Multifold Predictive Validation in ARMAX Time Series Models

Daniel PEÑA and Ismael SÁNCHEZ

This article presents a new procedure for multifold predictive validation in time series. The procedure is based on the so-called "filtered residuals," in-sample prediction errors evaluated in such a way that they are similar to out-of-sample ones. The filtered residuals are obtained from parameters estimated by eliminating from the estimation process the estimated innovations at the points to be predicted. Thus, instead of using the deletion of observations to validate the predictions, as in classical cross-validation, the procedure is based on deletion of the estimated innovations. It is proved that the filtered residuals are uncorrelated, up to terms of small order, with the in-sample innovations, a property shared with the out-of-sample residuals. The parameters needed for computing the filtered residuals can be obtained by estimating a model with innovational outliers at the points to be predicted. The proposed multifold predictive validation is asymptotically equivalent to an efficient model selection procedure. Some Monte Carlo evidence of the performance of the procedure is presented, and the application is illustrated in an example.

KEY WORDS: Cross-validation; Forecast accuracy; Model selection; Outlier; Prediction error; Predictive validation; Split-sample validation.

1. INTRODUCTION

One of the best-known methods for assessing the predictive ability of a model, or predictive validation, is by means of out-of-sample prediction errors. These out-of-sample prediction errors have been used in many problems, including selection among discriminant rules, estimation of the mean squared prediction error (MSPE), and selection of time series models. For independent data, a standard way to perform predictive validation with out-of-sample forecasts is cross-validation (Stone 1974; Allen 1974). Cross-validation is usually applied by dividing the data into two subsamples, the first used for model fitting and the second used for model validation. Then new partitions are selected, the process is repeated, and some criterion, such as the minimum MSPE, is used to estimate the prediction error of a given model (see, e.g., Burman 1989; Zhang 1993).

An important aspect in cross-validation is that the observations should alternate, or "cross," their roles. Each data point can then be used for estimating the parameters and for computing out-of-sample forecasts. However, with time series, we need to use the past to forecast the future, and this introduces clear restrictions in how the data can be used (Burman, Chow, and Nolan 1994). For this reason, prediction validation in time series is usually made through split-sample validation, in which the time series is divided in two subsamples, the first used for estimation and the second used to compute out-of-sample prediction errors (i.e., there is no "crossing").

The split-sample validation can be made in several ways (see, e.g., West 1996). The most popular scheme among practitioners is the so-called "rolling forecast," in which the splitting process is repeated, increasing recursively the estimation subsample observation by observation and decreasing the validation subsample accordingly. Then an estimation of the expected error criterion for a given horizon can be computed using the available prediction errors at that horizon. The out-of-sample prediction errors obtained by split-sample validation also have important drawbacks. First, the parameters of the model and the variance of the prediction error are both estimated with a fraction of the data, inducing a larger variance of the estimators. This added variance is called the "data-splitting variance." Second, the results of the split-sample validation depend on the initial partition, which is arbitrary. Third, the h-step-ahead prediction errors of rolling forecast are computed from models estimated with different sample sizes, and thus they cannot be compared easily.

When the objective of predictive validation is model selection, an alternative approach involves using model selection criteria. For ARMA(p, q) models, many of these criteria are based on minimization of functions of the form

$$G(p,q) = \ln \hat{\sigma}^2_{p,q} + (p+q)g(n), \tag{1}$$

where $p = 0, 1, \ldots, p^*$, $q = 0, 1, \ldots, q^*$, with p^* and q^* as some predetermined upper bounds, and $\hat{\sigma}^2_{p,q}$ is the maximum likelihood (ML) estimate of the residual variance of the fitted ARMA(p, q) model in a sample of size n. The penalty factor $g(n)$ is such that $g(n) \to 0$ when $n \to \infty$. If $g(n) = \ln(n)/n$, then (1) becomes the Bayes information criterion (BIC) (Schwartz 1978); if $g(n) = 2/n$, then we obtain the Akaike information criterion (AIC) (Akaike 1974), which is asymptotically equivalent to the final prediction error criterion (FPE) (Akaike 1969); and if $g(n) = c\ln(\ln n)/n$, then (1) is the Hannan and Quinn (1979) criterion (HQ). Some of these criteria have been generalized for h-step-ahead forecasting.

Model selection criteria are related to cross-validation and split-sample validation; in all cases we are assessing the prediction performance of the model in out-of-sample forecasting. Furthermore, Stone (1977), Rissanen (1986), Stoica, Eykhoff, Jansen, and Söderstrom (1986), and Kavalieris (1989) have shown the asymptotic equivalence between some cross-validation schemes and the AIC in selecting the order of an autoregression. Kavalieris (1989) has also shown the asymptotic equivalence between the rolling forecast and the BIC. Although asymptotically related, cross-validation and split-sample validation use information differently than model selection criteria. Cross-validation and split-sample validation are based on out-of-sample prediction errors, whereas model selection is based on some correction of the in-sample prediction errors. The in-sample prediction errors have the drawback of using the information twice, for estimating the parameters of the predictor and

Daniel Peña is Professor (E-mail: *daniel.pena@uc3m.es*) and Ismael Sánchez is Visiting Professor (E-mail: *ismael@est-econ.uc3m.es*), Department of Statistics, University of Carlos III de Madrid, Madrid, Spain.

© 2005 American Statistical Association
Journal of the American Statistical Association
March 2005, Vol. 100, No. 469, Theory and Methods
DOI 10.1198/016214504000000610

for computing the prediction errors. This data reuse decreases the possibility of detecting misspecifications and tends to select overparameterized predictors with lower values of residual variance, a problem analyzed by many authors (see, e.g., Efron 1986) and sometimes called "data-snooping bias" (White 2000). Thus some correction of the estimated in-sample variance is needed for model selection, as indicated by (1).

This article proposes an alternative multifold predictive validation procedure to evaluate the h-step-ahead prediction errors of a time series model. The procedure is based on the so-called "filtered residuals," which are in-sample prediction errors evaluated in such a way that they are similar to out-of-sample ones. This is obtained by eliminating from the estimation process the estimated innovations at the points to be predicted. It can be proved that with this proposal, the asymptotic covariance of the prediction and the innovation of the predicted point has order of magnitude $O(n^{-2})$, compared to $O(n^{-1})$ with classical residuals. Therefore, the data-snooping bias of traditional in-sample residuals is clearly diminished. When the prediction horizon and the number of initial values needed for the estimation are small compared to n, we can obtain almost as many prediction errors as the sample size, thus avoiding the data-splitting variance of the traditional out-of-sample prediction errors. Consequently, the estimated error variance does not suffer the loss of efficiency of rolling forecast or other split-sample validation procedures in time series. The estimated prediction error variance obtained with the filtered residuals is then very similar to the one obtained if we could compute n out-of-sample prediction errors. We show that the filtered residuals can be easily computed from a model that treats the points to be predicted as innovational outliers.

The rest of the article is organized as follows. In Section 2 we introduce the notation. In Section 3 we define the filtered residuals and prove that they are uncorrelated, up to terms of small order, with the in-sample innovations, a property shared with the out-of-sample prediction errors. We also prove that choosing among models by using the minimum mean squared filtered residuals (MSFRs) is equivalent to an efficient model selection procedure. In Section 4 we present a Monte Carlo study of the performance of the MSFR and illustrate the procedure with an application to real data. We give some final remarks in Section 5 and provide mathematical details and proofs of theorems in the Appendixes.

2. IN–SAMPLE, OUT–OF–SAMPLE, AND INTERPOLATED PREDICTION ERRORS

Let z_t follow an ARMAX model with known exogenous variables $x_{i,t}$, $i = 1, \ldots, k$, following the equation

$$\phi(B)z_t = \sum_{i=1}^{k} \eta_i(B)x_{i,t} + \theta(B)a_t, \tag{2}$$

where $\phi(B) = 1 - \sum_{j=1}^{p} \phi_j B^j$, $\eta_i(B) = \eta_{i0} - \sum_{j=1}^{s_i} \eta_{ij}B^j$, and $\theta(B) = 1 - \sum_{j=1}^{q} \theta_j B^j$. We assume that the random variables a_t are iid with mean 0 and variance σ^2 and that the roots of $\phi(B) = 0$, $\eta_i(B) = 0$, and $\theta(B) = 0$ are outside the unit circle. Let $\lambda = (\phi_1, \ldots, \phi_p, \eta_{10}, \ldots, \eta_{ks_k}, \theta_1, \ldots, \theta_q)'$ be the $m \times 1$ vector of structural parameters, with $m = p + k + \sum s_i + q$.

Suppose that an observed time series $\mathbf{Z}_n = (z_1, \ldots, z_n)'$ is represented by model (2). Let us denote the h-step-ahead forecast from origin t by $\hat{z}_{t+h} \equiv \hat{z}_{t+h}(\mathbf{Z}_t, \lambda) = E(z_{t+h}|\mathbf{Z}_t, \lambda)$, where the parameter vector λ is assumed known, $\mathbf{Z}_t = (z_1, \ldots, z_t)'$, $r \le t \le n - h$, and $z_1, \ldots, z_r$ is a set of initial values. Let us also define the population h-step ahead prediction error as $e_{t+h} = z_{t+h} - \hat{z}_{t+h}$. When λ is unknown, prediction errors can be defined in various ways. If $\hat{\lambda}_n = F(\mathbf{Z}_n)$ is some estimate of the parameters using the whole span of data, then the in-sample h-step-ahead forecast is $\hat{z}^{\text{in}}_{t+h} \equiv \hat{z}_{t+h}(\mathbf{Z}_t, \hat{\lambda}_n) = E(z_{t+h}|\mathbf{Z}_t, \hat{\lambda}_n)$, and the classical residuals, or in-sample one-step-ahead prediction errors, are given by $\hat{a}_{t+1} = z_{t+1} - \hat{z}^{\text{in}}_{t+1}$, $r \le t \le n - 1$. Similarly, the h residuals, or in-sample h-step-ahead prediction errors, are $\hat{e}^{\text{in}}_{t+h} = z_{t+h} - \hat{z}^{\text{in}}_{t+h}$, $r \le t \le n - h$, with $\hat{a}_{t+1} = \hat{e}^{\text{in}}_{t+1}$. By averaging the available in-sample prediction errors, we obtain an estimate of the in-sample MSPE as

$$\hat{V}^{\text{in}}(h) = \frac{\sum_{t=r}^{n-h} (\hat{e}^{\text{in}}_{t+h})^2}{n - h - r + 1}. \tag{3}$$

The out-of-sample prediction errors are based on the forecasts $\hat{z}^{\text{out}}_{t+h} \equiv \hat{z}_{t+h}(\mathbf{Z}_t, \hat{\lambda}_m) = E(z_{t+h}|\mathbf{Z}_t, \hat{\lambda}_m)$, where $\hat{\lambda}_m = F(\mathbf{Z}_m)$, $m \le t$ is some estimate of the parameters using a set of observations $\mathbf{Z}_m$ previous to z_{t+1} and are defined by $\hat{e}^{\text{out}}_{t+h} = z_{t+h} - \hat{z}^{\text{out}}_{t+h}(\mathbf{Z}_t, \hat{\lambda}_m)$, $r \le t \le n - h$. In this article we assume that $m = t$ and that all of the data up to z_t are included in the estimation of the parameters. This prediction procedure is known as rolling forecast (see, e.g., West 1996, for alternative schemes for obtaining out-of-sample prediction errors). Let n_h be the size of the initial estimation subsample. Then an estimate of the out-of-sample h-step-ahead MSPE is

$$\hat{V}^{\text{out}}(h) = \frac{\sum_{t=n_h}^{n-h} (\hat{e}^{\text{out}}_{t+h})^2}{n - h - n_h + 1}. \tag{4}$$

One advantage of $\hat{e}^{\text{out}}_{t+h}$ over $\hat{e}^{\text{in}}_{t+h}$ is that the former is free from the information in the observations to be predicted, $z_{t+1}, \ldots, z_{t+h}$. Another way to estimate the parameters without the effect of a block of observations is to assume that these observations are missing. Peña (1990) showed that the parameters obtained under this hypothesis are, for a large sample size, the same as those computed assuming additive outliers at these positions. The model that treats the block of observations $z_{T+1}, \ldots, z_{T+h}$ as missing is

$$\phi(B)\left(z_t - \sum_{j=1}^{h} w_j D_t^{(T+j)}\right) = \sum_{i=1}^{k} \eta_i(B)x_{i,t} + \theta(B)a_t, \tag{5}$$

where $D_t^{(t_0)} = 1$ if $t = t_0$ and $D_t^{(t_0)} = 0$ otherwise, and w_j, $j = 1, \ldots, h$, are the parameters corresponding to the variables $D_t^{(T+j)}$. Let $\hat{\lambda}_n^{\text{int}}$ be the estimated parameter vector λ in (5). The interpolated prediction errors are then given by $\hat{e}^{\text{int}}_{t+h} = z_{t+h} - \hat{z}^{\text{int}}_{t+h}$, $r \le t \le n - h$, with $\hat{z}^{\text{int}}_{t+h} \equiv \hat{z}_{t+h}(\mathbf{Z}_t, \hat{\lambda}^{\text{int}}_n)$. The estimator of the h-step-ahead MSPE using these errors is

$$\hat{V}^{\text{int}}(h) = \frac{\sum_{t=r}^{n-h} (\hat{e}^{\text{int}}_{t+h})^2}{n - h - r + 1}. \tag{6}$$

These prediction errors for $h = 1$ were used by Peña (1990) for building influence measures in time series. They are closely related to the conditional residuals (Haslett 1999) derived for

linear models with general covariance structure. It is important to note that, apart from the pure MA(h) case, the influence of the block of deleted observations is not completely discarded in these residuals, because the effect of the innovations $a_{T+1}, \ldots, a_{T+h}$ will be included in future observations when z_{T+k}, $k > h$ is correlated with the deleted observations.

3. FILTERED RESIDUALS

3.1 Definition

In this section we introduce a new type of prediction error in an ARMAX model that has some advantages with respect to the in-sample and out-of-sample prediction errors. Let $\mathbf{Z}_n$ be a vector of time series data following (2). To build intuition, consider first the AR(1) case, $z_t = \phi z_{t-1} + a_t$, with $|\phi| < 1$ and a_t white noise. Our interest is in evaluating the ability of this predictor to forecast future observations z_t, $t > n$, using the prediction errors computed in the sample $\mathbf{Z}_n$. Let z_{T+h}, $1 \le (T+h) \le n$ be an in-sample point to be predicted to estimate e_{T+h}. The out-of-sample approach to estimate e_{T+h} with information $\mathbf{Z}_T$ would first obtain the estimator $\hat{\phi}_T = \sum_2^T z_t z_{t-1} / \sum_2^T z_{t-1}^2$, then compute the predictor $\hat{z}_{T+h}^{\text{out}} = \hat{z}_{T+h}(\mathbf{Z}_T, \hat{\phi}_T) = \hat{\phi}_T^h z_T$, and finally obtain the estimated error $\hat{e}_{T+h}^{\text{out}} = z_{T+h} - \hat{\phi}_T^h z_T = e_{T+h} + (\phi^h - \hat{\phi}_T^h) z_T$. Note that the first and second terms are independent, because $e_{T+h} = a_{T+h} + \phi a_{T+h-1} + \cdots + \phi^{h-1} a_{T+1}$ and the innovations $a_{T+1}, \ldots, a_{T+h}$ are not included in the estimator $\hat{\phi}_T^h$. If we set, for example, $h = 1$, then it can be verified that $E(\hat{e}_{T+1}^{\text{out}})^2 \approx \sigma^2(1 + (T-1)^{-1})$. Because the MSPE of a new observation is $E(\hat{e}_{n+1})^2 \approx \sigma^2(1 + (n-1)^{-1})$, it can be concluded that the out-of-sample approach leads to an overestimation of this MSPE. To obtain the large-sample bias of $\hat{V}^{\text{out}}(1)$ as estimator of $E(\hat{e}_{n+1})^2$ we have that

$$E\big[\hat{V}^{\text{out}}(1)\big] \approx \sigma^2\left(1 + \frac{\sum_{t=n_1}^{n-1} 1/(t-1)}{n - n_1}\right) > \sigma^2\left(1 + \frac{1}{n-2}\right),$$

and the asymptotic bias of $\hat{V}^{\text{out}}(1)$ is positive. To better understand this bias, this expression can be approximated for large values of $n - n_1$ by

$$E\big[\hat{V}^{\text{out}}(1)\big] \approx \sigma^2\left(1 + \frac{\log(n) - \log(n_1)}{n - n_1}\right).$$

If we write n_1 as $n_1 = \alpha n$, $0 < \alpha < 1$, then the large-sample bias of $\hat{V}^{\text{out}}(1)$ can be written as

$$\text{Bias}\big[\hat{V}^{\text{out}}(1)\big] \approx \frac{\sigma^2}{n}\left(\frac{-\log\alpha}{1-\alpha} - 1\right),$$

which reveals that the positive bias will tend to increase exponentially as we reduce α, the portion of the sample used for estimation. If, on the other hand, we increase α, then the subsample $n(1-\alpha)$ used for the evaluation of forecasts will be smaller, increasing the variability of $\hat{V}^{\text{out}}(1)$. This is the difficult trade-off when using $\hat{V}^{\text{out}}$.

In contrast, the in-sample approach would first use an estimate $\hat{\phi}_n$ based on the n observations, then build $\hat{z}_{T+h}^{\text{in}} = \hat{z}_{T+h}(\mathbf{Z}_T, \hat{\phi}_n) = \hat{\phi}_n^h z_T$, and finally compute $\hat{e}_{T+h}^{\text{in}} = z_{T+h} - \hat{\phi}_n^h z_T = e_{T+h} + (\phi^h - \hat{\phi}_n^h) z_T$. The first and second terms are correlated, because $\hat{\phi}_n$ already contains, implicitly, the values $a_{T+1}, \ldots, a_{T+h}$. After some algebra, it can be verified that $E(\hat{e}_{T+1}^{\text{in}})^2 \approx \sigma^2(1 - (n-1)^{-1})$. The large-sample bias of $\hat{V}^{\text{in}}(1)$ as estimator of $E(\hat{e}_{n+1})^2$ is $\text{Bias}[\hat{V}^{\text{in}}(1)] = -2\sigma^2(n-1)^{-1}$, and it can be concluded that the in-sample approach underestimates the MSPE.

We could alternatively estimate the prediction error e_{T+h} by using an estimate that (a) does not include the information provided by $a_{T+1}, \ldots, a_{T+h}$, as in the case of $\hat{e}_{T+h}^{\text{out}}$, to avoid the bias of the in-sample residuals, but (b) does include the information provided by $a_{T+h+1}, \ldots, a_n$, as in $\hat{e}_{T+h}^{\text{in}}$, to improve the accuracy of the estimation. That is, instead of deleting observations to estimate the parameter, we would delete only the new information included in the observations to be predicted. Note that this is the idea behind cross-validation for independent data, because then deleting observations is equivalent to deleting the new information. However, for dependent data they are not equivalent, because the new information is just that which cannot be predicted from the past values of the time series. To avoid the new information, the parameter could be estimated by using a new time series of filtered values y_t that does not include the information provided by $a_{T+1}, \ldots, a_{T+h}$. To delete these innovations, we would first estimate the parameter as in the in-sample approach and compute $\hat{e}_{t+1}^{\text{in}} = z_{t+1} - \hat{\phi}_n z_t = \hat{a}_{t+1}$, $t \ge T+h$. Then we would use these residuals to build the filtered series y_t, as follows: For $t = 1, \ldots, T$, we have that $y_t = z_t$; for $t = T+1, \ldots, T+h$, we ignore the innovations and assume that $a_{T+1} = \cdots = a_{T+h} = 0$ and then $y_{T+j} = \hat{\phi}^j z_T$; and for $t > T+h$, the series y_t again takes into account the observed contemporaneous innovations, and then $y_t = \hat{\phi}_n y_{t-1} + \hat{a}_t$. A simpler procedure for disregarding the information provided by $a_{T+1}, \ldots, a_{T+h}$, which is faster to compute and can be interpreted as an iteration of the previous idea, is to assume that the innovations at these points are contaminated by outliers, which is equivalent to assuming innovative outliers at these positions. For instance, for the AR(1) model, this will imply estimating the model

$$z_t = \phi z_{t-1} + \sum_{l=1}^{h} w_l D_t^{(T+l)} + a_t,$$

which assumes innovative outliers at positions $T+1, \ldots, T+h$. It is well known (see, e.g., Chang, Tiao, and Chen 1988) that in this model $\hat{w}_l = z_{t+l} - \hat{\phi}^{\text{fil}} z_{t+l-1}$ and $\hat{\phi}^{\text{fil}} = \sum_A z_t z_{t-1} / \sum_A z_t^2$, where A is the set $(2, \ldots, T, T+h+1, \ldots, n)$. Thus in this model, $\hat{a}_{T+l} = 0$, $l = 1, \ldots, h$. Note that the information about a_{T+l} is not completely eliminated in the estimation of the parameter $\hat{\phi}^{\text{fil}}$, because $\hat{a}_{T+l}$ is only an estimation of the true innovation. However, as we prove in Theorem 1 in the next section, the prediction error $\hat{e}_{T+h}^{\text{fil}} = z_{T+h} - (\hat{\phi}^{\text{fil}})^h z_T$ has a similar behavior to the h-step-ahead out-of-sample prediction error. For instance, for $h = 1$, we have $\hat{e}_{T+1}^{\text{fil}} = a_{T+1} + (\phi - \hat{\phi}^{\text{fil}}) z_T$, and because we show that $E\{(\hat{\phi}^{\text{fil}} z_T) a_{T+1}\} = O(n^{-2})$, we have that $E(\hat{e}_{T+1}^{\text{fil}})^2 \approx \sigma^2(1 + (n-2)^{-1})$, and these filtered errors have an MSPE very close to the true one estimated with the complete sample, $E(\hat{e}_{n+1})^2$. Also, $E[\hat{V}^{\text{fil}}(1)] \approx \sigma^2(1 + (n-2)^{-1})$, and then the large-sample bias as estimator of $E(\hat{e}_{n+1})^2$ is $\text{Bias}[\hat{V}^{\text{fil}}(1)] = \sigma^2[(n-2)(n-1)]^{-1}$, which is of lower order of magnitude than its competitors. Theorem 2 in Section 3.3 extends this result to a more general situation.

In the general case, estimation of the parameters downweighting the information contained in the innovations $a_{T+1}, \ldots, a_{T+h}$ can be obtained by the model

$$\phi(B)z_t = \sum_{i=1}^{k} \eta_i(B)x_{i,t} + \theta(B)\left(a_t + \sum_{l=1}^{h} w_l D_t^{(T+l)}\right), \qquad (7)$$

where the $D_t^{(T+l)}$ are dummy variables as defined in (5) and w_l, $l = 1, \ldots, h$, are the parameters corresponding to these variables, $D_t^{(T+l)}$. Let us denote $\hat{\lambda}_n^{\text{fil}} = [(\hat{\phi}_n^{\text{fil}})', (\hat{\eta}_n^{\text{fil}})', (\hat{\theta}_n^{\text{fil}})']'$ as the parameter vector of ML or least squares (LS) estimates of λ using the model (7). Then, following Mann and Wald (1943), $\hat{\lambda}_n^{\text{fil}} \xrightarrow{p} \lambda$. The estimates $\hat{\lambda}_n^{\text{fil}}$ depend on T and h, but for simplicity we do not considered this in the notation. Let $\hat{z}_{T+h}^{\text{fil}} = \hat{z}_{T+h}(\mathbf{Z}_T, \hat{\lambda}_n^{\text{fil}})$ be the prediction of z_{T+h} from z_T using the estimated model

$$\hat{\phi}_n^{\text{fil}}(B)z_t = \sum_{i=1}^{k} \hat{\eta}_{ni}^{\text{fil}}(B)x_{i,t} + \hat{\theta}_n^{\text{fil}}(B)\hat{a}_t,$$

and let $\hat{e}_{T+h}^{\text{fil}} = z_{T+h} - \hat{z}_{T+h}^{\text{fil}}$ be the corresponding filtered residual. After estimating model (7) for $T = r, \ldots, n-h$, we have $n-h-r+1$ h-step-ahead filtered residuals. The average of these squared residuals is the MSFR, which will lead to the following estimate of the h-step-ahead MSPE:

$$\hat{V}^{\text{fil}}(h) = \frac{\sum_{t=r}^{n-h} (\hat{e}_{t+h}^{\text{fil}})^2}{n-h-r+1}. \qquad (8)$$

The relationship between filtered residual and innovative outliers makes computation of the MSFR straightforward. For $h = 1$, the computations are the same as in the standard procedure of checking for innovative outliers, and the estimation of w directly provides the value of the filtered residual. For $h > 1$, we just need to introduce a patch of h innovative outliers and use the estimated parameters to compute the h-step-ahead forecast. A Matlab program to compute the MSFR in a general ARMAX model can be downloaded from the authors' website.

3.2 Filtered Residuals as Out-Of-Sample Prediction Errors

The predictor $\hat{z}_{T+h}^{\text{in}}$, $r < (T+h) \le n$ verifies $E(\hat{z}_{T+h}^{\text{in}} \times a_{T+l}) \neq 0, l = 1, \ldots, h$, because of the influence of the innovations a_{T+l} in the estimation, whereas the out-of-sample predictor verifies $E(\hat{z}_{T+h}^{\text{out}} a_{T+l}) = 0$. The predictor $\hat{z}_{T+h}^{\text{fil}}$ is based on the estimator $\hat{\lambda}_n^{\text{fil}}$, which explicitly sets $\hat{a}_{T+l} = 0$, $l = 1, \ldots, h$. Because $\hat{a}_{T+l}$ are only estimates of the true innovations, their effect cannot be completely removed. However, we can prove that the influence of these innovations a_{T+l}, $l = 1, \ldots, h$, in $\hat{z}_{T+h}^{\text{fil}}$ is noticeably reduced. The following theorem shows that the covariance between the filtered predictor and the future innovations, $E(\hat{z}_{T+h}^{\text{fil}} a_{T+l})$, $l = 1, \ldots, h$, is of a lower order of magnitude than with the classical in-sample predictor. The proof is given in Appendix B.

Theorem 1. Let z_t follow the ARMAX model (2). Let $\hat{z}_{T+h}^{\text{in}}$ be the h-step-ahead predictor of z_{T+h}, where $r < (T+h) \le n$, and let $\hat{z}_{T+h}^{\text{fil}}$ be the filtered predictor. Then, for $l = 1, \ldots, h$,

$$\text{(a)} \quad E\{\hat{z}_{T+h}^{\text{in}} a_{T+l}\} = O(n^{-1})$$

and

$$\text{(b)} \quad E\{\hat{z}_{T+h}^{\text{fil}} a_{T+l}\} = O(n^{-2}).$$

This result then proves that the filtered residuals have similar properties to the out-of-sample prediction errors.

3.3 Estimation of the Out-Of-Sample MSPE With the Filtered Residuals and Efficient Model Selection

Let us denote by $V^{\text{pop}}(h) = E[(z_{n+h} - \hat{z}_{n+h})^2 | \mathbf{Z}_n, \hat{\lambda}_n]$ the out-of-sample h-step-ahead MSPE from predicting z_{n+h} with the ARMAX model (2) with parameter vector λ estimated with a root-n–consistent method using the whole sample Z_n. Estimation of $V^{\text{pop}}(h)$ is a rather challenging issue, because even an asymptotic approximation of a properly specified model is a nonlinear function of the parameter vector λ that changes with the horizon (see, e.g., Baillie 1980; Fuller and Hasza 1981; Yamamoto 1981; Kunitomo and Yamamoto 1985). In the ARMA model at $h = 1$, the asymptotic $V^{\text{pop}}(1)$ is

$$V^{\text{pop}}(1) = \sigma^2\left(1 + \frac{k}{n}\right) + O(n^{-3/2}), \qquad (9)$$

where k is the number of estimated parameters. Expression (9) is the basis of the FPE criterion; it is easy to handle because it needs only the estimation of σ^2. However, at larger horizons, $V^{\text{pop}}(h)$ depends on the dynamic of the model. For instance, in the case of a correctly specified AR(p) of mean 0, the asymptotic MSPE is (Fuller and Hasza 1981)

$$\begin{aligned} V^{\text{pop}}(h) = \sigma^2 &\sum_{i=0}^{h-1} (\mathbf{e}_p' \mathbf{B}^i \mathbf{e}_p)^2 \\ &+ \frac{\sigma^2}{n} \sum_{j=0}^{h-1} \sum_{s=0}^{h-1} (\mathbf{e}_p' \mathbf{B}^j \mathbf{e}_p)(\mathbf{e}_p' \mathbf{B}^s \mathbf{e}_p) \\ &\times \operatorname{trace}(\mathbf{B}^{h-1-j} \boldsymbol{\Gamma}_z \mathbf{B}^{h-1-s} \boldsymbol{\Gamma}_z^{-1}), \\ &+ O(n^{-3/2}), \end{aligned} \qquad (10)$$

where $\boldsymbol{\Gamma}_z = E[(z_t, z_{t-1})'(z_t z_{t-1})]$, $\mathbf{e}_p = (1, 0, \ldots, 0)'$ with dimension $p \times 1$ and

$$\mathbf{B} = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ & & \mathbf{I}_{p-1} & & \mathbf{0} \end{pmatrix},$$

where $\mathbf{I}_p$ is the identity matrix of size p and $\mathbf{0}$ is a vector of 0's of appropriate dimension. This expression requires the true parameter values $\mathbf{B}, \sigma^2$, and $\boldsymbol{\Gamma}_z$. An estimator of V^{pop} in this AR(p) case may be obtained from (10) by ignoring the remainder term and replacing $\mathbf{B}, \sigma^2$, and $\boldsymbol{\Gamma}_z$ by their estimators $\hat{\mathbf{B}}, \hat{\sigma}^2$, and $\hat{\boldsymbol{\Gamma}}_z$. Then a plug-in estimator is

$$\begin{aligned} \hat{V}^{\text{plug}}(h) = \hat{\sigma}^2 &\sum_{i=0}^{h-1} (\mathbf{e}_p' \hat{\mathbf{B}}^i \mathbf{e}_p)^2 \\ &+ \frac{\hat{\sigma}^2}{n} \sum_{j=0}^{h-1} \sum_{s=0}^{h-1} (\mathbf{e}_p' \hat{\mathbf{B}}^j \mathbf{e}_p)(\mathbf{e}_p' \hat{\mathbf{B}}^s \mathbf{e}_p) \\ &\times \operatorname{trace}(\hat{\mathbf{B}}^{h-1-j} \hat{\boldsymbol{\Gamma}}_z \hat{\mathbf{B}}^{h-1-s} \hat{\boldsymbol{\Gamma}}_z^{-1}). \end{aligned} \qquad (11)$$

In more general models, the estimator $\hat{V}^{\text{plug}}(h)$ requires computation of an even more complex function (see, e.g., the appendix

in Baillie 1980). Despite the difficulty in estimating $V^{\text{pop}}(h)$, its computation would provide very useful information for a variety of purposes, including extending the FPE criterion for $h > 1$ in a general ARMAX situation, building asymptotic confidence intervals, and assessing the predictability at horizon $h \geq 1$ of a series using R^2-like measures (see, e.g., Pierce 1979; Granger and Newbold 1986, p. 310; Diebold and Kilian 2001).

From the result in Theorem 1, the estimator $\hat{V}^{\text{fil}}(h)$, defined in (8) can be used as an estimate of $V^{\text{pop}}(h)$, without dismissing the possibility of using alternative loss functions. The MSFR is much easier to compute than $\hat{V}^{\text{plug}}(h)$, because no theoretical formulae are required, and has a very low bias, as shown in the following theorem.

Theorem 2. Let the conditions of Theorem 1 hold. Then

$$E[\hat{V}^{\text{fil}}(h)] = V^{\text{pop}}(h) + O(n^{-3/2}). \tag{12}$$

It is worth remarking that Theorem 2 relies on the ability of filtered residuals to mimic out-of-sample prediction errors, as shown in Theorem 1. A corollary of Theorem 2 is that, because $\hat{V}^{\text{fil}}(h)$ is asymptotically equivalent to $V^{\text{pop}}(h)$, minimizing the MSFR is asymptotically equivalent to selecting the most efficient predictor, at horizon h, among a set of competing models.

From Theorem 2 and expression (9), it can be seen that minimizing the MSFR is asymptotically equivalent to the FPE criterion for $h = 1$. Note also that the MSFR is, for pure autoregressive processes and $h = 1$, closely related to the cross-validation scheme described Stone (1974), Allen (1974), and Stoica et al. (1986). For instance, for pure autoregressive processes $\hat{V}^{\text{fil}}(1)$ is equivalent to the PRESS criteria (Allen 1974). The use of the MSFR for model selection, then can be interpreted as the generalization of the PRESS criteria to h-step prediction with ARMAX models.

4. EMPIRICAL COMPARISON

4.1 Preliminaries

In this section we illustrate the performance of the proposed procedure in finite samples using both a Monte Carlo experiment and some real data. We first compare the empirical bias and mean squared error (MSE) of alternative estimators of the population out-of-sample h-step-ahead MSPE ($V^{\text{pop}}(h)$). This experiment will allow to check how the asymptotic results of Theorem 2 apply in finite samples, comparing the bias of $\hat{V}^{\text{fil}}(h)$ with in-sample and out-of-sample approaches. The empirical results will also allow us to compare the precision of the alternative approaches, computing the MSE of the estimators, and showing the high inefficiency of the frequently used split-sample validation procedures. We use both univariate and ARMAX models in the comparison. It is important to find reliable predictive validation procedures in ARMAX models, because there is a higher risk of misspecification, falling into spurious relationships between the variables. We also include in the simulation a nonlinear ARMAX model, which can illustrate the behavior of the proposed procedure when the underline process does not belong to the ARMAX family.

In a second part of the experiment, we also evaluate the relative performance of the alternative estimators as validation criteria for model selection for a prediction horizon h of interest. We also include in the comparison popular criteria, such as AIC, BIC, HQ (with $c = 3$), and FPE. The alternative information criteria, as well as $\hat{V}^{\text{fil}}(h)$, have been built using asymptotic arguments, so it is useful to compare them in small and moderate samples. Finally, we illustrate some application of the proposed procedure with real data. To this aim, we have chosen a dataset that is easily available in the literature as the gas furnace data of Box and Jenkins (1976, p. 381).

Four types of prediction errors are considered in the comparison. The first are the in-sample prediction errors, $\hat{e}^{\text{in}}_{t+h}$, obtained from LS estimation of the predictor $\hat{z}^{\text{in}}_{t+h}$, and from them we compute two estimators of $V^{\text{pop}}(h)$, the average of squared prediction errors, $\hat{V}^{\text{in}}(h)$, and the average corrected by degrees of freedom, $\hat{V}^{\text{inc}}(h)$. For instance, for an AR(p), they are computed by

$$\hat{V}^{\text{in}}(h) = \frac{\sum_{t=p}^{n-h} (\hat{e}^{\text{in}}_{t+h})^2}{n-h-p+1} \tag{13}$$

and

$$\hat{V}^{\text{inc}}(h) = \frac{\sum_{t=p}^{n-h} (\hat{e}^{\text{in}}_{t+h})^2}{n-h-2p+1}. \tag{14}$$

The second type of prediction errors considered are out-of-sample prediction errors obtained using the rolling forecast. To compute these, the estimation subsample increases recursively, and the model is reestimated by LS including all data prior to the forecast origin. We compute two estimates of $V^{\text{pop}}(h)$, one from $\hat{e}^{\text{o50}}_{t+h}$, where the initial estimation subsample has size $[.5n]$ and $[\cdot]$ denotes the integer part, and one from $\hat{e}^{\text{o75}}_{t+h}$, where the initial estimation subsample has size $[.75n]$. The goal of including both prediction errors is to compare the effect of different initial subsamples. In both cases, the MSPE is estimated by averaging the available squared prediction errors, and the estimates are denoted by $\hat{V}^{\text{o50}}(h)$ and $\hat{V}^{\text{o75}}(h)$. In the case of an AR(p), they are given by

$$\hat{V}^{\text{o50}}(h) = \frac{\sum_{t=[.5n]}^{n-h} (\hat{e}^{\text{o50}}_{t+h})^2}{n-h-[.5n]+1} \tag{15a}$$

and

$$\hat{V}^{\text{o75}}(h) = \frac{\sum_{t=[.75n]}^{n-h} (\hat{e}^{\text{o75}}_{t+h})^2}{n-h-[.75n]+1}. \tag{15b}$$

The third and fourth types of prediction errors considered are the interpolated prediction errors $\hat{e}^{\text{int}}_{T+h}$ and the proposed filtered residuals $\hat{e}^{\text{fil}}_{T+h}$, which are also obtained using LS estimation. The corresponding MSPEs are estimated with (6) and (8), and we denote them by $\hat{V}^{\text{int}}(h)$ and $\hat{V}^{\text{fil}}(h)$.

We performed two experiments. In the first experiment, the true data-generating process (DGP) is the AR(3): $(1-2B)(1-.5B)(1-.7B)z_t = a_t$, with $a_t \sim \text{N}(0,1)$. In each simulated series, we fitted AR(p) models of order $p = 1, 2, \ldots, 6$. In the second experiment, the DGP is the model is $y_t = .9y_{t-1} + .7x_{1t} + .2x_{1t-1} + \gamma x_{1t}x_{1t-1} + a_t$ with $\gamma = 0$ (linear ARX case) and $\gamma = .1$ (nonlinear case). The independent variable x_{1t} is a sequence of iid random variables with distribution $x_{1t} \sim \text{N}(0,1)$.

In each replication of this model, we fit the following alternative linear models:

$$\text{M1}: \quad y_t = c + \alpha y_{t-1} + \beta_1 x_{1t} + \beta_2 x_{1t-1} + a_t,$$

$$\text{M2}: \quad y_t = c + \alpha y_{t-1} + \beta_1 x_{1t} + a_t,$$

$$\text{M3}: \quad y_t = c + \alpha y_{t-1} + \beta_1 x_{1t} + \beta_2 x_{1t-1} + \beta_3 x_{2t} + a_t,$$

$$\text{M4}: \quad y_t = c + \alpha y_{t-1} + \beta_1 x_{1t} + \beta_2 x_{1t-1} + \beta_3 x_{2t} + \beta_4 x_{3t} + a_t,$$

and

$$\text{M5}: \quad y_t = c + \sum_{i=1}^{5} \phi_p y_{t-p} + a_t,$$

where x_{2t} and x_{3t} are independent of x_{1t} but are also iid and N(0, 1).

For each Monte Carlo replication, we generated $200 + n + 5$ data values and dropped the first 200 data values to ensure stationary initial conditions. Of the remaining $n + 5$ data points, we used the first n to estimate the parameters of each model and considered the last five observations as future observations. We considered two sample sizes, $n = 25$ and 100. In the first experiment, the prediction horizon is $h = 1, 3$, and in the second experiment it is $h = 1, 3, 5$. For each experiment, model, and sample of size n, we computed the statistics $\hat{V}^{\text{in}}(h)$, $\hat{V}^{\text{inc}}(h)$, $\hat{V}^{\text{o50}}(h)$, $\hat{V}^{\text{o75}}(h)$, $\hat{V}^{\text{int}}(h)$, and $\hat{V}^{\text{fil}}(h)$, as well as the AIC, BIC, HQ, and FPE. For $h = 1$, we used the AIC, BIC, and HQ as in (1), with $\hat{V}^{\text{in}}(1)$ as the ML estimator of σ^2. The FPE uses $\hat{V}^{\text{inc}}(1)$ as an unbiased estimate of σ^2. We also used those criteria at $h > 1$ using (1) but replacing $\hat{V}^{\text{in}}(1)$ by $\hat{V}^{\text{in}}(h)$ and $\hat{V}^{\text{inc}}(1)$ by $\hat{V}^{\text{inc}}(h)$. By this, we intend not to propose new definitions of those criteria for h-step-ahead forecasting (see Hurvich and Tsai 1997 for generalizations of AIC along these lines), but rather to use them as simple approximate criteria for comparison purposes. We used remaining five observations to compute the population out-of-sample MSPE, $V^{\text{pop}}(h)$.

We ran each experiment twice. The first run was designed for estimation of the population MSPE for each model, $V^{\text{pop}}(h)$; the second, for comparing different estimates of this MSPE. In the first run, we made 100,000 replications. In each replication we used the n observations to estimate the competing AR(p) models in the first experiment and the models M1–M5 in the second experiment. With each estimated model, we predicted the observations $n + 1$ to $n + 5$. Then we used the five future observations to estimate the population out-of-sample MSPE by averaging the squared prediction errors. In this way we obtained the empirical MSPE for each model, reported in Tables 1 and 2 denoted by $V^{\text{pop}}(h)$. The figures in bold type correspond to the model with the lowest MSPE. In the second run, we generated 5,000 replications and obtained, in each replication, the estimates $\hat{V}^{\text{in}}(h)$, $\hat{V}^{\text{inc}}(h)$, $\hat{V}^{\text{o50}}(h)$, $\hat{V}^{\text{o75}}(h)$, $\hat{V}^{\text{int}}(h)$, and $\hat{V}^{\text{fil}}(h)$. By comparing these estimates with $V^{\text{pop}}(h)$, we can evaluate their bias, $[\hat{V}(h) - V^{\text{pop}}(h)]$, and MSE $[\hat{V}(h) - V^{\text{pop}}(h)]^2$, as estimators of $V^{\text{pop}}(h)$ by averaging these values over the 5,000 replications.

The results regarding bias and MSE of these statistics are also reported in Tables 1 and 2. We use all these estimators of $V^{\text{pop}}(h)$, together with AIC, BIC, HQ, and FPE, as model selection criteria and compute the proportion of times that each model has been selected by each criterion within the 5,000 replications. We summarize the conclusions of the experiments in the following sections. For the sake of clarity, and because HQ has similar performance to BIC in these experiments, we do not report their results. For the same reason, and to ease comparisons, in the second experiment we do not report the results on $\hat{V}^{\text{int}}(h)$ and of sample size $n = 25$, because they are qualitatively similar to those of the first experiment.

4.2 Bias and Mean Squared Error of the Competing Estimators

Table 1 gives the empirical bias of the alternative estimators of $V^{\text{pop}}(h)$ in the first experiment. As expected, $\hat{V}^{\text{in}}(h)$ has a large negative bias, followed by $\hat{V}^{\text{int}}(h)$ and $\hat{V}^{\text{inc}}(h)$. The negative bias grows with p. This is the aforementioned data-snooping bias, which will favor the selection of overparameterized models. In contrast, the estimators based on out-of-sample residuals $\hat{V}^{\text{o50}}(h)$ and $\hat{V}^{\text{o75}}(h)$ tend to have positive bias, especially in small samples. This positive bias grows with p and is due to the smaller sample size used in estimation of the parameters. This effect is part of the aforementioned data-splitting variance of the split-sample validation methods, which will favor the selection of smaller models. The bias of the proposed $\hat{V}^{\text{fil}}(h)$ is in general very low: it is the smallest for $n = 25$, and similar to $\hat{V}^{\text{o75}}(h)$ but larger than $\hat{V}^{\text{o50}}(h)$ for $n = 100$.

The importance of the bias can be better appreciated by noting that the differences between the true $V^{\text{pop}}(h)$ of the models under consideration are often smaller than the bias of the estimators. It is thus essential to use estimates with very low bias. For instance, in Table 1 with $n = 100$ and $h = 1$, the most efficient model is the AR(2), with $V^{\text{pop}}(1) = 1.01$, and the least efficient is the AR(6), with $V^{\text{pop}}(1) = 1.05$. The difference between the variance of both predictors is .04. However, if we use the estimate $\hat{V}^{\text{in}}(h)$ to compare these models, then the average value that we will obtain for the $V^{\text{pop}}(1)$ of the AR(2) will be 1.01 plus the bias (i.e., $1.01 - .046 = .964$), whereas for the AR(6), it will be $1.05 - .129 = .921 < .964$. Therefore, the AR(6) will be chosen. Also, by using $\hat{V}^{\text{inc}}(h)$, we would erroneously conclude that both models are equally accurate. However, if we use the proposed $\hat{V}^{\text{fil}}(h)$, then we will obtain $1.01 - .005 = 1.005$ for the AR(2) and $1.05 + .001 = 1.051$, clearly showing the advantage of the efficient predictor.

Table 2 shows the bias in the second experiment. The conclusions that we draw are similar to those of the first experiment in both the linear ($\gamma = 0$) and the nonlinear cases ($\gamma = .1$). The in-sample procedures $\hat{V}^{\text{in}}(h)$ and $\hat{V}^{\text{inc}}(h)$ have a large negative bias that increases with the size of the model. The bias is especially important in M5, where we fit an AR(5). The out-of-sample procedures $\hat{V}^{\text{o50}}(h)$ and $\hat{V}^{\text{o75}}(h)$ have positive bias, with $\hat{V}^{\text{o50}}(h)$ showing in general a lower performance, especially in the nonlinear case. The bias of the proposed $\hat{V}^{\text{fil}}(h)$ is the smallest of all the estimators compared.

Regarding the MSE, the results of the first experiment in Table 1 show that the out-of-sample estimates $\hat{V}^{\text{o50}}(h)$ and $\hat{V}^{\text{o75}}(h)$ have a very large MSE. This is also a consequence of the data-splitting variance and comes from the smaller number of prediction errors $\hat{e}^{\text{out}}_{t+h}$ that are used to build $\hat{V}^{\text{o50}}(h)$ and $\hat{V}^{\text{o75}}(h)$. When the sample size is small, $n = 25$, the estimator with minimum MSE is $\hat{V}^{\text{inc}}(h)$. Although the bias of this esti-

Table 1. Empirical Performance of Alternative Estimators of V^{pop} and Information Criteria When Fitting an AR(p) to a Sample of Size n of the Process $(1 - .2B)(1 - .5B)(1 - .7B)y_t = a_t$

			Bias						MSE						Probabilities of selecting each model							
h	p	V^{pop}	$\hat{V}^{in}$	$\hat{V}^{inc}$	$\hat{V}^{int}$	$\hat{V}^{o50}$	$\hat{V}^{o75}$	$\hat{V}^{fil}$	$\hat{V}^{in}$	$\hat{V}^{inc}$	$\hat{V}^{int}$	$\hat{V}^{o50}$	$\hat{V}^{o75}$	$\hat{V}^{fil}$	$\hat{V}^{inc}$	$\hat{V}^{int}$	$\hat{V}^{o50}$	$\hat{V}^{o75}$	BIC	AIC	FPE	$\hat{V}^{fil}$
$n = 25$																						
1	1	1.38	−.110	−.055	−.106	.011	**−.005**	**−.005**	.23	.24	**.22**	.50	.90	.25	.03	.02	.20	.24	.13	.06	.06	.09
	2	**1.03**	−.162	−.080	−.140	.048	.015	**.000**	.10	**.09**	.10	.23	.41	.11	**.35**	**.25**	**.57**	**.35**	**.58**	**.42**	**.51**	**.57**
	3	1.08	−.253	−.123	−.197	.113	.040	**.018**	.14	**.11**	.13	.33	.51	.14	.14	.12	.12	.13	.10	.12	.13	.14
	4	1.14	−.365	−.182	−.268	.235	.072	**.034**	.20	**.14**	.17	.57	.65	.18	.15	.15	.07	.11	.07	.11	.11	.10
	5	1.23	−.496	−.252	−.355	.540	.128	**.068**	.32	**.19**	.24	1.66	.90	.25	.15	.18	.03	.09	.05	.11	.09	.06
	6	1.34	−.661	−.346	−.460	4.730	.240	**.116**	.50	**.26**	.35	>	1.41	.38	.18	.28	.01	.08	.07	.18	.10	.04
3	1	5.79	−.874	−.640	−.767	.173	−.058	**.057**	8.09	8.45	**7.77**	24.5	39.0	11.5	.11	.05	.35	.31	.25	.14	.18	.21
	2	**4.98**	−1.381	−1.003	−1.212	.287	**−.059**	−.170	4.96	**4.73**	4.77	16.8	23.1	6.72	**.35**	**.17**	**.38**	**.23**	**.42**	**.35**	**.42**	**.43**
	3	5.23	−1.762	−1.150	−1.514	.936	.163	**.007**	6.17	**5.57**	5.76	31.6	29.4	8.93	.12	.12	.11	.12	.07	.09	.09	.10
	4	5.46	−2.204	−1.335	−1.851	2.186	.484	**.219**	7.85	**6.58**	7.01	76.9	40.4	12.6	.11	.15	.08	.11	.06	.09	.09	.10
	5	5.88	−2.823	−1.648	−2.294	6.543	1.111	**.629**	10.7	**8.13**	8.93	>	71.4	19.2	.10	.17	.05	.09	.06	.09	.08	.07
	6	6.40	−3.599	−2.071	−2.913	>	2.317	**1.375**	15.5	**10.5**	12.4	>	168	37.4	.21	.34	.03	.14	.14	.24	.14	.09
$n = 100$																						
1	1	1.33	−.011	.002	−.012	.015	**.011**	.016	**.06**	**.06**	**.06**	.12	.23	**.06**	.00	.00	.02	.09	.00	.00	.00	.00
	2	**1.01**	−.046	−.026	−.045	**.001**	−.003	−.005	**.02**	**.02**	**.02**	.04	.09	**.02**	**.34**	**.23**	**.60**	**.43**	**.89**	**.56**	**.57**	**.60**
	3	1.02	−.064	−.033	−.056	.008	**.000**	−.002	**.02**	**.02**	**.02**	.05	.09	**.02**	.17	.14	.17	.17	.08	.18	.18	.17
	4	1.03	−.085	−.043	−.070	.014	.002	**−.001**	**.03**	**.02**	**.03**	.05	.09	**.02**	.15	.15	.10	.11	.02	.10	.10	.10
	5	1.04	−.106	−.054	−.085	.019	.004	**.000**	**.03**	**.03**	**.03**	.05	.10	**.03**	.14	.18	.06	.09	.01	.08	.08	.07
	6	1.05	−.129	−.066	−.100	.025	.006	**.001**	.04	**.03**	**.03**	.05	.10	**.03**	.20	.30	.05	.11	.00	.08	.07	.06
3	1	5.56	−.186	−.130	−.165	.047	**.000**	.064	**1.95**	1.98	1.97	4.05	7.96	2.14	.01	.00	.16	.23	.09	.02	.02	.02
	2	**4.97**	−.459	−.363	−.437	**−.048**	−.106	−.124	1.20	**1.16**	1.16	2.31	4.56	1.19	**.41**	**.20**	**.45**	**.34**	**.70**	**.57**	**.58**	**.58**
	3	4.99	−.531	−.385	−.496	**−.007**	−.080	−.094	1.25	**1.18**	1.19	2.45	4.83	1.21	.16	.13	.15	.13	.06	.14	.14	.13
	4	5.04	−.631	−.435	−.582	**.014**	−.077	−.095	1.37	**1.25**	1.30	2.57	4.96	1.28	.12	.14	.09	.09	.03	.09	.09	.10
	5	5.10	−.742	−.494	−.678	**.039**	−.076	−.096	1.52	**1.33**	1.42	2.72	5.16	1.34	.10	.17	.07	.09	.02	.07	.07	.07
	6	5.16	−.845	−.544	−.764	.073	**−.062**	−.085	1.68	**1.40**	1.54	2.87	5.37	**1.40**	.20	.36	.08	.12	.01	.11	.10	.10

NOTE: Bias, MSE, and probabilities are based on 5,000 replications. MSPE is based on 100,000 replications. The symbol ">" means ">1,000."

Table 2. Empirical Performance of Alternative Estimators of V^{pop} and Information Criteria When Fitting Models M1–M5 to a Sample of Size 100 of the Process $y_t = .9y_{t-1} + .7x_{1t} + .2x_{1t-1} + \gamma x_{1t}x_{1t-1} + a_t$

			Bias					MSE					Probabilities of selecting each model						
h	Model	V^{pop}	$\hat{V}^{in}$	$\hat{V}^{inc}$	$\hat{V}^{o50}$	$\hat{V}^{o75}$	$\hat{V}^{fil}$	$\hat{V}^{in}$	$\hat{V}^{inc}$	$\hat{V}^{o50}$	$\hat{V}^{o75}$	$\hat{V}^{fil}$	$\hat{V}^{inc}$	$\hat{V}^{o50}$	$\hat{V}^{o75}$	BIC	AIC	FPE	$\hat{V}^{fil}$
$\gamma = 0$																			
1	M1	**1.04**	−.085	−.045	.025	.013	**−.003**	.03	**.02**	.05	.09	**.02**	**.46**	**.43**	**.34**	**.45**	**.56**	**.57**	**.57**
	M2	1.07	−.063	−.031	.019	.009	**−.002**	.03	**.02**	.05	.10	**.02**	.10	.36	.33	.51	.22	.22	.23
	M3	1.06	−.108	−.058	.029	.013	**−.004**	.03	**.02**	.05	.09	**.02**	.21	.12	.16	.03	.13	.12	.12
	M4	1.07	−.130	−.069	.034	.015	**−.004**	.04	.03	.05	.10	**.02**	.23	.09	.14	.01	.09	.09	.08
	M5	1.63	−.207	−.111	.060	.025	**−.010**	.09	**.06**	.13	.25	**.06**	.00	.00	.03	.00	.00	.00	.00
3	M1	**2.68**	−.406	−.308	.095	.043	**−.028**	.40	.35	.71	1.44	**.33**	**.40**	**.33**	**.29**	**.44**	**.47**	**.48**	**.48**
	M2	2.76	−.341	−.264	.085	.036	**.005**	.38	**.35**	.74	1.48	**.35**	.13	.35	.30	.47	.23	.23	.23
	M3	2.71	−.461	−.339	.107	.044	**−.029**	.45	.38	.74	1.48	**.34**	.21	.16	.17	.06	.16	.16	.16
	M4	2.74	−.512	−.366	.122	.052	**−.027**	.49	.40	.77	1.52	**.35**	.25	.15	.19	.03	.14	.13	.13
	M5	4.71	−.894	−.631	.265	**.087**	−.095	1.53	1.23	2.85	5.07	**1.14**	.01	.01	.06	.00	.00	.00	.00
5	M1	**3.82**	−.755	−.620	.201	.111	**−.028**	1.30	**1.18**	2.57	5.12	1.18	**.36**	**.30**	**.26**	**.44**	**.42**	**.43**	**.44**
	M2	3.93	−.658	−.551	.201	.102	**−.044**	1.22	**1.15**	2.66	5.17	1.24	.17	.30	.27	.44	.24	.24	.23
	M3	3.86	−.832	−.663	.218	.113	**−.031**	1.41	1.24	2.67	5.24	**1.21**	.21	.18	.17	.07	.17	.17	.17
	M4	3.90	−.906	−.704	.236	.122	**−.031**	1.52	1.30	2.77	5.35	**1.25**	.25	.18	.20	.04	.16	.15	.16
	M5	6.83	−1.562	−1.190	.515	.190	**−.152**	4.69	**4.00**	10.5	18.4	4.12	.01	.04	.10	.01	.01	.01	.00
$\gamma = .1$																			
1	M1	**1.05**	−.085	−.045	.028	.013	**−.001**	.03	**.02**	.05	.09	**.02**	**.47**	**.44**	**.36**	**.46**	**.57**	**.57**	**.56**
	M2	1.08	−.061	−.029	.024	.013	**−.004**	.03	**.02**	.05	.10	**.02**	.10	.34	.33	.51	.22	.22	.24
	M3	1.07	−.108	−.057	.034	.016	**−.002**	.03	.03	.05	.10	**.02**	.21	.12	.14	.03	.13	.13	.12
	M4	1.08	−.130	−.069	.038	.016	**−.003**	.04	.03	.05	.10	**.02**	.23	.09	.15	.01	.09	.08	.08
	M5	1.63	−.205	−.109	.062	.029	**−.007**	.09	.07	.13	.24	**.06**	.00	.00	.03	.00	.00	.00	.00
3	M1	**2.71**	−.415	−.317	.110	.039	**−.031**	.42	.37	.77	1.46	**.34**	**.40**	**.34**	**.30**	**.44**	**.47**	**.48**	**.47**
	M2	2.79	−.347	−.269	.101	.039	**−.003**	.38	**.35**	.78	1.49	**.35**	.14	.34	.30	.47	.23	.24	.25
	M3	2.74	−.470	−.346	.123	.047	**−.032**	.46	.39	.79	1.50	**.36**	.21	.15	.16	.06	.16	.16	.15
	M4	2.77	−.523	−.375	.135	.051	**−.031**	.51	.41	.81	1.54	**.36**	.24	.15	.19	.03	.13	.13	.13
	M5	4.72	−.886	−.622	.289	.127	**−.085**	1.53	1.24	2.87	5.08	**1.15**	.00	.01	.05	.00	.00	.00	.00
5	M1	**3.88**	−.783	−.647	.214	.080	**−.046**	1.35	1.22	2.71	5.09	**1.20**	**.37**	**.31**	**.28**	**.43**	**.43**	**.43**	**.43**
	M2	3.99	−.691	−.583	.209	.084	**−.018**	1.28	**1.20**	2.80	5.26	1.25	.17	.32	.26	.45	.25	.25	.25
	M3	3.92	−.856	−.686	.237	.095	**−.044**	1.46	1.28	2.78	5.23	**1.23**	.20	.16	.17	.08	.17	.16	.16
	M4	3.96	−.927	−.723	.257	.106	**−.039**	1.57	1.33	2.86	5.34	**1.26**	.25	.18	.20	.04	.15	.15	.15
	M5	6.89	−1.598	−1.224	.523	.221	**−.184**	4.82	**4.10**	10.8	18.9	4.13	.01	.03	.10	.00	.01	.01	.01

NOTE: Bias, MSE, and probabilities are based on 5,000 replications. V^{pop} is based on 100,000 replications. The symbol ">" means ">1,000."

mator is relatively large, its variance is the smallest because it includes the largest number of prediction errors, and this effect is very important for small sample size. When the sample size increases, $n = 100$, the bias become more important, and $\hat{V}^{\text{fil}}(h)$ has similar or better performance than $\hat{V}^{\text{inc}}(h)$. The same conclusions can be obtained from the transfer function experiment in Table 2 in both the linear and the nonlinear cases ($\gamma = .1$); the MSE is highly related to the number of prediction errors used in the estimator of $V^{\text{pop}}(h)$.

4.3 Comparison of Model Selection Criteria

Tables 1 and 2 also show the empirical performance of alternative model selection criteria for the first and second experiments. They report the proportion of times that each model was selected for each criterion. Thus the sum of the entries for each estimator for $p = 1, \ldots, 6$ in Table 1 and for M1, …,M5 in Table 2 must add up to 1. The row corresponding with the minimum $V^{\text{pop}}(h)$ is in boldface.

We can summarize the conclusions as follows:

1. The in-sample procedures $\hat{V}^{\text{inc}}(h)$ and $\hat{V}^{\text{int}}(h)$ show a strong tendency to choose a high-order p in the first experiment and a clearly overparameterized model in the second experiment (models M3 and M4). This result is consistent with the reported bias of these estimators.
2. In the contrast, the out-of-sample procedures $\hat{V}^{\text{o50}}(h)$ and $\hat{V}^{\text{o75}}(h)$ have a tendency toward smaller predictors, irrespective of their efficiency, in agreement with the reported bias of these estimators. The large MSE of $\hat{V}^{\text{o75}}(h)$ is also reflected in these tables as a lower capacity to discriminate between competing predictors; it can be seen that in both experiments, $\hat{V}^{\text{o75}}(h)$ has a tendency toward a uniformity in the probability of selecting the best predictor.
3. As can be expected, the BIC also has a tendency toward underfitting.
4. The tables show a high similarity between AIC, FPE, and the proposed $\hat{V}^{\text{fil}}(h)$ at $h = 100$.

In summary, $\hat{V}^{\text{fil}}(h)$ seems to be a good estimator of $V^{\text{pop}}(h)$ in finite samples. It clearly surpasses the traditional predictive validation criteria consisting on splitting the sample into an estimation subsample and a prediction one. Because $\hat{V}^{\text{fil}}(h)$ is just a sampling average of a function of the filtered prediction errors, it can be implemented using a different loss functions than the quadratic one, that is, the mean absolute deviation. This possibility gives $\hat{V}^{\text{fil}}(h)$ greater flexibility than its competitors.

4.4 An Example: Modeling the Gas Furnace Data

Box and Jenkins (1976, p. 381) built a transfer function model for the proportion of output CO_2 (y_t) as a function of the nonstochastic feed rate of methane (x_t) in a gas furnace. The data correspond to 296 readings at 9-second intervals. A general transfer function for these data using the Box–Jenkins notation would be

$$y_t = \frac{(\omega_0 - \omega_1 B - \cdots - \omega_s B^s)}{1 - \delta_1 B - \cdots - \delta_r B^r} x_{t-b} + \frac{1 - \theta_1 B - \cdots - \theta_q B^q}{1 - \phi_1 B - \cdots - \phi_p B^p} a_t. \quad (16)$$

Using their methodology, Box and Jenkins fitted the model $b = 3$, $s = 2, r = 1, q = 0, p = 2$. We denote this model by R1. Our purpose is to use the proposed predictive validation procedure to analyze the prediction performance of both R1 and some other alternative models, which can be considered as small departures from R1. The goal is to complement the identification procedure with an additional criterion based explicitly on forecasting performance. This table reports the value of $\hat{V}^{\text{fil}}(h)$ for $h = 1, 3, 7$ and, for comparison purposes, the respective values using $\hat{V}^{\text{o50}}(h)$. The table also shows the traditional AIC and BIC. For brevity, we report only the results of four alternative models (R2–R5) based on (16). The orders that are different from R1 are in bold type. We have also highlighted the cells that minimize the respective criterion. The estimation has been made with the nag routines g13bef and g13bjf as implemented in Matlab.

Table 3 shows that, according to the AIC and BIC, the preferred model would be R1. However, $\hat{V}^{\text{fil}}(h)$ shows that whereas R1 is the best predictor at $h = 1$, the predictions generated by R3 are more efficient at $h = 3, 5$. It can be seen that the less-efficient criterion $\hat{V}^{\text{o50}}(h)$ also selects R1 at $h = 1, 3$, but at longer horizons, it selects different models than $\hat{V}^{\text{fil}}(h)$. Table 3 also reports the results for comparing two univariate models for the output y_t, denoted by U1 and U2. Using the AIC and BIC, we would use and AR(4). However, $\hat{V}^{\text{fil}}(h)$ would select an AR(3) for $h = 5$.

From the results of the article, we can then use $\hat{V}^{\text{fil}}(h)$ as accurate estimates of the MSPE of the transfer function. [Note that $\hat{V}^{\text{o50}}(h)$ supplies very inflated estimates of MSPE.] We can use the estimates of MSPE for many purposes apart from model comparison. For instance, to build unconditional asymptotic prediction intervals. We can also obtain measures of the advantages of the transfer function with respect to the univariate model for different prediction horizons. By comparing $\hat{V}^{\text{fil}}(h)$ of R1 and U2 at $h = 1$, we can see that the transfer function

Table 3. Comparison of Alternative Transfer Funtions for the Gas Furnace Data

	Orders							$\hat{V}^{fil}(h)$			$\hat{V}^{o50}(h)$		
Model	b	s	r	p	q	AIC	BIC	$h=1$	$h=3$	$h=5$	$h=1$	$h=3$	$h=5$
R1	3	2	1	2	0	**−838.9**	**−813.0**	**.0622**	.433	.607	**.105**	**.796**	1.348
R2	3	2	**2**	2	0	−836.9	−807.4	.0627	.437	.610	.106	.798	1.348
R3	**2**	**1**	3	2	0	−837.3	−807.7	.0625	**.428**	**.605**	.106	.823	1.398
R4	3	2	1	**3**	0	−837.4	−807.9	.0637	.439	.677	.107	.800	1.352
R5	3	2	1	**1**	**1**	−801.4	−775.5	.0707	.490	.713	.121	.803	**1.277**
U1	0	0	0	3	0	−630.1	−615.4	.1215	1.917	**5.705**	.164	**2.237**	5.965
U2	0	0	0	4	0	**−642.0**	**−623.6**	**.1174**	**1.906**	5.717	**.159**	2.269	**5.943**

allows a reduction of 47% in the one-step-ahead MSPE with respect to the univariate model. Also, by comparing $\hat{V}^{\text{fil}}(h)$ of R3 and U1 at $h = 5$, the reduction in MSPE increases to 89%. Therefore, the transfer function for the gas furnace data is a much more useful tool at longer horizons than at $h = 1$.

5. CONCLUDING REMARKS

The identification of an ARMAX model is usually made after some amount of data mining, and hence the risk of building overidentified models is not negligible. The consequences in prediction of such overidentification are well documented. It is then not surprising that forecasters make extensive use of split-sample validation procedures to complement their analysis. The use of the filtered residuals defined in this article can then be a valuable alternative for performing such validation.

The proposed predictive validation procedure is simple to compute and provides a straightforward way to estimate the h-step-ahead prediction error of competing predictors. Its computation requires a similar effort as for the standard procedures of checking for outliers, and it can be implemented in a similar way. This predictive validation approach has clear intuitive appeal. Because the innovations of the h observations used to produce the h-step-ahead filtered prediction error are almost uncorrelated with the estimated predictor, it is closely related to a multifold (h-fold) cross-validation procedure. This makes the proposed predictive validation an effective tool to avoid the detection of spurious relationships.

APPENDIX A: PREVIOUS RESULTS

Let z_t follow the ARMAX model (2). For convenience, we express this model in VARX(1) form as

$$\mathbf{Y}_t = \mathbf{A}\mathbf{Y}_{t-1} + \mathbf{X}_{t+1} + \mathbf{U}_t, \tag{A.1}$$

where $\mathbf{Y}_t$, $\mathbf{X}_{t+1}$, and $\mathbf{U}_t$ are the following $m \times 1$ vectors, with $m = p + k + \sum s_i + q$, $\mathbf{U}_t = [a_t, \mathbf{0}'_{m-q-1}, a_t, \mathbf{0}'_{q-1}]$, $\mathbf{Y}_t = [z_t, z_{t-1}, \ldots, z_{t-p+1}, x_{1,t+1}, \ldots, x_{1,t+1-s_1}, \ldots, x_{k,t+1-s_k}, a_t, \ldots, a_{t-q+1}]'$, and $\mathbf{X}_{t+1} = [\mathbf{0}'_p, x_{1,t+1}, \mathbf{0}'_{s_1}, x_{2,t+1}, \mathbf{0}'_{s_2}, \ldots, x_{k,t+1}, \mathbf{0}'_{s_k}, \mathbf{0}'_q]'$, where $\mathbf{0}_p$ is a vector of 0's of dimension p, and $\mathbf{A}$ is the $m \times m$ block matrix:

$$\mathbf{A} = \begin{bmatrix} \mathbf{B} & \mathbf{C}_1 & \mathbf{C}_2 & \cdots & \mathbf{C}_k & \mathbf{D} \\ \mathbf{0} & \mathbf{E}_1 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{E}_2 & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{E}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{G} \end{bmatrix},$$

where $\mathbf{0}$ is a matrix with 0's of appropriate dimension; $\mathbf{B}$ is $p \times p$; $\mathbf{C}_i$, $i = 1, \ldots, k$, is $p \times (s_i + 1)$; $\mathbf{D}$ is $p \times q$; $\mathbf{E}_i$, $i = 1, \ldots, k$, is $(s_i + 1) \times (s_i + 1)$; and $\mathbf{G}$ is $q \times q$. These matrices have the following structure:

$$\mathbf{B} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ & & \mathbf{I}_{p-1} & & \mathbf{0} \end{bmatrix},$$

$$\mathbf{C}_i = \begin{bmatrix} \eta_{i,0} & -\eta_{i,1} & \cdots & -\eta_{i,s_i} \\ & & \mathbf{0} & \end{bmatrix},$$

$$\mathbf{E}_i = \begin{bmatrix} \mathbf{0} & 0 \\ \mathbf{I}_{s_i} & \mathbf{0} \end{bmatrix},$$

where $\mathbf{I}_n$ is the identity matrix of size n and $\mathbf{G}$ has the same structure as $\mathbf{E}_i$. The matrix $\mathbf{D}$ has the same structure as $\mathbf{C}_i$, but with the elements $[-\theta_1 - \theta_2 \cdots - \theta_q]$ in the first row. Because the first element of $\mathbf{X}_t$ is null, and using (A.1) recursively, we can express the observation z_{T+h} as (see Baillie 1980)

$$z_{T+h} = \alpha(h)'\mathbf{Y}_T + \sum_{l=1}^{h-1} \alpha(l)'\mathbf{X}_{T+h+1-l} + \sum_{l=0}^{h-1} \alpha(l)'\mathbf{U}_{T+h-l}, \tag{A.2}$$

where $\alpha(l)' = [\alpha_1(l), \ldots, \alpha_m(l)]' = \mathbf{c}'\mathbf{A}^l$, with $\mathbf{c} = (1, 0, \ldots, 0)'$ and dimension $(m \times 1)$. From the structure of $\mathbf{A}$ and $\mathbf{U}_t$, it can be verified that $\alpha(l)'\mathbf{U}_{T+h-l} = \psi_l a_{T+h-l}$, where $\psi(B) = (1 + \psi_1 B + \psi_2 B^2 + \cdots)$ is obtained from $\phi(B)\psi(B) = \theta(B)$. Let $\hat{\alpha}(l)$ be the estimation of $\alpha(l)$ obtained from ML estimation of model (2), based on the whole span of available data. Then the estimated predictor is $\hat{z}^{\text{in}}_{T+h} = \hat{\alpha}(h)'\mathbf{Y}_T + \sum_{l=1}^{h-1} \hat{\alpha}(l)'\mathbf{X}_{T+h+1-l}$, and the prediction error of this estimated predictor is

$$\begin{aligned} \hat{e}^{\text{in}}_{T+h} &= \sum_{l=0}^{h-1} \psi_l a_{T+h-l} + [\alpha(h) - \hat{\alpha}(h)]'\mathbf{Y}_T \\ &\quad + \sum_{l=1}^{h-1} [\alpha(l) - \hat{\alpha}(l)]'\mathbf{X}_{T+h+1-l} \\ &= \sum_{l=0}^{h-1} \psi_l a_{T+h-l} + \sum_{i=1}^{m} [\hat{\alpha}_i(h) - \alpha_i(h)]Y_T^{(i)} \\ &\quad + \sum_{l=1}^{h-1} \sum_{i=1}^{m} [\alpha_i(l) - \hat{\alpha}_i(l)]X_{T+h+1-l}^{(i)}, \end{aligned} \tag{A.3}$$

where $Y_t^{(i)}$ and $X_t^{(i)}$ are the ith row of the vectors $\mathbf{Y}_t$ and $\mathbf{X}_t$, and $\alpha_i(l)$ is the ith element of $\alpha(l)$. To evaluate $E(\hat{z}^{\text{in}}_{T+h} e_{T+h})$, we first prove the following lemma.

Lemma A.1. Let z_t follow model (2). Let λ_j be an element of the parameter vector λ and let $\hat{\lambda}_j$ be the ML or LS estimator. Then

$$E(\hat{z}^{\text{in}}_{T+h} e_{T+h}) = O\big(\max\big\{E\big[(\hat{\lambda}_j - \lambda_j)Y_T^{(i)} a_{T+h-l}\big]\big\}; \quad j, i = 1, \ldots, m;\ l = 1, \ldots, h;\big). \tag{A.4}$$

Proof. Let us denote by $\dot{\lambda}$ any vector of values in a closed region Λ of parameter points satisfying assumption (A.2) and containing λ in its interior. We can then write

$$\dot{a}_t = \dot{\phi}(B)\dot{\theta}(B)^{-1} z_t - \sum_{j=1}^{k} \dot{\eta}_j(B)\dot{\theta}(B)^{-1} x_{j,t}. \tag{A.5}$$

If the innovations sequence $\{a_t\}$ is normally distributed, then the log-likelihood function is

$$\log \dot{L} = \text{const} - \frac{n}{2} \log \dot{\sigma}^2 - \frac{1}{2\dot{\sigma}^2} \sum_{t=1}^{n} \dot{a}_t^2. \tag{A.6}$$

Because our interest is in the structural parameters λ, we use the concentrated likelihood on σ^2, with the ML estimation of this parameter being $\hat{\sigma}^2 = n^{-1} \sum \hat{a}_t^2$. Therefore, manipulation of the log-likelihood is based only on the last term in (A.6). Hence the proof will also be applicable to LS estimation without using the normality assumption. From (A.3), we can write

$$\begin{aligned} E(\hat{z}^{\text{in}}_{T+h} e_{T+h}) &= \sum_{l=0}^{h-1} \psi_l E\big\{[\hat{\alpha}(h) - \alpha(h)]'\mathbf{Y}_T a_{T+h-l}\big\} \\ &\quad + \sum_{l=0}^{h-1} \sum_{k=1}^{h-1} \psi_l E\big\{[\alpha(k) - \hat{\alpha}(k)]'\mathbf{X}_{T+h+1-k} a_{T+h-l}\big\}. \end{aligned}$$

By using a Taylor expansion on $\hat{\alpha}(h)$ around their true values, we obtain

$$E\{[\hat{\alpha}(h)-\alpha(h)]'Y_T a_{T+h-l}\} = \sum_{i=1}^{m}\sum_{j=1}^{m}\frac{\partial\alpha_i(h)}{\partial\lambda_j}E[(\hat{\lambda}_j-\lambda_j)Y_T^{(i)}a_{T+h-l}] \quad (A.7a)$$

$$+\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{r=1}^{m}\frac{1}{2}\frac{\partial\alpha_i(h)}{\partial\lambda_j\,\partial\lambda_r} \times E[(\hat{\lambda}_j-\lambda_j)(\hat{\lambda}_r-\lambda_r)Y_T^{(i)}a_{T+h-l}]+R_Y^*, \quad (A.7b)$$

where $R_Y^* = o(\max\{E[(\hat{\lambda}_j-\lambda_j)(\hat{\lambda}_r-\lambda_r)Y_T^{(i)}a_{T+h-l}]\})$. To analyze the order of magnitude of the first term in (A.7b) we have, applying the properties of the log-likelihood function $\log L$ (see, e.g., Pierce 1971; Tanaka 1984),

$$O\{E[(\hat{\lambda}_j-\lambda_j)(\hat{\lambda}_r-\lambda_r)Y_T^{(i)}a_{T+h-l}]\} = O\left\{\frac{1}{n^2}E\left[\left(\frac{\partial\log L}{\partial\lambda_j}\right)\left(\frac{\partial\log L}{\partial\lambda_r}\right)Y_T^{(i)}a_{T+h-l}\right]\right\}. \quad (A.8)$$

The term inside the expectation operator in (A.8) is a scalar obtained by the sum of products of combinations of random variables. It can be shown that except for a fixed number of terms that depends on the orders $p, q, s_1, \ldots, s_k$ and the horizon h, and not on n, the terms have null expectation. Then

$$O\left\{\frac{1}{n^2}E\left[\left(\frac{\partial\log L}{\partial\lambda_j}\right)\left(\frac{\partial\log L}{\partial\lambda_r}\right)Y_T^{(i)}a_{T+h-l}\right]\right\} = O(n^{-2}), \quad (A.9)$$

and thus $E\{[\hat{\alpha}(h)-\alpha(h)]'Y_T a_{T+h-l}\} = O(\max\{E[(\hat{\lambda}_j-\lambda_j)\times Y_T^{(i)}a_{T+h-l}]\})$. Using the same arguments, it also holds that $E\{[\alpha(k)-\hat{\alpha}(k)]'X_{T+h+1-k}a_{T+h-l}\} = O(\max\{E[(\hat{\lambda}_j-\lambda_j)Y_T^{(i)}a_{T+h-l}]\})$. Then $E(\hat{z}^{\text{in}}_{T+h}e_{T+h}) = O(\max_{j,i,l}\{E[(\hat{\lambda}_j-\lambda_j)Y_T^{(i)}a_{T+h-l}]\})$, and the lemma is proved.

APPENDIX B: PROOF OF THEOREM 1

B.1 Proof of Part (a)

By Lemma A.1 and using a Taylor series expansion of the log-likelihood function to approximate $(\hat{\lambda}_j-\lambda_j)$, we obtain

$$E(\hat{z}^{\text{in}}_{T+h}e_{T+h}) = O\left\{\max_{i,j,l}E\left(\frac{1}{n}\frac{\partial\log L}{\partial\lambda_j}Y_T^{(i)}a_{T+h-l}\right)\right\}.$$

For the sake of brevity, here we show only the elements corresponding to the parameters ϕ. It can be verified that the results hold for all of the elements λ_j. The first derivative is

$$\frac{1}{n}\frac{\partial\log L}{\partial\phi_i} = -\frac{1}{2\sigma^2}\sum_{t=1}^{n}2a_t\frac{\partial a_t}{\partial\phi_j} = \frac{\sum_{t=1}^{n}a_t u_{t-i}}{n\sigma^2}, \qquad i=1,\ldots,p.$$

Then $n^{-1}\sigma^{-2}\sum_{t=1}^{n}E(a_t u_{t-j}y_{T+1-i}a_{T+h-l}) = n^{-1}\sigma^{-2}E(a_{T+h-l}^2)\times E(u_{T+h-l-j}y_{T+1-i}) = O(n^{-1})$. Therefore, $E(\hat{z}^{\text{in}}_{T+h}e_{T+h}) = O\{\max_{j,i,l}E[(\hat{\lambda}_j-\lambda_j)Y_T^{(i)}a_{T+h-l}]\} = O(n^{-1})$, and the proof is completed. Because we are taking derivatives with respect to the structural parameters, and using the concentrated likelihood on σ^2, it can be verified that the results are also extended to LS estimation where the normality assumption is not needed.

B.2 Proof of Part (b)

Following similar arguments as those in previous sections, we have

$$E(\hat{z}^{\text{fil}}_{T+h}e_{T+h}) = O\left\{\max_{j,i,l}E[(\hat{\lambda}_j^{\text{fil}}-\lambda_j)Y_T^{(i)}a_{T+h-l}]\right\}, \quad (B.1)$$

where $\hat{\lambda}^{\text{fil}}$ maximizes the log-likelihood function $\log\dot{L}^{\text{fil}} = -(n/2)\times\log\dot{\sigma}^2 - (2\dot{\sigma}^2)^{-1}\sum_{t\neq T+1,\ldots,T+h}\dot{a}_t^2$ (or, equivalently, minimizes the LS $\sum_{t\neq T+1,\ldots,T+h}\dot{a}_t^2$). Now the new log-likelihood will show some differences in the derivatives. As before, and for the sake of brevity, we show the results for the parameters ϕ. The new derivatives are

$$\frac{1}{n}\frac{\partial\log L^{\text{fil}}}{\partial\phi_i} = -\frac{1}{2\sigma^2}\sum_{\substack{t=1\\ t\neq T+1,\ldots,T+h}}^{n}2a_t\frac{\partial a_t}{\partial\phi_j} = \frac{\sum_{t\neq T+1,\ldots,T+h}a_t u_{t-i}}{n\sigma^2}.$$

Because the difference between $\log L$ and $\log L^{\text{fil}}$ is in a finite number of terms in the sum of squared residuals, L^{fil} still has the same asymptotic properties. However, and contrary to the classical in-sample predictor, it can be shown that

$$\sum_{t\neq T+1,\ldots,T+h}E(a_t u_{t-j}Y_T^{(i)}a_{T+h-l}) = 0,$$

because of the omission of the innovations $a_{T+1},\ldots,a_{T+h}$. Similar results can be verified for the remaining parameters. Therefore, we should analyze terms of smaller order of magnitude. These terms are analyzed in (A.8). By (A.9), we then have that

$$O\left\{\max_{j,i,l}E[(\hat{\lambda}_j^{\text{fil}}-\lambda_j)Y_T^{(i)}a_{T+h-l}]\right\} = O(n^{-2}), \quad (B.2)$$

and the proof is completed.

APPENDIX C: PROOF OF THEOREM 2

Using similar arguments as those of Baillie (1980) and Kunitomo and Yamamoto (1985), we obtain

$$V^{\text{pop}}(h) = \sigma^2\sum_{l=0}^{h-1}\psi_l^2 + \frac{\sigma^2}{n}\left[\text{trace}\{\Omega_\alpha(h)\Gamma_Y\} + \sum_{l=1}^{h-1}\text{trace}\{\Omega_\alpha(l)\Gamma_X\} + 2\sum_{l=1}^{h-1}\text{trace}\{\Omega_\alpha(h,l)\Gamma_{XY}(h+1-l)\}\right] + O(n^{-3/2}),$$

where we denote $E\{[\alpha(h)-\hat{\alpha}(h)][\alpha(h)-\hat{\alpha}(h)]'\} = n^{-1}\sigma^2\Omega_\alpha(h) + O(n^{-3/2})$, $E(\mathbf{Y}_n\mathbf{Y}_n') = \Gamma_Y$, $\Gamma_X = E(\mathbf{X}_n\mathbf{X}_n')$, $E\{[\alpha(h)-\hat{\alpha}(h)][\alpha(l)-\hat{\alpha}(l)']\} = n^{-1}\sigma^2\Omega_\alpha(h,l) + O(n^{-3/2})$, and $\Gamma_{XY}(h+1-l) = E(\mathbf{X}_{n+h+1-l}\mathbf{Y}_n')$. In contrast, taking expectations to $(\hat{e}^{\text{fil}}_{t+h})^2$ and using (B.2), we obtain

$$E[(\hat{e}^{\text{fil}}_{t+h})^2] = \sigma^2\sum_{l=0}^{h-1}\psi_l^2 + E\{[\alpha(h)-\hat{\alpha}^{\text{fil}}(h)]'\mathbf{Y}_t\mathbf{Y}_t'[\alpha(h)-\hat{\alpha}^{\text{fil}}(h)]\} + \sum_{l=1}^{h-1}E\{[\alpha(l)-\hat{\alpha}^{\text{fil}}(l)]'\mathbf{X}_{t+h+1-l}\mathbf{X}_{t+h+1-l}'[\alpha(l)-\hat{\alpha}^{\text{fil}}(l)]\} + 2\sum_{l=1}^{h-1}E\{[\alpha(l)-\hat{\alpha}^{\text{fil}}(l)]'\mathbf{X}_{t+h+1-l}\mathbf{Y}_t'[\alpha(h)-\hat{\alpha}^{\text{fil}}(h)]\} + o(n^{-2}).$$

Because the asymptotic properties of $\hat{\alpha}^{\text{fil}}$ are the same as $\hat{\alpha}$, it holds that

$$E\left[\hat{V}^{\text{fil}}(h)\right] = \sigma^2 \sum_{l=0}^{h-1} \psi_l^2 + \frac{\sigma^2}{n}\left[\text{trace}\{\Omega_\alpha(h)\Gamma_Y\} + \sum_{l=1}^{h-1} \text{trace}\{\Omega_\alpha(l)\Gamma_X\} + 2\sum_{l=1}^{h-1} \text{trace}\{\Omega_\alpha(h,l)\Gamma_{XY}(h+1-l)\}\right] + O(n^{-3/2}),$$

and the proof is completed.

[Received March 2003. Revised April 2004.]

REFERENCES

Akaike, H. (1969), "Fitting Autoregressive Models for Prediction," *Annals of the Institute of Statistical Mathematics*, 21, 243–247.

——— (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, AC-19, 716–723.

Allen, D. M. (1974), "The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction," *Technometrics*, 16, 125–127.

Baillie, R. T. (1980), "Predictions From ARMAX Models," *Journal of Econometrics*, 12, 365–374.

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

Burman, P. (1989), "A Comparative Study of Ordinary Cross-Validation, *v*-Fold Cross-Validation, and the Repeated Learning-Testing Methods," *Biometrika*, 76, 503–514.

Burman, P., Chow E., and Nolan, D. (1994), "A Cross-Validatory Method for Dependent Data," *Biometrika*, 81, 351–358.

Chang, I., Tiao, G. C., and Chen, C. (1988), "Estimation of Time Series Parameters in the Presence of Outliers," *Technometrics*, 30, 193–204.

Diebold, F. X., and Kilian, L. (2001), "Measuring Predictability: Theory and Macroeconomic Applications," *Journal of Applied Econometrics*, 16, 657–669.

Efron, B. (1986), "How Biased Is the Apparent Error Rate of a Prediction Rule?" *Journal of the American Statististical Association*, 81, 461–470.

Fuller, W. A., and Hasza, D. P. (1981), "Properties of Predictors in Misspecified Autoregressive Time Series Models," *Journal of the American Statististical Association*, 76, 155–161.

Granger, C. W. J., and Newbold, P. (1986), *Forecasting Economic Time Series*, San Diego: Academic Press.

Hannan, E. J., and Quinn, B. G. (1979), "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society*, Ser. B, 41, 190–195.

Haslett, J. (1999), "A Simple Derivation of Deletion Diagnostic Results for the General Linear Model With Correlated Errors," *Journal of the Royal Statistical Society*, Ser. B, 61, 603–609.

Hurvich, C. M., and Tsai, C. (1997), "Selection of a Multistep Linear Predictor for Short Time Series," *Statistica Sinica*, 7, 395–406.

Kavalieris, L. (1989), "The Estimation of the Order of an Autoregression Using Recursive Residuals and Cross-Validation," *Journal of Time Series Analysis*, 10, 271–281.

Kunitomo, N., and Yamamoto, T. (1985), "Properties of Predictors in Misspecified Autoregressive Time Series Models," *Journal of the American Statististical Association*, 80, 941–950.

Mann, H. B., and Wald, A. (1943), "On the Statistical Treatment of Linear Stochastic Difference Equations," *Econometrica*, 11, 173–220.

Peña, D. (1990), "Influential Observations in Time Series," *Journal of Business and Economic Statistics*, 8, 235–241.

Pierce, D. A. (1971), "Least Squares Estimation in the Regression Model With Autoregressive-Moving Average Errors," *Biometrika*, 58, 2, 299–312.

——— (1979), "R^2 Measures for Time Series," *Journal of the American Statististical Association*, 74, 901–910.

Rissanen, J. (1986), "Order Estimation by Accumulated Prediction Errors," in *Essays in Time Series and Applied Processes* (special vol. 23A of *Journal of Applied Probability*), eds. Gani and M. B. Priestley, Sheffield, U.K.: The Applied Probability Trust, pp. 55–61.

Schwartz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Stoica, P., Eykhoff, P., Janssen, P., and Söderstrom, T. (1986), "Model-Structure Selection by Cross-Validation," *International Journal of Control*, 43, 1841–1878.

Stone, M. (1974), "Cross-Validation Choice and Assessment for Statistical Predictions," *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.

——— (1977), "An Asymptotic Equivalence of Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society*, Ser. B, 39, 44–47.

Tanaka, K. (1984), "An Asymptotic Expansion Associated With the Maximum Likelihood Estimators in ARMA Models," *Journal of the Royal Statistical Society*, Ser. B, 46, 1, 58–67.

West, K. D. (1996), "Asymptotic Inference About Predictive Ability," *Econometrica*, 64, 1067–1084.

White, H. (2000), "A Reality Check for Data Snooping," *Econometrica*, 68, 1097–1126.

Yamamoto, T. (1981), "Predictions for Multivariate Autoregressive Moving-Average Models," *Biometrika*, 68, 485–492.

Zhang, P. (1993), "Model Selection via Multifold Cross-Validation," *The Annals of Statistics*, 21, 299–313.