Empirical State Space Modelling with Application in Online Diagnosis of Multivariate Non-Linear Dynamic Systems

Jakobus Petrus Barnard

Dissertation presented for the Degree of Doctor of Philosophy in Engineering at the University of Stellenbosch.

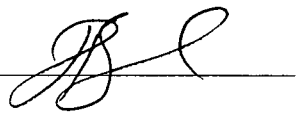
Promoter: Prof. C. Aldrich.

Co-promoter: Dr. M. Gerber.

- November 1999 -

DECLARATION

I, the undersigned, hereby declare that the work contained in this dissertation is my own original work and that I have not previously in its entirety or in part submitted it at any university for a degree.

Signature: 

Date: 2 - 02 - 2000

SYNOPSIS

System identification has been sufficiently formalized for linear systems, but not for empirical identification of non-linear, multivariate dynamic systems. Therefore this dissertation formalizes and extends non-linear empirical system identification for the broad class of non-linear multivariate systems that can be parameterized as state space systems. The established, but rather ad hoc methods of time series embedding and nonlinear modeling, using multi-layer perceptron network and radial basis function network model structures, are interpreted in context with the established linear system identification framework.

First, the methodological framework was formulated for the identification of non-linear state space systems from one-dimensional time series using a surrogate data method. It was clearly demonstrated on an autocatalytic process in a continuously stirred tank reactor, that validation of dynamic models by one-step predictions is insufficient proof of model quality. In addition, the classification of data as either dynamic or random was performed, using the same surrogate data technique. The classification technique proved to be robust in the presence of up to at least 10% measurement and dynamic noise.

Next, the formulation of a nearly real-time algorithm for detection and removal of radial outliers in multidimensional data was pursued. A convex hull technique was proposed and demonstrated on random data, as well as real test data recorded from an internal combustion engine. The results showed the convex hull technique to be effective at a computational cost two orders of magnitude lower than the more proficient Rocke and Woodruff technique, used as a benchmark, and incurred low cost (0.9%) in terms of falsely identifying outliers.

Following the identification of systems from one-dimensional time series, the methodological framework was expanded to accommodate the identification of nonlinear state space systems from multivariate time series. System parameterization was accomplished by combining individual embeddings of each variable in the multivariate time series, and then separating this combined space into independent components, using independent component analysis. This method of parameterization was successfully applied in the simulation of the above-mentioned autocatalytic process. In addition, the parameterization method was implemented in the one-step prediction of atmospheric NO_2 concentrations, which could become part of an environmental control system for Cape Town. Furthermore, the combination of the

embedding strategy and separation by independent component analysis was able to isolate some of the noise components from the embedded data.

Finally the foregoing system identification methodology was applied to the online diagnosis of temporal trends in critical system states. The methodology was supplemented by the formulation of a statistical likelihood criterion for simultaneous interpretation of multivariate system states. This technology was successfully applied to the diagnosis of the temporal deterioration of the piston rings in a compression ignition engine under test conditions. The diagnostic results indicated the beginning of significant piston ring wear, which was confirmed by physical inspection of the engine after conclusion of the test. The technology will be further developed and commercialized.

OORSIG

Stelselidentifikasie is wel genoegsaam ten opsigte van lineêre stelsels geformaliseer, maar nie ten opsigte van die identifikasie van nie-lineêre, multiveranderlike stelsels nie. In hierdie tesis word nie-lineêre, empiriese stelselidentifikasie gevolglik ten opsigte van die wye klas van nie-lineêre, multiveranderlike stelsels, wat geparameteriseer kan word as toestandveranderlike stelsels, geformaliseer en uitgebrei. Die gevestigde, maar betreklik ad hoc metodes vir tydreeksontvouing en nie-lineêre modellering (met behulp van multilaag-perseptron- en radiaalbasisfunksie-modelstrukture) word in konteks met die gevestigde lineêre stelselidentifikasieraamwerk vertolk.

Eerstens is die metodologiese raamwerk vir die identifikasie van nie-lineêre, toestandsveranderlike stelsels uit eendimensionele tydreekse met behulp van 'n surrogaatdata-metode geformuleer. Daar is duidelik by wyse van 'n outokatalitiese proses in 'n deurlopend geroerde tenkreaktor getoon dat die bevestiging van dinamiese modelle deur middel van enkelstapvoorspellings onvoldoende bewys van die kwaliteit van die modelle is. Bykomend is die klassifikasie van tydreekse as óf dinamies óf willekeurig, met behulp van dieselfde surrogaattegniek gedoen. Die klassifikasietegniek het in die teenwoordigheid van tot minstens 10% meetgeraas en dinamiese geraas robuust vertoon.

Vervolgens is die formulering van 'n bykans intydse algoritme vir die opspoor en verwydering van radiale uitskieters in multiveranderlike data aangepak. 'n Konvekse hulstegniek is voorgestel en op ewekansige data, sowel as op werklike toetsdata wat van 'n binnebrandenjin opgeneem is, gedemonstreer. Volgens die resultate was die konvekse hulstegniek effektief teen 'n rekenkoste twee grootte-ordes kleiner as die meer vermoënde Rocke en Woodruff-tegniek, wat as meetstandaard beskou is. Die konvekse hulstegniek het ook 'n lae loopkoste (0.9%) betreffende die valse identifisering van uitskieters behaal.

Na aanleiding van die identifisering van stelsels uit eendimensionele tydreekse, is die metodologiese raamwerk uitgebrei om die identifikasie van nie-lineêre, toestandsveranderlike stelsels uit multiveranderlike data te omvat. Stelselparameterisering is bereik deur individuele ontvouings van elke veranderlike in die multidimensionele tydreeks met die skeiding van die gesamenlike ontvouingsruimte tot onafhanklike komponente saam te span. Sodanige skeiding is deur middel van onafhanklike komponentanalise behaal. Hierdie metode van

parameterisering is suksesvol op die simulering van bogenoemde outokatalitiese proses toegepas. Die parameteriseringsmetode is bykomend in die enkelstapvoorspelling van atmosferiese NO_2-konsentrasies ingespan en sal moontlik deel van 'n voorgestelde omgewingsbestuurstelsel vir Kaapstad uitmaak. Die kombinasie van die ontvouingstrategie en skeiding deur onafhanklike komponentanalise was verder ook in staat om van die geraaskomponente in die data uit te lig.

Ten slotte is die voorafgaande tegnologie vir stelselidentifikasie op die lopende diagnose van tydsgebonde neigings in kritiese stelseltoestande toegepas. Die metodologie is met die formulering van 'n statistiese waarskynlikheidsmaatstaf vir die gelyktydige vertolking van multiveranderlike stelseltoestande aangevul. Hierdie tegnologie is suksesvol op die diagnose van die tydsgebonde verswakking van die suierringe in 'n kompressieontstekingenjin tydens toetstoestande toegepas. Die diagnostiese resultate het die aanvang van beduidende slytasie in die suierringe aangedui, wat later tydens fisiese inspeksie van die enjin met afloop van die toets, bevestig is. Die tegnologie sal verder ontwikkel en markgereed gemaak word.

ACKNOWLEDGEMENT

This research project has been a team effort in many ways and I am greatly indebted to some very dedicated and supportive people and institutions:

a) the Foundation for Research and Development, my sponsors,

b) Prof. Chris Aldrich and dr. Marius Gerber, my supervisors,

c) Juliana Steyl, the secretary of Prof. Aldrich,

d) Francois Gouws and Gregor Schmitz, two of my friends and co-students,

e) all my other friends and colleagues who had to bear with me through the bad times. Some of them still loves me nonetheless.

f) Prof. Henry Abarbanel. Parameterization of non-linear time series was done with the aid of the CspW software purchased by my supervisors under academic license from Applied Nonlinear Sciences (http://www.zweb.com/apnonlin/), directed by Prof. Abarbanel.

g) Prof. David Rocke. To implement the Rocke and Woodruff algorithm for outlier detection, I used C-code provided by Rocke and Woodruff on the STATLIB web-site (http://lib.stat.cmu.edu/jasasoftware/rocke), hosted by the Department of Statistics at the Carnegie Mellon University, United States of America.

h) Michael Small and Kevin Judd. The PL-MODEL software, developed by Michael and Kevin, was used for the implementation of radial basis function model structures.

i) Brad Barber. To construct the convex hulls, I used the Quick-hull (Qhull) algorithm (Barber et al., 1996) and software by Brad Barber and associates of the Geometry Center at the University of Minneapolis, United States of America.

j) Mathworks, for the use of their Matlab 4.2 and 5.3 software under academic license.

k) Microsoft SA, for the use of Office 97 software under their local academic program.

l) My parents, who made me and set me off on my course.

m) The person that I ultimately acknowledge is my Lord Jesus, who created me, from whom I received inspiration, and wisdom. God has truly been my provider in all regards.

Table of Contents

Table of Figures

Table of tables

1 INTRODUCTION

We have an inherent aspiration to embed observations of Nature as well as our creations in a pattern of some kind, to better understand, create and control. We also strive to ensure that our man-made systems remain in good working order and prefer to detect imminent failure before it strikes. Therefore we, the stewards of Nature, build mathematical models from observations. Engineers have a significant and very tangible influence on Nature through the systems and constructions they create. In essence, through the scientific method, we are locally transforming Nature into engineering systems of all kinds for various purposes.

Efficiency, reliability and maintenance are three major factors in the operation of engineering systems, ranging from plants such as chemical refineries, electrical power stations and manufacturing facilities, to aeronautical propulsion systems and all kinds of automotive vehicles. Cost-effectiveness and conservation of resources are directly affected by efficiency, while reliability and maintenance influence operational safety, continuity and cost. In most modern countries, statutory laws and regulations on environmental conservation often dictate operational boundaries for chemical plants and food-processing factories in terms of emissions of chemicals, effluent and noise. Certain American states have particularly strict regulations on the exhaust and fuel emissions from automotive vehicles, California being the prime example. Likewise, these laws and regulations necessitate proper operational diagnostics and maintenance of these engineering plants, systems and appliances. Prediction models to enhance efficient system management and online diagnosis to detect imminent system failure are therefore indispensable to operational continuity, safety as well as cost management and environmental protection.

Online diagnosis of complex systems rely strongly on some form of failure detection in order to maintain satisfactory performance and prolong system life. Often such detection depends on the operator's skill of interpreting rather rudimentary alarms and single state monitors, such as digital or analogue displays for temperatures and pressures. A more challenging, though frequent scenario is the simultaneous interpretation of multiple observations. One relevant area of such simultaneous interpretation is the diagnosis of sensors and controllers in diverse engineering systems. Another area of online system diagnostics, which often constitutes a significant problem, is the need to determine gradual, temporal system deterioration. Often this deterioration can only be reliably detected through simultaneous interpretation of the

history of dependent variables and a set of independent variables. An example is the temporal deterioration of an internal combustion engine, due to wear of the piston rings. Transportation enterprises and mining operations, among others, have fleets of vehicles or auxiliary power generators utilizing internal combustion engines. Neglecting the timely observation of a condition of failure in terms of ring wear will usually result in a costly overhaul of the cylinder bores. Maintenance could be scheduled more timely and more cost-effectively, if assistance could be provided in terms of the online interpretation of simultaneous system states. System diagnostics with regard to some form of automated state observation and failure detection are therefore required.

Most of the above-mentioned real world systems are dynamic, which means they exhibit temporal changing behaviour. Chemical, metallurgical and mechanical systems in particular can be high-dimensional and non-linear. Notwithstanding the current scientific knowledge-base, the complexity of these processes makes them difficult to understand, model, interpret and control. As a consequence, engineers often try to develop empirical dynamic process models for these systems direct from input-output data, rather than attempting to develop time consuming, expensive fundamental, analytical models. However, in developing these models several issues have to be addressed, such as the classification of process data, selection of model structure and order, system parameterization, stationarity of the data, handling of outliers and noise in the data, parameter estimation and model validation.

The foregoing issues have been sufficiently formalised for linear systems, but not for empirical identification of non-linear, multivariate dynamic systems. Therefore this dissertation proposes a formal methodological framework that combines empirical state space modeling and online system diagnosis of a specific class of non-linear, multivariate dynamic systems. Chapter 2 defines the system class that is of interest in this dissertation, reviews the formal methodological structure of linear system identification and reports on relevant empirical system identification methods within the context of this structure. Chapter 3 describes a methodology for identification of non-linear dynamics based on a one-dimensional time series observation of a system. The methodology involves classification of the time series using surrogate data techniques, parameterization of the system by way of time series embedding, and prediction of the time series with multi-layer perceptron network models. Particular attention is given to proper validation of the model with the assistance of a surrogate data technique. Chapter 4 addresses the fast detection of radial outliers in large multivariate data sets. A novel method for such detection is proposed, based on the

construction of convex hulls around the data. In Chapter 5 the system identification method described in Chapter 3 is expanded to multivariate time series. Parameterization by individual embedding of time series components is combined with independent component analysis to reconstruct the system state space. The methodology is applied to the simulation of an autocatalytic process and the prediction of NO_2 concentration in an environmental system. Chapter 6 describes the online diagnosis of temporal trends in critical system states. The methodology incorporates system identification techniques from the previous chapters to determine system failure statistically. The technique is applied to the diagnosis of piston ring wear in an internal combustion engine under laboratory test conditions. Chapter 7 follows with a discussion of results and conclusions.

2 CURRENT METHODOLOGY FOR EMPIRICAL, NON-LINEAR SYSTEM IDENTIFICATION

System identification is well defined for linear systems and described in several comprehensive publications (Ljung, 1987, Norton, 1986, Eykhoff, 1974). Linear, discrete time dynamic systems can be represented mathematically by a state equation and an output equation, in a number of state variables (Ogata, 1995), as:

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t \\ \mathbf{y}_t &= \mathbf{C}\mathbf{x}_t + \mathbf{D}\mathbf{u}_t \end{aligned} \tag{1}$$

where **x** is the state vector, **u** the input vector of independent variables, t the time and **y** the system output. Empirical identification of the above system from observed input-output data essentially requires solving for the constant coefficient matrices **A**, **B**, **C** and **D**, and validating the resultant model against some criterion. Established linear mathematical techniques sufficiently meet these requirements.

Non-linear systems on the other hand, are harder to identify. Methods for analysis of non-linear systems are class restricted, can give partial information and are cumbersome because non-linear behaviour is diverse and complex (Norton, 1986). This is especially so when system identification is done by fundamental analytical methods, as opposed to empirical methods. Standard system identification practice addresses the following aspects: data acquisition, noise reduction, selection of model structure and order, parameterization, parameter estimation and model validation. These aspects can be structured into the following methodological framework:

a) Model selection
b) Data acquisition
c) Parameter estimation
d) Model validation

This chapter sets out to establish a formal identification framework and terminology that will be used and evolved further in the dissertation. The chapter first defines the system class that is the focus of this dissertation and reviews current, applicable non-linear system identification techniques within the above methodological framework. Standard system identification terminology is used throughout. The notation of Ljung (1987) has been adopted

and suitably modified. Definitions of potentially ambiguous terms and expressions can be referenced under Terminology. Detailed descriptions of some concepts appear in alphabetical order in Appendix A.

2.1. Model selection

As was briefly mentioned in the introduction, system identification plays a pivotal role in a truly enormous range of engineering systems. Whenever algorithms are constructed for system identification, a model is invariably exploited (Wornell, 1995). Models may be explicit (when the class of systems is well defined) or implicit (when implicit assumptions are made with regard to the system producing the signal, such as assumptions on analytical smoothness). Some classes of systems are larger than others and generally models that apply to a smaller class tend to perform better than more complex models applicable to a larger class of systems - that is, each system model from the smaller set of systems generalize more accurately than systems models from larger classes. There are many systems that produce signals whose key characteristics are fundamentally different from those produced by conventional linear time-invariant systems. We are interested in the class of deterministic, non-linear dynamical systems that can be represented mathematically by a state equation in a number of state variables. Starting from some initial conditions, the system's state vector follows a trajectory with time that is confined to some closed subspace of the total available state space. The dynamic attractor, to which the trajectory thus converges, is a smooth, non-linear manifold of this state space and defines the true dynamics of the system (Thompson et al., 1995). In mathematical terms for discrete-time systems, the state equation is:

$$\mathbf{x}_{t+1} = \mathrm{f}[\mathbf{x}_t, \mathbf{u}_t] \tag{2}$$

where $\mathbf{x}$ is the state vector, $\mathbf{u}$ the input vector of independent variables and $\mathbf{f}$ the state transition function that maps the temporal evolution of $\mathbf{x}_t$ to $\mathbf{x}_{t+1}$. The output vector of dependent variables of the system is defined as:

$$\mathbf{y}_t = \mathrm{g}[\mathbf{x}_t, \mathbf{u}_t] \tag{3}$$

where $\mathrm{g}(\cdot)$ is a nonlinear function that projects $\mathbf{x}_t$ and $\mathbf{u}_t$ onto the output vector $\mathbf{y}_t$.

In the first part of system identification, the evolution of $\mathbf{x}_t$ is reconstructed from the observed system outputs as defined in section 2.1.3. The remaining steps of system identification focus on approximating $\mathbf{g}$ o $\mathbf{f}$ as $\hat{\mathrm{g}}(\cdot) : \mathbf{x}_t \rightarrow \mathbf{y}_{t+1}$ and validating the model.

2.1.1. Model class of $\hat{g}(\cdot)$

Several non-linear empirical model classes have been proposed as functional approximations for $g(\cdot)$. Examples are linear and non-linear polynomial regression, non-linear piecewise regression, regression trees (CART), multi-adaptive regression splines (MARS), kernel-based models such as radial basis function (RBF) neural networks as well as multi-layer perceptron (MLP) neural networks. MLP networks and RBF networks are strong functional approximations (Judd and Mees, 1995), because the optimal approximation error grows more slowly with dimension than for weak functional approximations. Examples of weak approximations are global and local linear approximations as well as global polynomials. Since the class of dynamic systems under discussion requires strong functional approximations, we shall focus on MLP networks and RBF networks. These methods have the added advantage that they are relatively easy to apply and are well supported by large commercial software systems, such as Matlab 4 and 5, G2 with NeurOn-line and Process Insights in the chemical process industries, and Neuroshell 2 in the financial markets. Mathematically the sets of MLP and RBF model structures can be defined respectively as:

$$\mathcal{M}^*_{FF}=\{\mathcal{M}_{FF}(\theta)|\theta \in D_M \subset \Re^d\} \tag{4}$$

$$\mathcal{M}^*_{RB}=\{\mathcal{M}_{RB}(\theta)|\theta \in D_M \subset \Re^d\} \tag{5}$$

where θ is the parameter vector of a model, d the order of the model, and D_M the set of possible model parameters. $\mathcal{M}_{FF}$ or $\mathcal{M}_{RB}$ denotes a model structure while $\mathcal{M}_{FF}(\theta)$ or $\mathcal{M}_{RB}(\theta)$ indicates a specific model for the estimated parameter vector θ.

The set of MLP model structures, $\mathcal{M}^*_{FF}$, are currently often implemented as non-linear regressors and classifiers (Rumelhart et al., 1994). Provided that the input space is carefully selected and the topology correctly specified, a MLP model, $\mathcal{M}_{FF}(\theta)$, can successfully simulate or predict multidimensional non-linear data (Funahashi, 1989). A MLP model structure is specified in terms of the model class, the topology and the nodal transfer (activation) function of each layer. The network is formed by interconnected nodes arranged in layers. Each node is a numerical processor, with a linear or non-linear transfer function. Examples of often used non-linear transfer functions are the bipolar sigmoidal or hyperbolic tangent function, of the form $\phi(\cdot) = [1-\exp(\cdot)]/[1+\exp(\cdot)]$ or the unipolar sigmoidal or logistic function, $\phi(\cdot) = 1/[1+\exp(\cdot)]$. Weights are assigned to the connections, which carry the output

of nodes forward from one layer to the next. In addition, a layer can also incorporate a bias constant. The weights and biases constitute the model parameters and are estimated by so-called training algorithms. The network topology typically comprises a linear input layer, one or two hidden non-linear layers, and a linear output layer. The number of nodes in the input layer is equal to the dimension of the input space, while the size of the output layer is equal to the dimension of the output space. A typical MLP network topology is shown in Figure 1. In this figure **X** and **Y** are the input and output spaces respectively, while w indicate a weight coefficient and b are a bias vector. For example, ${}_1w_{11}$ means the weight factor for the first input component to the first hidden node, while ${}_2w_{11}$ indicates the weight factor for the output from the first hidden node to the first node in the output layer.

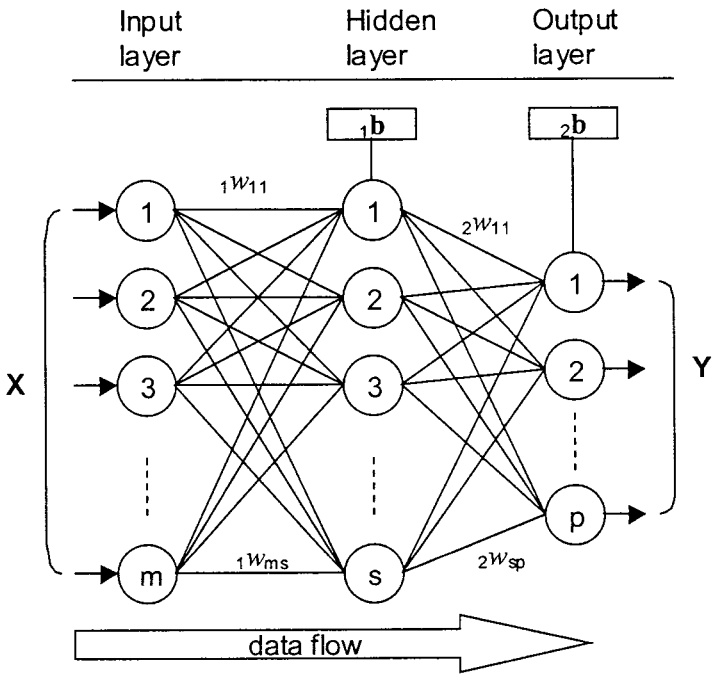


Figure 1 Schematic representation of a typical MLP network topology.

Variations on $\mathcal{M}^*_{FF}$ are the partially and fully recurrent networks. An example of the partially recurrent network is the Elman network. The output from the hidden layer of an Elman network is fed back to a set of extra input nodes, the so-called context nodes. Thus the network is able to reflect in its parameter space a diminishing history of the data, thereby gaining a capability to extract dynamic information from data. The fully recurrent network - a network with all nodes interconnected - is not applicable to work in this dissertation and therefore not treated here.

The set of RBF networks, $\mathcal{M}^*_{RB}$, offer an alternative to $\mathcal{M}^*_{FF}$ and were originally applied to strict interpolation in multidimensional space (Powell, 1987; Broomhead and Lowe, 1988). These networks have a superficially similar structure to MLP networks, except that the hidden layers tend to be much larger and often consist of Gaussian, rather than sigmoidal transfer functions. A hidden layer of radial basis function kernels with fixed parameters are placed at locations along the trajectory defined by the set of data vectors. Typical kernels are the Gaussian function, $\phi(\cdot) = \exp(-\|x-c_i\|/\beta^2)$, the thin-plate-spline function (Chen et al., 1991), $\phi(x) = \|x-c_i\|^2 \log(\|x-c_i\|)$, and multiquadratics (Hardy, 1971), $\phi(\cdot) = (x^2 + c^2)^{1/2}$, $c > 0$, $x \in \Re$, where c is the location center of the kernel and β the Gaussian spread coefficient. Kernel spread indicates the width of the kernel, that is, it determines the range of data over which a kernel is activated. The output layer is a linear combiner of the radial basis function outputs and only these connection weights are adjustable after placement of the kernels. The crux of selecting a $\mathcal{M}_{RB}$ lies in placing of the kernels (Judd and Mees, 1995) and is further discussed in section 2.3.5. Figure 2 shows an example of the placement of radial basis functions on a time series by a RBF network.

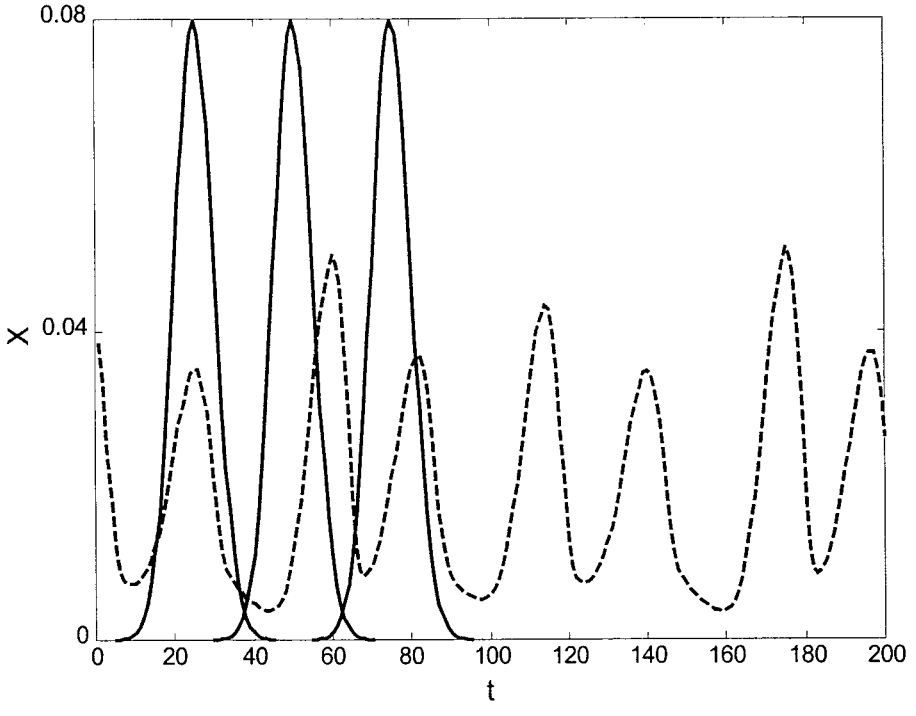


Figure 2 **Typical placement of radial basis functions on a time series, using an RBF network. Solid lines indicate RBF kernels and dotted lines, the data.**

An enhancement of $\mathcal{M}^*_{RB}$, the pseudo-linear radial basis function model, $\mathcal{M}^*_{PL}$ has been proposed by Judd and Mees (1995) and contains a combination of linear terms and Gaussian radial basis function terms.

2.1.2. Model order

Model order is defined as the number of model parameters of a model structure (Ljung, 1987). For $\mathcal{M}^*_{FF}$ and $\mathcal{M}^*_{RB}$, the model order depends on the dimension of input and output spaces, as well as the number of nodes in the hidden layer. Determining the order of a non-linear model can be approached in two ways (Judd and Mees, 1995):

a) approximating $\mathbf{g} \circ \mathbf{f}(\cdot)$ by determining the optimal estimation of model parameters of a certain model structure of specified order.

b) approximating $\mathbf{g} \circ \mathbf{f}(\cdot)$ by determining the optimal combination of model parameters of a preferred subclass of model structures.

According to the first approach, model order is usually determined iteratively by testing several models of increasing order for generalization against the Sum-Square-Error (SSE) norm, defined in section 2.3.1. In the presence of noise, it is possible to overfit by implementing a model of too high an order. Such a model will fit the training data well, but not generalize well, because the model also partially represent features of the noise component.

The second approach starts with a subclass of model structures and then optimizes model order by calculating, for example, Rissanen's minimum description length (MDL) for each model structure (Judd and Mees, 1995). According to this approach, the model parameters and model error are encoded as a bit stream of information. A more complex model will require more bits to encode than otherwise and so will a larger modelling error. The model structure corresponding with the lowest MDL is therefore optimal. This method presents a formalized structure to determining model order, as opposed to the first rather ad hoc approach.

Both above approaches are prominent in system identification and therefore applied in this thesis.

2.1.3. Parameterization

For linear models parameterization means selection of a certain state space representation (Ljung, 1987). For the class of non-linear state space models, parameterization introduces the concept of state space reconstruction.

Let the time series, $y_t = \mathbf{h}(\mathbf{x}_t)$, be the scalar observation of the output of a nonlinear state space system at time t. According to Takens (1981), one can reconstruct an equivalent representation of the system state space from a time series observation, $y \in \Re^n$, under the condition that the observation function $\mathbf{h}(\cdot)$ is smooth. Such a reconstruction is called an embedding of the observed time series by way of delay coordinates (equivalent state variables). The number of these coordinates is the embedding dimension, m and the time delay, k (in multiples of sample period) is the lag between each coordinate. A brief discussion of the theoretical background of the embedding of time series can be found in Osborne and Provenzale (1989).

The optimal time lag between the delay coordinates is usually determined by the average mutual information criterion (Frazer and Swinney, 1986), while the optimal number of coordinates is typically calculated using the method of false nearest neighbours (Kennel et al., 1992). Mutual information is an information statistic that estimates the probability to find a measurement again, given that the same measurement has been already been made. This statistic is calculated among all elements of the time series. The time lag is fixed heuristically at the point of the first minimum of mutual information for the time series. A full description of the technique of average mutual information appears in section A.1.

The method of false nearest neighbours involves iterating the embedding of the time series in space of increasing dimension. While unfolding the attractor in space of increasing dimension, the embedded points that are true neighbours can be progressively distinguished until, after reaching the optimal embedding dimension, no more additional false neighbours are discovered. False neighbours appear only because one views the attractor in space of too small a dimension, thereby mistaking two points for being neighbours. The nearness is expressed as the Euclidean distance between two points.

In an alternative approach both embedding lag and dimension can be calculated simultaneously by the method of false strands (Kennel et al., 1992), which is more robust against the effects of both measurement and dynamic noise in the data. This method considers neighboring strands of data instead of only single embedded points and determine whether all

points on these strands are true neighbors in a similar fashion as for pairs of single points, discussed above. Both false nearest neighbours and false strands are discussed in section A.3.

In mathematical terms, let $\mathbf{y} = [y_1, y_2, y_3, \ldots, y_n]$ be the observation vector of the output of a dynamic system. According to Takens (1981), an optimal embedding of $\mathbf{y}$ can be expressed as $\mathbf{x}_i = [y_{i+k(m-1)}, y_{i+k(m-1)-1}, y_{i+k(m-1)-2}, \ldots, y_i]$. The set $\{\mathbf{x}_t \in \Re^m, t = 1 \ldots n\}$ forms the trajectory of the embedding vector in state space and approximates the dynamic attractor asymptotically as $n \to \infty$. With reference to the state equation (**2**), $\mathbf{x}_t$ is the embedding vector, while $\mathbf{u}_t$ is the vector of independent variables. The reconstructed attractor is an implicit parameterization of the system state transition.

After reconstructing the dynamic attractor of the system, a one-step prediction model, $\mathcal{M}(\theta)$, is selected as $\hat{g} : \mathbf{x}_t \to \mathbf{y}_{t+1}$. Therefore full parameterization of non-linear state space systems is defined in terms of both $\{\mathbf{x}_t \in \Re^m\}$ and the parameter vector , θ, of the selected model structure $\mathcal{M}$.

2.2. Data acquisition

Since empirical system identification relies on the availability of sufficient, representative observations of the system, data acquisition is of paramount importance. A number of factors has to be considered: the dependent variables, independent variables, sampling period and the number of records that defines a stationary data set.

2.2.1. Selection of independent and dependent system variables

Independent and dependent variables (also called input and output variables) are selected based on *a priori* knowledge of the system. In order to determine which of several independent variables are correlated with the chosen dependent variables, one can apply second order statistics in the form of cross-correlation analysis to the observation space. While this is conclusive for linear systems, it is often misleading for non-linear systems (Abarbanel, 1996). Average cross mutual information (AXMI) at zero lag is a better indicator of non-linear cross-correlation. AXMI is closely related to AMI (refer to sections A.1 and A.2 for details on the definition of AMI and AXMI, respectively).

Where significant cross-correlation is suspected among the independent variables, the input space can not be determined only on the basis of strong direct correlation of independent

variables with the dependent variable. Rather, the full set of independent variables should be separated using, for example, principal component analysis and projected onto those principal components that declare a specified percentage (e.g. 99%) of the input variance. A more advanced separation method is Hyvärinen's independent component analysis (Hyvärinen, 1999), described in more detail in chapter 5. These techniques often result in a handsome reduction in input space dimensionality and therefore also model complexity. Mathematically the projection is expressed as:

$$\mathbf{S} = \mathbf{WX} \tag{6}$$

where **X** represents the independent variables, **W** the separation matrix and **S** the statistically independent components.

2.2.2. Sampling frequency

Highly non-linear systems appear random under linear analysis, even though they are actually deterministic (Farmer and Sidorowich, 1987). Since these systems are not periodic or even quasi-periodic, linear analysis, such as Fourier transforms, yields results similar to the linear analysis of broadband noise. The selection of a proper sampling frequency is therefore not straightforward. However, it still is possible to apply the Nyquist principle to determining sampling frequency, by stating that the highest sampling frequency should be between 2 and 10 times the frequency of the highest order interesting dynamics (Ljung, 1987). Interesting dynamics can be defined as that dynamical behavior of the observed system that one will attempt to model with a particular system parameterization and selected model structure. It is also the level of detail in the reconstructed dynamic attractor that one will attempt to describe using a particular model structure. In practice, this decision on sampling frequency is made in an iterative and rather subjective manner.

2.2.3. Classification of data

A major problem with empirical systems is to determine *a priori* whether deterministic dynamics underlie the data in the first place. To make matters worse, non-linear identification algorithms that calculate the system dimension from a time series do not always return an infinite value for stochastic processes (that have infinite dimension) as would be expected. Osborne and Provenzale (1989) have shown that stochastic data with power law spectra also yield correlation dimensions with finite values, so that statistics characterizing the

dimensionality of a system cannot be reliably used for the identification of determinism. Some stochastic processes generate so-called colored noise that have fractal curves in state space, but no dynamic attractors (Osborne and Provenzale, 1989). Non-linear identification algorithms cannot distinguish between fractal curves and fractal attractors. This can be problematic since reliable classification of the data as a first step in system identification is important, otherwise the resulting model will not generalize beyond the training data set.

A statistical approach to data classification is the method of surrogate data (Takens, 1993; Theiler and Pritchard, 1996; Theiler and Rapp, 1996). This method involves a null hypothesis against which the data are tested, as well as a discriminating statistic. The data are first assumed to belong to a specific class of dynamic processes. Surrogate data are subsequently generated, based upon the given data set, by using the assumed process. An appropriate discriminating statistic is calculated for both the surrogate and the original data (Theiler et al., 1992). If the calculated statistics of the surrogate and the original data are significantly different, then the null hypothesis that the process that has generated the original data is of the same class as the system that has generated the surrogate data, is rejected. By means of a trial-and-error elimination procedure, it is then possible to get a good idea of the characteristics of the original process. Refer to section A.7 for details.

2.2.4. Stationary data sets

Since any model is ultimately limited in its ability to extract information from data by the total information content in the data, it is imperative for successful implementation of a non-adaptive, global model to train on a representative, stationary data set. There are several means to determine stationarity of data, depending on the class of data. Random data can be tested for stationarity in terms of the invariance of first and second statistical moments of the data. Deterministic data extracted from a forced linear system should contain the longest forcing period to be regarded stationary. In non-linear terms, a data set $\{y \in \Re^n\}$ will be stationary if and only if it is a sufficient approximation of the dynamic attractor when properly embedded:

$$\{\mathbf{x}_t \in \Re^m\} \tag{7}$$

where $\Re^m$ is the m-dimensional space of real numbers. Simple, global statistics such as mean or variance are often unable to reliably indicate stationarity because they are not closely related to the geometric characteristics of the dynamic attractor traced by the state vector, $\mathbf{Z}$ =

[**X Y**], in state space. Invariant properties of dynamic attractors such as correlation dimension and Lyapunov exponents can determine non-linear stationarity because they are only invariant when the data set from which the attractor has been reconstructed, is stationary. Unfortunately these properties are often difficult to estimate reliably (Kennel, 1997; Eckmann and Ruelle, 1992; Parlitz, 1992).

Kennel (1997) reliably determined sufficient stationarity in non-linear dynamic data sets by way of a statistical test on a nearest neighbor analysis of embedded data. This method, however, involves embedding of the observed time series and, unfortunately, time series embedding of data with a large measurement and dynamic noise content can lead to non-optimal attractor reconstruction. Kennel reduced the negative influence of noise by determining nearest strands in an embedded time series instead of only nearest neighbours. This alternative approach enabled him to improve the reliability of finding a suitable embedding for his stationarity test.

2.3. Identification criteria and parameter estimation

Parameter estimation methods use a data set collected from a system to estimate the model parameters. A "good" model will produce acceptably insignificant prediction errors when applied to the observed data. The prediction error r_t is defined as:

$$r_{t,\theta} = y_t - \hat{y}_{t,\theta} \,. \tag{8}$$

where $\hat{y}_t$ is system output as estimated by the model using parameter vector θ and y_t is the observed system output. In order to quantify "insignificant errors", there are two approaches among others. One approach is based on a scalar criterion function that measures r_t. Another approach is to demand that r_t be uncorrelated with a given data sequence.

There are several parameters estimation methods, such as prediction error methods and the maximum likelihood method (Ljung, 1987). The MLP and RBF model structures that are implemented in this dissertation traditionally apply prediction errors methods.

When model performance is optimized by minimising the autocorrelation of the prediction error and the cross-correlation between prediction error and original output data, a simple whiteness test is sufficient for the identification of linear systems. However, for identification of non-linear systems the autocorrelation and cross correlation sequences cannot give any evidence of remaining nonlinear relationships, since any process can always be considered to

be a linear process with respect to its second-order statistics. Higher-order cumulants can give such evidence. For example, the third-order cumulant can be used to test for Gaussianity of the simulation error. Hinich (1982) proposed a zero-skewness test as a quantitative test for normality of a stationary data sample. If the prediction error passed this test, there is no significant linear or non-linear correlation content, which indicates that the model interpreted all the dynamic content in the original data. Refer to section A.5 for details.

The following aspects regarding identification criteria and parameter estimation are addressed in this section:

1. Model fitness norm
2. Pre-filtering
3. Outlier detection
4. Prediction horizon
5. Numerical estimation procedures.

2.3.1. Norm

In the estimation of $\mathcal{M}_{FF}$, $\mathcal{M}_{RB}$ and $\mathcal{M}_{PL}$ parameters using a prediction error method, the norm is the Sum Square Error of the model output and is defined in terms of θ and the data set $\mathbf{Z}$ as:

$$V_{\theta,Z} = \frac{1}{n}\sum_{t=1}^{n}\frac{1}{2}r_{t,\theta}^2 \tag{9}$$

where $\mathbf{Z} = [\mathbf{X}\ \mathbf{Y}]$, with $\mathbf{X} = \{\mathbf{x}_t \in \Re^m\}$ being the input space and $\mathbf{Y} = \{\mathbf{y}_t \in \Re^p\}$, the output space.

2.3.2. Noise reduction

Observations of system output can be contaminated with measurement noise as well as dynamic noise. In the presence of measurement noise ε_t, the output equation (3) can be expressed as:

$$\mathbf{y}_t = g[\mathbf{x}_t, \mathbf{u}_t] + \varepsilon_t \tag{10}$$

Measurement noise is caused by measurement inaccuracy, which reduces the signal to noise ratio of observations and impairs the ability to extract dynamic information through the simulation model.

Dynamic noise, δ_t can be mathematically expressed as:

$$\mathbf{x}_{t+1} = \mathrm{f}[\mathbf{x}_t, \mathbf{u}_t] + \delta_t \tag{11}$$

Dynamic noise results from inherent disturbances of the process that generates the dynamic orbit of the state vector in state space and tends to increase the model dimension or complexity. The objective of noise reduction is signal separation without *a priori* knowledge of the underlying noiseless dynamics or the noise distribution. For linear systems, filtering is a common solution to this problem. As pointed out earlier, non-linear data appear random under linear analysis, therefore linear filtering can seriously impair model validity by removing interesting high order dynamic information with the noise. Adaptive Moving Average filters, Volterra filters, bi-linear and multi-linear filters all address the filtering of non-linear systems with some success. Rauf and Ahmed (1997) presented a class of non-linear adaptive filters based on successive linearization (Rauf, 1997) for predictive modeling of chaotic systems.

Geometric filtering, another form of non-linear filtering, first embeds the observed time series in a high-dimensional phase space and then fits local linear models on the resultant geometric structure. For example, Kostelich and York (1988) demonstrated a method whereby points in a local neighbourhood on an attractor, reconstructed from the time series, were used to find a local approximation of the dynamics. The approximations were then used collectively to produce a new time series which is dynamically more consistent with the reconstructed attractor. Mees and Judd (1993) warned that geometric filtering can easily be misused through misunderstanding of the techniques. Since geometric filtering processes impose locally linear models on data, these processes can force the data to fit a linear model inappropriately if iterated enough times.

2.3.3. Outlier detection

Failure to remove outliers that result from systematic errors in process measurements can lead to gross distortion in models or decision support systems derived from these data. Yet, outliers that arise from actual process dynamics could reveal significant insight into the process mechanisms, once identified. Outliers in data are not traditionally treated as part of system identification and is usually dealt with separately. In this dissertation the handling of

outliers is included in the formal system identification methodology in the same sense as noise reduction. The handling of outliers is investigated and demonstrated in Chapter 4.

2.3.4. Prediction horizon

A dynamic system can be predicted in two basic ways: one-step or free-run. During one step prediction the input space is updated with the current observation and the model predicts the output one sampling step ahead. A variation on one-step prediction is long-step prediction, where the input stays fixed at the initial value and the model predicts output directly for any chosen multiple of sampling period ahead. During free-run prediction the input space is updated with the predicted output and the model predicts the output one sampling period ahead.

Non-linear deterministic systems, when chaotic, can have a finite prediction horizon regarding free-run prediction. Starting from any initial state within the basin of attraction, any free-run prediction for such a system will become ultimately unstable outside the basin of attraction that contains the dynamic attractor (Abarbanel, 1994). This rule applies unless the dynamics have been properly reconstructed from data observed in several adjacent basins of attraction. Even within a specific basin the highly non-linear character limits the free-run prediction horizon. For chaotic non-linear systems, the largest positive Lyapunov exponent of the attractor indicates the rate of divergence of neighboring trajectories and gives an indication of the upper limit of prediction accuracy (Abarbanel, 1994, 1996). Non-linear systems that are not chaotic do not have positive Lyapunov exponents and therefore have no fundamental upper limit for prediction.

We are interested in models that iteratively predict system output one step ahead. These models may also be used for free-run prediction, that is, the embedding, $\mathbf{x}_i$, is updated with predictions instead of observations while the model is running. The free-run prediction accuracy is ultimately limited by the size of the largest positive Lyapunov exponent of the system. Refer to section A.4 for details on the definition of Lyapunov exponents.

The number of Lyapunov exponents for a system equals the embedding dimension. Each exponent is invariant for the system. The positive exponents indicate how fast any certainty about a point on the state trajectory will be replaced by uncertainty. The negative exponents indicate the rate at which deviations from the attractor will dampen out and converge to the attractor.

Starting a prediction from index i and predicting in L steps until $i+L$, an upper bound on L, called the instability horizon (Abarbanel, 1996), is approximately given as:

$$t_L = \frac{\tau_s}{\lambda_1} \tag{12}$$

where τ_s is the sampling period of the observer function and λ_1 the first Lyapunov exponent. The initial error at index i scales as a function of the number of iterations L to predict from index i to $i+Lk$ and can be expressed in terms of the largest Lyapunov exponent by:

$$\varepsilon_L = e^{L\tau_s\lambda_1} \tag{13}$$

2.3.5. Parameter estimation

Parameter estimation is the process whereby information is extracted from the observed data and a model parameter vector estimated as:

$$\theta = \arg\min_{\theta \in D_{\mathcal{M}}} V_{\theta,Z} \tag{14}$$

where θ is the argument that minimizes the chosen norm, as defined in section 2.3.1.

There exist several algorithms to estimate the parameters of $\mathcal{M}_{FF}$ and all determine the gradient of the performance function (square error of network output). The simplest algorithm uses gradient descent to optimize the network weight and bias arrays. Conjugate gradient methods improve on basic gradient descent in terms of convergence speed. Newton's method improves again on conjugate gradient methods, but requires calculation of second derivatives (the Hessian), which is complex and expensive to calculate for MLP networks. Quasi-Newton methods avoid this and one of these, the Levenberg-Marquardt algorithm (Levenberg, 1944; Marquardt, 1963), approximates the Hessian as $\mathbf{H} = \mathbf{J}^T\mathbf{J}$, where $\mathbf{J}$ is the Jacobian, containing first derivatives of network output error with respect to network weights and biases. The gradient can be computed as $\mathbf{g} = \mathbf{J}^T\mathbf{e}$, where $\mathbf{e}$ is the network output error. The weights are updated at the end of a training iteration as $\mathbf{w}_{k+1} = \mathbf{w}_k - [\mathbf{J}^T\mathbf{J} + \mu\mathbf{I}]^{-1}\mathbf{J}^T\mathbf{e}$. The Levenberg-Marquardt algorithm converges faster than most other training algorithms (Hagan and Menhaj, 1994).

Parameter estimation for models structures $\mathcal{M}_{RB}$ and $\mathcal{M}_{PL}$ involves placement of radial basis function kernels with the specified spread coefficient after which the weights of the linear

output is determined by minimizing the sum-square-error norm. Kernel placement is traditionally done by a number of methods. Examples are placing of kernel centers at randomly selected data points, at random points in a bounding box, and at k-means cluster centroids. Chen et al. (1991) proposed an improved parameter estimation algorithm for the placement of $\mathcal{M}_{RB}$ kernel centers. For $\mathcal{M}_{PL}$, Judd and Mees (1995) found it advantageous to add Gaussian noise to the data so as to generate so-called chaperon centers that lie at the perimeter of the data region.

2.4. Model validation

Proper validation ensures the reliable application of the model on new observations from the same process. Model validation is usually based on statistical tests, such as the significance of R^2 or RMS criteria derived from actual and predicted values or empirical methods such as cross-validation or hold-out. Although statistical tests of model validity are the preferred approach (Rivals and Personnaz, 1999), they can unfortunately only be applied when the statistical properties of the system are known. Also, these statistics are static, global measures of correspondence between model and observation that do not evaluate local dynamic correspondence. Likewise, empirical methods such as cross-validation can perform poorly when used in the selection of linear models (Zhu and Rohwer, 1996; Goutte, 1997) and is highly unlikely to perform any better with non-linear models.

Model validation is often based on one-step ahead prediction, which is not necessarily a good indicator of the ability of the model to generalize the underlying (dynamic) process represented by the data (Zhu and Rohwer, 1996). The residue between prediction and observation of a linear system will be nearly white (linearly uncorrelated) if the model declares almost all information in the observations (Ljung, 1987). In the case of a non-linear system, non-linear correlation may still exist between simulation error and observation, which a whiteness test will not reveal.

A free-run prediction in which the model has to predict the long-term future behavior of the system, while being updated with succeeding predicted outputs instead of observed outputs, is a considerably more rigorous test of the validity of the model (Small and Judd, 1998). To achieve this, one has to generate a free-run time series with the model, reconstruct the dynamic attractor from these predicted values and characterize the attractor with some discriminating statistic. The dynamic attractor for the actual system is likewise reconstructed from the observed time series, and characterized. The reliability of the model can thus be assessed

systematically by comparing the discriminating statistic of the model with that of the experimental data. Refer to section A.7 for details on this so-called surrogate technique.

3 IDENTIFICATION OF NON-LINEAR SYSTEMS FROM A TIME SERIES.

The identification of the underlying dynamics of many process systems from experimental data is typically complicated, owing to a mixture of influences that cause erratic fluctuations in the time series. These influences can be notoriously difficult to disentangle. The development of process models is usually subject to considerable human judgement and can therefore be very unreliable. This is especially the case when the model priors are unknown and the model is validated empirically, such as with cross-validation or holdout methods. In this chapter, it is consequently shown by way of a case study that more reliable empirical identification of the large class of nonlinear state space systems is possible by applying a methodological system identification framework, as formulated in Chapter 2.

3.1. Methodology

The identification methodology formulated in Chapter 2 can be applied to a one-dimensional time series via the following procedure:

a) Classification of data with the help of a surrogate data method, as described in section 2.2.

b) Selecting a system output variable, $\mathbf{y}_t$, model structure $\mathcal{M}$ from $\{\mathcal{M}^*_{FF}, \mathcal{M}^*_{RB}\}$, model order d, parameterization by embedding of $\mathbf{y}_t$ as well as specification of model parameters θ, as described in section 2.1

c) Estimating model parameters, θ, as described in section 2.3.

d) Validating the model, $\mathcal{M}(\theta)$, using free-run prediction and non-linear surrogate data, as described in section 2.4.

Two aspects of system identification are not included in the above procedure. These are:

a) Handling of outliers in the data, for which a method will be introduced and demonstrated in Chapter 4.

b) Selection of a stationary length, for which a method will be introduced and demonstrated in Chapter 5.

3.2. Empirical identification of an autocatalytic process

This case study concerns an autocatalytic process in a continuous stirred tank reactor originally considered by Gray and Scott (1983, 1984) and subsequently investigated by Lynch (1992). The system is capable of producing self-sustained oscillations based on cubic autocatalysis with catalyst decay and proceeds mechanistically as follows.

$$\begin{aligned} \mathrm{A} + 2\mathrm{B} &\rightarrow 3\mathrm{B}, \quad -r_{\mathrm{A}} = k_1 c_{\mathrm{A}} c_{\mathrm{B}}^2 \\ \mathrm{B} &\rightarrow \mathrm{C}, \quad r_{\mathrm{C}} = k_2 c_{\mathrm{B}} \\ \mathrm{D} + 2\mathrm{B} &\rightarrow 3\mathrm{B}, \quad -r_{\mathrm{D}} = k_3 c_{\mathrm{D}} c_{\mathrm{B}}^2 \end{aligned} \tag{15}$$

where A, B, C and D are the participating chemical species and k_1, k_2, k_3 the rate constants for the chemical reactions. This process is represented by the following set of ordinary differential equations.

$$\begin{aligned} \frac{dX}{dt} &= 1 - X - aXZ^2 \\ \frac{dY}{dt} &= 1 - Y - bYZ^2 \\ \frac{dZ}{dt} &= 1 - (1+c)Z + daXZ^2 + ebYZ^2 \end{aligned} \tag{16}$$

where X, Y, and Z denote the dimensionless concentrations of species A, B and D, while a, b, c denote the Damköhler number for A, B and D respectively. The ratio of feed concentration of A to B is denoted by d and the similar ratio of D to B by e. The process is chaotic, with a well-defined attractor for specific ranges of the two parameters, d and e. For the settings: a = 18000; b = 400; c = 80; d = 1.5; e = 4.2, and initial conditions $[0,0,0]^{\mathrm{T}}$, the set of equations was solved by using a 5th order Runge Kutta numerical method over 100 simulated seconds. This gave approximately 10 000 observations, which were resampled with a constant sampling period of 0.01 s. The Y state was taken as the output variable. Figure 3 shows the attractor of the data reconstructed from the process states X, Y and Z.

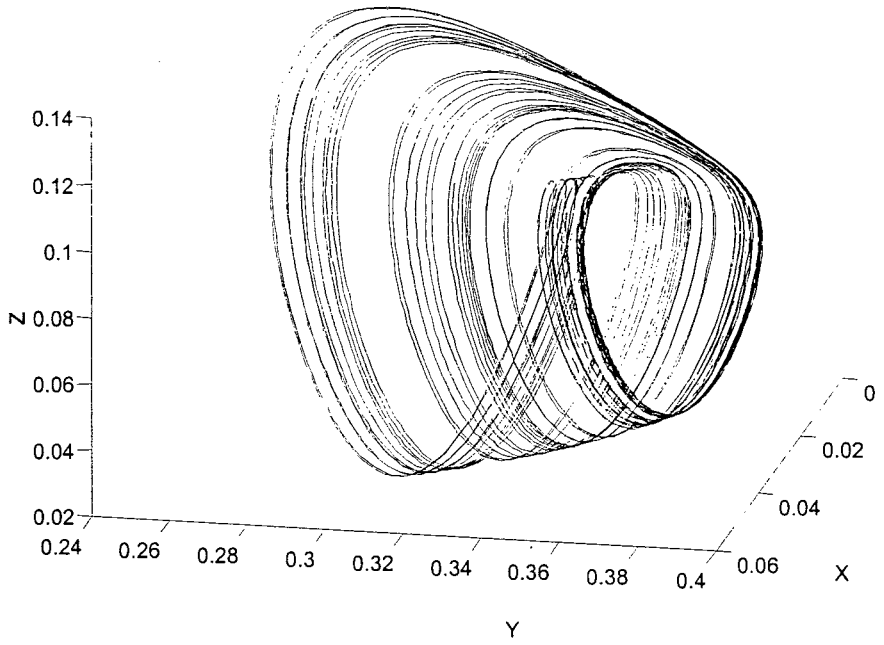
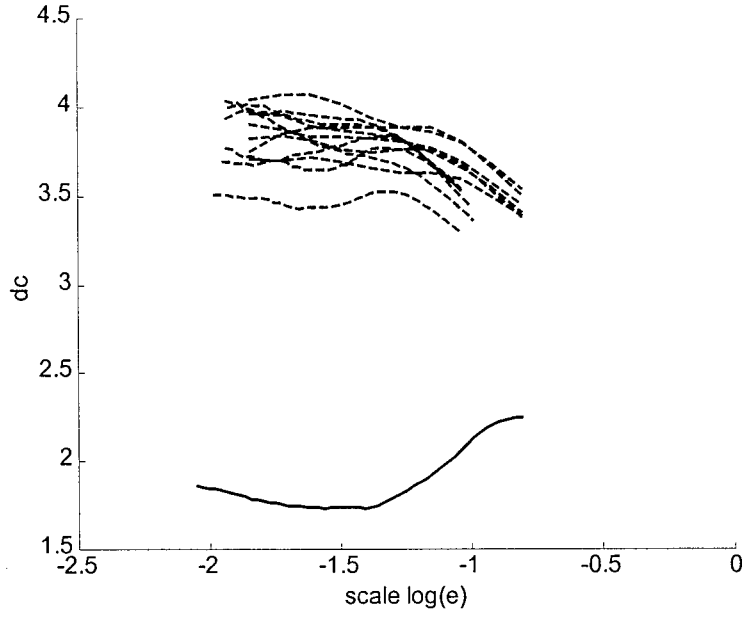


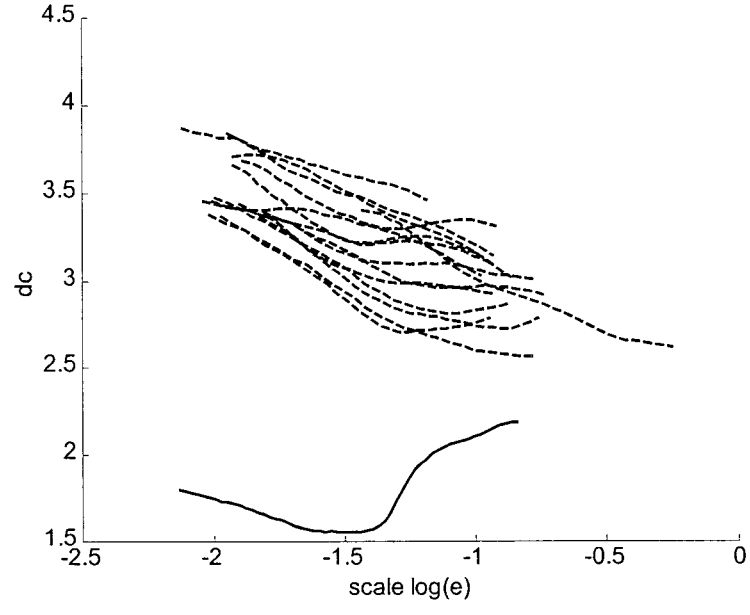
Figure 3 **Attractor of autocatalytic process constructed from process states *X*, *Y*, *Z*.**

Two different data sets were considered in order to assess the effect of the size of the data set on the identification method. The smaller of the two sets consisted of the first 2000 observations of the original data set of 10 000 observations, while the larger of the two sets consisted of the first 8000 observations. In each case the remainder of the data was used to validate subsequent models.

Classification of the data with type 2 surrogates (defined in section A.7) was performed first. This entailed calculation of the correlation dimension for each of the two data sets, as well as for 15 surrogate data sets generated from each data set. The results for both data sets are shown in Figure 4(a) and (b). The deterministic character of the data is evident from these figures. The shape of the correlation dimension curve for the observed correspond with the wide, two-dimensional ribbon shape of the attractor, curved in three dimensional space. The magnitude of the displacement between the correlation dimension curves of the surrogate data and that of the observed data is an indication of the difference in complexity between the random surrogates and the dynamic data. The difference in shape between the data in 4a and 4b is due to the possible non-stationarity of the data in the small data set compared to the larger data set. Stated differently, the correlation dimension algorithm was exposed to a more developed attractor in the larger data set. Embedding parameters were m=4, k = 10.



(a)



(b)

Figure 4 **Correlation dimension (d_c) vs. scale (log(e)) for Y-state (bottom curve) of autocatalytic process and its AAFT surrogates based on (a) the smaller data set, (b) the larger set.**

The next step involved the embedding of each of the training and validation data sets in an appropriate state space. By making use of the method of false nearest neighbours, both the smaller and the larger data set could be optimally embedded in a three-dimensional space (m = 3). Average Mutual Information analysis indicated a time lag, k = 7, between embedding variables. Two non-linear models were subsequently fitted to the data.

3.2.1. Multi-layer perceptron network model

The first two model structures consisted of multi-layer perceptron neural networks, each with an input layer of three nodes, a hidden layer with six bipolar sigmoidal nodes (activation functions of the form $g(\cdot)=[1-\exp(\cdot)]/[1+\exp(\cdot)]$) and a single linear output node. The parameter vectors, θ_1 and θ_2, of both $\mathcal{M}_{FF1}(\theta_1)$ and $\mathcal{M}_{FF2}(\theta_2)$ were estimated with the Levenberg-Marquardt algorithm, based on the smaller and larger data sets respectively. The optimal model order was determined via cross-validation on the test data, for both the smaller and the larger data set. For conciseness, $\mathcal{M}_{FF\theta1}$ and $\mathcal{M}_{FF\theta2}$ will indicate $\mathcal{M}_{FF1}(\theta_1)$ and $\mathcal{M}_{FF2}(\theta_2)$, respectively.

The $\mathcal{M}_{FF\theta1}$ was able to predict the data one-step ahead in the associated validation data set very accurately (R^2 = 0.999), as indicated in Figure 5. $\mathcal{M}_{FF\theta2}$ performed with the same degree of accuracy.

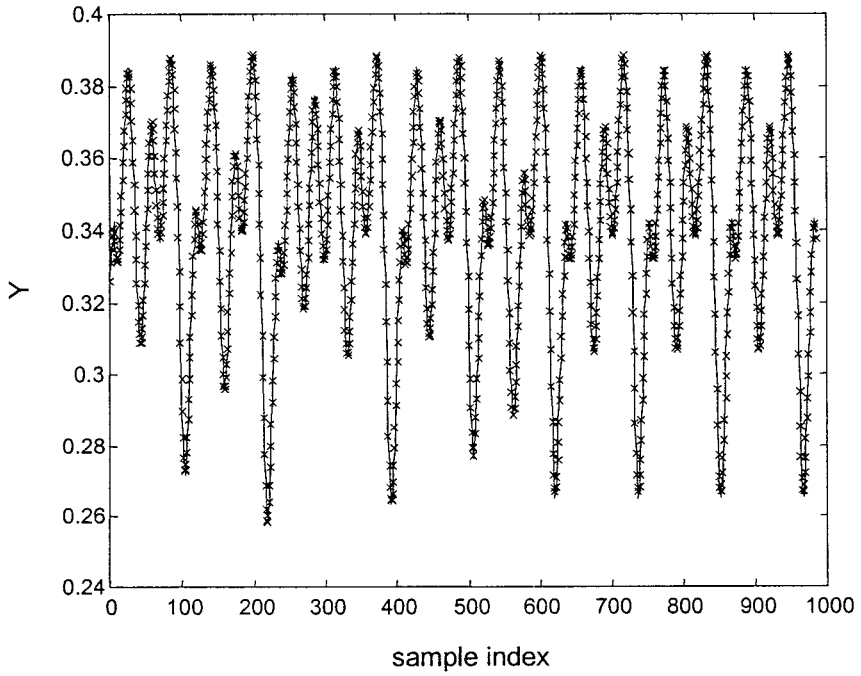
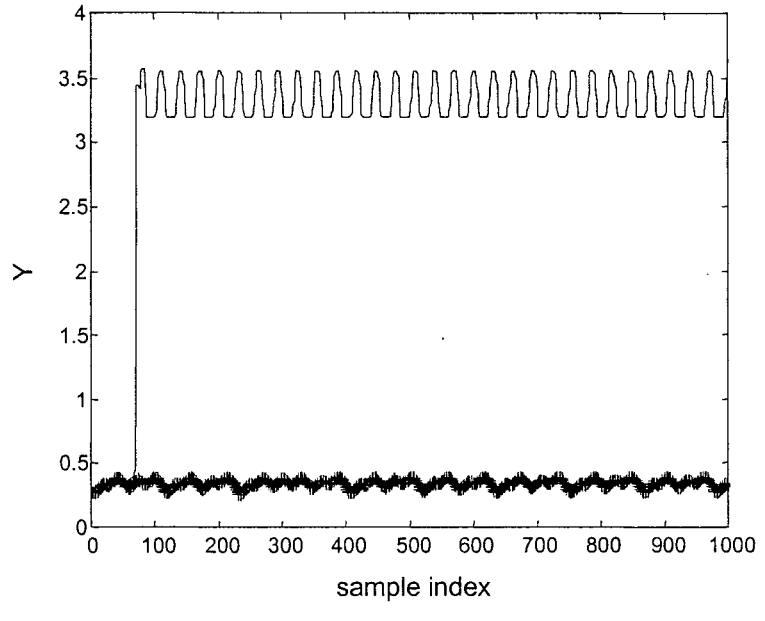


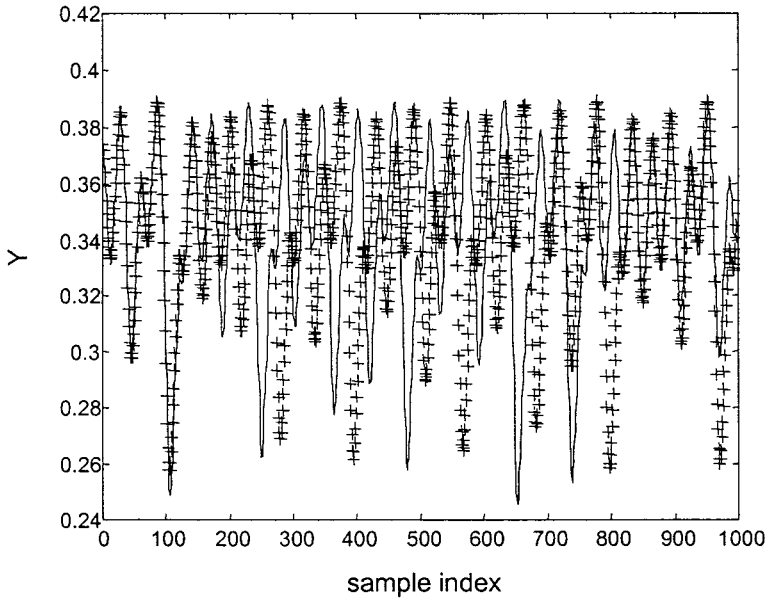
Figure 5 One-step prediction of autocatalytic Y-state (+ marker) vs. Y-state using a MLP network trained on the smaller data set.

The free-run predictions for the two data sets are shown in Figure 6(a) for $\mathcal{M}_{FF\theta1}$ and Figure 6(b) for $\mathcal{M}_{FF\theta2}$. As can be seen from Figure 6(a), $\mathcal{M}_{FF\theta1}$ could predict the data accurately in a free-run mode, up to about the 60th observation, after which it momentarily became unstable, then became a pseudo-periodic oscillation that grossly overestimated the actual values of the

observations. $\mathcal{M}_{FF\theta 2}$ performed significantly better, but after approximately 180 observations the predictions started to deviate significantly, as indicated in Figure 6(b). It is possible that the deviation between the output of $\mathcal{M}_{FF\theta 2}$ and the observed data may be attributed to the chaotic nature of the process, with the model merely being out of synchronization with the process.



(a)



(b)

Figure 6 **Free-run prediction of autocatalytic Y-state with MLP network models (x marker), (a) $\mathcal{M}_{FF\theta 1}$ and (b) $\mathcal{M}_{FF\theta 2}$.**

3.2.2. Pseudo-linear radial basis function model

The set of pseudo-linear radial basis function model structures previously proposed by Small and Judd (1998) was also fitted to the data sets by using variable embedding, Gaussian radial basis functions and an algorithm that optimizes model size via Rissanen's minimum description length (MDL) of the model during each iteration (Judd and Mees, 1995; Small and Judd, 1998a). The variable embedding strategies were restricted to a maximum of three-dimensional embeddings and maximum lag of nine sample periods.

The pseudo-linear radial basis function model structure, $\mathcal{M}_{PL}$, consisted of a combination of linear terms and a number of Gaussian radial basis function terms. The algorithm of Small and Judd (1998) determines the combination and number of these terms by using minimum description length as a criterion. $\mathcal{M}_{PL1}(\theta)$ based on the smaller set used 23 Gaussian kernels, while $\mathcal{M}_{PL2}(\theta)$ based on the larger data set used 18 Gaussian kernels. For conciseness, $\mathcal{M}_{PL\theta1}$ and $\mathcal{M}_{PL\theta2}$ are used instead of $\mathcal{M}_{PL1}(\theta)$ and $\mathcal{M}_{PL2}(\theta)$. Like the MLP network model, $\mathcal{M}_{PL\theta1}$ and $\mathcal{M}_{PL\theta2}$ was able to predict the data one-step ahead in the validation data associated very accurately ($R^2 = 0.999$), as indicated in Figure 7 (for $\mathcal{M}_{PL\theta1}$).

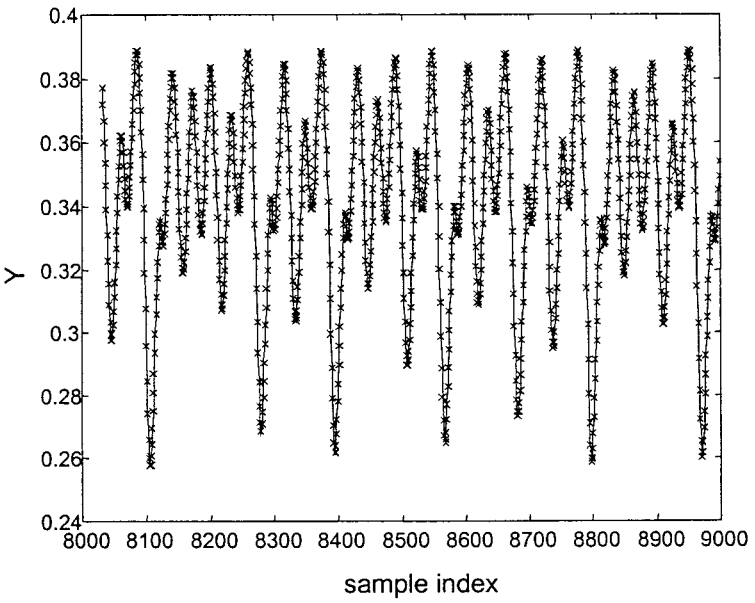
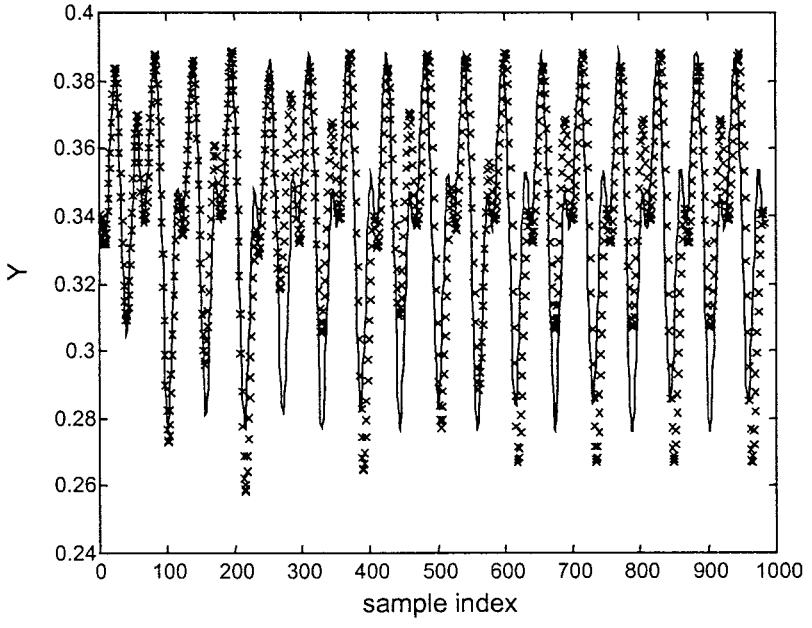


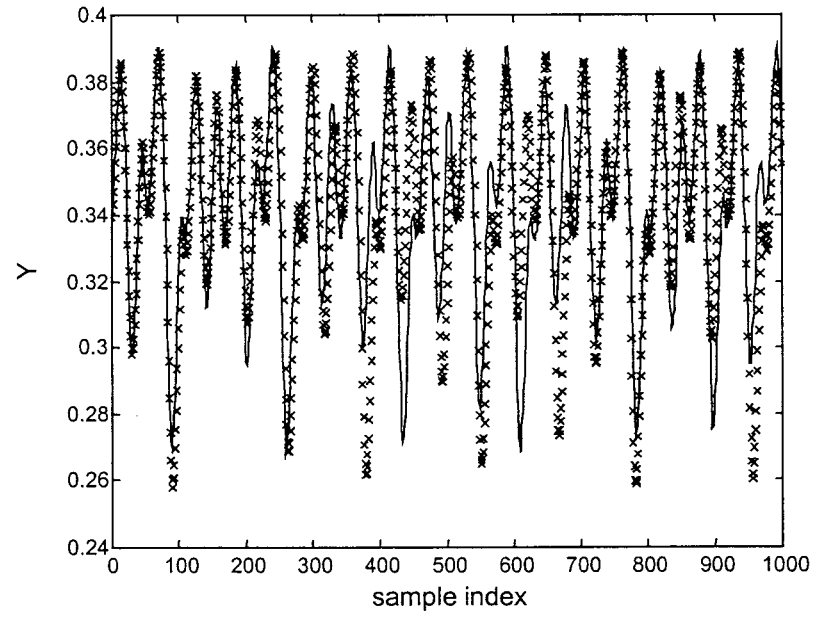
Figure 7 One-step prediction of observed autocatalytic Y-state with $\mathcal{M}_{PL\theta1}$ vs. the observed Y-state (x marker).

The free-run predictions are shown in Figure 8(a), using $\mathcal{M}_{PL\theta1}$ and Figure 8(b), using $\mathcal{M}_{PL\theta2}$. As can be seen from these figures, $\mathcal{M}_{PL\theta1}$ and $\mathcal{M}_{PL\theta2}$ could predict the data more accurately in

free-run mode than $\mathcal{M}_{FF\theta 1}$ or $\mathcal{M}_{FF\theta 2}$. This was especially so for the smaller data set, although $\mathcal{M}_{PL\theta 1}$ left the true attractor after 150 observations. $\mathcal{M}_{PL\theta 2}$ kept up with the attractor until after observation 200. This was at the expense of a much higher processing cost due to the larger data length. Both models managed to approximately follow the tendency of the attractor, but could not quite reach the bottom of the troughs and were also partially out of phase with the observed Y state.



(a)



(b)

Figure 8 **Free-run prediction of observed autocatalytic Y-state vs. Y-state (x marker), for (a) $\mathcal{M}_{PL\theta 1}$ and (b) $\mathcal{M}_{PL\theta 2}$.**

Based on these results, it is clear that analyses of one-step ahead predictions are comparatively poor indicators of the quality of the models and that the free-run predictions provide a better idea of the adequacy of the models representing the dynamics of the system. These analyses can be formalized by comparing the surrogate data derived from the models with the actual data. The results for $\mathcal{M}_{FF\theta 2}$ are shown in Figure 9. The results pertaining to $\mathcal{M}_{FF\theta 1}$ are not shown, owing to the obviously poor quality of the model shown in Figure 6(a). From Figure 9 it is clear, judged by the broken curve at the bottom that represents the data, that the model $\mathcal{M}_{FF\theta 2}$ (solid curves) has not captured the structure of the data completely, except in the large-scale region of the dynamic attractor ($\log \varepsilon_0 > -0.9$). The peculiar dip in the bottom curve just after $\log \varepsilon_0 = -0.9$ is due to numerical instability in the correlation dimension algorithm. In Figure 10(a), it can be seen that $\mathcal{M}_{PL\theta 1}$ has captured most of the large-scale structure ($\log \varepsilon_0 > -1.3$) of the dynamic attractor, except in the small-scale region (that is, the detail of the attractor, or high order dynamics). In contrast, $\mathcal{M}_{PL\theta 2}$ has evidently captured most of the dynamic structure of the data at all scales, as indicated by Figure 10(b) and is overall the best model investigated here. Take note that the correlation dimension curve representing the data is different from that in figure 4b, because it has been calculated for a smaller data set from a different section data (the test data). Embedding parameters were m=8, k=7.

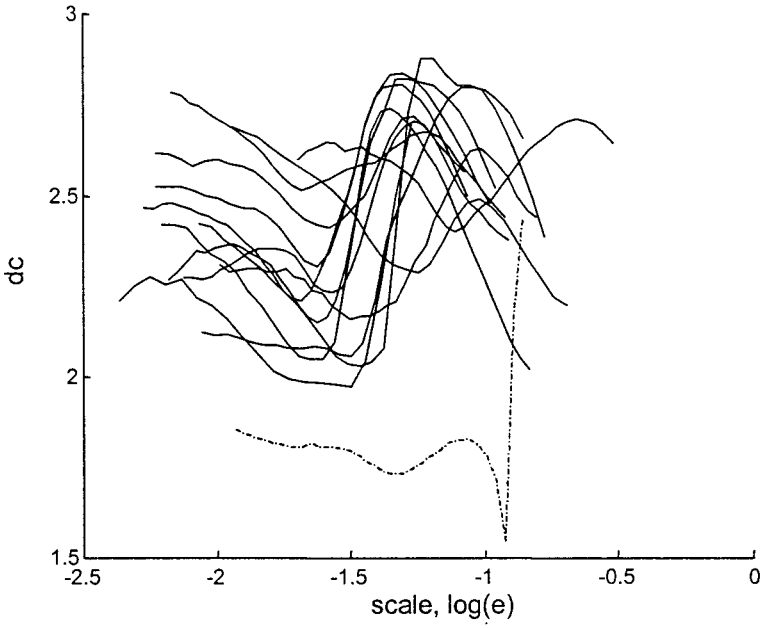
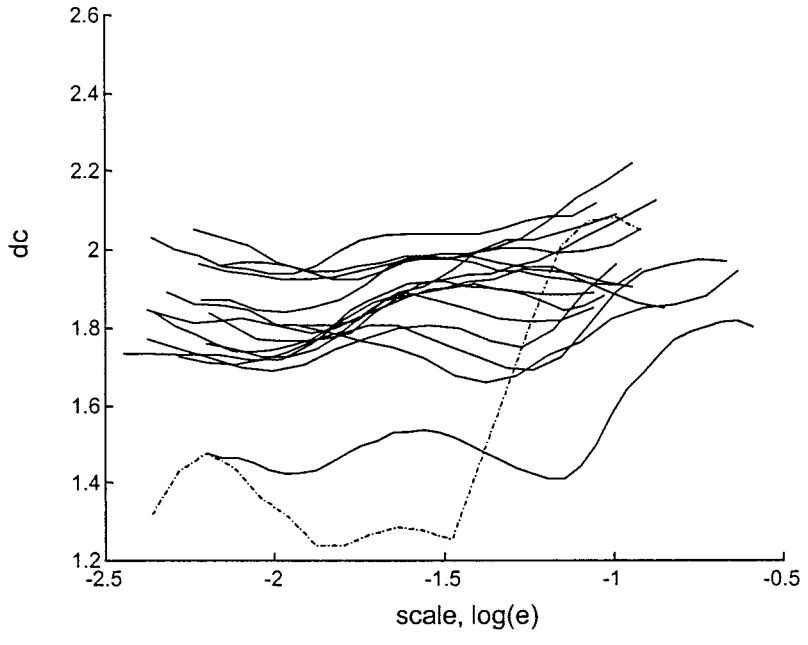
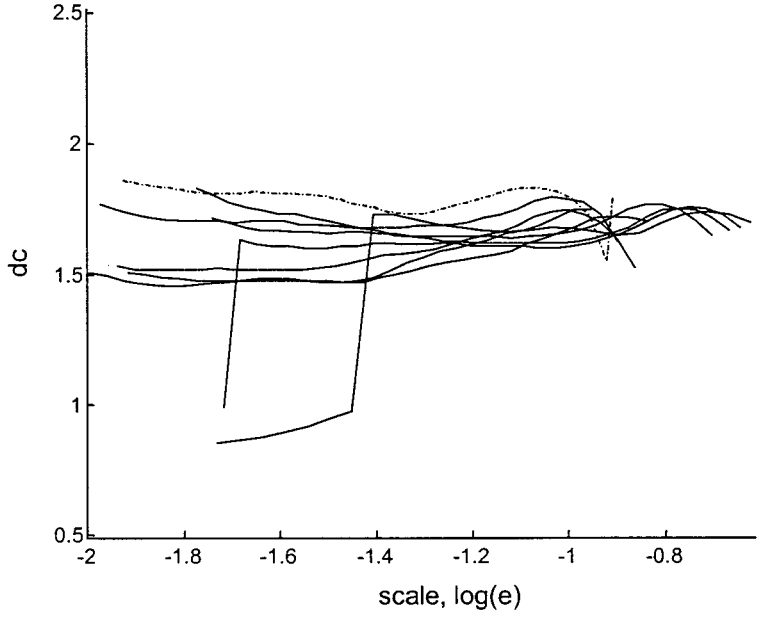


Figure 9 **Correlation dimension curves of non-linear surrogates of $M_{FF\theta 2}$ and that of the observed data (broken line, bottom) from the larger data set.**



(a)

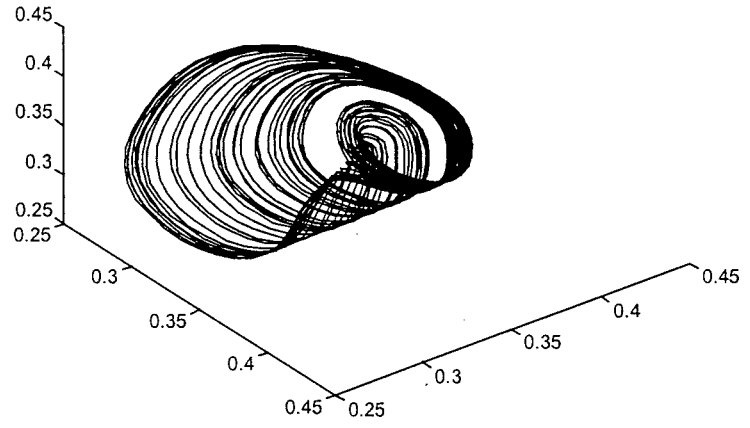


(b)

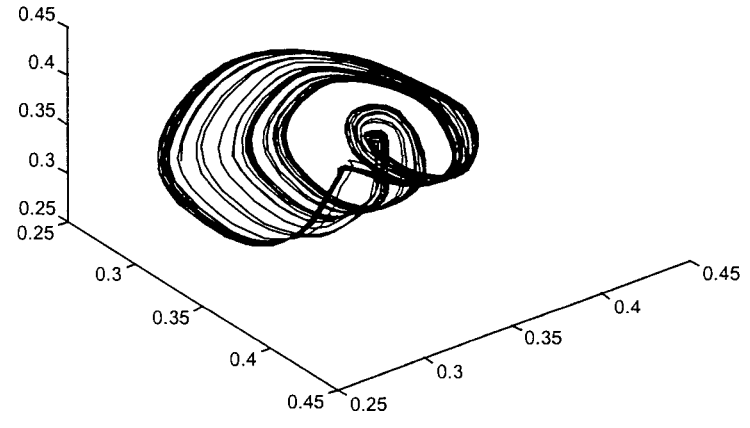
Figure 10 Correlation dimension curves of non-linear surrogates and that of the observed data (broken line, bottom), for (a) $\mathcal{M}_{PL\theta 1}$ and (b) $\mathcal{M}_{PL\theta 2}$.

Reconstructions of the dynamic attractor of the data, based on the actual data and the free-run predicted data ($\mathcal{M}_{PL\theta 2}$) are shown in Figure 11(a) and Figure 11(b). As can be seen from these figures, the two attractors are remarkably similar in appearance to each other and also to the attractor constructed from the *X*, *Y* and *Z* states, shown in Figure 3. This is confirmed by the

position of the correlation dimension curve for observed data amongst the cluster of non-linear surrogates in Figure 10(b).



(a)



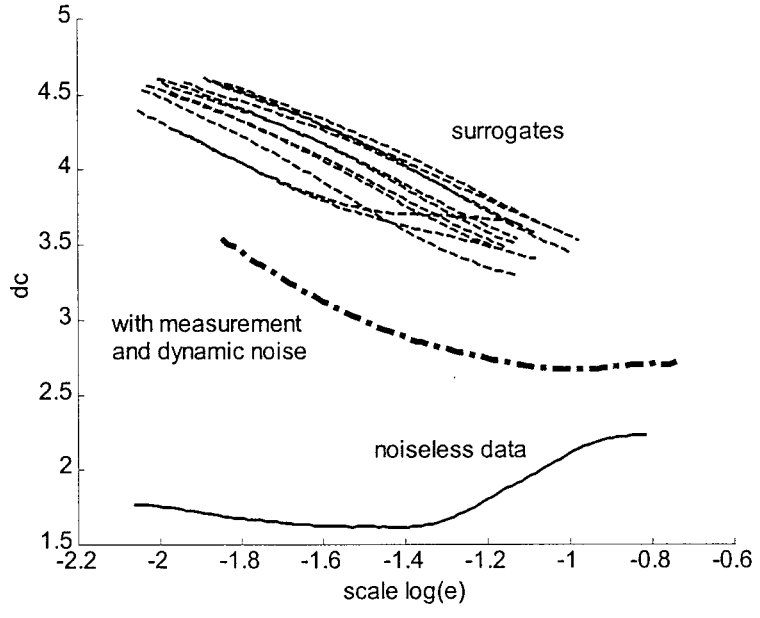
(b)

Figure 11 Dynamic attractor of autocatalytic process reconstructed from (a) the Y-state, and (b) the $\mathcal{M}_{PL\theta 2}$ free-run model of the Y-state.

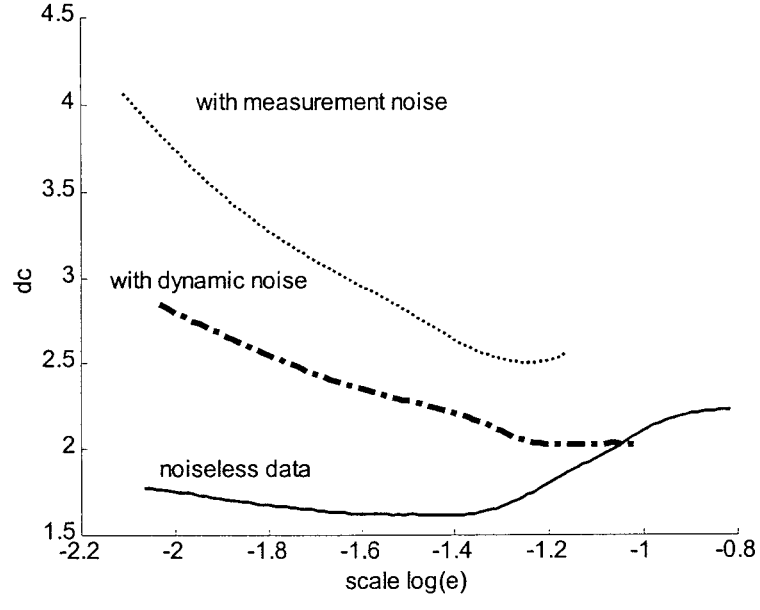
3.2.3. Effect of measurement and dynamic noise

To test the effectiveness of the identification method on noisy data, Gaussian measurement noise, as well as dynamic noise were added to the autocatalytic process data. The noise level was set at 0.1σ (10% of the sample standard deviation of the training data set) for

measurement noise. Dynamic noise was added by a modified one-step prediction of the training data with $\mathcal{M}_{PL\theta 2}$. Noise of 0.1σ (10% of the sample standard deviation of the training data set) was added to the i'th point, which was then included in the embedding for prediction of the $(i+1)$'th point.



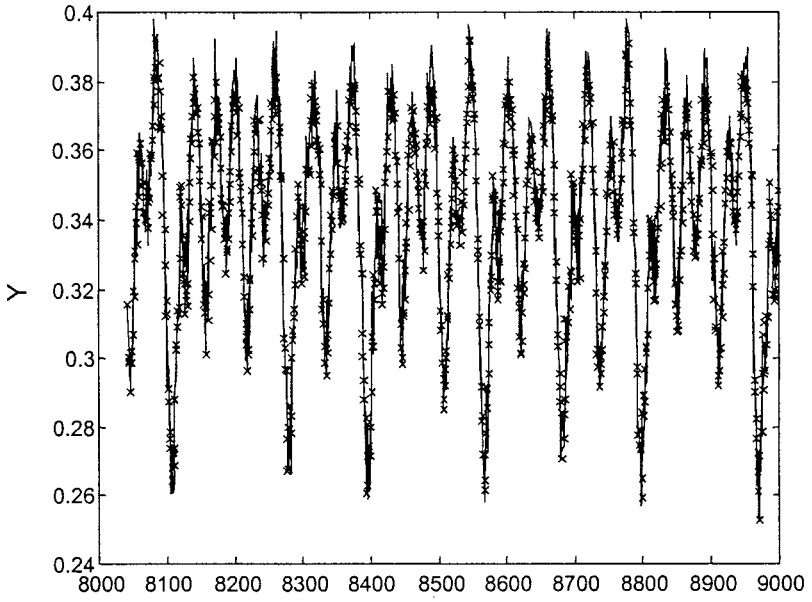
(a)



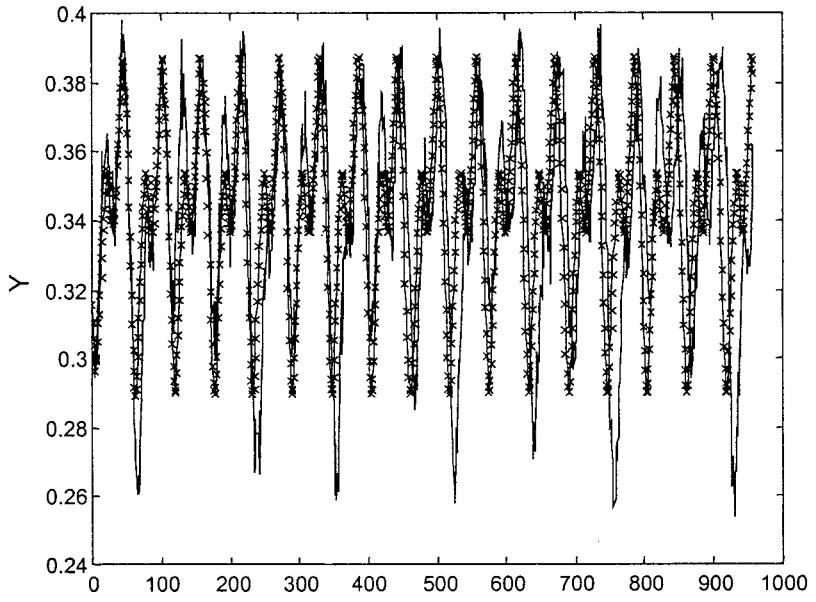
(b)

Figure 12 **(a) Correlation dimension curves for autocatalytic Y-state with noise (crosses) and its Type 2 surrogates, and for Y without noise (solid line). (b) Correlation dimension for Y (solid), with dynamic noise (dash-dot), or with measurement noise (dotted).**

In Figure 12(a), the correlation dimension curves of the noisy data, their associated surrogates and the noiseless data are shown, while the correlation dimension curves of the data with only measurement and only dynamic noise are shown in Figure 12(b). The correlation dimension curve for the data containing both dynamic and measurement noise appear between the noiseless data and the random surrogates in Figure 12(a). This is not surprising, since the noise increased the complexity of the attractor. However, it is interesting to note the relative positions of the small and large scale sections of the correlation dimension curves in Figure 12(b). For the data with only measurement noise, the whole correlation dimension curve has higher values than for the data without noise. On the other hand, the correlation dimension curve for the data with dynamic noise converges on that for the noiseless data at larger scales. A higher correlation dimension at small scales indicates more intricate microstructure in the attractor, or more complex detail dynamics, caused by dynamic noise. A higher correlation dimension at all scales, indicates an overall more complex structure in the attractor. Embedding parameters were m=8, k=7.



(a)



(b)

Figure 13 One-step predictions of autocatalytic Y-state (+ marker) with $\mathcal{M}_{PL\theta 3}$ vs. the Y-state with measurement and dynamic noise (a), and the free-run prediction of the same data with $\mathcal{M}_{PL\theta 3}$ (b).

Another model structure from the set $\mathcal{M}^*_{PL}$ was fitted to the noisy data set, resulting in the model $\mathcal{M}_{PL\theta 3}$. This model used 18 Gaussian kernels. Figure 13(a) shows the ability of $\mathcal{M}_{PL\theta 3}$ to predict the data with measurement and dynamic noise one step ahead, while Figure 13(b) shows the free-run predictions of $\mathcal{M}_{PL\theta 3}$. These figures indicate that though the model was

able to make acceptable one-step predictions, it failed to adequately capture the small-scale dynamics during free-run predictions by not following the peaks and troughs in the observed data.

3.3. Conclusions

In this chapter a proposed formal methodology for empirical non-linear system identification from an one-dimensional time series was successfully demonstrated on an autocatalytic process. Classification of data prior to model selection and parameter estimation, as well as validation and comparison of resultant models were performed using surrogate data methods. With surrogate data methods, model validation is based on criteria related to the topology of the system's dynamic attractor. Instead of comparing models based on single-valued statistical criteria, they can be compared on multiple scales of attractor topology by means of surrogate data methods. This allows better discrimination between models and can in principle also aid in the development of better models at different dynamic scales, where one model is not consistently better than the other over the entire range of scales.

From this investigation it is evident that different non-linear models may produce excellent one-step predictions, from which it may be very difficult to assess or compare the general validity of the models. For example, from the free-run predictions of the $\mathcal{M}_{FF}$ and $\mathcal{M}_{PL}$ models it is quite clear that the $\mathcal{M}_{PL}$ models was better able to capture the process dynamics from the smaller data set, than the $\mathcal{M}_{FF}$ models. This result demonstrates the inability of linear statistics, such as R^2, to truly measure the performance of non-linear models during validation. This conclusion can be extended to include cross-validation, in which R^2 plays the same role as in single-run validation, as used in this chapter.

As far as the application of surrogate data methods to the autocatalytic process in this investigation is concerned, the following can be concluded:

- Surrogate data are particularly valuable for the screening of data prior to model building. It is not always easy to determine the degree of determinism or stochasticity of real-world data, and this technique allows the engineer to inspect the data prior to building a model.

- Since the correlation dimension characterizes the topology of the attractor of the system in state space, it is a more rigorous criterion for the validation of dynamic process models than statistical or empirical criteria often used in practice, such as the R^2 statistic.

- Smaller data sets are less likely to represent the full-range of the dynamic behavior of a system and can therefore lead to the construction of less accurate global models. This can be readily assessed by use of surrogate data methods to visualize the performance of the system.

- Although a multilayer perceptron, as well as a pseudolinear radial basis function model were capable of similar, accurate one-step ahead prediction of a chaotic autocatalytic process, the pseudolinear radial basis function model was better able to capture the underlying dynamics of the system.

4 FAST OUTLIER DETECTION IN MULTI-DIMENSIONAL PROCESS DATA

In this chapter the problem of near-real-time detection and removal of up to 15% radial outliers from large data sets (10000 or more records) is investigated and a practical solution demonstrated. Radial outliers lie each in a different direction, with their means offset from that of the good data. There are several classes of outliers. Some classes are more difficult than others to detect successfully. In general the detection of multiple, multivariate outliers is particularly challenging. Among these, shift outliers, i.e. outliers that has a mean that is offset from that of the good data, but with the same covariance matrix, are the most difficult to detect. On the opposite end, radial outliers are easiest to detect. Outliers can also occur in multiple clusters with difficulty to detect proportional to the size of outlier clusters.

It often occurs that an outlier is not revealed by inspection of the individual components (limit checking) of a multivariate data vector, since for each component the extreme points are within acceptable limits. However when plotted together in a phase plot, the outlier becomes apparent since the vector containing the outlier components protrudes significantly from the neighbourhood of data vectors. For vectors, $\mathbf{x} \in \Re^m$ with $m > 3$, it is not possible to directly visualize and manually inspect for outliers. In the case of $m \leq 3$, human judgement has often proved to be subjective and quite inconsistent, hence many mechanistic means have been proposed so far to detect multiple outliers in multivariate process data. Hawkins (1980) has reported some multivariate outlier detection methods, for example principal component residues can be interpreted to identify outliers. In addition, principal component analysis can be applied in visualization techniques, where the data are projected onto two- or three-dimensional co-ordinates (Gnanadesikan, 1977). For example, two and three-dimensional scatter plots of data and the first principle component are often made. Alternatively, outliers can be detected by cluster analysis, without necessarily visualizing the data. The efficiency of some of these techniques can be severely limited when high-dimensional data are considered. Direct statistical approaches such as probability plots can facilitate discovery of outliers that distort location, scale and correlation estimates of data.

In order to distinguish between valid observations and outliers, it is required to estimate characteristic parameters of the distribution of observations as well as calculate a test statistic.

Often the probability distribution of the data is unknown, and consequently one attempts to find a robust estimator of the location and shape of the multivariate data. Some past methods include search algorithms to find the minimum volume ellipsoid (Hampel et al., 1986) or minimum covariance determinant (Rousseeuw and Leroy, 1987). Most methods are affine equivariant, which means they are invariant in terms of outcome under linear transformations. Rocke and Woodruff (1996) gave a comprehensive overview of multivariate outlier detection algorithms. In their paper, Rocke and Woodruff constructed a two-phase method that estimates the location and shape (phase one) and determines outliers by applying a χ^2 outlier criterion (phase two). In the first phase the location and shape are initially calculated by sequential point adding, using the minimum covariance determinant as initial estimation of shape. Finally shape is estimated using bi-weight M estimation. In the second phase a cutoff point is determined by simulation and the location and shape are updated using the (1-α) fraction of points falling within the cutoff region. Observations whose Mahalanobis square distance, using the updated location and shape, is larger than $\chi^2_{p;1-\alpha}$ (p is the dimension of the data) are rejected as outliers. Arguing that shift outliers are the hardest to locate, they showed successful results based mainly on this class of outliers. A practical limit to the ability of this algorithm, in terms of required data and processing time, is a maximum of about 35% outliers in 20-dimensional data. However, linearly extrapolating the results in table 5 of Rocke and Woodruff (1996) suggests that computational cost of this method may be prohibitively high for application to large data sets (size of order 10 000 records) often found in industry.

Points on the perimeter of a multivariate data space will by definition be coplanar with a convex hull constructed around the data, and so will be a fraction of potential outliers in the data. A novel method implementing this consequence is proposed to detect multivariate radial outliers in large data sets. The method is based on the use of convex hulls and the Mahalanobis distance. This method is compared to the Rocke and Woodruff algorithm (Rocke and Woodruff, 1996) in an example of elliptical random data and applied successfully to test data recorded on a Diesel automotive engine. The technique has the benefits of low computational cost with minimal operator input and can be implemented as a real-time outlier detection tool.

Detecting radial outliers in data does not require the sophisticated search for location and shape of generalized outlier detection algorithms. Also, a method was required that can be run on large data sets in real time or near real time (less than 10 seconds wall-clock time). It is

suggested that a convex hull can indicate the position of radial outliers and that the true location and shape of the data can be estimated after removing the hull from the data.

The indicated outliers are qualified as true or false using a test statistic based on statistical properties of the retained data.

4.1. Detecting radial outliers using convex hulls

Formally, the *convex hull* of a point set P is the smallest convex set that contains P. If P is finite, the convex hull defines a matrix A and a vector b such that for all x in P, $Ax+b <= [0]$. A convex set can be defined as the intersection of a set of half-spaces, that is the set of points on one side of a plane. It is possible to fit such a set of half spaces through the outer points of a data set so that it constitutes a convex set. The convex hull with maximum possible volume is co-planar with the outer data points in the set, and is by definition convex everywhere along its surface. For a detailed treatment of convex hulls, the reader may refer to O'Rourke (1994).

When one observes the data space, $\mathbf{X} \in \Re^m$, outliers tend to stand out from the local trend in the data. A convex set of the data space will be coplanar with at least some of the outlier vectors, but not necessarily all outliers, because less severe outliers may be excluded from the hull under the convex condition. Depending on the topology of the manifold along which the process state vector evolves, it is possible to improve the success of outlier detection by constructing the convex hull on the first difference of the data. This approach applies particularly to processes that can be described by a series of steady state set points with short transient regions, that is, the data are grouped into a number of distinct clusters, and fall along a piecewise smooth manifold. Inspecting first differences of such data, the incongruity of outliers with normal data will be more pronounced than when inspecting the data itself.

Detecting outliers in data requires a criterion by which an outlier can be qualified as a true outlier. Removal of valid data is undesirable - more so in small data sets (of order 1000 vectors or less). Since a convex hull does not distinguish between outliers and valid data along the perimeter of data space, a detection criterion can be of significant importance. Working either with or without first differences of the data, the reasoning regarding outlier detection runs as follows:

Assume a data set with radial outliers. Construct a convex hull around the data in $\Re^m$. It will go through at least some of the outliers. Remove the hull from the data. Let d^2 be the squared distance of each vertex on the convex hull to the location of the remaining data, and d_1^2 the

mean square distance of the remaining data to the location μ of the remaining data. If one compares d^2 to $d_1{}^2$, incorporating a sensitivity factor s, it should be possible to distinguish between true and false outliers.

As an indication of normalized distance between points and a data set, the Malahanobis distance between points **x** and **y** in $\Re^m$ is defined as:

$$d_{\Omega}^2(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T \Omega^{-1}(\mathbf{x}-\mathbf{y}) \tag{17}$$

where the metric Ω, is any p x p matrix related to the data. The choice of metric serves to normalize distances using a specific statistical property of the data. A commonly used metric is the covariance matrix. The more familiar Euclidean distance can be seen as the default Malahanobis distance that uses the identity matrix as metric. To qualify a vertex of the convex hull as an outlier, the convex hull was removed and the Malahanobis distance from the vertices to the location of the remaining data compared with the mean Malahanobis distance of the remaining data. A hull vertex qualifies as an outlier if the following inequality is satisfied:

$$d_{\Sigma}^2(Q_{0i},\mu_1) > s\overline{d_{\Sigma}^2(X_1,\mu_1)} \tag{18}$$

where $d^2(\cdot)$ is the Mahalanobis distance normalized with the covariance matrix of the remaining data, $Q_{0i} \in \Re^m$ a vertex of the convex hull, $X_1 \in \Re^m$ the remaining data and s a sensitivity factor. Since the Mahalanobis distance constitutes a spherical criterion, it will be optimal for spherical data spaces, but lead to identification of some false outliers for data spaces that are non-spherical, e.g. elliptically distributed data, like in the example below.

4.2. Procedure for radial outlier detection method

The procedure for radial outlier detection proposed here can be systematically performed in the following steps:

1. Scale the data, **X,** by standardizing to zero mean and unit standard deviation. Then normalize each component $\mathbf{X}_i$ to have unit length ($\|\mathbf{X}_i\| = 1$). Standardization and scaling ensure that all data components are of the same order of magnitude, otherwise a multidimensional outlier detection algorithm would disregard outliers in components of comparatively small magnitude.

2. Depending on the class of data, calculate the first difference, **X'**, of the data, **X**. As will be shown in this chapter, radial outlier detection of outliers in Gaussian data does not respond well when applied to first differences of the data. On the other hand, for data with a relatively smooth topology first differences enhances the efficiency of the proposed outlier detection method.
3. Construct a convex hull $\mathbf{Q}_0$ around **X'** (or **X**). The algorithm and code by Barber et al. (1996) can be used for this purpose.
4. Remove $\mathbf{Q}_0$ from **X'** (or **X**) to give $\mathbf{X}_1$.
5. Calculate the Mahalanobis distance from vertices of $\mathbf{Q}_0$ to the mean of $\mathbf{X}_1$.
6. Apply the criterion in equation (**18**), $d_{\Sigma}^{2}(Q_{0i},\mu_1) > s\overline{d_{\Sigma}^{2}(X_1,\mu_1)}$.

4.3. Demonstration of outlier detection method

The outlier detection method is demonstrated in two examples. The first is a normally distributed data set that has been generated so as to include a number of known outliers. The second example is a data set recorded of an internal combustion engine on a dynamo test bench, under controlled conditions. The outliers in this example were unknown before applying the detection method.

4.3.1. Random data containing outliers

To enable an objective evaluation of the outlier detection technique, a two-dimensional Gaussian data set was intersected with data describing an ellipse to produce a data set of 533 records. Six outliers were manually created by moving 6 arbitrary points in an *XY* phase plot to outside the elliptic boundary. None of the outliers could be detected as such from inspection of the *X* (Figure 14) and *Y* (Figure 15) components of the data set, since they were within the range {-2.5, +2.5} spanned by each component. However, in a phase plot (Figure 16) the outliers were clearly visible.

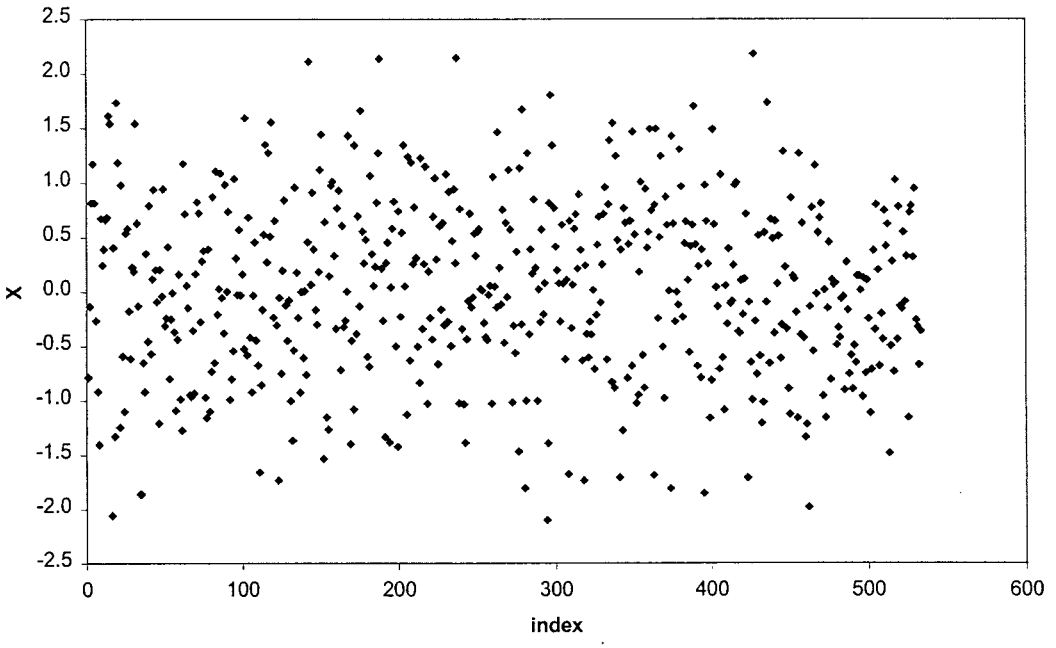


Figure 14 X component of random data set

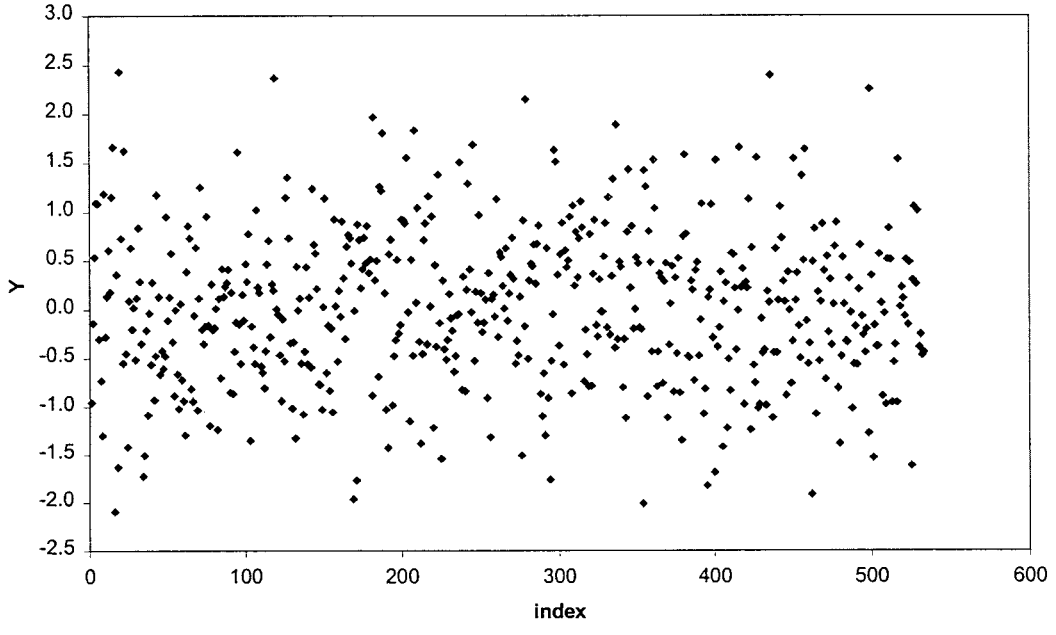


Figure 15 Y component of random data set

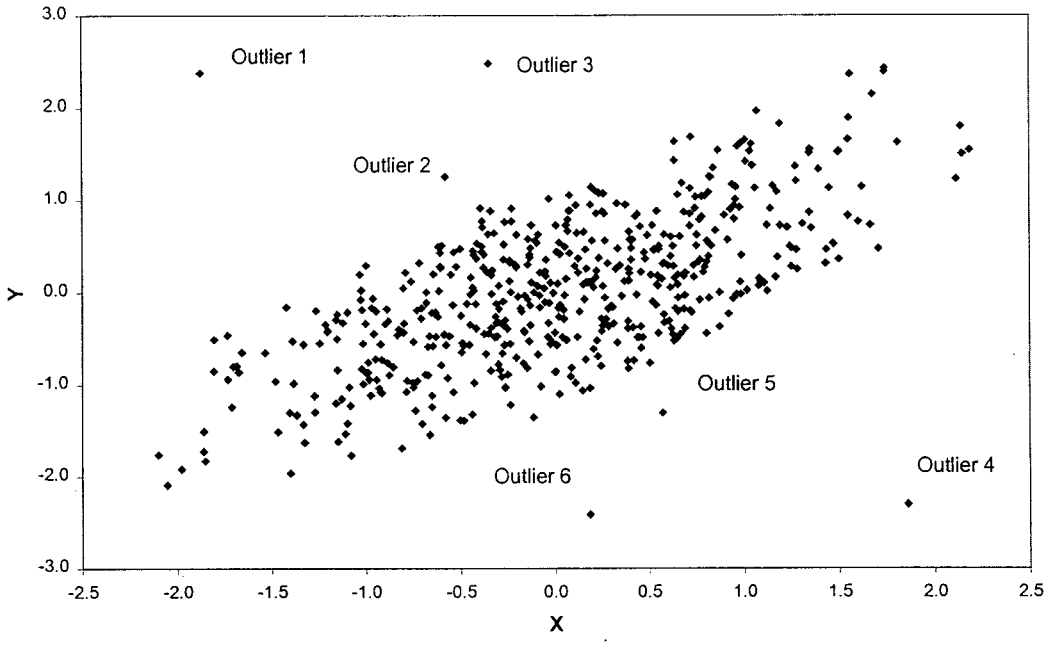


Figure 16 XY plot of random data showing manually added outliers.

The first approach to detecting the outliers dealt with first differences of the data. A convex hull was constructed around first differences of the data and then removed. The hull for the first iteration of the outlier detection procedure appears in Figure 17. Even though the outliers in the first differences of the data were correctly identified, these outliers did not necessarily correspond to outliers in the data themselves, as is apparent from Figure 18. Consequently, after two iterations only two out of the six outliers were removed from the first differences of the data, as seen in Figure 19. For random data, it appears that extreme values in first differences of data generally do not coincide well with extreme values in the data themselves.

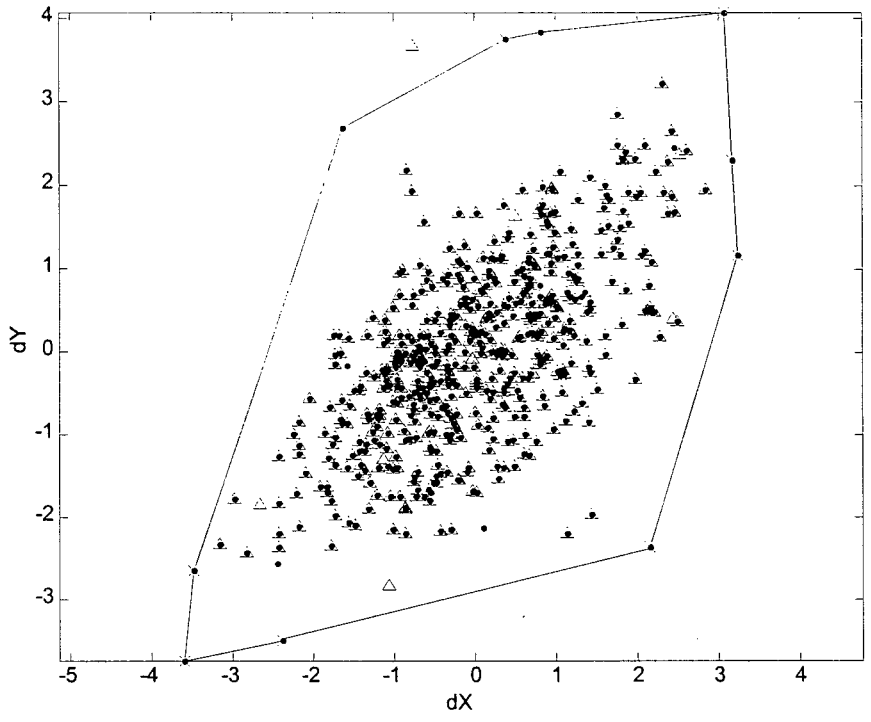


Figure 17 **Convex hull constructed on first difference of data during first iteration of the outlier detection algorithm. Triangles indicate first differences of the data set after removing the convex hull. Crosses indicate identified outliers.**

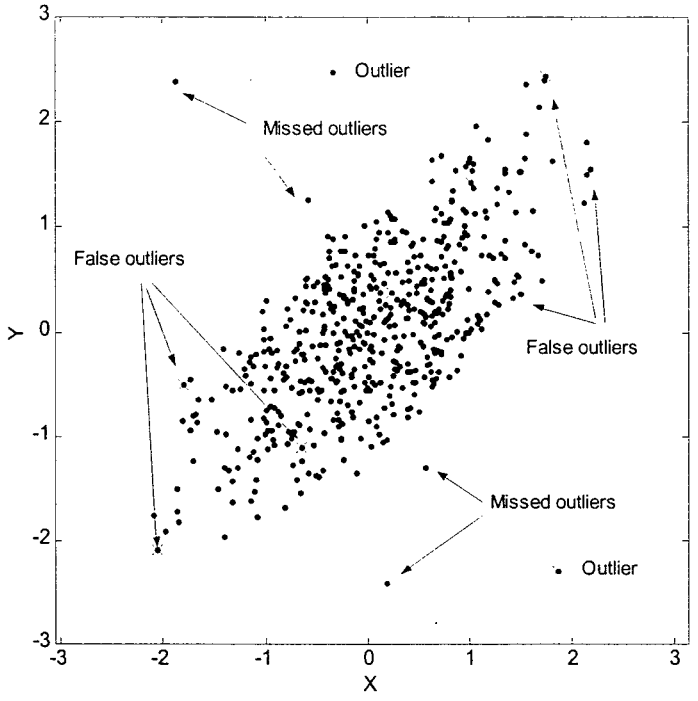


Figure 18 Results after first iteration of outlier detection procedure on first differences of random data.

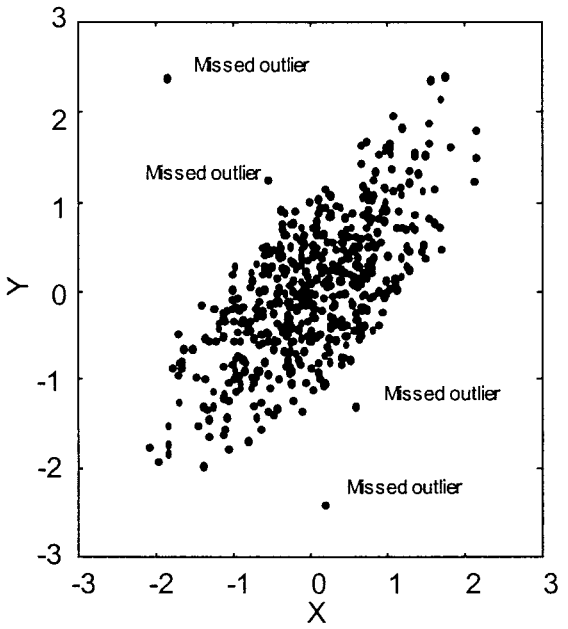


Figure 19 Results after second iteration of the convex hull outlier detection algorithm, based on first differences of random data.

When applied to the data, instead of first differences of the data, the outlier detection scheme converged quickly. After one iteration four out of the six outliers were detected correctly with

the two least severe outliers remaining (Figure 20).The cost of outlier detection was five false outliers identified at the ends of the long axis of the ellipse. This is due to the spherical outlier criterion applied to a data set that is very non-spherical in terms of distribution. A default sensitivity factor, s = 1, was used in all runs. However, since the motivation behind the convex hull method is fast detection in large data sets, this cost is very low at 0.9% of sample size. By comparison, the Rocke and Woodruff algorithm detected the same outliers as our technique, also failing to detect the two least severe outliers. Their algorithm incurred zero detection cost by not indicating any false outliers, as seen in Figure 21.

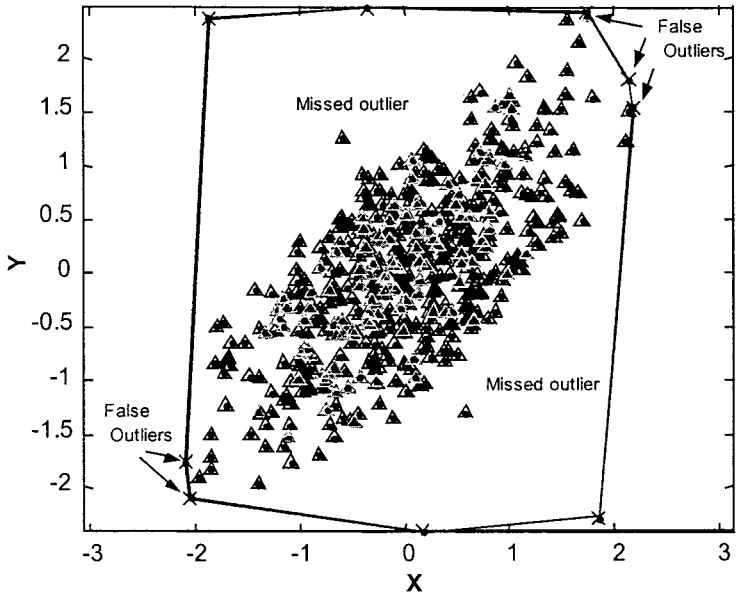


Figure 20 **Convex hull constructed around random data during first iteration of the convex hull outlier detection algorithm. Triangles indicate the data set after removing the convex hull. Crosses indicate identified outliers.**

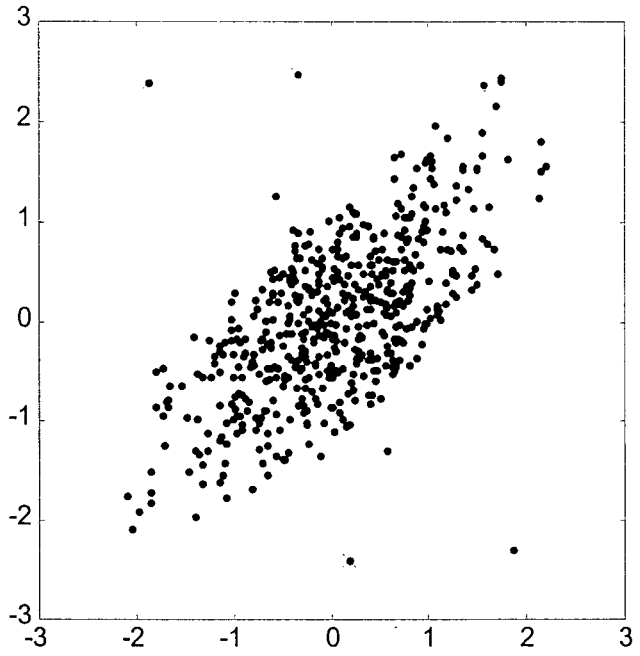


Figure 21 Outliers detected in random data with the Rock and Woodruff algorithm (indicated with crosses).

Since the motivation for this research was "real time outlier detection", the computational cost in terms of time required to process 10000 elliptically distributed random data points was investigated. The convex hull technique took 1.0 second to complete one iteration of detection while the Rocke and Woodruff technique needed 660 seconds, running ANSI C code for both algorithms under Microsoft® NT4 on an Intel® PII 400 Celeron™ processor with 256 MB RAM.

4.3.2. Internal combustion engine test data

As a second example, the convex hull technique was applied to data recorded during an endurance test of a Diesel engine under controlled laboratory conditions. Since the data would be used later to simulate some of the recorded engine states, it was important to remove the outliers. Also, since the outlier detection mechanism would eventually be required as part of real-time test monitoring software, it was important to use a method that requires minimal operator input.

To detect and remove as many outliers as possible, the outlier detection procedure was iterated twice. This provided the opportunity to detect repeated outliers as well as hidden outliers during the second pass. The data were first standardized to zero mean and unitary

standard deviation and then normalized so that each component had unit length. A convex hull was constructed around first differences of the data. Figure 22 shows the unscaled outliers, identified during the first pass using a default detection sensitivity, $s = 1$. In this data set, outliers indicated in the range between indices 30000 and 40000 are related to a disruption during the test schedule that caused a sudden fluctuation in the independent variables. Figure 23 shows the unscaled data after removing the outliers.

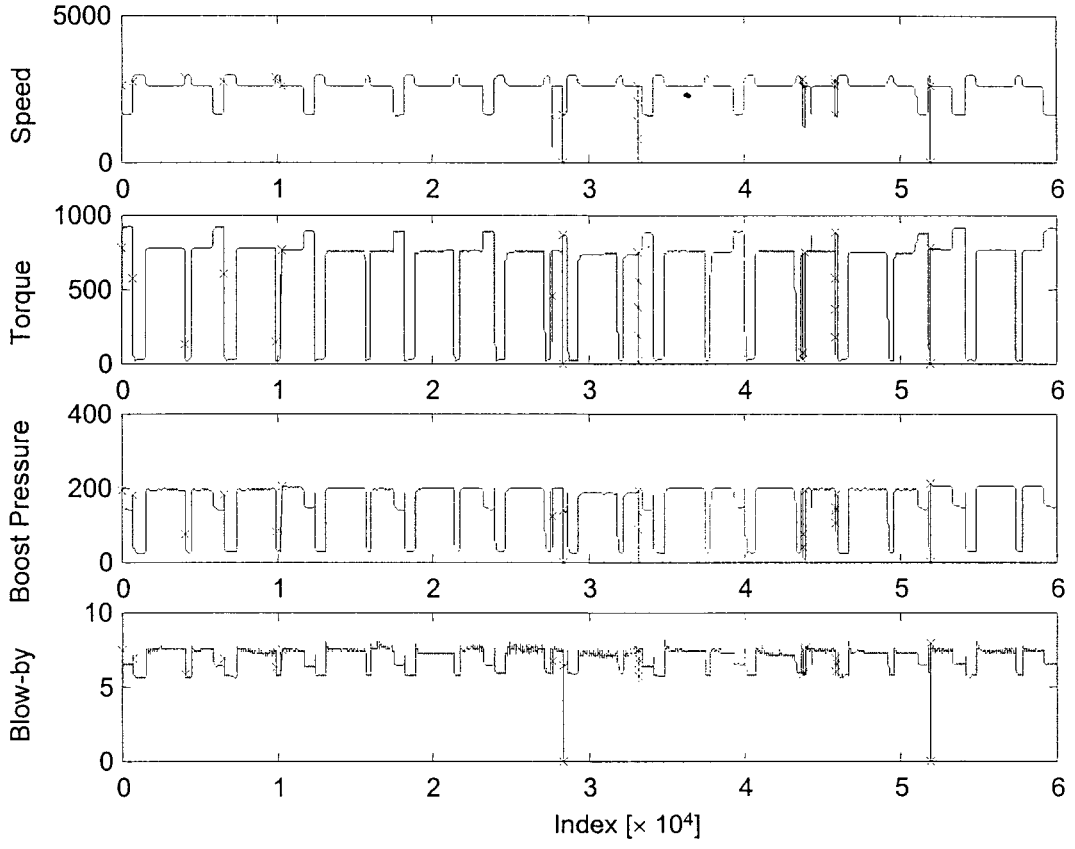


Figure 22 **Outliers detected by convex hull construction on first differences of engine data (detection sensitivity = 1). Crosses indicate outliers identified during the first iteration.**

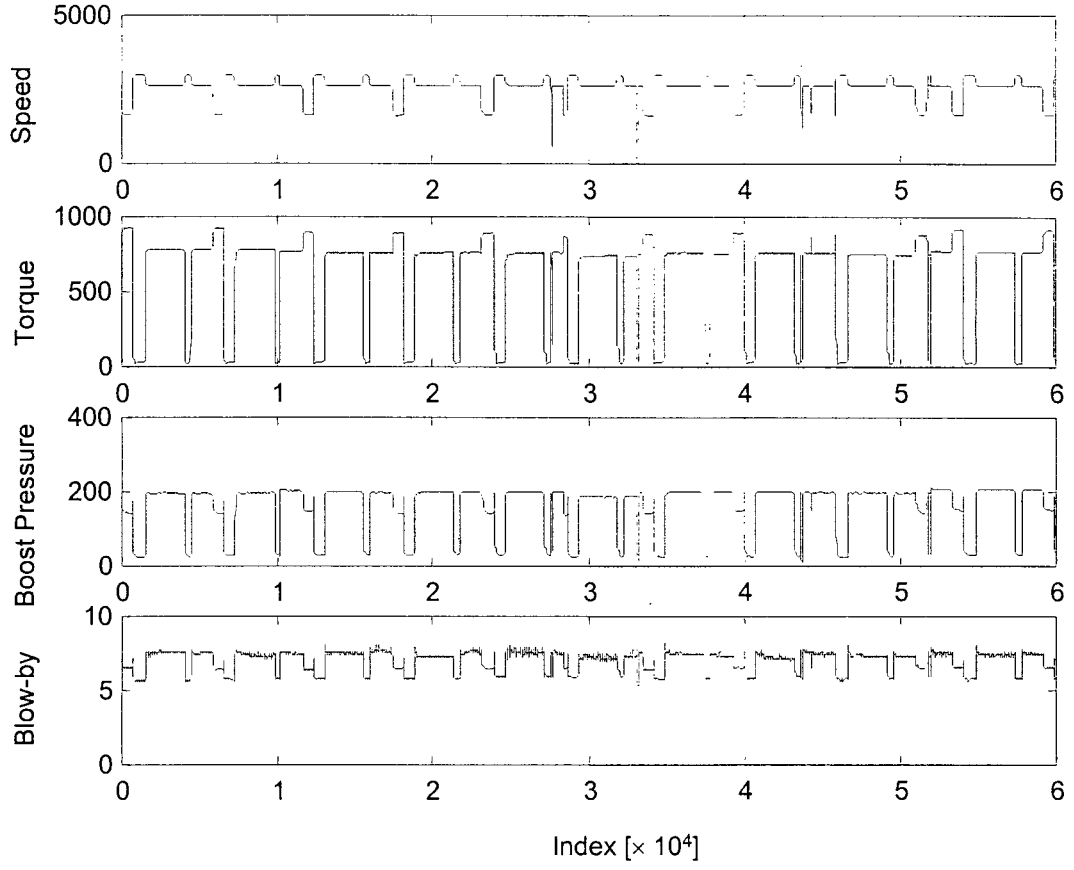


Figure 23 Engine data after removing outliers by two iterations of the convex hull method (the hulls were constructed on first differences of data, detection sensitivity = 1).

4.4. Conclusions

In this chapter the goal to develop a near real-time algorithm for detecting radial outliers was effectively achieved by way of an algorithm that incorporates convex hull construction and removal of points supporting the hull. The convex hull technique also requires no operator input beyond setting a detection sensitivity ratio, for which the default setting generally suffices. It runs fast with low computational and false outlier detection costs. The method was demonstrated successfully on artificial random data, as well as real test data. Our investigation into computational cost showed the convex hull technique to be two orders of magnitude faster than the Rocke and Woodruff algorithm. Compared to the Rocke and Woodruff algorithm, the cost in terms of number of false outliers detected was higher on the random data, owing to the application of a spherical outlier criterion to elliptically distributed data. The results for the engine data were satisfactory. Improvements to the procedure to minimize

false outlier detection could be to replace the outlier criterion by a ratio of distances derived as follows:

1. Construct a convex hull around the data and remove it.
2. Construct a second convex hull around the remainder of data and reduce the offset of each hyperplane in the convex set so that on average $(1-\alpha)$ of data fall inside the hull. Assume this to be the cut-off limit.
3. Calculate by suitable interpolation the vertices of the second hull that are co-directional with those of the first hull.
4. Identify true outliers as those vertices of the first hull for which the ratio of Malahanobis distances of co-directional vertices exceeds a set amount.

The above modifications to the proposed methodology will appear in future work in this field.

5 EMBEDDING OF MULTIDIMENSIONAL OBSERVATIONS.

In this chapter, a novel method is proposed to reconstruct the dynamic attractor of a dynamic, non-linear process from a multivariate time series observation of the process. Takens' embedding theory (1981) is combined with Independent Component Analysis (Hyvärinen, 1999) to transform individual embeddings of multidimensional process observations into a vector space of statistically independent vectors (state variables). The method is demonstrated on an analytically defined autocatalytic process and an air-pollution case study.

Embedding of observations in phase space is central to the analysis of non-linear time series. This topic has been discussed in section 2.1.3 and implemented in Chapter 3. It involves reconstructing the dynamic attractor for deterministic systems and mapping the attractor onto the time series, by using a non-linear model structure such as a multi-layer perceptron network or a radial basis function network.

The attractor can be reconstructed from a time-series by delay-coordinate embedding (Takens, 1981), and emulates the state space of the system. A state variable represents an independent energy state of a system, therefore the embedding variables should be statistically independent. The time-evolution of the state vector along a trajectory through state space forms the dynamic attractor of the system. In addition, an attractor with optimal structure expresses maximal information about the dynamic features of the process and this will benefit any regressor fitted to the data.

While embedding of one-dimensional observations (time-series) is well-established in applied mathematics, embedding of dynamic systems based on multi-dimensional observations has not been sufficiently formalized. It is not always possible to predict the time evolution of a system state from only a single observed variable. For example, the Lorenz system has three state variables, $x\ y\ z$, but $\dot{x} = f(x, y)$, while $\dot{z} = f(x, y, z)$, therefore z can not be properly predicted only from x, y or even (x,y). For empirical systems the situation is worse, since the observed variables required for a prediction model and the dynamic interdependencies are not obvious. Where only linear relationships exist among observed states (variables), one may use principle component analysis (PCA) to find a minimum subset of independent variables that declares a required percentage of total variance. For non-linear relationships, techniques such

as PCA often are insufficient, since they consider only static covariance. Also, PCA does not necessarily result in optimal projection of independent components (Friedman et al., 1974).

Multidimensional observations can be simulated by a model that maps *p-1* components onto the remaining component of an $\Re^p$ observation space. This does not always ensure optimal identification of the system that generated these observations. The Lorenz system is a case in point, where (x,y) is insufficient to predict z. Cao et al. (1998) proposed embedding all components of the multi-dimensional observations using an optimal Takens embedding for each component. The optimal values of embedding dimension for each component were found by minimizing prediction error of a nearest neighbor, locally constant predictor. Unfortunately Cao et al. did not indicate how to optimize embedding lag, which is crucial in reconstructing a representative attractor for practical systems. Optimization of embedding lag is especially complex if noise is present in the observations (Lai and Lerner, 1998).

The aim of finding the optimal embedding lag is to determine sufficiently independent embedding variables from the observed data, to serve as state variables in the dynamic attractor. Traditionally, embedding lag is calculated using a minimum mutual information criterion proposed by Frazer and Swinney (1986). To determine the optimum embedding lag for multidimensional observations, one has to find the mutual information between each point in any one component and all points in the other components. Calculating only static mutual information among observation components is similar to calculating the covariance matrix during PCA, and does not fully consider nonlinear dynamic correlation. Alternatively, creating a combined embedding space by the individual embedding of each observed component could lead to significant statistical dependence between some of the embedding (lag) variables. In this case the resultant manifold structure would not be optimally reconstructed from the observations.

A different approach is therefore proposed to embed multidimensional observations that avoids both linear approximations in finding embedding dimensions and potentially sub-optimal embedding lags. With this approach, each component of an observation space, $\mathbf{Y} \in \Re^p$, is treated as a one-dimensional time-series. Embed each component individually to generate a set of subspaces. Combine these subspaces to form a first approximation of the attractor in $\Re^{\Lambda} = [\Re^{m}_{1}\, \Re^{m}_{2}\, \Re^{m}_{3} \ldots\, \Re^{m}_{p}]$. Finally, separate the lag variables and optimize the structure of the attractor. This results in a reconstructed dynamic attractor based on the observation space. In addition the resultant attractor structure is an optimal projection of the original embedding variables.

There are several ways to separate a set of variables, that is, to remove possible statistical interdependence from a set of variables. Examples are :

1. Principle Component Analysis (PCA).
2. Projection Pursuit by PCA (Friedman et al., 1974).
3. Projection Pursuit by Independent Component Analysis (ICA) (Comon, 1994).
4. Blind Source Separation by Self-organizing maps (Pajunen et al., 1996).
5. Blind Source Separation by ICA (Hyvärinen, 1999).

Some of these methods will also project variables so as to optimize the resultant multidimensional structure (e.g. Projection Pursuit).

De-correlation and optimal projection by ICA as proposed by Hyvärinen (1999) is an acceptable median between PCA, on the one hand, and Blind Source Separation using the minimum description length (MDL) principle (Rissanen, 1989), on the other hand (Pajunen, 1998). PCA, though simple and computationally inexpensive, has limited ability for optimal projection and component separation, while Blind Source Separation using MDL, though more capable than ICA, is still computationally prohibitive in practical applications to classes of non-linearly mixed components.

In this chapter the identification of an autocatalytic and an environmental system is discussed. Both systems are multidimensional and therefore will be parameterized by various multivariate embedding strategies. In the first case study a parameterization will be devised for the simulation of one autocatalytic state in terms of the others. In the second case study a parameterization will be arranged to predict all observed states. Non-linear model structures will be fitted to map the embedding at time t to the observation at time t+1. After optimizing the model structures against the R^2 criterion, an iterative free-run prediction will be performed for the environmental system. That is, the embedding will be updated with predictions instead of observations while the model is running. The resultant prediction array will be then compared to the observations, first by inspection, and if very similar qualitatively, in terms of the correlation dimension statistic.

5.1. Multidimensional embedding methodology

System parameterization by the multidimensional embedding method proposed in this chapter, consists of the following steps:

1. Optimal embedding of individual observed components,
2. Scaling embedding components $\mathbf{x}_i$ as $\| \mathbf{x}_i \| = 1$,
3. Optimal projection of state space onto statistically independent state variables.
4. Selection of a non-linear model structure to approximate the output function of the system.

5.1.1. Optimal embedding of individual components

Each component of the multidimensional observation (time-series), is embedded using Takens' embedding. The embedding lag is determined by the minimum mutual information (AMI) critetion proposed by Frazer and Swinney (1986) and the embedding dimension by the false nearest neighbours (FNN) algorithm of Kennel (Kennel et al., 1992).

Let $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \cdots, \mathbf{Y}_N]^T$ be the array of p-dimensional observations, where $\mathbf{Y}_i = [y_{i1}, y_{i2}, y_{i3}, \cdots, y_{ip}]$, i = 1, 2, 3 ,..., N;. Multidimensional embedding of $\mathbf{Y}$ results in the following embedding matrix:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_{11} & \mathbf{x}_{12} & \cdots & \mathbf{x}_{1p} \\ \mathbf{x}_{21} & \mathbf{x}_{22} & \cdots & \mathbf{x}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_{N-J_0,1} & \mathbf{x}_{N-J_0,2} & \cdots & \mathbf{x}_{N-J_0,p} \end{bmatrix}, \tag{19}$$

where

$$\mathbf{x}_{ij} = \begin{bmatrix} y_{i+k_j(m_j-1),j} & y_{i+k_j(m_j-2)_j,j} & \cdots, & y_{i,j} \end{bmatrix} \quad i = 1 \ldots N, j = 1 \ldots p \tag{20}$$

and $J_0 = \max_{1 \le j \le M} k_j(m_j - 1) + 1$.

5.1.2. Optimal projection of initial embedding

Projection of the embedding $\mathbf{X}$ onto directions that optimize the structure of the attractor and statistically separate the embedding variables, is achieved by the following linear transformation:

$$\mathbf{S} = \mathbf{WX}, \tag{21}$$

where **S** is the optimal projection of the original embedding **X** and **W** is the separating matrix. The dimension of **S** may be less than that of **X** . Thus one may achieve optimal projection, reduction of dimensionality as well as independence of embedding variables.

To find **W**, one has to maximize the negentropy J_G of **X,** which is equivalent to minimizing the cross mutual information amongst components of **X,** under the constraint of decorrelation of components. Formally:

$$\text{maximize} \sum_{i=1}^{M} J_G(\mathbf{w}_i) \text{ wrt. } \mathbf{w}_i, \tag{22}$$

under the constraint,

$$E\left\{\left(\mathbf{w}_k^T\right)\left(\mathbf{w}_j^T\right)\right\} = \delta_{jk} \tag{23}$$

where

$$J_G(\mathbf{w}) = \left[E\left\{G\left(\mathbf{w}^T\mathbf{x}\right)\right\} - E\{G(v)\}\right]^2 \tag{24}$$

with G some non-quadratic, contrast function that estimates the probability density function of an independent component, c some insignificant constant, and v a standardized Gaussian variable. Each vector $\mathbf{w}_i$, is a row of matrix **W.**

Three practical choices of contrast functions (with their respective derivatives, indicated by lower-case function symbols) are listed below:

tanh-function, $G_1(u) = \frac{1}{a_1}\log\cosh(a_1 u)$, $g_1(u) = \tanh(a_1 u)$ **(25)**

Gaussian function, $G_2(u) = -\frac{1}{a_2}\exp(-a_2\frac{u^2}{2})$, $g_2(u) = u\exp(-a_2\frac{u^2}{2})$ **(26)**

Power 3 function, $G_3(u) = \frac{1}{4}u^4$, $g_3(u) = u^3$ **(27)**

A specific choice of contrast function serves to optimize performance of the ICA algorithm. G_1 is optimal for most data sets, G_2 for highly super-Gaussian data as well as when robustness against outliers is important, and G_3 for sub-Gaussian data with no outliers. (Super-Gaussian data has a distribution that "stands out" above the normal distribution curve $N(\mu,1)$, while sub-Gaussian data has a distribution that lies below the normal distribution curve $N(\mu,1)$.)

An important condition contained implicitly in equation (**24**) is the data may not be Gaussian, otherwise the maximization in equation (**22**) will not converge. However, should the data turn out to be Gaussian, or random for that matter, a state space parameterization would be excluded by classification (as discussed in 2.2.3). Thus choosing independent component analysis to separate the individual embeddings into subspaces containing independent state variables is not restrictive.

The maximization of equation (**22**) is done by a fixed-point algorithm developed by Hyvärinen (1999), which is a batch process (not gradient decent). It converges at least quadratically for practically any non-Gaussian distribution using any G and is roughly equivalent to projection pursuit because it estimates independent components by deflation. This means the independent components are calculated one at a time in succession.

5.1.3. Selection of a suitable model structure

After embedding the observations, a suitable associated model structure is selected and fitted to the data set, $\mathbf{Z} = [\mathbf{S}\ \mathbf{Y}]$.

The models fitted to the multidimensional embeddings were tested and validated by one-step as well as free-run predictions. In the first case study, one step predictions of the one data component were done, using perfect knowledge of the other data components. In the second case study, simultaneous one-step predictions of all data components as well as free-run predictions were performed. The R^2 statistic for one-step prediction versus observations was selected as linear criterion for embedding quality and model fitness. In addition, free-run predictions were compared by way of inspection. The prediction stability boundary was inferred from the estimation of the first global Lyapunov exponent and indicated the reasonable free-run prediction horizon that could be expected.

A prediction procedure for the above multi-channel embedding strategy can be formulated as follows:

1. Embed points $y_{i,j}$ through $y_{i+k(m\text{-}1),j}$ over p observation components, to form $\mathbf{x_i}$, applying equations (**19**) and (**20**).
2. Scale embedded components to unit vectors.
3. Separate embedded variables to form state variables, by applying equation (**21**), using $\mathbf{W}$ calculated from the data set used for estimation of the model parameters.

4. Predict $y_{i+k(m-1)+s,j}$ as $\hat{y}_{i+k(m-1)+s,j}$, where s is step size (normally one sampling unit).

5. For free-run predictions, use the prediction $\hat{y}_{i+k(m-1)+s,j}$ in the next embedding $\mathbf{x}_{i+1}$, instead of the observation $y_{i+k(m-1)+s,j}$.

6. Repeat from step 3 for the length of the prediction run.

Note that if p-1 observations are embedded to predict the p'th observation, then the above prediction scheme is not free-run but a simulation of y: y=f($\mathbf{x}$), since the p-1 observations are fully known at time t.

5.2. Application of the embedding method

The embedding method was applied to a well-defined chaotic autocatalytic process that also served as case study in section 3.2. In a second case study on air pollution data, the method was applied to the prediction of atmospheric NO_2 concentration.

Owing to available software, computer hardware and project time constraints at the time of this research, the set of MLP network model structures had to suffice for prediction of NO_2 concentration. Given the right resources, the set of radial basis function model structures is a very feasible alternative to the MLP network and will be investigated in future research on atmospheric pollution.

All the MLP network model structures implemented a single hidden layer of bipolar sigmoidal transfer functions, defined as $g(\cdot)=[1-\exp(\cdot)]/[1+\exp(\cdot)]$. The output layer consisted of bipolar linear transfer functions. The Levenberg-Marquardt algorithm was used for parameter estimation.

5.2.1. Autocatalytic process

The autocatalytic process in this case study has been formulated by Lynch (1992) and is the same system as defined in section 3.2 as a state space system in terms of the following set of differential equations:

$$\begin{aligned}\frac{dX}{dt} &= 1 - X - aXZ^2 \\ \frac{dY}{dt} &= 1 - Y - bYZ^2 \\ \frac{dZ}{dt} &= 1 - (1+c)Z + daXZ^2 + ebYZ^2\end{aligned} \qquad (28)$$

where, as before, X, Y, and Z denote the dimensionless concentrations of species A, B and D. For the settings: $a = 18000$; $b = 400$; $c = 80$; $d = 1.5$; $e = 4.2$, and initial condition $[0,0,0]^T$, the set of equations was solved over 100 simulated seconds using a 5th order Runge-Kutta numerical method. The result consisted of 23641 points that defined the evolution of the three states, X Y Z over the whole simulation period. These states were resampled by linear interpolation with a constant sampling period of 0.0033s to give 30000 records. The data was sampled at a higher rate than in Chapter 3 and the system was run over a longer time span, to generate enough data for the stationarity test. Alternatively, the sampling rate could have been kept as in Chapter 3, and the system just be run over a proportionately longer time span, to provide the minimum amount of data to test for stationarity. Figure 24 shows the true evolution of the three state variables (X, Y, Z), while Figure 25 shows the dynamic attractor traced by the state vector [X Y Z].

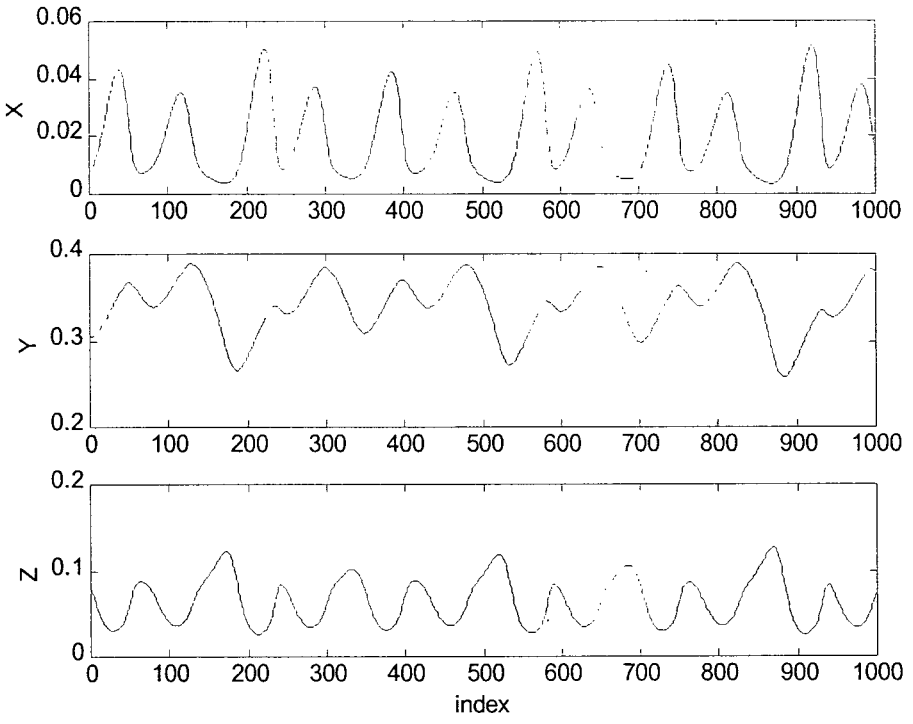


Figure 24 **Autocatalytic data, showing X, Y, and Z states resulting from a numerical solution for the process state equation.**

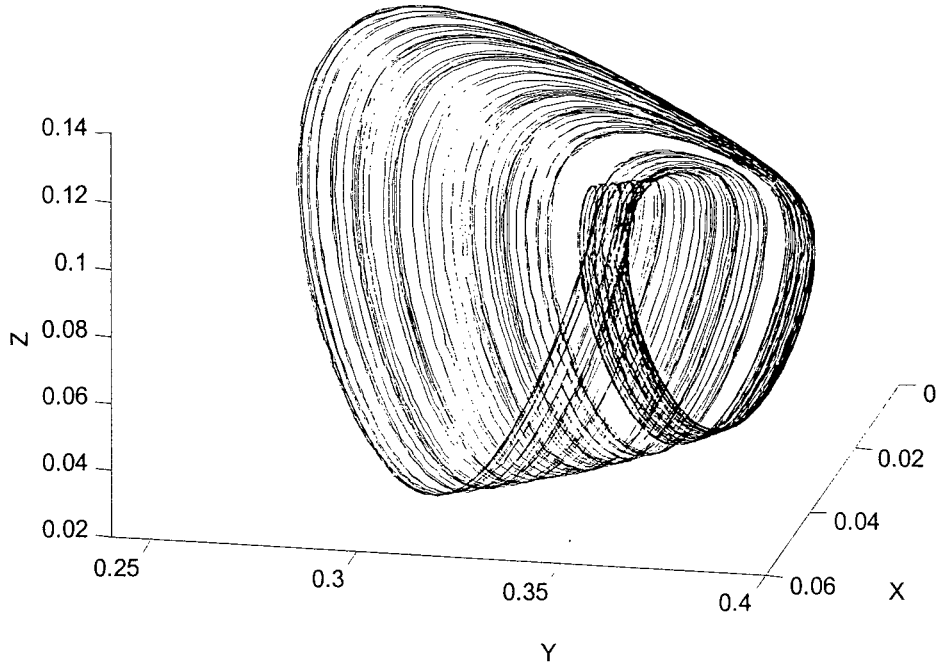


Figure 25 Dynamic attractor of autocatalytic process (first 10000 records), constructed from X_t, Y_t and Z_t, resulting from a numerical solution for the process state equation.

The aim of investigating the autocatalytic system was to predict one of the states from the other states with a non-linear predictor, using a suitable parameterization.

From inspection of the process state equations (**28**), it is apparent that $Z = F(X, Y, Z)$. Consequently a non-linear Multi-In-Single-Out model structure was selected from the set of feed-forward neural net models, $\mathcal{M}^*_{FF}$, to predict Z_{t+1} in terms of $[X\ Y]_t$.

The system was parameterized by individual embeddings of X and Y. AMI and FNN calculations for each of X and Y resulted in an embedding dimension set of {3,3} and lag set of {21,26}. The large embedding lags was due to the high sampling rate, which results in high linear correlation between adjacent data points. The individual embeddings were combined into $\Lambda_0 \in \Re^9$ and separated as $S_0 = W_0 \Lambda_0$, using ICA with the G2 contrast function (the only function for which the ICA algorithm converged). The following model structure was to be fitted:

$$\mathcal{M}_0: S_0(t) \rightarrow Z(t+1),\ \mathcal{M}_0 \in \mathcal{M}^*_{FF} \qquad (29)$$

The optimal model topology had a hidden layer of 8 bipolar sigmoidal nodes. The effect of separation by ICA was investigated in terms of a second parameterization consisting of the same embedding as above, but without the explicit separation of embedding variables. This model structure was called $\mathcal{M}_1$ and used the same model topology as $\mathcal{M}_0$.

After parameter estimation, the models were tested on records 25001 to 26000 records and validated on records 26001 to 27000 of the dataset. The test set 25001 to 26000 was unseen the first time it was used, however in the case of adjustments to optimize the model, the set was reused to test the model. On the other hand, the validation set, 26001 to 27000, were used only once - to validate the model after optimizing it- and was therefore unseen by the model until validation. The prediction of Z by $\mathcal{M}_0(\theta_0)$ operating on the validation data appears in Figure 26 and had an R^2 statistic of 0.9993. The prediction of Z by $\mathcal{M}_1(\theta_1)$ was very similar with an R^2 of 0.9963. No free-run prediction was performed because it is not defined for mappings of the form f : $X,Y \rightarrow Z$.

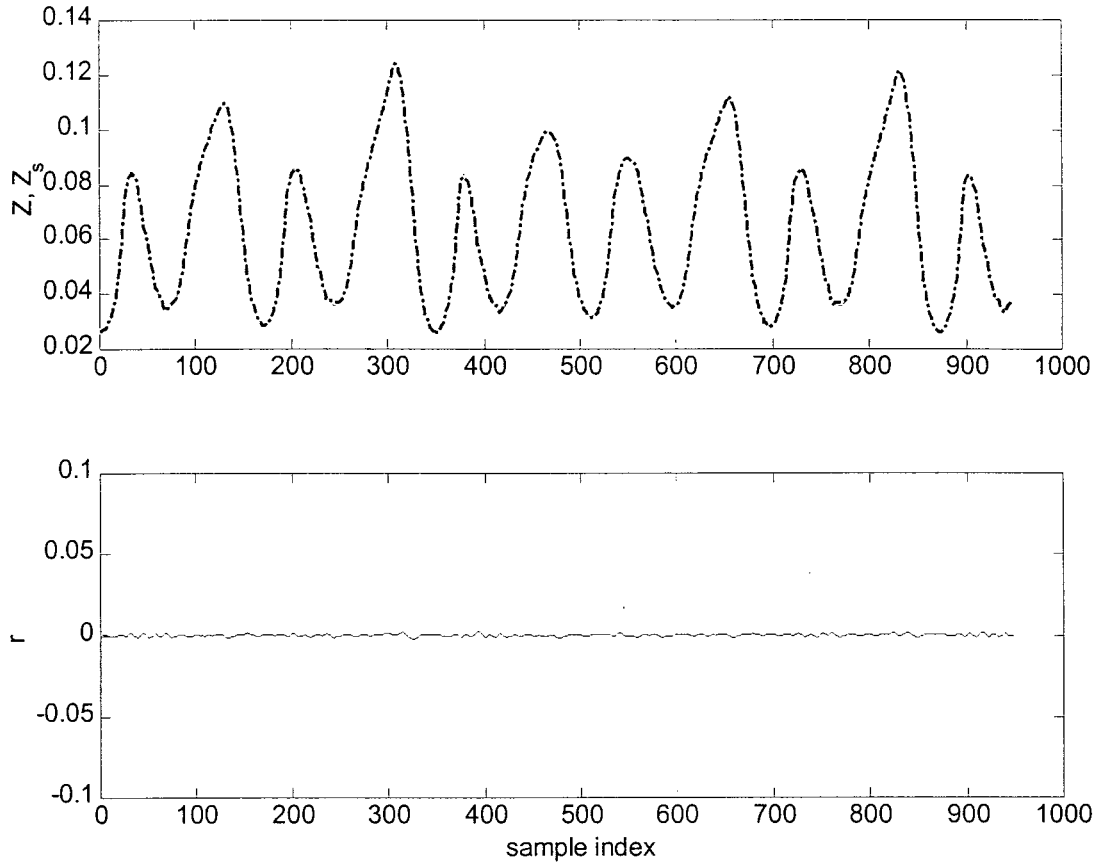


Figure 26 **Prediction of *Z*-state from *X* and *Y* taken from the validation data set. Z_s is the prediction (broken line) and *Z* the observation, while *r* is the prediction residue.**

5.2.2. NO_x-formation

Air pollution in terms of oxides of nitrogen (NO_x) is often a serious problem in metropolitan areas. NO and NO_2, collectively known as NO_x, are products of high-temperature, aerobic combustion. Combustion sources of NO_x are, among others, spark and compression ignition engines, oil-fueled power plants and tyre burning. Subsequent to the emission of these pollutants, they are subjected to solar radiation. The molecules absorb light and convert this energy into molecular energy. This photochemical reaction results in the formation of NO_2 from NO (Grobliki et al., 1981). Meteorological factors cause the NO_x pollutant to be transported and dispersed in the atmosphere. Not only are these pollutants a health hazard, but this reaction cycle is one of the precursors to photochemical smog (Dzubay, 1982).

There are essentially three main approaches to atmospheric dispersion modelling in general: the Eularian approach, the Lagrangian approach, and the statistical approach (Hassounah and Miller, 1994). The Eularian approach uses a continuity equation to develop a description of the physical and chemical processes that govern the relationships between emissions and the resulting concentrations. It is a rigorous model that addresses physical and chemical processes from first principles. It requires spatial and time resolved information of emission sources. Needless to say, it is extremely complex and requires vast amounts of input data and processing power.

The Lagrangian approach is the most frequently applied method. The motion of the pollutant particles in the atmosphere is modeled using a probabilistic description, and this in turn is used to derive expressions for pollutant concentration. The most commonly used probabilistic description is the Gaussian plume model. This model has a few simplifying assumptions that are restrictive (Turner, 1994). It assumes steady state conditions – the rate of emission is constant, the probability of the wind velocity is independent of time and location. Furthermore, the concentration of a pollutant along the vertical and crosswind axes, is assumed to be normally distributed.

With the statistical approach, statistical techniques are used to establish relationships between pollutant emissions, meteorological conditions and pollutant concentrations. The time series of the pollution data is correlated to synchronized emission and weather data using techniques such as multiple regression analysis, principal component analysis etc. However, most of these statistical models are linear. Hence, prediction of an evidently non-linear system is done using linear techniques. This is not ideal since a linear model will fail to identify important underlying non-linear dynamics.

During the past two decades, important progress has been made in the field of non-linear time series analysis. An empirical, non-linear regression model of the formation of NO_2 can be constructed based on synchronized observations of the chemical constituents and contributing environmental variables. The two main contributing environmental variables are usually taken to be ambient temperature and solar radiation. A predictive model would be valuable for metropolitan planning and management.

Air pollution in the Cape Town Metropolitan Area was selected as a basis for a case study. A data set was kindly provided by the Cape Town Metropolitan Scientific Services. The data set contained 8664 records of synchronized, hourly mean concentrations of NO, NO_2, ambient temperature and solar radiation for the city center of Cape Town, observed during 1996.

NO_2 concentration (Z_{NO2}) was chosen as the dependent variable with the aim to make acceptable one-step and free-run predictions in terms of NO concentration (X_{NO}) and solar radiation (X_{Es}). The first 6000 records was selected for fitting a non-linear model. Figure 27 shows NO_2 concentration. The sample size was determined by a stationarity test on Z_{NO2}, proposed by the author. The test involved dividing a data set of some initial size into two halves, Y_1 and Y_2. Then the joint histogram $P(Y_1,Y_2)$ for points of corresponding index in Y_1 and Y_2 was calculated. Figure 28 shows the evolution in the difference in centroid of the joint histogram $P(Y_1,Y_2)$ between iterations. The indicated point of sufficient stationarity was determined from convergence of the mean difference over a moving window of 64 iterations. Refer to section A.6 for details. A total of 452 outliers were detected and removed from the full observation space, using the convex hull technique (Barnard et al., 1999b) described in Chapter 4. There was no way to establish how many false outliers existed in the total number of indicated outliers. An side effect of running the outlier detection algorithm repeatedly, is to systematically reshape the data space towards a spherical geometry due to the spherical outlier criterion, therefore the algorithm was iterated only once on the data set. It was accepted that some removed data were in fact the result of the dynamics and therefore false outliers. The integrity of the technique as demonstrated in Chapter 4 was trusted and it was assumed that the percentage of false outliers was not excessive.

After removing outliers, the data was classified as either deterministic or stochastic, using the surrogate data method described in 3.1. Briefly this entailed that surrogate data were generated based on the Z_{NO2} data by randomizing the sample index and the Fourier spectrum. Correlation dimension curves were calculated for both data and surrogates, using the algorithm by Judd (1992), and mutually compared. The data appeared very random from inspection of the correlation dimension curves (Figure 29), but not entirely random, otherwise the curve for Z_{NO2} would have been located among those of the surrogate data. The consistently lower dimension values for Z_{NO2} indicated some inherent dynamics existed. In the absence of a known noise model, and given the risks of noise-reduction in non-linear systems (discussed in section 2.3.2) no explicit pre-filtering was applied to the data. The largest Lyapunov exponent was estimated as 0.974. The Lyapunov exponents were calculated, using the algorithm by Brown et. al. (1991) as implemented in the commercial software package, cspW, by Applied Nonlinear Sciences. Using equation (**12**), the prediction horizon was then determined:

$$t_L = \frac{\tau_s}{\lambda_1}$$
$$= \frac{1}{0.974}$$
$$= 1.02$$

This result implied a prediction horizon of 1.02 h , which meant that any model could expect difficulty under free-run prediction. Consequently, one-step predictions were the best one could hope for, unless some noise-reduction was applied.

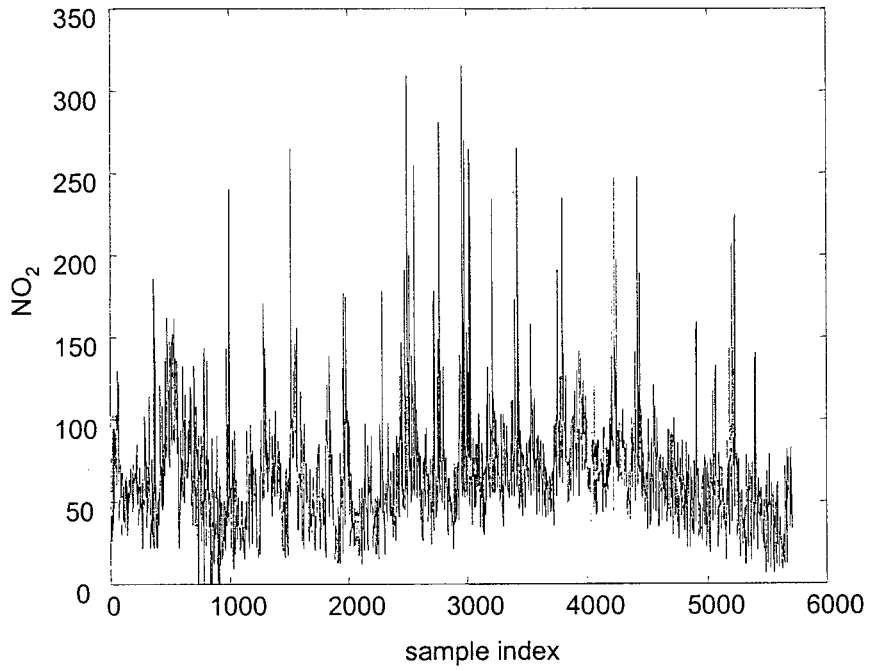


Figure 27 NO_2 data used for fitting a non-linear MLP model. No filtering was applied, but outliers were removed.

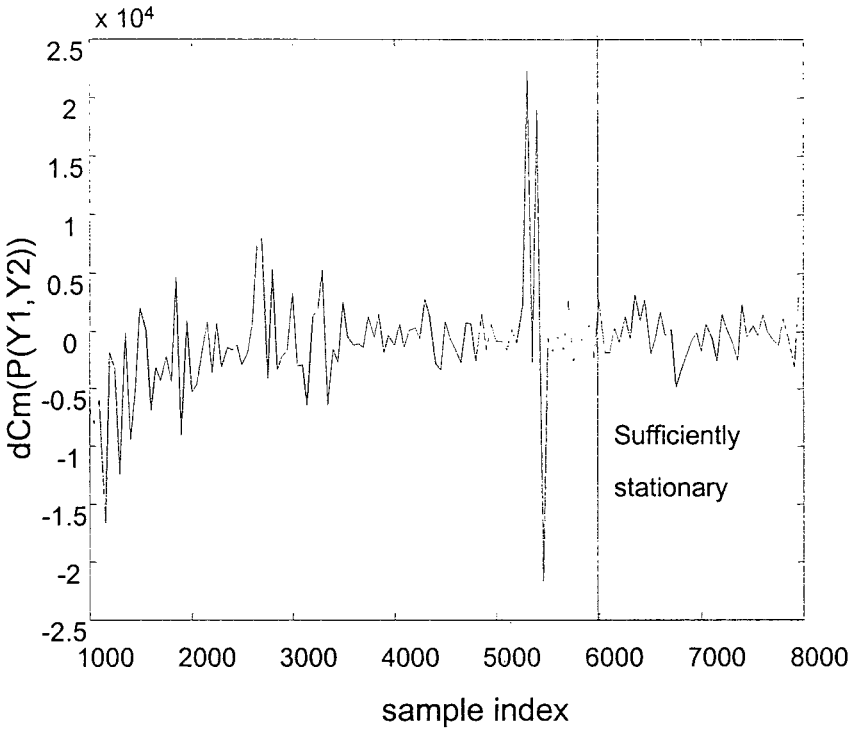


Figure 28 **Stationarity test on NO_2 concentration. The difference in the centroid of the joint probability matrix between iterations is indicated by $dC_m(P(Y_1,Y_2))$.**

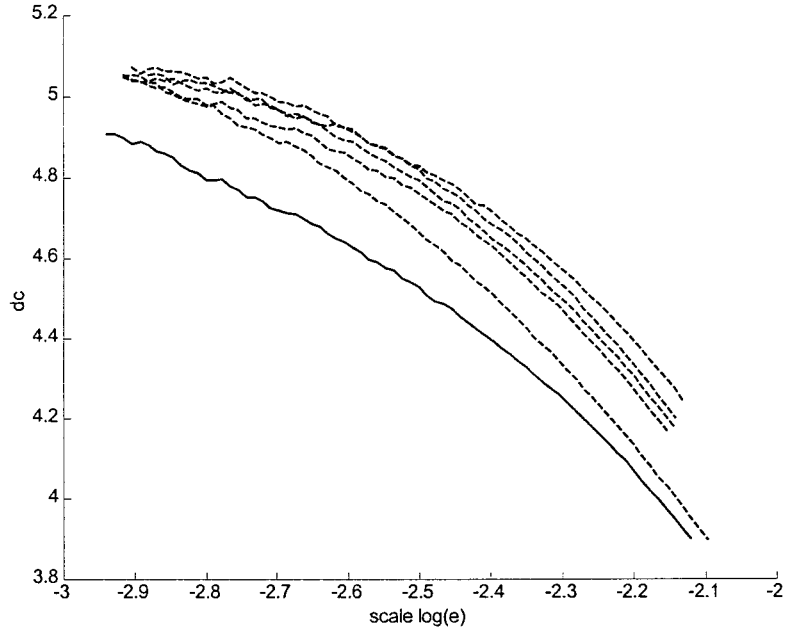


Figure 29 **Correlation dimension curves for NO_2 concentration (solid line) and non-linear transformed random surrogate data (dashed line) based on NO_2 concentration.**

A principal component analysis (Table I) indicated that environmental temperature had no major influence on Z_{NO2} and this component was consequently removed from the data set. Table I lists the loading of each principle component with respect to each variable in the data set. In addition, the column headings of the table include the eigenvalues as λ_i as well as the cumulative percentage of variance declared by each principal component. From inspection of the table it is clear that environmental temperature played a small role in terms of the first, second and third principle components, with loadings of 0.1040, 0.0677 and 0.1265 respectively. Cumulatively, the first three principal components declared 98% of the total variance in the data. The system was therefore redefined in terms of the three remaining observed variables: X_{NO}, X_{Es} and Z_{NO2}. Optimal parameterization of the system was investigated in terms of two alternative embedding strategies and associated model structures. Parameterization according to the first strategy implied embedding all observation components using optimal individual embeddings summarized in Table II. In terms of the second strategy only the Z_{NO2} was embedded and the observed X_{NO} and X_{Es} were included as one-dimensional time series (Table III). The motivation behind the first embedding strategy was that the three variables were all dependent observations of the microclimatic system and thus allowed an embedding of each variable. On the other hand, the second strategy treated X_{NO} and X_{Es} as control variables and thus independent, which did not justify embedding them. For conciseness, $\mathbf{M}_{2\theta}$ is used instead of $\mathbf{M}_2(\theta_2)$, and so forth.

Table I **Results from principal component analysis of air pollution data.**

	PC_1	PC_2	PC_3	PC_4
	λ_1: 9.065e-005	λ_2: 6.798e-005	λ_3: 1.604e-005	λ_4: 2.903e-006
	(51% variance)	(89% variance)	(98% variance)	(100% variance)
Z_{NO2}	0.3163	-0.3527	0.8722	-0.1213
X_{NO}	0.4135	-0.7841	-0.4575	0.0691
X_T	0.1040	0.0677	0.1265	0.9841
X_{Es}	0.8473	0.5060	-0.1178	-0.1093

In accordance with the first embedding strategy, the individual embeddings were combined as $\Lambda_2 \in \Re^{17}$ and separated as $S_2=W_2\Lambda_2$ by ICA with the G3 contrast function. The following model structure was to be fitted:

$$\mathcal{M}_2: S_2(t) \rightarrow Z_{NO2}(t+1),\ \mathcal{M}_2 \in \mathcal{M}^*_{FF} \tag{30}$$

In accordance with the second embedding strategy, individual embeddings were combined as $\Lambda_3 \in \Re^{7}$ and separated as $S_3=W_3\Lambda_3$. The associated model structure was:

$$\mathcal{M}_3: S_3(t) \rightarrow Z_{NO2}(t+1),\ \mathcal{M}_3 \in \mathcal{M}^*_{FF} \tag{31}$$

To test the influence of ICA, the system was again parameterized in accordance with the first embedding strategy, but without separation by ICA, resulting in the following model structure:

$$\mathcal{M}_4: \Lambda_4(t) \rightarrow Z_{NO2}(t+1),\ \mathcal{M}_4 \in \mathcal{M}^*_{FF} \tag{32}$$

Finally it was investigated if there was any advantage in embedding by specifying a fourth parameterization, without embedding of any variable. The associated model structure was:

$$\mathcal{M}_5: [X_{NO}\, X_{Es}] \rightarrow Z_{NO2}(t+1),\ \mathcal{M}_5 \in \mathcal{M}^*_{FF} \tag{33}$$

Table II **Embedding parameters for air pollution data: strategy 1**

	NO_2	NO	E_s
m	5	5	7
k	9	14	9

Table III **Embedding parameters for air pollution data: strategy 2**

	NO_2	NO	E_s
m	5	1	1
k	9	0	0

MLP networks with single hidden layers were selected as model structures in all cases. For $\mathcal{M}_{2\theta}$ each hidden layer initially consisted of 16 bipolar sigmoidal nodes (activation functions

of the form $g(\cdot)=[1-\exp(\cdot)]/[1+\exp(\cdot)]$). In addition, a bipolar linear output layer was used in each case. The network parameters were estimated by using the Levenberg-Marquardt algorithm. The initial model order (7 nodes) was selected on the basis of research (Lawrence et al., 1996) that indicated that the back-propagation algorithm will not get trapped in local minima of the error surface when one less node than the dimension of the input space is utilized. From this point onward in the optimization of the network, the number of nodes were doubled at each optimization iteration.Over-fitting was indicated by the network performing worse against the R^2 statistic during testing with 16 hidden nodes than when using 8 hidden nodes, while with 7 hidden nodes the network again performed sub-optimally. Therefore, 8 hidden nodes were taken as optimal. A more elaborate evaluation of model performance by surrogate techniques was impossible due to the high noise content of the data and the short prediction horizon calculated earlier. These conditions disabled free-run predictions over long enough periods for reliable application of a non-linear surrogate technique, which requires at least 1000 predicted points for sufficiently accurate estimation of correlation dimension. The optimal topologies for the other models were determined in similar fashion and are listed in Table IV.

Table IV **Optimal MLP network topologies for various parameterizations of the NO_x system.**

	$M_{2\theta}$	$M_{3\theta}$	$M_{4\theta}$	$M_{5\theta}$
Hidden Nodes	8	14	14	8

All models were tested by doing one-step and free-run predictions on observations 6001 to 7000 and validated on observations 7001 to 8000. These data sets were embedded and separated using the respective embedding parameters and separation matrices as calculated from the training data. The validation results for one-step and free-run predictions according to the first embedding strategy are shown in Table V and Figure 30, and the results according to the second strategy in Table V only.

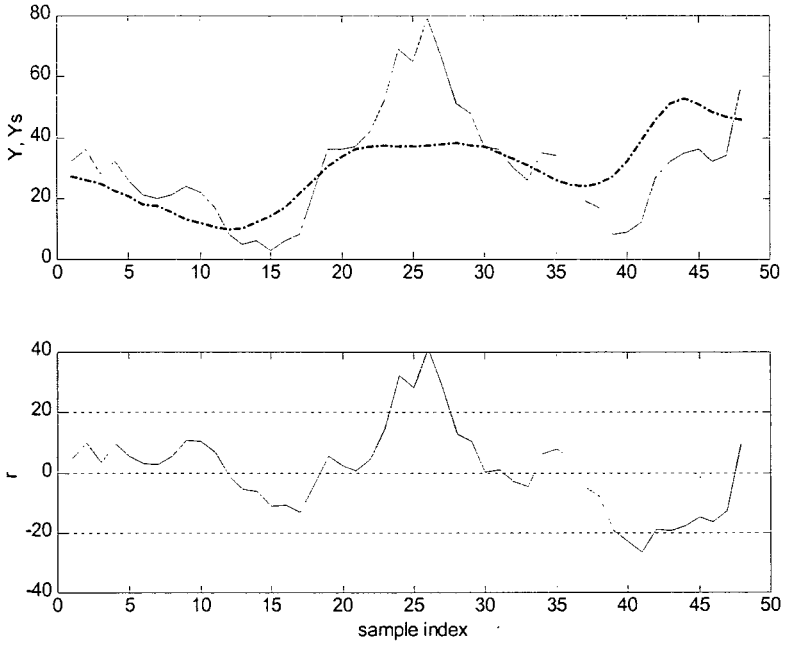
Table V R^2 **statistics for one-step prediction of** Z_{NO2} , X_{NO} , X_{Es}

	$\mathbf{M}_{2\theta}$	$\mathbf{M}_{3\theta}$	$\mathbf{M}_{4\theta}$	$\mathbf{M}_{5\theta}$
$\mathbf{Z}_{NO2}$	0.7937	0.7936	0.8244	0.7374
$\mathbf{X}_{NO}$	0.1989	0.2346	0.1741	0.1632
$\mathbf{X}_{Es}$	0.9315	0.9143	0.9267	0.8621

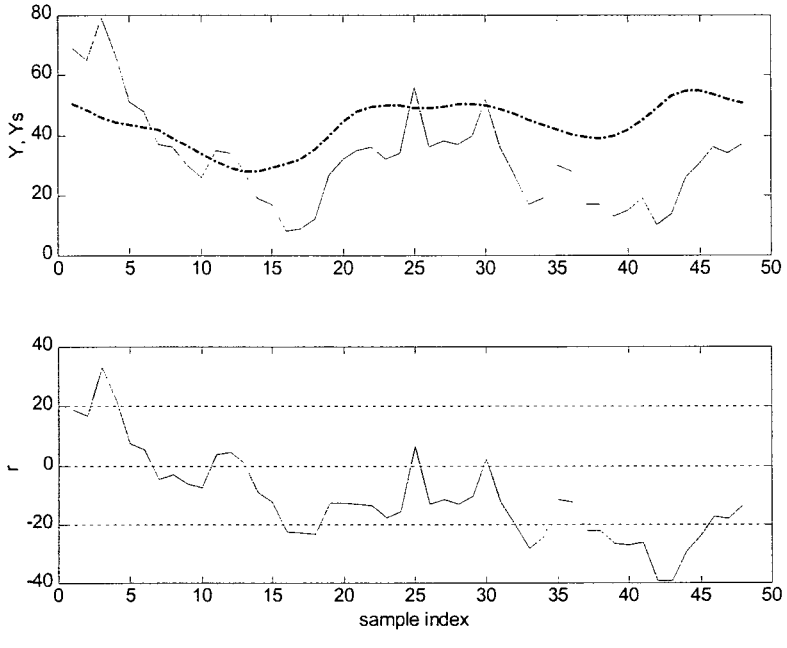
One-step predictions did not reveal significant difference between the embedding strategies, as can be seen from the R^2 statistics in Table V. For example, Z_{NO2} was predicted best by $\mathbf{M}_{4\theta}$, X_{NO} by $\mathbf{M}_{3\theta}$, and X_{Es} by $\mathbf{M}_{2\theta}$. However, parameterization without embedding clearly led to a less descriptive model of the system, as indicated by $\mathbf{M}_{5\theta}$ in Table V. Separation of embedding variables by ICA appears to have a slight negative influence on the linear correlation between prediction of Z_{NO2} with observed Z_{NO2}. To clarify whether this meant that $\mathbf{M}_{2\theta}$ and $\mathbf{M}_{3\theta}$ had failed to describe all determinism in the data, the prediction residual was subjected to the Hinich test (see A.5) and found to be Gaussian. This result indicated that the model did describe all the inherent dynamics and that the noise component is Gaussian. Consequently, the independent components of $\mathbf{S}_2$ were tested for randomness. $S_{2,7}$, $S_{2,16}$ and $S_{2,17}$ were found to be Gaussian with more than 70% confidence. On the other hand, none of the components of the Λ_4 were Gaussian, which implied that the noise was mixed with the signals. The conclusion was that separation by ICA after embedding actually managed to separate the noise components from the deterministic components. The inclusion of explicit noise components in the input space probably caused less optimal parameter estimation than when spreading the noise over all the input components as for $\mathbf{M}_{4\theta}$. Due to time constraints on the research project, the data were not reconstructed without the noise components, but this could be a sensible further investigation.

Iterative free-run prediction was attempted up to 48 steps ahead. Then the input to the model $\mathbf{M}_{2\theta}$ was reset with the embedding of Z_{NO2} , X_{NO} and X_{Es} at that index and another 48 steps were predicted. Results were poor, as expected from the short stability horizon, and are shown in Figure 30. Consequently no surrogate analysis was performed on the free-run prediction results. A more advanced model class and a suitable noise-reduction scheme should improve the extreme modeling situation. The results for one-step prediction of Z_{NO2} by $\mathbf{M}_{2\theta}$ over the first 200h of the validation data set appear in Figure 31. The mean prediction error normalized

with the mean observation of Z_{NO2} was 0.2481. It was concluded from these results that the model was capable of sufficiently accurate one-step predictions. The ability of the model to follow the peaks of the observed NO_2 concentration is particularly significant, since the maximum expected levels of NO_2 play an important role in environmental management and planning.



(a)



(b)

Figure 30 **Validation of a MLP network model by simultaneous free-run prediction of NO_2, NO and E_s, using embedding strategy 1. The result for NO_2 is shown here for the (a) first 48 hours, (b) next 48h. Y_s and Y are prediction (dashed line) and observation (solid line) respectively, while r is prediction residue.**

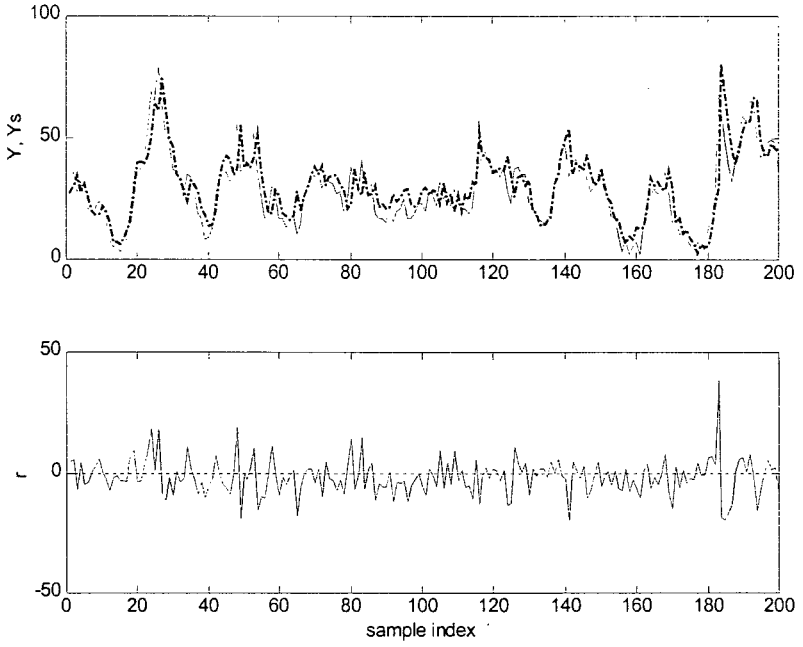


Figure 31 Validation of a MLP network model by simultaneous one-step prediction of NO_2, NO and E_s, using embedding strategy 1. The result for NO_2 is shown here for the first 200 hours. Y_s ,Y, and r are defined as above.

5.3. Conclusions

In this chapter a method for the embedding of a multidimensional time series was proposed. The method was demonstrated on two case studies, a chaotic autocatalytic process and the formation and dispersion of atmospheric NO_2. Results from the first case study suggested that the embedding method had enhanced the ability of a non-linear regressor to simulate a state of chaotic dynamics in terms of the other states. In the second case study the embedding method enabled acceptable one-step prediction of NO_2. The model based on a parameterization without embedding, performed significantly worse than the other two models that had used embedding strategies. In addition it became clear that multivariate embedding combined with separation of components by ICA enabled at least partial separation of noise components from the deterministic content.

No noise reduction was done prior to estimating the model parameters, so as to preserve small scale dynamics in the observations. Improvements in terms of the selection of a model structure in the first case study, and noise-reduction in the second case study, will probably

improve these results. The fundamental advantage of the multidimensional embedding technique has, however, been demonstrated here, especially in the presence of noise.

6 ONLINE DIAGNOSIS OF TEMPORAL TRENDS IN CRITICAL SYSTEM STATES

In the previous chapters techniques for the acquisition of a stationary data set, selection of system variables, system parameterization, noise-reduction, outlier detection, model parameter estimation and model validation have been addressed. In this chapter most of these techniques are assembled in a method for the on-line diagnosis of temporal trends in a critical state of an internal combustion engine. Such an application presents several challenges: multivariate non-linear modeling in the presence of both measurement and dynamic noise at unspecified levels, the stationarity of recorded data, outlier handling and the simultaneous on-line interpretation of several system variables.

The essence of the diagnosis of temporal deterioration in critical systems states lies in the simultaneous interpretation of more than one system variable and an estimation of the likelihood of finding the system in a certain state. For the interpretation of simultaneous states, a reliable and sufficiently accurate model that describes the relevant output behaviour of the system is required. In addition, the simulation error during system operation can be compared statistically to the simulation error that resulted from simulating the stationary data set from which the model parameters were estimated. This evaluation is, broadly spoken, a likelihood estimation of the system being in the current state and can be used to flag the operator when the current state is unlikely and therefore in a probable failure condition.

Automation of state observation in automotive engines has developed significantly during this decade. Several proprietary computerized control and data-logging systems exist in the automotive industry. However, these systems often fall short regarding simultaneous interpretation of measured engine states. Recently, the application of MLP network models of an internal combustion engine entered into engine control and diagnostics. Atkinson et al. (1998) developed MLP network models as virtual sensors for adaptive control of engine emissions as well as diagnostic purposes. A partially recurrent neural net was fitted to linearly filtered data recorded on spark and compression ignition engines. Grimaldi and Mariani (1997) experimented with various MLP networks in an investigation to model various measured engine states. Their research was ultimately aimed at meeting On Board II Diagnostic (OBD II) regulation by the California Air Resources Board on engine emission

control diagnostics. However, the important issue of data sample length, which governs model generalization, was not sufficiently addressed. Also, they used linear filters to attain good model fit, which could have removed interesting and subtle dynamic information (Kostelich and Yorke, 1988).

In this chapter, the simultaneous statistical interpretation of multivariate observations using a MLP network and independent component analysis are proposed. Subsequently, the diagnostic procedure is applied to data recorded of a compression ignition engine under test conditions.

6.1. Statistical interpretation of observation and simulation

The simulation model of a multivariate system proposed in this chapter is expressed mathematically as:

$$f : \mathbf{X} \rightarrow Y \tag{34}$$

which simulates the dependent state vector, Y as

$$\hat{Y} = f(\mathbf{X}) \tag{35}$$

Assume the system to be in either a normal operating condition or a failure condition, caused by some mechanism of temporal deterioration of the system. The model will be applied to detect failure of a system by simulating a critical dependent state, given the independent variables. Since the model will preferably not declare the measurement noise component of the data, and also may not declare some of the secondary dynamics, there will be significant temporal variance in the simulation error,

$$\mathbf{r} = Y - \hat{Y} \tag{36}$$

Therefore a joint statistical evaluation of the operational history of the observed dependent state and independent states is used to diagnose the system. The hypothesis is tested that the system is not in a failure condition, based on the probability to observe a given history of simulation error values. An internal combustion engine is used as subject in the following statistical reasoning.

Assume a sequence, S, of N independent, multidimensional observations of a running internal combustion engine. Let Y be an observation of a dependent variable and X an observation of

an independent variable, both from S. The total probability to simultaneously find Y in one specific region of S and X in another, can be expressed in terms of conditional probability:

$$P(Y) = \sum_{n=1}^{N} P(Y \mid X) P(X) \tag{37}$$

The above equation can be simplified and expressed in terms of the appropriate joint probability:

$$P(A) = \sum_{n=1}^{N} P(Y \cap X) \tag{38}$$

Instead of using a multidimensional X, the joint probability can be reformulated in terms of $\hat{Y}$ given in (**35**):

$$\begin{aligned} P(Y) &= \sum_{n=1}^{N} P(Y \cap \hat{Y}) \\ &\approx P(\mathbf{r}) \end{aligned} \tag{39}$$

where the probability distribution of the simulation error, $P(\mathbf{r})$ over N, approximates the joint probability.

The probability $P(\mathbf{r})$ can be implemented in terms of a hypothesis test. Formulate the null hypothesis that the engine is in normal operating condition. The null hypothesis will be accepted if the actual probability of observing the simulation error beyond a given limit percentile does not exceed the threshold probability of failure:

$$P(\mathbf{r} > \mathbf{r}_{P_L}) < P_{\Im}(\mathbf{r} > \mathbf{r}_{P_L}), \tag{40}$$

where subscript $\Im$ signifies threshold. If the null hypothesis is rejected, the engine enters a failure condition. To estimate $P_{\Im}(\mathbf{r}>\mathbf{r}_{\mathrm{PL}})$, set a limit percentile for the simulation error in the training data, which would indicate the upper bound of the normal operating region. Let a window of size N_w move along the training residue vector. Count the simulation error hits beyond the limit percentile and normalize with N_w. Average the vector of normalized scores over N–N_w. This mean, normalized simulation error score is an estimation of the probability to observe the simulation error beyond the limit percentile. Finally, multiply the warning and failure threshold factors with this probability to obtain the warning and failure threshold probabilities, respectively.

6.2. Diagnostic methodology

The diagnostic method proposed in this paper can be categorized conceptually into three stages, i.e. preprocessing, model construction, and application. Briefly stated, the procedure consists of the following steps:

1. Acquisition of a stationary data set.
2. Removal of outliers.
3. Estimation of the parameters of a non-linear, multivariate model, based on the data set.
4. Estimation of the probability of the simulation error to be in the extreme acceptable operating region, based on the data set used for model parameter estimation.

During operation of the engine, the model simulates the dependent variable while engine condition is diagnosed based on the statistical analysis of the simulation error in a moving time window of fixed size.

6.2.1. Preprocessing

The data as recorded can not be used directly to perform malfunction detection, but has to be preprocessed first. The aim of preprocessing is to convert ASCII data files into numerical arrays, standardize and scale the data arrays and construct input and output spaces for the simulation model. Stage one consists of the following steps:

1. Identify the dependent state, Y, and independent states, $\mathbf{X}=\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots \mathbf{x}_m\}$.
2. Determine the minimum size of the data set, $\mathbf{Z}=[\mathbf{X}\ Y]$, that is stationary in terms of Y and represents the engine under optimal performance conditions, as described in section A.6.
3. Standardize and scale data components so that $||Z_j|| = 1$ for $j = 1 \ldots m+1$.
4. Detect and remove all outliers from the training set, as defined in Chapter 4.
5. Separate $\mathbf{X}$ into independent components, $\mathbf{S}$, so that these declare 99% of variance, using equation (**21**) as described in sections 2.2.1 and 5.1.2.
6. Combine input and output spaces as the training set, $\mathbf{Z}=[\mathbf{S}\ Y] \mid \mathbf{Z} \in \Re^{s+p}$, $\mathbf{S}\in\Re^{s}$, $Y\in\Re^{p}$.
7. Standardize and scale $\mathbf{Z}$ so that $||\mathbf{z}_i|| = 1$ and $\bar{\mathbf{z}}_i = 0$.

6.2.2. Model selection and parameter estimation

In Stage two the model is fitted to the training data, tested for fitness and validated. Finally, the statistical criteria required for engine diagnostics are calculated. This stage consists of the following steps:

1. Fit a MLP network model structure to the training data as described in section 2.3.5. As a heuristic rule, set the initial number of hidden nodes to double the dimension of input space.
2. Test the model for fitness on part of the data not used for fitting the model as described in section A.5. Since the model will be a simulation model and not a predictive model, model performance can not be measured using free-run prediction and non-linear surrogate methods, since free-run predictions are not defined for pure simulations. R^2 is therefore an acceptable minimum criterion of model performance. Apply a R^2 of 0.90 as minimum fitness criterion and a successful Hinich test as the maximum fitness test. If the model fails the Hinich test and the R^2-test, attempt doubling the number of hidden nodes again, refit and retest the model. If the model still fails the Hinich test, but passes the R^2-test, double the number of hidden layers, keeping the previous number of hidden nodes. If the model again fails the Hinich test, but passes on a randomized index, keep this topology. The simulation error distribution is already approximately Gaussian, with only minor remaining correlated information.
3. Validate the model by simulating another section of data, unseen by the fitting algorithm and not used during testing of the model. Apply the criteria in step 2 (above) on the simulation error. Use a new validation set if any changes are consequently made to the model.
4. Determine warning and failure threshold probabilities as described in section 6.1.

6.2.3. Applying the model

In Stage three the model is applied to simulate the dependent state on-line and the simulation error is statistically evaluated in a moving time window. This stage consists of the following steps:

1. Start observing all relevant states until the moving time window of N_w samples is filled, then proceed at the sampling rate from step 2 onwards.

2. Detect and remove outliers through a single-iteration convex hull procedure.
3. Calculate $\mathbf{S} = \mathbf{WX}$ over the moving time window, using $\mathbf{W}$ from the training set.
4. Standardize and scale the data using the scaling vectors from the training set.
5. Count the observations in the moving time window of simulation error beyond P_L that was specified for the training data. Normalize this score with N_w to get an estimated $P(\mathbf{r} > \mathbf{r}_{PL})$ and compare with the threshold $P_{\Im}(\mathbf{r} > \mathbf{r}_{PL})$ calculated on the training data. Test the null hypothesis against the norm in equation (**40**). If the hypothesis is rejected, the engine enters failure state.

6.3. Diagnosing a Diesel engine under endurance testing

The diagnostic methodology was applied to data recorded during endurance testing of a commercial, turbo-charged and inter-cooled Diesel engine on an absorption dynamometer. The sampling period was set to three seconds in order to capture interesting transient behaviour after changeover between set points in the test schedule. A total of 751695 data records were observed during 626h of testing. All irregular events were logged by the test operator. The engine performed without apparent malfunction. The aim of the exercise was to detect online an imminent need for maintenance overhaul of the engine, due to excessive piston ring wear.

The "blow-by gas flow" of the engine was chosen as the dependent variable indicating piston ring wear from a total of thirteen observed states. Data directly following the first 120h run-in period were used for system identification. The minimum stationary data set size was determined by increasing the initial data sample of 100 records in increments of 100 records. The change in center of mass of the half samples' joint probability matrix was tested for Gaussianity over a moving time window of 128 iterations. The test for Gaussian randomness of the center of mass of the half samples' joint probability matrix is a very conservative criterion for stationarity under the formulation as used in this dissertation. High confidence levels such as 90 or 95% would result in over-estimation of the stationary length. This was established by calibrating the stationarity test during a tedious process of iteratively modeling the data using data sets of increasing size at each iteration and testing the generalization of the model on a later section of the recorded data. For the foregoing reasoning and from inspection of the results in Figure 32, a stationary length of 46200 records (38h) was accepted at a

moderate confidence level of 66%. All consequent system identification was performed using a data set of this length.

The selected independent states were engine speed, torque load, induction boost pressure, coolant temperature at heat exchanger outlet, engine oil temperature, diesel temperature and post-intercooler temperature (Figure 33). The normalized AXMI between these variables at zero lag and at a cut-off level of 0.50, indicated that engine speed, torque load, induction boost pressure, Diesel temperature and post-inter-cooler temperature were strongly correlated with the blow-by flow rate. After separating the initial input space of selected independent states, the first two independent components declared 99% of the variance, and formed the reduced input space to the simulation model.

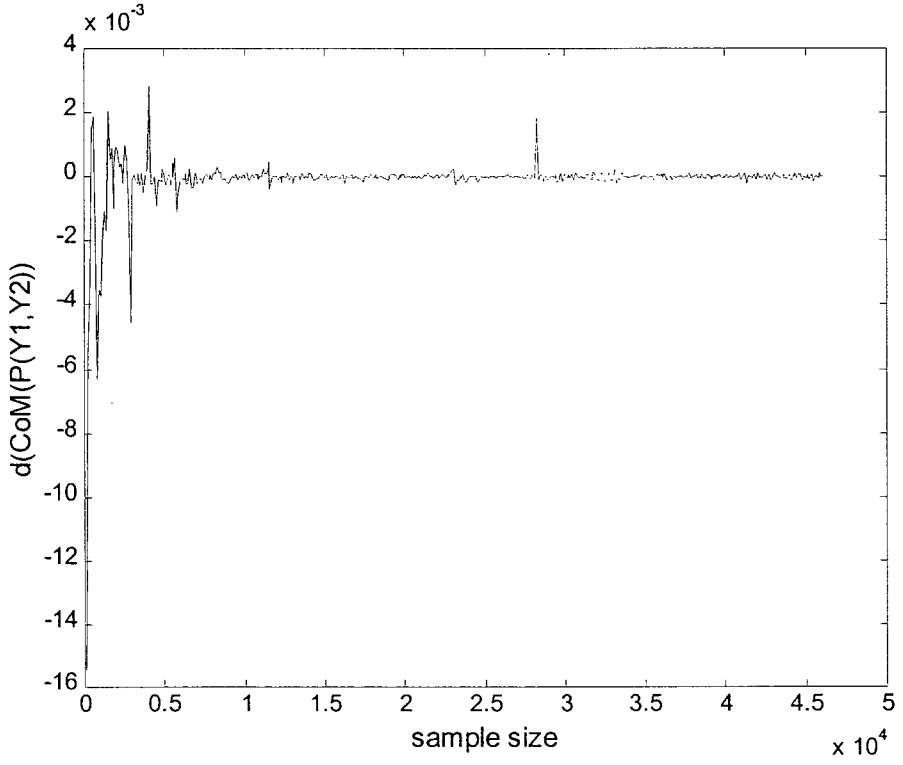


Figure 32 Change in center of mass of joint probability for half samples of increasing sample size.

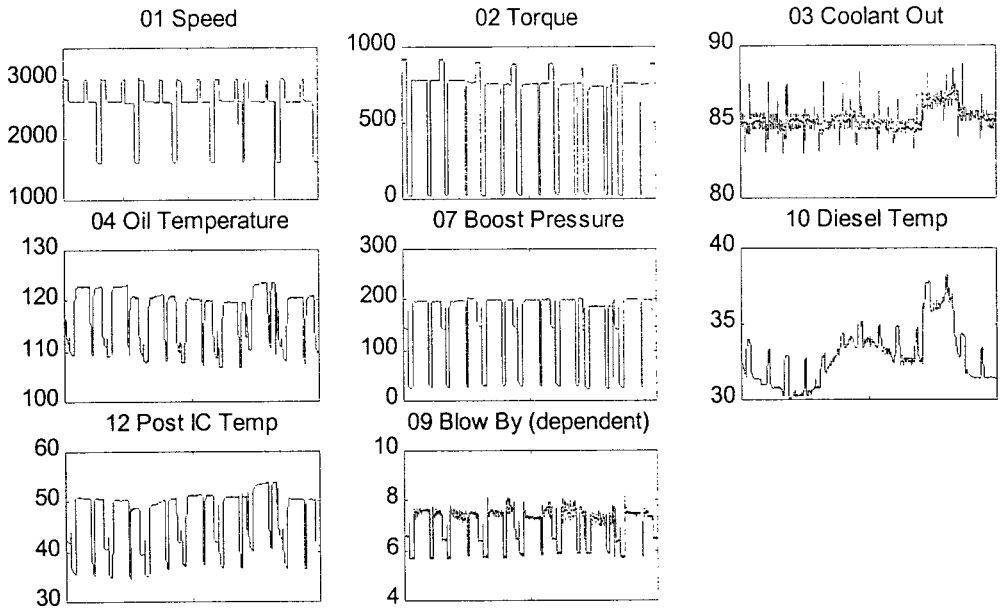


Figure 33 Independent and dependent observed states used in diagnostic model, after removing outliers.

Outliers were eliminated from the combined input and output spaces before reducing the input space. The outliers were caused by typical test disruptions as well as scheduled maintenance interruptions. After removing 91 outliers at detection sensitivity level 1.0, a feed-forward neural net was fitted to the remaining data. The heuristic rule mentioned in section 6.2.2, dictated that an initial four hidden nodes be used (twice the dimension of input space). The Levenberg-Marquardt training algorithm converged within four iterations.

Model fitness was evaluated first in terms of randomness in the simulation error, using the Hinich Gaussian test. After two trials of network optimization, a final topology with a double hidden layer with 4 nodes each was used. The null hypothesis that the simulation error was Gaussian was rejected, but accepted when the simulation error index was randomized, indicating a small remaining deterministic content in the simulation error. The R^2 statistic for the training data was 0.961. The model was tested on data from record 60000 to 80000, resulting in an R^2 statistic of 0.932. Model validation was done on data records 80000 to 90000, resulting in an R^2 of 0.926. The model rejected the noise and small scale dynamics, while it simulated the global, large scale dynamics very well. This is in accordance with the original intention not to use filtering to improve fit, but rather to rely on model structure, dimension and the training algorithm to reach an optimal model. The model was accepted, based on the overall result.

In order to test the effect of outliers on modelling accuracy, a feed-forward neural net with the same topology as the above neural net was fitted to the data without removing outliers. Because of the very low outlier content, the R^2 statistic for the training data was only slightly lower at 0.955.

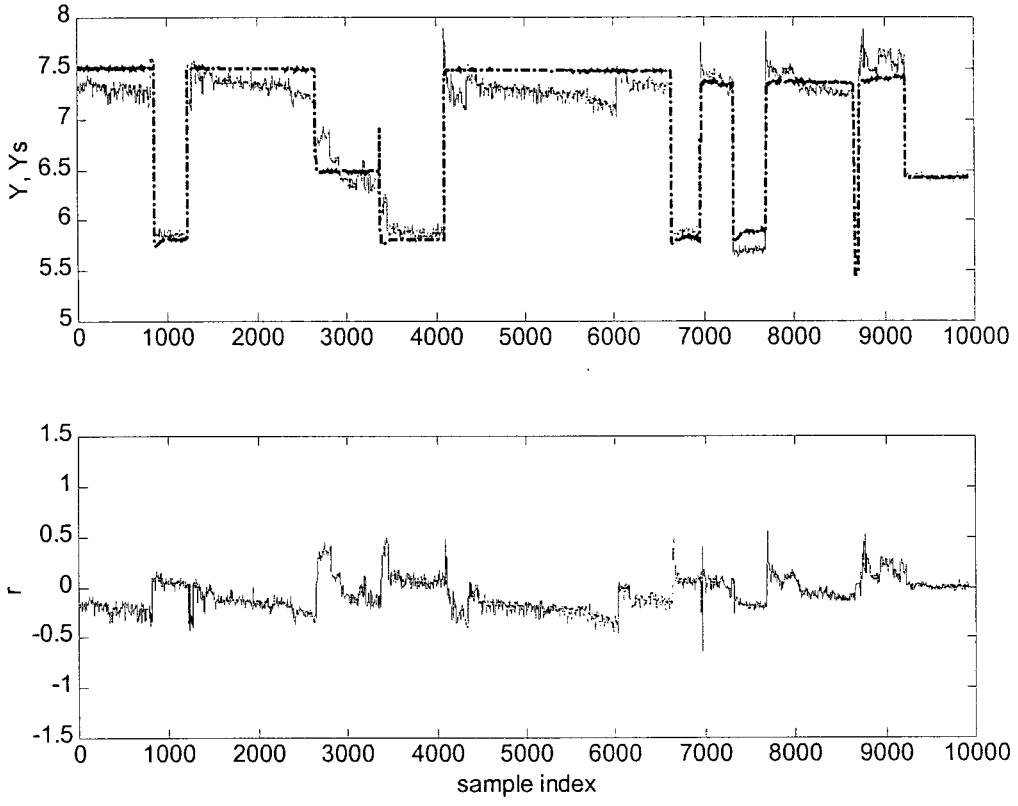


Figure 34 Top sub-plot: Simulation of blow-by gas flow (broken line) and observed blow-by from the validation data set. Bottom sub-plot: Simulation error.

Diagnosis of engine failure was first demonstrated on the same section of data as used for validation. This section, consisting of 10000 records, was divided into halves, which will be called half samples. The second half sample was modified so that the blow-by flow rate increased by a factor that changed along a linear ramp from 0 to a maximum of 1.3 at the end of the half sample (Figure 35). The aim was to detect the point in time where the diagnostic method would indicate failure condition, based on the modified observed and simulated blow-by gas flow data. The estimated probability of observations in the critical region was compared with the acceptable probability in this region inferred from the training data. The failure condition would be indicated when the probability of observations in the critical region

exceeded the acceptable probability. The lower boundary of the critical region was defined by the mean 99% percentile (P_{99}) of a 1000 point window moving along the training residue. The acceptable probability to observe blow-by gas flow values in this region was the normal cumulative density function, $1-N_{P_{99}}(\bar{\mu}_r,\bar{\sigma}_r)=0.00690$, with $P_{99}=0.220$, $\bar{\mu}_r=E[E[r]]=-0.00531$ and $\bar{\sigma}_r=E\left(\sqrt{E[(r-\bar{r})^2]}\right)=0.0914$. The simulation error in a 1000 point window conformed to the Hinich normality test after randomizing the sample index. Using the artificially induced failure condition, a caution threshold as well as a failure threshold was calibrated as $5\,N_{P_{99}}(\bar{\mu}_r,\bar{\sigma}_r)$ and $20\,N_{P_{99}}(\bar{\mu}_r,\bar{\sigma}_r)$ respectively. For the induced failure and threshold settings in this case study, the engine briefly visited the warning state between records 2800 and 3800 (before the induced failure) and entered the failure condition at just on record 7000 as shown in Figure 36.

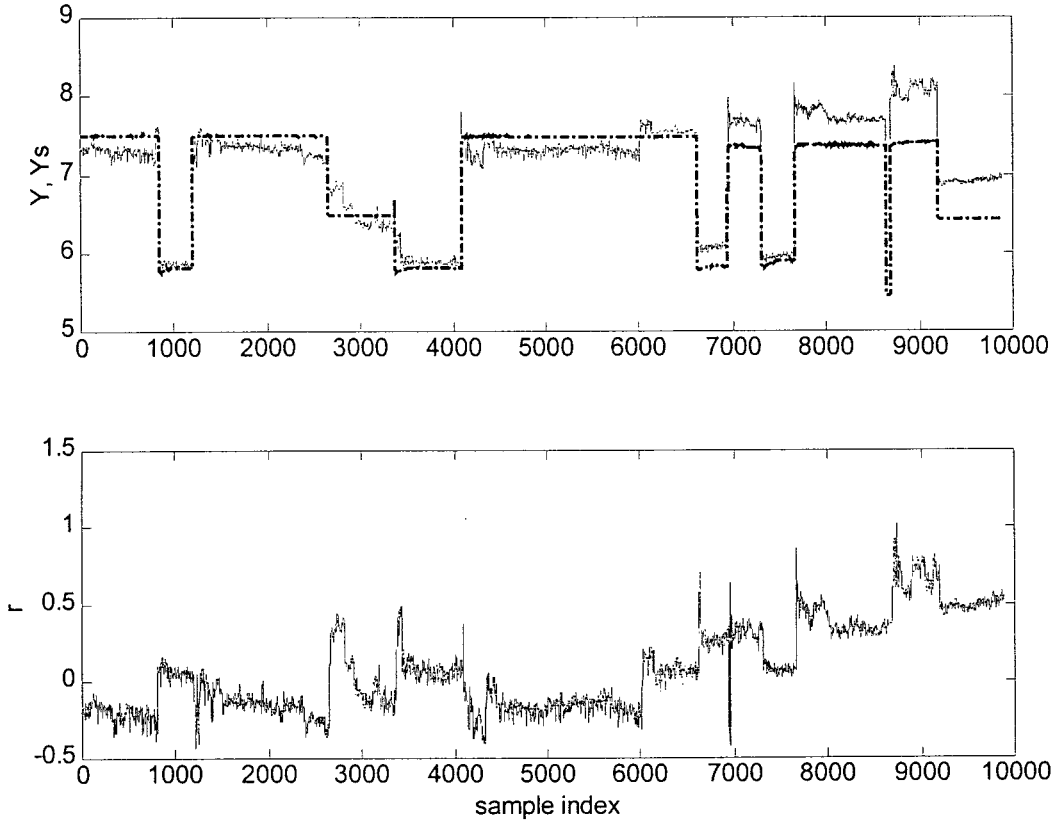


Figure 35 **Top sub-plot: Blow-by gas flow, observed (solid line) and simulated (dashed line), for artificially induced excessive blow-by gas flow. Bottom sub-plot: Simulation error.**

A second data set of 10 000 records at the end of the test was diagnosed in terms of blow-by gas flow. The diagnostic results in Figure 37 indicate that the blow-by gas flow briefly exceeded the caution threshold, which could be attributed to the piston rings starting to wear measurably. The short spell of higher than usual variation in simulation residue that occurred after the transient from low to high load, is an indication of increased instantaneous fluctuation in blow-by gas flow. On another segment from the second data sample the diagnostic method indicated a brief visitation to the failure zone, as shown in Figure 38.

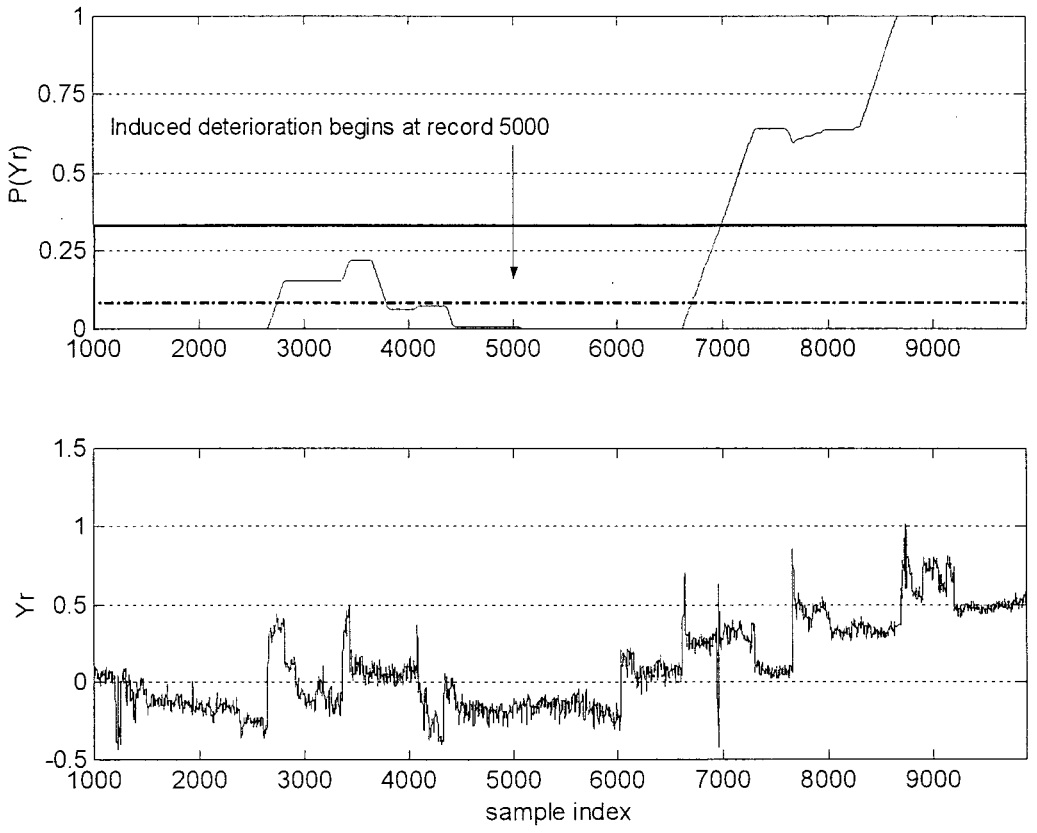


Figure 36 **Top sub-plot: Probability of simulation error, $P(Y_r)$, for artificially induced failure in terms of blow-by gas flow. Solid horizontal line is failure threshold and broken horizontal line is warning threshold. Bottom sub-plot: Simulation error, Y_r.**

On disassembly of the engine, it was found that significant piston ring wear had occurred, thus confirming the diagnosis. This result suggested that interpreting residual probabilities in the warning region together with the corresponding residual variance should be implemented as part of the diagnostic method.

6.4. Conclusions

An online diagnostic method to detect systematic temporal deterioration of an internal combustion engine has been successfully demonstrated in this chapter. A sufficiently stationary data size was determined, using a technique proposed in Chapter 5. The acquisition of a stationary data set contributed to the construction of a reliable simulation model of blow-by gas flow of a compression ignition engine. A statistical evaluation of the simulation error was developed and clearly indicated when the engine entered a failure state, given a predetermined failure threshold probability. Outlier detection by way of convex hull technology slightly improved the model generalization, and would be even more beneficial to the simulation of data that contains a larger outlier content than the data used in this investigation. In addition, the outlier detection algorithm requires no operator input and is fast enough to be implemented online during engine operation. Failure diagnosis was first demonstrated on an artificially deteriorating set of observations and subsequently, on a data sample from the end of the test, when evidence of wear could be expected to appear. In the first demonstration the diagnostic algorithm clearly indicated entry into failure state. The results of the second demonstration warned of imminent excessive blow-by gas flow, suggesting significant piston ring wear. This was confirmed by physical inspection after completion of the test.

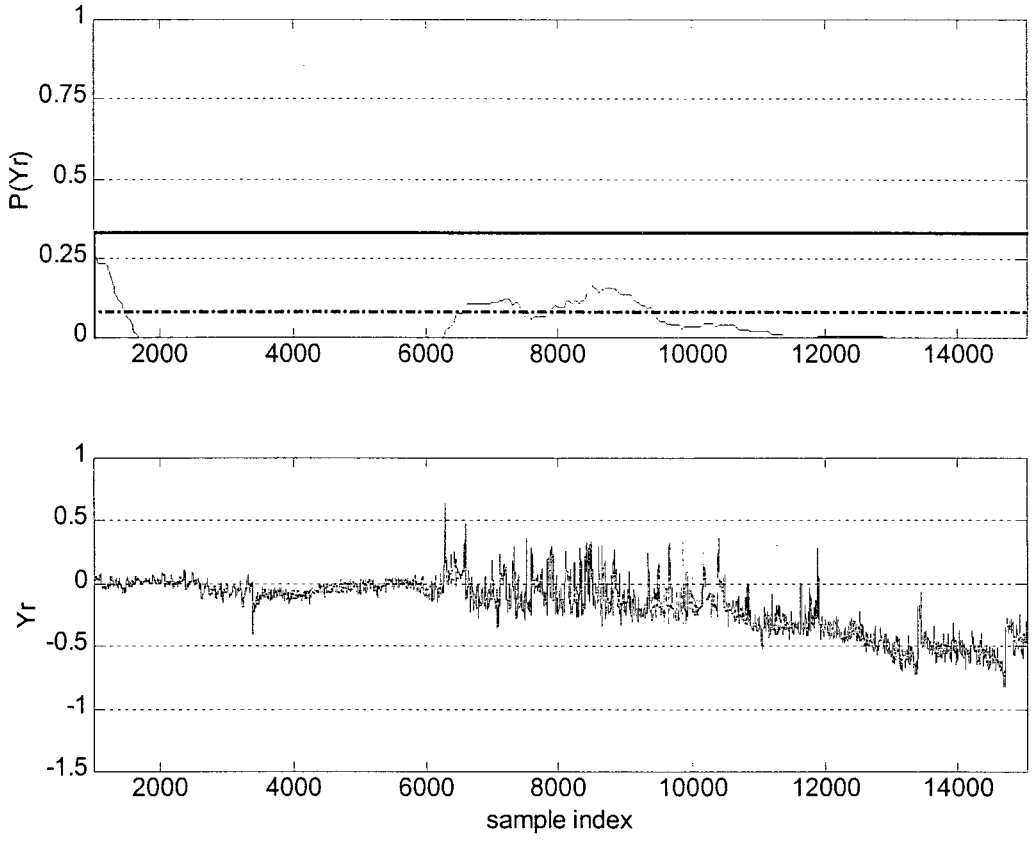


Figure 37 **Probability of simulation error (top), and simulation error (bottom), for blow-by gas flow data at 600 h with warning threshold at 0.0746 and failure threshold at 0.298. Note the sudden increase in residual variance between samples 7000 and 10000. (The sample indices are relative to the starting index of this data segment.)**

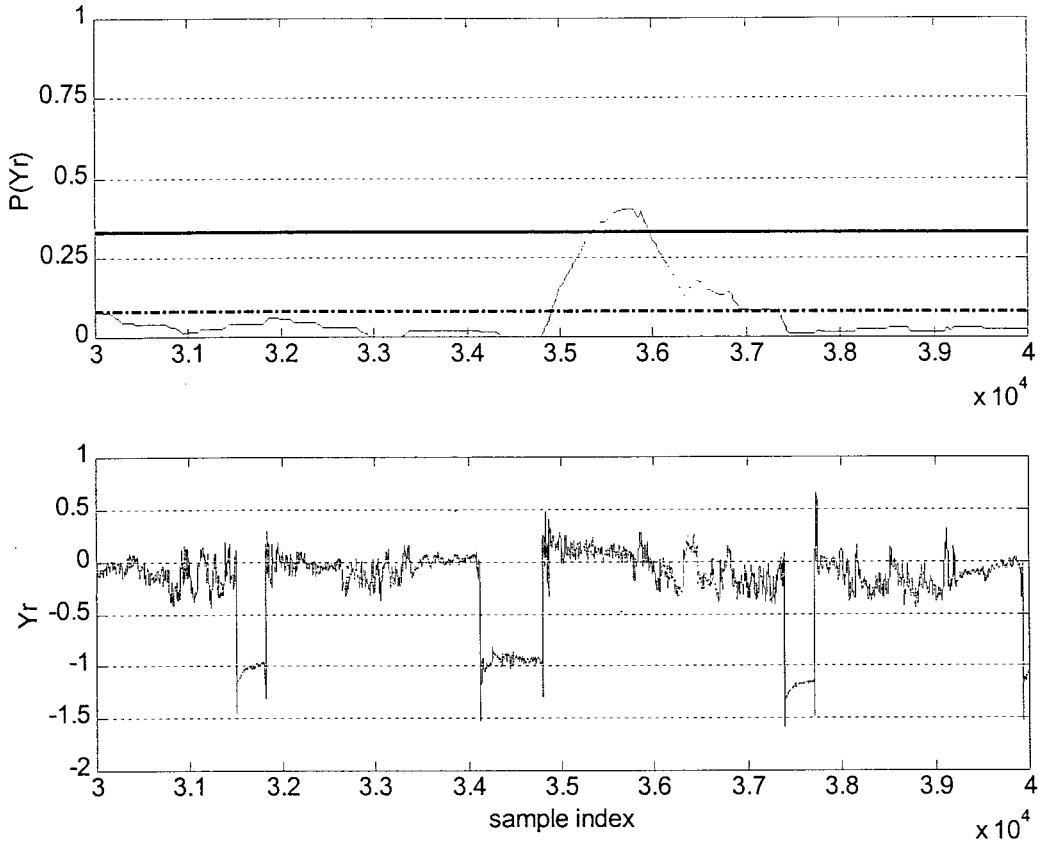


Figure 38 **Probability of simulation error (top), and simulation error (bottom), for blow-by gas flow data at 580 h with warning threshold at 0.0746 and failure threshold at 0.298. Note the brief visitation to the failure zone. (The sample indices are relative to the starting index of this data segment).**

7 CONCLUSIONS

In this dissertation I primarily endeavored to formalize and extend nonlinear system identification for the broad class of non-linear systems that can be parameterized as state space systems. It was shown in Chapter 2 that the established, but rather ad hoc methods of time series embedding and nonlinear modeling with MLP network and radial basis function model structures can be interpreted in context with the established linear system identification framework.

In Chapter 3 the methodological framework was formulated for the identification of non-linear state space systems from one-dimensional time series using a surrogate data method. In this chapter it was clearly demonstrated that validation of dynamic models by one-step predictions is insufficient proof of model quality. In the particular case study, a chaotic autocatalytic process, the R^2 statistic, generally used for measuring model fitness and performance during validation and cross-validation, were identical up to the third decimal (0.999) for two quite different model structures, a MLP and a RBF model respectively. On the other hand, free-run predictions used to generate non-linear surrogate data gave adamant proof of model quality, showing the superiority of the RBF model structure in this particular case. In addition, the classification of data as either dynamic or random was performed using the same surrogate data technique. Even when 10% measurement and dynamic noise were added to the original autocatalytic data, the classification technique still clearly distinguished the data from non-linearly transformed random data.

Chapter 4 the formulation of a nearly real-time algorithm for detection and removal of radial outliers in multidimensional data was pursued. A convex hull technique was proposed and demonstrated on random data as well as real test data recorded from an internal combustion engine. The results showed the convex hull technique to be effective at a computational cost two orders of magnitude lower than the more proficient Rocke and Woodruff technique. On the downside, the convex hull technique falsely indicated five outliers out of 533 observations (0.9%) when tested on an artificially generated random data set, while the Rocke and Woodruff technique identified no false outliers. This was a slight cost against the benefit of a simple and fast outlier detection algorithm.

In Chapter 5 the methodological framework was expanded for system identification as formulated in Chapter 3, to accommodate the identification of nonlinear state space systems

from multivariate time series. Specifically system parameterization was accomplished by combining individual embeddings of each variable in the multivariate time series, and then separating this combined space into independent components. This method of parameterization was successfully applied in the simulation of the autocatalytic process that was introduced in Chapter 3. In addition, the parameterization method was implemented in the one-step prediction of atmospheric NO_2 concentration, which would potentially become part of an envisaged environmental control system for Cape Town. Furthermore, the combination of the embedding strategy and separation by independent component analysis was able to isolate some of the noise components from the embedded data.

Chapter 6 aimed to implement the foregoing system identification methodology in the online diagnosis of temporal trends in critical system states. The methodology established in the previous chapters was supplemented by the formulation of a statistical likelihood criterion for simultaneous interpretation of multivariate system states. This technology was successfully applied to the diagnosis of the temporal deterioration of the pistons rings in an compression ignition engine under test conditions. The diagnostic results indicated the beginning of significant piston ring wear, which was confirmed by physical inspection of the engine after conclusion of the test. The technology will be further developed and commercialized.

Future activities resulting from this research include the following:

a) Expand the pseudo-linear radial basis function algorithm to accommodate multivariate embedding strategies.

b) Implement a Minimum Description Length algorithm to optimize MLP network model structures.

c) Improve the outlier detection method as outlined in section 4.4.

d) Develop an algorithm for noise-reduction based on results from Chapter 5 regarding ICA.

e) Further develop the diagnostic algorithm into prototype form and commercialize for implementation in industrial automotive and stationary engines. Tests that impose more complex transient behaviour over a broader operating range on an engine, will be conducted. The diagnostic technique will be applied to these data to refine the calibration of warning and failure thresholds.

In conclusion, my aspiration for knowledge and understanding of Nature has been greatly enhanced by the research that went into this dissertation, even beyond the boundaries of Engineering. I wish to speculate that our mental model of Nature might eventually shift from symbolic fundamental models to purely mathematical parameterizations - currently called "black box" models. How one will design systems this way is open to conjecture, since our tradition is to characterize and express properties of materials and systems in terms related to our existing symbolic models.

The truth is out there, and we endeavour to converge on it.

REFERENCES

Abarbanel, H. D. I., Gills, Z., Liu, C. and Roy, R. 1996. *Physical Review A*, **53,** 440.

Abarbanel, H. D. I.. 1994. *Nonlinearity and Chaos in Engineering Dynamics*, Wiley and Sons.

Abarbanel, H. D. I.. 1996. *Analysis of Observed Chaotic Data*, Springer-Verlag (New York).

Atkinson, C. M., Long, T. W. and Hanzevack, E. L. 1998. Virtual sensing: a neural network-based intelligent performance and emissions prediction system for on-board diagnostics and engine control. *SAE International Congress and Exposition* (Detroit).

Barber, C. B., Dobkin D.P., and Huhdanpaa, H.T. Dec. 1996. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, **22**(4), 469-483.

Barnard, J. P., Aldrich, C. A., and Gerber M. 1999a. Identification of dynamic process systems with surrogate data methods. Submitted to *American Institute of Chemical Engineering*.

Barnard, J. P., Aldrich, C. and Gerber, M. 1999b. Detecting outliers in large multivariate process data by using convex hulls. Submitted to *Technometrics*.

Broomhead, D. S. and Lowe, D. 1988. Multivariate functional interpolation and adaptive networks. *Complex Systems*, **2**, 321-355.

Bryant, P., Brown, R. and Abarbanel, H. D. I. 1991. Computing the Lyapunov spectrum of a dynamical system from observed time series. *Physical Review A*, **43**, 2787-2806.

Cao, L., Mees, A. and Judd, K. 1998. Dynamics from multivariate time series. *Physica D*, **121**, 75-88.

Chen, C. F., Cowan, N. and Grant, P. M. 1991. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, **2**(2), 303-309.

Comon, P. 1994. Independent Component Analysis - a new concept? *Signal Processing*, **36**, 287-314.

Dzubay, T. G. 1982. Visibility and Aerosol Composition in Houston, Texas. *Environmental Science and Technology,* **16**, pp514-525.

Eckmann, J. P. and Ruelle, D. 1992. Fundamental limitations for estimating dimensions and Lyapunov exponents in dynamical systems. *Physica D*, **56**.

Eykhoff, P. 1974. *System identification : parameter and state estimation*, Wiley, Chichester.

Farmer, J. D. and Sidorowich, J. J. 1987. Predicting chaotic time series. *Physical Review Letters*, **34**, 845.

Frazer, A. M., Swinney, H. L. 1986. Independent coordinates for strange attractors, *Physical Review Letters A*, **33**, 1134-1140.

Frazer, A.M. 1989. Reconstructing attractors form scalar time series: a comparison of singular system and redundancy criteria. *Physica D*, **34**, 391-404.

Frazer, A.M. and Swinney, H.L. 1986. Independent coordinates for strange attractors. *Physical Review Letters A,* **33**, 1134-1140.

Friedman, J. H. and Tukey, J. W. 1974. A projection pursuit algorithm for exploratory data nalysis. *Transactions on Computers*, **23C**(9), 881-889.

Funahashi, K. 1989. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, **2**, 183-192.

Gnanadesikan, R. 1977. Methods for statistical data analysis of multivariate observations, Wiley and Sons, 292-317.

Goutte, C. 1997. Note on free lunches and cross-validation. *Neural Computation*, **9**, 1245-1249.

Grassberger, P. and Procaccia, I. 1983. Characterization of strange attractors. *Physical Review Letters,* **50**, 346-349.

Grassberger, P., Hegger, R., Kantz., H., Schaffrath, C., and Schreiber, T. 1993. On noise reduction methods for chaotic data, *Chaos,* **3**.

Gray, P. and Scott, S. K. 1984. Autocatalytic reactions in the isothermal, continuous stirred tank reactor. Oscillations and instabilities in the system A + 2B $\rightarrow$ 3B, B $\rightarrow$ C. *Chemical Engineering Science*, **39**, 1087-1097.

Gray, P. and Scott, S. K.. 1983. Autocatalytic reactions in the isothermal, continuous stirred tank reactor. Isolas and other forms of multi-stability, *Chemical Engineering Science*, **38**, 29-43.

Grimaldi, C. N. and Mariani, F. 1997. On board diagnosis of internal combustion engines: a new model definition and experimental validation. SAE-No. 970211, Detroit: *SAE 1997 Transactions: Journal of Commercial Vehicles*, **2**, 21-29.

Grobliki, P. J., Wolff, G. T., and Countess, R., J. 1981. Visibility reducing species in the Denver "Brown Cloud". *Atmospheric Environment*, **15**(12), Pergammon, 2473-2484.

Hagan, M. T. and Menhaj, M. 1994. Training feed-forward networks with the Marquardt algorithm. IEEE Transactions on Neural Networks, **5**, no. 6, 989-993.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P.J. and Stahel, W.A. 1986. *Robust Statistics: The approach based on Influenced Functions*, John Wiley (New York).

Hardy, R. L. 1971. Multiquadratic equations of topography and other irregular surfaces. *Journal of Geophysics Research*, **76**,1905-1915.

Hassounah, M. I. and Miller, E. J. 1994. Modelling air pollution from road traffic: a review. *Traffic Engineering and Control*.

Hawkins, D. M. 1980. *Identification of outliers*, Chapman and Hall (London), 115-122.

Hinich, M. J. 1982. Testing for Gaussianity and linearity in a stationary time series. *Journal of Time Series Analysis*, **3**, 169-176.

Hyvärinen, A. 1999. Independent component analysis by minimization of Mutual Information. *IEEE Transactions on Neural Networks*. [submitted].

Judd, K. 1992. An improved estimator of dimension and some comments on providing confidence intervals. *Physica D*, **56**, 216-228.

Judd, K. 1994. Estimating dimension from small samples. *Physica D*, **71**, 421-429.

Judd, K. and Mees, A. 1995. On selecting models for non-linear time series. *Physica D*, **82**, 426-444.

Kennel, M. B. 1997. Statistical test for dynamical nonstationarity in observed time-series data. *Physical Review E*, **56**(1), 316-321.

Kennel, M. B., Brown, R. and Abarbanel, H. D. I. 1992. Determining minimum embedding dimension using a geometrical construction. *Physica Review A,* **45**, 3403-3411.

Kostelich, E. J. and Yorke, J. A. 1988. Noise reduction in dynamical systems. *Physical Review A,* **38**, 1649.

Lai, Y-C. and Lerner, D. 1998. Effective scaling region for computing the correlation dimension from chaotic time series. *Physica D*, **115**, 1-18.

Lawrence, W., Lee Giles, C. and Ah Chung Tsoi. 1996, What size neural network gives optimal generalization? Convergence properties of backpropagation. *Technical Report* UMIACS-TR-96-22 and CS-TR-3617, Institute for Advanced Computer Studies, University of Merryland.

Levenberg, K. 1944. A method for the solution of certain nonlinear problems in least squares. *Quart. Applied Mathematics*, **2**, 164-168.

Ljung, L. 1987. *System Identification*, Prentice-Hall (New Jersey).

Lynch, D. T. 1992. Chaotic behavior of reaction systems: parallel cubic autocatalators. *Chemical Engineering Science*, **47**(2), 347-355.

MacMurray, J. C. and Himmelblau, D. M. 1995. Modeling and control of a packed distillation column using artificial neural networks. *Computers and Chemical Engineering*, **19**(10), 1077-1088.

Marquardt, D. W. 1963. An algorithm for least squares estimation of nonlinear parameters. *Journal of SAIM*, **11**, 431-441.

Mees, A. I., and Judd, K. 1993. Dangers of geometric filtering. *Physica D,* **68**, 427-436.

Nikias, C. L. and Petropulu, A. 1993. *Higher-Order Spectra Analysis: A Nonlinear Signal Processing Framework*, Prentice-Hall (New Jersey).

Norton, J. P. 1986. *An Introduction to Identification*, Academic Press (London).

Ogata, K. 1995. *Discrete-Time Control Systems*, Prentice-Hall (New Jersey), 293-302.

O'Rourke, J. 1994. *Computational Geometry in C*, Cambridge University Press.

Osborne, A. R. and Provenzale, A. 1989. Finite correlation dimension for stochastic systems with power-law spectra. *Physica D*, **35**, 357-381.

Pajunen, P. 1998. *Procedure of Independence and Artificial Neural Networks Workshop*, 26.

Pajunen, P., Hyvärinen, A. and Karhunen, J. 1996. *1996 International Conference on Neural Information Processing (ICONIP'96)*, Hong Kong, 1207.

Parlitz, U. 1992. Identification of true and spurious Lyapunov exponents from time series. *International Journal of Bifurcation and Chaos*, **2**, 155-165.

Pham, D. T., Liu, X. and Oh, S. J. 1995. Dynamic system identification using Elman and Jordan neural networks. *Neural Networks for Chemical Engineers*, (ed. Bulsari, A.B.), Elsevier, (Netherlands), 573-591.

Powell, M. J. D. 1987. Radial basis functions for multivariate interpolation: a review. *Algorithms for Approximations*, Oxford, 143-167.

Rauf, F. 1997. *Non-linear Adaptive Filtering: A Unified Approach.* Ph.D. Thesis, Boston University (Boston).

Rauf, F. and Ahmed, H. 1997. New non-linear adaptive filters with applications to chaos. *International Journal of Bifurcation and Chaos*, **7**(8), 1791-1809.

Rissanen, J. 1989. Stochastic complexity in statistical inquiry. *World Scientific*, Singapore.

Rivals, I. and Personnaz, L. 1999. On cross validation for model selection. *Neural Computation*, **11**, 863-870.

Rock, D. M. and Woodruff, D. L. 1996. Identification of Outliers in Multivariate Data. *Journal of the American Statistical Association*, **47**(435), 27-42.

Rousseeuw, P. J. and Leroy, A. M. 1987. *Robust Regression and Outlier Detection*, John Wiley (New York).

Rumelhart, D. E., Widrow, B., and Lehr, M. 1994. The basic ideas in neural networks. *Communications of the ACM*, **37,** No 3, 87-91.

Sauer, T., Yorke, J.A., Casdagli, M. 1991. Embedology. *Journal of Statistical Physics*, **65**, 579-616.

Schreiber, T. and Schmitz, A. 1996. Improved surrogate data for non-linearity tests. *Physical Review Letters*, 635-638.

Small, M. and Judd, K. 1998a. Comparison of new non-linear modelling techniques with applications to infant respiration. *Physica D*, **117**, 283-298.

Small, M. and Judd, K. 1998b. Correlation dimension: a pivotal statistic for non-constrained realizations of composite hypotheses in surrogate data analysis. *Physica D*, **129**, 386-400.

Stefanovska, A., Strle, S. and Kroselj, P. 1997. On the overestimation of the correlation dimension. *Physical Letters A*, **235**, 24-30.

Su, H-T., McAvoy, T.J. and Werbos, P. 1992. Long term predictions of chemical processes using recurrent neural networks: A parallel training approach. *Industrial and Engineering Chemistry Research*, **31**, 1338-1352.

Subba Rao, T., Gabr, M. 1993. *An Introduction to Bispectral Analysis and Bilinear Time-Series Models*, Springer-Verlag (New York).

Takens, F. 1981. Detecting strange attractors in turbulence. *Lecture Notes in Mathematics*, **898**, Springer (Berlin), 366-381.

Takens, F. 1993. Detecting non-linearities in stationary time series. *International Journal of Bifurcations and Chaos*, **3**, 241-256.

Theiler, J. 1995. On the evidence for low-dimensional chaos in an epileptic encephalogram. *Physics Letters A*, **196**, 335-341.

Theiler, J. and Pritchard, D. 1996. Constrained realization Monte Carlo method for hypothesis testing, *Physica A*, **94**, 221-235.

Theiler, J. and Rapp, P.E. 1996. Re-examination of the evidence for low-dimensional non-linear structure in the human electroencephalogram. *Encephalography and Clinical Neurophysiology*, **98**, 213-222.

Theiler, J., Eubank, E., Longtin, A. and Galdrikian, B. 1992. Testing for non-linearity in time series: The method of surrogate data. *Physica*, **58D**, 77-94.

Thompson, J. M. T. and Bishop, S. R. 1995. *Nonlinearity and Chaos in Engineering Dynamics*, Wiley and Sons Ltd. (Chichester), 1-10.

Turner, D. B. 1994. *Atmospheric Dispersion Estimates – an introduction to dispersion modelling*, Lewis Publishers.

Wornell, G. W. 1996. *Signal Processing with fractals: A wavelet-based approach*. Prentice-Hall, Inc.

Zhu, H. and Rohwer, R. 1996. No free lunch for cross-validation. *Neural Computation*, **8**, 1421-1426.

TERMINOLOGY AND DEFINITION OF PARAMETERS

Here follows a table of symbols and expressions defined in the text of the thesis.

Simbol	Definition
AAFT	Amplitude adjusted Fourier transform
AMI	Average Mutual Information
AXMI	Average Cross Mutual Information
ICA	Independent Component Analysis
FNN	False Nearest Neighbours
$\mathcal{M}^*$	Model set
$\mathcal{M}$	Model structure
$\mathcal{M}_{FF}$	Multi-layer perceptron model structure
$\mathcal{M}_{RB}$	Radial basis function model structure
$\mathcal{M}_{PL}$	Pseudo-linear radial basis function model structure
$\mathcal{M}(\theta)$, $\mathcal{M}_\theta$	A model in terms of parameter vector θ
PCA	Principal Component Analysis
$\Re$	The set of real numbers
θ	Model parameter vector
μ	mean value
σ	standard deviation
MDL	Rissanen's Minimum Description Length
MLP	Multi-layer perception
$N(\cdot)$	Normal distribution
$P_\Im(\cdot)$	Threshold probability

RBF	Radial basis function
S	Space of independent components, from ICA
SSE	Sum-squared-error
W	Separating matrix from ICA
X	An embedding
Z	Data set consisting of input and output spaces
V	Norm
m	Embedding dimension
k	Embedding lag
n	Number of data records
r, r_t	Model simulation or prediction error
t	Time, [s]
$y(t), y_t$	Scalar system output
$\hat{y}(t), \hat{y}_t$	Output of model that approximates system output function

APPENDIX

A.1 Average Mutual Information (AMI)

Given the delay vector $\mathbf{x}_i = [y_{i+(m-1)\tau}, y_{i+(m-2)\tau}, y_{i+(m-3)\tau}, \ldots, y_i]$ for the time series $\mathbf{y} = [y_1, y_2, y_3, \ldots y_n]$ where m is the embedding dimension and n the size of the time series, an accurate reconstruction of system invariants like the attractor and Lyapunov coefficients are ensured by an embedding dimension of $m > 2d_A$, where d_A is the topological or fractal dimension of the attractor (Takens, 1981). Time lag, τ, is not specified in Taken's theorem, because it is based upon the assumption of unlimited noise-free data.

In the presence of noise the time lag becomes important. Too small a time lag or window results in to little information extracted from the data and an attractor contracted into the main diagonal line in state space. This phenomenon is called *redundance* (Sauer et al., 1991) and is based upon the relative effect of measurement error to the difference between data in a delay coordinate vector. The effect levels out because the attractor is finite in size, therefore likewise for the difference between data.

On the other hand, too large a time lag results in data without correlation to be included which causes an attractor with a false complex shape. This phenomenon is called *irrelevance* (Sauer et al., 1991). Underlying this phenomenon is the local divergence in time between adjacent trajectories of the state vector, which causes a dynamic error with increasing time lag. This error levels out due to the finite shape of the attractor.

The dominant method for calculating the time lag is the method of Mutual Information (Frazer and Swinney, 1986). The time lag is fixed heuristically at the point of the first minimum mutual information for the time series. This method estimate the probability to find a measurement again given that the same measurement has been already been made. This statistic is calculated among all elements of the timeseries.

Information theory is applied in the method of Mutual Information. Since strange attractors are ergodic with an asymptotic probability distribution, the required probabilities do exist and information theory is applicable. Measurements can be regarded as signals. Let S be a system of possible messages, $s_1, s_2, s_3, \ldots, s_n$ associated with the probabilities

$P_s(s_1), P_s(s_2), P_s(s_3), \ldots, P_s(s_N)$, with subscript s denoting a particular system. The average information gained from a measurement is indicated by the information entropy of the system,

$$H(S) = -\sum_i P_s(s_i)\log(P_s(s_i)) \tag{41}$$

The dependence of $x(t+\tau)$ on $x(t)$ can be expressed as the uncertainty of finding $x(t+\tau)$, given $x(t)$.

$$\begin{aligned} [s,q] &= [x(t), x(t+\tau)] \\ H(Q|s_i) &= -\sum_i P_{q|s}(q|s_i)\log\left[P_{q|s}(q|s_i)\right] \end{aligned} \tag{42}$$

The average uncertainty of finding $x(t+\tau)$ given $x(t)$ is calculated by averaging $H(Q|s_i)$ over s_i.

$$\begin{aligned} H(Q|S) &= -\sum_i P_s(s_i)H(Q|s_i) \\ &= -\sum_i P_{sq}(s_i,q_i)\log\left[\frac{P_{sq}(s_i,q_i)}{P_s(s_i)}\right] \\ &= H(S,Q) - H(S) \end{aligned} \tag{43}$$

The reduction in uncertainty about $x(t+\tau)$ gained from the measurement of $x(t)$ is the mutual information, which is

$$\begin{aligned} I(Q,S) &= H(Q) - H(Q|S) \\ &= I(S,Q) \\ &= \sum_i P_{sq}(s_i,q_i)\log\left[\frac{P_{sq}(s_i,q_i)}{P_s(s_i)P_s(q_i)}\right] \end{aligned} \tag{44}$$

The joint probability distribution, P_{sq}, is estimated by the joint histogram of s and q. In this dissertation the symbol k, denotes embedding lag as an unit of the sampling period of the system observer.

A.2 Average Cross Mutual Information (AXMI)

Average cross-mutual information is related to AMI defined in section A.1. A formal algorithm to calculate this statistic was first proposed by Fraser and Swinney (1986) for

determining embedding lag, used in time series embedding (Frazer 1989, Abarbanel 1995). The AXMI between the observation $x(t)$, at time t, and the observation y(t-k), at time (t-k), is

$$I_{xy}(k) = \sum P[x(t), y(t-k)] \log_2 \left\{ \frac{P[x(t), y(t-k)]}{P[x(t)]P[y(t-k)]} \right\} \tag{45}$$

where $P[\cdot]$ is the probability function in terms of x or y, $P[\cdot,\cdot]$, the joint probability function in x and y, t the time, and k some lag (multiples of sampling period, τ). Let R_{xy} be AXMI normalized with the average auto mutual information of the selected dependent state, $I_{xx}(0)$ as reference:

$$R_{xy} = I_{xy} / I_{xx} \tag{46}$$

For k=0, $R_{xy}(0)$ is a non-linear equivalent to linear cross-correlation.

A.3 False Nearest Neighbours and False Nearest Strands

According to Takens (1981) an attractor will be completely unfolded in an m-dimensional space given that $m > 2d_A$ where d_A is the topological or fractal dimension of the attractor. This requirement is only necessary, so a unambiguous technique is required to establish the minimum embedding dimension.

The false neighbours algorithm was developed for this purpose by Kennel et al., (1992). While unfolding the attractor in space of increasing dimension, the points that are true neighbours can be progressively distinguished until, after reaching the optimal embedding dimension, no more additional false neighbours are discovered.

False neighbours appear only because one views the attractor in space of too small a dimension, thereby mistaking two points for being neighbours. The nearness is expressed as the Euclidean distance between two points. A neighbour is classified in terms of the Euclidean distance being within a preset limit.

The technique will fail on data with a high noise content and fails ultimately on white noise. Alternatively, the False Nearest Strands technique (Kennel, 1995) is better suited for:

a) time series resulting from oversampling

b) using small time delays

c) sparsely populated regions of attractors

Oversampling can be heuristically identified by a mutual information time lag in excess of $\tau = 10\tau_s$, in which case the data may be oversampled.

Strand pairs are defined when temporally corresponding points or temporal iterates lie on two adjacent trajectories as nearest neighbours.

A.4 Lyapunov exponents

In chaotic (non-linear) systems two states that are nearly identical diverge from each other at an exponential rate which causes a sensitive dependence on initial conditions. The Lyapunov exponent characterizes this divergence. Let λ be the Lyapunov exponent, then the distance between adjacent trajectories after some time *t* will be

$$d(t) = d_0 2^{\lambda t} \tag{47}$$

When averaging this local divergence along the trajectory, one gets

$$\begin{aligned} d(t_1) &= d_1 \\ &\approx d_0 2^{\lambda(t_1 - t_0)} \end{aligned} \tag{48}$$

An overall Lyapunov exponent can be defined as:

$$\lambda = \frac{1}{t_N - t_0} \sum_{k=1}^{N} \log_2 \frac{d(t_k)}{d_0(t_{k-1})} \tag{49}$$

The calculation of Lyapunov exponents from data only can be treacherous and should be approached with circumspection (Brown et al., 1991; Parlitz, 1992; Abarbanel, 1996). A fairly reliable method to calculate Lyapunov exponents has been proposed by Brown et al. (1991) and was used in this research.

A.5 Model Fitness Test

A data model is normally tested for fitness according to residual correlation with the observed dependent state. However, the autocorrelation and cross correlation sequences cannot give any evidence of remaining nonlinear relationships, since any process can always be considered to be a linear process with respect to its second-order statistics. Higher-order cumulants can give such evidence. For example, the third-order cumulant can be used to test whether the simulation error is Gaussian. Hinich (1982) proposed a zero-skewness test as a quantitative test for normality of a stationary data sample. More specifically one tests the null

hypothesis that the estimation of bicoherence is zero at a calculated significance level. The bicoherence (or normalized bispectrum) can be estimated by the direct, Fast Fourier Transform-based algorithm (Subba Rao and Gabr, 1993; Nikias and Petropulu, 1993) and is defined in terms of the bispectrum, $B_{3y}(\bullet)$ and the power spectrum, $P_{yy}(\cdot)$,

$$B_{cy}(\omega_1,\omega_2)=\frac{B_{3y}(\omega_1,\omega_2)}{[P_{yy}(\omega_1)P_{yy}(\omega_2)P_{yy}(\omega_1+\omega_2)]^{1/2}} \tag{50}$$

The hypothesis test is based on the mean bicoherence power,

$$P_{bc}=\sum\left|B_{cy}(\omega_1,\omega_2)\right|^2 \tag{51}$$

The statistic, P_{bc}, is χ^2-distributed, with p degrees of freedom, which is a function of the Fast Fourier Transform length and a resolution parameter, c. For details refer to (Hinich, 1982). Finally, the probability that a χ^2 random variable with p degrees of freedom could exceed the value of P_{bc} is calculated. A high probability indicates that the null hypothesis should be accepted, that is the data sample has a normal distribution. If the residue scores a high probability in the test, it can be accepted as Gaussian and therefore the model is a proper representation of the data.

A test for the linear correlation between the model and the observation is the discriminating linear statistic, R^2, defined as:

$$R^2=1-\frac{1}{(n\text{-}1)\sigma_y^2}\sum(y-\hat{y})^2 \tag{52}$$

where y is the observed state, $\hat{y}$ the simulated state, σ the standard deviation of y and n the length of y. An arbitrary high value for R^2 (e.g. $R^2 > 0.90$) would indicate predicted and observed output that is sufficiently linearly correlated.

A.6 Stationarity Test

The proposed test for stationarity of a data set is as follows: suppose one would start with a time-series of size N, divide it into two halves (called half samples), Y_1 and Y_2, bin each half sample and compare the contents of each bin. Under the hypothesis that the half samples are mutually stationary, the joint probability to find data from each half sample in a bin of a given category should stay constant or vary only randomly, for increasing N. This would imply that

the topology of the joint probability matrix, $P(Y_1 \cap Y_2)$, converges as N increases towards the stationary size. The topology can be characterized in terms of a discriminating statistic, e.g. the center of mass C_m of the joint probability matrix. The value of N at which $\Delta C_m[P(Y_1,Y_2)]/\Delta N$ over the past N_w iterations is sufficiently small, will indicate the stationary size.

A.7 Surrogate Data

The method of surrogate data (Takens, 1993; Theiler and Pritchard, 1996; Theiler and Rapp, 1996) involves a null hypothesis against which the data are tested, as well as a discriminating statistic. The data are first assumed to belong to a specific class of dynamic processes. Surrogate data are subsequently generated, based upon the given data set, by using the assumed process. An appropriate discriminating statistic is calculated for both the surrogate and the original data (Theiler et al., 1992). If the calculated statistics of the surrogate and the original data are significantly different, then the null hypothesis that the process that has generated the original data is of the same class as the system that has generated the surrogate data, is rejected. By means of a trial-and-error elimination procedure, it is then possible to get a good idea of the characteristics of the original process.

More specifically, let $\mathbf{x} \in \Re^N$ be a time series consisting of N observations, ψ a specific hypothesis, $\Im_\psi$ the set of process systems consistent with the hypothesis, and $T:\Re^N \rightarrow U$ be a statistic that will be used to evaluate the hypothesis ψ that $\mathbf{x}$ was generated by some process $\Im \in \Im_\psi$. Generally the statistic $U \subset \Re$ and it will be possible to discriminate between the original data $\mathbf{x}$ and the surrogate data $\mathbf{x}_s$ consistent with the hypothesis given by the probability density of T, given $\Im$, i.e. $p_{T,\Im}(t)$.

A.7.1. Classes of hypotheses

Three classes of hypotheses are widely used. These are equivalent to the assumption that the data are identically, independently distributed noise (type 0), linearly filtered noise (type 1) and a monotonic non-linear transformation of linearly filtered noise (type 2).

Type 2 surrogates are also known as amplitude adjusted Fourier transform (AAFT) surrogates (Small and Judd, 1998). The procedure for generating type 2 surrogate data consists of the following steps:

i) Generation of a normally distributed data set **y**, reordered to have the same rank distribution as **x**, the observed (original) data set.

ii) Generation of a type 1 surrogate data set $\mathbf{y}_s$ from **y** (by phase-shuffling the Fourier transform of **y**).

iii) Finally, rank order $\mathbf{y}_s$ and replacing the amplitudes y_{sj} with that of x_i of corresponding rank.

A.7.2. Pivotal test statistics

Theiler (1995) has suggested that a distinction can be made between so-called pivotal and non-pivotal statistics. A test statistic T is considered to be pivotal, if the probability distribution $p_{T,\Im}$ is the same for all processes $\Im$ consistent with the hypothesis ψ, thus $p_{T,\Im}$ is invariant for all $\Im \in \Im_\psi$. Moreover, a distinction can be made between simple and composite hypotheses. If the set of all processes consistent with the hypothesis ($\Im_\psi$) is a singleton, then the hypothesis is simple. Otherwise, the hypothesis is composite and can be used not only to generate surrogate data consistent with a particular process $\Im$, but also to estimate $\Im \in \Im_\psi$. In fact, $\Im$ has to be specified when the hypothesis is composite, unless T is a pivotal statistic (Theiler, 1995).

Constrained realization (Schreiber and Schmitz, 1996) schemes can be employed when non-pivotal statistics are applied to composite hypotheses. That is, apart from generating surrogate data that represent typical realizations of a model of the system, the surrogate data should also be representative of a process yielding identical estimates of the parameters of the process when compared to the estimates of the process parameters obtained from the original data. Put in a different way, if $\Im_{est} \in \Im_\psi$ is the process estimated from the original data **x**, and $\mathbf{x}_s$ is a surrogate data set generated by $\Im' \in \Im_\psi$, then $\mathbf{x}_s$ is a constrained realization of $\Im_{est} \in \Im'$.

As an example, if ψ is the hypothesis that **x** is generated by linearly filtered independent identically distributed noise, then *non-constrained* surrogate data $\mathbf{x}_s$' can be generated from a Monte Carlo simulation based on the best linear model estimated from **x**. The data $\mathbf{x}_s$' can be constrained by shuffling the phases of the Fourier transform of the data, producing a set of random data $\mathbf{x}_s$" with the same power spectra (and autocorrelation) as the original data **x**. The autocorrelation, rank order statistics, non-linear prediction error, etc., would all be non-pivotal

test statistics characterizing dynamic manifold structures, since the distributions of these statistics would all depend on the form of the noise source and the type of linear filter. In contrast, the Lyapunov exponents and the correlation dimension (fractal dimension) would be pivotal test statistics, since the probability distributions of these quantities would be the same for all processes, regardless of the source of the noise or the estimated model. Since recent investigations have shown that Lyapunov exponents can be misleading in the presence of noise, the correlation dimension has gained favor as the pivotal statistic of choice.

A.7.3. Correlation dimension

The correlation dimension, d_c, is defined as follows.

$$d_c = \lim_{\varepsilon \to 0} \lim_{N \to \infty} \frac{\log C_N}{\log \varepsilon} \tag{53}$$

where C_N is the correlation function, defined by:

$$C_N(\varepsilon) = \binom{N}{2}^{-1} \sum_{0 \le i < j \le N} I\left(\left\| \upsilon_i - \upsilon_j \right\| < \varepsilon\right) \tag{54}$$

I(·) is a Heavyside function that returns one if the distance between point i and j is within ε, and zero otherwise, while N is the number of observations in the data set.

Reliable calculation of the correlation dimension is not as straightforward as first thought when the Grassberger-Procaccia algorithm appeared (Grassberger and Procaccia, 1983). Using this algorithm requires a linear scaling region to reliably calculate the correlation dimension. Noise strongly influences the approximation of the correlation dimension, according to Stefanovska et al. (1997). When working with measured (empirical) data, they stressed in particular the problem of using the Grassberger-Procaccia algorithm to obtain an adequate scaling region for a valid approximation of the correlation dimension. Lai and Lerner (1998) showed that this region is sensitive to the choice of the embedding lag. Linear correlation in the data set misleads the algorithm to falsely show convergence to some low dimension, which could then be misinterpreted for inherent low-dimensional dynamics (Judd, 1994).

Earlier, Judd (1992) pointed out the deficiencies of the Grassberger-Procaccia algorithm and proposed a different algorithm for calculation of the correlation dimension. This algorithm

replaces the requirement for a linear scaling region by fitting a polynomial of the order of the topological dimension in that region. It expresses the correlation dimension for inter-point distances below a specific scale ε_0. Instead of comparing estimates of the correlation dimension, one rather compares the clustering of correlation dimension curves calculated by the Judd algorithm. This allows the correlation dimension to be used for examining the macro- and microscale of the reconstructed dynamic attractor. For large data sets it asymptotically approaches the value of the true correlation dimension as ε_0 goes to zero. Also, the algorithm is not easily confused by linear correlation in the data (Judd, 1994).

Judd proposed that the correlation dimension be estimated as a function of scale ε_0 using the following equation, valid for $\varepsilon < \varepsilon_0$:

$$C_N(\varepsilon) \approx \varepsilon^{dc} q(\varepsilon) \tag{55}$$

where q(·) is a polynomial of order of the topological dimension.

Finally, accurate calculation of the correlation dimension depends on the minimum length of a time series. Stefanovska et al. (1997) have shown that too few points in a neighbourhood leads to overestimation of the correlation dimension when using the Grassberger-Procaccia algorithm. The Judd algorithm is less sensitive to the number of observations by an order of magnitude, compared to the Grassberger-Procaccia algorithm. In practical terms, a data set of approximately 1000 observations is usually sufficient for the Judd algorithm.