UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

J1171

(2)

REPORT DOCUMENTATION PAGE

AD-A214 873

DTIC
ELECTE
OCT 29 1989
B D

1b. RESTRICTIVE MARKINGS

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release; distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
F49620-88-C-0093

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Computational Engineering, Inc.

6b. OFFICE SYMBOL (If applicable)

7a. NAME OF MONITORING ORGANIZATION
Air Force Office of Scientific Research

6c. ADDRESS (City, State and ZIP Code)
14504 Greenview Dr., Suite 500
Laurel, MD 20708

7b. ADDRESS (City, State and ZIP Code)
Bolling AFB, DC 20332-6448

8a. NAME OF FUNDING/SPONSORING ORGANIZATION
AFOSR, ~~Bolling AFB~~

8b. OFFICE SYMBOL (If applicable)
NM

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
F49620 88 C 0093

8c. ADDRESS (City, State and ZIP Code)
Bldg 410
Bolling AFB, DC 20332-6448

10. SOURCE OF FUNDING NOS

PROGRAM ELEMENT NO	PROJECT NO	TASK NO	WORK UNIT NO
	3005	A1	

11. TITLE (Include Security Classification)
System Identification and Filtering of Nonlinear Controlled Markov Processes by Canonical Variate Analysis

12. PERSONAL AUTHOR(S)
Wallace E. Larimore

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Yr., Mo., Day)	15. PAGE COUNT
Final	FROM 8/88 TO 5/89	89 OCT 30	95

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB. GR.	

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

In this Phase I SBIR study, new methods are developed for the system identification and stochastic filtering of nonlinear controlled Markov processes. Currently available methods are restricted to very special forms or provide poor approximations to optimal procedures. The feasibility of using state space Markov process models and canonical variate analysis (CVA) for obtaining optimal nonlinear procedures for system identification and stochastic filtering is demonstrated. The theory of nonlinear CVA of Markov processes is developed in terms of a Hilbert space of nonlinear functions, and the multivariate nonlinear CVA is reduced to a sequential selection problem involving a univariate nonlinear CVA - the maximal correlation problem. The theory of maximal correlation, previously developed for

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS ☐

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Dr. Hajo Nachman

22b. TELEPHONE NUMBER (Include Area Code)
(202) 767-4938

22c. OFFICE SYMBOL
NM

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

19. ABSTRACT (continued)

Hilbert spaces of nonlinear functions, guarantees the existence of solutions to the multivariate CVA problem. A state space innovations representation for the Markov process is developed in terms of the canonical variable states. Extensions to the selection of a minimal rank state and interpretation of the canonical variables in terms of optimal normalizing transformations is developed. Computational algorithms are developed for determination of the canonical variable states, state space model fitting, and construction of nonlinear stochastic filters. The performance of the computational procedures are demonstrated on simulated data of the Lorenz chaotic attractor, a multiple equilibria nonlinear system, including process excitation noise. From observation of only one of the three states of the Lorenz attractor, the full dynamics of the system are determined. The filtered state estimate is accurate, and the identified nonlinear system has the same nonlinear character as the true process including chaos and multiple equilibria. These results considerably exceed the objectives of the Phase I study that involved only state affine processes, which cannot exhibit multiple equilibria and chaotic dynamics. Proposed Phase II follow-on research is discussed.

Final Report

SYSTEM IDENTIFICATION AND FILTERING C NONLINEAR CONTROLLED MARKOV PROCESS BY CANONICAL VARIATE ANALYSIS

October 27, 1989

COMPUTATIONAL ENGINEERING

CEI19-89-64/wl
J1171

Final Report

SYSTEM IDENTIFICATION AND FILTERING OF NONLINEAR CONTROLLED MARKOV PROCESSES BY CANONICAL VARIATE ANALYSIS

October 27, 1989

Prepared for

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
BOLLING AFB, DC

Under Contract No. F49620-88-C-0093

Prepared by

Wallace E. Larimore

COMPUTATIONAL ENGINEERING, INC.
14504 Greenview Drive, Suite 500
Laurel, MD 20708

FOREWORD

The research described in this report was performed by Computational Engineering, Inc. for the Air Force Office of Scientific Research under Contract No. F49620-88-C-0093, a Phase I Study funded under the Small Business Innovation Research Program. The Principal Investigator for the research was Dr. Wallace E. Larimore. Prof. John Baillieul of Boston University was a consultant to the study.

Contents

List of Figures

List of Tables

1 Overvie and Summary

The objective of this Phase I research is to demonstrate the feasibility of using state space Markov process models and canonical variate analysis (CVA) to obtain optimal nonlinear procedures for system identification and stochastic filtering. The emphasis in the study is to develop the major aspects of the theory and associated computational algorithms and to demonstrate the algorithms on simulated nonlinear data. Since this is an initial study in the feasibility of such methods, there are a number of further details and issues that need further study and are proposed for research in Phase II.

The results obtained in this study are very encouraging. The problems of system identification and stochastic filtering are of considerable importance in a number of aerospace systems as well as industrial processes. These problems are of long standing difficulty. The CVA approach is a new approach which is shown to generalize to very general nonlinear processes. The theory involves the use of operators on Hilbert spaces of nonlinear functions, and optimal solutions are shown to exist. The computational methods are a generalization of the singular value decomposition to Hilbert function spaces. The computations involve SVD type methods that are shown to work very well on the simulation data for the Lorenz chaotic attractor.

This section provides an overview and summary of the work. The topic is first introduced including a statement of the aspects of the problem that are of primary importance in this study. The technical objectives of this Phase I study are described, and the major results are briefly summarized. The activities of presenting results at technical conferences and submitting and publishing papers are discussed. Work proposed for Phase II is described, and finally the potential applications in DoD as well as in commercial processes are elaborated. In the rest of this report, the various topics are developed in sufficient detail to allow for the detailed evaluation of the feasibility of the CVA approach for system identification and stochastic filtering of nonlinear processes.

1.1 Introduction and Statement of the Problem

In recent years there has been considerable development of the field of nonlinear dynamical systems. Such systems include atmospheric dynamics, aerodynamic vehicle dynamics at high angles of attack, dynamics of rotating machinery and propellers, rotational dynamics of space craft, dynamics of robotic systems, and physiological dynamics of the heart and neurons of the brain (Marmarelis, 1989) to mention only a few. These recent developments have been primarily from the approach of deterministic nonlinear differential equations and simulation studies. The presently available methods for identification and filtering of stochastic nonlinear systems are very restrictive in the face of the rich class of nonlinear systems under current study using deterministic methods. A detailed review of the currently available methods for stochastic nonlinear systems is discussed in Section 2.

The problem addressed in this Phase I SBIR study is the development of system

identification and filtering method for stochastic nonlinear systems. A dominant feature of such nonlinear systems is that the system dynamics or filter of such a system are often infinite dimensional so that exact solution of the problem is not feasible. Even in the case where the process is finite state order, measurement noise results in an infinite dimensional filtering problem. Indeed, the problem is fundamentally an approximation problem, and useful solutions to the problem are basically efficient approximations to the infinite dimensional problem. Many of the previous approaches to the problem have been straight forward extensions of linear methods or nonlinear deterministic methods and have encountered formidable obstacles. Also, most of these previous approaches have dealt with the system state only indirectly or implicitly.

One of the fundamental aspects of many physical systems is the existence of a state which is a set of variables containing the memory of the system such as energy storing elements, i.e. position, velocity, rotational velocity, deflection of a structure, temperature of a fluid, etc. As will be seen, the fundamental difficulty is that a good approximation of the problem requires a good selection of an approximate state of the system. Given a good approximate state, the determination of a good identified model or stochastic filter becomes direct and simple. However, a poor choice of the state leads to poor results. The difficulty in previous approaches is the reliance on implicit methods that usually lead to open ended iteration involving nonlinear search using a prespecified model structure. Such a procedure usually has poor convergence properties with no known bound on the required computation. The procedure must be repeated for each candidate model structure, and for the nonlinear problem there are many more if not an infinite number of potential structures.

In the CVA approach, the approximation problem is formulated directly in terms of finding an optimal approximation of a specified state order. The states are to be determined as linear combinations of basis functions expressed as nonlinear functions of the past observations of the process. The states are computed in a CVA that determines the linear combinations of the nonlinear functions of the past that have predictive value for the future evolution of the process. The computation is fundamentally a linear procedure. Even in the Hilbert space of all possible nonlinear functions of the past, the solution to the CVA problem is shown to be the result of an iteration of projections on subspaces of nonlinear functions - a linear procedure for each iteration. It is shown that the optimal selection of states for a given order k, has the same first $k-1$ states as the solution for the case of finding only $k-1$ states. Thus the problem solution for state order k contains in canonical form the solution for all lower orders and the determination of the k-th state given the solution for the $k-1$ states can efficiently proceed in an iterative fashion. Since there is no *a priori* model structure imposed on the problem, the procedure will detect any significant model structure contained in the observed data.

More detailed background on the problem and approach to its solution are contained in Sections 2-4.

1.2 Phase I Objectives

The objectives originally proposed for the Phase I study were more narrowly defined in terms of state affine Markov processes. The results achieved are much more general and apply to general nonlinear Markov processes.

The objective of the Phase I study was to demonstrate the technical feasibility of using state affine Markov processes and canonical variate analysis for obtaining approximations to optimal nonlinear filters and for system identification. In particular, the objectives of the research are to demonstrate the feasibility to:

(1) Develop a general and complete theory of approximation of nonlinear Markov processes using state affine Markov processes

(2) Develop a general theory for canonical variate analysis of nonlinear processes

(3) Develop computational algorithms for canonical variate analysis of nonlinear processes and determine state affine Markov approximations

(4) Demonstrate prototype algorithms on simulation data of nonlinear processes.

1.3 Summary of Phase I Research Results

Not only were all of the objectives of the Phase I research accomplished, but the results were extended to general nonlinear systems including multiple equilibria systems that cannot be approximated by finite order state affine systems. This extension is a major accomplishment and considerably extends the applicability of the methods. Many important aerospace applications are multiple equilibria systems including rotational dynamics of spacecraft, rotational and buckling dynamics of shafts, and nonlinear aeroelastic dynamics of high performance aircraft and missiles.

In particular, the Phase I accomplishments are:

- Development of a general theory for representation of nonlinear Markov processes in a state space innovations form involving the canonical states. The predictive ability of the truncated state is related to the corresponding canonical correlations.
- The general theory of nonlinear canonical variate analysis is developed for nonlinear Markov processes on Hilbert spaces. The theory of maximal correlation is used to prove the existence of solutions to the problem.
- Computational algorithms and software are developed with particular attention to insuring numerical accuracy and stability.
- The software is demonstrated on the Lorenz chaotic attractor which is a multiple equilibria system. Reconstruction of the full state space from observation of only one of the three state coordinates is demonstrated even for the case of substantial process excitation noise. The identified system dynamics have the same character as the true nonlinear system.

A detailed discussion of the technical accomplishments is contained in Section 12.

In this Phase I study, the theory and computational algorithms are developed in outline form to demonstrate the feasibility of a comprehensive development in Phase II. The computational simulation and demonstration involved simplified models which will be replaced with more complex nonlinear systems in Phase II.

1.4 Paper and Conference Activities

During Phase I, several presentations were made at professional conferences and several papers prepared.

The paper "Generalized Canonical Variate Analysis of Nonlinear Systems" was published in the *Proceedings of the 27th IEEE Conference on Decision and Control*, Vol. 3, pp. 1720-5, and presented at the conference held December 7-9, 1988, Austin, TX. The paper was presented in the session Nonlinear Dynamics and Control of Aerospace Systems.

The presentation "Generalized Canonical Variate Analysis of Nonlinear Time Series" was made at the CBMS Regional Conference on Modern Computer Intensive Methods for Exploring and Modeling Multivariate Data held June 12-16, 1989, at George Washington University and jointly sponsored with the National Science Foundation. The conference discussed recent nonlinear regression methods for high dimensional spaces featuring Prof. Jerome Friedman of Stanford University as lecturer. The use of these methods for the identification of nonlinear Markov processes using CVA will be investigated in depth in the proposed Phase II research.

A paper is in preparation summarizing the theory developed in the Phase I study. This paper will be submitted to the *Annals of Statistics*. Also, an additional paper is in preparation summarizing the time series analysis procedure and the computational results of the simulation studies. This paper will be submitted to the *Journal of Time Series Analysis*.

1.5 Proposed Phase II Research

Work is proposed for Phase II research on system identification and stochastic filtering of nonlinear Markov processes using the method of canonical variate analysis. Previous work using other approaches has encountered substantial difficulties in achieving efficient and robust approximate solutions to these problems. The results of the Phase I research demonstrates the feasibility of obtaining accurate and computationally robust solutions to these problems using the CVA method.

The objective of the proposed Phase II research is to investigate in detail the theoretical, statistical, algorithmic, and computational aspects of the problem, and to implement prototype software for demonstrating the procedures on several types of data including rotating mechanical systems. This proposed work is an in depth investigation

into the aspects of the problem that have a direct bearing on the design and implementation of theoretically sound, statistically accurate and computational robust software. Further the computational and simulation studies will appraise the performance of the methods and algorithms.

The proposed Phase II research is a major collaborative effort involving as consultants some of the preeminent researchers in their respective fields. Prof. Jan C. Willems has made innovative contributions to the fundamental theory of dynamical systems including system identification. He will consult on the theoretical development of CVA on Hilbert spaces which is a major focus of the proposed study. Prof. Jerome H. Friedman has made major contributions in the field of nonlinear regression in high dimensional spaces. He will consult on the development of statistically efficient methods for fitting nonlinear dynamical models which is a central issue of the proposed work. Prof. John Baillieul has done extensive analysis of nonlinear systems particularly robotic and rotating mechanical systems, and has developed an extensive experimental facility in the Boston University Robotics Laboratory for data collection and comparison of theoretical models and experimental behavior. He will consult on the use of various models, simulated data, and experimental data in demonstrating the performance and interpreting the resulting identified nonlinear systems.

In addition, Dr. J. Doyne Farmer of Los Alamos National Laboratories has been one of the principal contributors to the development of the state reconstruction approach to modeling chaotic deterministic systems which is complementary to the CVA approach to identification and filtering of nonlinear stochastic systems. He heads a group of four people doing contract work from several agencies on the topic. There will be an informal collaboration between CEI and Los Alamos in exchanging data and computational methods as well as discussing the theoretical relationships between the two approaches.

The proposed Phase II work using CVA is a comprehensive program for the systematic and in-depth development of the dynamical systems theory, statistical estimation and computational methods, and demonstration on simulated and experimental data. The research will receive review and consulting by preeminent researchers in these fields. Further, the proposed study will involve experimental and computing facilities capable of demonstrating the performance of the developed methods and software on significant real systems and data sets.

The technical objectives proposed in Phase II for the modeling of nonlinear dynamical systems are:

(1) Further development of the theory of nonlinear canonical variables using operators on Hilbert spaces as was sketched in this Phase I study. Also further development of independent canonical variables to provide a solution to the problem of selecting states of minimal rank. Continue development of the optimal normalizing transformations and the entropy measure of approximate normality.

(2) Continue the development of the theory of approximation of nonlinear Markov processes including use of the prediction error of the future as measured by the

entropy. Also approximation measures appropriate for recursively defined Markov processes will be considered.

(3) Develop statistical inference theory and methods for the CVA procedure. This includes the development of adaptive nonlinear regression methods for high dimensional multivariate data and the development of related theory for assessing the accuracy of the resulting nonlinear Markov process models.

(4) Development of computational algorithms and prototype software for computational study. Particular attention will be given to the numerical accuracy and stability as well as the computational efficiency.

(5) Computational study using simulated and real data. Real data from robotic and rotating mechanical systems will be analyzed and compared to theoretical models of the dynamics. The behavior and performance of the developed methods and algorithms will be assessed.

The anticipated results of the proposed research is the development of system identification and stochastic filtering using canonical variate analysis. This will include a rigorous development of the mathematical theory of approximation in a Hilbert space setting for a general class of nonlinear controlled Markov processes. Markov processes in state space innovations form will be determined for approximation of optimal filters and nonlinear stochastic models. Near optimal choice of the state will be obtained using CVA on covariance data of the past and future. Computational methods will be developed and demonstrated on nonlinear data from computer simulation as well as laboratory experiments.

1.6 Potential Applications and Phase III Commercialization

1.6.1 Potential Commercial Applications

The chemical and industrial process industries as well as power systems involve a number of highly nonlinear processes that are difficult to identify, filter and control using methods of linear systems. These processes could benefit considerably from a general and automatic procedure for system identification and filtering of nonlinear systems.

Based upon CVA of linear systems, CEI is currently developing on-line system identification and adaptive control software for industrial processes. In a Phase I SBIR funded by NSF, linear system identification methods were applied for the local identification of several nonlinear processes. These included a stirred tank reactor and an autothermal reactor both of which are very nonlinear with the autothermal reactor having multiple equilibria. Present methods use local linearization for control system design because good identification of nonlinear models are not available.

Monitoring and signal enhancement of brain waves is of great importance in clinical analysis as well as neurosurgery. Many of these neural processes are known to be

nonlinear and exhibit chaotic behavior. This field could greatly benefit from nonlinear system identification and filtering methods. The development of accurate models of neural process dynamics has been greatly limited by the lack of good methods and computational algorithms for identifying nonlinear systems. CEI has been involved in the research and development of signal processing methods for a neurosurgery monitoring system using linear system identification and filtering methods.

Power generating plants have nonlinear dynamics of considerable importance to the control and safety of the systems (Zaborszky et al, 1988; Chiang et al, 1988). These dynamics are difficult to study accurately due to lack of available nonlinear system identification methods, and rely primarily upon simulation models constructed from first principles. One of the Phase III Funding Commitments on the NSF SBIR is for use of the linear CVA method for identification of locally linear models for control of a reboiler process in an electric power cogeneration plant. These reboilers are known to be highly nonlinear and will provide CEI with an opportunity to apply nonlinear CVA procedures following the completion of the proposed Phase II work.

1.6.2 Potential Federal Government Applications

A number of DoD systems involve nonlinear processes that require system identification or stochastic filtering. These include aircraft at high angles of attack, slewing dynamics of large space structures, adaptive control of robots, on-line reconfiguration of control systems for failed or battle damaged aircraft or space structures, and the dynamics of physiological neural networks. Accurate system identification and filtering of these processes would have considerable impact on the modeling, control, and understanding of these processes.

The advance of aircraft to greater maneuverability and higher angles of attack while minimizing weight requires a greater understanding of the nonlinear aeroelastic dynamics of the vehicle. While physical modeling of the dynamics are of great value, it is also necessary to verify that the physical models are accurate through wind tunnel and flight testing. Current methods for wind tunnel and flight testing involve primarily the determination of linearized models or simple nonlinear models. The inference of complex nonlinear models from transient or dynamic maneuvers is beyond the current state of the art. The CVA method for nonlinear system identification would be of considerable value in determine the complex aeroelastic dynamics of vehicles.

A notable area of nonlinear modeling is in spacecraft dynamics. The explorer satellite was initially spin stabilized about the desired axis that turned out to be unstable. As a result of constant momentum energy dissipation, the spacecraft settled into an end-over-end stable equilibrium (Kaplan, 1976). In the future, nonlinear system identification can play a major role in reconciling the theoretical models derived from first principles with the observed dynamics of experiments.

Another area of long standing difficulty is the dynamics of rotating machinery. The buckling of shafts of screw propellers on ships goes back a century. This is also a problem

for propeller aircraft and helicopters. Other problems arise in the nonlinear dynamics of turbine engines which can be very sensitive to stall in certain operating modes.

2 Background: Related Literature

This section describes previous work related to the problems of system identification and filtering of nonlinear stochastic processes. This related literature provides a context for the discussion of the CVA approach to these problems.

Fields related to the identification and filtering of nonlinear systems include:

- Volterra-Wiener kernels
- Quasi-linearization
- Nonlinear filtering methods
- Time-series using nonlinear difference equations
- State space reconstruction methods
- Nonlinear regression in high dimensions.

Below, these approaches to solving the nonlinear system identification and filtering problems are discussed along with the nature of the difficulties that have arisen.

2.1 Volterra-Wiener Kernel Methods

Descriptions and representations of nonlinear systems have been studied intensively in the past decade (Rugh, 1981). The approaches include Volterra-Wiener kernels, Laplace transforms, state space equations, and nonlinear difference equations in the inputs and outputs. While there are a number of equivalences among these different representations, some of them are much more difficult to use than others. A major issue that enters many of the problems of identification and filtering is the lack of a finite dimensional representation in most situations. This issue is fundamental to these problems, and the approach to the issue in this Phase I study is to develop efficient finite dimensional approximations to nonlinear systems that allow efficient computation of the solution to appropriate approximation problems.

Volterra-Wiener methods have been used extensively for modeling and identification of nonlinear systems (see the reviews Marmarelis and Marmarelis, 1978; Marmarelis, 1987). The identification approach primarily in use requires the stimulation of the system with white noise or a random impulse train which is not feasible for many nonlinear systems particularly in the presence of a control feedback. Averaging over a large number of samples is required to obtain a good approximation to the system dynamics. The resulting system description in terms of kernels can be of very high state dimension and is not easily used in many analyses. The resulting accuracy in identifying the nonlinear dynamics is very low as compared to parametric statistical procedures. Thus the possible input functions are very restricted, and considerable data is often required. Also the associated computation in using Volterra-Wiener methods is very

large. As a result of the low accuracy and large computational and data requirements, Volterra-Wiener methods have been largely restricted to low order kernels providing only low order approximations in cases where the system may be of high polynomial order.

2.2 Quasi-Linearization Methods

Many of the developments in the identification of nonlinear systems have treated the quasi-linear case where the system is represented as a time varying linear system linearized about the best state estimate at a given time. This approach has been pursued especially in the engineering community using the Kalman filter for evaluation of the likelihood function with the dynamics linearized about each state (Ljung, 1979; Schweppe, 1973; Larimore, 1981). Such quasi-linearization does not work well for systems that depart considerably from linearity. What is required is an approach that will capture the full nonlinear behavior of nonlinear systems.

2.3 Nonlinear Stochastic Filtering

One approach to generalization of the quasi-linearization method is to consider general nonlinear filtering of the observations. The topic of nonlinear stochastic filtering has received considerable attention during the last decade. The major difficulty encountered is that for nonlinear systems, the nonlinear filter is usually infinite dimensional even if the nonlinear system is finite dimensional. Finite computational methods are not available for solving the infinite dimensional filtering problem. Thus it is necessary to solve such problems only approximately. The current literature does not offer good approximation procedures for solution of the approximation problem. A major objective of this Phase I study is to provide a sound theoretical basis for efficient approximation in solving the nonlinear filtering problem.

A major implication of nonlinear stochastic filtering is that even if the original process is deterministic and finite state order, the addition of white noise in the measurements results in an infinite dimensional filter except in special cases. Specifically, suppose that y_t is the output of a deterministic nonlinear discrete time process of finite state order and the observations z_t are

$$z_t = y_t + n_t \tag{2-1}$$

with n_t white noise. The first issue is to determine the state of the process - what functions of the past have information for prediction of the future. If the "true" process y_t could be observed, then the state would be finite dimensional. However, for a simple processes involving a cubic nonlinearity, the state of the measurement process is infinite dimensional (Sussman, 1981; Hazewinkel and Marcus, 1981; Marcus, 1979, 1981; also see Section 7). This means no matter how many nonlinear functions of the past are considered, there is still additional information in the past data. Thus nonlinear filtering is fundamentally an approximation problem.

A closely related topic to the nonlinear filtering problem is the representation of a Markov process in terms of an innovations representation. An innovations representation directly provides an optimal nonlinear filter for a process. In light of the generally infinite dimensional nature of the nonlinear filter, the central issue is approximation of an innovations representation by a finite dimensional representation. As will be developed in this Phase I study, the CVA method will provide an optimal procedure for approximation of an innovations model and corresponding filter of a specified state order.

2.4 Time Series Analysis

Linear time series analysis methods have been used on nonlinear time series. These methods are not very successful because they do not consider the nonlinear terms. This becomes especially important in the forecasting problem. Nonlinear processes can be completely deterministic chaos which is very predictable far into the future, but appear as very noisy and unpredictable when analyzed using linear time series analysis. Thus it is necessary to generalize the time series procedure to include nonlinear models.

Nonlinear time series methods have developed considerably in the last decade. Much of the work has followed the development of linear time series making use of a nonlinear difference equation representation. One of the simplest nonlinear models is the bilinear ARMA form which has been widely studied (Granger and Andersen, 1978; Rao, 1981; Priestley, 1978). Models have been fitted to investigate the limit cycle behavior of nonlinear systems by Haggan and Ozaki (1980) using amplitude dependent AR models and by Tong and Lim (1980) using state dependent AR models.

More recent developments in nonlinear time series analysis include state dependent models (Haggan, Heravi, and Priestley, 1984; Priestley, 1980, 1987). Such models have the form of an ARMA model with the ARMA coefficients depending on the state of the process. The state is represented by delays in the observations which may be a very inefficient representation of the state as discussed in the next subsection. To allow for general nonlinear structures, many lags of the outputs must be included in the state leading to over parameterization of the ARMA model. As a result the model identification is statistically of lower accuracy and the computations may become numerically illconditioned (Gevers and Wertz, 1982). The state dependent ARMA coefficients are considered as time varying functions and estimated by *ad hoc* smoothing methods. The fundamental difficulty is the efficient choice of a state for representation of the past of the process. This is the *embedding* problem discussed in the next subsection and Section 7. The main point of departure of the CVA approach is that the efficient representation of that past of the process by a low dimensional state is addressed directly. This then permits statistically efficient and numerically well conditioned computation of the state space model.

The nonlinear autoregressive moving average with exogenous inputs (NARMAX) models are general polynomial difference equations in the past inputs and outputs of

the system (Leontaritis and Billings, 1985a, 1985b). These developments are primarily aimed toward modeling of a time series in the region of a single stability point, and the presence of multiple equilibria are not discussed. A related paper considers specifically the output affine model which can have only a single equilibrium point (Chen and Billings, 1988). Approximate likelihood functions for these models are expressed in terms of the difference equations, and nonlinear optimization methods are used. The optimization problems have all of the problems encountered in the linear case including illconditioning and lack of a global parameterization, and it often fails to converge in practice on moderately complex systems.

In special cases, however, nonlinear time series analysis methods have been very successful in giving considerably improved models and forecasts as compared to linear time series analysis. This usually requires that the parametric model structure used in the identification be sufficiently rich to capture the structure of the true process, and yet be economical in the use of parameters so that model accuracy is not lost in the estimation of unnecessary parameters. In fact, much of the problem in nonlinear system identification and filtering is the determination of low order, parametrically efficient model structures to obtain accurate models.

2.5 State Space Reconstruction Methods

A different approach growing out of the chaos literature is the state space reconstruction method (Crutchfield and McNamara, 1987; Crutchfield et al, 1982; Packard et al, 1980; Farmer and Sidorowich, 1987, 1988; Farmer et al, 1980, 1983). The method involves

- Embedding in a space of variables that are nonlinear functions of the observations
- Study of the resulting point cloud to determine a low dimensional manifold containing the trajectories and transition dynamics describing the time evolution of the process.
- Measurement of the approximation of the fitted process to the measured process. Entropy measures have been used.

The determination of the manifold of state space motion for determination of a minimal order state is difficult and closely related to the problem of nonlinear regression in high dimensional spaces discussed below. This has been applied primarily to deterministic processes or such processes with moderate measurement noise added. The case of noise excitation of the process so that the process is a Markov process appears to had little study from this point of view. The embedding space is chosen more or less by trial and error or by use of search methods which can have great difficulty in actually finding useful coordinates for the state space. If the choice of variables for embedding is poor, then the resulting model will be a poor predictor. The most advanced measure of model fit appears to be an entropy measure. Such entropy measures can be

extended to measures in a statistical inference setting as discussed in Section 9. The state reconstruction approach is perhaps closest in spirit to that taken in this study.

In a recent paper by Broomhead and King (1987), statistical methods of principal component analysis also known as Karhunen-Loeve expansion are applied to the delay coordinates to determine if the state space manifold is contained in a subspace of the delay coordinates. This gives some reduction of the dimensionality of the embedding coordinates particularly if there is substantial measurement noise. The principal component analysis is a selection of particular linear combinations of the delay coordinated to represent the state of the system.

For low noise, less reduction of the dimension of the embedding space results because there is generally curvature of the state manifold. Further reduction of the embedding requires methods for determining nonlinear transformations of the delay coordinates - a problem that apparently has not been systematically studied in the previous literature. This is a major point of departure of the present study which systematically investigates the determination of an efficient or even minimal state space for embedding of the state manifold. The nonlinear CVA method of the present study will be seen in Section 3 as a generalization of the principal components analysis used by Broomhead and King (1987). The nonlinear CVA generalization uses a predictive measure of fit error and represents states as nonlinear transformations of the delay coordinates.

Recent comparisons between the state reconstruction approach and linear time series analysis have been given in Farmer and Sidorowich (1988) comparing methods for forecasting and smoothing for deterministic chaotic processes observed in little or no observation noise. As would be expected, the nonlinear state reconstruction method does orders of magnitude better. A much more interesting comparison would be with *nonlinear* time series methods which can accurately identify and predict nonlinear processes. Making such a direct comparison is not straight forward since both of the methods involve a high degree of art in making the numerous arbitrary choices required in applying the methods.

A major weakness of the state reconstruction approach is the lack of any stochastic component in the dynamical process. If there is little noise or a large amount of observation data, then a simplistic approach to the noise will suffice. However, as discussed in Section 7, even if the nonlinear process is of finite state order, the addition of additive measurement noise will make the resulting process an infinite state order Markov process. The objective of the state reconstruction approach is obvious in that we wish to determine the most predictable "deterministic" components or states and describe the dynamics. However, all that is available are noisy data from an observation process that is infinite dimensional. The finite dimensional deterministic process cannot be reconstructed from the noisy measurements exactly, so that the approximation problem is to know when a good approximation has been obtained.

Actually the problem is fundamentally more difficult than this. As discussed in Section 7, the addition of additive measurement noise results in a stochastic process that fundamentally involves the propagation of the noise in the system dynamics. Fur-

thermore, the only way to obtain white prediction errors is to determine the state of the system and implement the optimal nonlinear filter. But solution in a simple way for the system dynamics and model of the process requires knowledge of the state that produces the white prediction errors.

The simple and attractive approach of the state reconstruction approach in the deterministic case becomes very complex and difficult in the stochastic case even for the elementary case of additive white observation noise. A fundamentally different approach is required in the stochastic case that considers the effect of random processes.

A statistical treatment of unpredictable components of a process also leads to a somewhat different philosophical view. In the deterministic chaos approach, it is conceptually attractive to view the world as fundamentally deterministic with the apparent random effects being only low dimensional projections of high order chaos. This is not a useful empirical approach to modeling stochastic processes since it presumes "hidden" variables, i.e., unobservable variables. If consideration is restricted to observable variables, then the best that is possible is the construction of states of processes from the observations and determination of the resulting dynamics as accurately as possible. The resulting error in prediction is then most usefully described as a purely unpredictable innovations random process.

A closely related problem is when one or only a few functions of the states are observed. If the nonlinear system is of high dimension, then some of the states will be so indirectly related to the observations that more error is introduced in trying to model them rather than replacing their effect with an excitation noise component in the model. Again, we are faced with an approximation problem.

2.6 Nonlinear Regression in High Dimensions

The topic of nonlinear regression cannot be avoided if the process has excitation noise, i.e., is a Markov process in a high dimensional space. There has been considerable work in the last decade in the area of nonlinear regression for static models as opposed to dynamic time series models (Brieman and Friedman, 1985; Buja, Hastie and Tibshirani, 1989; Koyak, 1987; Friedman and Stuetzle, 1981).

A very general procedure was developed by Brieman and Friedman (1985) for estimating optimal transformations for multiple regression using nonlinear functions. The alternating conditional expectation (ACE) algorithm is used for determining near optimal nonlinear transformations in a semiparametric way so that it is not necessary to specify the parametric form of the function to be fitted. The ACE algorithm involves the same projection operators as used in the Hilbert space theory of nonlinear canonical correlation. The CVA method extended to the nonlinear case in this Phase I study is much more general than the additive functional forms that are used in ACE, but the same basic theory applies to both procedures. ACE appears to be the present state-of-the-art in semiparametric nonlinear regression.

A related approach to fitting nonlinear regression models is the multivariate adap-

tive regression splines (MARS) algorithm (Friedman, 1989). In this approach, the multidimensional space is recursively partitioned and spline functions are fitted on each partition to construct a nonlinear function. The fit of the spline function to the observations is determined for each partition, and the best partition is determined. This is a statistical version of some of the methods currently used in the state reconstruction method.

The above methods appear to offer considerable advantage in developing statistically optimal methods for fitting nonlinear regression functions to obtain state space models. In Phase I, simple polynomial regression methods were used to demonstrate the feasibility of the CVA approach to nonlinear system identification and filtering. Computational aspects of adaptive nonlinear regression methods are discussed in some detail in Section 10, and in Phase II the use of these more advanced and optimal statistical methods is proposed.

3 CVA of Linear Systems

In this section, the results of CVA for linear systems is reviewed to provide the necessary background for extension to nonlinear Markov processes. The problem of the determination of the rank or state order of a linear Markov process is solved including the determination of the minimal rank state and a state space model of the system. The approach determines the linear functions of the past inputs and outputs that have linear predictive power for the future, and in particular the rank of such a predictive function. The analysis and computational method used to solve this problem is a canonical variate analysis (CVA) which is equivalent to a generalized singular value decomposition (SVD). The elegant solution to the problem in the linear case involves finite dimensional spaces and provides a prototype for the nonlinear case treated in the rest of the report where the spaces are fundamentally infinite dimensional.

3.1 Determination of Canonical States

In this section, the background to the canonical variate analysis method is presented, and a recent generalization is given that provides a completely general solution to the reduced rank stochastic prediction problem which is well defined statistically and computationally even when some or all of the various covariance matrices are singular (Larimore, 1989). Previous methods in the statistical literature do not address the general problem. The application of CVA to linear systems is contained in Larimore (1983b).

The analysis of canonical correlations and variates is a method of mathematical statistics developed by Hotelling (1936; also see Anderson, 1958). Concepts of canonical variables for representing random processes were explored by Gelfand and Yaglom (1959), Yaglom (1970), and Kailath (1974). The initial application of the canonical correlation analysis method to stochastic realization theory and system identification was done in the pioneering work of Akaike (1976, 1975, 1974a).

Consider the problem of choosing an optimal system or model of specified state order for use in predicting the future evolution of the process. Consider the past vector p_t consisting of past vector outputs y_t and vector inputs u_t before time t and the future vector f_t of vector outputs at time t or later so

$$p_t = (y_{t-1}^T, y_{t-2}^T, \ldots, u_{t-1}^T, u_{t-2}^T, \ldots)^T, \quad f_t = (y_t^T, y_{t+1}^T, \ldots)^T \tag{3-1}$$

For ease of development in this section, the vector processes y_t and u_t are assumed to be jointly stationary, and the covariance matrices among f and p are denoted as Σ_{ff}, Σ_{pp}, and Σ_{fp}.

For a specified number k, the major interest is in determining k linear combinations of the past p_t which allow the optimal prediction of the future f_t. The set of k linear combinations of the past p_t is denoted as a $k \times 1$ vector m_t and is considered as k-order memory of the past. The optimal linear prediction $\hat{f}_t$ of the future f_t, which is a function

of a reduced order memory m_t, is measured in terms of the prediction error

$$E\{\| f_t - \hat{f}_t \|^2_{\Lambda^\dagger}\} = E\{(f_t - \hat{f}_t)^T \Lambda^\dagger (f_t - \hat{f}_t)\} \tag{3-2}$$

where E is the expectation operation and Λ is an arbitrary positive semidefinite symmetric matrix so that the pseudoinverse $\Lambda^\dagger$ is an arbitrary quadratic weighting that is possibly singular. The optimal prediction problem is to determine an optimal k-order memory

$$m_t = J_k p_t \tag{3-3}$$

by choosing the k rows of J_k such that the optimal linear predictor $\hat{f}_t(m_t)$ based on m_t minimizes the prediction error (3-2).

As derived in Larimore (1986), the solution to this problem in the completely general case where the matrices Σ_{ff}, Σ_{pp}, and Λ may be singular is given by the generalized singular value decomposition as stated in the following theorem (see Van Loan, 1976, for a closely related generalized SVD).

Theorem 3-1: Consider the problem of choosing k linear combinations $m_t = J_k p_t$ of p_t for predicting f_t, such that (3-2) is minimized where Σ_{pp}, and Λ are possibly singular positive semidefinite symmetric matrices with ranks m and n respectively. Then the existence and uniqueness of solutions are completely characterized by the (Σ_{pp}, Λ)-generalized singular value decomposition which guarantees the existence of matrices J, L, and generalized singular values $\gamma_1, \cdots, \gamma_q$ such that

$$\begin{aligned} J\Sigma_{pp}J^T = I_m, \quad L\Lambda L^T = I_n, \\ J\Sigma_{pf}L^T = Diag(\gamma_1 \geq \cdots \geq \gamma_q > 0, \cdots, 0) \end{aligned} \tag{3-4}$$

The solution is given by choosing the rows of J_k as the first k rows of J if the k-th singular value satisfies $\gamma_k > \gamma_{k+1}$. If there are r repeated singular values equal to γ_k, then there is an arbitrary selection from among the corresponding singular vectors, i.e. rows of J. The minimum value is

$$\min_{rank(J_k\Sigma_{pp}J_k^T) = k} E\{\| f_t - \hat{f}_t \|^2_{\Lambda^\dagger}\} = tr\Lambda^\dagger\Sigma_{ff} - \gamma_1^2 - \cdots - \gamma_k^2 \tag{3-5}$$

This result not only gives a complete characterization of the solutions in selecting optimal predictors m_k from the past p_t for prediction of the future f_t, but the reduction in prediction error for all possible selections of order k is given simply in terms of the generalized singular values. This is of great importance since it avoids having to do a considerable amount of computation to determine what selection of order is appropriate in a given problem.

Different selections of the weighting matrix Λ can be used for different purposes. A number of classical reduced rank statistical analysis problems of static variables, i.e. with independence from 'time' to 'time', can be formulated and solved by the generalized CVA of Theorem 3-1. In the classical canonical correlation analysis problem, $\Lambda = \Sigma_{ff}$. In the principal components analysis problem, the 'past' and 'future' are the same space

and in addition $\Lambda = I$. A generalization of this with the 'past' and 'future' different is the principal component analysis of instrumental variables where $\Lambda = I$ (Rao, 1965). The only consideration of the case of singular covariance matrices for the canonical correlation analysis problem is by Khatri (1976). The solution is considerably more complicated and is not related to a computational procedure such as the SVD. Also it does not address the general CVA problem with Λ an arbitrary positive semidefinite matrix. For system identification, the use of the weighting matrix $\Lambda = \Sigma_{ff}$ results in a near maximum likelihood system identification procedure (Larimore, Mahmood and Mehra, 1984).

In the computational problem given finite data, the past and future of the process are taken to be finite of length d lags so

$$p_t^T = (y_{t-1}^T, \cdots, y_{t-d}^T, u_{t-1}^T, \cdots, u_{t-d}^T)^T \quad , \quad f_t^T = (y_t^T, \cdots, y_{t+d-1}^T)^T \tag{3-6}$$

Akaike (1976) proposed choosing the number d of lags by least squares autoregressive modeling using recursive least squares algorithms and choosing the number of lags as that minimizing the AIC criterion discussed below. This insures that a sufficient number of lags are used to capture all of the statistically significant behavior in the data. This procedure is easily generalized to include the case with inputs u. In the model identification problem, by using the weighting matrix $\Lambda = \Sigma_{ff}$ the identified system is close to the maximum likelihood estimation solution (Larimore, Mahmood, and Mehra, 1984). The generalized SVD of Theorem 3-1 determines a transformation J of the past that puts the state in a canonical form so that the memory $m_t = J_k p_t$ contains the states ordered in terms of their importance in modeling the process. The optimal memory for a given order k then corresponds to selection of the first k states.

Computational aspects of the SVD are discussed in Golub (1969), and the problem of canonical variate analysis is discussed in Bjorck and Golub (1973). Algorithms for computation of the SVD on systolic arrays are developed in Brent and Luk (1985), and extensions to the generalized SVD is given in Larimore and Luk (1988).

3.2 State Space Model Fitting

The CVA method in conjunction with entropy based multiple decision procedures allows the determination of the fit of the various state space models, and the selection of the best model state order.

Consider the general case of the reduced order filtering and modeling problem: given the past of the related random processes u_t and y_t, we wish to model and predict the future of y_t by a k-order state x_t and state-space structure of the form

$$x_{t+1} = \Phi x_t + G u_t + w_t \tag{3-7}$$

$$y_t = H x_t + A u_t + B w_t + v_t \tag{3-8}$$

where x_t is the state and w_t and v_t are white noise processes that are independent with covariance matrices Q and R respectively. A special case of the reduced-order filtering

problem is the transfer function approximation problem where u_t and y_t are the input and output processes and an approximate state-space model is desired.

To decide on the model state order or model structure, recent developments based upon entropy or information measures are used. Such methods were originally developed by Akaike (1973) and involve the use of the Akaike Information Criterion (AIC) for deciding the appropriate order of a statistical model. The AIC for each order k is defined by

$$AIC(k) = -2 \log p(Y^N, U^N; \hat{\theta}_k) + 2M_k \tag{3-9}$$

where p is the likelihood function, based on the observations (Y^N, U^N) at N time points, with the maximum likelihood parameter estimates $\hat{\theta}_k$ using a k-order model with M_k parameters. The model order k is chosen corresponding to the minimum value of AIC(k). A predictive inference justification of the use of an information or entropy based criterion such as AIC is given in Larimore (1983a) based upon the fundamental statistical principles of sufficiency and repeated sampling. The number of parameters M_k in the state space model (3-7) and (3-8) is determined by the general state space canonical form as in Candy et al (1979) and is far less than the number of elements in the various state space matrices.

Once the optimal k-order memory m_t is determined, state-space equations of the form (3-7) and (3-8) for approximating the process evolution are easily computed by a simple multiple regression procedure (Larimore, 1983b). Since the CVA system identification procedure involves the state space model form, it has the major advantage that the model is globally identifiable so that the method is statistically well-conditioned in contrast to ARMA modeling methods (Gevers and Wertz, 1982). Furthermore, since the computations are primarily a SVD, the computations are numerically stable and accurate with an upper bound on the required computations. Thus the method is completely reliable, and has been demonstrated as such in tests involving reidentification of the system dynamics tens of thousands of times. From the theory of the CVA method (Larimore, Mahmood and Mehra, 1984), it can be shown that there are no difficulties such as biased estimates caused by the presence of a correlated feedback signal.

4 Technical Approach: CVA of Nonlinear Systems

The CVA approach taken in this study is the extension of the linear CVA method to nonlinear systems. In this section the technical approach of this study is outlined and the relationships between the various topics and the previous literature discussed. Subsequent sections treat these various topics in considerable detail.

The approach is to use CVA to determine best selection for a given order k of the states of the nonlinear system as nonlinear functions of the past inputs and outputs of the system that minimize a measure of prediction error. For much of this initial Phase I study, the functions used are linear combinations of powers and products of the past inputs and outputs. For a given choice of the state order, CVA determines the linear combinations to select as states that give the best least squares linear prediction of the future outputs. The same computational structure is used for the identification of a nonlinear system as for a linear system. The major change is the addition of the powers and products to the 'past' of the process. The nonlinear CVA procedure determines an optimal selection of nonlinear coordinates for the states of the process which are orthogonal where the prediction accuracy due to each state is additive. A much more general formulation of the nonlinear CVA problem is also developed in terms of optimal normalizing transformations and minimal order realizations. Computation methods using the ACE algorithm and demonstration on nonlinear process data is proposed for Phase II study.

Once the choice of state order is determined, the state space model is given by nonlinear regression. This approach has a number of advantages over other system identification methods even in the case of linear systems. The state order is determined first including the actual states as functions of the past observations of the process. This involves primarily a singular value decomposition which is numerically accurate and stable. The nonlinear procedure is also well conditioned so that the resulting algorithm for system identification is computationally well conditioned.

The state space Markov model is in innovations form so that the corresponding optimal nonlinear filter for the approximating state space model is immediate and exact. The degree of approximation of the nonlinear filter corresponds to the degree of approximation of the process by the state space Markov model.

4.1 Nonlinear Markov Processes

A central aspect of the approach to nonlinear system identification and filtering is the use of a nonlinear Markov process model, and in particular the concept of the process state. For many physical systems, it is known from the physics of the problem that the process approximately satisfies a finite set of dynamical system equations. The variables of these equations involve the memory or energy storing variables of the system. If the states of a system can be determined, then all of the information from the past about the future evolution of the process is available.

In a general nonlinear process, the evolution of the state as well as the observations are nonlinearly related and in general the noise disturbances of the process are nongaussian. This is the fundamental complication that has foiled past attempts to extend to nonlinear processes the current methods that have been so successful on linear processes. The most successful nonlinear method to date is that of state reconstruction. However, state reconstruction appears to have a major shortcomings if there is significant noise in the measurements or process itself. Also these methods may have great difficulty on high or infinite dimensional processes since the state is not chosen economically in terms of dimension.

The approach taken in this study considers the completely general case where the process may be nonlinear with possibly substantial measurement or process excitation noise. Also the process may in fact be infinite dimensional. Thus this study address a very general class of nonlinear processes that have not been successfully treated in the past. The main restriction appears to relate to the smoothness of the process dynamics relative to the observed data - if abrupt jumps in the dynamics occur and relatively little data is present, then it may not be possible to resolve such abrupt changes in the dynamics. The theory is completely general, but the identification of the process dynamics may require substantial data.

The model form determined is the state space innovations representation. This form has been much discussed in the literature and appears to be the most robust description of a system from a numerical point of view. Also the innovations representation provides explicitly the optimal nonlinear filter for estimating the system state and future evolution of the observed process. Many of the methods for system analysis, control, estimation and filtering are available for the state space form.

4.2 Nonlinear Canonical Variables

The major problem to be addressed is the approximation of state space innovations models and filters of nonlinear systems. This has been a very difficult problem because of the implicit nature of the state which is not directly observable. That it is usually infinite dimensional for nonlinear processes is a major complication in other approaches.

The CVA approach gives a direct method for determination of an optimal state of a specified dimension for approximation of the system. This is optimal whether the system is deterministic or stochastic, or whether the system is finite or infinite dimensional.

The canonical variables or canonical state of the system give the states in order of their ability to use past information for prediction of the future. The canonical states are mutually orthogonal so that the prediction error due to each is additive. Furthermore, the best $k+1$ canonical states are obtained by adding the $k+1$-st state to the first k canonical states which are themselves optimal. Thus efficient iterative computation is possible.

The theory of CVA for nonlinear processes is developed on the Hilbert space of nonlinear functions of the past and future of a process. The selection of an optimal

state of a given finite order is intrinsically an infinite dimensional problem, even if only a finite set of past observations are considered. This problem is first reduced to a sequential selection problem where the canonical states are selected sequentially one at a time.

The problem of selecting just one additional state is the maximal correlation problem that has been studied in the Hilbert space setting in detail. As a result, the nonlinear CVA problem is shown to have a solution for a general class of nonlinear processes. The existence of the solution involves certain projection or conditional expectation operators. These operators are involved in implementing the alternating conditional expectation (ACE) algorithm. In the most general formulation of the nonlinear CVA problem, pairs of canonical variables are computed using ACE.

4.3 Minimal State Rank

For linear gaussian processes, orthogonality of the state is all that is required to guarantee that the canonical states are minimal order, i.e. that attempting to fit more than the true process state order will detect no further information in the past of the process. This is not true for nonlinear processes. Mutually orthogonal or uncorrelated random variables may be functionally or statistically dependent if the random variables are not Gaussian. The question of choosing a state of minimal rank or order is investigated. Although orthogonality is a useful and computationally convenient way to reduce redundancy among variables, it is only partially useful in nonlinear processes.

First the concept of local minimal rank is defined rigorously in terms of the concepts of differential topology. If the past/future map is constant rank, then a minimal rank state exists at least locally. The global minimal realization problem for deterministic systems is also discussed. Unfortunately these approachs do not provide computational or constructive methods for determining such a minimal rank state.

The theory of maximal correlation gives the result that two vector random variables X and Y are independent if and only if the maximal correlation $\rho^*(X, Y)$ is zero, i.e. for any nonlinear function g and h the correlation $\rho(g(X), h(Y))$ is zero. Thus the condition that the random variables be independent rather than uncorrelated is equivalent to the random variables having zero maximal correlation.

An *independent* CVA is defined where the requirement of orthogonality of the canonical variables or states is replaced by the requirement of statistical independence. This insures that the states are not functionally or statistically dependent. Approaches to computation of such an independent CVA are discussed.

4.4 Optimal Normalizing Transformations

At first glance, maximizing correlation appears to be an arbitrary criterion particularly in the context of nonlinear estimation and prediction. Some optimal properties of canonical variables as approximate normalizing transformations are studied in some

detail.

In the univariate Gaussian case, the nonlinear functions maximizing the correlation are in fact *linear* functions. This result does not generalize to the multivariate case with the requirement that canonical variables be orthogonal, but it does generalize if the canonical variables are *independent*. This suggests yet another reason for requiring independence of the canonical variables.

The interpretation of the CVA as an optimal normalizing transformation is developed. An entropy measure of approximate normality of random variables is developed, and it is shown that an independent CVA produces optimal normalizing transformations.

4.5 Computational Methods

Computational methods are developed for implementing the nonlinear CVA. For ease of demonstration in this Phase I study, detailed implementation is restricted to polynomial basis functions for the Hilbert space. Polynomials up to a particular degree in past lags of the inputs and outputs are used as basis functions. The CVA is then implemented in terms of these basis functions. The nonlinear regressions for determination of the state transition function and the state/measurement transformation are computed using polynomial basis functions. These computations are direct and are used to demonstrate the feasibility of developing workable computational solutions to the nonlinear system identification and filtering problems.

More statistically optimal computational methods of adaptive nonlinear regression for implementing the nonlinear CVA are also considered and are proposed for detailed development in the proposed Phase II work. One of the most successful methods for adaptive nonlinear regression in high dimensional spaces is the ACE algorithm. The theory of the ACE algorithm is very closely related to the theory of maximal correlation. The algorithm actually implements the projection operators used in the proof of existence of nonlinear functions maximizing the correlation. A related approach using multivariate adaptive regression splines (MARS) is also discussed.

4.6 Computational Demonstration

To demonstrate the feasibility of the CVA approach, the developed computational algorithms are applied to the Lorenz chaotic attractor with process excitation noise included. The presence of process noise of significant magnitude presents a much more difficult identification problem than the cases of no process noise and low additive white measurement noise considered in most previous studies. The observation process is taken as one of the states of the Lorenz attractor.

The canonical variables are shown to contain all of the state information of the original three states of the Lorenz attractor. This is demonstrated by transforming the canonical states to the states of the Lorenz attractor. Based upon the identified canon-

ical states, the state transition and output transformations are computed by nonlinear regression. This provides an innovations representation for the observation process.

The identified nonlinear system is used to simulate states and outputs for the canonical states and measurements. The trajectories of the simulated process have the same character as the original 'true' process.

5 Nonlinear CVA Problem on Hilbert Spaces

The extension of CVA to nonlinear systems requires a more general setting such as a Hilbert space of nonlinear functions. In this section, the general problem is posed in a Hilbert space setting and reduced to a sequence of scalar maximal correlation problems.

5.1 Hilbert Space of Nonlinear Functions

Let $X = (x_1, \ldots, x_m)^T$ and $Y = (y_1, \ldots, y_n)^T$ be two sets of random variables defined on a probability space Ω with respect to the probability measure μ and let P denote the induced probability measure on R^M and R^N respectively. The random variables X and Y will play the role of the past p_t and future f_t of previous sections while offering a simplicity in the notation where the notion of time t is not required in the present section.

For a given positive integer r, consider the space $\mathcal{F}_X^r$ and $\mathcal{F}_Y^r$ of all Borel measurable r-dimensional vector functions $f(X) = (f_1(X), \ldots, f_r(X))^T$ and $g(Y) = (g_1(Y), \ldots, g_r(Y))$ satisfying

$$Ef(X) = 0, \quad Eg(Y) = 0 \tag{5-1}$$

The space is a linear vector space on which we define the inner product

$$< f, g >= trEf(X)g^T(Y) = E\sum_{i=1}^{r} f_i(X)g_i^T(Y) = tr\Sigma_{fg} \tag{5-2}$$

where for zero mean random functions as in (5-1) the covariance matrix notation $\Sigma_{fg} = Ef(X)g(Y)^T$ is used. The pseudonorm is given by

$$\|< f, f >\| = < f, f >^{1/2} \tag{5-3}$$

The spaces $\mathcal{F}_X^r$ and $\mathcal{F}_y^r$ are separable Hilbert spaces under the inner product.

5.2 Projection and Conditional Expectation

In this section, projection operators on the Hilbert space are expressed in terms of conditional expectations and are shown to be the solution of an optimal prediction problem. Such a prediction problem is central to the nonlinear CVA problem discussed in the following subsections.

Consider two random variables u and v defined on a probability space with respect to a probability measure P. Then the conditional expectation $E(v|u)$ of v with respect to u is a random variable for each value of u and satisfies

$$\int_A E(v|u)dP = \int_A vdP \quad \text{for every } A \in \mathcal{A}_u \tag{5-4}$$

where $\mathcal{A}_u$ is the least sigma algebra of measurable sets of Ω for which u is measurable. In particular

$$E(E(v|u)) = Ev \tag{5-5}$$

and for any Borel measurable function $g(u)$,

$$E(g(u)v|u) = g(u)E(v|u) \tag{5-6}$$

The concept of projection on a subspace provides the optimal solution to linear problems. Let P_K denote the operator of orthogonal projection on a subspace K. Then for any $h \in \mathcal{H}$

- $h_k = P_K h$ if and only if for some $h_k \in K$ and for all $g \in K$, $< h, g >=< h_k, g >$.
- $h_k = P_K h$ if and only if for some $h_k \in K$

$$\min_{g \in K} \| h - g \| = \| h - h_k \| . \tag{5-7}$$

The conditional expectation has optimal properties as a projection operator in the Hilbert space $\mathcal{F}$.

Theorem 5-1 (Optimal Projection): The optimal projection $P_{\mathcal{F}_u} v$ of the random variable v on the Hilbert subspace $\mathcal{F}_u$ of nonlinear functions of the vector of random variables u is the conditional expectation

$$P_{\mathcal{F}_u} v = E(v|u) \tag{5-8}$$

Proof: From the definition of projection, $E(v|u)$ is the projection of v on $\mathcal{F}_u$ if for all $g \in \mathcal{F}_u$, $< v, g >=< E(v|u), g >$. But this last expression is equivalent to

$$E[vg(u)] - E[E(v|u)g(u)] = E[E(vg(u)|u)] - E[E(v|u)g(u)] = 0 \tag{5-9}$$

using (5-5) and (5-6).

5.3 Multivariate Nonlinear CVA Problem

First consider the problem where f and g are fixed functions and we wish to find the nonlinear function $\hat{g}(f(X))$ such the relative prediction error

$$\| g(Y) - \hat{g}(f(X)) \|_{\Sigma_{gg}^{\dagger}} \doteq E\{[g(Y) - \hat{g}(f(X))]^T \Sigma_{gg}^{\dagger} [g(Y) - \hat{g}(f(X))]\} \tag{5-10}$$

is minimum where $(\dagger)$ denotes the pseudoinverse operation. As shown in the previous section, the solution is given by the conditional expectation projection operator

$$\hat{g} = E(g|f). \tag{5-11}$$

Now with the optimal prediction $\hat{g}$ given by (5-11), consider the r-rank nonlinear prediction problem of finding an r-dimensional nonlinear function $f(X)$ of X and r-dimensional nonlinear function $g(Y)$ of Y so as to minimize (5-10). Specifically consider the following minimization problem.

Rank r Nonlinear Prediction Problem: For a given positive integer r, find r-dimensional vector functions $f(X)$ and $g(Y)$ minimizing the relative prediction error

$$\max_{(f,g)} \| g(Y) - \hat{g}(f(X)) \|_{\Sigma_{gg}^{\dagger}} \tag{5-12}$$

where $\hat{g} = E(g|f)$.

5.4 Reduction to Sequential Selection Problem

In this section, the multivariate nonlinear CVA problem is reduced to a sequence of scalar problems each involving maximizing the correlation between a pair of nonlinear functions.

Since $f(X)$ enters the problem only through $\hat{g}(f(X))$, it is sufficient to include the nonlinearity of $\hat{g}$ in the function $f(X)$. In this case for a particular $g(Y)$, the optimal prediction is given by

$$\hat{g}(f(X)) = \Sigma_{gf}\Sigma_{ff}^{\dagger} f(X) \tag{5-13}$$

and the prediction error is

$$\| g(Y) - \hat{g}(f(X)) \|_{\Sigma_{gg}^{\dagger}} = tr\Sigma_{gg}^{\dagger}|(\Sigma_{gg} - \Sigma_{gf}\Sigma_{ff}^{\dagger}\Sigma_{fg}| \tag{5-14}$$

The above measure is invariant to linear transformations of both f and g. Thus we are free to impose the constraints

$$\Sigma_{ff} = \Sigma_{gg} = I_r \tag{5-15}$$

so that f and g are each orthonormal sets of functions. Then the prediction error reduces to

$$tr(I_r - \Sigma_{gf}\Sigma_{fg}) = r - tr(\Sigma_{gf}\Sigma_{fg}) = r - \sum_{i,j=1}^{r} < f_i(X), g_j(Y) >^2 \tag{5-16}$$

which is the sum of the squares of the elements of the covariance matrix Σ_{fg}.

The optimal prediction problem then reduces to finding f and g to solve the maximization problem

$$\max_{\Sigma_{ff} = \Sigma_{gg} = I_r} tr(\Sigma_{gf}\Sigma_{fg}) \tag{5-17}$$

Note that in the univariate case of $r = 1$, the problem is precisely the maximal correlation problem to be discussed in the next section.

One further simplification occurs by doing a SVD on the covariance matrix Σ_{fg}

$$\Sigma_{fg} = UDV^T, \quad D = Diag(d_1, \geq \ldots \geq d_k > 0, \ldots, 0) \tag{5-18}$$

Then the prediction error reduces to

$$r - \sum_{i=1}^{k} d_i^2 \tag{5-19}$$

which is the sum of the squared canonical correlations d_i. The optimal prediction problem is then given by the solution of

$$\max_{\substack{\Sigma_{ff} = \Sigma_{gg} = I_r \\ \Sigma_{fg} = Diag}} \sum_{i=1}^{r} d_i^2 \tag{5-20}$$

For a given fixed R, let f^R and g^R denote the R-dimensional functions maximizing (5-20) for $r = R$. Then it is easily shown that the maximum of (5-20) is achieved for all $r < R$ with f^r and g^r the first r components of f^R and g^R respectively and with D^r the first r components of D^R (Larimore, 1989). Thus the problem is transformed to the sequence of R one dimensional problems for $r = 1, \ldots, R$. In particular, this is summarized in the following theorem

Theorem 5-2 (Sequential Selection): The vector functions f and g giving an optimal solution to the nonlinear prediction problem (5-12) are given sequentially by the following procedure: For each r, find the pair of functions (f_r, g_r) such that they are orthogonal to the previously selected functions $f^{r-1} = (f_1, \ldots, f_{r-1})^T$ and $g^{r-1} = (g_1, \ldots, g_{r-1})^T$ respectively and maximize the correlation, i.e. such that

$$d_r = \max_{\substack{\Sigma_{f^{r-1}f_r} = \Sigma_{g^{r-1}g_r} = 0 \\ \Sigma_{f_r f_r} = \Sigma_{g_r g_r} = 1}} \Sigma_{f_r g_r} \tag{5-21}$$

6 Maximal Correlation and Projection Operators

The solution to the nonlinear CVA problem was reduced in Section 5 to the problem of finding scalar functions $f(X)$ and $g(Y)$ of the respective sets X and Y of random variables such that the correlation is maximized. In this section, properties of the maximal correlation are investigated and related to projection operators on Hilbert spaces. We follow primarily the development of Renyi (1959; see also Csaki and Fischer, 1960, 1963).

6.1 Maximal Correlation

Consider random vectors $X = (x_1, \ldots, x_m)$ and $Y = (y_1, \ldots, y_n)$

The *maximal correlation* of X and Y is defined as

$$\rho^*(X,Y) = \sup_{f,g} \rho(f(X), g(Y)) = \sup_{\substack{f,g \\ \| f \| = 1 \\ \| g \| = 1}} E[f(X)g(Y)] \tag{6-1}$$

where f and g run over all Borel measurable functions with zero mean, i.e. $Ef = Eg = 0$, and $\rho(f,g)$ is the correlation coefficient given in this case by $E[f(X)g(Y)]$.

The maximal correlation satisfies the following properties:

- $\rho(X,Y)$ is defined for any pair of random vectors X and Y, neither of them being a constant with probability 1.
- $\rho^*(X,Y) = \rho^*(Y,X)$.
- $0 \leq \rho^*(X,Y) \leq 1$.
- $\rho^*(X,Y) = 0$ if and only if X and Y are *stochastically independent.*
- $\rho^*(X,Y) = 1$ if there is a *strict dependence* between X and Y, i.e. $f(X) = g(Y)$ for some nonzero Borel measurable functions f or g. The converse requires some additional conditions.
- *Invariance.* Under 1-1 onto Borel-measurable transformations f and g, $\rho^*(f(X), g(Y)) = \rho^*(X,Y)$.
- If the joint distribution of X and Y is normal, then $\rho^*(X,Y)$ is the maximum canonical correlation, i.e. the sup is achieved by considering only *linear functions f and g.*

6.2 Projection Operators and the Maximum

When the maximum is attained for some pair of functions f and g, then certain operator equations involving projections are satisfied. These operator equations are used in the next subsection to establish the existence of functions f and g attaining the maximum.

Now consider the situation when the optimal solution exists to the CVA problem (5-17). Then there exist $f_0(X)$ and $g_0(Y)$ such that

$$E(g_0(Y)|X) = \Sigma_{g_0 f_0} f_0(X) \tag{6-2}$$

recalling the normalization (5-15), and similarly

$$E(f_0(X)|Y) = \Sigma_{f_0 g_0} g_0(Y) \tag{6-3}$$

Furthermore by taking the conditional expectation of (6-3) with respect to X and expressing the result in terms of projections operators we have

$$P_{\mathcal{F}_X} P_{\mathcal{F}_Y} f_0 = E(E(f_0(X)|Y)|X) = \Sigma_{f_0 g_0} \Sigma_{g_0 f_0} f_0(X) \tag{6-4}$$

and similarly

$$P_{\mathcal{F}_Y} P_{\mathcal{F}_X} g_0 = E(E(g_0(Y)|X)|Y) = \Sigma_{g_0 f_0} \Sigma_{f_0 g_0} g_0(Y) \tag{6-5}$$

Thus the optimal solutions f_0 and g_0 are eigenvectors of a successive projection operator. The function f_0 is projected successively on the function space $\mathcal{F}_Y$ and $\mathcal{F}_X$ and similarly for g_0. The eigenvectors f_0 have the common eigenvalue $\Sigma^2_{g_0 f_0}$ which is equal to the squared maximal correlation.

To determine the existence of an optimal solutions f_0 and g_0 to (6-4) and (6-5), we study the pair of operator equations

$$Af = D^2 f \tag{6-6}$$

$$Bg = D^2 g \tag{6-7}$$

where A and B are the operators defined on the left hand side of (6-4) and (6-5) respectively. We wish to determine under what conditions there exist solutions f_0 and g_0 to these operator equations with D equal to the maximal correlation.

6.3 Existence of the Maximum

The existence of the maximum is insured by results from the theory of operators on Hilbert spaces. We will show that the operators A and B are bounded, selfadjoint and continuous under the assumption that the distributions are regular as defined below. Such operators must then have eigenfunction solutions f and g attaining the maximum eigenvalue D^2.

First we prove the following.

Theorem 6-1: The operators A and B are bounded selfadjoint linear operators such that the maximum eigenvalue is $(\rho^*(X,Y))^2$, the squared maximal correlation.

Proof: We deal primarily with the operator A with the result applying also to B. To show boundedness of the operator A, consider

$$\begin{aligned} \| v \|^2 &= Ev^2 = EE([v - E(v|u) + E(v|u)]^2|u) \\ &= EE([v - E(v|u)]^2|u) + E\{E([v - E(v|u)|u)E(v|u)\} + E[E(v|u)]^2 \\ &\geq E[E(v|u)]^2 = \| P_{\mathcal{F}_u} v \|^2 \end{aligned} \tag{6-8}$$

from which it follows that $\| Af \| \leq \| f \|$ so A is bounded.

Now if f_1 and f_2 are in $P_{\mathcal{F}_X}$, then from (5-5) and (5-6)

$$\begin{aligned} < Af_1, f_2 > &= E[E(E(f_1(X)|Y)|X)f_2(X)] \\ &= E[E(f_2(X)E(f_1(X)|Y)|X)] = E[f_2(X)E(f_1(X)|Y)] \\ &= E[E(f_2(X)E(f_1(X)|Y)|Y)] \\ &= E[E(f_2(X)|Y)E(f_1(X)|Y)] = < f_1, Af_2 > \end{aligned} \tag{6-9}$$

where the last equality follows by the symmetry in f_1 and f_2. Thus A is selfadjoint.

To determine the maximum eigenvalue of A, consider for any $f \in P_{\mathcal{F}_X}$ and $g \in P_{\mathcal{F}_Y}$ with $\| f \| = \| g \| = 1$

$$E^2(f(X)g(Y)) = E^2(E(f(X)|Y)g(Y)) \leq E(E^2(f(X)|Y)) = < Af, f > \tag{6-10}$$

using (5-6) and the Schwarz inequality. Thus with λ the maximum eigenvalue of A we have $(\rho^*(X,Y))^2 \leq \lambda$.

Now if $f \in \mathcal{F}_X$ and $\| f \| = 1$ and letting $g(Y) = E(f(X)|Y)$ gives $E^2(g(Y)) = E(f(X)g(Y)) \leq DE^2(g(Y))$ which implies $E^2(g)Y)) \leq D$ so that

$$< Af, f > = E(g(X)E(f(X)|Y)) = E(f(X)g(Y)) \leq DE^2(g(Y)) \leq D^2 \tag{6-11}$$

Thus $(\rho^*(X,Y))^2 \geq \lambda$, and from the above we have $(\rho^*(X,Y))^2 = \lambda$ which proves the theorem.

To show that A and B are completely continuous, some restrictions on the distribution of the random variables are required.

Definition 6-2: The dependence between the vector random variables X and Y is *regular* if the joint distribution $P(X,Y)$ is absolutely continuous with respect to the direct product $P_{X \cdot Y}$ of the marginal distributions P_X and P_Y where the direct product distribution is defined by

$$P_{X \cdot Y}(x \in X \text{ and } y \in Y) = P_X(x \in X)P_Y(y \in Y) \tag{6-12}$$

If the dependence between X and Y is regular, then the probability density $k(x,y)$ of the joint distribution $P(X,Y)$ relative to the product distribution $P(X)P(Y)$ exists as a Radon-Nikodyn derivative satisfying

$$P_{X*Y}(C) = \int\int_C k(x,y)dP_X(x)dP_Y(y) \tag{6-13}$$

for every Borel set C in the product space $X * Y$.

If X and Y are regular, then the conditional expectation operator has a useful expression as an integral operator with kernel $k(x,y)$ (Csaki and Fischer, 1963)

$$E(f(X)|Y) = \int f(x)k(x,y)dP(x) \tag{6-14}$$

The continuity of the operators A and B is shown in the following theorem.

Theorem 6-3: If the dependence between X and Y is regular and $Ek^2(x,y)$ is finite, then the operators A and B are completely continuous.

Proof: In proving this theorem, the operators A and B are expressed as integral operators involving $k(x,y)$. The respective kernals of the integral operators are then shown to be square integrable and hence completely continuous.

From (6-4) and (6-14), the operator A has the expression

$$Af(x) = \int f(u)\int k(u,v)k(x,v)dP_Y(v)dP_X(u) \tag{6-15}$$

which is an integral operator with kernel given by the inner integral

$$h(u,x) = \int k(u,v)k(x,v)dP_Y(v) \tag{6-16}$$

Now the integral operator (6-15) is completely continuous if the kernei $h(u,x)$ is square integrable, i.e. if

$$\int\int h^2(u,x)dP_X(u)dP_X(x) < \infty \tag{6-17}$$

Using the Schwarz inequality

$$h^2(u,x) = [\int k(u,v)k(x,v)dP_Y(v)]^2 \leq \int k^2(u,v)dP_Y(v)\int k^2(x,v)dP_y(v) \tag{6-18}$$

so that from the square integrability of $k(x,y)$ we have

$$\int\int h^2(u,x)dP_X(u)dP_X(x) \leq [\int\int k^2(x,y)dP_X(x)dP_Y(y)]^2 = [Ek^2(x,y)]^2 < \infty \tag{6-19}$$

7 Nonlinear Markov Processes

In this section, various aspects and representations of Markov processes are developed. First some major distinctions between deterministic and stochastic processes are discussed. The fundamental properties of the state of a Markov process are reviewed. The selection of coordinates for the state are developed in terms of the canonical variate analysis. Given a state for a Markov process, the development of the state space innovations representation is immediate. The special case of a state affine representation is described.

7.1 Representations of Deterministic and Stochastic Processes

There are a number of distinctions required in going from representations of deterministic nonlinear systems or stochastic linear systems to nonlinear stochastic systems. The finite dimensionality of the problem is lost except in special cases as described in this section.

There has been considerable interest and work in recent years on identification and modeling of deterministic nonlinear processes particularly chaotic processes. The approach has involved primarily delay coordinates

$$(y_{t-1}, \ldots, y_{t-k}) \tag{7-1}$$

of the observations y_t. It is shown in Takens (1981) that the if r is the state dimension of the process, then it is sufficient to choose the number of delays $k = 2r + 1$ to insure that all of the state information is contained in the k delays (7-1). Most of the work in the field has been done from this point of view with some consideration given to numerical conditioning. The choice of the time delay τ involved in sampling and the use of principal components to improve the conditioning has been considered by Broomhead and King (1987). Also the use of mutual information for choice of the delay coordinates is discussed in Fraser and Swinney (1986).

Major difficulties appear when the process is not deterministic but involves noise. Consider what is perhaps the simplest case of a linear system of autoregressive type (AR) of the form

$$y_t = a_1 y_{t-1} + \cdots + a_p y_{t-p} + e_t = \hat{y}_t + e_t \tag{7-2}$$

where e_t is white noise. If y_t is observed directly, then the problem of determining the coefficients a_i is a linear regression problem in the y_{t-i} with independent noise e_t. The optimal linear least squares prediction $\hat{y}_t$ of the observation y_t as a function of the past is a linear function

$$\hat{y}_t = f(y_{t-1}, \ldots, y_{t-p}) \tag{7-3}$$

of only p delay coordinates.

Consider the case where white noise n_t is added to y_t itself in (7-2) so that the observed process is

$$z_t = y_t + n_t. \tag{7-4}$$

Since the variables y_t are not observable in this case, to express the model strictly in terms of the observable variables z_t requires use of an ARMA structure of the form

$$z_t = a_1 z_{t-1} + \cdots + a_p z_{t-p} + e_t b_1 e_{t-1} + \cdots + b_q e_{t-q} + e_t \tag{7-5}$$

with $q = p$ and where e_t is a white noise process (Box and Jenkins, 1976). The optimal linear least squares estimate $\hat{z}_t$ is still a linear function of the past z_{t-i}, but it involves the infinite past observations $z_{t-1}, z_{t-2}, \ldots$. Thus a finite number of delay coordinates are no longer sufficient. In addition, the prediction error

$$d_t = z_t - \hat{z}_t, \tag{7-6}$$

is now correlated in time. Furthermore, fitting the coefficients a_i and b_j of this ARMA model from the observations z_t is no longer a linear problem.

The ARMA model can be represented in state space form by p states, but the states are not expressible as a function of the finite past. Thus even in the case of a linear autoregression observed in additive measurement noise, the state is not exactly expressible in terms of a finite number of delay coordinates. Thus we are fundamentally stuck with an approximation problem.

In the general case of a nonlinear stochastic system, the finite dimensionality of the state is lost with the addition of white measurement noise except in special cases (Sussman, 1981; Hazewinkel and Marcus, 1981; Marcus, 1979, 1981). Consider for example the one state nonlinear process

$$x_{t+1} = x_t + \tau^{-1} x_t^3 + v_t \tag{7-7}$$

$$y_t = x_t \tag{7-8}$$

with the state observed directly and where v_t is white noise and τ is the sampling time interval. If white measurement noise n_t is added

$$z_t = y_t + n_t \tag{7-9}$$

where z_t is the noisy measurement, then the process z_t is no longer a process with a finite dimensional state. By this we mean that there is no finite dimensional state $\hat{x}_t$ expressible as a function of the past outputs $z_{t-1}, z_{t-2}, \ldots$ that contains all of the information for prediction of the future $z_t, z_{t+1}, \ldots$. A central issue is thus how to choose a good approximation to the process when a finite number of delays in the measured output does not capture the state information.

The problem of representation of a nonlinear process by a state space model is very closely related to the nonlinear stochastic filtering problem. There was a lot of activity in nonlinear stochastic filtering around 1980 exploring the possibility of finite dimensional filters for nonlinear processes (Hazewinkel and Willems, 1981). The consensus of that work was that finite dimensional filters only exist in special cases. The effect of this negative result has been a relatively low level of activity in the field. Effective finite dimensional approximation methods have not been developed for nonlinear filtering.

Thus it is seen that the identification and filtering problems for nonlinear Markov processes are fundamentally more difficult than for either linear Markov processes or nonlinear deterministic processes. Except in special cases, for nonlinear Markov processes a finite number of delay coordinates or even a finite number of states does not capture all of the information in the past for prediction of the future. In the sections below, approximate representations are developed using CVA which have an optimality property in terms of finding an optimal state of a prescribed order. In addition, the canonical states can be computed recursively in order, or for a given order the optimal choice for all lower orders is also given.

7.2 Past/Future Markov Property

In this section, the fundamental property of the state of a Markov processes is reviewed.

For simplicity, consider first purely stochastic processes with no observed deterministic input to the system. A fundamental property of a nonlinear, strict sense Markov process of finite state order is the existence of a finite dimensional state x_t which is a nonlinear function of the past p_t

$$x_t = C_t(p_t) \tag{7-10}$$

with $C_t(\cdot)$ a nonlinear function. The state x_t has the property that the conditional probability of the future f_t conditioned on the past p_t is identical to that of the future f_t conditioned on the finite dimensional state x_t so

$$P\{f_t|p_t\} = P\{f_t|x_t\} \tag{7-11}$$

Thus, only a finite amount of information from the past is relevant to the future evolution of the process.

To extend this concept to processes involving deterministic controls or inputs, the effects of future inputs must first be removed from the future outputs. Let v_t denote the future inputs $v_t^T = (u_t^T, u_{t+1}^T, \ldots,)$ and consider the conditional random variable $f_t|v_t$. Then the process is a *controlled Markov processes* of order k if there exists a k-order state such that the conditional distribution of $f_t|v_t$ given the past p_t is identical to the conditional distribution of $f_t|v_t$ given the state x_t so

$$P\{(f_t|v_t)|p_t\} = P\{(f_t|v_t)|x_t\} \tag{7-12}$$

This is equivalent of the statement that

$$P\{f_t|v_t, p_t\} = P\{f_t|v_t, x_t\} \tag{7-13}$$

7.3 Selecting Nonlinear Coordinates for the Past

The canonical variables give optimal selection of nonlinear coordinates of the past for embedding the state of a nonlinear Markov process. Two approaches to the determination of canonical variables are discussed in this report. The first approach is the use

of polynomial functions as basis functions to approximate the Hilbert space with the canonical variables determined as linear combinations of the polynomials. The second approach involves the implicit determination of the canonical variables using projection operators and is discussed in detail in Section 10.

The first approach, discussed in this section, proceeds in two stages. First a set of polynomial basis functions are selected for representation of the past, and second the canonical variables are determined as linear combinations of the polynomial basis functions.

The selection of polynomial basis functions can be done conveniently using a procedure for modeling a nonlinear autoregression with exogenous input (NARX) process. What is to be determined are the polynomial terms to use for the basis functions. This involves both the variables of the past including time lag and the associated powers to be used. By contrast, the basis used in the state reconstruction method is just the variables of the past back so many lags which may not be a very economical choice. The use of the NARX modeling gives a computationally inexpensive procedure for a good if not nearly optimal way for selecting a finite basis of a given order for the past.

The use of the NARX modeling has the interpretation of fitting a NARX model using least squares, or if the one step prediction error were assumed to be gaussian then the procedure is maximum likelihood. The process is assumed to be generated by a NARX model of the form

$$y_t = \sum_{\iota \in I} a_\iota p_t^\iota + n_t \tag{7-14}$$

where a_ι are coefficients and n_t is white gaussian noise with covariance matrix Σ. I is a set of indices ι denoting the powers corresponding to the components of the past p_t as defined in (7-25).

For efficiently comparing the various possible terms to include in the NARX model, a subset regression procedure can be used. A given maximum degree of the polynomial can be selected along with the maximum number of lags in the past variables. The leaps and bounds algorithm efficiently determines which of the terms need to be included in the polynomial basis functions (Furnival and Wilson, 1974). For comparing among models, the Akaike information criterion is useful and is an optimal order selection procedure in the case that the process is gaussian (see Section 3).

The selected polynomial basis functions provide a basis for describing the canonical variables as linear functions of polynomials in the past. These nonlinear canonical variables given the optimal selection of nonlinear coordinates of the past for embedding the state space manifold. This is optimal in terms of the prediction error of the future for a given selection of state order requiring orthogonality of the state variables. As discussed in Section 8, orthogonality of the canonical variables leads to functional and statistical redundancy in the selection of the canonical variables. The requirement of statistical independence in place of orthogonality solves this problem and leads to a minimal order realization of the nonlinear Markov process if such exists.

Following sections of the report discuss in detail the theory and computational pro-

cedures for selecting the state of a process as the canonical variables. In the next subsection, the construction of a state space model for the process is developed assuming that the state has been determined by CVA.

7.4 State Space Innovation Representation

Now suppose that the state x_t is given from CVA as in the previous subsection, and we wish to obtain the generally nonlinear state equations describing the state evolution and observed output from the observed inputs and unobserved disturbances. First we define, for a given selection for the state x_t of the process, the *innovations* process which is the error in the optimal nonlinear prediction $E(y_t|x_t)$ of the process y_t from the state x_t given by

$$\nu_t = y_t - E(y_t|x_t) \tag{7-15}$$

Then the following shows that the vector (ν_t, u_t, x_t) is a state at time $t+1$ and thus the state evolution can be obtained as a nonlinear function of these variables.

Theorem 7-1. Suppose that the joint and marginal densities among p_t, f_t, v_t, u_t, and y_t are nonzero. Then the state at time $t+1$ is a function of x_t, u_t, and y_t, and the state evolves as

$$x_{t+1} = \phi(x_t, u_t, \nu_t) \tag{7-16}$$

where the innovation process ν_t is an orthogonal increment process orthogonal to $(p_t, u_t)^{[\infty]}$ defined by

$$y_t = \mu_t(x_t, u_t) + \nu_t \tag{7-17}$$

where $\mu_t(x_t, u_t)$ is the projection of y_t on $\mathcal{F}_{(x_t,u_t)}$.

Proof: From the Hilbert space projection theorem, the present output y_t decomposes as in (7-17) where the innovation ν_t is orthogonal to the subspace $\mathcal{F}_{(x_t,u_t)}$. Thus from this, functions of (y_t, x_t) can be replaced by functions of (ν_t, u_t, x_t). The conditional probability at time $t+1$ can be expressed as

$$\begin{aligned}
p(f_{t+1}|v_{t+1}, p_{t+1}) &= p(f_{t+1}|v_{t+1}, y_t, u_t, p_t) && (7\text{-}18)\\
&= \frac{p(f_{t+1}, y_t, p_t|v_t)}{p(y_t, p_t)} && (7\text{-}19)\\
&= \frac{p(f_{t+1}, y_t|v_t, p_t)p(p_t)}{p(y_t|p_t)p(p_t)} && (7\text{-}20)\\
&= \frac{p(f_{t+1}, y_t|v_t, x_t)p(x_t)}{p(y_t|x_t)p(x_t)} && (7\text{-}21)\\
&= p(f_{t+1}|v_{t+1}, y_t, u_t, x_t) && (7\text{-}22)\\
&= p(f_{t+1}|v_{t+1}, \nu_t, u_t, x_t) && (7\text{-}23)
\end{aligned}$$

where (7-21) follows from (7-13), (7-22) from (7-19) through (7-21) with p_t replaced by x_t, and (7-23) from (7-17). Thus the variables (ν_t, u_t, x_t) provide a state for the process at time t and the state evolves functionally in the form (7-16).

The importance of this is in the evolution of the state equations. Let x_{t+1} be a minimal order state. Then from the above, the variables (ν_t, u_t, x_t) generate the subspace $\mathcal{F}_{(\nu_t, u_t, x_t)}$ containing the state x_{t+1} so that x_{t+1} can be found by projection

$$x_{t+1} = E(x_{t+1}|(\nu_t, u_t, x_t)) = \phi_t(\nu_t, u_t, x_t) \tag{7-24}$$

using the conditional expectation operator $E\{\cdot|\cdot\}$. The structures of the nonlinear stochastic model and filters are illustrated respectively in Figures 7-1 and 7-2.

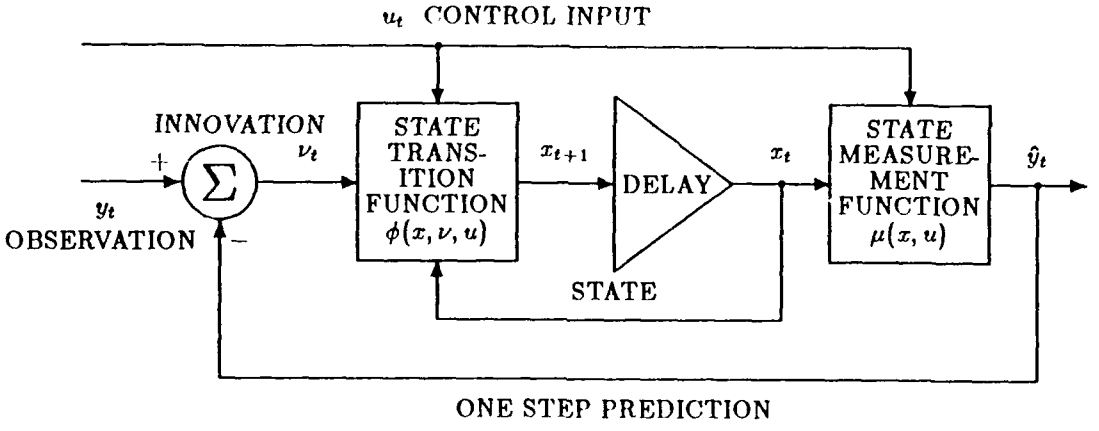


Figure 7-1: State Space Innovations Filter

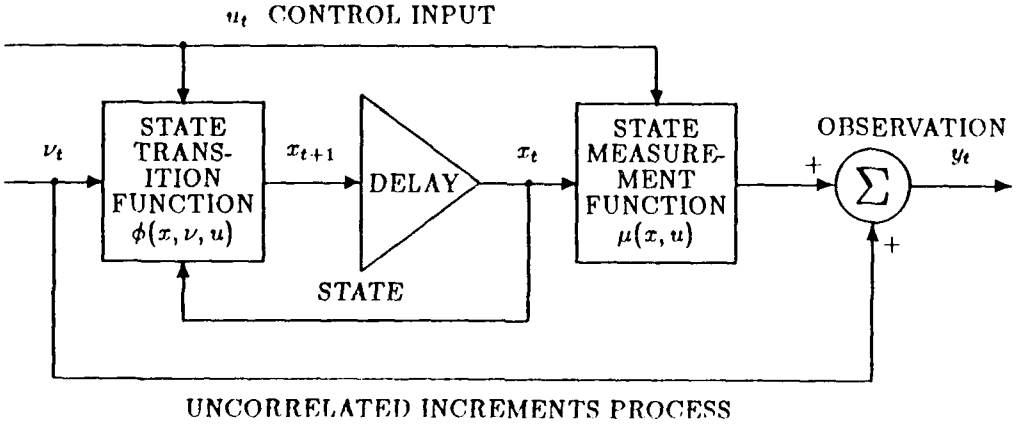


Figure 7-2: State Space Innovations Process Model

7.5 State Affine Representation

In this subsection, state affine models for nonlinear processes are discussed. The work originally proposed for study in Phase I was based to a significant degree on state affine

models because the theory is fairly well developed at this point. In the course of this Phase I study, the methods developed were shown to apply to very general nonlinear processes so that the focus has shifted to the much more general case which of course includes the state affine process as a special case.

State affine processes are processes such that there exists a state space representation with the output a linear function of the state and the state transition function also a linear function of the state. This implies that the entire future f_t outputs of the process as a function of the affine state x_t is a linear function. As a consequence, the state affine structure of the processes affords a number of very nice essential linear properties. Unfortunately, this linearity is also the problem with state affine system - they are essentially linear systems and cannot exhibit the very rich phenomena of nonlinear systems such as chaos and multiple equilibria.

The direct extension of deterministic state affine processes to state affine Markov processes is problematic for several reasons. First, approximation of deterministic processes by state affine processes utilizes the approximation of a bounded nonlinear functions on a finite interval by polynomial functions. This is made precise in the following theorem suggested by Sontag (1979) and stated by Fliess and Normand-Cyrot (1982) with a modified proof given in Diaz (1986).

Theorem 7-2: On a finite time interval and with bounded inputs, any input-output map in which the output depends continuously on the past inputs can be approximated arbitrarily well by state affine systems.

A second difficulty is that random processes in general do not have bounded inputs but are often of bounded variance. The general Hilbert space setting of CVA provides a simple solution to the problem. An affine state is any state that is linearly related to the future of the process. Thus if a process is affine and if a CVA between the nonlinear past and linear future is done, an affine state will be obtained. However, for more general nonlinear Markov processes, such a CVA will not in general produce an affine state as demonstrated in the simulation study of the Lorenz attractor in Section 11.

To be more specific, consider the two function space $\mathcal{L}_f$ and $\mathcal{F}_p$ of linear functions of the future f_t and nonlinear functions of the past p_t respectively. Then the theory of maximal correlation guarantees that the canonical variables for such a CVA exist and minimize the prediction error of the future. Now if the state affine process is finite dimensional, then the canonical variables will be a function of the affine state and so the canonical variables will also be finite dimensional. From the state space innovations representation, the process has the state space form (7-16) and (7-17) were in addition the functions ϕ_t and μ_t are linear in the affine state x_t. In particular, the state affine process can be written explicitly in terms of polynomial functions.

First some notation is developed. Consider a process with vector inputs $u_t = (u_{1t}, \ldots, u_{mt})^T$ and vector outputs $y_t = (y_{1t}, \ldots, y_{nt})^T$ indexed by time t. The notation $(y_t, u_t)^\iota$ is defined as the product of powers of components of y_t and u_t

$$(y_t, u_t)^\iota = y_{1t}^{i_1} \cdots y_{nt}^{i_n} u_{1t}^{j_1} \cdots u_{mt}^{j_m} \tag{7-25}$$

where the vector index $\iota = (i_1, \ldots, i_n, j_1, \ldots, j_m)$ denotes the powers of the product and ranges over some set I of powers. The *degree* of such a power product term is defined as

$$d(\iota) = i_1 + \cdots + i_n + j_1 + \cdots + j_m \tag{7-26}$$

Define the past p_t as the past inputs u_t and outputs y_t and the future f_t as present and future outputs

$$p_t^T = (y_{t-1}, y_{t-2}, \ldots, u_{t-1}, u_{t-2}, \ldots)^T, \quad f_t^T = (y_t, y_{t+1}, \ldots)^T \tag{7-27}$$

and let $p_t^{[r]}$ be the vector of powers and products of the components of past inputs and outputs with degree $d(\iota)$ no greater than r so

$$p_t^{[r]} = (y_{1[t-1]}, y_{1[t-2]}, \ldots, u_{1[t-1]}, u_{2[t-1]}, \ldots, (p_t)^\iota, \ldots)^T \tag{7-28}$$

The *nonlinear past* $p_t^{[\infty]}$ is the infinite vector including product terms of all orders. In practice only finite approximations of various orders are considered.

Suppose that the affine functions ϕ_t and μ_t have continuous derivatives up to the order q so that they can be expanded to that order in a taylor series. Then the affine Markov process $\overline{y}_t$ can be expressed in innovations form for approximating the true process y_t as

$$x_{t+1} = Ax_t + Bvec\{(u_t, \nu_t)^{[q]} x_t^T\} + C(u_t, \nu_t)^{[q]} \tag{7-29}$$

$$\overline{y}_t = Hx_t + Dvec(u_t^{[q]} x_t^T) + Eu_t^{[q]} + \overline{\nu}_t \tag{7-30}$$

where $\overline{\nu}_t = y_t - \overline{y}_t$ is an approximation to the innovations process ν_t and $vecM$ denotes the vector $(m_1^T, ..., m_n^T)$ with m_i the ith column of the matrix M. The direct feed through terms D and E in the output equations involve only u_t and not ν_t from the definition of ν_t. These terms are present only when there is an instantaneous affect of u_t on y_t.

The corresponding approximate nonlinear innovations filter representation is

$$\tilde{y}_t = Hx_t + Dvec(u_t^{[q]} x_t^T) + Eu_t^{[q]} \tag{7-31}$$

$$\tilde{\nu}_t = y_t - \tilde{y}_t \tag{7-32}$$

$$x_{t+1} = Ax_t + Bvec\{(u_t, \tilde{\nu}_t)^{[q]} x_t^T\} + C(u_t, \tilde{\nu}_t)^{[q]} \tag{7-33}$$

The output $\tilde{y}_t$ of the state affine filter is considered as an approximation to the optimal filter output $\hat{y}_t = E(y_t | \overline{p}_t^{[\infty]})$, and the variable $\tilde{\nu}_t = y_t - \tilde{y}_t$ is an approximation to the innovations processes ν_t.

The structure of the stochastic model and filter are also illustrated in Figures 7-1 and 7-2. The right hand sides of equations (7-29) and (7-30) are represented by the state affine functions $g(x, \nu, u)$ and $h(x, u)$, respectively, which are linear in x but may be nonlinear in u and ν. For linear systems, these functions are linear in u and ν as well. The stochastic model and filter have the same structure as in the linear case except that these functions may be nonlinear.

8 Minimal State Rank

In this section, the fundamental problem of minimal realization of a nonlinear system is discussed. This problem is to construct a state space realization where the order of the state vector is minimal. First this topic is explored in terms of the rank of the past/future map using results from differential topology. These results only apply to local mappings whereas what is needed in the case of nonlinear systems is a global mapping involving a state of minimal order.

By restricting consideration to nonlinear systems that are globally observable, the problem can be approached via canonical variate analysis. Results on existence and uniqueness of minimal realizations for deterministic nonlinear systems are reviewed. A modification of the CVA procedure gives a global approach to determining a state of minimal order. Replacing the requirement for orthogonality of the canonical variables with mutual independence avoids any functional or statistical dependence among the canonical variables.

8.1 Local Rank of the State Manifold

In this section, the notion of the local rank of a system is discussed for a *deterministic system*. The notion of the local rank of the state space of a nonlinear system can be defined in terms of the rank of a mapping from past to future. First the various mappings are defined, then the definition of the rank of a function is reviewed.

Define the *input/output* map $f_t = a(p_t, q_t)$ as the future outputs f_t as a function of the past inputs and outputs p_t and future inputs $q_t = (u_{t+1}, u_{t+2}, \ldots)$. Next define the *past/future* map $b : p_t \rightarrow f_t$ as the input/output map with future inputs q_t fixed

$$f_t = b(p_t) = a(p_t, q_t) \tag{8-1}$$

The *Rank of a function* is the maximum of the rank of the derivative maximized over the domain p_t.

$$Rank(b) = \max_{p_t} Rank\left(\frac{\partial b(p_t)}{\partial p_t}\right) \tag{8-2}$$

Now consider a neighborhood of a point and suppose that the past/future map has constant rank in the neighborhood. The construction of such mappings locally is a central issue in differential topology considered in the rank theorem. A differentiable map with a differentiable inverse is called a *diffeomorphism*. A diffeomorphism defined on some open neighborhood U of a given point x is called a *local diffeomorphism* at x. Local diffeomorphisms at x can be regarded simply as invertible nonlinear changes of local coordinates near x.

The rank theorem of differential topology is a generalization of the inverse function theorem that states conditions under which a differentiable map can be transformed locally into a linear map by a smooth change of coordinates of the domain and range variables.

Rank Theorem: Let E and F be finite dimensional vector spaces and let W be an open subset of E that contains the point x_0. Let f be a continuously differentiable map from W to F. Let the derivative of f at a point $s \in W$ be $Df(x)$. If the derivative $Df(x)$ has constant rank for all $x \in W$ then

(a) There exists a set V, an open subset of E, such that $V \subset W$ and another set V^* an open subset of F, such that $f(V) \subset V^*$. There also exist two diffeomorphisms d_1 and d_2 where $d_1 : V \to E$ and $d_2 : V^* \to F$.

(b) The restriction of $f|V$ of the mapping f to the set V is equal to

$$f|V = d_2^{-1} \circ Df(x_0) \circ d_1, \tag{8-3}$$

A consequence of the rank theorem is that for any past/future function $b : p_t \to f_t$ of rank n, there exists a *past/state* map $c : p_t \to x_t$ of rank n and a *state/future* 1-1 nonsingular map $d : x_t \to f_t$ such that for a local neighborhood

$$b = d^{-1} \circ c \tag{8-4}$$

that is, $f_t = b(p_t) = d^{-1}(x_t) = d^{-1}(c(p_t))$. The variables x_t have all of the properties of a state.

The rank theorem gives a local result, that is, for a given point it guarantees the existence of a mapping from the past to a set of variables equal in number to the rank of the past/future function and that agrees with the state in a neighborhood of the given point. Unfortunately, in general this state cannot be extended to a global state. Under additional restrictions, the global minimal realization has the same rank as the local realizations, but not in general.

8.2 Minimal Realization of Deterministic Systems

The fundamental problem of minimal realization of nonlinear systems is to construct a minimal state space realization where the order of the state is minimal. For a general nonlinear system, the order of a minimal state may not equal the rank of the input/output function. This may be the case if the process state is imbedded in a manifold that locally has a lower dimension than it has globally such as is true for a sphere.

The form of (8-4) is closely related to the CVA problem of finding transformations of the past p_t and future f_t such that the transformed variables are maximally correlated. The transformations c and d respectively play analogous roles. To generalize this notion requires a condition on the observability of the state of a system. In particular we are interested in considering the global problem of minimal realization.

The state space realization of a process is *globally observable* if distinct states x_t imply distinct future outputs f_t with the future inputs q_t fixed. If a state space representation is not globally observable, then some states that are distinct in the representation are

not observably distinct in terms of the system outputs. In such a case, the inverse function from the future to the state is not defined as a single valued function. Thus global observability is equivalent to the existence of an inverse function from the future to the state.

The problem of existence and uniqueness of minimal realizations of nonlinear systems has been investigated extensively for deterministic nonlinear systems. Under suitable conditions that the state space system has various analytic properties (Sussmann, 1977; Jakubczyk, 1980; Gauthier and Bornard, 1982), there exists a minimal realization that is globally observable and controllable and is unique up to isomorphism. While such results are of great theoretical value, they give little guidance for construction of minimal realizations in practice.

In the following section, the use of CVA for the determination of the transformations c and d is discussed. This extends to the case of a stochastic process.

8.3 Minimal Rank and Independent CVA

CVA provides a very useful procedure for construction of states for a nonlinear process. Such a state vector however is not of minimal order. Orthogonality is not sufficient to exclude redundancy among the canonical variables. What is required is independence rather than orthogonality. The construction of independent canonical variables is discussed.

For nonlinear processes, the number of canonical variables with nonzero canonical correlations is not equal to the order of the minimal order state as it is for linear processes. The problem is that although a nonlinear function $e(g_1)$ of the first canonical variable g_1 is exactly predictable by a nonlinear function of g_1, $e(g_1)$ may not be orthogonal to g_1. Thus there is considerable redundancy in the canonical variables, i.e. some nonlinear functions of different canonical variables will be highly correlated.

On the other hand the concept of minimal rank in the choice of the state involves functional independence between the different state components. The functional independence is expressed in the linear independence of the rows of the partial derivative matrix of the functions. CVA does not require functional independence of the canonical variables, but only orthogonality.

Suppose that two canonical variables g_1 and g_2 were such that there was no functional redundancy between them in the sense that for any functions e and f, $e(g_1)$ and $f(g_2)$ were uncorrelated. This is equivalent to the statement that the maximal correlation is zero, i.e. $\rho^*(g_1, g_2) = 0$, which from Section 6 is the case if and only if g_1 and g_2 are stochastically independent random variables. Thus we seek a stochastically independent canonical variate analysis (ICVA), i.e. replace the requirement of orthogonality with that of stochastic independence.

This would require that the canonical variables are mutually independent rather than just orthogonal. In particular, in the notation of Section 6, the two sets X and Y

of random variables are independent if and only if $\rho^*(X,Y) = 0$. Thus for an *independent CVA*, in the Sequential Selection Theorem of Section 5 replace the orthogonality condition $< g^{(r-1)}, g_r >= 0$ with the mutual independence condition $\rho^*(g^{(r-1)}, g_r) = 0$

Further work is needed to establish conditions under which such a development will lead to a solution. In particular, some regularity conditions are required so that at each step after the choice of g_i, there exist $M - i$ independent generators so that nonlinear functions of them span the subspace orthogonal to $\mathcal{F}_{g^{(i-1)}}$. Then the canonical variables will be minimal with rank equal to that of the state space.

In the CVA approach developed in the rest of this report, the canonical variables are constructed sequentially by orthogonalization of a set of variables with respect to the canonical variables already constructed. What is needed is an analogous procedure for the construction of a set of random variables that are independent of a given set. The theorem below is such a procedure. In Phase II, the construction of independent canonical variables is proposed for study.

In particular, we wish to show that for any sets X and Y of random variables, there exists a set $Z(X,Y)$ of random variables such that X and Z are mutually independent and span the same space as X and Y. One version of such a theorem is given as follows:

Theorem 8-1: If the density $k(x,y)$ of the joint distribution $P_{X,Y}(x,y)$ with respect to the product $P_X(x)P_Y(y)$ of the marginals exists, is continuous and nonzero, then there exists a transformation $Z(X,Y)$ such that the map: $(X,Y) \rightarrow (X,Z)$ is 1-1 and X and Z are mutually independent.

Proof: From the hypothesis of the theorem, the probability densities exist and are nonzero. Since $p(x,z) = p(z|x)p(x)$ and x and z are independent if and only if $p(x,z) = p(z)p(x)$, it follows that x and z are independent if and only if $p(z|x) = p(z)$, i.e. if and only if the conditional density is equal to the marginal density. Thus we construct a transformation of y to a variable z such that this is true. Let $F_W(w) = P(W < w)$ be the cumulative distribution function for the random variable W. Consider the conditional density $p(Y|x)$ which can be transformed to the uniform random variable u by

$$u(x,y) = F_{Y|x}(y) \tag{8-5}$$

and transformed back to the random variable z with the same density as the marginal $p(y)$ by

$$z(x,y) = F_Y^{-1}(u(x,y)) \tag{8-6}$$

The cumulative distribution of $Z|x$ of Z for fixed x is

$$F_{Z|x}(z) = F_Y(z), \tag{8-7}$$

the marginal density of Y which does not depend upon x. The marginal density of Z is

$$p_Z(z) = \int p_{Z,X}(z,x)dx = \int p_{Z|X}(z|x)p_X(x)dx \tag{8-8}$$

$$= \int p_Y(z)p_X(x)dx = p_Y(z) = p_{Z|x}(z|x) \tag{8-9}$$

By construction, the map: $(X,Y) \rightarrow (X,Z)$ is 1-1 which proves the theorem.

For the construction of independent canonical variables, at each step of the CVA procedure, the subspace independent of the previous canonical variables would need to be constructed. In fact this is given simply by the variables Z above as stated in the following theorem:

Theorem 8-2: The random variables Z independent of X generate the subspace $\mathcal{M}$ of the function space $\mathcal{F}_{(X,Y)}$ that is independent of the variables X, i.e. $\mathcal{M} = \mathcal{F}_Z$.

Proof of the following Lemma is sufficient to prove the theorem.

Lemma 8-3: Let X, Y, and Z be sets of random variables such that Z is independent of X and X,Y is a 1-1 function of X,Z, then any random variables in $\mathcal{F}_{X,Y}$ independent of X is a function only of Z and not X.

Proof: Since the map: $(X,Y) \rightarrow (X,Z)$ is 1-1, any function in $\mathcal{F}_{(X,Y)}$ can be considered as a function of (X,Z). We will show that any function $g(X,Z)$ is a function of only Z. Define the function $\bar{g}(X,Z) = E(g|X)$, i.e. the conditional expectation of g given X, and define

$$\tilde{g} = \begin{cases} 1 \text{ if } g - \bar{g}(X,Z) \neq 0 \\ 0 \text{ otherwise} \end{cases} \tag{8-10}$$

The set of point $S = \{X,Z : \tilde{g}(X,Z) = 1\}$ is the set of points for which $g(X,Z)$ is not equal to the conditional expectation $\bar{g}$, i.e. where $g(X,Z)$ as a function of X is not a constant.

Now suppose that the probability of S is strictly positive, and let S' be the projection of S on X so $S' = \{X : \tilde{g}(X,Z) = 1 \text{ for some } Z\}$. Now since by assumption $P(S) > 0$, we also have $P(S') > 0$. If $P(S') = 1$ then redefine S as a subset such that $P(S') < 1$ and redefine $\tilde{g}(X,Z) = 0$ for X,Z not in S. Define the function of X

$$p(X) = \begin{cases} 1/P(S') \text{ if } X \in S' \\ -1/(1-P(S')) \text{ otherwise} \end{cases} \tag{8-11}$$

so that $Ep(X) = 0$. Also note that $E\bar{g}(X,Z) = Eg(X,Z) - E[E(g|X)] = 0$. Now since S' is the projection of S we have that

$$E[p(X)\tilde{g}(X,Z)] = P(S)/P(S') > 0 \tag{8-12}$$

Since $p(X)$ is strictly a function of X and $\tilde{g}(g)$ is strictly a function of g, this contradicts that the maximal correlation of X and g is zero. Thus the assumption that $P(S) > 0$ is false so that

$$g(X,Z) = E(g|X) \tag{8-13}$$

with probability 1 and hence g is a function only of Z.

Thus the use of conditional and marginal distributions provides a means of constructing independent random variables. From the equivalence of mutual independence and zero maximal correlation, it may be possible to use the orthogonality of all functions

of the two sets to define implicit algorithms such as the ACE algorithm for constructing mutually independent random variables.

A further issue to be addressed in future work is the minimal rank issue. Conditions under which the transformation to mutually independent random variables is a 1-1 transformation need to be established. The minimal rank of the state space will be guaranteed if any transformation to a set of mutually independent random variables has an inverse transformation to the original variables which holds except on a set of probability zero. This condition also enters into the discussion in the next section. These topics are proposed for detailed study in detail in Phase II.

9 Optimal Normalizing Transformations

The canonical variables have a more general optimality property than maximal correlation. It is shown that the canonical variables are the result of transforming the original variables to variables that are closest to normality. The measure of closeness to normality involves an entropy measure. Thus a first order approximation of the distribution of the canonical variables would be the normal distribution.

9.1 Nonlinear CVA of Normalizable Variables

In this section, the independent nonlinear CVA is shown to give the same result on jointly normal random variables as the linear orthogonal CVA. This gives a natural generalization of maximal correlation to the multivariate case when the distribution is normal.

A starting point that suggests a much more basic relationship is the following (Lancaster, 1966).

Theorem 9-1: If X and Y are jointly normal variables, then the maximal correlation occurs for *linear* transformations $g(X)$ and $h(Y)$, and if the maximal correlation is positive then strictly nonlinear transformations will strictly decrease the correlation.

The result does not generalize to the multivariate case for the reasons discussed in the previous section - statistically or functionally dependent variables may be orthogonal in the case of nonlinear transformations of the variables. In this section, the term linear canonical variables and linear CVA will mean the usual CVA considering only linear functions of the random variables. The strongest multivariate result appears to be that given in Lancaster (1966)

Theorem 9-2: Let X and Y be jointly normal random vectors, and let the nonlinear transformations $g_i(X)$ and $h_j(Y)$ be recursively defined so that $E(g_i g_j) = E(h_i h_j) = 0$ for $i < j$. Then g_1 and h_1 have maximal correlation if they are respectively the first pair of linear canonical variables; and if for $i > 1$ we have $\rho_i > \rho_1^2$, then the maximal correlation of g_i and h_i are given respectively by the ith pair of linear canonical variables.

The condition $\rho_i > \rho_1^2$ is sufficient to insure that nonlinear functions that are orthogonal to the previously defined canonical variables will not have large enough correlation. If however the condition of orthogonality is replaced by that of independence or equivalently zero maximal correlation, then the following multivariate generalization is obtained.

Theorem 9-3: Let X and Y be jointly normal random vectors of dimensions k and ℓ respectively, and let the nonlinear transformations $g_j(X)$ and $h_j(Y)$ be recursively defined such that $\rho^*(g_i, g_j) = \rho^*(h_i, h_j) = 0$ for $i < j$. Then the functions g_j and h_j have maximal correlation if they are respectively the jth pair of linear canonical variables c_j and d_j. For $\rho_j > 0$, g_j and h_j are strictly linear functions respectively of $c_1, \ldots, c_k$ and $d_1, \ldots, d_\ell$.

Proof: The theorem is proven by induction. From Theorem 9-1 the first pair g_1, h_1 coincides with the usual first pair of canonical variables. We show that if it is true for $j-1$ then it is also true for j. For $j > 1$, the condition $\rho^*(g_i, g_j) = \rho^*(h_i, h_j) = 0$ for $i < j$ is equivalent to that of independence of g_j and h_j with previously chosen variables g_i and h_i. By hypothesis, for $i < j$ we have $g_i = c_i$ and $h_i = d_i$ so that g_j is independent of c_i and h_j is independent or d_i.

From Lemma 8-3, the above implies that $g_j(C)$ is a function of only $c_j, \ldots, c_k$ and similarly h_j is a function of only $d_j, \ldots, d_\ell$. The maximal correlation of g_j and h_j can be determined by restriction to nonlinear functions of $c_j, ..., c_k$ and $d_j, ..., d_\ell$ respectively. From Theorem 9-1, the maximal correlation of the normal random variables $c_j, ..., c_k$ and $d_j, ..., d_\ell$ is attained by the linear canonical variables c_j and d_j respectively. From Theorem 9-1 it is apparent that for $\rho_j > 0$, g_j and h_j are strictly linear functions respectively of the linear canonical variables $c_1, \ldots, c_k$ and $d_1, \ldots, d_\ell$. This proves the induction from $j-1$ to j, and the theorem is proven.

From this theorem, the canonical variables from a linear CVA on jointly normal variables will also satisfy an independent nonlinear CVA. Also the independent canonical variables with positive canonical correlations have the same uniqueness properties as the linear CVA. However nonlinear canonical variables with zero canonical correlation may be nonlinear functions of the normal variables since only independence is required. For prediction of one set X of variables by another set Y of variables, only the variables with positive canonical correlation are involved. If a set of random variables are *normalizable*, that is if there exists a 1-1 transformation to a set of normal random variables, then the nonlinear canonical variables needed for prediction of X from Y are normally distributed as stated in the following theorem.

Theorem 9-4: Let X and Y be vector random variables such that there exists a 1-1 transformation to jointly normal random vectors. Then the sets of canonical variables $g_1, ..., g_r$ and $h_1, ..., h_r$ respectively with positive canonical correlations $p_1 \geq ... \geq p_r > 0$ are jointly normally distributed, and the optimal prediction of X from Y is expressible in terms of the linear prediction of $h_1, ..., h_r$ from $g_1, ..., g_r$.

Thus in the case of joint normality, of all possible transformations the canonical variables relevant to prediction are normally distributed and the corresponding prediction problem among the canonical variables is linear.

9.2 Mutual Information and Approximation

In this section, a broader interpretation of canonical variables is obtained. This is important since the correlation coefficient as a measure of dependence appears to be very simplistic and to miss much of the complexity of arbitrary distributions.

The previous section addresses the case where there exists a 1-1 transformation to a set of jointly normal random variables. But what about the more general case where there may not exist such a transformation. The correlation coefficient has an interpretation in terms of mutual information and approximating normal distributions.

As noted by Gelfand and Yaglom (1959), the mutual information between two Gaussian random variables c and d is

$$E\{ln(\frac{p(c,d)}{p(c)p(d)})\} = -\frac{1}{2}ln(1-\rho_{cd}^2) \tag{9-1}$$

so that maximizing the correlation is tantamount to maximizing the mutual information.

To generalize this procedure to nonlinear processes requires a number of new concepts and results. The linear case depends completely on the Gaussian assumption which is violated in the nonlinear case. Also, in the nongaussian case the expression for the mutual information no longer holds. However, the basic computational procedure of the CVA method is minimizing the function (9-1) of the correlation coefficient based upon second moments of the variables which is computationally direct and efficient. Thus it is an attractive computational procedure if reasons for its use in nonlinear processes are developed.

Consider arbitrary random variables c and d with some arbitrary joint density function $p(c,d)$. Consider the normal densities $n(c,d)$, $n(c)$, and $n(d)$ using the first and second order moments μ_c, μ_d, and σ_{cc}, σ_{dd}, σ_{cd} respectively of the true distribution of c and d. For simplicity of notation, the means μ_c and μ_d will be assumed to be zero which will not affect the results below. Now consider the expected mutual information of the normal densities with expectation E_p taken with respect to the true density $p(c,d)$

$$\begin{aligned} E_p\{ln(\frac{n(c,d)}{n(c)n(d)})\} &= E_p\{ln(\frac{n(d|c)}{n(d)})\} \\ &= -\frac{1}{2}E_p\{ln2\pi(\sigma_{dd}-\sigma_{dc}\sigma_{cc}^{-1}\sigma_{cd}) \\ &\quad +\frac{(d-\sigma_{dc}\sigma_{cc}^{-1}c)^2}{(\sigma_{dd}-\sigma_{dc}\sigma_{cc}^{-1}\sigma_{cd})} - ln2\pi\sigma_{dd} - \frac{d^2}{\sigma_{dd}}\} \\ &= -\frac{1}{2}ln(1-\frac{\sigma_{cd}^2}{\sigma_{cc}\sigma_{dd}}) = -\frac{1}{2}ln(1-\rho_{cd}^2) \end{aligned} \tag{9-2}$$

Thus the function (9-2) of ρ_{cd} has the interpretation as the average over the true density of the mutual information of the hypothesized normal densities $n(c,d)$, $n(c)$, and $n(d)$ based upon the true covariances. The presence of the true density p in the expectation may at first appear to be a problem since it is often not encountered in information arguments, however in terms of the statistical inference problem it will play the role of the "truth" or data. As with relative or Kullback information (see Kullback, pp. 6, 1959; Larimore, 1983a), consider a relative measure of approximation in terms of the *expected relative mutual information* between the true density $p(c,d)$ and an approximating normal density $n(c,d)$ expressed as

$$\int p(c,d)\{ln\frac{p(c,d)}{p(c)p(d)} - ln\frac{n(c,d)}{n(c)n(d)}\}dcdd \tag{9-3}$$

Now suppose that $c(x)$ and $d(y)$ are transformations of the random variables x and y respectively. The first term in the integrand is the expected mutual information which

is invariant if c and d are 1-1 transformations. Thus (9-3) is a function only of the second term. From (9-2), minimizing the measure (9-3) in the case that c and d are 1-1 is equivalent to maximizing the correlation coefficient ρ_{cd}.

Now we consider the case where the transformations c and d are multivariate transformations of the random vectors x and y respectively. For the first term in (9-3) to be constant, the transformations c and d must be 1-1. This suggests yet another generalization of CVA under the additional constraint that the transformations c and d be 1-1. Suppose that in addition the transformations c and d are required to produce pairs of random variables (c_i, d_i) that are independent of (c_j, d_j) for $i \neq j$. Then the first term of (9-3) would be a constant and the second term would decompose into the sum of terms (9-2) involving each of the pairs of components c_i and d_i of the transformed variables, i.e.

$$E_p\{ln(\frac{n(c,d)}{n(c)n(d)})\} = -\frac{1}{2}\sum_i ln[1-\rho^2(c_i,d_i)] \qquad (9\text{-}4)$$

This is very close to the construction of independent canonical variables as described in Section 8.

9.3 Generality of Independent CVA

In the literature of maximal correlation and nonlinear canonical variables (see Section 6), there appears to be no consideration of the multivariate case were orthogonality is replaced by independence or equivalently zero maximal correlation. From several points of view, this appears to be the natural nonlinear generalization for the multivariate case:

- In the case of the multivariate normal distribution where orthogonality is equivalent to independence, it is independence that generalizes to the nonlinear case not orthogonality (see Theorem 9-3).
- Independence removes the functional and statistical redundancy present in the nonlinear case as discussed in Section 8. This provides the basis for studying the selection of a minimal order state.
- The mutual information measure generalizes to the multivariate case if the canonical variables are independent.

The computational aspects of the independent CVA appears to be significantly greater since it involves the construction of independent random variables rather than orthogonal ones. Further study on this topic is needed as in the proposed Phase II research.

10 Computational Methods

In this section, several implementations of the CVA computations are discussed. The general nature of the computational problems are first described. An implementation using polynomial basis functions is developed and was programmed and applied to simulated data from the Lorenz chaotic attractor in the next section. Computational methods of nonlinear regression in high dimensional spaces are discussed and proposed for implementation in the proposed Phase II research. Such methods are likely to be more statistically efficient and adaptive. For completeness, computations specific to the state affine model are developed.

10.1 Computational Problems

The major computational problems to be solved are the computation of the CVA between past and future and the determination of the state transition and state output functions of the state space innovations model. These problems are discussed in general in this section, and their solutions using polynomial basis functions or adaptive nonlinear methods are described in following sections.

The Hilbert space theory and the theory of maximal correlation provide a very general structure for nonlinear CVA of nonlinear Markov processes. In this context the canonical variables can be determined sequentially as pairs of variables achieving the maximal correlation. For any pair of Borel measurable functions $g(f)$ and $h(p)$ of the past p and future f where both functions have finite variance, the covariance function $\Sigma_{g(f),h(p)}$ is uniquely defined. This sequence of canonical variables is uniquely defined by such a covariance function up to uniqueness of the canonical correlations. Thus if the covariance function was known exactly, we would only need to find a computational procedure for computing such a sequence of canonical variables. Note that there is no requirement that the canonical variables even be continuous functions.

The other major computation required is the fitting of the state space innovations model

$$x_{t+1} = \phi(x_t, u_t, \nu_t) \tag{10-1}$$

$$y_t = \mu_t(x_t, u_t) + \nu_t \tag{10-2}$$

provided by Theorem 10-1 which is a very useful model for a nonlinear Markov process. With the state given as a nonlinear function of the past observations by CVA, the state transition function ϕ and state measurement transformation μ can be determined by nonlinear regression. There are numerous ways of fitting such nonlinear regression models.

The major difficulty in solving the above computational problems is that only a finite sample of observations are available. As a result, the covariance function and second moments of arbitrary nonlinear functions are not exactly known. The major issue is then the approximate determination of the canonical variables and nonlinear regression functions from a finite sample of observations.

For ease of computation, in this Phase I study a polynomial regression procedure was used. In the proposed Phase II research, more adaptive nonlinear regression methods for high dimensional spaces will be investigated.

10.2 Polynomial Basis Functions

The Hilbert space is approximated as being generated by a finite number of polynomial functions in a given number of lags of the past inputs and outputs of the process. The required computation increases rapidly as the number of polynomial terms increases as shown in Table 10-1. The growth is much more rapid in the polynomial degree than in the number of lag variables.

A first issue is the selection of the number of lags and the polynomial degree to use in the CVA. In the state reconstruction method, the state is usually represented as simply some number of past observations. Representation of the Hilbert space as a polynomial in the past inputs and outputs can be viewed as representing the states as polynomials in the past. An upper bound on the number of polynomial terms required is obtained by fitting nonlinear autoregressive process (NAR) models to the process

$$y_t = \sum_{\iota \in I} a_\iota p_t^\iota + n_t \tag{10-3}$$

where the index $\iota = (i_1, \ldots, i_k)$ prescribes the powers of the respective components of the first k elements of the past vector p_t (see (7-25) and (7-26) for notation). The NAR model structure I is the set of indices ι for which the coefficients of the NAR model are nonzero. The set I specifies the model structure of a particular NAR model by the number of lags in the past of p_t and the particular polynomial terms in the model.

The number of lags in the past and the degrees of the polynomial terms that adequately describe the past of the process for prediction of the future can be determined adaptively by subset selection methods of nonlinear regression. For each model structure I up to some degree d and number of lags ℓ in the past, the computation of a measure of fit is desired. The measure of fit could be an entropy type measure such as the Akaike AIC which in many cases reduces to a sum of squares criterion. The problem is that the number of model structures I grows exponentially with k and ℓ. To avoid this problem the Leaps and Bounds algorithm of Furnival and Wilson (1974) can be used. This algorithm searches the tree of model structures starting from the most general including all terms up to some degree d and number of lags ℓ. Most of the branches will be "pruned" very soon because deletion of a product term in the sum (10-3) will give a drastic increase in the prediction error. For nonlinear models, it is also necessary to include all lower order terms of a power term to insure that the model is invariant to linear transformation of the data (Peixoto, 1987). Such a subset selection procedure determines an adequate set of lags of the past and polynomial terms in these variables which contain all of the statistically useful information in the data. This procedure using the AIC trades off the bias in using too low an order NAR model against the additional variability introduced in using too high an order.

The above procedure determines a NAR model (10-3) for optimal or near optimal modeling of the present observations y_t by a sum of monomial terms in the past. What is needed is the information in the past for prediction of the future

$$f_t = (y_t, y_{t+1}, \ldots) \tag{10-4}$$

If only ℓ lags in the past are needed for optimal prediction, then only outputs up to $y_{t+\ell}$ are affected by the past p_t. The prediction of y_{t+j} from the past p_{t+j} is given by

$$\hat{y}_{t+j} = \sum_{\iota \in I} a_\iota p^\iota_{t+j} \tag{10-5}$$

Thus prediction of y_{t+j} as a function of p_t involves only the factors in the terms p^ι_{t+j} of (10-5) that are a function only of p_t, i.e. expressed as

$$p^\iota_{t+j} = g(y_{t+j}, \ldots, y_t) p^\tau_t \tag{10-6}$$

All such terms p^τ_t for $j \leq \ell$ and $\iota \in I$ need to be included as basis functions for the function space $\mathcal{F}_{p_t}$ of nonlinear functions of the past p_t for prediction of the future f_t.

The above procedure is very useful when only prediction of the future f_t is of interest rather than the more general case of nonlinear function of the future f_t. In any case, we suppose that a basis for the nonlinear past has been chosen or determined. Similarly, suppose that a basis for the future has been chosen. In the simulation example, the future itself is chosen. Then a CVA is done between the functions of the past and the functions of the future. This is simply a linear CVA involving the nonlinear functions of the past and future. First the nonlinear functions of the past and future are evaluated for each time t and treated as the data in a linear CVA. The canonical variables are expressed as linear combinations of the nonlinear functions of the past and future.

The canonical variables of the past are the candidate state of the nonlinear Markov process. The canonical correlations can be inspected to determine if there is an obvious choice of model state order. An example of this is given in the simulation example in Section 11. For each selection of model state order, the state transition function ϕ (7-16) and state output function μ (7-17) are fitted by polynomial regression. Here again an adaptive polynomial regression procedure would be very useful. The polynomial regressions for different degree polynomial terms can be fitted by a Leaps and Bounds algorithm. The resulting model gives a state space innovations model for the nonlinear Markov process.

10.3 Adaptive Nonlinear Methods

In this subsection, the use of adaptive nonlinear computational methods is discussed. These methods directly address issues of fitting nonlinear models in high dimensional spaces for observational data. Computational algorithms and software for adaptive nonlinear procedures were not implemented in this Phase I study, but have considerable potential and are proposed for detailed study and implementation in the Phase II research.

While nonlinear polynomial regression is useful in low dimensional spaces or where the structure of the functions is low order polynomial in form, for general functions in high dimensional spaces the results are generally poor. There has been considerable development of adaptive nonlinear methods in the past decade. These methods have produced spectacular improvements over traditional nonadaptive methods. The theoretical basis for some of these methods is very closely related to the theory of maximal correlation and nonlinear CVA on Hilbert spaces. Some of these methods can be directly applied to the CVA problem, while others will require some further research and development for successful extension to CVA of nonlinear Markov processes. In this section, various methods and approaches of adaptive nonlinear regression in high dimensional spaces are discussed in the context of the nonlinear CVA problem, and areas of particular promise for the extension to the nonlinear CVA problem are indicated.

The problem considered by adaptive nonlinear methods is the determination of nonlinear stochastic models for finite, i.e. incomplete, samples of a random process. While virtually all of the literature concentrates on static problems as opposed to dynamic systems involving time evolution, the same fundamental issues must also be addressed in the time series case. In fact the introduction of the time variable only increases the severity of the problem.

The fundamental problem stated simply is that in high dimensional spaces the number of parameters required for an accurate model grow exponentially with the space dimension whereas the available sample size in most practical problems does not increase proportionately. As a result the error in the model due to statistical variability increases dramatically and often is so large as to render the model useless in practice. The remedies developed in the past decade are to adaptively determine the parameters that are really necessary in the model and to exclude the others so that errors in their estimation do not corrupt the model.

Two approaches that have particular potential for nonlinear CVA are additive modeling with adaptive regression splines and multivariate adaptive regression splines (MARS). An additive model has the form

$$y = f(x_1, \ldots, x_n) = \sum_{i=1}^{n} f_i(x_i) \tag{10-7}$$

where the i-th function $f_i(x_i)$ is a function of the single variable x_i. This is a generalization of linear functions which permits nonlinearity in a single variables but not nonlinear functions jointly of two or more variables. The additive models have been studied much more than other methods of nonlinear adaptive regression and much more theory is available.

For additive models, adaptive regression splines have been very successful in adaptively determining spline models for the functions $f_i(x_i)$. The basic approach is to adaptively fit splines for each of the univariate functions. For each of the functions $f_i(x_i)$, knot locations are adaptively added and deleted until a near optimal cubic spline is determined. The criterion for goodness of fit is based upon a generalized cross validation (GCV) measure.

The functions $z_i = f_i(x_i)$ can be considered as nonlinear transformations of the variables x_i respectively that best produces a linear model in the transformed variables z_i in a least squares sense. More generally the problem can be considered of simultaneously transforming the variable y by $g(y)$ so that the function

$$E[g(y) - \sum_{i=1}^{n} f_i(x_i)]^2 / var g(y) \tag{10-8}$$

is minimized. That is we seek to find optimal transformations $[g^*(y), \{f_i^*(x_i)\}_1^n]$ that minimize the fraction of the variance $varg(y)$ of $g(y)$.

The alternating conditional expectation (ACE) algorithm (Breiman and Friedman, 1985) is used to determine the functions $[g^*(y), \{f_i^*(x_i)\}_1^n]$. ACE effectively does the minimization by alternating between the two minimization problems:

- Given $g(y)$, find functions $\{f_i(x_i)\}_1^n$ minimizing (10-8)
- Given functions $\{f_i(x_i)\}_1^n$, find $g(y)$ minimizing (10-8)

In actual computation, only a sample are available so that the expectation of (10-8) is replaced by the sample average. The optimization is done using adaptive regression splines to obtain a smooth function.

The theory of the ACE algorithm is closely related to the theory of maximal correlation (Breiman and Friedman, 1985). Let $X = (x_1, \ldots, x_n)$ and $Y = y$ and consider the function space $\mathcal{A}_X$ of *additive* functions of X of the form (10-7) and the function space $\mathcal{F}_Y$. Then the problem of finding the optimal transformations minimizing (10-8) is equivalent to the maximal correlation problem

$$\max_{g \in \mathcal{F}_Y, f \in \mathcal{A}_X} \rho(f(X), g(Y)) \tag{10-9}$$

This is a special case of the maximal correlation problem where Y is a single variable and the function $f(X)$ is restricted to be of additive form in the variables X.

The ACE algorithm works in general for finding the maximal correlation no matter what function spaces are involved. Thus it can be used in general for finding optimal transformations for CVA by sequentially maximizing the correlation. The difficult part of the problem is in replacing the expectation with the sample average. The problem of too many parameters and too few data require the use of adaptive methods that are extended to more general nonlinear models than additive models, e.g. including interactions between the variables involving simultaneous functions of several variables that are not additive. Such generalizations have been developed in the MARS algorithms.

The inclusion of higher order interactions in the functional form requires a fundamentally different approach than that of additive models. The n-dimensional space involving the independent variables X is recursively partitioned into pieces specifying knots of a spline function on the space. If the function changes rapidly in a region, then

the algorithm will partition the space more finely in the region. The same strategy is used in determining the inclusion of higher order interactions of the variables.

The above methods have the elements of adaptive nonlinear regression needed for solving the CVA problem. The ACE algorithm is an implementation involving alternating between the projection operators (6-4) and (6-5). This will work in the general CVA context for finding the function maximizing the correlation. Only an orthonormalization is required at each step before finding the next pair of canonical variables using the ACE algorithm. In the case of independent CVA, a computational procedure is required to construct random variables independent of the preceding canonical variables. Further study of this case is needed.

10.4 Computation of Affine Models from Covariances

In the originally proposed research, the state affine model was central to the development. Subsequent Phase I research resulted in very general procedures discussed in the subsections above. In this subsection, the modeling of state affine models is discussed in particular.

The state affine representation provides an economical representation as compared to other methods. Note that the state affine representation involves only powers and products among the state and inputs at a *fixed* time. An alternative approach is the use of nonlinear difference equations (Billings and Leontaritis, 1982; Diaz and Desrochers, 1987) which equate the output at y_t to a polynomial in the past involving powers and products among the components at *many different* times. For multi-input multi-output systems, the state affine form accounts for a considerable economy in the use of parameters for the representation of nonlinear processes and a corresponding improvement in statistical accuracy of the identified process. In Diaz and Desrochers (1987), a model is identified in difference equation form and then converted to state affine form which is of much simpler form.

In terms of the formulation of CVA analysis of the previous section, the state or memory m_t of the system of order k is chosen from linear combinations of $p_t^{[r]}$, the powers and products of the past of degree no greater than r, as given by

$$m_t = J_k p_t^{[r]} \tag{10-10}$$

where J_k is rank k. The degree of r is chosen sufficiently large so that no information for predicting f_t is lost. CVA gives the k statistically significant linear combinations m_t of $p_t^{[r]}$ for linear prediction of the future f_t. Thus CVA is computed on the covariance matrices among $p_t^{[r]}$ and the future f_t.

From the affine Markov innovation representation (7-29) and (7-30), the resulting state space model will have the form

$$m_{t+1} = Am_t + Bvec\{(u_t, \nu_t)^{[q]} m_t^T\} + C(u_t, \nu_t)^{[q]} \tag{10-11}$$

$$y = Hm_t + Dvec(u_t^{[q]} m_t^T) + Eu_t^{[q]} + \nu_t \tag{10-12}$$

where ν_t is an uncorrelated increments process and $vec(\cdot)$ is defined in (7-29). Following the methods outlined in Larimore (1983b), the matrices A, B, C, D, E, and H can be determined by linear regression in the power product terms in (10-11) and (10-12). Error in the regression gives an estimate of the covariance R_t of the innovations process ν_t, and higher order moments can be estimated. For nonlinear processes, the innovations covariance R_t is not necessarily stationary since the noise processes may be involved in the process in a nonlinear manner in the feedback of y_t in the state equation. A regression of the innovations covariance R_t on magnitudes of the terms $\{(u_t, \nu_t)^{[q]} m_t^T\}$ would provide a useful parameterization.

For the determination of model state order, recent developments in the selection of model order and structure based upon entropy or information can be used. Such methods were originally developed by Akaike (1973, 1974b) and involve use of the Akaike Information Criterion (AIC) for deciding the appropriate order of a statistical model. The AIC for each state order k is defined by

$$AIC(k) = -2 \log p(Y^N, U^N; \hat{\theta}_k) + 2M_k \tag{10-13}$$

where p is the likelihood function, based on the observations (Y^N, U^N) at N time points, with the maximum likelihood parameter estimates $\hat{\theta}_k$ using a k-order model with M_k parameters. The model order k is chosen with the minimum value of AIC(k). A predictive inference justification of the use of an information or entropy based criterion such as AIC is given in Larimore (1983a) based upon the fundamental principles of sufficiency and repeated sampling. In choosing model order, the risks of introducing bias into the model in choosing too low an order must be weighed against introducing additional variability in choosing too high an order (Larimore and Mehra, 1985).

The number of parameters M_k in the state space model (22) and (23) is determined by the general state space canonical form as in Candy et al (1979) as

$$M_k = 2kn + km + nl + n(n+1)/2 \tag{10-14}$$

where k and n are the dimensions of the state and output vectors m_t and y_t respectively as in the case of linear systems; however for the nonlinear case the number of inputs m becomes $(k+1)Dim\{(u_t, \nu_t)^{[q]}\}$, the total number of elements in the vectors $vec\{(u_t, \nu_t)^{[q]} m_t^T\}$ and $(u_t, \nu_t)^{[q]}$. If the direct feed through term is present, then the term nl is present where l is $(k+1)Dim(u_t^{[q]})$. If higher order moments of ν_t are estimated, then the term $n(n+1)/2$ is modified appropriately to reflect the number of parameters estimated.

The interpretation of the criterion is different since the process is no longer gaussian. The AIC measure is still a quadratic measure of the multi-step prediction error in terms of the measured second moment of the variable being predicted. The usual expression for the AIC assumes that the innovations noise v_t is distributed as a gaussian process. More exact computation of the AIC is possible using higher order moments of the innovations process ν_t. An AIC optimal model can be determined for the set of models indexed by (k, q), where k and q are the state order and the polynomial degree respectively.

The stochastic filtering problem is solved efficiently by approximation using CVA as follows. The covariance matrices among f_t and $p_t^{[r]}$ are calculated for a finite number of lags in f_t and p_t using any method that is available such as propagating moments using a difference or state equation model of the process. These covariance matrices are then used in the CVA. For any given order of approximation, the one-step prediction error for the corresponding state affine CVA model is computed to determine the desired state order k and polynomial degree q depending on the desired filtering accuracy.

Number of Variables	maximum degree				
	1	2	3	4	5
1	2	3	4	5	6
2	3	6	10	15	21
3	4	10	20	35	56
4	5	15	35	70	126
5	6	21	56	126	252
6	7	28	84	210	462
7	8	36	120	330	792
8	9	45	165	495	1287
9	10	55	220	715	2002
10	11	66	286	1001	3003
11	12	78	364	1365	4368
12	13	91	455	1820	6188
13	14	105	560	2380	8568
14	15	120	680	3060	11628
15	16	136	816	3876	15504
16	17	153	969	4845	20349
17	18	171	1140	5985	26334
18	19	190	1330	7315	33649
19	20	210	1540	8855	42504
20	21	231	1771	10626	53130
21	22	253	2024	12650	65780
22	23	276	2300	14950	80730
23	24	300	2600	17550	98280
24	25	325	2925	20475	118755
25	26	351	3276	23751	142506
26	27	378	3654	27405	169911
27	28	406	4060	31465	201376
28	29	435	4495	35960	237336
29	30	465	4960	40920	278256
30	31	496	5456	46376	324632

Table 10-1: Number of Polynomial Terms

11 Simulation Results

In this section, the CVA procedure is demonstrated on the Lorenz attractor with process excitation noise. The full 3-dimensional dynamics of the Lorenz system is obtained by observing only one component of the 3-dimensional state. The canonical variables are computed and displayed, and a state space innovations model identified. The adequacy of the canonical variables are demonstrated by transforming to the original state variables of the Lorenz attractor and comparing the estimated states with the true states. Finally the state space equations governing the canonical states are determined by nonlinear regression and used to simulate trajectories for the identified system. These simulated trajectories have a similar character to that of the true process demonstrating the feasibility of using CVA for accurate identification of nonlinear state space models for multiple equilibria nonlinear Markov processes.

11.1 Lorenz Attractor

The Markov process considered is the Lorenz attractor (Lorenz, 1963) with process excitation noise. The state equations of the Lorenz attractor are of the form

$$\begin{aligned} \dot{x}^{(1)} &= \sigma\left(x^{(2)} - x^{(1)}\right) \\ \dot{x}^{(2)} &= \rho x^{(1)} - x^{(2)} - x^{(1)}x^{(3)} \\ \dot{x}^{(3)} &= -\beta x^{(3)} + x^{(1)}x^{(2)} \end{aligned} \tag{11-1}$$

The values of the parameters used in the simulation are $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$ which results in the much studied chaos of the system. The differential equations are discretized with $\Delta t = 0.01$, and white process noise is added to the state equations so that the discrete time equations used for simulation become

$$x_{t+1}^{(1)} = x_t^{(1)} + \Delta t\sigma\left(x_t^{(2)} - x_t^{(1)}\right) + n_t^{(1)} \tag{11-2}$$

$$x_{t+1}^{(2)} = x_t^{(2)} + \Delta t\left[\rho x_t^{(1)} - x_t^{(2)} - x_t^{(1)}x_t^{(3)}\right] + n_t^{(2)} \tag{11-3}$$

$$x_{t+1}^{(3)} = x_t^{(3)} - \Delta t\left[\beta x_t^{(3)} + x_t^{(1)}x_t^{(2)}\right] + n_t^{(3)} \tag{11-4}$$

The noise covariance matrix of the white process excitation noise $(n_t^{(1)}, n_t^{(2)}, n_t^{(3)})^T$ used in the simulation is $10^{-4} \times I_3$ with I_3 the 3 dimensional identity matrix. The presence of process excitation noise provides a much more difficult identification problem since the process no longer is exactly predictable given exact arithmetic. Most studies of identification of chaos consider only the presence of additive white noise which can be reduced by simple averaging of the observations. The time correlation introduced by the nonlinear process dynamics presents a much more difficult problem for identification.

For system identification, the measurement observation data is $y_t = x_t^{(1)}$, the first component of the discretized Lorenz process, which is shown in Figure 11-1. The sections below show that the entire 3-dimensional dynamics of the process can be reconstructed from the measured first component.

11.2 Canonical Variate Analysis

The measurements y consisting of only the first component $x^{(1)}$ of the Lorenz attractor are used to compute nonlinear functions of the past as basis functions for canonical variate analysis. The past p_t consists of functions that are powers and products of up to degree three in the first three lags $(y_{t-1}, y_{t-2}, y_{t-3})$ of the measurements y so that functions of the form

$$f_{i_1,i_2,i_3}(y_{t-1}, y_{t-2}, y_{t-3}) = y_{t-1}^{i_1} y_{t-2}^{i_2} y_{t-3}^{i_3} \quad \text{for } i_1 + i_2 + i_3 \leq 3 \tag{11-5}$$

are considered. There are 20 such basis functions. The future f_t is the vector of outputs up to 20 lags into the futu: e so

$$f_t = (f_t, \ldots, f_{t+20})^T \tag{11-6}$$

A canonical variate analysis of sample covariances of the past and future is given in Table 11-1. Note that the canonical correlations drop until a floor is hit at 0.078, and from this point on the canonical correlations fall off slowly. This is typical behavior of sample canonical correlations and most likely the canonical correlations less than or equal to 0.078 are not statistically significant. This suggests that there are 7 significant canonical variables. In the discussion below, the *canonical state* is chosen as the first 5 canonical variables to reduce the computation since the contribution of the two additional canonical variables is small.

In terms of the chaos literature on state reconstruction, the canonical states provide a set of variables for embedding the process. The CVA provides a systematic procedure for selecting the embedding variables so that nothing is missed. The canonical variables have the property that they are orthogonal so that the information in the different variables is uncorrelated and the predictive improvement in the future is additive. Also from Section 5, the canonical variables are optimal in the sense that for any k the first k canonical variables provide the best prediction of the future of any possible choice.

11.3 State Reconstruction of the Lorenz Attractor

While the canonical states have optimal properties for embedding the dynamics of the observations, they in general do not directly relate to the states of a process that may be of interest. In the present case, it is important to assess the ability of the canonical states c to predict the full 3-dimensional motion of the Lorenz attractor state x.

To obtain an estimate of the original state x of the Lorenz attractor, an approximate nonlinear transformation is constructed from the canonical states c to x. The transformation $g(c)$ is constructed by polynomial regression on polynomials in the states c up to degree 6. The estimated state $\hat{x}$ in the original 3-dimensional coordinates of the Lorenz attractor is

$$\hat{x} = g(c) \tag{11-7}$$

Phase plane plots of pairs of the components of the Lorenz attractor state x are shown in Figure 11-2 for original "true" states (solid line) and the reconstructed state estimates $\hat{x}$ (dashed line). Note that components $x^{(1)}$ and $x^{(2)}$ are estimated very accurately in that the reconstructed trajectories based upon CVA provided by the regression of the canonical states on the Lorenz states is very accurate. The estimate of the component $x^{(3)}$ is much noisier, but on the average provides a good estimate of the true value of $x^{(3)}$. The exception is for values of $x^{(3)}$ below about 20 that appear to have been distorted. There are several reasons for this distortion:

- From equation (11-2), the variable $x^{(3)}$ is not directly observed by $y_t = x_t^{(1)}$ but only indirectly through $x_{t-1}^{(2)}$.
- From equation (11-3), the variable $x^{(3)}$ is related to $x^{(2)}$ through the term $x_{t-1}^{(1)} x_{t-1}^{(3)}$ so the effect of $x^{(3)}$ is proportional to the magnitude of $x^{(1)}$.

The distortion in Figure 11-2 is seen to be most severe for $x^{(1)}$ and $x^{(3)}$ small and for $x^{(2)}$ and $x^{(3)}$ small. Either of these conditions will cause poor identifiability in those regions of the trajectory from the discussion above. More precise state reconstruction will require more data. Another alternative is to measure more components of the process. Apart form this, however, the state estimate provides a very faithful reconstruction of the process although it is somewhat noisy. The noisy estimate is to be expected since only one of the components of the process was used for measurement data.

The accurate state reconstruction demonstrates that the canonical variables contain a great deal of the state information in the measurements. In theory, CVA provides an automatic selection of variables for embedding the process that will loose the least information using a given number of canonical variables. The ability of the canonical variables to reconstruct the full 3-dimensional dynamics of the Lorenz attractor based upon measuring only one component is illustrated very graphically in Figure 11-2.

To demonstrate the CVA method on the Lorenz attractor the choice of basis functions for the CVA was somewhat arbitrary. Several improvements are possible to avoid the arbitrary choice and to increase the accuracy. The accuracy includes bias, which is systematic error resulting in an inadequate number of basis functions, and variability in the sample. Stepwise regression methods can be employed to determine a sufficient number of basis functions, and that there are no significant improvements with the addition of more functions.

Another approach to improved accuracy in the selection of the canonical variables is the direct construction of the nonlinear functions that define the transformation to canonical variables. This approach has been developed in the literature of nonlinear regression in high dimensional spaces and is solved in particular by the alternating conditional expectation (ACE) algorithm (see Brieman and Friedman, 1985). The ACE algorithm constructs a spline function for the transformation that adaptively determines the knot locations so as to minimize the statistical error of the solution. This approach appears to be one of the most promising and is proposed in Phase II.

11.4 State Space Model Identification

In the above paragraphs, the canonical variables are shown to determine an adequate state for the process. In following paragraphs, a state space model is identified by nonlinear regression using the canonical state.

Theorem 7-1 in Section 7 provides a state space innovations model of a Markov process. If the state for the process is given, then it is only necessary to determine the nonlinear output function μ, innovations process ν_t, and the state transition function ϕ by nonlinear regression.

The function μ is considered as a 6th degree polynomial in the canonical state c_t. The error is the innovation process ν_t. Then the transition function ϕ is considered as a 6th degree polynomial in the canonical state and innovations. The polynomial coefficients are found by nonlinear regression of the canonical state c_t one step ahead on power product terms in c_t and ν_t.

To investigate the dynamics of the resulting identified state space model, data was simulated using the model with no process excitation noise. Phase plane plots of pairs of the components of the canonical state variables c_t are shown in Figure 11-3 for the reconstructed states (dashed line) and for states simulated from the identified model (solid lines). Note that the trajectories constructed from the observed data have the same dynamical character as those simulated from the identified model. Only the first three components are shown here since they contain nearly all of the energy.

The two trajectories are interleaved until the jump to the second equilibrium where the simulated trajectory becomes a stable limit cycle. It is known (Manneville and Pomeau, 1980) for the Lorenz attractor (11-1), that a change in the value of the ρ parameter can result in a change from chaotic motion to a stable limit cycle. Thus a stable limit cycle is qualitatively consistent with the Lorenz attractor (11-1). Some inaccuracy of the identified model is to be expected since the data length is very short relative to the total number of parameters estimated in the model fitting. Each component of the state transition function is a polynomial in 84 terms so that for the 5 canonical states c there are a total of 420 parameters as compared with 1000 observations.

The phase plane trajectories shown in Figure 11-3 of the canonical states determined by CVA are very similar to the principal components obtained by Broomhead and King (1986) for the same Lorenz attractor model. This is not surprising since principal components analysis is a special case of CVA. In the present case, CVA weighting of the future is a sum of squares of the prediction error for each future lag which is closely related to the principal component weighting of the sum of squares of regression error of the past outputs. Also, apparently the nonlinear terms do not play a large role for the Lorenz attractor in terms of the embedding coordinates, i.e. the states can be chosen reasonably efficiently using linear functions of the past. The CVA approach however is much more general in this respect and can be expected to perform much better in cases requiring nonlinear embedding to obtain efficient state realizations.

11.5 High Noise Case

To demonstrate the performance of the CVA method with much greater noise, the Lorenz process (11-1) was simulated with the process noise covariance matrix for $(n_x, n_y, n_z)^T$ as $10^{-2} \times I_3$, i.e. with the process noise variance larger by two order of magnitude. The same procedure for computational analysis of the data was use as in the case of lower noise described above.

The first component of the state $x_t^{(1)}$ used for the measurement in the system identification is shown in Figure 11-4. Note that the presence of the noise on the process is very noticeable unlike the case of lower noise in Figure 11-1.

A canonical variate analysis of sample covariances of the past and future are given in Table 11-2. Note that the canonical correlations drop until a floor is hit at 0.1782, and from this point on the canonical correlations fall off slowly. This is typical behavior of sample canonical correlations and most likely the canonical correlations less than or equal to 0.1782 are not statistically significant. This suggests that there are 5 statistically significant canonical variables. In the discussion below, the canonical state is chosen as the first 5 canonical variables. Note that, as compared with the low noise case, there are two fewer statistically significant canonical correlations due to the higher noise.

Phase plane plots of pairs of the components of the Lorenz attractor state x are shown in Figure 11-5 for original "true" state (solid line) and the reconstructed state estimate $\hat{x}$ (dashed line). Note that the estimated state $\hat{x}$ has the same character as in the low noise case but the estimation error is greater. The true state (solid line) has clearly discernable perturbations due to the process noise which was not the situation in the low noise case. By observing the behavior on sections of the trajectories that are nearly perpendicular to an axis, it is apparent how variable each of the reconstructed states are. From such an observation, the errors in the reconstructed states increase in the order $x^{(1)}$, $x^{(2)}$, $x^{(3)}$.

For illustration, phase plane plots of pairs of all five of the canonical states are plotted in Figure 11-6. Note that the canonical states are much more noisy than for the case of lower noise in Figure 11-3. The phase plane plots in Figure 11-6 confirm the canonical correlation analysis that the canonical states have successively less and less information. A state space model was identified and used to simulate phase plane trajectories. Because of the high noise, the identified model had much less of the character of the Lorenz model and settled to a single stability point. Due to the high noise, much more data would be needed to obtain an accurately identified model with phase plane trajectories of the same character as the true model.

The results of applying CVA to the high noise case confirm that CVA can indeed reconstruct good estimates of the Lorenz process states. These estimates are much more noisy, but still contain the information for the three dimensional Lorenz process. The first three canonical states for the high noise case in Figure 11-6 have the same dynamical character as for the low noise case of Figure 11-3.

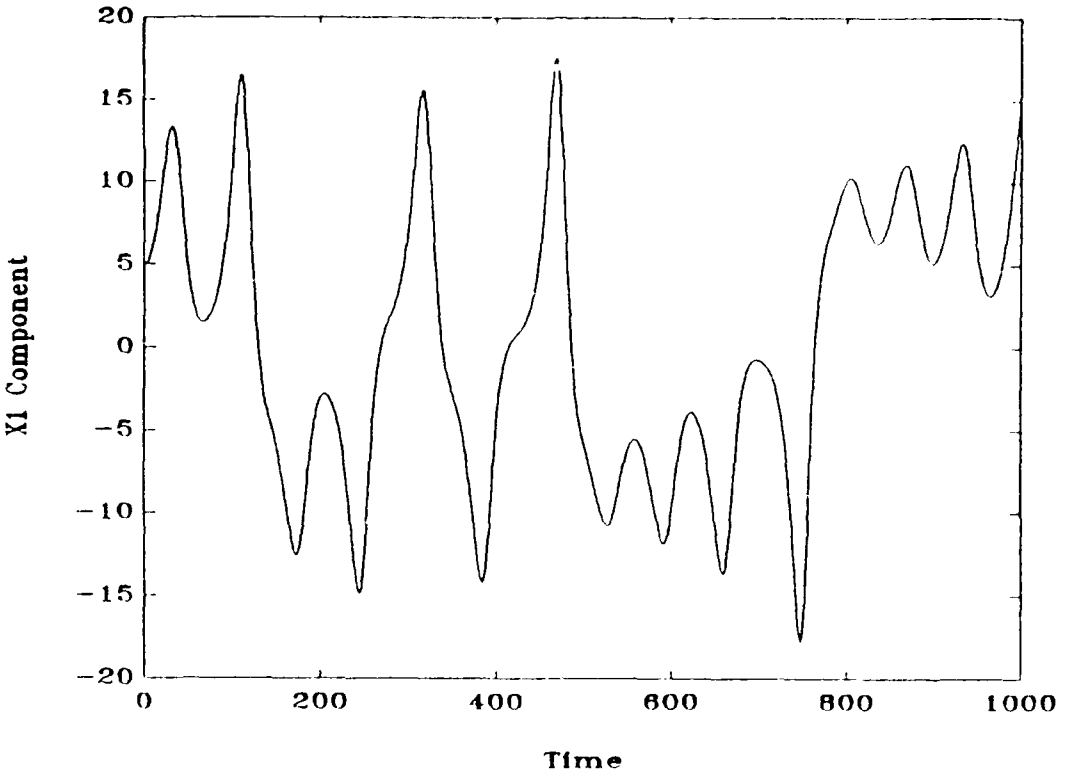
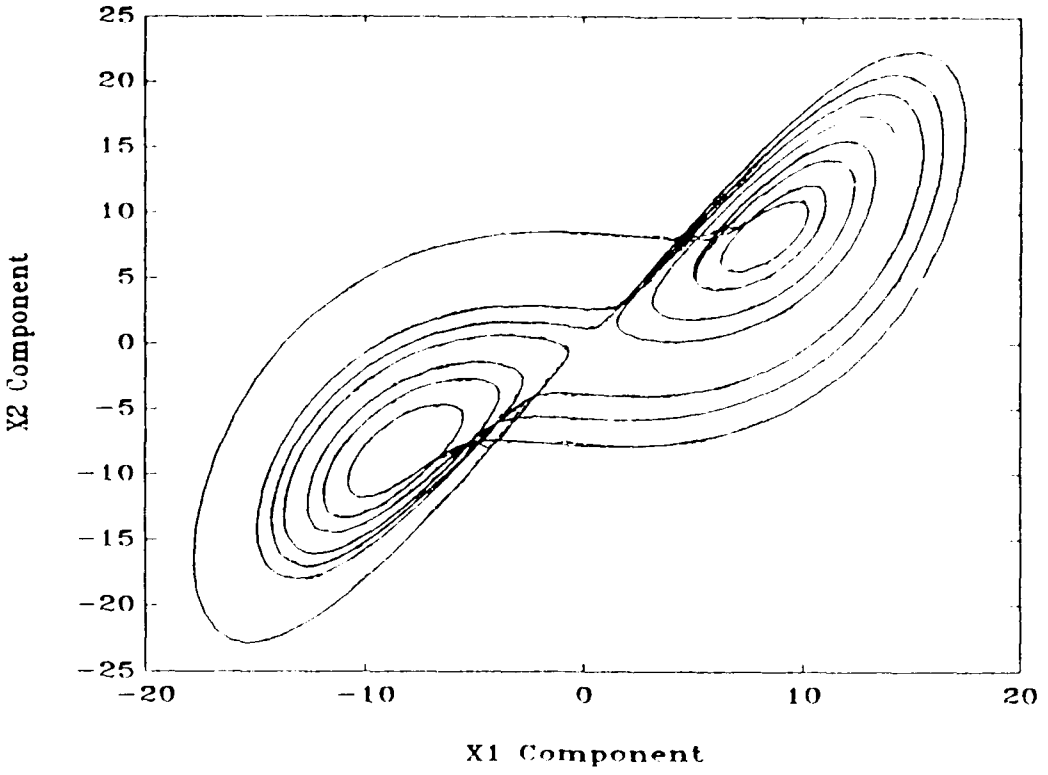


Figure 11-1: Data used in State Reconstruction and Model Identification

Index	Canonical Correlation
1	0.999998
2	0.999463
3	0.989650
4	0.959988
5	0.712619
6	0.372002
7	0.260169
8	0.078201
9	0.070837
10	0.057570
11	0.030183
12	0.027724
13	0.022471
14	0.016929
15	0.002708
16	0.002070
17	0.000632
18	0.000338

Table 11-1: Canonical Correlations



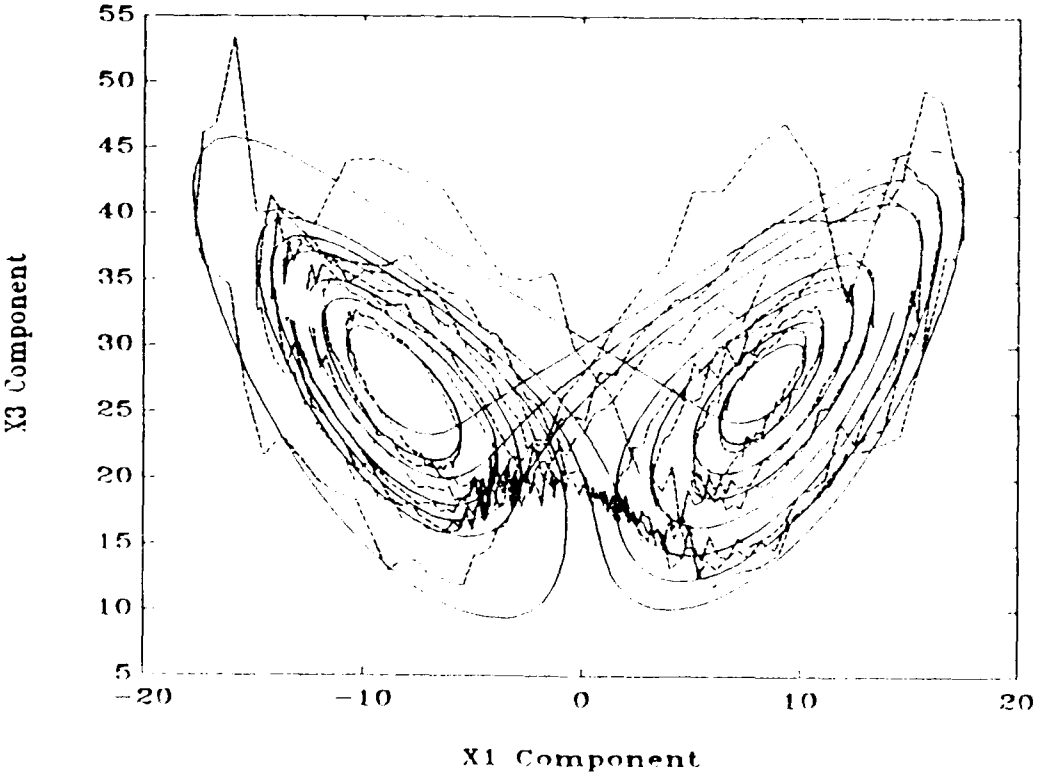


Figure 11-2: Phase Space of True vs. Reconstructed Process

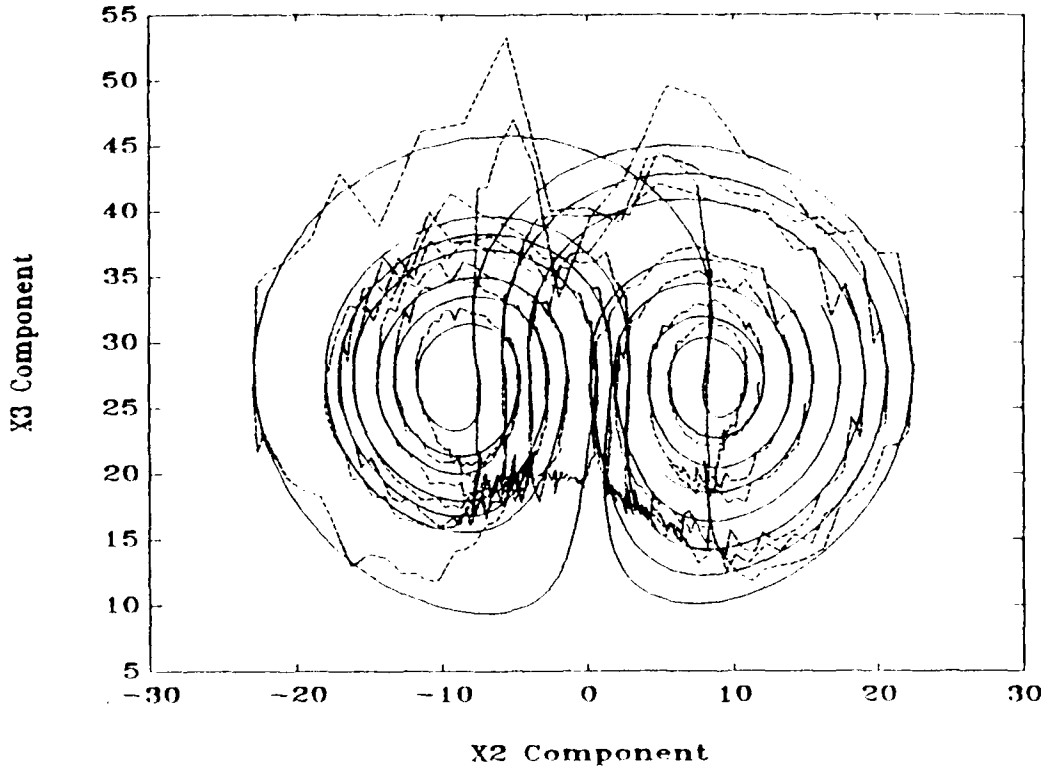


Figure 11-2: Phase Space of True vs. Reconstructed Process (continued)

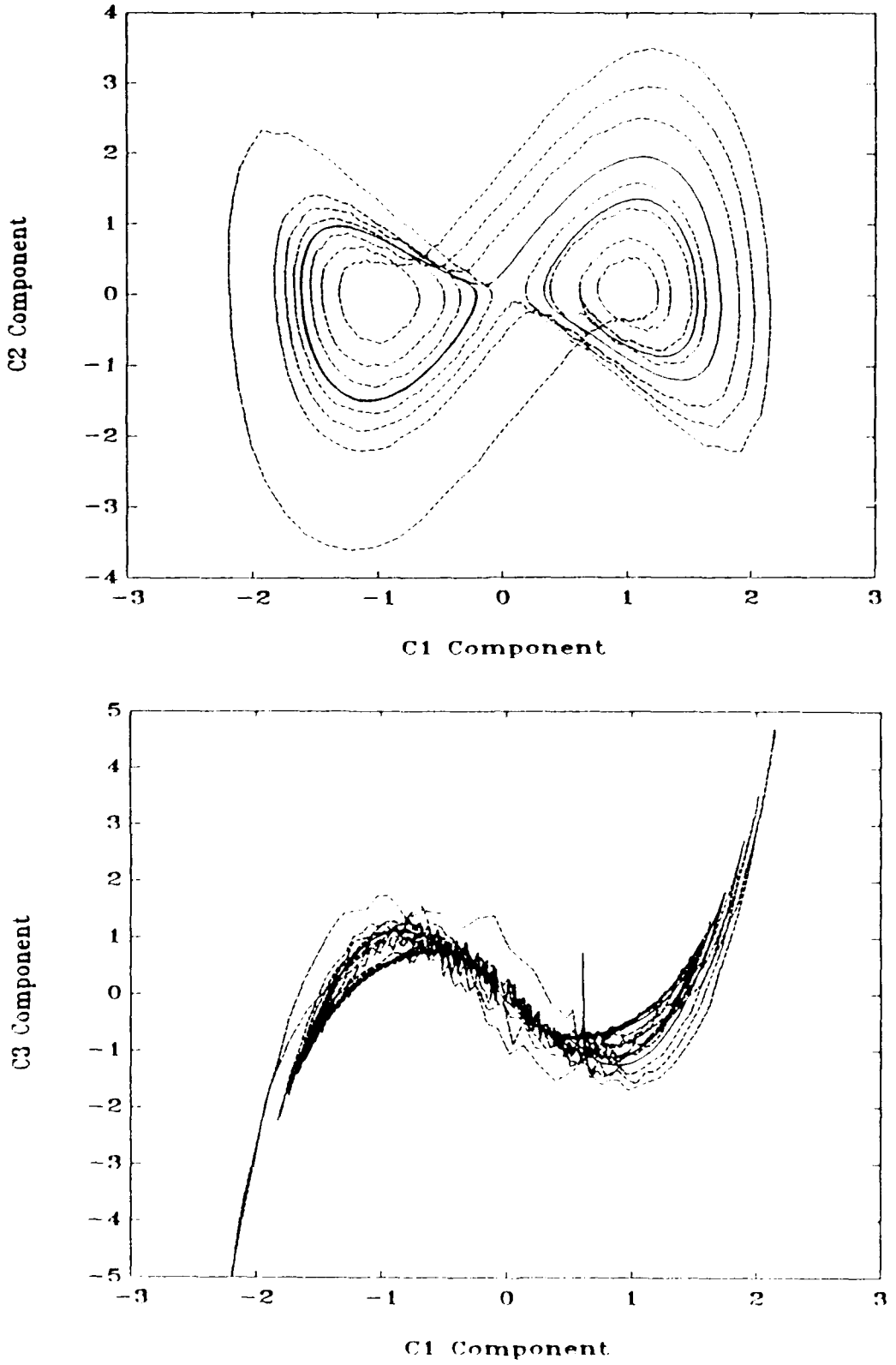


Figure 11-3: Phase Space of Observed vs. Identified Process

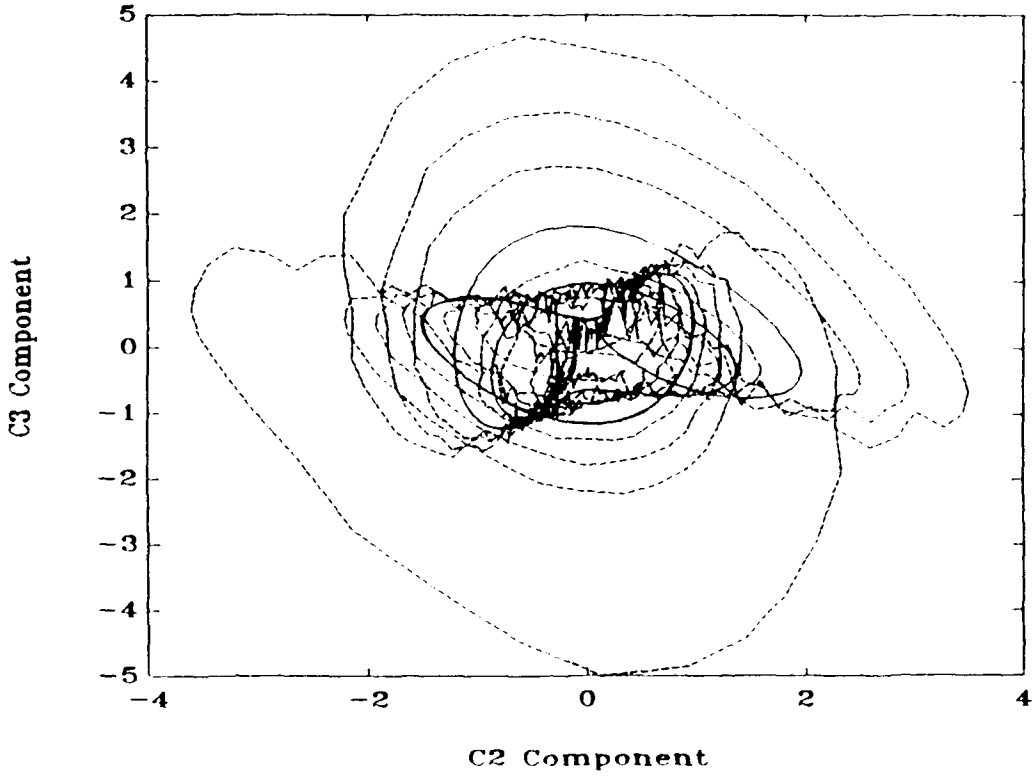


Figure 11-3: Phase Space of Observed vs. Identified Process (continued)

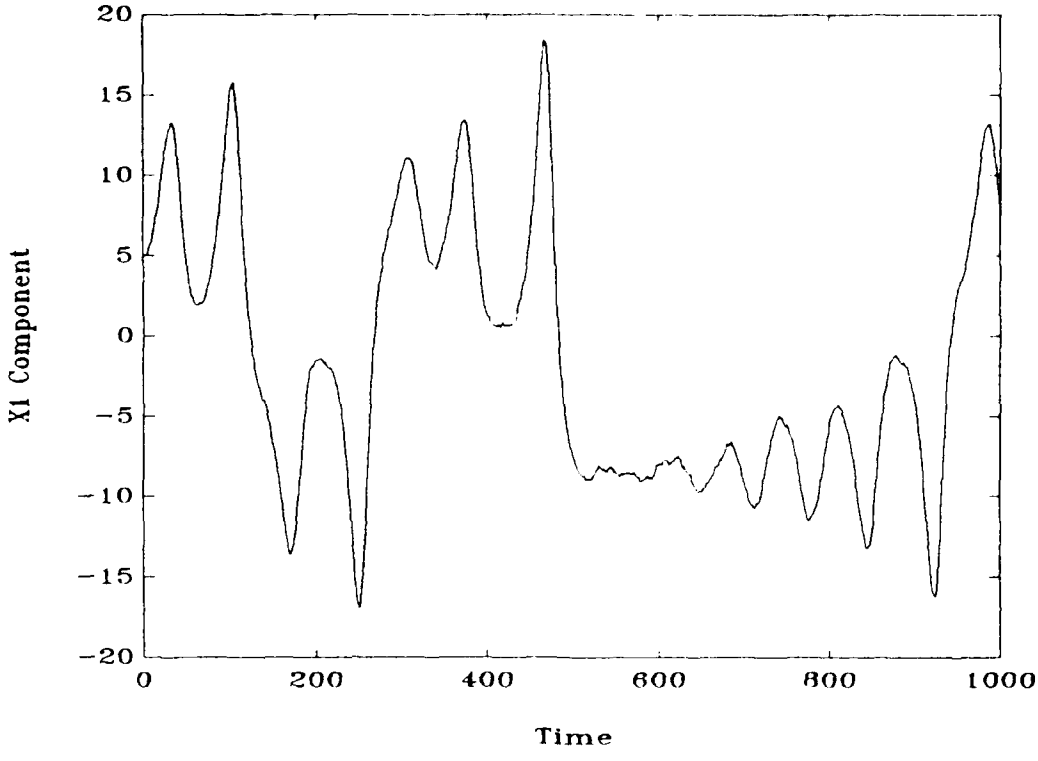


Figure 11-4: Data Used in State Reconstruction

Index	Canonical Correlation
1	0.9999
2	0.9746
3	0.9043
4	0.6062
5	0.3022
6	0.1782
7	0.1626
8	0.1539
9	0.1309
10	0.0969
11	0.0940
12	0.0827
13	0.0686
14	0.0581
15	0.0461
16	0.0149
17	0.0102
18	0.0041
19	0.0011

Table 11-2: Canonical Correlations for High Noise Case

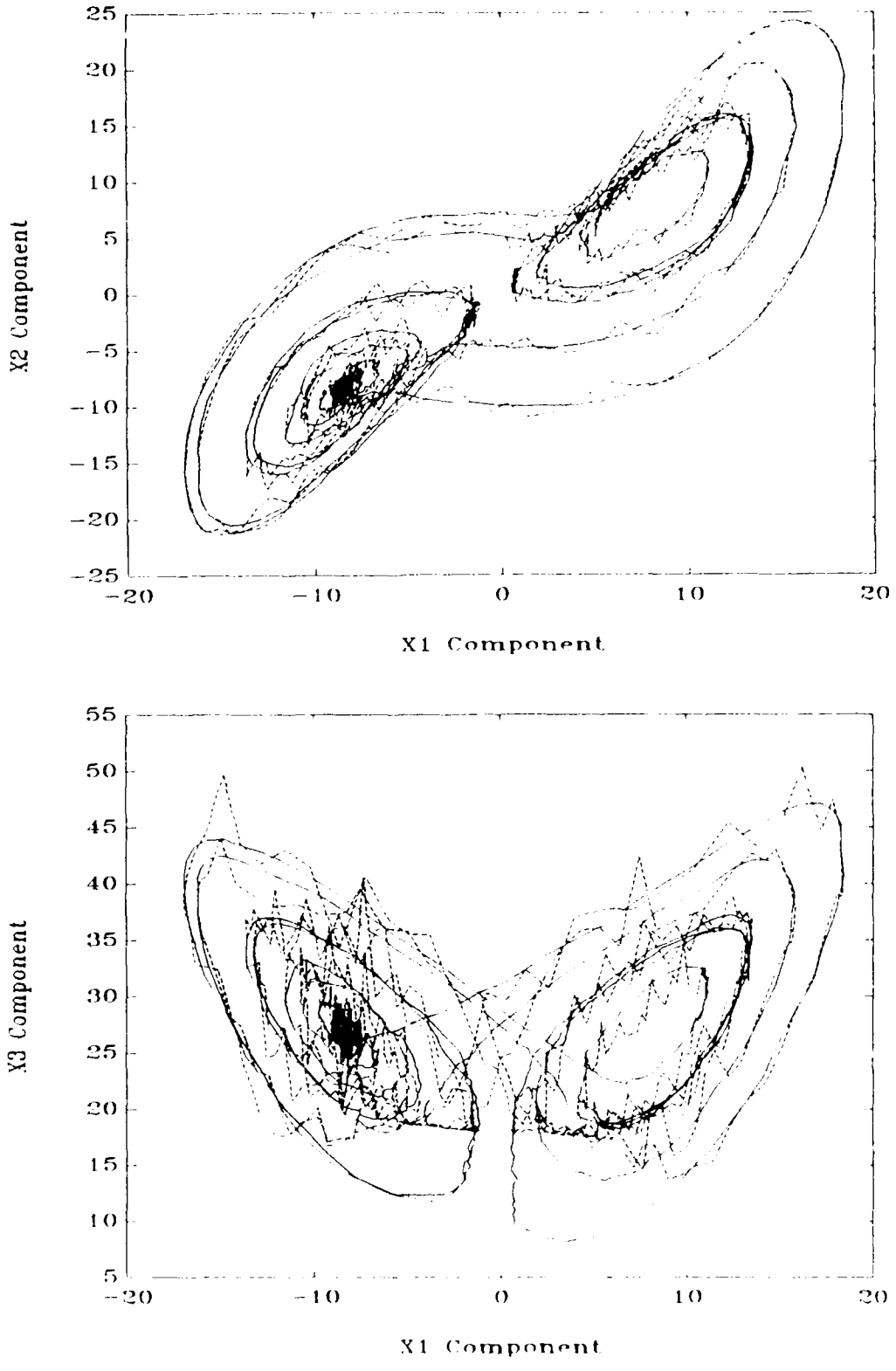


Figure 11-5: Phase Space of True vs. Reconstructed Process

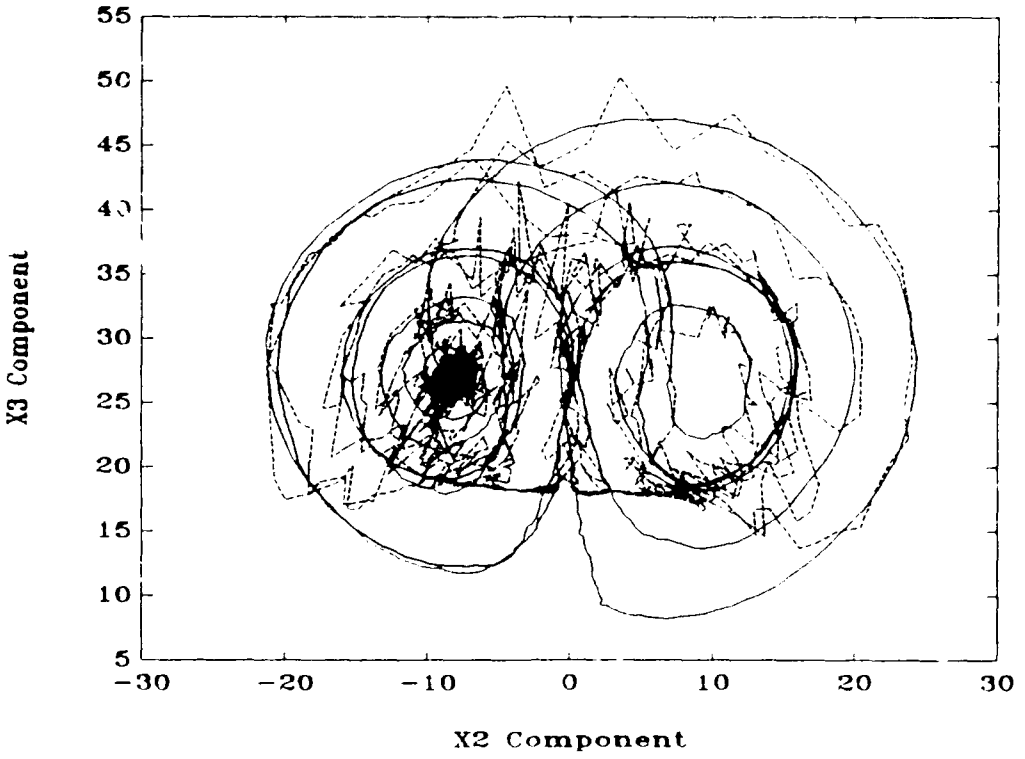


Figure 11-5: Phase Space of True vs. Reconstructed Process (continued)

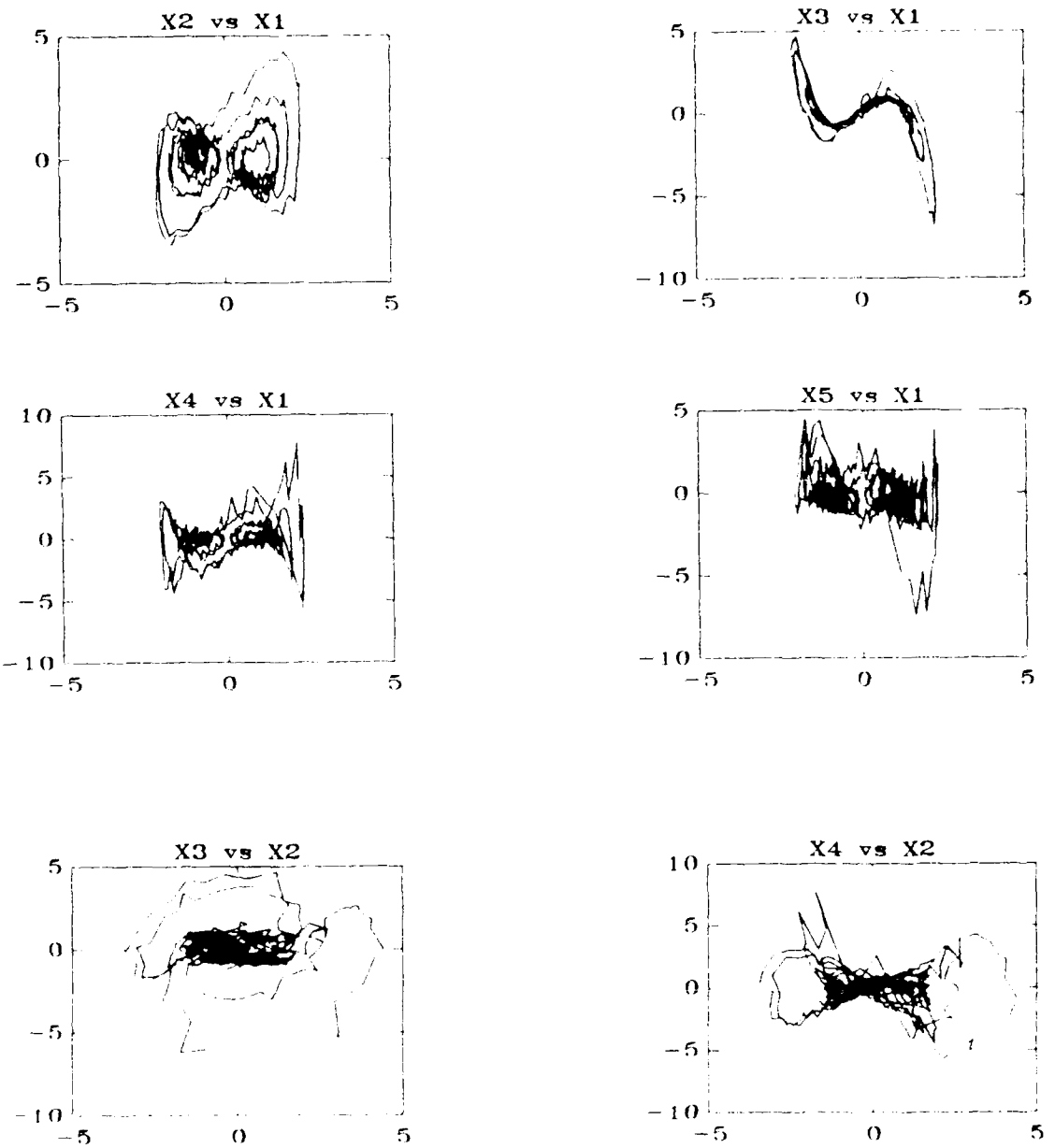


Figure 11-6: Phase Space of Canonical Variable States

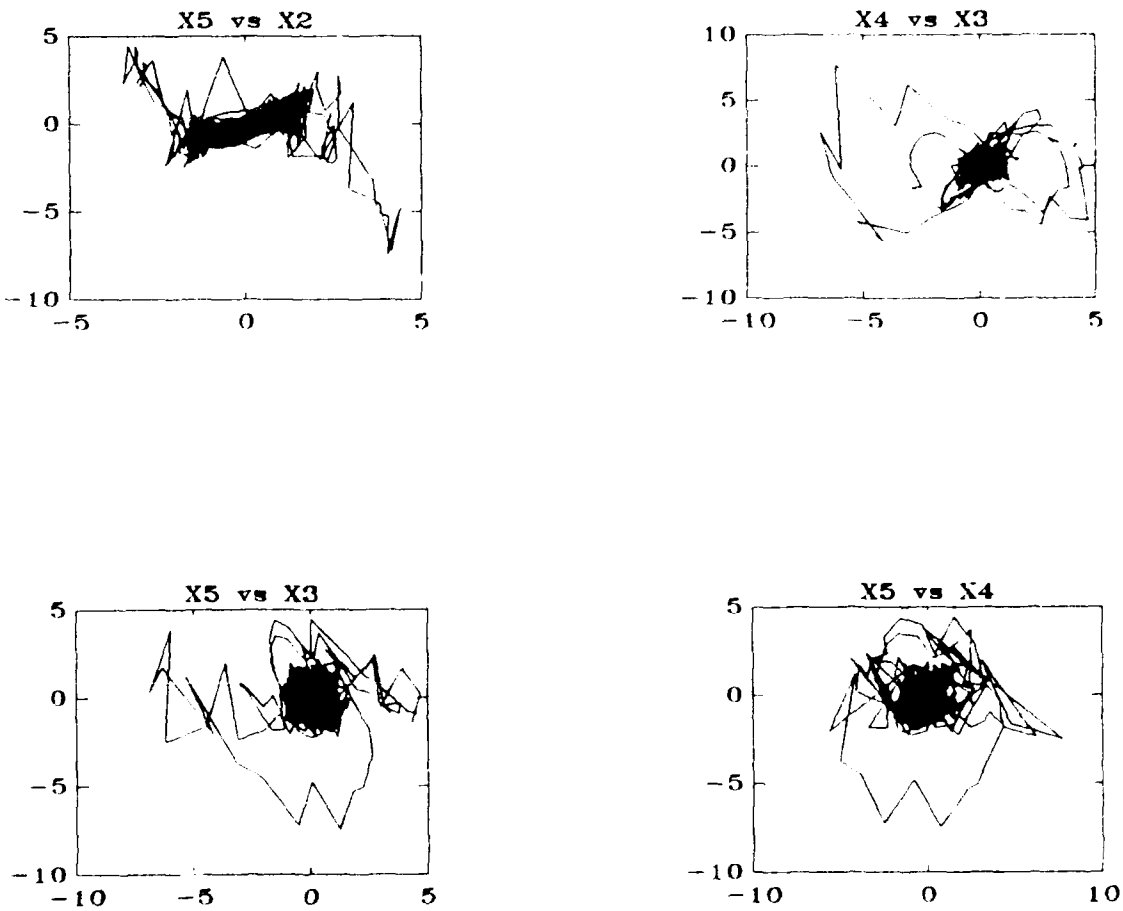


Figure 11-6: Phase Space of Canonical Variable States (continued)

12 Technical Results and Conclusions

In this section, the results and conclusions of the Phase I study are discussed.

12.1 Literature Review

In the course of this study, a number of related fields are investigated to determine if the methods and concepts are relevant to the topic and approach of this study. A number of valuable relationships are found. These concern primarily the topics of nonlinear stochastic filtering, nonlinear time series analysis, the state space reconstruction method, and nonlinear regression in high dimensions.

The literature of nonlinear stochastic filtering has investigated in detail the problem of nonlinear filtering using finite dimensional state space filters. This problem is also central to the identification of finite dimensional state space models for nonlinear Markov processes. The literature has a sizable number of papers that come to the following conclusions:

- Except for special cases there do not exist finite dimensional filters for optimal estimation and prediction.
- In general it requires an infinite number of states to optimally filter or predict.
- For the case of a simple cubic nonlinearity, the addition of additive white noise results in the optimal nonlinear filter being infinite dimensional.
- The one case where finite dimensional filters do exist is for affine systems. However, such systems do not allow multiple equilibria which considerably restricts the type of nonlinear behavior that can be considered.
- Substantial work has been done on minimal order realization of nonlinear deterministic systems which is primarily concerned with existence proofs.

Since the publication of these somewhat negative results about a decade ago, there has been little activity on the topic of finite dimensional approximation of nonlinear stochastic systems.

Nonlinear time series analysis and nonlinear system identification has had some success with the use of nonlinear difference equations, primarily of polynomial form.

- Nonlinear difference equations of AR, ARMA and state space forms have been fitted to observational data with some success.
- The problem of identifiability and numerical conditioning present difficulties in addition to those present in the linear case.
- The main results concern higher order approximation of the dynamics in a neighborhood of a single equilibrium point.

- Much of the problem of nonlinear time series analysis involves the determination of parametrically efficient nonlinear structures for modeling which is closely related to nonlinear regression in high dimensions.

State space reconstruction methods have been successful in fitting state space models from deterministic nonlinear systems including chaotic systems with multiple equilibria.

- If the state order of a deterministic system is order r, then an embedding of the state space requires at most $2r + 1$ lags of the outputs.
- For deterministic systems with small additive white noise, good fits have been reported for large samples.
- From the nonlinear stochastic filtering literature, it is known that for additive white measurement noise, the optimal filter is no longer finite dimensional.
- The problem of process noise that excites the nonlinear system dynamics appears not to have been addressed
- More optimal nonlinear embeddings and formulation of the selection of the state space as an approximation problem for the stochastic problem have not been addressed.

The spirit of the state reconstruction method is close to the CVA approach. Much of this study can be viewed as a means of extending the state reconstruction method to nonlinear stochastic processes.

Nonlinear regression in high dimensional spaces has been a field of considerable activity in the last decade. There have been a number of notable success in devising methods for adaptively fitting nonlinear models that are static, i.e. with no dynamical aspect.

- The major problem is that the number of possible parameters grows much faster than the sample size so that parametrically efficient methods are required.
- Methods for the adaptive selection of the model structure have been developed that are much more efficient parametrically than nonadaptive procedures.
- Recursive partitioning methods have been used to determine locally the regions where more parameters are required for modeling the function
- The alternating conditional expectation (ACE) algorithm for fitting nonlinear functions in a semiparametric or adaptive nonparametric way involves projection operators of the theory of maximal correlation central to CVA

The methods of nonlinear regression in high dimensional spaces has much to offer the further development and computational implementation of CVA of nonlinear processes and is proposed for Phase II research.

In addition to the above topics, the literature of maximal correlation was studied in depth and used in the development of a general theory of CVA on Hilbert spaces. This theory includes processes much more general than the state affine processes originally proposed for study. The above connections between CVA of nonlinear processes and the related literature provide a number of very useful ideas that are investigated in this Phase I study or are proposed for further research in Phase II.

12.2 Theory of CVA for Nonlinear Markov Processes

A general theory of CVA for nonlinear Markov processes is developed. This includes the development of nonlinear CVA on a Hilbert space of nonlinear functions involving the theory of maximal correlation between two nonlinear functions. The Markov structure of the problem is studied, and a number of Markov structures are discussed including the state space innovations representation. Minimal rank of the state is investigated and found to require the condition of independence in place of orthogonality. The relationship with normalizing transformations is discussed.

In the development of the a general theory for CVA:

- The CVA problem is formulated as a nonlinear prediction problem on a Hilbert space of nonlinear functions.
- The solution is reduced to a sequential selection problem requiring solution of a *maximal correlation problem at each step.*
- The solution for a given order D contains the solutions for any lower order r which is obtained by restriction to the first r canonical variables.
- The theory of maximal correlation gives the optimal nonlinear functions in terms of projection operators with the existence of the solution based upon the theory of operators on Hilbert spaces.
- The projection operators are precisely those involved in the alternating conditional expectation (ACE) algorithm for nonlinear regression in high dimensions.

The theory provides a successive approximation procedure for prediction depending on the number of canonical variables to be used. The procedure is rigorously developed on the Hilbert space of nonlinear functions, and computation is related to the ACE algorithm.

Nonlinear controlled Markov processes were developed:

- The approach of deterministic systems using a finite number of lags of the past does not work for nonlinear stochastic processes which are infinite dimensional except in special cases.

- The selection of a finite number of nonlinear polynomial functions of the past as a basis for approximation of a nonlinear Markov process is related to fitting nonlinear autoregressive (NAR) processes.
- The problem of optimal embedding by choosing states as nonlinear transformations of the past is precisely the CVA problem. Minimal embedding requires use of the independent CVA.
- The existence of a state space innovations representation is proved where the state transition and state output transformations can be determined by regression.
- The special case of a state affine Markov process where the state transition and state output transformations are affine in structure is developed.

The relationship between minimal rank of the state, independence, and approximate normality are developed:

- A local approach to rank of the state does not appear to be useful for fitting global nonlinear models.
- Global observability guarantees that unique states produce unique sequences of future outputs of the process.
- As a result of nonlinearities, the canonical variables generally do not provide a minimal rank state for a nonlinear Markov process due to redundancies present in spite of the orthogonality of canonical variables.
- Replacing orthogonality of canonical variables by independence of canonical variables rules out deterministic or stochastic redundancies.
- A procedure analogous to orthonormalization is developed to transform one set of random variables so that the resulting random variables are independent of a given set of random variables.

An independent CVA, i.e. CVA under the constraint of mutual independence of the canonical variables replacing orthogonality, is shown to optimally transform variables to approximate normality:

- In the case of the multivariate normal distribution where orthogonality is equivalent to independence, it is independence that generalizes to the nonlinear case not orthogonality.
- The mutual information measure generalizes to the multivariate case if the canonical variables are independent.

In the literature of maximal correlation and nonlinear canonical variables there appears to be no consideration of the multivariate case were orthogonality is replaced by independence or equivalently zero maximal correlation. From several points of view, this appears to be the natural nonlinear generalization for the multivariate case.

12.3 Computational Methods and Simulation Results

Computational methods are developed and applied to simulated data of a nonlinear Markov process to demonstrate the feasibility of system identification and stochastic filtering. The major computational problem is the computation of the nonlinear CVA to determine state variables and the computation of the nonlinear regressions to obtain the state transition and the state output transformations. The approach of polynomial basis functions is implemented because of the simplicity of the method. The more statistically accurate and adaptive method of nonlinear regression in high dimensional spaces is discussed and proposed for detailed in Phase II. Data are simulated using the Lorenz attractor chaotic system. Canonical variate analysis gives canonical variables for the process states which are compared with the true states. The identified state space model is used to simulate trajectories of the process and compared with the behavior of the true process trajectories.

Computational algorithms are developed and implemented based upon the polynomial basis function approach:

- To determine the nonlinear polynomial functions to use for representing the past, algorithms for fitting nonlinear autoregressive (NAR) models are developed.
- For data analysis, a fixed NAR structure is used
- Nonlinear CVA is done by implementing linear CVA on the nonlinear functions of the NAR model
- The state transition and state output functions are computed by polynomial regression.

The use of methods for nonlinear regression in high dimensions is discussed for study in the proposed Phase II research:

- The ACE algorithm provides a means of computing functions with maximal correlation in a semiparametric way.
- The recursive partitioning method can be used to recursively fit splines so that refinement of the algorithm occurs only locally where it is needed.
- These nonlinear regression methods are expected to be much more adaptive and to achieve higher accuracy through greater parametric efficiency.

Computer simulation of observational data is used for system identification and stochastic filtering:

- The Lorenz chaotic attractor which has two equilibrium points provides an excellent test for the nonlinear CVA method.

- Process noise is included in the simulation of the Lorenz dynamics which introduces correlated noise in the measurements and considerably complicates the identification problem.
- System identification and stochastic filtering is done using only one component of the 3-dimensional Lorenz process.

Nonlinear CVA is computed:

- The past is represented as polynomial basis functions and the future is taken as future observations.
- The canonical variables exhibit a noise floor typical of the case for linear CVA.

The Lorenz states are constructed from the canonical states:

- A transformation from the canonical states to the Lorenz states is determined by nonlinear regression.
- The first two canonical variables have nonlinear dynamics qualitatively very similar to the first two states of the Lorenz attractor, but higher canonical variables are related to the Lorenz states in a vary nonlinear way.
- There is distortion of the dynamics of $x^{(3)}$ for small values possibly due to the marginal observability of $x^{(3)}$.
- The use of the more adaptive methods of nonlinear regression in high dimensions is expected to significantly reduce the distortion.

State space innovations models are fitted for the canonical states:

- The state transition function ϕ and state output function μ are determined by polynomial regression.
- The identified state space model is used to simulate trajectories that are very similar to the trajectories of the truth model.

12.4 Comparison with Other Methods

In this section, the CVA approach is compared primarily with the state reconstruction method which has raised a number of important issues and provided some valuable insights.

The CVA approach directly addresses the issue of determination of a state for a nonlinear process directly from the observational data without making assumptions about the functional form of the nonlinear dynamics. From a review of the literature, such an approach appears to be taken only by the state reconstruction method and applies only to deterministic systems with little or no observation noise. For such

deterministic systems including chaotic systems, the state reconstruction method has been very successful. However, there appear to be a number of difficulties to the direct extension of the state reconstruction method to nonlinear Markov processes:

- Nonlinear Markov processes involve infinite dimensional filter states in all but very special cases.
- The use of lagged past observations is probably quite inefficient for representing the past of some nonlinear processes.
- Large inefficiencies will result if the statistically significant states are not determined as in the CVA procedure.
- Computation of the state transition function would require the use of statistical regression methods.

In comparison, the CVA approach address a number of issues that have been recently raised in the state reconstruction literature:

- How can optimal embeddings, i.e. nonlinear functions of the past,

  be chosen to represent the past of a stochastic process?
- Given such an embedding, how can a low dimensional state and associated transition function be chosen to approximate the memory and dynamics of the process?
- What are appropriate measures of approximation of a stochastic system?

The CVA approach gives a general solution to these problems:

- An optimal nonlinear embedding of a chosen order is obtained by CVA.
- The optimal selection of the states of a given dimension are given by the canonical variables.
- The measure of approximation is given by the relative mutual information which gives a measure of error between the mutual information for the true and the approximating normal distribution for the canonical variables.
- The canonical variable transformations are optimal normalizing transformations.

The theory for CVA is developed rigorously on Hilbert spaces of nonlinear functions and the projection operators provide powerful computational methods for implementation. The related fields of nonlinear stochastic filtering and nonlinear time series analysis do not appear to have directly addressed these issues.

The approach of CVA to system identification and stochastic filtering appears to provide a unique and powerful approach to the solution in both theory and computation of these difficult problems. In the proposed Phase II research, these problems will be developed in detail and the success of the resulting methods will be fully evaluated.

13 REFERENCES

Akaike, H. (1976). "Canonical Correlation Analysis of Time Series and the Use of an Information Criterion," *System Identification: Advances and Case Studies*, R.K. Mehra and D.G. Lainiotis, eds., New York: Academic Press, pp. 27-96.

Akaike, H. (1975), "Markovian Representation of Stochastic Processes by Canonical Variables," *SIAM J. Control.*, Vol. 13, pp. 162-173

Akaike, H. (1974a), "Stochastic Theory of Minimal Realization," *IEEE Trans. Automatic Control*, Vol. 19, pp. 667-674.

Akaike, H. (1974b), "A New Look at Statistical Model Identification," *IEEE Trans. Automatic Control*, Vol. 19, pp. 667-674.

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," *2nd International Symposium on Information Theory*, Eds. B.N. Petrov and F. Csaki, pp. 267-281. Budapest: Akademiai Kiado.

Anderson, T.W. (1958), *An Introduction to Multivariate Statistical Analysis*, New York: Wiley.

Billings, S.A., and I.J. Leontaritis (1982). "Parameter Estimation Techniques for Nonlinear Systems," *Identification and System Parameter Estimation*, Sixth IFAC Symposium, G.A. Bekey Ed., Vol. 1, pp. 427-32. Washington, D.C.: McGregor & Warner.

Box, G.E.P. and G.M. Jenkins (1976), *Time Series Analysis Forecasting and Control*, San Francisco: Holden-Day

Brent, R., and F. Luk (1985), "The Solution of Singular-Value and Symmetric Eigenvalue Problems on Multiprocessor Arrays," *SIAM J. Scientific and Statistical Computing*, Vol. 6, pp. 69-84.

Brieman, L., and Friedman, J.H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," *J. of the Amer. Stat. Assoc.*, Vol. 80, pp. 580-5597.

Broomhead, D.S. and G.P. King (1987), "Extracting Qualitative dynamics from Experimental Data," *Physica*, Vol. 20D, pp. 217.

Buja, A. T. Hastie and R. Tibshirani (1989), "Linear Smoothers and Additive Models", *Annals of Statistics*, Vol. 17, No. 2. pp. 453-555.

Candy, J.V., Bullock, T.E., and Warren, M.E. (1979), "Invariant Description of the Stochastic Realization," *Automatica,* Vol. 15, pp. 493-5.

Chen, S. and S.A. Billings (1988), "Recursive Maximum Likelihood Identification of a Nonlinear Output-Affine Model", *Int. J. Control*, Vol. 48, No. 4, pp. 1605-1629.

Chiang, H.D., M.W. Hirsch, and F.F. Wu (1988), "Stability Regions of Nonlinear Autonomous Dynamical Systems," *Trans. Automatic Control*, Vol. 33, pp. 16-27.

Crutchfield, J.P., J.D. Farmer, and B.A. Huberman (1982), "Fluctuations and Simple Chaotic Dynamics", *Physics Report*, Vol. 92, pp. 45.

Crutchfield, J P. and B.S. McNamara (1987), "Equations of Motion from a Data Series", in *Complex Systems*, Vol. 1, pp. 417-452.

Csaki, P. and J. Fischer (1963), "On the General Notion of Maximal Correlation", *Magyar Tud. Akad. Mat. Kutato Int. Kozl.*, Vol. 8, pp. 27-51.

Csaki, P. and J. Fischer (1960), "Contributions to the Problem of Maximal Correlation", *Publ. Math. Inst. Hung. Acad. Sci.*, vol. 5, pp. 325-337.

Diaz, H. (1986). "Modeling of Nonlinear Systems from Input-Output Data," Ph.D. Thesis, Electrical, Computer, and Systems Engineering Department, Rensselaer Polytechnic Institute, Troy, NY.

Diaz, H., and A.A. Desrochers (1987). "Modeling of Nonlinear Discrete Time Systems From Input-Output Data," *Proc. 10th World IFAC Congress*, held July 27-31, Munich.

Farmer, J.D. and J. Sidorowich (1987), "Predicting Chaotic Time Series", LANL preprint LA-UR-87-1502.

Farmer, J.D., E. Ott, and J.A. Yorke (1983), "The Dimension of Chaotic Attractors", *Physica*, Vol. 7D, pp. 153.

Farmer, J.D., J.P. Crutchfield, H. Froehling, N.H. Packard and R.S. Shaw (1980), "Power Spectra and Mixing Properties of Strange Attractors", *Annals of the New York Academy of Sciences*, Vol. 357, pp. 453.

Fliess, M. and D. Normand-Cyrot (1982). "On the Approximation of Nonlinear Systems by Some Simple State-Space Models," *Identification and System Parameter Estimation*, Sixth IFAC Symposium, G.A. Bekey Ed., Vol. 1, pp. 433-6. Washington, D.C.: McGregor & Warner.

Fraser, A.M. and H.L. Swinney (1986), "Independent Coordinates for Strange Attractors from Mutual Information," *Physical Review*, Vol. 33a, pp. 1134-1140.

Friedman, J.H. (1989), CBMS Regional Conference on Modern Computer Intensive Methods for Exploring and Modeling Multivariate Data, held June 12-16, 1989, George Washington University, unpublished lecture notes - book in writing.

Friedman, J.H. and B.W. Silverman (1989), "Flexible Parsimonious Smoothing and Additive Modeling", *Technometrics*, Vol. 31, No. 1. pp 3-21.

Friedman, J.H. and W. Stuetzle (1981), "Projection Pursuit Regression", *J. of the Amer. Stat. Assoc.*, Vol 76, No. 376.

Furnival, G.M., and R.C. Wilson, Jr. (1974), "Regression by Leaps and Bounds," *Technometrics*, Vol. 16, pp. 499-511.

Gauthier, J.P. and Bornard, G. (1986), "Global Realizations of Analytic Input-Output Mappings," *SIAM J. on Control and Optimization*, Vol. 24, No. 3, pp. 509-521.

Gelfand, I.M., and Yaglom, A.M. (1959), "Calculation of the Amount of Information About a Random Function Contained in Another Such Function," *Amer. Math. Soc. Trans.*, Series (2), Vol. 12, pp. 199-236 (original *Usp. Mat. Nauk.*, Vol. 12, 3-52, 1956).

Gevers, M. and Wertz, V. (1982), "On the Problem of Structure Selection for the Identification of Stationary Stochastic Processes", *Sixth IFAC Symposium on Identification and System Parameter Estimation*, Eds. G. Bekey and G. Saridis), Washington D.C.: McGregor & Werner, pp. 387-92.

Golub, G.H. (1969). *Matrix Decompositions and Statistical Calculations, Statistical Computation*, R.C. Milton and J.A. Nelder, eds., New York: Academic Press, pp. 365-379.

Granger, C.W.J. and Andersen, A.P. (1978), *An Introduction to Bilinear Time Series Models*, Cottingen: Vandenhoek and Ruprecht.

Haggan, V., Heravi, S.M. and Priestley, M.B. (1984), "A Study of the Application of State-Dependent Models in Non-linear Time Series Analysis," *J. Time Ser. Anal.*, Vol. 5, No. 2, pp. 69-102.

Haggan, V. an Ozaki, T. (1981), "Modelling Non-linear Random Vibrations Using an Amplitude-Dependent Autoregressive Time Series Model," *Biometrika*, Vol. 68, pp. 189-196.

Hazewinkel, M. and Marcus, S.I. (1981), "Some Results and Speculations on the Role of Lie Algebras in Filtering," in *Stochastic Systems: The Mathematics of Filtering and Identification and Applications*, M. Hazewinkel and J.C. Willems (eds.), Boston:Reidel.

Hazewinkel, M. and Willems, J.C. (1981), *Stochastic Systems: The Mathematics of Filtering and Identification and Applications*, Boston:Reidel.

Hotelling, H. (1936). "Relations Between Two Sets of Variates", *Biometrika*, Vol. 28, pp. 321-377.

Jakubczyk, B. (1986), "Existence and Uniqueness of Realizations of Nonlinear Systems," *SIAM J. Control and Opt.*, Vol. 24, No. 2.

Kailath, T. (1974), "A View of Three Decades of Linear Filter Theory," *IEEE Trans. Info. Theory*, Vol. 20, pp. 146-181.

Khatri, C.G. (1976), "A Note on Multiple and Canonical Correlation for a Singular Covariance Matrix," *Psychometrika*, Vol. 41, pp. 465-70.

Kaplan, M.H. (1976), *Modern Spacecraft Dynamics and Control*, New York: Wiley.

Koyak, R.A. (1987), "On Measuring Internal Dependence in a Set of Random Variables", *Annals of Statistics*, Vol. 15, No. 3, pp. 1215-1228.

Kullback, S. (1959). *Information Theory and Statistics*, Dover.

Lancaster, H.O. (1969), *The Chi-squared Distribution*, John Wiley & Sons, New York.

Lancaster, H.O. (1966), "Kolmogorov's Remark on the Hotelling Canonical Correlations," *Biometrika*, Vol. 53, pp. 585-588.

Larimore, W.E. (1989), "A Unified View of Reduced Rank Multivariate Prediction Using a Generalized Singular Value Decomposition" Submitted for Publication.

Larimore, W.E., and F.T. Luk (1988), "System Identification and Control Using SVDs on Systolic Arrays," *SPIE Symposium on Innovative Science and Technology, Proc. of Conference on High Speed Computing*, Vol. 880, January, 1988.

Larimore, W.E. and R.K. Mehra (1985), "The Problem of Overfitting Data," *Byte*, Vol. 10, pp. 167-80.

Larimore, W.E., S. Mahmood and R.K. Mehra (1984), "Multivariable Adaptive Model Algorithmic Control", *Proc. Conference on Decision and Control*, Eds. A.H. Haddad and M. Polis, Vol. 2, pp. 675-80. New York: IEEE.

Larimore, W.E. (1983a). "Predictive Inference, Sufficiency, Entropy, and an Asymptotic Likelihood Principle", *Biometrika*, Vol. 70. pp. 175-81.

Larimore, W.E. (1983b). "System Identification, Reduced-Order Filtering and Modeling Via Canonical Variate Analysis", *Proc. 1983 American Control Conference*, H.S. Rao

and T. Dorato, Eds., pp. 445-51. New York: IEEE.

Larimore, W.E. (1981), "Small Sample Methods for Maximum Likelihood Identification of Dynamical Processes," *Applied Time Series Analysis, Proceedings of the Fifth International Time Series Meeting*, Eds. O.D. Anderson and M.R. Perryman, pp. 167-71. Amsterdam: North-Holland.

Leontaritis, I.J. and S.A. Billings (1985), "Input-Output Parametric Models for Nonlinear Systems, Part I: Deterministic Nonlinear Systems", *Int. J. Control*, Vol. 41, No. 2, pp. 303-328.

Leontaritis, I.J. and S.A. Billings (1985), "Input-Output Parametric Models for Nonlinear Systems, Part II: Stochastic Nonlinear Systems", *Int. J. Control*, Vol. 41, No. 2, pp. 329-344.

Ljung, L. (1979), "Asymptotic Behavior of the Extended Kalman Filter as a Parameter Estimator for Linear Systems", *IEEE Trans. Auto. Control*, Vol. 24, pp. 36-50.

Marmarelis, V.Z. (1989), *Advanced Methods of Physiological System Modeling - Volume 2*, Plenum Publishing Corp.: New York.

Marmarelis, V.Z. (1987). "Nonlinear and Nonstationary Modeling of Physiological Systems - An Overview," In *Advanced Methods of Physiological System Modeling*, Vol. I, V.Z. Marmarelis, Ed., Published by Biological Simulations Resource, University of Southern California, Los Angeles, CA. pp. 1-24.

Marmarelis, P.Z., and V.Z. Marmarelis (1978). *Analysis of Physiological Systems: The White-Noise Approach*, New York: Plenum Press.

Packard, N.H. J.P. Crutchfield, J.D. Farmer, and R.S. Shaw (1980), "Geometry from a Time Series", *Physical Review Letters*, Vol. 45, pp. 712.

Peixoto, J.L. (1987), "Hierarchical Variable Selection in Polynomial Regression Models," *The American Statistician*, Vol. 41, No. 4, pp. 311.

Priestley, M.B. (1987), "New Developments in Time-Series Analysis," In M.L. Puri, J.P. Vilapiana, and W. Wertz, editors, *New Perspectives in Theoretical and Applied Statistics*, John Wiley and Sons.

Priestley, M.B. (1980), "State Dependent Models: A General Approach to Nonlinear Time Series Analysis," *J. of Time Series Analysis*, Vol. 1, pp. 47-71.

Rao, C.R. (1965), "The Use and Interpretation of Principal Component Analysis in Applied Research," *Sankhya (A)*, Vol. 26, pp. 329-58

Renyi, A. (1959), "On Measures of Dependence", *Acta. Math. Acad. Sci. Hungar.*, Vol. 10, pp. 441-451.

Rugh, W.J. (1981), *Nonlinear Systems Theory*. Baltimore: John Hopkins Univ. Press.

Schweppe, F.C. (1973), *Uncertain Dynamic Systems*, Englewood Cliffs, N.J.: Prentice Hall.

Sontag, E.D. (1979), "Realization Theory of Discrete-Time Nonlinear Systems: Part I - The Bounded Case," *IEEE Trans. on Circuits and Systems*, Vol. 26.

Subba Rao, T. (1981), "On the Theory of Bilinear Time Series Models," *J.R. Statist. Soc. B*, Vol. 43, pp. 244-255.

Sussman, H.J. (1981), "Rigorous Results on the Cubic Sensor Problem," in *Stochastic Systems: The Mathematics of Filtering and Identification and Applications*, M.

Hazewinkel and J.C. Willems (eds.), D. Reidel Publishing Company.

Sussman, H.J. (1977), "Existence and Uniqueness of Minimal Realizations of Nonlinear Systems," *Math. Systems Theory*, Vol. 10, pp. 263-284.

Takens, F. (1981), "Detecting Strange Attractors in Fluid Turbulence," in D. Rank and L.-S. Young, editors, *Dynamical Systems and Turbulence*, Springer-Verlag, Berlin.

Tong, H. and K.S. Lim (1980), "Threshold Autoregression, Limit Cycles and Cyclical Data," *J. of the Royal Stat. Soc. B*, 42(3), pp. 245-292.

Van Loan, C.F. (1976), "Generalizing the Singular Value Decomposition," *SIAM J. Numer. Anal.*, Vol. 13, pp. 76-83.

Yaglom, A.M. (1970), "Outline of Some Topics in Linear Extrapolation of Stationary Random Processes," *Proc. Fifth Berkeley Symp. Math. Stat. and Prob.*, Berkeley, California, California Press, pp. 259-278

Zaborszky, J., G. Haung, B Zeng, and T.C. Leung (1988), "On the Phase Portrait of a Class of Large Nonlinear Dynamic Systems Such as the Power System," *Trans. Automatic Control*, Vol. 33, pp. 4-15.