Selection Between Models Through Multi-Step-Ahead Forecasting

Tucker McElroy[1] and David Findley[1,*]
U.S. Census Bureau

[1] Statistical Research Division, U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-9100

∗ Corresponding author E-mail address: david.f.findley@census.gov

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 mean square forecast error. 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. The simpler variance estimate, which ignores estimation uncertainty, can be rather straightforwardly calculated for any pair of ARIMA models with the same differencing operator, and its broad consistency property offers improvement to what are known as tests of Diebold and Mariano (1995) type.

[1] Statistical Research Division, U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-9100

∗ Corresponding author. E-mail address: david.f.findley@census.gov

Keywords. ARIMA models; Diebold-Mariano tests; Misspecified models; Model selection; Parameter estimation effects; Time series

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 the effect 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 (5) 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 2.2, whose Theorem 2 provides the CLT and consistent estimator of its variance needed for our main test statistic (11). Section 2.3 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 also 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 non-nested model comparison test statistic of Vuong (1989).

Section 3 contains a simulation-based power study of our test statistics and an empirical study of their application to competing models for series from Box, Jenkins and Reinsel (1994) and Brockwell and Davis (2002). There are fewer results for statistics with variance estimates that account for parameter estimation uncertainty because we currently have complete algebraic formulas only for moving average models, meaning formulas in which algebraic expressions are available for all occurring derivatives of functions of model spectral densities with respect to parameters. But our limited results for this variance estimate suggest that, with average sample sizes and for low order models, its value differs negligibly from the value of the simpler variance estimate that does not

account for parameter estimation uncertainty. For the latter estimate, we also present results for AR and ARMA models and associated difference-stationary models, including results for a form of Diebold-Mariano test.

Our numerical results do not include a size study of the test statistics. We explain why this is difficult to do and prove in a Remark in Section 3 that the size study of Diebold and Mariano (1995) is invalid.

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. The stationary series $W_t = \delta(B) Y_t$ is assumed to be Gaussian with mean zero and to be purely nondeterministic. Thus its spectral density $\tilde{f}$ is log integrable and generates its autocovariances via

$$\gamma_j(\tilde{f}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \tilde{f}(\lambda) e^{ij\lambda} \, d\lambda,$$

a formula that shows our convention with the constant 2π. The Toeplitz matrix of these autocovariances is denoted $\Gamma(\tilde{f})$, i.e. $\Gamma_{jk}(\tilde{f}) = \gamma_{j-k}(\tilde{f})$. The order of $\Gamma(\tilde{f})$ is the number of W_t used.

The statistics we use to measure multi-step-ahead forecast performance will be called goodness-of-fit (gof) diagnostics because of their similarity to the gof diagnostics of McElroy and Holan (2009). We first motivate their formulas by deriving asymptotic connections with average squared within-sample forecast error and with mean square forecast error. We then present our tests for comparing two competing models on the basis of forecast performance.

2.1 Multi-Step-Ahead Forecasting

We start by reviewing some basic forecasting results for nonstationary Y_t. 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, $1 - d \leq t \leq n$ denote the available data. Set $\xi(z) = 1/\delta(z)$ expressed as a power series with coefficients ξ_j. We define $[\xi]_0^{h-1}(z) = \sum_{j=0}^{h-1} \xi_j z^j$. For any $1 \leq h < n$ and any $1 \leq t \leq n - h$, we have $Y_{t+h} = [\xi]_0^{h-1}(B) W_{t+h} + \sum_{j=0}^{d-1} c_{j,h} Y_{t-j}$, where the coefficients $c_{j,h}$ depend only on the coefficients of $\delta(z)$, see Bell (1984,

p. 650). 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

$$\hat{Y}_{t+h|t} = [\xi]_0^{h-1}(B)\hat{W}_{t+h|t} + \sum_{j=0}^{d-1} c_{j,h} Y_{t-j}. \tag{1}$$

Consequently, the forecast errors are given by $Y_{t+h} - \hat{Y}_{t+h|t} = [\xi]_0^{h-1}(B)\left(W_{t+h} - \hat{W}_{t+h|t}\right)$.

To motivate our gof diagnostic, we start with the forecasts $\hat{W}_{t+h|t}$ 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 $W_{t+h} - W_{t+h|t} = [\Psi]_0^{h-1}(B)\Psi^{-1}(B)W_{t+h} = [\Psi]_0^{h-1}(B)\varepsilon_t$, this forecast error is a moving average process of order (at most) $h-1$, as is also the error process of the forecasts $Y_{t+h|t} = [\xi]_0^{h-1}(B)W_{t+h|t} + \sum_{j=0}^{d-1} c_{j,h} Y_{t-j}$,

$$\begin{aligned} Y_{t+h} - Y_{t+h|t} &= \sum_{j=0}^{h-1} \xi_j B^j \left(W_{t+h} - W_{t+h|t}\right) \\ &= \sum_{j=0}^{h-1} \xi_j B^j [\Psi]_0^{h-j}(B)\Psi^{-1}(B) W_{t+h} \\ &= \sum_{j=0}^{h-1} \xi_j B^j [\Psi]_0^{h-1-j}(B)\varepsilon_{t+h}, \end{aligned} \tag{2}$$

because each of the polynomials $B^j[\Psi]_0^{h-1-j}(B)$ has degree $h-1$.

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 $j \ge t$. Similarly, denoting the filter in (2) by

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

the truncated filter forecast errors $\hat{\varepsilon}_t^{(h)} = Y_t - \hat{Y}_{t|t-h}$ are given by

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

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 $\Psi\left(e^{-i\lambda}\right) = 1 + \sum_{j=1}^{\infty} \psi_j e^{-ij\lambda}$ satisfying

$\int_{-\pi}^{\pi} \log\left|\Psi\left(e^{-i\lambda}\right)\right| d\lambda = 0$ and such that the model spectral density is equal to $\sigma^2\left|\Psi\left(e^{-i\lambda}\right)\right|^2$ for some $\sigma^2 > 0$, see Theorem VII of Pourahmadi (2001, p. 68). (For an ARMA model considered for W_t, with AR and MA polynomials $\alpha(B)$ and $\beta(B)$ respectively, $\Psi(B) = \beta(B)\,\alpha^{-1}(B)$.) 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 $\tilde{f}(\lambda) = \sigma^2\left|\Psi\left(e^{-i\lambda}\right)\right|^2$ for some $\sigma^2 > 0$, then the series ε_t will not be white noise and $\sum_{j=0}^{h-1} \xi_j B^j \varepsilon_t$ will generally not be a moving average process of order $h-1$.

One measure 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 predictors or from the standard finite-past predictors discussed, for example, in Section 3.3.1 of Findley, Pötscher and Wei (2004). With either predictor, for an invertible ARIMA model for example, Proposition 4.1 of Findley (1991a) shows that, as $n \to \infty$, this average converges in probability to the mean square infinite-past forecast error, $(1/2\pi)\int_{-\pi}^{\pi}\left|\eta^{(h)}\left(e^{-i\lambda}\right)\right|^2 \tilde{f}(\lambda)\, d\lambda$, at the rate $O_p\left(n^{-1/2}\right)$, as does also the expression on the right in the approximation

$$\frac{1}{n}\sum_{t=1}^{n}[\hat{\varepsilon}_t^{(h)}]^2 \approx \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\eta^{(h)}\left(e^{-i\lambda}\right)\right|^2 I(\lambda)\, d\lambda, \tag{4}$$

in which I denotes the periodogram of W_t, $t = 1, 2, \ldots, n$, see Lemma 3.3.1 of Taniguchi and Kakizawa (2002). Thus the error of (4) is of order at most $O_p\left(n^{-1/2}\right)$.

We will use the expression on the right in (4) as a model gof diagnostic for h-step-ahead forecasting. Note that taking $h = 1$ and dividing through by the innovation variance yields the log Whittle likelihood in the stationary case, excluding the term involving the log of the determinant. In recognition of these similarities, with the column vector $W = (W_1, \ldots, W_n)'$ we define the following quadratic form gof diagnostic:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} g(\lambda)\, I(\lambda)\, d\lambda = \frac{1}{n} W'\Gamma(g)W, \tag{5}$$

with $g(\lambda) = \left|\eta^{(h)}\left(e^{-i\lambda}\right)\right|^2$, the squared gain of the model's h-step forecast error filter. Now we embed (5) into a framework similar to that of McElroy and Holan (2009) extended to handle two or more models.

2.2 GOF Diagnostics

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 $(-\pi, \pi)\backslash\{0\}$. To match the generality of McElroy and Holan (2009), in our general result, Theorem 1 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

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

which has the form of the l.h.s. of (5). 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 $\theta = \left(\theta^{(1)}, \ldots, \theta^{(l)}\right)$ with values in the Cartesian product of the L compact parameter spaces $\Theta^{(i)}$. The Quasi-Maximum-Likelihood estimates (QMLEs) of these models are defined as the minimizers of $D(f_{\theta^{(i)}}, I)$, where $D(k, h)$ is the Kullback-Leibler (KL) discrepancy (see Dahlhaus and Wefelmeyer (1996)):

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

Likewise, the corresponding pseudo-true values $\tilde{\theta}^{(i)}$ are, by definition, minimizers of $D(f_{\theta^{(i)}}, \tilde{f})$ over $\theta^{(i)} \in \Theta^{(i)}$, where $\tilde{f}$ is the true spectral density. We need the following conditions, which are simple extensions of those found in McElroy and Holan (2009):

1. W_t is stationary, mean zero and Gaussian.

2. Each $\Theta^{(i)}$ for $i = 1, 2, \cdots, L$ is compact and convex.

3. Each $\tilde{\theta}^{(i)}$, the pseudo-true value of the parameter in the i-th model, exists uniquely and lies in the interior of $\Theta^{(i)}$.

4. The model spectral densities $f_{\theta^{(i)}}(\lambda)$ are twice continuously differentiable in $\theta^{(i)}$, and are continuous in λ.

5. The weighting functions $g_{\theta^{(i)},i}(\lambda)$ are twice continuously differentiable in $\theta^{(i)}$, and are continuous in λ.

6. The matrices $M_f(\theta^{(i)})$, which are the Hessians of the KL discrepancy between $f_{\theta^{(i)}}$ and $\tilde{f}$, are each nonsingular at $\theta^{(i)} = \tilde{\theta}^{(i)}$.

7. The first derivatives of the model spectral densities are uniformly bounded and bounded away from zero (in λ).

In assumption 5 we see that each weighting function g_i depends on i and $\theta^{(i)}$ but not on any other parameter vectors. 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 Theorems 1 and 2 of McElroy and Holan (2009), but gives a slightly different treatment more appropriate for our situation; moreover it considers the case of L models.

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

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

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

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

These entries are defined in terms of the following quantities:

$$\begin{aligned}
p_{\theta^{(i)},i}(\lambda) &= f_{\theta^{(i)}}^{-2}(\lambda) b'_{\theta^{(i)},i} M_f^{-1}(\theta^{(i)}) \nabla_{\theta^{(i)}} f_{\theta^{(i)}}(\lambda) \\
b_{\theta^{(i)},i} &= \frac{j_i!}{2\pi} \int_{-\pi}^{\pi} \tilde{f}^{j_i}(\lambda) \nabla_{\theta^{(i)}} g_{\theta^{(i)},i}(\lambda)\, d\lambda \\
M_f(\theta^{(i)}) &= \nabla_{\theta^{(i)}} \nabla'_{\theta^{(i)}} D(f_{\theta^{(i)}}, \tilde{f}).
\end{aligned}$$

We are interested in the case of two fitted models for the data whose forecast performance we want 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 (5) 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 $f_{\theta^{(i)}}(\lambda) = \sigma^2_{(i)} \left|\Psi_{\theta^{(i)}}\left(e^{-i\lambda}\right)\right|^2$ with $\sigma_{(i)}$ not functionally related to how $\theta^{(i)}$ determines $\Psi_{\theta^{(i)}}\left(e^{-i\lambda}\right)$, $i = 1, 2$. Then application of Theorem 1

with $j_1 = 1 = j_2$ shows that

$$\left\{\sqrt{n}\left(Q_n(I,\hat{g}_i,\hat{\theta}^{(i)}) - Q_n(\tilde{f},\tilde{g}_i,\tilde{\theta}^{(i)})\right)\right\}_{i=1,2} \stackrel{\mathcal{L}}{\Longrightarrow} \mathcal{N}\left(0, W(\tilde{\theta})\right). \tag{6}$$

The entries of the asymptotic variance matrix are as follows:

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

for $k, l = 1, 2$. 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. Our null hypothesis is that the models have the same asymptotic mean square forecast error,

$$H_0 : Q(\tilde{f},\tilde{g}_1,\tilde{\theta}^{(1)}) = Q(\tilde{f},\tilde{g}_2,\tilde{\theta}^{(2)}). \tag{7}$$

Therefore, we consider the statistic

$$\sqrt{n}\left(Q(I,\hat{g}_1,\hat{\theta}^{(1)}) - Q(I,\hat{g}_2,\hat{\theta}^{(2)})\right). \tag{8}$$

Applying the vector $(1,-1)$ to the joint convergence (6), we obtain the first assertion of the following result:

Theorem 2 *Under the assumptions of Theorem 1 plus the assumption that the logs of the model spectral densities are integrable, we have*

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

where $V = W_{11}(\tilde{\theta}) - 2W_{12}(\tilde{\theta}) + W_{22}(\tilde{\theta})$ *has the formula*

$$V = \frac{1}{\pi}\int_{-\pi}^{\pi} \tilde{f}^2\left(\tilde{g}_1 + \tilde{p}_1 - \tilde{g}_2 - \tilde{p}_2\right)^2 d\lambda, \tag{9}$$

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

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

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* $\tilde{p}_1$ *and* $\tilde{p}_2$.

Under H_0 the asymptotic mean is zero, because $Q(\tilde{f},\tilde{g}_1,\tilde{\theta}^{(1)}) - Q(\tilde{f},\tilde{g}_2,\tilde{\theta}^{(2)})$ is the integral of

$$\tilde{f}\left(\tilde{g}_1 - \tilde{g}_2\right). \tag{10}$$

So if we define the test statistic

$$T_{\widehat{V}} = \left(\widehat{V}/n\right)^{-1/2} \left(Q(I, \hat{g}_1, \hat{\theta}^{(1)}) - Q(I, \hat{g}_2, \hat{\theta}^{(2)})\right), \tag{11}$$

then when $V > 0$ and the assumptions of Theorem 2 hold, $T_{\widehat{V}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$.

A test based on $T_{\widehat{V}}$ will have adequate power for distinguishing between the h-step forecasting performance of the two models when $\sqrt{n}$ times integral of (10) is relatively large. 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 the one-step-ahead forecast errors from each model. The function (10) includes the multi-step-ahead performance of each model through the forecast error filter functions $\eta^{(h)}_{\tilde{\theta}^{(i)}}(z) = \sum_{j=0}^{h-1} \xi_j z^j [\Psi_{\tilde{\theta}^{(i)}}]_0^{h-j}(z) \Psi^{-1}_{\tilde{\theta}^{(i)}}(z)$ used to define $g_i = \left|\eta^{(h)}_{\tilde{\theta}^{(i)}}\left(e^{-i\lambda}\right)\right|^2$, $i = 1, 2$.

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.

2.3 The Case of Constant Parameters

Meese and Rogoff (1988) seems to be the earliest article in which the limiting distribution of the difference of average multi-step forecast squared errors (the l.h.s. of (4)) from two models is obtained for a stationary Gaussian time series W_t, along with an estimate of its variance, and thereby also a test statistic for the null hypothesis of equal limiting mean square h -step forecast error of the two models. 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 $\varepsilon_t^{(h)}\left(\theta^{(1)}\right)$ and $\varepsilon_t^{(h)}\left(\theta^{(2)}\right)$ of forecasts from the infinite past from models whose parameters $\theta^{(1)}$ and $\theta^{(2)}$ are constant rather than estimated. (Because the parameters used in practice are always estimated, the constant parameters should be regarded as the pseudo-true values to which the estimates converge when Assumption 3 holds.) Unaware of the work of Meese and Rogoff, for the null hypothesis (7) and a large class of stationary time series models, Findley (1991a) obtained a limiting distribution equivalent to theirs for the errors $\hat{\varepsilon}_t^{(h)}\left(\theta^{(i)}\right)$ from the standard *finite-past* predictors defined by constant parameters $\theta^{(i)}, i = 1, 2$,

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

but provided no estimator for the limiting variance V_{MR}. Meese and Rogoff's formula for V_{MR} will be presented below in (14) and shown to have the value

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

This is the variance that Theorem 2 yields for the constant parameter case,

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

because the terms in (9) involving derivatives with respect to the parameters drop out. Theorem 2 provides a consistent estimate of V_c,

$$\begin{aligned} \widehat{V}_c &= \frac{1}{2\pi} \int_{-\pi}^{\pi} I^2 \left(g_1 - g_2\right)^2 d\lambda \\ &= \sum_{j,k=-n+1}^{n-1} \hat{\gamma}_j \hat{\gamma}_k \left\{ \Gamma_{j-k}(g_1^2) + \Gamma_{j-k}(g_2^2) - 2\,\Gamma_{j-k}(g_1 g_2) \right\}, \end{aligned} \tag{12}$$

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

$$T_{\widehat{V}_c} = \left(\widehat{V}_c / n \right)^{-1/2} \left(Q(I, g_1, \theta^{(1)}) - Q(I, g_2, \theta^{(2)}) \right). \tag{13}$$

The simplifications of the proof of Theorem 2 that result from using constant parameters show that $T_{\widehat{V}_c}$ has an $\mathcal{N}\left(0,1\right)$ 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 non-nested model comparison test statistic of Vuong (1989) in Findley (1990) if the $h = 1$ instance of $\widehat{V}_c$ 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 autocovariance matrices on the right in (12) can be calculated by identifying the coefficients of the ARMA models whose spectral densities are g_1^2, g_2^2 and $g_1 g_2$. This requires identification of the coefficients ξ_j, $0 < j \leq h-1$ from the identities implied by $\xi\left(z\right)\delta\left(z\right) = 1$ and then doing polynomial multiplications. With the coefficients, the autocovariance matrices in (12) that determine $\widehat{V}_c$ can be obtained from standard algorithms. In the numerical studies below, the parameters treated as constant are maximum likelihood estimates from $W_1, \ldots, W_n$. The calculation of $\widehat{V}$ is much more complex because of its terms that involve derivatives.

To present Meese and Rogoff's formula for V_{MR} and their estimate $\widehat{V}_{MR}$ as described for general h by Diebold and Mariano (1995), set $v_t = \varepsilon_t^{(h)}\left(\theta^{(1)}\right) + \varepsilon_t^{(h)}\left(\theta^{(2)}\right)$ and $w_t = \varepsilon_t^{(h)}\left(\theta^{(1)}\right) - \varepsilon_t^{(h)}\left(\theta^{(2)}\right)$

and observe that $[\varepsilon_t^{(h)}\left(\theta^{(1)}\right)]^2 - [\varepsilon_t^{(h)}\left(\theta^{(2)}\right)]^2 = v_t w_t$. Thus, with E denoting expectation, the null hypothesis (7) is equivalent to $Ev_t w_t = 0$, and V_{MR} is the asymptotic variance of $n^{-1/2}\sum_{t=1}^{n} v_t w_t$, whose well known formula is, in the Gaussian case,

$$V_{MR} = \sum_{r=-\infty}^{\infty} \left[\gamma_{vv}(r)\,\gamma_{ww}(r) + \gamma_{vw}(r)\,\gamma_{vw}(-r)\right]. \tag{14}$$

In Subsection A.4 of the Appendix, we verify that

$$V_{MR} = V_c. \tag{15}$$

Motivated by the fact discussed above that with correct models, the h-step-head 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_{MR} defined by

$$\widehat{V}_{MR} = \sum_{r=-h+1}^{h-1} \left(1 - \frac{|r|}{n}\right) \left[\hat{\gamma}_{vv}(r)\,\hat{\gamma}_{ww}(r) + \hat{\gamma}_{vw}(r)\,\hat{\gamma}_{vw}(-r)\right], \tag{16}$$

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 $\widehat{V}_{MR}$ converges to

$$V_h = \sum_{r=-h+1}^{h-1} \left[\gamma_{vv}(r)\,\gamma_{ww}(r) + \gamma_{vw}(r)\,\gamma_{vw}(-r)\right],$$

which will generally differ from V_{MR} unless the models contain the correct model as a special case. However, in the correct model situation, with true parameter values, $w_t = 0$ and $V_h = V_{MR} = 0$, so the test statistic proposed by these authors does not have a limit distribution.

In the empirical results of the next section, for uniformity, the Meese and Rogoff test statistic is taken as $T_{MR} = \left(\hat{V}_{MR}/n\right)^{-1/2}\left(Q(I, g_1, \theta^{(1)}) - Q(I, g_2, \theta^{(2)})\right)$, which differs from their actual statistic to the extent of the approximation error in (4). However, it has the same $\mathcal{N}(0, V_c/V_h)$ limit distribution when $V_h \neq 0$.

3 Numerical Studies

We have not been able to do size studies of the statistic $T_{\widehat{V}}$ because the only examples known to us of pairs of incorrect models for which the null hypothesis (7) is satisfied are autoregressive models like those described in Section 3 of Findley (1991b), and we have not obtained the algebraic gradient and Hessian formulas needed to calculate $\widehat{V}$ for autoregressions. The autoregressive models in each of these pairs have different parameter uncertainty effects asymptotically, which should yield

$V > 0$, but they have identical forecast error processes and weighting functions, $g_1 = g_2$, so $V_c = 0$. Thus they cannot be used for size or other studies of $T_{\widehat{V}_c}$ because $\widehat{V}_c$ converges to converge to 0. Consequently, we present no size studies. We begin with an example to illustrate the concepts of Section 2 and to support a small simulation study of power. Then we apply $T_{\widehat{V}}$ and, more extensively, $T_{\widehat{V}_c}$ to published time series.

Remark. The simulation results presented as size studies of T_{MR} in Section 3 of Diebold and Mariano(1995) are not valid for this purpose. They are based on the assumption that a W_t can exist that has two incorrect models whose 2-step-ahead forecast errors processes are distinct invertible MA(1) processes with the same MA coefficient. However, for models with $|\xi_1| < 1$, for a given MA(1) polynomial $\theta(B) = 1 + \theta_1 B$ with $|\theta_1| < 1$, the MA(1) process is unique, being given by $\theta(B)\varepsilon_t$ where ε_t is the innovations process of W_t. Indeed, for any $h \geq 1$, if the zeroes of $[\Psi]_0^{h-1}(z)$ and $\theta(z)$ lie in $\{|z| > 1\}$, then from $[\Psi]_0^{h-1}(B)\Psi^{-1}(B)W_t = \theta(B)e_t$, we have $W_t = \tilde{\Psi}(B)e_t$ with $\tilde{\Psi}(B) = \Psi(B)\left([\Psi]_0^{h-1}(B)\right)^{-1}\theta(B)$. Because the inverse of $\tilde{\Psi}(B)$ is causal, $\tilde{\Psi}^{-1}(B) = \theta^{-1}(B)[\Psi]_0^{h-1}(B)\Psi^{-1}(B) = 1 + \sum_{j=1}^{\infty}\tilde{\phi}_j B^j$, 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 Power Study

We will be interested in the quantity $Q(\tilde{f}, g_1, \tilde{\theta}^{(1)}) - Q(\tilde{f}, g_2, \tilde{\theta}^{(2)})$ discussed in Section 2, which we call the "Asymptotic Mean Square Forecast Error Discrepancy" (AMSFED). We work from the following theoretical model results.

Example 1. Let $h = 1$ and suppose that model 1 is white noise, whereas model 2 is an MA(1). Further, suppose that the true process is an MA(1), so that model 2 is correctly specified. Then $\tilde{f}(\lambda) = |1 + \tilde{\theta}e^{-i\lambda}|^2\tilde{\sigma}^2$, and the pseudo-true value for the innovation variance of model 1 is $\sigma_1^2 = (2\pi)^{-1}\int_{-\pi}^{\pi}\tilde{f}(\lambda)\,d\lambda = \tilde{\sigma}^2(1 + \tilde{\theta}^2)$. For the second model, the pseudo-true values are the same as truth. So $g_1(\lambda) = 1$ and $g_2(\lambda) = \left(\tilde{\sigma}^2/\sigma_2^2\right)|1 + \theta_2 e^{-i\lambda}|^{-2}$ with $\sigma_2 = \tilde{\sigma}$ and $\theta_2 = \tilde{\theta}$. Hence the AMSFED equals $\tilde{\theta}^2\tilde{\sigma}^2$.

In our simulation study of power for $h = 1$ of $T_{\widehat{V}}$ and $T_{\widehat{V}_c}$, we used MA(1) models as data generating processes with three moving average parameter choices, .1,.3, and .6, with unit innovation variance and with sample sizes $n = 100, 200, 500$. We fitted both WN and an MA(1) to these processes, so by Example 1 the AMSFED is $\tilde{\theta}^2$. After 1,000 Monte Carlo simulations of a particular DGP, we obtained the empirical distribution of the test statistic (8) and its estimated

standard error. The power results of Table 1 were based off an upper one-sided test using the 5% standard normal critical value of 1.645. The effect of using the alternative variance estimate $\widehat{V}_c$ is small, and is sometimes beneficial ($\tilde{\theta} = .6$, $n = 100$) – it is not the case that the $\widehat{V}_c$ always becomes larger or smaller than $\widehat{V}$. Of course, its effect is more pronounced in smaller samples. The power of the procedure is excellent for $\tilde{\theta} = .3, .6$ and very poor for $\tilde{\theta} = .1$, where the WN and MA(1) are almost equally good fits.

Because the $h = 1$ forecast errors of a nonstationary model coincide with those of the differenced series, see (3), these power results also apply when model 1 is ARIMA(0,d,0) and model 2 is ARIMA(0,d,1) for any $d \geq 1$.

3.2 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, Jenkins, and Reinsel (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). We fitted ARIMA models to these series, and tested gof for one- and two-step-ahead forecasting ($h = 1, 2$). We tested various model pairs, using either $T_{\widehat{V}}$ and $T_{\widehat{V}_c}$ or $T_{\widehat{V}_c}$ and T_{MR}. With rare exceptions, $T_{\widehat{V}}$ and $T_{\widehat{V}_c}$ differed negligibly; this is illustrated by the comparison of ARMA component models of white noise for model 1 of an MA(1) for model 2 (in the framework of Section 2.2), using $d = 1$ and $h = 1$, see Table 2. (Other model comparisons not presented gave similarly close results.)

More extensive model comparisons were made with $T_{\widehat{V}_c}$ and T_{MR}. The results are summarized in Table 3. We only considered nonstationary models and only compared models with the same order of differencing. For simplicity of reference, we refer to the models by their stationary component, e.g. an ARIMA(1,1,1) model is referred to as an ARMA(1,1) model even though it is the forecast errors of the nonstationary model that are considered. This distinction is important only for $h > 1$, as the formula (3) shows.

For Series A (197 observations) the recommended nonstationary model in Box, Jenkins, and

Reinsel (1994) is an ARIMA(0,1,1). In terms of stationary components, we used an MA(1) as model 2, testing against an MA(2), ARMA(1,1), AR(1), and AR(2) for model 1. Negative values in the table (when statistically significant) indicate that model 1 is favored, whereas positive values indicate that the recommended MA(1) is favored. The results indicate that an MA(2) is marginally preferable to the MA(1), but the MA(1) is significantly superior to an ARMA(1,1), AR(1), or AR(2). Results for $h = 1$ and $h = 2$ are similar, but there are some conspicuous differences between the values of $T_{\widehat{V}_c}$ and T_{MR} (e.g., AR(2) with $h = 1$).

The situation for Series B (369 observations) is similar; the recommended models are an ARIMA (0,1,0) and an ARIMA(0,1,1). As model 1, we compared the same set of stationary components used for Series A to an MA(1) for model 2. (We also tested a white noise versus an MA(1), obtaining non-significant support for the latter with $T_{\widehat{V}_c} = .876$ for $h = 1$). None of the statistics is significant. This is a tricky case, since the MA(1) parameter estimate is .085 with standard error .051, so that it is not significantly different from zero. Being so close to a white noise model, the statistics have a hard time distinguishing it from other models that contain white noise as a special case.

Next, Series C (226 observations) has an ARIMA(1,1,0) and an ARIMA(0,2,2) as the recommended models. We make the stationary component comparisons as above. For $d = 1$, the MA(1) model is solidly rejected in favor of an MA(2), AR(1), or AR(2). Generally the $h = 2$ conclusions are similar to those for $h = 1$, although in some cases T_{MR} differs conspicuously from $T_{\widehat{V}_c}$. For the $d = 2$ case, we take instead an MA(2) as model 2 and obtain weak support across the board for this choice against the MA(1), ARMA(1,1), AR(1), and AR(2) models. Again, there are some conspicuous differences between $T_{\widehat{V}_c}$ and T_{MR}, and taking $h = 2$ instead of $h = 1$ has a sizeable impact on the statistics in each case. With $d = 2$ and $h = 2$, we have $[\xi]_0^{h-1}(B) = 1 + 2B$ (from $(1 - z)^{-2} = 1 + 2z + \cdots$). So $\delta(B)$ may be having more impact in this case than when $d = 1$ and $[\xi]_0^{h-1}(B) = 1 + B$.

Finally, Series D (78 observations) has a recommended model (Brockwell and Davis, 2002) of ARIMA(0,1,2). We compare the stationary components MA(2), ARMA(1,1), AR(1), and AR(2) as model 1 against MA(1) as model 2. There is marginal support for the MA(2) over the MA(1), and the same is true for the AR(1) and AR(2). The MA(1) is significantly preferred over the ARMA(1,1). Results are extremely close for $h = 1$ and $h = 2$. Again $T_{\widehat{V}_c}$ is consistently smaller than T_{MR} showing that $\widehat{V}_c$ is consistently larger than $\widehat{V}_{MR}$, just as happens with the $d = 1$ results for Series C.

4 Concluding Summary

For testing the null hypothesis that two stationary or difference-stationary models for a series have the same asymptotic mean square h-step forecast error, we have introduced two new test statistics, conceptualized as goodness-of-fit diagnostics in the sense of McElroy and Holan (2009). The important novelty of the statistics resides in the broadly consistent variance estimates whose square roots standardize them, which differ in whether parameter estimation effects are addressed or not. The new theoretical results of Theorems 1 and 2 show that both statisics converge in law to a standard normal distribution. The terms of the variance estimate $\widehat{V}$ that account for parameter estimation involve first and second derivatives of functions of the model spectral densities. The tedious calculations required to obtain algebraic formulas for these derivatives are given in the Appendix for the simplest case, that of moving average models (after appropriate differencing) and can doubtless also be done for AR and ARMA models. For the small set of simulated and real series results presented for moving average models, the numerical differences were mostly negligible between this variance estimate and the much more easily calculated variance estimate $\widehat{V}_c$ that treats the model parameters as constant.

For the constant parameter case, our results support the use of $\widehat{V}_c$ in preference to the usually inconsistent variance estimate $\widehat{V}_{MR}$ recommended by Meese and Rogoff (1988) and Diebold and Mariano (1995). In our small empirical study presented in Section 3, the values of $\widehat{V}_{MR}$ were more variable than those of $\widehat{V}_c$ and led to a different model choice in one case.

Acknowledgements. 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.

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

Appendix

A.1 Proof of Theorem 1.

For each i we have

$$\begin{aligned}
&Q_n(I^{j_i}, \hat{g}_i, \hat{\theta}^{(i)}) - j_i! Q_n(\tilde{f}^{j_i}, \tilde{g}_i, \tilde{\theta}^{(i)}) \\
&= \left(Q_n(I^{j_i}, \hat{g}_i, \hat{\theta}^{(i)}) - j_i! Q_n(\tilde{f}^{j_i}, \hat{g}_i, \hat{\theta}^{(i)}) \right) + j_i! \left(Q_n(\tilde{f}^{j_i}, \hat{g}_i, \hat{\theta}^{(i)}) - Q_n(\tilde{f}^{j_i}, \tilde{g}_i, \tilde{\theta}^{(i)}) \right).
\end{aligned}$$

The first term expands to

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

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

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

For the second term we have

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

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

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

and hence

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

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

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

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

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

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

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

with the variance matrix given by

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

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 $W = W(\tilde{\theta})$. $\square$

A.2 Implementation Details

We discuss in some detail the MA(q) model, describing how the various quantities in the test statistic (8) are computed. That is, we suppose that both models, after having been suitably differenced, are given by moving average models (of different order). As mentioned in section 2.1, the differencing operator is assumed to be correctly specified but the stationary moving average models may both be incorrect. We will consider one model at a time, which will make the notation less cumbersome. A treatment of AR and ARMA models would be similar, although some aspects related to derivative calculations would be considerably more difficult. So let $\theta' = (\theta_1, \theta_2, \cdots, \theta_q, \theta_{q+1})$, where θ_{q+1} is the innovation variance. The moving average polynomial is $\Psi(B)$, and $f_\theta(\lambda) = \Psi(e^{-i\lambda})\Psi(e^{i\lambda})\theta_{q+1}$. We need to compute the first h coefficients of $\xi(z)$ from $1/\delta(z)$, i.e., $\xi_0, \xi_1, \cdots, \xi_{h-1}$, as well as the gradient of the polynomial $\Phi^{(h)}(B) = \sum_{k=0}^{h-1} \xi_k B^k \sum_{l=0}^{h-1-k} \theta_l B^l$. First note that the kth coefficient of this polynomial is given by $\phi_k = \sum_{l=0}^{k} \theta_l \xi_{k-l}$. It is helpful to write $\xi_k = 0$ when $k < 0$. Then the derivative of $\Phi^{(h)}(B)$ with respect to θ_j (for $1 \le j \le q$) is given by $\sum_{k=j}^{h-1} \xi_{k-j} B^k$ if $j \le h - 1$, and zero otherwise. It is very simple to determine the weighting function g, and by (5) we can compute (8) by determining $\{\gamma_k(g)\}$. To that end, consider an ARMA process Z_t given

by $\Psi(B)Z_t = \Phi^{(h)}(B)\rho_t$, where ρ_t is standard white noise. Then the autocovariances of $\{Z_t\}$ are exactly $\{\gamma_k(g)\}$; of course QMLs or MLEs should be substituted for all parameter values.

Computing the asymptotic variance under H_0 is much more complicated. Below we use the notation $\zeta^\sharp = \zeta + \overline{\zeta}$ for any complex number ζ, with $\overline{\zeta}$ the complex conjugate of ζ. Beginning with the vector $b_{\theta^{(k)},k}$, it is consistently estimated by substituting I for $\tilde{f}$, and furthermore is also a consistent estimate when evaluated at QMLs or MLEs. Let $d(\lambda) = |\Phi^{(h)}(e^{-i\lambda})|^2\theta_{q+1}$, and $g(\lambda) = d(\lambda)/f(\lambda)$. Then the needed derivatives of g are obtained from

$$\frac{\partial}{\partial\theta_j^{(k)}} d_{\theta^{(k)}}(\lambda) = \begin{cases} \left(\sum_{k=j}^{h-1} \xi_{k-j} e^{-i\lambda k} e^{-i\lambda j}\Phi^{(h)}(e^{i\lambda})\right)^\sharp \theta_{q_k+1}^{(k)} & j \le h-1 \\ 0 \quad h \le j \le q_k \\ |\Phi^{(h)}(e^{-i\lambda})|^2 \quad j = q_k+1 \end{cases}$$

$$\frac{\partial}{\partial\theta_j^{(k)}} f_{\theta^{(k)}}(\lambda) = \begin{cases} \left(e^{-i\lambda j}\Psi(e^{i\lambda})\right)^\sharp \theta_{q_k+1}^{(k)} & j \le q_k \\ |\Psi(e^{-i\lambda})|^2 \quad j = q_k+1 \end{cases}$$

together with $\nabla_{\theta^{(k)}} g_{\theta^{(k)}} = f_{\theta^{(k)}}^{-1}\nabla_{\theta^{(k)}} d_{\theta^{(k)}} - f_{\theta^{(k)}}^{-2} d_{\theta^{(k)}}\nabla_{\theta^{(k)}} f_{\theta^{(k)}}$. Note that $Q(I, \nabla_{\theta^{(k)}} g_{\theta^{(k)}}, \theta^{(k)})$ is computed via $n^{-1}W'\Gamma(\nabla_{\theta^{(k)}} g_{\theta^{(k)}})W$ as in (5), which entails computing inverse Fourier Transforms of $\nabla_{\theta^{(k)}} g_{\theta^{(k)}}$. Note that when $h > q_k$ we have $\nabla_{\theta^{(k)}} g_{\theta^{(k)},k} \equiv 0$. If instead $h \le q_k$ then

$$\frac{\partial}{\partial\theta_j^{(k)}} g_{\theta^{(k)}}(\lambda) = |\Psi(e^{-i\lambda})|^{-2} \cdot \begin{cases} \left(\sum_{k=j}^{h-1} \xi_{k-j} e^{-i\lambda k} e^{-i\lambda j}\Phi^{(h)}(e^{i\lambda})\right)^\sharp \theta_{q_k+1}^{(k)} \\ \quad -|\Phi^{(h)}(e^{-i\lambda})|^2 |\Psi(e^{-i\lambda})|^{-2}\left(e^{-i\lambda j}\Psi(e^{i\lambda})\right)^\sharp \qquad j \le h-1 \\ -|\Phi^{(h)}(e^{-i\lambda})|^2 |\Psi(e^{-i\lambda})|^{-2}\left(e^{-i\lambda j}\Psi(e^{i\lambda})\right)^\sharp \quad h \le j \le q_k \\ 0 \quad j = q_k+1 \end{cases}$$

Since we need all the inverse Fourier Transforms (FTs) corresponding to this function, we have for integer t and $j \ne q_k + 1$

$$\gamma_t\left(\frac{\partial}{\partial\theta_j^{(k)}} g_{\theta^{(k)}}\right) = \sum_{k=0}^{h-1}\xi_{k-j}\sum_{l=0}^{h-1}\theta_l^{(k)}\left(\gamma_{t-k+l}\left(|\Psi(e^{-i\cdot})|^{-2}\right) + \gamma_{t+k-l}\left(|\Psi(e^{-i\cdot})|^{-2}\right)\right)1_{\{j\le h-1\}}$$
$$-\sum_{l=0}^{q}\theta_l^{(k)}\left(\gamma_{t+l-j}\left(\frac{|\Phi^{(h)}(e^{-i\cdot})|^2}{|\Psi(e^{-i\cdot})|^4}\right) + \gamma_{t-l+j}\left(\frac{|\Phi^{(h)}(e^{-i\cdot})|^2}{|\Psi(e^{-i\cdot})|^4}\right)\right).$$

In this way, the resulting estimate $\hat{b}_{\theta^{(k)},k}$ is computed, and we have $\hat{b}_{\hat{\theta}^{(k)},k} \xrightarrow{P} b_{\tilde{\theta}^{(k)},k}$.

Next we consider the Hessian of the KB discrepancy, which under H_0 is not the same as the

Fisher information matrix, unfortunately:

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

For the moving average model, the mixed partial derivatives are

$$\frac{\partial^2 f_{\theta^{(k)}}(\lambda)}{\partial\theta_i^{(k)}\theta_j^{(k)}} = \begin{cases} \left(e^{i\lambda(i-j)}+e^{-i\lambda(i-j)}\right)\theta_{q+1}^{(k)} & i,j\neq q_k+1 \\ e^{-i\lambda j}\Psi(e^{i\lambda})+e^{i\lambda j}\Psi(e^{-i\lambda}) & i=q_k+1, j\neq i \\ 0 \quad i=j=q_k+1. \end{cases}$$

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

$$f_{\theta^{(k)}}^{-1}(\lambda)\frac{\partial^2 f_{\theta^{(k)}}(\lambda)}{\partial\theta_i^{(k)}\theta_j^{(k)}} - f_{\theta^{(k)}}^{-2}(\lambda)\frac{\partial f_{\theta^{(k)}}(\lambda)}{\partial\theta_i^{(k)}}\frac{\partial f_{\theta^{(k)}}(\lambda)}{\partial\theta_j^{(k)}} = -\begin{cases} \frac{e^{i\lambda(i+j)}}{\Psi^2(e^{i\lambda})}+\frac{e^{-i\lambda(i+j)}}{\Psi^2(e^{-i\lambda})} & i,j\neq q_k+1 \\ 0 \quad i \text{ or } j = q_k+1, \text{ but } i\neq j \\ (\theta_{q_k+1}^{(k)})^{-2} & i=j=q_k+1, \end{cases}$$

which integrates to zero unless $i=j=q_k+1$. Now for the terms involving $\tilde{f}$,

$$\begin{aligned} &2f_{\theta^{(k)}}^{-3}(\lambda)\frac{\partial f_{\theta^{(k)}}(\lambda)}{\partial\theta_i^{(k)}}\frac{\partial f_{\theta^{(k)}}(\lambda)}{\partial\theta_j^{(k)}} - f_{\theta^{(k)}}^{-2}(\lambda)\frac{\partial^2 f_{\theta^{(k)}}(\lambda)}{\partial\theta_i^{(k)}\theta_j^{(k)}} \\ &= f_{\theta^{(k)}}^{-1}(\lambda)\begin{cases} 2\frac{e^{i\lambda(i+j)}}{\Psi^2(e^{i\lambda})}+2\frac{e^{-i\lambda(i+j)}}{\Psi^2(e^{-i\lambda})}+\frac{e^{i\lambda(i-j)}+e^{-i\lambda(i-j)}}{\Psi(e^{-i\lambda})\Psi(e^{i\lambda})} & i,j\neq q_k+1 \\ \frac{e^{-i\lambda j}}{\Psi(e^{-i\lambda})\theta_{q_k+1}^{(k)}}+\frac{e^{i\lambda j}}{\Psi(e^{i\lambda})\theta_{q_k+1}^{(k)}} & i=q_k+1, j\neq i \\ \frac{e^{-i\lambda i}}{\Psi(e^{-i\lambda})\theta_{q_k+1}^{(k)}}+\frac{e^{i\lambda i}}{\Psi(e^{i\lambda})\theta_{q_k+1}^{(k)}} & j=q_k+1, i\neq j \\ 2[\theta_{q_k+1}^{(k)}]^{-2} & i=j=q_k+1. \end{cases} \\ &= \begin{cases} \frac{2e^{i\lambda(i+j)}\Psi^2(e^{-i\lambda})+2e^{-i\lambda(i+j)}\Psi^2(e^{i\lambda})}{|\Psi(e^{-i\lambda})|^6\theta_{q_k+1}^{(k)}}+\frac{e^{i\lambda(i-j)}+e^{-i\lambda(i-j)}}{|\Psi(e^{-i\lambda})|^4\theta_{q_k+1}^{(k)}} & i,j\neq q_k+1 \\ \frac{e^{i\lambda j}\Psi(e^{-i\lambda})+e^{-i\lambda j}\Psi(e^{i\lambda})}{|\Psi(e^{-i\lambda})|^4\theta_{q_k+1}^{(k)}} & i=q_k+1, j\neq i \\ \frac{e^{i\lambda i}\Psi(e^{-i\lambda})+e^{-i\lambda i}\Psi(e^{i\lambda})}{|\Psi(e^{-i\lambda})|^4\theta_{q_k+1}^{(k)}} & j=q_k+1, i\neq j \\ 2[\theta_{q+1}^{(k)}]^{-2}f_{\theta^{(k)}}^{-1}(\lambda) & i=j=q_k+1. \end{cases} \end{aligned}$$

We need the inverse Fourier Transforms of this expression. Letting $\Psi^2(B)=\sum_{l=0}^{2q_k}\kappa_l B^l$, the tth

inverse FT (for $t \geq 0$) of the above is then

$$\begin{cases} 2[\theta_{q+1}^{(k)}]^2 \sum_{l=0}^{2q} \kappa_l \left(\gamma_{t+i+j-l}(f_{\theta^{(k)}}^{-3}) + \gamma_{t-i-j+l}(f_{\theta^{(k)}}^{-3}) \right) \\ \qquad +\theta_{q+1}^{(k)} \left(\gamma_{t+i-j}(f_{\theta^{(k)}}^{-2}) + \gamma_{t-i+j}(f_{\theta^{(k)}}^{-2}) \right) \quad i,j \neq q_k+1 \\ \theta_{q+1}^{(k)} \sum_{l=0}^{q} \theta_l^{(k)} \left(\gamma_{t-j+l}(f_{\theta^{(k)}}^{-2}) + \gamma_{t+j-l}(f_{\theta^{(k)}}^{-2}) \right) \quad i = q_k+1, j \neq i \\ \theta_{q+1}^{(k)} \sum_{l=0}^{q} \theta_l^{(k)} \left(\gamma_{t-i+l}(f_{\theta^{(k)}}^{-2}) + \gamma_{t+i-l}(f_{\theta^{(k)}}^{-2}) \right) \quad j = q_k+1, i \neq j \\ 2[\theta_{q+1}^{(k)}]^{-2} \gamma_t(f_{\theta^{(k)}}^{-1}(\lambda)) \quad i = j = q_k+1. \end{cases}$$

These values are then used to fill out the Toeplitz matrix $\Gamma \left(2f_{\theta^{(k)}}^{-3} \frac{\partial f_{\theta^{(k)}}}{\partial \theta_i^{(k)}} \frac{\partial f_{\theta^{(k)}}}{\partial \theta_j^{(k)}} - f_{\theta^{(k)}}^{-2} \frac{\partial^2 f_{\theta^{(k)}}}{\partial \theta_i^{(k)} \theta_j^{(k)}} \right)$, with a separate matrix for each choice of i and j. Then in order to estimate the ijth entry of $M_f(\theta^{(k)})$, we substitute I for $\tilde{f}$, so that

$$[\widehat{M_f(\theta^{(k)})}]_{ij} = \frac{1}{n} W' \Gamma \left(2f_{\theta^{(k)}}^{-3} \frac{\partial f_{\theta^{(k)}}}{\partial \theta_i^{(k)}} \frac{\partial f_{\theta^{(k)}}}{\partial \theta_j^{(k)}} - f_{\theta^{(k)}}^{-2} \frac{\partial^2 f_{\theta^{(k)}}}{\partial \theta_i^{(k)} \theta_j^{(k)}} \right) W - 1_{\{i=j=q_k+1\}} [\theta_{q+1}^{(k)}]^{-2}.$$

This is a consistent estimate of the Hessian when QMLs or MLEs are substituted for the parameters. Noting that $R'\Gamma(v)R = (2\pi)^{-1} \int_{-\pi}^{\pi} v(\lambda) I^2(\lambda)\, d\lambda$ for any bounded $v(\lambda)$ and R the $2n-1$ vector of sample autocovariances (i.e., $R_j = n^{-1} \sum_{t=1}^{n-|j-n|} W_t W_{t+|j-n|}$), it follows that

$$\begin{aligned} \widehat{W}_{kl}(\theta) &= R' \Gamma \left(g_{\theta^{(k)}} g_{\theta^{(l)}} \right) R \\ &\quad - \widehat{b}'_{\theta^{(l)},l} \widehat{M_f(\theta^{(l)})}^{-1} \left\{ R' \Gamma \left(g_{\theta^{(k)}} f_{\theta^{(l)}}^{-2} \frac{\partial f_{\theta^{(l)}}}{\partial \theta_j^{(l)}} \right) R \right\}_{j=1}^{q_l+1} \\ &\quad - \widehat{b}'_{\theta^{(k)},k} \widehat{M_f(\theta^{(k)})}^{-1} \left\{ R' \Gamma \left(g_{\theta^{(l)}} f_{\theta^{(k)}}^{-2} \frac{\partial f_{\theta^{(k)}}}{\partial \theta_j^{(k)}} \right) R \right\}_{j=1}^{q_k+1} \\ &\quad + \widehat{b}'_{\theta^{(l)},l} \widehat{M_f(\theta^{(l)})}^{-1} \left\{ R' \Gamma \left(f_{\theta^{(l)}}^{-2} f_{\theta^{(k)}}^{-2} \frac{\partial f_{\theta^{(l)}}}{\partial \theta_i^{(l)}} \frac{\partial f_{\theta^{(k)}}}{\partial \theta_j^{(k)}} \right) R \right\}_{i,j=1}^{q_l+1,q_k+1} \widehat{M_f(\theta^{(k)})}^{-1} \widehat{b}_{\theta^{(k)},k}. \end{aligned}$$

Here q_k and q_l are the MA orders for models k and l, where $k, l = 1, 2$. Using the above derivations and letting $\Phi^{(h),(k)}$ denote the $\Phi^{(h)}$ polynomial pertaining to model k for $k = 1, 2$, we have

$$\frac{g_{\theta^{(k)}}(\lambda)}{f_{\theta^{(l)}}^2(\lambda)} \frac{\partial}{\partial \theta_j^{(l)}} f_{\theta^{(l)}}(\lambda) = \frac{|\Phi^{(h),(k)}(e^{-i\lambda})|^2}{|\Psi^{(k)}(e^{-i\lambda})|^2 f_{\theta^{(l)}}^2(e^{-i\lambda})} \cdot \begin{cases} \left(e^{-i\lambda j} \Psi^{(l)}(e^{i\lambda}) \right)^{\sharp} \theta_{q_l+1}^{(l)} & 1 \leq j \leq q_l \\ |\Psi^{(l)}(e^{-i\lambda})|^2 & j = q_l+1. \end{cases}$$

Letting $v^{(k),(l)}$ denote the function $\frac{|\Psi^{(h),(k)}(e^{-i\cdot})|^2}{|\Psi^{(k)}(e^{-i\cdot})|^2 |\Psi^{(l)}(e^{-i\cdot})|^4}$, we have the tth inverse FT is given by

$$\begin{cases} [\theta_{q_l+1}^{(l)}]^{-1} \sum_{m=0}^{q_l} \theta_m^{(l)} \left(\gamma_{t+m-j} \left(v^{(k),(l)} \right) + \gamma_{t-m+j} \left(v^{(k),(l)} \right) \right) & 1 \leq j \leq q_l \\ [\theta_{q_l+1}^{(l)}]^{-2} \gamma_t \left(v^{(k),(l)} |\Psi^{(l)}(e^{-i\cdot})|^2 \right) & j = q_l+1. \end{cases}$$

Finally we have

$$
\begin{aligned}
&f_{\theta^{(l)}}^{-2}(\lambda) f_{\theta^{(k)}}^{-2}(\lambda) \frac{\partial f_{\theta^{(l)}}(\lambda)}{\partial \theta_i^{(l)}} \frac{\partial f_{\theta^{(k)}}(\lambda)}{\partial \theta_j^{(k)}} \\
&= \begin{cases}
\frac{\theta_{q+1}^{(l)} \theta_{q+1}^{(k)}}{f_{\theta^{(l)}}^2(\lambda) f_{\theta^{(k)}}^2(\lambda)} \left(e^{-i\lambda i} \Psi^{(l)}(e^{i\lambda})\right)^{\sharp} \left(e^{-i\lambda j} \Psi^{(k)}(e^{i\lambda})\right)^{\sharp} & i \leq q_l,\, j \leq q_k \\
\frac{\theta_{q_k+1}^{(k)}}{\theta_{q_l+1}^{(l)} f_{\theta^{(k)}}^2(\lambda) f_{\theta^{(l)}}(\lambda)} \left(e^{-i\lambda j} \Psi^{(k)}(e^{i\lambda})\right)^{\sharp} & j \leq q_k,\, i = q_l + 1 \\
\frac{\theta_{q_l+1}^{(l)}}{\theta_{q_k+1}^{(k)} f_{\theta^{(l)}}^2(\lambda) f_{\theta^{(k)}}(\lambda)} \left(e^{-i\lambda i} \Psi^{(l)}(e^{i\lambda})\right)^{\sharp} & i \leq q_l,\, j = q_k + 1 \\
\frac{1}{f_{\theta^{(l)}}(\lambda) f_{\theta^{(k)}}(\lambda) \theta_{q+1}^{(l)} \theta_{q+1}^{(k)}} & i = q_l + 1,\, j = q_k + 1,
\end{cases}
\end{aligned}
$$

from which the tth inverse FTs are obtained:

$$
\begin{cases}
\theta_{q+1}^{(l)} \theta_{q+1}^{(k)} \sum_{m=0}^{q_l} \sum_{p=0}^{q_k} \theta_m^{(l)} \theta_p^{(k)} [\gamma_{t-i-j+m+p}(f_{\theta^{(l)}}^{-2} f_{\theta^{(k)}}^{-2}) + \gamma_{t+i+j-m-p}(f_{\theta^{(l)}}^{-2} f_{\theta^{(k)}}^{-2}) \\
\qquad + \gamma_{t-i+j+m-p}(f_{\theta^{(l)}}^{-2} f_{\theta^{(k)}}^{-2}) + \gamma_{t+i-j-m+p}(f_{\theta^{(l)}}^{-2} f_{\theta^{(k)}}^{-2})] \quad i \leq q_l,\, j \leq q_k \\
\frac{\theta_{q_k+1}^{(k)}}{\theta_{q_l+1}^{(l)}} \sum_{m=0}^{q_k} \theta_m^{(k)} \left(\gamma_{t-j+m}(f_{\theta^{(k)}}^{-2} f_{\theta^{(l)}}^{-1}) + \gamma_{t+j-m}(f_{\theta^{(k)}}^{-2} f_{\theta^{(l)}}^{-1})\right) \quad j \leq q_k,\, i = q_l + 1 \\
\frac{\theta_{q_l+1}^{(l)}}{\theta_{q_k+1}^{(k)}} \sum_{m=0}^{q_l} \theta_m^{(l)} \left(\gamma_{t-i+m}(f_{\theta^{(l)}}^{-2} f_{\theta^{(k)}}^{-1}) + \gamma_{t+i-m}(f_{\theta^{(l)}}^{-2} f_{\theta^{(k)}}^{-1})\right) \quad i \leq q_l,\, j = q_k + 1 \\
\frac{\gamma_t(f_{\theta^{(l)}}^{-1} f_{\theta^{(k)}}^{-1})}{\theta_{q+1}^{(l)} \theta_{q+1}^{(k)}} \quad i = q_l + 1,\, j = q_k + 1.
\end{cases}
$$

From the preceding discussion, $\widehat{W}_{kl}(\widehat{\theta}) \xrightarrow{P} W_{kl}(\widetilde{\theta})$, and therefore is used to normalize the diagnostic given in (8).

A.3 Proof of Theorem 2

The asymptotic normality follows from Theorem 1, and by the discussion preceding Theorem 2. The consistency of $\widehat{V}$ follows from conditions 2, 3, 4, 5, and 6 (as well as condition 7 if we are considering MLEs instead of QMLs), together with Lemma 3.1.1 of Taniguchi and Kakizawa (2000). □

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

Applying Parseval's identity, (14) 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 $\varepsilon_t^{(h)}\left(\theta^{(j)}\right)$, $j = 1, 2$:

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

where (suppressing the λ argument)

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

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 $V_{MR} = V_c$ follows.

References

Bell, W., 1984. Signal extraction for nonstationary time series. Ann. Statist. 12, 646 –664.

Box, G., Jenkins, G., Reinsel, G., 1994. Time Series Analysis: Forecasting and Control, 3rd Edition. 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.

Dahlhaus, R., Wefelmeyer, W., 1996. Asymptotically optimal estimation in misspecified time series models. Ann. Statist. 16, 952–974.

Diebold, F., Mariano, R., 1995. Comparing predictive accuracy. Journal of Business and Economics 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., 1991a. Convergence of finite multistep predictors from incorrect models and its role in model selection. Note di Matematica XI, 145–155.

Findley, D., 1991b. Counterexamples to Parsimony and BIC. Ann. Inst. Statist. Math. 43, 505-514.

Findley, D., Pötscher, B., Wei, C., 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.

Pourahmadi, M., 2001. Foundations of Time Series Analysis and Prediction Theory. Wiley, New York.

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.

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

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

	DGP MA Parameter Values		
n	.1	.3	.6
100	4.6, 1.9	40.3, 37.5	97.7, 98.4
200	5.0, 3.2	83.2, 82.5	100.0, 100.0
500	16.3, 14.8	99.9, 99.9	100.0, 100.0

Table 1: Power for various DGPs in percentages, grouped by sample size and true MA parameter value. In each cell are two power results: on the left those for the test statistic $T_{\widehat{V}}$ (11), on the right those for $T_{\widehat{V}_c}$ (13). The power results are null hypothesis rejection rates for the upper one-sided test at α level 5%. They show no practical differences between the two statistics. As expected, there is little power at parameter value .1.

	Series			
Stationary Components	A	B	C	D
WN vs MA(1)	3.371, 3.390	.890, .876	3.276, 3.298	1.698, 1.623

Table 2: Test statistics of white noise as model 1 against an MA(1) as model 2, using the test statistics $T_{\widehat{V}}$ (11) and $T_{\widehat{V}_c}$ (13) on the left and right of the comma, respectively. For all four series one differencing ($d = 1$) is used and $h = 1$. Only for series D would the difference in values lead to different decisions, with WN rejected only by $T_{\widehat{V}}$ with a one-sided test at at α level 5%

	Alternative Stationary Components vs Model 2's			
Series ($d = 1$)	MA(2)	ARMA(1,1)	AR(1)	AR(2)
A, h = 1	-.999, -1.099	3.605, 3.151	2.526, 1.832	2.520, 1.690
A, h = 2	-1.225, -1.037	3.841, 3.290	3.397, 2.855	2.867, 2.563
B, h = 1	-.097, -.094	-.091, -.088	-.014, -.014	-.341, -.355
B, h = 2	.295, .284	.300, .290	.382, .372	-.724, -.780
C, h = 1	-2.844, -4.129	1.840, 3.587	-2.782, -5.056	-2.786, -5.072
C, h = 2	-2.737, -4.168	-.303, -1.145	-2.381, -3.683	-2.386, -3.694
D, h = 1	-1.205, -1.629	1.925, 2.752	-1.262, -1.842	-1.183, -1.693
D, h = 2	-1.123, -1.409	1.922, 2.568	-1.200, -1.577	-1.082, -1.434
Series ($d = 2$)	MA(1)	ARMA(1,1)	AR(1)	AR(2)
C, h = 1	.694, .610	2.796, 1.882	.729, .623	.758, .650
C, h = 2	1.253, 1.172	1.859, 1.657	1.286, 1.196	1.153, 1.081

Table 3: Test statistics of the indicated model (as model 1) against an MA(1) (as model 2), using $T_{\widehat{V}_c}$ and T_{MR} on the left and right of the comma, respectively. For all four series one differencing ($d = 1$) is utilized, and for series C also two differencings ($d = 2$), in which case the indicated models are tested against an MA(2) as model 2. The statistic T_{MR} is more variable, and it indicates signicant superiority for the two AR models for series D when $h = 1$, which $T_{\widehat{V}_c}$ does not.