The Selection of Time Series Models Perhaps with Regressors

Marco Riani*
Dipartimento di Economia, Università di Parma, Italy
and
Anthony C. Atkinson†, Department of Statistics,
London School of Economics, London WC2A 2AE, UK

April 27, 2010

Abstract

We develop a C_p statistic with a known distribution for the selection of models for stationary and non-stationary time series, including ARIMA and structural models, that may include regressors. To provide a unified framework we use the state-space approach which readily incorporates explanatory variables. We exemplify use of the statistic through generalized ARMA modelling of a series of UK one-day ahead electricity prices that includes four explanatory variables. A plot of our time series C_p statistic is highly informative about the choice of model.

Keywords: AIC; C_p plot; electricity prices; F distribution; Kalman filter; state-space formulation

1 Introduction

In this paper we derive a statistic with known distribution for the choice of a time series model with, or without, regressor variables. Our example is an ARMA model with four explanatory variables

There is a vast literature on methods for selection of non-nested models. In AIC (Akaike, 1974) the maximized log-likelihood is penalized by the number of

*e-mail: mriani@unipr.it
†e-mail: a.c.atkinson@lse.ac.uk.

parameters in the model. For Gaussian regression models AIC becomes Mallows' C_p (Mallows, 1973) when the nuisance parameter σ^2 is estimated from a full model containing sufficiently many terms to provide an unbiased estimate. For Gaussian regression models the distribution of C_p is the linear function of an F-distribution given in (8). There is no simple non-asymptotic expression for the distribution of AIC.

Book length treatments of the properties and applications of these and other procedures include McQuarrie and Tsai (1998) and, more recently, Konishi and Kitagawa (2008) and Claeskens and Hjort (2008). Tong (2001, §9) gives references to methods solely for time series. These include the use of C_p for time series with autoregressive components. However, in no case is there any extension of the standard results to the class of model of interest here, that of Gaussian state-space models with added regressor variables.

In §2 we briefly review AIC, particularly for regression models, including its close relationship with C_p and give distributional results for C_p. Our main theoretical result is in §3 where we use the state-space representation to extend AIC and C_p to stationary and non-stationary time series. We prove a theorem giving the distribution of our new C_p statistic. The analysis of simulated data in §4 reveals the structure of the time series C_p plot. Our analysis of electricity pricing data in §5 shows how this structure leads to a clear and well-informed choice between models. We conclude with a short discussion of numerical and statistical matters. The proof of our theorem is in the Appendix.

One important aspect of our results is that we provide a criterion with known distribution function for the choice of a time series model. A second is that we extend standard criteria to the choice of time series models with regressors. A numerical consequence is that, through use of the Kalman filter, we provide a novel and simple way of calculating both AIC and C_p for time series data. Our analysis of data on electricity pricing shows how informative our procedure can be.

2 AIC and C_p in Regression

The loglikelihood of n observations y, a function of the $u \times 1$ vector of parameters β is $L(\beta; y)$. If $\hat{\beta}$ is the maximum likelihood estimator of β, AIC is defined as

$$AIC = -2L(\hat{\beta}; y) + 2u. \tag{1}$$

That model is selected for which AIC is a minimum. It would be natural to use p as the number of parameters, but this is a paper about the analysis of time series and our notation is intended to allow for the discussion of general ARMA(p, q) models with regressors.

We first consider regression without a time-series structure. For the linear multiple regression model $y = X\beta + \epsilon$, X is an $n \times d$ full-rank matrix of known constants, with ith row x_i^T. The normal theory assumptions are that the errors ϵ_i are i.i.d. $N(0, \sigma^2)$. The residual sum of squares from fitting this model to the data is R_d and, for known σ^2,

$$AIC_\sigma = n\log(2\pi) + n\log\sigma^2 + R_d/\sigma^2 + 2d. \tag{2}$$

If, as is usually the case, σ^2 is not known, the maximum likelihood estimator is

$$\hat{\sigma}^2 = R_d/n. \tag{3}$$

With this internal estimate of σ^2 the criterion (2) becomes

$$AIC_I = n\log(2\pi) + n\log\{R_d/n\} + n + 2d, \tag{4}$$

a form often used in the selection of non-nested time series models with normally distributed errors (Tong, 2001, §9).

When σ^2 is estimated, the total number of parameters in the model is $d + 1$. However, it is self-evident that the ranking of models by (2) or (4) is unaffected by addition of the same constant to the value of the statistic for each model.

In the selection of regression variables σ^2 is estimated from a large regression model with $n \times d^+$ matrix X^+, $d^+ > d$, of which X is submatrix. The unbiased estimator of σ^2 comes from regression on all d^+ columns of X^+ and can be written

$$s^2 = R_{d^+}/(n - d^+). \tag{5}$$

With this estimate the criterion (2) is

$$AIC = n\log(2\pi) + n\log\{R_{d^+}/(n - d^+)\} + (n - d^+)R_d/R_{d^+} + 2d. \tag{6}$$

Since both n and s^2 are fixed, the variable factors are the value of d and the regressors that are being considered. Then choice of the model minimizing (6) is identical to the choice of model minimizing

$$C_p = R_d/s^2 - n + 2d = (n - d^+)R_d/R_{d^+} - n + 2d. \tag{7}$$

One interpretation of C_p (Mallows, 1973) is that it provides an estimate of the scaled mean squared error of prediction at the n observational points from the model of interest, provided the full model with d^+ parameters yields an unbiased estimate of σ^2. Then $\mathrm{E}\{R_d\} = (n-d)\sigma^2$, $\mathrm{E}(s^2) = \sigma^2$ and $\mathrm{E}(C_p)$ is approximately d.

As we illustrate in §5 it helps not merely to select models with small values of C_p but also to calibrate those values against their distribution. The distribution of

C_p is given, for example, by Mallows (1973) and by Gilmour (1996). From (7) we require the distribution of the ratio of two nested residual sums of squares. It is straightforward to show that the required distribution is

$$C_p \sim (d^+ - d)F + 2d - d^+, \qquad \text{where} \qquad F \sim F_{d^+-d,n-d^+}. \tag{8}$$

In short, if

$$F^* \sim F_{\nu_1,\nu_2}, \qquad \mathrm{E}\,(F^*) = \nu_2/(\nu_2 - 2). \tag{9}$$

Then, from (8),

$$\mathrm{E}\,(C_p) = d + 2\frac{d^+ - d}{n - d^+ - 2}. \tag{10}$$

As $n \to \infty$, $\mathrm{E}\,(C_p) \to d$. Gilmour comments that when $n - d^+$ is small, E(C_p) can be appreciably greater than d. Hurvich and Tsai (1989) use Taylor series expansions to find a correction for this bias for very small n.

3 Model Selection in Stationary and Non-stationary Time Series

The state-space form provides a unified representation of a wide range of linear Gaussian time series models including ARIMA models, time-varying regression models, dynamic linear models and unobserved components time series models (Anderson and Moore (1979), Harvey (1989), Durbin and Koopman (2001)). An advantage, in the present context, is that it is relatively straightforward to extend the preceding results on C_p and its distribution to time series with a Gaussian structure. The Gaussian state-space formulation consists of a transition equation and a measurement equation which can be formalized as:

$$\alpha_{t+1} = d_t + T_t\alpha_t + H_t\epsilon_t \qquad \alpha_1 \sim N(a, P) \tag{11}$$

$$y_t = c_t + Z_t\alpha_t + G_t\epsilon_t \qquad \epsilon_t \sim NID(0, I), \tag{12}$$

where $NID(\mu, \Sigma)$ denotes an independent sequence of normally distributed random vectors with mean μ and variance covariance matrix Σ. The Markovian structure of the transition equation is an effective way of describing the serial correlation structure of the multivariate time series y_t. Although these equations are written in a general form, in our application we consider only univariate series. The measurement equation relates the time series to a signal $(c_t + Z_t\alpha_t)$ and a vector of disturbances ϵ_t. The deterministic matrices T_t, Z_t, H_t and G_t are customarily referred to as system matrices and they are usually sparse. The vectors d_t and c_t are fixed and can be used to incorporate known patterns into the model, otherwise they are always zero. As is well known, the Kalman filter is a recursive

algorithm for the evaluation of moments of the normal distribution of the state vector α_{t+1} conditional on the data set $Y_t = \{y_1, \ldots, y_t\}$, that is

$$a_{t+1} = E(\alpha_{t+1}|Y_t) \qquad P_{t+1} = \text{cov}\,(\alpha_{t+1}|Y_t)$$

for $t = 1, 2, \ldots, T$. More precisely, the Kalman filter is given by the following set of equations:

$$\begin{aligned} v_t &= y_t - c_t - Z_t a_t \\ F_t &= Z_t P_t Z_t' + G_t G_t' \\ K_t &= (T_t P_t Z_t' + H_t G_t') F_t^{-1} \\ a_{t+1} &= d_t + T_t a_t + K_t v_t \\ P_{t+1} &= T_t P_t T_t' + H_t H_t' - K_t F_t K_t'. \end{aligned}$$

Since we focus our attention on univariate time series, the matrix F_t becomes a scalar and will be written as f_t.

The generalization to stationary time series, without regression, of equation (2) is

$$AIC_\sigma = T\log(2\pi) + \sum_{t=1}^{T} \log f_t + \sum_{t=1}^{T} \frac{v_t^2}{f_t} + 2(p+q), \tag{13}$$

where $p+q = u$ denotes the number of parameters of the current model (excluding the scale parameter in the selection matrices G_t and H_t). For example, with an ARMA(1,1) $u = 2$.

If we concentrate the scale parameter out of the likelihood and use the superscript ^{c} to denote the scaled version of the measurement equation (12)

$$y_t = Z_t \alpha_t + G_t^c \epsilon_t^c, \qquad \epsilon_t^c \sim N(0, \sigma_*^2) \qquad \sigma_*^2 > 0,$$

with unknown variance σ_*^2. The state space form (11) and (12) still applies with $G_t = \sigma_* G_t^c$ and $H_t = \sigma_* H_t^c$. Equation (13) becomes

$$AIC_\sigma = T\log(2\pi) + T\log\sigma_*^2 + \sum_{t=1}^{T} \log f_t + \sum_{t=1}^{T} \frac{v_t^2/f_t}{\sigma_*^2} + 2(p+q). \tag{14}$$

The maximum likelihood estimator of σ_*^2 is

$$\hat{\sigma}_*^2 = \frac{1}{T} \sum_{t=1}^{T} \frac{v_t^2}{f_t^c}. \tag{15}$$

Hence, the generalization of equation (4) is

$$AIC_I = T\left\{\log(2\pi) + 1\right\} + T\log\hat{\sigma}_*^2 + \sum_{t=1}^{T} \log f_t^c + 2(p+q). \tag{16}$$

So far we have assumed a model for stationary time series and no regressor effects so we have implicity assumed that $a = 0$ and P is of known form. If we let d denote the number of elements in the state vector which have a diffuse initial distribution, the estimate of the scale factor becomes:

$$\hat{\sigma}_*^2 = \frac{1}{T-d} \sum_{t=d+1}^{T} \frac{v_t^2}{f_t^c}. \tag{17}$$

Usually (for example, Durbin and Koopman (2001)) d is the number of non-stationary elements plus the number of fixed regression effects in the state vector. However, when we consider stationary models, d retains its interpretation from §2.

With d diffuse elements equation (14) becomes

$$AIC_\sigma = T\log(2\pi) + (T-d)\log\sigma_*^2 + \sum_{t=d+1}^{T} \log f_t^c + \sum_{t=d+1}^{T} \frac{v_t^2/f_t^c}{\sigma_*^2} + 2u, \tag{18}$$

where $u = d+p+q$. For example, if to the ARMA(1,1) model we add time varying monthly trigonometric seasonality, with the same variance σ_ω^2 for each harmonic, the number of non-stationary elements d in the state vector is 11. However, the total number of extra parameters is 12, since we also need to count σ_ω^2. With $p = q = 1$ we obtain $u = 14$. Similarly, if we add two stochastic time varying explanatory variables to the ARMA(1,1) model the number of extra parameters is 4 and $u = 6$. If the two explanatory variables were to be deterministic, then $u = 4$.

As in regression, in order to compute AIC the estimate of σ_*^2 which is used is the one based on the current model which is being fitted. In this case, if we use the estimate in (17) AIC becomes:

$$AIC_I = T\log(2\pi) + (T-d)\left(\log\hat{\sigma}_*^2 + 1\right) + \sum_{t=d+1}^{T} \log f_t^c + 2u. \tag{19}$$

In addition, in the state space literature u is replaced by $u+1$, that is by the number of all the parameters in the model (including the one which can be concentrated out of the likelihood). See, for example, Commandeur and Koopman (2007).

A natural extension of equation (6) to time series is

$$C_p^T = T\log(2\pi) + (T-d)\log\hat{\sigma}_{u^+}^2 + \sum_{t=d+1}^{T} \log f_t^{c^+} + \frac{\sum_{t=d+1}^{T} v_t^2/f_t^c}{\hat{\sigma}_{u^+}^2} + 2u \tag{20}$$

where

$$\hat{\sigma}_{u^+}^2 = \frac{\sum_{t=d^++1}^{T} v_t^{2^+}/f_t^{c^+}}{(T-d^+)},$$

is the estimate of σ_*^2 based on the full model. The total number of parameters in the full model is written as $u^+ = p^+ + q^+ + d^+$, that is excluding σ^2, and the summation from $d^+ + 1$ is for f^+ and v^+ from the full model. We require that the model with u parameters is nested within that with u^+. We then not only require that $u^+ > u$, but also the conditions on the component parameters that $d^+ \geq d, p^+ \geq p$ and $q^+ \geq q$.

The choice of the model minimizing (20) is identical to the choice of model minimizing

$$C_p^T = \sum_{t=d+1}^{T} \log f_t^c + \frac{\sum_{t=d+1}^{T} v_t^2/f_t}{\hat{\sigma}_{u^+}^2} - \sum_{t=d^++1}^{T} \log f_t^{c^+} - T + 2u. \qquad (21)$$

The one-step-ahead prediction variances f_t^c and $f_t^{c^+}$ rapidly tend to one as t increases and their logarithms thus tend to zero. For regression models f_t and f_t^+ depend on submatrices of the matrices of explanatory variables X_T and X_T^+, but not on the observations y_t. More precisely, in regression $f_t^c = 1 + x_t'(X_{t-1}'X_{t-1})^{-1}x_t$, where X_{t-1} is the matrix which contains the values of the explanatory variables up to time $t-1$. Their effect on C_p^T is to add a constant to the right hand side of (21). Although this constant will depend on the model being fitted, it is small. For time series models f_t^c and $f_t^{c^+}$ depend weakly on the data through the estimated parameters contained in the system matrices of the transition (11) and measurement equations (12), but again the values of the logarithms are small and decrease rapidly with t. For example, for MA(1) with $|\theta| < 1$, convergence to the steady state value of one is exponentially fast (Anderson and Moore (1979, pp. 79–80)). To obtain a statistic with known distribution we ignore the sums of these terms and take

$$C_p = \frac{\sum_{t=d+1}^{T} v_t^2/f_t}{\hat{\sigma}_{u^+}^2} - T + 2u. \qquad (22)$$

For the full model with u^+ parameters, $C_p = 2u^+ - d^+$.

Distributional results about C_p assume that the full model with u^+ parameters yields an unbiased estimate of σ_*^2 and that all the extra parameters not contained in the reduced model are unnecessary. In the appendix we show that

$$C_p \sim (u^+ - u)F + 2u - d^+ \qquad \text{where} \qquad F = F_{u^+-u,T-d^+}. \qquad (23)$$

Methods parallel to those of §2 show that

$$E(C_p) = u + \frac{(u^+ - d^+)(T - d^+)}{T - d^+ - 2} + \frac{2(d^+ - u)}{T - d^+ - 2}.$$

As $T \to \infty$ we obtain

$$E(C_p) = u + (u^+ - d^+) = u + p^+ + q^+.$$

Thus the expected value of the statistic, for large T, depends on the total number of parameters in the reduced model and on the number of stochastic parameters in the full model. This however is a constant when comparing different reduced models, so that the penalty in comparisons is just u, as it is d for regression models. In neither case does the parameter for the error variance, which is concentrated out in the time series application, affect the distribution of the statistic.

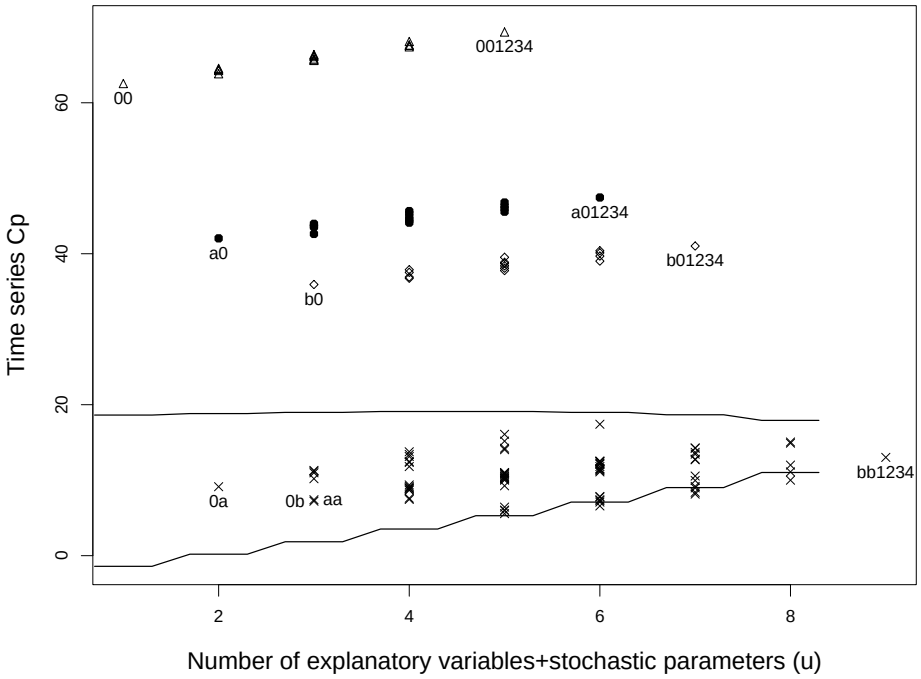

Figure 1: Time series C_p plot for simulated MA(1) process ($0\,a$) with $\theta = 0.9$. Large model $b\,b\,1\,2\,3\,4$ plus a constant (u^+ = 9). $\times$ MA and ARMA models with explanatory variables; $\diamondsuit$ AR(2) models with explanatory variables; $\bullet$ AR(1) models with explanatory variables; $\triangle$ regression models. See Table 1 for notation. Bands 1% and 99% points of (23). The importance of selecting the correct stochastic model is evident

4 The Structure of Time Series C_p Plots: an example with simulated data

We first look at the structure of C_p time series plots for a simulated example. In these plots we label the models with a notation of the form "$p\,q\,i_1\,i_2...$", where p and q denote the order of the autoregressive and moving average models and the i_j denote those regression variables that are included in the model. Further,

Table 1: Notation used in the figures for ARMA models with regressors

Notation	Model and Regressors
$a0$	AR(1)
$0a$	MA(1)
$0b$	MA(2)
$aa34$	ARMA(1,1) x_3 x_4

we denote the increasing values of p and q as $0, a, b$ etc. Some examples are in Table 1.

We simulated 100 observations from an MA(1), that is 0 a, with $\theta = 0.9$, that included a constant, equal to 5, and four explanatory variables that were unrelated to the time series. The explanatory variables were independently distributed $\mathcal{N}(0, 1)$ and σ^2 = 1. In our calculation of C_p the maximum model was $b\,b\,1\,2\,3\,4$, all models containing a constant. So the maximum number of parameters $u^+ = 9$. Figure 1 shows the resulting plot of C_p for all models containing at least two parameters.

The times series C_p plot shows four bands of values corresponding to different families of stochastic models. The family with smallest C_p values, marked with crosses in the figure, falls within the band of the 1% and 99% points of the F distribution (23). The simplest model is the MA(1) without explanatory variables, that is 0 a, from which the data were simulated. Reading upwards, the next two models in this band, with three parameters, are $a\,a$ and 0 b, with C_p values around two lower than their special case 0 a. The remaining four models, with higher C_p values, are 0 $a\,i$, i = 1,...,4; that is MA(1) models including one of the explanatory variables. The models with more parameters in this group ($u \geq 4$) are all at least ARMA(1,1) or MA(2) with explanatory variables. For the maximum model with $u^+ = 9$, the value of C_p is 13, agreeing with the special case of (22).

The second series of C_p values in the plot, shown by diamonds, are for AR(2) models including explanatory variables. The next band is for AR(1) models also including such variables. The highest band of all, the triangles, is for pure regression models without any time series component.

An exciting feature of this plot is that the bands sort the models into clear groups with differing stochastic structure. It is clear from the figure that we need at least an MA(1) model and that the improvements from including explanatory variables are negligible. In the provision of this information the time series C_p plot is very different from the C_p plot for regression, for example Figure 1 of Atkinson

and Riani (2008), in which the form is that of the series of values for one of the sets of models with the same stochastic structure in Figure 1.

5 The Day–Ahead Price of Electricity

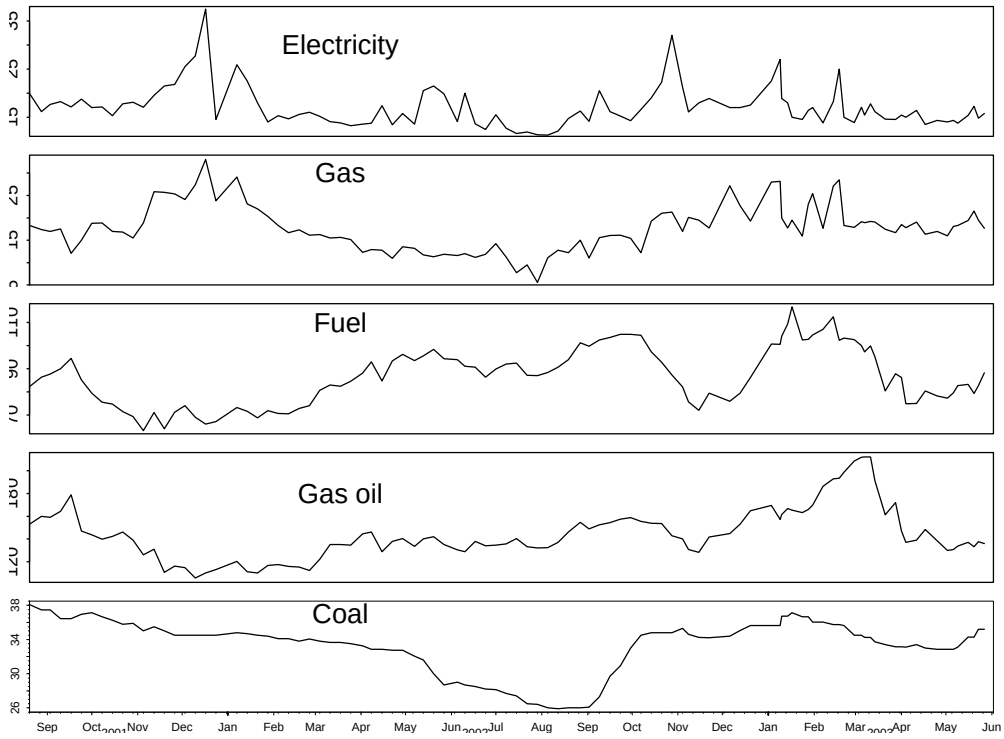

Figure 2: Electricity data. Time series plots of the price of electricity y and the four fuel price series used as explanatory variables x_1 to x_4. All models with low C_p combine the inclusion of x_1 with some stochastic structure

A consequence of the dismantling of centralised electricity generating and supply monopolies by privatisation has been the development of a variety of markets in electricity. The first day-ahead electricity market was established in England and Wales in 1990. The history of this process in those countries, followed by Scotland, is outlined by Weron (2006, pp. 9–11). A characteristic of the time series of such prices is that they combine a stochastic process with explanatory variables. A complicated example is given by Koopman et al. (2007).

We analyse one hundred observations on the UK day-ahead baseload power price. To avoid the difference in price structure for different days (see §6) we only consider data for Mondays from the 20th of August 2001 to the 26th of May 2003. As explanatory variables we take the prices of four fuels that can be used in the

generation of electricity. We can also expect that there will be correlation between adjacent prices that will necessitate time series modelling. The data are available from Datastream. The labels of the variables in Figure 2, their definitions and codes are:

y Electricity.
UK day-ahead baseload power price, £/MWh (AADET00)

x_1 Gas.
UK day-ahead gas price for NBP, pence/therm (NGAAA00)

x_2 Fuel.
NWE barges heavy fuel oil 3.5% FOBR, £/mt (PUABC00)

x_3 Gas oil.
NWE barges gas oil 0.2% FOBR, £/mt (POAAG00)

x_4 Coal.
Coal CIF ARA 90-day, £/mt (CSABG00).

MWh: megawatt hour. NBP: National Balancing Point (UK). NWE: North West Europe. FOBR: FOB Rotterdam. mt: metric tonne. ARA: Amsterdam Rotterdam Antwerp. For further information see `www.platts.com`.

Plots of the five time series of these variables are given in Figure 2. The values of the response show appreciable correlated variability. Of the explanatory variables the price of gas shows the greatest variability over the time period and the price of coal the least.

We again calculated our time series C_p statistic from 100 observations with the maximum model $b\,b\,1\,2\,3\,4$, all models containing a constant. The resulting time series C_p plot is in Figure 3. This leads to banding of the type we have seen before and a clear identification of a suitable model.

The lowest band in the figure, plotted as dots, is for models with some time series structure that also include x_1. With three parameters the model with smallest C_p is for $a\,0\,1$, the AR(1) model including regression on x_1. The other good model in this class is $0\,a\,1$, that is MA(1) with the same regression. For four parameters the best model is the combination of these two, the ARMA(1,1) with x_1. However, there is no evidence that this model is to be preferred as moving from $a\,0\,1$ to $a\,a\,1$ causes a 0.3 increase in the value of C_p. The next band of symbols, the diamonds, lie around the upper boundary of the F distribution. These are models with no time series structure that include regression on x_1. Above these, the filled triangles are models with time series structure that exclude x_1. Finally, the worst set of models, represented by crosses, are regressions excluding x_1.

The C_p value for $a\,0\,1$ is 4.22, against 8.04 for $0\,a\,1$. The clear conclusion is that, within this class of models, an AR(1) model with regression on x_1 is appropriate. We tried extending the class by including lag-one versions of all

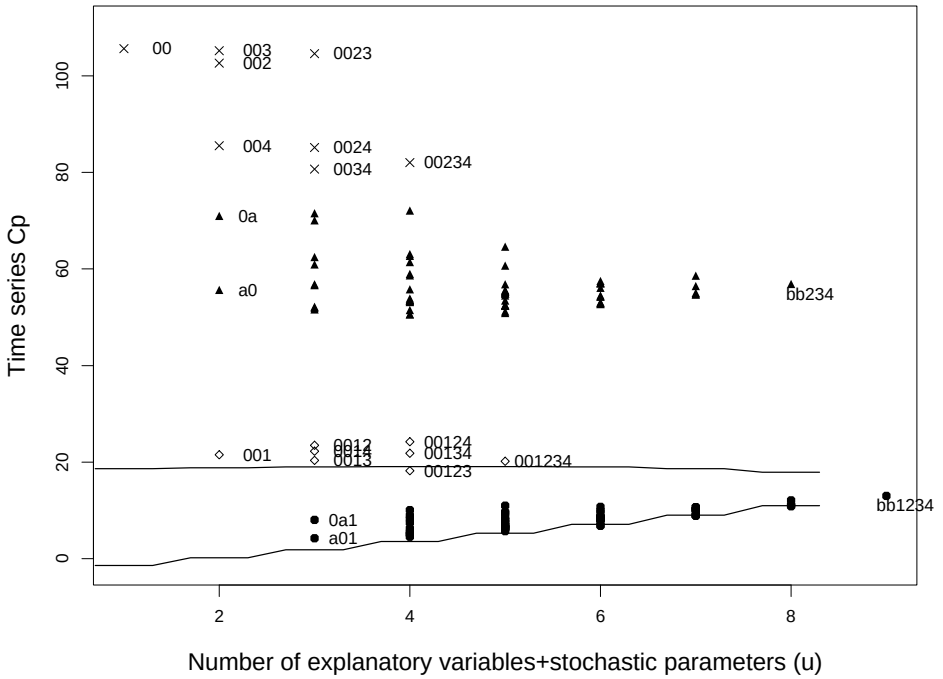

Figure 3: Time series C_p plot for electricity data with large model $b\,b\,1\,2\,3\,4$ plus a constant (u^+ = 9). • AR, MA and ARMA models including x_1; $\diamondsuit$ regression models including x_1; ▲ AR, MA and ARMA models without x_1; $\times$ regression models without x_1. See Table 1 for notation. Bands 1% and 99% points of (23).

four explanatory variables. These did not lead to any further models with small C_p values.

Post facto it is not perhaps surprising that the price of gas is the only important explanatory variable. In 2001, that is at the beginning of our period, gas accounted for approximately 37% of electricity production in the UK, with coal supplying about 1/3. Gas is burnt at the point of electricity consumption, so that the two one-day ahead prices can be expected to be related, whereas coal may be stored for several months before use. The irrelevance of lagged variables is also explained by the day-ahead nature of the two series. However, what is not at all obvious without an analysis of the kind presented here, is what stochastic model should be combined with the regression structure.

6 Discussion

In our calculations we used the latest version of the library SSFPACK (Koopman et al., 2008) in conjunction with the Ox programming language of Doornik (2001). The likelihoods were maximized with Doornik's routine maxBFGS.

Our analysis of the electricity price data was guided by the limits of the F distribution (23) superimposed on Figure 3. We have performed extensive simulations to check this distribution and found good agreement between theory and empirical results for a variety of AR models. The only problems that arose were when the parameters of the reduced model were such that the time series was very close to non-stationarity. When we performed simulations for the comparison of MA models, we again usually obtained numerical results similar to those for AR models. However, occasionally we obtained distributions with very long tails.

It is well-known, for example Shephard (1993), that maximum likelihood estimates of the parameters in MA models can have extremely long-tailed distributions when the model is close to non-invertibility. Our experience with MA models has been that the starting points for numerical maximization need to be chosen with some care, even when the models to be fitted are invertible, to avoid iterations of the numerical algorithm straying close to regions of non-invertibility. We stress that these numerical problems arise in the repeated estimation of parameters in large simulations. We had no numerical problems with the relatively few maximizations needed in the analysis of the electricity price data.

There is an increasing literature on the analysis of high-frequency electricity data. For example, Koopman et al. (2007) model daily electricity spot prices in four markets and Dordonnat et al. (2008) model hourly electricity load in France. Such representative works differ from ours in that, by working with one day a week readings, we avoid the changes in stochastic volatility that are a feature of rapidly sampled data. In both papers model building is by considering the properties of estimated coefficients. In regression this reduces to examination of t statistics for individual explanatory variables. In the presence of many variables (Dordonnat et al. (2008) have 27 for the structure of their time series) this procedure can be fraught, since eliminating a variable correlated with others can cause large changes in the remaining t statistics. Our analysis shows the clarity obtained from use of a model selection criterion with known distributional properties.

The procedure we have developed applies not only to ARIMA models but in general to structural time series models, such as the choice of local level, local linear trend or basic structural model (plus stochastic cycle). It has also not escaped our attention that C_p is an aggregate statistic, based on all the data. For regression, Atkinson and Riani (2008) use the forward search (Hadi (1992); Atkinson and Riani (2000); Atkinson et al. (2004)) to determine how the choice of a regression model using C_p is affected by groups of observations. Although the numerical procedure is more complicated, related methods could be applied to our C_p statistics for time series to illuminate the dependence of model choice on individual observations, breaks in structure and on anomalous patches of observations in the time series. Indeed, analysis of more recent periods of the electricity price data introduces just such features.

Finally, in our opinion, the mechanical use of C_p is to be avoided. Any model selected by use of C_p should be subject to customary statistical checks, such as tests of the significance of the terms in the model.

Acknowledgements.

We are grateful to Prof. Luigi Grossi of the University of Verona for introducing us to the challenge of analysing electricity price data and for his helpful comments and discussion. We are also grateful to Prof. Siem Jan Koopman for making available to us an advance version of SsfPack3.0.

Appendix: distribution of C_p in stationary and non-stationary time series models

To obtain the distribution of C_p in the case where all important regressors, and stochastic parameters are included in the model we assume, without loss of generality, that $\beta_{d+1} = \cdots = \beta_d^+ = 0$, $\phi_{p+1} = \cdots = \phi_{p^+} = 0$ and $\theta_{q+1} = \cdots = \phi_{q^+} = 0$, i.e. that all extra parameters not belonging to the reduced model are unimportant. Then:

$$\begin{aligned} C_p &= \frac{\sum_{t=d+1}^{T} v_t^2/f_t}{\hat{\sigma}_{u+}^2} - T + 2u \\ &= (T-d^+)\frac{\sum_{t=d^++1}^{T} \frac{v_t^{2^+}}{f_t^{c^+}} + SS(d^+-d, p^+-p, q^+-q)}{\sum_{t=d^++1}^{T} \frac{v_t^{2^+}}{f_t^{c^+}}} - T + 2u, \end{aligned}$$

where $SS(d^+-d, p^+-p, q^+-q) = SS(\beta_{d+1}, \ldots, \beta_{d^+}, \phi_{p+1}, \ldots, \phi_{p^+}, \theta_{q+1}, \ldots, \theta_{q^+} | \beta_1, \ldots, \beta_d, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$ is the contribution to the sum of squares of one-step-ahead standardized prediction residuals from fitting the additional terms in the full model to the residuals of the reduced model.

$$C_p = (T-d^+)\left\{1+\frac{SS(d^+-d,p^+-p,q^+-q)}{\sum_{t=d^++1}^{T}\frac{v_t^{2+}}{f_t^{c+}}}\right\}-T+2u \quad (24)$$

$$= (T-d^+)\frac{SS(d^+-d,p^+-p,q^+-q)}{\sum_{t=d^++1}^{T}\frac{v_t^{2+}}{f_t^{c+}}}+2u-d^+ \quad (25)$$

$$= (u^+-u)\frac{U/(u^+-u)}{V/(T-d^+)}+2u-d^+. \quad (26)$$

In regression the distributions of the equivalent sums of squares do not depend on estimation of σ^2. Here the same result holds for estimation of σ_*^2. If we ignore the slight effect of estimation of the variance ratios in (11) on these distributions we can take $U \sim \chi^2_{u^+-u}$ and $V \sim \chi^2_{T-d^+}$ with U and V independent. Hence

$$C_p \sim (u^+-u)F+2u-d^+ \qquad \text{where} \qquad F = F_{u^+-u,T-d^+}.$$

In §6 we mentioned some evidence that this distributional result for AR models holds to a high accuracy, even when T is not very large.

References

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

Anderson, B. D. O. and J. B. Moore (1979). *Optimal Filtering*. Englewood Cliffs, N. J.: Prentice Hall.

Atkinson, A. C. and M. Riani (2000). *Robust Diagnostic Regression Analysis*. New York: Springer–Verlag.

Atkinson, A. C. and M. Riani (2008). A robust and diagnostic information criterion for selecting regression models. *Journal of the Japanese Statistical Society 38*, 3–14.

Atkinson, A. C., M. Riani, and A. Cerioli (2004). *Exploring Multivariate Data with the Forward Search*. New York: Springer–Verlag.

Claeskens, G. and N. L. Hjort (2008). *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.

Commandeur, J. J. F. and S. J. Koopman (2007). *An Introduction to State Space Time Series Analysis*. Oxford: Oxford University Press.

Doornik, J. A. (2001). *Ox 3.0: Object-oriented matrix programming language (4th ed.)*. London: Timberlake Consultants Press.

Dordonnat, V., S. J. Koopman, M. Ooms, A. Dessertaine, and J. Collet (2008). An hourly periodic state space model for modelling French national electricity load. *International Journal of Forecasting 24*, 566–587.

Durbin, J. and S. J. Koopman (2001). *Time Series Analysis by State Space Models*. Oxford: Oxford University Press.

Gilmour, S. G. (1996). The interpretation of Mallows's C_p-statistic. *The Statistician 45*, 49–56.

Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, Series B 54*, 761–771.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Hurvich, C. M. and C.-L. Tsai (1989). Regression and time series model selection in small samples. *Biometrika 76*, 297–307.

Konishi, S. and G. Kitagawa (2008). *Information Criteria and Statistical Modeling*. New York: Springer-Verlag.

Koopman, S. J., M. Ooms, and M. A. Carnero (2007). Periodic seasonal reg-ARFIMA-GARCH models for daily electricity spot prices. *Journal of the American Statistical Association 102*, 16–27.

Koopman, S. J., N. Shephard, and J. A. Doornik (2008). *Statistical Algorithms for Models in State Space Form: SsfPack 3.0*. London: Timberlake Consultants Press.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics 15*, 661–675.

McQuarrie, A. D. R. and C. L. Tsai (1998). *Regression and Time Series Model Selection*. Singapore: World Scientific.

Shephard, N. (1993). Distribution of the ML estimator of a MA(1) and a local level model. *Econometric Theory 9*, 377–401.

Tong, H. (2001). A personal journey through time series in Biometrika. *Biometrika 88*, 195–218.

Weron, R. (2006). *Modeling and Forecasting Electricity Loads and Prices: a statistical approach*. New York: Wiley.