1

An Unified Approach To Model Selection, Discrimination, Goodness of fit And Outliers In Time Series

Daniel Peña and Pedro Galeano

Departamento de Estadística, Universidad Carlos III de Madrid, Spain
Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela, Spain

Abstract:

This article presents an unified approach of several procedures in time series. First, we show that quadratic discrimination provides a unifying approach for deriving model selection criteria in the ARMA framework. Second, we establish a connection between model selection criteria and a goodness of fit test. Finally, we show that the outlier detection problem can be seen as a particular case of model selection. Therefore, the problems of model selection, discrimination, goodness of fit tests and outliers can be treated under the same principles.

Keywords and phrases: Akaike information criterion; Bayesian information criterion; Discrimination; Goodness of fit tests; Outliers

1.1 Introduction

Model selection criteria is one of the most popular tools for selecting the model that better fits the data among a set of candidates . Although these criteria have been derived from very different points of view, it is usual to split them into two different groups. The first group is formed by the consistent criteria, which, under certain conditions and the assumption that the data come from a model with a finite number of parameters, asymptotically select the true one. Two consistent criteria are the Bayesian information criterion (BIC), derived by Schwarz (1978), which approaches the posterior probabilities of the models, and the Hannan-Quinn criterion (HQC), derived by Hannan and Quinn (1979), which was designed to be a consistent criterion with the fastest convergence rate to the true model. The second group is formed by the efficient criteria, which, under certain conditions and the assumption that the data come from a model with an infinite number of parameters, asymptotically select the model which produces the least mean square prediction error. Three efficient criteria are the final prediction error criterion (FPE), derived by Akaike (1969), which selects the model that minimizes the one step ahead square prediction error, the Akaike information criterion (AIC), derived by Akaike (1973), which is an estimator of the expected Kullback-Leibler divergence between the true and the fitted model, and the corrected Akaike information criterion (AICc), derived by Hurvich and Tsai (1989), which is a bias correction form of the AIC that appears to work better in small samples.

All these criteria have the general form:

$$MSC = -2 \times (\log \text{maximized loglikelihood}) + r \times C(T, r),$$

where r is the number of estimated parameters of the model, T is the sample size of the data, and $C(T,r) = \log(T)$, for the BIC, $C(T,r) = 2c \log\log(T)$ with $c > 1$, for the HQC, $C(T,r) = \frac{T}{r} \log(\frac{T+r}{T-r})$ for the FPE, $C(T,r) = 2$ for the AIC, $C(T,r) = \frac{2T}{T-r-1}$ for the AICc.

The discrimination problem appears when it is known that the data may belong to one of several known populations and the objective is to classify the data into one of these populations. When the data are Gaussian distributed, the classic solution to this problem is to classify the data by using either the standard or the Bayesian quadratic discrimination rule. The first purpose of this paper is to see that both quadratic discrimination rules for ARMA time series models provides a way of deriving model selection criteria such as the AIC, AICc and BIC criteria, establishing a connection between discrimination and model selection in linear Gaussian time series.

Goodness of fit tests are a useful tool for checking whether the data are reasonable well fitted by a chosen model. These tests proceed by using a test statistic, which is compared with some value of the statistic assuming the model is correct and a given confidence level, which takes into account the potential loss incurred if the model is rejected. Thus, it is reasonable to think that model selection criteria and the goodness of fit tests are closely related. The second purpose of this paper is to analyze the relationships between model selection criteria and a goodness of fit test statistic for linear Gaussian time series proposed by Peña and Rodríguez (2006).

The real data are often affected by the presence of outliers which may have serious effects on statistical analysis in many different ways (see, Barnett and Lewis, 1993). The most usual method to detect the presence of outliers is the use of test statistics based on the Mahalanobis distances between the data and some estimate of the center of the data distribution. The third intent of this paper is to show that the detection of outliers in a linear Gaussian time series can be seen as a kind of model selection problem and that model selection criteria provide some objective rules for outlier detection.

The rest of this paper is organized as follows. Section 2 briefly presents the class of autoregressive moving average Gaussian time series models. Section 3 shows the connection between the quadratic discriminant rules in linear Gaussian time series and model selection criteria from the maximum likelihood and Bayesian approaches. Sections 4 shows the connection between the goodness of fit test proposed by Peña and Rodríguez (2006) and model selection criteria. Finally, section 5 shows that the problem of outlier detection in time series can be seen as a model selection problem which provides new suitable solutions to this problem.

1.2 ARMA Time Series Models

In what follows, assume that a time series given by $x = (x_1, ..., x_T)'$ has been generated by the autoregressive moving average Gaussian process, ARMA(p, q), if it follows the equation:

$$x_t - \phi_1 x_{t-1} - \ldots - \phi_p x_{t-p} = a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q}, \tag{1.1}$$

where a_t is a sequence of independent Gaussian distributed random variables with zero mean and variance $\sigma^2_{p,q}$. The ARMA(p, q) model, denoted by $M_{p,q}$, has the $(p + q + 1) \times 1$ vector of parameters $\alpha_{p,q} = \left(\beta'_{p,q}, \sigma^2_{p,q}\right)'$ where $\beta_{p,q} = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$, is assumed to be causal, invertible, stationary and such that the polynomials $1 - \phi_1 B - \ldots - \phi_p B^p$ and $1 - \theta_1 B - \ldots - \theta_q B^q$ have no common roots.

The likelihood function of x under the model $M_{p,q}$ is given by:

$$p\left(x \mid M_{p,q}\right) = (2\pi)^{-\frac{T}{2}} \left|\Sigma_{p,q}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} x' \Sigma^{-1}_{p,q} x\right),$$

where $\Sigma_{p,q}$ is the $T \times T$ covariance matrix of x under the model $M_{p,q}$, which can be written as $\Sigma_{p,q} = \sigma^2_{p,q} Q_{p,q}$, where $Q_{p,q}$ is a $T \times T$ matrix which only depends on the parameters $\beta_{p,q}$. The vector of innovations can be written as $a_{p,q} = L^{-1}_{p,q} x$, where $Q_{p,q} = L_{p,q} L'_{p,q}$ is the Cholesky decomposition of $Q_{p,q}$.

The maximum likelihood estimators of the vector of parameters $\alpha_{p,q}$ of the model $M_{p,q}$ are denoted by $\widehat{\alpha}_{p,q} = \left(\widehat{\beta}'_{p,q}, \widehat{\sigma}^2_{p,q}\right)'$ and are obtained after maximizing the log-likelihood of x under the model $M_{p,q}$, given by:

$$\log p\left(x \mid M_{p,q}\right) = -\frac{T}{2} \log 2\pi - \frac{1}{2} \log \left|\Sigma_{p,q}\right| - \frac{1}{2} x' \Sigma^{-1}_{p,q} x.$$

The estimated covariance matrix of x under the model $M_{p,q}$ is written as $\widehat{\Sigma}_{p,q} = \widehat{\sigma}^2_{p,q} \widehat{Q}_{p,q}$, where $\widehat{Q}_{p,q}$ is the matrix $Q_{p,q}$ with $\beta_{p,q}$ replaced by $\widehat{\beta}_{p,q}$. The vector of residuals of the fit can be written as $\widehat{a}_{p,q} = \widehat{L}_{p,q}^{-1} x$, where $\widehat{Q}_{p,q} = \widehat{L}_{p,q} \widehat{L}_{p,q}'$ is the Cholesky decomposition of $\widehat{Q}_{p,q}$.

1.3 Quadratic Discrimination of ARMA Time Series Models

The discrimination problem in time series can be stated as follows (see, Galeano and Peña, 2000). Suppose it is known that the time series $x = (x_1, ..., x_T)'$ has been generated by one of the models $M_{p,q}$, in which $p \in \{0, ..., p_{\max}\}$ and $q \in \{0, ..., q_{\max}\}$, where $p_{\max}$ and $q_{\max}$ are some fixed upper bounds. The objective of discrimination is to select the true data generating model of the time series x, which is denoted by M_{p_0,q_0} and has the $(p_0 + q_0 + 1) \times 1$ vector of parameters $\alpha_{p_0,q_0} = \left(\beta_{p_0,q_0}', \sigma_{p_0,q_0}^2\right)'$. This is equivalent to consider the set of hypothesis $M_{p,q} : x \in N_T\left(0, \Sigma_{p,q}\right)$. The standard quadratic classification rule will select the model which maximizes,

$$p\left(x \mid M_{p,q}\right) = (2\pi)^{-\frac{T}{2}} \left|\Sigma_{p,q}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} x' \Sigma_{p,q}^{-1} x\right), \tag{1.2}$$

while the standard quadratic Bayesian classification rule will select the model which maximizes,

$$p\left(M_{p,q}\right) p\left(x \mid M_{p,q}\right) = p\left(M_{p,q}\right) (2\pi)^{-\frac{T}{2}} \left|\Sigma_{p,q}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} x' \Sigma_{p,q}^{-1} x\right), \tag{1.3}$$

where $p\left(M_{p,q}\right)$ is the prior probability of the model $M_{p,q}$.

In practice, the vector of parameters $\alpha_{p,q}$ may be considered as unknown. Following a maximum likelihood approach, if the unknown parameters, $\alpha_{p,q}$, are replaced by its maximum likelihood estimates, $\widehat{\alpha}_{p,q}$, the rule (1.2) will always choose the model with the largest number of parameters. A first attempt to avoid this problem is to select the model that maximizes:

$$E_{\alpha_{p_0,q_0}}\left[\log p(y|\widehat{\alpha}_{p,q})\right] = \int \log p(y|\widehat{\alpha}_{p,q}) p(y|\alpha_{p_0,q_0}) dy,$$

which will be the model that maximizes the expectation with respect to future observations generated by the true model with parameters α_{p_0,q_0}. Note that this approach takes into account the uncertainty about new observations but not the uncertainty in the parameter estimates. Galeano and Peña (2007*a*) showed that:

$$\begin{aligned} E_{\alpha_{p_0,q_0}}\left[\log p(y|\widehat{\alpha}_{p,q})\right] = & -\frac{T}{2}(\log 2\pi + 1) - \frac{1}{2}\log\left|\widehat{\Sigma}_{p,q}\right| \\ & -\frac{T(p+q+1)}{T-(p+q+1)-1} + O_p(1), \end{aligned}$$

which include terms that have the same order, $O_p(1)$, that the penalty term and can not be avoided. Thus, it is necessary to take into account the uncertainty about the parameter estimates, which can be done by taking also the expectation with respect to the distribution of the estimate, $\widehat{\alpha}_{p,q}$. This leads to select the model which attains the largest value of:

$$E_{\widehat{\alpha}_{p,q}}\left[E_{\alpha_{p_0,q_0}}\left[\log p(y|\widehat{\alpha}_{p,q})\right]\right] = \int\int \log p(y|\widehat{\alpha}_{p,q})p(y|\alpha_{p_0,q_0})dyd\widehat{\alpha}_{p,q} \qquad (1.4)$$

where $\widehat{\alpha}_{p,q}$ and y are assumed to be independent. Thus, the rule selects the model that maximizes the expected value with respect to the two sources of uncertainty: the distribution of future observations and the distribution of the estimate. The rule (1.4) can be written as follows (see, Galeano and Peña, 2007):

$$\begin{aligned} E_{\widehat{\alpha}_{p,q}}\left[E_{\alpha_{p_0,q_0}}\left[\log p(y|\widehat{\alpha}_{p,q})\right]\right] &= -\frac{T}{2}\left(\log 2\pi + 1\right) - \frac{1}{2}\log\left|\widehat{\Sigma}_{p,q}\right| \\ &- \frac{T\left(p+q+1\right)}{T-\left(p+q+1\right)-1} + o(1), \end{aligned}$$

which is equivalent to the expression of the AICc criterion for ARMA models derived by Hurvich, Shumway and Tsai (1990). Note also that the rule (1.4) selects the model which minimizes the expected Kullback-Leibler divergence to the true one, which is given by:

$$E_{\widehat{\alpha}_{p,q}}\left[E_{\alpha_{p_0,q_0}}\left[\log\frac{p(y|\alpha_{p_0,q_0})}{p(y|\widehat{\alpha}_{p,q})}\right]\right] = \int\int \log\frac{p(y|\alpha_{p_0,q_0})}{p(y|\widehat{\alpha}_{p,q})}p(y|\alpha_{p_0,q_0})dyd\widehat{\alpha}_{p,q},$$

and was approximated by Akaike (1973) to derive the AIC criterion. Thus, both the AIC and the AICc may be derived by the standard quadratic discriminant rule.

On the other hand, the Bayesian approach of computing the posterior probabilities of each model takes automatically into account the two sources of uncertainty previously discussed. In fact, the log-likelihood, $\log p\left(x \mid M_{p,q}\right)$, can be written as follows by using the Laplace approximation (see, Galeano and Peña, 2007*a*):

$$\begin{aligned} \log p(x|M_{p,q}) &= \frac{1}{2}\left(p+q+1-T\right)\log\left(2\pi\right) - \frac{1}{2}(p+q+1)\log\left(T\right) \\ &- \frac{1}{2}\log\left|\widehat{\Sigma}_{p,q}\right| - \frac{1}{2}T + \log p(\widehat{\alpha}_{p,q}|M_{p,q}) + O_p(1), \end{aligned}$$

which, taking the same prior probabilities for all the set of candidate models, leads to the expression of the BIC criterion of ARMA models. Therefore, the Bayesian quadratic classification rule (1.3) leads to the BIC criterion proposed by Schwarz (1978).

In summary, it has been shown that the quadratic discriminant rules allow to derive model selection criteria such as the AIC, AICc and BIC, which shows the connection between discrimination and model selection in linear Gaussian time series. Thus, the model selection problem can been seen as a kind of discrimination analysis which allow to present an unified approach of criteria proposed in the literature from different points of view.

1.4 Goodness of fit for ARMA Time Series Models

The goodness of fit tests in time series work as follows. After selecting a model to fit the time series $x = (x_1, ..., x_T)'$, a goodness of fit test checks whether the data are reasonable well fitted by the chosen model by using a test statistic which measures the quality of the fit. Thus, although goodness of fit appears to be related with model selection criteria in some way, they are not identical procedures. Some goodness of fit test for linear time series are the proposed by Ljung and Box (1978), Monti (1994), Velilla (1994) and Peña and Rodríguez (2003). In this section, we analyze the connection between both problems in the particular case of the goodness of fit test for linear Gaussian time series proposed by Peña and Rodríguez (2006). These authors used the log of the determinant of the $T \times T$ autocorrelation matrix of the estimated residuals $\widehat{a}_{p,q}$, which is denoted by $\widehat{R}_{p,q}$, for testing goodness of fit in ARMA time series. This autocorrelation matrix is defined as follows:

$$\widehat{R}_{p,q} = \frac{\widehat{a}_{p,q}\widehat{a}'_{p,q}}{T\widehat{\sigma}^2_{p,q}}.$$

On the other hand, taking into account that $x = \widehat{L}_{p,q}\widehat{a}_{p,q}$, the sample covariance matrix of x can be written as:

$$\frac{xx'}{T} = \frac{\widehat{L}_{p,q}\widehat{a}_{p,q}\widehat{a}'_{p,q}\widehat{L}'_{p,q}}{T} = \widehat{\sigma}^2_{p,q}\widehat{L}_{p,q}\widehat{R}_{p,q}\widehat{L}'_{p,q},$$

which shows that,

$$\left|\frac{xx'}{T}\right| = \left|\widehat{\sigma}^2_{p,q}\widehat{L}_{p,q}\widehat{R}_{p,q}\widehat{L}'_{p,q}\right| = \left|\widehat{\Sigma}_{p,q}\right| \left|\widehat{R}_{p,q}\right|, \tag{1.5}$$

because $\left|\widehat{\Sigma}_{p,q}\right| = \left|\widehat{\sigma}^2_{p,q}\widehat{L}_{p,q}\widehat{L}'_{p,q}\right|$. Now, taking logs in (1.5), the following expression holds:

$$\log\left|\frac{xx'}{T}\right| = \log\left|\widehat{\Sigma}_{p,q}\right| + \log\left|\widehat{R}_{p,q}\right|.$$

Thus, all the considered model selection criteria can be written in terms of the goodness of fit test of Peña and Rodríguez (2006) as follows:

$$MSC\left(M_{p,q}\right)=\log\left|\frac{xx'}{T}\right|-\log\left|\widehat{R}_{p,q}\right|+(p+q+1)\,C\left(T,p+q+1\right).$$

Therefore, taking into account that the sample covariance matrix of the time series x is constant for all the candidate models, any of the model selection criterion will select the model which have a significatively larger value of the goodness of fit test statistic proposed by Peña and Rodríguez (2006) but penalized by the number of parameters. This shows two interesting facts. First, the model selected by a model selection criterion is not always the model with the most significant goodness of fit statistic. Second, the term $\log\left|\widehat{\Sigma}_{p,q}\right|$ can also be seen as a measure of the goodness of fit of the model $M_{p,q}$ to the series x.

1.5 Outliers In ARMA Time Series Models

Outliers in time series can arise for several reasons. First, outliers may be gross errors such as measurement, recording and typing mistakes. Second, outliers may be real data which are somehow suspicious or surprising as they not follow the same pattern that the rest of observations and may be caused for unknown events. The presence of outliers in time series can seriously affect the estimation of the parameters of the model and produce poor forecasts. Since the seminal paper of Fox (1972), outliers in time series have received considerable attention and several papers have analyzed their effects and proposed methods for their detection in univariate linear time series. See for instance, Tsay (1986), Chang, Tiao and Chen (1988), Chen and Liu (1993), Le, Martin and Raftery (1996), Luceño (1998), and Sánchez and Peña (2003), among others. Much of these works have been focused on the framework of statistical hypothesis testing. In particular, the procedure proposed by Chen and Liu (1993) is widely used and has been implemented in several time series packages used by practitioners, such as TRAMO and SCA. This and other procedures rely on the use of likelihood ratio tests with critical values are obtained via simulation depending on different sample sizes and models. In this section, we show that the outlier detection problem can be formulated as a model selection problem, which can be solved by using model selection criteria. These criteria provide objective rules to decide whether a set of observations are outliers or not, avoiding the use of simulation to obtain critical values.

Let $x=(x_1,\ldots,x_T)'$ be a time series generated by an ARMA(p,q) process as in (1.1), where, for simplicity, the orders p and q are assumed known. Assume that instead of observing x, we observe a time series $y = (y_1, \ldots, y_T)'$ defined as follows:

$$y_t = \begin{cases} x_t & t \neq t_1, \ldots, t_m \\ x_t + w_t & t = t_1, \ldots, t_m \end{cases},$$

where m is the number of outliers in the time series, $t_1, \ldots, t_m$ are their locations, such that $1 \leq t_1 < \cdots < t_m \leq T$, and $w_{t_1}, \ldots, w_{t_m}$ are their sizes. Note that m can be as large as T, the sample size.

In practice, either the parameters of the ARMA(p, q) model and the number, locations and sizes of the outliers are unknown, and have to be estimated from the data. In this section, it is shown that the outlier detection problem can be stated as a model selection problem, for which model selection criteria can be applied. For that, three alternative approaches are analyzed.

The first approach is the following. Let $M_{t_1,\ldots,t_m}$ be the ARMA(p, q) model with m outliers with locations at the vector time indices $t_1, \ldots, t_m$. The problem of joint estimation of the model parameters, number of outliers, their locations and their sizes can be now stated as the selection of the true model among the set of candidate ones, which include the model without outliers, denoted by M, the T models with one outlier, denoted by $M_1, \ldots, M_T$, etc... In general, there are $\binom{T}{m}$ candidate models with m outliers with all the possible $\binom{T}{m}$ locations of the m outliers, so that, the total number of candidate models is:

$$\binom{T}{0} + \binom{T}{1} + \cdots + \binom{T}{T} = 2^T.$$

Note that it is assumed that the set of candidate models includes the true one. Thus, as the objective is to select the true model, appears to be more suitable to use the BIC criterion, which is the most widely used consistent criteria. Therefore, given the model $M_{t_1,\ldots,t_m}$, which assumes m outliers at locations $t_1, \ldots, t_m$, the parameters to estimate are the $(p + q + 1) \times 1$ vector of unknown parameters of the ARMA(p, q) model, $\alpha_{p,q}$, and the $m \times 1$ vector of unknown sizes of the outliers, $w_{t_1}, \ldots, w_{t_m}$. In summary, the model have $p+q+m+1$ unknown parameters, that are included in the $(p + q + m + 1) \times 1$ vector $\theta_{t_1,\ldots,t_m} = \left(\alpha'_{p,q}, w_{t_1}, \ldots, w_{t_m}\right)'$. The maximum likelihood estimators of the vector of parameters $\theta_{t_1,\ldots,t_m}$ of the model $M_{t_1,\ldots,t_m}$ are denoted by $\widehat{\theta}_{t_1,\ldots,t_m} = \left(\widehat{\alpha}'_{p,q}, \widehat{w}_{t_1}, \ldots, \widehat{w}_{t_m}\right)'$ and are obtained after maximizing the log-likelihood given by:

$$\log p\left(y \mid M_{t_1,\ldots,t_m}\right) = \tag{1.6}$$
$$= -\frac{T}{2}\log 2\pi - \frac{1}{2}\log\left|\Sigma_{t_1,\ldots,t_m}\right| - \frac{1}{2}\left(y - w_{t_1,\ldots,t_m}\right)' \Sigma^{-1}_{t_1,\ldots,t_m}\left(y - w_{t_1,\ldots,t_m}\right),$$

where $\Sigma_{t_1,\ldots,t_m}$ is the $T \times T$ covariance matrix of y under the model $M_{t_1,\ldots,t_m}$ and $w_{t_1,\ldots,t_m}$ is a $T \times 1$ vector whose components are, the outliers sizes at the components $t_1, \ldots, t_m$, and are 0, elsewhere. The solution provided by the BIC for this problem is to choose the model which minimizes:

$$BIC\left(M_{t_1,\ldots,t_m}\right) = -2\log p\left(y \mid \widehat{\theta}_{t_1,\ldots,t_m}\right) + (p+q+m+1) \times \log T,$$

where $\log p\left(y \mid \widehat{\theta}_{t_1,\ldots,t_m}\right)$ is the maximized log-likelihood of y under the $M_{t_1,\ldots,t_m}$ model. Galeano and Peña (2007*b*) shows that the BIC above defined tends to select models with many outliers although the series have no one. The reason of this behavior appears to be that the number of candidate models is 2^T which is much larger than the sample size T, even for small values of T. This conclusion also holds if the number of possible outliers in the sample is bounded to be less than T.

The second approach is the following. Let M_m be the ARMA(p,q) model with m outliers at unknown locations $t_1, \ldots, t_m$. In this case, the problem of joint estimation of the model parameters, number of outliers, their locations and their sizes can also be now stated as the selection of the true model among the set of candidates ones, which include the model without outliers, denoted by M, the model with one outlier, denoted by M_1, etc... In general, there is only one candidate model with m outliers. Thus, the total number of candidate models is $T+1$. The model M_m, which assumes m outliers at unknown locations $t_1, \ldots, t_m$, have the following parameters to estimate: the $(p+q+1)\times 1$ vector of parameters of the ARMA(p,q) model, $\alpha_{p,q}$, the $m \times 1$ vector of sizes of the outliers, $w_{t_1}, \ldots, w_{t_m}$ and the $m\times 1$ vector of locations of the outliers, $t_1, \ldots, t_m$. In summary, the model have $p+q+2m+1$ unknown parameters, that are included in the $(p+q+2m+1)\times 1$ vector $\theta_m = \left(\alpha'_{p,q}, w_{t_1}, \ldots, w_{t_m}, t_1, \ldots, t_m\right)'$. The maximum likelihood estimators of the vector of parameters θ_m of the model M_m are denoted by $\widehat{\theta}_m = \left(\widehat{\alpha}'_{p,q}, \widehat{w}_{t_1}, \ldots, \widehat{w}_{t_m}, \widehat{t}_1, \ldots, \widehat{t}_m\right)'$, and are obtained after maximizing the log-likelihood of y under the M_m model given by:

$$\log p\left(y \mid M_m\right) = -\frac{T}{2}\log 2\pi - \frac{1}{2}\log\left|\Sigma_m\right| - \frac{1}{2}\left(y - w_m\right)' \Sigma_m^{-1}\left(y - w_m\right),$$

where Σ_m is the $T \times T$ covariance matrix of y under the model M_m and w_m is a $T \times 1$ vector whose components are, the outliers sizes at the components $t_1, \ldots, t_m$, and are 0, elsewhere. The BIC will choose the model which minimizes:

$$BIC\left(M_m\right) = -2\log p\left(y \mid \widehat{\theta}_m\right) + (p+q+2m+1) \times \log T,$$

where $\log p\left(y \mid \widehat{\theta}_m\right)$ is the maximized log-likelihood of y under the M_m model. Although the number of candidate models have been substantially reduced, Galeano and Peña (2007*b*) showed that the BIC tends to select models with no outliers even with for series affected by outliers. Thus, the BIC also fails to consistently selects the true model with this approach. Note that the BIC approximation of the posterior probability of each model does not verify the conditions given by Kass, Tierney and Kadane (1990) in the sense that the locations are not differentiable parameters. Thus the BIC approximation of this problem does not have a theoretical justification.

The third possibility is the following. The idea is to take into account the hierarchical structure of the problem. In other words, first, make inference on the number of outliers in the sample, m, then, on the locations of the outliers given m, $t_1, \ldots, t_m|m$, and finally, on the parameters given m and $t_1, \ldots, t_m$, $\alpha_{p,q}, w_{t_1}, \ldots, w_{t_m}|t_1, \ldots, t_m, m$. Thus, the focus of interest are the posterior probabilities, $p\left(m|y\right)$, $p\left(t_1, \ldots, t_m|y, m\right)$ and $p\left(\alpha_{p,q}, w_{t_1}, \ldots, w_{t_m}|t_1, \ldots, t_m, m, y\right)$.

First, it can be shown that the marginal distribution of the number of outliers given the sample is given by:

$$p\left(m|y\right) = \frac{\sum\limits_{t_1,\ldots,t_m} \pi\left(t_1, \ldots, t_m, m\right) p\left(y|M_{t_1,\ldots,t_m}\right)}{\sum\limits_{i=0}^{T} \sum\limits_{t_1,\ldots,t_i} \pi\left(t_1, \ldots, t_i, i\right) p\left(y|M_{t_1,\ldots,t_i}\right)},$$

where $\pi\left(t_1, \ldots, t_m, m\right)$ is the prior distribution of the parameters $t_1, \ldots, t_m, m$ and $p\left(y|M_{t_1,\ldots,t_m}\right)$ is the distribution of the time series y given the model $M_{t_1,\ldots,t_m}$, as defined in the first approach. On the other hand, the posterior probability of model $M_{t_1,\ldots,t_m}$ given the data is given by:

$$p\left(M_{t_1,\ldots,t_m}|y\right) = \frac{\pi\left(t_1, \ldots, t_m, m\right) p\left(y|M_{t_1,\ldots,t_m}\right)}{\sum\limits_{i=0}^{T} \sum\limits_{t_1,\ldots,t_i} \pi\left(t_1, \ldots, t_i, i\right) p\left(y|M_{t_1,\ldots,t_i}\right)},$$

which implies that:

$$p\left(y|M_{t_1,\ldots,t_m}\right) = \frac{\sum\limits_{i=0}^{T} \sum\limits_{t_1,\ldots,t_i} \pi\left(t_1, \ldots, t_i, i\right) p\left(y|M_{t_1,\ldots,t_i}\right)}{\pi\left(t_1, \ldots, t_m, m\right)} p\left(M_{t_1,\ldots,t_m}|y\right).$$

This shows that the posterior probability of the number of outliers, $p\left(m|y\right)$, can be written as follows:

$$p\left(m|y\right) = \sum_{t_1,\ldots,t_m} p\left(M_{t_1,\ldots,t_m}|y\right),$$

and it is not necessary to specify the prior probabilities $\pi\left(t_1, \ldots, t_m, m\right)$.

Second, the BIC of the model $M_{t_1,\ldots,t_m}$ is an approximation of minus two times the posterior probability given the time series y. In other words,

$$-2 \log p\left(M_{t_1,\ldots,t_m}|y\right) \simeq BIC\left(M_{t_1,\ldots,t_m}\right) =$$
$$= -2 \log p\left(y \mid \widehat{\theta}_{t_1,\ldots,t_m}\right) + \left(p+q+m+1\right) \times \log T,$$

so that,

$$p\left(M_{t_1,\ldots,t_m}|y\right) \simeq \exp\left(-\frac{BIC\left(M_{t_1,\ldots,t_m}\right)}{2}\right). \tag{1.7}$$

Therefore, using the BIC approximation (1.7), $p\left(m|y\right)$ can be approximated as follows:

$$\begin{aligned} p\left(m|y\right) &\simeq \sum_{t_1,\ldots,t_m} \exp\left(-\frac{BIC\left(M_{t_1,\ldots,t_m}\right)}{2}\right) = \\ &= \exp\left(-\frac{(p+q+m+1)}{2}\log T\right) \sum_{t_1,\ldots,t_m} p\left(y \mid \widehat{\theta}_{t_1,\ldots,t_m}\right), \end{aligned}$$

where $\widehat{\theta}_{t_1,\ldots,t_m}$ are the maximum likelihood estimates of the vector of parameters of the model $M_{t_1,\ldots,t_m}$, $\theta_{t_1,\ldots,t_m} = \left(\alpha'_{p,q}, w_{t_1}, \ldots, w_{t_m}\right)'$ obtained after maximizing $\log p\left(y \mid M_{t_1,\ldots,t_m}\right)$ given in (1.6).

Finally, the BIC for the number of outliers m can be defined as the approximation of $-2\log p\left(m|y\right)$ as follows:

$$\begin{aligned} BIC\left(m\right) &= -2\log\left(\exp\left(-\frac{(p+q+m+1)}{2}\log T\right)\sum_{t_1,\ldots,t_m} p\left(y \mid \widehat{\theta}_{t_1,\ldots,t_m}\right)\right) = \\ &= -2\log\left(\sum_{t_1,\ldots,t_m} p\left(y \mid \widehat{\theta}_{t_1,\ldots,t_m}\right)\right) + (p+q+m+1)\log T. \end{aligned}$$

In summary, the number of outliers is selected as the number which provides the minimum value of $BIC\left(m\right)$. After that, inference on the vector of unknown locations, $t_1, \ldots, t_m$, is done by using the distribution $p\left(M_{t_1,\ldots,t_m}|y, m\right)$, which can be written as follows:

$$\begin{aligned} p\left(M_{t_1,\ldots,t_m}|y, m\right) &= \frac{p\left(y|M_{t_1,\ldots,t_m}\right)}{\sum\limits_{t_1,\ldots,t_m} p\left(y|M_{t_1,\ldots,t_m}\right)} = \frac{p\left(M_{t_1,\ldots,t_m}|y\right)}{\sum\limits_{t_1,\ldots,t_m} p\left(M_{t_1,\ldots,t_m}|y\right)} \simeq \\ &\simeq \frac{\exp\left(-\frac{BIC\left(M_{t_1,\ldots,t_m}\right)}{2}\right)}{\sum\limits_{t_1,\ldots,t_m} \exp\left(-\frac{BIC\left(M_{t_1,\ldots,t_m}\right)}{2}\right)} = \frac{p\left(y \mid \widehat{\theta}_{t_1,\ldots,t_m}\right)}{\sum\limits_{t_1,\ldots,t_m} p\left(y \mid \widehat{\theta}_{t_1,\ldots,t_m}\right)}. \end{aligned}$$

Thus, the estimates of the unknown locations, $t_1, \ldots, t_m$, are the ones that attains the largest value of $p\left(y \mid \widehat{\theta}_{t_1,\ldots,t_m}\right)$. Finally, estimation of the vector of parameters once that m and $t_1, \ldots, t_m$ have been selected is carried out with the maximum likelihood estimates $\widehat{\theta}_{t_1,\ldots,t_m} = \left(\widehat{\alpha}'_{p,q}, \widehat{w}_{t_1}, \ldots, \widehat{w}_{t_m}\right)'$.

Galeano and Peña (2007*b*) presented a detailed treatment on the advantages of using this approach. Also, in order to avoid the computation of the maximum likelihood estimates of all the models, these authors proposed an algorithm that only requires to compute the maximum likelihood estimates of the models with largest posterior probabilities.

References

1. Akaike, H. (1969). Fitting autoregressive models for prediction, *Annals of the Institute of Statistical Mathematics,* **21**, 243–247.

2. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, In *Proceedings of the 2nd International Symposium on Information Theory*, pp. 267–281, Akademiai Kiadó, Budapest.

3. Barnett, V. and Lewis, T. (1993). *Outliers in Statistical Data, 3 ed*, John Wiley and Sons, Chichester.

4. Chang I., Tiao, G. C. and Chen, C. (1988). Estimation of time series parameters in the presence of outliers, *Technometrics,* **3**, 193–204.

5. Chen, C. and Liu, L. (1993). Joint estimation of model parameters and outlier effects in time series, *Journal of the American Statistical Association*, **88**, 284–297.

6. Fox, A. J. (1972). Outliers in time series, *Journal of the Royal Statistical Society B,* **34**, 350–363.

7. Galeano, P. and Peña, D. (2000). Multivariate analysis in vector time series, *Resenhas*, **4**, 383-403.

8. Galeano, P. and Peña, D. (2007*a*). On the connection between model selection criteria and quadratic discrimination in ARMA time series models, *Mimeo.*

9. Galeano, P. and Peña, D. (2007*b*). Outlier detection by Bayesian Information Criterion, *Mimeo.*

10. 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.

11. Hurvich, C. M., Shumway, R. and Tsai, C. (1990). Improved estimators of Kullback-Leibler information for autoregressive model selection in small samples, *Biometrika,* **77**, 709–719.

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

13. Kass, R., Tierney, L. and Kadane, J. (1990) The validity of posterior expansions based on Laplace's method, In *Bayesian and likelihood methods in Statistics and Econometrics*, (Ed., S. Geisser, J. S. Hodges, S. J. Press and A. Zellner), pp. 473–478, Amsterdam: New Holland.

14. Le, N. D., Martin, R. D. and Raftery, A. E. (1996). Modeling flat stretches, bursts, and outliers in time series using mixture transition distribution models, *Journal of the American Statistical Association*, **91**, 1504–1515.

15. Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models, *Biometrika,* **65**, 297–303.

16. Luceño, A. (1998). Detecting possibly non-consecutive outliers in industrial time series, *Journal of the Royal Statistical Society B,* **60**, 295–310.

17. Monti, A. C. (1994). A proposal for residual autocorrelation test in linear models, *Biometrika,* **81**, 776–780.

18. Peña, D. and Rodríguez, J. (2003). A powerful portmanteau test of lack of fit for time series, *Journal of the American Statistical Association,* **97**, 601–619.

19. Peña, D. and Rodríguez, J. (2006). The log of the determinant of the autocorrelation matrix for testing goodness of fit in time series, *Journal of the Statistical Planning and Inference,* **136**, 2706–2718.

20. Sánchez, M. J. and Peña, D. (2003). The identification of Multiple Outliers in ARIMA models, *Communications in Statistics: Theory and Methods,* **32**, 1265–1287.

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

22. Tsay, R. S. (1986). Time Series Model Specification in the Presence of Outliers, *Journal of the American Statistical Association*, **81**, 132–141.

23. Velilla, S. (1994). A goodness-of-fit test for autoregressive moving-average models based on the standardized sample spectral distribution of the residuals, *Journal of Time Series Analysis*, **15**, 637–647.