A test for improved multi-step forecasting

John Haywood[*] and Granville Tunnicliffe Wilson[†]

29 November 2004

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 estimation, for some specified lead-time. The test statistic is based on a score function, defined as the derivative of the multi-step forecast error variance with respect to the model parameters, and evaluated at parameters estimated using the single-step criterion. We show that the test has acceptable size properties for higher lead times, when applied to the exponentially weighted moving average predictor, and investigate how its power varies with the lead time, under the simple ARMA(1,1) alternative. We also demonstrate the high power of the test when it is applied to a process generated as the sum of a stochastic trend and cycle plus noise, which has been modelled by a high order autoregression. We use frequency domain methods which give insight into the derivation and sampling properties of the test, but note that the test statistic may be expressed as a quadratic form in the residual sample autocorrelations. The test is illustrated on two real time series, which demonstrate its wide applicability.

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

1 Introduction

This paper addresses a dilemma that arises in the construction of multi-step predictions of discrete time series. One approach is first to fit a parametric process model by minimising the sum of squares of in-sample single-step forecast errors, or equivalently, by maximum likelihood estimation, or one of the several, closely related, approximations to this criterion. 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. This result

[*] School of Mathematics, Statistics & Computer Science, Victoria University of Wellington, PO Box 600, Wellington, NZ.

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

is often termed the *plug-in* or *iterated* multi-step (IMS) prediction. A second 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 dilemma, setting out the problem, and presenting the statistical issues. 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 minimise 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 Haywood and Tunnicliffe-Wilson (1997) 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 generalised method of moments (GMM) is used in place of OLS for autoregressive predictors (Chevillon and Hendry, 2004). 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) and for autoregressive models by Bhansali (1996). See also 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 careful model selection, prior to estimation by maximum likelihood, followed by diagnostic checking applied to the estimated residuals. However, 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 (2004) 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.

In this paper 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 estimation, for some specified lead-time. Such a test provides one solution to the dilemma of choice between an IMS predictor based on a parametric time series model, and its corresponding (parametric) DMS predictor. As with the portmanteau statistics, it may be applied immediately following maximum likelihood estimation; our test statistic is, in fact, a low rank quadratic form in the residual sample autocorrelations. 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 prediction at that lead time, even if there is some evidence from the portmanteau statistic, of unspecific model inadequacy. If the result is significant, the user can expect an immediate gain in forecast accuracy, at that lead time, by re-estimation of the model parameters. They may also be prompted to investigate ways in which the model can be modified.

We now illustrate the test with 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. 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 Haywood and Tunnicliffe Wilson (1997, page 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. Choosing a maximum lag of 25, the Ljung-Box portmanteau statistic (Ljung and Box, 1978) 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.

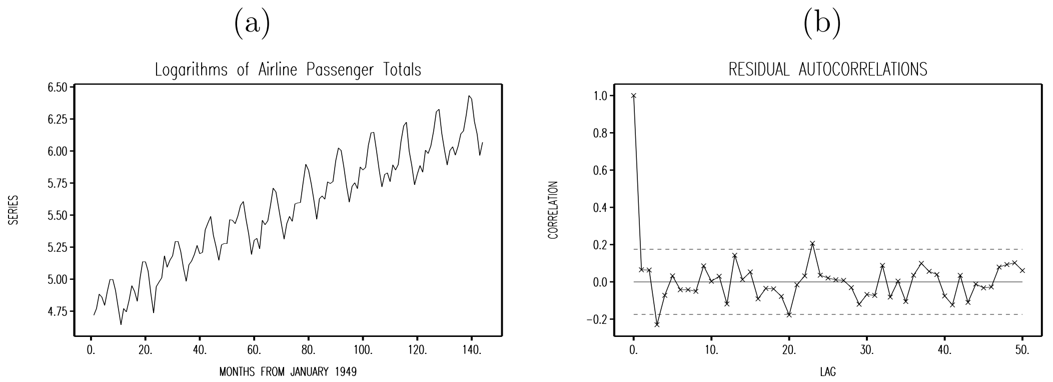

Figure 1: (a) The logarithm of the series of monthly airline passenger totals and (b) the sample residual autocorrelations following estimation of a seasonal structural model.

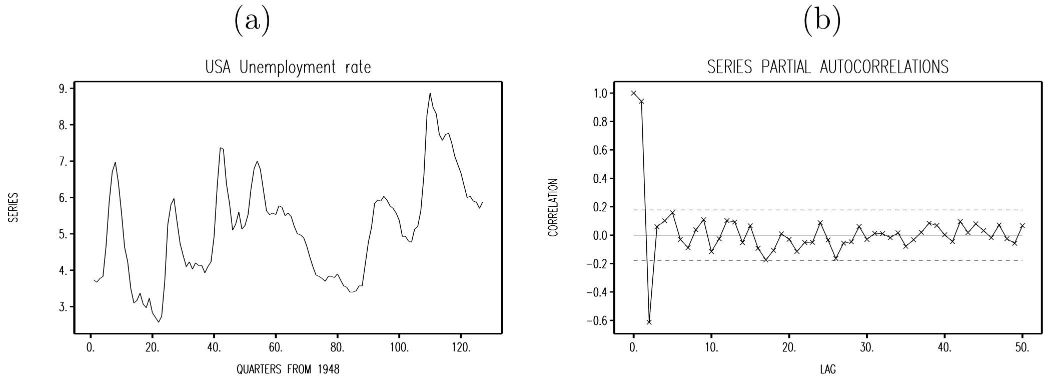

Figure 2: (a) The USA unemployment rate from 1948 to 1979, together with (b) the sample partial autocorrelations of the square root transformed series.

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 d.f. with a p-value of 7.5%. On applying our test with a lead time of one year (four quarters), we obtained a p-value of 4%, indicating a significant potential for forecast improvement.

We derive our test statistic, and its sampling properties, in the next section of this paper. In Section 3 we carry out an empirical study of its size when applied to the IMA(1,1)

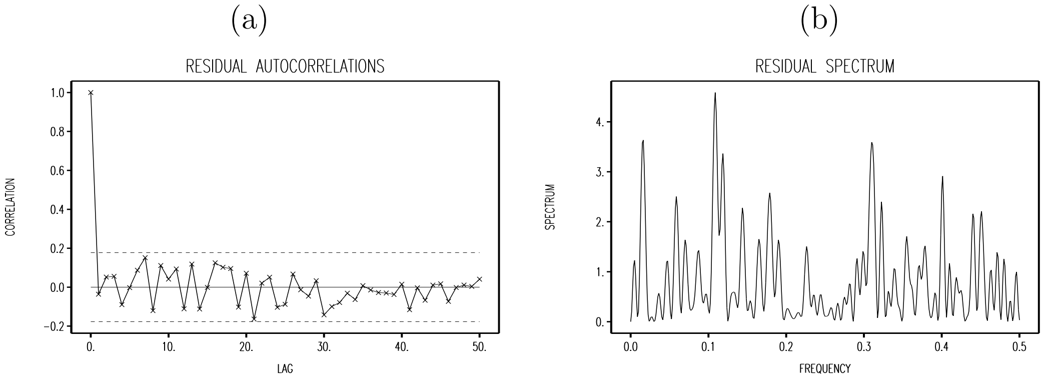

Figure 3: (a) The residual sample autocorrelations and (b) the sample spectrum of the residuals, from fitting an AR(3) model to the transformed USA unemployment rate series.

model, and its power against the alternative of 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 of 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 4 is motivated by the foregoing example of the USA unemployment series and the use of autoregressive predictors. We seek to identify the model mis-specification that might explain the significant result in that example. Based on this we formulate a model for use in a simulation exercise to confirm the size properties of our test in the context of a higher order autoregressive predictor, and its power against a realistic alternative. We give some brief concluding comments in Section 5.

2 A score-type test

Haywood and Tunnicliffe Wilson (1997), hereafter referred to as HTW, presented a frequency domain approach to estimating time series model parameters by minimising the multi-step ahead prediction error sum of squares. The score statistic that we now investigate is proposed in the discussion of that paper. In recent years there has been renewed appreciation of frequency domain methods for the estimation and testing of parametric time series models (e.g. Harvey 1989, Section 4.3), using what is often termed the Whittle likelihood. HTW show how these methods provide insight for the multi-step estimation problem, and we exploit the asymptotic independence of sample spectral ordinates to derive the properties of our test statistic, which may, however, also be expressed as a low rank quadratic form in the residual sample autocorrelations. The score statistic is evaluated at the parameters estimated by minimising the single-step forecast error variance, so that the model need not be re-estimated by determining the parameters that minimise the multi-step forecast error variance. We also find that the asymptotic distributional properties of the test, under the null hypothesis, are reliable, even at higher lead times.

Standard score function test statistics that approximate likelihood ratio statistics, are quadratic forms in the score, having chi-squared distributions, and being invariant to locally linear parameter transformations. But, except in the case of a single parameter model, we have a choice, in our context, of two distinct invariant statistics. One is formed from the score and its asymptotic variance matrix, and approximates the Mahalanobis length of the change in parameter values that results from multi-step re-estimation of the model. It has an asymptotic chi-squared distribution. The other is formed from the score and the local Hessian matrix, so as to approximate the proportional reduction in the multi-step error variance that results from the re-estimation. Its asymptotic distribution is a weighted sum of chi-squared variables on one degree of freedom. This is the test that we favour, for its greater reliability and power.

We introduce the statistic by briefly reviewing and extending some of the methodology and notation introduced by HTW. We assume that the 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. Suppose that the model for w_t has a spectrum $S(f)$, for $0 \leq f \leq \frac{1}{2}$, which has a linear form with coefficients $\beta_1, \ldots, \beta_k$:

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

where S_β is the row vector of component functions $S_i(f)$, and β is the column vector of coefficients. This is easily extended to a wider class of spectra which are products or ratios of this form, which encompasses all ARMA models.

Let the infinite moving average representation of the model for w_t be

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

where e_t, with variance σ^2, is the white noise linear innovation of w_t. Then the model spectrum for w_t is given by

$$S(f) = \sigma^2|\psi\{\exp(i\omega)\}|^2 = |\nu\{\exp(i\omega)\}|^2, \tag{3}$$

where $\nu(B) = \sigma\psi(B)$ and $\omega = 2\pi f$. The infinite moving average representation of the integrated process $x_t = (1-B)^{-d}w_t$ is then given as

$$x_t = (1-B)^{-d}\psi(B)e_t \overset{\text{def}}{=} \Psi(B)e_t. \tag{4}$$

Now define the truncated operator

$$T(B) = \sigma\left(\Psi_0 + \Psi_1 B + \cdots + \Psi_{L-1}B^{L-1}\right). \tag{5}$$

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, may be expressed as

$$e_t(L) = \frac{T(B)}{\nu(B)}w_{t+L}. \tag{6}$$

The mean sum of squares (MSS) of L-step ahead prediction errors then has the frequency domain approximation

$$F_L(\beta) = 2\int_0^{\frac{1}{2}} G(f)\frac{S^*(f)}{S(f)}df, \tag{7}$$

in terms of the model spectrum, the *gain* function

$$G(f) = |T\{\exp(i\omega)\}|^2 \tag{8}$$

and the sample spectrum of w_t:

$$S^*(f) = \frac{1}{n}\left|\sum_{t=1}^{n} w_t \exp(i\omega t)\right|^2. \tag{9}$$

The minimisation of F_L with respect to β requires the iterative solution of equations

$$H\delta = g, \tag{10}$$

for the parameter corrections δ, where g is the negative of the derivative of F_L with respect to β, and H is an estimate of the Hessian of F_L. Expressions for g and H are given in HTW.

The test statistic which we propose, is

$$q = g'H^{-1}g, \tag{11}$$

evaluated at the (approximate MLE) parameters $\hat{\beta}$ which minimise $F_L(\beta)$ for lead time $L = 1$. (In HTW we in fact proposed $\frac{1}{2}q$). Let $\tilde{\beta}$ be the parameters which minimise $F_L(\beta)$ for a specified value of $L > 1$. The value of $\frac{1}{2}q$ is an approximation (under a locally quadratic assumption) to $F_L(\hat{\beta}) - F_L(\tilde{\beta})$, the reduction that may be achieved in the L-step prediction error MSS, by re-estimation of the parameters. Expressions for g and H are given by first forming the functions of frequency $Y = S^*/S$, $X_i = S_i/S$ and $Z_i = K(X_i)$, where K is a linear operator, depending on the model parameters, that is defined in HTW. Then the elements of g and H are

$$g_i = 2\int_0^{\frac{1}{2}} Y(f)\, \Re Z_i(f) df, \qquad H_{i,j} = 2\int_0^{\frac{1}{2}} X_i(f) \Re Z_j(f) df. \tag{12}$$

In practice these integrals are evaluated by finite sums over the discrete harmonic frequencies $f_r = r/n$, $0 \le r \le n/2$. Retaining for notational convenience the same symbols, we define the vector Y with elements $Y_r = Y(f_r)$ and matrices X and Z with elements $X_{r,i} = X_i(f_r)$, $Z_{r,i} = \Re Z_i(f_r)$. Then we take, with $h = 1/n$,

$$g = 2hZ'Y, \qquad H = 2hZ'X. \tag{13}$$

We remark on a minor but useful modification, that terms in the finite sum that correspond to frequencies $f = 0$ and $f = \frac{1}{2}$ should be weighted by $1/2$. This ensures that the finite sum approximates the integral exactly for functions that are linear combinations of $\cos 2\pi\kappa f/n$ and $\sin 2\pi\kappa f/n$ for $\kappa = 0, 1, \ldots, n-1$. This modification is implemented by dividing the first row, and the last row if n is even, of Y, X and Z, by $\sqrt{\frac{1}{2}}$. This step also has the advantage of ensuring that all elements of Y have the same asymptotic variance under the true parameter values. We shall later use a vector of ones, modified in the same way.

To employ the statistic q, we shall need a consistent estimate of its distribution. First, however, we must draw attention to the fact that H is singular, and explain how we overcome this. In HTW we noted that $\beta' g = 0$ and $\beta' H = 0$ for the values of β used to construct g and H. The solution of $H\delta = g$ for the step δ therefore contains an arbitrary multiple of β. The reason for this singularity is that F is invariant to any scaling of S, and hence of β. In HTW we proposed a normalisation of β to obtain a unique estimate. However, because $q = g'\delta$, and $g'\beta = 0$, the value of the score statistic is not changed by the addition of any arbitrary multiple of β to δ. We therefore employ a generalised inverse of H in (11). In terms of its eigen-decomposition $H = WNW'$, we take $H^{-1} = WMW'$, where for the non-zero eigen-values, $M_{ii} = 1/N_{ii}$, but $M_{ii} = 0$ for $N_{ii} = 0$.

In the remainder of this section we investigate the sampling properties of q, and show that it may be approximated as a weighted sum of $(k-1)$ independent chi-squared variables C_i on 1 degree of freedom

$$q \approx \sum_{i=1}^{k-1} D_{i,i}\, C_i. \tag{14}$$

Our development begins by expressing, from (11),

$$q = 4h^2 \hat{Y}' Z H^{-1} Z' \hat{Y}, \tag{15}$$

where we have written $\hat{Y}$ in place of Y, to emphasise that q is formed at $\beta = \hat{\beta}$, so that $Y = \hat{Y} = S^*/\hat{S}$, the sample spectrum of the residuals from approximate maximum likelihood estimation of the model. Thus q is a quadratic form in the residual sample spectrum and hence also a quadratic form in the residual sample autocorrelations. We develop an approximate formula for this residual sample spectrum, as follows. In large samples, for which the likelihood may be accurately approximated by a quadratic about the true value of β, a single iteration only, of weighted least squares, is required, starting from β, to obtain the approximate MLE $\hat{\beta}$. Thus $\hat{\beta} = (X'X)^{-1}X'Y$, where Y is the sample spectrum S^*/S of the true model innovations e_t. Consequently we can, in large samples, take the weighted fitted values as

$$\frac{\hat{S}}{S} = X(X'X)^{-1}X'Y \overset{\text{def}}{=} UY. \tag{16}$$

We further approximate

$$\begin{aligned} \hat{Y} \; = \; \frac{S^*}{\hat{S}} \; &= \; 1 + \frac{S^* - \hat{S}}{\hat{S}} \\ &= \; 1 + \frac{S^* - \hat{S}}{S}\frac{S}{\hat{S}} \\ &= \; 1 + (I - U)Y\frac{S}{\hat{S}} \\ &\approx \; 1 + (I - U)Y \\ &= \; 1 + PY. \end{aligned} \tag{17}$$

In (17) we have first substituted for $\hat{S}/S$ from (16), then approximated $S/\hat{S}$ by its limit in probability, which is 1. The error in this approximation is neglected as being of second order. We obtain our final expression for q by noting, from HTW, that $Z'1 = 0$, where 1 is the modified vector of ones, and also $P\,1 = 0$, giving

$$q \approx 4h^2(Y-1)'PZH^{-1}Z'P(Y-1) = (Y-1)'ADA'(Y-1), \tag{18}$$

where D is a $(k-1)\times(k-1)$ diagonal matrix, and A is an orthonormal $m\times(k-1)$ matrix. These are constructed by using singular value decomposition to express $2h\,P\,Z = L\,\Delta\,R$, diagonalising the $k\times k$ matrix $\Delta R'H^{-1}R\Delta$ as EDE', and setting $A = L\,E$. The rank of D is at most $k-1$, by inheritance from H, so the conforming dimensions of D and A can be correspondingly reduced.

Now D and A are functions of $\hat{\beta}$. The diagonals of D will be consistently estimated constants, and the columns of A are consistently estimated functions of frequency. Each element of $A'(Y-1)$ is an orthonormal linear function of the mean-corrected spectral ordinates of white noise with unit variance, and so by the central limit theorem is asymptotically standard normal. The squares of these elements, C_i, are therefore each asymptotically chi-squared on 1 degree of freedom. A further statistic, mentioned above, is the invariant measure of the magnitude of the parameter step δ. This is equivalently, and more conveniently, considered as the invariant measure of the magnitude of the score $g = 2hZ'P(Y-1)$. It is given by $r = g'W^{-1}g$, where $W = \operatorname{var}(g)$, and a generalised inverse is used for W. We can express $r = (Y-1)'AA'(Y-1)$, the same as for q, but without the weighting. Its distribution is therefore chi-squared with degrees of freedom equal to the rank of A, which is at most $(k-1)$.

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

We here suppose that the series x_t is fitted by an IMA(1,1) model, so that $w_t = (1-B)x_t$ has spectrum parameterised as

$$S = \beta_1 + \beta_2(2 - 2\cos w), \tag{19}$$

which corresponds to taking x_t to be the sum of a random walk with innovation variance β_1 and independent white noise with variance β_2. In this case H is of rank 1, so the test statistic q is simply of the form $D_{1,1}C_1$. We can therefore take $q/D_{1,1}$ to be asymptotically chi-squared on 1 degree of freedom. For the rank 1 case, $q/D_{1,1}$ is the same as the r statistic. It may be shown that for this example, the statistic is asymptotically the square of a linear combination of the sample residual autocorrelations, with the coefficient of the autocorrelation at lag j being proportional to the coefficient of z^j in

$$(1-\eta z)^{-1}\left\{(1-\eta)(z+z^2+\cdots+z^{L-1})+z^L\right\}, \tag{20}$$

where η is the moving average parameter in the IMA(1,1) representation of the model. If η is close to 1, most of the weight will be placed on residual sample autocorrelations around lag L, but if η is close to 0, the weight will be spread out over the first L lags.

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 under the alternative hypothesis of an ARMA(1,1) model. 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 test based upon re-estimation of the model parameters. 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 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^2_J and D^2_J $(J = L-1)$. 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 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 of those considered by Tiao and Xu (1993) in their Table 3, where comparable maximum powers for the Ljung-Box statistic 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

Lead time	Nominal size	Length of series	η 0	0.2	0.4	0.6	0.8
2	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
	5 %	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	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
	5 %	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	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
	5 %	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 proposed test statistic at various lead times, series lengths and parameter values, with the $(1, \eta)$ model. 10,000 replications for each tabulated entry.

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. Our accurate or conservative empirical size may explain 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, such as $\phi \approx \theta$, 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

Nominal	Lead	(ϕ, θ)					
size	time	$(0.1, 0.0)$	$(0.4, 0.1)$	$(0.7, 0.4)$	$(0.9, 0.0)$	$(0.9, 0.4)$	$(0.95, 0.3)$
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
5 %	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 proposed test statistic at various lead times. 10,000 replications with series of length 200 for each tabulated entry. Testing the $(1, \eta)$ estimated model against the (ϕ, θ) data generating process.

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, page 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 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

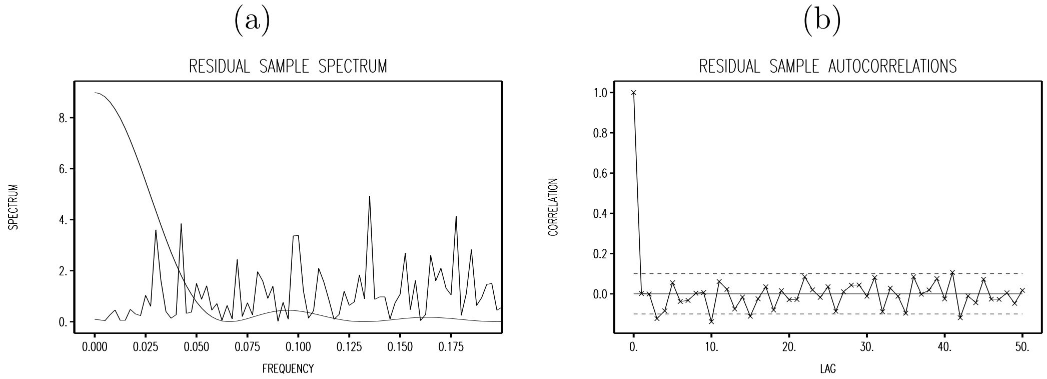

Figure 4: Analysis of the residuals obtained from MLE of an IMA(1,1) model for an AR(1) process with parameter 0.9. (a) The lower frequencies of the residual sample spectrum, with the (scaled) gain function $G(f)$ used to weight the spectrum in the MSS of 15 step-ahead prediction errors, (b) the sample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

at low lags. The (scaled) gain function $G(f)$ used in the expression (7) 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 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 a maximum lag of 21, were 0.18 and 0.30.

4 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 that 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 around frequency 0.05 (period 5 years). However, the fitted spectrum fails to resolve these peaks. Motivated by this we have investigated the power of our test for simulated series which might reflect a similar structure. We chose as an alternative model, a series composed of the sum of three independent components: a near unit root AR(1) process, an AR(2) process with a stochastic cycle of period 25, and white noise. Thus $x_t = u_t + v_t + z_t$ where $u_t = \phi u_{t-1} + a_t$, with $\phi = 0.99$ and var $(a_t) = 1.0$; $v_t = 2r\cos(\lambda)v_{t-1} - r^2 v_{t-2} + b_t$ with $r = 0.98$,

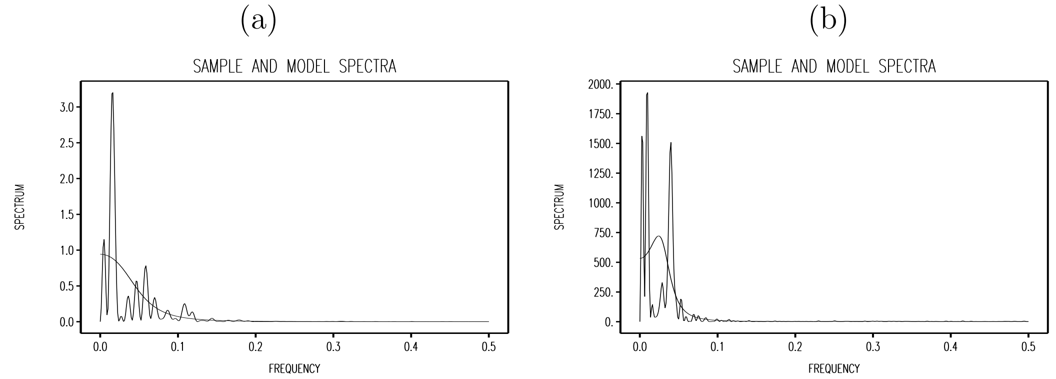

Figure 5: (a) The sample spectrum of the USA unemployment series with the AR(3) model spectrum superimposed; (b) the sample spectrum of a simulated series with an AR(6) model spectrum superimposed.

$\lambda = 2\pi/25$ 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 would be important to detect inadequacy of predictions. The frequency domain form of the AR(6) model spectrum S is given by the reciprocal linear expression $1/S = R = \sum_{j=0}^{6} \beta_j \cos(j\omega) = R_\beta \beta$. The frequency domain procedures of Section 2 only required modification similar to the use of an adjusted response in generalized linear models with a reciprocal link (see McCullagh and Nelder, 1989, Section 2.5).

The results of these simulations are shown in Table 3. The empirical sizes are shown for the nominal sizes of 10% and 5%, for the tests based on both our statistics q and r, and for the Ljung-Box test statistic calculated using a maximum lag of 20. For the q statistic the empirical sizes are slightly conservative for lead time 8, more so for lead time 16. The r statistic is so seriously over sized at lead time 8, as to make it unreliable for use in testing. We believe that the reason lies in the poor conditioning (apart from the exact singularity) of the variance matrix of the score g, used in its construction, and are unwilling to recommend its general use. The Ljung-Box test is slightly over-sized.

Lead time	Nominal size	q Statistic	r Statistic	Ljung-Box
8	10%	0.095	0.140	0.113
	5 %	0.045	0.093	0.056
16	10%	0.084	0.072	0.113
	5 %	0.035	0.036	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 applying it to the alternative model 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	10%	0.377	0.246
	5 %	0.274	0.144
16	10%	0.722	0.246
	5 %	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.

5 Conclusion

We have presented a diagnostic test for improved multi-step forecasting, which has wide applicability, together with reliable size and good power in a range of examples. The potential reduction in the multi-step forecast error that may be gained from re-estimation of the model parameters, justifies a test which is sensitive to this possibility. The methodology of the test also provides insight into the 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,

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.

Chevillon, G. and Hendry, D. F. (2004). Non-parametric direct multi-step estimation for forecasting economic processes. *Int. Journal of Forecasting*, to appear.

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

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.

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. (2004). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Unpublished Report*, Department of Economics, Harvard University and the NBER.

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.

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