Working Paper 01-11
Statistics and Econometrics Series 07
January 2001

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

NEW IN-SAMPLE PREDICTION ERRORS IN TIME SERIES WITH APPLICATIONS

Daniel Peña and Ismael Sánchez*

Abstract

This article introduces two new types of prediction errors in time series: the filtered prediction errors and the deletion prediction errors. These two prediction errors are obtained in the same sample used for estimation, but in such a way that they share some common properties with out of sample prediction errors. It is proved that the filtered prediction errors are uncorrelated, up to terms of magnitude order $O(T^{-2})$, with the in sample innovations, a property that share with the out-of-sample prediction errors. On the other hand, deletion prediction errors assume that the values to be predicted are unobserved, a property that they also share with out-of-sample prediction errors. It is shown that these prediction errors can be computed with parameters estimated by assuming innovative or additive outliers, respectively, at the points to be predicted. Then the prediction errors are obtained by running the procedure for all the points in the sample of data. Two applications of these new prediction errors are presented. The first is the estimation and comparison of the prediction mean squared errors of competing predictors. The second is the determination of the order of an ARMA model. In the two applications the proposed filtered prediction errors have some advantages over alternative existing methods..

Keywords: Forecast accuracy; Model comparison; Order selection; Prediction errors; Time series.

*Peña, Department of Statistics and Econometrics; Universidad Carlos III de Madrid, Getafe (Madrid), e-mail: dpena@est-econ; Sánchez, Department of Statistics and Econometrics; Universidad Carlos III de Madrid, Avda. de la Universidad, 30, 28911 Leganés (Madrid); e-mail: ismael@est-econ.uc3m.es.

1 Introduction

Prediction errors are of clear interest in every stage of time series model building, as well as in model comparison. There are several procedures for identification of time series models based on information criteria constructed with the residual variance associated with a fitted model (see, for instance, Koreisha & Pukkila 1995, and references therein). Also, model comparison and model selection are mainly done through the comparison of out-of sample predictive accuracy (Chong and Hendry, 1986; Diebold and Mariano, 1995; West, 1996; Ashley, 1998; West and McCracken, 1998;. Clark and McCracken, 1999; White, 2000), as well as the comparison of in-sample residuals (Ashley et al., 1980).

The above mentioned procedures use either in-sample or out-of-sample prediction errors. In-sample prediction errors have the drawback of using the same information twice: for estimating the parameters of the model and for computing the prediction errors. This data reuse decreases the possibility of detecting misspecifications, a problem called *data-snooping bias,* (Lo and MacKinlay, 1990; White, 2000) and tends to select overparameterized models with lower values of residual variance.

Out-of-sample prediction errors, on the other hand, can avoid such data-snooping bias because the subsample used for prediction contains independent information to evaluate the model. However, they can only be obtained in a portion of the sample, inducing a larger variance in the statistics built with them than with the in-sample prediction errors. For instance, we may wrongly conclude that the out-of-sample predictive performance of several predictors is comparable. Besides, as a portion of the sample is used for estimation, the sampling variability of the estimates will also increase, leading to select models that are too parsimonious, and therefore suboptimal. This added variance induced by splitting the sample of data will be called *data-splitting variance*. Thus, in-sample and out-of-sample prediction errors can have opposite effects, and we have to choose between procedures that incur on data-snooping bias or data-splitting variance.

This article proposes two alternative procedures to evaluate the prediction errors of a time series model. Both are based on in-sample forecasting errors, but evaluated in such a way that the information of the points to be predicted is avoided in the estimation of the model. The first proposal are the filtered prediction errors.

They are computed by assuming that the innovation at the point to be predicted is equal to zero. This is equivalent to build a model that treats the points to be predicted as innovational outliers. It can be proved that the asymptotic covariance of the prediction and the innovation of the point to be predicted is of magnitude order $O(T^{-2})$, whereas with classical residuals it is $O(T^{-1})$. Consequently, the resulting prediction errors are very close to out-of-sample ones. Therefore, since the influence of the predicted points in the predictions is marginal (and asymptotically null), the data-snooping bias is clearly diminished. The filtered prediction errors are obtained by running the procedure for all the points in the sample of data. The second proposal are the deletion prediction errors. They are computed by assuming that the points to be predicted are missing values. This is equivalent to build a model that treats the points to be predicted as additive outliers. As they do not take into account in the estimation the value that we want to forecast, again they are closer to out-of-sample residuals than the ordinary residuals. We can obtain almost as many filtered or deleted prediction errors as the sample size, and thus in both cases the data-splitting variance is avoided.

The rest of the paper is organized as follows. Section 2 introduces notation and definitions for the in-sample and out-of-sample prediction errors. Section 3 defines the filtered prediction errors and Section 4 the deletion prediction errors. These two types of prediction errors are compared in Section 5 to estimate the prediction mean squared error, and in Section 6 they are used for model selection. Finally, Section 7 includes some final remarks.

2 In sample and out-of sample prediction errors

Suppose that an observed time series, $Z_n = (z_1, ..., z_n)'$ is represented by a model with parameter vector $\boldsymbol{\lambda}$. Let us denote the $h-$steps ahead forecasts from observation z_t generated by this model by

$$\widehat{z}_{t+h}\left(\boldsymbol{\lambda}\right) \equiv \widehat{z}_{t+h}(\boldsymbol{\lambda}, Z_t),$$

where the parameter vector is assumed to be known, $Z_t = (z_{r+1}, ..., z_t)'$, $r < t < n$, is the vector of past values of the time series required by the predictor $\widehat{z}_{t+h}$, and $z_1, ..., z_r$ is a set of initial values. Let us also define the population prediction errors as

$$e_{t+h} = z_{t+h} - \widehat{z}_{t+h}(\boldsymbol{\lambda}).$$

When $\boldsymbol{\lambda}$ is unknown, the prediction errors can be defined in various ways. If $\widehat{\boldsymbol{\lambda}}_n = F(Z_n)$ is some estimate of the parameters using the whole span of data, we obtain the $h-$steps ahead forecasts

$$\widehat{z}_{t+h}\left(\widehat{\boldsymbol{\lambda}}_n\right) \equiv \widehat{z}_{t+h}(\widehat{\boldsymbol{\lambda}}_n, Z_t),$$

and the classical residuals defined by

$$\hat{a}_{t+1} = z_{t+1} - \widehat{z}_{t+1}(\widehat{\boldsymbol{\lambda}}_n), \qquad r+1 \leq t \leq n-1.$$

These residuals are also the in-sample one step ahead forecast errors. The residuals depends on the estimation method, and we assume that parameters are estimated by minimizing $\sum_{r+1}^{n} \hat{a}_t^2$, that is by least squares (LS); or by maximum likelihood (ML). If we are interested in h–step ahead forecasts, we can define the $h-$residuals or in-sample h–step ahead prediction errors by

$$\hat{e}_{t+h}^{\text{in}} = z_{t+h} - \hat{z}_{t+h}(\widehat{\boldsymbol{\lambda}}_n), \qquad r \leq t \leq n-h, \tag{1}$$

with $\hat{a}_{t+1} = \hat{e}_{t+1}^{\text{in}}$. In this situation, the parameter vector could alternatively be estimated using a criteria related with the horizon, say $\hat{\boldsymbol{\lambda}}_n \equiv \hat{\boldsymbol{\lambda}}_{n,h} = F_h(Z_n)$, for instance minimizing $\sum \hat{e}_{t+h}^2$ instead of $\sum \hat{e}_{t+1}^2$. The resulting residuals can be used in adaptive forecasting, see i.e. Tiao and Tsay (1994).

Another type of residuals of common use in time series analysis are the predictive residuals or out of sample prediction errors. They are defined by

$$\hat{e}_t(h,m,j) = z_{t+h} - \widehat{z}_{t+h}(\widehat{\boldsymbol{\lambda}}_m, Z_j), \qquad r \leq t \leq n-h,$$

where $\widehat{\boldsymbol{\lambda}}_m = F(Z_m)$, $m \leq t$ is some estimation of the parameters using a set of observations Z_m previous to z_{t+1}, and Z_j, $m \leq j \leq t$, are the past values that are assumed known in order to forecast z_{t+h}. These residuals are often used for checking the forecast precision of the fitted model with out of sample data. Note that the parameters used to build the forecasts have been estimated without including the data to be forecasted. The most important case of out of sample prediction errors are with $m = j = t$,

$$\hat{e}_{t+h}^{\text{out}} \equiv \hat{e}_t(h,t,t) = z_{t+h} - \widehat{z}_{t+h}(\widehat{\boldsymbol{\lambda}}_t, Z_t), \qquad r \leq t \leq n-h \tag{2}$$

in which all the data previous to $z_{t+1}$are included in the computation of the parameters (see, i.e., West, 1996, for alternative schemes to obtain out-of-sample residuals).

3 Filtered prediction errors

Let z_t, $t = 1, ..., n$, be the observed time series following the process and let z_{T+h}, $1 \leq (T+h) \leq n$, be the point to be predicted from z_T. For simplicity we consider, first, that the model is the AR(1), $z_t = \phi z_{t-1} + a_t$, where $|\phi| < 1$ and a_t is white noise. The optimal predictor is, then, $\widehat{z}_{T+h}(\phi, Z_T) = \phi^h z_T$ and the population prediction error is e_{T+h}, where it holds that $e_{T+h} = a_{T+h} + \phi a_{T+h-1} + \cdots + \phi^{h-1} a_{T+1}$.

The out-of-sample approach to estimate e_{T+h} would first estimate the parameter $\hat{\phi}_T = F(Z_T)$, based on the first T observations, obtain the predictor $\hat{z}^{\text{out}}_{T+h} = \widehat{z}_{T+h}(\hat{\phi}_T, Z_T) = \hat{\phi}^h_T z_T$ and then compute the prediction error $\hat{e}^{\text{out}}_{T+h} = z_{T+h} - \hat{\phi}^h_T z_T = e_{T+h} + (\phi^h - \hat{\phi}^h_T) z_T$, where the first and second term are independent. On the other hand, in the in-sample approach, we first estimate $\hat{\phi}_n = F(Z_n)$ with the n observations, then build $\hat{z}^{\text{in}}_{T+h} = \widehat{z}_{T+h}(\hat{\phi}_n, Z_T) = \hat{\phi}^h_n z_T$ and compute the prediction error $\hat{e}^{\text{in}}_{T+h} = z_{T+h} - \hat{\phi}^h_n z_T = e_{T+h} + (\phi^h - \hat{\phi}^h_n) z_T$ where now the first and second term are correlated because $\hat{\phi}_n$ already contains, implicitly, the values $a_{T+1}, ..., a_{T+h}$.

We could alternatively estimate the prediction error by using an estimate that (i) does not include the information provided by $a_{T+1}, ..., a_{T+h}$, as in the case of $\hat{e}^{\text{out}}_{T+h}$, and (ii) includes the information provided by $a_{T+h+1}, ..., a_n$, in order to improve the accuracy of the estimations. This can be done by building a filtered series y_t (non-observable) that is free from the effect of the innovations $a_{T+1}, ..., a_{T+h}$, as follows. For $t = 1, ..., T$; we have that $y_t = z_t$. For $t = T+1, ..., T+h$; the series y_t should ignore the information of the corresponding innovations assuming that $a_{T+1} = \cdots = a_{T+h} = 0$. Finally, for $t > T+h$, the series y_t would take again into account the contemporaneous innovations. Then,

$$\begin{aligned} y_t &= z_t = \phi y_{t-1} + a_t \ ; \quad t = 2, ..., T, \\ y_{T+1} &= \phi z_T = \phi y_T, \\ &\vdots \\ y_{T+h} &= \phi^h z_T = \phi^h y_T, \\ y_t &= \phi y_{t-1} + a_t \ ; \quad t = T+h+1, ..., n. \end{aligned} \tag{3}$$

Figure 1 shows an example of this filtered series to evaluate the five-step-ahead prediction error from $t = 25$. The observed series is a realization of an AR(1) with $\phi = 0.9$ and $n = 50$. The solid line represents the observed

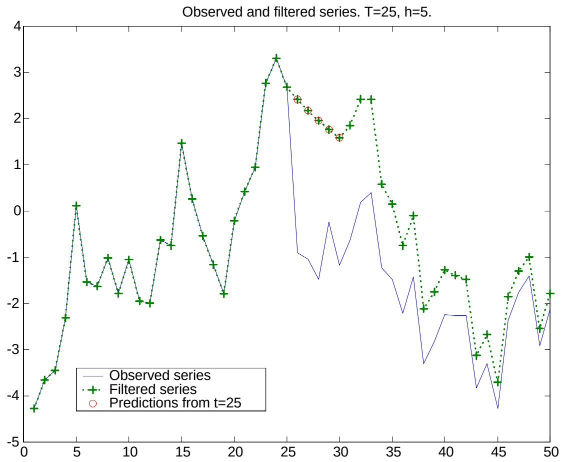

Figure 1: Observed and filtered series from the model $z_t = 0.9z_{t-1} + a_t$, to estimate $e_{25}(5)$.

series and the filtered series is represented by the dotted line with the $(+)$ symbol. The filtered series for $t = 1, ..., 25$ is still the original series. For $t = 26, ..., 30$, the filtered series are the predictions from $t = 25$, and are represented by circles. For $t = 31, ..., n$, the filtered series is obtained following (3).

This filtered series can be extended to a more general ARMA case. Let us assume that z_t represents deviations from some mean μ and that it admits the ARMA(p, q) representation:

$$\phi(B)z_t = \theta(B)a_t, \tag{4}$$

where B is the backshift operator; $\phi(B) = (1-\sum_{i=1}^{p} \phi_i B^i)$ and $\theta(B) = (1-\sum_{i=1}^{q} \theta_i B^i)$ are polynomial operators on B such that $\phi(B) = 0$ and $\theta(B) = 0$ have all their roots outside the unit circle, and a_t be a sequence of independent identically distributed random variables with zero mean and variance σ^2. Let us define

$$\varepsilon_t = a_t - \sum_{j=1}^{h} a_{T+j} D_t^{(T+j)}$$

where $D_t^{(t_0)} = 1$ if $t = t_0$ and $D_t^{(t_0)} = 0$ otherwise. The sequence ε_t verifies $\varepsilon_t = a_t$ if $t = 1, ..., T, T + h + 1, ..., n$ and $\varepsilon_t = 0$ otherwise. The filtered series that do not contain the information of the innovations $a_{T+1}, ..., a_{T+h}$ is, therefore,

$$y_t = \phi^{-1}(B)\theta(B)\varepsilon_t = \psi(B)\varepsilon_t, \tag{5}$$

where $\psi(B) = \phi^{-1}(B)\theta(B)$. The filtered series y_t follows, then, the model

$$\phi(B)y_t = \theta(B)\left(a_t - \sum_{j=1}^{h} a_{T+j} D_t^{(T+j)}\right). \tag{6}$$

Under the assumption of normality of the disturbances, the parameters of the filtered series can be estimated by ML. Let us denote $\boldsymbol{\phi}\equiv (\phi_1, ..., \phi_p)'$, $\boldsymbol{\theta} \equiv (\theta_1, ..., \theta_q)'$. From (5), and taking initial values $\mathbf{y}_0 \equiv (y_1 = z_1, y_2 = z_2, ..., y_p = z_p)$ and $\mathbf{a}_0 \equiv (a_p = 0, a_{p-1} = 0, ..., a_{p-q+1} = 0)$, the conditional log-likelihood can be expressed for a given parameter values $\dot{\boldsymbol{\phi}}, \dot{\boldsymbol{\theta}}$ as:

$$\sum_{\substack{i=p+1\\ i\neq T+1,...,T+h}}^{n} l(y_i|y_{i-1}, ..., y_1, \mathbf{a}_0, \dot{\boldsymbol{\phi}}, \dot{\boldsymbol{\theta}}) = -\frac{n-p-h}{2}\ln 2\pi - \frac{n-p-h}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}S(\dot{\boldsymbol{\phi}}, \dot{\boldsymbol{\theta}}), \tag{7}$$

where

$$S(\dot{\boldsymbol{\phi}}, \dot{\boldsymbol{\theta}}) = \sum_{\substack{i=p+1\\ i\neq T+1,...,T+h}}^{n} \dot{a}_i^2. \tag{8}$$

Following Chang *et al* (1988), the expression (7) coincides with the concentrated log-likelihood of a model of the original observed series z_t with innovational outliers at $t = T+1, ..., T+h$. Then, the ML estimation of the parameters free of the effects of $a_{T+1}, ..., a_{T+h}$ can easily be obtained by assuming innovational outliers at points $T+1, ..., T+h$, which leads to the model

$$\phi(B)z_t = \theta(B)\left(a_t + \sum_{i=1}^{h} w_i D_t^{(T+i)}\right). \tag{9}$$

Alternatively, a LS estimator could be obtained by minimizing the squared sum (8). Let us denote $\hat{\boldsymbol{\lambda}}_n^{(IO)} = \left(\hat{\boldsymbol{\phi}}_n^{(IO)}, \hat{\boldsymbol{\theta}}_n^{(IO)}\right)'$, where $\hat{\boldsymbol{\phi}}_n^{(IO)} \equiv \left(\hat{\phi}_1^{(IO)}, ..., \hat{\phi}_p^{(IO)}\right)'$ and $\hat{\boldsymbol{\theta}}_n^{(IO)}\equiv \left(\hat{\theta}_1^{(IO)}, ..., \hat{\theta}_q^{(IO)}\right)'$, to the parameter vector of ML or LS estimates of $\boldsymbol{\lambda} \equiv (\boldsymbol{\phi}, \boldsymbol{\theta})'$ in (9), where the subscript indicates that the estimator is based in the whole span of data and the superscript shows that the estimation has been made treating the points to be predicted as innovational outliers. Then, following Mann and Wald (1943),

$$\hat{\boldsymbol{\lambda}}_n^{(IO)} \xrightarrow{p} \boldsymbol{\lambda}.$$

Although the estimates are different at each T and h, this aspect is not considered, for simplicity, in the notation. Let $\hat{z}_{T+h}^{\text{filter}} = \hat{z}_{T+h}\left(\hat{\boldsymbol{\lambda}}_n^{(IO)}, Z_T\right)$ be the prediction of z_{T+h} from z_T using the estimated model $\hat{\phi}_n^{(IO)}(B)z_t = \hat{\theta}_n^{(IO)}(B)\hat{a}_t$, and let $\hat{e}_{T+h}^{\text{filter}}$ be the corresponding estimated prediction error. This prediction error will be denoted as filtered prediction errors since it is obtained from a series that has been filtered out from the innovations of the observations to be predicted. It can alternatively be calculated from the estimates $\hat{w}_i$ in (9), since it holds that these estimates are the predictions of the innovations $a_{T+1}, ..., a_{T+h}$, and also that $\hat{a}_{T+1} = \cdots = \hat{a}_{T+h} = 0$. Then,

$$\hat{e}_{T+h}^{\text{filter}} = z_{T+h} - \hat{z}_{T+h}\left(\hat{\boldsymbol{\lambda}}_n^{(IO)}, Z_T\right) = \hat{w}_h + \hat{\psi}_1^{(IO)}\hat{w}_{h-1} + \cdots + \hat{\psi}_{h-1}^{(IO)}\hat{w}_1, \tag{10}$$

After estimating model (9) for $T = p+1, ..., n$; we will have $n-p-h$ h-steps ahead filtered prediction errors. Applying the results in Peña (1990) it can be shown that, for $h = 1$,

$$\hat{e}^{\text{filter}}_{T+1} = \frac{\hat{e}^{\text{in}}_{T+1}}{(1-d_{T+1})}, \tag{11}$$

where d_{T+1} is the so-called leverage of the observation. Therefore, as $1/n \leq d_{T+1} \leq 1$ we have that $\left|\hat{e}^{\text{filter}}_{T+1}\right| \geq \left|\hat{e}^{\text{in}}_{T+1}\right|$. Expression (11) helps to give a geometric interpretation of the difference between classical residuals and the proposed filtered prediction errors. It is well known that the higher the leverage d_t of an observation z_t, the larger its influence on the parameter estimation and, hence, the larger the data-snooping bias in predicting that observation with the estimated predictor. Expression (11) says that each residual should be corrected by that leverage in order to avoid the effect of the observation in the estimation.

4 Deletion prediction errors

A second natural way to estimate the parameters free of the effect of the observations $z_{T+1,...,}z_{T+h}$ is. to assume that these observations are missing values. Peña (1987) showed that the parameters obtained under this hypothesis are, for large sample size, the same as those obtained assuming additive outliers at these positions. Thus, we can obtain parameters estimates that do not contain the values $z_{T+1,...,}z_{T+h}$, by estimating the model

$$z_t = \sum_{j=1}^{h} \omega_j D_t^{(T+j)} + \psi(B)a_t,$$

or, equivalently,

$$\phi(B)\left(z_t - \sum_{j=1}^{h} \omega_j D_t^{(T+j)}\right) = \theta(B)a_t. \tag{12}$$

Let us denote by $\hat{\boldsymbol{\lambda}}_n^{(AO)} = \left(\hat{\boldsymbol{\phi}}_n^{(AO)}, \hat{\boldsymbol{\theta}}_n^{(AO)}\right)'$, with $\hat{\boldsymbol{\phi}}_n^{(AO)} \equiv \left(\hat{\phi}_1^{(AO)}, ..., \hat{\phi}_p^{(AO)}\right)'$ and $\hat{\boldsymbol{\theta}}_n^{(AO)} \equiv \left(\hat{\theta}_1^{(AO)}, ..., \hat{\theta}_q^{(AO)}\right)'$, to the parameter vector of ML or LS estimates of $\boldsymbol{\lambda} \equiv (\boldsymbol{\phi}, \boldsymbol{\theta})'$ in (12), where the subscript indicates that the estimator is based in the whole span of data and the superscript shows that the estimation has been made treating the points to be predicted as additive outliers. Then,

$$\hat{\boldsymbol{\lambda}}_n^{(AO)} \xrightarrow{p} \boldsymbol{\lambda}.$$

We define then the deletion prediction errors by

$$\hat{e}^{\text{del}}_{T+h} = z_{T+h} - \hat{z}^{\text{del}}_{T+h}, \qquad r \leq T \leq n-h,$$

with $\widehat{z}^{\text{del}}_{T+h} = \widehat{z}_{T+h}\left(\hat{\boldsymbol{\lambda}}^{(AO)}_{n}, Z_T\right)$. These prediction errors, for $h = 1$, where used by Peña (1990) for building influence measures in time series. They are closely related with the conditional residuals (Haslett and Hayes, 1998; Hasslet, 1999), derived for linear models with general covariance structure, and with applications in many fields (see e.g. Cressi, 1991). However, these last residuals are computed assuming that the covariance matrix is known, or estimated using the whole sample, and so they are less useful to obtain an estimation of $h-$steps ahead prediction errors free of the effect of a set of observations.

5 Application to the estimation of MSPE

5.1 General considerations

In this section, we apply the proposed prediction errors, filtered and deletion, to the evaluation of the mean squared prediction error (MSPE) and to the comparison of competing predictors. It is well known that the comparison of in-sample MSPEs tend to favor highly parameterized models and for this reason forecast-accuracy comparison is usually carried out by splitting the sample and estimating the predictors in the first part and evaluating the MSPEs on the second. This out-of-sample comparison should also be made with caution, since there is the danger of incurring in just the opposite effect. Namely, the larger sampling variability induced for splitting the sample will tend to mask actual features.

The potential advantage of estimating the MSPEs with the filtered prediction errors comes from the following result, that it is proved in the Appendix.

Proposition 1 Let $z_t = \phi z_{t-1} + a_t$, with $|\phi| < 1$ and a_t is a sequence of martingale differences with zero mean and variance σ^2. Let $\hat{z}^{in}_{T+1}$ be the one-step ahead prediction of z_T, $1 < T+1 \leq n$, using the LS estimator $\hat{\phi}_n$, and let $\hat{z}^{filter}_{T+1}$ be the prediction using the estimator $\hat{\boldsymbol{\phi}}^{(IO)}_n$ that minimizes (8). Then

$$\begin{array}{ll} a) & E\{\hat{z}^{in}_{T+1} a_{T+1}\} = O(T^{-1}), \\ b) & E\{\hat{z}^{filter}_{T+1} a_{T+1}\} = O(T^{-2}). \end{array}$$

This result indicates that the information that the predictor $\hat{z}^{\text{filter}}_{T+1}$ contains about the future innovation a_{T+1}, due to the estimation process, is much lower than the usual in-sample predictor $\hat{z}^{\text{in}}_{T+h}$ .The key element in the proof of this proposition is that (8) explicitly excludes the residuals corresponding to the point to be predicted. Therefore, it seems reasonable to conjecture similar results (with appropriate moments restrictions on a_t) in more general situations.

When using deletion prediction error, we explicitly get rid of the observations to be predicted $z_{T+1}, ..., z_{T+h}$. For, instance, in the AR(1) case and for $h = 1$, we would use the estimator (obtained with a recursive procedure)

$$\widehat{\phi}^{(AO)} = \frac{\sum_{t=2}^{n} x_t x_{t-1}}{\sum_{t=2}^{n} x_t^2},$$

where $x_t = z_t$ for $t \neq T+1$, and $x_{T+1} = \widehat{\phi}^{(AO)}(1 + \widehat{\phi}^{(AO)2})^{-1}(z_{T+1} + z_{T-1})$. The estimated parameter does not contain explicitly the data z_{T+1}. However, z_{T+k} , $k > 1$, is correlated with z_{T+1} and, therefore, its information is not completely discarded. Hence, if the process has a strong serial correlation, the predictor $\hat{z}_{T+h}^{\text{del}}$ can still incur in a significative data-snooping bias.

5.2 Empirical comparison of MSPEs in nested models

In this subsection, we compare in a simulation study the properties of the estimated MSPE of competing predictors using different types of prediction errors. The first type is the in-sample prediction error, $\hat{e}_{t+h}^{\text{in}}$, obtained from the LS estimation of the predictor $\hat{z}_{t+h}^{\text{in}}$. From these prediction errors we compute two estimators of MSPE : the average of squared prediction errors, $\hat{V}^{\text{in}}(h)$, and the corrected by degrees of freedom average, $\hat{V}^{\text{in-c}}(h)$. For instance, for an AR(p) they are

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

$$\hat{V}_p^{\text{in-c}}(h) = \hat{V}_p^{\text{in}}(h)\frac{n-h-p+1}{n-h-2p+1}. \tag{14}$$

The second type is the out-of-sample prediction error. For computing them, the estimation subsample increases recursively, and the models are re-estimated by LS in order to include all past data prior to the forecast origin. We compute two estimators of MSPE. We denote $\hat{e}_{t+h}^{\text{out-50}}$ to the out-of-sample prediction error when the initial estimation subsample is the 50% of the total sample, and $\hat{e}_{t+h}^{\text{out-75}}$ when the initial estimation subsample is the 75% of the total sample. In both cases, the MSPE is estimated by averaging the squared prediction errors, and the estimates are denoted as $\hat{V}^{\text{out-50}}(h)$ and $\hat{V}^{\text{out-75}}(h)$, respectively. In the case of an

AR(p) they are

$$\hat{V}_p^{\text{out-50}}(h) = \frac{\sum_{t=[0.5n]}^{n-h} \left(\hat{e}_{t+h}^{\text{out-50}}\right)^2}{n-h-[0.5n]+1}, \tag{15a}$$

$$\hat{V}_p^{\text{out-75}}(h) = \frac{\sum_{t=[0.75n]}^{n-h} \left(\hat{e}_{t+h}^{\text{out-75}}\right)^2}{n-h-[0.75n]+1}, \tag{15b}$$

where $[\cdot]$ represents the integer part. Finally, the third and fourth type of prediction errors are the filtered prediction errors, $\hat{e}_{T+h}^{\text{filter}}$ and the deletion prediction errors $\hat{e}_{T+h}^{\text{del}}$, that are also obtained using LS estimation and the MSPE is computed by averaging the available squared prediction errors. The corresponding estimators are

$$\hat{V}_p^{\text{filter}}(h) = \frac{\sum_{t=p}^{n-h} \left(\hat{e}_{t+h}^{\text{filter}}\right)^2}{n-h-p+1}, \tag{16}$$

$$\hat{V}_p^{\text{del}}(h) = \frac{\sum_{t=p}^{n-h} \left(\hat{e}_{t+h}^{\text{del}}\right)^2}{n-h-p+1}. \tag{17}$$

In the first experiment, the following AR(2) model is used: $(1-0.9B)(1-\phi B)z_t = a_t$, where $a_t \sim N(0,1)$, and $\phi = 0.3, 0.2, 0.1, 0.0$. Two competing predictors, an AR(1) and an AR(2), estimated by LS, are used to generate predictions at $h = 1, ..., 5$ (although only $h = 1, 5$ will be reported). In each replication, we generate a random sample of the process of size 205. The first 100 observations are dropped to assure stationary initial conditions. The subsequent $n = 100$ observations are used to estimate the predictors and the prediction errors. With these prediction errors we estimate the MSPE for each predictor using expressions (13) to (17) and we will denote by $\hat{V}_p$ $(p = 1, 2)$ to these estimates when derived from the *AR(p)* model. Finally, the last 5 observations are used to evaluate the actual out-of-sample prediction errors of the estimated predictors. The experiment has been run two times. In the first run, the population out-of-sample MSPE of the estimated predictors is estimated by averaging the squared errors, at each horizon, of predicting those five out-of-sample observations along 100,000 replications. Let us denote by V_p $(p = 1, 2)$ to these empirical MSPEs, at each horizon. In the second run, we generate 20,000 replications and obtain, in each replication, the estimates (13) to (17). With these 20,000 set of estimates we estimate some features like bias and mean squared error (MSE) of $\hat{V}_p$ using V_p as population values.

Tables 1 and 2 summarize the results. The first four columns of Table 1 shows the empirical bias, for $h = 1$ and $h = 5$, of $\hat{V}_p$. This empirical bias is computed as the sampling average of $\hat{V}_p - V_p$ along the 20,000 replications. Columns fifth to eighth show the empirical mean squared error (MSE) of these estimators $\hat{V}_p$ as the average of $\left(\hat{V}_p - V_p\right)^2$ along the replications. From this table we can extract the following conclusions:

1. Columns one to four show that, as expected, the estimator $\hat{V}_p^{\text{in}}$ has a large negative bias that it is larger in the model with more parameters. This confirms the well known result that $\hat{V}_p^{\text{in}}$ underestimate the true MSPE. This effect is partly alleviated with $\hat{V}_p^{\text{in-c}}$ and $\hat{V}_p^{\text{del}}$. However, their negative bias is still large. Regarding $\hat{V}_p^{\text{in-c}}$, it should be noted that its correction for degrees of freedom, which is designed to diminish the bias of the one-step-ahead prediction error, can no longer be appropriate if $h > 1$. It is well known (see, i.e., Fuller and Hasza, 1981) that the bias of the in-sample estimator of MSPE depends for $h > 1$ on the dynamic structure of the model, and cannot be avoided just by a degrees-of-freedom correction. It can then be said that the performance of the estimator $\hat{V}_p^{\text{in-c}}$ at $h > 1$ is highly model-dependent. This can explain why the relative behavior of the bias of $\hat{V}_p^{\text{in-c}}$ with respect to $\hat{V}_p^{\text{in}}$ is better at $h = 1$. On the other hand, estimators $\hat{V}_p^{\text{out-50}}$, $\hat{V}_p^{\text{out-75}}$, and $\hat{V}_p^{\text{filter}}$ significantly reduces the bias. This bias is, in general, lower for $\hat{V}_p^{\text{out-50}}$. It can be concluded that $\hat{V}_p^{\text{in}}$, $\hat{V}_p^{\text{in-c}}$, and $\hat{V}_p^{\text{del}}$ have high negative bias and $\hat{V}_p^{\text{out-50}}$, $\hat{V}_p^{\text{out-75}}$, and $\hat{V}_p^{\text{filter}}$ have low bias. Therefore, as expected from the theory, the proposed $\hat{V}_p^{\text{filter}}$ significantly reduces the data snooping bias with respect to other in-sample procedures.

2. Columns five to eight show that the estimators based on out-of-sample prediction errors have very large MSE, whereas estimators that use the whole span of data are, as expected, more accurate. At $h = 1$, these estimators have comparable MSE with a ranking that depends on the model. At $h = 5$, the estimators with lower MSE are $\hat{V}_p^{\text{in-c}}$ and $\hat{V}_p^{\text{in}}$. At $h = 5$, the relative increment of MSE of $\hat{V}_p^{\text{filter}}$ and $\hat{V}_p^{\text{del}}$ with respect to $\hat{V}_p^{\text{in}}$ and $\hat{V}_p^{\text{in-c}}$ can be explained by the added variability provoked by the intervention variables in the estimation.

We can conclude that the estimator $\hat{V}_p^{\text{filter}}$ offers a good compromise between bias and MSE. Estimators with some lower bias ($\hat{V}_p^{\text{out-50}}$ and $\hat{V}_p^{\text{out-75}}$) incur into a very large MSE, and estimators with some lower MSE, incur into large negative bias.

Table 2 shows the ability of the different estimators of MSPE to detect the more efficient predictor, based on the previous experiment. Let us denote $d = V_2 - V_1$ to the population MSPE differential. This value is estimated using the empirical values V_p obtained with the 100,000 replications of the first run of the experiment. Positive values of d means that the estimated AR(1) is the efficient predictor, whereas negative values means that the AR(2) is more efficient. It can be seen in Table 2 that for $\phi = 0.3$ and $\phi = 0.2$ the AR(2) seems to be more efficient, whereas for $\phi = 0.1$ and $\phi = 0.0$, the AR(1) would be preferred. For each replication of the second run of the experiment, we compute the differential $\hat{d} = \hat{V}_2 - \hat{V}_1$, for each estimator of MSPE. Columns one and two of Table 2 show, for each estimator and value of ϕ, the sampling average of the 20,000 replications of $\hat{d}$ (denoted as $E(\hat{d})$). Columns three and four show the empirical average difference $E(\hat{d} - d)$. This figure is the sampling average of $\hat{d} - d$ along the 20,000 replications. Negative values of $E(\hat{d} - d)$ means that there is a bias toward the AR(2), and positive values represents a bias toward the AR(1). In a similar fashion, columns five and six report the empirical MSE of $\hat{d}$. Finally, the last two columns report the proportion of times that the AR(2) predictor was found to have lower empirical MSPE (and therefore would be preferred), for each estimator, than the AR(1). A good estimator of MSPE should have high values in these two last columns for $\phi = 0.3$ and $\phi = 0.2$, but low values for $\phi = 0.1$ and $\phi = 0.0$. The following conclusions can be obtained from this table:

- $\hat{V}_p^{\text{in}}$, has a biased toward the AR(2). In the last two columns, the proportion of times that AR(2) is found to be more efficient than the AR(1) is very high. This is especially relevant with $\phi = 0.1$ and $\phi = 0.0$, where the AR(2) predictor is no longer efficient. This is in agreement with columns 1 to 4 of Table 1 that showed that $\hat{V}_p^{\text{in}}$ always incurs in negative bias in the estimation of the MSPE. Besides, this negative bias is much larger in the AR(2). As a consequence, the estimated loss differential $\hat{d}$ has, on average, large negative values (columns 1 and 2) which are always lower than the actual ones (columns 3 and 4). Then, although the estimators of MSPE and the differential are very efficient (see columns 5 to 8 in Table 1 and columns 5 and 6 in Table 2), the high negative bias leads to the perception that the AR(2) is more efficient than it actually is.

- Estimators $\hat{V}_p^{\text{in-c}}$ and $\hat{V}_p^{\text{del}}$ produce a very efficient estimation of d with lower bias than $\hat{V}_p^{\text{in}}$ (columns 3 to 6). At $h = 1$ their behavior is very similar and they still incur in high negative bias. As a consequence, they choose the AR(2) with probability larger than 0.5 when $\phi = 0.1$, whereas the efficient predictor is the AR(1). Also, when $\phi = 0$, they still chose the AR(2) in more than 35% of the replications. At $h = 5$, their relative behavior differs. Regarding $\hat{V}_p^{\text{del}}$, it should be noted that this estimator can still incur in data-snooping bias, since the effect of the innovations at the predicted points is not efficiently discarded.

Besides, since such an effect will depend on the dynamic of the process, the performance of $\hat{V}_p^{\text{del}}$ can be highly model-dependent. This model-dependency can also explain the different relative behavior of this estimator at $h = 1$ with respect to $h = 5$.

- The proposed $\hat{V}_p^{\text{filter}}$ has a high tendency to choose the more efficient predictor (see columns 7 and 8). This can be explained by the combination of low values of $E(\hat{d} - d)$ and $E\left[(\hat{d} - d)^2\right]$. When $\phi = 0.3$ this estimator leads to chose the more efficient AR(2) in a proportion similar to other in-sample estimators. When $\phi = 0.0$, this estimator leads to chose the AR(1) in a proportion similar to out-of-sample estimators. Besides, when $\phi = 0.1$, this estimator leads to chose the AR(1) both at $h = 1$ and $h = 5$.

- The out-of-sample estimators $\hat{V}_p^{\text{out-50}}$ and $\hat{V}_p^{\text{out-75}}$ tend to favor the AR(1) model. This is a consequence of the mentioned data-splitting variance: due to the lower amount of available data, the more parsimonious model seems to be more efficient than they would be if the whole span of data where used. The empirical $E\left[(\hat{d} - d)^2\right]$ is high, especially at $h = 5$. At $h = 1$, their tendency to the AR(1) is confirmed in column 7. For instance, when $\phi = 0.3$, these estimator chose the AR(1) almost 50% of the times. At $h = 5$ the high variability (large values of $E\left[(\hat{d} - d)^2\right]$) can explain the low discriminating power of these estimators with respect to $h = 1$ (column 7 and 8). Therefore, although their ability to detect the efficient predictor is bigger than with the in-sample residuals, their performance is poorer than the proposed $\hat{V}_p^{\text{filter}}(h)$.

To better understand the differences between the alternative estimators we have performed two more experiments where there are more than two competing models. In the first experiment, the competing predictors are an AR(1), an AR(2), and an AR(3). The true model is the AR(2): $(1 - 0.9B)(1 - \phi B)y_t = a_t$, with $\phi = 0.3, 0.1$. In the last experiment the selection is an AR(p) with $p = 1, 2, ..., 5$, when the true model is the AR(3): $(1 - 0.9B)(1 - 0.5B)(1 - \phi B)y_t = a_t$, with $\phi = 0.5, 0.1$. The experiments have the same structure than the previous one. The population MSPE of each estimated predictor is estimated with 100,000 replications and the proportion of times that each predictor is found to have lower estimated MSPE, for each estimator, is made with 20,000 replications. Results for the first experiment are summarize in Table 3, Panel A. From this table, the following can be concluded:

- When $\phi = 0.3$, the best predictor is the AR(2). The proposed $\hat{V}_p^{\text{filter}}$ is the estimator that chooses this predictor with highest probability (0.698 at $h = 1$ and 0.505 at $h = 5$). The second choice for this estimator is, at $h = 1$ the AR(3), which is also the second best predictor. However, at $h = 5$ it select the AR(1) and the AR(3) with similar proportion. The remaining estimators select the AR(2) as best predictor with lower probability than $\hat{V}_p^{\text{filter}}$ for quite different reasons. The in-sample estimators $\hat{V}_p^{\text{in}}$ and $\hat{V}_p^{\text{in-c}}$ have a higher preference for the overparameterized AR(3), especially $\hat{V}_p^{\text{in}}$; whereas the out-of-sample estimators have a higher preference for the parsimonious AR(1). The out-of-sample estimators clearly select the AR(1) as their second best predictor even though it is the least efficient. The behavior of the estimator $\hat{V}_p^{\text{del}}$ is similar to $\hat{V}_p^{\text{in-c}}$ at $h = 1$. At $h = 5$, however, this estimator has very low discriminating power.

- When $\phi = 0.1$, the best predictor is the AR(1). In this case, the estimators that chooses this predictor with highest probability at $h = 1$ are $\hat{V}_p^{\text{out-50}}$ and then the proposed $\hat{V}_p^{\text{filter}}$. At $h = 5$, the estimator that chooses the AR(1) with highest probability is $\hat{V}_p^{\text{filter}}$. The out-of-sample estimators have, in this case, two reasons to choose the AR(1): it is very efficient and it is more parsimonious. This explains why they have better relative performance with respect to the case with $\phi = 0.3$. The remaining in-sample estimators completely fail, at $h = 1$, to select the efficient predictor, due to their bias toward the more parameterized (and inefficient) models. However, at $h = 5$, $\hat{V}_p^{\text{in-c}}$ and $\hat{V}_p^{\text{del}}$ show a preference toward the AR(1), whereas $\hat{V}_p^{\text{in}}$ still prefers the overparameterized AR(3). The relative behavior of $\hat{V}_p^{\text{in-c}}$ and $\hat{V}_p^{\text{del}}$ with h is, therefore, different with $\phi = 0.3$ and $\phi = 0.1$. This behavior is confusing and can be explained by the high model-dependency of these estimators. Therefore, in a real situation, it could be difficult to foresee the performance of these estimators.

The results for the last experiment are summarized in Table 4, Panel A. For the sake of brevity, only $h = 1$ is reported. The conclusions of this experiment are similar to the previous one. With $\phi = 0.5$ the more efficient predictor is the AR(3) and the proposed $\hat{V}_p^{\text{filter}}$ is the estimator that select this model with highest probability. The remaining in-sample estimators have worse performance for their tendency to select overparameterized models. Conversely, the out-of-sample estimators show a larger tendency to chose the more parsimonious models, although they are less efficient. With $\phi = 0.1$, the AR(2) is more efficient. As before, the best estimators are $\hat{V}_p^{\text{out-50}}$ and $\hat{V}_p^{\text{filter}}$. The remaining estimators show a much lower discriminating power.

6 Application to order selection

Many criteria for selecting the true orders p, q in ARMA(p, q) models have been proposed, see i.e. Choi (1992). The motivation for using these criteria is, again, to avoid the bias of in-sample prediction errors toward overparameterized models. Some of these criteria are based on minimization of functions of the form

$$G(p, q) = n \ln \hat{\sigma}^2_{p,q} + (p + q) g(n), \tag{18}$$

where $p = 0, 1, , ..., p^*$, $q = 0, 1, ..., q^*$, with p^*, q^* some pre-determined upper bounds; and $\hat{\sigma}^2_{p,q}$ is an estimate of the residual variance of the fitted ARMA(p, q) model. The term $g(n)$ is a penalty factor to discourage the fitting of models with too many parameters. If $g(n) = \ln(n)$, then (18) becomes the BIC (Schwartz 1978); if $g(n) = 2$, we get the AIC (Akaike 1974); if $g(n) = c \ln (\ln n)$ expression (18) is the Hannan and Quinn's (1979) criterion (HQ). Another criterion is to minimize the Final Prediction Error (FPE) of Akaike (1969). The FPE is the following estimate of the one-step MSPE:

$$\text{FPE} = \hat{\sigma}^2_{p,q} \frac{n + p + q}{n - p - q}.$$

The main goal of the penalty function $g(n)$ in (18) is to avoid the tendency of overparameterization of $\hat{e}^{\text{in}}_{T+1}$. Since $\hat{e}^{(IO)}_{T+1}$ and $\hat{e}^{(AO)}_{T+1}$ are also designed to avoid this effect, it is interesting to compare the capabilities of these prediction errors in order selection, where the selected orders will be the ones that achieves lower MSPE at $h = 1$, estimated with the proposed prediction errors. In a similar fashion, FPE is designed to select the model with lower MSPE at $h = 1$, but avoiding the downward bias of the residual variance. Again, FPE shares this feature with the proposed prediction errors.

In order to compare the order selection criteria and the proposed prediction errors we have included these criteria in the second experiment shown in Tables 3 and 4 Panel B (HQ criterion uses the value $c = 3$). In Panel B of these tables, each figure is the percentage of times that a model is selected by each criterion. From this table we can extract the following conclusions:

- FPE and AIC have similar performance and also similar to $\hat{V}^{\text{filter}}_p$. They tend to favor the most efficient predictor.

- BIC and HQ have also similar behavior and also similar to $\hat{V}^{\text{out-50}}_p$ and $\hat{V}^{\text{out-75}}_p$, showing a tendency to underfit the model. This tendency is more acute when $\phi = 0.1$.

We can conclude that $\hat{V}_p^{\text{filter}}$ at $h = 1$ is also useful as a order selection criterion, similar to FPE or AIC. Besides, the filtered prediction error can be applied to build an efficient predictor at any horizon whereas FPE and AIC only supply information related with one step ahead prediction.

7 Conclusions

In this article, we have introduced two new types of prediction errors. The motivation is to find in-sample prediction errors with similar properties than out-of-sample ones. We have shown that the filtered prediction errors, that are computed with parameter values estimated without taking into account the innovation at each point, have several advantages for time series analysis. These prediction errors are easily computed by estimating the parameters assuming and innovative outlier at each sample point. We have seen that these prediction errors have good performance (1) for computing an efficient estimate of the forecast accuracy of the model, (2) for comparing alternative predictors at a given horizon, and (3) for order selection. The deleted prediction errors, that are very useful for building measures of influence on time series and have found applications outside the time series field, are less useful for these purposes. Although our analysis is limited to some applications in univariate time series analysis, the proposed prediction errors can also be of interest in many other time series applications. Further research will undoubtedly extend the applicability of the proposed procedures.

8 Acknowledgment

This research has been sponsored by DGES (Spain) under project PB-96-0111, PB96-0339 and the Cátedra BBVA de Calidad .

A Proof of proposition 1

a) The predictor with the LS estimator $\hat{\phi}_n$ is $\hat{z}_{T+1}^{in} = \hat{\phi}_n z_T = z_{T+1} - \hat{a}_{T+1}$, where $\hat{a}_{T+1}$ is the LS residual. Then $E(\hat{\phi}_n z_T a_{T+1}) = E(z_{T+1}a_{T+1}) - E(\hat{a}_{T+1}a_{T+1})$, with $E(z_{T+1}a_{T+1}) = \sigma^2$. Let $Z = (z_1, z_2, ..., z_{n-1})'$, $U = (a_2, a_3, ..., a_n)'$, and $\hat{U} = (\hat{a}_2, \hat{a}_3, ..., \hat{a}_n)'$. Then, by the properties of OLS estimation: $\hat{U} = MU$, where $M = I_{n-1} - Z(Z'Z)^{-1}Z'$, and I_{n-1} is the identity matrix of size $n-1$. Then, it can be verified that $\hat{a}_{T+1} = a_{T+1} -$ $\left(\sum_{t=2}^{n} z_{t-1}a_t / \sum_{t=2}^{n} z_{t-1}^2\right) z_T$. Therefore,

$$E(\hat{a}_{T+1}a_{T+1}) = \sigma^2 - n^{-1}E\left(\hat{\gamma}_z^{-1}\sum_{t=2}^{n} z_{t-1}a_t z_T a_{T+1}\right),$$

where $\hat{\gamma}_z = n^{-1}\sum_{t=2}^{n} z_{t-1}^2$. Applying a Taylor expansion of $\hat{\gamma}_z^{-1}$ around the population value γ_z^{-1} it can be obtained that

$$E\left(\hat{\gamma}_z^{-1}\sum_{t=2}^{n} z_{t-1}a_t z_T a_{T+1}\right) = \sigma^2 + O\left[E\left\{(\hat{\gamma}_z - \gamma_z)\sum_{t=2}^{n} z_{t-1}a_t z_T a_{T+1}\right\}\right],$$

where it holds that

$$E\left\{(\hat{\gamma}_z - \gamma_z)\sum_{t=2}^{n} z_{t-1}a_t z_T a_{T+1}\right\} = n^{-1}\sum_{s=2}^{n}\sum_{t=2}^{n} E\left(z_{s-1}^2 z_{t-1}a_t z_T a_{T+1}\right) - \gamma_z^2\sigma^2. \tag{19}$$

Noting that when $s = t = T+1$ the first term on the right hand side of (19) is null, it can be written that

$$n^{-1}\sum_{s=2}^{n}\sum_{t=2}^{n} E\left(z_{s-1}^2 z_{t-1}a_t z_T a_{T+1}\right) = n^{-1}\sum_{s=2}^{n} E\left(z_{s-1}^2 z_T^2 a_{T+1}^2\right) + n^{-1}\sum_{\substack{s=2\\ s\neq T+1}}^{n}\sum_{\substack{t=2\\ t\neq T+1}}^{n} E\left(z_{s-1}^2 z_{t-1}a_t z_T a_{T+1}\right). \tag{20}$$

It can easily be seen that $n^{-1}\sum_{s=2}^{n} E\left(z_{s-1}^2 z_T^2 a_{T+1}^2\right) = O(1)$. The second term on the right hand side of (20) is more involved. It can be seen, however, that, if $s < T+1$ and $t < T+1$ it is null. The same result is obtained if $s < T+1$ and $t > T+1$. Therefore,

$$\sum_{\substack{s=2\\ s\neq T+1}}^{n}\sum_{\substack{t=2\\ t\neq T+1}}^{n} E\left(z_{s-1}^2 z_{t-1}a_t z_T a_{T+1}\right) = \sum_{s=T+2}^{n}\sum_{t=2}^{T} E\left(z_{s-1}^2 z_{t-1}a_t z_T a_{T+1}\right)$$
$$+ \sum_{s=T+2}^{n}\sum_{t=T+2}^{n} E\left(z_{s-1}^2 z_{t-1}a_t z_T a_{T+1}\right). \tag{21}$$

It can be verified that

$$\sum_{s=T+2}^{n}\sum_{t=2}^{T} E\left(z_{s-1}^2 z_{t-1}a_t z_T a_{T+1}\right) = \sum_{s=T+2}^{n}\sum_{t=2}^{T} E\left\{\left(\sum_{i=0}^{\infty}\phi^i a_{s-1-i}\right)\left(\sum_{j=0}^{\infty}\phi^j a_{s-1-j}\right)\right.$$
$$\left.\times\left(\sum_{l=0}^{\infty}\phi^l a_{t-1-l}\right)a_t\left(\sum_{k=0}^{\infty}\phi^k a_{T-k}\right)a_{T+1}\right\} = 2\sum_{s=T+2}^{n}\sum_{t=2}^{T}\phi^{2s-2t-2}\sum_{l=0}^{\infty}\phi^{2l}E\left(a_t^2 a_{T+1}^2 a_{t-1-l}^2\right).$$

Since $E\left(a_t^2 a_{T+1}^2 a_{t-1-l}^2\right) = O(1)$, $\sum_{l=0}^{\infty} \phi^{2l} = O(1)$ and, as $n \to \infty$, $\sum_{s=T+2}^{n} \sum_{t=2}^{T} \phi^{2s-2t-2} = O(1)$, it holds that the first term at the right hand side of (21) is $O(1)$. Similarly, the last term in (21) verifies

$$\sum_{s=T+2}^{n} \sum_{t=T+2}^{n} E\left(z_{s-1}^2 z_{t-1} a_t z_T a_{T+1}\right) = 2 \sum_{t=T+2}^{n} \sum_{s=t+1}^{n} \phi^{2s-2(T+1)-1} \sum_{k=0}^{\infty} \phi^{2k} E\left(a_t^2 a_{T+1}^2 a_{T-k}^2\right).$$

Since

$$\lim_{n\to\infty} \sum_{t=T+2}^{n} \sum_{s=t+1}^{n} \phi^{2s-2(T+1)-1} = \lim_{n\to\infty} \frac{\phi}{\left(1-\phi^2\right)^2} \left[\phi^2 - \phi^{2(n-T-1)} \left\{\left(1-\phi^2\right)(n-T-1) + \phi^2\right\}\right] = 0,$$

it holds that the last term in (21) is null, and hence

$$n^{-1} \sum_{\substack{s=2 \\ s\neq T+1}}^{n} \sum_{\substack{t=2 \\ t\neq T+1}}^{n} E\left(z_{s-1}^2 z_{t-1} a_t z_T a_{T+1}\right) = O\left(n^{-1}\right). \tag{22}$$

Therefore $E\left\{\left(\hat{\gamma}_z - \gamma_z\right) \sum_{t=2}^{n} z_{t-1} a_t z_T a_{T+1}\right\} = O(1),$. and then $E(\hat{\phi}_n z_T a_{T+1}) = E(\hat{z}_{T+1}^{in} a_{T+1}) = O(n^{-1})$

.

b) Now the predictor is $\hat{z}_{T+1}^{\text{filter}} = \hat{\boldsymbol{\phi}}_n^{(IO)} z_T$, where $\hat{\boldsymbol{\phi}}_n^{(IO)}$ is the OLS estimation of ϕ in the model $z_t = \phi z_{t-1} + w h_t + \varepsilon_t$, with $h_t = 1$ if $t = T+1$ and $h_t = 0$ if $t \neq T+1$. Let us denote as $\hat{w}$ to the OLS estimator of w in the previous model. Then, it can be verified that $\hat{w} = z_{T+1} - \hat{\boldsymbol{\phi}}_n^{(IO)} z_T \equiv \hat{a}_{T+1}^{\text{filter}}$ is the one-step ahead prediction error of the predictor $\hat{z}_{T+1}^{\text{filter}}$. Let us denote $h = (h_2, h_3, ..., h_n)'$, $Z_2 = (z_2, z_3, ..., z_n)'$, and $X = [Z_1 \; h]$. Then $\hat{a}_{T+1}^{\text{filter}} = \left(h' M_f h\right)^{-1} h' M_f Z_2 \;=\; a_{T+1} + \; \left(h' M_f h\right)^{-1} h' M_f U$; where $M_f = I_{n-1} - X(X'X)^{-1}X'$. After some algebra, it can be verified that $\hat{a}_{T+1}^{\text{filter}} = a_{T+1} - \; \left(\sum_{t\neq T+1} z_{t-1} a_t / \sum_{t\neq T+1} z_{t-1}^2\right) z_T$. Therefore, $E(\hat{\boldsymbol{\phi}}_n^{(IO)} z_T a_{T+1}) = \sigma^2 - E(\hat{a}_{T+1}^{\text{filter}} a_{T+1})$, with

$$E(\hat{a}_{T+1}^{\text{filter}} a_{T+1}) = \sigma^2 - n^{-1} E\left(\hat{\gamma}_{fz}^{-1} \sum_{t\neq T+1} z_{t-1} a_t z_T a_{T+1}\right),$$

where $\hat{\gamma}_{fz} = n^{-1} \sum_{t\neq T+1} z_{t-1}^2$. Then, applying a Taylor expansion of $\hat{\gamma}_{fz}^{-1}$ around the true value γ_z^{-1} it can be obtained that

$$E\left(\hat{\gamma}_{fz}^{-1} \sum_{t=2}^{n} z_{t-1} a_t z_T a_{T+1}\right) = \gamma_z^{-1} E\left(\sum_{t\neq T+1} z_{t-1} a_t z_T a_{T+1}\right) + O\left[E\left\{\left(\hat{\gamma}_{fz} - \gamma_z\right) \sum_{t\neq T+1} z_{t-1} a_t z_T a_{T+1}\right\}.\right]$$

It can be checked that $E\left(\sum_{t\neq T+1} z_{t-1} a_t z_T a_{T+1}\right) = 0$. In order to see it, the following decomposition can be used:

$$\begin{aligned}\sum_{t\neq T+1} E\left(z_{t-1}a_t z_T a_{T+1}\right) &= \sum_{t=1}^{T} E\left(z_{t-1}a_t z_T a_{T+1}\right) + \sum_{t=T+2}^{n} E\left(z_{t-1}a_t z_T a_{T+1}\right) \\ &= \sum_{t=1}^{T} E\left(\left(\sum_{i=0}^{\infty}\phi^i a_{t-1-i}\right) a_t \left(\sum_{j=0}^{\infty}\phi^j a_{T-i}\right) a_{T+1}\right) \\ &\quad + \sum_{t=T+2}^{n} E\left(\left(\sum_{i=0}^{\infty}\phi^i a_{t-1-i}\right) a_t \left(\sum_{j=0}^{\infty}\phi^j a_{T-i}\right) a_{T+1}\right) \\ = \sum_{t=1}^{T} E\left(\phi^{T-t}\left(\sum_{i=0}^{\infty}\phi^i a_{t-1-i}\right) a_{T+1}a_t^2\right) &+ \sum_{t=T+2}^{n} E\left(\phi^{t-T-2}a_{T+1}^2\left(\sum_{j=0}^{\infty}\phi^j a_{T-i}\right) a_t\right) \\ &= \sum_{t=1}^{T}\phi^{T-t}\sum_{i=0}^{\infty}\phi^i E\left(a_{t-1-i}a_{T+1}a_t^2\right) + \sum_{t=T+2}^{n}\phi^{t-T-2}\sum_{i=0}^{\infty}\phi^i E\left(a_{T-i}a_{T+1}^2 a_t\right),\end{aligned}$$

where it can easily be seen that both $E\left(a_{t-1-i}a_{T+1}a_t^2\right)$ and $E\left(a_{T-i}a_{T+1}^2 a_t\right)$ are null. Then, by (22),

$$E\left(\hat{\gamma}_{fz}^{-1}\sum_{t=2}^{n} z_{t-1}a_t z_T a_{T+1}\right) = n^{-1}O\left[\sum_{s\neq T+1}\sum_{t\neq T+1} E\left\{z_{s-1}^2 z_{t-1}a_t z_T a_{T+1}\right\}\right] = O(n^{-1}).$$

Therefore, $E(\hat{a}_{T+1}^{\text{filter}}a_{T+1}) = \sigma^2 + O(n^{-2})$ and, hence, $E(\hat{\boldsymbol{\phi}}_n^{(IO)} z_T a_{T+1}) = E(\hat{z}_{T+1}^{\text{filter}}a_{T+1}) = O(n^{-2})$ and the proposition holds

References

[1] Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst .Statist. Math.*, **21**, 343-7.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE. Trans. Aut. Contr.* **19**, 203-17.

[3] Ashley, R. (1998). A new technique for postsample model selection and validation. *J. of Econ. Dynam. Control,* **22**, 647–65.

[4] Ashley, R. Granger, C.W.J., & Schmalensee, R. (1980). Advertising and aggregate consumption: an analysis of causality. *Econometrica*, **48**, 1149–67.

[5] Clark, T.E. & McCracken, M.W. (1999). Tests of equal forecast accuracy and encompassing for nested models. Manuscript. Federal Reserve Bank of Kasnsas City.

[6] Chang, I., Tiao, G.C. & Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics,* **30**, 2, 193–204.

[7] Chong Y.Y. & Hendry, D.F. (1986). Econometric evaluation of linear macro-economic models. *Rev. Econ. Studies*, **53**, 671–690.

[8] Choi, B.S. (1992). *ARMA model identification.* New York. Springer-Verlag.

[9] Cressi, N. (1991). *Statistics for spatial data.* New York:Wiley

[10] Diebold, F.X. & Mariano, R.S. (1995). Comparing predictive accuracy. *J. Bus. & Econ. Statis*, **13**, 253–263.

[11] Fuller, W.A. & Hasza, D.P. (1981). Properties of predictors in misspecified autoregressive time series models. *J. Am Statist. Assoc.*, **76**, 155–161.

[12] Hannan E. J. & Quinn, B. J. (1979). The determination of the order of an autoregression" *J. R. Statist. Soc B,* **41**, 190-95.

[13] Haslett, J. & Hayes, K. (1998). "Residuals for the linear model with general covariance Structure" *J. R. Statist. Soc B,* **60**,201-205

[14] HASLETT, J. (1999). A simple derivation of deletion diagnostic results for the general linear model with correlated errors" *J. R. Statist. Soc B,* **61**, 603-9.

[15] KOREISHA, S.G. & PUKKILA, T. (1995). A comparison between different order-determination criteria for identification of ARIMA models. *J. Bus. & Econ. Statist.,* **13**, 127–131.

[16] LO, A.W. & MACKINLAY, A.C. (1990). Data snooping biases in tests of financial assets. *Rev.Financial Studies*, **3**, 431–68.

[17] MANN, H.B. & WALD, A. (1943). On the statistical treatment of linear stochastic difference equations. *Econometrica*, **11**, 173–220.

[18] PEÑA, D. (1987). Measuring the importance of outliers in ARIMA models". In *New Perspectives in Theoretical and Applied Statistics,* 109–18. Ed. Puri et al. John Wiley.

[19] PEÑA, D. (1990). "Influential observations in time series". *J. Bus. & Econ. Statis.,* **8**, 235–41

[20] SCHWARZ, G.(1978). Estimating the dimension of a model. *Ann. Statist.*, **6**,461-64

[21] TIAO, C.G. & TSAY (1994). Some advances in non-linear and adaptive modelling in time-series. *J. Forecasting* , **13**, 109-31

[22] WEST, K.D. (1996). Asymptotic inference about predictive ability. *Econometrica*, **64**, 1067–84.

[23] WEST, K.D. & MCCRACKEN, M.W. (1998). Regression-based tests of predictive ability,' *Int. Econ. Rev.*, **39**, 817–40.

[24] WHITE, H. (2001). A reality check for data snooping. *Econometrica*, forthcoming.

Table 1: Empirical properties of alternative estimators of the MSPE of an AR(1) and an AR(2) at horizon h. True model: $(1-0.9B)(1-\phi B)y_t = a_t$.

	Empirical bias				Empirical MSE			
	B_1		B_2		MSE$_1$		MSE$_2$	
	$h=1$	$h=5$	$h=1$	$h=5$	$h=1$	$h=5$	$h=1$	$h=5$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\phi = 0.3$								
$\hat{V}_p^{\text{in}}(h)$	-0.0281	-0.4623	-0.0495	-0.6016	0.0279	3.4140	0.0224	2.7782
$\hat{V}_p^{\text{in-c}}(h)$	-0.0172	-0.3984	-0.0291	-0.4778	0.0279	3.4275	0.0216	2.7508
$\hat{V}_p^{\text{out-50}}(h)$	-0.0028	0.0244	0.0011	0.0495	0.0570	7.7192	0.0443	6.7899
$\hat{V}_p^{\text{out-75}}(h)$	-0.0044	-0.0496	-0.0034	-0.0303	0.1115	14.8833	0.0868	13.6216
$\hat{V}_p^{\text{del}}(h)$	-0.0195	-0.2412	-0.0335	-0.2249	0.0280	3.5039	0.0219	2.9311
$\hat{V}_p^{\text{filter}}(h)$	-0.0065	-0.0009	-0.0088	-0.0784	0.0285	3.9047	0.0220	3.1077
$\phi = 0.2$								
$\hat{V}_p^{\text{in}}(h)$	-0.0277	-0.3620	-0.0496	-0.4630	0.0236	2.0595	0.0223	1.8169
$\hat{V}_p^{\text{in-c}}(h)$	-0.0172	-0.3108	-0.0292	-0.3614	0.0236	2.0664	0.0216	1.8036
$\hat{V}_p^{\text{out-50}}(h)$	-0.0029	0.0157	0.0010	0.0327	0.0482	4.6335	0.0442	4.4166
$\hat{V}_p^{\text{out-75}}(h)$	-0.0049	-0.0474	-0.0037	-0.0365	0.0942	8.9899	0.0865	8.8339
$\hat{V}_p^{\text{del}}(h)$	-0.0183	-0.1721	-0.0316	-0.1714	0.0237	2.1348	0.0218	1.9333
$\hat{V}_p^{\text{filter}}(h)$	-0.0069	-0.0157	-0.0088	-0.0613	0.0241	2.3479	0.0219	2.0420
$\phi = 0.1$								
$\hat{V}_p^{\text{in}}(h)$	-0.0279	-0.2910	-0.0495	-0.3633	0.0215	1.3039	0.0224	1.2352
$\hat{V}_p^{\text{in-c}}(h)$	-0.0177	-0.2489	-0.0291	-0.2785	0.0215	1.3073	0.0217	1.2292
$\hat{V}_p^{\text{out-50}}(h)$	-0.0036	0.0118	0.0009	0.0253	0.0437	2.9637	0.0443	2.9982
$\hat{V}_p^{\text{out-75}}(h)$	-0.0052	-0.0291	-0.0036	-0.0225	0.0859	5.8621	0.0868	6.0253
$\hat{V}_p^{\text{del}}(h)$	-0.0179	-0.1266	-0.0295	-0.1357	0.0216	1.3655	0.0219	1.3198
$\hat{V}_p^{\text{filter}}(h)$	-0.0078	-0.0264	-0.0088	-0.0483	0.0219	1.4837	0.0220	1.3968
$\phi = 0.0$								
$\hat{V}_p^{\text{in}}(h)$	-0.0290	-0.2408	-0.0496	-0.2910	0.0208	0.8514	0.0224	0.8615
$\hat{V}_p^{\text{in-c}}(h)$	-0.0188	-0.2055	-0.0291	-0.2190	0.0207	0.8525	0.0217	0.8589
$\hat{V}_p^{\text{out-50}}(h)$	-0.0040	0.0085	0.0007	-0.0192	0.0421	2.0054	0.0443	2.0831
$\hat{V}_p^{\text{out-75}}(h)$	-0.0058	-0.0227	-0.0036	-0.0190	0.0830	4.0377	0.0869	4.1875
$\hat{V}_p^{\text{del}}(h)$	-0.0180	-0.0962	-0.0274	-0.1099	0.0209	0.9018	0.0219	0.9216
$\hat{V}_p^{\text{filter}}(h)$	-0.0090	-0.0350	0.0089	-0.0386	0.0211	0.9660	0.0220	0.9778

Note: $B_1 = E(\hat{V}_1 - V_1)$; $B_2 = E(\hat{V}_2 - V_2)$; V_2 and V_1 are estimated with 100,000 replications. MSE$_1$ = $E\left[(\hat{V}_1 - V_1)^2\right]$; MSE$_2 = E\left[(\hat{V}_2 - V_2)^2\right]$; where $E(\cdot)$ are sampling averages with 20,000 replications.

Table 2: Empirical properties of alternative estimators of the MSPE of an AR(1) and an AR(2) at horizon h. True model: $(1-0.9B)(1-\phi B)y_t = a_t$.

MSPE differential: AR(2)-AR(1)

	$E(\hat{d})$		$E(\hat{d}-d)$		$E\left\{(\hat{d}-d)^2\right\}$		$\Pr(\hat{V}_2 < \hat{V}_1)$	
	$h=1$	$h=5$	$h=1$	$h=5$	$h=1$	$h=5$	$h=1$	$h=5$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\phi=0.3$								
$\hat{V}_p^{\text{in}}(h)$	-0.0924	-0.3092	-0.0214	-0.1393	0.0048	0.1905	0.981	0.780
$\hat{V}_p^{\text{in-c}}(h)$	-0.0829	-0.2493	-0.0119	-0.0794	0.0045	0.1744	0.945	0.714
$\hat{V}_p^{\text{out-50}}(h)$	-0.0671	-0.1448	0.0039	0.0251	0.0087	0.6629	0.774	0.612
$\hat{V}_p^{\text{out-75}}(h)$	-0.0700	-0.1505	0.0010	0.0194	0.0173	1.4061	0.704	0.608
$\hat{V}_p^{\text{del}}(h)$	-0.0850	-0.1536	-0.0140	0.0163	0.0047	0.1250	0.948	0.627
$\hat{V}_p^{\text{filter}}(h)$	-0.0734	-0.2474	-0.0024	-0.0775	0.0045	0.1851	0.894	0.704
$d=$	-0.0710	-0.1699						
$\phi=0.2$								
$\hat{V}_p^{\text{in}}(h)$	-0.0464	-0.1422	-0.0219	-0.1010	0.0024	0.0714	0.901	0.701
$\hat{V}_p^{\text{in-c}}(h)$	-0.0365	-0.0919	-0.0120	-0.0507	0.0021	0.0623	0.782	0.570
$\hat{V}_p^{\text{out-50}}(h)$	-0.0206	-0.0242	0.0004	0.0170	0.0040	0.2641	0.583	0.541
$\hat{V}_p^{\text{out-75}}(h)$	-0.0233	-0.0303	0.0012	0.0109	0.0080	0.5456	0.573	0.569
$\hat{V}_p^{\text{del}}(h)$	-0.0378	-0.0405	-0.0133	0.0007	0.0023	0.0437	0.789	0.486
$\hat{V}_p^{\text{filter}}(h)$	-0.0265	-0.0869	-0.0020	-0.0457	0.0021	0.0663	0.666	0.555
$d=$	-0.0245	-0.0412						
$\phi=0.1$								
$\hat{V}_p^{\text{in}}(h)$	-0.0199	-0.0564	-0.0216	-0.0723	0.0013	0.0255	0.732	0.585
$\hat{V}_p^{\text{in-c}}(h)$	-0.0096	-0.0137	-0.0113	-0.0296	0.0009	0.0207	0.527	0.412
$\hat{V}_p^{\text{out-50}}(h)$	0.0061	0.0294	0.0044	0.0135	0.0016	0.0973	0.344	0.449
$\hat{V}_p^{\text{out-75}}(h)$	0.0034	0.0225	0.0017	0.0066	0.0033	0.1928	0.400	0.501
$\hat{V}_p^{\text{del}}(h)$	-0.0099	0.0068	-0.0116	-0.0091	0.0010	0.0155	0.531	0.374
$\hat{V}_p^{\text{filter}}(h)$	0.0011	-0.0108	-0.0006	-0.0267	0.0009	0.0226	0.379	0.400
$d=$	0.0017	0.0159						
$\phi=0.0$								
$\hat{V}_p^{\text{in}}(h)$	-0.0101	-0.0219	-0.0206	-0.0502	0.0008	0.0097	0.613	0.515
$\hat{V}_p^{\text{in-c}}(h)$	0.0002	0.0146	-0.0103	-0.0137	0.0005	0.0075	0.366	0.326
$\hat{V}_p^{\text{out-50}}(h)$	0.0150	0.0393	0.0045	0.0110	0.0009	0.0407	0.214	0.381
$\hat{V}_p^{\text{out-75}}(h)$	0.0116	0.0289	0.0011	0.0006	0.0017	0.0762	0.296	0.436
$\hat{V}_p^{\text{del}}(h)$	0.0011	0.0121	-0.0094	-0.0162	0.0005	0.0075	0.356	0.348
$\hat{V}_p^{\text{filter}}(h)$	0.0106	0.0246	0.0001	-0.0037	0.0005	0.0091	0.226	0.308
$d=$	0.0105	0.0283						

Note: $d = V_2 - V_1$, estimated with 100,000 replications; $\hat{d} = \hat{V}_2 - \hat{V}_1$. $E(\hat{d})$, $E(\hat{d}-d)$, and $E\left[(\hat{d}-d)^2\right]$ are sampling averages of 20,000 replications.

Table 3: Proportion of times that each model has lower empirical MSPE using different estimators of MSPE. Sample size $n = 100$. True model: $(1 - 0.9B)(1 - \phi B)y_t = a_t$.

	$\phi = 0.3$						$\phi = 0.1$					
	AR(1)		AR(2)		AR(3)		AR(1)		AR(2)		AR(3)	
	$h = 1$	$h = 5$	$h = 1$	$h = 5$	$h = 1$	$h = 5$	$h = 1$	$h = 5$	$h = 1$	$h = 5$	$h = 1$	$h = 5$
Panel A												
MSPE(h)	1.092	6.321	1.025	6.166	1.035	6.218	1.021	4.163	1.025	4.181	1.035	4.213
$\hat{V}_p^{\text{in}}(h)$	0.010	0.144	0.386	0.390	0.604	0.466	0.146	0.321	0.282	0.250	0.572	0.429
$\hat{V}_p^{\text{in-c}}(h)$	0.040	0.255	0.603	0.471	0.357	0.274	0.375	0.523	0.333	0.233	0.292	0.244
$\hat{V}_p^{\text{out-50}}(h)$	0.216	0.359	0.600	0.414	0.184	0.227	0.608	0.480	0.260	0.289	0.132	0.231
$\hat{V}_p^{\text{out-75}}(h)$	0.277	0.360	0.485	0.374	0.238	0.266	0.522	0.419	0.277	0.296	0.201	0.285
$\hat{V}_p^{\text{del}}(h)$	0.036	0.331	0.560	0.395	0.404	0.274	0.372	0.550	0.325	0.191	0.303	0.259
$\hat{V}_p^{\text{filter}}(h)$	0.093	0.263	0.698	0.505	0.209	0.232	0.560	0.542	0.286	0.246	0.154	0.212
Panel B												
AIC	0.091		0.695		0.214		0.559		0.289		0.152	
FPE	0.086		0.691		0.223		0.556		0.288		0.156	
HQ	0.265		0.681		0.054		0.841		0.138		0.021	
BIC	0.266		0.681		0.053		0.842		0.137		0.021	

Table 4: Proportion of times that each model has lower empirical MSPE using different estimators of MSPE at horizon $h = 1$ and sample size $n = 100$. True model: $(1 - 0.9B)(1 - 0.5B)(1 - \phi B)y_t = a_t$.

$h = 1$	$\phi = 0.5$					$\phi = 0.1$				
	AR(1)	AR(2)	AR(3)	AR(4)	AR(5)	AR(1)	AR(2)	AR(3)	AR(4)	AR(5)
Panel A										
MSPE(1)	2.525	1.073	1.033	1.044	1.055	1.393	1.023	1.032	1.043	1.054
$\hat{V}_p^{\text{in}}(1)$	0.000	0.017	0.191	0.234	0.558	0.000	0.115	0.127	0.223	0.535
$\hat{V}_p^{\text{in-c}}(1)$	0.000	0.089	0.429	0.217	0.264	0.000	0.396	0.198	0.177	0.229
$\hat{V}_p^{\text{out-50}}(1)$	0.000	0.282	0.490	0.128	0.100	0.024	0.643	0.170	0.089	0.074
$\hat{V}_p^{\text{out-75}}(1)$	0.008	0.295	0.392	0.154	0.151	0.080	0.483	0.186	0.121	0.130
$\hat{V}_p^{\text{del}}(1)$	0.000	0.042	0.278	0.234	0.446	0.000	0.265	0.181	0.206	0.348
$\hat{V}_p^{\text{filter}}(1)$	0.000	0.195	0.529	0.155	0.121	0.000	0.636	0.178	0.098	0.088
Panel B										
AIC	0.000	0.192	0.526	0.162	0.120	0.000	0.629	0.179	0.099	0.093
FPE	0.000	0.181	0.516	0.166	0.137	0.000	0.611	0.181	0.103	0.105
HQ	0.000	0.461	0.480	0.046	0.013	0.001	0.903	0.071	0.019	0.006
BIC	0.000	0.461	0.480	0.046	0.013	0.001	0.903	0.071	0.019	0.006