ELSEVIER

International Journal of Forecasting 19 (2003) 229–239

international journal of forecasting

www.elsevier.com/locate/ijforecast

Statistically significant forecasting improvements: how much out-of-sample data is likely necessary?☆

Richard Ashley*

Department of Economics, Virginia Tech (VPI), 3027 Pamplin Hall, Blacksburg, VA 24061-0316, USA

Abstract

Testing the out-of-sample forecasting superiority of one model over another requires an a priori partitioning of the data into a model specification/estimation (‘training’) period and a model comparison/evaluation (‘out-of-sample’ or ‘validation’) period. How large a validation period is necessary for a given mean square forecasting error (MSFE) improvement to be statistically significant at the 5% level? If the forecast errors from each model are NIID and these errors are independent of one another, then the 5% critical points for the *F* distribution provide the answer to this question. But even optimal forecast errors from well-specified models can be serially correlated. And forecast errors are typically substantially crosscorrelated. For such errors, a validation period in excess of 100 observations long is typically necessary in order for a 20% MSFE reduction to be statistically significant at the 5% level. Illustrative applications using actual economic data are given.

Keywords: Postsample forecasting; Statistical testing

1. Introduction

Out-of-sample forecasting is known to be a rigorous check of the statistical adequacy of a model. Indeed, in the case of time series data, poorly specified models usually provide inferior out-of-sample forecasts than do naive — e.g., constant growth rate — models. Even reasonably well specified models typically forecast less well than their fit to the sample data would lead one to expect.

Thus, even a modest improvement in the out-of-sample mean square forecast error (MSFE) compared to that of an alternative formulation is generally taken to be strong evidence in favor of a model. Ratios of MSFE ratios are typically subject to substantial amounts of sampling error, however, so the question arises of whether a particular observed ratio is or is not significantly less than one.

This question has been explicitly considered in several strands of the time series forecasting literature. One approach begins with the observation (Morgan, 1939–1940; Granger & Newbold, 1977) that the difference in the squares of two time series equals the product of the time series formed by adding and subtracting these series:

$$x_t^2 - y_t^2 = (x_t + y_t)(x_t - y_t)$$

☆A more detailed version of this paper is available as VPI Economics Department Working Paper #E-00-06.

*Tel.: +1-540-231-6220; fax: +1-540-231-5097.

E-mail address: ashleyr@vt.edu (R. Ashley).

0169-2070/01/$ – see front matter © 2001 International Institute of Forecasters. Published by Elsevier Science B.V. All rights reserved.
PII: S0169-2070(01)00139-X

so that, if $\{x_t\}$ and $\{y_t\}$ is each a sequence of zero mean forecast errors,

$$\mathrm{var}(x_t) - \mathrm{var}(y_t) = \mathrm{cov}(x_t + y_t, x_t - y_t).$$

Recognizing that actual out-of-sample forecast error series are typically cross-correlated, serially correlated, and biased, this notion was later developed (Ashley, Granger & Schmalensee, 1980; Ashley, 1981) into a usable, albeit somewhat cumbersome, test of the null hypothesis that $\mathrm{MSFE}(x_t) = \mathrm{MSFE}(y_t)$ based on regressing $x_t - y_t$ against $x_t + y_t$. The attractiveness of this test is diminished, however, by its inherent limitation to settings in which a squared error loss function is appropriate and by the fact that it is only asymptotically justified, whereas feasible sequences of validation period forecasting errors are typically rather short.

Diebold and Mariano (1995) provide several tests of out-of-sample forecasting accuracy which at least relax the restriction to squared error losses. All of their tests are based on the observed loss differential, $g(x_t) - g(y_t)$, where $g(\cdot)$ is some specified loss function on the forecast errors. One of their tests is based on the estimated spectral density of the mean observed loss differential, evaluated at frequency zero. They also suggest a non-parametric test (the standard sign test) for assessing whether the median loss differential exceeds zero. This latter test would be exact for short model validation periods if $\{x_t\}$ and $\{y_t\}$ were both identically and independently distributed, but all of their tests are only asymptotically justified where (as is often the case) one or both forecast error series is serially correlated. For example, as noted below, serial correlation is common in multi-step-ahead forecasts; indeed, optimal h-step-ahead ARMA model forecasts are well known to be MA($h-1$). Even one-step-ahead forecasts are often serially correlated due to model mis-specification.[1]

In a third approach, Ashley (1998) develops a two-stage bootstrap algorithm for testing the null hypothesis that both forecast error series yield equal expected losses. This test allows for non-quadratic loss functions and for errors which are at once non-gaussian and both contemporaneously and serially correlated. The bootstrap approach — replacing the (unknown) true population distribution of variates by their (observed) empirical distribution — is itself only asymptotically justified, however.

That is the motivation for the second stage of the bootstrap algorithm: in this stage the bootstrap is used to estimate the degree of uncertainty which should be assigned to the first stage bootstrap inferences due to the fact that the number of validation period forecast errors available is not large. Thus, in this approach, the significance level at which the null hypothesis (of equal expected losses for both forecast error series) can be rejected is typically estimated (using the bootstrap) one hundred times, allowing one to use the interquartile range of these one hundred significance levels as a measure of the uncertainty in the median significance level.

Thus, for a given sequence of T out-of-sample forecast errors from each of two models ($\{x_t\}$ and $\{y_t\}$, $t = 1 \ldots T$) and a given loss function (g), it is now possible to more or less credibly test the null hypothesis that $E\{g(x_t)\} = E\{g(y_t)\}$. Regardless of which test one chooses, however, one must first select a value for T, the number of observations which will be withheld from the model specification/estimation process for use as a model validation period, but it is not clear how to make this selection.

The present paper addresses the resulting practical issues:

1. Given that one expects (hopes) to obtain an out-of-sample mean square forecasting error (MSFE) improvement of, say, 20%, how long a model validation period will be necessary, on average, in order for this MSFE improvement to be statistically significant at the 5% level?
2. Supposing that one can only ‘afford’ to withhold, say, $T = 40$ observations for model validation, how large an out-of-sample MSFE improvement will likely be necessary in order to conclude that this improvement was significant at the 5% level?

If the two forecast error series are known to be

[1] West (1996); West and McCracken (1998); Stock and Watson (1999); McCracken (2000); and Chao, Corradi and Swanson (2000) provide alternative asymptotically justified tests. Which, if any, of these large-sample tests is appropriate for use — e.g., correctly sized — in small model validation samples is not directly relevant here: the present paper examines how much out-of-sample data is necessary in order to reject the null hypothesis that $\mathrm{MSFE}(x_t) = \mathrm{MSFE}(y_t)$ at the 5% level in a selection of simple settings in which the distribution (under the null) of the sample MSFE ratio itself can be obtained by simulation.

gaussian, zero mean, i.i.d., *and* independent of each other, then these questions would be immediately answerable by reference to the usual tables of the 5% critical points of the F distribution. But these assumptions are typically substantially violated by actual sequences of forecast errors. For one thing, optimal multi-step-ahead forecasts are known to be serially correlated — e.g., see Granger and Newbold (1977, p. 121). And modest amounts of serial correlation are also commonly observed in sequences of one-step-ahead forecast errors, due to model misspecification and model instability. Moreover, forecast errors *even from optimal forecasts made using well-specified models* can be (and often are) notably contemporaneously cross-correlated.

Therefore, these two questions are addressed below using simulation methods for sequences of serially correlated, contemporaneously crosscorrelated forecast errors; both gaussian and non-gaussian error distributions are considered. The mechanics of generating pairs of simulated forecast error series with these characteristics are detailed in Section 2; simulated 5% critical points for testing the null hypothesis of equal MSFE for these series are given in Section 3.

These critical points are noticeably larger than one might expect, indicating that MSFE reductions well in excess of 20% will typically be necessary in order to be significant at the 5% level, even with model validation periods 80 to 100 observations in length. Put another way, these calculations indicate that a model validation period in excess of 100 observations long is typically necessary in order for a 20% MSFE reduction to be statistically significant at the 5% level. Illustrative examples using forecasting errors from models based on actual economic data are given in Sections 4 and 5.

2. Generation of correlated forecast error series

In order to calculate how large an observed MSFE ratio is required to be able to reject the null hypothesis that the population ratio is one, sequences of data with the specified serial and cross correlations are obtained via numerical simulation. For simplicity, each of these 'forecast error' series is generated with zero population mean, so the MSFE ratio reduces to a ratio of variances. (Forecast bias — while it can in some cases be important — is not ordinarily the main source of forecast MSE. Consequently, and since inference on second moments requires far more data than inference on first moments, the issue of how long a model validation sample needs to be usually hinges crucially on the ability to detect differences in error variance rather than differences in mean error.)

Both gaussian and non-gaussian (Student's t and truncated gaussian) variates are explicitly considered so as to gauge the sensitivity of the results to the shape of the tails of the error distribution. The generation of correlated AR(1) errors is described in some detail below; since optimal h-step-ahead forecast errors are known to follow an MA($h-1$) process, analogous results are also given for MA(2) errors.

In particular, let $\{x_t;\ t = 1\ldots T\}$ denote one series ('the out-of-sample forecast errors made by Model X') and let $\{y_t;\ t = 1\ldots T\}$ denote the corresponding 'out-of-sample forecast errors made by Model Y.' These forecast errors are assumed to be both contemporaneously crosscorrelated and serially correlated; specifically:

$$\rho_x = \mathrm{corr}(x_t, x_{t-1}) \tag{1}$$

$$\rho_y = \mathrm{corr}(y_t, y_{t-1}) \tag{2}$$

$$\rho = \mathrm{corr}(x_t, y_t) \tag{3}$$

The time series $\{x_t\}$ and $\{y_t\}$ are then generated from the stochastic recursion equations

$$x_t = \rho_x x_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim \text{i.i.d.}(0, 1) \tag{4}$$

$$y_t = \rho_y y_{t-1} + \gamma(\eta_t + w\,\varepsilon_t)$$
$$\eta_t \sim \text{i.i.d.}(0, 1) \tag{5}$$

where $\mathrm{corr}(\varepsilon_t,\ \eta_t) = 0$ and 'i.i.d.(0,1)' denotes 'identically and independently distributed with zero mean and unit variance.' The constants γ and w are chosen below so that $\mathrm{corr}(x_t,\ y_t) = \rho$ and so that (corresponding to the relevant null hypothesis) $\mathrm{var}(x_t) = \mathrm{var}(y_t)$.

Since γ and w are constants, it follows from Eqs. (4) and (5) that $\gamma(\eta_t + w\varepsilon_t)$ is identically and independently distributed with mean zero and variance equal to $\gamma^2\ (1 + w^2)$. Thus, $\{x_t\}$ and $\{y_t\}$ are ordinary (albeit correlated) AR(1) processes. Eqs.

(1) and (2) therefore follow from well-known results on such processes — e.g., Box and Jenkins (1976, pp. 56) or Granger and Newbold (1977, pp. 15).

Explicit expressions for γ and w imposing the restrictions that $\mathrm{corr}(x_t, y_t)=\rho$ and that $\mathrm{var}(x_t)=\mathrm{var}(y_t)$ follow from the MA(∞) forms of these two processes:

$$x_t = \sum_{j=0}^{\infty} \rho_x^j \varepsilon_{t-j} \tag{6}$$

$$y_t = \sum_{j=0}^{\infty} \rho_y^j \gamma(\eta_{t-j} + w\varepsilon_{t-j}) \tag{7}$$

which directly imply that

$$\mathrm{var}(x_t) = \frac{1}{1-\rho_x^2} \tag{8}$$

$$\mathrm{var}(y_t) = \frac{\gamma^2(1+w^2)}{1-\rho_y^2} \tag{9}$$

$$\mathrm{cov}(x_t, y_t) = \frac{\gamma w}{1-\rho_x\rho_y} \tag{10}$$

The resulting expression for the squared correlation between x_t and y_t yields

$$w = \sqrt{\frac{(1-\rho_x\rho_y)^2\rho^2}{(1-\rho_x^2)(1-\rho_y^2)-(1-\rho_x\rho_y)^2\rho^2}} \tag{11}$$

as the value of w for which the squared correlation between x_t and y_t equals ρ^2. Note that not every (ρ_x, ρ_y, ρ) combination yields a real value for w; this reflects the fact that it is not possible for x_t and y_t to be too highly correlated if ρ_x differs substantially from ρ_y.

The relevant null hypothesis here is that $\mathrm{var}(x_t)=\mathrm{var}(y_t)$. Solving Eqs. (8) and (9) for the value of γ which this implies,

$$\gamma = \sqrt{\frac{(1-\rho_y^2)}{(1+w^2)(1-\rho_x^2)}} \tag{12}$$

Thus, using Eqs. (11) and (12) to obtain the parameters w and γ, realizations of $\{x_t;\ t=1\ldots T\}$ and of $\{y_t;\ t=1\ldots T\}$ for a given serial and cross correlation structure specified by (ρ_x, ρ_y, ρ) can be obtained by recursively simulating Eqs. (4) and (5) using generated random variates for the innovation series, $\{\varepsilon_t\}$ and $\{\eta_t\}$. As noted at the beginning of this section, both gaussian and non-gaussian innovations are used in the simulations reported below. The gaussian variates were obtained using the Box–Muller method, as described in Press, Flannery and Teukolsky (1986, pp. 202–3). Truncated gaussian variates (truncated at ±2 standard deviations) were obtained using the acceptance/rejection method and scaling appropriately to restore unit variance. Variates from the Student's t distribution with 5 degrees of freedom were obtained directly from its definition, by generating in each case both a unit normal variate and an independent $\chi^2(5)$ variate.

Since optimal h-step-ahead forecast errors are known to follow an MA($h-1$) process, it is worth noting that, from Eqs. (6) and (7), the AR(1) processes considered above are essentially equivalent to large order MA processes with geometrically declining weights. Variates from MA(2) processes with non-declining weights were obtained as follows, letting

$$x_t = \varepsilon_t + \theta_x\varepsilon_{t-1} + \theta_x\varepsilon_{t-2} \tag{13}$$

$$y_t = \gamma_{ma}(\eta_t + w_{ma}\varepsilon_t) + \theta_y\gamma_{ma}(\eta_{t-1} + w_{ma}\varepsilon_{t-1}) + \theta_y\gamma_{ma}(\eta_{t-2} + w_{ma}\varepsilon_{t-2}) \tag{14}$$

So as to make the results based on these MA(2) processes more easily comparable to the AR(1) results, θ_x is chosen so that the R^2 for the x_t process equals that of an AR(1) process with given parameter ρ_x. This implies that

$$\theta_x = \sqrt{\frac{\rho_x^2}{2(1-\rho_x^2)}} \tag{15}$$

with an analogous relation obtaining for θ_y. Choosing w and γ, as before, to force the squared correlation between x_t and y_t to equal ρ^2 and the variance of x_t to equal that of y_t yields

$$w_{ma} = \sqrt{\frac{(1+2\theta_x^2)(1+2\theta_y^2)\rho^2}{(1+2\theta_x\theta_y)^2-(1+2\theta_x^2)(1+2\theta_y^2)\rho^2}} \tag{16}$$

and

$$\gamma_{ma} = \sqrt{\frac{(1+2\theta_x^2)}{(1+w_{ma}^2)(1+2\theta_y^2)}} \tag{17}$$

3. Simulation results and interpretation

Tables 1 through 4 report 5% critical points for the distribution of s_x^2/s_y^2, the ratio of the sample variance of one postsample forecast error series — $\{x_t;\ t = 1 \ldots T\}$ — to that of another — $\{y_t;\ t = 1 \ldots T\}$ — for given serial and cross correlations (ρ_x, ρ_y, and ρ) and validation period length, T. These critical points are calculated by simulating N pairs of forecast error series as described in the previous section, computing the sample variance ratio for each, and obtaining the 95% fractile of the N resulting sample variance ratios. N was increased until (at $N = 10^6$) the results stabilized.

In each table critical points are calculated using forecast error series generated with negligible (ρ_x, $\rho_y = 0.00$), moderate (ρ_x, $\rho_y = 0.50$), and severe (ρ_x, $\rho_y = 0.90$) levels of positive serial correlation; negative serial correlation is not considered since it is so atypical. Since the critical points are invariant to the sign of the cross correlation (ρ), results are given only for non-negative values of this parameter with no loss of generality. Critical values are presented in the tables for all feasible values of (ρ_x, ρ_y, and ρ) with ρ equal to 0.00, 0.50, and 0.90; a few combinations — e.g., (0.00, 0.90, 0.90) — are omitted since it is not possible for two series with such disparate degrees of serial correlation to be this highly correlated with one another.

Table 1 reports the results for forecast error series based on gaussian innovations. Since the sample variance in essence squares the forecast errors, the symmetry of the gaussian distribution is of little concern here. In contrast, the robustness of the results to varying assumptions as the shape of the tails of the innovation distributions is of considerable concern; consequently, Tables 2 and 3 report analogous results where the underlying innovations are gaussian variates truncated at ±2 standard deviations or Student's t variates with 5 degrees of freedom.[2]

For Tables 1 through 3, the forecast error serial correlation considered is of AR(1) form, which corresponds to an infinite order MA process with geometrically declining weights. Since optimal h-step-ahead forecasts are known to follow an MA($h-1$) process and, more broadly, to examine the robustness of the results to a departure from this particular pattern of serial correlation, results are given in Table 4 for forecast errors generated from an MA(2)

[2] These innovations are the $\{\varepsilon_t\}$ and $\{\eta_t\}$ series driving Eqs. (4) and (5) or (13) and (14). Note that, since the forecast errors $\{x_t\}$ and $\{y_t\}$ are weighted sums of these innovations, they follow the same distribution as the innovations only in the gaussian case.

Table 1
Out-of-sample error variance ratios — 5% critical points. Gaussian innovations — AR(1) processes

ρ_x	ρ_y	$\lvert\rho\rvert$	$T = 10$	$T = 20$	$T = 40$	$T = 80$	$T = 160$
0.00	0.00	0.00	3.18	2.16	1.70	1.45	1.29
0.00	0.50	0.00	4.36	2.67	1.94	1.57	1.36
0.00	0.90	0.00	16.66	8.11	4.52	2.86	2.05
0.50	0.00	0.00	2.75	2.08	1.71	1.48	1.32
0.50	0.50	0.00	3.75	2.54	1.93	1.59	1.39
0.50	0.90	0.00	14.19	7.60	4.41	2.85	2.06
0.90	0.00	0.00	1.14	1.33	1.54	1.59	1.50
0.90	0.50	0.00	1.54	1.60	1.69	1.68	1.55
0.90	0.90	0.00	5.67	4.47	3.52	2.72	2.12
0.00	0.00	0.50	2.77	1.96	1.59	1.38	1.25
0.00	0.50	0.50	3.68	2.38	1.80	1.49	1.32
0.50	0.00	0.50	2.33	1.86	1.58	1.40	1.28
0.50	0.50	0.50	3.22	2.27	1.79	1.51	1.33
0.50	0.90	0.50	10.35	6.00	3.77	2.56	1.91
0.90	0.50	0.50	1.19	1.31	1.47	1.50	1.43
0.90	0.90	0.50	4.69	3.78	3.05	2.40	1.91
0.00	0.00	0.90	1.69	1.41	1.27	1.18	1.12
0.50	0.50	0.90	1.84	1.52	1.35	1.23	1.16
0.90	0.90	0.90	2.31	2.04	1.81	1.58	1.40

Table 2
Out-of-sample error variance ratios — 5% critical points. Truncated gaussian innovations — AR(1) processes

ρ_x	ρ_y	$\lvert\rho\rvert$	$T = 10$	$T = 20$	$T = 40$	$T = 80$	$T = 160$
0.00	0.00	0.00	2.75	1.93	1.56	1.36	1.24
0.00	0.50	0.00	3.85	2.41	1.81	1.50	1.32
0.00	0.90	0.00	14.96	7.59	4.38	2.82	2.03
0.50	0.00	0.00	2.42	1.88	1.59	1.40	1.28
0.50	0.50	0.00	3.35	2.34	1.82	1.53	1.35
0.50	0.90	0.00	12.85	7.11	4.28	2.81	2.04
0.90	0.00	0.00	1.04	1.27	1.49	1.56	1.49
0.90	0.50	0.00	1.41	1.53	1.66	1.65	1.55
0.90	0.90	0.00	5.24	4.30	3.50	2.72	2.12
0.00	0.00	0.50	2.61	1.86	1.52	1.33	1.22
0.00	0.50	0.50	3.56	2.29	1.75	1.46	1.30
0.50	0.00	0.50	2.25	1.80	1.53	1.37	1.26
0.50	0.50	0.50	3.09	2.18	1.73	1.47	1.31
0.50	0.90	0.50	10.30	5.96	3.75	2.56	1.91
0.90	0.50	0.50	1.19	1.32	1.47	1.52	1.45
0.90	0.90	0.50	4.59	3.75	3.07	2.44	1.95
0.00	0.00	0.90	1.71	1.41	1.26	1.18	1.12
0.50	0.50	0.90	1.85	1.53	1.34	1.23	1.16
0.90	0.90	0.90	2.32	2.04	1.81	1.59	1.42

Table 3
Out-of-sample error variance ratios — 5% critical points. Student's *t* innovations — AR(1) processes

ρ_x	ρ_y	$\lvert\rho\rvert$	$T = 10$	$T = 20$	$T = 40$	$T = 80$	$T = 160$
0.00	0.00	0.00	4.50	2.98	2.23	1.81	1.54
0.00	0.50	0.00	5.99	3.59	2.48	1.93	1.61
0.00	0.90	0.00	22.03	10.12	5.39	3.29	2.29
0.50	0.00	0.00	3.79	2.79	2.18	1.80	1.55
0.50	0.50	0.00	5.05	3.34	2.43	1.92	1.61
0.50	0.90	0.00	18.41	9.35	5.20	3.24	2.28
0.90	0.00	0.00	1.51	1.68	1.82	1.79	1.66
0.90	0.50	0.00	1.99	1.99	2.00	1.89	1.71
0.90	0.90	0.00	7.18	5.35	4.07	3.02	2.30
0.00	0.00	0.50	3.12	2.28	1.83	1.56	1.39
0.00	0.50	0.50	3.93	2.61	1.98	1.63	1.42
0.50	0.00	0.50	2.47	2.03	1.73	1.52	1.37
0.50	0.50	0.50	3.54	2.55	2.00	1.67	1.45
0.50	0.90	0.50	10.37	6.08	3.82	2.60	1.94
0.90	0.50	0.50	1.19	1.34	1.50	1.52	1.46
0.90	0.90	0.50	5.00	4.01	3.23	2.55	2.02
0.00	0.00	0.90	1.66	1.40	1.27	1.19	1.13
0.50	0.50	0.90	1.81	1.51	1.34	1.23	1.16
0.90	0.90	0.90	2.29	2.03	1.80	1.58	1.40

process with equal weights on each of the two lagged innovations. So as to enhance the comparability of these results to those of the previous tables, in each case the two MA coefficients are chosen so as to yield an MA(2) process with R^2 equivalent to that of an AR(1) process with ρ_x or ρ_y equal to 0.00, 0.50, or 0.90.

The top row of Table 1 corresponds to the case of

Table 4
Out-of-sample error variance ratios — 5% critical points. Gaussian innovations — MA(2) processes

ρ_x	ρ_y	$\lvert\rho\rvert$	$T=10$	$T=20$	$T=40$	$T=80$	$T=160$
0.00	0.00	0.00	3.17	2.16	1.70	1.44	1.29
0.00	0.50	0.00	4.19	2.58	1.90	1.55	1.35
0.00	0.90	0.00	5.59	3.06	2.09	1.64	1.40
0.50	0.00	0.00	2.87	2.12	1.72	1.48	1.32
0.50	0.50	0.00	3.76	2.51	1.90	1.57	1.37
0.50	0.90	0.00	4.95	2.94	2.08	1.66	1.42
0.90	0.00	0.00	2.90	2.19	1.77	1.52	1.35
0.90	0.50	0.00	3.78	2.58	1.96	1.61	1.40
0.90	0.90	0.00	4.92	3.00	2.13	1.69	1.45
0.00	0.00	0.50	2.76	1.96	1.59	1.38	1.25
0.00	0.50	0.50	3.55	2.31	1.76	1.47	1.31
0.50	0.00	0.50	2.44	1.89	1.59	1.40	1.27
0.50	0.50	0.50	3.24	2.25	1.77	1.50	1.33
0.50	0.90	0.50	4.00	2.47	1.84	1.52	1.34
0.90	0.50	0.50	3.01	2.17	1.73	1.48	1.32
0.90	0.90	0.50	4.15	2.63	1.95	1.60	1.39
0.00	0.00	0.90	1.69	1.41	1.26	1.18	1.12
0.50	0.50	0.90	1.85	1.52	1.34	1.23	1.16
0.90	0.90	0.90	2.13	1.65	1.41	1.27	1.18

normally, identically, and independently distributed forecast errors from each model and an assumption that the errors made by one model are uncorrelated with those made by the other model. Here simulations are not actually necessary: for a validation period of length T, the relevant critical point is that of the $F(T-1,\ T-1)$ distribution. In this case we see that a validation period length of $T=10$ or $T=20$ will almost always be inadequate — in order to be significant at the 5% level one model's forecast error variance would need to be two to three times smaller than the other! Even at $T=80$, a 45% error variance reduction is required for significance at the 5% level.

Are similar error variance reductions necessary in the more realistic situation where the forecast errors are substantially correlated with one another, and perhaps serially correlated as well? Looking at the remaining rows of Table 1, the critical points clearly do vary with (ρ_x, ρ_y, ρ), but are in many cases even larger than for totally uncorrelated errors. The most notable exceptions are where ρ_x is 0.90 and T is very small (due to severe small-sample biases in s_x^2 with such data) and, more importantly, where ρ is fairly high. Nevertheless, even where ρ is 0.90, these results indicate that *substantial* variance reductions are usually necessary at $T=20$: in order to be significantly different at the 5% level, the sample variance of one forecast error series ordinarily needs to be at least 40% to 70% smaller than the other. Not until $T=80$ is a 20% error variance reduction significant at the 5% level, and then only if ρ is 0.90 and the serial correlation in both error series is modest.[3]

Turning to the results using non-gaussian innovations — truncated gaussian or Student's t with five degrees of freedom — the results are substantially similar, especially for highly correlated forecast errors. Apparently, over the parameter ranges considered here, the form and extent of the serial and cross correlation structure in the forecast errors is much more important than the tail shape of the underlying innovation distributions, even for small T.

[3] Mizrach (1992) shows analytically how the density function for the variance ratio of two sample variances depends on ρ^2 in the special case where $\rho_x = \rho_y = 0$. Ironically, the ρ^2 sensitivity of this density — which Mizrach (correctly) identifies as causing serious size distortion in statistical tests when the Theil U-statistic is (incorrectly) taken to be distributed as an F variate — here causes a smaller MSFE reduction to be significant at the 5% level. Note also that s_x^2 is biased in small samples for $\rho_x \neq 0$; that is why the critical point drops in some cases as N rises from 10 to 20.

Table 4 lists the results using MA(2) forecast errors. As one might expect from the foregoing, these critical points differ — in many cases considerably — from the analogous results using AR(1) processes of similar 'strength.' In particular, for small T the largest MA(2) critical points are substantially smaller and the smallest MA(2) critical points are substantially larger than for the AR(1) processes. This is probably due to differences in the small-sample biases in s_x^2 and s_y^2 using the MA(2) instead of the AR(1) process. The basic conclusions to be drawn from the MA(2) process critical point results, however, are essentially identical to those obtained with AR(1) errors: with validation periods ten to twenty periods long, one sample variance must still be at least several times smaller than that of another in order for the difference to be statistically significant at the 5% level unless the errors are quite strongly correlated, in which case a 40% to 60% variance reduction is still necessary.

4. Example: Univariate stock price forecasting models

Guerard and Thomakos (2000) compute several forecasts for the logarithmic growth rate of U.S. stock prices, as measured by the Dow Jones Industrial Average. They compare the accuracy of forecasts from an ARMA(0,1) model to that obtained from two naive models: a constant growth rate model and a a no-change (zero growth rate) model.

In all three cases the one-step-ahead forecast errors were computed using a rolling 64 quarter sample period, initially 1969I–1974IV; the resulting 96 quarter out-of-sample forecasting period was 1975I–1998IV. The forecast errors from all three models appear to be covariance stationary and serially uncorrelated.

Since the observed mean square forecasting error of the ARMA forecasts is smaller than that of either set of naive forecasts, Guerard and Thomakos proceeded to test the statistical significance of these MSE reductions using both the Diebold–Mariano and the bootstrap tests alluded to in Section 1 above. Test results on their forecast error series, in each case based on a squared error loss function, are summarized in Table 5.

Even with 96 out-of-sample observations, the 8% MSE reduction of the ARMA model forecasts compared to the constant growth rate model forecasts is not significant at even the 10% level on either test. In contrast, the 16% MSE reduction provided by the ARMA model compared to the no-change model is significant at the 6% level on the Diebold–Mariano test and at the 5% to 10% level on the bootstrap test.[4]

[4] The Ashley (1998) bootstrap test computes 100 significance levels so as to quantify the uncertainty in the inference significance level induced by the finite length of the validation period used. What is quoted here is the 25% and 75% fractiles of these 100 significance levels — i.e., in this case, the middle 50 computed significance levels lie between 0.05 and 0.10.

Table 5
ARMA(0,1) versus naive stock return forecasts

	Constant growth rate naive model	'No-change' naive model
Model validation period	1975I to 1998IV	1975I to 1998IV
N	96	96
Out-of-sample MSE reduction	8%	16%
ARMA-naive error correlation	0.83	0.83
Bootstrap test results		
median significance level	0.22	0.08
50% confidence[a]	[0.17, 0.26]	[0.05, 0.10]
Diebold–Mariano test results		
Significance level	0.22	0.06

[a] Endpoints are the 25 and 75% fractiles of the 100 significance levels generated by the Ashley (1998) bootstrap test algorithm for testing the MSE ratio.

Thus, the 16% MSE drop apparently is sufficient for statistical significance, at least at the 10% level.

Are these the kind of test results one might expect, given the simulated 5% critical points tabulated in the previous section? In addition to gaussianity, those critical points are for error variance ratios rather than MSE ratios, but bias is not an important factor in the sample MSE for any of these forecasts: the bias^2/MSE ratio is only 0.006, 0.033, and 0.114 for the ARMA, constant growth rate, and no-change models, respectively. In each case, the sample correlation of the ARMA forecast errors with the naive model errors is 0.83. Using the results given in Section 2 to generate a large number of gaussian sample pairs, each with sample length 96 and with (ρ_x, ρ_y, ρ) equal to (0.00, 0.00, 0.83), 5% of these pairs yielded sample variance ratios in excess of 1.21. Thus, assuming that the sample correlations are reasonably accurate and that the forecast errors are zero mean, white, and gaussian, a 21% MSE drop would be required for significance at the 5% level. This is consistent with an actual 16% MSE drop yielding significance at the 5% to 10% level on the statistical tests.

This example underscores a sobering feature of the results in Section 3 above: even a rather lengthy model validation period of nearly 100 observations is still barely sufficient to detect as significant an MSE drop of only 15% to 20%.

5. Example 2: Multivariate wheat market forecasting models

Robledo, Zapata and McCracken (2000) specify and estimate two cointegrated VAR models for the U.S. wheat market. These models are derived from a dynamic econometric model proposed by Chambers and Just (1981) in which the market for wheat (production, domestic wheat consumption, inventories, and exports) is linked to the U.S. macroeconomy. Robledo et al. (2000) analyze the forecasting effectiveness of two such models, one measuring U.S. wheat prices using the price of #2 soft red winter wheat at Chicago and the other using the price of #1 soft red winter wheat at Kansas City.

They compare rolling forecasts of a number of variables over several model validation periods and using a number of statistical tests, including the Diebold–Mariano and bootstrap tests. The particular result of interest here is a comparison of the domestic consumption forecast from the two models over the sixteen quarter period extending from 1996I to 1999IV.

The mean square error over this period in forecasting domestic consumption using the model based on Kansas City wheat prices is 19% smaller than that obtained using the model based on Chicago wheat prices. This MSE reduction is significant at the 0.3% to 3% level using the bootstrap test.[5] Since these errors appear to be zero mean, serially uncorrelated, and highly cross-correlated ($\rho = 0.99$), the 5% critical point for testing the significance of this MSE reduction based on the gaussian simulation formulas developed in Section 2 is only 1.14. Thus, it is not surprising that the observed 19% MSE reduction is apparently significant at the 5% level using the bootstrap test.

That is not the point of this example, however. When they plot these two forecast error series versus time, Robledo et al. find that these time series are quite likely not gaussian at all due to an outlying observation in each error series in 1996III. (The forecast errors for the model based on Chicago wheat prices are plotted in Fig. 1; a plot of the errors based on Kansas City wheat prices is quite similar.) In each case, the outlier is significant at the 0.02% level on the usual *t* test.

Thus, the gaussianity assumption underlying the 5% critical point calculation discussed above is probably not a useful approximation for these data. More importantly, however, this observation also implies that **no** statistical test of a null hypothesis involving the MSE for **either** model is sensibly applicable over this model validation period because **any** such test implicitly assumes that the observations in each forecast error series are at least

[5] As noted above, the Ashley (1998) bootstrap test quantifies the uncertainty in the inference significance level due to the finite length of the model validation period used by computing 100 significance levels; what is quoted here is the 25 and 75% fractiles of these 100 significance levels — i.e., the middle 50 computed significance levels lie between 0.003 and 0.030. Diebold–Mariano test results are not quoted since there are only 16 observations.

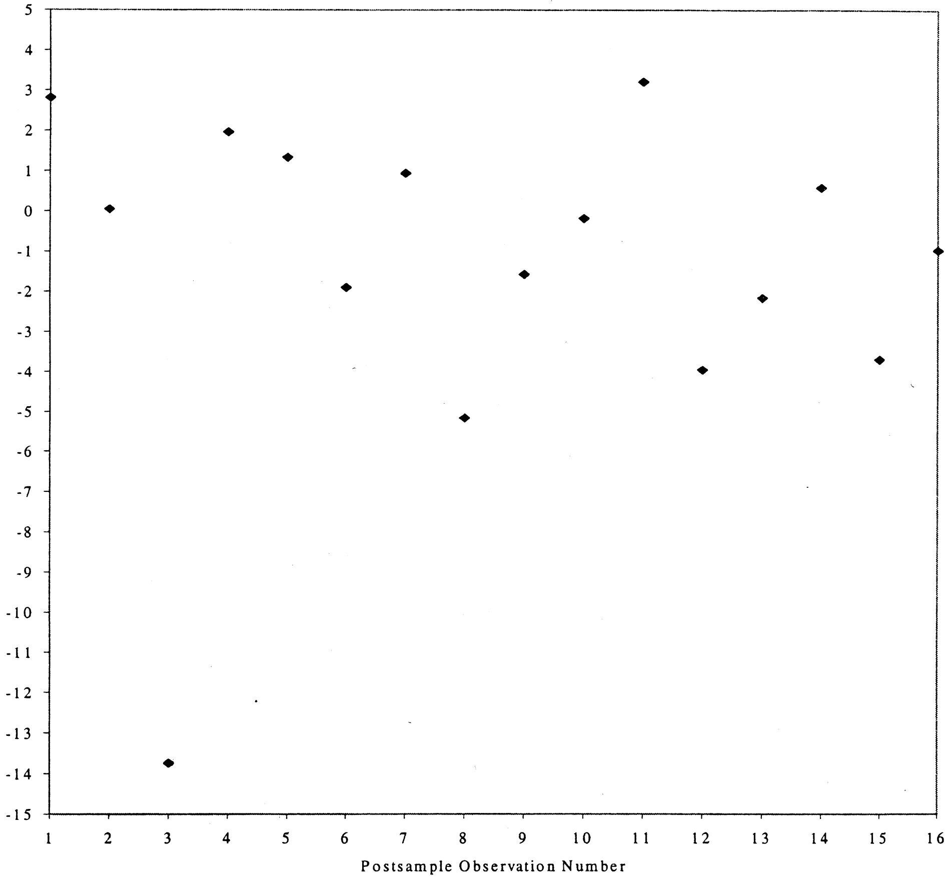

Fig. 1. Domestic wheat consumption forecast errors, 1996I–1999IV model based on Chicago wheat prices.

identically distributed, an assumption which is extremely suspect in this instance.[6]

Indeed, when this single pair of forecast error observations is dropped from the model validation period, the observed MSE drop — still seemingly substantial at 15% — is now only significant at the 3% to 14% level (with a median of 8%) on the bootstrap test. Since the sample crosscorrelation between the two forecast error series is now somewhat smaller, the 5% variance ratio critical point based on the gaussian simulation formulas from Section 2 rises to 1.26. Thus, consistent with the results from the bootstrap test, one might expect to need a 26% MSE drop with a model validation period like this in order for the reduction to be significant at the 5% level.

In addition to illustrating the wisdom of always first plotting the data one uses in any statistical test, this example also illustrates the general principle that any statistical result which seems too good to be true — in this case, a 19% MSE improvement being significant at the 5% level with only 16 out-of-sample observations — probably isn't.

[6] Of course, one can always interpret an outlying observation as an ordinary realization from a highly non-gaussian distribution. The bootstrap test, being non-parametric, appropriately accounts for such non-gaussianity in sufficiently large samples. In small samples, the presence of an outlier will yield a large dispersion in the bootstrap inference significance levels, as measured by the variance of these significance levels around their mean. However, since outlying observations are easily detectable by other — e.g., graphical — means, it is usually preferable to quantify the small-sample uncertainty in the bootstrap inferences using (as was done above) the interquartile range of the bootstrap inference significance levels.

6. Conclusions

The results reported here provide a kind of 'reality check' as to what sort of out-of-sample forecast error variance ratio (or MSFE ratio) is necessary or, if you like, how long a model validation period is necessary, in order for an observed forecasting improvement to be statistically significant at the 5% level.

This study has focused on how much model validation data is inherently necessary for successful inference, given that one could simulate the actual distribution of the relevant sample variance ratio. While the conclusions to be drawn from these simulations are necessarily limited to the distributions and processes considered, one result that emerges rather clearly is that model validation periods much less than 40 observations in length are not likely to be adequate. Even for model validation periods of length 40 to 80, it seems evident that the 25% to 35% forecast error variance reduction that one would ordinarily be quite satisfied to obtain from a model is not likely to be statistically significant unless the errors are substantially cross correlated and only modestly autocorrelated. Substantial cross correlation between out-of-sample forecast error series is actually common, however, so inference using model validation periods of length 40 to 80 is not hopeless.

In summary, as desirable as out-of-sample model validation is, it apparently is simply not feasible unless one can afford to devote to it at least 40 to 80 observations — and more if the errors are not expected to be strongly cross correlated. Indeed, it appears that mean square forecasting error reductions in excess of 20% are typically necessary for significance at the 5% level even with 100 out-of-sample observations.

References

Ashley, R. (1981). Inflation and the distribution of price changes across markets. *Economic Inquiry*, *XIX*, 650–660.

Ashley, R. (1998). A new technique for postsample model selection and validation. *Journal of Economic Dynamics and Control*, *22*, 647–665.

Ashley, R., Granger, C. W. J., & Schmalensee, R. (1980). Advertising and aggregate consumption: an analysis of causality. *Econometrica*, *59*, 817–858.

Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day: San Francisco.

Chambers, R. G., & Just, R. E. (1981). Effects of exchange rate changes on US agriculture: a dynamic analysis. *American Journal of Agricultural Economics*, *1*, 32–46.

Chao, J., Corradi, V., & Swanson, N. (2000). An out of sample test for Granger causality. Unpublished manuscript.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*, 253–265.

Granger, C. W. J., & Newbold, P. (1977). *Forecasting Economic Time Series*. Orlando, Florida: Academic Press.

Guerard, J. B., & Thomakos, D. D. (2000). 'Naïve', ARIMA, transfer function and VAR models: a comparison of forecasting performance. Unpublished manuscript.

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

Mizrach, B. (1992). The distribution of the Theil U-statistic in bivariate normal populations. *Economics Letters*, *38*, 163–167.

Morgan, W. A. (1939–1940). A test for the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika*, *31*, 13–19.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1984). *Numerical Recipes*. Cambridge: Cambridge University Press.

Robledo, C. W., Zapata, H. O., & McCracken, M. W. (2000). The predictive ability of two models of the U.S. wheat market. Unpublished manuscript.

Stock, J. H., & Watson, M. W. (1999). Forecasting Inflation. Working Paper 7023, National Bureau of Economic Research.

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

West, K. D., & McCracken, M. W. (1998). Regression based tests of predictive ability. *International Economic Review*, *39*, 817–840.

Biography: Richard ASHLEY (Ph.D. from UCSD, 1976) is Professor of Economics at Virginia Tech (V.P.I.) in Blacksburg, Virginia. His research interests include the statistical analysis of economic forecasts, the detection and modeling of nonlinear serial dependence in time series, and the detection and modeling of frequency dependence in econometrically estimated relationships.