A test for improved multi-step forecasting

John Haywood* and Granville Tunnicliffe Wilson†

October 2, 2008

Abstract

We propose a general test of whether a time series model, with parameters estimated by minimising the single-step forecast error sum of squares, is robust with respect to multi-step prediction, for some specified lead-time. The test may be applied to a, possibly seasonal, ARIMA model using the parameters and residuals following maximum likelihood estimation. It is based on a score statistic, evaluated at these estimated parameters, that measures the sensitivity of the multi-step forecast error variance with respect to the parameters. We derive the large sample properties of the test and show by a simulation study that it has acceptable small sample size properties for higher lead times when applied to the IMA(1,1) model that gives rise to the exponentially weighted moving average predictor. We investigate the power of the test when the IMA(1,1) model has been fitted to an ARMA(1,1) process. Further, we demonstrate the high power of the test when an autoregression is fitted to a process generated as the sum of a stochastic trend and cycle plus noise. We use frequency domain methods for the derivation and sampling properties of the test, and to give insight into its application. The test is illustrated on two real series, and an R function for its general application is available on the website of the first author.

Keywords: Diagnostic statistic; Model robustness; Multi-step prediction; Time series

1 Introduction

In this paper we propose a test of whether a (possibly seasonal) ARIMA model with parameters estimated by maximum likelihood, is robust with respect to multi-step prediction, for some specified lead-time. The test may be carried out using the model residuals and estimated parameters. If the result is significant, the test statistic provides an estimate of the reduction in multi-step forecast error variance, at the specified lead time, that might

*School of Mathematics, Statistics and Operations Research, Victoria University of Wellington, PO Box 600, Wellington 6140, NZ.

†Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK.

be achieved by re-fitting the model parameters so as to minimise the sum of squares of multi-step errors. The test is easily extended to provide these re-fitted parameters, using the method of Haywood and Tunnicliffe Wilson (1997), hereafter referred to as HTW.

We propose this test as a practical solution to a problem that is a current topic of debate; of deciding the relative merits of two distinct approaches to multi-step forecasting. The first approach is to fit a parametric process model by minimizing the sum of squares of in-sample single-step forecast errors, or equivalently by maximum likelihood estimation. Multi-step forecasts, from the end of the series, are then derived from this model. For a linear model the expected future value may be obtained quite simply, by successively using the single step predictor, treating each prediction as if it were an observed value. The result is often termed the *plug-in* or *iterated* multi-step (IMS) prediction. The alternative approach is to construct a *direct* multi-step (DMS) prediction as a specified function of the observations, and to choose the coefficients in this predictor by minimising the sum of squares of the multi-step forecast errors.

Bhansali (1999) reviews many of the earlier contributions to the analysis of this debate, setting out the problem, and presenting the statistical issues. Chevillon (2007) provides an exhaustive overview of the current state of this research. The most common DMS predictor, termed a *non-parametric* predictor, is of the autoregressive type, i.e. a linear combination of a finite set of recent process observations. However, every parametric process model gives rise to an IMS predictor, for example the widely used multi-step exponentially weighted moving average (EWMA) predictor may be viewed as the predictor from an IMA(1,1) model. A parametric DMS predictor can therefore be constructed by using the IMS predictor of a process model *but* with the parameters determined to minimize the sum of squares of in-sample *multi-step*, rather than single step, forecast errors. Bhansali (1999) gave an example of such a predictor (see also Stoica and Soderstrom, 1984), and HTW show how they can be constructed and their properties evaluated for a wide range of parametric models.

The advantage of the IMS prediction is that one model suffices for all lead-times, and whether one is forecasting a future level, trend or cumulative total. It is also more efficient in the sense of mean square forecast error, provided the process model is correctly specified, although the efficiency of DMS predictions can be restored if generalized method of moments (GMM) is used in place of OLS for autoregressive predictors (Chevillon and Hendry, 2005). The advantage of DMS prediction, as a robust method, is clear when IMS prediction is

applied using a mis-specified model, as discussed by Findley (1983). However, we envisage that our test is used mainly when any such mis-specification is subtle, because care has been taken to identify a plausible model. See Bhansali (1996) for discussion of the advantages of DMS for autoregressive models, Kang (2003) for a recent practical investigation, and Ing (2003), for a comparative treatment of the asymptotics for stationary autoregressive forms of IMS and DMS predictors.

It is good practice in time series modelling to guard against model mis-specification that might disadvantage the IMS predictor, by following maximum likelihood estimation of a carefully selected model with diagnostic checking applied to the estimated residuals. However, model selection is still an uncertain procedure and portmanteau diagnostics may give no clear indication that a subtly mis-specified model could be improved in any particular way. Yet, as our examples illustrate, in such circumstances substantial reduction in forecast variance may still be possible, at some high lead time. For these reasons, considerable attention continues to be devoted to resolving the problem of choice between these two types of predictors. In a recent applied contribution, Marcellino, Stock and Watson (2006) present the results of a large scale exercise to compare the two approaches using out of sample assessment of the forecasts. They only considered autoregressive predictors for both IMS and DMS methods, and examined cases where the number of terms in the predictor was fixed, or chosen by an information criterion. Two of the points among their conclusions are, that the IMS method is preferred for a large class of about 80% of their series, and that the DMS is preferred in the other cases when a low order model has been used. However, "most if not all of the advantage" of the DMS is eliminated if the number of terms in the model is increased. They indicate that a model including a moving average term might be better in these cases.

The test we have proposed provides a practical approach to this problem. As with the portmanteau statistics, it may be applied immediately following maximum likelihood estimation. However, the test is valuable to the user because of its more specific nature. If the result of the test is not significant, for a given lead time, it provides re-assurance of the adequacy of the model parameters for IMS prediction at that lead time. If the result is significant, the user can be sure that an immediate gain in forecast accuracy can be achieved, at that lead time, by re-estimation of the model parameters for DMS prediction. We do not propose it as a general test of model adequacy; indeed, it is specific in what it tests, and there is no guarantee, were it applied for some arbitrary lead time, that it would detect general

inadequacy of a mis-specified model. However, if applied in the circumstances we describe, following careful model selection, estimation and checking, it can be a valuable new tool.

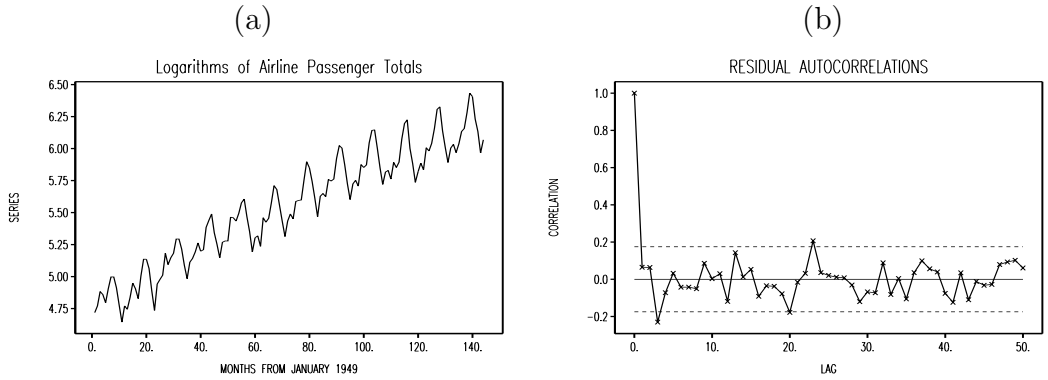

Figure 1: (a) The logarithm of the series of monthly airline passenger totals and (b) the sample residual autocorrelations following estimation of a seasonal structural model.

To illustrate the test we here present two real examples, the first concerning a seasonal moving average type of model; the second a non-seasonal autoregression. Consider first the familiar monthly "Airline" series of Box and Jenkins (1976). The logarithm of the series is shown in Figure 1 (a) and the sample residual autocorrelations shown in Figure 1 (b), following (approximate maximum likelihood) estimation of the structural model with seasonal component described by HTW (p.249). See also Harvey (1989). There is some evidence of model inadequacy in that the sample autocorrelations at lags 3 and 23 lie outside the nominal 95% limits. The Ljung-Box portmanteau statistic (Ljung and Box, 1978) based on residual autocorrelations up to lag 25, was 31.6 on 22 d.f., corresponding to a p-value of 8.4%. Applying our test for lead time 6 gave a highly significant result, with a p-value of 0.3%. From the value of the test statistic we also estimate a potential reduction in forecast error variance of 12% at that lead time, that may be achieved by re-estimation of the model parameters.

For our second example we take the series of quarterly seasonally adjusted USA unemployment rate for the period 1948 to 1979, shown in Figure 2, together with its sample partial autocorrelation function, calculated following a square root transformation (a Box-Cox parameter of 0.48 was estimated by maximum likelihood). A third order autoregressive model was selected for this series using the AIC (Akaike, 1973). The residual autocorrelation and spectrum are shown in Figure 3. There is no obvious sign of model inadequacy, though the Ljung-Box portmanteau test statistic for the first 20 residual autocorrelations is 25.98 on 17

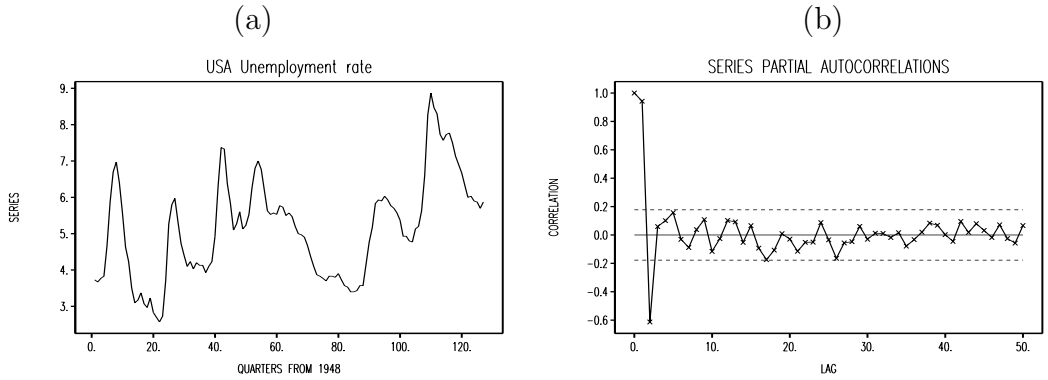

Figure 2: (a) The quarterly seasonally adjusted USA unemployment rate from 1948 to 1979, together with (b) the sample partial autocorrelations of the square root transformed series.

d.f. with a p-value of 7.5%, and that using the first 30 values is 39.97 on 27 d.f. with p-value of 5.1%. On applying our test with a lead time of 30 quarters ($7\frac{1}{2}$ years), we obtained a p-value of 0.4%, and the test statistic indicated a potential reduction of 30% in the multistep forecast error variance. Fitting the AR(3) DMS predictor actually gave a reduction of 43%, but the statistical significance of this can not be directly assessed, because the regression errors are highly correlated.

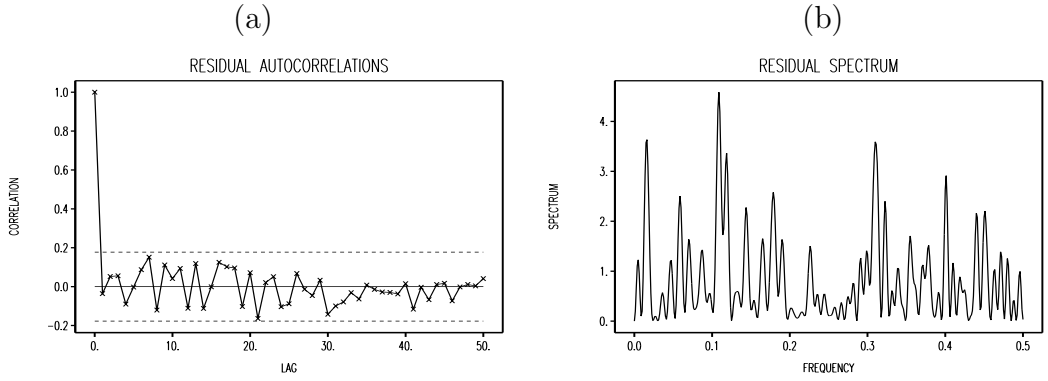

Figure 3: (a) The residual sample autocorrelations and (b) the sample spectrum of the residuals, from fitting an AR(3) model to the transformed USA unemployment rate series.

In Section 2 we review and extend the methodology and notation introduced by HTW, concerning the use of a frequency domain multi-step criterion for model estimation. Our test statistic can be viewed as a natural precursor to such model estimation and it is valuable because of its known sampling properties, which we derive in Section 3 of this paper. In Section 4 we carry out an empirical study to verify its small sample size when applied to the

IMA(1,1) model, and evaluate its power to reject the IMA(1,1) model when it has been fitted to an ARMA(1,1) process. This is motivated by an original investigation by Cox (1961) into the robustness of the EWMA for multi-step prediction, which was more recently taken up by Tiao and Xu (1993). They carried out a test for a significant change in the moving average parameter when the model was re-estimated to minimise the sum of squares of multi-step forecast errors. They show that their test can have greater power than the Ljung-Box test in this context. However, the parameter in this case is non-linear, and their test encounters a practical limitation when applied to lead times greater than two, because of the tendency of the multi-step estimate to take the unit value on the boundary of the parameter space. Our test statistic is well defined for higher lead times and we are able to illustrate the effects on the power, of the choice of lead time. The investigation in Section 5 is motivated by the foregoing example of the USA unemployment series and the use of autoregressive predictors. We postulate a type of model mis-specification that might explain the significant result in that example. Based on this we design a simulation exercise to confirm the size properties of our test in the context of an autoregressive model, and its power when an autoregression has been fitted to a series that is the sum of irregular and cyclical components.

2 A frequency domain multi-step criterion

Model estimation using a frequency domain multi-step criterion was the subject of HTW. Frequency domain methods for the estimation of time series models have received renewed interest in recent years, e.g. in the estimation of structural models, Harvey (1989, Section 4.3), and parameters of long-memory processes, Beran (1994, ch.6). See also Haywood and Tunnicliffe Wilson (2000a,b). These methods are based on the sample spectrum of the observations, using what is termed the Whittle likelihood. An advantage of spectrum based estimation is that it can provide greater insight into the statistical properties than time domain methods, as HTW shows for the multi-step estimation problem. This is because of the asymptotic independence of the sample spectrum ordinates, which we exploit in this paper to derive the properties of our test statistic. Nevertheless, the test is straightforward to implement following standard time domain estimation of an ARIMA model.

Our test statistic was proposed in the discussion of HTW, and is evaluated following estimation of the model parameters by minimizing the single-step forecast error variance, equivalent in large samples to using maximum likelihood. Our null hypothesis is that a

correctly specified model has been fitted, and we show in this paper that the size properties of the statistic are reliable, even at higher lead times, under this hypothesis. The statistic may also be expressed as a low rank quadratic form in the usual residual sample autocorrelations, upon which the familiar portmanteau statistics of Box and Pierce (1970) and Ljung and Box (1978) are also based. It detects model dependent features that characterize specific departures from white noise residuals. The statistic is formed from a local gradient (or score) and Hessian matrix, so as to approximate the reduction in the multi-step error variance that would result from the re-estimation of the model to minimize that criterion. We show that its asymptotic distribution, under the null hypothesis, is a weighted sum of independent chi-squared variables on one degree of freedom. The models that we use for illustration in this paper are equivalent to moving average and autoregressive models, though differently parameterized. We explain in Appendix C how the test is implemented for conventionally parameterized ARIMA models.

We here review the frequency domain model estimation and the multi-step criterion, extending some of the methodology and notation introduced by HTW. We assume that the observed time series under consideration, x_t, may be modelled as an (possibly) integrated process that requires differencing of order $d \geq 0$ to yield a stationary process w_t with autocovariance function γ_k. The assumption that w_t is Gaussian is sufficient for our development, but we make reference to much more general conditions in Appendix A. We also suppose that the model for w_t has a spectrum $S(f) = \gamma_0 + 2\sum_{k=1}^{\infty} \gamma_k \cos(2\pi k f)$, for $-\frac{1}{2} \leq f \leq \frac{1}{2}$, which has a linear form with coefficients $\beta_1, \ldots, \beta_k$:

$$S(f;\boldsymbol{\beta}) = \sum_{i=1}^{k} \beta_i S_i(f), \tag{1}$$

where $\boldsymbol{\beta}$ is the column vector of coefficients. Our null hypothesis is that this model specification is valid, with the orders d and k and the functions $S_i(f)$ known, but the coefficients $\boldsymbol{\beta}$ unknown. A simple example is presented later in (19). In general (1) encompasses the standard moving average model, though not with the usual parameterization, and the structural models of Harvey (1989). Our theoretical treatment will be limited to this form of model, but we explain in Appendix C how we can extend these models to reciprocals, ratios and products of such a linear form, which encompasses non-seasonal and seasonal ARIMA models. For now we assume, as is true for all the examples we have come across, that $S_i(f)$ are smooth functions of f. We further assume that w_t can be represented by a model in this class for which $S(f;\boldsymbol{\beta})$ is bounded away from zero, and restrict the parameter space to

ensure this condition.

It is important that we explain how the linear model will first be fitted to the sample spectrum, $S^*(f)$, of a finite record of $w_t, t = 1, \ldots, n$:

$$S^*(f) = \frac{1}{n} \left| \sum_{t=1}^{n} w_t \exp(i2\pi ft) \right|^2 . \tag{2}$$

Fitting is done by approximate maximum likelihood estimation of $\boldsymbol{\beta}$ using the 'Whittle likelihood', e.g. see Beran (1994, ch.6) and Taniguchi and Kakizawa (2000, sections 3.1 and 6.2):

$$\int_{f=-\frac{1}{2}}^{\frac{1}{2}} \{\log S(f;\boldsymbol{\beta}) + S^*(f)/S(f;\boldsymbol{\beta})\}\, df. \tag{3}$$

This is an approximation to $-2/n$ times the log likelihood, and is minimised to estimate $\boldsymbol{\beta}$. In practice a finite sum over the harmonic frequencies $f_j = j/n$ for $j = 0, \ldots, n-1$ is used:

$$\sum_{j=0}^{n-1} \{\log S(f_j;\boldsymbol{\beta}) + S^*(f_j)/S(f_j;\boldsymbol{\beta})\}\,, \tag{4}$$

which approximates to minus twice the log likelihood. We will use both the integral form, such as (3), of the expressions we use in our development below, and also, where appropriate, the summation form such as (4). See Appendix A for a discussion of their relationship and underlying assumptions. Expressions for the large sample results we present are naturally couched in terms of integrals. However, the summation form of the likelihood (4) admits a heuristic argument for its use. It corresponds to that obtained by treating the sample spectrum as data, with the distributional properties that, for $0 < f_j < \frac{1}{2}$, $S^*(f_j)$ are realized values of independent exponential random variables with mean $S(f_j)$. For the anomalous frequencies $f_j = 0$ and (if n is even) $f_j = \frac{1}{2}$, the distribution is gamma with shape parameter $\frac{1}{2}$, but still with expectation $S(f_j)$. Because $S^*(f_{n-j}) = S^*(f_j)$ all except these anomalous frequencies are duplicated in (4), which then conveniently results in the correct expression for minus twice the log likelihood.

Part of the derivation of the large sample properties of our test statistic, presented in Appendix B, will be based on this argument, because it enables us to treat the fitting problem as an example of Generalised Linear Modelling (GLM), McCullagh and Nelder (1989). This is pointed out, for example, by Diggle (1990, p.124) and Beran (1994, ch.6). The re-weighted least squares (RLS) algorithm then provides a practical procedure for estimation of model (1), which was extended in HTW to multi-step estimation. This algorithm is described

now because of its place in the construction of our test statistic and the development of its properties.

Let $\boldsymbol{\beta}$ denote a general parameter value, $\boldsymbol{\beta}_0$ the true value under the null hypothesis and $\hat{\boldsymbol{\beta}}$ the value that minimises (4). The model components (regressors) $S_i(f)$ are fitted to the sample spectrum (response) $S^*(f)$, after they have both been (inversely) weighted by the fitted response $S(f;\boldsymbol{\beta})$ from the previous iteration. The weighted response and regressor functions are therefore:

$$Y(f;\boldsymbol{\beta}) = S^*(f)/S(f;\boldsymbol{\beta}) \quad \text{and} \quad X_i(f;\boldsymbol{\beta}) = S_i(f)/S(f;\boldsymbol{\beta}). \tag{5}$$

Corresponding to these, by evaluation at the harmonic frequencies, we have the $n\times 1$ response vector $\boldsymbol{Y}$ and $n \times k$ regression matrix $\boldsymbol{X}$ with elements $\boldsymbol{Y}_j = Y(f_j;\boldsymbol{\beta})$ and $\boldsymbol{X}_{j,i} = X_i(f_j;\boldsymbol{\beta})$. The weighted regression step provides a new parameter as a solution to the least squares equations $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}_{\text{new}} = \boldsymbol{X}'\boldsymbol{Y}$. For the first step the regression is unweighted and thereafter a sequence $\boldsymbol{\beta}_s$, $s = 1, 2, \ldots$ is generated, with $\boldsymbol{Y}$ and $\boldsymbol{X}$ updated at each step. An iterate may have to be scaled back towards the previous value if it yields $S(f;\boldsymbol{\beta})$ with a negative value for some f, but with such a correction the sequence is found in practice to converge rapidly to the global minimum. The estimate then satisfies $\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{X}}'\hat{\boldsymbol{Y}}$, where we have emphasised, by using $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{Y}}$, their implicit dependence upon $\hat{\boldsymbol{\beta}}$. We will refer to $\hat{\boldsymbol{\beta}}$ as the GLM estimates of model (1), but as they maximize the Whittle likelihood, they are also (approximate) maximum likelihood estimates.

To define our test statistic we first review the development (as in HTW) of an expression, given in (9) below, for the mean sum of squares (MSS) of the L-step ahead prediction errors which may be realised by applying the model (1) to the given series x_t. We first express

$$S(f;\boldsymbol{\beta}) = \sigma^2|\psi\{\exp(i2\pi f)\}|^2 \tag{6}$$

so that the moving average representation of the model for w_t is

$$w_t = (1 + \psi_1 B + \psi_2 B^2 + \cdots)e_t = \psi(B)e_t, \tag{7}$$

where $e_t = \psi(B)^{-1}w_t$, with variance σ^2, is the white noise linear innovation (or one-step forecast error) process of w_t under the assumption of model (1).

Taking z to be a complex variable lying in the unit circle, expand $\Psi(z) = (1-z)^{-d}\psi(z) = 1 + \sum_{k=0}^{\infty}\Psi_k z^k$ and define $\Psi_{L-1}(z) = 1 + \sum_{k=0}^{L-1}\Psi_i z^k$. Then the error $e_t(L)$ in the multi-step

prediction of x_{t+L} made at time t, that results from applying this model, is expressed as

$$e_t(L) = e_{t+L} + \Psi_1 e_{t+L-1} + \cdots + \Psi_{L-1} e_{t+1} = \Psi_{L-1}(B) e_{t+L} = \Psi_{L-1}(B)\psi(B)^{-1} w_{t+L}. \quad (8)$$

The required mean sum of squares (MSS) of the in-sample L-step ahead prediction errors $e_t(L)$ based on model (1), then has the frequency domain approximation in terms of the sample spectrum, as

$$F_L(\boldsymbol{\beta}) = \int \frac{|\Psi_{L-1}\{\exp(i2\pi f)\}|^2}{|\psi\{\exp(i2\pi f)\}|^2} S^*(f) df = \int G_L(f;\boldsymbol{\beta}) \frac{S^*(f)}{S(f;\boldsymbol{\beta})} df, \quad (9)$$

where $G_L(f;\boldsymbol{\beta}) = \sigma^2 |\Psi_{L-1}\{\exp(i2\pi f)\}|^2$. The limits in these and all subsequent integrals are taken to be those in (3). The expression (9) was used by HTW as a criterion in place of (3) for estimation of $\boldsymbol{\beta}$ to minimise the L-step prediction error sum of squares. When $L = 1$ it furnishes the same estimates as (3).

3 A score-type test

Our test statistic $\hat{q}$ is defined in (10) below, in terms of a k-dimensional negative gradient (or score) vector $\boldsymbol{g}(\boldsymbol{\beta})$ and (expected) Hessian matrix $\boldsymbol{H}(\boldsymbol{\beta})$ of the criterion $F_L(\boldsymbol{\beta})$. The motivation is that $\frac{1}{2}\hat{q}$ is an approximation (under a locally quadratic assumption) to the (further) reduction that may be achieved in the L-step prediction criterion $F_L(\boldsymbol{\beta})$ from its value at the GLM (or maximum likelihood) estimate $\hat{\boldsymbol{\beta}}$ to its overall minimum as a function of $\boldsymbol{\beta}$. Under the null hypothesis that the model is correctly specified, we expect the reduction $\hat{q}$ to be small. Correspondingly, we expect $\hat{q}$ to be large if the model is mis-specified in some way that permits an appreciable reduction in the multi-step prediction sum of squares by re-adjustment of $\boldsymbol{\beta}$ from $\hat{\boldsymbol{\beta}}$. This is the departure from the null hypothesis to which the test is sensitive. Let the values of the gradient and Hessian at $\hat{\boldsymbol{\beta}}$ be respectively $\hat{\boldsymbol{g}}$ and $\hat{\boldsymbol{H}}$. The objective of this section is to explain the construction of $\hat{q}$, and of the critical values of its distribution under the null hypothesis. We start by defining

$$\hat{q} = \hat{\boldsymbol{g}}' \hat{\boldsymbol{H}}^{-1} \hat{\boldsymbol{g}}. \quad (10)$$

We review the expressions for $\boldsymbol{g}$ and $\boldsymbol{H}$, as given in HTW, because they are required to establish the large sample properties of $\hat{q}$. First we define a new set of linear functions $Z_i(f;\boldsymbol{\beta})$ of $X_i(f;\boldsymbol{\beta})$ as

$$Z_i(f;\boldsymbol{\beta}) = 2\,\mathrm{Re}\big(\sigma^2 \Psi_{L-1}\{\exp(-i2\pi f)\}\,[\Psi_{L-1}\{\exp(i2\pi f)\} X_i(f;\boldsymbol{\beta})]_L^\infty\big)\,, \quad (11)$$

where Re(.) indicates the real part and the operator $[\ .\ ]_L^\infty$ removes all the Fourier coefficients with index less than L from its argument. Its numerical implementation using the discrete Fourier transform is very efficient. Of course this requires a finite Fourier series approximation, but for the smooth bounded functions that arise in practice, the order of the discrete transform can be chosen to ensure that the error is negligible. We will note for later use that $\int Z_i(f)df = 0$. We then have the elements of $\boldsymbol{g}(\boldsymbol{\beta})$ and Hessian matrix $\boldsymbol{H}(\boldsymbol{\beta})$ given by

$$\boldsymbol{g}_i = \int Z_i(f)Y(f)df \quad \text{and} \quad \boldsymbol{H}_{i,j} = \int Z_i(f)X_j(f)df. \tag{12}$$

Although the expression for $\boldsymbol{H}$ is not symmetric, it may be expressed in a form which demonstrates its symmetry (see HTW p.244). In practice we use summation forms of the expressions in (12); see Appendix A of this paper.

The following theorem enables us to determine critical values of the test statistic.

Theorem Under the null hypothesis that w_t is a stationary Gaussian process with a spectrum belonging to the class (1), with parameter $\boldsymbol{\beta}_0$, and bounded away from zero,

$$\tfrac{1}{2}n\hat{q} \xrightarrow{d} \sum_{i=1}^{k-1} d_i C_i,$$

where C_i are independent chi-squared random variables with 1 d.f. and d_i are the non-zero eigenvalues of $\boldsymbol{H}^{-1}\boldsymbol{V} - \boldsymbol{F}^{-1}\boldsymbol{H}$, the $k \times k$ matrices $\boldsymbol{V}$ and $\boldsymbol{F}$ having elements defined by

$$\boldsymbol{V}_{i,j} = \int Z_i(f)Z_j(f)df \quad \text{and} \quad \boldsymbol{F}_{i,j} = \int X_i(f)X_j(f)df. \tag{13}$$

Remark. All these matrices are evaluated at $\boldsymbol{\beta}_0$. However, they are all continuous functions of $\boldsymbol{\beta}$, so consistent estimates of d_i may be determined by substituting $\hat{\boldsymbol{\beta}}$. A consistent critical value c may therefore be calculated for a test with specified size, which rejects the null when q exceeds c. Note that d_i are also the eigenvalues of the $k \times k$ symmetric matrix $\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}'$ where $\boldsymbol{R}'\boldsymbol{R} = \boldsymbol{V} - \boldsymbol{H}\boldsymbol{F}^{-1}\boldsymbol{H}$.

Proof

We first express $\hat{\boldsymbol{g}}$ in the summation form (see Appendix A)

$$\hat{\boldsymbol{g}} = n^{-1}\hat{\boldsymbol{Z}}'(\hat{\boldsymbol{Y}} - 1) \tag{14}$$

where $\hat{\boldsymbol{Z}}$ is the $n \times k$ matrix with elements $\boldsymbol{Z}_{j,i} = Z_i(f_j;\boldsymbol{\beta})$, evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$. The property $\int Z_i(f)df = 0 \Rightarrow \boldsymbol{Z}'1 = 0$, allows us to subtract 1 from $\hat{Y}$. The essential part of the proof is contained in the following lemma.

Lemma

(i)

$$\hat{\boldsymbol{H}}^{-1} = \boldsymbol{H}^{-1} + o_p(1) \tag{15}$$

and

(ii)

$$n^{\frac{1}{2}}\hat{\boldsymbol{g}} = n^{-\frac{1}{2}}\boldsymbol{W}'(\boldsymbol{Y} - 1) + o_p(1), \tag{16}$$

where the $k \times n$ matrix $\boldsymbol{W}' = \boldsymbol{Z}' - \boldsymbol{H}\boldsymbol{F}^{-1}\boldsymbol{X}'$ and all quantities on the RHS are evaluated at $\boldsymbol{\beta}_0$. Furthermore $n^{-1}\boldsymbol{W}'\boldsymbol{W} = \boldsymbol{V} - \boldsymbol{H}\boldsymbol{F}^{-1}\boldsymbol{H}$.

The proof of this lemma is presented in Appendix B, but we remark here that the correction to $\boldsymbol{Z}$, in the expression for $\boldsymbol{W}$, is to allow for the difference between the true and residual spectra, respectively $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$. This corresponds precisely to the correction to the properties of the sample autocorrelation function required to establish the distribution of the portmanteau statistic of Box and Pierce (1970).

To complete the proof of the theorem, express $\boldsymbol{V} - \boldsymbol{H}\boldsymbol{F}^{-1}\boldsymbol{H} = \boldsymbol{R}'\boldsymbol{R}$, where $\boldsymbol{R}$ is an upper right Choleski factor, so that $\boldsymbol{W} = \boldsymbol{Q}\boldsymbol{R}$ where the $n \times k$ matrix $\boldsymbol{Q}$ satisfies $\boldsymbol{Q}'\boldsymbol{Q} = n\,\boldsymbol{I}$. Further, form the eigenvalue decomposition of $\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}' = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}'$ where $\boldsymbol{D}$ is diagonal with elements d_i and $\boldsymbol{U}$ is $k \times k$ and orthonormal. Set $\boldsymbol{Q}\boldsymbol{U} = \boldsymbol{A}$ which is $n \times k$ and satisfies $\boldsymbol{A}'\boldsymbol{A} = n\,\boldsymbol{I}$. Furthermore, the elements of $\boldsymbol{A}$ are given by $\boldsymbol{A}_{j,i} = A_i(f_j; \boldsymbol{\beta_0})$ where $A_i(f)$ are orthonormal functions, well defined by the foregoing construction in terms of $X_i(f)$ and $Z_i(f)$. Then from the Lemma, $\frac{1}{2}n\hat{q}$ differs by $o_p(1)$ from

$$\tfrac{1}{2}n^{-1}(\boldsymbol{Y} - 1)'\boldsymbol{W}\boldsymbol{H}^{-1}\boldsymbol{W}'(\boldsymbol{Y} - 1) = \tfrac{1}{2}n^{-1}(\boldsymbol{Y} - 1)'\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}'(\boldsymbol{Y} - 1) = \sum_{i=1}^{k} d_i\, \tfrac{1}{2}u_i^2, \tag{17}$$

where

$$u_i = n^{-\frac{1}{2}} \sum_j A_{j,i}\,(\boldsymbol{Y}_j - 1) = n^{\frac{1}{2}} \int A_i(f; \boldsymbol{\beta}_0)\ \{Y(f; \boldsymbol{\beta}_0) - 1\}\, df. \tag{18}$$

Lemma 3.1.1 of Taniguchi and Kakizawa (2000, p.56) may now be applied directly (see the discussion in Appendix A) to the quantities u_i to show that they have an asymptotic jointly normal distribution with covariance matrix $2\,\boldsymbol{I}$. Consequently $\frac{1}{2}\,u_i^2$ are asymptotically independent chi-squared variables each on one degree of freedom.

4 Assessment of the statistic for the IMA(1,1) model

We here suppose that the Gaussian series x_t is the sum of a random walk with innovation variance β_1 and independent white noise with variance β_2. There is an IMA(1,1) representation of $(1-B)x_t = (1-\eta B)e_t$, where $\sqrt{\beta_1/\beta_2} = (1-\eta)/\sqrt{\eta}$. The spectrum of $w_t = (1-B)x_t$ is parameterized using β_1 and β_2 as

$$S = \beta_1 + \beta_2\{2 - 2\cos(2\pi f)\}. \tag{19}$$

In this case H is of rank 1, so we can take $\frac{1}{2}n\hat{q}/d_1$ to be asymptotically chi-square with 1 d.f. We carried out a simulation study to investigate the accuracy, under the null hypothesis, of the distributional approximation we have presented for this statistic. Our aim was to assess the reliability of the size of the test for a range of parameter values and lead times. Further, we investigate the power of the test when the IMA(1,1) model is fitted to an ARMA(1,1) process. This exercise was motivated in large part by the original consideration by Cox (1961) of the robustness of the IMA(1,1) model, for prediction of an ARMA(1,1) process, and the more recent investigation by Tiao and Xu (1993), of a chi-squared test statistic based upon re-estimation of the parameter η of the IMA(1,1) model. Our scaled statistic $\frac{1}{2}n\hat{q}/d_1$ may be considered a local version of this, and the two should be asymptotically equivalent. For the combinations of nominal size, lead time, value of η and series length given in Table 1, 10,000 replications were performed to obtain the empirical sizes of our test statistic, displayed in the table.

The empirical size of the test statistic is accurate, or conservative, at all combinations of parameter values and lead times reported in Table 1, including $L = 10$. Thus there is no evidence of the excessive skewness (compared to the relevant asymptotic chi-squared distribution) reported by Tiao and Xu (1993) for all $L > 2$, which gave infeasible large positive size distortions for their proposed statistics, T_J^2 and D_J^2 ($J = L-1$). While our test size does improve with series length, differences are often quite small and the test appears reasonably sized at moderate lead times, certainly for series of length 100 or greater.

Table 2 presents the empirical power of the test when it was applied to the Gaussian ARMA(1,1) process as an alternative data generating model, to which the IMA(1,1) was fitted. Again, 10,000 replications were performed for series generated using each of six combinations of ARMA parameters ($\phi > \theta \geq 0$) at five lead times and two nominal sizes, with all series of length 200. Our six combinations of ARMA parameters form a subset

Lead time	Nominal size	Length of series	η 0	0.2	0.4	0.6	0.8
2	0.10	50	0.094	0.096	0.091	0.091	0.099
		100	0.094	0.096	0.093	0.094	0.093
		200	0.099	0.097	0.092	0.093	0.097
	0.05	50	0.046	0.046	0.043	0.040	0.054
		100	0.046	0.046	0.045	0.045	0.047
		200	0.049	0.048	0.046	0.043	0.047
4	0.10	50	0.088	0.083	0.079	0.080	0.076
		100	0.091	0.091	0.089	0.089	0.085
		200	0.095	0.094	0.094	0.094	0.094
	0.05	50	0.040	0.038	0.039	0.036	0.039
		100	0.044	0.041	0.041	0.042	0.043
		200	0.047	0.047	0.045	0.046	0.047
10	0.10	50	0.075	0.069	0.066	0.063	0.053
		100	0.082	0.082	0.083	0.081	0.071
		200	0.090	0.093	0.087	0.088	0.085
	0.05	50	0.041	0.036	0.032	0.027	0.021
		100	0.043	0.041	0.041	0.037	0.032
		200	0.045	0.043	0.044	0.041	0.041

Table 1: Empirical sizes for the test statistic at various lead times, series lengths and parameter values, with the $(1, \eta)$ IMA(1,1) model. 10,000 replications for each tabulated entry.

of those considered by Tiao and Xu (1993) in their Table 3, where comparable maximum powers for the Ljung-Box statistic, maximised over lags from 2 to 21, are also given. Tiao and Xu gave empirical powers only for 10% nominal size and only for lead time 2, due to the excessive size distortions noted above.

Some general patterns are suggested for our test statistic: power decreases with lead time for 'moderate' values of the AR parameter, $\phi < 0.5$ say, while power is maximised at medium lead times, $2 < L < 10$ for higher values of ϕ. When compared to Tiao and Xu's (1993, Table 3) statistic T_1^2, our test has comparable empirical power for some cases, such as $\phi = 0.9$, but appears less powerful in others. We believe that our accurate or conservative empirical size explains that difference. For example, Tiao and Xu report empirical sizes for T_1^2 of 16.5% and 11.7%, at nominal sizes of 10% and 5%, with series of length 100 and $\eta = 0.8$.

To interpret the observed patterns in the power of our test statistic, it is necessary to consider when the IMA(1,1) model could reasonably be expected to fit well a series generated by an ARMA(1,1) process. With η near unity the IMA(1,1) model can fit well a process that is close to white noise. Hence combinations of ARMA parameters that impose little structure,

Nominal size	Lead time	(ϕ, θ)					
		$(0.1, 0.0)$	$(0.4, 0.1)$	$(0.7, 0.4)$	$(0.9, 0.0)$	$(0.9, 0.4)$	$(0.95, 0.3)$
0.10	2	0.431	0.905	0.557	0.174	0.163	0.108
	4	0.262	0.799	0.714	0.268	0.265	0.136
	6	0.200	0.643	0.706	0.326	0.318	0.148
	8	0.150	0.538	0.666	0.351	0.339	0.154
	10	0.127	0.463	0.610	0.363	0.347	0.151
0.05	2	0.329	0.840	0.437	0.101	0.091	0.056
	4	0.186	0.712	0.599	0.162	0.165	0.072
	6	0.129	0.537	0.583	0.196	0.195	0.079
	8	0.090	0.425	0.534	0.197	0.203	0.074
	10	0.068	0.341	0.458	0.186	0.194	0.066

Table 2: Empirical powers for the test statistic at various lead times. 10,000 replications with series of length 200 for each tabulated entry. Testing the $(1, \eta)$ IMA(1,1) estimated model against the (ϕ, θ) ARMA(1,1) data generating process.

such as $\phi \approx \theta$ (not all reported in this paper), can be well modelled by the IMA(1,1). Also, the IMA(1,1) model can, reasonably well, fit an ARMA(1,1) process with ϕ close to one. Generated processes that are more clearly stationary within the sample period, with moderate values of ϕ, display reasonably rapid reversion to the mean. In such cases power will be greatest at low lead times, where the divergence between the autocorrelation structures beyond lag 1 is marked but forecast uncertainty is still moderate. Conversely, for generated processes that display more persistence, with ϕ close to one, power will increase with lead time, since a greater horizon is required to differentiate between stationary and integrated observed behaviour. However, at long lead times, such as $L = 10$, power tends to reduce, even for ϕ close to one, because of the limited information available for discrimination. The figure given on page 246 of HTW shows how parameter estimation of the IMA(1,1) model, to minimise the MSS of multi-step errors, focuses on information at lower frequencies.

To emphasise this point we carried out one further investigation by simulation of an alternative ARMA(1,1) process of length 400, with $\phi = 0.9$ and $\theta = 0$, and performing the test with a lead time $L = 15$. The greatest advantage from re-estimating the IMA model to minimise the MSS of multi-step errors, a reduction by a substantial factor of two (Tiao and Xu, 1993, p.630), is approached under these circumstances. The MLE of η will be close to $1 - \phi = 0.1$, so that most weight is placed upon recent observations in the EWMA, whereas for a high lead time prediction a value of η close to 1.0 is optimal, so that the forecast is close to the mean of the stationary autoregression. The residuals from MLE of the mis-specified

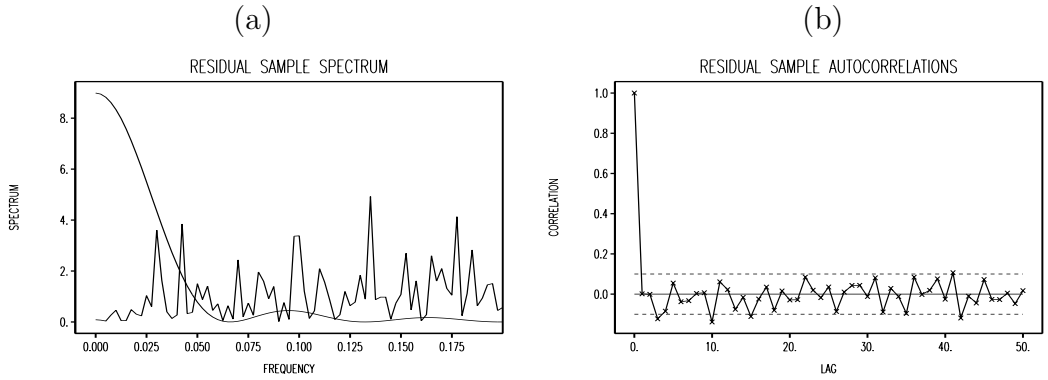

Figure 4: Analysis of the residuals obtained from MLE of an IMA(1,1) model for an AR(1) process with parameter 0.9. (a) The lower frequencies of the residual sample spectrum, with the (scaled) gain function $G_L(f)$ used to weight the same sample spectrum in the MSS of 15 step-ahead prediction errors, (b) the sample autocorrelation function of the residuals.

IMA(1,1) model, contain, however, only a limited amount of information to point to the large swing in the moving average parameter needed for high lead time prediction. From a typical realisation, the evidence may be seen in the rapid reduction in the ordinates of the residual sample spectrum below frequency 0.02, as shown in Figure 4 (a). The corresponding effect may be seen in the residual autocorrelations in Figure 4 (b), as a slight downward bias at low lags. The (scaled) gain function $G_L(f)$ used in the expression (9) for the MSS of 15 step-ahead prediction errors, is shown as a line on Figure 4 (a). By placing weight on the frequencies below 0.05, our test focuses on the information in the lower 5 to 10 spectral ordinates, that is available, in a series of this length, for detecting the lack of uniformity in the residual spectrum. In 10,000 replications, our $\hat{q}$ statistic demonstrated powers of 0.56 and 0.80 for tests of respective size 5% and 10%, showing that the information supporting model re-estimation, though limited, can be detected by this test. The corresponding powers of the Ljung-Box test, using residual autocorrelations up to lag 21, were 0.18 and 0.30.

5 Assessment of the statistic for an autoregressive model

To introduce this section, reconsider the AR(3) model fitted to the seasonally adjusted series of quarterly USA unemployment, in the introduction. Figure 5 (a) shows the sample spectrum of this series, together with the spectrum of the fitted model. The sample spectrum appears to have distinct low frequency peaks, a sharp one close to frequency zero and a broad peak over the frequency range 0.03 to 0.06 (period 4 to 7 years). However, the fitted spectrum

fails to resolve these peaks. Motivated by this we have investigated the power of our test

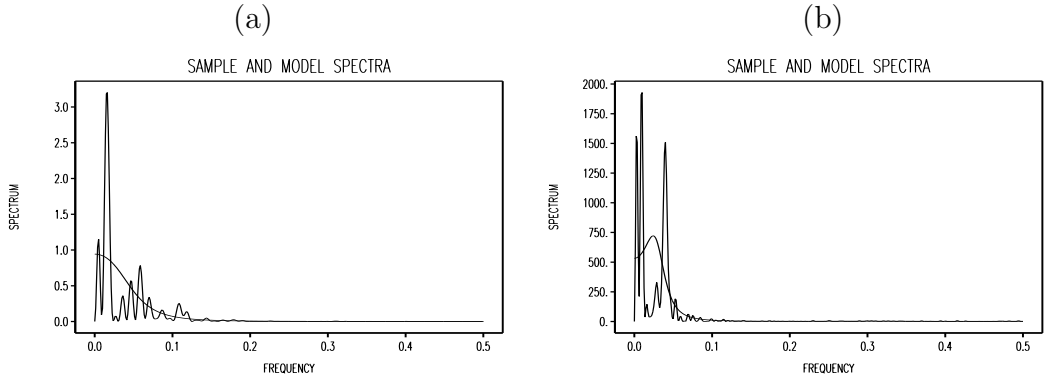

Figure 5: (a) The sample spectrum of the USA unemployment series with the AR(3) model spectrum superimposed; (b) the sample spectrum of a simulated series with an AR(6) model spectrum superimposed.

for simulated series which might reflect a similar structure. We chose as an alternative model, a series composed of the sum of three independent Gaussian components: a near unit root AR(1) process, an AR(2) process with a spectrum peak approximately centered on the frequency 0.04, corresponding to a period of 25, and a white noise series. Thus $x_t = u_t + v_t + z_t$ where $u_t = \phi u_{t-1} + a_t$, with $\phi = 0.99$ and $\text{var}\,(a_t) = 1.0$; $v_t = 2r\cos(\lambda)v_{t-1} - r^2 v_{t-2} + b_t$ with $r = 0.98$, $\lambda = 2\pi/25$ and $\text{var}\,(b_t) = 0.3^2$; and z_t is uncorrelated with variance 1.0. To check the size of the test, we required an approximating autoregressive model of suitable order. From a simulation of the process x_t, use of the AIC selected an autoregressive approximation of order 6. We therefore determined the AR(6) model $\phi(B)x_t = e_t$ that provided the minimum variance one-step ahead predictor of x_t. This had coefficients $\phi_1 = 0.9177$, $\phi_2 = 0.2455$, $\phi_3 = -0.0069$, $\phi_4 = -0.0892$, $\phi_5 = -0.0919$, $\phi_6 = -0.0290$, and $\text{var}\,(e_t) = 3.6451$.

We simulated 10,000 samples of time series with length 200 from this AR(6) model, fitted an AR(6) model to each sample in the frequency domain, and evaluated the test statistic for lead times $L = 8$ and $L = 16$. These were chosen as substantial fractions of the period of the stochastic cycle, at which it might be considered important to detect inadequacy of predictions. The construction of the test statistic for this model is described in Appendix C.

The results of these simulations are shown in Table 3. The empirical sizes are shown for the nominal sizes of 10% and 5%, for our test statistic q, and for the Ljung-Box test statistic calculated using residual autocorrelations up to lag 20. For the q statistic the empirical sizes are slightly conservative for lead time 8, more so for lead time 16. The Ljung-Box test is

slightly over-sized.

Lead time	Nominal size	q Statistic	Ljung-Box
8	0.1	0.095	0.113
	0.05	0.045	0.056
16	0.1	0.084	0.113
	0.05	0.035	0.056

Table 3: Empirical size of tests of model adequacy for given lead time and nominal size, for an AR(6) null model. 10,000 replications for each tabulated entry.

We proceeded to estimate the power of the tests for rejecting the AR(6) model, when fitting it to the alternative process with independent components described above, using the same series length and number of replications. Table 4 shows that for the test of nominal size 5% based on the q statistic, applied for lead time 16, an empirical power of over 60% was achieved against the alternative. In comparison, the Ljung-Box statistic is very much less powerful. When applied for lead time 8, the power of the q statistic was lower than for lead time 16, but still noticeably more powerful than the (over-sized) Ljung-Box test.

Lead time	Nominal size	q Statistic	Ljung-Box
8	0.1	0.377	0.246
	0.05	0.274	0.144
16	0.1	0.722	0.246
	0.05	0.620	0.144

Table 4: Empirical powers of tests of model adequacy for given lead time and nominal size, for an AR(6) model fitted to the alternative process with independent components described in the text. 10,000 replications for each tabulated entry.

We can gain some further insight into these results, by calculation of the prediction error variances of the alternative process under the true model, and the best approximating AR(6) DMS predictor. The prediction error variances for lead times 1, 8 and 16, using the true model, are respectively 3.28, 24.51 and 31.82. The error variances that are achieved at the same lead times, using the AR(6) model that minimises the one-step prediction error variance, are respectively 3.65, 46.15 and 72.61. These are clearly much greater than those of the true model, for the lead times greater than 1. The error variances that can be achieved, using, for each respective lead time, the AR(6) DMS predictor, are 3.65, 40.93 and 38.14. This shows that the potential reduction in prediction error variance that may be achieved is relatively small at lead time 8, from 46.15 to 40.93. However it is very substantial at lead

time 16, from 72.61 to 38.14, which is not much greater than the value of 31.82 achieved by the true model. This accords with the results of the power investigation, that suggests it is quite difficult to detect potential forecast improvement using a lead time of 8 in this case, but an important potential gain is detected at a lead time of 16. Investigation of several realisations confirmed the indications gained from the example of USA unemployment. The residual autocorrelations and spectrum do not show up any clear lack of fit of the AR(6) model but, as Figure 5 (b) shows, the fitted model spectrum typically fails to resolve the peaks in the sample spectrum of the series.

6 Conclusion

The potential reduction in the multi-step forecast error that may be gained from re-estimation of the parameters of a time series model, justifies a test which is sensitive to this possibility. We have presented a widely applicable test for this purpose, derived its asymptotic properties and validated its small sample properties in important applications. The methodology of the test also provides insight into statistical features of the residual spectrum that are associated with a significant outcome.

References

Akaike, H. (1973) A new look at statistical model identification. *IEEE Transactions on Automatic Control*, **AC-19**, 716–723.

Beran, J. (1994) *Statistics for Long Memory Processes*. London: Chapman and Hall.

Bhansali, R. J. (1996) Asymptotically efficient autoregressive model selection for multistep prediction. *Ann. Inst. Statist. Math.*, **48**, 577–602.

Bhansali, R. J. (1999) Parameter estimation and model selection for multistep prediction: a review. In S. Gosh (Ed.), *Asymptotics, Nonparametrics and Time Series*, 201–225. New York: Marcel Dekker.

Box, G. E. P. and Jenkins, G. M. (1976) *Time Series Analysis, Forecasting and Control*, Revised Edition. California: Holden-Day.

Box, G. E. P. and Pierce, D. A. (1970) Distribution of residual autocorrelations in autoregressive-integrated moving average time series models, *J. Am. Statist. Ass.*, **65**, 1509–1526.

Chevillon, G. (2007) Direct multi-step estimation and forecasting. *Journal of Economic Surveys*, **21**, 746–785.

Chevillon, G. and Hendry, D. F. (2005) Non-parametric direct multi-step estimation for forecasting economic processes. *Int. Journal of Forecasting*, **21**, 201–218.

Cox, D. R. (1961) Prediction by exponentially weighted moving averages and related methods. *J. R. Statist. Soc.* B, **23**, 414–422.

Diggle. P. J. (1990) *Time Series: A Biostatistical Introduction.* Oxford: Oxford University Press.

Fahrmeir, L. and Kaufmann, H. (1985) Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.*, **13**, 342–368.

Findley, D. F. (1983) On the use of multiple models for multi-period forecasting. *Proceedings of Business and Economic Statistics Section*, 528–531, American Statistical Association.

Harvey, A. C. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: Cambridge University Press.

Haywood, J. and Tunnicliffe Wilson, G. (1997) Fitting time series models by minimizing multistep-ahead errors: a frequency domain approach. *J. R. Statist. Soc.* B, **59**, 237–254.

Haywood, J. and Tunnicliffe Wilson, G. (2000a) Selection and estimation of component models for seasonal time series. *Journal of Forecasting*, **19**, 393–417.

Haywood, J. and Tunnicliffe Wilson, G. (2000b) An improved state space representation for cyclical time series. *Biometrika*, **87**, 724–726.

Ing, C.-K. (2003) Multistep prediction in autoregressive processes. *Econometric Theory*, **19**, 254–279.

Kang. I.-B. (2003) Multi-period forecasting using different models for different horizons: an application to U.S. economic time series data. *Int. Journal of Forecasting*, **19**, 387–400.

Ljung, G. M. and Box, G. E. P. (1978) On a measure of lack of fit in time series models. *Biometrika*, **65**, 297–303.

Marcellino, M., Stock, J. H. and Watson, M. W. (2006) A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, **135**, 499–526.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

Stoica, P. and Soderstrom, T. (1984) Uniqueness of estimated k-step prediction models of ARMA processes. *Syst. Control Letters*, **4**, 325–331.

Taniguchi, M. and Kakizawa, Y. (2000) *Asymptotic Theory of Statistical Inference for Time Series*. New York: Springer-Verlag.

Tiao, G. C. and Xu, D. (1993) Robustness of maximum likelihood estimates for multi-step predictions: the exponential smoothing case. *Biometrika*, **80**, 623–641.

Appendix A

Our Theorem is expressed in terms of the vector and matrices $\boldsymbol{g}$, $\boldsymbol{H}$, $\boldsymbol{V}$ and $\boldsymbol{F}$ defined using integral formulae in (12) and (13). However, for our proofs we prefer to use, in place of these, definitions such as (14) in terms of summation formulae. This has the advantage that we can use, throughout the proof in Appendix B, familiar matrix notation for the manipulation of projections and other decompositions which arise from the GLM estimation and its extension to the multi-step context. In particular we use it to express $n^{-1}\boldsymbol{Z}'\boldsymbol{X} = \boldsymbol{H}$, $n^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{F}$ and $n^{-1}\boldsymbol{Z}'\boldsymbol{Z} = \boldsymbol{V}$. It also corresponds precisely to the discrete computational procedures used to implement the test. The disadvantage is that summation formulae for $\boldsymbol{H}$, $\boldsymbol{V}$ and $\boldsymbol{F}$ involve the number of spectral ordinates n, and are strictly only valid for large n. However, as we note in the following, the discrepancy between an integral and summation formula decreases exponentially with n, justifying its neglect relative to the stochastic errors.

The elements of $\boldsymbol{H}$, $\boldsymbol{V}$ and $\boldsymbol{F}$ are typically integrals of an analytic function $A(f)$ with Fourier coefficients a_j, $-\infty < j < \infty$ which satisfy $|a_j| < \alpha\,\rho^j$ for some $0 \leq \rho < 1$. The discrete Fourier coefficients of $A(f)$ are then the aliased values $\tilde{a}_j = a_j + \sum_{k=1}^{\infty}(a_{j+kn} + a_{j-kn})$, $0 \leq j \leq (n-1)$. Now

$$\int_{f=-\frac{1}{2}}^{\frac{1}{2}} A(f)df = a_0 \quad \text{and} \quad n^{-1}\sum_{j=0}^{n-1} A(f_j) = \tilde{a}_0,$$

so these integral and summation forms differ by $O(\rho^n)$.

The elements of $\boldsymbol{g}$ in (12) are integrals of the product of an analytic function with the

sample spectrum, which may be expressed:

$$\int_{f=-\frac{1}{2}}^{\frac{1}{2}} A(f)S^*(f)df = \sum_{j=-(n-1)}^{(n-1)} a_j C_j, \tag{20}$$

where C_j is the usual sample autocovariance of w_t. The sum is finite because C_j is a finite sequence, the aliased coefficient of which is just $C_j + C_{j-n}$. The summation form may be expressed, using $C_{-j} = C_j$ and taking, for $-n < j < 0$, $\tilde{a}_j$=$\tilde{a}_{j+n}$, as

$$n^{-1}\sum_{j=0}^{n-1} A(f_j)S^*(f_j) = \sum_{j=0}^{(n-1)} \tilde{a}_j(C_j + C_{j-n}) = \sum_{j=-(n-1)}^{n-1} \tilde{a}_j C_j. \tag{21}$$

The difference between the integral and summation forms are again $O(\rho^n)$, and converge to zero much more rapidly than terms of $o_p(1)$ in the stochastic approximations we use.

In practice, in all our computations, the integral form may be replaced by twice the integral from 0 to $\frac{1}{2}$, and the summation form by twice the sum of the terms for f_j between 0 and $\frac{1}{2}$, plus the term for f_0 and, if n is even, the term for $f_{\frac{1}{2}n}$. We could approximate the integrals to arbitrary accuracy using a sum over a larger number of divisions than n, but this makes little difference to our simulation results. However, the selection of the harmonic frequencies enables us to use, in Appendix B, asymptotic properties of estimates of generalized linear models based on the large sample statistical properties of the sample spectrum at these frequencies. The completion of the proof of the theorem in the text of the paper refers to Lemma 3.1.1 of Taniguchi and Kakizawa (2000), for which two derivations are indicated, one depending on detailed distributional properties of the sample spectra, which are necessary for statistics which are functionals of the whole spectrum. The other is based upon representing linear functionals of the sample spectrum as linear combinations of the sample covariances as in (20) and (21) above. These two approaches are valid under very general, but different, conditions which are stated in the Lemma, and to which we refer the reader for general conditions which may be applied to our process w_t.

Appendix B

We remark first that $\hat{q}$ may also be expressed as $\hat{\boldsymbol{g}}'\boldsymbol{\delta}$where $\boldsymbol{\delta}$ is a solution of $\hat{\boldsymbol{H}}\boldsymbol{\delta} = \hat{\boldsymbol{g}}$. However, as noted in HTW, this solution is in general unique only up to a multiple of the value $\boldsymbol{\beta}$ used to construct $\boldsymbol{g}$ and $\boldsymbol{H}$; a consequence of the fact that, by construction, $\sum \beta_i X_i(f) \equiv 1$ and $\sum \beta_i Z_i(f) \equiv 0$, so that $\boldsymbol{\beta}'\boldsymbol{g} = 0$ and $\boldsymbol{\beta}'\boldsymbol{H} = 0$. The particular choice of solution does not

affect the value of q, but to overcome this singularity, we propose that whenever the inverse of $\boldsymbol{H}(\boldsymbol{\beta})$ is required, it is replaced by the inverse of the augmented matrix $\boldsymbol{H}(\boldsymbol{\beta})+\alpha\boldsymbol{\beta}\boldsymbol{\beta}'$, for some $\alpha > 0$, which is, in general, strictly positive definite. It is readily checked (premultiply the equations by $\boldsymbol{\beta}'$) that the solution $\boldsymbol{\delta}$ of $(\boldsymbol{H} + \alpha\boldsymbol{\beta}\boldsymbol{\beta}')\boldsymbol{\delta} = \boldsymbol{g}$ is actually orthogonal to $\boldsymbol{\beta}$ and is consequently also a solution of $\boldsymbol{H}\boldsymbol{\delta} = \boldsymbol{g}$. We will assume hereafter that $\boldsymbol{H}$ has been augmented in this manner in the definition (10) and other expressions involving $\boldsymbol{H}^{-1}$, but, for convenience of notation, we will not explicitly show this. We will shortly show that $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}_0$, so the first part of the lemma follows because the (augmented) matrix $\boldsymbol{H}$ is a continuous function of $\boldsymbol{\beta}$ and bounded away (in norm) from zero.

Proof of the second part of the lemma requires a careful assessment of the properties of the estimate $\hat{\boldsymbol{\beta}}$ obtained from the GLM equations. We use a result of Fahrmeir and Kaufmann (1985, p.353), that relates $\hat{\boldsymbol{\beta}}$ to the Fisher information $\boldsymbol{X}'\boldsymbol{X}$ and score $\boldsymbol{X}'(\boldsymbol{Y}-1)$ at $\boldsymbol{\beta}_0$. We express this result as $n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) = o_p(1)$, where $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{Y}-1)$. We are assuming that for $0 \le j \le \frac{1}{2}n$, $\boldsymbol{Y}_j = S^*(f_j)/S(f_j)$, are independent Gamma random variables with mean 1 and variance 1, except that for $j = 0$ and (if n is even) $j = \frac{1}{2}n$, the variance is $\frac{1}{2}$. Consequently, $n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_0) = O_p(1)$ and $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = O_p(n^{-\frac{1}{2}})$, so that $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}_0$. We now need to show that

$$n^{-\frac{1}{2}}\hat{\boldsymbol{Z}}'\left(\frac{\boldsymbol{S}^* - \hat{\boldsymbol{S}}}{\hat{\boldsymbol{S}}}\right) = n^{-\frac{1}{2}}\boldsymbol{Z}'\left(\frac{\boldsymbol{S}^* - \tilde{\boldsymbol{S}}}{\boldsymbol{S}}\right) + o_p(1), \tag{22}$$

where, from (14), the LHS is $n^{\frac{1}{2}}\hat{\boldsymbol{g}}$. For simplicity of notation, we are using $\boldsymbol{S}^*$, $\hat{\boldsymbol{S}}$, $\tilde{\boldsymbol{S}}$ and $\boldsymbol{S}$ for the vectors with elements $S^*(f_j)$, $S(f_j;\hat{\boldsymbol{\beta}})$, $S(f_j;\tilde{\boldsymbol{\beta}})$ and $S(f_j;\boldsymbol{\beta}_0)$. Similarly, $\hat{\boldsymbol{Z}}$ and $\boldsymbol{Z}$ have elements $Z_i(f_j;\hat{\boldsymbol{\beta}})$ and $Z_i(f_j;\boldsymbol{\beta}_0)$. The fractions in brackets, here and elsewhere, are elementwise ratios. To verify (22), split both sides into two components, by expressing, on the left, $\boldsymbol{S}^* - \hat{\boldsymbol{S}} = (\boldsymbol{S}^* - \boldsymbol{S}) - (\hat{\boldsymbol{S}} - \boldsymbol{S})$, and on the right $\boldsymbol{S}^* - \tilde{\boldsymbol{S}} = (\boldsymbol{S}^* - \boldsymbol{S}) - (\tilde{\boldsymbol{S}} - \boldsymbol{S})$. The difference between the first components on the left and right may be expressed:

$$n^{-\frac{1}{2}}\left[\boldsymbol{S}\left(\frac{\hat{\boldsymbol{Z}}}{\hat{\boldsymbol{S}}} - \frac{\boldsymbol{Z}}{\boldsymbol{S}}\right)\right]'(\boldsymbol{Y}-1). \tag{23}$$

Now $\boldsymbol{S}$ and $\boldsymbol{Z}$ are both linear in $\boldsymbol{\beta}$, so we may linearize

$$\frac{\hat{\boldsymbol{Z}}}{\hat{\boldsymbol{S}}} - \frac{\boldsymbol{Z}}{\boldsymbol{S}} = \boldsymbol{\rho}(\boldsymbol{\beta}_0)\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + O\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^2\right], \tag{24}$$

where the matrix $\boldsymbol{\rho}$ is the derivative of $\boldsymbol{Z}/\boldsymbol{S}$. Then (23) becomes

$$n^{-\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\,(\boldsymbol{S}\boldsymbol{\rho})'\,(\boldsymbol{Y}-1) + n^{-\frac{1}{2}}O_p(n^{-1})\sum|\boldsymbol{Y}_j - 1|\,. \tag{25}$$

From the statistical properties of $\boldsymbol{Y}$, $(\boldsymbol{S\rho})'(\boldsymbol{Y}-1)$ is $O_p(n^{\frac{1}{2}})$ and $\sum|\boldsymbol{Y}_j-1|$ is $O_p(n)$. But $(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$ is $O_p(n^{-\frac{1}{2}})$, so the overall magnitude of the difference is $O_p(n^{-\frac{1}{2}})$.

The second component on the left approximates the second on the right as:

$$\begin{aligned} & n^{-\frac{1}{2}}\left(\frac{\hat{\boldsymbol{Z}}}{\hat{\boldsymbol{S}}}\right)'(\hat{\boldsymbol{S}}-\boldsymbol{S}) \\ = \; & n^{-\frac{1}{2}}\left(\frac{\boldsymbol{Z}}{\boldsymbol{S}}+O_p(n^{-\frac{1}{2}})\right)'\left([\tilde{\boldsymbol{S}}-\boldsymbol{S}]+o_p(n^{-\frac{1}{2}})\right) \\ = \; & n^{-\frac{1}{2}}\left(\frac{\boldsymbol{Z}}{\boldsymbol{S}}\right)'(\tilde{\boldsymbol{S}}-\boldsymbol{S})+n^{-\frac{1}{2}}\,n\,o_p(n^{-\frac{1}{2}})+n^{-\frac{1}{2}}O_p(n^{-\frac{1}{2}})\,n\,O_p(n^{-\frac{1}{2}})+o_p(n^{-\frac{3}{2}}) \\ = \; & n^{-\frac{1}{2}}\left(\frac{\boldsymbol{Z}}{\boldsymbol{S}}\right)'(\tilde{\boldsymbol{S}}-\boldsymbol{S})+o_p(1), \end{aligned} \tag{26}$$

where we have used the fact that $(\hat{\boldsymbol{S}}-\tilde{\boldsymbol{S}})=o_p(n^{-\frac{1}{2}})$ and $(\tilde{\boldsymbol{S}}-\boldsymbol{S})=O_p(n^{-\frac{1}{2}})$, and that there are n terms in the sums involved. To complete the proof, from (22) we express

$$\begin{aligned} n^{-\frac{1}{2}}\boldsymbol{Z}'\left(\frac{\boldsymbol{S}^*-\tilde{\boldsymbol{S}}}{\boldsymbol{S}}\right) &= n^{-\frac{1}{2}}\boldsymbol{Z}'(\boldsymbol{Y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}}) \\ &= n^{-\frac{1}{2}}\left(\boldsymbol{Z}'\boldsymbol{Y}-\boldsymbol{Z}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\right)=n^{-\frac{1}{2}}\boldsymbol{W}'(\boldsymbol{Y}-1), \end{aligned} \tag{27}$$

where $\boldsymbol{W}'=\boldsymbol{Z}'-\boldsymbol{Z}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'=\boldsymbol{Z}'-\boldsymbol{H}\boldsymbol{F}^{-1}\boldsymbol{X}'$. We can subtract 1 from $\boldsymbol{Y}$ because $\boldsymbol{Z}'1=0$ and $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'1=\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}_0=\boldsymbol{X}\boldsymbol{\beta}_0=1$. We can also write $\boldsymbol{W}'=\boldsymbol{Z}'\boldsymbol{P}$, where $\boldsymbol{P}=\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is a symmetric projection, so that $n^{-1}\boldsymbol{W}'\boldsymbol{W}=n^{-1}\boldsymbol{Z}'\boldsymbol{P}^2\boldsymbol{Z}=n^{-1}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{Z}=n^{-1}\boldsymbol{W}'\boldsymbol{Z}=\boldsymbol{V}-\boldsymbol{H}\boldsymbol{F}^{-1}\boldsymbol{H}$. Note also that $\boldsymbol{Z}\boldsymbol{\beta}=0\Rightarrow\boldsymbol{W}\boldsymbol{\beta}=0$, so $\boldsymbol{W}$ has at most rank $k-1$.

Appendix C

The test statistic can be conveniently calculated for a general (possibly seasonal) ARIMA model which has been fitted using standard time domain methods. An R function which implements this is available from the authors; its arguments are the maximum likelihood parameter estimates and the residual series. These are used as follows to calculate the quantities needed to carry out the test.

The weighted response $Y(f;\boldsymbol{\beta})=S^*(f)/S(f;\boldsymbol{\beta})$ is taken as the sample spectrum of the residual series. The weighted frequency domain regressors $X_i(f;\boldsymbol{\beta})=S_i(f)/S(f;\boldsymbol{\beta})$ are determined from a linearisation of the model spectrum $S(f;\boldsymbol{\beta})$. They are the derivatives

of $\log S(f;\boldsymbol{\beta})$. The test statistic is invariant to re-parameterisation, so at this point the logarithm of the spectrum of the ARMA part (omitting integration) can be expressed as

$$\log S(f;\boldsymbol{\beta}) = \log(\sigma_e^2) + \log\{\alpha_0 + \sum_{j=1}^{q} \alpha_j \, 2\, \cos(2\pi j f)\} - \log\{\beta_0 + \sum_{j=1}^{p} \beta_j \, 2\, \cos(2\pi j f)\},$$

where, taking ϕ_j and θ_j to be the estimated ARMA model parameters, and $\phi_0 = \theta_0 = -1$, $\alpha_j = \sum_{k=0}^{q-k} \theta_k \theta_{j+k}$ and $\beta_j = \sum_{k=0}^{p-k} \phi_k \phi_{j+k}$. Setting $S_{MA}(f) = \{\alpha_0 + \sum_{j=1}^{q} \alpha_j \, 2\, \cos(2\pi j f)\}$ and $S_{AR}(f) = \{\beta_0 + \sum_{j=1}^{p} \beta_j \, 2\, \cos(2\pi j f)\}$, the required linearised regressors for the MA and AR parts are respectively $2\, \cos(2\pi j f)/S_{MA}(f)$, $j = 1,\ldots, q$ and $-2\, \cos(2\pi j f)/S_{AR}(f)$, j=1,..., p. The regressor for the term $\log(\sigma_e^2)$ may be omitted, because it is constant, and the elements of g and H involving this term are all zero. Then H will in general be invertible.

The transformation (11) of the regressors, from $X_i(f;\boldsymbol{\beta})$ to $Z_i(f;\boldsymbol{\beta})$, further requires the function $\Psi_{L-1}\{\exp(-i2\pi f)\}$, for which the coefficients Ψ_k may be directly calculated from the fitted ARIMA model. The above procedure extends easily to multiplicative seasonal ARIMA models of Box and Jenkins (1976). The regressors are augmented by seasonal moving average and autoregressive terms similar to the non-seasonal terms, but replacing $\cos(2\pi j f)$ with $\cos(2\pi j s f)$, where s is the seasonal period.

Because estimated ARIMA model parameters can suffer high collinearity, in implementing our function we have replaced the simple matrix formulas with numerically stable forms using singular value decomposition. We have carried out checks that the procedure is highly reliable, and gives results very close to those based on frequency domain estimation of the models.