges in probability to $W_{kl}(\tilde{\theta})$, and therefore is used to normalize the diagnostic given in (12).

As a final note, we can easily extend our methods to so-called "gap" models. These are ARMA models where some subset of the coefficients are fixed ahead of time to chosen values. In this case the corresponding derivatives are zero, and the expressions for b and the Hessian matrix are simplified. Letting J denote a selection matrix such that $J\theta$ consists only of the nonfixed parameters, we can replace b by Jb and M_f^{-1} by $JM_f^{-1}J'$—and similarly in the expressions for each $\hat{W}_{kl}$—for each of the two models. This technique will provide the correct uncertainty formulas.

A.3. Derivation of (7)

The integrand of $(1/2\pi)\int_{-\pi}^{\pi}|\eta^{(h)}(e^{-i\lambda})|^2 I(\lambda)\,d\lambda$ is n^{-1} times the squared modulus of

$$\sum_{j=0}^{\infty}\eta_j^{(h)}e^{-ij\lambda}\sum_{t=1}^{n}W_t e^{-it\lambda}=\sum_{k=1}^{\infty}c_k e^{-ik\lambda},$$

where $c_k=\sum_{j=0}^{k-1}\eta_j^{(h)}W_{k-j}$ if $1\le k\le n$ and $c_{n+q}=\sum_{j=0}^{n-1}\eta_{j+q}^{(h)}W_{n-j}$ for $q\ge 1$. By Parseval's identity,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}|\eta^{(h)}(e^{-i\lambda})|^2 I(\lambda)\,d\lambda=\frac{1}{n}\sum_{k=1}^{n}c_k^2+\frac{1}{n}\sum_{q=1}^{\infty}c_{n+q}^2,$$

with $n^{-1}\sum_{k=1}^{n}c_k^2=n^{-1}\sum_{t=1}^{n}[\hat{\varepsilon}_t^{(h)}]^2$ for the truncated filter forecast errors (4). Thus, if the truncated filter forecasts are used, the difference of the two measures in (7) is given by $n^{-1}\sum_{q=1}^{\infty}c_{n+q}^2$, which we will show is $O_p(n^{-1})$ by proving that

$$E\left(\sum_{q=1}^{\infty}c_{n+q}^2\right)=\sum_{q=1}^{\infty}Ec_{n+q}^2<\infty. \tag{A.3}$$

If the finite-past forecasts detailed in Findley et al. (2004, Section 3.2.1) are used, then under (5), we have $n^{-1}\sum_{k=1}^{n}c_k^2-n^{-1}\sum_{t=1}^{n}[\hat{\varepsilon}_t^{(h)}]^2=o_p(n^{-1/2})$. This can be verified by a simplification of the proof of Proposition 5.1 of Findley (1991a), which shows under (5) that differences between the averages of squared finite-sample and infinite-past forecast errors are of order $o_p(n^{-1/2})$. Thus, for finite-past forecast errors also, (7) will follow from (A.3).

To verify (A.3), we observe that, by the invertibility assumption for the ARMA model for W_t, there exist $0<v<1$ and $K>0$ such that

$$\begin{aligned}Ec_{n+q}^2&=\sum_{j,k=0}^{n-1}\eta_{j+q}^{(h)}\eta_{k+q}^{(h)}\gamma_{|j-k|}(\tilde{f})\le\sum_{j,k=0}^{n-1}|\eta_{j+q}^{(h)}||\eta_{k+q}^{(h)}||\gamma_{|j-k|}(\tilde{f})|\le K|\gamma_0(\tilde{f})|\sum_{j,k=0}^{n-1}v^{2q+j+k}\\&=Kv^{2q}|\gamma_0(\tilde{f})|\left(\sum_{j=0}^{n-1}v^j\right)^2<Kv^{2q}|\gamma_0(\tilde{f})|\left(\frac{1}{1-v}\right)^2,\end{aligned}$$

from which (A.3) follows immediately, and thereby also (7).

A.4. Derivation of $V_{c,MR}=V_c$

Applying Parseval's identity, (17) can be reformulated in terms of the spectral and cross spectral densities $f_{vv}(\lambda)$, $f_{ww}(\lambda)$ and $f_{vw}(\lambda)$ of v_t and w_t, and then in terms of $\tilde{f}(\lambda)$ and the transfer functions $H_j(\lambda)=\eta_{\theta^{(j)}}^{(h)}(e^{-i\lambda})$ of the forecast error filters defining $\varepsilon_t^{(h)}(\theta^{(j)})$, $j=1,2$:

$$V_{c,MR}=\frac{1}{2\pi}\int_{-\pi}^{\pi}[f_{vv}(\lambda)f_{ww}(\lambda)+f_{vw}^2(\lambda)]\,d\lambda=\frac{1}{2\pi}\int_{-\pi}^{\pi}\tilde{f}^2(\lambda)H(\lambda)\,d\lambda, \tag{A.4}$$

where (suppressing the λ argument)

$$\begin{aligned}H&=(H_1+H_2)(\overline{H}_1+\overline{H}_2)(H_1-H_2)(\overline{H}_1-\overline{H}_2)+(H_1+H_2)^2(\overline{H}_1-\overline{H}_2)^2\\&=2(|H_1|^2-|H_2|^2)^2+2(|H_1|^2-|H_2|^2)(H_2\overline{H}_1-\overline{H}_2H_1)=2(g_1-g_2)^2+2(g_1-g_2)(H_2\overline{H}_1-\overline{H}_2H_1).\end{aligned}$$

After multiplication by the even function $\tilde{f}^2$, the final term on the right remains an odd (and imaginary) function. Thus its integral is zero and $V_{c,MR}=V_c$ follows.

References

Bell, W., 1984. Signal extraction for nonstationary time series. Annals of Statistics 12, 646–664.
Box, G., Jenkins, G., Reinsel, G., 1994. Time Series Analysis: Forecasting and Control, third ed. Prentice-Hall, Englewood Cliffs.

Brockwell, P., Davis, R., 1991. Time Series: Theory and Methods. Springer-Verlag, New York.
Brockwell, P., Davis, R., 2002. Introduction to Time Series and Forecasting. Springer-Verlag, New York.
Chen, W., Deo, R., 2000. On the integral of the squared periodogram. Stochastic Processes and their Applications 85, 159–176.
Clark, T.E., McCracken, M.W., 2001. Tests of forecast accuracy and encompassing for nested models. Journal of Econometrics 105, 85–110.
Clark, T.E., McCracken, M.W., 2005. Evaluating direct multi-step forecasts. Econometric Reviews 24, 369–404.
Dahlhaus, R., Wefelmeyer, W., 1996. Asymptotically optimal estimation in misspecified time series models. Annals of Statistics 16, 952–974.
Diebold, F., Mariano, R., 1995. Comparing predictive accuracy. Journal of Business and Economic Statistics 13, 253–263.
Findley, D., 1990. Making difficult model comparisons. SRD Research Report no. RR90/11, U.S. Census Bureau 〈http://www.census.gov/srd/www/byname.html〉.
Findley, D.F., 1991a. Convergence of finite multistep predictors from incorrect models and its role in model selection. Note di Matematica XI, 145–155 〈http://www.census.gov/ts/papers/convergence.pdf〉.
Findley, D.F., 1991b. Counterexamples to parsimony and BIC. Annals of the Institute of Statistical Mathematics 43, 509–514 〈http://www.census.gov/ts/papers/counterexamples.pdf〉.
Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C., Chen, B.C., 1998. New capabilities and methods of the X-12-ARIMA seasonal adjustment program. Journal of Business & Economic Statistics 16, 127–177 (with discussion).
Findley, D.F., Pötscher, B.M., Wei, C.-Z., 2004. Modeling of time series arrays by multistep prediction or likelihood methods. Journal of Econometrics 118, 151–187.
McElroy, T., Holan, S., 2009. A local spectral approach for assessing time series model misspecification. Journal of Multivariate Analysis 100, 604–621.
Meese, R., Rogoff, K., 1988. Was it real? The exchange rate-interest differential relation over the modern floating-rate period. Journal of Finance 43, 933–948.
Ploberger, W., 1982. Slight misspecifications for linear systems. In: Fechtinger, G., Kali, P. (Eds.), Operations Research in Progress. Riedel, Dordrecht, pp. 413–424.
Pourahmadi, M., 2001. Foundations of Time Series Analysis and Prediction Theory. Wiley, New York.
R Development Core Team, 2008. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna 〈http://www.R-project.org〉.
Rivers, D., Vuong, Q., 2002. Model selection tests for nonlinear dynamic models. Econometrics Journal 5, 1–39.
Taniguchi, M., Kakizawa, Y., 2000. Asymptotic Theory of Statistical Inference for Time Series. Springer-Verlag, New York.
Vuong, Q., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 307–333.
West, K., 1996. Asymptotic inference about predictive ability. Econometrica 64, 1067–1084.