Monograph:
Leontaritis, I.J. and Billings, S.A. (1986) Model Selection and Validation Methods for Nonlinear Systems. Research Report. Acse Report 292 . Dept of Automatic Control and System Engineering. University of Sheffield

[redacted]

MODEL SELECTION AND VALIDATION METHODS FOR NONLINEAR SYSTEMS

I. J. Leontaritis B.Sc., M.Sc., Ph.D.

S. A. Billings Ph.D., B.Eng., C.Eng.,
MIEE, M.Inst.Mc., AFIMA.

Department of Control Engineering,
University of Sheffield,
Mappin Street,
Sheffield. S1 3JD.

March 1986

Research Report No. 292

Abstract

The theory of hypothesis testing is used to select a model with the correct structure and the relation o[illegible] such a method to the AIC and FPE criteria is investigated. Parame[illegible]alidation and correlation validation methods are develo[illegible]d for [illegible]r difference equation models. Several shortcomings of [illegible]nal me[illegible]pecially when applied to nonlinear systems are de[illegible]

1. Introduction

The coupled problems of selecting a parametric model of the correct structure and of validating the created model to ensure that it is acceptable are studied for nonlinear difference equation models. |Leontaritis and Billings, 1985, and Billings and Voon, 1986a|.

Initially, the likelihood ratio test |Goodwin and Payne, 1977| is described for the case where the test is between two parametric models one being a restriction of the other. The likelihood ratio test is then extended in section 3 to work with a prediction error estimator |Ljung and Soderstrom 1983, Billings and Voon 1986a| to produce a new test called the log determinant ratio test. An expression relating the power of the two tests is derived to enable a comparison to be made between them. An efficient way of calculating the two tests is provided and a comparison with the traditional F-test |Astrom and Eykhoff(1973)| is included.

The selection of one model from many competing models is considered in section 4. The requirement of non-conflicting pairwise comparisons leads to the criterion which the finally selected model must minimize. The criterion depends on a significance level, the probability of accepting a model with one more parameter than the true model. The AIC criterion |Akaike 1974a,b, Priestley 1981| corresponds to the derived criterion for a particular significance level and thus an explanation of the problems related to the use of the AIC criterion is included. The other objective criterion for multiple model selection, the FPE criterion |Akaike 1969, Priestley 1981| is also considered. It is shown to be asymptotically equivalent to the AIC criterion and thus to correspond to the same significance level. Consistemt criteria that correspond to significant levels which depend on the number of data points are briefly discussed. Stepwise backward elimination of parameters and stepwise forward inclusion of parameters |Draper and Smith 1981| are also described as practical methods of selecting a reduced model.

In section 5, two distinct model validation methods are discussed. Initially a parametric model validation method is described. Next the non-parametric correlation validation methods are defined rigorously as proper hypothesis tests and compared with the traditional correlation validation methods which employ the autocorrelation function of the residuals and the cross-correlation function between the inputs and the residuals |Box and Jenkins 1976| . The application of the correlation tests to the non-linear case and the comparison with the parametric validation methods is discussed and simulated examples are included to demonstrate the application of the results.

2. Hypothesis testing

There are many occasions when one is concerned not only with the estimation of a parameter vector but with the selection of one of two models that can describe the data. The problem can be formulated as a statistical hypothesis testing problem |Kendall and Stuart (1967)| . The key ideas of the classical theory of hypothesis testing follow.

Two hypotheses are always involved in the problem. The null hypothesis denoted by H_o and the alternative hypothesis denoted by H_1. The null hypothesis is the one that is not rejected unless the data provide strong evidence that it is not correct. The null hypothesis thus corresponds to the established situation which the data try to prove incorrect. Two types of possible error in the decision can be made.

Type 1 Error: Reject the null hypothesis when it is actually true.

Type II Error: Accept the null hypothesis when it is actually false.

The decision of accepting or rejecting the null hypothesis is based on the partition of the sample space of the data y into two separate regions. If the realization of the data y is in the one region, the null hypothesis is accepted and if it is in the other, the null hypothesis is rejected. The probability of committing a type I error is called the level of significance of the test and is denoted by α and the probability of committing a type II error is denoted by β.

$$\alpha = \Pr \text{ (type I error)} \qquad (1)$$

$$\beta = \Pr \text{ (type II error)}$$

The quantity $1-\beta$ is called the power of the test. The basic idea behind hypothesis testing is that the partition of the sample space of the data y is done in such a way that the level of significance is a given small number and the power of the test $1-\beta$ is as large as possible. Thus it is guaranteed that the probability of rejecting the null hypothesis if it is true is very small ($=\alpha$) and the probability of accepting the null hypothesis if it is not true is as small as possible ($=\beta$).

The test that determines whether the null or the alternative hypothesis is selected is usually based on a statistic (a function of the data y). The probability density function of the statistic under the assumption that the null hypothesis is correct is calculated and the region where the statistic takes values is divided in two distinct regions. The one is called the acceptance region and the other the critical region. If the realization of the statistic falls within the acceptance region the null hypothesis H_o is accepted and if it falls within the critical region the alternative H_1 is accepted. The probability of committing a type I error, the level of significance of the test α, is calculated from the probability density function of the statistic under the null hypothesis, as the area under the density function for the critical region. The power of the test $1-\beta$ can only be calculated if the alternative hypothesis is a specific one. In that case the probability density function of the statistic under the alternative hypothesis can be calculated and the area under the density function for the acceptance region gives the probability of a type II error, β.

3. The likelihood ratio and log determinant ratio test

A very important statistic for hypothesis testing is the ratio $\lambda(y)$ of the maximum values of the likelihood function under the alternative and the null hypothesis

$$\lambda(y) = \frac{p(y|\hat{\theta}_1)}{p(y|\hat{\theta}_o)} \qquad (2)$$

where $\hat{\theta}_1$ is the maximum likelihood estimate of θ under the alternative hypothesis H_1 and $\hat{\theta}_o$ is the maximum likelihood estimate of θ under the null hypothesis H_o. The test based on this statistic is called likelihood ratio test [Goodwin and Payne 1977]. If the ratio is large then the data are more plausible under the alternative hypothesis than under the null hypothesis.

The null hypothesis considered here is that s out of the n_θ components of the parameter vector θ take a specific value. Let the parameter vector θ be rearranged so that

$$\theta = \begin{bmatrix} a \\ b \end{bmatrix} \tag{3}$$

where b is a column vector of dimension s and a is a column vector of dimension $n_\theta - s$. The null hypothesis is that the vector b is equal to a specific vector b* and the alternative hypothesis is that the vector b is unrestricted

$$\begin{aligned} H_o &: b = b^* \\ H_1 &: b \neq b^* \end{aligned} \tag{4}$$

The vector b* is usually taken as a zero vector and the null hypothesis represents a reduced model with s of the parameters equal to zero. The alternative hypothesis represents the full model with all the parameters present. The purpose of the test is to find if there is significant statistical evidence that the more complicated full model gives a better explanation of the data than the simpler reduced model. The statistic that is actually used is

$$d(y) = 2\log \lambda(y) \tag{5}$$

Under the null hypothesis H_o: b=b*, the statistic d(y) converges to a chi-square distribution with s degrees of freedom for the data sequence length $N \to \infty$ i.e.

$$d(y) = 2 \log \lambda(y) \to \chi^2(s) \tag{6}$$

The proof contains some results that will be needed later, therefore a full but not entirely rigorous version of the proof is given.

$$L(\theta) = -\log p(y|\theta) \tag{7}$$

The Hessian of $L(\theta)$ at the unrestricted maximum likelihood estimate $\hat{\theta}_1$ tends to the information matrix M. Thus for θ near θ_1 it approximately holds that

$$L(\theta) = L(\hat{\theta}_1) + \frac{1}{2}(\theta-\hat{\theta}_1)^\tau M(\theta-\hat{\theta}_1)$$

$$= L(\hat{\theta}_1) + \frac{1}{2} \begin{bmatrix} a-\hat{a}_1 \\ b-\hat{b}_1 \end{bmatrix}^\tau \begin{bmatrix} M_{aa} & M_{ab} \\ M_{ab}^\tau & M_{bb} \end{bmatrix} \begin{bmatrix} a-\hat{a}_1 \\ b-\hat{b}_1 \end{bmatrix}$$

$$= L(\hat{\theta}_1) + \frac{1}{2}(a-\hat{a}_1)^\tau M_{aa}(a-\hat{a}_1) + (a-\hat{a}_1)^\tau M_{ab}(b-\hat{b}_1) + \frac{1}{2}(b-\hat{b}_1)^\tau M_{bb}(b-\hat{b}_1) \tag{8}$$

Assuming the null hypothesis is correct the restricted and unrestricted estimates get close to each other when $N\to\infty$ and thus the restricted estimate $\hat{\theta}_o$ minimizes the function $L(\theta)$ given in (eqn.8). The restricted estimate is the vector $\hat{\theta}_o = [\hat{a}_o^T \; b^{*T}]^T$. The minimization of (eqn.8) is done with the parameter b fixed at its assumed true value of b and with a as an independent variable. Differentiating (eqn.8) with respect to a and equating to zero, the minimizing value of a, $\hat{a}_o$, is

$$\hat{a}_o = \hat{a}_1 - M_{aa}^{-1} M_{ab} (b^* - \hat{b}_1) \tag{9}$$

Substituting this value in (eqn.8) gives

$$L(\hat{\theta}_o) = L(\hat{\theta}_1) + \frac{1}{2}(b^* - \hat{b}_1)^T [M_{bb} - M_{ab}^T M_{aa}^{-1} M_{ab}] (b^* - \hat{b}_1) \tag{10}$$

But

$$d(y) = 2\log \lambda(y) = 2\log \frac{p(y|\hat{\theta}_1)}{p(y|\hat{\theta}_o)} = 2L(\hat{\theta}_o) - 2L(\hat{\theta}_1) \tag{11}$$

and from (eqn.10)

$$d(y) = (b^* - \hat{b}_1)^T [M_{bb} - M_{ab}^T M_{aa}^{-1} M_{ab}] (b^* - \hat{b}_1) \tag{12}$$

The asymptotic distribution of $\hat{\theta} - \theta$, where $\hat{\theta}$ is the maximum likelihood estimate and θ is the true value of the parameter, is a normal distribution with zero mean and covariance matrix M^{-1}, the inverse of the information matrix M. Here $\hat{b}_1$ is the unrestricted maximum likelihood estimate of the assumed true value b^*. The vector $\hat{b}_1 - b^*$ is thus asymptotically normally distributed with covariance matrix the lower right partition of M^{-1}. The lower right partition of M^{-1}, from the matrix inversion theorem, equals $(M_{bb} - M_{ab}^T M_{aa}^{-1} M_{ab})^{-1}$. The random vector $(\hat{b}_1 - b^*)^T [M_{bb} - M_{ab}^T M_{aa}^{-1} M_{ab}] (\hat{b}_1 - b^*)$ is thus asymptotically distributed as a chi-square distribution with s degrees of freedom (the dimension of the vector b). Thus from (eqn.12)

$$d(y) = 2L(\hat{\theta}_o) - 2L(\hat{\theta}_1) \to \chi^2(s) \tag{13}$$

Now that the distribution of the statistic d(y) has been determined, the critical and the acceptance region, for a specific level of significance α, can be defined. Let the critical value of the chi-square distribution with s degrees of freedom, for a level of significance α, be called $k_\alpha(s)$. The $k_\alpha(s)$ is such that the area under the chi-square density function to the left of $k_\alpha(s)$ is $1-\alpha$ and to the right is α. If d(y) is greater than $k_\alpha(s)$ (for say $\alpha=0.05$) there is strong evidence against the null bypothesis and it is thus rejected.

It has been pointed out that the power of the test $1-\beta$ should be as large as possible so that the probability of accepting the null hypothesis when it is not true is small. The power of the test can only be calculated if the alternative hypothesis H_1: $b \neq b^*$ is made specific, that is, the true value of b is not b^* but b^o. The asymptotic distribution of d(y), assuming that the alternative hypothesis is correct is now a non-central chi-square distribution with s degrees of freedom and non-centrality parameter h where

$$h = (b^* - b^o)^T [M_{bb} - M_{ab}^T M_{bb}^{-1} M_{ab}] (b^* - b^o) \tag{14}$$

The power $1-\beta$ of the test is the area under the non-central chi-square distribution $\chi^2(s,h)$ to the right of the critical point $k_\alpha(s)$ |Goodwin & Payne (1977)|.

Suppose that the unrestricted maximum likelihood estimate $\hat{\theta}_1$ and the value of $L(\hat{\theta}_1)$ is calculated. The information matrix M can be estimated as the Hessian of the function $L(\theta)$ at the point $\hat{\theta}_1$. The restricted estimate $\hat{\theta}_o$ does not actually need to be calculated from the original data because a very good approximation of $\hat{\theta}_o$ and $L(\hat{\theta}_o)$ is given by equations (9) and (10). This way the value of the statistic $d(y)=2L(\hat{\theta}_o)-2L(\hat{\theta}_1)$ given by equation (12) can be evaluated very easily and the likelihood ratio test for several different hypotheses can be done with minimum effort. In practice equation (12) is not actually used in the numerical evaluation of d(y) because it uses the inverse of the matrix M_{aa} which is time consuming to calculate. The square root methods and the Householder orthogonal transformation give a very elegant and efficient numerical solution to the problem.

The proof that the statistic d(y) is asymptotically distributed as a chi-square distribution with s degrees of freedom was based on only two assumptions. First, that the unrestricted estimate $\hat{\theta}_1$ is asymptotically normally distributed with covariance matrix a matrix M^{-1} and second, that the Hessian of the function $L(\theta)$ at $\hat{\theta}_1$ is asymptotically equal to M. In the case of the prediction error method [Ljung and Soderstrom 1983] both these facts hold where the information matrix M is repaced by the inverse of the asymptotic covariance matrix of the estimator $H=p_2^{-1}$ and the function $L(\theta)$ by the function $NJ_2(\theta)$, where N is the number of data points, $J_2(\theta) = \frac{1}{2} \log \det Q(\theta)$,

$$Q(\theta) = \frac{1}{N} \sum_{t=1}^{N} \varepsilon(t,\theta) \; \varepsilon(t,\theta)^T$$

and $\varepsilon(t,\theta)$ represents the prediction errors . The statistic

$$d(y) = 2NJ_2(\hat{\theta}_o) - 2NJ_2(\hat{\theta}_1) \tag{15}$$

where $\hat{\theta}_o$ and $\hat{\theta}_1$ are the restricted and the unrestricted prediction error estimates is thus also asymptotically distributed as a chi-square distribution with s degrees of freedom. The statistic (eqn.15) becomes

$$d(y) = 2NJ_2(\hat{\theta}_o)-2NJ_2(\hat{\theta}_1)=N\log \det Q(\hat{\theta}_o)-N\log \det Q(\hat{\theta}_1)$$

$$=N\log \frac{\det Q(\hat{\theta}_o)}{\det Q(\hat{\theta}_1)} \tag{16}$$

The test based on the statistic d(y) in eqn(16) will be called the log determinant ratio (LDR) test.

The price paid for using the prediction error method is that the power of the log determinant ratio test is smaller than the likelihood ratio test. Indeed the asymptotic distribution of the statistic d(y), assuming that the alternative hypothesis $b=b^o$ is correct, is again a non-central chi-square distribution with s degrees of freedom and non-centrality parameter h equal to

$$h = (b^*-b^o)^T \left[H_{bb} - H_{ab} H_{bb}^T H_{ab}\right] (b^*-b^o) \tag{17}$$

where $H=p_2^{-1}$, and p_2 is the asymptotic covariance matrix associated with the prediction error estimator. The matrix H is always smaller than the optimum information matrix M. Thus $H<M$ and consequently $H^{-1}>M^{-1}$. The lower right partition of H^{-1} should also be greater or equal to the lower right partition of M^{-1}. From the matrix inversion theorem these partitions are

$$(H_{bb} - H_{ab}^T H_{bb}^{-1} H_{ab})^{-1} \geq (M_{bb} - M_{ab}^T M_{bb}^{-1} M_{ab})^{-1} \tag{18}$$

Thus

$$(H_{bb}-H_{ab}^{\tau}H_{bb}^{-1}H_{ab}) \leq (M_{bb}-M_{ab}^{\tau}M_{bb}^{-1}M_{ab}) \tag{19}$$

and

$$(b^*-b^o)^{\tau}[H_{bb}-H_{ab}^{\tau}H_{bb}^{-1}H_{ab}](b^*-b^o) \leq (b^*-b^o)^{\tau}[M_{bb}-M_{ab}^{\tau}M_{bb}^{-1}M_{ab}](b^*-b^o) \tag{20}$$

The non-centrality parameter h in the case of the log determinant ratio test is thus always smaller than the non-centrality parameter of the likelihood ratio test for any b^o. Given a level of significance α and thus also a critical value $k_\alpha(s)$, the area under the non-central chi-square distribution to the right of $k_\alpha(s)$ is the power of the test $1-\beta$. The area of the non-central chi-square distribution to the right of $k_\alpha(s)$ increases for an increasing non-centrality parameter h and thus the likelihood ratio test has higher power than the log determinant ratio test. The two tests are however identical for Gaussian innovations. The loss of power by using the prediction error method is usually small since the matrix H is not much smaller than the optimum information matrix M for reasonable bell-shaped density functions of the innovations.

The value of the statistic d(y) and of $\hat{\theta}_o$ can be very easily approximately evaluated using formulae similar to (eqns.9) and (12). They are

$$\hat{a}_o=\hat{a}_1-H_{aa}^{-1}H_{ab}(b^*-\hat{b}_1) \tag{21}$$

$$d(y) = (b^*-\hat{b}_1)^{\tau}[H_{bb}-H_{ab}^{\tau}H_{aa}^{-1}H_{ab}](b^*-\hat{b}_1) \tag{22}$$

where H is the Hessian of $NJ_2(\theta)$ at the minimum $\hat{\theta}_1$.

In the very special case of ordinary least squares with normally distributed data, the traditional test to discriminate between two models is the F-test |Goodwin and Payne (1977)|. The F-test can thus only be used for single-output systems. The statistic used by the F-test is

$$\frac{Q(\hat{\theta}_o)-Q(\hat{\theta}_1)}{Q(\hat{\theta}_1)} \quad \frac{N-n_\theta}{s} \tag{23}$$

where here the sample variance $Q(\theta)$ is scalar. Under the restricted conditions of ordinary least squares it can be shown that the statistic (eqn.23) is $F(s,N-n_\theta)$ distributed for any data length N. It has been proposed that the statistic (eqn.23) is still asymptotically F-distributed for $N\to\infty$ even if the restrictive conditions of the ordinary least squares are not assumed |Astrom and Eykhoff (1973)|. The F-test has since become a traditionally used test in system identification. It will be shown here that the F-test and the log determinant ratio test are asymptotically equivalent. The log determinant ratio test can then be regarded as a generalization of the F-test to the multivariable case. Also the F-test is in this way proved to be correct under the very general conditions assumed for the log determinant ratio test. Let

$$Q_o=Q(\hat{\theta}_o)$$
$$Q_1=Q(\hat{\theta}_1) \tag{24}$$

The log determinant ratio test accepts the null hypothesis if

$$N \log \frac{Q_o}{Q_1} < k_\alpha(s) \tag{25}$$

where $k_\alpha(s)$ is the critical value of the chi-square distribution with s degrees of freedom and significance level α. The condition in eqn.(25) can be written

$$\frac{Q_o}{Q_1} < \exp(k_\alpha(s)/N) \tag{26}$$

and for $N\to\infty$ (eqn.26) is asymptotically equivalent to

$$\frac{Q_o}{Q_1} < 1 + \frac{1}{N} k_\alpha(s) \tag{27}$$

The F-test accepts the null hypothesis if

$$\frac{Q_o - Q_1}{Q_1} \quad \frac{N-n_\theta}{s} < k_\alpha(s, N-n_\theta) \tag{28}$$

where $k_\alpha(s,N-n_\theta)$ is the critical value of the F-distribution with degrees of freedom s and $N-n_\theta$. It can be shown theoretically and checked by the tables giving the critical values of the chi-squared and the F-distribution that

$$sk_\alpha(s,\infty)=k_\alpha(s) \tag{29}$$

Thus asymptotically for $N\to\infty$ condition (eqn.28) becomes

$$\frac{Q_o - Q_1}{Q_1} (N-n_\theta) < k_\alpha(s) \tag{30}$$

or since asymptotically $N-n_\theta = N$

$$\frac{Q_o}{Q_1} < 1+\frac{1}{N} k_\alpha(s) \tag{31}$$

which is exactly the condition of the log determinant ratio test. It is known that the two tests are asymptotically equivalent Soderstrom (1977). Here it is proved in a different way that they correspond to the same significance level. In practice, for data of length N>100 and a not very large number of parameters, the two tests are numerically almost identical. For instance if $\alpha=0.05$, N=123, $n_\theta=3$, s=2, then the determinant ratio test is from eqn (26)

$$\frac{Q_o}{Q_1} < 1.0499 \tag{32}$$

and the F-test is from eqn (28)

$$\frac{Q_o}{Q_1} < 1.0511 \tag{33}$$

4. Multiple selection methods

The situation where a selection has to be made between two models only is rather restrictive. The usual case is that there are many different models, each with its own parameter vector, and a single model must finally be selected. Here it is assumed that all the models are special cases of a full model with parameter vector θ. Every other model then is like the full model but with

some of the elements of the parameter vector θ equal to zero. The first problem that has to be solved is the assignment of non-conflicting significance levels to possible pairwise comparisons of models. For instance assume the full model to have four parameters, $\theta_1,\theta_2,\theta_3,\theta_4$. Let the model M_{134} be the one with parameters θ_1, θ_3 and θ_4. The rest of the reduced models are denoted accordingly. Suppose the model M_{134} is compared with M_{13} and out of the two the model M_{13} is selected. The model M_{13} is now compared with M_1 and M_1 is selected. If the model M_{134} was compared directly with M_1 the model M_1 should have been selected. The critical points for the several tests that can be performed have to be chosen so that conflicts do not appear. Assume no particular preference to any of the available parameters. Suppose that a comparison between a model and another one with a parameter vector reduced by one is done. The critical point on this occasion is k(1). In the case of the parameter vector being reduced by two, the critical point is k(2). The only way that no conflict occurs is to choose the critical points such that $k(2)=2k(1), k(3)=3k(1), \ldots$ etc. Now let two models have parameter vectors θ_1 and θ_2 with dimensions n_{θ_1} and n_{θ_2}. Assume also $n_{\theta_1}<n_{\theta_2}$ and $s=n_{\theta_2}-n_{\theta_1}$. The model with parameter vector θ_1 is selected according to the likelihood ratio test if

$$2L(\theta_1)-2L(\theta_2)<k(s)=sk(1)=(n_{\theta_2}-n_{\theta_1})k(1)=n_{\theta_2}k(1)-n_{\theta_1}k(1) \tag{34}$$

or if

$$2L(\theta_1)+n_{\theta_1}k(1)<2L(\theta_2)+n_{\theta_2}k(1) \tag{35}$$

The model that is selected amongst all the several competing models is the one that minimizes the criterion

$$C = 2L(\theta)+n_\theta k(1) \tag{36}$$

where θ is the parameter vector of the particular model and n_θ its dimension. The term $n_\theta k(1)$ is the one that takes into account the complexity of the model and penalizes the ones with a large number of parameters. In fact the full model makes the term $2L(\theta)$ minimum but the number of parmeters used is also the largest and the term $n_\theta k(1)$ assumes its highest value. Criteria that take into account the complexity of the model are said to follow the 'principle of parsimony'. The value of the criterical point k(1) is left to be decided In |Akaike (1974b)| the value of 2 is proposed and the criterion eqn.(36) with k(1)=2 is known as Akaike's information criterion (AIC).

$$AIC=2L(\theta)+2n_\theta \tag{37}$$

The AIC criterion was motivated by probabilistic information arguments for the case where the true system is very complex and an approximating model needs to be found. It is however shown here that it is a special case of the hypothesis testing method for a particular choice of significance level. The AIC criterion has been criticized because it has been shown that it may consistently overestimate the true parameter vector |Shibata (1976)|. Different values of the constant k(1) have been proposed to overcome such problems |Bhansali and Downham (1977)|. A proper understanding of the reasons behind these modifications can be obtained if the AIC criterion is considered as a hypothesis testing criterion with a specific significance level. The significance level of the AIC criterion is given by the significance level of the chi-square distribution

with s degrees of freedom and critical value $k_\alpha(s)=2s$, where s is the difference in the number of parameters of the two models to be compared. The significance levels of the AIC criterion for s = 1,2,...,20 are given in table 1

s	1	2	3	4	5	6	7	8	9	10
α	0.156	0.135	0.111	0.091	0.074	0.061	0.050	0.042	0.035	0.029

s	12	14	16	18	20
α	0.020	0.014	0.010	0.007	0.005

Table 1

The first observation is that the value of the significance level for s=1 is very large. This means that the probability of selecting a model with one more parameter than the true parameter vector is not insignificant. The significance levels for s=2,3,4,5,6 are still relatively large but for larger values of s they become acceptably small. Thus the AIC criterion does not have an insignificant probability of accepting a model with 1,2,3,4,5 or even 6 more parameters than the true model. The first and obvious way to decrease the significance levels is to select a critical value for s=1 larger than 2. The most commonly used significance levels are 0.05 or 0.01. The critical values of the chi-square distribution with one degree of freedom for significance levels 0.05 and 0.01 are 3.841 and 6.635 respectively. Thus the criterion eqn (36) with k(1) equal to or higher than 3.841 will reduce the probability of selecting a model with one more parameter than the true model to an insignificant level. A very convenient value for k(1) is 4 so that the significance level for s=1 is α=0.0456. The choice of a value of k(1) greater than 2 in the criterion eqn (36) has been proposed but without connections to significance levels. The significance levels for s>1 are always smaller than the one for s=1 and thus for the choice k(1)=4, they are also insignificant.

The disadvantage of all model selection criteria of the form (eqn.36) is that they assume an a priori knowledge of the probability density function of the data. In the system identification context this means that the real probability density function of the innovations must be known. This of course is a very severe restriction. The prediction error method was created so that this restriction is relaxed and the performance is only slightly reduced for reasonable bell-shaped density functions. The test between two models for the prediction error method is the same as in the likelihood ratio test with the log likelihood function $L(\theta)$ replaced by $NJ_2(\theta)$. In the case of selection between many models the criterion the best model must minimize is the equivalent of (eqn.36).

$$C = N\log\det Q(\theta) + n_\theta k \quad (1) \qquad (38)$$

The choice of the significance level k(1) follows the same arguments as in the case of the likelihood ratio tests and it is thus again reasonable to accept k(1)=4 for practical situations. The only disadvantage of criterion (eqn.38) is that the power of the tests might now be slightly lower and thus a larger number of data are needed to discriminate between different models. In the case where the distribution of the data is Gaussian the two criteria are exactly equivalent.

Akaike has proposed another criterion that must be minimized by the finally

selected model. It is called Final Prediction Error (FPE) criterion [Akaike 1969, Priestley 1981] and it is

$$FPE=\frac{N+n_\theta}{N-n_\theta}\ Q(\theta) \tag{39}$$

Obviously it is a criterion for single-output systems only. This criterion is actually asymptotically equivalent to the AIC criterion adapted to the prediction error method, i.e. the criterion (eqn.38) with k(1)=2. In fact assume that the two competing models have parameter vectors θ_1 and θ_2 of dimension n_{θ_1} and n_{θ_2}. Let

$$\begin{aligned} Q_1&=Q(\hat{\theta}_1) & n_1&=n_{\theta_1} \\ Q_2&=Q(\hat{\theta}_2) & n_2&=n_{\theta_2} \end{aligned} \tag{40}$$

The model with parameter vector θ_1 is selected according to the prediction error AIC criterion when

$$N\log Q_1+2n_1 < N\log Q_2+2n_2 \quad \leftrightarrow$$

$$N\log\frac{Q_1}{Q_2} < 2s \quad \leftrightarrow$$

$$\frac{Q_1}{Q_2} < \exp(2s/N) \tag{41}$$

where $s=n_2-n_1$

Equation (41) for large N asymptotically becomes

$$\frac{Q_1}{Q_2} < 1+2\frac{s}{N} \tag{42}$$

The model with parameter vector θ_1 is selected according to the FPE criterion when

$$\frac{N+n_1}{N-n_1}\ Q_1 < \frac{N+n_2}{N-n_2}\ Q_2 \quad \leftrightarrow$$

$$\frac{Q_1}{Q_2} < \frac{N-n_1}{N+n_1}\ \frac{N+n_2}{N-n_2} \tag{43}$$

which asymptotically for large N is equal to (eqn.42). Thus the prediction error adapted AIC criterion and the FPE criterion are equivalent and so the FPE criterion corresponds to the same significance levels chosen by the AIC criterion. Consequently, the FPE criterion also has a significance probability that it will choose a more complicated model than the true one. If the FPE criterion is to be equivalent to criterion (eqn.38), it must be transformed to

$$FPE^* = \frac{2N+k(1)n}{2N-k(1)n} \; Q(\theta) \qquad (44)$$

The criteria AIC and FPE are called objective criteria because they do not depend on a subjective choice of significance levels and the model selection can be done completely automatically. However it has been demonstrated here that they correspond to a hypothesis testing criterion with particular significance levels and thus their objectivity is as good as the choice of significance levels they correspond to. The significance levels of the objective criteria are not particularly well chosen and they can consistently over-estimate the number of needed parameters.

An alternative approach to the hypothesis testing method of model selection can be provided by Bayesian methods. This approach is analysed in |Kasyap (1977)|. The Bayesian approach requires several assumptions which, it may be argued, are no more plausible than the choice of the significance level required by the hypothesis testing approach. Also the results of the Bayesian method can be interpreted by the hypothesis testing method when the significance levels of the pairwise comparisons are chosen in a non-conflicting manner.

The hypothesis testing method is traditionally based on the F-test with the same significance level for all the pairwise comparisons between models. Such a method creates inconsistencies which triggered Akaike to create his 'objective' criteria AIC and FPE and Kasyap to investigate the Bayesian approach. The choice of non-conflicting significance levels avoids this problem and provides a simple and elegant solution based only on the hypothesis testing theory.

The probability of accepting a model with less parameters than in the true parameter vector can be high for a small number of data. The power of the likelihood ratio and the log determinant ratio tests increases for an increasing number of data points. In order to make sure that no fewer than the necessary parameters are accepted, the power of the test for a specific alternative hypothesis has to be calculated and if it is not high enough more data are actually needed to increase the power of the test. It is in the nature of hypothesis testing that only if there exists enough evidence the null hypothesis is rejected. In the case of a small number of data there exists a tendency of accepting the null hypothesis since the data do not provide enough evidence against it. A model is consequently rejected in favour of another one with more parameters only if the data provide enough evidence that the more complicated model is significantly better. In hypothesis testing a decrease in the significance level α has the consequence that the probability of type II error β increases and the power of the test $1-\beta$ decreases. If for a particular choice of significance level α the probability β is found to be extremely small, the significance level α can be decreased so that β increases and a more balanced proportion of types of error is obtained. In the case of model selection a large number of data points has the consequence that the power of the test can be very high thus a reduction on the significance level can be made. The critical point k(1) can thus be increased as the number of data points N increases. Criteria where the critical point k(1) is a function of the number of data points N have been proposed for the case where the order of a linear model is to be estimated. They are known as consistent criteria because asymptotically the probability of selecting the wrong order is zero. One such choice proposed in |Kasyap (1980)| and |Rissanen (1979)| is $k(1)=\log(N)$. Another one proposed in |Schwartz (1978)| is $k(1) = \frac{1}{2}\log(N)$. In practical situations where the number of data points is not extremely large, so that the power of the test is not very small, the choice of k(1) equal to 4 is a reasonable choice which was found to work well in linear and non-linear identification.

The procedure followed in order to find the best parametric model is: decide which is the most complicated model to be considered and let the parameter vector of this full model be θ of dimension n_θ. Consider all the reduced models with s of the elements of the full parameter vector θ forced to be zero (or some other convenient value). The number of all such models for some s is

$$C(n_\theta,s) = \frac{n_\theta!}{s!(n_\theta-s)!} \qquad \text{where } 0 \leq s \leq n_\theta \tag{45}$$

The number of all the competing models is

$$\sum_{s=0}^{n_\theta} C(n_\theta,s)=2^{n_\theta} \tag{46}$$

The criterion C of (eqn.36) or (38) is calculated for all the 2^{n_θ} models and the model that minimizes C is chosen as the best one. The criterion C has thus to be evaluated 2^{n_θ} times. For small values of n_θ this is not a prohibitively large number of times since the criterion can be approximately evaluated extremely quickly using (eqn.12) or (22). For a slightly large number of parameters n_θ the value of 2^{n_θ} becomes excessive. For instance for n_θ=10, 2^{n_θ} =1024, for n_θ=15, 2^{n_θ}=32768 and for n_θ= 30, 2^{n_θ}=1.07*10^9. Such a method of choosing the best model is called a combinatorial method.

One way of reducing the number of times the criterion C has to be calculated in the combinatorial method is to calculate the minimum of C sequentially for the classes s=0,1,2,...,n_θ and stop when it is found that the minimum value of C for some class is higher than the minimum value for the previous one. It is hoped that the value of C for classes with even less parameters is larger. Another way is to start with the class with no parameters and keep on increasing the number of parameters, i.e., consider the sequence of classes s=n_θ,n_θ-1,...,0. When a class is found that has a minimum value of C higher than the previous class, no more classes are considered. Again, it can only be hoped that the optimum class has been found. The number of times the criterion C has to be evaluated can still be prohibitively large particularly when fitting nonlinear models and other approximate solutions have to be found. Two very popular methods in the field of multivariate regression analysis are the Stepwise Backward Elimination (SBE) of parameters and the Stepwise Forward includion (SFI) of parameters |Draper and Smith 1981|.

The SBE method first of all calculates the criterion C for the full model (s=0). Then it considers all the models with 1 less parameter than the full model and calculates the criterion C for all of them. If the minimum value of C for all these models is greater than that of the full model it stops and accepts the full model, the model that minimizes C is accepted and the parameter that does not exist in this model is deleted and it is never considered again. The next class of models that is considered is the one with 2 parameters less than the full model, one of them being the already deleted one. The minimum value of the criterion for this class is calculated again and compared with the minimum value of the criterion of the previously considered class. If it is smaller, the parameter that is missing in the model that minimizes C is deleted and will never be considered again. The method carries on in this fashion until some class of the considered models has a minimum value of C greater than the previously considered class. The advantage of the SBE method is that the maximum number of times the criterion C needs to be evaluated is $1+n_\theta(n_\theta+1)/2$, an extremely large reduction from 2^{n_θ}

The SFI method starts from the model with no parameters ($s=n_\theta$) and calculates the criterion C for this model. Then it considers the models with 1 parameter ($s=n_\theta-1$) and calculates the minimum value of C for all these models. Again if the minimum value of C is less than that of the model with no parameters, the parameter of the model that minimized C is included and it will always belong to the models considered in the future. The method carries on including parameters until some class of considered models has a minimum value of C greater than the previously considered class. The maximum number of times the criterion C needs to be evaluated is again $1+n_\theta(n_\theta+1)/2$.

The SBE and SFI methods do not always give the optimum model with the minimum value of the criterion C provided by the combinatorial method. If however they both choose the same model there is a certain amount of confidence that this is also the optimum model. Stepwise Regression is a combination of SBE and SFI which has been shown to perform extremely well in conjunction with a prediction error estimator |Billings and Voon 1986, Billings and Fadzil 1985|.

There are some cases where the set of considered models are nested classes of models with an increasing parameter vector for the classes with increasing complexity. An example of such models is the set of linear models with increasing order. If the assumption is made that the model within a class with all the parameters present gives the criterion C its minimum value, the selection of the most appropriate class is straight forward. The criterion C is calculated for the full model of every class and then the class with the minimum value of C is selected. This is the common way of using the criterion C for the estimation of the order of a linear system.

5 Model validation methods

The general approach used in the model selection methods is to choose a full model with a parameter vector θ and then calculate the value of θ that minimizes the loss function $L(\theta)$ or $NJ_2(\theta)$ and the value of the Hessian of the loss function at the minimum. Then all the models with a reduced parameter vector are considered and the one out of them that minimizes criterion C is selected as the best one. The calculation of the criterion C for all the reduced models can easily be done once the Hessian of the full model is known. Amongst all the models that are special cases of the full model it is thus certain that the selected one is actually the best. This does not guarantee though that another model even more complicated than the full model may not be an even better one. One way of overcoming this difficulty is to choose the full model as a model so complicated that an even more complicated one would not be desirable. This might prove a very inappropriate solution. First, such a full model contains too many parameters and the minimization of the loss function might be too costly to perform. Second, an over-parametrized model has inevitably many parameters that are linearly dependent. The Hessian of the loss function thus becomes almost singular and the numerical methods employed to minimize the loss function have to be used with extreme care. Third, the number of reduced models for a full model with many parameters might become so large that the selection of one of them would be impossible. A very complicated full model then is not the correct solution. A comparison between one model and an extension of this model must be done without calculating the actual minimum of the loss function for the extended model. In this way a model can be shown to be better than any other extended model and thus validated to be the best possible one. Such an approach to validation was first considered in |Bohlin (1978)| and it is called the parametric validation method.

The traditional methods for validating a model are correlation type methods where again it is not necessary to perform the estimation of an extended model. Parametric and the correlation validation methods are discussed next.

5.1 Parametric validation method

Let the parameter vector of a model be a vector a and the parameter vector of an extended model be $[a^T b^T]^T$. The restriction of the extended model when the vector b takes the specific value b*, is the original model. It is usually taken that b*=0. The parameter vector of the extended model is then $\theta_1=[a^T b^T]^T$ and the parameter vector of the original model is the vector θ_1 with b restricted to the value b*. It is then $\theta_o=[a^T b^{*T}]^T$. Hypothesis testing is used again to decide whether the extended model is significantly better than the simpler original one or not. The null hypothesis is that the original model is correct and the alternative hypothesis is that the extended model is correct. Thus

$$\begin{aligned} H_o &: b=b^* \\ H_1 &: b\neq b^* \end{aligned} \tag{47}$$

Under the null hypothesis, the statistic d(y) of eqns (11) or (16) is asymptotically distributed as a chi-square distribution with s degrees of freedom, where s is the dimension of the vector b. Let the statistic d(y) correspond to the likelihood ratio test. For the log determinant ratio test the only difference is that the loss function $L(\theta)$ is replaced by $NJ_2(\theta)$. The statistic $d(y)=2L(\hat{\theta}_o)-2L(\hat{\theta}_1)$ has to be calculated where $L(\hat{\theta}_o)$ is the minimum of the loss function $L(\theta)$ under the null hypothesis and $L(\hat{\theta}_1)$ the minimum under the alternative hypothesis. The difference between the model selection and the validation problem is that in the latter case the minimization of the loss function is performed for the restricted model only and the value of the loss function for the extended model has to be approximated. Assuming the null hypothesis is correct, i.e. b=b*, the vector $\hat{\theta}_1=[\hat{a}_1^T \hat{b}_1^T]^T$ and the vector $\hat{\theta}_o=[\hat{a}_o^T b^{*T}]^T$ are very close to each other for a large number of data. It then approximately holds

$$L(\theta_1)=L(\hat{\theta}_o)+\left[\frac{\partial L}{\partial \theta}\right]_{\hat{\theta}_o}(\theta_1-\hat{\theta}_o)+\frac{1}{2}(\theta_1-\hat{\theta}_o)^T\left[\frac{\partial^2 L}{\partial \theta^2}\right]_{\hat{\theta}_o}(\theta_1-\hat{\theta}_o) \tag{48}$$

The value of θ_1 that minimizes $L(\theta_1)$ in eqn.(48) can be found by equating the derivative of $L(\theta_1)$ to zero. It is then

$$\hat{\theta}_1=\hat{\theta}_o-\left[\frac{\partial^2 L}{\partial \theta}\right]^{-1}_{\hat{\theta}_o}\left[\frac{\partial L}{\partial \theta}\right]^T_{\hat{\theta}_o} \tag{49}$$

and substituting eqn.(49) in (48) the statistic d(y) is found to be

$$d(y)=2L(\hat{\theta}_o)-2L(\hat{\theta}_1)=\left[\frac{\partial L}{\partial \theta}\right]_{\hat{\theta}_o}\left[\frac{\partial^2 L}{\partial \theta^2}\right]^{-1}_{\hat{\theta}_o}\left[\frac{\partial L}{\partial \theta}\right]^T_{\hat{\theta}_o} \tag{50}$$

Let the Hessian of $L(\theta)$ at the point $\hat{\theta}_o$ be partitioned as

$$\left[\frac{\partial^2 L}{\partial \theta^2}\right]_{\theta_o}=\begin{bmatrix} M_{aa} & M_{ab} \\ M_{ab}^T & M_{bb} \end{bmatrix} \tag{51}$$

Since $\hat{\theta}_o$ is restricted minimum of $L(\theta)$ when $b=b^*$, it holds that

$$\left[\frac{\partial L}{\partial a}\right]_{\hat{\theta}_o} = 0 \tag{52}$$

Thus, using the matrix inversion theorem, equation (50) becomes

$$d(y) = \left[\frac{\partial L}{\partial b}\right]_{\hat{\theta}_o} \left[M_{bb} - M_{ab}^T M_{aa}^{-1} M_{ab}\right]^{-1} \left[\frac{\partial L}{\partial b}\right]_{\hat{\theta}_o}^T \tag{53}$$

If the gradient and the Hessian of the loss function $L(\theta)$ are calculated at the restricted minimum $\hat{\theta}_o = [\hat{a}^T \; b^{*T}]^T$ the value of the statistic $d(y)$ is given by eqn.(53). The statistic $d(y)$ can then be compared with $sk(1)$ and the null hypothesis is accepted if $d(y)<sk(1)$. The value of $k(1)$ can be taken, as discussed above, equal to 4. If the null hypothesis is rejected, i.e., the extended model is better than the original one, the estimate $\hat{\theta}_1$ given in eqn.(49) and the value of $L(\hat{\theta}_1)$ given by eqn.(50) can be very far from the true ones. The estimation then of $\hat{\theta}_1$ has to be made using the original data y. This method can thus be used only for validation and not as a quick way of extending an already estimated model. It gives however a good starting point for the minimization of the loss function $L(\theta)$ of the extended model using the original data y. In the case of ordinary least squares where the loss function $L(\theta)$ is exactly quadratic, equation (49) gives the exact estimate of the extended model. The calculation of $d(y)$ using eqn.(53) does not actually require the inverse of the Hessian since the powerful and accurate square root methods can be employed. The validation test that has been described can be proved to have maximum power for long sets of data |Bohlin (1978)|.

5.2 Correlation based validation

The traditional validation method used in linear systems identification is to test that the autocorrelation function of the residuals for the created model $\varepsilon(t)$ is an impulse and that the cross-correlation function between the residuals $\varepsilon(t)$ and the input $u(t)$ is zero. The formulation of this method to a properly defined statistical test and the generalization to the non-linear case is considered next. Several shortcomings of the traditional formulation of the correlation tests for both linear and non-linear systems are discussed. A comparison with the maximum power parametric validation tests described previously is also inclused. The methods presented here are alternatives to the tests given in Billings and Voon (1983,1986b)

It is well known that the innovations $e(t)$ must satisfy $E[e(t)|y^{t-1},u^t] = 0$ where for an r-input m-output system

$$\begin{aligned} u(t) &= [u_1(t), u_2(t), \ldots u_r(t)]^T \\ y(t) &= [y_1(t), y_2(t) \ldots y_m(t)]^T \\ \text{and} \quad u^t &= [(u(t))^T, (u(t-1))^T, \ldots (u(1))^T]^T \\ y^t &= [(y(t))^T, (y(t-1))^T, \ldots (y(1))^T]^T \end{aligned} \tag{54}$$

or equivalently

$$E[e(t) \mid e^{t-1}, u^t] = 0 \tag{55}$$

Let the vector x^t contain all the outputs, innovations and inputs up to the time t, i.e.

$$x^t = \begin{bmatrix} y^{t-1} \\ e^{t-1} \\ u^t \end{bmatrix} \tag{56}$$

Conditions (54) and (55) can be collectively written as

$$E[e(t) \mid x^t] = 0 \tag{57}$$

Conditional expected values are difficult to test directly. Alternative conditions however, derived from eqn(57), can easily be tested. They are basically conditions of zero cross-correlation between innovations and functions of previous data. The data up to time t are the vectors y^{t-1} and u^t since the vector e^{t-1} is actually a function of the data y^{t-1} and u^t.

Let the covariance matrix R of the innovations be factorized as

$$E[e(t)e^T(t) \mid x^t] = S^T S \tag{58}$$

where S is a square root of the covariance matrix of the innovations R. The innovation process e(t) can be normalized to have unit covariance matrix. Let

$$w(t) = S^{-T} e(t) \tag{59}$$

where the stochastic process w(t) has unit covariance matrix. Now let some matrix Z(t) have every element dependent only on the vector x^t, i.e.

$$Z(t) = Z(x^t) \tag{60}$$

The matrix Z(t) is of dimension sxm where m is the dimension of the output vector y(t) or equivalently the dimension of the innovation vector e(t). The matrix Z(t) is assumed to satisfy the law of large numbers for the time average

$$\frac{1}{N} \sum_{t=1}^{N} Z(t)Z^T(t) = \Gamma^T \Gamma \tag{61}$$

Clearly this is not a restrictive assumption. The following random vector can be defined

$$\mu = \frac{1}{N} \sum_{t=1}^{N} Z(t)w(t) \tag{62}$$

It is also assumed that the central limit theorem holds for eqn(62). This assumption is also non-restrictive. The first result given in [Bohlin (1978)] is that the random variable w(t) is asymptotically normal with zero mean and covariance matrix equal to $\frac{1}{N} \Gamma^T \Gamma$. In fact from eqn(62)

$$
\begin{aligned}
E[\mu] &= \frac{1}{N}\sum_{t=1}^{N} E[Z(t)w(t)] \\
&= \frac{1}{N}\sum_{t=1}^{N} E[Z(x^t)W(t)] \\
&= \frac{1}{N}\sum_{t=1}^{N} E[Z(x^t)\, E[w(t)\ |x^t]] \\
&= 0
\end{aligned}
\tag{63}
$$

since $E[w(t)\ |x^t]=0$ and where the property of the double expected value |Papoulis (1965)| was used in eqn.(63). Also

$$
\begin{aligned}
E[\mu\mu^\tau] &= E\left[\frac{1}{N}\sum_{t=1}^{N} Z(t)w(t)\ \frac{1}{N}\sum_{k=1}^{N} w^\tau(s)Z^\tau(k)\right] \\
&= N^{-2}\sum_{t=1}^{N}\sum_{k=1}^{N} E[Z(t)w(t)w^\tau(k)Z^\tau(k)] \\
&= N^{-2}\sum_{t=1}^{N}\sum_{k=1}^{N} E[E[S(t)w(t)w^\tau(k)Z^\tau(k)\ |x^{\max(t,k)}]] \\
&= N^{-2}\sum_{t=1}^{N}\sum_{k=1}^{N} E[Z(t)\ E[w(t)w^\tau(k)\ |x^{\max(t,k)}]Z^\tau(k)] \\
&= N^{-2}\sum_{t=1}^{N} E[Z(t)\ Z^\tau(t)]
\end{aligned}
\tag{64}
$$

since $E[w(t)w^\tau(k)\ |x^{\max(t,k)}]=0$ for $t\neq k$. Thus from eqn.(64)

$$
E[\mu\mu^\tau] = \frac{1}{N}\,\Gamma^\tau\Gamma \tag{65}
$$

The random variable μ is thus asymptotically normal with zero mean and covariance matrix given by eqn.(65). The random vector μ can be normalized and the random vector

$$
\rho = \sqrt{N}\ \ \Gamma^{-\tau}\mu \tag{66}
$$

has asymptotically zero mean and unit covariance. The variable $d=\rho^\tau\rho$ is then asymptotically chi-square distributed with s degrees of freedom where s is a dimension of the vector μ. A validation test can then be based on the statistic d where

$$
d=\rho^\tau\rho=N\ \mu^\tau\ (\Gamma^\tau\Gamma)^{-1}\mu \tag{67}
$$

The null hypothesis is that the given input-output data are generated by a particular model and the alternative hypothesis is that the data are not generated by that model. The residuals for the particular model are calculated and the value of the statistic d in eqn.(67) is found. If the value of the statistic d is outside the acceptance region for a given level of significance α, the model is not an acceptable one. The acceptance region for a given level of significance α is the region

$$d<k_{\alpha}(s) \tag{68}$$

where $k_{\alpha}(s)$ is the critical value of the chi-squared distribution with s degrees of freedom.

The problem that still has to be investigated is how to choose the matrix $Z(x^t)$. First a comparison between the test eqn(68) and the traditional correlation tests for the linear systems is presented. Such a comparison will also give some clues about the proper choice of the matrix $Z(x^t)$. In order to make the arguments easier to understand, the single-input single-output case only is discussed, but everything is also valid for the multivariable case.

In the single-output case, the matrix Z(t) is a row vector and the realization of the statistic $d=\rho^{\tau}\rho$ for a particular model can be written as

$$d=\rho^{\tau}\rho=N(1/\sum_{t=1}^{N}\varepsilon^2(t))(\sum_{t=1}^{N}Z(t)\varepsilon(t))^{\tau}[\sum_{t=1}^{N}Z(t)Z^{\tau}(t)]^{-1}(\sum_{t=1}^{N}Z(t)\varepsilon(t)) \tag{69}$$

where $\varepsilon(t)$ represent the residuals.

For instance if the matrix Z(t) is

$$Z(t)=\begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix} \tag{70}$$

then

$$d=N(1/\sum\varepsilon^2(t))\begin{bmatrix}\sum z_1(t)\varepsilon(t) & \sum z_2(t)\varepsilon(t)\end{bmatrix}\begin{bmatrix}\sum z_1(t)z_1(t) & \sum z_1(t)z_2(t) \\ \sum z_1(t)z_2(t) & \sum z_2(t)z_2(t)\end{bmatrix}^{-1}\begin{bmatrix}\sum z_1(t)\varepsilon(t) \\ \sum z_2(t)\varepsilon(t)\end{bmatrix} \tag{71}$$

where every sum is from t=1 to t=N. The traditional validation method in linear system identification consists of the task of calculating the autocorrelation of the residuals and the cross-correlation between the residuals and the inputs and checking if they fall within a confidence interval, usually $\pm 2/\sqrt{N}$. It is evident then that the traditional method corresponds to a choice of the matrix $Z(x^t)$ equal to

$$Z(x^t)=[\varepsilon(t-1),\ \varepsilon(t-2),\ \ldots\ ,\ \varepsilon(t-t_d)]^{\tau} \tag{72}$$

and

$$Z(x^t)=[u(t),\ u(t-1),\ \ldots\ ,\ u(t-t_d)]^{\tau} \tag{73}$$

where t_d is the maximum delay considered in the correlations. The validation test described before is thus different from the traditional method. The traditional method can be analysed in order to investigate if it is actually a properly defined test or just provides an indication that the model is not acceptable.

Let the variables w(t) be as before, the normalized residuals

$$w(t)=\frac{\varepsilon(t)}{[\frac{1}{N}\sum_{t=1}^{N}\varepsilon^2(t)]^{1/2}} \tag{74}$$

and the random vector μ then is

$$\mu_1 = \frac{1}{N}\left[\sum_{t=1}^{N}\varepsilon(t-1)w(t),\ \ldots,\ \sum_{t=1}^{N}\varepsilon(t-t_d)w(t)\right]^\tau \tag{75}$$

for $Z(x^t)$ in eqn(72) and

$$\mu_2 = \frac{1}{N}\left[\sum_{t=1}^{N}u(t)w(t),\ldots,\ \sum_{t=1}^{N}u(t-t_d)w(t)\right]^\tau \tag{76}$$

for $Z(x^t)$ in eqn(73). Using eqns(65) and (61), the vector μ_1 has a diagonal covariance matrix because the innovations are uncorrelated with previous innovations. If the model is correct, it then asymptotically holds

$$\Gamma^\tau\Gamma = \left(\frac{1}{N}\sum_{t=1}^{N}\varepsilon^2(t)\right) I \tag{77}$$

where I is the unit matrix. The vector μ_1 is thus orthogonal and the normalization is simply done by dividing by the standard deviation. The vector

$$\rho_1 = \sqrt{N}\,\frac{I}{\sum_{t=1}^{N}\varepsilon^2(t)}\left[\sum_{t=1}^{N}\varepsilon(t-1)\varepsilon(t)\ \sum_{t=1}^{N}\varepsilon(t-2)\varepsilon(t)\ \ldots\ \sum_{t=1}^{N}\varepsilon(t-t_d)\varepsilon(t)\right]^\tau \tag{78}$$

is thus normally distributed with unit covariance matrix and consequently every element of the vector ρ_1 is independent from the others and with unit covariance. The traditional validation method consists of testing that every element of the vector ρ_1 is, for some level of significance α, in the region $\pm k_\alpha$ where k_α is the critical value of the normal distribution with level of significance α. If α=0.0456, then k_α=2 and the traditional test requires

$$-\frac{2}{\sqrt{N}} < \frac{\sum_{t=1}^{N}\varepsilon(t-k)\varepsilon(t)}{\sum_{t=1}^{N}\varepsilon^2(t)} < \frac{2}{\sqrt{N}} \qquad \text{for } k = 1,2,\ \ldots\ ,t_d \tag{79}$$

This test is an extremely severe one. If for instance the maximum delay is t_d=50, the probability of accepting the true model is only $(1-\alpha)^{50}$=0.0969. This test is thus so severe that it has only 9.7% probability of accepting the true model. In practice it has been known that a few correlations of the autocorrelation function of the residuals can be expected to exceed the confidence limits even if the model is the correct one. All the careful authors have never claimed that the traditional test is a proper statistical test but that, as it is put in |Astrom (1980)|, the calculation of the correlation function is revealing. Some practical rules have been devised to overcome this difficulty. A rule of thumb has sometimes been used which says that in practice if the first five correlations lay within the confidence limit, the model is acceptable. Let us investigate how good such a rule is as a statistical test. The probability of accepting the true model for α=0.0456 is now $(1-\alpha)^5$=0.792 and thus such a test is still extremely severe since the probability of rejecting the true model is almost 21%, an unacceptably large probability. In practice experienced engineers

can detect a non acceptable correlation function from too many correlations outside the confidence limit. Such experience is not actually needed if the proper chi-square test is employed. This test required that

$$\rho_1^T \rho_1 < k_\alpha(s) \tag{80}$$

where $k_\alpha(s)$ is the critical value of the chi-square distribution with $s=t_d$ degrees of freedom. The chi-square test has been introduced in |Bohlin (1971)| and made more recently in |Bohlin (1978)|.

The traditional cross-correlation test between residuals and inputs has exactly the same problems. Every element of the vector μ_2 in eqn.(76) is individually normally distributed and the traditional test for $\alpha=0.0456$ is

$$-\frac{2}{\sqrt{N}} < \frac{\sum_{t=1}^{N} u(t-k)\varepsilon(t)}{\left(\sum_{t=1}^{N} \varepsilon^2(t)\right)^{1/2}\left(\sum_{t=1}^{N} u^2(t)\right)^{1/2}} < \frac{2}{\sqrt{N}} \tag{81}$$

The covariance matrix of μ_2 is not diagonal unless the input signal is a white noise signal. To calculate the statistic d of the chi-square test, the covariance matrix of the vector Z(t) must be calculated first and using eqns(61) and (67) the test variable $\rho_2^T\rho_2$ can be found. The inversion of the matrix $\Gamma^T\Gamma$ can easily be done if it is decomposed first by the Cholesky factorization so that the square root matrix Γ is an upper triangular matrix. |Bierman (1977)|.

A comparison of the correlation chi-square tests where the matrix $Z(x^t)$ is chosen as in eqns(72) and (73) with the maximum power parametric tests, was first done in |Bohlin (1978)|. It was found that the correlation tests can in the case of long time constants have a much lower power compared with the parametric tests. In more ordinary cases however, the correlation tests are not low powered tests. Traditionally the vector $Z(x^t)$ may consist of residuals or inputs only. The vector x^t however also contains the outputs y^{t-1}. If the tests with $Z(x^t)$ that consist of residuals or inputs have small power, the tests with $Z(x^t)$ that consist of outputs may have a considerably higher power. Define the vector

$$Z(x^t) = [y(t-1),\ y(t-2),\ \ldots,\ y(t-t_d)]^T \tag{82}$$

For linear systems it was found by simulation that correlation tests with $Z(x^t)$ as in eqns(72), (73) or (82) have not dramatically lower power than the maximum power parametric tests. In some cases however, the inclusion of the test based on the vector eqn(82) is important in order to make the correlation tests comparable with the optimum parametric validation tests.

It can be argued that since the conditions eqn(54) and (55) are equivalent, the vector x^t need only consist of either residuals and inputs or outputs and inputs. This is correct but the tests based on an x^t that consists of residuals and inputs may have smaller power compared with the tests based on an x^t that consists of outputs. Thus a $Z(x^t)$ as in eqn(82) should also be considered.

In the case of non-linear systems, the validation based on correlation tests leads to the problem of the choice of the vector $Z(x^t)$. It is no longer sufficient to choose the elements of the vector $Z(x^t)$ as past residuals, inputs or outputs. The elements of the vector $Z(x^t)$ must be generalized to include non-linear functions of the elements of the vector x^t, and one particular solution to this is given in Billings and Voon |1986b|. The type of non-linear functions

that can be used is arbitrary but a very satisfactory choice is the monomials of the elements of the vector x^t. Only monomials will be used here as elements of the vector $Z(x^t)$ since they are easy to compute and they are found to perform very well. In practice the most obvious generalization of the traditional correlation validation method for linear systems where $Z(x^t)$ is as in eqn(72) and (73), are the correlation tests $Z(x^t)$ that have as elements powers of past residuals or inputs, i.e.

$$Z(x^t)= [(\varepsilon(t-1))^i, (\varepsilon(t-2))^i, \ldots , (\varepsilon(t-t_d))^i]^\tau \tag{83}$$

or

$$Z(x^t)= [(u(t))^i, (u(t-1))^i, \ldots , (u(t-t_d))^i]^\tau \tag{84}$$

where i=1,2,...

However if the validation of a model is based only on correlation tests with $Z(x^t)$ in the form (83) or (84), it is possible that the validation tests will have insignificant power. Such a situation is particularly noticeable in the case of non-linear odd systems (systems that behave in a symmetrical way for positive and negative inputs) excited by a white noise input signal. The problem can be solved by creating a more general vector $Z(x^t)$, as for instance

$$Z(x^t)=[\varepsilon^i(t-1)\varepsilon^j(t-2),\varepsilon^i(t-2)\varepsilon^j(t-3), \ldots ,\varepsilon^i(t-t_d)\varepsilon^j(t-t_d-1)]^\tau \tag{85}$$

or

$$Z(x^t)=[u^i(t)u^j(t-1), u^i(t-1)u^j(t-2),\ldots, u^i(t-t_d)u^j(t-t_d-1)]^\tau \tag{86}$$

where i,j=1,2,...

Even more complicated monomials can be used as for instance $u^i(t)u^j(t-1)u^k(t-2)$. Such choices of $Z(x^t)$ can detect for instance that a linear model is wrong when the non-linear system is odd and the input is a white noise signal. Another way to generalize the vector $Z(x^t)$ in eqn(83) and (84) is to use it as it was proposed for the linear case, the outputs of the system as well as the inputs and the residuals. Correlation tests with vectors $Z(x^t)$ that include the outputs were found by simulation to have, quite often, the highest power for non-linear models. Of course they are always inferior to the maximum powered parametric tests but they prove to be in a position to discriminate between the correct and the wrong models in most cases. The vector $Z(x^t)$ can thus initially be chosen as

$$Z(x^t) = [y^i(t-1), y^i(t-2), \ldots , y^i(t-t_d)]^\tau \tag{87}$$

or more generally as

$$Z(x^t) = [m(t), m(t-1), \ldots ,m(t-t_d)]^\tau \tag{88}$$

where m(t) is a monomial of elements of the vector x^t, for instance it could be

$$m(t)=y^3(t-1)u^2(t-1)u(t-2)\varepsilon(t-1) \tag{89}$$

Several types of monomials m(t) must be tried before it can be said that a model is properly validated by the correlation validation tests.

In practice the correlation based tests derived in Billings and Voon |1983, 1986b| are augmented with the tests described above to provide the user with a powerful combination of techniques which indicate which terms have been omitted from the model |Billings and Voon 1986a, Billings and Fadzil 1985|.

6. Simulation Results

Let the system to be identified have input u(t) and output y(t) where

$$y(t)=0.8y(t-1)+0.2u(t-1)-0.8e(t-1)+0.1y^3(t-1)-0.05y(t-1)u^2(t-1)-$$
$$0.2y(t-1)u(t-1)e(t-1)+e(t) \qquad (90)$$

and where e(t) is a Gaussian white sequence of variance equal to 0.05. The system will be called S_1. An input-output data sequence of 500 points was generated. The input signal was an independent sequence of uniform distribution between -3 and 3. The inputs and outputs of the system for the first 100 points are illustrated in figure 1.

A prediction error method assuming the correct model structure was used to identify the system S_1 from the input-output data record. Newton's method with line search was used to minimize the loss function. The first input and output data points were used as the initial input and output while the initial residual was taken equal to zero. The rest of the data points were used for the creation of the loss function. The initial value of the parameter vector was the ordinary least squares solution. In five iterations the minimum was found. The inverse of the Hessian of the loss function at the minimum is the estimate of the covariance matrix of the parameter vector estimate. The diagonal elements of this matrix are the variances of the individual parameters. The estimates of the parameters and their standard deviations are

1	y(t-1)**1 =	0.7997E+00(±0.4725E-02)
2	u(t-1)**1 =	0.1988E+00(±0.9044E-03)
3	e(t-1)**1 =	-0.8017E+00(±0.2712E-01)
4	y(t-1)**3 =	0.1091E+00(±0.9563E-02)
5	y(t-1)**1*u(t-1)**2 =	-0.5115E-01(±0.1250E-02)
6	y(t-1)**1+u(t-1)**1*e(t-1)**1 =	-0.1928E+00(±0.3528E-01)

where the coefficients of the terms on the left-hand side have estimated values and variances given at the right-hand side. The method produces unbiased estimates since the true values of the parameters are at least within 2 standard deviations of the estimated values. The response of the estimated model and the residuals for the first 100 points are illustrated in figure 1.

The same input-output data were used to estimate the parameters of a linear model with only three parameters.

The loss function was minimized as before and the estimates of the parameters and their standard deviations are

1	y(t-1)**1 =	0.7058E+00 (±0.6280E-02)
2	u(t-1)**1 =	0.1964E+00 (±0.2227E-02)
3	e(t-1)**1 =	- 0.3617E+00 (±0.4240E-01)

The input, the output of the system, the output of the estimated model and the residuals for the first 100 points are illustrated in figure 2. It is obvious that there is a large bias of the estimated values of the parameters, since the true values of the parameters are not within a few standard deviations

of the estimated values of the parameters. This bias is actually beneficial if a linear model is to be used, otherwise the model would behave extremely badly for large values of inputs. The fact that the model used was incorrect must however be detected by the model validation methods. The non-parametric correlation validation method was employed first to detect whether the model was satisfactory.

The traditional correlations between the residuals and past residuals and inputs are given in figure 3. The confidence interval is the 95% confidence interval. The chi-square correlation tests described in section 5 for the same correlations are given in figure 4. It can be seen from figure 3 that some correlations are outside the confidence interval. The chi-square correlation tests however confirm that these values of the correlations are acceptable and thus correlations of the residuals with past residuals and inputs cannot detect that the linear model is biased. This was to be expected since the tested model is the best possible linear one.

Correlations of the residuals with powers of past residuals and inputs are attempted next. Such correlations are the direct generalization of the correlations used for linear systems. The chi-square tests for three such correlations are given in figure 5. The power of these tests is small and they cannot detect that the linear model is a wrong model.

More general correlations can be considered however and the chi-square tests based on them are given in figures 6 and 7. The first correlation test in figure 6 shows that the correlations of the residuals with monomials which include several delayed inputs can detect that the linear model is inadequate. The second test in figure 6 shows that the correlations of the residuals with monomials that include the output of the system provide, in this case, the correlation test with the maximum power. The rest of the tests in figures 6 and 7 demonstrate that several other correlation tests are not in a position to detect that the linear model is incorrect.

The parametric validation tests are however the most powerful tests. The value of the statistic d(y) in eqn(53) for the inclusion of the terms $y^3(t-1)$, $y(t-1)u^2(t-1)$ or $y(t-1)u(t-1)e(t-1)$ are

inclusion of the term	value of the statistic d(y)
$y^3(t-1)$	7.8998
$y(t-1)u^2(t-1)$	450.2393
$y(t-1)u(t-1)e(t-1)$	0.2760

The parametric validation tests confirm that the model expanded with the first or the second of the above terms is better than the original linear one since the value of the statistic d(y) is greater than 4. If the first two of the above terms are included in the model and the statistic d(y) is computed again for the inclusion of the third term, the value of d(y) is found to be equal to 41.8130, which correctly indicates that the third term must be included in the model.

Another model used to identify the system S_1 was an over-parametrized model with 13 parameters that corresponds to a third order polynomial expansion of the model

$$y(t)=q\left[y(t-1),u(t-1),\ e(t-1)\right]+e(t) \qquad (91)$$

with only the odd order terms present. The input-output data sequence was the same as in the previous cases. The loss function was minimized as before and the estimates of the parameters and their standard deviations are

1	y(t-1)**1 =	0.7996E+00	(±0.5549E-02)
2	u(t-1)**1 =	0.2006E+00	(±0.2271E-02)
3	e(t-1)**1 =	-0.7220E+00	(±0.6938E-01)
4	y(t-1)**3 =	0.1174E+00	(±0.1129E-01)
5	y(t-1)**2*u(t-1)**1 =	-0.5559E-02	(±0.3859E-02)
6	y(t-1)**2*e(t-1)**1 =	-0.1774E+00	(±0.1318E+00)
7	y(t-1)**1*u(t-1)**2 =	-0.5201E-01	(±0.1330E-02)
8	y(t-1)**1*u(t-1)**1*e(t-1)**1 =	-0.2045E+00	(±0.4387E-01)
9	y(t-1)**1*e(t-1)**2 =	-0.1232E+01	(±0.9800E+00)
10	u(t-1)**3 =	0.4212E-05	(±0.3273E-03)
11	u(t-1)**2*e(t-1)**1 =	-0.5457E-02	(±0.1055E-01)
12	u(t-1)**1*e(t-1)**2 =	-0.9669E-01	(±0.2446E+00)
13	e(t-1)**3 =	-0.2969E+01	(±0.4989E+01)

The input, the output of the system, the output of the estimated model and the residuals for the first 100 points are given in figure 8. There is no bias of the estimated value of the parameters since they are all within a few standard deviations from their correct value. Such a model can however be greatly reduced since the system S_1 contains only 6 out of the 13 terms of the model. The Stepwise Backward Elimination (SBE) of parameters was first used to reduce the model. Akaike's information criterion (AIC) and the criterion C eqn(36) with k(1)=4 for the reduction of every term is

Total Number of eliminated parameters	No of eliminated parameter	AIC of reduced model - AIC of full model	C of reduced model - C of full model	Standard Deviation of the residuals
1	10	-0.2000E+01	-0.4000E+01	0.4874E-01
2	12	-0.3831E+01	-0.7831E+01	0.4875E-01
3	11	-0.5602E+01	-0.1160E+02	0.4876E-01
4	13	-0.7237E+01	-0.1524E-02	0.4878E-01
5	9	-0.7537E+01	-0.1754E+02	0.4886E-01
6	5	-0.7833E+01	-0.1983E+02	0.4895E-01
7	6	-0.5949E+01	-0.1995E+02	0.4914E-01
8	8	0.2271E+02	0.6708E+01	0.5067E-01

The minima of the two criteria are underlined. The first observation is that Akaike's criterion does not reduce the model to the correct one. It keeps an extra term, the term $y^2(t-1)e(t-1)$ (No 6). It is known that the AIC criterion tends to over-estimate the number of necessary parameters. The criterion C with k(1)=4 gives the correct model. The Stepwise Forward Inclusion (SFI) of parameters is also used to reduce the model. Akaike's information criterion and the criterion C with k(1)=4 is given for the inclusion of every term

Total Number of included parameters	No of included parameter	AIC of reduced model - AIC of full model	C of reduced model - C of full model	Standard Deviation of the residuals
1	1	0.5413E+05	0.5410E+05	very big
2	2	0.3460E+04	0.3438E+04	0.1597E+01
3	7	0.1266E+04	0.1246E+04	0.1769E+00
4	3	0.1614E+03	0.1434E+03	0.5835E-01
5	4	0.2270E+02	0.6702E+01	0.5067E-01
6	8	-0.5950E+01	-0.1995E+02	0.4914E-01
7	6	-0.7833E+01	-0.1983E+02	0.4895E-01
8	5	-0.7538E+01	-0.1754E+02	0.4886E-01
9	9	-0.7237E+01	-0.1524E+02	0.4878E-01

The SFI method gives exactly the same final model as the SBE method. Again Akaike's criterion over-estimates the number of the necessary parameters. The slight numerical difference in the value of the criteria that correspond to the same reduced models in the case of the SBE and SFI methods is caused by the completely different numerical methods employed in the two cases. The SBE method is faster because the decomposed Hessian of the loss function of every reduced model is used for the next elimination of a parameter but it is slightly inaccurate because of accumulation of numerical errors. The SFI method is slower because the Hessian of the loss function of every reduced model is calculated from the original undecomposed Hessian of the full model but numerically more accurate because every reduced model does not depend on the decomposed Hessian of the reduced model of the previous step.

7. Conclusions

The problems of model selection and model validation have been studied for both linear and nonlinear systems. The problem of selecting one out of two models was formulated as a hypothesis testing problem. If the maximum likelihood estimation method is used, the selection of the best model is provided by the likelihood ratio test and if the prediction error estimation method is used, the selection of the best model is provided by a test given the name log determinant ratio test. An efficient method of approximately calculating the two tests, using the Hessian of the loss function at the minimum has been presented. The traditionalF-test for single-input single-output systems was shown to be asymptotically equivalent to the log determinant ratio test. The selection of one model among many competing models was then considered. It was shown that the significance levels for the many possible pairwise

comparisons must be chosen in a restricted way if conflicts are to be avoided. The criterion which the finally selected model must minimize was thus derived. Akaike's AIC criterion corresponds to the criterion derived for a specific significance level. The relatively high significance level that corresponds to the AIC criterion provided an explanation for the tendency of this criterion to overestimate the number of required parameters. The other objective criterion, the FPE criterion, was shown to be asymptotically equivalent to the AIC subjective. The objective criteria AIC and FPE were proposed so that they could perform the model selection in a completely automatic way, independent of subjection choices of significance levels. However, it has been demonstrated that they actually correspond to a specific significance level and thus they just predetermine the significance level that should be used. The methods of stepwise forward inclusion of parameters and of stepwise backward elimination of parameters were described as methods of selecting a reduced model when the number of competing models is excessive.

Model validation methods have been studied and a parametric validation method was derived and its relation with the approximate calculation of the likelihood ratio or the log determinant ratio test was shown. The correlation validation methods were then studied. The chi-square tests were described and compared with the traditional linear correlation tests involving the autocorrelation function of the residuals and the cross-correlation function between residuals and inputs. An extension of the traditionally used correlations was also provided by considering correlations between residuals and past inputs. The correlation tests for linear systems were then modified to suit non-linear systems. It was shown that correlations of the residuals with non-linear functions of past outputs are very important in the validation of non-linear systems.

8. Acknowledgements

One of the authors (SAB) gratefully acknowledges financial support for the work presented above from the Science and Engineering Research Council .

9. References

Akaike, H.(1969): Fitting autoregressive models for prediction; Ann Inst. Statist. Math. 21, 243-247.

Akaike, H.(1974a): Stochastic theory of minimal realization; IEEE Trans. Aut. Contr. AC-19, 667-674.

Akaike, H.(1974b): A new look at statistical model identification; IEEE Trans. Auto. Contr. AC-19, 716-723.

Astrom, K.J., Eykhoff, P. (1971): System identification - a survey; Automatica, 7, 123-162.

Astrom, K.J.(1980): Maximum likelihood and prediction error methods; Automatica, 16, 551-574.

Bierman, G.J.(1977): Factorization methods for discrete sequential estimation; Academic Press, New York.

Billings S.A., Fadzil M.B. (1985): The practical identification of nonlinear systems: 7th IFAC Symp. Ident. & Syst. Par. Est, 155-160.

Billings S.A., Voon W.S.F. (1986a): A prediction error and stepwise regression algorithm for nonlinear systems: Int.J. Control (to appear).

Billings S.A. Voon W.S.F. (1986b): Correlation based model validity tests for nonlinear systems: Int.J.Control (to appear)

Billings S.A. Voon W.S.F. (1983): Structure detection and model validity tests in the identification of nonlinear systems: Proc. IEE Part D, 130, 193-199.

Bhansali R.J, Downham D.Y. (1977): Some properties of the order of an autoregressive model selected by a generalisation of the FPE criterion: Biometrica, 64, 547-551.

Bohlin T. (1971): On the problems of ambiguities in maximum likelihood identification: Automatica, 7, 199-210.

Bohlin T (1978): Maximum power validation of models without higher order fitting: Automatica, 14, 137-146.

Box G.E.P., Jenkins G.M. (1976): Time Series Analysis: Holden Day, USA.

Draper N.R. Smith H. (1981): Applied Regression Analysis, Wiley.

Goodwin G.C. Payne R.L. (1977): Dynamic System Identification, Experiment Design and Data Analysis: Academic Press.

Kasyap R.L, (1977): A Baysian comparison of the different classes of dynamic models using empirical data; IEEE Trans. Auto Contr, AC-22, 715-727.

Kasyap R.L. (1980): Inconsistency of the AIC rule for estimating the order of autoregressive models; IEEE Trans. Auto Contr, AC-25, 996-998.

Kendall M.G., Stuart A. (1967): The advanced theory of statistics, Vol 2, Griffin, London.

Leontaritis I.J., Billings S.A. (1985): Input output parametric models for nonlinear systems. Part I Deterministic nonlinear systems. Part II Stochastic nonlinear systems; Int. J. Control 41, 303-344.

Ljung L, Soderstrom T. (1983): Theory and practice of recursive identification, MIT Press.

Papoulis A. (1965). Probability, Random Variables and Stochastic Processes; McGraw-Hill.

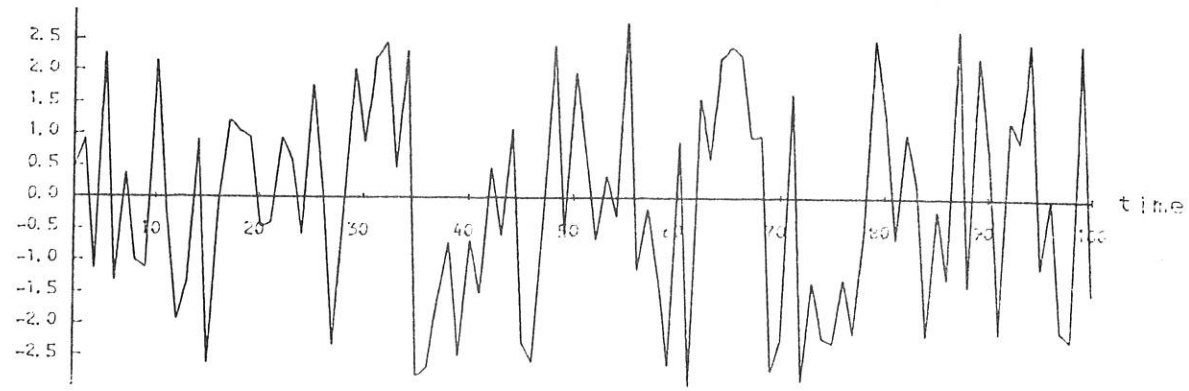
Priestley M.B., (1981): Spectral Analysis and Time Series; Academic Press.

Rissanen J, (1979): Shortest data description and consistency of order estimates of an ARMA Process; in A. Benousson and J.L.Lions (Eds) Int. Symp. on Systems Optimization and Analysis; Springer Verlag.

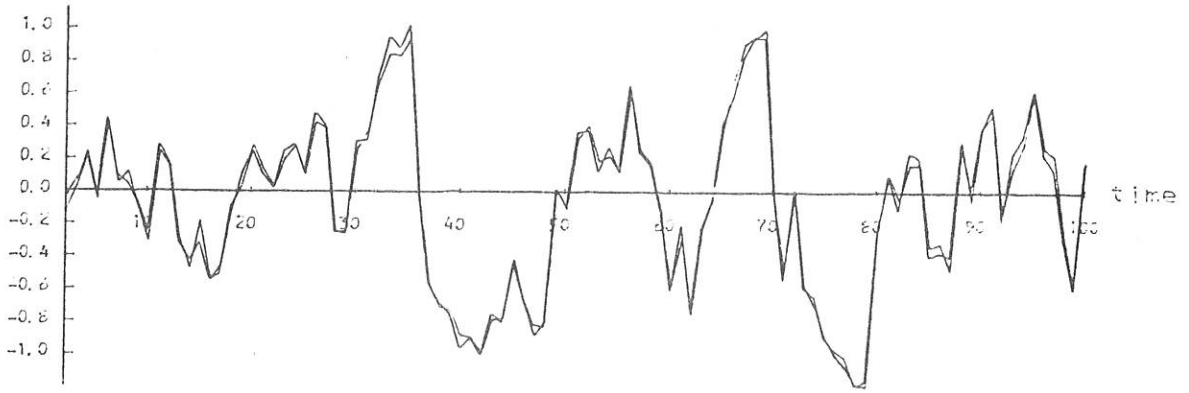
Schwartz G. (1978): Estimating the dimension of a model: Ann Statist, 6, 461-464

Schibata R, (1976): Selection of the order of an autoregression model by Akaike's information criterion; Biometrica, 63, 117-126.

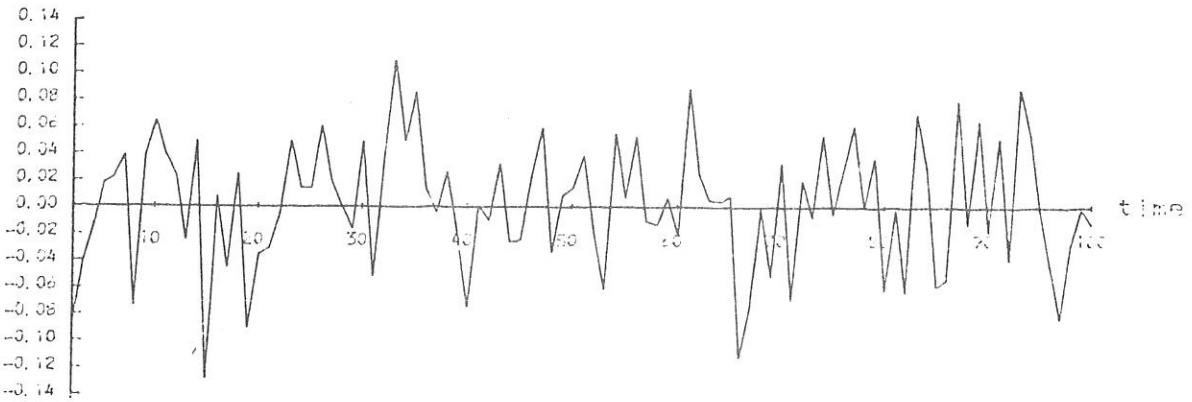
Soderstrom T, (1977): On model structure testing in system identification; Int. J. Control, 26, 1-18.



Input

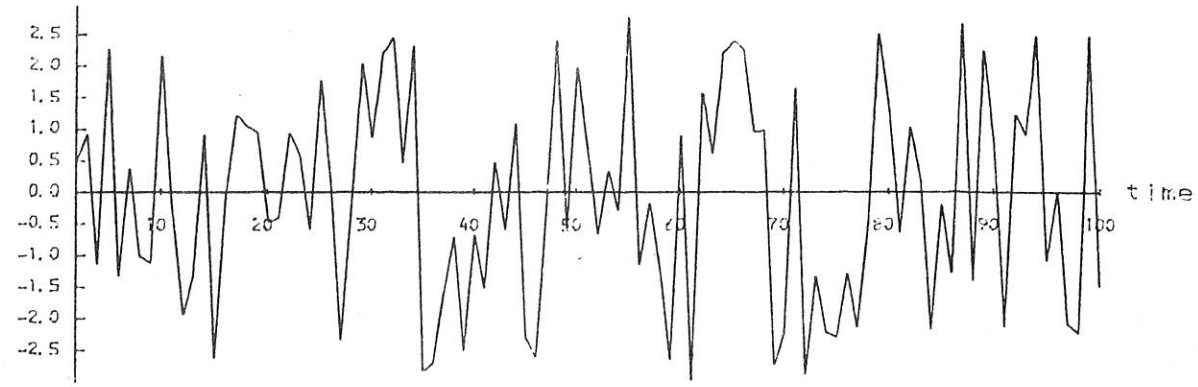


——— Output of the system
——— Output of the model

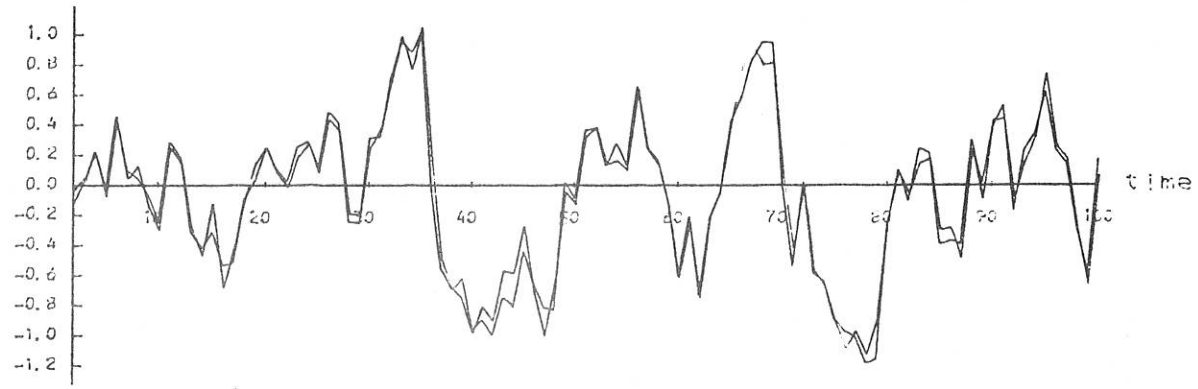


Residuals

Figure 1. System S_i - with estimated non linear model.



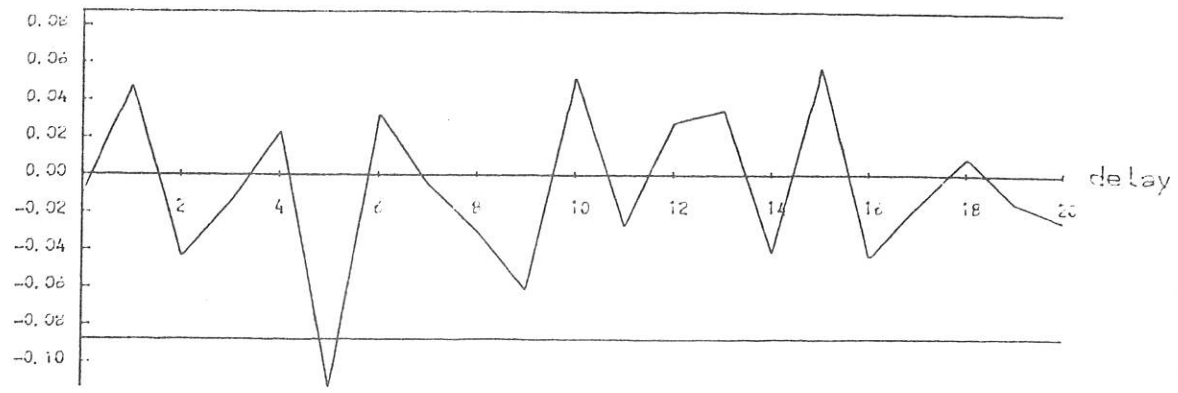
Input



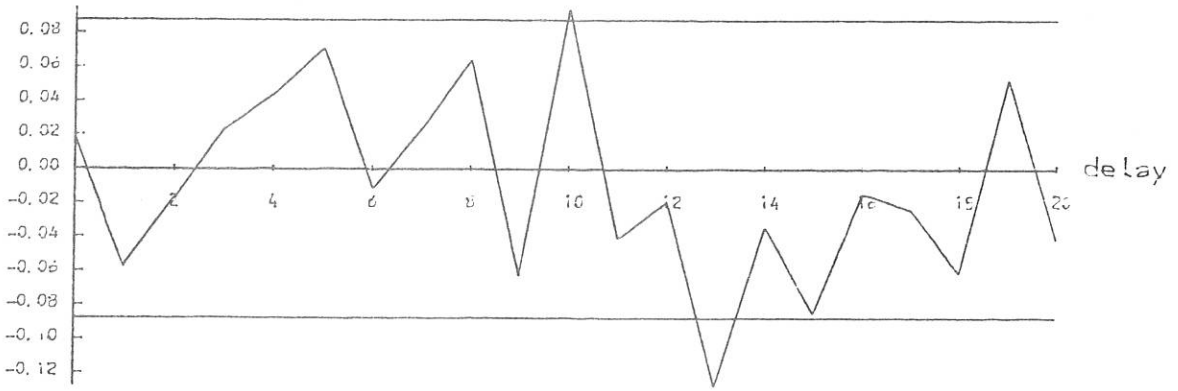
—— Output of the system
—— Output of the model

Residuals

Figure 2. System S_1 - with estimated linear model.



Correlation of e(t) with e(t-1-delay)



Correlation of e(t) with u(t-1-delay)

Figure 3. Traditional linear covariance tests.

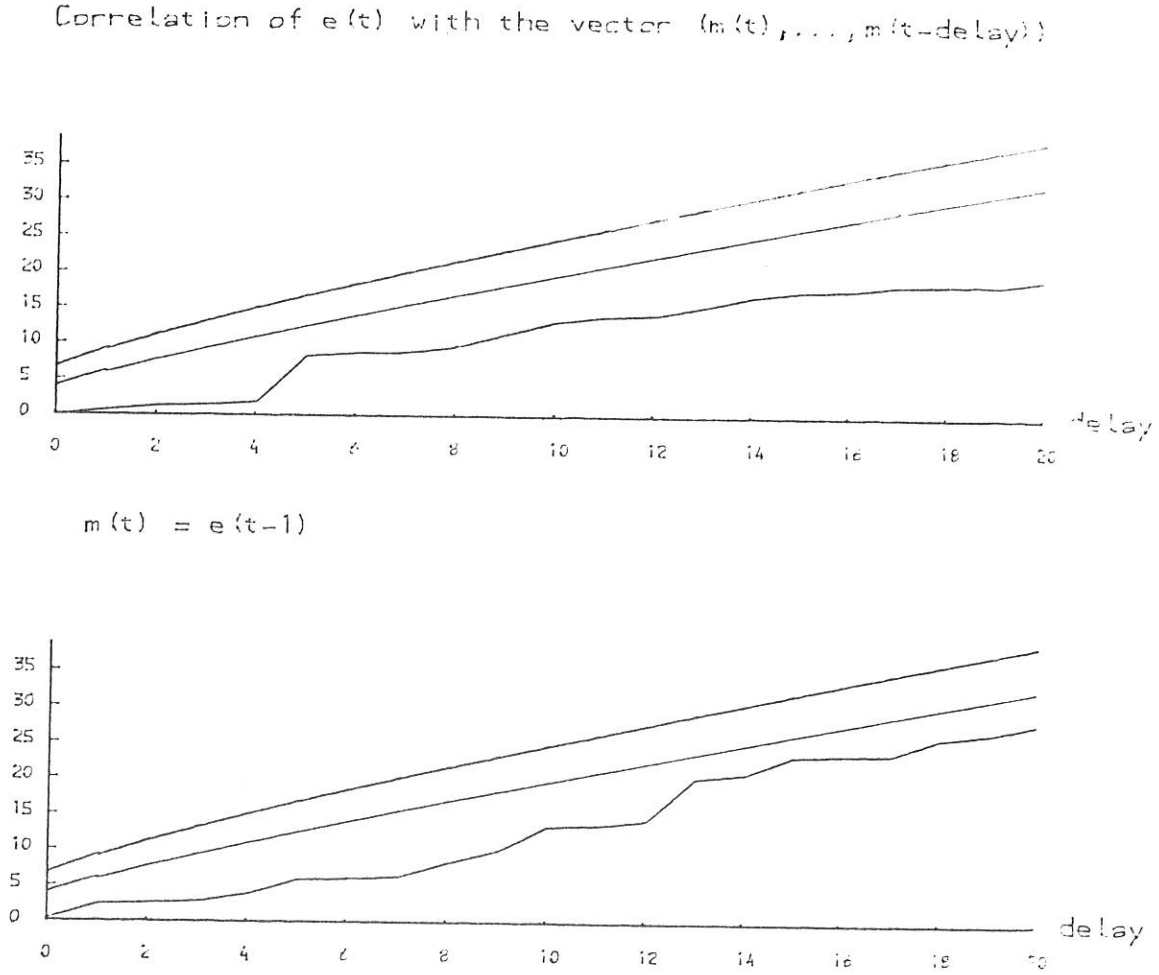


Figure 4. Chi-square correlation tests.

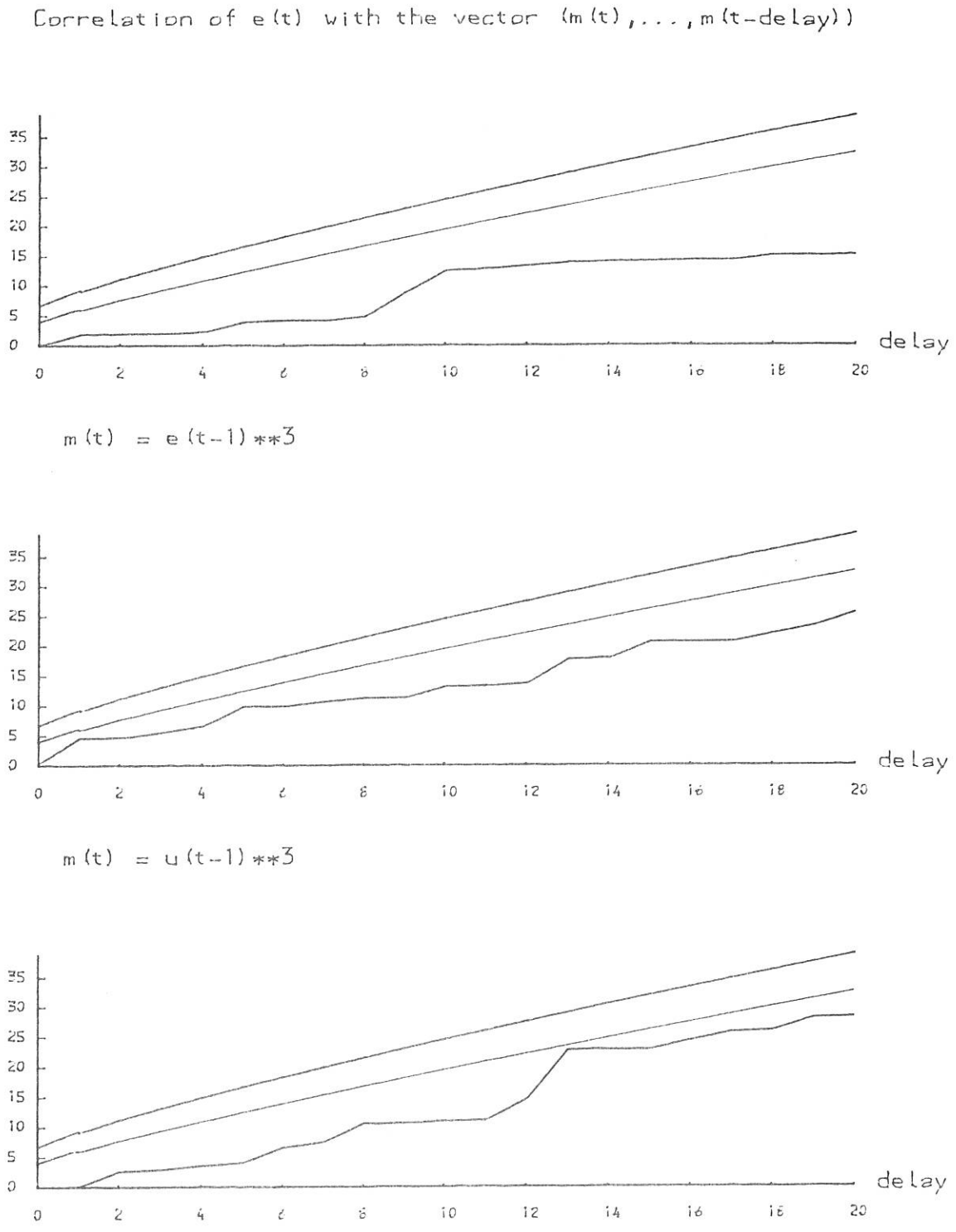
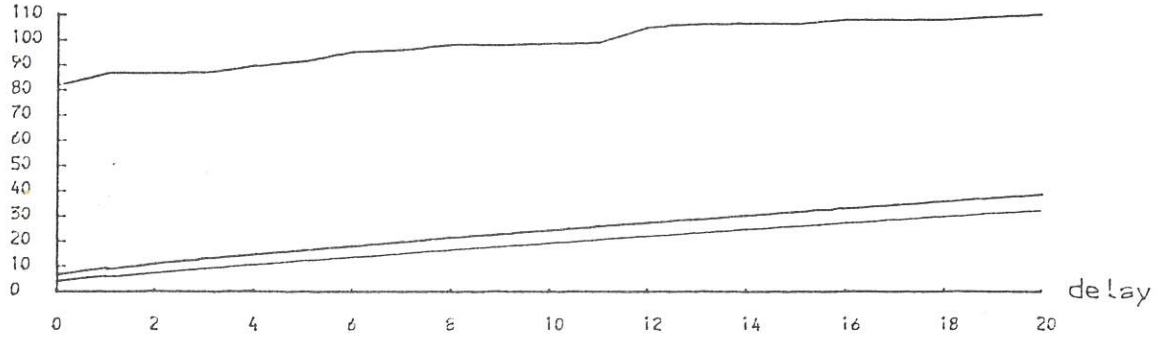
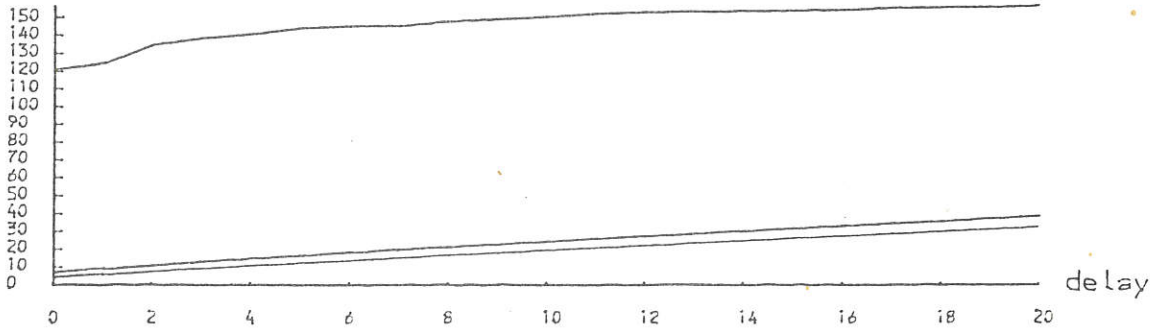


Figure 5. Chi-square Correlation Tests.

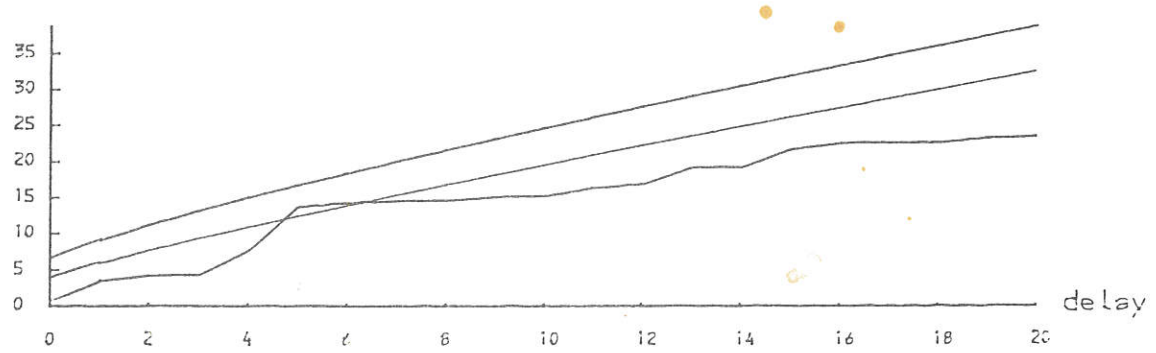
Correlation of e(t) with the vector (m(t),...,m(t-delay))



m(t) = u(t-1)**2*u(t-2)



m(t) = y(t-1)*u(t-1)**2



m(t) = y(t-1)**2*u(t-1)

——— 95% confidence limit ——— 99% confidence limit

Figure 6. Chi-Square Correlation Tests.

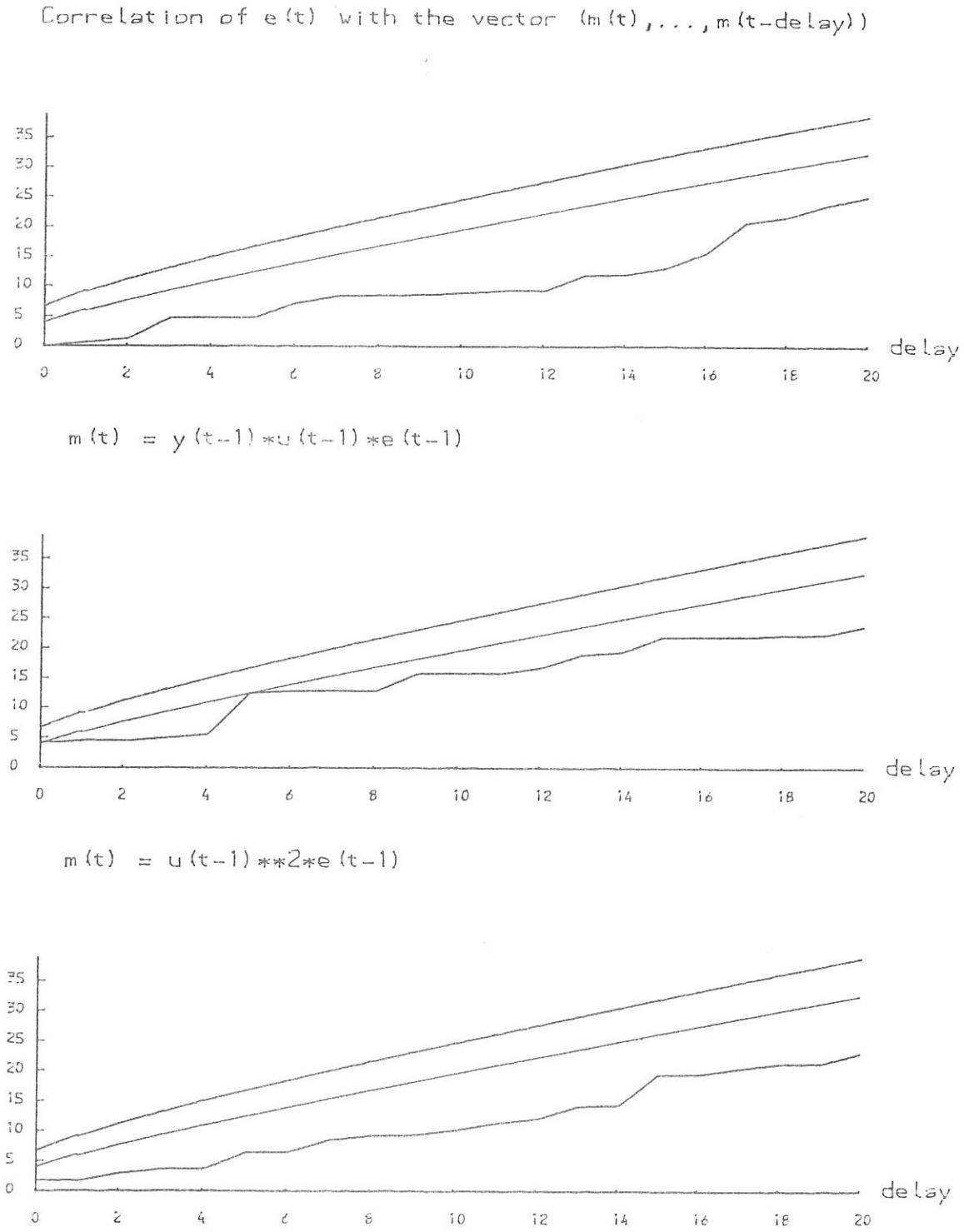


Figure 7. Chi-Square Correlation Tests.