KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Economics and Applied Economics

Prediction focussed model selection for autoregressive models

Gerda Claeskens, Christophe Croux and Johan Van Kerckhoven

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 0611

Prediction Focussed Model Selection for Autoregressive Models

Gerda Claeskens [1,2], Christophe Croux [1], and Johan Van Kerckhoven [1]
Katholieke Universiteit Leuven

Summary

In order to make predictions of future values of a time series, one needs to specify a forecasting model. A popular choice is an autoregressive time series model, where the order of the model is chosen by an information criterion. We propose an extension of the Focussed Information Criterion (FIC) for model-order selection with focus on a high predictive accuracy (i.e. the mean squared forecast error is low). We obtain theoretical results and illustrate in a simulation study that this FIC can outperform classical order selection criteria in the setting with one series to predict and a different series for parameter estimation. We also demonstrate, via a simulation study and some real data examples, that in the practical setting of only one available time series, the performance of the FIC is comparable to the performance of other information criteria.

Key words: autoregressive order selection, focussed information criterion, prediction, time series

1 Introduction

In many fields of applied research (e.g. economics, demographics, ...), a variable is observed over time, and the researcher wishes to model the time-structure of the data and predict future values of the variable. This modelling consists of two important parts: first, the general trend over time is modelled and seasonal effects are identified, and then the dynamic structure of the resulting stationary series is investigated. In this paper we are concerned

[1]ORSTAT and University Center for Statistics, K.U. Leuven, Naamsestraat 69, B-3000 Leuven, Belgium
[2]Author to whom correspondence should be addressed. e-mail: gerda.claeskens@econ.kuleuven.be
Phone: +32-16-32.6993 Fax: +32-16-32.6732

with the latter. A popular choice is the autoregressive model,

$$Z_t = \phi_1(p)Z_{t-1} + \cdots + \phi_p(p)Z_{t-p} + \varepsilon_t(p), \tag{1}$$

which predicts the stationary variables Z_t by its lagged variables. Model (1) is an autoregressive model of order p, abbreviated as AR(p)-model. The variables Z_t have been centered by their average, and the $\varepsilon_t(p)$ are zero mean, white noise innovation terms. Modelling the time series can serve many purposes, but usually the goal is to make accurate predictions of the series in the unobserved future.

We focus on making forecasts of the series h steps beyond the last observation. Generally, the accuracy of these forecasts depends on the autoregressive order p of the model used, in other words on how far in the past we look in order to model the series. If we restrict ourselves to only the recent past, p small, then we might fail to capture more long-term influences. Conversely, if we include the far past, p large, then the accuracy of the predictions will suffer because of the chosen model's complexity. Hence, a balance between completeness and simplicity must be chosen and a commonly used method of selecting an appropriate AR-order is by computing the value of an information criterion for each candidate model, and selecting the model with the best value of the criterion.

In this paper, we propose an adapted version of the Focussed Information Criterion (abbreviated FIC) as defined in Claeskens & Hjort (2003). The main novel aspects are the application to time series and that we allow the maximal order of the autoregressive model to increase slowly to infinity as the length of the series increases. We also provide a bound on the rate of this increase by adapting a theorem in Portnoy (1985) to the time series setting. This result is needed because, originally, the theory behind FIC was developed for the case where the maximal number of variables in the model, or in this case the maximal considered autoregressive order, remains constant. We develop these ideas in the setting of two independent realisations of the data generating process, hereby following Shibata (1980), Bhansali (1996), and Lee & Karagrigoriou (2001). This framework is described in Section 2, where we also provide an extension to the more realistic case of only one realisation of the data generating process. Section 3 contains the derivation of the Focussed Information Criterion.

In Section 4 we report the results of a simulation study. We compare the efficiency in

mean squared error sense of the models selected by FIC, and two well-known criteria: AIC (Akaike, 1974) and BIC (Schwarz, 1978), also sometimes called SIC. First, the two-series setting is discussed, where AIC has been proven to be an asymptotically efficient criterion (Shibata, 1980, Bhansali, 1996, Lee & Karagrigoriou, 2001). We expect FIC to perform very well in this setting since, unlike AIC and BIC, it uses information from both available time series. Moreover, FIC is constructed to minimise the estimated MSE of the predictions. Additionally, we also performed a simulation study in the one-series setting. Recently, AIC was proven to be also asymptotically efficient in this situation (Ing & Wei, 2005).

To illustrate the practical use of FIC, we compare in Section 5 the performance of the aforementioned criteria on two real data examples. In Section 6, we provide some extensions to the ideas presented in this paper such as the application of the FIC to simultaneously select a subset of regression variables and the autoregressive order of the error terms, as in Shi & Tsai (2004). Finally, we summarise and make some concluding remarks in Section 7.

2 Model Setting

In this section we state the model setting, and define h-step ahead predictions of a time series (as in Shibata, 1980, and Bhansali, 1996). The true time series is a realisation of an AR(∞)-process, and we approximate this by a finite order autoregressive model. We first assume that we have a univariate time series $\{y_t\}$ available, where $t = 1, \ldots, T$, and that we want to make a prediction of this series at time-horizon h. We denote this prediction $\hat{y}_{T+h}$. We also assume that we have a second series $\{x_t\}$ available of the same length T. These two series are assumed to be independent realisations of the same length T of a stochastic process $\{Z_t\}$, with the following dependency structure,

$$Z_t = \varepsilon_t + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots . \tag{2}$$

We assume that the innovation terms ε_t are independent and identically normally distributed, with mean 0 and variance σ^2. We also assume that the autoregression coefficients ϕ_i are absolutely summable (that is $\sum_i |\phi_i| < \infty$), and that the associated power series

$$\Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots$$

converges and is different from zero for $|z| \leq 1$. Our goal is to select the best approximating autoregressive model of order p, with $0 \leq p \leq p_T$, using the series $\{x_t\}$. Here we allow the maximal considered AR-order, denoted by p_T, to depend on T. This is done because one typically fits a time series model of a higher order if the length of the series increases. Next, we use this selected model to make a h-step ahead forecast for the series $\{y_t\}$. In practice, however, the user often has only one time series $\{x_t\}$ available. In such case, we make a h-step ahead prediction of the series $\{x_t\}$ itself. Our results are valid for both situations: one series and two series. For notational simplicity, we continue to work in the two-series setting. The results for the single-series setting are obtained by setting $\{y_t\}$ equal to $\{x_t\}$.

There are two methods to make the h-step ahead forecast $\hat{y}_{T+h}$. The first method is the direct method, which assumes that we estimate different models

$$Z_t = \phi_1(p,h)Z_{t-h} + \cdots + \phi_p(p,h)Z_{t+1-h-p} + \varepsilon_t(p,h) \tag{3}$$

for each horizon h. The $\varepsilon_t(p,h)$ are assumed to have zero mean and variance $\sigma^2(p,h)$. We forecast the series $\{y_t\}$ at horizon h by $\hat{y}_{T+h} = \hat{\phi}_1(p,h)y_T + \cdots + \hat{\phi}_p(p,h)y_{T+1-p}$. Here, the parameters $\phi_i(p,h)$ are estimated using ordinary least squares. This would make little difference as opposed to using the maximum likelihood estimator, especially for large T, while the ML-estimator would complicate the computations. The second method is the plug-in method. This is the more common approach and follows immediately from the estimates of model (2). Here, we compute recursively

$$\hat{y}_{T+h}(p) = \hat{\phi}_1(p)\hat{y}_{T+h-1}(p) + \cdots + \hat{\phi}_p(p)\hat{y}_{T+h-p}(p) \tag{4}$$

with $\hat{y}_t(p) = y_t$ for $t \leq T$. Once again, the parameter estimates $\hat{\phi}_i(p)$ are obtained using OLS. Observe that both methods are identical for $h = 1$. In the main part of this paper we make predictions using the direct method, however, see Section 6.1 for the plug-in method. The main advantage of using the direct method and not the plug-in method is shown in Bhansali (1996). He showed that the lower bound on the MSE of predictions obtained via the direct method method is lower than for the plug-in method. Also, he showed that for the direct method this lower bound can be achieved, which is not the case for the plug-in method.

Even in the situation of two independent series $\{x_t\}$ and $\{y_t\}$, the classical information criteria depend only on the series $\{x_t\}$. As a consequence, the extra information contained

in the series $\{y_t\}$ is ignored during the model selection step. In contrast, the Focussed Information Criterion (FIC), as originally proposed by Claeskens & Hjort (2003) and further explored and adapted in the rest of the paper, takes this extra information into account.

3 The Focussed Information Criterion

In this section we propose an extended version of the Focussed Information Criterion (abbreviated FIC) as defined in Claeskens & Hjort (2003). The idea of the FIC is that an information criterion should take into account the purpose of the statistical analysis, by trying to estimate the MSE of the estimate of a focus parameter. In our setting, this focus parameter is the h-step ahead prediction of the time series. In this extension, we allow the number of variables to increase towards infinity with the sample size. In time series analysis we select an AR(p)-model which fits the available data best, with $0 \leq p \leq p_T$. Recall that p_T is the maximal autoregressive order, depending on the length of the series. We allow the number of variables to increase to infinity by letting the maximal autogressive order increase as the length of the time series increases. Using an adaptation of a theorem in Portnoy (1985), we obtain an upper bound for this rate of increase such that the FIC theory still holds. Aim is to predict the series $\{y_t\}$, based on an AR-model estimated from $\{x_t\}$.

At this point, we introduce some notation for the "direct" model (3). First, denote the vectors $x_t(p,h) = (x_{t-h}, \ldots, x_{t+1-h-p})'$, $\phi(p,h) = \big(\phi_1(p,h), \ldots, \phi_p(p,h)\big)'$, and $y(p) = (y_T, \ldots, y_{T+1-p})'$. The OLS-estimates based on the series $\{x_t\}$ of the parameters $\phi(p,h)$ are $\hat{\phi}(p,h)$. Consequently, the h-step ahead prediction of the series $\{y_t\}$ is $\hat{y}_{T+h} = \hat{\phi}(p,h)'y(p)$ if $1 \leq p \leq p_T$, and $\hat{y}_{T+h} = 0$ for $p = 0$. Because our goal is to make this prediction as accurate as possible, we take as focus parameter $\mu(p,h) = \phi(p,h)'y(p)$.

Our goal is now to construct an information criterion aimed at selecting the model yielding the "best" estimate for the focus parameter from the $p_T + 1$ possible AR(p)-models. "Best" is defined in the sense of having the lowest mean squared forecast error. If we select the order p too low, the h-step ahead prediction of the series $\{y_t\}$ will be biased. On the other hand, choosing p too high, will inflate the variance of the prediction. Therefore, we need to select p such that the h-step ahead prediction has at the same time a small bias and a small variance.

To define the Focussed Information Criterion, we will assume that we work in the same setting as in Claeskens & Hjort (2003). In particular, the results for the FIC apply in a local misspecification setting where the true, or optimal, values of the focus parameters are $\mu_{\text{true}} = \delta(p_T)'y(p_T)T^{-1/2}$. Here, δ is a fixed (though unknown) vector of infinite length, of which for practical purposes the first p_T components are used, which are denoted by $\delta(p_T)$. A similar setup is assumed for le Cam's contiguity results, the local asymptotic normality, and in calculations under local alternatives for hypothesis testing problems. Let $J_{T,\text{full}}$ be the estimated $p_T \times p_T$ information matrix of the AR$\big(p_T\big)$-model, the largest model under consideration, and assume that this matrix is of full rank. Since we will use straightforward OLS-estimation for the parameters, this matrix can be estimated by

$$\hat{J}_{T,\text{full}} = \frac{\hat{R}(p_T, h)}{\hat{\sigma}^2(p_T, h)}.$$

Here

$$\hat{R}\big(p_T, h\big) = \frac{1}{T + 1 - h - p_T} \sum_{t=p_T+h}^{T} x_t(p_T, h) x_t(p_T, h)' \tag{5}$$

is the estimated autocovariance matrix of order p_T of the series $\{x_t\}$, and $\hat{\sigma}^2(p_T, h)$ is the estimated variance of the residuals after OLS-estimation. On the other hand, $R(p_T)$ is the true autocovariance matrix of order p_T, and $\sigma^2(p_T, h)$ the true variance of the error terms. Using the ML-estimator would increase the complexity of the information matrix in the finite sample setting, while OLS and ML lead to the same limit expression for $J_{T,\text{full}}$. We define the matrices $\hat{K}_{T,p} = \hat{\sigma}^2(p_T, h)\hat{R}(p, h)^{-1}$, and

$$\hat{M}_{T,p} = \hat{\sigma}^2(p_T, h) \begin{pmatrix} \hat{R}(p, h)^{-1} & 0 \\ 0 & 0 \end{pmatrix} \text{ of dimension } p_T \times p_T.$$

Finally, define

$$D_T = \hat{\delta}(p_T, h) = \sqrt{T}\hat{\phi}(p_T, h).$$

The following proposition states the limit distribution of the estimated focus parameter. This result is the cornerstone of the Focussed Information Criterion when applied in this setting. The proof is found in the Appendix. Using similar notation as in Claeskens & Hjort (2003), we set

$$\delta(p_T, h) = \sqrt{T}\phi(p_T, h)$$

and

$$\delta(p,h) = \begin{pmatrix} R(p,h)^{-1} & 0 \\ 0 & 0 \end{pmatrix} R(p_T,h)\delta(p_T,h).$$

Proposition 1 *Take h fixed and let $\hat{\mu}(p,h) = \hat{\phi}(p,h)'y(p)$ be the h-step ahead forecast of the true value μ_{true}. Under conditions (A1), (A2), (A3) listed in the appendix and if*

$$\frac{p_T\sqrt{\log T}}{T} \to 0 \text{ as } T \to \infty,$$

then we have for every $0 \leq p < \infty$

$$\sqrt{T}(\hat{\mu}(p,h) - \mu_{\text{true}}) \xrightarrow{d} \Lambda_p, \text{for } T \to \infty, \tag{6}$$

where Λ_p is normally distributed with mean and variance given by

$$\lambda_p = \mathbb{E}[\Lambda_p] = \lim_{T\to\infty} y(p_T)'\big(\delta(p,h) - \delta(p_T,h)\big) \tag{7}$$

$$\sigma_p^2 = \text{Var}(\Lambda_p) = y(p)'R(p,h)^{-1}y(p) \lim_{T\to\infty} \sigma^2(p_T,h). \tag{8}$$

This proposition does not assume that the time series $\{x_t\}$ and $\{y_t\}$ are independent. In fact, the results remain valid for $y_t = x_t$, stating the proposition for the single-series setting, but conditional on the observed data.

Hjort & Claeskens (2003) prove (although not specific for time series) that the proposition holds for a finite maximal AR-order p_T. The additional condition on the rate of increase of p_T is a result of an adaptation of Theorem 3.2 in Portnoy (1985), which is formulated as Lemma 1 in the appendix, where also the proof of Proposition 1 may be found. The distribution of Λ_p in (6) is normal, with non-zero mean due to the local misspecification setting in which we work.

The distribution of Λ_p is the key result upon which the FIC is constructed. Specifically, the limiting distribution has mean squared error

$$\begin{aligned} r(p) = &\lim_{T\to\infty} y(p_T)'\big(\delta(p,h) - \delta(p_T,h)\big)\big(\delta(p,h) - \delta(p_T,h)\big)'y(p_T) \\ &+ y(p)'R(p,h)^{-1}y(p) \lim_{T\to\infty} \sigma^2(p_T,h). \end{aligned}$$

The FIC estimates this risk quantity for each AR-order p under consideration. To estimate $r(p)$, we estimate the unknown $R(p_T,h)$ and $\sigma^2(p_T,h)$ by $\hat{R}(p_T,h)$, see (5), and

$\hat{\sigma}^2(p_T,h)$. We also unbiasedly estimate the quantity $\delta(p_T,h)\delta(p_T,h)'$ by $\hat{\delta}(p_T,h)\hat{\delta}(p_T,h)' - \hat{\sigma}^2(p_T,h)\hat{R}(p_T,h)^{-1}$, where we used that $\mathrm{Cov}\big(\hat{\delta}(p_T,h)\big) = \sigma^2(p_T,h)R(p_T,h)^{-1}$. Finally, we drop the limit of T tending to infinity. After some algebraic manipulation we get,

$$\begin{aligned}\hat{r}(p) &= \Big\{y(p_T)'\big(\hat{\delta}(p,h) - \hat{\delta}(p_T,h)\big)\Big\}^2 + 2\hat{\sigma}^2(p_T,h)y(p)'\hat{R}(p,h)^{-1}y(p) \\ &\quad - \hat{\sigma}^2(p_T,h)y(p_T)'\hat{R}(p_T,h)^{-1}y(p_T).\end{aligned}$$

If we add $\hat{\sigma}^2(p_T,h)y(p_T)'\hat{R}(p_T,h)^{-1}y(p_T)$, which is independent of p, we arrive at the more compact expression for the FIC:

$$\mathrm{FIC}_p = \Big\{y(p_T)'\big(\hat{\delta}(p,h) - \hat{\delta}(p_T,h)\big)\Big\}^2 + 2\hat{\sigma}^2(p_T,h)y(p)'\hat{R}(p,h)^{-1}y(p). \tag{9}$$

We then select the AR-order p with the smallest value for the FIC_p.

4 Simulations

We present the results of a simulation study to examine the performance of FIC compared to AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion), both in the two-series setting and in the one-series setting. We provide the precise details of the simulation settings, and give an interpretation of the results.

Recall that we estimate the parameters and select the AR-order using one series $\{x_t\}$. We also assume that the actual prediction is done on a different series $\{y_t\}$, independent of $\{x_t\}$, though with the same stochastic structure. This is a similar setup as in Shibata (1980), Bhansali (1996) and Lee & Karagrigoriou (2001). In practical applications however, such a situation does not occur. Usually there is only a single time series available, and model selection, as well as parameter estimation and prediction have to be done using this single time series.

We performed simulation experiments to compare the performance of FIC with the more classical AIC (Akaike, 1974) and BIC (Schwarz, 1978). In all the studies, the true data-generating process is an ARMA(1,1)-model

$$Z_t = \phi Z_{t-1} + \varepsilon_t + \eta\varepsilon_{t-1},$$

where $\varepsilon_t \sim \mathcal{N}(0,1)$ i.i.d., and both ϕ and η take values in $\{-0.9, -0.7, \ldots, 0.9\}$. The stationarity and invertibility conditions on the parameters in this model reduce to $|\phi| < 1$ and $|\eta| < 1$. Hence, the ARMA(1,1)-model has an AR(∞)-representation. We let both parameters vary to examine whether or not the relative performance of the different information criteria depends on the values of these parameters. Note that although the true data-generating process is an ARMA(1,1)-model, this model is not included in the group of considered models, i.e. autoregressive models of finite order. Hence, the selected model will always be the "best" approximating model among the candidate autoregressive models.

In the first simulation experiment we generate for each setting a series $\{x_t\}$ of length $T = 200$, which we use for both model order selection and parameter estimation. Then, for each of the $M = 1000$ simulation runs, we generate an independent series $\{y_t\}$ for which we make a prediction. This series $\{y_t\}$ is generated up to length $T+h$ to allow an out-of-sample estimate of the prediction accuracy of the h-step ahead forecast of $\{y_t\}$. Based on the series $\{x_t\}$, we select the model as in (3) for $0 \leq p \leq p_T$, and $h = 2$, which yields the "best" finite-order AR approximation of the series $\{x_t\}$. We have chosen the maximal order $p_T = 20$ here. For each simulation run, the selection is done by AIC, BIC and FIC. Once the model is selected, a h-step ahead forecast is made of this series $\{y_t\}$ using the parameter-estimates from the selected model of $\{x_t\}$. This forecast is denoted by $\hat{y}_{T+h}^{(j)} = y(p_{h,j})'\hat{\phi}(p_{h,j}, h)$, where j is the number of the simulation run, and $p_{h,j}$ is the AR-order of the model selected for the h-step ahead forecast in simulation run j. Note that for AIC and BIC the order of the selected model is always independent of the series $\{y_t\}$ whereas for FIC the order of the selected model will also depend on this series.

The second simulation experiment is similar to the first, except that the h-step ahead forecast of the series $\{y_t\}$ is executed with the plug-in method. Hence, we select and estimate the model (1):

$$Z_t = Z_t(p)'\phi(p) + \varepsilon_t(p),$$

which fits the series $\{x_t\}$, and use the estimated parameters to predict the series $\{y_t\}$ using the recursion formula,

$$\hat{y}_{T+h} = \hat{\phi}_1 \hat{y}_{T+h-1} + \cdots + \hat{\phi}_p \hat{y}_{T+h-p},$$

where $\hat{y}_t = y_t$ for $t \leq T$ and where $\hat{y}_{T+i}$ is derived using the same relation as above, see also (4).

For each simulation setting in the experiments above, we present the mean squared error of the h-step ahead prediction of the series $\{y_t\}$, where the prediction is performed using the models selected by (i) AIC, (ii) BIC, and (iii) FIC. We define the Mean Squared Error by

$$MSE(\hat{y}_{T+h}) = \frac{1}{M}\sum_{j=1}^{M}(\hat{y}_{T+h}^{(j)} - y_{T+h})^2,$$

with $\hat{y}_{T+h}^{(j)}$ as defined above, and with y_{T+h} the true generated value of the series $\{y_t\}$. We also define the Relative Mean Squared Error as

$$rMSE(\hat{y}_{T+h}, \text{xIC}_1, \text{xIC}_2) = \frac{MSE(\hat{y}_{T+h,\text{xIC}_1})}{MSE(\hat{y}_{T+h,\text{xIC}_2})}, \tag{10}$$

where $\hat{y}_{T+h,\text{xIC}_1}$ and $\hat{y}_{T+h,\text{xIC}_2}$ are the h-step ahead predictions of the series $\{y_t\}$ made with models chosen by respectively xIC_1 and xIC_2 as information criteria. If the relative MSE is smaller than 1, it means that xIC_1 selects models with a lower MSE for the h-step ahead prediction than xIC_2

Tables 1 and 2 present the simulated relative MSEs of the models selected by FIC with respect to AIC (relative $MSE(\hat{y}_{T+h}, \text{FIC}, \text{AIC})$, top tables), and those with respect to BIC (relative $MSE(\hat{y}_{T+h}, \text{FIC}, \text{BIC})$, bottom tables). A first examination of these tables shows that they have a very similar structure. More precisely, we see that in the region where $|\phi+\eta|$ is smaller than 1, FIC and AIC select models with about the same MSE for the 2-step ahead predictor, since the values are around 1. When $|\phi + \eta|$ is large (top left and bottom right corner of each table), we see that FIC selects models with a lower MSE for prediction of the 2-step ahead estimator than both AIC and BIC. We also observe that the larger $|\phi + \eta|$ is, the better the FIC performs as compared to AIC and BIC. Intuitively speaking, $|\phi + \eta|$ large means that more coefficients of the AR(∞)-representation of the ARMA(1,1)-model are significantly different from zero. The results of Table 1 (upper sub-table) are graphically represented in Figure 1. We clearly observe a large area in the (ϕ, η) parameter space where FIC performs much better than AIC (values below 1). In the center part of the graph, both methods perform approximately equally well (values around 1).

We can explain why FIC selects models with a lower or equal MSE than AIC with the following intuitive argument. Assume that we keep the two time series $\{x_t\}$ and $\{y_t\}$ fixed. By construction, FIC is, up to a constant, an unbiased estimator of the asymptotic MSE for

ϕ/η	-0.9	-0.7	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.7	0.9
-0.9	0.239	0.325	0.270	0.316	0.335	0.384	0.529	0.654	0.881	1.072
-0.7	0.582	0.721	0.735	0.709	0.770	0.978	1.032	1.022	1.051	1.104
-0.5	0.881	0.875	0.892	0.984	0.935	1.001	1.018	1.048	1.011	1.019
-0.3	0.998	1.069	1.022	1.093	1.001	1.136	1.044	1.029	1.020	0.979
-0.1	1.091	1.087	1.076	1.089	1.036	1.099	1.033	1.050	1.059	1.024
0.1	1.020	1.101	1.046	1.069	1.032	1.071	1.067	1.075	1.053	1.106
0.3	0.992	1.131	1.036	1.047	1.027	1.011	1.001	0.997	0.982	1.100
0.5	1.203	1.025	1.032	1.042	1.015	0.995	0.980	0.893	0.874	0.824
0.7	1.016	1.090	1.010	0.911	0.852	0.782	0.768	0.639	0.681	0.606
0.9	1.079	0.945	0.707	0.514	0.400	0.383	0.371	0.296	0.242	0.247

ϕ/η	-0.9	-0.7	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.7	0.9
-0.9	0.239	0.325	0.270	0.316	0.335	0.384	0.529	0.654	0.881	1.072
-0.7	0.582	0.721	0.735	0.709	0.770	0.978	1.032	1.022	1.051	1.104
-0.5	0.881	0.875	0.892	0.984	0.935	1.001	1.018	1.048	1.011	1.019
-0.3	0.998	1.069	1.022	1.093	1.001	1.136	1.044	1.029	1.020	0.979
-0.1	1.091	1.087	1.076	1.089	1.036	1.099	1.033	1.050	1.059	1.024
0.1	1.020	1.101	1.046	1.069	1.032	1.071	1.067	1.075	1.053	1.106
0.3	0.992	1.131	1.036	1.047	1.027	1.011	1.001	0.997	0.982	1.100
0.5	1.203	1.025	1.032	1.042	1.015	0.995	0.980	0.893	0.874	0.824
0.7	1.016	1.090	1.010	0.911	0.852	0.782	0.768	0.639	0.681	0.606
0.9	1.079	0.945	0.707	0.514	0.400	0.383	0.371	0.296	0.242	0.247

Table 1: Ratios of mean squared errors for the 2-step ahead prediction of the series $\{y_t\}$, with model order selection using series $\{x_t\}$, and prediction according to the *direct method*. An ARMA(1,1)-process generated both series $\{x_t\}$ and $\{y_t\}$. The autoregression parameter ϕ can be found in the leftmost column, and the moving average parameter η is indicated in the top row. The upper table shows the $rMSE(\cdot, \mathrm{FIC}, \mathrm{AIC})$, the lower table shows the $rMSE(\cdot, \mathrm{FIC}, \mathrm{BIC})$, as defined in (10). If this ratio is smaller than 1, the FIC selects a model with lower MSE for prediction.

ϕ/η	-0.9	-0.7	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.7	0.9
-0.9	0.249	0.275	0.283	0.297	0.338	0.415	0.467	0.710	0.947	1.102
-0.7	0.649	0.655	0.673	0.749	0.799	0.841	0.907	0.974	1.005	1.076
-0.5	0.875	0.886	0.863	0.892	0.914	1.074	1.020	1.010	1.072	1.048
-0.3	1.036	1.002	0.982	1.068	1.002	1.013	1.020	1.051	1.063	1.042
-0.1	1.090	1.061	1.023	1.007	1.032	1.035	1.036	1.024	1.134	1.054
0.1	1.097	1.035	1.029	1.005	1.022	1.014	1.081	0.999	1.076	1.068
0.3	1.046	1.000	1.059	1.104	1.122	1.014	1.000	1.005	1.072	1.004
0.5	0.988	1.035	1.048	1.059	1.038	0.941	0.937	0.939	0.914	0.968
0.7	0.997	1.110	0.991	0.888	0.818	0.792	0.702	0.724	0.703	0.626
0.9	1.041	0.863	0.647	0.491	0.390	0.336	0.327	0.246	0.270	0.242

ϕ/η	-0.9	-0.7	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.7	0.9
-0.9	0.249	0.275	0.283	0.297	0.338	0.415	0.467	0.710	0.947	1.102
-0.7	0.649	0.655	0.673	0.749	0.799	0.841	0.907	0.974	1.005	1.076
-0.5	0.875	0.886	0.863	0.892	0.914	1.074	1.020	1.010	1.072	1.048
-0.3	1.036	1.002	0.982	1.068	1.002	1.013	1.020	1.051	1.063	1.042
-0.1	1.090	1.061	1.023	1.007	1.032	1.035	1.036	1.024	1.134	1.054
0.1	1.097	1.035	1.029	1.005	1.022	1.014	1.081	0.999	1.076	1.068
0.3	1.046	1.000	1.059	1.104	1.122	1.014	1.000	1.005	1.072	1.004
0.5	0.988	1.035	1.048	1.059	1.038	0.941	0.937	0.939	0.914	0.968
0.7	0.997	1.110	0.991	0.888	0.818	0.792	0.702	0.724	0.703	0.626
0.9	1.041	0.863	0.647	0.491	0.390	0.336	0.327	0.246	0.270	0.242

Table 2: Ratios of mean squared errors for the 2-step ahead prediction of the series $\{y_t\}$, with model order selection using series $\{x_t\}$, and prediction according to the *plug-in method*. An ARMA(1,1)-process generated both series $\{x_t\}$ and $\{y_t\}$. The autoregression parameter ϕ can be found in the leftmost column, and the moving average parameter η is indicated in the top row. The upper table shows the $rMSE(\cdot, \mathrm{FIC}, \mathrm{AIC})$, the lower table shows the $rMSE(\cdot, \mathrm{FIC}, \mathrm{BIC})$, as defined in (10). If this ratio is smaller than 1, the FIC selects a model with lower MSE for prediction.

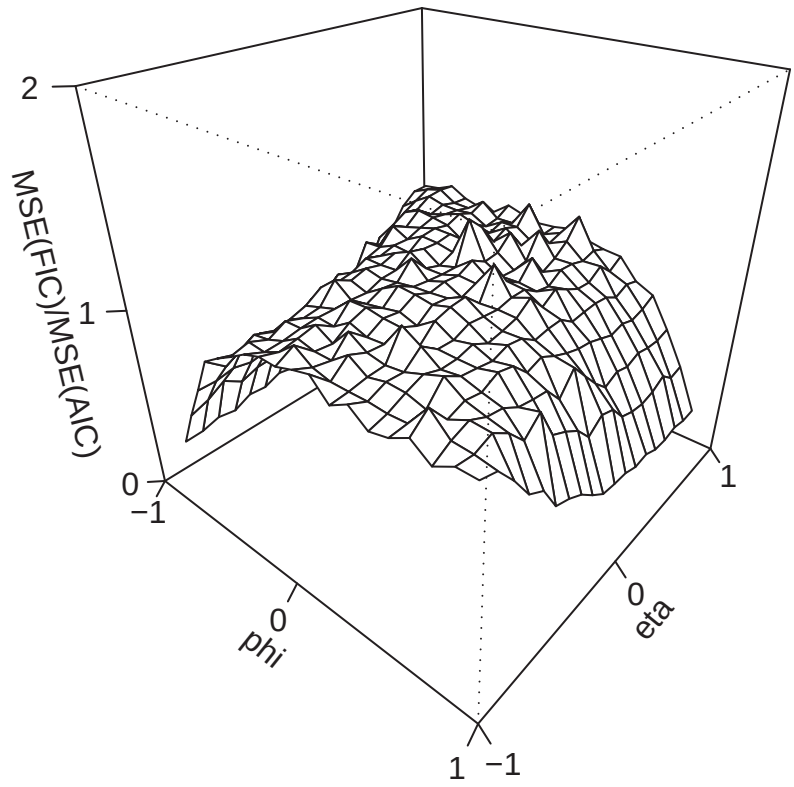

Figure 1: 3D-surface plot for the ratios of mean squared errors for the 2-step ahead prediction of the series $\{y_t\}$, with model order selection using series $\{x_t\}$, and where prediction is done using the direct method. An ARMA(1,1)-process generated both series $\{x_t\}$ and $\{y_t\}$. The autoregression parameter ϕ can be found on the phi axis, and the moving average parameter η is indicated on the eta axis. The surface shows the $rMSE(\cdot, \text{FIC}, \text{AIC})$ as defined in (10). If this ratio is smaller than 1, the FIC selects a model with the lower MSE for prediction.

the h-step ahead prediction. The model with the smallest value of FIC is selected. Or in other words, we select the model with the smallest estimated mean squared forecast error.

In most applications however, we do not have 2 different time series available with the same stochastic structure, and we have to construct our forecasts using the same series as that we use to estimate the AR-parameters. A recent paper by Ing & Wei (2005), demonstrates that AIC is asymptotically efficient in this one-series setting. We investigate this more realistic situation of only one time series in the third and fourth simulation study. Therefore, in the third simulation study we generate, for each of the $M = 1000$ simulation runs, a series $\{x_t\}$ of length $T = 200$, and we select an AR(p)-model of the form (3) with $0 \leq p \leq p_T$, and $h = 2$, which best fits this series. For each simulation run the selection is done by (i) AIC, (ii) BIC and (iii) FIC. Once the model is selected, we make a h-step ahead forecast of this series $\{x_t\}$, using the direct method as described above, employing the parameter estimates from the selected model. This forecast is denoted by $\hat{x}_{T+h}^{(j)} = x_{T+h}(p_{h,j}, h)'\hat{\phi}(p_{h,j}, h)$, where j is the number of the simulation run, and $p_{h,j}$ is the AR-order of the model selected for the h-step ahead forecast in run j. The fourth simulation experiment proceeds along the same lines as the third setting, with the exception that now, for the h-step ahead prediction of the series $\{x_t\}$, we use the plug-in method. The results of these simulation experiments are summarised in Tables 3 and 4.

These tables show the relative MSE of the 2-step ahead prediction in the models selected by FIC, compared to those selected by AIC (top tables) and compared to those obtained by BIC (bottom tables). Table 3 shows the results using the direct method for forecasting is done, and Table 4 shows the results for forecasting using the plug-in method. The favourable results for FIC, obtained in the two time series setting (Tables 1 and 2), no longer apply. All three criteria FIC, AIC, and BIC now perform similarly. This is observed more clearly in Figure 2, which is a graphical representation. Across the entire range of the (ϕ, η) parameter space, the surface is almost flat with values close to 1, from which we conclude that none of the model selection methods significantly outperforms the others in mean squared prediction error sense. Another observation is that the relative MSEs do not change significantly whether we use the direct method or the plug-in method. Finally, comparing the MSEs of the direct and the plug-in method with each, with model selection done using AIC, we found that the methods perform comparably well in the two-series case. This was also true for

ϕ/η	-0.9	-0.7	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.7	0.9
-0.9	0.994	0.997	0.990	0.992	1.010	1.008	0.999	1.006	1.024	1.058
-0.7	0.994	1.013	1.019	1.027	1.017	1.033	1.015	1.029	1.037	1.052
-0.5	0.985	0.985	0.997	1.036	1.014	1.050	1.030	1.041	1.019	1.041
-0.3	1.017	1.023	1.018	1.012	1.033	1.036	1.038	1.056	1.035	1.030
-0.1	1.041	1.049	1.047	1.058	1.035	1.059	1.034	1.047	1.052	1.031
0.1	1.026	1.049	1.044	1.020	1.039	1.040	1.036	1.029	1.023	1.045
0.3	1.043	1.052	1.062	1.019	1.031	1.035	1.019	1.017	1.012	1.043
0.5	1.016	1.043	1.042	1.036	1.021	1.017	1.009	1.035	1.050	1.024
0.7	1.027	1.042	1.054	1.011	0.997	1.011	1.006	1.024	1.002	1.020
0.9	1.022	1.015	1.022	1.030	1.000	1.025	0.999	1.003	1.003	1.006

ϕ/η	-0.9	-0.7	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.7	0.9
-0.9	1.001	1.003	1.003	1.001	1.004	1.010	1.014	1.014	1.056	1.064
-0.7	1.027	1.036	1.022	1.035	1.022	1.044	1.029	1.051	1.038	1.069
-0.5	1.013	1.004	1.022	1.044	1.024	1.057	1.049	1.050	1.032	1.088
-0.3	1.050	1.073	1.050	1.031	1.067	1.044	1.053	1.066	1.069	1.060
-0.1	1.053	1.069	1.067	1.093	1.055	1.073	1.062	1.054	1.074	1.047
0.1	1.054	1.075	1.065	1.035	1.051	1.062	1.047	1.040	1.036	1.068
0.3	1.056	1.063	1.074	1.033	1.045	1.062	1.042	1.045	1.020	1.079
0.5	1.048	1.078	1.059	1.060	1.029	1.029	1.006	1.062	1.081	1.056
0.7	1.075	1.066	1.074	1.022	1.008	1.024	1.023	1.037	1.017	1.039
0.9	1.036	1.039	1.021	1.038	1.001	1.030	1.008	1.009	1.014	1.008

Table 3: Ratios of mean squared errors for the 2-step ahead prediction of the series $\{x_t\}$, with model order selection using the same series, and prediction according to the *direct method*. An ARMA(1,1)-process generated the series $\{x_t\}$. The autoregression parameter ϕ can be found in the leftmost column, and the moving average parameter η is indicated in the top row. The upper table shows the $rMSE(\cdot, \mathrm{FIC}, \mathrm{AIC})$, the bottom table shows the $rMSE(\cdot, \mathrm{FIC}, \mathrm{BIC})$, as defined in (10). If this ratio is smaller than 1, the FIC selects a model with lower MSE for prediction.

ϕ/η	-0.9	-0.7	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.7	0.9
-0.9	0.999	1.041	1.030	1.033	1.010	1.026	1.039	1.063	1.036	1.031
-0.7	0.997	1.010	1.026	1.010	1.032	1.026	1.026	1.023	1.050	0.998
-0.5	1.000	1.032	1.022	1.021	1.030	1.043	1.012	1.059	1.035	1.032
-0.3	1.016	1.032	1.010	1.029	1.020	1.027	1.046	1.022	1.023	1.027
-0.1	1.002	1.003	1.006	1.002	1.020	1.053	1.048	1.038	1.022	1.012
0.1	0.988	1.022	1.021	1.048	1.039	1.031	1.005	1.017	1.009	1.012
0.3	0.999	1.023	1.009	1.037	1.032	1.039	1.015	1.049	1.018	0.985
0.5	1.028	1.045	1.012	1.023	1.057	1.038	1.046	1.032	1.023	0.988
0.7	1.025	1.034	1.047	1.033	1.079	1.016	1.016	1.041	1.023	1.000
0.9	1.062	1.036	1.026	1.026	1.041	1.025	1.030	1.049	1.015	1.007

ϕ/η	-0.9	-0.7	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.7	0.9
-0.9	1.020	1.042	1.040	1.070	1.029	1.029	1.051	1.049	1.028	1.034
-0.7	1.002	1.019	1.050	1.041	1.044	1.039	1.016	1.041	1.061	1.008
-0.5	1.011	1.057	1.021	1.032	1.060	1.058	1.031	1.071	1.033	1.039
-0.3	1.028	1.051	1.016	1.035	1.048	1.035	1.052	1.045	1.048	1.024
-0.1	1.032	1.000	1.012	1.019	1.032	1.067	1.054	1.052	1.025	1.030
0.1	1.016	1.030	1.024	1.070	1.042	1.042	1.018	1.033	1.011	1.029
0.3	1.024	1.019	1.011	1.040	1.050	1.052	1.029	1.052	1.038	0.997
0.5	1.018	1.075	1.030	1.028	1.062	1.046	1.062	1.039	1.043	0.991
0.7	1.002	1.046	1.057	1.029	1.095	1.021	1.012	1.053	1.022	1.011
0.9	1.068	1.011	1.029	1.039	1.056	1.044	1.049	1.060	1.012	1.010

Table 4: Ratios of mean squared errors for the 2-step ahead prediction of the series $\{x_t\}$, with model order selection using the same series, and prediction according to the *plug-in method*. An ARMA(1,1)-process generated the series $\{x_t\}$. The autoregression parameter ϕ can be found in the leftmost column, and the moving average parameter η is indicated in the top row. The upper table shows the $rMSE(\cdot, \mathrm{FIC}, \mathrm{AIC})$, the bottom table shows the $rMSE(\cdot, \mathrm{FIC}, \mathrm{BIC})$, as defined in (10). If this ratio is smaller than 1, the FIC selects a model with lower MSE for prediction.

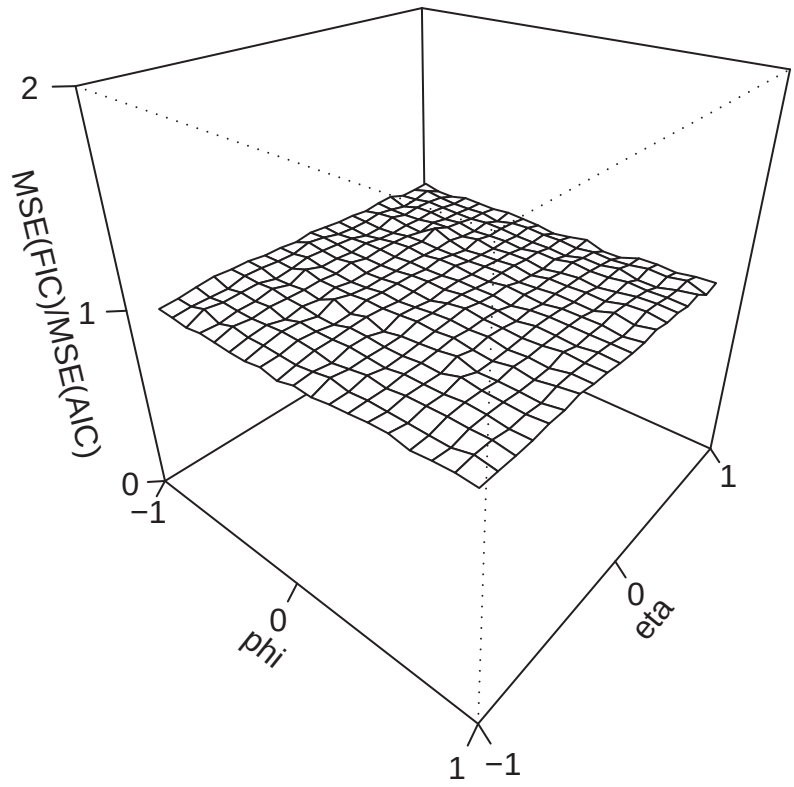

Figure 2: 3D-surface plot for the ratios of mean squared errors for the 2-step ahead prediction of the series $\{x_t\}$, with model order selection using the same series, and where prediction is according to the *direct method*. An ARMA(1,1)-process generated the series $\{x_t\}$. The autoregression parameter ϕ can be found on the phi axis, and the moving average parameter η is indicated on the eta axis. The surface shows the $rMSE(\cdot, \text{FIC}, \text{AIC})$ as defined in (10). If this ratio is smaller than 1, the FIC selects the models with the lower MSE for prediction.

the single-series setting, but only when the autoregression parameter ϕ was not close to ± 1. When ϕ is close to 1 in absolute value, we see that the direct method for prediction yields higher MSE than the plug-in method.

5 Real data applications

In this section we compare the performances of AIC, BIC, and FIC on two real datasets. The datasets used are the monthly US liquor sales data (Diebold, 2001, p. 54), and monthly life insurance data (Data available at the URL:
http://www.econ.kuleuven.be/public/NDBAE06/courses/dynmodels/ASSVIE.XLS).
The life insurance data goes from January 1964 to December 1980, and denotes the net number of new personal life insurances for a large insurance company. Since the theory above is developed for stationary series, we first remove the trend and seasonality effects. First, we take the logarithm of the series to make the variance of the innovation terms constant over time. Next, we take the first differences to remove the trend, and take seasonal differences to remove the seasonality effects, such that we have a stationary series. Out-of-sample h-step ahead forecasting is used to estimate the mean squared errors for each of the three information criteria, this for horizons $h = 1, \ldots, 5$. More precisely, we start with the first half of the series $\{x_t\}$, that is $1 \leq t \leq T/2$, and make a prediction of $x_{T/2+h}$. We then add the next observation, $x_{T/2+1}$, and based on $\{x_t\}$, $1 \leq t \leq T/2+1$, make a prediction of $x_{T/2+1+h}$. This process is then repeated until we use all observations up to and including x_{T-h} to predict x_T. We choose the maximal AR-orders of the models equal to $p_T = 15$. Next, we perform a pairwise comparison of the estimated MSEs for each h, and test whether there are significant differences. The MSEs are estimated as

$$MSE = \frac{1}{T/2+1} \sum_{t=T/2}^{T-h} (x_{t+h} - \hat{x}_{t+h})^2.$$

The comparison proceeds by the Diebold-Mariano test (Diebold, 2001, p. 293-294), which is basically a type of paired t-test for equality of means. In this case however, the data consists of squared residuals, one group for each information criterion. As it is likely that there is serial correlation in these residuals, special care must be taken to determine the standard error used in computing the t-values.

(a)	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$
$MSE(AIC)$	1.153	1.516	1.509	1.566	1.630
$MSE(BIC)$	1.392	1.715	1.713	1.781	1.855
$MSE(FIC)$	1.176	1.504	1.528	1.591	1.666
			Diebold-Mariano test results		
AIC–FIC	–0.818 (0.413)	0.312 (0.755)	–0.314 (0.753)	–0.740 (0.459)	–0.549 (0.583)
BIC–AIC	2.735 (0.006)	2.447 (0.014)	2.618 (0.009)	2.652 (0.008)	3.086 (0.002)
BIC–FIC	2.971 (0.003)	4.003 (0.000)	2.696 (0.007)	2.545 (0.011)	1.931 (0.053)

(b)	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$
$MSE(AIC)$	94.51	149.46	134.68	120.46	113.94
$MSE(BIC)$	76.77	119.71	117.13	120.32	118.06
$MSE(FIC)$	84.74	137.28	124.18	118.36	120.91
			Diebold-Mariano test results		
AIC–FIC	1.892 (0.059)	1.300 (0.194)	1.006 (0.314)	0.180 (0.857)	–0.560 (0.575)
BIC–AIC	–2.086 (0.037)	–2.212 (0.027)	–1.888 (0.059)	–0.016 (0.987)	0.390 (0.697)
BIC–FIC	–0.807 (0.420)	–1.191 (0.234)	–0.549 (0.583)	0.142 (0.887)	–0.227 (0.820)

Table 5: Comparison of the information criteria FIC, AIC, and BIC. The table contains the estimated mean squared forecast errors ($\times 10^{-3}$), and t-values (p-values) of the Diebold-Mariano test. Results are given in (a) for the US Liquor sales data, and in (b) for the life insurance data.

Table 5 shows the estimated mean square forecast errors for the different prediction horizons h and the different order selection criteria. It also shows the t-values and corresponding p-values for the Diebold-Mariano tests. The upper table shows the resulting values for the US liquor sales time series, and the bottom table shows the corresponding results for the Life Insurance time series. A positive t-value means that the first criterion leads to predictions with a higher MSE than the second criterion.

For the US liquor sales time series we observe that there are no significant differences in performance between AIC and FIC. On the other hand, the BIC performs significantly worse than both AIC and FIC. For the Life Insurance time series FIC performs slightly, but not significantly, better than AIC. The BIC is not significantly better than the FIC. To conclude, we can see that the three different information criteria perform equally well on the two examples considered, hereby confirming the results of Section 4.

6 Extensions

In this section we list three extensions of the main ideas in this paper. First we explain how the results need to be adapted in order to derive similar results for prediction with the plug-in method. Second, we provide the expression for FIC when the impulse response is the focus parameter. Third, we obtain an FIC definition for simultaneous selection of regression variables and the autoregressive order of the error terms.

6.1 Using plug-in methods

Direct prediction results in a h-step ahead predictor which is a linear combination of the parameter estimates. Therefore Proposition 1 is applicable. In contrast, the plug-in method leads to a predictor which is a polynomial of order h of the parameter estimates, see equation (4). In order to derive the distribution of the predictor in each candidate model, the first main step is to show that a suitably scaled version of $\sqrt{T}\Big(g\big(\hat{\phi}(p_T)\big) - g\big(\phi_{\text{true}}(p_T)\big)\Big)' y(p_T)$ has an asymptotic normal distribution, where $g(\hat{\beta}) = \Big(g_1\big(\hat{\phi}(p_T)\big), \ldots, g_{p_T}\big(\hat{\phi}(p_T)\big)\Big)'$ with $g_i\big(\hat{\phi}(p_T)\big)$ a polynomial of degree h in $\hat{\phi}_1(p_T), \ldots, \hat{\phi}_{p_T}(p_T)$. The argument in the appendix shows why this is the case in our setting. We then proceed by computing its limiting mean

squared error, and by estimating this quantity in an unbiased way. This estimator is the FIC, which is then computed for each candidate autoregressive order p. In our setting, it has the same form as the FIC for the direct method (9), but with $y(p_T)$ replaced by the recursively defined

$$\hat{\omega}_h(p_T) = \hat{m}_h(p_T) + \hat{\Omega}_h(p_T)\phi(p_T).$$

Here $\hat{m}_h(p_T) = (\hat{y}_{T+h-1}(p_T), \ldots, \hat{y}_{T+h-p_T}(p_T))$ with $\hat{y}_{T+i}$ defined as in (4). Also, $\hat{\Omega}_h(p_T) = (\hat{\omega}_{h-1}(p_T), \ldots, \hat{\omega}_{h-p_T}(p_T))$ where $\hat{\omega}_i(p_T) = 0$ for $i \leq 0$. The $y(p)$ in expression (9) are replaced by a vector containing the first p elements of $\hat{\omega}_h(p_T)$. This yields the FIC we have used in the simulations of Section 4 for the plug-in method for prediction. The model finally selected is, as before, the model with the lowest value of FIC.

6.2 Focus on the impulse response

Up to now the goal was to select the autoregressive order p to obtain the h-step ahead predictor with the smallest value of the FIC. Here we change focus to the impulse response at lag τ, denoted $\imath(\tau)$. This situation, using FIC to select the best AR-order for making estimates of the impulse response function at a certain lag, has been investigated via a simulation study in Hansen (2005). Here we give a theoretical justification for the use of FIC in this setting.

We use the same notation as in Section 2. The focus parameter μ introduced in Section 3 gets replaced by $\mu = \imath(\tau)$. The plug-in method based on model (1) leads to the following estimated focus parameter

$$\hat{\mu} = \hat{\imath}(\tau) = \hat{\phi}_1(p)\hat{\imath}(\tau - 1) + \cdots + \hat{\phi}_p(p)\hat{\imath}(\tau - p),$$

where $\hat{\imath}(\tau) = 0$ for $\tau < 0$ and $\hat{\imath}(0) = 1$.

From this expression it is clear that estimating the impulse response of a time series at lag τ is a special case of a τ-step ahead prediction, applied to a time series with 0 on every time t, except for a 1 on time T, where the parameter estimators are constructed from the given time series $\{x_t\}$. With this observation, the results of Proposition 1 are readily applicable for the impulse response as a focus parameter. If the observation made in Section 6.1 holds, this guarantees the correctness of the FIC as an unbiased estimator of the limiting mean

squared error in the case of a growing number of parameters. The expression of FIC_p for impulse response is as given in the previous subsection, though now with $y_T = 1$ and $y_t = 0$ for $1 \leq t < T$.

6.3 Simultaneous selection of regression variables and the AR order

Up to now we considered stationary time series with zero mean. We implicitly assumed that the trend and the seasonality effects of this series were removed beforehand. Furthermore, we made the assumption that the series $\{x_t\}$ has been regressed on another explicative time series. We worked with the residual series. This is the most commonly used technique when estimating and predicting time series: first identify and fit the deterministic component, and then determine the error-structure. However, if the identification of the deterministic component includes a variable selection step, Golan et al. (1996, Chapter 10) illustrate that the classical variable selection criteria perform poorly if the residual errors do not satisfy the uncorrelatedness assumption. Recently, Shi & Tsai (2004) proposed an alternative selection criterion which simultaneously selects the regression variables for inclusion in the model and the autoregressive order of the error terms.

In a similar spirit, we can employ the Focussed Information Criterion to perform simultaneous selection of the regression variables to include in the deterministic part and of the AR-order of the model errors. Assume that we have a time series $\{y_t\}$ and explicative series $\boldsymbol{x}_t = \big(\{x_{t,1}\}, \ldots, \{x_{t,k}\}\big)'$, and that the data are generated from the following model

$$y_t = \boldsymbol{x}_t'\beta + u_t \quad \text{with} \quad u_t = \phi_1 u_{t-1} + \cdots + \phi_P u_{t-P} + \varepsilon_t, \tag{11}$$

where the errors ε_t are independent and identically normally distributed with mean 0 and variance σ^2 for $t = P{+}1, \ldots, T$, and where $U_P = (u_1, \ldots, u_P)'$ is distributed as $\mathcal{N}(0, \sigma^2 R(P))$. The log-likelihood function under the model (11) is then (omitting constants not depending on the model):

$$\ell(\beta, \Phi, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2}\log|R(P)| - \frac{1}{2\sigma^2}(Y - X\beta)'\Sigma^{-1}(Y - X\beta),$$

where $\Phi = (\phi_1, \ldots, \phi_P)'$, $Y = (y_1, \ldots, y_T)'$, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)'$, $U = (u_1, \ldots, u_T)'$, and $\Sigma = \mathrm{Cov}(U)/\sigma^2$. Note that Σ and $R(P)$ depend on Φ. The expressions for $|R(P)|$ and Σ^{-1}

can be found in Ljung & Box (1979). To facilitate the derivations, we condition on the first P observations, and write the conditional log-likelihood function as

$$\begin{aligned}
&\ell(\beta, \Phi, \sigma^2 \mid x_1, \ldots, x_P, y_1, \ldots, y_P) \\
&\quad = \; -\frac{n}{2}\log\sigma^2 - \frac{1}{2}\log|R(P)| - \frac{1}{2\sigma^2}\sum_{t=P+1}^{T}\Big(y_t - \boldsymbol{x}_t'\beta - \sum_{i=1}^{P}\phi_i(y_{t-i} - \boldsymbol{x}_{t-i}'\beta)\Big)^2. \quad (12)
\end{aligned}$$

From this expression, we derive the estimated $(k+P)\times(k+P)$ information matrix $J_{T,\text{full}}$. This matrix has components

$$\begin{aligned}
\big(J_{T,\text{full}}\big)_{\beta_i,\beta_j} &= -\frac{1}{T-P}\cdot\frac{\partial^2\ell(\cdot)}{\partial\beta_i\partial\beta_j}\Big|_{\hat\beta,\hat\Phi,\hat\sigma^2} \\
&= \frac{1}{(T-P)\hat\sigma^2}\sum_{t=P+1}^{T}\Big(x_{t,i} - \sum_{l=1}^{P}\hat\phi_l x_{t-l,i}\Big)\Big(x_{t,j} - \sum_{l=1}^{P}\hat\phi_l x_{t-l,j}\Big), \\
\big(J_{T,\text{full}}\big)_{\phi_i,\phi_j} &= -\frac{1}{T-P}\cdot\frac{\partial^2\ell(\cdot)}{\partial\phi_i\partial\phi_j}\Big|_{\hat\beta,\hat\Phi,\hat\sigma^2} \\
&= \frac{1}{T-P}\cdot\frac{\partial^2\log|R(P)|}{\partial\phi_i\partial\phi_j}\Big|_{\hat\Phi} \\
&\qquad + \frac{1}{(T-P)\hat\sigma^2}\sum_{t=P+1}^{T}(y_{t-i} - x_{t-i}'\hat\beta)(y_{t-j} - x_{t-j}'\hat\beta), \text{ and} \\
\big(J_{T,\text{full}}\big)_{\beta_i,\phi_j} &= -\frac{1}{T-P}\cdot\frac{\partial^2\ell(\cdot)}{\partial\beta_i\partial\phi_j}\Big|_{\hat\beta,\hat\Phi,\hat\sigma^2} \\
&= \frac{1}{(T-P)\hat\sigma^2}\sum_{t=P+1}^{n}\Big\{x_{t-j,i}\Big(y_t - x_t'\hat\beta - \sum_{l=1}^{P}(y_{t-l} - x_{t-l}'\hat\beta)\Big) \\
&\qquad + (y_{t-j} - x_{t-j}'\hat\beta)\Big(x_{t,i} - \sum_{l=1}^{P}\hat\phi_l x_{t-l,i}\Big)\Big\}.
\end{aligned}$$

For S a subset of $\{1,\ldots,k\}$ and $0\le p\le P$, let $\pi_{S,p}$ a projection matrix of dimension $(|S|+p)\times(k+P)$ mapping any vector $\nu = (\nu_{1,1},\ldots,\nu_{1,k},\nu_{2,1},\ldots,\nu_{2,P})'$ onto $(\nu_S,\nu_{2,1},\ldots,\nu_{2,p})'$, where ν_S has components $\nu_{1,i}$ with $i\in S$. Denote $K_{T,S,p} = (\pi_{S,p}J_{T,\text{full}}\pi_{S,p}')^{-1}$ and $M_{T,S,p} = \pi_{S,p}'K_{T,S,p}\pi_{S,p}$. The focus parameter in the FIC is the h-step ahead forecast, using the plug-in method for prediction:

$$\mu(\hat\beta,\hat\Phi) = x_{T+h}'\hat\beta + \hat\phi_1(\hat y_{T+h-1} - x_{T+h-1}'\hat\beta) + \cdots + \hat\phi_P(\hat y_{T+h-P} - x_{T+h-P}'\hat\beta),$$

and denote ω the vector with components

$$\omega_{1,i} = -\left.\frac{\partial\mu(\beta,\Phi)}{\partial\beta_i}\right|_{\hat{\beta},\hat{\Phi}} = -x_{T+h,i} - \sum_{j=1}^{P}\hat{\phi}_j\left(\left.\frac{\partial\hat{y}_{T+h-j}}{\partial\beta_i}\right|_{\hat{\beta},\hat{\Phi}} - x_{T+h-j,i}\right) \quad 1\le i\le k$$

$$\omega_{2,j} = -\left.\frac{\partial\mu(\beta,\Phi)}{\partial\phi_j}\right|_{\hat{\beta},\hat{\Phi}} = -\sum_{i=1}^{P}\left(\hat{y}_{T+h-i} - x'_{T+h-i}\hat{\beta} + \left.\frac{\partial\hat{y}_{T+h-i}}{\partial\phi_j}\right|_{\hat{\beta},\hat{\Phi}}\right) \quad 1\le j\le P,$$

where $\hat{y}_t = y_t$ and hence $(\partial\hat{y}_t)/(\partial\beta_i) = (\partial\hat{y}_t)/(\partial\phi_j) = 0$ for $t \le T$.

Combining these ingredients leads to

$$\mathrm{FIC}_{S,p} = \omega'(I - M_{T,S,p}J_{T,\mathrm{full}})\hat{\delta}\hat{\delta}'(I - J_{T,\mathrm{full}}M_{T,S,p})\omega + 2\omega M_{T,S,p}\omega,$$

where $\hat{\delta} = \sqrt{T}(\hat{\beta},\hat{\Phi})'$. The model with the smallest value of FIC is selected. This version of FIC can simultaneously select a subset of the explicative variables $x_{t,1},\ldots,x_{t,k}$ and the autoregressive order p of the error term, where $0 \le p \le P$.

7 Concluding Remarks

In this paper we extended the FIC mechanism to allow for an increasing number of parameters as the sample size increases. We specifically worked inside the framework of h-step ahead prediction of time series using an AR-model, with the direct method for prediction. We illustrated, via simulations, that FIC selects models which give predictions with a lower MSE than that of AIC in some areas of the parameter space. The best results are obtained in the case where two independent time series are available, one series is used to estimate the parameters, the other series to predict. In practical applications, where only one time series is available, a simulation study and two real time series examples have shown that the performance of the FIC is comparable to that of the classical information criteria. This simulation study also demonstrated that the relative mean squared errors for the plug-in method for prediction are quite comparable to those of the direct method. We gave a theoretical justification for Hansen's (2005) use of the FIC for the impulse response. An extension to simultaneous selection of regression variables and autoregressive order is promising to explore more in-depth.

Acknowledgement

This research has been supported by the Research Fund K.U.Leuven and the Fund for Scientific Research Flanders (Contract numbers G.0594.05 and G.0542.06).

Appendix

Assumptions

We make the following assumptions on the series $\{x_t\}$ and $\{y_t\}$:

(A1) The maximum and minimum eigenvalues of $X'X$ satisfy (for constants $B > 0$ and $b > 0$)

$$\lambda_{\max}(X'X) \leq BT; \qquad \lambda_{\min}(X'X) \geq bT,$$

where $X = \big(x_{p_T+h+1}(p_T, h), \ldots, x_T(p_T, h)\big)'$.

(A2) We define $\alpha_t(p_T, h) = (X'X)^{-1/2} x_t(p_T, h)$, for $t = p_T + h + 1, \ldots, T$. Then uniformly in t_1 and t_2,

$$\alpha_{t_1}(p_T, h)' \alpha_{t_2}(p_T, h) = \mathcal{O}(p_T/T).$$

(A3) $\|y\big(p_T\big)\| = \mathcal{O}(\sqrt{p_T})$, and $\max\{|x_t(p_T, h)' y(p_T)| : t = p_T+h+1, \ldots, T\} = \mathcal{O}(p_T\sqrt{\log T})$.

These assumptions on the time series $\{x_t\}$ have an intuitive explanation. Assumption (A1) amounts to having an empirical autocovariance matrix which is bounded for all lengths T, and for which the inverse exists and is bounded. (A2) states that there are no outlying observations of the time series, and (A3) limits the extent of the dependency between the series $\{x_t\}$ and $\{y_t\}$.

We first prove the following lemma, which is an adaptation of Theorem 3.2 in Portnoy (1985) for the setting in which we work.

Lemma 1 *Under assumptions (A1), (A2), and (A3), and the condition* $p_T\sqrt{\log T}/T \to 0$ *for* $T \to \infty$*, the following result holds,*

$$y(p_T)'\big(\hat{\delta}(p_T, h) - \delta_{\text{true}}(p_T, h)\big) \left(\frac{1}{v\sigma}\right) \to_d \mathcal{N}(0, 1) \text{ for } T \to \infty$$

where $v^2 = y(p_T)'(X'X)^{-1}y(p_T)$ *and* σ^2 *as in model* (2)*.*

The proof follows the same lines as the proof of Theorem 3.2 in Portnoy (1985).

Proof. Let $b(p_T) = (X'X)^{-1/2}y(p_T)$. Then we can write that

$$v^2 = \|b(p_T)\|^2 \quad \text{and} \quad y(p_T)'\big(\hat{\delta}(p_T,h) - \delta_{\text{true}}(p_T,h)\big)\begin{pmatrix}1\\ v\sigma\end{pmatrix} = \frac{b(p_T)'\hat{\theta}}{\|b(p_T)\|}$$

with $\theta = \frac{1}{\sigma}(X'X)^{1/2}\big(\hat{\delta}(p_T,h) - \delta_{\text{true}}(p_T,h)\big)$. It suffices to show that for $\|b(p_T)\| = 1$, $b(p_T)'\hat{\theta} \to_d \mathcal{N}(0,1)$. So, assume that $\|b(p_T)\| = 1$. For OLS estimation and normally distributed error terms, Lemma 3.4 of Portnoy (1985) is applicable, and gives that

$$b(p_T)'\hat{\theta} = \frac{1}{\sigma}b(p_T)'\sum_{t=p_T+h}^{T}\alpha_t(p_T,h)'b(p_T)\varepsilon_t(p_T,h).$$

Using the definition of $b(p_T)$ and assumptions (A1) and (A2), we find $\alpha_t(p_T,h)'b(p_T) = x_t(p_T,h)'(X'X)^{-1}y(p_T) = \frac{c}{T}x_t(p_T,h)'y(p_T)$ for some constant c. Using assumption (A3) and the constraint on p_T, we then arrive at $\max_t|\alpha_t(p_T,h)b(p_T)| = \mathcal{O}(p_T\sqrt{T}/T) \to 0$ as $T \to \infty$. With

$$\sum_{t=p_T+h}^{T}\big(\alpha_t(p_T,h)'b(p_T)\big)^2 = \|b(p_T)\|^2 = 1,$$

the Central Limit Theorem implies that $b(p_T)'\hat{\theta} \to_d \mathcal{N}(0,1)$ as $T \to \infty$, and the lemma holds. □

Proof of Proposition 1

Proof. For h-step ahead prediction

$$\begin{aligned}
&\sqrt{T}\big(\hat{\mu}(p,h) - \mu_{\text{true}}(p_T,h)\big)\\
&\quad= \sqrt{T}\big(\hat{\mu}(p,h) - \hat{\mu}_{\text{true}}(p,h)\big) + \sqrt{T}\big(\hat{\mu}_{\text{true}}(p,h) - \hat{\mu}_{\text{true}}(p_T,h)\big)\\
&\quad= \sqrt{T}\big(\hat{\phi}(p,h) - \phi(p,h)\big)'y(p) + \sqrt{T}\big(\phi(p,h)'y(p) - \phi(p_T,h)'y(p_T)\big).
\end{aligned}$$

The first term converges in distribution to a normal distribution. This follows by application of the limiting result in Hjort & Claeskens (2003, Lemma 3.3), where the maximal order is equal to p finite. The second term converges to a constant, since $\phi(p_T,h)'y(p_T)$ is $\mathcal{O}_p(1/\sqrt{T})$. Hence, for each p fixed, the proposition holds.

However, the proposition must also hold for a growing number of time series components:

$$\begin{aligned}\sqrt{T}\big(\mu(p_T,h)-\mu_{\text{true}}(p_T,h)\big) &= \sqrt{T}\big(\hat{\phi}(p_T,h)-\phi(p_T,h)\big)'y(p_T)\\ &= \big(\hat{\delta}(p_T,h)-\delta(p_T,h)\big)'y(p_T).\end{aligned}$$

Lemma 1 proves that this converges to a normal distribution as $T\to\infty$, and the proposition holds. □

Proof of Extension 6.1

Assume that h is the fixed prediction horizon and assume that p_T, the maximal AR-order of the considered models, satisfies the condition in Proposition 1. We also assume that $h \le p_T$. Then recursive substitution reveals that

$$\begin{aligned}\hat{\mu}(p_T,h) &= \hat{\phi}_1(p_T)\hat{\mu}(p_T,h-1)+\cdots+\hat{\phi}_{h-1}(p_T)\hat{\mu}(p_T,1)\\ &\quad+\hat{\phi}_h(p_T)y_T+\cdots+\hat{\phi}_{p_T}(p_T)y_{T+h-p_T}\\ &= \big[\hat{\phi}_h(p_T)+\tilde{g}_1\big(\hat{\phi}(p_T)\big)\big]y_T+\cdots+\big[\hat{\phi}_{p_T}(p_T)+\tilde{g}_{p_T-h}\big(\hat{\phi}(p_T)\big)\big]y_{T+h-p_T}\\ &\quad+\tilde{g}_{p_T-h+1}\big(\hat{\phi}(p_T)\big)y_{T+h-p_T-1}+\cdots+\tilde{g}_{p_T}\big(\hat{\phi}(p_T)\big)y_{T-p_T},\end{aligned}$$

where we used that $\hat{\mu}(p_T,-i)=y_{T-i}$ for $i\ge 0$. In this expression, $\tilde{g}_i\big(\hat{\phi}(p_T)\big)$ for $1\le i\le p_T$ are polynomials of degree h in $\hat{\phi}_1(p_T),\ldots,\hat{\phi}_{p_T}(p_T)$ without a constant term or a first degree term. Since $\hat{\phi}(p_T)=\hat{\delta}(p_T)/\sqrt{T}$, it can be verified easily that $\tilde{g}_i\big(\hat{\phi}(p_T)\big)=\mathcal{O}_p(1/T)$ for all $1\le i\le p_T$. We use this to rewrite the expression $\sqrt{T}(\hat{\mu}(p_T,h)-\mu(p_T,h))$ as

$$\sqrt{T}(\hat{\mu}(p_T,h)-\mu(p_T,h))=\sqrt{T}\Big(\sum_{i=0}^{p_T-h}\hat{\phi}_{h+i}(p_T)y_{T-i}\Big)+\sqrt{T}\Big(\sum_{i=0}^{p_T}\tilde{g}\big(\hat{\phi}(p_T)\big)y_{T-i}\Big),$$

where $\mu(p_T,h)$ is the true value of the plug-in estimator. From the previous, we see that the second term is $\mathcal{O}_p(1/\sqrt{T})$ and hence will have no contribution in the limit. We can then apply the same reasoning as in the proof of Proposition 1, but with $\tilde{y}(p_T)=(0,\ldots,0,y_T,\ldots,y_{T+h-p_T})'$ of length p_T, which proves the validity of Extension 6.1. □

References

Akaike, H. (1974). A new look at statistical model identification. *I.E.E.E. Transactions on Automatic Control*, **19**, 716–723.

Bhansali, R. J. (1996). Asymptotically efficient autoregressive model selection for multistep prediction. *Annals of the Institute of Statistical Mathematics*, **48**, 577–602.

Brockwell, P. J. and Davis, R. A. (1995). *Time Series: Theory and Methods*, 2nd Edition. Springer, New York.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion [with discussion]. *Journal of the American Statistical Association*, **98**, 900–916.

Diebold, F. X. (2001). *Elements of Forecasting*, 2nd Edition. South-Western.

Golan, A., Judge, G. and Miller D. (1996). *Maximum Entropy Economics: Robust Estimation with Limited Data*. John Wiley & Sons, New York.

Hansen, B. E. (2005). Challenges for econometric model selection. *Econometric Theory*, **21**, 60–68.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators [with discussion]. *Journal of the American Statistical Association*, **98**, 879–899.

Ing, C.-K. and Wei, C.-Z. (2005). Order selection for same-realization prediction in autoregressive processes. *Annals of Statistics*, **33**, 2423–2474.

Lee, S. and Karagrigoriou, A. (2001). An asymptotically optimal selection of the order of a linear process. *Sankhyā: The Indian Journal of Statistics*, **63**, 93–106.

Ljung, G. M. and Box, G. E. P. (1979). The likelihood function of stationary autoregressive-moving average models. *Biometrika*, **66**, 265–270.

Portnoy, S. (1985). Asymptotic behaviour of M estimators of p regression parameters when p^2/n is large; II. Normal approximation. *The Annals of Statistics*, **13**, 1403–1417.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Shi, P. and Tsai, C.-L. (2004). A joint regression variable and autoregressive order selection criterion. *Journal of Time Series Analysis*, **25**, 923–941.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics*, **8**, 147–164.