ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

Empirical Bayesian Model Selection for Autoregressive Processes

P. Mariyappan and P. Arumugam

1. Research scholar Manonmaniyam Sunadaranar university,Thirnelveli, Department of statistics, Mahabararhi Engineering College, Vasudevanur.
2. Asso. Professor in statistics, Annamalai university, Annamalai Nager, Chidhambaram.

Manuscript Info

Manuscript History:

Received: 15 July 2013
Final Accepted: 19 July 2013
Published Online: August 2013

Key words:
Bayes Factor, prior distribution, posterior probability, AR model.

Abstract

In this paper, we study the Autoregressive (AR) model of order p for model selection by using the Bayesian approach. A numerical study has been carried out to illustrate the developed model.

1. INTRODUCTION

The AR (p), or autoregressive (AR) model of order p, for a time series $\{x_t\}$ is defined by

$$x_t = \emptyset_1 x_{t-1} + \cdots + \emptyset_p x_{t-p} + \varepsilon_{t,} \qquad (1)$$

Where $\{\varepsilon_1, \varepsilon_2, \ldots\ldots\}$ is a sequence of independent $N(0, \sigma_\varepsilon^2)$ random variables.This model is useful for the forecasting and control of time series, as well as for the estimation of functional such as the spectrum or the amount of energy in a given frequency band.To use this model, a value for p must be specified. Because there is rarely a direct physical motivation for the AR(p) model (1), this choice must be based on the data. There has been much work on ways of making this choice (e.g., de Gooijer, Abraham, Gould, and Robinson 1985), with particular emphasis on automatic model selection criteria such as the Akaike information criterion (AIC; Akaike 1973) and the Bayes information criterion (BIC; Schwarz 1978). Let $y_t = x_t + v_t$where $v_t$is an iid with variance σ_v^2 and x_t has variance σ_x^2. Then the lag-l correlations of the processes $\{y_t\}$ and $\{x_t\}$ denoted by ρ_l^Y and ρ_l^X, satisfy

$$\rho_l^Y = \rho_l^X(1 - R), \quad l = 1,2, \ldots \ldots \ldots$$

where $R = \frac{\sigma_v^2}{(\sigma_X^2 + \sigma_V^2)}$. Therefore, as σ_V^2 increases to ∞ (i.e., as R increases to 1), ρ_l^Y decreases to zero. Thus model selection based on the empirical autocorrelation and partial autocorrelation functions. Both AIC and BIC are monotone functions of the prediction error variance σ_ε^2, usually estimated by the maximum likelihood estimate (MLE) $\hat{\sigma}_\epsilon{}^2$. Here we propose a new approach to the comparison of AR models that attempts to overcome the difficulties associated with model uncertainty. This consists of calculating the posterior probabilities of the competing AR*(p)* models in a way that is Empirical to outliers and then obtaining the predictive distributions of quantities of interest, such as future observations or characteristics of the spectrum, as a weighted average of the

conditional predictive distribution given each of the models. To obtain the posterior probabilities, we calculate the Bayes factors, or ratios of posterior to prior odds, for each of a set of pairwise model comparisons. The basic idea is explained in Section 2 in the context of AR models. In Section 3 we introduce the idea of Empirical Bayes factors, obtained by replacing the likelihood for model (1) by a Empirical likelihood following Martin (1981). This Empirical likelihood has two key ingredients. The first is a Empirical dictor that provides Empirical location or centering for the predictive distribution, along with an associated Empirical scale. The Empirical predictor and associated scale are obtained using the Empirical filtering algorithm of Masreliez (1975) and Martin (1979). The second ingredient is a bounded and continuous likelihood-type loss function that replaces the non-Empirical sum of squared residuals in the Gaussian likelihood. Here Empirical refers to Empirical of the Bayes factors against the outliers in the observed data, not against the prior distribution as often referred to in the literature. Computation of the Bayes factors using the Empirical likelihood requires integration over the parameter space. Because this is analytically difficult, we follow Raftery (1988) and use the Laplace method for integrals (Tierney and Kadane 1986). In doing so, we reparameterize the model (1) in terms of the partial autocorrelation coefficients and modify the Laplace method to take into account the finiteness of the parameter space.

2. BAYES FACTORS FOR TIME SERIES

2.1 Bayes Factors and Accounting for Model Uncertainty :

Fruitful approaches to statistical problems often involve postulating a class of probability models and comparing these models on the basis of how well they predict the observed data. The Bayesian approach to the problem of inference in the presence of several competing models is based on posterior model probabilities. If the class consists of the (p + 1) models $M_{0,}$,….M_p, then the posterior probability of the model M_p given data D is

$$p(M_p|D) = \frac{p(D|M_p)p(M_p)}{\sum_{l=0}^{K} p(D|M_l)p(M_l)} \qquad (2)$$

In Equation (2), $p(M_p)$ is the prior probability of model M_p and $p(D|M_p)$ is its integrated likelihood, defined by

$$p(D|M_p) = \int_{\ominus_p} p(D|\theta_p, M_p)p(\theta_p|M_p)d\theta_p \qquad (3)$$

where θ_p is the (vector) parameter of model M_p, $p(\theta_p|M_p)$ is its prior distribution, $p(D|\theta_p, M_p)$ is the likelihood, and θ_p is the parameter space. Pairwise comparisons are based on the posterior odds ratio

$$\frac{p(M_p|D)}{p(M_l|D)} = \left[\frac{p(D|M_p)}{p(D|M_l)}\right]\left[\frac{p(M_p)}{p(M_l)}\right] = B_{pl}\lambda_{pl},$$

where B_{pl} is the Bayes factor for M_p against M_l and λ_{pl} is the corresponding prior odds. If M_p is nested within M_l, then the data D favor M_p if $B_{pl} > 1$, whereas they provide evidence for M_l if $B_{pl} < 1$, Jeffreys (1961, app. B) suggested that the evidence for the larger model be considered strong if $B_{pl} < 10^{-1}$, and conclusive if $B_{pl} < 10^{-2}$. The posterior probabilities can be recovered using the equation. $p(M_p|D) = \left[B_{0p}\lambda_{0p}\{1 + \sum_{l=1}^{p}(B_{0l}\lambda_{0l})^{-1}\}\right]^{-1}$

This framework yields solutions to the estimation, prediction, and decision-making problems that take into account uncertainty about the order of the AR, unlike model selection methods that condition on a single selected model. If Δ is a quantity of interest, such as a property of the spectrum, the next observation, or the utility of a course of action, then its posterior distribution given the data D is evaluated by combining all models considered; that is,

$$P(\Delta|\ D) = \sum_{p=0}^{n} p(\Delta|\ M_{p,}D)p(M_p\ |\ D) \qquad (4)$$

This equation was first given by Leamer (1978, p.117) and was proposed explicitly as a solution to the decision-making problem in the time series context in equation (5.1) of Poskitt (1988).A simple approximation for $p(D\ |\ M_p)$, introduced by Schwarz (1978), is

$$\text{Log } p(D\ |\ M_p) \approx \log p(D\ |\ M_p,\ \theta_p) - \tfrac{1}{2}\text{d log n}, \qquad (5)$$

where is the MLE of $\hat{\theta}_p$ and d and n are the numbers of parameters and observations. We refer to Equation (5) as the BIC approximation; its error is O(1) (Kass and Raftery 1995). Choosing the order that maximizes the right side of Equation (5) is the much-used BIC model selection procedure. Akaike (1983) wrote that, asymptotically,

$$\text{Log } p(D\ |\ M_p) \approx \log p(D\ |\ M_p,\ \theta_p) - d, \qquad (6)$$

which we call the AIC approximation. This is true only if prior information increases at the same rate as the information in the data, which is unrealistic in most applications. Nevertheless, the procedure of choosing the order that maximizes the right side of Equation (6) has been much used, and so we including it in our comparison and examples. (For a review of Bayes factors, see Kass and Raftery 1995).

2.2. Bayes Factors for Autoregressive Processes :

We now apply the Bayesian framework to the model comparison problem where the data $y^T = (x_1, \ldots x_T)$ are from a stationary Gaussion AR(p) process defined by (1); Let M_p denote the Gaussian AR(p) model. To obtain the posterior probabilities, we need to evaluate the integrated likelihood, $\text{p}(y^T\ |\ M_p)$, p=0,…P, which is given by Equation (3). The log-likelihood function of the data given the model and its parameters is

$$\text{Log p}(y^T\ |\ \theta_{p,}M_p)\ = -\frac{T}{2}log(2\pi) - \frac{1}{2}\sum \log f_t^2 - \frac{1}{2}\sum\left(\frac{x_{t-}\hat{x}_t^{t-1}}{f_t}\right)^2.$$

where $\hat{x}_t^{t-1} = \text{E}(x_t/y^{t-1}), y^{t-1} = (x_{t-1,}\, x_{t-2,\ldots,}x_1)$ and $f_t^2 = E(x_t - \hat{x}_t^{t-1})^2$ Thus $\hat{x}_t^{t-1}$ is the conditional mean of x_t given the data up to time $(t-1)$ and f_t^2 is the corresponding conditional variance; both can be found using the Kalman filter (Harvey 1981). There are several difficulties with the evaluation of the integral (3). The constraints on the parameters $(\emptyset_1, \ldots\ \emptyset_p)$ that ensure stationarity are complicated. We avoid this difficulty by reparameterizing in terms of the first p partial autocorrelations $\theta_p = (\pi_1, \ldots., \pi_p)$. The parameter space θ_p is then just the hypercube $(-1, 1)^p$ and the mapping that transforms the $(\emptyset_1, \ldots, \emptyset_p)$ such that the process is stationary to $(\pi_1, \ldots, \pi_p)$ is one-to-one and onto $\ominus_p$ and both it and its inverse are continuously differentiable (Barndorff-Nielsen and Schou 1973; Ramsey 1974). The integral (3) cannot be evaluated analytically, and so we approximate it using the Laplace method for integrals (Tierney and Kadane 1986). The Laplace method was applied to Bayes factors by Raftery (1988) in the context of generalized linear models. The Laplace method is modified here to take into account the finiteness of the parameter space, as follows. Let $\text{g}(\theta)$ be a real-valued function from R^p to R, where θ is a p-dimensional vector. A Taylor series expansion of $g(\theta)$ at θ_0 yields

$$g(\theta) \approx g(\theta_0) + (\theta - \theta_0)'(\nabla_{g\theta_0}) + \frac{1}{2}\,(\theta - \theta_0)'(\nabla^2{}_{g\theta_0})(\theta - \theta_0)$$

where $(\nabla_{g\theta_0})$ and $(\nabla^2{}_{g\theta_0})$ are the gradient and Hessian of $g(\theta)$ evaluated at θ_0. Let θ_0 be the mode of $g(\theta)$. Then

$$\int_{\Theta} exp \quad [g(\theta)]d\theta$$

$$\approx \int_{\Theta} \exp\left[g(\theta_0) + \frac{1}{2}(\theta - \theta_0)'(\nabla^2 g\theta_0)(\theta - \theta_0)\right] d\theta$$

$$= \exp[g(\theta_0)] \int_{\theta} exp \quad \left[\frac{1}{2}(\theta - \theta_0)'(\nabla^2 g\theta_0)(\theta - \theta_0)\right] d\theta$$

$$= \exp[g(\theta_0)] \mid -\nabla^2 g\theta_0 \mid^{1/2} \left(2\pi^{p/2}\right) \int_{\Theta} \emptyset(\theta) d\theta \ , \qquad (7)$$

where $\emptyset(\theta)$ is the p-dimensional multivariate normal density with mean (θ_0) and variance-covariance matrix $[-\nabla^2 g\theta_0]^{-1}$

Applying the approximation (7) to the integral (3), with $\theta_p = (\pi_1, \ldots \pi_p)$ as the k-dimensional vector of partial autocorrelations and with

$$g(\theta_p) = \log[p(y^T \mid \theta_p, M_p) p(\theta_p \mid M_p)],$$

$$p(y^T \mid M_p) \approx (2\pi)^{\frac{p}{2}} \mid \nabla^2 g\theta^*_p \mid^{\frac{1}{2}} p(y^T) \mid \theta^*_p, M_p) p(\theta^*_p \mid M_p)$$

$$\times \int_{(-1,1)^p} \emptyset(\theta_p) d\theta_p, \qquad (8)$$

where θ^*_p is the value of θ_p that maximizes $g(\theta_p)$. The integral on the right side of Equation (8) is evaluated by Monte Carlo integration. Arguments similar to those of Tierney and Kadane (1986) show that the error of the approximation (8) is $O(T^{-1})$.Thus for a good approximation of the integrated like-lihood $p(y^T \mid M_p)$, all we need are the posterior mode of θ_p and the Hessian of the log-likelihood function, $\log[p(y^T \mid \theta_p, M_p)]$ at that point. A natural parameterization for AR models is in terms of the partial autocorrelations, when the parameter space is the hypercube $(-1,1)^p$. When little prior information is available, a reasonable "noninformative" prior is uniform in the partial autocorrelations; this is proper, and so difficulties with improper priors do not arise. With this prior, the posterior mode is equal to the MLE.

3. Empirical Bayes Factors for Autoregressive Models

3.1. A Empirical Likelihood for Autoregressive Models :

We now consider the model comparison problem for AR processes with additive outliers. Suppose that the data y^T=($y_1, \ldots y_T$) are generated by the model.

$$y_t = x_t + Z_t W_t, \qquad (9)$$

where $\{x_t\}$ follows Equation (1), $\{W_t\}$ is a sequence of observations from a generating distribution whose variance is much larger than σ^2_ϵ and $\{Z_t\}$ is a 0 – 1 process with P$[Z_t = 1] = \gamma$ being the fraction of outliers in the data. When $Z_t = 1, y_t$ is called an additive outlier.

Our approach to the comparison of different AR orders in the model (9) is to use a Empirical likelihood that approximates the likelihood of the (unobserved) series$\{x_t\}$. Following Martin (1981), this is defined as

$$\text{Log } \tilde{p}(y^T \mid M_p, \theta_p) = -\frac{T}{2} log \quad (2\pi) - \frac{1}{2}\sum_{t=1}^{T} \log S_t^2 - \frac{1}{2}\sum_{t=1}^{T} \rho\left(\frac{y_t - \tilde{x}_t^{t-1}}{S_t}\right) \qquad (10)$$

In Equation (10), $\tilde{x}_t^{t-1}$ and S_t are Empirical estimates of the conditional mean and standard deviation of x_t given $x_1, \ldots x_{t-1}$, found by Empirical filtering as described in Section 3.2. The function is chosen to be bounded and continuous so as to ensure that one observation does not have a large influence on the likelihood function and that small changes do not produce large changes in the likelihood function. Here we use the function.

$$\rho(x) = x^2 \text{ if } |x| \le a.$$
$$= a^2 \text{ if } |x| > a.$$

The observations whose prediction residuals $y_t = \hat{x}_t^{t-1}$ are large compared to their predictive standard deviations s_t. Here we use a = 2.5 as the tuning constant, so that an observation is censored once its prediction residual is more than 2.5 times its predictive standard deviation. A Empirical integrated likelihood $\tilde{p}(y^T \mid M_p)$ is defined by replacing $\dot{p}(y^T \mid M_p, \theta_p)$ with $\tilde{p}(y^T \mid M_p, \theta_p)$ in (8).

3.2. Empirical Filtering :

` To calculate the robust Bayes factors, we need the prediction location and scale of the observations, $\hat{x}_{t|t-1}$ and S_t .We obtain these using the robust filter of Maserliez (1975) and Martin (1979). The model (1) and (9) can be written in state-space form as

$$X_t = \Gamma x_{t-1} + \varepsilon_t$$

And
$$y_t = HX_t + v_t, \qquad (11)$$

Where $v_t \equiv Z_t W_t$ denotes the outlier-generating component, X_t and ε_t have dimesion p, Γ is a p × p matrix, and H is a 1 × p matrix, defined by

$$\Gamma = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix}$$

$$x_t^T = (x_t, x_{t-1}, \ldots, x_{t-p+1})$$

H = (1, 0, ...0)

$$\varepsilon_t^T = (\varepsilon_t, 0, \ldots 0).$$

We denote the state prediction density by $f(X_t \mid y^{t-1})$; this is assumed to exist for t ≥ 1. The observation prediction density is $f(y_t \mid y^{t-1})$. The conditional mean of X_t given y^t is denoted by $\hat{X}_t = E\left(x_t / y^t\right)$.

When ε_t and v_t in (11) are Gaussian the computation of $\hat{X}_t = E\left(x_t / y^t\right)$ yields the kalman filter recursion equation. Unfortunately, $\hat{X}_t$ is hard to calculate exactly when v_y is non-gaussian, except in a few special cases such as that of stable random variables (stuck1976). But there is a simplifying assumption that does allow calculation of $\hat{X}_t$ (Marseliez 1975) –that the state predictor density is Gaussian, namely.

$$f(X_t \mid y^{t-1}) = N\left(X_t; \hat{X}_t^{t-1}, M_t\right),$$

Where $N(¨, \mu, \sum)$ denotes the multivariate normal density with mean μ and covariance matrix ∑ and

$$M_t = E\left\{\left(X_t - \hat{X}_t^{t-1}\right)\left(X_t - \hat{X}_t^{t-1}\right)^T \mid y^{t-1}\right\}$$

Is the conditional covariance matrix for the state prediction error. given this, satisfies the recursion

$$\hat{X}_t = \hat{X}_t^{t-1} + M_t H^T \psi_t(y_t), \qquad (12)$$

$$M_{t+1} = \Gamma P_t \Gamma^T + Q, \qquad (13)$$

and
$$P_t = M_t - M_t H^t \psi'_t(y_t) H M_t, \qquad (14)$$

$$\psi_t(y_t) = -\left(\frac{\partial}{\partial y_t}\right) log \quad f_y(y_t \mid y^{t-1})$$

is the score function for the observation prediction density $f_y(y_t \mid y^{t-1})$. The matrix Q is the covariance matrix of $\in_t$ that is equal to σ_ϵ^2 at the (1,1) position and to zero every where else, $\hat{X}_t^{t-1} = \Gamma \hat{X}_{t-1}$ and

$$\psi'_t(y_t) = -\left(\frac{\partial}{\partial yt}\right)\psi_t(y_t).$$

The density $f_y(y_t \mid y^{t-1})$ is generally intractable when outliners are present. Thus it is difficult to obtain the ψ function. But, as noted by martin (1979) ψ and ψ' can be well appropriately chosen bounded continuous functions. Boundedness ensures that y_t does not have an un bounded influence on $\hat{X}_t$ and continuity ensures that small changes in y_t do not produce large changes in $\hat{X}_t$ that hampel's two- part redescending function caused little bias in outlier- free situations while providing good robustness towards outliers. thus here we use hampel's two –part redescending function,

$$\begin{aligned}
\psi(y) &= y, \qquad |y| \le \alpha c, \\
&= \alpha(c - y)/(1 - \alpha), \qquad \alpha c < y \le c, \\
&= -\alpha(c + y)/(1 - \alpha), \qquad -c \le y < -\alpha c, \\
&= 0, \quad |y| > c,
\end{aligned}$$

with αc = 2.5 and c = 4.0. That is, observation with prediction residuals (divided by their predictive standard deviations) in the interval are downweighted linearly , and those with prediction residuals greater than 4 are given zero weight. To ensure boundedness and continuity of ψ' martin and su (1985) also recommended that ψ' be replaced by the weight function $w(z) = \psi(z)/z$.

Let S_t^2 be the (1,1) element of M_t. Then the recursions (12) –(14) may be replaced by

$$\hat{X}_t = \Gamma \hat{X}_{t-1} + \frac{m_t}{s_t^2} S_t \psi\left(\frac{r_t}{s_t}\right),$$

$$M_{t+1} = \Gamma P_t \Gamma^T + Q,$$

$$P_t = M_t - \omega\left(\frac{r_t}{s_t}\right)\frac{m_t m_t^T}{s_t^2}$$

and where m_t is the first column of M_t and r_t is the observation prediction residual

$$r_t = y_t - H\hat{X}_t^{t-1}.$$

REFERENCES

1. Barndorrf-Nielsen, O., and Schou, G. (1973), "On the parameterization of Autoregressive Models by partial Autocorrelations," Journal of Multivariate Analysis, 3, 408-419.
2. Akaike,H. (1973), " Information Theory and an Extension of the Maximum Likelihood Principal,"In the Second International Symposium on Information Theory, eds.by B.N.Petrov and F.Csake, Akademiai Kiado, Hungary,pp. 267-281.
3. Brubacher, S.R.(1974), " Time series outlier Dectection and Modeling With Interpolation,"Bell Laboratories technical memo.
4. Box,G.E.P., and Jenkins, G.M. (1976) ,Time series Analysis Forecasting and control (2nd ed.), San Francisco: Holden-Day.

5. Jeffreys, H.(1961), Theory of probability (3rd ed.), Oxford, U.K.: Oxford University press.
6. Harvey, A.C. (1981), Time series Models, New York: Halsted press.
7. Kass, R.E., and Raftery, A.E.(1985) "bayes Factores," Journal of the American Statistical Assocication, 90,773-795.
8. Le, N.D., Martin, R.D., and Raftery , A.E.(1990), " Modeling outliers, Bursts and Flat stretches In time Series Using Mixture Transition Distribution (MTD) mModels," Technical Report 194, University of Wasington, Dept.of statistics.
9. Leamer,E.E. (1978), Specification Searches: Ad Hoc inference With Non-experimental data,New York: John Wiley.
10. Madigan,D., and Raftery, A.E.(1994), " Model selection and Accounting for Model Uncertainty In Graphical Models using Occam's Window," Journal of the American Statistical Ass0ciation, 89,1335-1346.