This is a repository copy of *Identification of Nonlinear Rational Systems Using A Prediction-Error Estimation Algorithm*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/77049/

Monograph:
Billings, S.A. and Chen, S.I (1987) Identification of Nonlinear Rational Systems Using A Prediction-Error Estimation Algorithm. Research Report. Acse Report 317 . Dept of Automatic Control and System Engineering. University of Sheffield

IDENTIFICATION OF NONLINEAR RATIONAL SYSTEMS USING

A PREDICTION-ERROR ESTIMATION ALGORITHM

S.A. Billings and S. Chen

Department of Control Engineering

University of Sheffield

Mappin Street

Sheffield S1 3JD

U.K.

IDENTIFICATION OF NONLINEAR RATIONAL SYSTEMS USING A PREDICTION-ERROR ESTIMATION ALGORITHM

Abstract

Identification of discrete-time nonlinear stochastic systems which can be represented by a rational input-output model is considered. A prediction-error parameter estimation algorithm is developed and a criterion is derived using results from the theory of hypothesis testing to determine the correct model structure. The identification of a simulated system and a heat exchanger are included to illustrate the algorithms.

1. Introduction

The most important type of nonlinearity, when no strong threshold effects or other discontinuities are dominant, is given by multiplicative effects. This gives rise to the notion of polynomial input-output maps or response functions, in which present output values are expressed as the sums and products of past input values. The realization of polynomial response functions for deterministic discrete-time nonlinear systems has been investigated in detail by Sontag (1979). One of the striking results provided in Sontag's work is that a polynomial response function is finitely realizable and bounded iff it satisfies a globally valid output-affine difference equation. Such a system can be realized by a state-affine state-space model and Cyrot-Normand and Dang Van Mien (1980) used this as a basis to develop a method of identifying a series of linear signal dependent models which could then be combined to yield an estimate of the nonlinear state-affine state-space model. Billings, Korenberg and Chen (1987) presented an alternative approach by using the output-affine model directly and developed an orthogonal least squares estimation algorithm. A later study introduced the nonlinear stochastic output-affine model and derived a prediction-error estimator to provide optimal parameter estimates from noisy output measurements (Chen and Billings, 1987).

The finite realizable condition alone is necessary and sufficient for a polynomial response function to satisfy a rational difference equation (Sontag, 1979). The output-affine model can therefore be considered as a subset of the class of rational input-output models. Furthermore the general rational model can be considered as a particular expansion of the NARMAX (Nonlinear ARMAX) model (Leontaritis and Billings, 1985) and this should provide a useful alternative to the polynomial expansion of the NARMAX model which has previously been studied in the literature (Billings and Voon, 1986b).

The present study introduces a stochastic rational input-output model as a basis for nonlinear systems identification. A prediction-error estimation routine (Goodwin and Payne, 1977) and a log determinant ratio test derived previously for the output-affine model (Chen and

Billings, 1987) are extended to the rational model case as an efficient method of fitting parsimonious rational models to nonlinear systems. Model validity tests are briefly discussed and examples are included to illustrate the results. The single-input single-output case is studied throughout although the results can be extended to multi-input multi-output systems.

2. The rational input-output model

The function that describes the input-output behaviour of a system is of primary importance in systems theory because this is all an external observer can see. Let U be the input set and Y be the output set with the assumption that U and Y are normed vector spaces. Denote as U^+ the union of all U^t, $t \geq 1$, where U^t is the set of all sequences $(u(1),...,u(t))$ of length t: $U^t = \{(u(1),...,u(t)): u(i) \in U, i=1,...,t\}$. A response function that describes the input-output behaviour of the system can then be defined as $f: U^+ \to Y$. For every $t=1,2,...,$ the response function $f_t: U^t \to Y$ is a different function since the domain U^t is a different one. A strictly causal response function is one where, for each t, $f_t(u(1),...,u(t))$ is independent of $u(t)$. f is said to be finitely realizable iff it has a finite dimensional state-space realization.

f is a polynomial response function if for each t, f_t is a polynomial of finite degree on all variables, although this degree may tend to ∞ as $t \to \infty$. It is known that the polynomial response function f of a discrete deterministic nonlinear system is finitely realizable iff it satisfies the rational difference equation (Sontag, 1979)

$$a(\hat{y}(t-1),...,\hat{y}(t-r),u(t-1),...,u(t-r))\hat{y}(t) = b(\hat{y}(t-1),...,\hat{y}(t-r),u(t-1),...,u(t-r)) \qquad (1)$$

where $\hat{y}(t) \in Y$ is the output of the deterministic system at time t, $a(.)$ and $b(.)$ are polynomials of finite degree and r is the order of the system. This rational difference equation is an input-output model valid globally.

An equivalent model for stochastic systems can be derived by considering the input-output maps based on a prediction-error formulation or innovation approach. Let $u(t)$ and $y(t)$ be the input and output of a stochastic discrete-time causal system at time t respectively. Assume that the observation of the system starts from time 1 and denote

$$y^t = (y(1),...,y(t))^\tau \qquad (2)$$
$$u^t = (u(1),...,u(t))^\tau \qquad (3)$$

The conditional probability density function of $y(t)$ given all the past inputs and outputs u^t and y^{t-1}

$$p(y(t)|y^{t-1},u^t) \qquad (4)$$

for $t=1,2,...,$ completely specifies the system (Leontaritis and Billings, 1985). The output $y(t)$ can be written in a prediction-error or innovation form

$$y(t) = \hat{y}(t) + e(t) \tag{5}$$

where

$$\hat{y}(t) = E[y(t)|y^{t-1},u^t] = f_t(y^{t-1},u^t) \tag{6}$$

is the prediction of the output at time t and

$$e(t) = y(t) - \hat{y}(t) = y(t) - f_t(y^{t-1},u^t) \tag{7}$$

is the prediction error or innovation. Let

$$e^t = (e(1),...,e(t))^\tau \tag{8}$$

The elements of the vector e^{t-1} can be evaluated from the vectors y^{t-1} and u^{t-1} using eqn.(7) iteratively. In a similar manner, the vector y^{t-1} can be obtained from e^{t-1} and u^{t-1}. Therefore, the information contained in (y^{t-1},u^t) is equivalent to that of (e^{t-1},u^t) and

$$p(y(t)|y^{t-1},u^t) = p(y(t)|e^{t-1},u^t) \tag{9}$$

The prediction $\hat{y}(t)$ can thus alternatively be given by

$$\hat{y}(t) = E[y(t)|e^{t-1},u^t] = f_t^*(e^{t-1},u^t) \tag{10}$$

f_t^* can be considered as the response function of a deterministic system where the input is $(u(t),e(t))^\tau$ and the output is $\hat{y}(t)$. If this response function is polynomial and finitely realizable, it follows as a consequence of Sontag's results that the following input-output model describes the system

$$a^*(\hat{y}(t-1),...,\hat{y}(t-r),u(t-1),...,u(t-r),e(t-1),...,e(t-r))\hat{y}(t)$$
$$= b^*(\hat{y}(t-1),...,\hat{y}(t-r),u(t-1),...,u(t-r),e(t-1),...,e(t-r)) \tag{11}$$

or equivalently

$$y(t) = \frac{b(y(t-1),...,y(t-r),u(t-1),...,u(t-r),e(t-1),...,e(t-r))}{a(y(t-1),...,y(t-r),u(t-1),...,u(t-r),e(t-1),...,e(t-r))} + e(t) \tag{12}$$

The rational input-output model (12) can also be derived by an alternative route. Under the assumptions that the response function f is finitely realizable and a linearized model would exist if the system were operated close to an equilibrium point, the system can be represented by the difference equation model (Leontaritis and Billings, 1985)

$$\hat{y}(t) = F^*(\hat{y}(t-1),...,\hat{y}(t-r),u(t-1),...,u(t-r)) \tag{13}$$

in a region around the equilibrium point, where F^* is some nonlinear function. An equivalent model for stochastic systems can be derived as follows by repeating the same procedure given in the above paragraph

$$y(t) = F(y(t-1),...,y(t-r),u(t-1),...,u(t-r),e(t-1),...,e(t-r)) + e(t) \tag{14}$$

Eqn.(14) is known as the NARMAX model. Whilst this model may not be globally valid the response function of the system is no longer limited to be polynomial.

In order to use model (14) as a basis for identification, a means of parametrization is required. Define

$$x_i(t) = \begin{cases} y(t-i) & i=1,\dots,r \\ u(t-i+r) & i=r+1,\dots,2r \\ e(t-i+2r) & i=2r+1,\dots,n \end{cases} \tag{15}$$

where $n = 3r$. If F is analytic, formally it can be expressed as an infinite power series

$$F(x_1(t),\dots,x_n(t)) = \sum_{j=0}^{\infty} \Big\{ \sum_{\substack{k_1+\dots+k_n=j \\ k_i \geq 0,\, i=1,\dots,n}} \theta_{k_1 \cdots k_n} \cdot \prod_{i=1}^{n} (x_i(t))^{k_i} \Big\} \tag{16}$$

In practice, a truncated series has to be used and an approximate polynomial model is given as

$$y(t) = \sum_{j=0}^{L} \Big\{ \sum_{\substack{k_1+\dots+k_n=j \\ k_i \geq 0,\, i=1,\dots,n}} \theta_{k_1 \cdots k_n} \cdot \prod_{i=1}^{n} (x_i(t))^{k_i} \Big\} + e(t) \tag{17}$$

The degree of the polynomial L depends on the shape of nonlinear function F and the requirements of accuracy. The justification for using model (17) is Weierstrass' theorem which states that: Any continuous $F: X \to R$ can be uniformly approximated with arbitrary accuracy by an n-variate polynomial, where X is any closed bounded subset of Euclidean space R^n. The parameters in model (17) can now be estimated using a least squares based or a prediction-error algorithm and several structure determination routines can be employed to select significant terms in eqn.(17) giving rise to a parsimonious model (Billings and Voon, 1984, 1986b; Korenberg, Billings and Liu, 1987; Leontaritis and Billings, 1987b).

If instead a rational approximant of F is used, then the system can be represented approximately as

$$y(t) = \frac{\displaystyle\sum_{j=0}^{L_1} \Big\{ \sum_{\substack{k_1+\dots+k_n=j \\ k_i \geq 0,\, i=1,\dots,n}} \alpha_{k_1 \cdots k_n} \cdot \prod_{i=1}^{n} (x_i(t))^{k_i} \Big\}}{\displaystyle\sum_{j=0}^{L_2} \Big\{ \sum_{\substack{k_1+\dots+k_n=j \\ k_i \geq 0,\, i=1,\dots,n}} \beta_{k_1 \cdots k_n} \cdot \prod_{i=1}^{n} (x_i(t))^{k_i} \Big\}} + e(t) \tag{18}$$

which is equivalent to eqn.(12). For the same number of parameters, a rational approximant can achieve a more varied shape of curves of F than a polynomial approximant. In this sense, the rational model (18) is 'parsimonious' compared with the polynomial model (17). The rational model has another advantage. Some nonlinear time series are inherently stable. A nonexplosive model can always be achieved by a suitable choice of L_1 and L_2 in eqn.(18). To illustrate this, consider the simplest case

$$y(t) = F(y(t-1)) + e(t)$$

If $|F(y)| \leq y$, the model is nonexplosive. An example is

$$y(t) = \frac{0.8 + 0.9y^2(t-1)}{1 + y^2(t-1)} y(t-1) + e(t)$$

In any case, since the rational functions include the polynomials as a special case, it seems

possible that more efficient approximations can be formed by searching in this wider class. Some attractive properties of rational approximants defined from power series eqn.(16) can be found in (Chisholm and McEwan, 1974).

3. Prediction-error estimation and structure determination

The rational model (12) or more specifically eqn.(18) is not linear-in-the-parameters and hence, the linear least squares principle is not applicable. However, the prediction-error estimation method is a general estimation method and can readily be applied to the model of eqn.(18). Arrange the parameters in eqn.(18) into a vector θ, where the dimension of θ is

$$n_\theta = \sum_{i=0}^{L_1} n_i + \sum_{i=0}^{L_2} n_i; \quad n_0=1; \quad n_i = n_{i-1}(3r+i-1)/i, \ i\geq 1 \tag{19}$$

For a given value of the parameter vector θ, the residual is given as

$$\varepsilon(t,\theta) = y(t) - \frac{b(y(t-1),...,y(t-r),u(t-1),...,u(t-r),\varepsilon(t-1,\theta),...,\varepsilon(t-r,\theta))}{a(y(t-1),...,y(t-r),u(t-1),...,u(t-r),\varepsilon(t-1,\theta),...,\varepsilon(t-r,\theta))} \tag{20}$$

The loss function will be chosen as

$$J(\theta) = \frac{1}{2}\log \det Q(\theta) = \frac{1}{2}\log Q(\theta) \tag{21}$$

in the scalar case where

$$Q(\theta) = \frac{1}{N}\sum_{t=1}^{N}\varepsilon^2(t,\theta) \tag{22}$$

and N is the data length.

The i-th element of gradient vector $\nabla_\theta J = \dfrac{\partial^\tau J}{\partial \theta}$ and the approximate ij-th element of Hessian matrix $H = \dfrac{\partial^2 J}{\partial \theta^2}$ are given by (Goodwin and Payne, 1977)

$$\frac{\partial J}{\partial \theta_i} = \frac{1}{NQ(\theta)}\sum_{t=1}^{N}\varepsilon(t,\theta)\frac{\partial\varepsilon(t,\theta)}{\partial\theta_i}, \quad i=1,...,n_\theta \tag{23}$$

$$\frac{\partial^2 J}{\partial \theta_i \partial \theta_j} = \frac{1}{NQ(\theta)}\sum_{t=1}^{N}\frac{\partial\varepsilon(t,\theta)}{\partial\theta_i}\frac{\partial\varepsilon(t,\theta)}{\partial\theta_j}, \quad i,j=1,...,n_\theta \tag{24}$$

The derivatives of the residuals can be obtained by differentiating eqn.(20)

$$\frac{\partial\varepsilon(t,\theta)}{\partial\theta_i} = -\frac{1}{a(.)}\frac{\partial b(.)}{\partial\theta_i} + \frac{b(.)}{a^2(.)}\frac{\partial a(.)}{\partial\theta_i}, \quad i=1,...,n_\theta \tag{25}$$

It is difficult to give a more specific formula for the derivatives of the residuals but each derivative can be calculated once r, L_1 and L_2 are specified. For example, a first order system with a linear $a(.)$ and a quadratic $b(.)$ is written as

$$y(t)=\frac{b(y(t-1),u(t-1),\varepsilon(t-1,\theta))}{a(y(t-1),u(t-1),\varepsilon(t-1,\theta))}+\varepsilon(t,\theta)=\frac{1}{\theta_{11}+\theta_{12}y(t-1)+\theta_{13}u(t-1)+\theta_{14}\varepsilon(t-1,\theta)}\{\theta_1+\theta_2 y(t-1)+\theta_3 u(t-1)+$$

$$\theta_4\varepsilon(t-1,\theta)+\theta_5 y^2(t-1)+\theta_6 y(t-1)u(t-1)+\theta_7 y(t-1)\varepsilon(t-1,\theta)+\theta_8 u^2(t-1)+\theta_9 u(t-1)\varepsilon(t-1,\theta)+\theta_{10}\varepsilon^2(t-1,\theta)\}+\varepsilon(t,\theta)$$

and

$$\frac{\partial \varepsilon(t,\theta)}{\partial \theta_1} = -\frac{1}{a(.)}\{1+(\theta_4+\theta_7 y(t-1)+\theta_9 u(t-1)+2\theta_{10}\varepsilon(t-1,\theta))\frac{\partial \varepsilon(t-1,\theta)}{\partial \theta_1}\}+\frac{b(.)}{a^2(.)}\theta_{14}\frac{\partial \varepsilon(t-1,\theta)}{\partial \theta_1}$$

etc.

The prediction-error estimator produces an estimate $\hat{\theta}$ of the parameter vector by minimizing the loss function (21)

$$\hat{\theta} = arg \min_{\theta} J(\theta) \qquad (26)$$

Since the gradient and the approximate Hessian of $J(\theta)$ are provided, the minimisation of $J(\theta)$ can be performed very efficiently using Newton's method. The optimisation procedure involving Newton's method is well docummented elsewhere and will not be detailed here. The direction vector $d = -H^{-1}\nabla_\theta J$ required at each step of the minimisation is usually obtained from

$$W^\tau W \, d = -\nabla_\theta J \qquad (27)$$

rather than direct inversion, where W is the square root of H and is upper triangular. When the model is over-parametrized, the Hessian tends to become almost singular and the square root of the Hessian becomes very difficult to calculate numerically. In this case, a diagonal matrix ρI, where ρ is a small positive scalar, can be added to the Hessian before the factorization. That is

$$W^\tau W = H + \rho I \qquad (28)$$

The choice of the initial value θ^0 for the parameter vector imposes a small problem. Unlike the case of the polynomial model, θ^0 cannot be chosen as the zero vector. A solution is to assume initially $a(.)=1$ and to use the least squares estimate to provide some values of θ^0. The rest of θ^0 may then be set to zero.

The determination of the structure or which terms to include in the final model is essential since model (18) can easily involve an excessive number of terms. Many of them may actually be redundant and can be removed with little effect. A parsimonious model is highly desired if the model is to be employed in controller design, prediction and other applications.

One possible approach to the problem of how to select one of two models that can describe the data is to use results from hypothesis testing. Two hypotheses, the null hypothesis and the alternative hypothesis, will be involved in the problem. The test that determines whether the null or alternative hypothesis is selected is usually based on a statistic. In the situation where parameter estimates are provided by the prediction-error estimator, the statistic can be chosen as

$$d(y) = N \log \frac{\det Q(\hat{\theta}_o)}{\det Q(\hat{\theta}_1)} \qquad (29)$$

where $\hat{\theta}_o$ and $\hat{\theta}_1$ are the prediction-error estimates under the null and the alternative hypotheses respectively. The test based on this statistic is referred to as the log determinant ratio test (Leontaritis and Billings, 1987a). In the scalar case, the statistic (29) reduces to

$$d(y) = N \log \frac{Q(\hat{\theta}_o)}{Q(\hat{\theta}_1)} \tag{30}$$

If the null hypothesis is that s elements of θ are equal to zero and the alternative hypothesis is that they are unrestricted, then under the assumption of the null hypothesis, the statistic $d(y)$ is asymptotically distributed as a chi-squared distribution with s degrees of freedom

$$d(y) \rightarrow \chi^2(s) \tag{31}$$

Denote $k_\alpha(s)$ the critical value of the chi-squared distribution with s degrees of freedom and a given significant level α. The test involves calculating the value of $d(y)$ and comparing it with $k_\alpha(s)$. The null hypothesis is accepted if $d(y) < k_\alpha(s)$. If $d(y) \geq k_\alpha(s)$ there is strong evidence against the null hypothesis and it is thus rejected.

The situation where a selection is made between two models is rather restricted. The usual case is that the selection has to be made from many different models. Incorporating the requirement of non-conflicting pairwise comparisons with the above test results in a C-criterion (Leontaritis and Billings, 1987a) which the model selected from all the competing models must minimize

$$C = N \log \det Q(\overline{\theta}) + \overline{n}_\theta \, k_\alpha(1) \tag{32}$$

or

$$C = N \log Q(\overline{\theta}) + \overline{n}_\theta \, k_\alpha(1) \tag{33}$$

in the present scalar case, where $\overline{\theta}$ is the parameter vector of the particular model and $\overline{n}_\theta$ its dimension. Akaike's information criterion (AIC) corresponds to this criterion with a particular $k_\alpha(1) = 2$. It is well-known that AIC may consistently overestimate the true parameter vector and a more reasonable choice of $k_\alpha(1)$ has been shown to be 4 (Leontaritis and Billings, 1987a).

In order to use the C-criterion for determining the correct structure of a rational model, the model reduction method (Leontaritis and Billings, 1987b) can be employed. Initially, the most complicated model is defined by specifying r, L_1 and L_2 in eqn.(18). This model is referred to as the full model. All the other models to be considered are reduced models of the full model with some of the parameters of the full model set to zero. The model that minimizes the C-criterion is regarded as the best one and is selected as the final model.

Inspite of the disadvantage that the Hessian of the loss function for the full model may become almost singular and therefore the precaution illustrated in eqn.(28) for calculating the square root is required, there are many numerical advantages of using the model reduction

approach. In particular, the value of the C-criterion for a reduced model can be evaluated extremely efficiently once the square root of the full model Hessian is known without using the prediction-error estimator to compute the corresponding Hessian at each stage. Details of the numerical implementation using the Householder orthogonal transformation to compute the values of the C-criterion has been given in (Leontaritis and Billings, 1987b; Chen and Billings, 1987) and therefore will not be repeated here. The number of reduced models is the combinatorial sum of the dimension of the full parameter vector. It can easily become prohibitively large. In practice, the stepwise backward elimination (SBE) method (Draper and Smith, 1981) can be employed to overcome this difficulty at the cost that the final model selected may only be a suboptimal one.

4. Correlation-based model validation

If the full model does not include all the significant terms of the true system, the model reduction approach cannot find the correct structure of the system and the final model selected will not be a good representation of the system. Model validation should be able to reveal such a deficient situation. In general, model validation forms the final stage of any identification procedure.

A statistical test known as the chi-squared test has been introduced by Bohlin (1978) and used for nonlinear model validation by Leontaritis and Billings (1987a). Define

$$x^t = \begin{bmatrix} y^{t-1} \\ e^{t-1} \\ u^t \end{bmatrix} \tag{34}$$

as the vector that contains all the outputs, innovations and inputs up to time t, and

$$w(t) = e(t)/\sigma \tag{35}$$

as the normalized innovation where

$$\sigma^2 = E[e^2(t)|x^t] \tag{36}$$

Let the vector $Z(t)$ be a function of x^t

$$Z(t) = Z(x^t) \tag{37}$$

such that the central limit theorem holds

$$\mu = \frac{1}{N} \sum_{t=1}^{N} Z(t)w(t) \tag{38}$$

and the law of large number holds for the time average

$$\Gamma^\tau \Gamma = \frac{1}{N} \sum_{t=1}^{N} Z(t)Z^\tau(t) \tag{39}$$

It can be shown that the random vector μ is asymptotically normal with mean zero and covariance matrix $\Gamma^\tau \Gamma/N$ (Bohlin, 1978). The normalized random vector

$$\xi = \sqrt{N}\Gamma^{-\tau}\mu \qquad (40)$$

is thus asymptotically normal with zero mean and unit covariance. The variable

$$\eta = \xi^\tau\xi = N\mu^\tau(\Gamma^\tau\Gamma)^{-1}\mu \qquad (41)$$

is then asymptotically chi-squared distributed with s degrees of freedom where s is the dimension of $Z(t)$.

The problem of validating an estimated model can be regarded as one of statistical testing where the null hypothesis is that the input-output data are generated by the model and the alternative hypothesis is that the data are not generated by the model. The value of the statistic η can be calculated by replacing the innovation $e(t)$ with the residuals $\varepsilon(t,\hat{\theta})$. If this value is within the acceptance region for a given level of significance α, that is

$$\eta < k_\alpha(s) \qquad (42)$$

the model is accepted. Otherwise the model is not an acceptable one.

The difficulty in implementing the above test is how to choose the elements of $Z(t)$. For nonlinear systems, the elements of $Z(t)$ must include nonlinear functions of x^t. Which type of nonlinear functions are better than the others, however, is not clear. Leontaritis and Billings (1987a) considered the following type of $Z(t)$

$$Z(t) = (m(t),...,m(t-s+1))^\tau, \quad s=1,...,t_d \qquad (43)$$

where $m(t)$ is a monomial of elements of x^t, for instance it could be selected as

$$m(t) = y^{i_y}(t-1)y^{j_y}(t-2)y^{k_y}(t-3)u^{i_u}(t-1)u^{j_u}(t-2)u^{k_u}(t-3)\varepsilon^{i_\varepsilon}(t-1,\hat{\theta})$$
$$\cdot \varepsilon^{j_\varepsilon}(t-2,\hat{\theta})\varepsilon^{k_\varepsilon}(t-3,\hat{\theta}), \quad i_y, j_y, k_y, i_u, j_u, k_u, i_\varepsilon, j_\varepsilon, k_\varepsilon \geq 0 \qquad (44)$$

and t_d is the maximum delay considered in the chi-squared correlation test. It remains unanswered how many types of monomials $m(t)$ must be tested before it can be said that a model is properly validated. It is in this area where experience can help.

Alternatively, a simple correlation test can be used to validate an estimated model (Billings and Voon, 1986a). The test involves calculating five sampled correlation functions

$$\left.\begin{array}{ll} \hat{\Phi}_{\varepsilon\varepsilon}(k) & k \neq 0 \\ \hat{\Phi}_{u\varepsilon}(k) & \forall k \\ \hat{\Phi}_{\varepsilon(\varepsilon u)}(k) & k \geq 0 \\ \hat{\Phi}_{u^{2\prime}\varepsilon}(k) & \forall k \\ \hat{\Phi}_{u^{2\prime}\varepsilon^2}(k) & \forall k \end{array}\right\} \qquad (45)$$

where $u^{2\prime}(t)=u^2(t)-\overline{u^2(t)}$ and the bar indicates time averaging. If these correlation functions fall within the (95%) confidence interval $\pm 1.96/\sqrt{N}$, the model is regarded as adequate. As pointed out by Leontaritis and Billings (1987a), the nature of this correlation test is revealing. It may not be viewed as a properly defined statistical test since the test may become extremely severe. However, this correlation test is very easy to implement. Experience has shown that if the test

is used in conjunction with a prediction-error algorithm, it often gives the experimenter a great deal of information regarding deficiencies in the fitted model, and helps to indicate which terms should be included in the model to improve the fit.

5. Simulation study

Example 1.

This is a simulated example. An input-output sequence of 1000 points was generated using the model

$$y(t) = \frac{y(t-1)+u(t-1)+y(t-1)u(t-1)+y(t-1)e(t-1)}{1+y^2(t-1)+u(t-1)e(t-1)}+e(t)$$

The input $u(t)$ was an independent sequence of uniform distribution with mean zero and variance 1.0 and the noise $e(t)$ was a Gaussian white sequence with zero mean and variance 0.01. The data was partitioned as estimation and testing sets. The estimation set consisted of the first 600 points of the data.

A linear model with $r=2$ was first used to fit the data. The prediction-error estimates and their standard deviations are given in TABLE 1. A clear difference between the one-step ahead prediction of the model and the measured output can be seen in Fig. 1 where unacceptably large residuals are also observed. The simple correlation tests and two typical chi-square tests are shown in Fig.s 2 and 3 respectively. Notice that if only the traditional linear covariance tests $\hat{\Phi}_{\varepsilon\varepsilon}(k)$ and $\hat{\Phi}_{u\varepsilon}(k)$ had been used, the experimenter would have been misled into believing that the estimated linear model was adequate. The higher-order correlation tests correctly revealed the deficiency in the fitted model. A similar and more visual result was obtained using the chi-square tests. When the monomial $m(t)$ was linear, the statistic was well inside the acceptance region. However, using nonlinear $m(t)$ clearly showed that the linear model was inadequate.

Next, a rational model of $L_1=L_2=2$ and $r=1$ was used. The full model had 20 terms. The estimates and standard deviations are shown in TABLE 2. The SBE routine was then employed to simplify the full model. The elimination procedure is given in TABLE 3. It is seen that the C-criterion deleted all the redundant parameters in the model while AIC over-estimated the number of required parameters and retained three redundant terms. Notice that the final estimated model shown in TABLE 4. is correct. In fact, the system is

$$y(t) = \frac{y(t-1)+u(t-1)+y(t-1)u(t-1)+y(t-1)e(t-1)}{1+y^2(t-1)+u(t-1)e(t-1)}+e(t)$$

$$= \frac{\alpha y(t-1)+\alpha u(t-1)+\alpha y(t-1)u(t-1)+\alpha y(t-1)e(t-1)}{\alpha+\alpha y^2(t-1)+\alpha u(t-1)e(t-1)}+e(t)$$

This indicates that the parameters in a rational model can, in general, be identified uniquely

only within a scalar, that is, parameter vector θ is indistinguishable from $\alpha\theta$, $\alpha \neq 0$. This phenomenon however poses no problem to the prediction-error estimator, the model reduction routine, or to the applications of the model. If there is a prior knowledge that a term exists in the system, then this parameter may be fixed and the estimates will be unique. As expected, the results of the model validation shown in Fig.s 4, 5 and 6 confirm that the estimated final model is correct.

Example 2. A heat exchanger

A detailed description of this process and the experiment design can be found in (Billings and Fadzil, 1985). One of loops in the process has been shown to be nonlinear and NARMAX polynomial models have been fitted using different estimators and structure determination techniques (Billings and Fadzil, 1985; Liu Korenberg, Billings and Fadzil, 1987).

The authors (Liu et al, 1987) used an orthogonal least squares estimator to determine the model structure and parameter estimates. TABLE 5, Fig.s 7, 8 and 9 give the results of the prediction-error estimation and model validation using their model structure. These results show that this nonlinear model is adequate.

In the present study, a rational model has been fitted to the same nonlinear loop. The model structure was determined using the SBE routine. The final estimates are given in TABLE 6. Model validity tests shown in Fig.s 10, 11 and 12 indicate that this rational model is also satisfactory.

Finally, as a comparison, a linear model with $r=5$ was fitted to the same data set. The results are given in TABLE 7, Fig.s 13, 14 and 15. Again, the model validity results show that the traditional linear validity tests give very misleading information.

6. Conclusions

If the response of the system is dominated by nonlinear characteristics it is often necessary to use a nonlinear model and this immediately raises the problem of what class of models to use. The rational representation provides a concise parametric model for a wide class of discrete-time nonlinear stochastic systems. It includes the linear model, the NARMAX polynomial model and the output-affine model as special cases. Unified computer software can therefore be developed to identify these various models.

A combined algorithm incorporating a prediction-error estimator and structure determination routine has been developed for the rational model. The application to both simulated and real data has been demonstrated.

Model validity tests for nonlinear systems have been briefly discussed and it has been shown that the traditional model validity techniques developed for linear systems often provide incorrect information when applied to nonlinear systems

References

[1] Billings, S.A., and M.B. Fadzil (1985). The practical identification of systems with nonlinearities. *Proc. of the 7th IFAC Symp. on Identification and System Parameter Estimation*, York, U.K., pp155-160.

[2] Billings, S.A., M.J. Korenberg and S. Chen (1987). Identification of nonlinear output-affine systems using an orthogonal least squares algorithm. *Int. J. Systems Sci.*, (to appear).

[3] Billings, S.A., and W.S.F. Voon (1984). Least squares parameter estimation algorithms for non-linear systems. *Int. J. Systems Sci.*, Vol.15, No.6, pp601-615.

[4] Billings, S.A., and W.S.F. Voon (1986a). Correlation based model validity tests for non-linear models. *Int. J. Control*, Vol.44, No.1, pp235-244.

[5] Billings, S.A., and W.S.F. Voon (1986b). A prediction-error and stepwise-regression estimation algorithm for non-linear systems. *Int. J. Control*, Vol.44, No.3, pp803-822.

[6] Bohlin, T.(1978). Maximum-power validation of models without higher-order fitting. *Automatica*, Vol.14, pp137-146.

[7] Chen, S., and S.A. Billings (1987). A prediction-error estimation algorithm for nonlinear output-affine systems. *Int. J. Control*, (to appear).

[8] Chisholm, J.S.R., and J. McEwan (1974). Rational approximants defined from power series in N variables. *Proc. R. Soc. Lond.*, Series A, Vol.336, pp421-452.

[9] Cyrot-Normand, D., and H. Dang Van Mien (1980). Nonlinear state-affine identification methods: applications to electrical power plants. *Preprints of IFAC Symp. on Auto. Control in Power Generation, Distribution and Protection.* Pretoria, South Africa, 1980, pp449-462.

[10] Draper, N.R., and H. Smith (1981). *Applied Regression Analysis.* Wiley, New York.

[11] Goodwin, G.C., and R.L. Payne (1977). *Dynamic System Identification: Experiment Design and Data Analysis.* Academic Press, New York.

[12] Korenberg, M.J., S.A. Billings, and Y.P. Liu (1987). An orthogonal parameter estimation algorithm for nonlinear stochastic systems. *Int. J. Control,* (to appear).

[13] Leontaritis, I.J., and S.A. Billings (1985). Input-output parametric models for non-linear systems, Part I: deterministic non-linear systems; Part II: stochastic non-linear systems. *Int. J. Control,* Vol.41, No.2, pp303-344.

[14] Leontaritis, I.J., and S.A Billings (1987a). Model selection and validation methods for non-linear systems. *Int. J. Control,* Vol.45, No.1, pp311-341.

[15] Leontaritis, I.J., and S.A. Billings (1987b). A prediction error estimator for nonlinear stochastic systems. *Int. J. Systems Sci.,* (to appear).

[16] Liu, Y.P., M.J. Korenberg, S.A. Billings, and M.B. Fadzil (1987). The nonlinear identification of a heat exchanger. To be presented at *the 26th IEEE Conference on Decision and Control,* Dec.9-11, 1987, Los Angeles, U.S.A..

[17] Sontag, E.D. (1979). *Polynomial Response Maps,* Lecture Notes in Control and Information Sciences 13. Springer Verlag, Berlin.

TABLE 1. Estimates and standard deviations (linear model for Example 1)

terms	estimates	standard deviations
constant	-0.11785E+0	0.30845E-1
$y(t-1)$	-0.32321E-1	0.26154E+0
$y(t-2)$	0.26010E+0	0.12237E+0
$u(t-1)$	0.59244E+0	0.17170E-1
$u(t-2)$	0.28211E+0	0.15713E+0
$e(t-1)$	0.31393E+0	0.26464E+0
$e(t-2)$	-0.63846E-1	0.70689E-1

TABLE 2. Estimates and standard deviations (full model for Example 1)

	terms	estimates	standard deviations
numerator	$constant$	0.72790E-2	0.90390E-2
	$y(t-1)$	0.10154E+1	0.15747E+0
	$u(t-1)$	0.10037E+1	0.15559E+0
	$e(t-1)$	0.11100E+0	0.57900E-1
	$y^2(t-1)$	-0.54036E-2	0.15971E-1
	$y(t-1)u(t-1)$	0.98884E+0	0.15434E+0
	$y(t-1)e(t-1)$	0.10151E+1	0.18164E+0
	$u^2(t-1)$	0.54017E-2	0.93588E-2
	$u(t-1)e(t-1)$	0.21821E-1	0.87255E-1
	$e^2(t-1)$	0.18346E+0	0.23565E+0
denominator	$constant$	0.97625E+0	0.15133E+0
	$y(t-1)$	-0.25374E-1	0.27148E-1
	$u(t-1)$	0.14922E-1	0.15863E-1
	$e(t-1)$	-0.47052E-1	0.12369E+0
	$y^2(t-1)$	0.99396E+0	0.15458E+0
	$y(t-1)u(t-1)$	0.35260E-1	0.18588E-1
	$y(t-1)e(t-1)$	0.17339E+0	0.12884E+0
	$u^2(t-1)$	0.15069E-1	0.76022E-2
	$u(t-1)e(t-1)$	0.92783E+0	0.15389E+0
	$e^2(t-1)$	-0.50969E+0	0.28047E+0

TABLE 3. Model reduction using SBE (full model for example 1)

elimination step	eliminated parameter	AIC value	C-criterion value	standard deviation of residuals
		-0.27184E+4	-0.26784E+4	0.10001E+0
1	$u(t-1)e(t-1)$	-0.27204E+4	-0.26824E+4	0.10001E+0
2	$y^2(t-1)$	-0.27223E+4	-0.26863E+4	0.10002E+0
3	$u^2(t-1)$	-0.27240E+4	-0.26900E+4	0.10004E+0
4	$constant$	-0.27255E+4	-0.26935E+4	0.10008E+0
5	$u(t-1)$ *	-0.27273E+4	-0.26973E+4	0.10010E+0
6	$e^2(t-1)$	-0.27286E+4	-0.27006E+4	0.10016E+0
7	$e(t-1)$ *	-0.27293E+4	-0.27033E+4	0.10027E+0
8	$y(t-1)$ *	-0.27297E+4	-0.27057E+4	0.10040E+0
9	$y(t-1)e(t-1)$ *	-0.27302E+4	-0.27082E+4	0.10053E+0
10	$y(t-1)u(t-1)$ *	_-0.27303E+4_	-0.27103E+4	0.10069E+0
11	$u^2(t-1)$ *	-0.27300E+4	-0.27120E+4	0.10088E+0
12	$e(t-1)$	-0.27302E+4	-0.27142E+4	0.10103E+0
13	$e^2(t-1)$ *·	-0.27294E+4	_-0.27154E+4_	0.10127E+0
14	$y(t-1)e(t-1)$	-0.27075E+4	-0.26955E+4	0.10330E+0

*: parameter in denominator.

TABLE 4. Estimates and standard deviations (final model for Example 1)

	terms	estimates	standard deviations
numerator	$y(t-1)$	0.61074E+0	0.15673E+0
	$u(t-1)$	0.61245E+0	0.15716E+0
	$y(t-1)u(t-1)$	0.60968E+0	0.15646E+0
	$y(t-1)e(t-1)$	0.58427E+0	0.15361E+0
denominator	$constant$	0.60820E+0	0.15609E+0
	$y^2(t-1)$	0.61379E+0	0.15772E+0
	$u(t-1)e(t-1)$	0.55981E+0	0.14569E+0

TABLE 5. Estimates and standard deviations of polynomial model (Example 2)

terms	estimates	standard deviations
constant	0.22474E+1	0.38315E+0
$y(t-1)$	0.10568E+1	0.18920E+0
$y(t-2)$	-0.18571E+0	0.16181E+0
$u(t-1)$	0.50474E+0	0.57263E-1
$e(t-1)$	-0.13618E+0	0.42356E-1
$e(t-2)$	0.79003E-2	0.39261E-1
$e(t-3)$	0.22674E+0	0.37954E-1
$e(t-4)$	0.48573E-1	0.38295E-1
$e(t-5)$	-0.14145E+0	0.38217E-1
$y^2(t-1)$	-0.21461E-1	0.19541E-1
$y(t-1)y(t-2)$	0.83798E-2	0.18673E-1
$y(t-1)u(t-1)$	-0.28540E-1	0.63780E-2
$u^2(t-1)$	-0.12705E-1	0.19391E-2

TABLE 6. Estimates and standard deviations of rational model (Example 2)

	terms	estimates	standard deviations
numerator	constant	0.48816E+1	0.37165E+1
	$y(t-2)$	-0.36175E+0	0.30592E+0
	$u(t-2)$	0.49541E-1	0.41992E-1
	$e(t-3)$	0.31704E+0	0.24977E+0
	$e(t-4)$	0.12353E+0	0.10699E+0
	$e(t-5)$	-0.21385E+0	0.17224E+0
	$y^2(t-1)$	0.14087E+0	0.10669E+0
	$y(t-1)u(t-1)$	-0.98255E-1	0.78460E-1
	$u^2(t-1)$	-0.51901E-1	0.40601E-1
denominator	constant	0.96123E+0	0.73080E+0
	$y(t-2)$	-0.37864E-1	0.32970E-1
	$u(t-1)$	-0.13988E+0	0.10997E+0
	$y^2(t-1)$	0.10151E-1	0.76705E-2

TABLE 7. Estimates and standard deviations of linear model (Example 2)

terms	estimates	standard deviations
constant	0.10124E+1	0.66118E+0
$y(t-1)$	0.86731E+0	0.18316E+0
$y(t-2)$	-0.12683E+0	0.18985E+0
$y(t-3)$	0.45364E+0	0.16745E+0
$y(t-4)$	-0.36701E+0	0.12420E+0
$y(t-5)$	0.58619E-1	0.62505E-1
$u(t-1)$	0.26513E+0	0.91180E-2
$u(t-2)$	-0.52428E-1	0.51516E-1
$u(t-3)$	-0.68724E-2	0.43055E-1
$u(t-4)$	-0.10280E+0	0.31954E-1
$u(t-5)$	-0.29881E-1	0.25234E-1
$e(t-1)$	-0.11213E+0	0.18618E+0
$e(t-2)$	0.15859E-1	0.14955E+0
$e(t-3)$	-0.27407E+0	0.11371E+0
$e(t-4)$	0.25822E-1	0.82294E-1
$e(t-5)$	-0.11156E+0	0.57652E-1

Fig.1. System and estimated linear model of Example 1

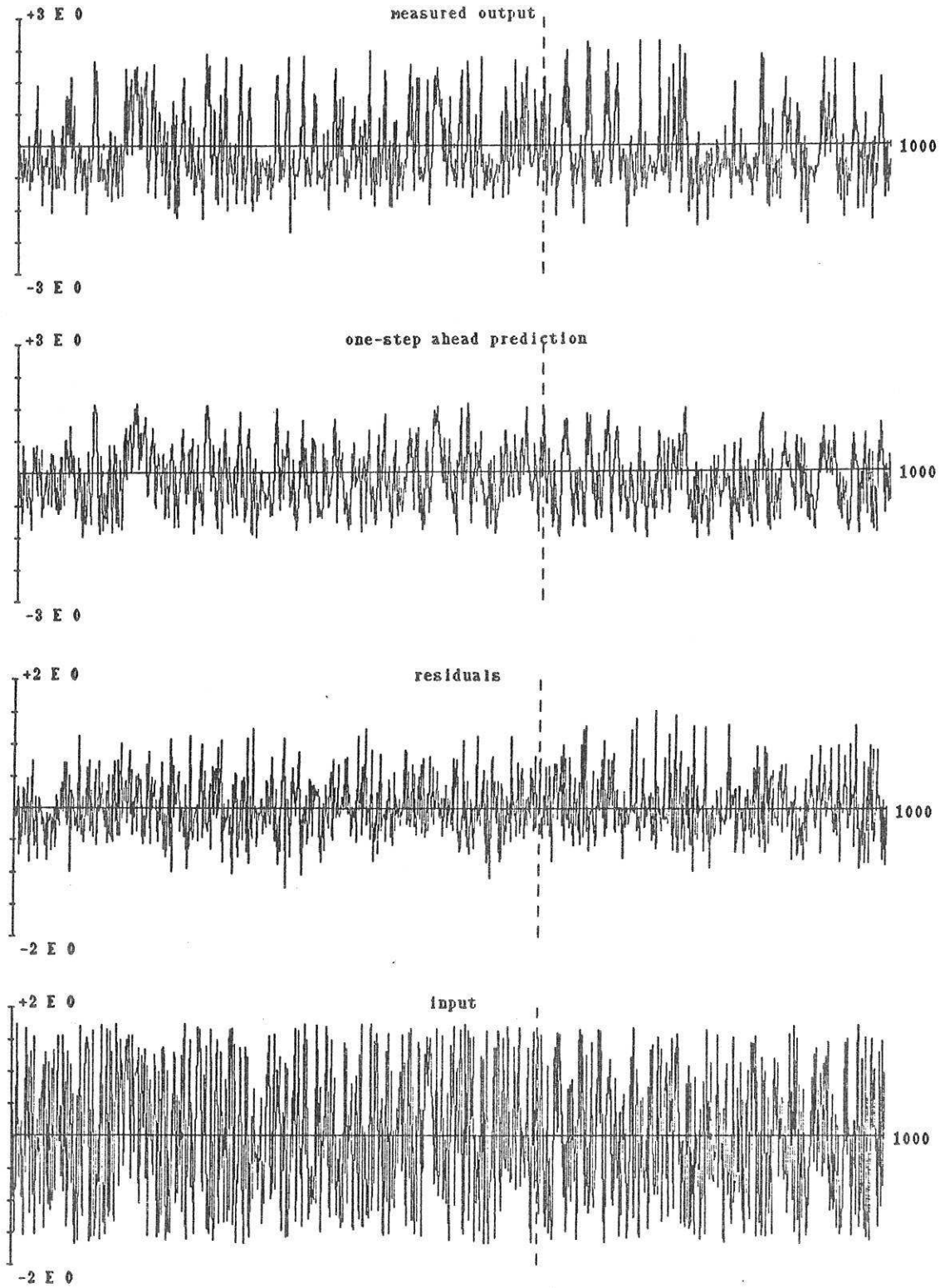**Fig.2. Simple correlation tests (linear model for Example 1)**
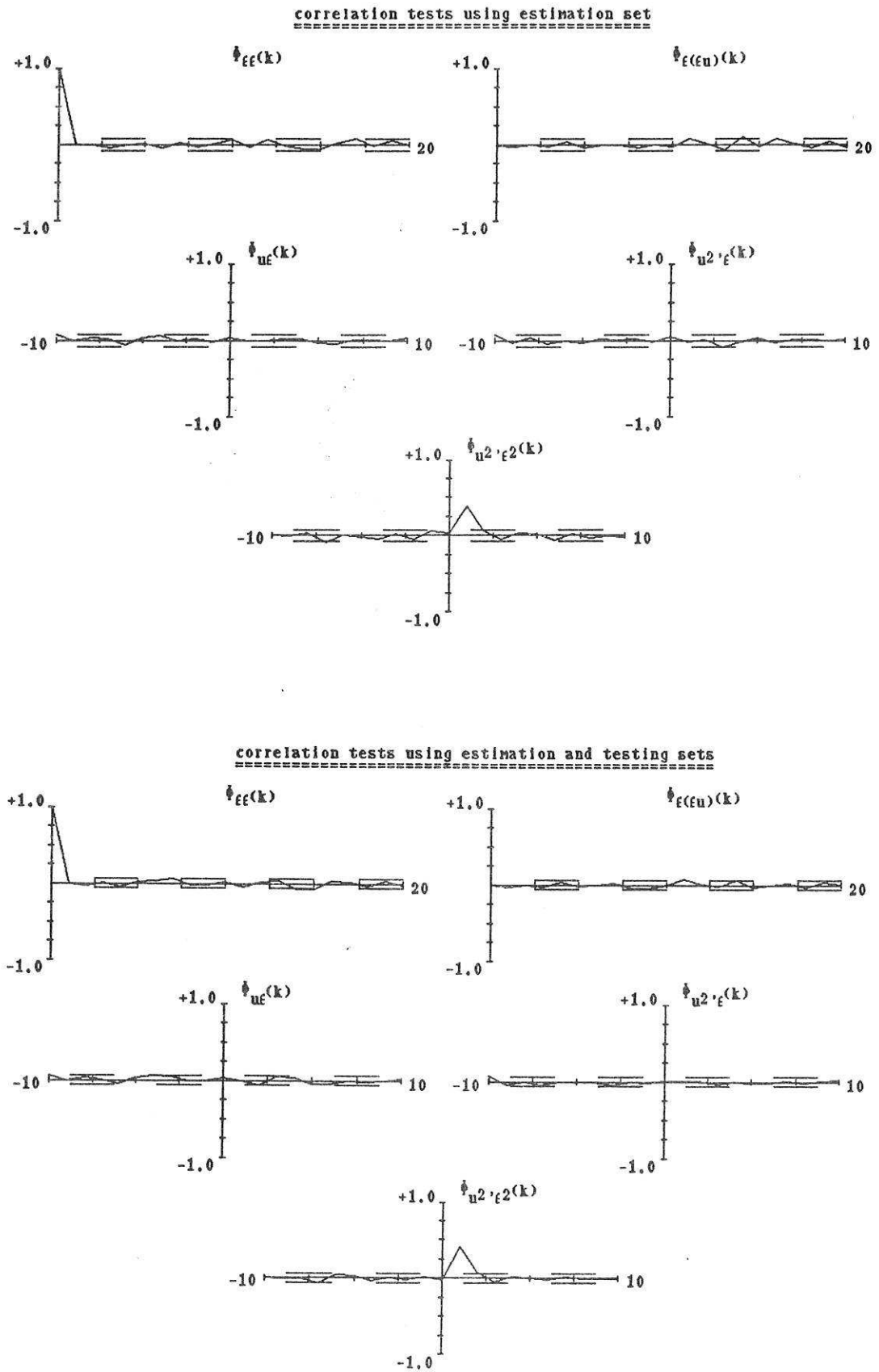
Fig.3. Chi-square statistical tests (linear model for Example 1)

chi-square correlation test

$m(t)=y(t-1)**1$

correlation using estimation set

correlation using estimation and testing sets

chi-square correlation test

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

correlation using estimation set

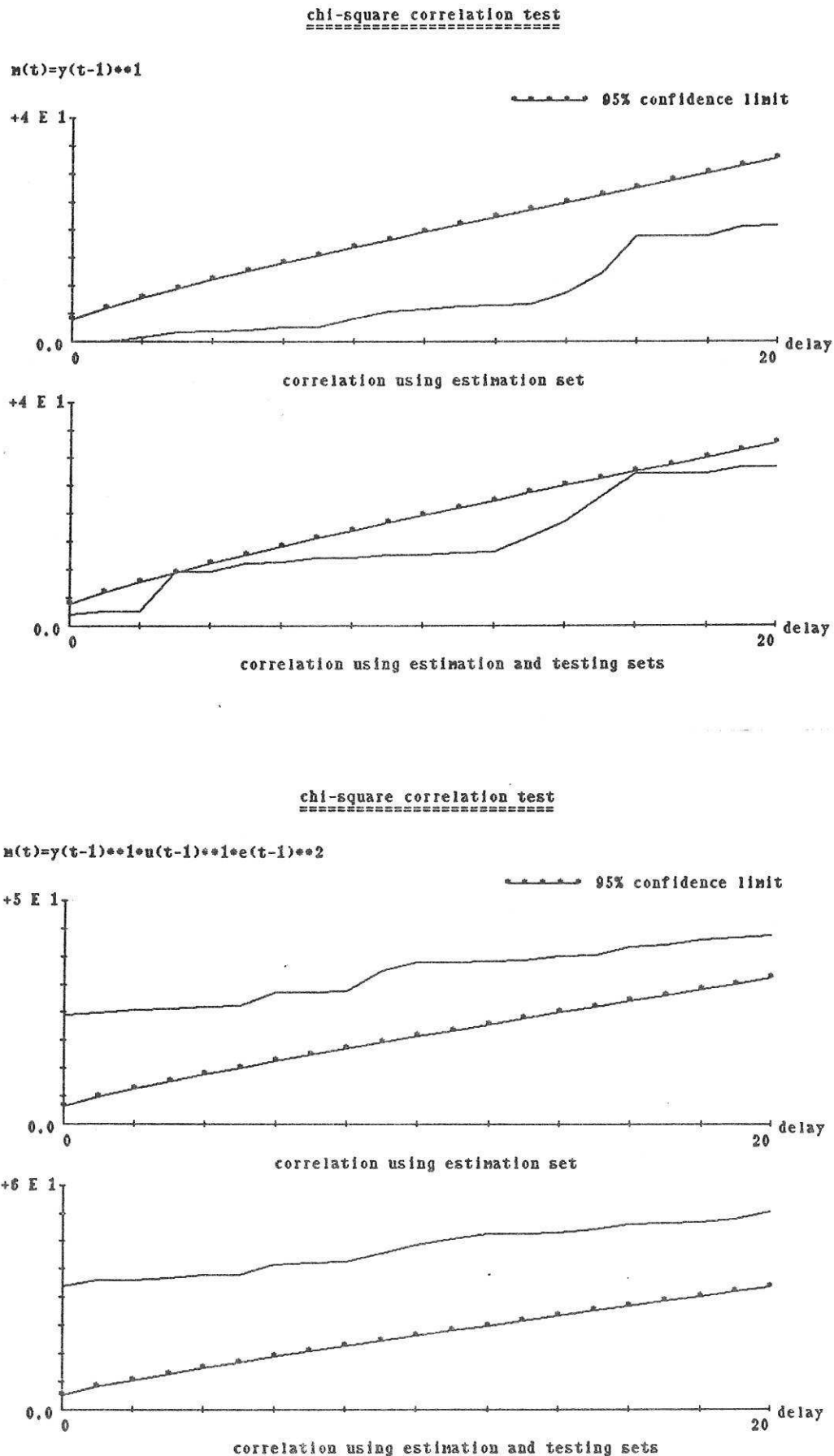correlation using estimation and testing sets

Fig.4. Predicted outputs and residuals of final model (Example 1)
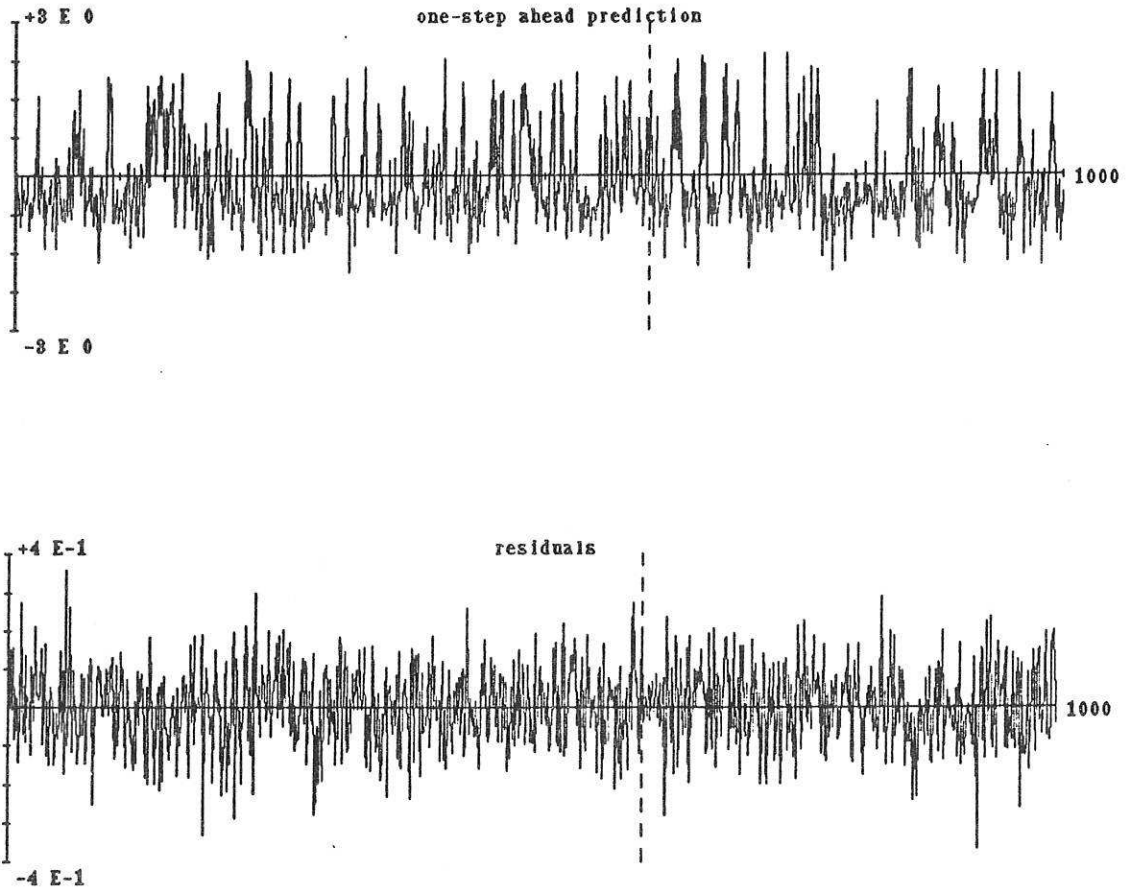
Fig.5. Simple correlation tests (final model for Example 1)

correlation tests using estimation set

$\hat{\phi}_{\epsilon\epsilon}(k)$ $\hat{\phi}_{\epsilon(\epsilon u)}(k)$ $\hat{\phi}_{u\epsilon}(k)$ $\hat{\phi}_{u2',\epsilon}(k)$ $\hat{\phi}_{u2',\epsilon 2}(k)$

correlation tests using estimation and testing sets

$\hat{\phi}_{\epsilon\epsilon}(k)$ $\hat{\phi}_{\epsilon(\epsilon u)}(k)$ $\hat{\phi}_{u\epsilon}(k)$ $\hat{\phi}_{u2',\epsilon}(k)$ $\hat{\phi}_{u2',\epsilon 2}(k)$
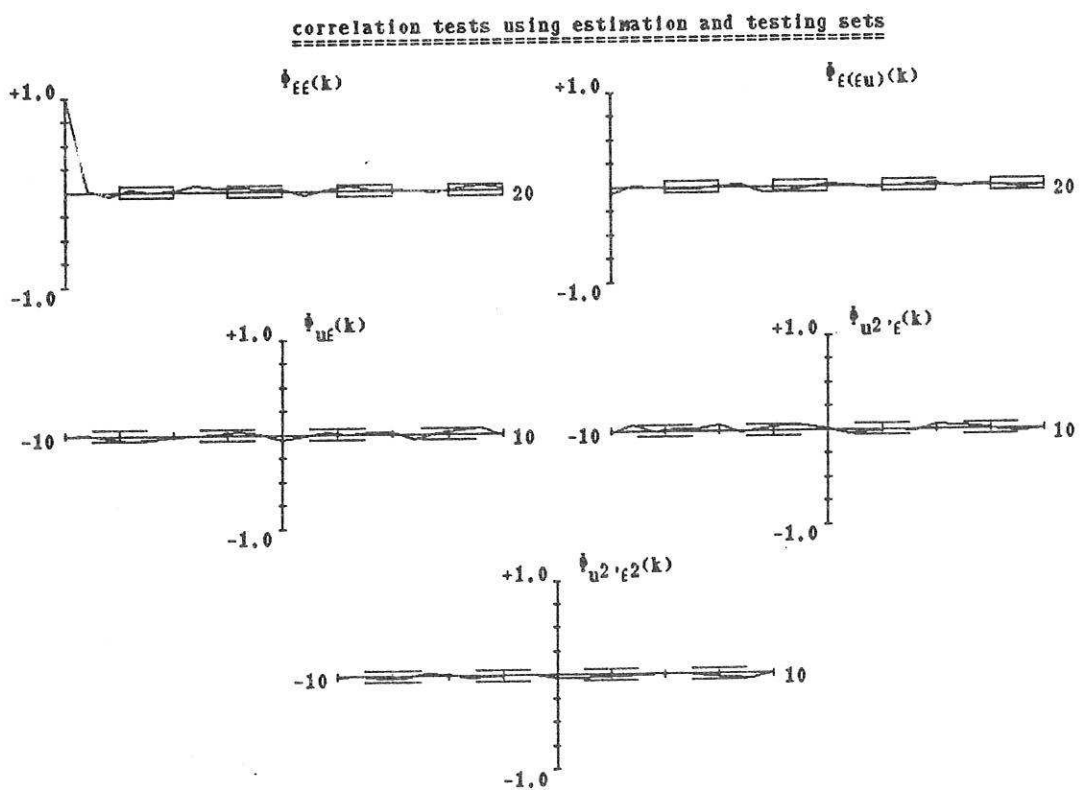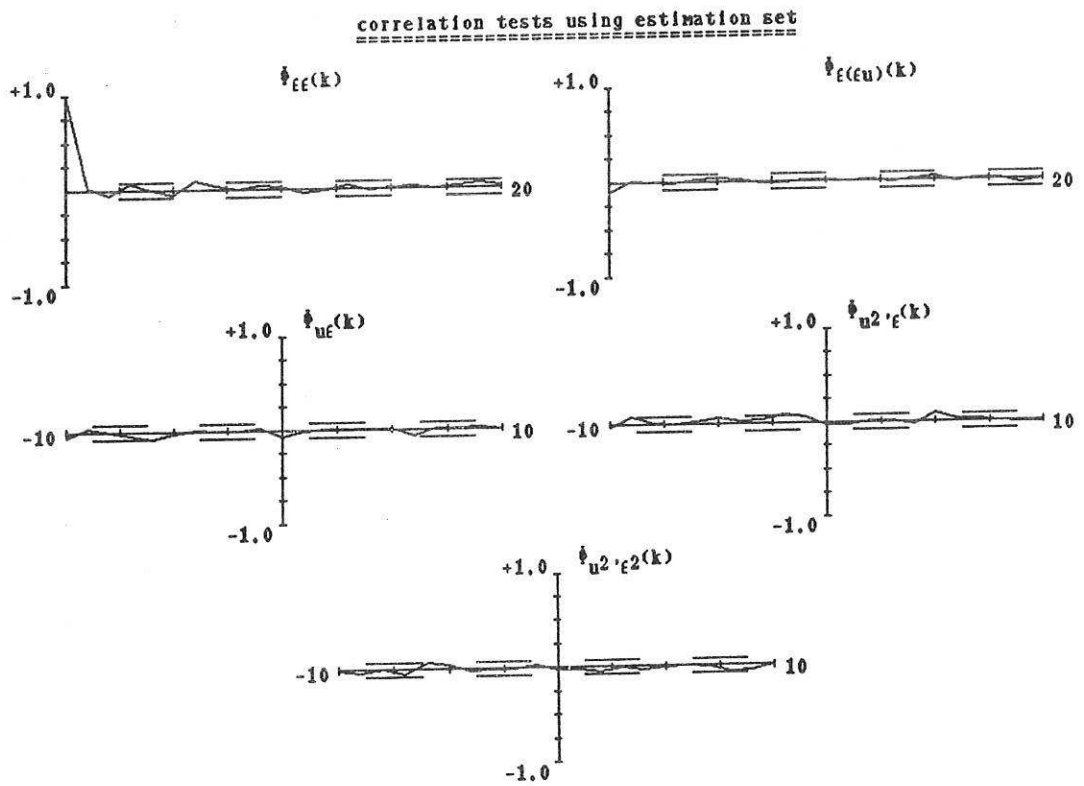
Fig.6. Chi-square statistical tests (final model for Example 1)

chi-square correlation test

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

correlation using estimation set

correlation using estimation and testing sets

chi-square correlation test

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

correlation using estimation set

correlation using estimation and testing sets

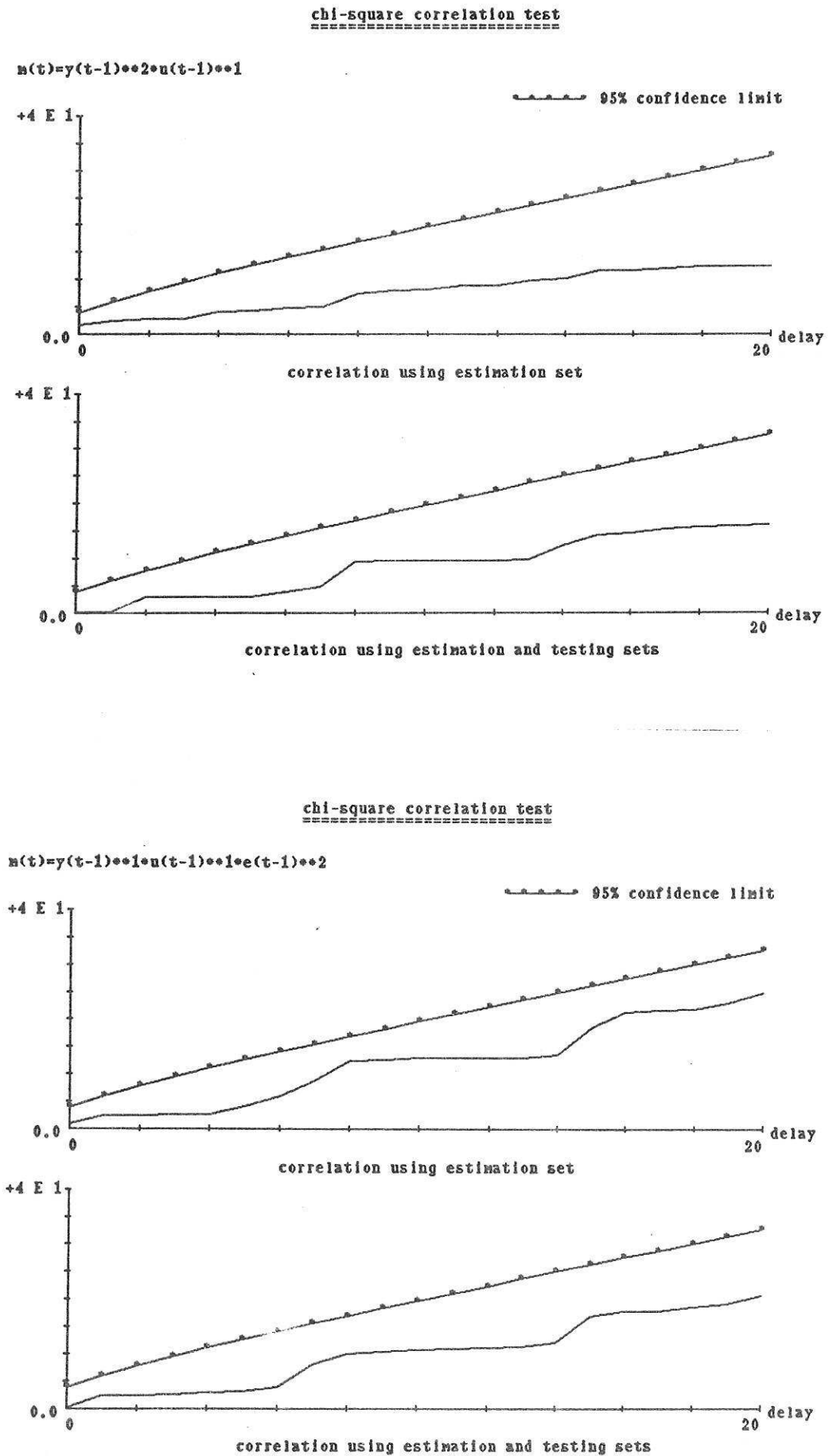Fig.7. System and estimated polynomial model of Example 2
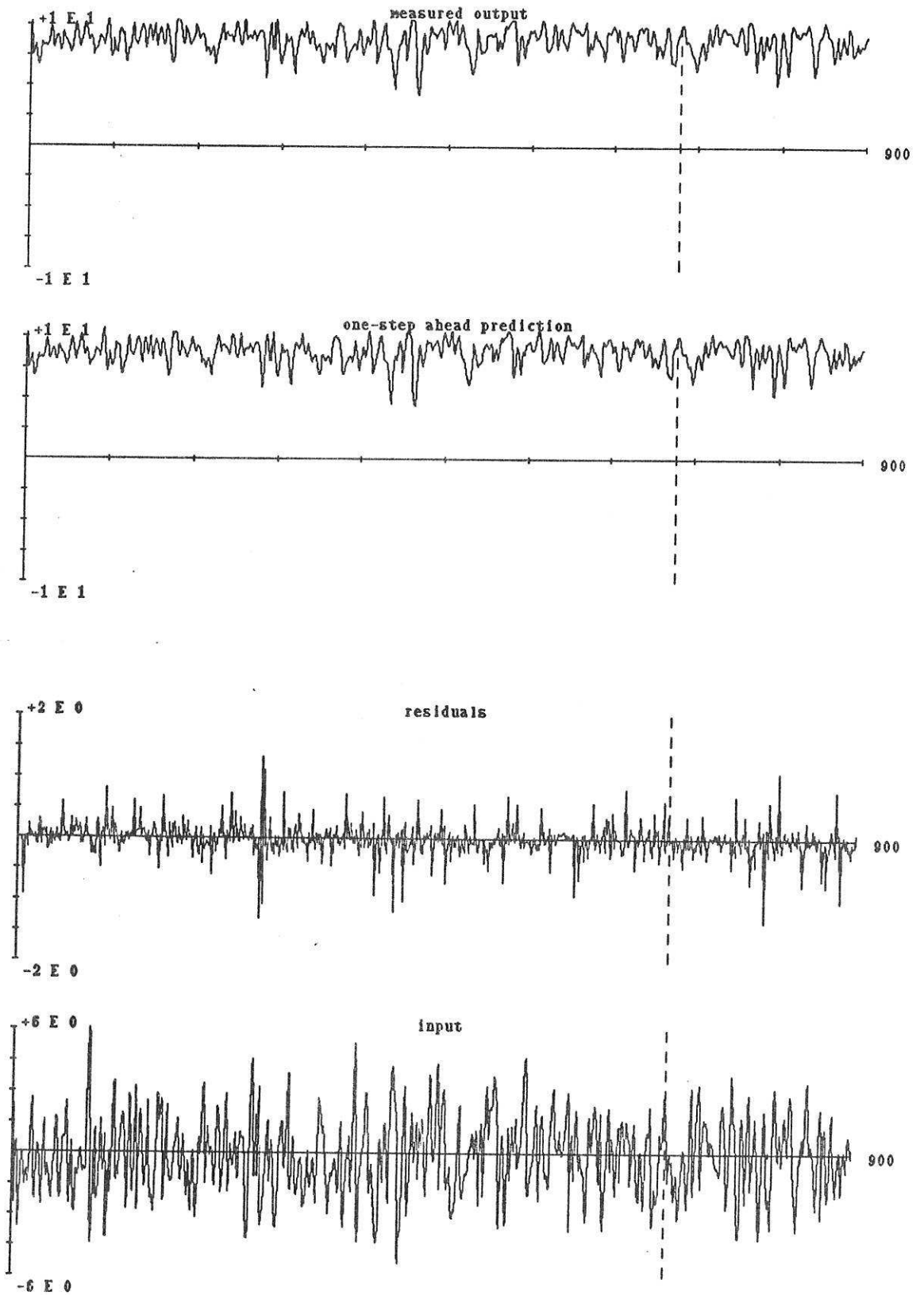
Fig.8. Simple correlation tests (polynomial model for Example 2)

correlation tests using estimation set

correlation tests using estimation and testing sets
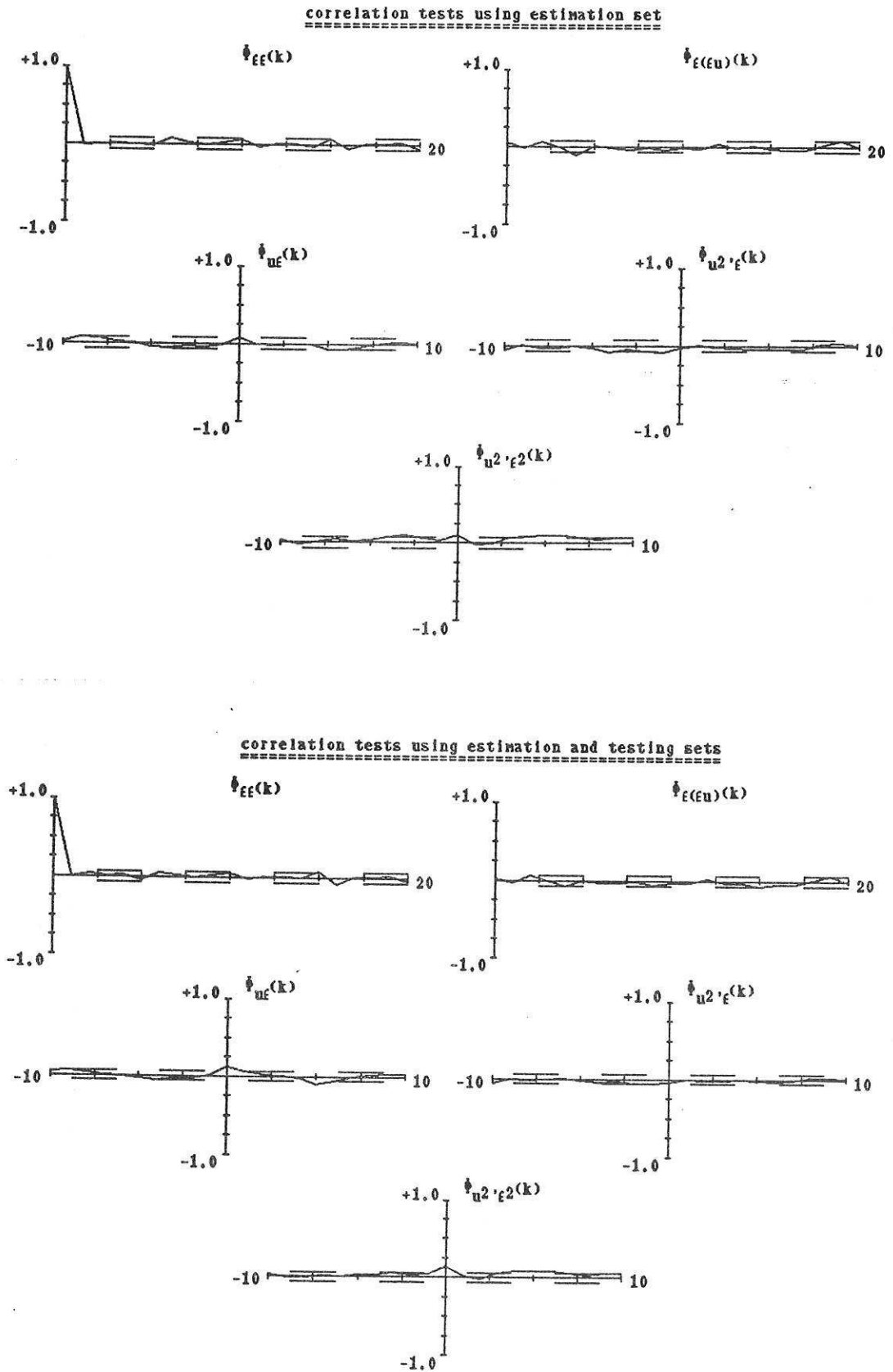
Fig.9. Chi-square statistical tests (polynomial model for Example 2)

chi-square correlation test

n(t)=y(t-1)**1

95% confidence limit

+4 E 1

0.0

0   delay   20

correlation using estimation set

+4 E 1

0.0

0   delay   20

correlation using estimation and testing sets

chi-square correlation test

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

95% confidence limit

+4 E 1

0.0

0   delay   20

correlation using estimation set

+4 E 1

0.0

0   delay   20

correlation using estimation and testing sets

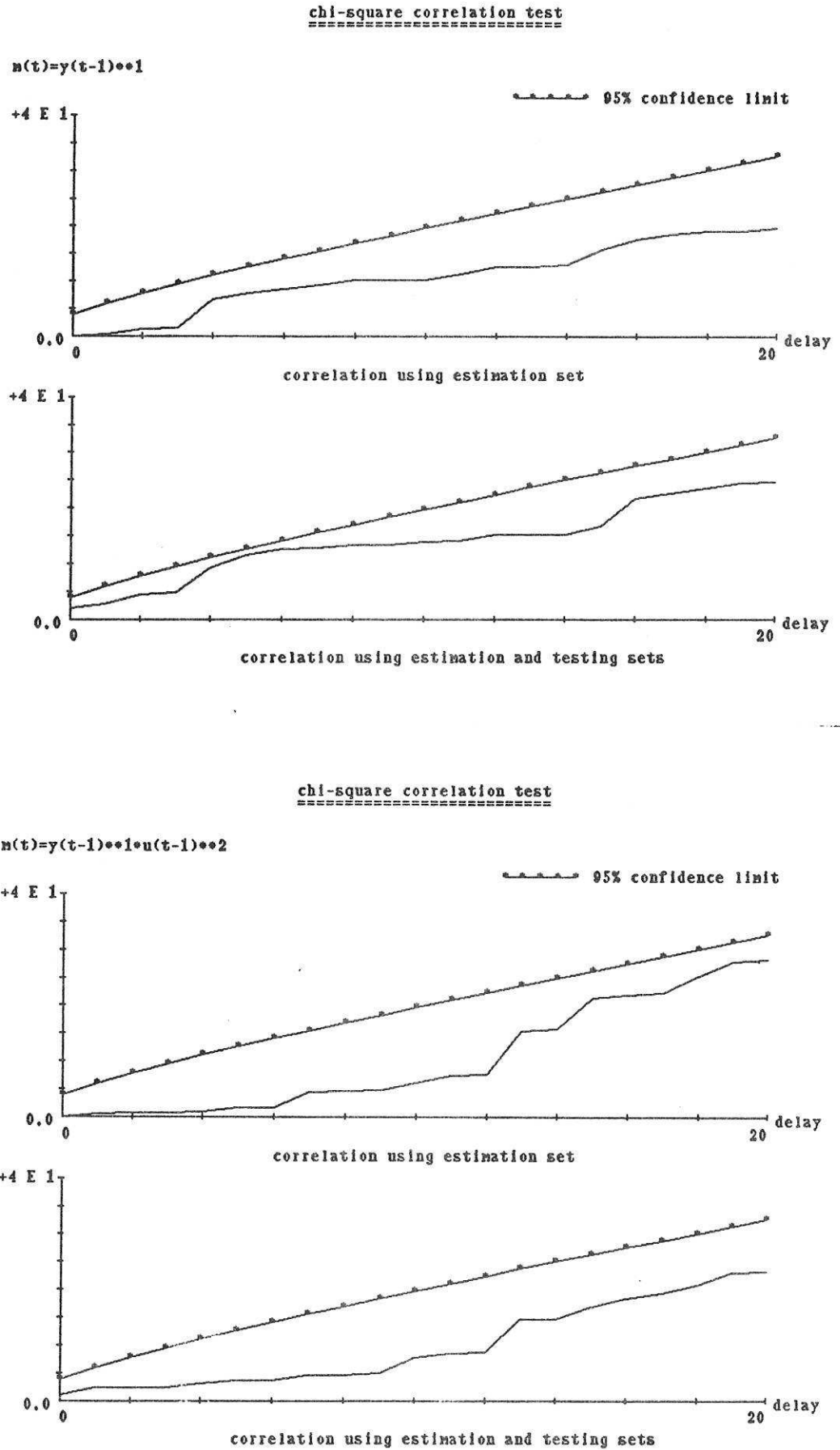**Fig.10. Predicted outputs and residuals of rational model (Example 2)**
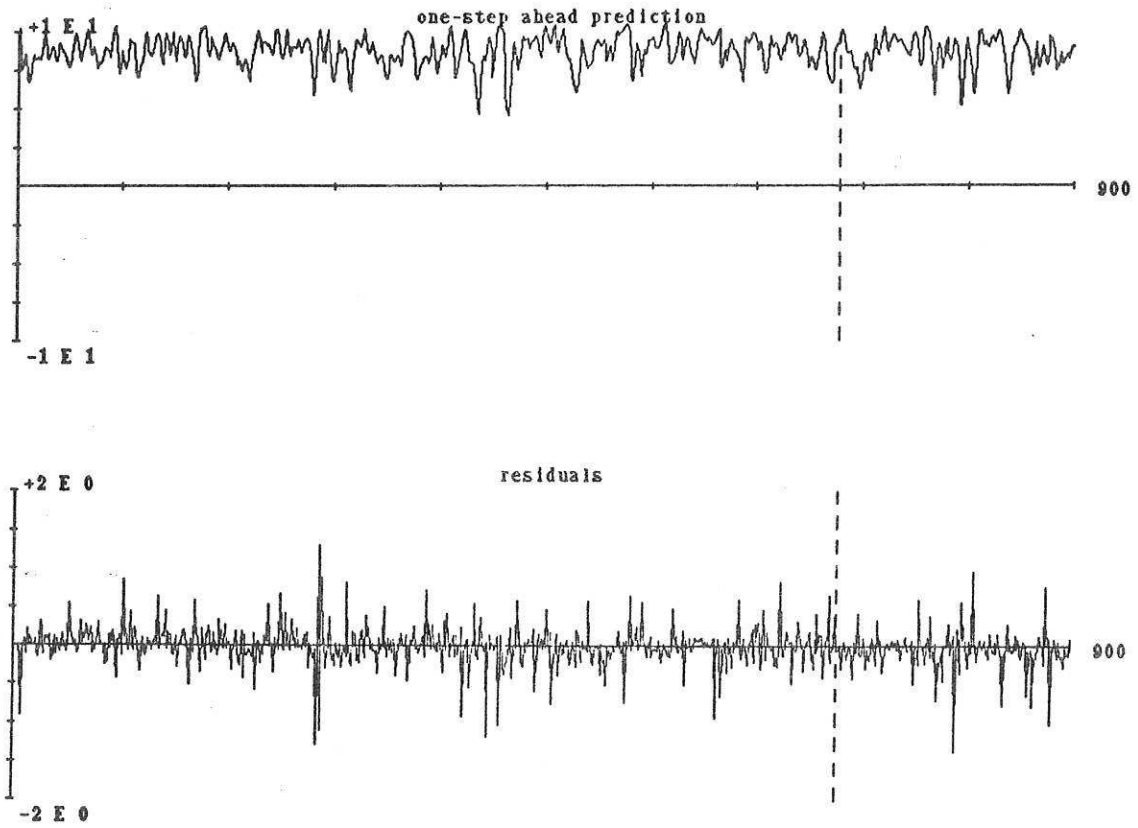
Fig.11. Simple correlation tests (rational model for Example 2)

correlation tests using estimation set

$\phi_{\epsilon\epsilon}(k)$

$\phi_{\epsilon(\epsilon u)}(k)$

$\phi_{u\epsilon}(k)$

$\phi_{u2'\epsilon}(k)$

$\phi_{u2'\epsilon2}(k)$

correlation tests using estimation and testing sets

$\phi_{\epsilon\epsilon}(k)$

$\phi_{\epsilon(\epsilon u)}(k)$

$\phi_{u\epsilon}(k)$

$\phi_{u2'\epsilon}(k)$

$\phi_{u2'\epsilon2}(k)$
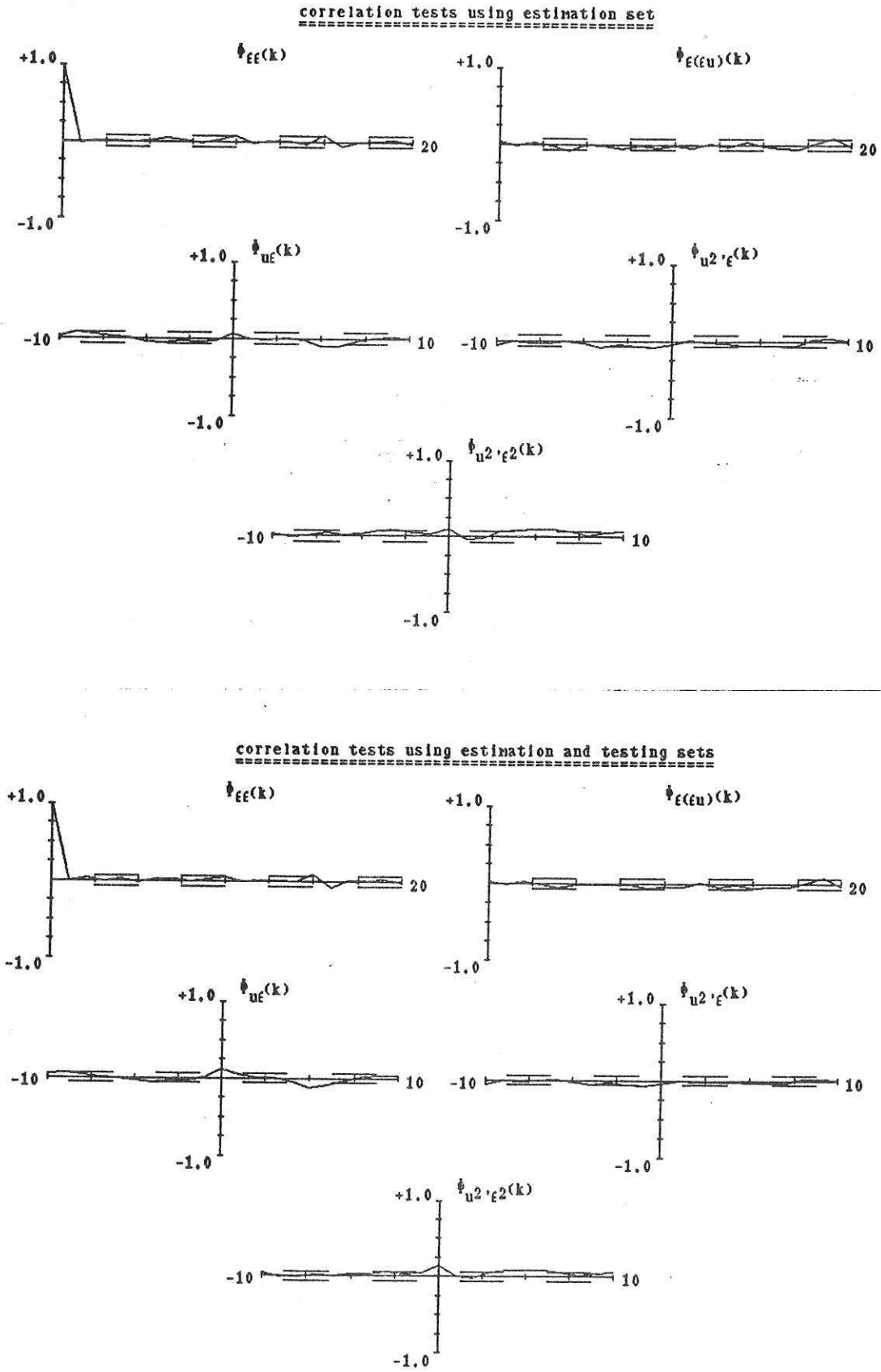
Fig.12. Chi-square statistical tests (rational model for Example 2)

chi-square correlation test

n(t)=y(t-1)**1

+4 E 1

95% confidence limit

0.0

0   20 delay

correlation using estimation set

+4 E 1

0.0

0   20 delay

correlation using estimation and testing sets

chi-square correlation test

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

+4 E 1

95% confidence limit

0.0

0   20 delay

correlation using estimation set

+4 E 1

0.0

0   20 delay

correlation using estimation and testing sets

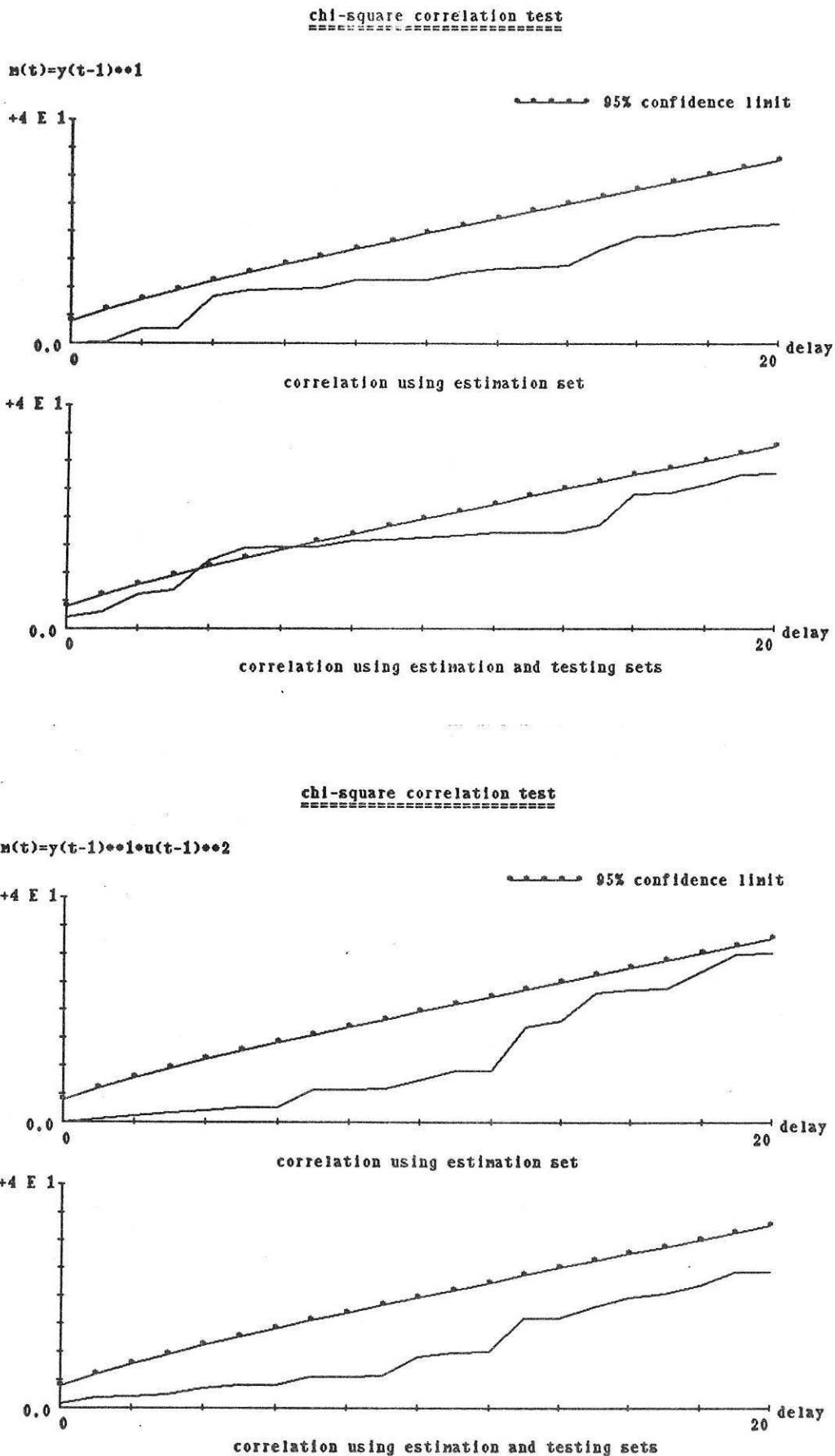Fig.13. Predicted outputs and residuals of linear model (Example 2)