HIGHER-ORDER STATISTICS AND MODEL ORDER DETERMINATION

By

Adnan Al-Smadi

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in
Electrical Engineering

August, 1995

Nashville, Tennessee

Approved:

[Signatures]

Date:

[Signatures]

20 July '95
CHAPTER I

INTRODUCTION

During recent years, there has been a lot of interest in using higher-order statistics (such as cumulants) in signal processing and system identification problems. There are several reasons behind this interest [1],[2],[3],[4],[5],[43]. First, higher-order cumulants are blind to all kinds of Gaussian processes, hence cumulants suppress additive colored Gaussian noise. Therefore if the signal to be analyzed is contaminated by additive Gaussian noise, the noise will vanish in the cumulant domain. Second, cumulants are useful for identifying nonminimum phase systems or for reconstructing nonminimum phase signals if the signals are non-Gaussian. That is because cumulants preserve the phase information of the signal. Third, cumulants are useful for detecting and characterizing the properties of nonlinear systems.

Higher-order statistics have many advantages over autocorrelation based methods. Many real-world signals are non-Gaussian, and until recently researchers typically treated these signals as if they were Gaussian due to lack of appropriate analytical tools. Under the Gaussian assumption, the information contained in the power spectrum is obtained from a second order analysis (e.g., the autocorrelation sequence) [4] which does not preserve phase information. Hence, a nonminimum phase system will be identified as being minimum phase [5]. If the input sequence is Gaussian, then the true zeros of the system that are outside the unit circle cannot be correctly identified [6]. Therefore, when using cumulants the input sequence must be zero-mean,
independent identically distributed (i.i.d.), and non-Gaussian.

In geophysical (seismic) and biological (e.g., brain) signals, the purpose of processing the signal is to detect and characterize these signals from measurement sensor data. Since these signals are nonlinear, for example involving quadratic or cubic terms, the use of power spectrum (autocorrelation) based techniques cannot distinguish nonlinear relationships from independent signals with the same resonance conditions [3].

In system identification where a given sequence represents the output of an autoregressive moving average (ARMA) process, the estimation of the proper ARMA model order and parameters is an important problem. Similarly, in communications, control, and signal processing, the problem of fitting the correct ARMA model to the available data sequence is of fundamental interest. For example, in spectral analysis, the accuracy of the frequency estimates depends on the estimated order of the prediction filter [51]. “Experience have shown that model of incorrect order can cause serious problems in control design” [71]. Also the problem of ARMA order determination plays a significant role in many areas such as speech processing and seimology, where forecasting is extremely important [52]. However, the problem of selecting the proper ARMA model order is a difficult one and has never been solved satisfactorily [5],[46],[47].

The existing ARMA model order determination methods can be classified into two basic groups as information criterion methods and linear algebra methods [47], with most techniques in existence being based on second order analysis [7],[8],[9],[10],[11]. Akaike proposed a criterion for selecting the model order called the final prediction error (FPE) [12]. A second criterion was also proposed by Akaike called an information
theoretic criterion (AIC) [10], [42]. This method is based on the expected variance of the prediction error. An alternative information criterion was proposed by Rissanen and Schwarz [8],[9],[13],[14], [44]. This algorithm is based on selecting the order that minimizes the description length, and is known as the minimum description length (MDL) principle. Model order determination methods that use FPE, AIC, and MDL include some penalty function terms. To use such methods, one has to select many possible ARMA orders, estimate the corresponding ARMA coefficients, and then calculate the maximum likelihood estimates of the prediction error variance, $\sigma^2$, of the input for all pairs of the orders $(p, q)$ such that $0 \leq p \leq p_{\text{max}}$ and $0 \leq q \leq q_{\text{max}}$. The $(p, q)$ pair resulting in the lowest value of the selected criterion is considered to be the best estimate of the correct model order. Hence, the ARMA coefficients must be estimated for each $(p, q)$ pair of the ARMA model. Therefore, the use of these techniques is often computationally expensive and in some cases inconsistent [69]. Kashyap [49] established theoretically the inconsistency of the AIC rule for choosing the order of an AR model. He claimed that the AIC rule is statistically inconsistent in the sense that the probability of error in picking the correct order does not go to zero even when $N$ tends to infinity. On the other hand, the MDL information criterion is statistically consistent for the selection of the AR model order [44],[53].

Other techniques than FPE, AIC, and MDL have been proposed. Ribeiro and Moura [41] developed the $LD^2$-ARMA identification algorithm which combines an order selection scheme with a linear, dual, decoupling algorithm for estimation of the autoregressive (AR) and moving average (MA) components. Beguin et al. [46] proposed a theoretical approach for solving the ARMA model order estimation problem called the corner method. His method is based on the autocorrelation function. A
technique was proposed by Chan and Wood [15] to determine the ARMA orders in one pass. It uses the Gram-Schmidt orthonormalization procedure to determine the linear dependency of the columns of a matrix. In their algorithm, an autocorrelation matrix is formed from the output sequence of an ARMA process, then the linear dependency of each column is checked with the Gram-Schmidt orthonormalization procedure. The first column at which linear dependency occurs is the AR order. The next independent column is used to calculate the MA order. Zhang [47] found that this algorithm is not very practical because it is difficult for one to check for the linear independence of the column vectors of the sample correlation matrix when the data length is short.

Nikias [16], [17] proposed an algorithm for estimating the MA model order based on parametric modeling of the third order cumulants of the sequence, and causal and anticausal ARMA models. The magnitude and phase response are expressed in terms of the AR parameters of two of the ARMA models. The AR part of the causal ARMA model contains the minimum phase component of the system, while the AR part of the anticausal ARMA model contains the maximum phase component. An overdetermined linear system of equations composed of the third order cumulants is formed and solved to obtain both sets of AR parameters. The MA model order is estimated based on a matrix rank determination. It should be noted that this algorithm requires prior estimation of the parameters of the system. Nikias and Pan [18] proposed a procedure to estimate the MA model order similar to the one above, but using fourth order cumulants. Alshebeili [73] proposed a method for estimating the MA model order by minimizing a cost function defined in terms of a 2-D slice of output fourth order cumulants.
Giannakis and Mendel [19] proposed an algorithm for estimating the order of an AR model-based on the rank of a matrix formed using third order cumulants. They assumed that the ARMA orders are unknown, but they also assumed some knowledge of the upper bounds. The order of the AR model is the rank of an extended matrix of cumulants. For a robust rank determination, the singular value decomposition (SVD) is applied. In this case, the AR model order equals the maximum number of nonzero singular values. However, since the extended matrix contains third order cumulant estimates computed from the data, most of the singular values will be nonzero. Hence, the largest drop between two successive singular values of the extended matrix is used to identify the AR order. Zhang and Zhang [39] proposed an algorithm to determine the MA model order of a general ARMA process using cumulants assuming the AR order and parameters are known. Similar to Giannakis and Mendel, they also estimated the rank of an extended matrix of cumulants using the SVD. Notice that in many of these higher order statistics approaches, one has to estimate the AR order and the AR parameters before the MA order can be estimated.

Recently, Liang et al. [7] proposed an algorithm for estimating the ARMA model order that does not require prior estimation of the model parameters. Their algorithm is derived from the MDL technique and is based on the minimum eigenvalues of a family of covariance matrices. They calculate a table of minimum eigenvalues for all possible values of $p$ and $q$ and then search for the corner where the minimum eigenvalue drops quickly. Also Davila and Chiang [70] have developed a new algorithm for estimating the model order that looks at input/output data covariance matrix eigenvectors. The method uses the noise subspace eigenvectors of the sample data covariance matrix. These two approaches are very new and yield a level of