ISSN 1440-771X

AUSTRALIA

DEPARTMENT OF ECONOMETRICS
AND BUSINESS STATISTICS

Empirical Information Criteria for Time Series Forecasting Model Selection

Md B Billah, R.J. Hyndman and A.B. Koehler

Working Paper 2/2003

Empirical information criteria for time series forecasting model selection

Md.Baki Billah[1], Rob J. Hyndman[1], Anne B. Koehler[2] 29 January 2003

Abstract: In this paper, we propose a new Empirical Information Criterion (EIC) for model selection which penalizes the likelihood of the data by a function of the number of parameters in the model. It is designed to be used where there are a large number of time series to be forecast. However, a bootstrap version of the EIC can be used where there is a single time series to be forecast. The EIC provides a data-driven model selection tool that can be tuned to the particular forecasting task.

We compare the EIC with other model selection criteria including Akaike's Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (BIC). The comparisons show that for the M3 forecasting competition data, the EIC outperforms both the AIC and BIC, particularly for longer forecast horizons. We also compare the criteria on simulated data and find that the EIC does better than existing criteria in that case also.

Keywords: Exponential smoothing; forecasting; information criteria; M3 competition; model selection.

JEL classification: C53, C52, C22.

[1]Department of Econometrics & Business Statistics, Monash University, Clayton, VIC 3800, Australia
[2]Department of Decision Sciences & Management Information Systems, Miami University, Oxford, OH 45056, USA.
This research was supported by the Business & Economics Forecasting Unit, Department of Econometrics & Business Statistics, Monash University.

1 Introduction

In many industrial applications a large number of series need to be forecast on a routine basis; examples include production planning and inventory management. In the last few decades many forecasting models have been developed. The forecaster may either select one appropriate model for all series under consideration, or may use a general selection methodology which will select the appropriate model for each series from a group of competitive models. The appropriate choice of forecasting model has the potential for major cost savings through improved accuracy.

Information criteria provide a simple method to choose from a range of competing models. However, it is not clear which information criterion is best for a given forecasting task, and any one information criterion does not perform well for all forecasting model selection problems (see, for example, Billah et al., 2001; Mills & Prasad, 1992; Hurvich & Tsai, 1991). The arguments for and against each information criteria are usually highly theoretical, and it is not clear how to proceed in practice. The performance of an information criterion may depend on a number of factors such as the models in the choice set, forecasting horizons, and the series being forecast. Thus, the practitioner is confronted with a problem: which information criterion is the best for selecting an appropriate forecasting model for each time series?

We overcome these problems by proposing a data-driven information criterion that we call the Empirical Information Criterion (EIC). The EIC can be tuned to the particular forecasting task. This new criterion uses information from a large number of roughly similar time series to calibrate the EIC appropriately.

Suppose we have a single time series of length n and N possible models from which to choose. We can choose amongst these models using an information criterion, defined as a penalized log-likelihood:

$$\text{IC} = \log L(\widehat{\theta}) - f(n, q), \tag{1.1}$$

where $\log L(\widehat{\theta})$ is the maximized log-likelihood function, θ is the q-vector of unknown free parameters and $f(n, q)$ is the corresponding penalty function. The model with the largest value of IC is the chosen model.

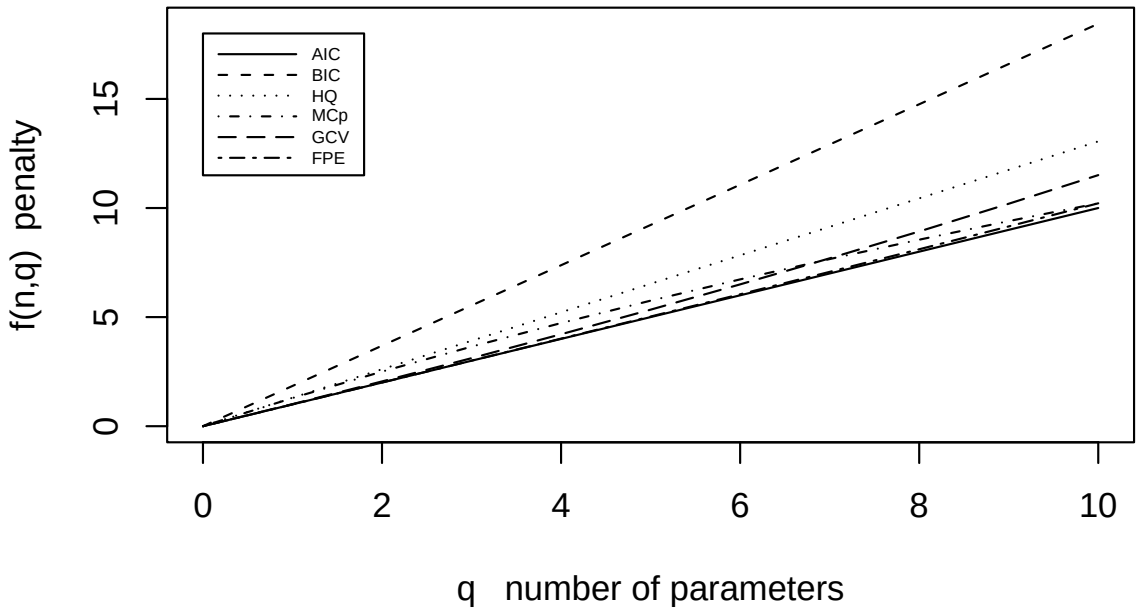

Figure 1: *Penalty functions for six different information criteria.*

Six commonly-used information criteria are Akaike's Information Criterion (AIC; Akaike, 1973), the Bayesian Information Criterion (BIC; Schwarz, 1978), Hannan & Quinn's criterion (HQ; Hannan & Quinn, 1979), Mallows' Criterion (MCp; Mallows, 1964), the Generalized Cross Validation criterion (GCV; Golub et al., 1979) and the Finite Prediction Error criterion (FPE; Akaike, 1970). The penalty functions of these criteria are as follows:

Criterion	**Penalty Function**
AIC	q
BIC	$q\log(n)/2$
HQ	$q\log(\log(n))$
MCp	$n\log(1+2q/r)/2$
GCV	$-n\log(1-q/n)$
FPE	$(n\log(n+q)-n\log(n-q))/2$

where $r = n - q^*$ and q^* is the number of free parameters in the smallest model that nests all models under consideration. Figure 1 shows the penalty functions for the six criteria for $n = 40$ and $q^* = 10$.

Any of these six information criteria may be used for automatic selection among competing forecasting models (e.g., Hyndman et al., 2002). However, rather than using a fixed penalty function $f(n,q)$, we estimate the penalty function for the particular forecasting task, using an

ensemble of similar time series.

The plan of this paper is as follows. We introduce the EIC in Section 2. Section 3 describes the application of the EIC to the M3 forecasting competition data using exponential smoothing models, and we show that it performs better than the existing information criteria. We apply the bootstrap version of the EIC in Section 4 which is applicable when there is only one series to be forecast. The paper ends with some concluding remarks in Section 5.

2 Two new empirical information criteria

Suppose we have m time series that are 'similar' to each other. Let $y_{t,j}$ be the tth observation of the jth series ($j = 1, \ldots, m$ and $t = 1, \ldots, n_j$). We denote the jth series by $\boldsymbol{y}_j$ and the ensemble of series by $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m]$.

This situation can arise when we have a large inventory of m products for which sales need to be forecast on a regular basis. The m series form the ensemble used to compute the penalty function.

Alternatively, we can fit an initial model to the series of interest (chosen using the AIC for example), and then generate m bootstrap series from the fitted model. In this case, the bootstrap series form the ensemble for estimating the penalty function which is then applied to the original series of interest. This approach has some similarities with the algorithms proposed by Grunwald & Hyndman (1998) and Chen et al. (1993), although these authors were considering model selection in other contexts.

A number of different forecast evaluation criteria could be used for selecting the penalty function. Because we are particularly interested in forecast accuracy, we shall use the mean absolute percentage error (MAPE) as the evaluation criterion.

We consider two different forms of EIC, one involving a non-linear penalty function and the other involving a linear penalty function. Both assume that the penalty does not depend on the length of the series. This is not as restrictive as it first appears because we are estimating the penalty function based on time series that are similar. Thus all m series will usually be of the same or similar length.

The non-linear EIC (NLEIC) has $f(n,q) = k_q q$ where k_q is the penalty weight for a model with q parameters. Thus

$$\text{NLEIC} = \log L(\widehat{\theta}) - k_q q.$$

The model with the largest NLEIC is the chosen model. If q_i is the number of parameters for the ith model ($i = 1, 2, \ldots, N$), then the penalty weights k_{q_i}, need to be estimated from the ensemble of m series. Without loss of generality we let the first model have the fewest parameters and assume $k_{q_1} = 0$.

Figure 1 shows that the commonly used penalty functions are all close to linear over the range of interest. For larger n, they become more linear (and MCp, GCV and FPE all converge to AIC). This suggests that it may be beneficial to restrict attention to Information Criteria where $f(n,q) = kq$ and k is a constant across all models considered. We call this the Linear Empirical Information Criterion (LEIC):

$$\text{LEIC} = \log L(\widehat{\theta}) - kq,$$

and the value of k is estimated from the ensemble of m series.

For both EIC, each series in the ensemble is divided into two segments: the first segment consists of $n_j^* = n_j - H$ observations; the second segment consists of the last H observations. The value of H needs to be chosen by the forecaster according to what is appropriate for the particular series of interest. A common choice will be to set H to the largest forecast horizon required.

2.1 Penalty estimation for LEIC

For the LEIC, we need to select a value for k using the m series in $\boldsymbol{Y}$. Small changes in k will not usually result in a change in the selected model. Therefore this is not a smooth optimization problem.

We consider values of k between 0.25 and $2\log(n)$ in steps of size δ. This range of values is wider enough to contain all of the commonly used penalty functions. We have found that $\delta = 0.25$ works well in practice.

The steps for estimating k for the LEIC are as follows.

Step 1: Model estimation

1a) For each of the m series, use the first n_j^* observations to estimate the parameters in each of the N competing models using maximum likelihood estimation.

1b) Record the maximized log-likelihoods for all estimated models.

Step 2: Penalty estimation

2a) For each trial value of k, select a model for each time series by using LEIC.

2b) For each value of k and for each forecast horizon h, calculate the MAPE across the m time series to obtain

$$\text{MAPE}(h,k) = \frac{100}{m}\sum_{j=1}^{m} \left|y_{n_j^*+h} - \hat{y}_{n_j^*}(h)\right| / y_{n_j^*+h},$$

where $\hat{y}_{n_j^*}(h)$ is the h-step ($h = 1,\ldots,H$) ahead forecast for the model selected for the jth series.

2c) Select a value of $k^{(h)}$ by minimizing MAPE(h,k) over the grid of k values. Thus, a $k^{(h)}$ is selected for each forecast horizon h ($h = 1,\ldots,H$).

2d) We obtain the final value for k by averaging the H values of $k^{(h)}$:

$$k = \frac{1}{H}\sum_{h=1}^{H} k^{(h)}. \tag{2.1}$$

We then use the selected k value to find the best model for each series $\boldsymbol{y}_j$ (using all n_j observations) and produce forecasts from these chosen models.

2.2 Penalty estimation for NLEIC

The estimation of the penalty for the NLEIC is similar except that we have to choose a k_q value for each unique q in $\{q_2,\ldots,q_N\}$. In this non-linear case, there is no reason for the values of k_q to remain positive. Consequently, we consider values of k_q between $-2\log(n)$ and $2\log(n)$ in steps of size δ. Assuming all values of $\{q_1,\ldots,q_N\}$ are unique, and that $(\xi-1)\delta \leq 4\log(n) < \xi\delta$, then there are ξ values of k_{q_i} in the grid, for each i. Thus there are ξ^{N-1} possible sets of $\{k_{q_2},\ldots,k_{q_N}\}$.

Step 1 is same for both LEIC and NLEIC. Step 2 for the NLEIC is as follows.

Step 2: Penalty Estimation

2a) For each trial set of $k_{q_2}, \ldots, k_{q_N}$ select a model for each time series by using NLEIC.

2b) For each set of $k_{q_2}, \ldots, k_{q_N}$ and each forecast horizon h, calculate the MAPE across the m time series to obtain

$$\text{MAPE}(h; k_{q_2}, \ldots, k_{q_N}) = \frac{100}{m} \sum_{j=1}^{m} \left| y_{n_j^*+h} - \hat{y}_{n_j^*}(h) \right| / y_{n_j^*+h}$$

where $\hat{y}_{n_j^*}(h)$ is the h-step $(h = 1, \ldots, H)$ ahead forecast for the model selected for the jth series.

2c) Select a value of $\{k_{q_2}^{(h)}, \ldots, k_{q_N}^{(h)}\}$ by minimizing $\text{MAPE}(h; k_{q_2} \ldots, k_{q_N})$ over the grid of $k_{q_2}, \ldots, k_{q_N}$. Thus, a set $\{k_{q_2}^{(h)}, \ldots, k_{q_N}^{(h)}\}$ is selected for each h, $h = 1, \ldots, H$.

2d) A final value of k_{q_i} is obtained by averaging the H values of $k_{q_i}^{(h)}$ as follows:

$$k_{q_i} = \frac{1}{H} \sum_{h=1}^{H} k_{q_i}^{(h)}. \tag{2.2}$$

We then use the selected set $\{k_{q_1}, \ldots, k_{q_N}\}$ in (2.2) to find the best model for each series $\boldsymbol{y}_j$ (using all n_j observations) and produce forecasts from this chosen model.

We advocate a grid search in these algorithms because the MAPE function is complicated and relatively ill-behaved. However, it does lead to high computational time which increases sharply with the number of parameters and so can be extremely high for small δ. For a large number of parameters, the simulated annealing algorithm of Goffe et al. (1994) can be used instead.

Variations on the algorithm can be obtained by replacing the MAPE criterion by some other criteria. For example, mean absolute error (MAE), mean square error (MSE) and root mean square error (RMSE). (But note that these three assume all m series are on the same scale.)

2.3 Bootstrap EIC

Both the LEIC and NLEIC assume we have a suitable ensemble of m series to use in calibrating the penalty function. However, frequently only one series will be available. In this case, a bootstrap approach may be used.

Chen et al. (1993) proposed a bootstrap approach to estimate a suitable penalty function for selecting the order p of an AR(p) model. Using a similar approach, we generate the additional m series in the ensemble using a bootstrap.

We assume that the series to be forecast is stationary. If this is not so, it should be made stationary through transformations and differences.

Let $\boldsymbol{y}_0$ denote the one series of interest, and let it be of length n. Then the steps for the bootstrap EIC are as follows.

Step 0: Bootstrap sample generation

0a) Fit a high order AR(p) model to $\boldsymbol{y}_0$, the series of interest, and calculate the residuals $\boldsymbol{z} = \{z_1, z_2, \ldots, z_n\}$.

0b) Generate m bootstrap samples of size n from the residuals $\boldsymbol{z}$. Then, generate m samples of size n from the fitted AR(p) model using the m bootstrap samples of residuals as the errors.

Then either the LEIC or NLEIC can be applied to obtain the optimal penalty functions. These penalty functions can then be applied to $\boldsymbol{y}_0$ to obtain a new model for the series of interest. The candidate models should all be stationary in this case; they need not be restricted to AR models.

Chen et al. (1993) show that if the true series is an AR model of order less than p and the candidate models are autoregressive models, then this procedure coupled with the LEIC can produce a consistent estimate of the order of the model.

3 Example 1: Non-seasonal exponential smoothing models and the M3 data

Exponential smoothing methods are widely used in forecasting sales of individual products for inventory control, production scheduling and production planning (Brown, 1959; Winters, 1960; Brown, 1963; Brown, 1967; Gardner, 1985; Makridakis & Wheelwright, 1989; Makridakis & Hibon, 1991). These methods have been shown to perform very well for forecasting (Makridakis et al., 1982; Makridakis et al., 1993; Fildes et al., 1998; Hyndman et al., 2002). As there are many exponential smoothing methods available, using only one method for all time series under study may not give good accuracy (see Fildes, 1989). It is anticipated that selecting a method (from a group of competing methods) to suit each individual series improves forecasting accuracy. Hyndman et al. (2002) describe 24 such exponential smoothing methods and provide state space models for each of them. This allows the likelihood of each model to be computed and allows penalized likelihood model selection to be used.

In this application, we apply the non-seasonal exponential smoothing models to the 3003 time series that were part of the M3 competition (Makridakis & Hibon, 2000). First, we outline the underlying state space model of various exponential smoothing methods used in the plausible group. Then, we discuss the computations and results.

3.1 State space models

A class of state space models has been shown to underpin the exponential smoothing methods (Ord et al., 1997). The linear state space models have the following form:

$$y_t = H\boldsymbol{x}_{t-1} + e_t, \tag{3.1}$$

$$\boldsymbol{x}_t = F\boldsymbol{x}_{t-1} + Ge_t, \tag{3.2}$$

where $\boldsymbol{x}_t$ is an unobserved state variable, e_t is a disturbance term that is independently and normally distributed with mean zero and variance σ^2, and F, G and H are coefficient matrices. Equation (3.1) is called the observation equation and equation (3.2) is called the state equation. When computing the various information criteria, the number of parameters includes the

unknown elements of F, G and H as well as the elements of the initial state vector $\boldsymbol{x}_0$.

The following are some special cases of such state space models (see Hyndman et al., 2002, for details):

Model 1: Local Level Model (LLM):

$y_t = \ell_{t-1} + e_t$, where $\ell_t = \ell_{t-1} + \alpha e_t$ is the local level at time t and α is the exponential smoothing parameter. This underpins the simple exponential smoothing (SES) method.

Model 2: Local Level Model with Drift (LLMD):

$y_t = \ell_{t-1} + b + e_t$, where $\ell_t = \ell_{t-1} + b + \alpha e_t$ is the local level at time t, b is the drift and α is the exponential smoothing parameter. This underpins the SES with drift method. Hyndman & Billah (2003) show that the LLMD is identical to the Theta method of Assimakopoulos & Nikolopoulos (2000) which performed well in the M3 competition of Makridakis & Hibon (2000). Hence, this method is of considerable interest to forecast practitioners.

Model 3: Local Trend Model (LTM):

$y_t = \ell_{t-1} + b_{t-1} + e_t$, where $\ell_t = \ell_{t-1} + b_{t-1} + \alpha e_t$, $b_t = b_{t-1} + \beta e_t$. Here, b_t is the growth rate with exponential smoothing parameter β. It underpins Holt's method.

Model 4: Damped Trend Model (DTM):

$y_t = \ell_{t-1} + b_{t-1} + e_t$, where $\ell_t = \ell_{t-1} + b_{t-1} + \alpha e_t$, $b_t = \phi b_{t-1} + \beta e_t$, and ϕ is the damped parameter. It underpins damped exponential smoothing. The LTM is a special case of DTM.

It is not difficult to see how each of these models can be written in the state space form (3.1) and (3.2). The four models have 2, 3, 4 and 5 parameters respectively.

The h-step ahead point forecasts for the LLM, LLMD, LTM and DTM are given, respectively,

by

$$\widehat{y}_n(h) = \widehat{\ell}_n, \tag{3.3}$$

$$\widehat{y}_n(h) = \widehat{\ell}_n + h\widehat{b}, \tag{3.4}$$

$$\widehat{y}_n(h) = \widehat{\ell}_n + h\widehat{b}_n, \tag{3.5}$$

$$\text{and} \quad \widehat{y}_n(h) = \widehat{\ell}_n + \widehat{b}_n \sum_{i=0}^{h-1} \widehat{\phi}^i, \tag{3.6}$$

where $\widehat{\ell}_n$, $\widehat{b}$, $\widehat{b}_n$ and $\widehat{\phi}$ are maximum likelihood estimates of ℓ_n, b, b_n and ϕ respectively.

3.2 Calculations and results

For the annual data in the M3 competition, the above models are used in this paper as the competitive models. Previous studies (e.g., Makridakis et al., 1982; Makridakis & Hibon, 2000) show that for seasonal data, the deseasonalized exponential smoothing methods do better than their corresponding seasonal versions, particularly for monthly data. Therefore, for seasonal data, the deseasonalized versions of these methods are used. The seasonal data are deseasonalized using the ratio-to-moving average method (Makridakis, Wheelwright & Hyndman, 1998) and the forecasts are re-seasonalized before calculating the MAPE.

Estimates of parameters are obtained by maximizing the conditional log-likelihood as described in Ord et al. (1997). The likelihood depends on $\boldsymbol{x}_0$ and the parameters α, β and ϕ. Constrained optimization was employed to obtain the values of $\boldsymbol{x}_0$ and parameters that maximize the log-likelihood conditional on $\boldsymbol{x}_0$.

We treat the annual, quarterly and monthly data separately. All series in the set of annual time series are used as the ensemble for calibrating the penalty function for these series. Similarly for the quarterly and monthly series.

Each series is divided into two segments: the training set and the test set. The jth time series ($j = 1, \cdots, m$) has n_j observations in the training set and H observations in the test set. For annual, quarterly and monthly data, the values for H are 6, 8 and 18, respectively. The training set is further divided into two subsets. The jth time series has n_j^* observations in the first subset and H observations in the second subset. The data in the training sets are used to estimate the

penalty functions (Steps 1 and 2 in Section 2) for each series.

For the LEIC, the penalties are then averaged across the forecast horizons to obtain a single penalty function for use with annual data, another with monthly data and a third with quarterly data. Similarly, the penalties for NLEIC are averaged to obtain just one for each of annual, monthly and quarterly data.

We compare the LEIC and NLEIC obtained in this way with the six other criteria outlined in Section 1. Each selected model is used to forecast the values in the test set, and the APE is computed for each forecasting horizon.

The MAPEs from the M3 competition are presented in Tables 1 to 3. The results show that both LEIC and NLEIC perform better than all existing information criteria (NLEIC is better than LEIC). Among the existing information criteria BIC is the best. The performances of the criteria AIC, BIC, HQ, MCp, GCV and FPE are not the same, particularly for yearly data where, as compared to quarterly and monthly data, the series sizes are usually smaller. The strength of LEIC and NLEIC is that they work well for all model selection problems. The estimated penalty weights are presented in Table 4.

Figure 2 shows the penalty functions for AIC, LEIC and NLEIC. The estimated penalty weights for LEIC are larger than unity and hence compared to AIC, LEIC penalizes larger models more heavily. The estimated penalties for NLEIC are highly non-linear. The non-linear form is similar for all three data types with a maximum at $q = 4$. This consistency demonstrates the stability of our procedure. Since the NLEIC has much higher values of k_{q_i} for the Local Trend Model than the other models, it has high penalty and will be less likely to be chosen than other models.

Methods	Forecasting Horizons						Average	
	1	2	3	4	5	6	1 to 4	1 to 6
AIC	8.2	12.9	22.0	23.8	28.6	29.7	18.8	22.2
BIC	8.1	12.8	21.8	23.3	28.0	29.4	16.5	20.6
HQ	8.2	12.7	21.9	23.6	28.3	29.7	16.6	20.7
MCp	8.2	12.9	22.0	23.8	28.5	29.6	16.7	20.8
GCV	8.1	12.8	21.8	23.5	28.3	29.2	16.6	20.6
FPE	8.2	13.0	22.1	24.0	28.9	30.1	16.8	21.1
LEIC	8.5	13.0	21.7	21.9	26.0	26.6	16.3	19.6
NLEIC	8.4	12.7	21.4	21.4	25.3	25.8	16.0	19.2

Table 1: *Average MAPE for the annual M3 competition data*

Methods	Forecasting Horizons								Average		
	1	2	3	4	5	6	7	8	1 to 4	1 to 6	1 to 8
AIC	5.2	7.9	8.3	9.3	10.5	13.7	13.0	14.1	7.7	9.2	10.3
BIC	5.2	8.0	8.3	9.3	10.5	13.7	13.0	14.0	7.7	9.2	10.3
HQ	5.2	8.0	8.3	9.3	10.5	13.7	13.0	14.2	7.7	9.2	10.3
MCp	5.2	7.9	8.3	9.3	10.5	13.7	13.0	14.2	7.7	9.2	10.3
GCV	5.2	7.9	8.3	9.3	10.5	13.7	13.0	14.1	7.7	9.2	10.3
FPE	5.2	8.1	8.4	9.5	10.9	14.4	13.8	14.9	7.8	9.4	10.6
LEIC	5.1	8.0	8.2	9.1	10.0	13.2	12.2	13.3	7.6	8.9	9.9
NLEIC	5.0	7.8	7.8	8.8	9.5	12.6	11.6	12.6	7.4	8.6	9.5

Table 2: *Average MAPE for the quarterly M3 competition data*

Methods	Forecasting Horizons								Average			
	1	2	3	4	5	8	12	18	1 to 4	1 to 8	1 to 12	1 to 18
AIC	15.1	13.9	15.7	18.1	14.7	15.6	16.6	21.9	15.7	15.6	16.0	17.6
BIC	15.1	13.8	15.5	17.8	14.6	15.3	16.1	21.8	15.6	15.4	15.7	17.4
HQ	15.2	13.9	15.7	18.0	14.7	15.6	16.6	21.9	15.7	15.6	16.0	17.6
MCp	15.1	13.9	15.7	18.1	14.7	15.6	16.6	21.9	15.7	15.6	16.0	17.6
GCV	15.1	13.9	15.7	18.1	14.7	15.6	16.6	21.9	15.7	15.6	16.0	17.6
FPE	15.2	13.9	15.7	18.0	14.8	15.6	16.6	21.9	15.7	15.6	16.0	17.7
LEIC	15.1	13.9	15.4	17.7	14.5	15.0	15.9	21.4	15.5	15.3	15.6	17.2
NLEIC	15.1	13.9	15.4	17.7	14.5	15.1	15.9	21.4	15.5	15.3	15.6	17.2

Table 3: *Average MAPE for the monthly M3 competition data*

Methods	Estimated Weights	Data Types		
		Annual	Quarterly	Monthly
LEIC	k	3.08	2.59	3.94
NLEIC	k_{q_2}	1.92	1.75	1.03
	k_{q_3}	4.41	3.62	3.47
	k_{q_4}	1.42	0.69	1.75

Table 4: *Estimated weights for the M3 competition data*

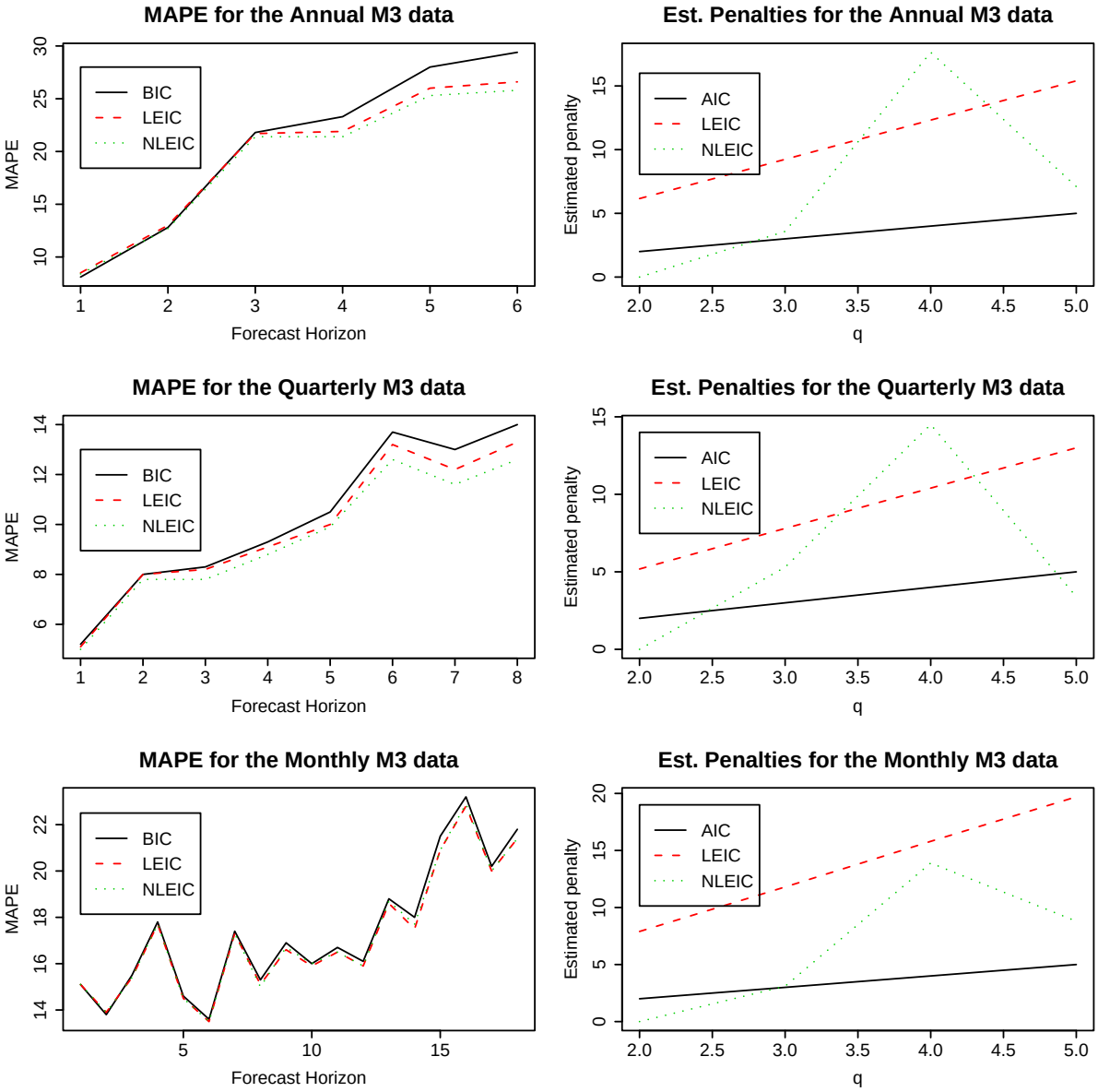

Figure 2: *MAPE and estimated penalty functions for LEIC and NLEIC.*

4 Example 2: Bootstrap EIC applied to simulated data

To test the procedure on simulated data, we generated 500 series from the AR(2) model, $y_t = 1.2y_{t-1} - 0.5y_{t-2} + e_t$, for sample sizes $n = 20$, 30 and 50. The first $n^* = n - 6$ observations were used for estimating the candidate models AR(1), AR(2) and AR(3) and the last $H = 6$ observations were used for computing the RMSE. For each series the penalty values were estimated using the bootstrap LEIC and NLEIC. The forecast RMSE for the models selected by AIC, LEIC and NLEIC are calculated and presented in Table 5. The penalty weights for these simulations are given in Table 6 and plotted in Figure 3; these demonstrate that the LEIC penalty is very close to AIC for this problem. The results in Tables 5 show that for larger n the NLEIC does substantially better than either AIC or LEIC. For $n = 20$ there is little difference between the methods.

Again, the nonlinear penalty functions (Figure 3) are very similar for all sample sizes. This suggests that the penalty functions are determined by the nature of the data (in this case AR(2)), which supports our general philosophy of allowing the entire ensemble of similar data to determine the nature of the penalty function.

	Method	**Forecasting Horizons**						**Average**	
		1	2	3	4	5	6	1 to 4	1 to 6
	NLEIC	1.17	1.95	2.31	2.26	2.24	2.23	1.92	2.03
$n = 20$	LEIC	1.11	1.85	2.23	2.35	2.37	2.28	1.88	2.03
	AIC	1.13	1.87	2.24	2.35	2.38	2.30	1.90	2.04
	NLEIC	1.09	1.72	2.10	2.16	2.09	2.03	1.77	1.87
$n = 30$	LEIC	1.16	1.82	2.07	2.15	2.19	2.20	1.80	1.93
	AIC	1.18	1.82	2.09	2.17	2.23	2.23	1.82	1.95
	NLEIC	1.03	1.65	1.95	2.08	2.15	2.18	1.68	1.84
$n = 50$	LEIC	1.11	1.74	2.06	2.26	2.25	2.07	1.79	1.92
	AIC	1.10	1.73	2.06	2.26	2.23	2.05	1.79	1.90

Table 5: *Average RMSE for the simulated data*

Methods	Estimated Weights	Sample size n		
		20	30	50
LEIC	k	0.851	0.915	1.096
NLEIC	k_{q_2}	-1.593	-1.514	-1.467
	k_{q_3}	-0.833	-0.698	-0.632

Table 6: *Estimated average weights for the simulated data*

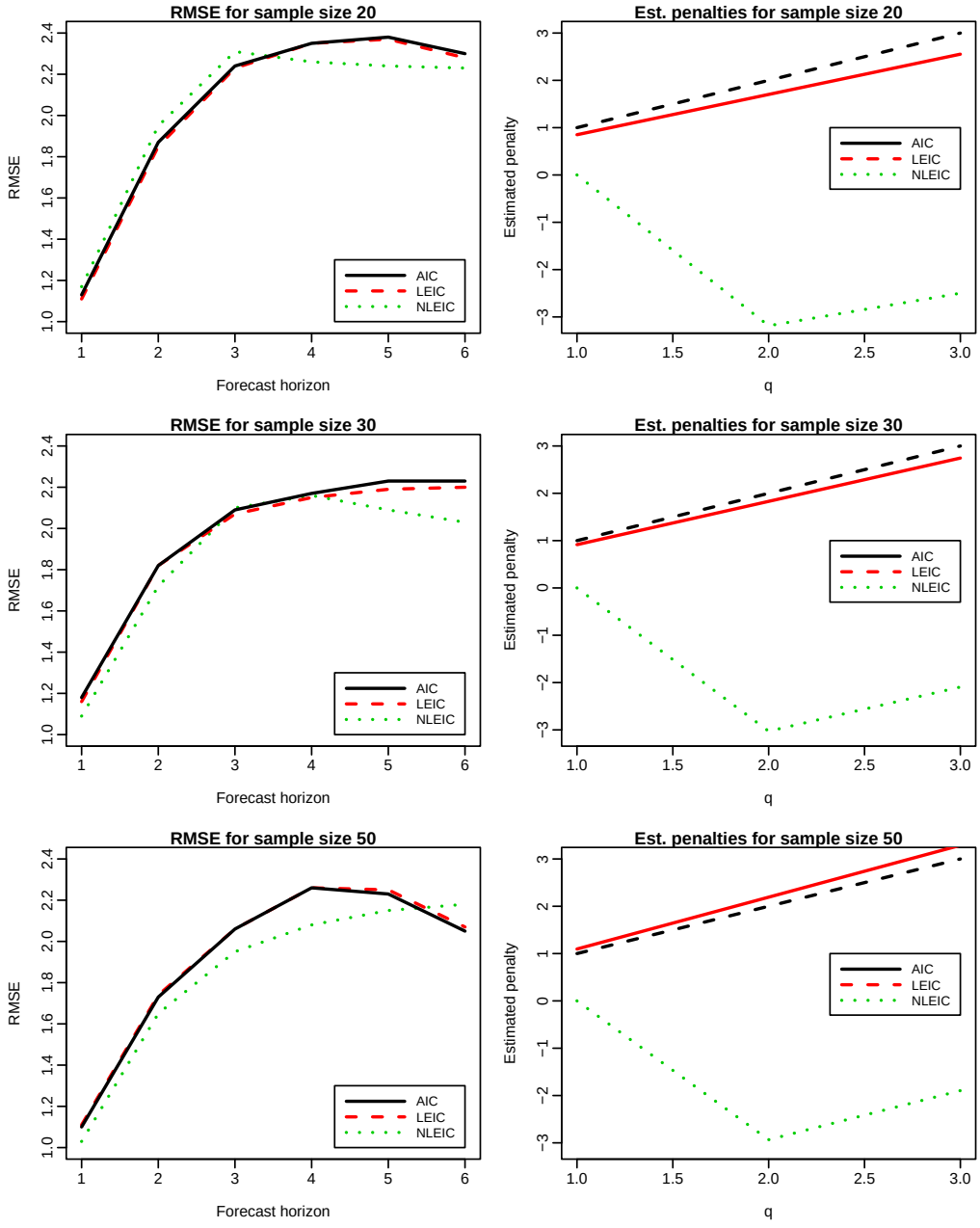

Figure 3: *RMSE and estimated penalty functions for AIC, LEIC and NLEIC for the simulated data.*

5 Conclusions

We have proposed an automatic forecasting model selection algorithm when a large number of series need to be forecast on a routine basis. The methodology is based on a penalized likelihood criterion and is data-adaptive in the sense that the penalty function is determined by the data to be forecast. Thus, the penalty level is tuned to the attributes of the series being forecast. We have proposed a linear and a non-linear version of EIC; both were shown to perform better than all six standard existing information criteria on real and simulated data. The non-linear EIC in particular gives substantial improvement in forecast accuracy over other comparable methods.

References

AKAIKE, H. (1970) Statistical predictor identification, *Annals of Institute of Statistical Mathematics*, **22**, 203–217.

AKAIKE, H. (1973) Information theory and an extension of the maximum likelihood principle, in B.N. Petrov and F. Csaki (eds.), *Second International Symposium on Information Theory*, Akademiai Kiado: Budapest, 267–281.

ASSIMAKOPOULOS, V. and NIKOLOPOULOS, K. (2000) The theta model: a decomposition approach to forecasting. *International Journal of Forecasting* **16**, 521–530.

BILLAH, M.B., KING, M.L., KOEHLER, A.B. and SNYDER, R.D. (2001) Exponential smoothing model selection for forecasting, Working paper, Department of Econometrics and Business Statistics, Monash University, Australia.

BROWN, R.G. (1959) *Statistical Forecasting for Inventory Control*, McGraw Hill: New York.

BROWN, R.G. (1963) *Smoothing, Forecasting and Prediction of Discrete Time Series*, Prentice-Hall: Englewood Cliffs.

BROWN, R.G. (1967) *Decision Rules for Inventory Management*, Holt, Rinehart and Winston: New York.

CHEN, C., DAVIS, R.A., BROCKWELL, P.J. and BAI, Z.D. (1993) Order determination for autoregressive processes using resampling methods, *Statistics Sinica*, **3**, 481–500.

FILDES, R. (1989) Evaluation of aggregate and individual forecast method selection rules, *Man-*

agement Science, **35**, 1056–1065.

FILDES, R., HIBON, M., MAKRIDAKIS, S. and MEADE, N. (1998) Generalizing about univariate forecasting methods: further empirical evidence, *International Journal of Forecasting*, **14**, 339–358.

GARDNER, E.S. (1985) Exponential smoothing: the state of the art, *Journal of Forecasting*, **4**, 1–28.

GOFFE, W.L., FERRIER, G.D. and ROGERS, J. (1994) Global optimization of statistical functions with simulated annealing, *Journal of Econometrics*, **60**, 65–99.

GOLUB, G.H., HEATH, M. and WAHBA, G. (1979) Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics*, **21**, 215–223.

GRUNWALD, G.K. and HYNDMAN, R.J. (1998) Smoothing non-Gaussian time series with autoregressive structure, *Computational Statistics & Data Analysis*, **28**, 171–191.

HANNAN, E.J. and QUINN, B.G. (1979) The determination of the order of an autoregression, *Journal of the Royal Statistical Society, Series B*, **41**, 190–195.

HYNDMAN, R.J. and BILLAH, M.B (2003) Unmasking the Theta Method, *International Journal of Forecasting*, to appear.

HYNDMAN, R.J., KOEHLER, A.B., SNYDER, R.D. and GROSE, S. (2002) A state space framework for automatic forecasting using exponential smoothing methods, *International Journal of Forecasting*, **18**(3), 439–454.

HURVICH, C.M. and TSAI, C.L. (1991) Bias of the corrected AIC criterion for underfitted regression and time series models, *Biometrika*, **78**, 499–509.

MAKRIDAKIS, S., ANDERSEN, A., CARBONE, R., FILDES, R., HIBON, M., LEWANDOWSKI, R., NEWTON, J., PARZEN, E. and WINKLER, R. (1982) The accuracy of extrapolation (time series) methods: results of a forecasting competition, *Journal of Forecasting*, **1**, 111–153.

MAKRIDAKIS, S., CHATFIELD, C., HIBON, M., MILLS, T., ORD, J.K. and SIMMONS, L.F. (1993) The M2-Competition: a real-time judgmentally based forecasting study, *International Journal of Forecasting*, **9**, 5–22.

MAKRIDAKIS, S. and HIBON, M. (1991) Exponential smoothing: the effect of initial values and loss functions on post-sample forecasting accuracy, *International Journal of Forecasting*, **7**, 317–330.

MAKRIDAKIS, S. and HIBON, M. (2000) The M3-Competitions: results, conclusions and impli-

cations, *International Journal of Forecasting*, **16**, 451–476.

MAKRIDAKIS, S. and WHEELWRIGHT, S.C. (1989) *Forecasting Methods for Management*, 5th edition, John Wiley and Sons: New York.

MAKRIDAKIS, S. WHEELWRIGHT, S.C. and HYNDMAN, R.J. (1998) *Forecasting Methods and Applications*, 3rd edition, John Wiley & Sons: New York.

MALLOWS, C.L. (1964) Choosing variables in a linear regression: a graphical aid, presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas.

MILLS, J.A. and PRASAD, K. (1992) A comparison of model selection criteria, *Econometric Reviews*, **11**, 201–233.

ORD, J.K., KOEHLER, A.B. and SNYDER, R.D. (1997) Estimation and prediction for a class of dynamic nonlinear statistical models, *Journal of the American Statistical Association*, **92**, 1621–1629.

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

WINTERS, P.R. (1960) Forecasting sales by exponentially weighted moving averages, *Management Science*, **6**, 324–342.