

AUTOMATIC MODELING METHODS FOR UNIVARIATE SERIES

Víctor Gómez and Agustín Maravall

AUTOMATIC MODELING METHODS FOR UNIVARIATE SERIES

Víctor Gómez and Agustín Maravall

Banco de España - Servicio de Estudios
Documento de Trabajo nº 9808

In publishing this series the Banco de España seeks to disseminate studies of interest that will help acquaint readers better with the Spanish economy.

The analyses, opinions and findings of these papers represent the views of their authors: they are not necessarily those of the Banco de España.

The Banco de España is disseminating some of its principal reports via INTERNET and INFOVÍA.

The respective WWW server addresses are:
<http://www.bde.es> and <http://www.bde.inf>.

ISSN: 0213-2710

ISBN: 84-7793-597-1

Depósito legal: M. 10665-1998

Imprenta del Banco de España

_____ Abstract

In this article, a unified approach to automatic modeling for univariate series is presented. First, ARIMA models and the classical methods for fitting these models to a given time series are reviewed. Second, some objective methods for model identification are considered and some algorithmical procedures for automatic model identification are described. Third, outliers are incorporated into the model and an algorithm for automatic model identification in the presence of outliers is proposed. Finally, the previous algorithm is extended to cope with missing observations, trading day and Easter effects, intervention and regression effects, and transfer functions.

CONTENTS

- 1 INTRODUCTION
 - 1.1 ARIMA Models
 - 1.2 The Three Stages of Box and Jenkins' Methodology
 - 1.3 The Difficulties of Model Identification
 - 1.4 The Desirability of Automatic Modeling Methods
- 2 CLASSICAL MODEL IDENTIFICATION METHODS
 - 2.1 Box and Jenkins' Method
 - 2.2 The Inverse Autocorrelation Method
 - 2.3 Subjectivity of the Classical Methods
 - 2.4 The Difficulties With Mixed ARMA Models
- 3 OBJECTIVE MODEL IDENTIFICATION METHODS
 - 3.1 Unit Root Testing
 - 3.2 Penalty Function Methods
 - 3.3 Pattern Identification Methods
 - 3.4 Uniqueness of the Solution and the Purpose of Modeling
- 4 TOOLS FOR AUTOMATIC MODEL IDENTIFICATION
 - 4.1 Tests for the Log-Level Specification
 - 4.2 Regression Techniques for Estimating Unit Roots
 - 4.3 The Hannan and Rissanen's Method
 - 4.3.1 Computation of $BIC_{p,q}$
 - 4.3.2 Optimization of $BIC_{p,q}$
 - 4.4 Liu's Filtering Method
 - 4.5 The LTF method for Transfer Function Identification

- 5 AUTOMATIC MODELING METHODS IN THE PRESENCE OF OUTLIERS
 - 5.1 Outlier Definition
 - 5.2 Algorithms for Automatic Outlier Detection and Correction
 - 5.2.1 Estimation and Preadjustment for the Effect of an Outlier
 - 5.2.2 The Case of Multiple Outliers
 - 5.2.3 Estimation of the Standard Deviation σ of the Residuals
 - 5.3 Estimation and Filtering Techniques to Speed up the Algorithms
 - 5.4 The Need to Robustify Automatic Modeling Methods
 - 5.5 An Algorithm for Automatic Model Identification in the Presence of Outliers

- 6 AN AUTOMATIC PROCEDURE FOR THE GENERAL REGRESSION-ARIMA MODEL IN THE PRESENCE OF OUTLIERS, SPECIAL EFFECTS AND, POSSIBLY, MISSING OBSERVATIONS
 - 6.1 Missing Observations
 - 6.2 Trading Day and Easter Effects
 - 6.3 Intervention and Regression Effects
 - 6.4 Transfer Functions

REFERENCES

1 Introduction

1.1 Arima Models

Many of the time series found in industry, economics, social sciences and many other scientific fields can be represented, at least as a first approximation, by means of a stationary linear process, possibly after having applied to the data some suitable transformation and differencing. In this article, the term stationary means wide sense stationary, not strict stationary. See Brockwell and Davis (1992), pp. 11-14. Other synonyms for wide sense stationary are second order stationary and weakly stationary. The process $\{z_t\}$ is stationary if the mean and the autocovariances are given by

1. $E(z_t) = \mu$
2. $E[(z_{t-j} - \mu)(z_t - \mu)] = \gamma_j, \quad j = 0, 1, \dots$

For a stationary process, the autocorrelation function is given by

$$\rho_j = \frac{\gamma_j}{\gamma_0}, \quad j = 0, 1, \dots$$

According to a famous theorem due to Wold, every non-deterministic stationary process $\{z_t\}$ admits the representation

$$z_t = \mu + \sum_{j=0}^{\infty} \psi_j a_{t-j},$$

where $\{a_t\}$ is a sequence of uncorrelated variables of mean zero and variance σ_a^2 (white noise) and $\sum_{j=0}^{\infty} |\psi_j|^2 < \infty$. Modifying σ_a^2 if necessary, we can suppose $\psi_0 = 1$.

It follows from the theorem of Stone-Weierstrass of functional analysis, that the function $f(z) = \sum_{j=0}^{\infty} \psi_j z^j$ can be approximated with any degree of accuracy by a quotient of polynomials

$$f(z) = \sum_{j=0}^{\infty} \psi_j z^j \approx \frac{1 + \theta_1 z + \dots + \theta_q z^q}{1 + \phi_1 z + \dots + \phi_p z^p}$$

This suggests that every non-deterministic stationary process can be approximated by an AutoRegressive-Moving-Average (ARMA) process:

$$z_t + \phi_1 z_{t-1} + \dots + \phi_p z_{t-p} = C + a_t + \theta_1 a_{t-1} + \dots + \theta_q a_{t-q}. \quad (1)$$

If B is the backshift operator $Bz_t = z_{t-1}$, this expression can be written more compactly as $\phi(B)z_t = C + \theta(B)a_t$, or $\phi(B)(z_t - \mu) = \theta(B)a_t$, where $\phi(B) = 1 + \phi_1 B + \dots + \phi_p B^p$ and $\theta(B) = 1 + \theta_1 B + \dots + \theta_q B^q$. The process (1) is denoted by ARMA(p, q). It can be shown that (1) is stationary if, and only if, all the roots of the autoregressive polynomial $\phi(B)$ lie outside the unit circle. See Brockwell and Davis (1992).

With seasonal time series, multiplicative models are often used

$$\phi(B)\Phi(B^s)z_t = C + \theta(B)\Theta(B^s)a_t,$$

where s is the number of seasons.

ARIMA models are nonstationary models such that after some degree of differencing they become stationary ARMA models. The I in the acronym stands for “integrated”, because integration is the inverse operation of differencing. It turns out that ARIMA models can be successfully used in practice to represent many time series, especially in economics. The general seasonal multiplicative ARIMA model is

$$\phi(B)\Phi(B^s)\nabla^d\nabla_s^D z_t = C + \theta(B)\Theta(B^s)a_t, \quad (2)$$

where, C is a constant, s is the number of seasons, $d = 0, 1, 2$, $D = 0, 1$, $\nabla = 1 - B$ is a regular difference and $\nabla_s = 1 - B^s$ is a seasonal difference. Instead of z_t , it may be necessary to use $\log(z_t)$, or some other transformation, to stabilize the variance of the series. If p and P are the degrees of the autoregressive polynomials $\phi(B)$ and $\Phi(B)$, and q and Q those of the moving average polynomials $\theta(B)$ and $\Theta(B)$, model (2) is denoted as a multiplicative $(p, d, q)(P, D, Q)_s$ model.

An ARIMA model (2) is said to be invertible if all the roots of the moving average polynomial $\theta(B)\Theta(B^s)$ lie outside the unit circle. If this is the case, assuming $C = 0$ for simplicity, the model can be written as $z_t + \sum_{j=1}^{\infty} \pi_j z_{t-j} = a_t$.

ARIMA models are widely used for the following reasons, among others,

1. they can represent many processes with a parsimonious model -
2. they can be extended to incorporate the modeling of deterministic effects (interventions), outliers, trading day and Easter effects, etc.
3. a well established procedure for modeling has been developed

1.2 The Three Stages of Box and Jenkins' Methodology

ARIMA models became very popular after the publication of the seminal book by Box and Jenkins (1976), whose first edition was published in 1970. In this book, a systematic procedure was for the first time proposed to handle the problem of time series modeling by means of iterative cycles consisting of

- i) Model identification
- ii) Model estimation
- iii) Diagnostic checking

At the same time, in Box and Jenkins' book of 1976 some computer programs were given to implement this methodology with the aid of digital computers. In this way, given the big computing power of computers, the user was for the first time able to apply a new and powerful machinery which would prove very useful in the field of time series analysis. The arrival of personal computers contributed further to increase enormously the demand for these new tools.

At the identification stage one chooses a particular model from the class (2), that is, one selects values for p, d, q, P, D and Q . Classical tools for model identification include plotting the data over time; the sample autocorrelation function, the sample inverse autocorrelation function and the sample partial autocorrelation function.

At the estimation stage of the model building cycle, the coefficients of the identified model are estimated using some nonlinear optimization routine. Usually, three estimation methods for ARMA models are considered

- Conditional least squares
- Unconditional least squares
- Exact maximum likelihood

Fast recursive procedures for likelihood evaluation are available. This, together with the speed of modern computers, makes it advisable to use the exact maximum likelihood method.

Two recursive procedures for likelihood evaluation deserve special mention. These are the innovations algorithm of Brockwell and Davis (1992) and the fast Kalman filter algorithm of Morf, Sidhu and Kailath

(1974), as described by Pearlman (1980) and improved by Mélard (1984).

Example Suppose that we want to estimate the parameters of the MA(1) model $z_t = a_t + \theta a_{t-1}$, $t = 1, \dots, N$, where a_t is distributed $N(0, \sigma^2)$, and let $z = (z_1, \dots, z_N)'$ be the observed series and $\text{Var}(z) = \sigma^2 \Sigma$. Then, the log-likelihood is, up to a constant,

$$\lambda(z) = -\frac{1}{2} \{N \ln(\sigma^2) + \ln |\Sigma| + z' \Sigma z / \sigma^2\}.$$

The maximum likelihood estimator of σ^2 is $\hat{\sigma}^2 = z' \Sigma z / N$ and the σ^2 -maximized log-likelihood $l(z)$ is, up to a constant,

$$l(z) = -\frac{1}{2} \{N \ln(\hat{\sigma}^2) + \ln |\Sigma|\}.$$

Then,

- the conditional least squares method minimizes $z' \Sigma z$, under the assumption that $a_0 = 0$. That is, it minimizes $\sum_{t=1}^N a_t^2$, where the a_t are recursively obtained from $a_t = z_t - \theta a_{t-1}$, $a_0 = 0$.
- the unconditional least squares method minimizes $z' \Sigma z$, but estimating by maximum likelihood the initial observation a_0 . Box and Jenkins did that using the so-called “backcasting”. Modern recursive methods, like the Kalman filter or the innovations algorithm, yield

$$\begin{aligned} \hat{a}_{t+1} &= z_{t+1} - \frac{\theta}{D_t} \hat{a}_t \\ D_{t+1} &= 1 + \theta^2 - \frac{\theta^2}{D_t}, \end{aligned}$$

with the initial conditions $\hat{a}_1 = z_1$ and $D_1 = 1 + \theta^2$. Then, the method minimizes $z' \Sigma z = \sum_{t=1}^N (\hat{a}_t^2 / D_t)$. It can be shown that $D_t \rightarrow 1$ as $t \rightarrow \infty$.

- the exact maximum likelihood method minimizes $z' \Sigma z |\Sigma|^{1/N}$, where, with the above notation, $z' \Sigma z = \sum_{t=1}^N (\hat{a}_t^2 / D_t)$ and $|\Sigma| = \prod_{t=1}^N D_t$.

There is a need for good starting values for the optimization method used to estimate the model parameters. One good option is the

Hannan–Rissanen’s method, which will be described later, see Hannan and Rissanen (1982). It is computationally cheap, because it uses only linear regressions, and it provides estimates that are close to the exact maximum likelihood estimates.

Finally, diagnostic checks are applied to determine whether or not the chosen model adequately represents the data. If this is not the case, the whole iterative cycle of identification, estimation, and diagnostic checking is repeated until a satisfactory model is obtained.

There are many ways in which we can check the adequacy of the model we have estimated. The most basic assumption made in ARIMA models is that the errors $\{a_t\}$ are independently and normally distributed.

The residuals $\{\hat{a}_t\}$ can be checked in a number of ways. To name a few, we can inspect the plot of $\{\hat{a}_t\}$, examine the sample autocorrelation and partial autocorrelation function of $\{\hat{a}_t\}$, use the Ljung–Box statistic, check the squared residuals, perform some tests of randomness, check for normality, etc.

1.3 The Difficulties of Model Identification

The modeling procedure proposed by Box and Jenkins (1976) was by no means a process that could be fully automated with the help of computers. In particular, step i) of the procedure, which involves model identification, was rather an art which required an expert in time series analysis to carry it out. It was certainly the most difficult of the three steps.

We will describe later the most common tools used nowadays for ARIMA model identification but, in spite of the advances that have taken place in the last two decades, we can say that step i) continues to be the most difficult of the three steps. Besides, we have to be aware of the fact that the majority of time series encountered in practice usually have outliers, which makes even more difficult the modeling procedure.

We will also review later the most important methods for automatic detection and correction of outliers that are currently in use. As far as ARIMA model identification is concerned, the presence of outliers can make it very difficult due to the important biases induced by the outliers in the parameter estimates and in the sample autocorrelation and partial autocorrelation functions. For this reason, any

good strategy for ARIMA model identification has to account for the presence of outliers.

1.4 The Desirability of Automatic Modeling Methods

There are many reasons to try to automate as much as possible the ARIMA model identification stage, but they can be basically reduced to two. The first one is that one should eliminate as much as possible all mundane and mechanical chores, which can be performed by the computer, thus increasing the analyst's productivity. If the user is an accomplished analyst, he may invest more of his precious time on troublesome data sets that he has to model. On the contrary, if he is not an expert in time series models, he can use a powerful methodology that he couldn't even dream of using before. The second reason has to do with the objectivity of the identification stage, since it is desirable that this stage be not subject to heuristic methods and ad-hoc procedures that vary with each time series expert. For example, if a National Statistical Office has to produce some statistical data which require the modeling of some time series datasets and an expert is involved in the production process who uses subjective techniques, it may be criticized for publishing data which are neither objective nor reproducible.

2 Classical Model Identification Methods

In this article, we review the classical model identification methods, with particular emphasis in the Box and Jenkins' method. It is pointed out that the classical methods are subjective, in the sense that the results depend to a great degree on the analyst's experience and background. In addition, it is noted that the tools proposed by Box and Jenkins for model identification work well for pure moving average or pure autoregressive models, but not so well for mixed ARMA models.

2.1 Box and Jenkins' Method

As stated previously, in the model identification stage we try to determine “appropriate” orders for p, d, q, P, D and Q of the $\text{ARIMA}(p, d, q) \times (P, D, Q)_s$ model, possibly after the application of some suitable transformation to the data to stabilize the variance. The transformations used belong to the class of power transformations

$$z_t^{(\lambda)} = \begin{cases} (z_t^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \ln z_t & \text{if } \lambda = 0 \end{cases}$$

of Box and Cox (1964). It is very often the case that only the logarithmic transformation ($\lambda = 0$) is considered, versus the alternative of no transformation ($\lambda = 1$). Economists usually prefer to use the logarithmic transformation to linearize the relationship between the input variables and the output in a regression. In this way, the parameters can be interpreted as the elasticity between the variables.

Assuming that the series has been transformed if necessary by means of a Box-Cox transformation, the first step in the Box and Jenkins' method is to determine whether or not to difference the series. To achieve this, apart from the plot of the data, the sample autocorrelation function (SACF)

$$c_j = \frac{\sum_{t=1}^{N-j} (z_t - \bar{z})(z_{t+j} - \bar{z})}{\sum_{t=1}^N (z_t - \bar{z})^2}, \quad 0 \leq j < N,$$

where $\bar{z} = \sum_{t=1}^N z_t / N$, is an indispensable tool.

If a process requires some differencing to induce stationarity, it can be shown that the sample autocorrelations will typically behave as a very smooth function, and thus not die out rapidly, at high lags. The failure of the SACF to die out at high lags thus indicates that differencing is required. This is so even though the first few sample autocorrelations need not necessarily be large.

After having determined the stationary transformation, that is the operator $\nabla^d \nabla_s^D$ in model (2), the following step is to find an ARMA model for the stationary series $u_t = \nabla^d \nabla_s^D z_t$. The SACF and sample partial autocorrelation function (SPACF) are the fundamental tools to obtain the orders of the autoregressive and moving average parts. The partial autocorrelation function is defined by $\alpha_1 = \text{Corr}(u_2, u_1)$ and

$$\alpha_j = \text{Corr}[u_{j+1} - E(u_{j+1}|u_2, \dots, u_j), u_1 - E(u_1|u_2, \dots, u_j)], \quad j \geq 2$$

where $E(\cdot|\cdot)$ denotes orthogonal projection and Corr stands for correlation. It can be shown that the partial autocorrelations α_j can be estimated by successive $\text{AR}(j)$ fittings. That is, if we fit

$$u_t + \hat{\phi}_{j1}u_{t-1} + \cdots + \hat{\phi}_{jj}u_{t-j} = \hat{C} + a_t$$

for $j = 1, 2, \dots$, then $\hat{\alpha}_j = \hat{\phi}_{jj}$ is the SPACF.

It can also be shown that if $\{u_t\}$ is a pure autoregressive model, i.e., $q = Q = 0$ in (2), the theoretical partial autocorrelation function dies out in an exponential or sinusoidal fashion and has the cut-off feature, which means that the partial autocorrelations α_j will obey $\alpha_k = 0$ for all $k > p + P$. Also, if $\{u_t\}$ is a pure moving average model, i.e., $p = P = 0$ in (2), the theoretical autocorrelation function dies out in an exponential or sinusoidal fashion and has the cut-off feature, which means that the autocorrelations ρ_j will obey $\rho_k = 0$ for all $k > q + Q$.

Estimated autocorrelations can have rather large variances and can be highly autocorrelated with each other. Therefore, one has to be careful when working with the SACF because detailed adherence to the theoretical autocorrelation function cannot be expected. It may be the case that moderately large estimated autocorrelations occur after the theoretical autocorrelation function has damped out, or apparent ripples and trends may be present in the SACF which have no basis in the theoretical function.

If $\{u_t\}$ is a pure moving average process, the elements of the sequence $\{\sqrt{N}c_j\}$, $j > q + Q$, are asymptotically normal with mean 0 and variance $\hat{\sigma}^2(c_j) = 1 + 2 \sum_{k=1}^{q+Q} c_k^2$, $j > q + Q$. Then, an approximate 100(1 - α) percent confidence interval of ρ_j is $\left(-z_{\alpha/2}\hat{\sigma}(c_j)/\sqrt{N}, z_{\alpha/2} \times \hat{\sigma}(c_j)/\sqrt{N}\right)$, $j > q + Q$, where $z_{\alpha/2}$ is the z -value such that the area $\alpha/2$ lies to its right in the standard normal probability density.

If $\{u_t\}$ is a pure autoregressive process, then the elements of the sequence $\{\sqrt{N}\hat{\phi}_{jj}\}$, $j > p + P$, are asymptotically independent random variables with mean 0 and variance 1. Therefore, an approximate 100(1 - α) percent confidence interval of ϕ_{jj} is $\left(-z_{\alpha/2}/\sqrt{N}, z_{\alpha/2}/\sqrt{N}\right)$, $j > p + P$.

For mixed ARMA models, the identification can be more difficult if only the SACF and SPACF of the series are available. Box and Jenkins (1976) provided some information on how to determine the orders of the autoregressive and moving average parts from "reading"

the SACF and SPACF of a stationary series. However, this approach is usually not very effective in practice.

2.2 The Inverse Autocorrelation Method

The inverse autocorrelation function (IACF) is the autocorrelation function associated with the reciprocal of the spectral density of a stationary time series, assuming the model is invertible. This means that if $\{z_t\}$ follows the invertible ARMA model $\phi(B)z_t = \theta(B)a_t$ model, then the inverse autocorrelation function of this model is the autocorrelation function (ACF) of the dual model $\theta(B)z_t = \phi(B)a_t$. See Cleveland (1972) and Chatfield (1979) for more details on the inverse autocorrelation function.

Note that if the original model is a pure autoregressive model, then the IACF is an ACF corresponding to a pure moving average model. Thus, the IACF has the cut-off feature and behaves similarly to the partial autocorrelation function.

To estimate the sample IACF (SIACF), first a high-order autoregressive model is fit to the data by means of the Yule-Walker equations. These are best solved using the Durbin-Levinson algorithm, which will be described later in this article. Then, the SIACF is calculated as the autocorrelation function that corresponds to a moving average model which has as moving average polynomial the previously estimated autoregressive polynomial. The order of the long autoregressive polynomial can be chosen by using the BIC or some other penalty function criterion of the type that will be considered in the next section.

The SIACF generally indicates subset and seasonal autoregressive models better than the SPACF. However, more calculations are necessary to obtain the SIACF than to obtain the SPACF.

It can be shown that the SIACF is asymptotically normally distributed, so that confidence intervals can be constructed.

The confidence interval of the IACF depends heavily on the estimation method and the asymptotic variance of the IACF estimate is more complicated than that of the partial autocorrelation function (PACF) estimate, which is $1/N$. Therefore, there is more statistical error when estimating a confidence interval of the IACF than of the PACF.

2.3 Subjectivity of the Classical Methods

The Box and Jenkins' method for model identification relies heavily on the inspection of plots of data over time and the inspection of the graphs of the sample autocorrelation and partial autocorrelation functions. As we saw in the last section, the sample inverse autocorrelation function can also be used for model identification.

However, these last tools can be effective to identify pure autoregressive or pure moving average models, but not so effective with mixed ARMA models. Besides, the determination of the stationary transformation, that is, the numbers d and D in (2), can be very difficult.

With the exception of very few cases in which the data show a very distinctive pattern, it is usually rather difficult to identify a model for the series at hand. For example, given a sample of finite length, it may be extremely difficult to distinguish between the nonstationary model $(1 - .7B)\nabla z_t = a_t$ and the process $(1 - 1.704B + .706B^2)z_t = (1 - .715B)(1 - .989B)a_t$, which has very similar coefficients but for which the autoregressive polynomial has all its zeros outside the unit circle.

Therefore, it is most probable that several time series experts who use the Box and Jenkins' method, when confronted with the same data, will specify different models. This makes the whole process of classical model identification dependent on the person who applies the techniques. The more experience he has, the more likely he is to choose an adequate model for the series. This subjectivity is inherent in the classical model identification methods.

2.4 The Difficulties With Mixed Arma Models

It has already been mentioned in the last section that the sample autocorrelation, partial autocorrelation and inverse autocorrelation functions can effectively identify pure autoregressive and pure moving average models. On the other hand, when both the degrees of the autoregressive polynomial $(p + P)$ and the moving average polynomial $(q + Q)$ are not zero, the previous functions are much more difficult to interpret. In this case, other model identification methods, different from the classical methods, are called for.

The difficulty of identifying mixed ARMA models is further increased when seasonality is also present in the time series at hand.

Several major advances have been made in the last two decades to identify ARIMA models for non-seasonal time series. Among these, we can mention the extended autocorrelation function and the smallest canonical correlation methods developed by Tsay and Tiao (1984, 1985). These methods are very informative in the identification of ARIMA models for non-seasonal time series, but they are less successful when they are directly applied to seasonal time series. It is to be noted that these methods can also be used with nonstationary series.

Since the early 1970s, some penalty function methods have been proposed for ARMA model identification. These methods can be used with seasonal time series and their popularity is constantly increasing. The reason for this is that they can be effective, computationally cheap and objective. However, although some results have been extended to nonstationary series, these methods are in principle only applicable to stationary series.

3 Objective Model Identification Methods

In this section, we deal with objective model identification methods, in contrast to the subjective model identification methods considered in the last section. First, in order to obtain the degrees d and D of the stationary transformation in (2), we can use unit root tests. Then, several methods can be applied to identify an ARMA model for the stationary (differenced) series. We review in this section the penalty function and the pattern identification methods. Both of these methods are regarded as objective. A good reference for ARMA model identification is the book by Choi (1992).

3.1 Unit Root Testing

In the Box and Jenkins methodology, the decision concerning the need for differencing is based upon the characteristics of the plot of the data and of its sample autocorrelation function. For example, failure of this last function to die out sufficiently quickly indicates that differencing is required.

In recent times, there has been a growing interest in more formal inference procedures concerning the appropriateness of differencing

operators in the model. Since all the roots of the differencing operators $\nabla = 1 - B$ and $\nabla_s = 1 - B^s$ lie on the unit circle, testing for differencing is usually referred to as unit root testing.

It is interesting to note that, as Dickey and Pantula (1987) point out, the results obtained by several authors suggest that overdifferencing is not a problem as far as forecasting is concerned. However, there appear to be uses for unit root tests in investigating some economic hypothesis. The practical implication of this is that when one is interested in the routine treatment of many series with forecasting purposes, one should not care very much about whether or not some of the series are overdifferenced. It is our practical experience that much the same thing happens with regard to model based seasonal adjustment. We can say that overdifferencing is compensated by moving average parameters that go to unity.

We will not review here the vast amount of existing literature concerning unit root testing. The reader can consult, for example, Reinsel (1997) and the references therein. We will content ourselves with making a few remarks on existing procedures.

The two “classical” unit root tests of Dickey–Fuller and Phillips–Perron tend to exhibit rather poor behaviour in the presence of certain types of serial correlation. See the Monte Carlo analysis of Schwert (1989).

When there is no seasonality in the series at hand and only regular differences, that is, differences of the form ∇^d , are considered, it seems that the sequential testing procedure suggested by Dickey and Pantula (1987) is the best strategy to follow. According to these authors, only tests that compare a null hypothesis of k unit roots with an alternative of $k - 1$ unit roots are considered. In the sequential procedure, one should start with the largest k under consideration and work down, that is, decrease k by one each time the null hypothesis is rejected.

The situation is different for seasonal time series. In this case, further research is needed and no general agreement exists on how to proceed as far as unit root testing is concerned. It seems that a generalization of the Dickey and Pantula (1987) approach to the seasonal case would be an interesting topic to investigate.

3.2 Penalty Function Methods

In the identification stage, once the differencing orders d and D in (2) have been obtained for the nonstationary series $\{z_t\}$, the problem remains of finding an ARMA model for the differenced series $u_t = \nabla^d \nabla_s^D z_t$.

Since the early 1970s, some procedures to determine the orders k and i of an ARMA(k, i) model have been proposed which minimize a function of the form

$$P(k, i) = \ln \hat{\sigma}_{k,i}^2 + (k + i) \frac{C(M)}{M}, \quad k \leq K, \quad i \leq I, \quad (3)$$

where $\hat{\sigma}_{k,i}^2$ is the maximum likelihood estimate of the variance of the white noise variance, $C(M)$ is some function of the number of observations M of the series, and K and I are upper bounds for the orders, usually imposed a priori. Because $\hat{\sigma}_{k,i}^2$ decreases as the orders increase, it cannot be a good criterion to choose the orders minimizing it. This is the reason why the penalty term $(k + i)C(M)/M$ is included. The penalty function identification methods are regarded as objective.

If $C(M)$ in (3) is replaced with 2, we obtain the famous AIC criterion, which stands for Akaike's Information Criterion. Other possible choices are $C(M) = \ln M$, which corresponds to the BIC (Bayesian Information Criterion), and $C(M) = 2\ln(\ln M)$, which gives the HQ criterion (Hannan and Quinn). The BIC criterion imposes a greater penalty term than does AIC.

One criterion for selection of AR(p) models is the FPE (Final Prediction Error) criterion, which is given by $\text{FPE}(p) = \{1 + (p/M)\}\hat{\sigma}_p^2$.

The BIC criterion estimates the orders of an ARMA model consistently, whereas the AIC does not. However, this is not a reason to prefer BIC instead of AIC because consistency is based on the assumption that there is a "true" ARMA model for the series and this is doubtful proposition. Models are artificial constructs and probably there is no such a thing as a true model.

It is our practical experience and also the experience of some other authors, like, for example, Lütkepohl (1985), that the BIC criterion works better in practice than AIC, in terms of selecting more often the original model when working with simulated series and selecting models with a better fit when working with real series.

Although the penalty function methods are in principle computationally expensive, because they need maximum likelihood estimates

for all possible ARMA models, there are methods, like the Hannan-Rissanen's method, which use cheaper estimates based on linear regression techniques only. Also, in the case of multiplicative seasonal ARMA models, it will be seen that it is possible to further reduce the computational burden by proceeding sequentially. That is, by iterating between selections of the regular and of the seasonal parts.

The penalty function methods can also be used to identify vector ARMA models. The penalty functions to use with multivariate data are direct generalizations of the ones for the univariate case. This is a great advantage, not shared by many of the other identification methods.

3.3 Pattern Identification Methods

Since the early 1980s, some methods have been applied for determining the orders of an ARMA process which use the extended Yule Walker equations. For the ARMA(p, q) process (1), these last equations are given by

$$\gamma_j = -\phi_1\gamma_{j-1} - \dots - \phi_p\gamma_{j-p}, \quad j = q+1, q+2, \dots,$$

where γ_j , $j = 0, 1, \dots$, is the covariance function of the process. These methods are often called pattern identification methods. See Choi (1992). It is to be noted that, contrary to Choi's remark about penalty function methods being computationally exorbitant and pattern identification methods being computationally cheap, it will be shown later in this article that the proposed sequential application of Hannan-Rissanen's method, which is based on the BIC criterion, for stationary seasonal models is computationally cheap and can be very effective.

The pattern identification methods are so called because they are based on certain functions which give rise to two-way arrays with distinctive patterns. For each ARMA(p, q) model, the corresponding two-way array shows a unique pattern. Using the sample analog of this two-way array, an ARMA model is identified by looking for a theoretical pattern which is closely resembled by the sample one. Among the many pattern identification methods which have been proposed in the literature, we can mention the R and S array method by Gray, Kelley and McIntire (1978), the Corner method by Beguin, Gourieroux and Monfort (1980), the extended sample autocorrelation method by Tsay and Tiao (1984) and the smallest canonical correlation method

by Tsay and Tiao (1985). These last two methods can be effective with non-seasonal time series and can also be used with nonstationary series. However, the R and S array method and the Corner method, which can only be used with stationary series, do not seem to be very useful even for data with no seasonality. The Corner method has been applied to identification of transfer function models by Liu and Hanssens (1982).

3.4 Uniqueness of the Solution and the Purpose of Modeling

In the identification stage of model building, it is often the case that there are several models for which the fit is acceptable. For example, if the BIC criterion is used, there may be a very small difference between the BIC of an AR(2) model and the BIC of an ARMA(1, 1) model. In this case, we can probably use any of these two competitive models to model the data.

When some competitive models exist, one should try to choose the more parsimonious one. That is, the one with less parameters. On occasion, it may be useful to choose models that are also balanced. This means that the degree of the autoregressive part, included the nonstationary transformation, equals the degree of the moving average part. Balanced models are useful when one is going to perform model based seasonal adjustment.

In summary, models should be considered as artificial constructs, which are useful for certain purposes, but are only a crude approximations to reality. In this respect, the criteria used to select models, especially when some competing models exist, depend on the applications. Some criteria may be good for forecasting, but not so good for signal extraction, for example. One should always have in mind that usually there is not a unique solution to the identification problem.

4 Tools for Automatic Model Identification

In this section, some practical procedures will be presented for automatic model identification. The emphasis is in the word practical, so

that the methods presented will aim at simplicity, efficiency and speed when applied to real data.

We will start with two tests which are proposed for the log-level specification. The first one is based on the maximum likelihood principle when the series is supposed to follow the model $(0, 1, 1)(0, 1, 1)_s$. This is the airline model of Box and Jenkins (1976). The reason why we choose that model in this and other tests later in this article is that it encompasses many other models and is a model very often found in practice. The second test is based on a range-mean regression which is robust against outliers.

We will then review the two-stage method proposed by Gómez (1997) to estimate unit roots. After that, the Hannan-Rissanen's method, hereafter referred to as HR's method, is reviewed. This method is used to identify an ARMA model for the stationary (differenced) series. It is based on the BIC criterion and is computationally cheap. Some improvements to HR's method, proposed by Gómez (1997), will be described. Finally, the method proposed by Liu for ARIMA model identification and the LTF method for Transfer Function Identification are reviewed.

4.1 Tests for the Log-Level Specification

The first test for the log-level specification is based on the maximum likelihood estimation of the parameter λ in the Box-Cox transformations. We fit an airline model with mean to the data, first in logs ($\lambda = 0$) and then without logs ($\lambda = 1$). Let $z = (z_1, \dots, z_N)'$ be the differenced series and let T be a transformation of the data, which can be any of the Box-Cox transformations. It is assumed that $T(z)$ is normally distributed with mean zero and $\text{Var}(T(z)) = \sigma^2 \Sigma$. Then, the logarithm of the density function $f(z)$ of z is

$$\ln(f(z)) = k - \frac{1}{2} \{ N \ln(\sigma^2) + \ln |\Sigma| + T(z)' \Sigma^{-1} T(z) / \sigma^2 + \ln(1/J(T))^2 \},$$

where k is a constant and $J(T)$ is the jacobian of the transformation. Considering the parameter λ in the T transformation fixed, the previous density function is maximized first with respect to the other model parameters. It is easy to see that σ^2 can be concentrated out of this function by replacing it with the maximum likelihood estimator $\hat{\sigma}^2 =$

$T(z)' \Sigma^{-1} T(z) / N$. The concentrated function is

$$l(z) = -\frac{1}{2} \{ N \ln(T(z)' \Sigma^{-1} T(z)) + \ln(1/J(T))^2 \} + \dots,$$

where the dots indicate terms that do not depend on the value of λ associated with the transformation T . After having maximized with respect to all model parameters different from λ , we maximize with respect to λ . Denoting the sum of squares $T(z)' \Sigma^{-1} T(z)$ by $S(z, T)$, the maximum likelihood principle leads to the minimization of the quantity $S(z, T)(1/J(T))^{2/N}$. It is easy to see that $(1/J(T))^{1/N}$ is the geometric mean in the case of the logarithmic transformation, and unity in the case of no transformation. Therefore, the test compares the sum of squares of the model without logs with the sum of squares multiplied by the square of the geometric mean in the case of the model in logs. Logs are taken in case this last function is the minimum.

The other test is the following. First, the data is divided into groups of successive observations of length l equal to a multiple of the number of observations per year s . That is, the first group is formed with the first l observations, the second group is formed with observations $l + 1$ to $2l$, etc. Then, for each group, the observations are sorted and the smallest and largest observations are rejected. This is done for protection against outliers. With the other observations, the range and mean are computed. Finally, a range-mean regression of the form $r_i = \alpha + \beta m_i + u_i$ is performed. The criterion is the slope β . If β is greater than a specified value, logs are taken.

4.2 Regression Techniques for Estimating Unit Roots

Let the observed series $\{z(t)\}$ follow the ARIMA(p, d, q) model

$$\phi(B)\delta(B)(z(t) - \mu) = \theta(B)a(t), \quad (4)$$

where $\phi(B) = 1 + \phi_1 B + \dots + \phi_p B^p$, $\delta(B) = 1 + \delta_1 B + \dots + \delta_d B^d$ and $\theta(B) = 1 + \theta_1 B + \dots + \theta_q B^q$ are polynomials in the backshift operator B of degrees p, d and q , $\{a(t)\}$ is a i.i.d. $N(0, \sigma^2)$ sequence of random variables and μ is the mean of the process. The roots of $\delta(B)$ are assumed to lie on and those of $\phi(B)$ outside the unit circle, so that the process $w(t) = \delta(B)z(t)$ follows a stationary ARMA(p, q)

process. As mentioned earlier, most economic series follow so-called multiplicative seasonal models, where

$$\begin{aligned}\delta(B) &= \nabla^d \nabla_s^D, \\ \phi(B) &= \phi_r(B) \phi_s(B^s), \\ \theta(B) &= \theta_r(B) \theta_s(B^s),\end{aligned}\tag{5}$$

s is the number of observations per year, $\nabla^d = (1 - B)^d$ and $\nabla_s^D = (1 - B^s)^D$. In practice, for economic time series, the inequalities $0 \leq d \leq 2$ and $0 \leq D \leq 1$ hold. For simplicity, we will use in the rest of the Section the notation (4), even for multiplicative seasonal models. We will make specific reference to these models when necessary.

In the following, a procedure to obtain the differencing orders is reviewed which is based on the estimation of unit roots. This last procedure is the first step of an automatic model identification method which has been proposed by Gómez (1997) and is implemented in programs TRAMO and SEATS, see Gómez and Maravall (1996). The estimation of the unit roots is done by first estimating autoregressive models of the form

$$(1 + \phi_1 B + \phi_2 B^2)(1 + \Phi B^s)(z(t) - \mu) = a(t),\tag{6}$$

where $\{z(t)\}$ is the observed series, s is the number of observations per year, μ in the mean of the process, B is the backshift operator, $Bz(t) = z(t - 1)$, and $\{a(t)\}$ is a sequence of i.i.d. $N(0, \sigma^2)$ random variables. Then, the series is differenced using the differencing orders given by the unit roots obtained after estimating (6) and an ARMA(1, 1) \times (1, 1) $_s$ model with mean, that is, a model of the form

$$(1 + \phi B)(1 + \Phi B^s)(x(t) - \mu) = (1 + \theta B)(1 + \Theta B^s)a(t),\tag{7}$$

is fitted to the differenced series $\{x(t)\}$. If any new unit roots appear after estimating (7), the differencing orders are properly increased and a new model (7) is fitted. The process is continued until no more unit roots are found. Then, the residuals of the last estimated model are used to decide whether to specify a mean for the model or not. The choice of models (6) and (7) will be justified later.

Suppose that the series $\{z(t)\}$ follows model (4), where it is assumed $\mu = 0$ to simplify matters. Then, by theorems 3.2 and 4.1 of Tiao and Tsay (1983), the ordinary least squares, henceforth OLS,

estimators obtained from an $AR(k)$ regression, where $k \geq d$, asymptotically verify

$$\hat{\Phi}_k(B) \doteq \hat{\delta}(B)\hat{\phi}_m(B),$$

where \doteq denotes asymptotic equivalence in probability, $m = k - d$ and $\hat{\Phi}_k(B)$, $\hat{\delta}(B)$ and $\hat{\phi}_m(B)$ are, respectively, the polynomials estimated by OLS in the autoregressions

$$\begin{aligned}\Phi_k(B)z(t) &= a(t), \\ \delta(B)z(t) &= a(t), \\ \phi_m(B)w(t) &= a(t),\end{aligned}$$

where $w(t) = \delta(B)z(t)$ is a stationary process that follows the $ARMA(p, q)$ model $\phi(B)w(t) = \theta(B)a(t)$ and the subindex in $\Phi_k(B)$ and $\phi_m(B)$ denotes the polynomial degree. In addition, the equality $\hat{\delta}(B) = \delta(B) + O_p(N^{-1})$ holds, where N is the series length.

The practical implication of this result is that if we perform an autoregression of order greater than or equal to the (unknown) degree of the polynomial $\delta(B)$, we obtain a consistent estimate of $\delta(B)$ as a component of $\hat{\Phi}_k(B)$. If we specify a model of the form $AR(2) \times (1)_s$, for $\Phi_k(B)$, we cover the cases $\delta(B) = 1$, $\delta(B) = \nabla$, $\delta(B) = \nabla_s$, $\delta(B) = \nabla\nabla_s$ and $\delta(B) = \nabla^2\nabla_s$, which are the ones of most applied interest.

In the case of non-seasonal models, where $\delta(B) = \nabla^d$ and $0 \leq d \leq 2$ is assumed, if we specify an $AR(2)$ model, all important cases are covered.

Based on the previous considerations, the algorithm to identify the differencing polynomial is the following.

- I) Specify a model of the form $AR(2) \times (1)_s$, with mean, given by equation (6) if the process is multiplicative seasonal, or an $AR(2)$ model with mean, also given by (6) but without the second factor, if the process is regular. This autoregressive process is estimated using HR's method, which will be described later, unless the user decides to use unconditional least squares. If the roots estimated with HR's method lie outside the unit circle, the autoregression is estimated again using unconditional least squares. A root is considered to be a unit root if its modulus is greater than a specified value, which by default is .97. Go to II).
- II) In addition to the differencing degrees identified in I) as a result of the estimated unit roots, a model of the form $ARMA(1, 1) \times$

$(1, 1)$, with mean for seasonal series, or a model $\text{ARMA}(1, 1)$ with mean for non-seasonal series, is specified. Letting $x(t)$ be the series that results from differencing $z(t)$ with the differencing polynomial obtained after the estimation of the initial autoregression, the equations for these models are given by (7) in the seasonal case, and by (7) without the factors involving B^s in the regular case. The model is estimated using HR's method or exact maximum likelihood, depending on the option chosen by the user, and if any of the estimated autoregressive parameters is close to 1, the degree of differencing is increased accordingly. A parameter is considered to be close to 1 if its modulus is greater than a specified value, which by default is .88. To avoid cancelation of terms in the model, the absolute value of the difference between each autoregressive parameter and its corresponding moving average parameter should be greater than .15. For multiplicative seasonal models, it is not possible to pass from 0 differencing to $\nabla\nabla$, directly. If this happens, the roots of the autoregressive polynomial obtained in I) are considered again, the one with greatest modulus is chosen, and the series is differenced accordingly. If the series has been differenced in this step, repeat II). Otherwise, go to III).

- III) Using the residuals of the last estimated model, it is decided whether to specify a mean for the model of the series or not depending on the significance of the estimated residual mean. Stop.

The $\text{ARMA}(1,1) \times (1,1)$, model used in II) is very flexible and constitutes a generalization of the airline model of Box and Jenkins (1976). For stationary series, it approximates well many of the ARMA models encountered in practice. When it is used with nonstationary series, it can detect autoregressive unit roots which have not been detected by the autoregressive model used I). Imagine, for example, a model of the form $(1 - B)(z(t) - \mu) = (1 - .8B)a(t)$, where the autoregressive and the moving average part almost cancel out. In this case, an $\text{ARMA}(1,1)$ model would probably estimate the unit root better than an $\text{AR}(2)$ model.

Consider now the case of a regression model with ARIMA errors. The question naturally arises as to whether the previous analysis is still valid and if, in consequence, the procedure just described is also

applicable in this case. In this respect, the results of Tsay (1984), pp. 119–120, allows one to state that, under very general conditions, it is possible to work with the original series in order to identify the differencing polynomial.

4.3 The Hannan and Rissanen’s Method

After having obtained the stationary transformation, the next step in the model building process is the identification of an $\text{ARMA}(p, q)$ model for the differenced series, possibly corrected for outliers and other regression effects. We will start by considering that there are neither outliers nor other regression effects and we will extend the results later in this article to the general case.

In the following, the HR’s method and a procedure to identify $\text{ARMA}(p, q)$ models based on it are reviewed. This last procedure is the second step of an automatic model identification method which has been proposed by Gómez (1997) and is implemented in programs TRAMO and SEATS, see Gómez and Maravall (1996). The HR’s method is a penalty function method based on the BIC criterion, where the estimates of ARMA model parameters are computed by means of linear regressions. Therefore, these estimates are computationally cheap, although it can be shown that the estimators have similar properties to those obtained by maximum likelihood. See Hannan and Rissanen (1982).

Let $z = (z(1), \dots, z(N))'$ the observed series, which follows model (4), where we assume $\mu = 0$ for simplicity. After $\delta(B)$ has been identified, we can compute the differenced series $w(t) = \delta(B)z(t)$, $t = d + 1, \dots, N$, which follows the $\text{ARMA}(p, q)$ model

$$\phi(B)w(t) = \theta(B)a(t), \quad (8)$$

where $\phi(B)$, $\theta(B)$ and $\{a(t)\}$ are like in (4). If the model is multiplicative seasonal, the decomposition (5) holds. In order to avoid notational problems, let the differenced series be $w = (w(1), \dots, w(N - d))'$. If the orders of the fitted model (8) are (p, q) , the BIC statistic is

$$\text{BIC}_{p,q} = \log(\hat{\sigma}_{p,q}^2) + (p + q) \log(N - d)/(N - d), \quad (9)$$

where $\hat{\sigma}_{p,q}^2$ is the maximum likelihood estimator of σ^2 . The criterion estimates the orders (p, q) by choosing (\hat{p}, \hat{q}) which minimizes (9).

The method just described to choose the orders, which is based on the traditional BIC criterion, is computationally expensive because one has to perform a nonlinear optimization for each (p, q) to compute $\hat{\sigma}_{p,q}^2$. For this reason, Hannan and Rissanen (1982) propose to perform the estimation using linear regression techniques in three steps, although the third step is used to compute estimators of the ARMA model chosen by the BIC criterion which have similar properties to the maximum likelihood estimators. Therefore, only the first two steps are used to select the orders (p, q) .

4.3.1 Computation of $\text{BIC}_{p,q}$

In the first step of HR's method, which takes place only if there is a moving average part ($q > 0$), estimates $\hat{a}(t)$ of the innovations $a(t)$ in (8) are obtained by fitting a long autoregressive model to the series. That is, given a big n , the $\hat{a}(t)$ are computed using

$$\hat{a}(t) = \sum_{j=0}^n \hat{\phi}_n(j)w(t-j), \quad \hat{\phi}_n(0) = 1, \quad t \geq 1,$$

where $w(t) = 0$ if $t \leq 0$ and the $\hat{\phi}_n(j)$ are computed using Durbin-Levinson's algorithm. This last algorithm consists of first estimating the sample autocorrelations

$$c(t) = \frac{1}{N-d} \sum_{s=1}^{N-d-t} w(s)w(s+t)$$

and then recursively computing the $\hat{\phi}_n(j)$ using the equations

$$\begin{aligned} \hat{\phi}_n(n) &= - \sum_{j=0}^{n-1} \hat{\phi}_{n-1}(j)c(n-j)/\hat{\sigma}_{n-1}^2, \\ \hat{\phi}_n(j) &= \hat{\phi}_{n-1}(j) + \hat{\phi}_n(n)\hat{\phi}_{n-1}(n-j), \\ \hat{\sigma}_n^2 &= \{1 - \hat{\phi}_n^2(n)\}\hat{\sigma}_{n-1}^2, \quad \hat{\phi}_1(1) = c(1)/c(0), \quad \hat{\sigma}_0^2 = c(0). \end{aligned}$$

In the procedure proposed by Gómez (1997), the value of n is chosen to be $n = \max\{\lceil \log^2(N-d) \rceil, 2\max\{p, q\}\}$, where (p, q) are the orders of the ARMA model for which the BIC is being computed and $\lceil \log^2(N-d) \rceil$ is the integer part of $\log^2(N-d)$. This choice is based on the fact

that Hannan and Rissanen (1982), p. 88, assume that n is greater than $\log(N - d)$, but not greater than $\log^b(N - d)$, for some $b < \infty$.

In the second step of HR's method, given the orders (p, q) , first the parameters of model (8) are estimated by minimizing

$$S(p, q) = \sum_{t=m}^{N-d} \left\{ \sum_{j=0}^p \phi_j w(t-j) - \sum_{j=1}^q \theta_j \hat{a}(t-j) \right\}^2, \quad (10)$$

where $m = \max\{p+1, q+1\}$ and $\phi_0 = 1$. Then, the estimator $\hat{\sigma}_{p,q}^2$ is computed by the formula $\hat{\sigma}_{p,q}^2 = S(p, q)/(N - d)$ and the BIC $_{p,q}$ statistic is computed using (9).

In the procedure proposed by Gómez (1997), the following modifications are made. If there is no moving average part ($q = 0$), the estimation of the parameters of the ARMA model finishes here. Note that, in this case, the estimates obtained for the autoregressive part coincide with the ones obtained by OLS.

If there is a moving average part ($q > 0$), the estimators $\hat{\phi}_j$ and $\hat{\theta}_j$, obtained by minimizing (10), are consistent but have a bias and, therefore, they are not asymptotically efficient. In order to obtain bias-corrected, consistent and asymptotically efficient estimators, see Zhao-Guo (1985), first form

$$\tilde{a}(t) = - \sum_{j=1}^q \tilde{\theta}_j \tilde{a}(t-j) + \sum_{j=0}^p \tilde{\phi}_j w(t-j), \quad t \geq 1,$$

where $\tilde{a}(t) = 0$ and $w(t) = 0$ if $t \leq 0$. Then put

$$\eta(t) = - \sum_{j=1}^p \tilde{\phi}_j \eta(t-j) + \tilde{a}(t), \quad \xi(t) = - \sum_{j=1}^q \tilde{\theta}_j \xi(t-j) + \tilde{a}(t), \quad t \geq 1,$$

where $\eta(t) = 0$ and $\xi(t) = 0$ if $t \leq 0$. Finally, regress $\tilde{a}(t)$ on $-\eta(t-j)$, $j = 1, \dots, p$, and $\xi(t-j)$, $j = 1, \dots, q$. The estimated regression coefficients are added to the estimators $\hat{\phi}_j$ and $\hat{\theta}_j$ to obtain the desired estimators $\tilde{\phi}_j$ and $\tilde{\theta}_j$.

When there are a moving average part ($q > 0$) and an autoregressive ($p > 0$), Gómez (1997) proposes to obtain better estimates of the moving average part by repeating the previous procedure with the series filtered with the autoregressive filter. That is, the series is first filtered with the autorregressive filter $\hat{\phi}(B)$, which has been estimated in the two previous steps, to obtain the series $x(t) = \hat{\phi}(B)w(t)$. Then,

the series $x(t)$, which asymptotically follows the model $x(t) = \theta(B)a(t)$ and, therefore, does not have an autoregressive part, is subject to the two previous steps.

Once the parameter estimates of model (8) have been obtained for some orders (p, q) , the estimator $\hat{\sigma}_{p,q}^2$ is needed to compute the $\text{BIC}_{p,q}$ statistic. In the procedure proposed by Gómez (1997), the residuals $r(t)$, $t = 1, \dots, n = \max\{p, q\}$ of the series $w(t)$ are first computed using a fast Kalman filter routine based on the algorithm of Morf, Sidhu and Kailath (1974). Then, the rest of the residuals $r(t)$, $t = n + 1, \dots, N - d$ are recursively obtained using the difference equation (8). Finally, the estimator $\hat{\sigma}_{p,q}^2$ is computed by the formula

$$\hat{\sigma}_{p,q}^2 = \frac{1}{N-d} \sum_{t=1}^{N-d} r^2(t),$$

and the $\text{BIC}_{p,q}$ statistic is computed using (9).

4.3.2 Optimization of $\text{BIC}_{p,q}$

After having described the algorithm to compute $\text{BIC}_{p,q}$ for each (p, q) , we now review the algorithms used by HR's method and the procedure proposed by Gómez (1997) to obtain the optimal model of the form (8). In HR's method, the model is chosen as that $\text{ARMA}(\bar{p}, \bar{q})$ model for which $\text{BIC}_{\bar{p}, \bar{q}}$ is minimum among all $\text{ARMA}(p, q)$ models satisfying $p \leq P$ and $q \leq Q$, where P and Q are fixed upper bounds. These authors recommend to search first among models with $p = q$ and refine the search later.

To describe the procedure proposed by Gómez (1997), suppose the general case, where the series follows a multiplicative seasonal model given by (5). In practice, it is assumed that the orders of the $\text{ARMA}(p_r, q_r) \times (p_s, q_s)_s$ model followed by the series verify $0 \leq p_r, q_r \leq 3$ and $0 \leq p_s, q_s \leq 2$, and the BIC statistic should be computed for all these combinations. Since the resulting number of combinations is high, the search is performed sequentially. The algorithm is

- I) Specify first an $\text{ARMA}(3,0)$ model for the regular part. Then, compute the BIC statistic for models where the seasonal part verifies $0 \leq p_s, q_s \leq m_s$, and choose the minimum. The number m_s is chosen by the user, the default value being 1.

- II) Fix the seasonal part to that chosen in I), compute the BIC statistic for models where the regular part verifies $0 \leq p_r, q_r \leq m_r$, and choose the minimum. The number m_r is chosen by the user, the default value being 3.
- III) Fix the regular part to that chosen in II), compute the BIC statistic for models where the seasonal part verifies $0 \leq p_s, q_s \leq m_s$, and choose the minimum. The number m_s is that of I).

The previous algorithm allows for a substantial reduction in computing time and, however, the results obtained with it are very satisfactory. Once the previous algorithm has finished, and in order to avoid the tendency of BIC to overparametrize, especially in the seasonal part, the smallest five BIC are first ordered in ascending order. Then, the first one is compared to the other four and if the difference in absolute value is less than a certain number and the biggest of the two BIC corresponds to a more parsimonious seasonal part, this last one is chosen. Among all the BIC that satisfy this condition, the one that corresponds to the more parsimonious part is chosen, provided that the seasonal part exists ($p_s > 0$ or $q_s > 0$). The procedure also favours balanced models (models where the degrees of the autoregressive and the moving average parts coincide).

In the previous algorithm, if the parameters estimated for an ARMA model are such that the roots of the autoregressive or the moving average polynomials lie within the unit circle, this fact is considered as an indication of model inadequacy and the model is rejected.

The tentative model ARMA(3,0) specified in I) of the previous algorithm seems to be robust and the sequential search of the algorithm has given very satisfactory results in all performed tests of the proposed procedure, with real and simulated series.

If there is a mean or other regression effects in model (8), the procedure proposed by Gómez (1997) obtains first OLS estimators of the regression parameters. Then, these effects are subtracted from the differenced series before computing the parameter estimates of model (8) and also before computing the residuals $r(t)$ needed in the computation of $\hat{\sigma}_{p,q}^2$ and the BIC statistic.

4.4 Liu's Filtering Method

Recently, the SCA software package has incorporated a module for automatic ARIMA model identification, called "SCA-Expert". This module uses a procedure based on the filtering method proposed by Liu (1989) and certain heuristic rules. Briefly, this method consists of the following.

1. Examine first the sample autocorrelation functions (SACF) of $z(t)$, $(1 - B)z(t)$, $(1 - B^s)z(t)$ and $(1 - B)(1 - B^s)z(t)$ to assert the differencing orders and to see whether seasonality is present. After that, examine the SACF of the properly differenced series. If an obvious seasonal ARIMA model can be specified from the SACF, stop. Otherwise, go to the following step. Denote by $y(t)$ the differenced series.
2. If an obvious tentative model cannot be deduced from the SACF of $y(t)$, estimate an intermediate model of the type $\text{ARMA}(1, 1) \times (1, 1)_s$. If no one of the autoregressive parameters has is close to 1, generate the series $R(t)$ and $S(t)$, which are the result of filtering $y(t)$ with the $\text{ARMA}(1, 1)_s$ and $\text{ARMA}(1, 1)$ models that make up the intermediate model.

If any of the autoregressive parameters is close to 1, then difference properly. After differencing, a new intermediate model of the same type is estimated and new $R(t)$ and $S(t)$ series are generated.

3. Use the sample autocorrelation and partial autocorrelation functions, as well as the extended sample autocorrelation function, of $R(t)$ to identify an ARMA model adequate for the $R(t)$ series.
4. In order to identify a model for $S(t)$, the SACF of $S(t)$ can be used. If a model is not clear for $S(t)$, examine also the estimated parameters for the seasonal part in the intermediate model and use them to specify a model for $S(t)$.

This procedure can be criticized for using heuristic rules that are not documented.

4.5 The LTF method for Transfer Function Identification

The general form of the single-input, single-output transfer function model can be expressed as

$$z_t = C + \frac{\omega(B)}{\gamma(B)}x_t + n_t, \quad (11)$$

where $\{n_t\}$ follows an ARIMA model, $\omega(B) = (\omega_0 + \omega_1 B + \dots + \omega_{m-1} B^{m-1})B^b$ and $\gamma(B) = 1 + \gamma_1 B + \dots + \gamma_r B^r$.

In practice, the number of terms in $\omega(B)$ is small and the value of r is usually 0 or 1. The rational operator $\nu(B) = \omega(B)/\gamma(B)$ can be written as $\nu(B) = \sum_{i=0}^{\infty} \nu_i B^i$.

The linear transfer function (LTF) identification method proposed by Liu and Hanssens (1982) can be described as follows. The transfer function model (11) can be approximated by the following linear model

$$z_t = C + (\nu_0 + \nu_1 B + \nu_2 B^2 + \dots + \nu_k B^k)x_t + n_t, \quad (12)$$

where k is a sufficiently large number. Based on the representation (12), the method consists of the following three steps

1. Initially estimate (12) for a “sufficiently” large value of k and a “reasonable” approximation for n_t . These are discussed below.
2. Examine the estimates of the parameters in the model for n_t and the residuals in the fitted equation. The estimated parameters may indicate that differencing is necessary. The residuals are used to discover any gross discrepancies in the model.
3. Use the estimated ν_j to determine the form of the transfer function. This is done by means of the corner method of Beguin, Gouriéroux and Monfort (1980). In addition, examine the residuals $\hat{n}_t = z_t - \hat{C} - \sum_{i=0}^k \nu_i x_{t-i}$ from the fitted model to determine an appropriate ARIMA model for n_t .

There are two key elements in the LTF method. These are the choice of k and the proxy used for n_t in (12). The latter is usually an autoregressive, if there is no seasonality, or multiplicative seasonal autoregressive model. The choice for k is somewhat arbitrary. There should be enough weights to account for the longest lagged response between input and output. The choice depends also on the sample size.

5 Automatic Modeling Methods in the Presence of Outliers

Many time series encountered in practice have outlying observations. These may be due to errors in the data, strikes, changes in regulations, etc. The presence of outliers can make extremely difficult the process of model identification. For this reason, any automatic model identification method has to incorporate some kind of outlier treatment.

In this section, we start first with the definition of an outlier. Second, after examining some algorithms for outlier treatment, we review the method proposed by Gómez (1997) for automatic outlier detection and correction. Third, some estimation and filtering techniques are reviewed which are used to speed up the algorithms of the previous methods. Fourth, some reasons are given for the need to robustify automatic modeling methods. Finally, an algorithm is proposed for automatic model identification in the presence of outliers.

5.1 Outlier Definition

When analysing time series data, it is not unusual to find outlying observations due to uncontrolled or unexpected interventions, like strikes, major changes in political or economic policy, the occurrence of a disaster, gross errors, and so forth. Since ARIMA models, which are frequently used in time series modeling, are designed to grasp the information of processes with a homogeneous memory pattern, the presence of outlying observations or structural changes may influence the efficiency and goodness of fit of these models. See, for example, Abraham and Box (1979), Chen and Tiao (1990), Tsay (1986), and Guttman and Tiao (1978).

The traditional approach to handle the problem of outliers, once it is assumed that a proper ARIMA model has been correctly identified for the series, consists of first identifying the location and the type of outliers and then use the intervention analysis proposed by Box and Tiao (1975). This procedure requires that a time series expert first examines the data and then, with the help of some time series software, analyses the sample autocorrelation and partial autocorrelation functions of the residuals, graphs of the series and the residuals, etc. For this reason, it is important to try to find some procedure which automates as much as possible the process of detection and correction

of outliers. Among the first steps in this direction, we can mention the procedures of Chang, Tiao and Chen (1988), Hillmer, Bell and Tiao (1983), and Tsay (1988). These procedures are quite effective in detecting the locations and estimating the effects of large isolated outliers. However, the problem is not solved because

- 1) The presence of outliers may result in an incorrectly specified model
- 2) Even if the model is appropriately specified, outliers may still produce important biases in parameter estimates
- 3) Some outliers may not be identified due to a masking effect

The method proposed by Tsay (1986) is an important contribution to solve the problem of model identification in the presence of outliers. Chen and Liu (1993) have proposed a method for outlier treatment that tries to solve problems 2) and 3). This method works rather satisfactorily, although it presents some aspects that can be improved. For example, the method uses exact maximum likelihood estimation several times, thus making it computationally expensive. It does not use exact residuals. The algorithm is too complicated. Multiple regressions are performed without filtering the data and the columns of the design matrix by an “exact” filter, like the Kalman filter. It uses instead a conditional filter. Finally, the method uses a sort of backward elimination procedure to select the “best regression equation”, instead of a stepwise procedure which is more robust.

The method proposed by Gómez (1997) for the detection and correction of outliers attempts to solve problems 2) and 3) in such a way that the above mentioned aspects in the procedure of Chen and Liu (1993) are improved, as will be described later. In addition, if it is used in the algorithm proposed at the end of this Section, together with the algorithm for automatic model identification proposed by Gómez (1997), the proposed algorithm constitutes an alternative procedure to that of Tsay (1986) to solve the problem of automatic model identification in the presence of outliers that complements and may improve considerably Tsay’s procedure.

Suppose first that there are no regression effects. We will extend the results to the general case later. Let the series $\{z(t)\}$ follow the $\text{ARIMA}(p, d, q)$ model given by (4), where it is understood that if the model is multiplicative seasonal, (5) holds. We will assume for simplicity that $\mu = 0$ and we will use the notation (4), referring to (5)

when necessary. To model the effect of an outlier, consider the model

$$z^*(t) = z(t) + \omega\nu(B)I_t(T), \quad (13)$$

where $\nu(B)$ is a quotient of polynomials in B , $\{z(t)\}$ is the outlier free series, $I_t(T) = 1$ if $t = T$ and $I_t(T) = 0$ otherwise, is an indicator function to refer to the time in which the outlier takes place and ω represents the magnitude of the outlier. Four types of outliers will be considered

IO: $\nu(B) = \theta(B)/(\delta(B)\phi(B))$,

AO: $\nu(B) = 1$,

TC: $\nu(B) = 1/(1 - \delta B)$,

LS: $\nu(B) = 1/(1 - B)$.

The acronyms stand for innovational outlier (IO), additive outlier (AO), temporary change (TC) and level shift (LS). The value of δ is considered fixed and is made equal to 0.7. For a more detailed discussion on the nature and motivation for these outliers, see Chen and Tiao (1990), Fox (1972), Hillmer, Bell and Tiao (1983), and Tsay (1988). These four outliers correspond to some simple types of outliers. More complicated outliers can be usually approximated by combinations of these four types.

5.2 Algorithms for Automatic Outlier Detection and Correction

We will start by considering that there is only one outlier. After having described how the effect of the outlier can be estimated and adjusted, the case of multiple outliers will be considered. Finally, the algorithm proposed by Gómez (1997) will be reviewed.

5.2.1 Estimation and Adjustment for the Effect of an Outlier

Suppose that the parameters in model (4) are known, the observed series is $z^* = (z^*(1), \dots, z^*(N))'$, the outlier free series is $z = (z(1), \dots, z(N))'$ and put $Y = (\nu(B)I_1(T), \dots, \nu(B)I_N(T))'$. Then, (13) can be written as the regression model with ARIMA errors

$$z^* = Yw + z. \quad (14)$$

To simplify the exposition, we will assume that z in (14) follows an ARMA model or, what amounts to the same thing, $\delta(B) = 1$ in (4). If this is not the case, we would work with the series obtained by differencing z^* , z and Y in (14). Let $\text{Var}(z) = \sigma^2\Omega$ and $\Omega = LL'$, with L lower triangular, the Cholesky decomposition of Ω . Premultiplying (14) by L^{-1} , the following ordinary least squares model is obtained

$$L^{-1}z^* = L^{-1}Yw + L^{-1}z. \quad (15)$$

Letting $r = L^{-1}z$, the equality $\text{Var}(r) = \sigma^2 I_N$ holds and vector r is the residual vector of the series (not observed). If we let the estimated residuals be $r^* = L^{-1}z^*$ and write $X = L^{-1}Y$, (15) can be written as

$$r^* = Xw + r. \quad (16)$$

If Y is 0 in (14), the model would be $z^* = z$ and if we applied the Kalman filter to this model, we would obtain $L^{-1}z^*$. This result, which is a standard result of control theory, allows us to see the Kalman filter as an algorithm that, applied to any vector v instead of z^* , yields $L^{-1}v$. Therefore, if we apply the Kalman filter to the vector of observations z^* and to the vector Y , we can move from (14) to (15) or, what amounts to the same thing, from (14) to (16).

We can estimate ω by OLS in (16) to obtain

$$\tilde{\omega} = (X'X)^{-1}X'r^*, \quad (17)$$

where the estimator variance is $\text{Var}(\tilde{\omega}) = (X'X)^{-1}\sigma^2$. To test the null hypothesis that there is no outlier at $t = T$, we can use the statistic

$$\tau = (X'X)^{1/2}\tilde{\omega}/\sigma, \quad (18)$$

which is distributed $N(0, 1)$ under the null.

In practice, the parameters of model (4) will not be known and they will have to be estimated. Under these circumstances, the usual procedure consists of estimating first the parameters of model (4) by exact maximum likelihood, as if there were no outliers, and then using instead of (17) and (18) their sample counterparts

$$\hat{\omega} = (\hat{X}'\hat{X})^{-1}\hat{X}'\hat{r}^*, \quad \hat{\tau} = (\hat{X}'\hat{X})^{1/2}\hat{\omega}/\hat{\sigma},$$

which are obtained by replacing in (17) and (18) the unknown parameters with their estimates. It can be shown that $\hat{\tau}$ is asymptotically

equivalent to τ . See Chang, Tiao and Chen (1988), p. 196. Each matrix X and, therefore, $X'X$ depends on the type of the outlier.

To see whether there is an outlier at $t = T$, the four estimators $\hat{\omega}_{IO}(T)$, $\hat{\omega}_{AO}(T)$, $\hat{\omega}_{TC}(T)$ and $\hat{\omega}_{LS}(T)$ are first computed, along with the statistics $\hat{\tau}_{IO}(T)$, $\hat{\tau}_{AO}(T)$, $\hat{\tau}_{TC}(T)$ and $\hat{\tau}_{LS}(T)$, where the subindex refers to the outlier type. Then, as proposed by Chang, Tiao and Chen (1988), the statistic $\lambda_T = \max\{|\hat{\tau}_{IO}(T)|, |\hat{\tau}_{AO}(T)|, |\hat{\tau}_{TC}(T)|, |\hat{\tau}_{LS}(T)|\}$ is used. If $\lambda_T > C$, where C is a predetermined critical value, then there is the possibility of an outlier of the type given by the subindex of the statistic $\hat{\tau}$ for which the maximum is obtained.

Since the time $t = T$ at which the outlier occurs is unknown in practice, the criterion based on the likelihood quotient of Chang y Tiao (1983), leads to repeat the previous operation for $t = 1, \dots, N$ and compute $\lambda = \max_t \lambda_t = |\hat{\tau}_{tp}(T)|$, where tp can be IO, AO, TC or LS. If $\lambda > C$, then there is an outlier of type tp at T .

Once the type of an outlier at $t = T$ is known, the series and the residuals can be corrected for its effect using (13) and (16).

Up to now, we have assumed that τ^* and X were computed by means of an “exact” filter, which was the Kalman filter. This is the correct thing to do, since the number of observations in a time series is always finite and we cannot apply the semi-infinite filter, given by the inverse of the series model $\pi(B) = 1 + \pi_1 B + \pi_2 B^2 + \dots = \phi(B)\delta(B)/\theta(B)$, to (13) to obtain

$$\pi(B)z_t^* = \omega[\pi(B)\nu(B)I_t(T)] + a(t), \quad t = 1, \dots, N,$$

instead of (15). In practice, the usual procedure consists of truncating the filter $\pi(B)$ and disregarding some observations at the beginning of the series. See Chen and Liu (1993), p. 285. In the procedure proposed by Gómez (1997), the residuals are filtered with an exact filter to obtain τ^* and the filter $\pi(B)$ is used to filter the vector Y in (14).

5.2.2 The Case of Multiple Outliers

When multiple outliers are present, we should use instead of (13) the model

$$z^*(t) = z(t) + \sum_{i=1}^k \omega_i \nu_i(B) I_i(t_i). \quad (19)$$

As shown in Chen and Liu (1993), the estimators of the ω_i obtained simultaneously using (19), can be very different from the ones obtained by an iterative process using the results of the previous Section. That is, by obtaining first $\hat{\omega}_1$, then $\hat{\omega}_2$, etc. For this reason, it is important that every algorithm for outlier detection performs at some point multiple regressions to detect spurious outliers and correct the bias produced in the estimators sequentially obtained.

In order to estimate the parameters in the multiple regressions, when the parameters of the ARIMA model (4) are assumed to be known, the algorithm proposed by Gómez (1997) uses first the Kalman filter like when we moved from (14) to (15). Then, the estimators of the ω_i and the corresponding statistics are computed using (17) and (18). This is done in an efficient manner, using the QR algorithm and Housholder transformations. A more detailed description will be given at the end of this Section.

5.2.3 Estimation of the Standard Deviation σ of the Residuals

When outliers are present in the series, the usual sample estimator can overestimate δ . For this reason, it is advisable to use a robust estimator. In the procedure proposed by Gómez (1997) the estimator used is the MAD estimator, defined by

$$\hat{\sigma} = 1.483 \times \text{median}\{|r^*(t) - \tilde{r}^*|\},$$

where \tilde{r}^* is the median of the estimated residuals $r^* = L^{-1}z^*$. The parameters of the model were assumed to be known in the previous formula. If they were unknown, they would be replaced with their estimates, as usual.

For outlier treatment, the procedure proposed by Gómez (1997) assumes that the orders (p, d, q) of model (4) are known and it proceeds iteratively. In the first stage, outliers are detected one by one and the model parameters are modified after each outlier has been detected. When no more outliers have been detected, the procedure goes to the second stage, where a multiple regression is performed. The outliers with the lowest t -value is discarded and the procedure goes back to the first stage to iterate.

The procedure used to incorporate or reject outliers is similar to the stepwise regression procedure for selecting the “best” regression

equation. This results in a more robust procedure than that of Chen and Liu (1993), which uses “backward elimination” and may therefore detect too many outliers in the first stage of the procedure.

Up to now, we have supposed that there were no regression effects, but it is easy to incorporate these effects into the procedure. Let the series follow the regression model with ARIMA errors

$$z(t) = y'(t)\beta + \nu(t), \quad t = 1, \dots, N, \quad (20)$$

where $\beta = (\beta_1, \dots, \beta_k)'$ is the vector containing the regression parameters, which may include the mean as the first component, $\{z(t)\}$ is the observed series, $\{y(t)\}$ are the vectors containing the regression variables and $\{\nu(t)\}$ follows the ARIMA model (4) with $\mu = 0$. Then, the algorithm proposed by Gómez (1997) for automatic detection and correction of outliers, described in detail, is the following

Initialization

If there are any regression variables in the model, included the mean, the regression coefficients are estimated by OLS and the series is corrected for their effects.

Stage I: Detection and estimation of outliers one by one

- I.1) The ARIMA parameters are estimated, using Hannan–Rissanen’s method, and the series is corrected for all regression effects present at the time, included the outliers so far detected. If desired by the user, exact maximum likelihood can be used for estimation, instead of HR’s method.
- I.2) Considering the estimates of the ARIMA parameters obtained in I.1 as fixed, the regression coefficients are estimated by GLS and their t statistics are computed. To this end, the fast algorithm of Morf, Sidhu and Kailath (1974) is used, followed by the QR algorithm. New estimated residuals are obtained.
- I.3) With the estimated residuals obtained in I.2, the robust MAD estimator of the standard deviation of the residuals is computed.
- I.4) If $u = (u_{d+1}, \dots, u_N)'$, where d is the degree of the differencing operator, denotes the differenced series, the statistics $\hat{\tau}_{IO}(t)$, $\hat{\tau}_{AO}(t)$, $\hat{\tau}_{LS}(t)$ and $\hat{\tau}_{TC}(t)$ are computed for $t = d + 1, \dots, N$.

To this end, the residuals computed in I.2 and the MAD obtained in I.3 are used. Let, for each $t = d + 1, \dots, N$, $\lambda_t = \max\{|\hat{r}_{IO}(t)|, |\hat{r}_{AO}(t)|, |\hat{r}_{TC}(t)|, |\hat{r}_{LS}(t)|\}$. If $\lambda = \max_t \lambda_t = |\hat{r}_{tp}(T)| > C$, where C is a pre-selected critical value, then there is a possible outlier of type tp at T . The subindex tp can be IO, AO, TC or LS. If no outlier has been found the first time the algorithm passes through this point, then stop. The series is free from outlier effects. If no outlier has been found, but it is not the first time that the algorithm passes through this point, then go to II.1. If, on the contrary, an outlier has been found, then correct the series for all regression effects, using the estimates obtained in I.2 and the last outlier coefficient estimate obtained while computing λ , and go back to I.1 to iterate.

Stage II: Multiple Regression

Using the estimates of the multiple regression and their t statistics obtained the last time the algorithm passed through I.2, check whether there are any outliers with a t statistic $< C$, where C is the same critical value than in I.4. If there aren't any, stop. If, on the contrary, there are some, then remove the one with the lowest absolute t -value and go back to I.2 to iterate.

5.3 Estimation and Filtering Techniques to Speed up the Algorithms

To estimate the regression parameters of a regression model with ARIMA errors, when the autoregressive and moving average parameters of the ARIMA model are assumed to be known, the procedure proposed by Gómez (1997) uses the following algorithm. Let the observed series $z = (z(1), \dots, z(N))'$ follow the regression model with ARIMA errors

$$z = Y\beta + u, \quad (21)$$

where $\beta = (\beta_1, \dots, \beta_k)'$ is the vector containing the regression parameters, which may include the mean as the first component, Y is an $N \times k$ matrix of full column rank and u follows the ARIMA model (4) with $\mu = 0$, which is supposed to be known. After differencing z , the

columns of Y and u in (21), it is obtained that

$$y = X\beta + v, \quad (22)$$

where $y = (y(d+1), \dots, y(N))'$, X is an $(N-d) \times k$ matrix, the components of $v = (v(d+1), \dots, v(N))'$ follow the ARMA model $\phi(B)v(t) = \theta(B)a(t)$ and it is assumed that the degree of the differencing polynomial $\delta(B)$ is d .

If $\text{Var}(v) = \sigma^2\Omega$ and $\Omega = LL'$, with L lower triangular, is the Cholesky decomposition of Ω , then, premultiplying (22) by L^{-1} , it is obtained that

$$L^{-1}y = L^{-1}X\beta + L^{-1}v, \quad (23)$$

which is an OLS model. As described in Section 5.2.1, the Kalman filter can be applied to y and the columns of the X matrix to move from (22) to (23). The Kalman filter algorithm used is the fast algorithm of Morf, Sidhu and Kailath (1974). β can now efficiently estimated in (23) by means of the QR algorithm. This last algorithm produces an orthogonal matrix Q such that $Q'L^{-1}X = (R', 0)'$, where R is a nonsingular upper triangular matrix. Partitioning $Q' = (Q_1', Q_2')'$ conforming to $(R', 0)'$, one can move from (23) to

$$\begin{aligned} Q_1' L^{-1} y &= R\beta + Q_1' L^{-1} v \\ Q_2' L^{-1} y &= \quad + Q_2' L^{-1} v, \end{aligned}$$

from which $\hat{\beta} = R^{-1}Q_1' L^{-1}y$ and $\hat{\sigma}^2 = y'(L^{-1})'Q_2'Q_2' L^{-1}y / (N-d-k)$ are easily obtained. The Q matrix is obtained by means of Housholder transformations.

5.4 The Need to Robustify Automatic Modeling Methods

The presence of outliers in the series can affect tremendously all automatic model identification procedures, starting with the specification of unit roots and ending with the identification of an ARMA model for the differenced series. For this reason, there is a need to robustify automatic modeling methods. This can be achieved by the following scheme. Specify first a robust model for the series. This model could be the airline model, since, as mentioned earlier, it encompasses many models and is a model very often found in practice. Then, use this

model to detect and correct the series for outlier effects. The critical value at this stage should not be low because we want to correct the series for the effects of the biggest outliers, which are the outliers that can distort most the automatic model identification procedure. With the series corrected for the outlier effects detected with the airline model, apply the automatic model identification procedure. With the model identified by this last procedure, specify a lower critical value and detect and correct the series for outliers. This cycle can be repeated several times until a satisfactory model is found. Usually, two iterations are enough.

5.5 An Algorithm for Automatic Model Identification in the Presence of Outliers

Taking into account the procedure proposed by Tsay (1986) and the previous considerations on how it could be improved, an algorithmical procedure is proposed (this procedure is implemented in programs TRAMO and SEATS, see Gómez and Maravall, 1996) which, briefly described, is the following.

- a) *Preliminary tests.* If desired by the user, the procedure can test for the log-level specification, Trading Day and Easter effects. These last two tests are performed using the default model (airline model).
- b) *Initialization.* If the user wants the series to be corrected for outliers, accept the model specified by the user (the default model is the airline model) and go to step 3. Otherwise, go to step 1. The critical value C for outlier detection can be either entered by the user or specified by the procedure. In this last case, the value of C is chosen depending on the length of the series.
- c) *Step 1.* If the user has specified the differencing orders and whether there should be a mean in the model, go to step 2. Otherwise, the series is first corrected for all regression effects, if any. Then, using the corrected series, the differencing orders for the ARIMA model are automatically obtained and, also automatically, it is decided whether to specify a mean for the series or not. Go to step 2.
- e) *Step 2.* Perform automatic identification of an $\text{ARMA}(p, q)$ model for the differenced series, corrected for all outliers and other re-

gression effects, if any. If the user wants to test for Trading Day and Easter effects and any of these effects was specified in the preliminary tests, check whether the specified effects are significant for the new model. If the user wants to correct the series for outliers, go to step 3. Otherwise, stop.

- d) *Step 3.* Assuming the model known, perform automatic detection and correction of outliers using C as critical value. If a stop condition is not satisfied, perhaps decrease the critical value C and go to step 1.

In the previous algorithm, the procedures for obtaining the differencing orders, automatic model identification and automatic detection and correction of outliers are the ones proposed by Gómez (1997), which have been described in previous sections. The tests for the log-level specification are the ones considered in the previous section. The Trading Day and Easter effects, as well as tests for their presence in the model, will be described in detail in the next section.

6 An Automatic Procedure for the General Regression–Arima Model in the Presence of Outliers, Special Effects and, Possibly, Missing Observations

In this section, the algorithm for automatic model identification in the presence of outliers of last section is extended to the case in which there are missing observations. The algorithm was seen to handle any kind of regression effect. Special effects, like Trading Day and Easter effects are considered in detail, as well as intervention and other regression effects. Tests for the presence of Trading Day and Easter effects are given. Finally, some guidelines as to how transfer functions can be incorporated into the model are given.

6.1 Missing Observations

The procedure proposed in the last section for automatic model identification in the presence of outliers can be extended easily to the case of missing observations. Missing observations are treated as additive

outliers. This implies that we can work with a complete series, because the missing values are first assigned tentative values. Then, after the model has been estimated, the difference between the tentative value and the estimated regression coefficient is the interpolated value. See Gómez, Maravall and Peña (1997) for details.

Since we work with a complete series (there are no holes in it), we can use same algorithms described previously for automatic model identification and for automatic detection and correction of outliers. The tentative values assigned to the missing observations are the semisum of the two adjacent values.

6.2 Trading Day and Easter Effects

Traditionally, six variables have been used to model the trading day effect. These are: (no. of *Mondays*) - (no. of *Sundays*), ..., (no. of *Saturdays*) - (no. of *Sundays*).

The motivation for using these variables is that it is desirable that the sum of the effects of each day of the week cancel out. Mathematically, this can be expressed by the requirement that the trading day coefficients β_j , $j = 1, \dots, 7$, verify $\sum_{j=1}^7 \beta_j = 0$, which implies $\beta_7 = -\sum_{j=1}^6 \beta_j$.

Sometimes, a seventh variable, called the length-of-month variable, is also included. This variable is defined as $m_t - \bar{m}$, where m_t is the length of the month (in days) and $\bar{m} = 30.4375$ is the average month length.

There is the possibility of considering a more parsimonious modeling of the trading day effect by using one variable instead of six. In this case, the days of the week are first divided into two categories: working days and non-working days. Then, the variable is defined as (no. of (*M, T, W, Th, F*)) - (no. of (*Sat, Sun*) $\times 5/2$).

Again, the motivation is that it is desirable that the trading day coefficients β_j , $j = 1, \dots, 7$ verify $\sum_{j=1}^7 \beta_j = 0$. Since $\beta_1 = \beta_2 = \dots = \beta_5$ and $\beta_6 = \beta_7$, we have $5\beta_1 = -2\beta_6$.

The Easter variable models a constant change in the level of daily activity during the d days before Easter. The value of d is usually supplied by the user.

The variable has zeros for all months different from March and April. The value assigned to March is equal to $p_M - m_M$, where p_M is the proportion of the d days that fall on that month and m_M is the

mean value of the proportions of the d days that fall on March over a long period of time. The value assigned to April is $p_A - m_A$, where p_A and m_A are defined analogously. Usually, a value of $m_M = m_A = 1/2$ is a good approximation.

Since $p_A - m_A = 1 - p_M - (1 - m_M) = -(p_M - m_M)$, the sum of the effects of both months, March and April, cancel out, a desirable feature.

Since Trading Day and Easter effects are modeled by means of regression variables, a possible test for these effects is the following. If no model has been identified, specify an airline model with mean. Otherwise, use the identified model. Then, using the differenced series y , apply first the Kalman filter to move from model (22) to model (23), where β is the vector of regression parameters, that includes the Trading Day and/or Easter parameters. Since model (23) is an OLS model, we can use an ordinary F -test to test if all Trading Day parameters are zero or not. A t -test can be used to test if the Easter parameter is zero.

6.3 Intervention and Regression Effects

Intervention variables are regression variables that are used to model certain abnormal effects, like strikes, major changes in economic policy, natural disasters, etc. See Box and Tiao (1975).

Examples of intervention variables are

- impulses
- level shifts
- temporary changes
- ramps

These variables usually consist of sequences of ones and zeros. Other regression effects, like economic variables thought to be related to the observed series, can also be incorporated.

6.4 Transfer Functions

Transfer functions can be incorporated into the model by means of a procedure similar to the LTF method. First, consider some lags of the input series as regressors in all the process of automatic model

identification in the presence of outliers. Once an adequate model for the output series has been identified, use the corner method to specify a rational function in the backshift operator B which fits the pattern of coefficients of the lags of the input series.

After the identification stage has been completed, the estimation stage can be performed, for which a model of the input series is needed.

References

1. Abraham, B. and Box, G. E. P., (1979), "Bayesian Analysis of Some Outlier Problems in Time Series", *Biometrika*, **66**, 229-236.
2. Beguin, J. M., Gouriéroux, C. and Monfort, A., (1980), "Identification of a Mixed Autoregressive Moving Average Process: The Corner Method", in *Time Series*, O. D. Anderson, Ed., North-Holland, Amsterdam, 423-436.
3. Box, G.E.P., and Cox, D.R., (1964), "An Analysis of transformations", *Journal of the Royal Statistical Society, Series B*, **26**, 211-243.
4. Box, G.E.P., and Jenkins, G.M., (1976), *Time Series Analysis, Forecasting and Control*, Holden-Day, San Francisco.
5. Box, G.E.P., and Tiao, G.C., (1975), "Intervention Analysis with Applications to Economic and Environmental Problems", *Journal of the American Statistical Association*, **70**, 70-79.
6. Brockwell, P., and Davis, R., (1992), *Time Series: Theory and Methods, Second Edition*, Springer-Verlag, Berlin.
7. Chang, I., Tiao, G.C. (1983), "Estimation of Time Series Parameters in the Presence of Outliers", Technical Report 8, University of Chicago, Statistics Research Center.
8. Chang, I., Tiao, G.C. and Chen, C., (1988), "Estimation of Time Series Parameters in the Presence of Outliers", *Technometrics*, **30**, 193-204.
9. Chatfield, C., (1979), "Inverse Autocorrelations", *Journal of the Royal Statistical Society, Series A*, **142**, 363-377.

10. 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.
11. Chen, C. and Tiao, G. C., (1990), "Random Level Shift Time Series Models, ARIMA Approximation, and Level Shift Detection", *Journal of Business and Economic Statistics*, **8**, 170-186.
12. Choi, B., (1992), *ARMA Model Identification*, Springer Verlag, New York.
13. Cleveland, W.S., (1972), "The Inverse Autocorrelations of a Time Series and Their Applications", *Technometrics*, **14**, 277-298.
14. Dickey, D. A. and Pantula, S. G., (1987), "Determining the Order of Differencing in Autoregressive Processes", *Journal of Business and Economic Statistics*, **5**, 455-461.
15. Fox, A. J., (1972), "Outliers in Time Series", *Journal of the Royal Statistical Society, Series B*, **34**, 350-363.
16. Gómez, V., (1997), "Automatic Model Identification in the Presence of Missing Observations and Outliers", Mimeo, Ministerio de Economía y Hacienda, Dirección General de Análisis y Programación Presupuestaria, Madrid.
17. Gómez, V., and Maravall, A., (1996), "Programs TRAMO and SEATS", Instructions for the User (Beta Version: September 1996), Working Paper No. 9628, Bank of Spain.
18. Gómez, V., Maravall, A., and Peña, D. (1996), "Missing Observations in ARIMA Models: Skipping Strategy Versus Additive Outlier Approach", *Journal of Econometrics* (Forthcoming).
19. Gray, H. L., Kelley, G. D. and McIntire, D. D., (1978), "A new Approach to ARMA Modeling", *Communications in Statistics*, **B7**, 1-77.
20. Guttman, I. and Tiao, G. C., (1978), "Effect of Correlation on the Estimation of a Mean in the Presence of Spurious Observations", *The Canadian Journal of Statistics*, **6**, 229-247.
21. Hamilton, D. C. and Watts, D. G., (1978), "Interpreting Partial Autocorrelation Function of Seasonal Time Series Models", *Biometrika*, **65**, 135-140.
22. Hannan, E. J. and Rissanen, J., "Recursive Estimation of Mixed Autoregressive-Moving Average Order", *Biometrika*, **69**, 81-94.

23. Ljung, G. M. and Box, G. E. P., (1979), "The Likelihood Function of Stationary Autoregressive-Moving Average Models", *Biometrika*, **66**, 265-270.
24. Liu, L. M., (1989), "Identification of Seasonal ARIMA Models Using a Filtering Method", *Communications in Statistics*, **18**, 2279-2288.
25. Liu, L. M. and Hanssens, D. M., (1982), "Identification of Multiple-Input Transfer Function Models", *Communications in Statistics*, **A11**, 297-314.
26. Lütkepohl, H., (1985), "Comparison of Criteria for Estimating the Order of a Vector Autoregressive Process", *Journal of Time Series Analysis*, **6**, 35-52.
27. Mélard, G., (1984), "A Fast Algorithm for the Exact Likelihood of Autoregressive-Moving Average Models", *Applied Statistics*, **35**, 104-114.
28. Morf, M., Sidhu, G.S., and Kailath, T., (1974), "Some New Algorithms for Recursive Estimation on Constant, Linear, Discrete-Time Systems", *IEEE Transactions on Automatic Control*, **AC-19**, 315-323.
29. Pearlman, J.G., (1980), "An Algorithm for the Exact Likelihood of a High-Order-Autoregressive-Moving Average Process", *Biometrika*, **67**, 232-233.
30. Reinsel, G. C., (1997), *Elements of Multivariate Time Series Analysis*, second edition, Springer Verlag, New York.
31. Schwert, G. W., (1989), "Tests for Unit Roots: A Monte Carlo Investigation", *Journal of Business and Economic Statistics*, **7**, 147-159.
32. Tiao, G. C. and Tsay, R. S., (1983), "Consistency Properties of Least Squares Estimates of Autoregressive Parameters in ARMA Models", *The Annals of Statistics*, **11**, 856-871.
33. Tsay, R. S., (1984), "Regression Models With Times Series Errors", *Journal of the American Statistical Association*, **79**, 118-124.
34. Tsay, R. S., (1986), "Time Series Model Specification in the Presence of Outliers", *Journal of the American Statistical Association*, **81**, 132-141.

35. Tsay, R. S., (1988), "Outliers, Level Shifts, and Variance Changes in Time Series", *Journal of Forecasting*, **7**, 1-20.
36. Tsay, R. S. and Tiao, G. C., (1984), "Consistent Estimates of Autoregressive Parameters and Extended Sample Autocorrelation Functions for Stationary and Nonstationary ARMA models", *Journal of the American Statistical Association*, **79**, 84-96.
37. Tsay, R. S. and Tiao, G. C., (1985), "Use of Canonical Analysis in Time Series Model Identification", *Biometrika*, **72**, 299-316.
38. Zhao-Guo, C., (1985), "The Asymptotic Efficiency of a Linear Procedure of Estimation for ARMA Models", *Journal of Time Series Analysis*, **6**, 53-62.

WORKING PAPERS (1)

- 9406 **Juan José Dolado, José Manuel González-Páramo y José M.ª Roldán:** Convergencia económica entre las provincias españolas: evidencia empírica (1955-1989).
- 9407 **Ángel Estrada e Ignacio Hernando:** La inversión en España: un análisis desde el lado de la oferta.
- 9408 **Ángel Estrada García, M.ª Teresa Sastre de Miguel y Juan Luis Vega Croissier:** El mecanismo de transmisión de los tipos de interés: el caso español.
- 9409 **Pilar García Perea y Ramón Gómez:** Elaboración de series históricas de empleo a partir de la Encuesta de Población Activa (1964-1992).
- 9410 **F. J. Sáez Pérez de la Torre, J. M.ª Sánchez Sáez y M.ª T. Sastre de Miguel:** Los mercados de operaciones bancarias en España: especialización productiva y competencia.
- 9411 **Olympia Bover and Ángel Estrada:** Durable consumption and house purchases: Evidence from Spanish panel data.
- 9412 **José Viñals:** Building a Monetary Union in Europe: Is it worthwhile, where do we stand, and where are we going? (The Spanish original of this publication has the same number.)
- 9413 **Carlos Chuliá:** Los sistemas financieros nacionales y el espacio financiero europeo.
- 9414 **José Luis Escrivá and Andrew G. Haldane:** The interest rate transmission mechanism: Sectoral estimates for Spain. (The Spanish original of this publication has the same number.)
- 9415 **M.ª de los Llanos Matea y Ana Valentina Regil:** Métodos para la extracción de señales y para la trimestralización. Una aplicación: Trimestralización del deflactor del consumo privado nacional.
- 9416 **José Antonio Cuenca:** Variables para el estudio del sector monetario. Agregados monetarios y crediticios, y tipos de interés sintéticos.
- 9417 **Ángel Estrada y David López-Salido:** La relación entre el consumo y la renta en España: un modelo empírico con datos agregados.
- 9418 **José M. González Mínguez:** Una aplicación de los indicadores de discrecionalidad de la política fiscal a los países de la UE.
- 9419 **Juan Ayuso, María Pérez Jurado and Fernando Restoy:** Is exchange rate risk higher in the E.R.M. after the widening of fluctuation bands? (The Spanish original of this publication has the same number.)
- 9420 **Simon Milner and David Metcalf:** Spanish pay setting institutions and performance outcomes.
- 9421 **Javier Santillán:** El SME, los mercados de divisas y la transición hacia la Unión Monetaria.
- 9422 **Juan Luis Vega:** Is the ALP long-run demand function stable? (The Spanish original of this publication has the same number.)
- 9423 **Gabriel Quirós:** El mercado italiano de deuda pública.
- 9424 **Isabel Argimón, José Manuel González-Páramo y José María Roldán:** Inversión privada, gasto público y efecto expulsión: evidencia para el caso español.
- 9425 **Charles Goodhart and José Viñals:** Strategy and tactics of monetary policy: Examples from Europe and the Antipodes.
- 9426 **Carmen Melcón:** Estrategias de política monetaria basadas en el seguimiento directo de objetivos de inflación. Las experiencias de Nueva Zelanda, Canadá, Reino Unido y Suecia.
- 9427 **Olympia Bover and Manuel Arellano:** Female labour force participation in the 1980s: the case of Spain.

- 9428 **Juan María Peñalosa:** The Spanish catching-up process: General determinants and contribution of the manufacturing industry.
- 9429 **Susana Núñez:** Perspectivas de los sistemas de pagos: una reflexión crítica.
- 9430 **José Viñals:** ¿Es posible la convergencia en España?: En busca del tiempo perdido.
- 9501 **Jorge Blázquez y Miguel Sebastián:** Capital público y restricción presupuestaria gubernamental.
- 9502 **Ana Buisán:** Principales determinantes de los ingresos por turismo.
- 9503 **Ana Buisán y Esther Gordo:** La protección nominal como factor determinante de las importaciones de bienes.
- 9504 **Ricardo Mestre:** A macroeconomic evaluation of the Spanish monetary policy transmission mechanism.
- 9505 **Fernando Restoy and Ana Revenga:** Optimal exchange rate flexibility in an economy with intersectoral rigidities and nontraded goods.
- 9506 **Ángel Estrada and Javier Vallés:** Investment and financial costs: Spanish evidence with panel data. (The Spanish original of this publication has the same number.)
- 9507 **Francisco Alonso:** La modelización de la volatilidad del mercado bursátil español.
- 9508 **Francisco Alonso y Fernando Restoy:** La remuneración de la volatilidad en el mercado español de renta variable.
- 9509 **Fernando C. Ballabriga, Miguel Sebastián y Javier Vallés:** España en Europa: asimetrías reales y nominales.
- 9510 **Juan Carlos Casado, Juan Alberto Campoy y Carlos Chuliá:** La regulación financiera española desde la adhesión a la Unión Europea.
- 9511 **Juan Luis Díaz del Hoyo y A. Javier Prado Domínguez:** Los FRAs como guías de las expectativas del mercado sobre tipos de interés.
- 9512 **José M.ª Sánchez Sáez y Teresa Sastre de Miguel:** ¿Es el tamaño un factor explicativo de las diferencias entre entidades bancarias?
- 9513 **Juan Ayuso y Soledad Núñez:** ¿Desestabilizan los activos derivados el mercado al contado?: La experiencia española en el mercado de deuda pública.
- 9514 **M.ª Cruz Manzano Frias y M.ª Teresa Sastre de Miguel:** Factores relevantes en la determinación del margen de explotación de bancos y cajas de ahorros.
- 9515 **Fernando Restoy and Philippe Weil:** Approximate equilibrium asset prices.
- 9516 **Gabriel Quirós:** El mercado francés de deuda pública.
- 9517 **Ana L. Revenga and Samuel Bentolila:** What affects the employment rate intensity of growth?
- 9518 **Ignacio Iglesias Araúzo y Jaime Esteban Velasco:** Repos y operaciones simultáneas: estudio de la normativa.
- 9519 **Ignacio Fuentes:** Las instituciones bancarias españolas y el Mercado Único.
- 9520 **Ignacio Hernando:** Política monetaria y estructura financiera de las empresas.
- 9521 **Luis Julián Álvarez y Miguel Sebastián:** La inflación latente en España: una perspectiva macroeconómica.
- 9522 **Soledad Núñez Ramos:** Estimación de la estructura temporal de los tipos de interés en España: elección entre métodos alternativos.
- 9523 **Isabel Argimón, José M. González-Páramo y José M.ª Roldán Alegre:** Does public spending crowd out private investment? Evidence from a panel of 14 OECD countries.

- 9524 **Luis Julián Álvarez, Fernando C. Ballabriga y Javier Jareño:** Un modelo macroeconómico trimestral para la economía española.
- 9525 **Aurora Alejano y Juan M.ª Peñalosa:** La integración financiera de la economía española: efectos sobre los mercados financieros y la política monetaria.
- 9526 **Ramón Gómez Salvador y Juan J. Dolado:** Creación y destrucción de empleo en España: un análisis descriptivo con datos de la CBBE.
- 9527 **Santiago Fernández de Lis y Javier Santillán:** Regímenes cambiarios e integración monetaria en Europa.
- 9528 **Gabriel Quirós:** Mercados financieros alemanes.
- 9529 **Juan Ayuso Huertas:** Is there a trade-off between exchange rate risk and interest rate risk? (The Spanish original of this publication has the same number.)
- 9530 **Fernando Restoy:** Determinantes de la curva de rendimientos: hipótesis expectacional y primas de riesgo.
- 9531 **Juan Ayuso and María Pérez Jurado:** Devaluations and depreciation expectations in the EMS.
- 9532 **Paul Schulstad and Ángel Serrat:** An Empirical Examination of a Multilateral Target Zone Model.
- 9601 **Juan Ayuso, Soledad Núñez and María Pérez-Jurado:** Volatility in Spanish financial markets: The recent experience.
- 9602 **Javier Andrés e Ignacio Hernando:** ¿Cómo afecta la inflación al crecimiento económico? Evidencia para los países de la OCDE.
- 9603 **Barbara Dluhosch:** On the fate of newcomers in the European Union: Lessons from the Spanish experience.
- 9604 **Santiago Fernández de Lis:** Classifications of Central Banks by Autonomy: A comparative analysis.
- 9605 **M.ª Cruz Manzano Frías y Sofía Galmés Belmonte:** Credit Institutions' Price Policies and Type of Customer: Impact on the Monetary Transmission Mechanism. (The Spanish original of this publication has the same number.)
- 9606 **Malte Krüger:** Speculation, Hedging and Intermediation in the Foreign Exchange Market.
- 9607 **Agustín Maravall:** Short-Term Analysis of Macroeconomic Time Series.
- 9608 **Agustín Maravall and Christophe Planas:** Estimation Error and the Specification of Unobserved Component Models.
- 9609 **Agustín Maravall:** Unobserved Components in Economic Time Series.
- 9610 **Matthew B. Canzoneri, Behzad Diba and Gwen Eudey:** Trends in European Productivity and Real Exchange Rates.
- 9611 **Francisco Alonso, Jorge Martínez Pagés y María Pérez Jurado:** Weighted Monetary Aggregates: an Empirical Approach. (The Spanish original of this publication has the same number.)
- 9612 **Agustín Maravall and Daniel Peña:** Missing Observations and Additive Outliers in Time Series Models.
- 9613 **Juan Ayuso and Juan L. Vega:** An empirical analysis of the peseta's exchange rate dynamics.
- 9614 **Juan Ayuso Huertas:** Un análisis empírico de los tipos de interés reales *ex-ante* en España.
- 9615 **Enrique Alberola Ila:** Optimal exchange rate targets and macroeconomic stabilization.

- 9616 **A. Jorge Padilla, Samuel Bentolila and Juan J. Dolado:** Wage bargaining in industries with market power.
- 9617 **Juan J. Dolado and Francesc Marnol:** Efficient estimation of cointegrating relationships among higher order and fractionally integrated processes.
- 9618 **Juan J. Dolado y Ramón Gómez:** La relación entre vacantes y desempleo en España: perturbaciones agregadas y de reasignación.
- 9619 **Alberto Cabrero and Juan Carlos Delrieu:** Construction of a composite indicator for predicting inflation in Spain. (The Spanish original of this publication has the same number.)
- 9620 **Una-Louise Bell:** Adjustment costs, uncertainty and employment inertia.
- 9621 **M.ª de los Llanos Matea y Ana Valentina Regil:** Indicadores de inflación a corto plazo.
- 9622 **James Conklin:** Computing value correspondences for repeated games with state variables.
- 9623 **James Conklin:** The theory of sovereign debt and Spain under Philip II.
- 9624 **José Viñals and Juan F. Jimeno:** Monetary Union and European unemployment.
- 9625 **María Jesús Nieto Carol:** Central and Eastern European Financial Systems: Towards integration in the European Union.
- 9626 **Matthew B. Canzoneri, Javier Vallés and José Viñals:** Do exchange rates move to address international macroeconomic imbalances?
- 9627 **Enrique Alberola Ila:** Integración económica y unión monetaria: el contraste entre Norteamérica y Europa.
- 9628 **Víctor Gómez and Agustín Maravall:** Programs TRAMO and SEATS.
- 9629 **Javier Andrés, Ricardo Mestre y Javier Vallés:** Un modelo estructural para el análisis del mecanismo de transmisión monetaria: el caso español.
- 9630 **Francisco Alonso y Juan Ayuso:** Una estimación de las primas de riesgo por inflación en el caso español.
- 9631 **Javier Santillán:** Política cambiaria y autonomía del Banco Central.
- 9632 **Marcial Suárez:** Vocabula (Notas sobre usos lingüísticos).
- 9633 **Juan Ayuso and J. David López-Salido:** What does consumption tell us about inflation expectations and real interest rates?
- 9701 **Víctor Gómez, Agustín Maravall and Daniel Peña:** Missing observations in ARIMA models: Skipping strategy versus outlier approach.
- 9702 **José Ramón Martínez Resano:** Los contratos DIFF y el tipo de cambio.
- 9703 **Gabriel Quirós Romero:** Una valoración comparativa del mercado español de deuda pública.
- 9704 **Agustín Maravall:** Two discussions on new seasonal adjustment methods.
- 9705 **J. David López-Salido y Pilar Velilla:** La dinámica de los márgenes en España (Una primera aproximación con datos agregados).
- 9706 **Javier Andrés and Ignacio Hernando:** Does inflation harm economic growth? Evidence for the OECD.

- 9707 **Marga Peeters:** Does demand and price uncertainty affect Belgian and Spanish corporate investment?
- 9708 **Jeffrey Franks:** Labor market policies and unemployment dynamics in Spain.
- 9709 **José Ramón Martínez Resano:** Los mercados de derivados y el euro.
- 9710 **Juan Ayuso and J. David López-Salido:** Are *ex-post* real interest rates a good proxy for *ex-ante* real rates? An international comparison within a CCAPM framework.
- 9711 **Ana Buisán y Miguel Pérez:** Un indicador de gasto en construcción para la economía española.
- 9712 **Juan J. Dolado, J. David López-Salido and Juan Luis Vega:** Spanish unemployment and inflation persistence: Are there phillips trade-offs?
- 9713 **José M. González Mínguez:** The balance-sheet transmission channel of monetary policy: The cases of Germany and Spain.
- 9714 **Olympia Bover:** Cambios en la composición del empleo y actividad laboral femenina.
- 9715 **Francisco de Castro and Alfonso Novales:** The joint dynamics of spot and forward exchange rates.
- 9716 **Juan Carlos Caballero, Jorge Martínez y M.^a Teresa Sastre:** La utilización de los índices de condiciones monetarias desde la perspectiva de un banco central.
- 9717 **José Viñals y Juan F. Jimeno:** El mercado de trabajo español y la Unión Económica y Monetaria Europea.
- 9718 **Samuel Bentolila:** La inmovilidad del trabajo en las regiones españolas.
- 9719 **Enrique Alberola, Juan Ayuso and J. David López-Salido:** When may peseta depreciations fuel inflation?
- 9720 **José M. González Mínguez:** The back calculation of nominal historical series after the introduction of the european currency (An application to the GDP).
- 9721 **Una-Louise Bell:** A Comparative Analysis of the Aggregate Matching Process in France, Great Britain and Spain.
- 9722 **Francisco Alonso Sánchez, Juan Ayuso Huertas y Jorge Martínez Pagés:** El poder predictivo de los tipos de interés sobre la tasa de inflación española.
- 9723 **Isabel Argimón, Concha Artola y José Manuel González-Páramo:** Empresa pública y empresa privada: titularidad y eficiencia relativa.
- 9724 **Enrique Alberola and Pierfederico Asdrubali:** How do countries smooth regional disturbances? Risksharing in Spain: 1973-1993.
- 9725 **Enrique Alberola, José Manuel Marqués and Alicia Sanchís:** Unemployment persistence, Central Bank independence and inflation performance in the OECD countries. (The Spanish original of this publication has the same number.)
- 9726 **Francisco Alonso, Juan Ayuso and Jorge Martínez Pagés:** How informative are financial asset prices in Spain?
- 9727 **Javier Andrés, Ricardo Mestre and Javier Vallés:** Monetary policy and exchange rate dynamics in the Spanish economy.
- 9728 **Juan J. Dolado, José M. González-Páramo y José Viñals:** A cost-benefit analysis of going from low inflation to price stability in Spain.

- 9801 **Ángel Estrada, Pilar García Perea, Alberto Urtasun y Jesús Briones:** Indicadores de precios, costes y márgenes en las diversas ramas productivas.
- 9802 **Pilar Álvarez Canal:** Evolución de la banca extranjera en el período 1992-1996.
- 9803 **Ángel Estrada y Alberto Urtasun:** Cuantificación de expectativas a partir de las encuestas de opinión.
- 9804 **Soyoung Kim:** Monetary Policy Rules and Business Cycles.
- 9805 **Víctor Gómez and Agustín Maravall:** Guide for using the programs TRAMO and SEATS.
- 9806 **Javier Andrés, Ignacio Hernando and J. David López-Salido:** Disinflation, output and unemployment: the case of Spain.
- 9807 **Olympia Bover, Pilar García-Perea and Pedro Portugal:** A comparative study of the Portuguese and Spanish labour markets.
- 9808 **Víctor Gómez and Agustín Maravall:** Automatic modeling methods for univariate series.

(1) Previously published Working Papers are listed in the Banco de España publications catalogue.

Queries should be addressed to: Banco de España
Sección de Publicaciones. Negociado de Distribución y Gestión
Telephone: 338 51 80
Alcalá, 50. 28014 Madrid