LEARNING AUTOREGRESSIVE MODEL ORDER FOR MULTI-STEP AHEAD PREDICTIONS

Mikhail Prokopenko

CSIRO Division of Information Technology,
Locked Bag 17, North Ryde, NSW 2113, Australia
E-mail: mikhail@syd.dit.csiro.au, Phone: [+61 2] 325-3144, Fax: [61 2] 325-3101

ABSTRACT: This paper addresses the problem of learning an order of an autoregressive (AR) model with multi-step ahead prediction properties. It reviews various aspects of learning of a model order (model selection problem) and the techniques originated in fields related to Computational Learning Theory and developed in the context of time series analysis. The paper presents a new selection criterion (*"pattern residuals' interdependence measure estimator"* - PRIME), briefly describes a computational learning algorithm, and compares the multi-step ahead predictions for actual data obtained by different models. In order to derive a criterion for the selection of an AR model with sufficient predictive power we have examined the interdependence among residuals obtained from a given training set (an observed time series), a time series modelled by an autoregressive-moving average (ARMA) process, and a corresponding pattern extracted from the original data by a smoothing filter. The filtered series is considered as a deterministic pattern of "true series". A mean expected log-likelihood function of a correlation coefficient between the residuals is proposed as a measure of the residuals' interdependence. The comparison between models selected by the proposed criterion and well-known criteria provides favourable evidence for the strength of the PRIME criterion.

1. INTRODUCTION.

Fundamental principles of Computational Learning Theory have strong interconnections with traditional methodology developed in the fields of applied mathematics (interpolation and function approximation), engineering (signal filtering and pattern recognition), statistics (regression analysis, ARMA modelling, classification). At the same time the paradigms of computer learning intersect with more recent (and historically distinct) research in artificial intelligence (machine learning) and the neurosciences (neural networks). All these approaches have the same goal: to build a system that can learn from its environment. In particular, the system is allowed to see a set of examples and must develop a hypothesis (a model) that can explain those examples [8]. This process of "detecting and perhaps formally expressing an apparent regularity or pattern in a body of data about many particular instances or events" can be called inductive learning [18] and covers such topics as the acquisition of general theories from particular data, and "explanation" as a way of making data less surprising or more probable under a plausible theory.

It is well known in time series analysis that the aim of model selection is not only to check the adequacy of a model, or to identify a system but also to obtain a good predictor [15]. The same idea supports the development of learning systems which attempt to extract useful prediction rules by processing data taken from the past - in other words, based only on cases for which the values of both inputs and outputs of a real system have been determined (a given

training set). If the environment in which the system operates changes dynamically or/and the input information is uncertain or otherwise deficient then an accepted hypothesis (a selected model) can quickly become inadequate. These more realistic scenarios suggest the consideration of models capable of providing an effective multi-step ahead prediction.

The general selection problem for time series [9] can be reduced to a model selection, eg., selection of an autoregressive model order, if a given class of candidate spectrum estimates which can be computed from the data consists entirely of finite-parameter models, eg., autoregressions. Hence estimating the orders of p and q of an appropriate ARMA(p,q) model for given time series data is the most challenging problem in time series analysis [13].

Usually "a bivariate process" $\{y_t, x_t\}$ $(t = 0, \pm 1, ...)$ is defined in order to discuss the important class of prediction problems considered in time series analysis: how can we best predict y_t from $\{x_s, s \leq t\}$? More precisely, "If $y_t = x_{t+v}$, $v > 0$, then the problem is that of predicting the 'future' of x_t on the basis of its past". If $x_t = z_t + e_t$, where z_t is the signal and e_t the noise and $y_t = z_{t+v}$ then "for $v = 0$ the problem is that of 'signal extraction', for $v > 0$ that of predicting the signal and for $v < 0$ that of interpolating the signal, in the presence of noise" [4]. The collective term estimation can be used for referring to such problems. In the early 1940's, Wiener [19] and Kolmogorov [11] solved an important problem in random process theory, that of providing an estimate of a random signal process on the basis of observation of the signal process additively corrupted by noise. Their solution was dependent on the assumption of stationarity, ergodicity, and knowledge of the entire past of observed process. The end result of their investigation was the specification of the weighting function of the optimal estimator (filter). A sequential form of the solution known as the Kalman filter was obtained later by Kalman [10]. The Kolmogorov-Wiener prediction theory of stationary stochastic processes leads to frequency-domain formulae which require power spectrum estimates for their practical implementation. The unifying force in the time-domain and frequency-domain approaches was the celebrated Wold decomposition theorem stating that any stationary process can be uniquely represented as the sum of linearly deterministic process and purely nondeterministic, or an MA(∞) process, and these components are mutually uncorrelated [20].

The most influential work on time series forecasting in the 1960's involved ideas of Wold decomposition was carried out by Box and Jenkins [5], who formulated the class of ARMA models and developed a model selection strategy. An ARMA(p,q) process is generated by the equation

$$\mathbf{z_t = \sum^{p}_{j=1} a_j z_{t-j} + \varepsilon_t + \sum^{q}_{i=1} b_i \varepsilon_{t-i}}$$

where ε_t is a zero-mean white noise. The basis of the Box-Jenkins forecasting method is formed by recognition of statistical patterns. Parzen [12] mentioned the importance of pattern recognition characterising the following problem - "from what has been observed, infer significant characteristics of the process generating the data such as significant time lags, significant frequencies, extractable signals, and noise". In other words the task of forecasting "demands local structure and alternative parametrization of $\{X_t\}$" [13].

The model selection strategy advocated by Box and Jenkins consists of three stages: identification, estimation and diagnostic checking. At the first stage the principle of parsimony [5] is of major importance. It is worth noting that the interpretation of learning as a competition [18] uses the similar idea of complexity: other things being equal, simple theories are to be preferred to complex ones (Occam's Razor). Another analogy between learning and model selection is that a degree of fit to the known data is also a critical criterion. "The primary objective of the concept

learner is to infer a classification rule that describes the target concept. While attempting to achieve the primary objective, the secondary objective is to infer a classification rule that is as close as possible to the target concept" [17]. If ARMA model is a classification rule, then a model order selection and parameters estimation correspond to inferring the rule "closest" to the target concept (true model).

Returning to the problem of identification, it is necessary to mention that a typical way of realising balancing behaviour (between overfitting risks and underfitting risks) is to select an order p (AR) and an order q (MA) which minimise

$$\mathbf{T + \alpha\,(p + q),} \qquad \mathbf{(1)}$$

where T is one of test statistics. The second term can be considered as a penalty term for the complexity of the model, and a term for compensating the random fluctuation of T [15]. There exist several so called criterion procedures: Akaike Information Criterion (AIC) [1,2], Schwarz Bayesian Criterion (SBC) [16], or ϕ [7]. All criteria are of the form of (1) with

$$\mathbf{T = - 2\ log(maximum\ likelihood)} \qquad \mathbf{(2)}$$

The most controversial point [15] is how to choose α in (1), which is 2 in AIC, log n in SBC, and c loglog n for some c > 2 in ϕ. The choice of α depends on the aim of the selection (prediction, identifying, etc.). And "the conclusion that the optimum value of α depends, in a complicated way, on unknown parameters is unhelpful for the analysis of data" [3].

The well-known AIC provides an asymptotically efficient solution to this problem but suffers some serious drawbacks. In particular, it often selects models of very high dimension if these are included as candidates [9]. Moreover, the AIC procedure is not consistent. A bias-corrected generalisation of AIC had been derived "to avoid these difficulties by providing a more nearly unbiased estimate than AIC of the expected Kullback-Leibler information" [9]. On the other hand since the criterion involves σ^2, the one-step prediction variance, the identified model may possess certain optimality properties for one-step ahead predictions. However, if the goal is multi-step ahead predictions, then it is plausible that the criterion must be modified to reflect this goal [13].

2. WELL-FITTED AR MODEL AND A FILTERED SERIES AS A PATTERN OF TRUE SERIES.

It is well known that orders of an AR representation of true and observed series must be the same for well-fitted model. The main objective of the section is to use the idea that a filtered series or deterministic pattern (a signal extracted from a signal corrupted by noise) has to have the same order of AR components as a well-fitted AR model and is expected to be very close to a true series (of probably infinite dimension). In other words, a finite-dimensional model should serve as an approximation to the infinite-dimensional truth [9]. Consideration of a linear filter in the form

$$\mathbf{y_t \equiv L(x_t) = \sum_{j=0}^{k} c_j\, x_{t-j}} \qquad \mathbf{(3)}$$

(a series y_t is formed by a linear combination of terms of a series x_t) provides an analogy between filtering of a series and AR(p) process generated by the equation

$$z_t = \sum_{j=1}^{p} a_j z_{t-j} + \varepsilon_t, \qquad (4)$$

where ε_t is a zero-mean white noise: "if z_t is an AR(p) process, it may be described in the following way: an appropriate finite backward-looking filter applied to z_t will produce a white noise series" [6]. The expression (3) describes a linear filter with non-recursive structure. A linear filter in the form

$$y_t \equiv L^r(x_t) = \alpha_0 x_t + \sum_{l=1}^{k} \alpha_l y_{t-l} \qquad (5)$$

is a recursive filter (the superscript r is for recursion). The difference between (3) and (5) is that the current value of series y_t in (5) is a recursively convoluted function of the current value of series x_t and past values of the filtered series y_t unlike (3) where y_t is determined merely by the past values of the unfiltered series x_t. There exists a transformation of a recursive filter (5) to a non-recursive one, if the order k of filter (5) is the same as in (3) [14]. The transformation is given by $c_0 = \alpha_0$ and k recursive equations:

$$c_l = \sum_{i=1}^{l} \alpha_i c_{l-i} \qquad (6)$$

for l = 1, ... , k, which set the dependency between coefficients c_l of $L(x_t)$ and α_l of $L^r(x_t)$.

In order to construct a linear filter $L(x_t)$ of order k representing a deterministic pattern of true series we associate $x_t = z_t + e_t$, where x_t is a series of observed data and z_t is a true series, with some certain model of ARMA(p,q) process, namely with a model where p = k. As a first step of model selection an attempt should be made to approximate the process by an autoregression of order p, i.e. an AR(p) process. Consider an AR(p) model (4). In general, the white noise disturbance term ε_t differs from the noise $e_t = x_t - z_t$, which emerged when the realisation x_t was observed from an unknown true stochastic process z_t, and includes also an unknown stochastic term (shock) δ_t: $\varepsilon_t = e_t + \delta_t$. Hence the expression (4) can be transformed to

$$z_t = \sum_{j=1}^{p} a_j z_{t-j} + x_t - z_t + \delta_t$$

or

$$z_t - (1/2)\delta_t = (1/2)x_t + (1/2) \sum_{j=1}^{p} a_j z_{t-j} \qquad (7)$$

The intuitive explanation of this representation is that in the case when the nondeterministic component of the model ε_t coincides with the error term e_t or, in other words, does not include the stochastic shock δ_t the true series can be decomposed into two equally-weighted components: the current observed value and the past values of true series.

Define a filtered series y_t as a deterministic pattern of the true series:

$$y_{t-s} = \begin{cases} z_{t-s} - (1/2)\,\delta_{t-s}, & s = 0 \\ z_{t-s}, & s > 0 \end{cases} \qquad (8)$$

Then we can rewrite (7) in the form

$$y_t = (1/2)x_t + (1/2) \sum_{j=1}^{p} a_j y_{t-j}, \qquad (9)$$

and compare (9) with $L^r(x_t)$ (5), where k = p. The comparison yields immediately that $\alpha_0 = ½$ and $\alpha_j = (1/2)a_j$. So coefficients c_l of the smoothing filter $L(x_t)$ (3) are to be calculated by the transformation (6) applied to the half-values of AR(p) coefficients.

Since $\delta_t = 2\ (z_t - y_t)$, the specification error ε_t can be decomposed as

$$\boldsymbol{\varepsilon_t = e_t + \delta_t = (x_t - z_t) + 2\ (z_t - y_t) = (x_t - y_t) + (z_t - y_t).}$$

By construction $L(x_t) = L^r(x_t) = y_t$ and the estimate of ε_t is given then by

$$\boldsymbol{\varepsilon_t^{hat} = (x_t - L(x_t)) + (X_t^{hat} - L(x_t)),} \qquad \textbf{(10)}$$

where X_t^{hat} denotes estimate of z_t by the well-fitted AR model (4).

3. DERIVATION OF THE "PRIME" CRITERION AND ORTHOGONALITY ASSUMPTION.

After introducing a principle of statistical model building based on the maximum likelihood estimate (MLE) Akaike extended this classical principle and derived the minimum information theoretic criterion estimate that allowed him to explicitly formulate the problem of statistical identification as a problem of estimation and completely eliminate "the need of the subjective judgement required in the hypothesis testing procedure for the decision on the levels of significance" [2]. In the well-known work [1] instead of using the expected squared prediction error of a future observation, Akaike has adopted the mean information

$$\boldsymbol{I(g,f(x \mid \theta)) = E\ [\ \log g(x) - \log f(x \mid \theta)\ ]}$$

for discrimination between the density function $g(\bullet)$ of the true model and the density function $f(\bullet)$ of the model used for prediction, where $f(x \mid \theta)$ is a parametric family of density functions and θ is vector of parameters. Akaike has defined AIC of θ by

$$\boldsymbol{AIC(\theta_{MLE}) = (-2)\ \log(\text{maximum likelihood}) + 2k,} \qquad \textbf{(11)}$$

where the number k of independently adjusted parameters within the model is added to correct the downward bias [2].

The PRIME (*"pattern residuals' interdependence measure estimator"*) criterion can not be viewed entirely as an information criterion. However, it incorporates a likelihood and a penalty term to take away some of the "natural growth" that absolute value of log(likelihood) has for large orders. In other words, the criterion has the structure similar to (1)-(2). The basic idea behind the derivation of the PRIME criterion [14] lies in geometric interpretation of the problem of construction of a linear predictor for a multi-step ahead forecast. Geometrically, linear modelling (regression, ARMA model, optimal filtering, etc.) is finding a point on a plane (or a hyperplane) "closest" to the given point in n-dimensional space Ω (n - sample size). The point in space is dependent variable X_t, the plane is determined by independent vectors X_{t-j} (j = 1 ... k), and the point closest to X_t on the X-plane is X_t^{hat}. The vector connecting the point X_t to the X-plane is determined by k model parameters. It is well known that problem of estimation X_t^{hat} depends on whether or not X-plane is well-identified. When two or more X_{t-j} vectors are too closely related (multicollinearity), the resulting plane is unstably determined in the sense that a small change in one of the vectors X_{t-j} can twist the X-plane substantially and can lead to a widely varying estimates of parameters. On the other hand, if the vectors X_{t-j} are orthogonal the X-plane and identified parameters are stable.

Let Ω^* be (n+m)-dimensional space, such that Ω is the subspace of Ω^*. Denote the series X_t in Ω^* as X_t^*. Since our goal is multi-step ahead predictions, i.e. finding $X_t^{*\,hat}$ in (n+m)-dimensional space Ω^* based on information and the structure of n-dimensional space Ω, we can interpret X_t^{hat} as the projection of $X_t^{*\,hat}$ onto Ω. A linear combination of X_t and X_{t-j} (j = 1 ... k), for example, the linear filter $L(x_t)$ (3) applied to observed data X and representing a deterministic pattern of the true series, also can be interpreted as the projection of the true series from Ω^*. So if filter $L^*(x_t)$ represents true series in Ω^* we have to assume that the projections X_t^{hat} and $L(x_t)$ of $X_t^{*\,hat}$ and $L^*(x_t)$, respectively, are well-defined in terms of the observed series X_t. The proposed criterion is based on the assumption that stable multi-step ahead forecast requires the well-identified plane created by points X_t, X_t^{hat}, and $L(x_t)$ in the Ω space. In other words, vectors $(L(x_t) - X_t)$ and $(L(x_t) - X_t^{hat})$ should be orthogonal, if X_t^{hat} is the estimate by a well-fitted AR model. The equality $(u \times v)/(\| u \| \times \| v \|) = \cos \beta$, where $\| u \| = \sum^n_{i=1} u_i^2$ denotes Euclidean norm, and $\times$ means inner product, holds for every pair of vectors u and v, and the angle β formed by them. The "orthogonality" assumption leads to the condition

$$\mathbf{(L(x_t) - X_t) \times (L(x_t) - X_t^{hat}) / (\| L(x_t) - X_t \| \, \| L(x_t) - X_t^{hat} \|) = 0, \quad (12)}$$

since $\cos \beta = 0$.

Assuming both residuals vectors centred: $E(L(x_t) - X_t) = 0$, $E(L(x_t) - X_t^{hat}) = 0$, condition (12) can be rewritten as

$$\mathbf{\rho = corr(L(x_t) - X_t, L(x_t) - X_t^{hat}) = 0,} \quad \mathbf{(13)}$$

for correlation coefficient ρ between residuals $L(x_t) - X_t^{hat}$ and $L(x_t) - X_t$ constructed after ARMA modelling and data filtering. It is worth noting that the estimates of both residuals terms in right-hand side of the expression (10) have to be uncorrelated under the orthogonality assumption (12). So the presented approach can be supported by the decomposition (10) of the specification error ε_t into two uncorrelated components.

The PRIME procedure estimates expected log likelihood of the correlation coefficient ρ for each model and selects the model which attains the minimum expected log likelihood. Unlike the AIC procedure which maximises expected log likelihood of the model distribution, the PRIME procedure minimises expected log likelihood of the bivariate distribution of two orthogonal residuals. Obviously the likelihood of such a coexistence has to be minimal under aforementioned assumptions. The bivariate log likelihood of correlation coefficient ρ can be formed from the standardised bivariate normal distribution:

$$\mathbf{dF = [1/(2\pi(1-\rho^2)^{1/2})] \exp\{-1/(2(1-\rho^2))(x^2 - 2\rho xy + y^2)\}dxdy; \qquad -\infty \le x, y \le \infty; \quad |\rho| < 1}$$

and is defined by

$$\mathbf{\mathit{l}(\rho) = -(n-k)\log(2\pi) - (n-k)\log(1-\rho^2)/2 - 1/(2(1-\rho^2)) \left(\sum^n_{i=k+1} x_i^2 - 2\rho \sum^n_{i=k+1} x_i y_i + \sum^n_{i=k+1} y_i^2\right)}$$

Let the maximum likelihood estimator of the correlation coefficient in sample of n-k be ρ_k. Then after taking expectation E of $l(\rho)$ and using the properties $E(x^2) = E(y^2) = 1$, $E(xy) = \rho$ [14], we can formulate the problem of model selection as

$$\min_{\rho_k} E\mathit{l}(\rho_k) = \min \{ -(n-k)\log(2\pi) - (n-k)\log(1-\rho_k^2)/2 - (n-k) + k \}, \qquad (14)$$

where the additional k corrects the bias analogously to (11). The object of the optimisation in (14) can be inverted which leads to the final expression for the PRIME criterion:

$$\max_{\rho_k} E\mathit{l}(\rho) = \max \{(n-k)\log[4\pi^2(1-\rho_k^2)] + 2n - 4k\}. \qquad (15)$$

It is clear that the maximum attains when $\rho_k = 0$.

4. EXPERIMENTAL RESULTS AND CONCLUDING REMARKS.

A computational algorithm of the PRIME procedure [14] involves construction of a smoothing filter by the transformation (6) applied to AR coefficients for every candidate model. When both AR model and filtered series are obtained, the algorithm calculates ρ_k and PRIME value for a given order k. Finally, it selects the model orders corresponding to local maxima of (15). One potential benefit of the PRIME procedure is that obtained "smoothed" filter estimates of the observed data do not have to be adapted for each step of prediction.

In order to provide evidence for the strength of the PRIME criterion two methods have been utilised. Firstly, the PRIME criterion was compared with AIC and SBC by a Monte Carlo simulation with 100 replications of the Gaussian AR(3) model with 200 observations. The results of the Monte Carlo simulation generate an empirical distribution of the proposed estimator and provide evidence of appropriate identification of data. The PRIME procedure selects the 3rd order by maximising (15). With growth of order k the PRIME value is steadily decreasing. The AIC demonstrates similar behaviour, although after correct selection of the 3rd order as a global minimum it indicates local minima for large k. It is well known that AIC has a tendency to select longer lags than the true number of lags. SBC obviously does not have such fluctuations. That on the other hand might be the reason for slightly less predictive power for the multi-step ahead forecast of SBC than AIC - in order to predict many steps ahead a model has to include low-frequency components (longer lags). Also the PRIME provides sharper results in indication of the global optimum than AIC and SBC. Similar results were obtained in simulations with AR models of higher order.

Secondly, multi-step ahead post-sample forecasts have been generated for actual data* by models identified by PRIME criterion and AIC, and an h-step mean squared error of prediction has been estimated ($1 \le h \le 10$) for both models. The comparison showed that the alternative model (selected by PRIME criterion) has a larger predictive power than the basic one (selected by AIC): the h-step mean squared error of prediction was smaller uniformly for the alternative model.

The comparison between the PRIME criterion and well-known criteria (AIC, SBC) provides favourable evidence for the proposed criterion. The two utilised methods demonstrated that the PRIME procedure takes more into account longer lags if these are included as candidates than SBC does, although it does not have a tendency to overestimate model as AIC. Additional research is required in order to work out more rigorous approach to the orthogonality assumption. A derivation of a special procedure for a proper identification of MA components seems to be

* U.S. civilian labor force for the period: January, 1969 - October, 1993 (288 observations). The source of data is the U.S. Bureau of Labor Statistics.

necessary as well. Nevertheless, the presented results show that the PRIME criterion can be used as an appropriate tool for learning (selection of) an AR model order.

REFERENCES:

1. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle, in B.N. Petrov and F.Csaki, eds., *2nd International Symposium on Information Theory*, Akademia Kiado, Budapest, 1973, pp. 267-281.
2. Akaike, H. A New Look at the Statistical Model Identification, *IEEE Transactions on Automatic Control*, vol. AC-19, No. 6, 1974, pp. 716-723.
3. Atkinson, A.C. Likelihood Ratios, Posterior Odds and Information Criteria, *Journal of Econometrics* 16, 1981, pp. 15-20.
4. Bhansali, R.J. and Karavellas, D. Wiener Filtering (with emphasis on frequency-domain approaches), in David R. Brillinger and Paruchuri R. Krishnaiah (eds.*), Time Series in the Frequency Domain. Handbook of Statistics*, vol.3, 1983, pp. 1-19.
5. Box, G.E.P. and Jenkins, G.M. *Time Series Analysis: Forecasting and Control*, revised edn. San Francisco: Holden Day, 1976.
6. Granger, C.W.J. and Newbold, P. *Forecasting economic time series*.Academic Press,1986.
7. Hannan, E.J. and Quinn, B.G. The determination of the order of an autoregression. *J. Roy. Statist. Soc*. Ser. B 41, 1979, pp. 190-195.
8. Hanson, S.J., Petsche, T., Kearns, M., and Rivest, R.L. *Computational Learning Theory And Natural Learning Systems*. Volume II. The MIT Press, 1994.
9. Hurvich C.M. Selection of time series models and spectrum estimates using a bias-corrected generalisation of AIC, in *New Direction in Time Series Analysis*, v. 46, 1990, pp. 155-168.
10. Kalman, R.E. A New Approach to Linear Filtering and prediction problems, *Journal of Basic Engineering*, 82D, 1960, pp. 35-45.
11. Kolmogorov, A.N. Interpolation and Extrapolation. *Bull. Acad. Sci. USSR*, Ser. Math. 5, 1941, pp. 3-14.
12. Parzen, E. Some Recent Advances in Time Series Modeling. *IEEE Transactions on Automatic Control*, vol. AC-19, No. 6, 1974, pp. 723-730.
13. Pourahmadi, M. Fundamental Roles of the Idea of Regression and Wold Decomposition in Time Series, in *New Direction in Time Series Analysis*, v.46, 1990, pp.287-314.
14. Prokopenko, M.I. *An Autoregressive Model Order Selection Based on Estimation of Pattern Residuals Interdependence*, M.A. thesis, University of Missouri - Columbia, UMI, Ann Arbor, 1994.
15. Shibata, R. Various Model Selection Techniques in Time Series Analysis, in E.J. Hannan, P.R. Krishnaiah and M.M. Rao (eds.*), Time Series in the Time Domain. Handbook of Statistics*, vol.5, 1985, pp. 179-187.
16. Schwarz, C. Estimating the dimension of a model. *Ann. Statist.* 6, 1978, pp. 461-464.
17. Utgoff, P.E. *Machine Learning of Inductive Bias*. Kluwer Academic Publishers, 1986.
18. Wallace, C.S. *Machine Learning. Lecture presented at the Second Australian Theory Day*, UNSW, Sydney, 17.02.1995.
19. Wiener, N. *Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*, Wiley, New York, 1949.
20. Wold, H. *A Study in the Analysis of Stationary Time Series*, 2nd ed. Uppsala: Almquist and Wicksell, 1954.