Algorithm: ARFIT — A MATLAB Package for Estimation and Spectral Decomposition of Multivariate Autoregressive Models

Tapio Schneider
Princeton University
and
Arnold Neumaier
Universität Wien

ARFIT is a collection of MATLAB routines for modeling multivariate time series by autoregressive (AR) models. It provides tools for all stages of the model identification process: statistics that aid in the selection of the model order, fast algorithms for least squares estimation of parameters, modules that produce approximate confidence regions for the estimated parameters, and routines for diagnostic checking of a fitted model. Furthermore, methods for the spectral analysis of AR models are implemented, and, as a supplement, ARFIT contains a program for the Monte-Carlo simulation of AR processes.

Categories and Subject Descriptors: G.3 [**Mathematics of Computing**]: Probability and Statistics—*statistical computing, statistical software*; I.6 [**Computing Methodologies**]: Simulation and Modeling; I.6.4 [**Simulation and Modeling**]: Model Validation and Analysis; J.2 [**Computer Applications**]: Physical Sciences and Engineering—*Earth and atmospheric sciences*

General Terms: Algorithms

Additional Key Words and Phrases: confidence regions, least squares, MATLAB, model identification, multivariate autoregressive process, order selection, parameter estimation, spectral decomposition

1. INTRODUCTION

ARFIT is a collection of MATLAB routines for the identification of multivariate *autoregressive models of order* p, short AR(p) models, defined by

$$v_\nu = w + \sum_{l=1}^{p} A_l v_{\nu-l} + \varepsilon_\nu, \qquad \varepsilon_\nu = \text{noise}(C), \qquad \nu = 1, \ldots, N. \tag{1}$$

Such models are appropriate for representing a large class of stationary time series of m-dimensional *state vectors* v_ν ($\nu = 1-p, \ldots, N$), observed at equally-spaced instants ν. Here, $\varepsilon_\nu = \text{noise}(C)$ is shorthand for saying that the ε_ν are uncorrelated

Name: Tapio Schneider
Affiliation: AOS Program, Princeton University
Address: P.O. Box CN710, Princeton, NJ 08544, U.S.A., email: tapio@splash.princeton.edu
Name: Arnold Neumaier
Affiliation: Institut für Mathematik, Universität Wien
Address: Strudlhofgasse 4, A-1090 Wien, Austria, email: neum@cma.univie.ac.at

m-dimensional random vectors with zero mean and covariance matrix C. The m-dimensional parameter vector w of intercept terms is included to allow for a nonzero mean

$$\langle v_\nu \rangle = (I - A_1 - \cdots - A_p)^{-1} w$$

of the time series, the symbol $\langle\cdot\rangle$ standing for the expectation value operation.

In the model identification process, one wants to estimate from the observed time series the order p and, for given p, the parameter matrices $A_1, \ldots, A_p$, C, and the intercept vector w. The adequacy of the fitted model for representing the data must then be assessed. ARFIT provides tools for all phases of this model identification process, with particular emphasis on fast algorithms for large data sets. Since in many applications understanding and interpretation of the fitted model is facilitated by studying spectral information computed from the model parameters, we also implemented algorithms for the spectral decomposition of AR models. A module for the simulation of AR processes is included as a useful aid in testing the estimation procedures that form the core of this package.

Several reasons led us to implement the presented algorithms in the form of MATLAB modules, so called M-files. MATLAB provides a widely used and very convenient environment for handling and analyzing data. The standard MATLAB package, possibly complemented by the *Signal Processing Toolbox* [Krauss et al. 1994], contains numerous tools for computing statistics such as spectra from time series data, and thanks to MATLAB's visualization capabilities, these statistics are easily displayed. ARFIT supplements the available functions, and thus enhances the data analysis capabilities of MATLAB.

Detailed descriptions of the employed algorithms for selection of a model's order and estimation of its parameters appear in a companion paper (Neumaier and Schneider [1997], hereafter referred to as NS) and shall not be repeated here. In Section 2, we briefly review a statistic that aids in diagnostic checking of a fitted model's adequacy, and we give recommendations on the use of the ARFIT modules in fitting AR models to data.

Our MATLAB code provides online documentation, is extensively annotated, and its internal variable names are closely related to the notation in NS and in the present note. Therefore, we refrain from a detailed description of the ARFIT modules. Instead, in Section 3, we give a brief overview on the purposes of the M-files that form the package, being confident that the implementations are easily understood by directly studying the source code and its enclosed documentation.

2. MODEL IDENTIFICATION

The implemented techniques for selecting the model order p and for estimating the parameters $A_1, \ldots, A_p$, C, and w were already described in NS. Given a lower bound $p_{\min}$ and an upper bound $p_{\max}$ for the model order p, the ARFIT module **ar** returns approximations to the logarithm of Akaike's Final Prediction Error (FPE; see Akaike [1971]), to Schwarz' Bayesian Criterion (SBC; see Schwarz [1978]), and to a modification of Schwarz' Bayesian Criterion (MSC) described in NS. Usually, the model order is estimated as the minimizer of one of these (or other, similar) criteria. Which criterion to employ depends on the use one wishes to make of the fitted model. However, numerical experiments reported in NS suggest that MSC

chooses the correct model order most frequently. Therefore, when in doubt as to what order selection criterion to employ, we recommend the use of MSC.

The known order selection criteria obtain their theoretical justification only from asymptotic considerations in the limit of infinite sample sizes. Numerical experiments, such as those reported by Lütkepohl [1985] or in NS, suggest that for finite samples the model order estimated by minimizing one of the order selection criteria is not always reliable. Thus, in particular for short time series, the estimate obtained by minimizing an order selection criterion should be regarded as tentative. In practice, one usually finds several values of p for which the employed order selection criterion is close to its minimum, and all the corresponding models should be considered as candidates for the "best-fitting model". In the following, we describe a test for discriminating between these models.

Having obtained a sequence of autoregressive models that nearly optimize the employed order selection criterion, two related questions still remain: First, one must determine if any of the estimated models adequately represents the data; second, one has to choose a model from the possibilities deemed adequate. As one approach to evaluating the adequacy of the model one can check the validity of the assumptions made in its derivation. The principal assumption made in (1) is that the noise vectors ε_ν have zero mean and are uncorrelated. To verify this assumption, the (asymptotic) uncorrelatedness of the residuals

$$\hat{\varepsilon}_\nu = v_\nu - \sum_{l=1}^{p} \hat{A}_l v_{\nu-l}, \qquad \nu = 1, \ldots, N \tag{2}$$

can be tested using estimates $\hat{R}(l)$ of the lag l correlation matrices with entries

$$\hat{R}_{ij}(l) = \frac{\hat{c}_{ij}(l)}{\sqrt{\hat{c}_{ii}(0)\hat{c}_{jj}(0)}}, \qquad l = 1, \ldots, k,$$

where

$$\hat{c}(l) = \sum_{\nu=l+1}^{N} (\hat{\varepsilon}_{\nu-l} - \hat{\mu})(\hat{\varepsilon}_\nu - \hat{\mu})^T$$

and

$$\hat{\mu} = \frac{1}{N} \sum_{\nu=1}^{N} \hat{\varepsilon}_\nu.$$

(The superscript $\hat{}$ refers to quantities that were estimated from data.) Li and McLeod [1981] showed that under the null hypothesis of model adequacy, for Gaussian noise, and for sufficiently large k the quantity

$$Q_k = N \sum_{l=1}^{k} x_{\hat{R}(l)}^T (\hat{R}(0)^{-1} \otimes \hat{R}(0)^{-1}) x_{\hat{R}(l)} + \frac{m^2 k(k+1)}{2N} \tag{3}$$

is asymptotically χ^2-distributed with

$$f = m^2(k-p)$$

degrees of freedom. Here, x_A consists of the components of the matrix A, arranged as a vector by stacking adjacent columns, the superscript ^{T} denotes transposition,

and $P \otimes Q$ is the *Kronecker product* of P and Q (as returned by the MATLAB function **kron**). The maximum lag k up to which residual correlation matrices are computed should be chosen such that for a model of order $l > k$ the AR parameter matrices A_l (often called the lag l *partial autocorrelation matrices* [Tiao and Box 1981]) are consistent with zero. In practice, for low-order processes the choice $k = 20$ should suffice, though if in doubt, it is preferable to choose k rather too large than too small.

From the asymptotic distribution of the Li-McLeod statistic Q_k, it follows that the hypothesis of the residuals' uncorrelatedness is rejected with approximate significance level β if

$$Q_k > \chi^2 \tag{4}$$

where χ^2 is a solution of

$$\beta = 1 - \Phi\left(\frac{f}{2}, \frac{\chi^2}{2}\right)$$

and

$$\Phi(\alpha, x) = \frac{1}{\Gamma(\alpha)} \int_0^x t^{\alpha-1} e^{-t}\, dt$$

is the incomplete gamma function. Equivalently, one rejects the hypothesis of the residuals' uncorrelatedness if the estimated significance level

$$\hat{\beta} \equiv 1 - \Phi\left(\frac{f}{2}, \frac{Q_k}{2}\right) \tag{5}$$

satisfies $\hat{\beta} \leq \beta$, where β indicates the probability that the test rejects a true hypothesis. As a typical value one may choose $\beta = 0.05$.

We will refer to this extension of the univariate *portmanteau test* (cf. Box and Jenkins [1970] and Box and Pierce [1970]) to the multivariate case as the modified *Li-McLeod portmanteau (LMP) test*.

It must, however, be noted that the uncorrelatedness of the residuals is only a necessary and not a sufficient condition for the validity of the model. Therefore, the main purpose of the LMP test is to disqualify unsatisfactory models from consideration, but passing the LMP test does not guarantee model adequacy. For further tests of model adequacy see, e.g., Brockwell and Davis [1991] or Tiao and Box [1981].

It is crucial to perform diagnostic tests of model adequacy such as the LMP test before trusting an analysis of a model's dynamical structure. Attempting to gain an understanding of the dynamical relationships between the components of the state vectors v_ν by analyzing (e.g., by a spectral decomposition) an inadequate model can yield misleading results. For example, Tiao and Box [1981] describe a case in which an AR model of inappropriately low order exhibits spurious feedback mechanisms that are not present in an adequate model of higher order.

While the abovementioned order selection criteria are easily computed along with the least squares estimates, testing the residuals for uncorrelatedness requires considerable additional effort. For computational convenience, one therefore evaluates the order selection criteria first, and then computes the Li-McLeod statistic only for those models that nearly optimize the order selection criterion one wishes to employ.

Summarizing, we arrive at the following model identification procedure:

(i) Given the time series v_ν and bounds $p_{\min}$ and $p_{\max}$ on the order, compute the order selection criteria for a sequence of models of order $p_{\min} \le p \le p_{\max}$. For further analysis, select models that optimize or nearly optimize the employed order selection criterion and compute their respective model parameters. The module **ar** performs these tasks.

(ii) For the selected models, compute the LMP significance levels $\hat{\beta}$ from the time series of residuals (2). Within the class of autoregressive models, accept as the best-fitting model the one that maximizes $\hat{\beta}$. The best-fitting model does not provide an appropriate representation of the data if it fails the LMP test, i.e., if $\hat{\beta} \le \beta$ with $\beta = 0.05$, say. This step can be carried out using the routine **arres**.

3. M-FILE SUMMARY

MATLAB modules come in the form of M-files with the generic filename *filename*.**m**. Omitting the filename extension **.m**, the following M-files constitute ARFIT:

ar. Given a minimum and a maximum model order $p_{\min}$ and $p_{\max}$, respectively, **ar** uses the fast QR factorization algorithm detailed in NS to compute the order selection criteria FPE, SBC, and MSC for AR models of order $p_{\min} \le p \le p_{\max}$. Also returned are least squares estimates of the parameters $A_1, \ldots, A_{p_{\rm opt}}$, C, and w for a model of order $p_{\rm opt}$, chosen as the minimizer of a user-defined order selection criterion.

arres. Diagnostic checking of a fitted model's residuals. **arres** computes the time series of residuals (2) for a fitted model and returns the significance level $\hat{\beta}$ of the LMP statistic (5) to test the residuals for uncorrelatedness.

arconf. Computes confidence intervals for the estimates of the AR parameter matrices $A_1, \ldots, A_{p_{\rm opt}}$, the intercept vector w, and the covariance matrix C.

armode. Spectral decomposition of AR process. For a fitted AR model, **armode** computes eigenmodes and their associated periods and damping rates, as well as confidence intervals for the estimated spectral information.

acf. Plots the sample autocorrelation function of an univariate time series. In checking a fitted model diagnostically, **acf** may be used to examine whether the residuals are correlated. (**acf** calls **xcorr** from MATLAB's *Signal Processing Toolbox*.)

arsim. Monte-Carlo simulation of AR processes (cf., e.g., Lütkepohl [1993, Appendix D]).

arfit. Demonstration of modules contained in the ARFIT package.

The modules above are those with which the user will typically interact. However, some of these modules depend on lower-level routines, which are also part of ARFIT:

arqr. QR factorization for an AR model. **arqr** is called by **ar**.

arord. Approximate order selection criteria for a sequence of AR models. **arord** is called by **ar**.

adjph. Multiplies a complex vector by a phase factor such that its real and imaginary parts are orthogonal and the norm of the real part is at least that of the

imaginary part. **adjph** is called by **armode** to normalize the eigenmodes of the AR model.

Placing the M-files in a directory that is accessible by MATLAB and invoking MATLAB's online help function,

help *filename*,

will display detailed information on the M-file *filename*, including the synopsis of its calling sequence. The script **arfit** serves to demonstrate the main features of the M-files listed above. It illustrates our recommendations on the use of ARFIT by guiding the user through the whole model identification process for a specific example with simulated data.

To become familiar with ARFIT and to test its installation, we suggest to run the demo script **arfit** first and then to explore details of the implementations by invoking MATLAB's online help function.

At the URL http://solon.cma.univie.ac.at/~neum/software/arfit/, the ARFIT package is electronically available through the WWW.

REFERENCES

AKAIKE, H. 1971. Autoregressive model fitting for control. *Ann. Inst. Statist. Math. 23*, 163–180.

BOX, G. E. P. AND JENKINS, G. M. 1970. *Time Series Analysis: Forecasting and Control.* Holden-Day, San Francisco.

BOX, G. E. P. AND PIERCE, D. A. 1970. Distribution of the residual autocorrelations in autoregressive-integrated moving average time series models. *J. Amer. Statist. Ass. 65*, 1509–1526.

BROCKWELL, P. J. AND DAVIS, A. 1991. *Time Series: Theory and Methods* (2nd ed.). Springer, New York.

KRAUSS, T. P., SHURE, L., AND LITTLE, J. N. 1994. *Signal Processing Toolbox.* The MathWorks, Inc., Natick, Massachussetts.

LI, W. K. AND MCLEOD, A. I. 1981. Distribution of the residual autocorrelations in multivariate ARMA time series models. *J. R. Stat. Soc. B 43*, 231–239.

LÜTKEPOHL, H. 1985. Comparison of criteria for estimating the order of a vector autoregressive process. *J. Time Series Anal. 6*, 35–52.

LÜTKEPOHL, H. 1993. *Introduction to Multiple Time Series Analysis* (2nd ed.). Springer-Verlag, Berlin.

NEUMAIER, A. AND SCHNEIDER, T. 1997. Multivariate autoregressive and Ornstein-Uhlenbeck processes: Estimates for Order, Parameters, Spectral Information, and Confidence Regions. Submitted to *ACM Trans. Math. Softw.*

SCHWARZ, G. 1978. Estimating the dimension of a model. *Ann. Statistics 6*, 461–464.

TIAO, G. C. AND BOX, G. E. P. 1981. Modeling multiple time series with applications. *J. Am. Stat. Assoc. 76*, 802–816.