Identification of multi-class linear and nonlinear systems

by

K. M. Tsang[†]

and

S. A. Billings[*]

[†]Department of Electrical Engineering
Hong Kong Polytechnic
Hung Hom
Kowloon
Hong Kong

[*]Department of Automatic Control and Systems Engineering
University of Sheffield
Sheffield S1 4DU
UK

Research Report No. 431

May 1991

Abstract

An algorithm for the identification of multi-class systems which can be described by a class of models over different operating regions is presented. The algorithm involves partitioning the raw data set using discriminant functions followed by parameter estimation. An orthogonal least squares algorithm coupled with a backward elimination procedure are employed for the parameter estimation and data partitioning processes. Provided the data elements are linearly separable, the proposed algorithm will correctly partition the data into the respective classes and parameter estimation algorithms can then be applied to estimate the models associated with each different class. Simulation studies are included to illustrate the algorithm.

1. Introduction

In the practical world, systems which cannot be described by just one global mathematical model or physical function are not uncommon and a set or class of models may provide a better description of these systems. The characteristics of piece-wise linear functions, hysteresis elements, coulomb friction elements are examples of systems which can be described by a class of models or functions. Fitting one global model to these elements may give unsatisfactory results and the fitted model may fail to capture the behaviour of the underlying system.

The usefulness of threshold models which can be viewed as a subset of multi-class systems has been demonstrated in the literature [1]. Nonlinear phenomena such as limit cycles, jump resonance, higher harmonics, subharmonics as well as hysteresis effects have been observed in simulated piece-wise linear models. For example, the characteristic of the anode current versus the grid voltage of a triode valve, which is a saturation type function, can be approximated by a set of piece-wise linear functions. Hysteresis effects, which turn up in many servo-systems such as gear chains, engines and motors, can also be described by some form of piece-wise functions [1]. Estimation algorithms are therefore required which can identify systems of this type.

The objective of this paper is to present a new algorithm for identifying multi-class systems. The first task is to correctly partition the data elements into their respective classes and then to apply parameter estimation to determine the model appropriate to each class. Provided the data captured can be correctly classified into the respective classes, and the different classes of models are linearly separable, models describing the data in each class can be easily identified using some of the well developed parametric identification algorithms [2,3]. Simulation studies are included to illustrate the concepts.

2. Partitioning of multi-class systems

For a well defined multi-class system, a decision rule is used to partition the space $\Omega \subset \mathbf{R}^m$ into n regions Ω_i, $i=1,2,\ldots,n$ where n is the number of classes in the system. An object $\boldsymbol{x} \in \mathbf{R}^m$ is classified as belonging to class i if it lies in the region Ω_i. Since the aim is to partition the space Ω containing the measurement vector $\boldsymbol{x}$ into n regions, each region should ideally contain objects from only a single class. The discriminant functions which are widely used in data analysis and partitioning can be adopted in multi-class system identification as a data classification tool.

Consider a well defined two-class system, if the two classes Ω_1 and Ω_2 are linearly separable, there exists a discriminant function $d(\boldsymbol{x})$ such that

$$\begin{aligned} d(\boldsymbol{x}) > k &\rightarrow \boldsymbol{x} \in \Omega_1 \\ d(\boldsymbol{x}) < k &\rightarrow \boldsymbol{x} \in \Omega_2 \end{aligned} \tag{1}$$

where $d(.)$ is a linear or nonlinear function operating on the variable $\boldsymbol{x}$ and k is a threshold value. For example, if $d(\boldsymbol{x})$ is a linear function,

$$d(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m \tag{2}$$

where the βs are some constant parameters or if $d(\boldsymbol{x})$ is an ℓ'th order polynomial type nonlinear function,

$$\begin{aligned} d(\boldsymbol{x}) = \beta_0 &+ \beta_1 x_1 + \ldots + \beta_m x_m \\ &+ \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \ldots \\ &\vdots \\ &+ \beta_{11\ldots1} x_1^{\ell} + \ldots \end{aligned} \tag{3}$$

The two-class problem of eqn.(1) can be extended to the general n-class case [4] as n-1 two-class decision surfaces or discriminant functions can be formed such that the first decision function $d_1(\boldsymbol{x})$ separates Ω_1 from $\Omega_2, \ldots, \Omega_n$; the second decision function $d_2(\boldsymbol{x})$ separates Ω_2 from $\Omega_3, \ldots, \Omega_n$; and so on as illustrated in Fig. 1.

2.1 Multi-class linear dynamical systems

Linear difference equations which are widely used in the modelling of linear dynamical systems can be used as a basis for constructing multi-class linear dynamical systems. Incorporating a set of discriminant functions with a set of linear difference equations produces a multi-class linear model. Hence extending the two-class system an n-class linear system can be described by n linear difference equations and n-1 discriminant functions

$$y(t) = f_n(\boldsymbol{x}; d(\boldsymbol{x})) \tag{4}$$

where $\boldsymbol{x} = [y(t-1) \ldots y(t-n_y)\ u(t-1) \ldots u(t-n_u)]$, $y(t)$, $u(t)$ are the system output and input, n_y and n_u are the maximum lags in the output and input respectively, $f_n(.)$ is an n-class linear function and $d(.)$ is a discriminant function. Expanding eqn.(4) gives

$$y(t) = \begin{cases} \theta_0^1+\theta_1^1 y(t-1)+\ldots+\theta_{n_y+n_u}^1 u(t-n_u); & \beta_0^1+\beta_1^1 y(t-1)+\ldots+\beta_{n_y+n_u}^1 u(t-n_u) > k_1 \\ \theta_0^2+\theta_1^2 y(t-1)+\ldots+\theta_{n_y+n_u}^2 u(t-n_u); & \beta_0^1+\beta_1^1 y(t-1)+\ldots+\beta_{n_y+n_u}^1 u(t-n_u) < k_1 \\ & \textit{and} \\ & \beta_0^2+\beta_1^2 y(t-1)+\ldots+\beta_{n_y+n_u}^2 u(t-n_u) > k_2 \\ \vdots & \\ \theta_0^n+\theta_1^n y(t-1)+\ldots+\theta_{n_y+n_u}^n u(t-n_u); & \beta_0^1+\beta_1^1 y(t-1)+\ldots+\beta_{n_y+n_u}^1 u(t-n_u) < k_1 \\ & \textit{and} \\ & \beta_0^2+\beta_1^2 y(t-1)+\ldots+\beta_{n_y+n_u}^2 u(t-n_u) < k_2 \\ & \vdots \\ & \beta_0^{n-1}+\beta_1^{n-1} y(t-1)+\ldots+\beta_{n_y+n_u}^{n-1} u(t-n_u) < k_{n-1} \end{cases}$$

where the superscripts on the parameter θ and β denote the class number of the associated model and k_i, i-1,...,n-1 are some constant values. Hence a two-class linear system with first order dynamics can be expressed as

$$y(t) = \begin{cases} \theta_0^1+\theta_1^1 y(t-1)+\theta_2^1 u(t-1); & \beta_0^1+\beta_1^1 y(t-1)+\beta_1^1 u(t-1) > k \\ \theta_0^2+\theta_1^2 y(t-1)+\theta_2^2 u(t-1); & \beta_0^1+\beta_1^1 y(t-1)+\beta_2^1 u(t-1) < k \end{cases}$$

2.2 Multi-class nonlinear systems

Recent results in approximation and realisation theory have produced nonlinear difference equation models that provide concise representations of nonlinear sampled data systems and which have been used as a basis for identification [5,6]. Extending these nonlinear difference models provides a foundation for building multi-class nonlinear systems. Incorporating a set of nonlinear difference equations with a set of discriminant functions produces an n-class nonlinear model which can be expressed as

$$y(t) = f_n^l(\boldsymbol{x}; d(\boldsymbol{x}))$$

$$= \begin{cases} \theta_0^1+\theta_1^1 y(t-1)+\ldots+\theta_{n_y+n_u}^1 u(t-n_u); & \beta_0^1+\beta_1^1 y(t-1)+\ldots+\beta_{n_y+n_u}^1 u(t-n_u) \\ \quad +\theta_{11}^1 y^2(t-1)+\ldots & \quad +\beta_{11}^1 y^2(t-1)+\ldots > k_1 \\ \vdots & \\ \theta_0^n+\theta_1^n y(t-1)+\ldots+\theta_{n_y+n_u}^n u(t-n_u); & \beta_0^1+\beta_1^1 y(t-1)+\ldots+\beta_{n_y n_u}^1 u(t-n_u) \\ \quad +\theta_{11}^n y^2(t-1)+\ldots & \quad +\beta_{11}^1 y^2(t-1)+\ldots < k_1 \\ & \vdots \\ & \beta_0^{n-1}+\beta_1^{n-1} y(t-1)+\ldots+\beta_{n_y+n_u}^{n-1} u(t-n_u) \\ & \quad +\beta_{11}^{n-1} y^2(t-1)+\ldots < k_{n-1} \end{cases} \qquad (5)$$

where $f_n^l(.)$ is an n-class nonlinear function of l degree of nonlinearity, $d(.)$ is a linear or nonlinear discriminant function, $\boldsymbol{x} = [y(t-1)\ldots y(t-n_y)\ u(t-1)\ldots u(t-n_u)]$, $y(t)$, $u(t)$ are the system output and input, and n_y, n_u are the maximum lags in the output and input respectively. A two-class nonlinear system with first order dynamics and a second order nonlinearity can for example be expressed as

$$y(t) = \begin{cases} \theta_0^1+\theta_1^1 y(t-1)+\theta_2^1 u(t-1)+\theta_{11}^1 y^2(t-1) & \beta_0^1+\beta_1^1 y(t-1)+\beta_2^1 u(t-1)+\beta_{11}^1 y^2(t-1) \\ \quad +\theta_{12}^1 y(t-1)u(t-1)+\theta_{22}^1 u^2(t-1); & \quad +\beta_{12}^1 y(t-1)u(t-1)+\beta_{22}^1 u^2(t-1) > k \\ \theta_0^2+\theta_1^2 y(t-1)+\theta_2^2 u(t-1)+\theta_{11}^2 y^2(t-1) & \beta_0^1+\beta_1^1 y(t-1)+\beta_2^1 u(t-1)+\beta_{11}^1 y^2(t-1) \\ \quad +\theta_{12}^2 y(t-1)u(t-1)+\theta_{22}^2 u^2(t-1); & \beta_{12}^1 y(t-1)u(t-1)+\beta_{22}^1 u^2(t-1) < k \end{cases}$$

3. System identification and data partitioning

Assuming that the discriminant functions $d(\boldsymbol{x})$ in eqns. (4) and (5) have known forms with unknown parameters, the design problem then becomes one of estimating the coefficients of $d(\boldsymbol{x})$ to optimise the partition rule eqn. (1). Redefine the two-class problem as

$$\begin{aligned} d(\boldsymbol{x}) = 0 &\rightarrow \boldsymbol{x} \in \Omega_1 \\ d(\boldsymbol{x}) \neq 0 &\rightarrow \boldsymbol{x} \in \Omega_x \end{aligned} \qquad (6)$$

That is the data elements are partitioned into two classes Ω_1 and $\Omega_x = \Omega_2 U \Omega_3 U \ldots U \Omega_n$, one which belongs to Ω_1 and one which does not belong to Ω_1. $d(\boldsymbol{x}) = 0$ then becomes the model describing the data in the subspace Ω_1. It is clear that the decision surface which arises from the discriminant function $d(\boldsymbol{x}) = 0$ is a hyperplane and its corresponding coefficients β can only be determined when the two classes can be separated by such a surface.

Define the misclassification rate for the data set in Ω_1 as the cost function

$$J_N(\theta_N) = \frac{1}{N} \sum_{t=1}^{N} \zeta^2(t;\theta_N) \tag{7}$$

where N is the number of data records for analysis, $\theta_N = [\theta_0\ \theta_1\ \ldots]^T$ is the estimated parameter vector for the model in Ω_1,

$$\zeta(t;\theta_N) = y(t) - \hat{y}(t;\theta_N) \tag{8}$$

is the prediction error of the fitted model in subspace Ω_1 and $\hat{y}(t;\theta_N)$ is the predicted output of the fitted model based on the estimated parameters θ_N. The problem of classification then becomes the minimisation of the cost function eqn.(7). From the definition of the partition rule eqn.(6), it is clear that only points which are misclassified and included in the estimation contribute to the cost function $J_N(\theta_N)$ and they do so to an extent proportional to the square of their distance from the hyperplane $d(\boldsymbol{x}) = 0$, the decision surface or in this case the correct model describing the data in Ω_1.

Reformulate the model describing the data elements in Ω_1 for eqns (4) or (5) to give the general representation

$$y(t) = \sum_{i=1}^{m} \theta_i p_i(t) + \zeta(t) \tag{9}$$

where θ_i, $i=1,\ldots,m$ represent the m real unknown parameters associated with the variables $p_i(t) = \{1, y(t-1), y(t-2), \ldots, y^2(t), \ldots\}$ and $\zeta(t)$ represents some modelling error. The objective of system identification and data partitioning is to estimate the unknown parameters in eqn.(9) while minimising the cost function eqn.(7). The orthogonal least squares estimation algorithm [7] which has been found to be an efficient procedure for identifying unknown linear and nonlinear systems will be used as the basis for the parameter estimation. The algorithm involves transforming eqn.(9) into an auxiliary equation

$$y(t) = \sum_{i=1}^{m} g_i w_i(t) + \zeta(t) \tag{10}$$

where g_i, $i=1,\ldots,m$ are some constant coefficients and $w_i(t)$, $i=1,\ldots,m$ are constructed to be orthogonal over the data records such that

$$\overline{w_j(t)w_i(t)} = 0 \quad for \quad j \neq i \tag{11}$$

where overbar $\overline{}$ denotes time average. Orthogonal data records can be constructed using the formula [7]

$$\begin{aligned} w_1(t) &= p_1(t) \\ w_i(t) &= p_i(t) - \sum_{j=1}^{i-1} \alpha_{ji} w_j(t) \quad , j < i \\ \alpha_{ji} &= \frac{\overline{w_j(t)p_i(t)}}{\overline{w_j^2(t)}} \quad , j = 1,\ldots,i-1 \end{aligned} \tag{12}$$

and the orthogonal parameters $\hat{g}_i$, $i=1,\ldots,m$ can be obtained according to the formula

$$\hat{g}_i = \frac{\overline{y(t)w_i(t)}}{\overline{w_i^2(t)}} \tag{13}$$

The original system parameters θ_i, $i=1,\ldots,m$ can then be recovered from eqn.(13) as

$$\begin{aligned} \theta_m &= \hat{g}_m \\ \theta_k &= \hat{g}_k - \sum_{i=k+1}^{m} \alpha_{ki}\hat{\theta}_i \quad , k=m-1,\ldots,1 \end{aligned} \tag{14}$$

The error reduction ratio [7],

$$err_i = \frac{\hat{g}_i^2 \overline{w_i^2(t)}}{\overline{y^2(t)}} \times 100 \quad , i=1,\ldots,m$$

which is a by product of the orthogonal least squares estimation algorithm, can provide information regarding the significance of variables $w_i(t)$, $i=1,\ldots,m$ in the system model. A large value for the error reduction ratio indicates the significance of the associated variable and err can therefore can be used as a structure detection tool. Combining the orthogonal estimator and the error reduction ratio test into a forward regression procedure gives a powerful estimation algorithm for system identification.

Full details are given elsewhere [7]. The sum of the error reduction ratios can also act as an indicator of how close the fitted model is to the original system model. If the sum of the error reduction ratios is close to 100, this may be a good indication that the fitted model is very close to the true model of the system because almost all the output power has been captured by the fitted model.

For large N,

$$\begin{aligned}
\frac{1}{N}\sum_{t=1}^{N} y(t)w_i(t) &\simeq \frac{1}{N-1}\sum_{t=1}^{N-1} y(t)w_i(t) &&\simeq \overline{y(t)w_i(t)} \\
\frac{1}{N}\sum_{t=1}^{N} w_k(t)p_i(t) &\simeq \frac{1}{N-1}\sum_{t=1}^{N-1} w_k(t)p_i(t) &&\simeq \overline{w_k(t)p_i(t)} \\
\frac{1}{N}\sum_{t=1}^{N} w_i^2(t) &\simeq \frac{1}{N-1}\sum_{t=1}^{N-1} w_i^2(t) &&\simeq \overline{w_i^2(t)}
\end{aligned} \tag{15}$$

such that

$$\hat{G}_N \simeq \hat{G}_{N-1} \tag{16}$$

where the subscript on $\hat{G}$ denotes the number of data points used for the estimation and $\hat{G} = [\hat{g}_1 \ \ldots \ \hat{g}_m]^T$. For large N, assuming that the small change in the estimated parameters from $\hat{G}_N$ to $\hat{G}_{N-1}$ is so small that they are virtually the same and the difference between the prediction error terms will be insignificant such that

$$\zeta(t;\hat{G}_N) \simeq \zeta(t;\hat{G}_{N-1}) \ , \quad t=1,\ldots,N-1 \tag{17}$$

where

$$\zeta(t;\hat{G}) = y(t) - \sum_{i=1}^{m} \hat{g}_i w_i(t)$$

The misclassification rate or the cost function of eqn.(7) can therefore be written as

$$J_N(\hat{G}) = \frac{(N-1)J_{N-1}(\hat{G}) + \zeta^2(N;\hat{G})}{N}$$

or

$$J_{N-1}(\hat{G}) = J_N(\hat{G}) - \frac{\zeta^2(N;\hat{G}) - J_N(\hat{G})}{N-1} \tag{18}$$

The objective of data classification is to minimise the misclassification rate eqn.(18) after some selection or elimination processes. If the data sequence is rearranged such that the measurement which contributes the maximum cost is positioned at N

$$|\zeta(N;\hat{G})| > |\zeta(t;\hat{G})| \ , \quad t=1,\ldots,N-1 \tag{19}$$

then it is possible to consider the stepwise elimination process of excluding the measurement which contributes $\zeta(N;\hat{G})$ from the estimation. That is the measurement $\boldsymbol{x}$ which contributes $\zeta(N;\hat{G})$ is classified as not belonging to the class Ω_1 compared to the rest of the measurements. From eqn.(19)

$$N\zeta^2(N;\hat{G}) > \sum_{t=1}^{N} \zeta^2(t;\hat{G})$$

or

$$\zeta^2(N;\hat{G}) > J_N(\hat{G}) \tag{20}$$

Since $J_N(\hat{G})$ and $J_{N-1}(\hat{G})$ are always real and positive and the result of eqn.(20) indicates that the term $\dfrac{\zeta^2(N;\hat{G}) - J_N(\hat{G})}{N-1}$ in eqn.(18) is a positive real number, this induces

$$J_{N-1}(\hat{G}) < J_N(\hat{G}) \tag{21}$$

Hence if this elimination process is repeated, the cost function should be monotonically decreasing. If the system under investigation is a well defined two-class system and the two classes are linearly separable, then in a finite number of steps, the cost function will go to zero after all the misclassified data have been excluded from the estimation and the estimated model will converge to the true model describing the data in the subspace Ω_1.

After initial fitting, data $\boldsymbol{x} \in \Omega_x$ not falling on the hyperplane described by the model

$$y(t) - \sum_{i=1}^{m} \hat{g}_i w_i(t) - \sum_{i=1}^{m} \theta_i p_i(t) \quad \forall \quad \boldsymbol{x} \in \Omega_1$$

will undergo the same parameter estimation and data partition processes again and this will result in a model describing the data elements in Ω_2. If the system is of the multi-class, this operation can continue indefinitely until the data can no longer be partitioned.

4. Stochastic systems

If the multi-class system under investigation is stochastic, the discriminant function is less well defined and the linearly separable property may no longer hold. There may then be areas of overlapping hyperplane between different classes. For low signal to noise ratios, the separation of classes is poor because there will be a large overlap of areas between classes. However, if the signal to noise ratio is high, the overlapping areas will be small and a large number of data records can still be partitioned and classified with some modifications to the discriminant function of eqn.(6).

For stochastic processes, the misclassification rate or the cost function of

eqn.(6) will not go to zero and the selection rule $d(\boldsymbol{x}) = 0$ can no longer be used. However with the proposed elimination procedure, the convergence properties of the cost function eqn.(21) should still be valid. If the system is stochastic, the cost function will not converge to zero and the data elimination process will continue indefinitely. Hence a new stopping rule has to be devised. The Akaike Information Criterion (AIC) [8], which is widely used in statistical model identification, can be readily applied as an indicator for the termination of the data elimination process. Consider the AIC criterion

$$AIC = N \log(\hat{e}^2) + 2n_\theta \tag{22}$$

where $\hat{e}^2$ and n_θ are the estimated variance of the prediction error sequence and number of parameters in the fitted model respectively. Instead of investigating the model structure of the estimated system with AIC, the number of data points falling within the subspace Ω_i, $i=1,2,\ldots,$ is being investigated. Therefore, in the stepwise elimination process, if there is no further improvement in the AIC value, the data measurements remaining in the parameter estimation and data elimination processes are considered to belong to the same class and the elimination process is stopped. That is

$$\text{if } AIC_j < AIC_{j+1}, \text{ then stop.} \tag{23}$$

The subscript on AIC in eqn.(23) denotes the stage of the data elimination process or the number of data records that have been eliminated. As the cost function will not converge to zero, the discriminant function of eqn.(6) has to be modified. Consider a new discriminant function for stochastic systems

$$\begin{aligned} |d(\boldsymbol{x})| &\le k \rightarrow \boldsymbol{x} \in \Omega_1 \\ |d(\boldsymbol{x})| &> k \rightarrow \boldsymbol{x} \in \Omega_x \end{aligned} \tag{24}$$

Equation (24) explains that if the absolute value of the discriminant function for a measurement vector $\boldsymbol{x}$ is greater than a certain threshold k, then this element is considered as not belonging to the family and excluded from the estimation. In this case, the discriminant function will be defined as

$$d(\boldsymbol{x}) = y(t) - \sum_{i=1}^{m} \hat{g}_i w_i(t) = \zeta(t;\hat{G})$$

or

$$d(\boldsymbol{x}) = y(t) - \sum_{i=1}^{m} \theta_i p_i(t) = \zeta(t;\theta) \tag{25}$$

The threshold k in eqn. (24) will therefore be the absolute value of the largest prediction error obtained from eqn.(25) for data measurements falling within Ω_1.

5. Algorithm for the identification of multi-class systems

The algorithm for the identification of multi-class linear and nonlinear systems can be summarised as follows.

a.) Select n_u and n_y. Select l if the system is nonlinear and n_e if the system is stochastic where n_e is the order of the noise term.

b.) Use the orthogonal estimation algorithm, eqns.(12) and (13) or the forward regression algorithm [7] to fit an initial model to the whole data set.

c.) With the estimated parameter vector $\hat{G}$, calculate the prediction error sequence

$$\zeta(t;\hat{G}) = y(t) - \sum_{i=1}^{m} \hat{g}_i w_i(t) \tag{26}$$

d.) Sort through the prediction error sequence $\zeta(t;\hat{G})$, $t=1,2,3,\ldots$. Identify the location of the data record which contributes the maximum cost to the misclassification rate eqns.(7) or (18) and eliminate it from the estimation process.

e.) Reduce the number of data for estimation by 1 and repeat procedures b.), c.) and d.) until the cost function of eqn.(7) or (18) goes to zero. Or in the case of stochastic systems, there is no further improvement in the AIC function eqn.(22).

f.) Use eqn.(14) to reconstruct the actual system model and the corresponding discriminant function.

g.) Procedures a.), b.), c.), d.), e.) and f.) can be re-applied to process the excluded data records (data in Ω_x) until no further classification or partitioning can be obtained. Actually, data in Ω_x can be preprocessed before repeating the procedures. The data can be partition into two separate data segments Ω_a and Ω_b first using the discriminant function obtained in f.) such that

$$\begin{aligned} d(\boldsymbol{x}) > 0 &\rightarrow \boldsymbol{x} \in \Omega_a \\ d(\boldsymbol{x}) < 0 &\rightarrow \boldsymbol{x} \in \Omega_b \end{aligned} \tag{27}$$

because the hyperplane described by the discriminant function $d(\boldsymbol{x}) = 0$ partitioned the excluded data set into two classes. If the system is stochastic, the partition rule then becomes

$$\begin{aligned} d(\boldsymbol{x}) > k &\rightarrow \boldsymbol{x} \in \Omega_a \\ d(\boldsymbol{x}) < -k &\rightarrow \boldsymbol{x} \in \Omega_b \end{aligned} \tag{28}$$

6. Illustrated examples

The operation and the effectiveness of the proposed algorithm are best illustrated by examples. Three simulated examples are included. These consist of the identification of a multi-class linear system, the identification of a multi-class nonlinear system and the identification of a multi-class stochastic system.

6.1 Deterministic multi-class linear system

Consider a piece-wise linear system (S_1) described by the equation

$$y(t) = \begin{cases} u(t-1) + 0.5y(t-1) - 0.5 , & y(t-1) > 1 \\ u(t-1) , & |y(t-1)| \le 1 \\ u(t-1) + 0.5y(t-1) + 0.5 , & y(t-1) < -1 \end{cases} \tag{29}$$

A zero mean Gaussian white noise of variance 1, $u(t) \sim N(0,1)$, was used to excite the model and 1000 pairs of input-output data records were collected for the identification of the system. Figure 2 shows the input-output data. Using the orthogonal forward regression algorithm, a linear first order model was initially fitted to the 1000 data records and the model was found to be

$$\begin{matrix} y(t) = & 0.004124 & + 0.161941y(t-1) & + 1.00516u(t-1) \\ & (0.00155) & (2.61167) & (96.1864) \end{matrix} \tag{30}$$

where the numbers in brackets underneath the equation denote the error reduction ratio of the associated variables. The first 50 model predicted outputs of eqn.(30) are shown in Fig. 3. Clearly, a rather poor output prediction of the system was obtained. This was further supported by the sum of the error reduction ratios which was equal to 98.7996. Because the system is noise free this indicates that around 1.2% of the output power was not captured by the fitted linear model eqn.(30). The cost function of this fitted linear model was equal to 0.0131515. When a nonlinear model of first order dynamics and fifth order nonlinearity was fitted to the data, Fig. 4 shows that a far superior model predicted output was obtained compared to the fitted linear model of eqn.(30). The corresponding fitted nonlinear model was given as

$$\begin{matrix} y(t) = & - 0.04472y(t-1) & + 0.99980u(t-1) & + 0.099386y^3(t-1) \\ & (0.0331) & (96.1864) & (3.5627) \\ & - 0.00647y^5(t-1) & & \\ & (0.1790) & & \end{matrix} \tag{31}$$

The performance of the nonlinear model compared to the linear model was confirmed by the sum of the error reduction ratio of 99.9612 and the small cost of 0.00042517. For the nonlinear model, the unmodelled output power was less than 0.04% which was far less than that obtained for the linear model which was around 1.2%.

Using the new multi-class algorithm, with the best linear model eqn.(30) as the initial model produced the parameter vector profiles and corresponding cost function shown in Figs. 5 and 6 respectively. The cost function is converging and the change of parameter vector at each iteration is small. After 354 iterations, 354 data elements were eliminated from the partition and estimation processes, and the cost function was reduced to zero (Fig. 6). This clearly demonstrated that all the misclassified data records that were initially included in the estimation had been correctly eliminated and the remaining data records all belonged to the same class. The final estimated model for the 646 data records was given by

$$y(t) = u(t-1) \quad \forall \boldsymbol{x} \in \Omega_1 \tag{32}$$

(100)

and the corresponding discriminant function was

$$d(\boldsymbol{x}) = y(t) - u(t-1) = 0 \rightarrow \boldsymbol{x} \in \Omega_1 \tag{33}$$

Procedure g.) was then followed. Using the fitted model of eqn.(32), a prediction error sequence

$$\zeta(t;\theta) = y(t) - u(t-1) \tag{34}$$

was evaluated and the 354 excluded data records were partitioned into two groups according to the partition rule of eqn.(27).

$$\begin{aligned} d(\boldsymbol{x}) - \zeta(t;\theta) > 0 &\rightarrow \boldsymbol{x} \in \Omega_a \\ d(\boldsymbol{x}) - \zeta(t;\theta) < 0 &\rightarrow \boldsymbol{x} \in \Omega_b \end{aligned} \tag{35}$$

After this partition process, 170 data records were classified to Ω_a and 184 data records were classified to Ω_b. Procedures a.) to f.) were then re-applied to the data records in segments Ω_a and Ω_b. Analysis of the data records in these segments revealed that they could not be further partitioned because both of cost functions resulting from the initial parameter estimation were equal to zero. The two identified models were given by

$$
\begin{aligned}
y(t) &= u(t-1) + 0.5y(t-1) - 0.5 \quad \forall \boldsymbol{x} \in \Omega_a \\
&\quad (88.939) \quad (9.364) \quad (1.697) \\
y(t) &= u(t-1) + 0.5y(t-1) + 0.5 \quad \forall \boldsymbol{x} \in \Omega_b \\
&\quad (90.818) \quad (7.636) \quad (1.546)
\end{aligned} \tag{36}
$$

Substituting the values of y(t) obtained in eqn.(36) into the discriminant function eqn.(35) produces a new set of discriminant function

$$
\begin{aligned}
y(t-1) > 1 &\rightarrow \boldsymbol{x} \in \Omega_a \\
y(t-1) < -1 &\rightarrow \boldsymbol{x} \in \Omega_b
\end{aligned} \tag{37}
$$

Equations (32), (33), (35), (36) and (37) describe the characteristic of the system. Combining the discriminant functions eqns.(33) and (37) with the fitted linear models of eqn.(32) and (36) gives the identified multi-class linear system

$$
y(t) = \begin{cases} u(t-1) + 0.5y(t-1) - 0.5 , & y(t-1) > 1 \\ u(t-1) , & |y(t-1)| \le 1 \\ u(t-1) + 0.5y(t-1) + 0.5 , & y(t-1) < -1 \end{cases} \tag{38}
$$

which clearly coincides with the true characteristic of the original system and the predicted output of this fitted model should therefore coincide with the original system. The superiority of this fitted piece-wise linear model over the fitted linear and nonlinear model can also be confirmed by the sum of error reduction ratios of a 100 and a cost function of zero.

6.2 Deterministic multi-class nonlinear system

Consider a two-class nonlinear system (S_2) described by the equation

$$
y(t) = \begin{cases} 0.8y(t-1)+u(t-1) & ; \quad 0.3y(t-2)+0.2y^2(t-1) \\ +0.2y^2(t-1)-0.5u^2(t-1) & \quad -0.5u^2(t-1) \le 1 \\ \\ 2+0.5y(t-1)+u(t-1) & ; \quad 0.3y(t-1)+0.2y^2(t-1) \\ & \quad -0.5u^2(t-1) > 1 \end{cases} \tag{39}
$$

A zero mean Gaussian white noise of variance 1 was used to excite the system, and 1000 pairs of input-output data records were collected. Figure 7 shows the input-output records.

A linear first order model was initially fitted to the 1000 data records and the model was found to be

$$\begin{matrix} y(t) = -0.05403 + 0.64400y(t-1) + 0.96727u(t-1) \\ \qquad\quad (0.0752) \qquad (41.3878) \qquad\quad (25.5192) \end{matrix} \tag{40}$$

The cost function for this fitted model was 1.27359. The model predicted output for this fitted linear model is shown in Fig. 8 which indicates that the model is very poor. The poor performance of this model is undoubtedly due to the fact that only 66.7822% of the total system output power has been captured.

When a nonlinear model of first order dynamics and second order nonlinearity was fitted to the data, the resulting estimate was given as

$$\begin{matrix} y(t) = \quad 0.08786 \quad + \quad 0.61125y(t-1) \quad + \quad 1.01352u(t-1) \\ \qquad\quad (0.1117) \qquad\qquad (41.3878) \qquad\qquad (25.5192) \\ +0.09238y^2(t-1) - 0.00448y(t-1)u(t-1) - 0.47827u^2(t-1) \\ \qquad (16.3821) \qquad\qquad (0.0017) \qquad\qquad (11.9436) \end{matrix} \tag{41}$$

The cost function for this fitted nonlinear model was reduced to 0.179511 and a far superior output prediction (Fig. 9) was obtained. The sum of the error reduction ratios reached 95.3462 and the cost function was far less than for the linear model.

Taking eqn.(41) as an initial estimate for the proposed multi-class identification algorithm, Figs. 10 and 11 show profiles of the cost function and the parameter vector for the first 153 iterations operating on the initial data set. After 153 data records had been eliminated from the estimation, the cost function reduced to zero and the sum of the error reduction ratios reached 100. This clearly indicated that the remaining data measurements belonged to the same class and all the misclassified data had been eliminated. The final estimate for the 847 data records was given as

$$\begin{matrix} y(t) = 0.8y(t-1) + \quad u(t-1) \quad + 0.2y^2(t-1) - 0.5u^2(t-1) \\ \qquad\quad (10.7752) \quad (42.5885) \quad\quad (3.5994) \quad\quad (43.0369) \end{matrix} \tag{42}$$

and the discriminant function was given as

$$d(\boldsymbol{x}) = y(t) - 0.8y(t-1) - u(t-1) - 0.2y^2(t-1) + 0.5u^2(t-1) = 0 \rightarrow \boldsymbol{x} \in \Omega_1 \tag{43}$$

A noise sequence was then formed based on eqn.(42)

$$\zeta(t;\theta) = y(t)-0.8y(t-1)-u(t-1)-0.2y^2(t-1)+0.5u^2(t-1) \tag{44}$$

and the 153 eliminated data were partitioned into two parts according to the rule

$$\begin{aligned} d(\boldsymbol{x}) - \zeta(t;\theta) > 0 &\rightarrow \boldsymbol{x} \in \Omega_a \\ d(\boldsymbol{x}) - \zeta(t;\theta) < 0 &\rightarrow \boldsymbol{x} \in \Omega_b \end{aligned} \tag{45}$$

According to eqn.(45), 3 data records fell back to Ω_1, 37 were classified to Ω_a and 113 were classified to Ω_b. The multi-class identification algorithm was then re-applied to the data records belonging to Ω_a and Ω_b. Analysis of the records revealed that they could not be further partitioned because both of the cost functions for the initial estimates became equal to zero. The two identified models were given by

$$\begin{aligned} y(t) = &\underset{(45.4046)}{2} + \underset{(41.5567)}{0.5y(t-1)} + \underset{(13.0387)}{u(t-1)} \quad \forall \boldsymbol{x} \in \Omega_a \\ y(t) = &\underset{(21.9168)}{2} + \underset{(73.5808)}{0.5y(t-1)} + \underset{(4.5024)}{u(t-1)} \quad \forall \boldsymbol{x} \in \Omega_b \end{aligned} \tag{46}$$

and the sum of the error reduction ratios for both model were equal to 100. Substituting eqn.(46) into the discriminant function of eqn.(45) produces

$$\begin{aligned} 0.3y(t-1)+0.2y^2(t-1)-0.5u^2(t-1) < 2 &\rightarrow \boldsymbol{x} \in \Omega_a \\ 0.3y(t-1)+0.2y^2(t-1)-0.5u^2(t-1) > 2 &\rightarrow \boldsymbol{x} \in \Omega_b \end{aligned} \tag{47}$$

Since both of the models fitted to the segments Ω_a and Ω_b were exactly the same, they could be combined together forming a single class $\Omega_2 = \Omega_a \cup \Omega_b$ and the discriminant function of eqn.(47) was modified to

$$0.3y(t-1)+0.2y^2(t-1)-0.5u^2(t-1) \neq 2 \rightarrow \boldsymbol{x} \in \Omega_2 \tag{48}$$

Combining the discriminant functions of eqn.(43) and (48) with the fitted models of eqn.(42) and (46) gives the two-class nonlinear model

$$y(t) = \begin{cases} 0.8y(t-1)+u(t-1) & ; \quad 0.3y(t-1)+0.2y^2(t-1) \\ +0.2y^2(t-1)-0.5u^2(t-1) & \quad -0.5u^2(t-1) - 2 \\ \\ 2+0.5y(t-1)+u(t-1) & ; \quad 0.3y(t-1)+0.2y^2(t-1) \\ & \quad -0.5u^2(t-1) \neq 2 \end{cases} \tag{49}$$

which clearly coincides with the true characteristic of the original system. The superiority of this fitted multi-class nonlinear model over the fitted single-class linear model and the single class nonlinear model was confirmed by the sum of the error reduction ratios which was equal to 100 and the cost function which was equal to zero.

6.2 Stochastic multi-class linear system

Consider the same multi-class linear system and the same data records described in section 6.1 but with the output of the system corrupted by some output additive noise, $e(t) \sim N(0,0.0001)$, which were uncorrelated with the input $u(t)$. The signal to noise ratio for this noise corrupted system was around 80dB. The corrupted output for the stochastic system (S_3) was defined as

$$z(t) = y(t) + e(t) \tag{50}$$

Substituting $y(t) = z(t) - e(t)$ into eqn.(29) gives

$$z(t) = \begin{cases} u(t-1)+0.5z(t-1)-0.5 & ; \quad z(t-1)-e(t-1) > 1 \\ \quad -0.5e(t-1)+e(t) & \\ \\ u(t-1) & ; \quad |z(t-1)-e(t-1)| \leq 1 \\ \\ u(t-1)+0.5z(t-1)+0.5 & ; \quad z(t-1)-e(t-1) < -1 \\ \quad -0.5e(t-1)+e(t) & \end{cases}$$

The best first order dynamical linear model with a first order noise model for the 1000 records was given by

$$\begin{aligned} z(t) = & \underset{(0.00128)}{0.003744} + \underset{(2.61251)}{0.161437}z(t-1) + \underset{(96.179)}{1.00484}u(t-1) \\ & +\underset{(0.00225)}{0.043524}\zeta(t-1) + \zeta(t) \end{aligned} \tag{51}$$

The sum of the error reduction ratio for this fitted linear model was equal to 98.795 and the cost function was equal to 0.0131968. Figure 12 shows the first 50 model predicted outputs of the fitted linear model of eqn.(51).

Using the proposed multi-class algorithm and following similar procedures as described in sections 6.1 and 6.2, profiles of the cost function, AIC and the parameter vector for the first 354 iterations are shown in Figs. 13, 14 and 15 respectively. The cost function was monotonically decreasing while the AIC reached a minimum when 321 data records were eliminated. The AIC of Fig. 14 illustrates that there was no further improvement in the information content of the fitted model even when more data records were eliminated. The profile of the parameter vector (Fig. 15) also indicates that there was no significant differences within the parameter vector even if more than 321 data records were eliminated from the estimation. The final estimate for the 679 data records was found to be

$$\begin{aligned} z(t) = & \underset{(0.000020)}{-0.000443} + \underset{(0.000294)}{0.003070}z(t-1) + \underset{(99.9898)}{0.999879}u(t-1) \\ & -0.006762\zeta(t-1) + \zeta(t) \qquad \forall \boldsymbol{x} \in \Omega_1 \end{aligned} \tag{52}$$

$$(0.000050)$$

The cost function for this fitted linear model was equal to 0.0001049 and the corresponding AIC was equal to -6213. The sum of the error reduction ratios was equal to 99.9901. For the selected 679 data records, a prediction error sequence was formed based on eqn.(52) and the absolute value of the maximum prediction error term was found to be 0.03241. Combining eqn.(52) with the absolute value of the maximum prediction error term forms the discriminant function describing the data elements in Ω_1

$$\begin{aligned} d(\boldsymbol{x}) = & z(t)+0.000443-0.003070z(t-1)-0.999879u(t-1)+0.006762\zeta(t-1) \\ & -\zeta(t) \leq 0.03241 \rightarrow \boldsymbol{x} \in \Omega_1 \end{aligned} \tag{53}$$

The 321 eliminated data records were further partitioned into two separate groups according to the partition rule of eqn.(28).

$$\begin{aligned} d(\boldsymbol{x}) - \zeta(t) > 0.03241 &\rightarrow \boldsymbol{x} \in \Omega_a \\ d(\boldsymbol{x}) - \zeta(t) < -0.03241 &\rightarrow \boldsymbol{x} \in \Omega_b \end{aligned} \tag{54}$$

Partition rule eqn.(54) classified 157 data records to Ω_a and 164 data records to Ω_b. The parameter estimation and data elimination processes were then re-applied. Profiles of the cost function, AIC and the parameter vector for the first 50 iterations operating on data belonging to Ω_a are shown in Figs. 16, 17 and 18 respectively. Even though the cost function Fig. 16 was converging during the data elimination process, Fig. 17 shows that the AIC

reached a minimum when zero data records were eliminated. Hence all the 157 data records were classified to Ω_a and the fitted model was given by

$$\begin{aligned} z(t) = & \underset{(1.5737)}{-0.498066} + \underset{(10.4659)}{0.498635}z(t-1) + \underset{(87.9508)}{0.999249}u(t-1) \\ & \underset{(0.0018)}{-0.472959}\zeta(t-1) + \zeta(t) \end{aligned} \qquad \forall\ \boldsymbol{x} \in \Omega_a \quad (55)$$

The corresponding cost function and AIC were 0.000092 and -1451 respectively and the sum of the error reduction ratios was 99.9922.

Figures 19, 20 and 21 show profiles of the cost function, AIC and the parameter vector for the first 50 iterations operating on the 164 data records belonging to Ω_b. Figure 20 illustrates that the AIC reached a minimum when one data record was eliminated from the parameter estimation even though the cost function was decreasing monotonically. However, when the only selected record was eliminated from the estimation, the improvement in the AIC and the change in the estimated parameter vector were very small. Therefore, all 164 data records were classified to Ω_b and the fitted linear model to the 164 data records was given by

$$\begin{aligned} z(t) = & \underset{(1.3603)}{0.499744} + \underset{(8.8510)}{0.500375}z(t-1) + \underset{(89.7783)}{1.0001}u(t-1) \\ & \underset{(0.0020)}{-0.463664}\zeta(t-1) + \zeta(t) \end{aligned} \qquad \forall\ \boldsymbol{x} \in \Omega_b \quad (56)$$

The corresponding cost function and AIC were 0.000097 and -1507.5 respectively and the sum of the error reduction ratios was 99.9916.

Substituting the values of $z(t)$ obtained in eqns. (55) and (56) into the determinant function of eqn.(54) yields

$$\begin{aligned} &0.495565z(t-1)-0.00063u(t-1)-0.466197\zeta(t-1) > 0.530033 \rightarrow \boldsymbol{x} \in \Omega_a \\ &0.497305z(t-1)+0.000221u(t-1)-0.456902\zeta(t-1) < -0.532597 \rightarrow \boldsymbol{x} \in \Omega_b \end{aligned} \quad (57)$$

Combining the discriminant functions of eqns.(53) and (57) with the fitted linear models of eqns.(52), (55) and (56) gives the multi-class linear model

$$z(t) = \begin{cases} -0.498066 + 0.498635z(t-1) \\ +0.999249u(t-1) - 0.472959\zeta(t-1) \\ +\zeta(t) \end{cases} ; \quad 0.495565z(t-1) - 0.00063u(t-1) - 0.466197\zeta(t-1) > 0.530033$$

$$\begin{cases} 0.499744 + 0.500375z(t-1) \\ +1.0001u(t-1) - 0.463664\zeta(t-1) \\ +\zeta(t) \end{cases} ; \quad 0.497305z(t-1) + 0.000221u(t-1) - 0.456902\zeta(t-1) < -0.532597 \qquad (58)$$

$$\begin{cases} -0.000443 + 0.003070z(t-1) \\ +0.999879u(t-1) - 0.006762\zeta(t-1) \\ +\zeta(t) \end{cases} ; \quad \begin{cases} 0.495565z(t-1) - 0.00063u(t-1) \\ -0.466197\zeta(t-1) \le 0.530033 \\ \textit{and} \\ 0.497305z(t-1) + 0.000221u(t-1) \\ -0.456902\zeta(t-1) \ge -0.532597 \end{cases}$$

Comparing eqn.(58) with eqn.(29) shows the structure of the fitted model eqn.(58) is very close to the original. Figure 22 shows the first 50 model predicted outputs of the fitted multi-class linear model eqn.(58) superimposed on the actual output of the system.

The power of the output additive noise was increased by 40dB, such that $e(t) \sim N(0, 0.01)$ and the new multi-class identification algorithm was again applied to the 1000 collected records. The signal to noise ratio in this system reduced to around 40dB. Referring this system as S_4, the initial best fitted first order dynamical linear model with a first order noise term was given by

$$\begin{aligned} z(t) = {} & \underset{(0.00001)}{0.000380} + \underset{(2.57414)}{0.162093}z(t-1) + \underset{(95.3694)}{1.00361}u(t-1) \\ & \underset{(0.00842)}{-0.064780}\zeta(t-1) + \zeta(t) \end{aligned} \qquad (59)$$

The sum of the error reduction ratios for this fitted linear model was equal to 97.9519 and the cost function was 0.0225264. The first 50 model predicted output for this fitted linear model eqn.(59) was shown in Fig. 23. The application of the multi-class identification algorithm produced profiles of the cost function, AIC and the parameter vector shown in Figs. 24, 25 and 26 respectively. The AIC reached a minimum when 90 data records had been eliminated from the estimation and the fitted model for the selected 910 data records was given by

$$
\begin{aligned}
z(t) = \quad & -0.003519 \quad + 0.107152z(t-1) + 0.998038u(t-1) \\
& (0.00126) \qquad (0.86511) \qquad (9.9846) \\
& -0.041216\zeta(t-1) + \zeta(t) \\
& (0.00338)
\end{aligned}
\qquad \forall \boldsymbol{x} \in \Omega_1 \qquad (60)
$$

The cost function for this fitted linear model was equal to 0.0121633 and the corresponding AIC was equal to -4004.49. The sum of the error reduction ratio was equal to 98.8543. A prediction error sequence was then formed based on eqn.(60). For the selected 910 data records, the absolute value of the maximum prediction error term was found to be 0.254194. Combining eqn.(60) with the threshold value 0.254194 produced the discriminant function describing the data elements in Ω_1

$$
\begin{aligned}
d(\boldsymbol{x}) = {} & z(t)+0.003519-0.107152z(t-1)-0.998038u(t-1)+0.041216\zeta(t-1) \\
& -\zeta(t) \leq 0.254194 \rightarrow \boldsymbol{x} \in \Omega_1
\end{aligned}
\qquad (61)
$$

The 90 eliminated data records were further partitioned into two segments according to the partition rule

$$
\begin{aligned}
d(\boldsymbol{x}) - \zeta(t) > 0.254194 &\rightarrow \boldsymbol{x} \in \Omega_a \\
d(\boldsymbol{x}) - \zeta(t) < -0.254194 &\rightarrow \boldsymbol{x} \in \Omega_b
\end{aligned}
\qquad (62)
$$

Eqn. (62) classified 50 data records to Ω_a and 40 data records to Ω_b. The multi-class identification algorithm was then re-applied. Profiles of the cost function, AIC and the parameter vector for the first 20 iterations operating on data belonging to Ω_a and Ω_b are shown in Figs. 27, 28 and 29 respectively. Even though the cost functions were decreasing monotonically, the corresponding AIC fluctuated and the change in the parameter vector at each iteration was rather high. This was probably caused by the small number of data records in both estimation processes which meant the assumption on the consistency of the parameter vector was no longer valid. According to the selection rule, 1 data records would be eliminated from the estimation for data records belonging to Ω_a and 9 would be eliminated from Ω_b (Fig. 28). Since the number of records eliminated in both operations (according to eqn.(62)) were small and in each case the number of records was also small, the effects of including them in the estimation instead of eliminating them was investigated. This produced the models

$$
\begin{aligned}
z(t) = {} & 0.287318 + 0.160788z(t-1) + 1.00663u(t-1) \\
& (2.21366) \qquad (22.0097) \qquad (75.0844) \qquad \forall \boldsymbol{x} \in \Omega_a \\
& +0.161784\zeta(t-1) + \zeta(t) \\
& (0.02327)
\end{aligned} \tag{63}
$$

$$
\begin{aligned}
z(t) = {} & -0.309377 + 0.172679z(t-1) + 0.961068u(t-1) \\
& (16.6499) \qquad (3.12086) \qquad (79.4311) \qquad \forall \boldsymbol{x} \in \Omega_b \\
& -0.326336\zeta(t-1) + \zeta(t) \\
& (0.0941485)
\end{aligned}
$$

The sum of the error reduction ratios, the cost function and the AIC for the model fitted to Ω_a were 99.3311, 0.00907981 and -227.085 and 99.2959, 0.115988 and -170.224 for the model fitted to Ω_b respectively. Combining eqn.(63) with (62) produces a set of new discriminant functions

$$
\begin{aligned}
& 0.053636z(t-1)+0.008592u(t-1)+0.203\zeta(t-1) > -0.036643 \rightarrow \boldsymbol{x} \in \Omega_a \\
& 0.065527z(t-1)-0.03697u(t-1)+0.367552\zeta(t-1) < 0.055183 \rightarrow \boldsymbol{x} \in \Omega_b
\end{aligned} \tag{64}
$$

Combining the discriminant functions of eqns.(61) and (64) with the fitted linear model of eqns.(60) and (63) produces the multi-class linear model

$$
z(t) = \begin{cases}
0.287318+0.160788z(t-1) \\
+1.00663u(t-1)+0.161784\zeta(t-1) \\
+\zeta(t)
\end{cases} ; \quad 0.053636z(t-1)+0.008592u(t-1) +0.203\zeta(t-1) > -0.036643
$$

$$
\begin{cases}
-0.309377+0.172679z(t-1) \\
+0.961068u(t-1)-0.326336\zeta(t-1) \\
+\zeta(t)
\end{cases} ; \quad 0.065527z(t-1)-0.03697u(t-1) +0.367552\zeta(t-1) < 0.055183 \tag{65}
$$

$$
\begin{cases}
-0.003519+0.107152z(t-1) \\
+0.998038u(t-1)-0.041216\zeta(t-1) \\
+\zeta(t)
\end{cases} ; \quad \begin{cases}
0.053636z(t-1)+0.008592u(t-1) \\
+0.203\zeta(t-1) \leq -0.036643 \\
\textit{and} \\
0.065527z(t-1)-0.03697u(t-1) \\
+0.367552\zeta(t-1) \geq 0.055183
\end{cases}
$$

Comparing eqn.(65) with eqn.(29) shows the structure of the original system has been lost. This might be attributed to the low signal to noise ratio because the discriminant functions are not well defined in this case and a significant number of data records could not be linearly separated. Figure 30 shows the first 50 model predicted outputs for the fitted multi-class linear model eqn.(65) superimposed on the actual output of the system. A

rather poor performance was obtained.

From the high and low signal to noise ratio examples, a number of remarks can be drawn. The algorithm performed well when the signal to noise ratio was high at 80dB and the structure of the original system was recovered with the proposed algorithm. The performance of the algorithm deteriorated rapidly as the signal to noise ratio decreased probably because there were large areas of overlapping hyperplane and consequently the classification of data records using the discriminant became unclear. The linearly separable property is then no longer valid and the structure of the original system is lost. This interpretation is reinforced by the fact that only 90 out of the 354 data records were isolated from the first data partitioning process when the signal to noise ratio was poor.

7. Conclusions

An algorithm for the identification of multi-class systems has been developed based on the use of discriminant functions for partitioning the data records. Providing the system is well defined and linearly separable, the algorithm can correctly classify the data into the respective classes. Model estimates corresponding to each class can then be obtained using any of the linear or nonlinear parametric identification algorithms. For less well defined systems, the performance of the algorithm deteriorates rapidly as the signal to noise ratio decreases. This is probably due to the fact that there is an increasing area of overlapping hyperplane such that a large number of data records are no longer linearly separable and procedures which overcome this deficiency are currently under investigation.

7. References

1. Tong H. [1983]: Threshold models in non-linear time series analysis, Lecture notes in statistics 21, Springer-Verlag.
2. Astrom K.J. and Eykhoff P. [1971]: System identification - a survey, Automatica 7, 123-162.
3. Billings S.A. [1985]: An overview of nonlinear systems identification, 7th IFAC symposium on identification and system parameter estimation, 725-729.
4. Hand D.J. [1981]: Discrimination and classification, John Wiley & Sons.
5. Billings S.A. and Fadzil M.B. [1985]: The practical identification of system with nonlinearities, 7th IFAC symposium on identification and system parameter estimation, 155-160.
6. Leontaritis I.J. and Billings S.A. [1985]: Input-output parametric models for nonlinear systems: Part I - Deterministic nonlinear systems, Part II - Stochastic nonlinear systems, International Journal of Control, vol.41, 303-344.
7. Korenberg M.J., Billings S.A. and Liu Y.P. [1988]: An orthogonal parameter estimation algorithm for nonlinear stochastic systems, International Journal of Control, 48, 193-210.
8. Akaike H. [1974]: A new look at statistical model identification, IEEE Trans. AC-19, 716-723.

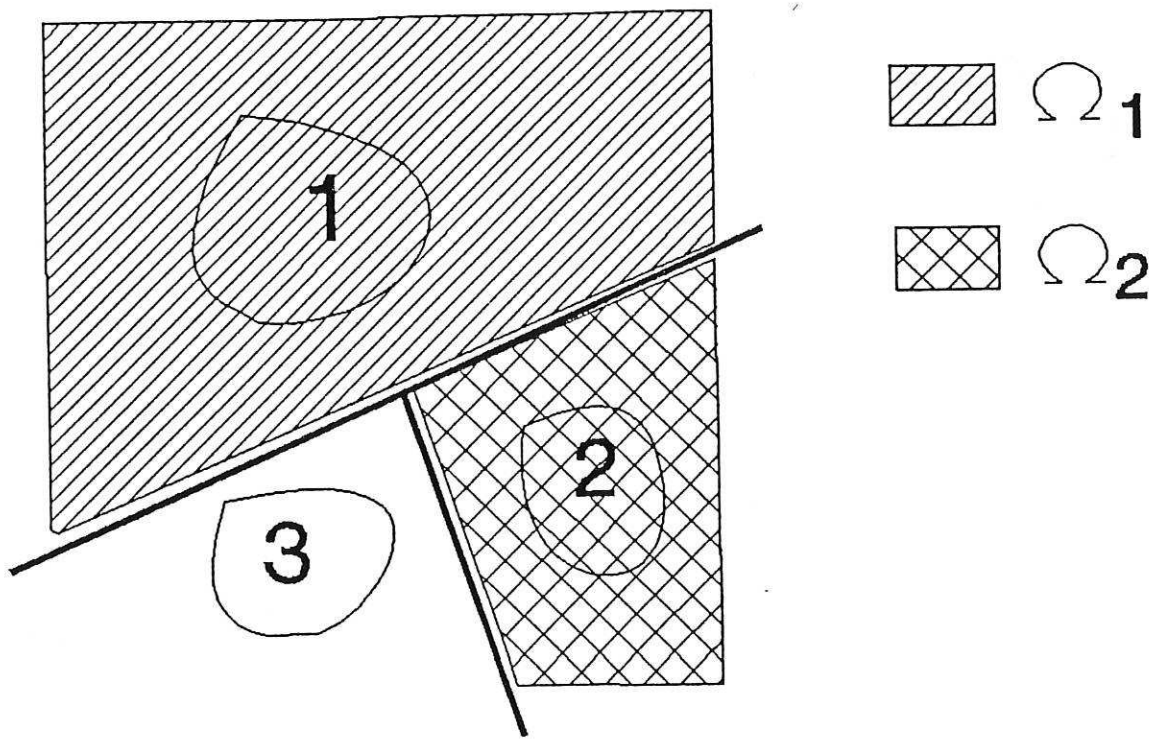

Figure 1. n-1 two-class decision surfaces, each separating Ω_i from $(\Omega_{i+1}, \ldots, \Omega_n)$, i-1, . . . , n.

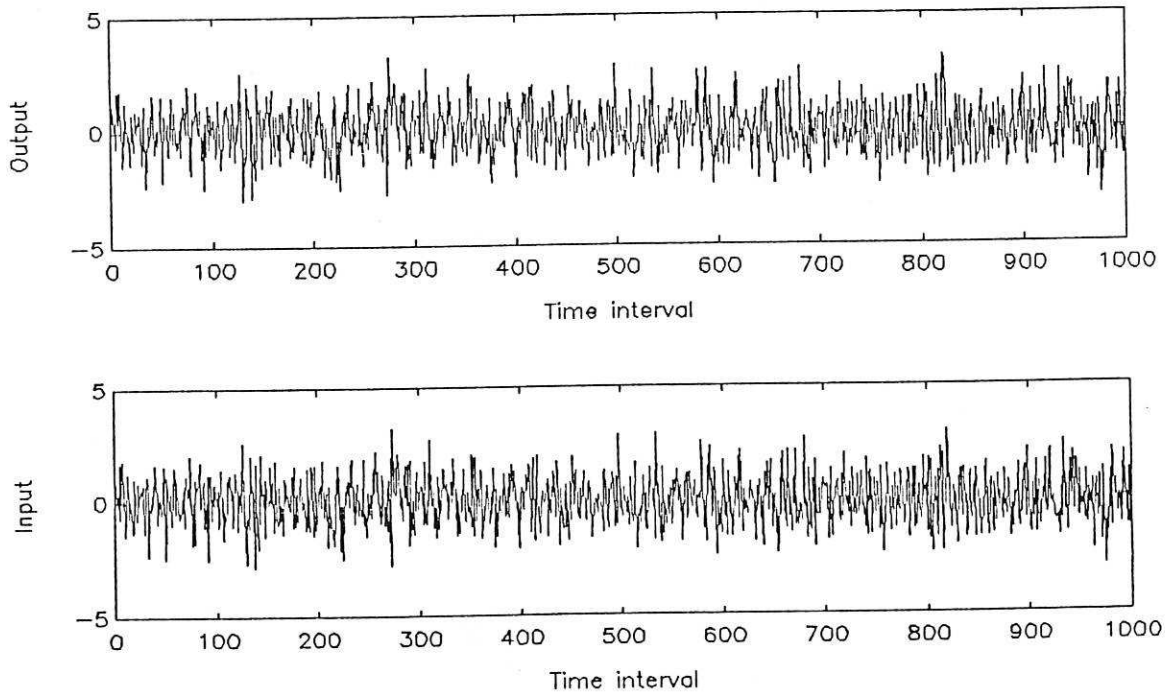

Figure 2. Input and output data records for the piece-wise linear system (S_1)

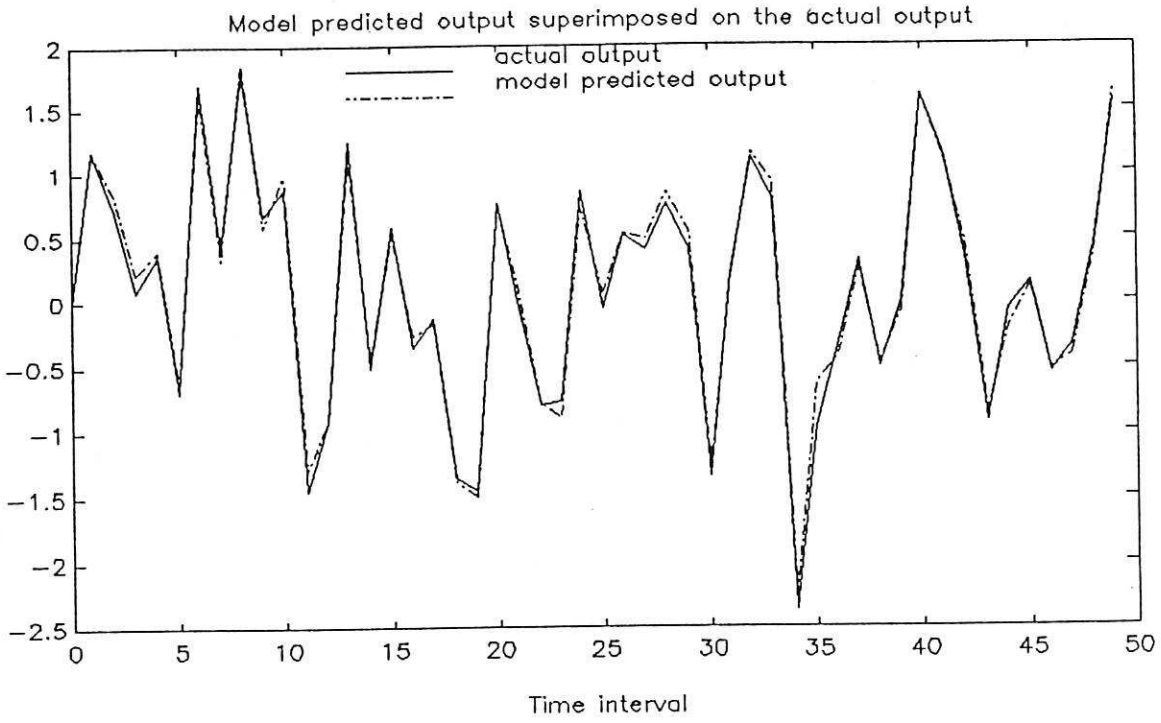

Figure 3. Model predicted output of the fitted linear model, S_1

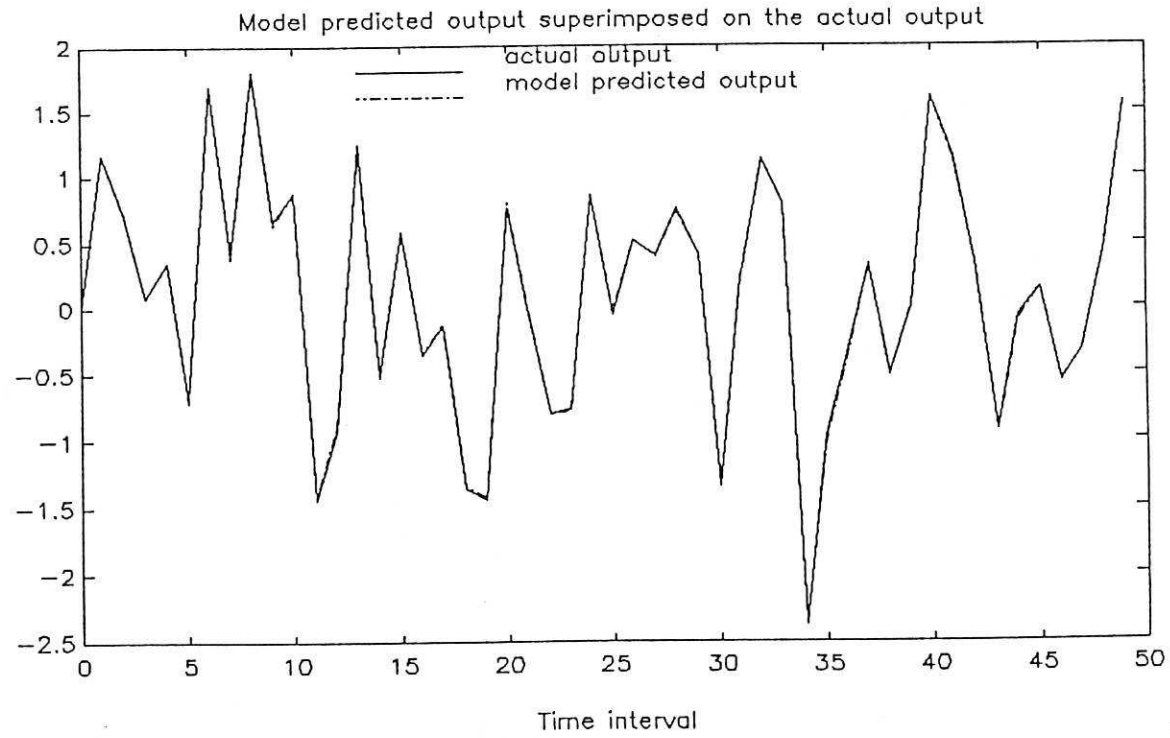

Figure 4. Model predicted output of the fitted nonlinear model, S_1

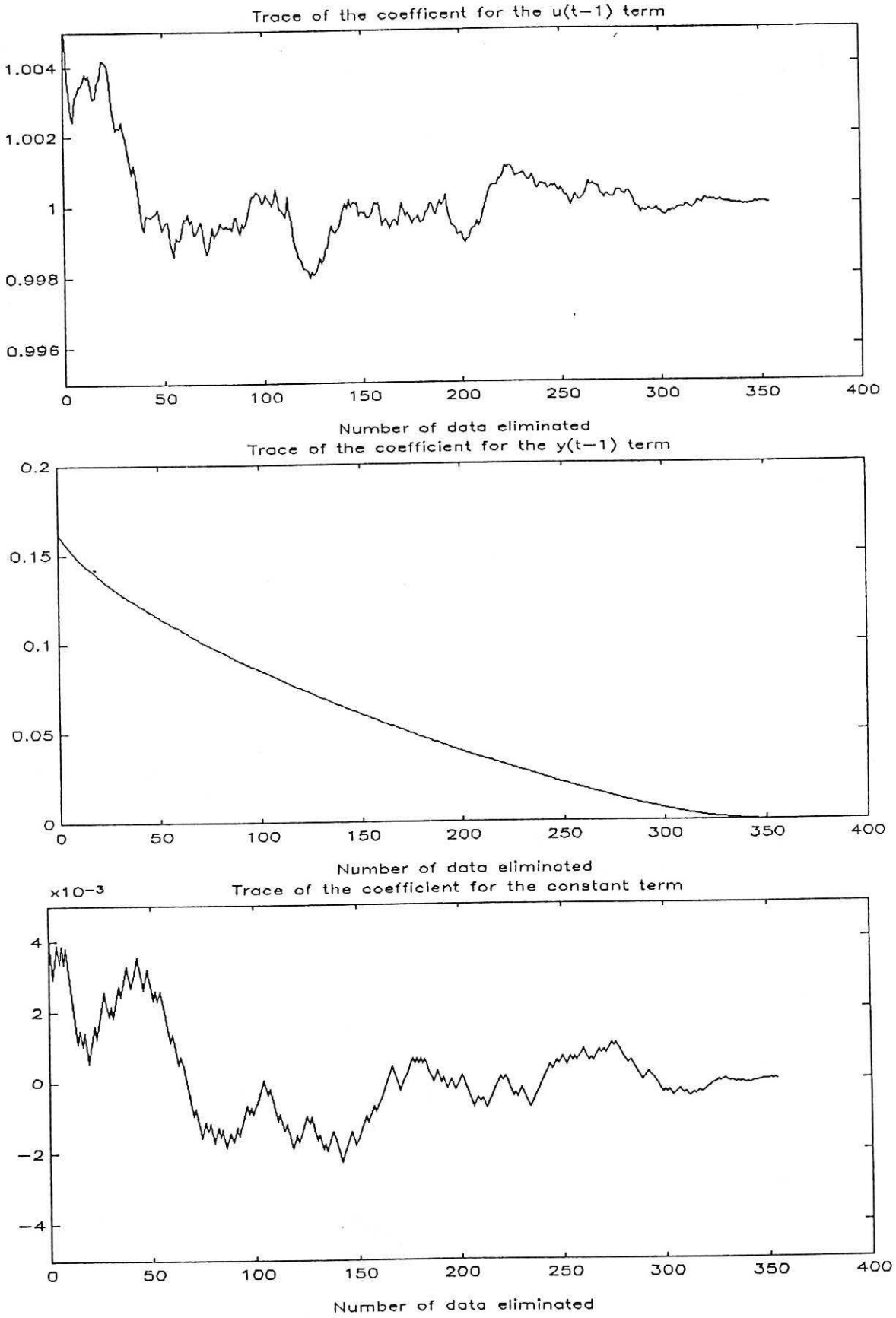

Figure 5. Profile of the parameter vector for S_1, Ω_1

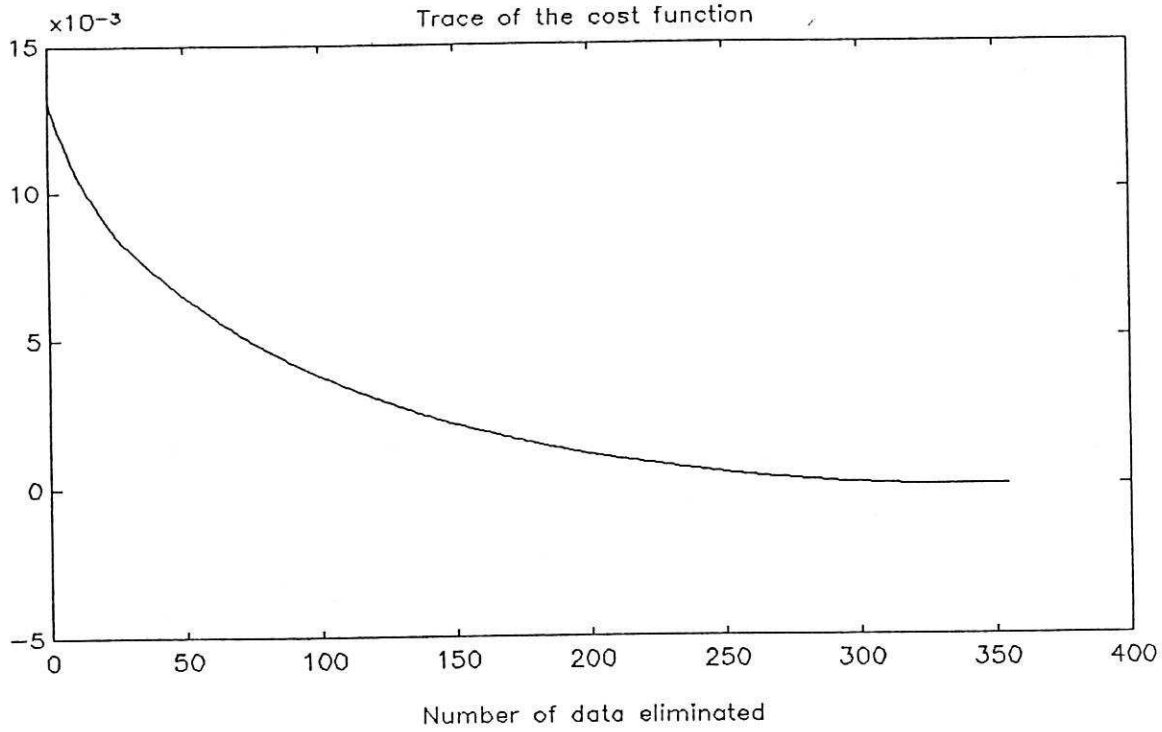

Figure 6. Profile of the cost function for S_1

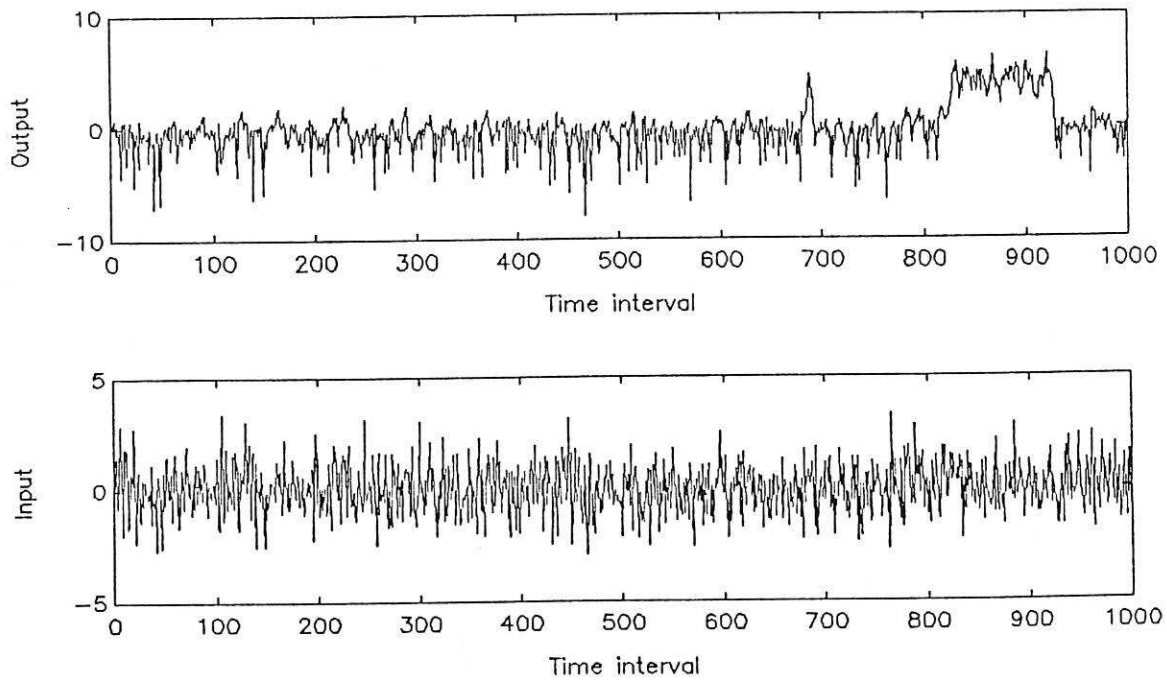

Figure 7. Input and output data records for S_2

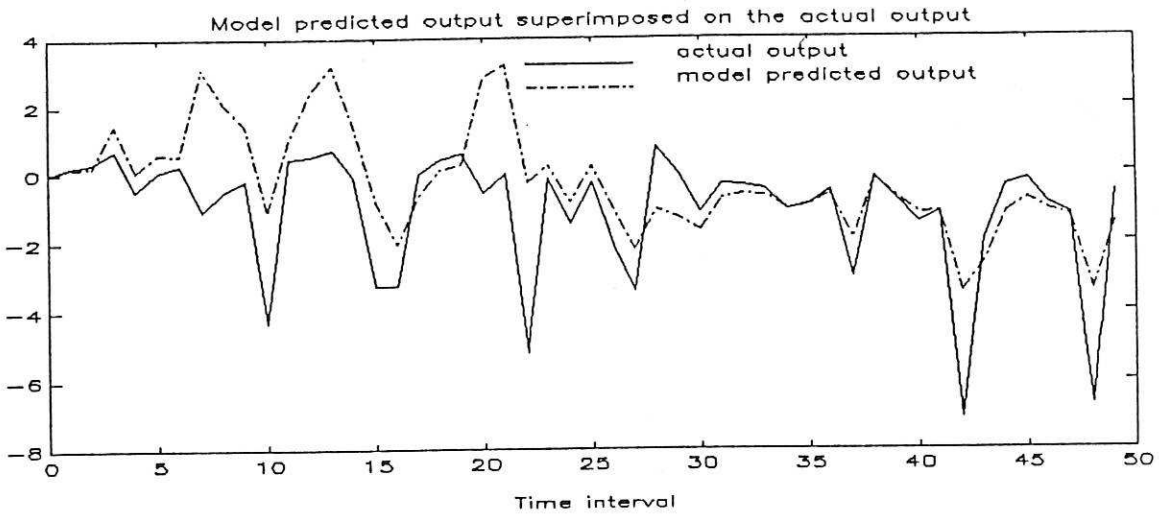

Figure 8. Model predicted output of the fitted linear model, S_2

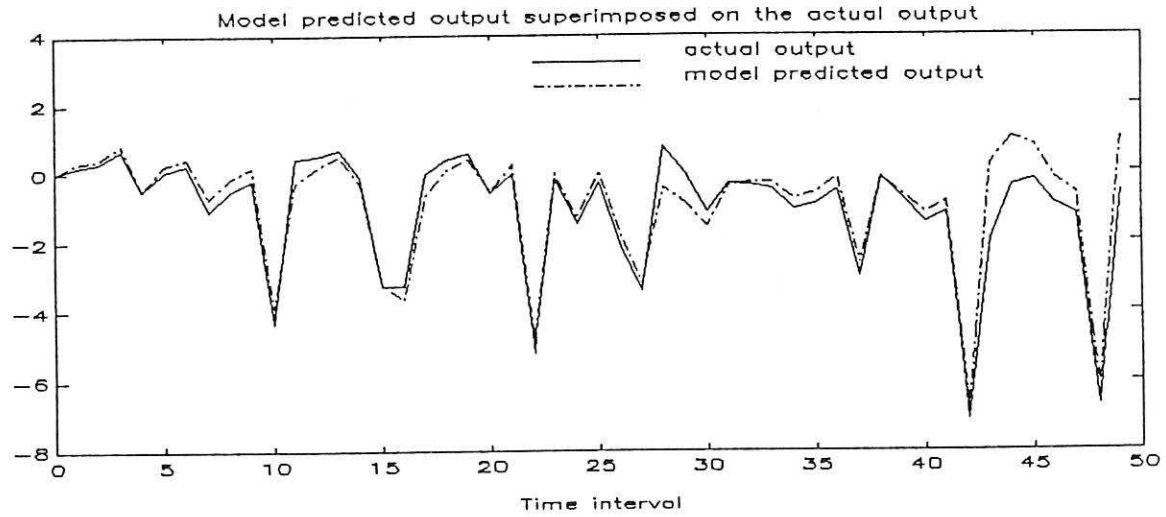

Figure 9. Model predicted output of the fitted nonlinear model, S_2

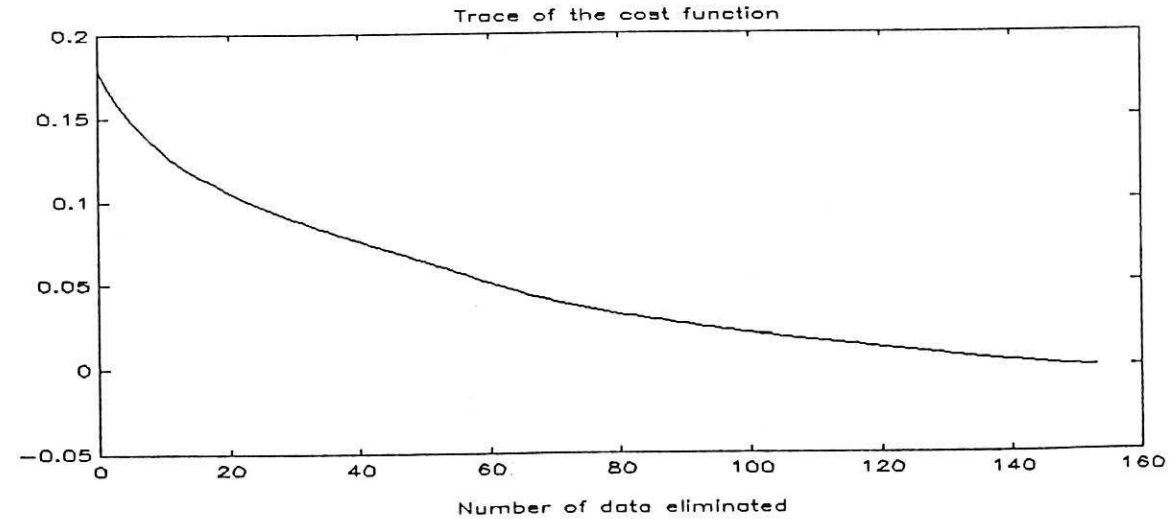

Figure 10. Profile of the cost function for S_2, Ω_1

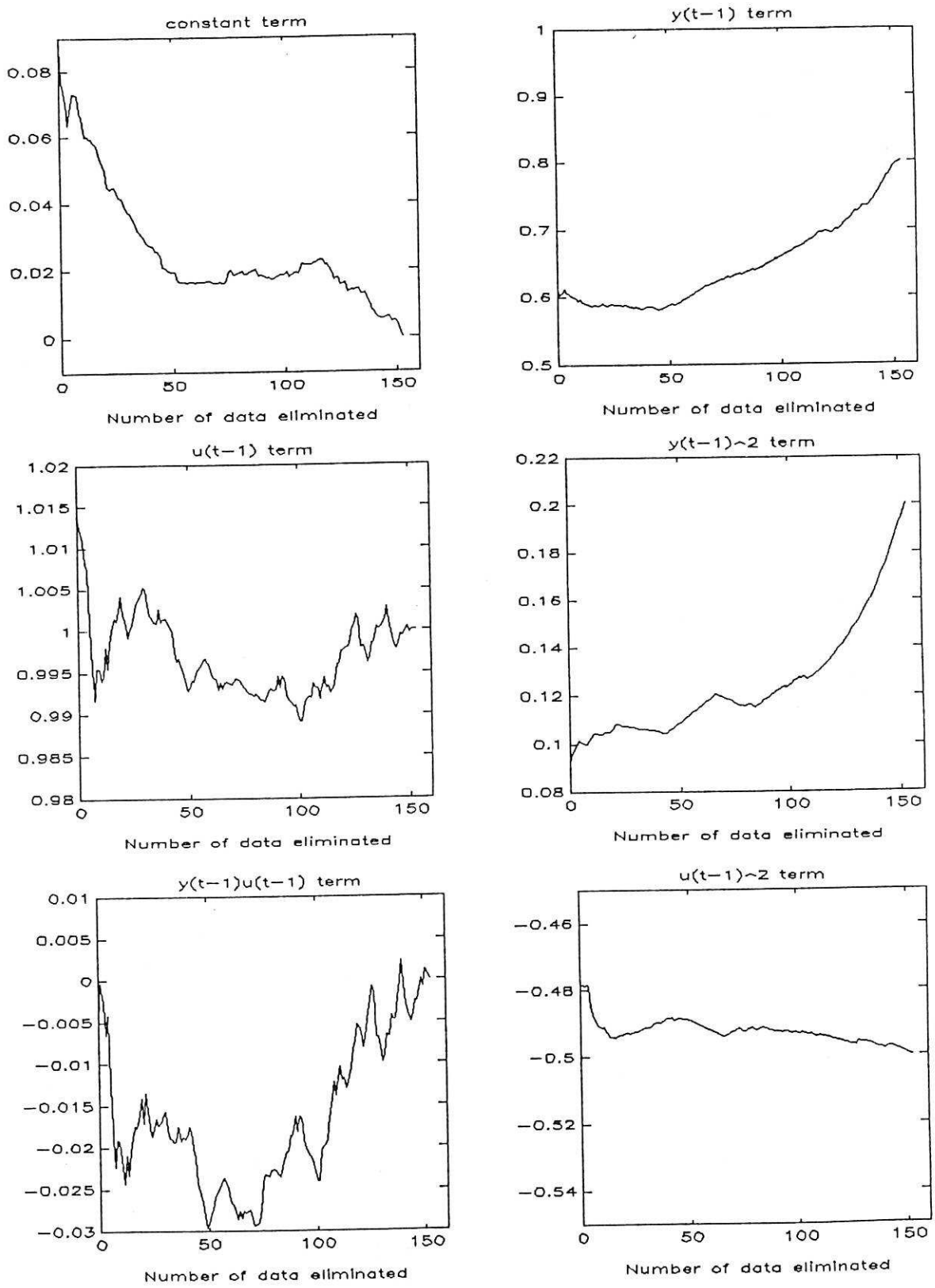

Figure 11. Profile of the parameter vector for S_2, Ω_1

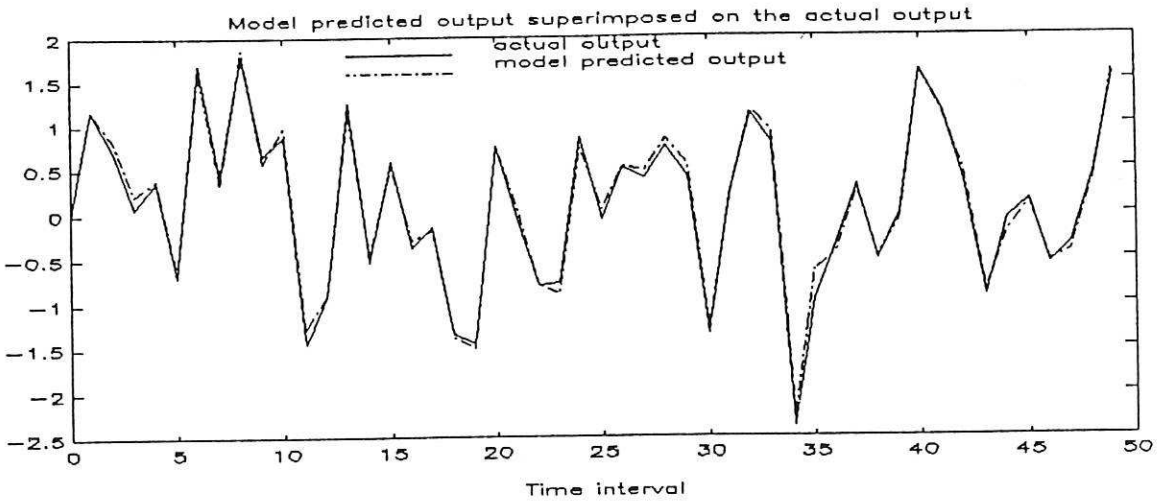

Figure 12. Model predicted output of the fitted linear model, S_3

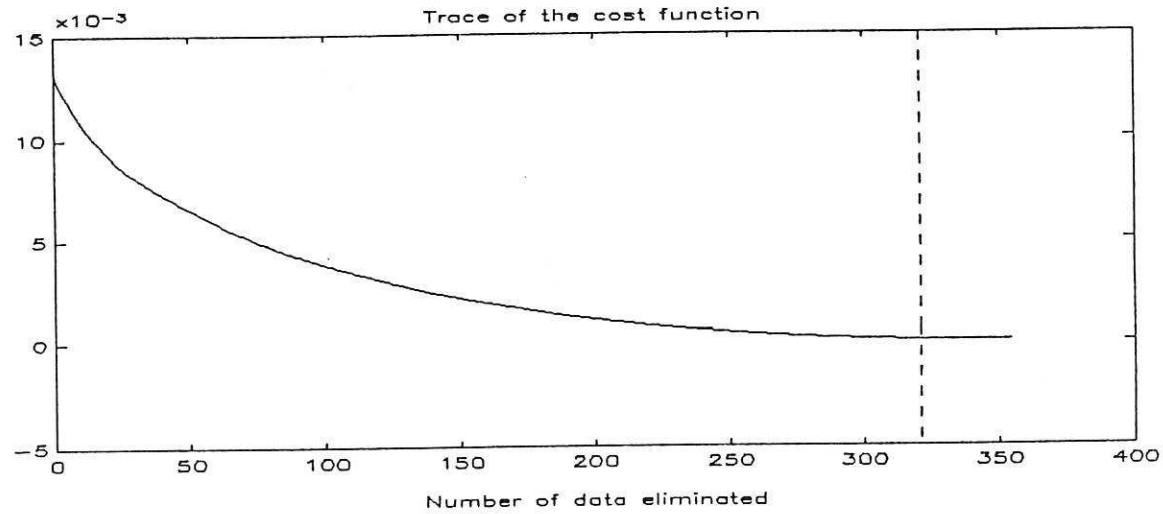

Figure 13. Profile of the cost function for S_3, Ω_1

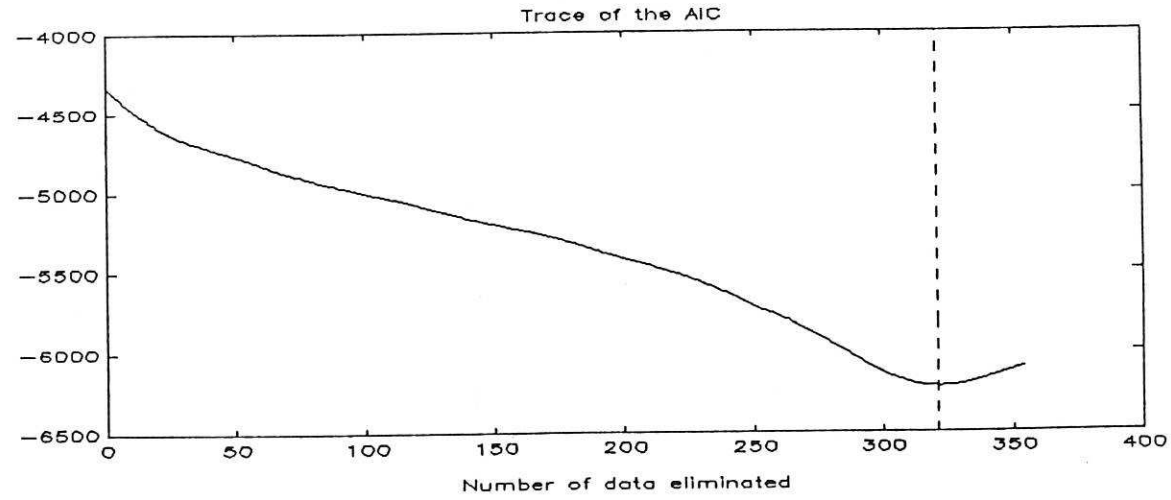

Figure 14. Profile of the AIC for S_3, Ω_1

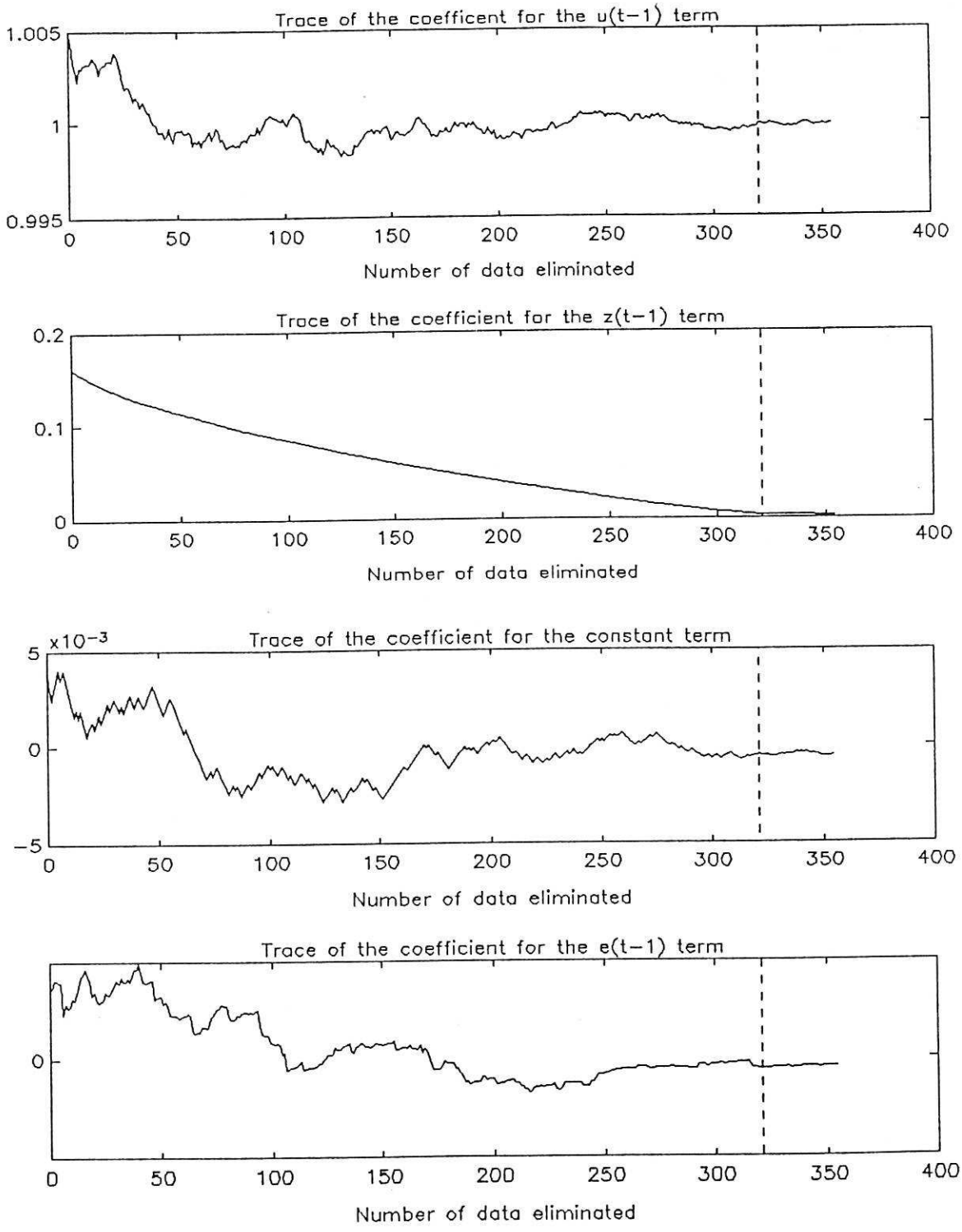

Figure 15. Profile of the parameter vector for S_3, Ω_1

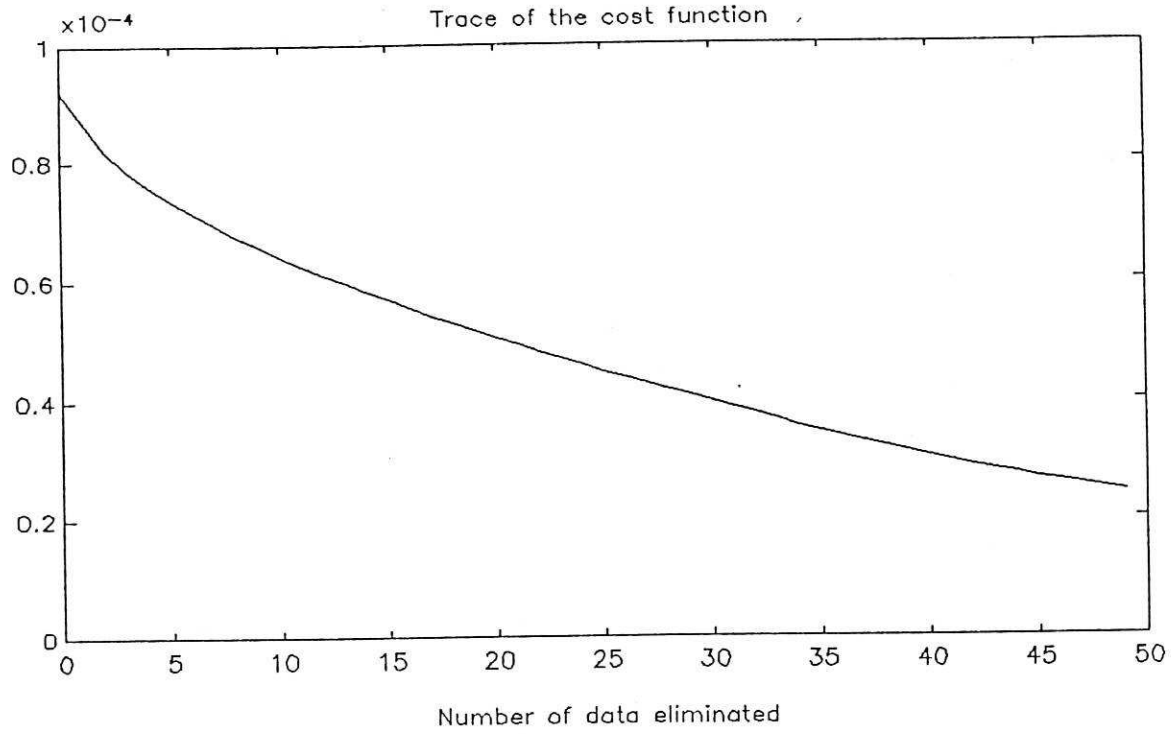

Figure 16. Profile of the cost function for S_3, Ω_a

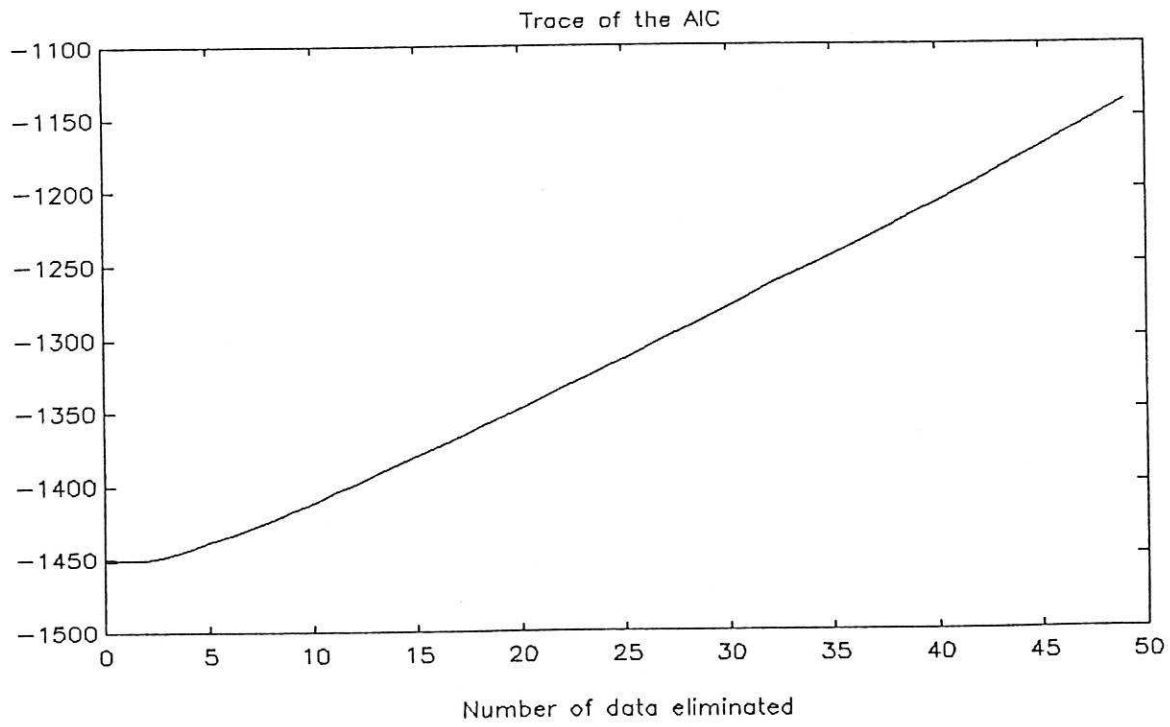

Figure 17. Profile of the AIC for S_3, Ω_a

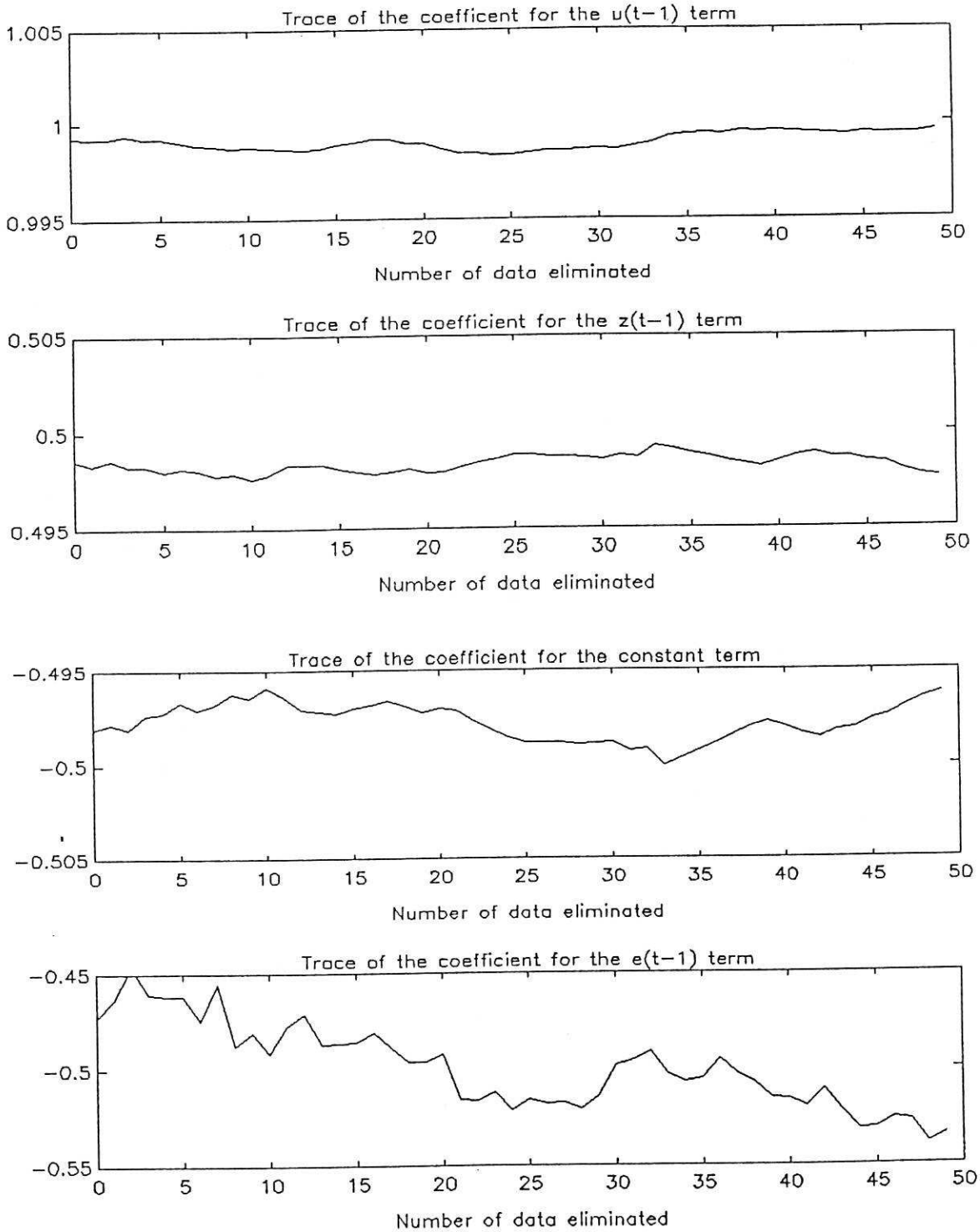

Figure 18. Profile of the parameter vector for S_3, Ω_a

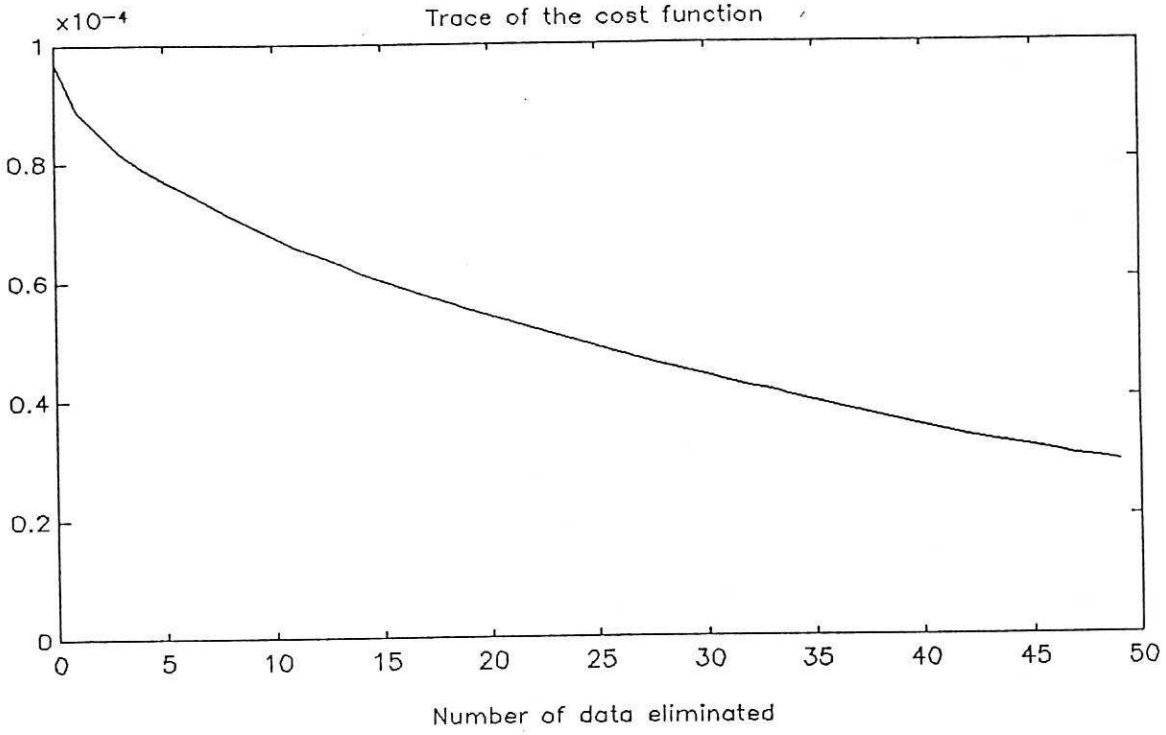

Figure 19. Profile of the cost function for S_3, Ω_b

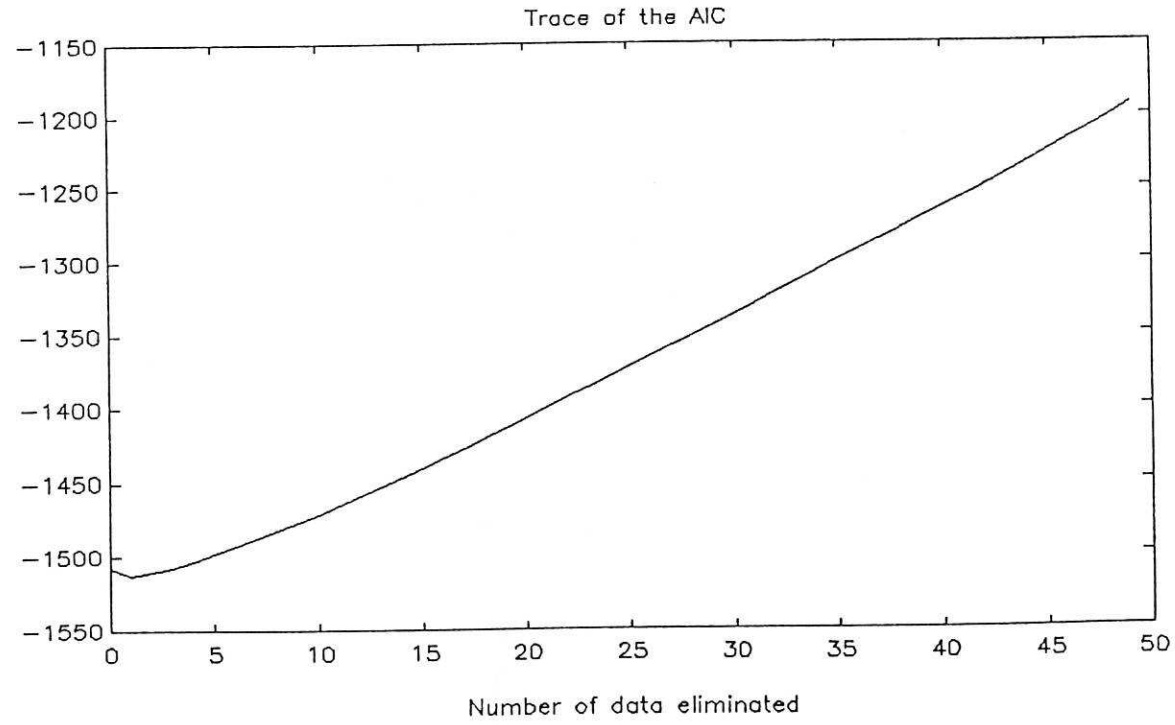

Figure 20. Profile of the AIC for S_3, Ω_b

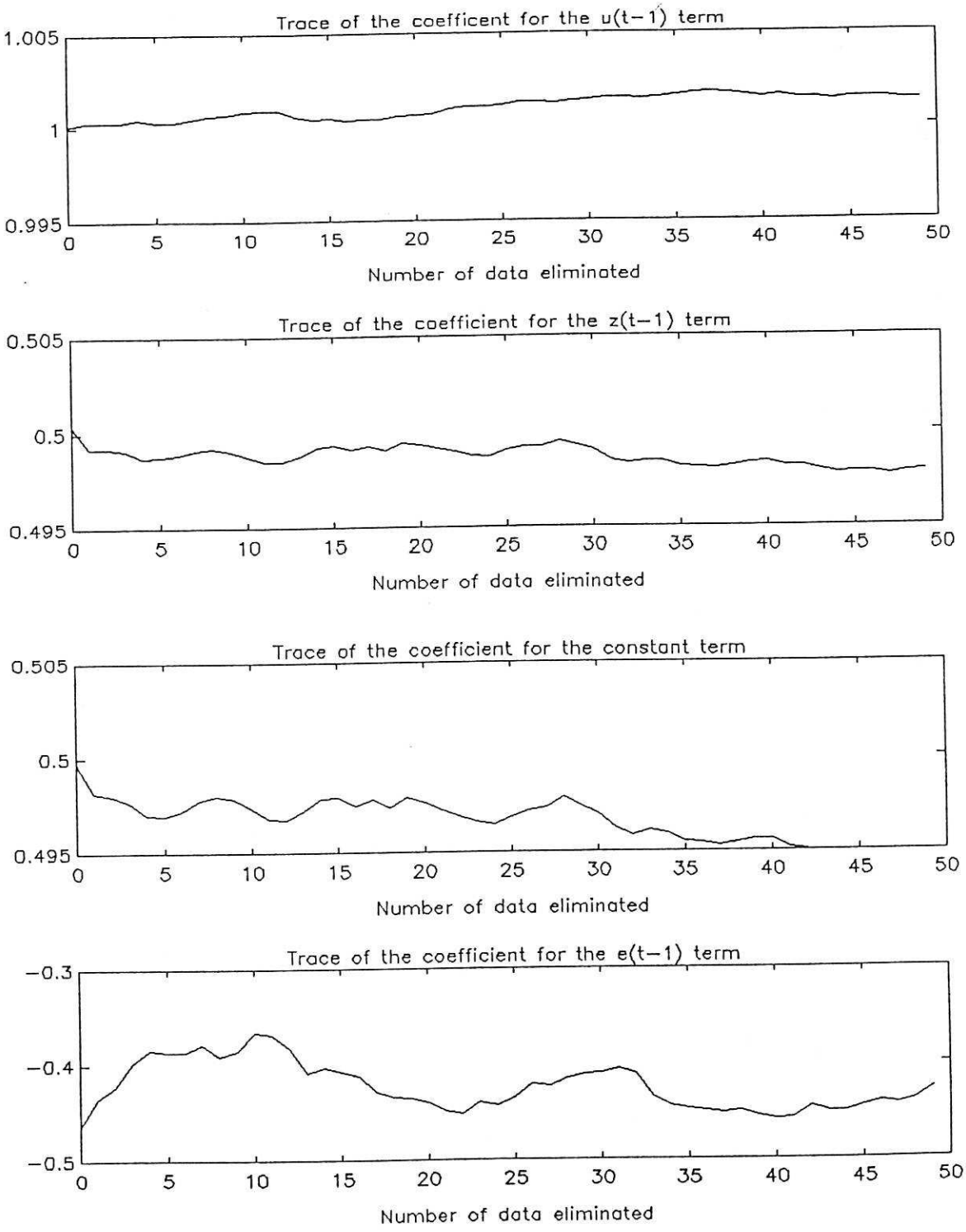

Figure 21. Profile of the parameter vector for S_3, Ω_b

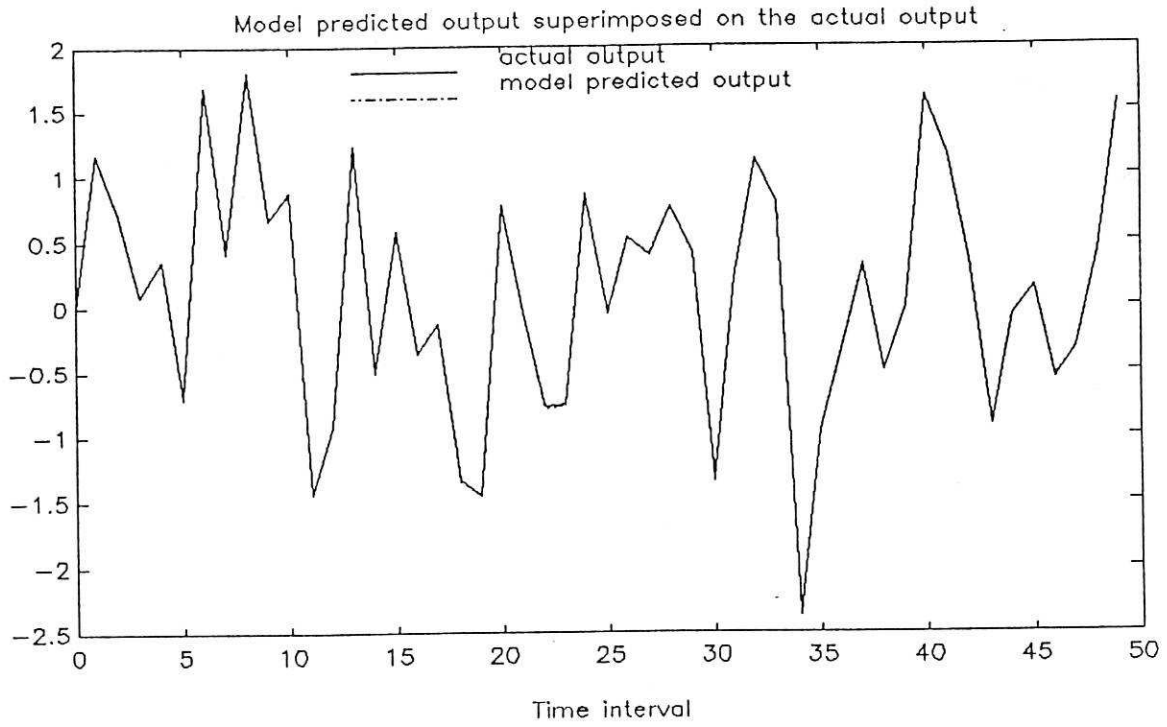

Figure 22. Model predicted output of the fitted multiclass linear model for S_3

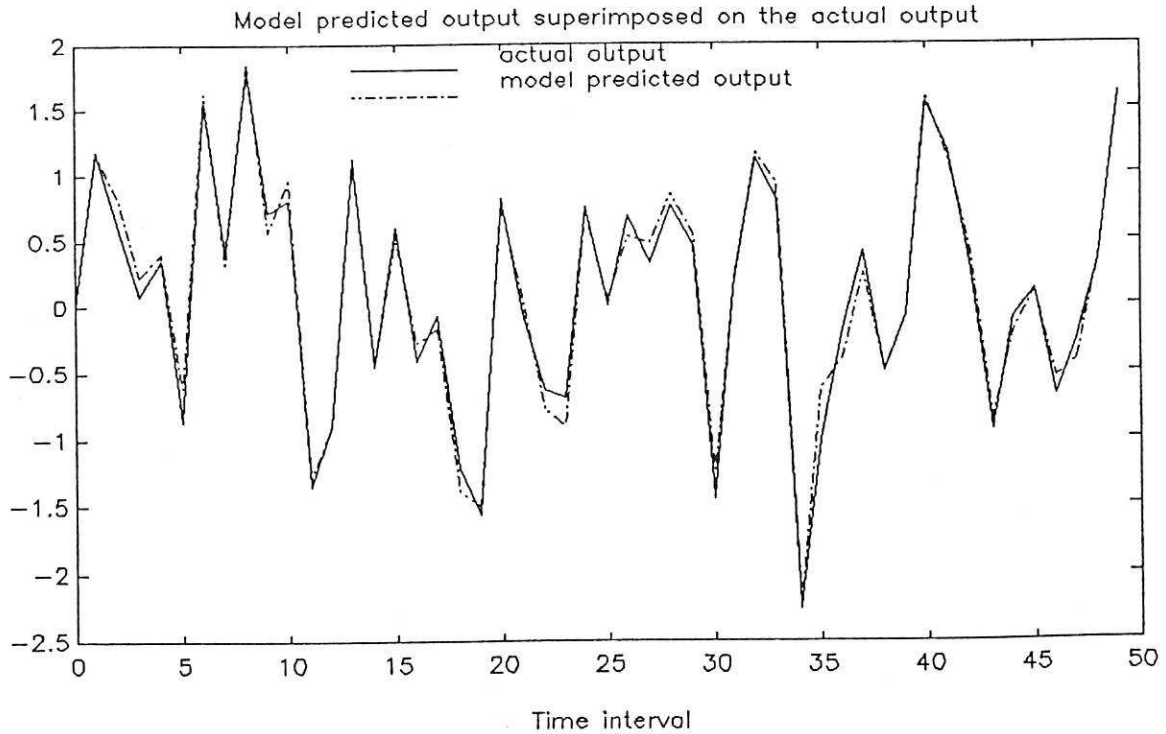

Figure 23. Model predicted output of the fitted linear model for S_4

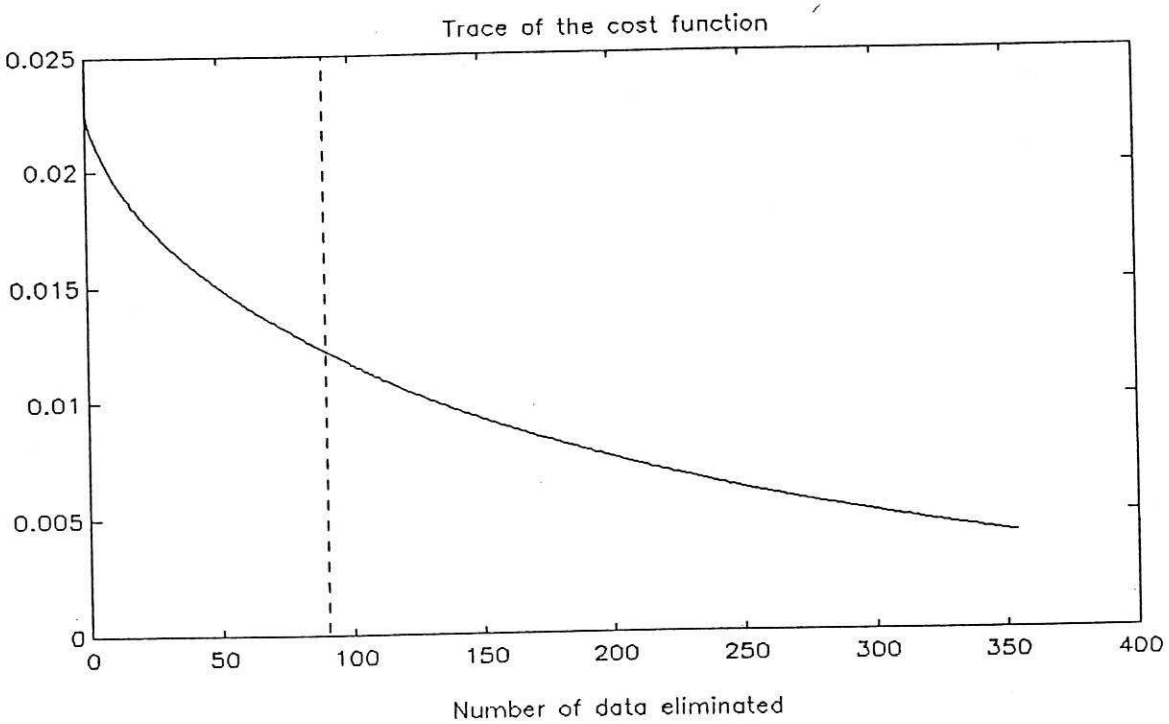

Figure 24. Profile of the cost function for S_4, Ω_1

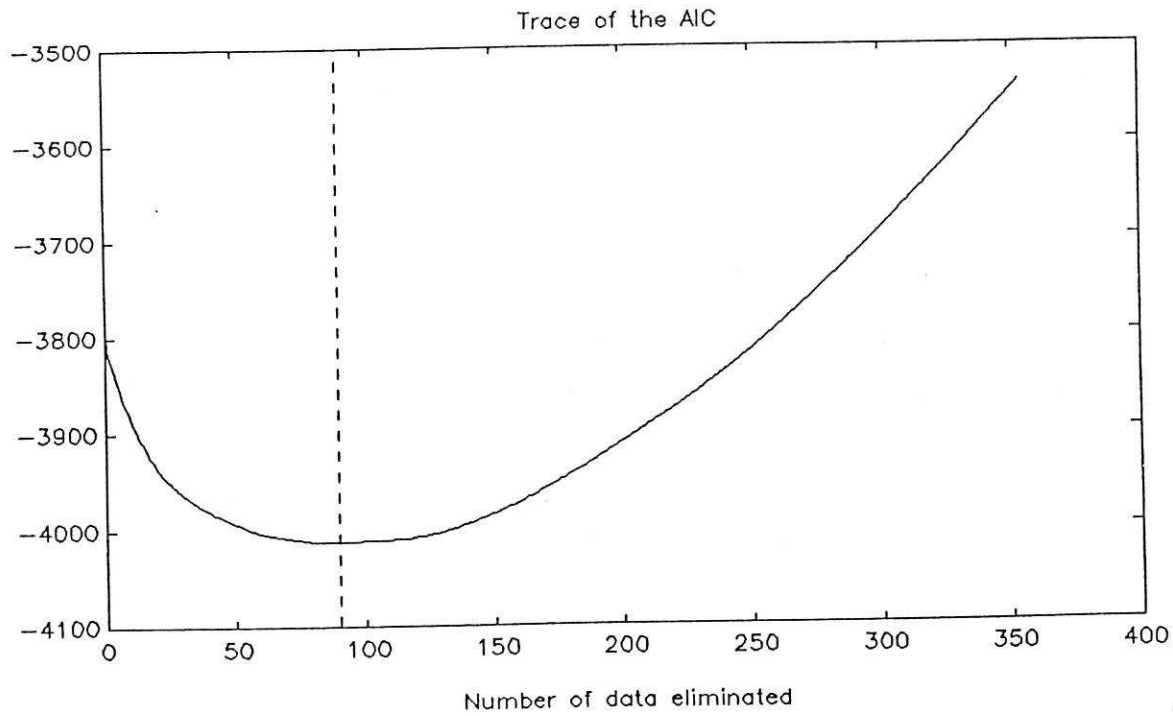

Figure 25. Profile of the AIC for S_4, Ω_1