Maximum Likelihood Subspace Identification for Linear, Nonlinear, and Closed-loop Systems

(Invited Paper)

Wallace E. Larimore
Adaptics, Inc, 1717 Briar Ridge Road, McLean, VA 22101 USA
Phone: 703 532-0062, Fax: 703 536-3319, Email: larimore@adaptics.com

Abstract— This tutorial paper presents a first principles development of subspace system identification (ID) using a fundamental statistical approach. This includes basic concepts of reduced rank modeling of ill-conditioned data to obtain the most appropriate statistical model structure and order using optimal maximum likelihood methods. These principles are first applied to the well developed subspace ID of linear dynamic models; and using recent results, it is extended to closed-loop linear systems and then general nonlinear closed-loop systems.

The fundamental statistical approach gives expressions of the multistep likelihood function for subspace identification of both linear and nonlinear systems. This leads to direct estimation of the parameters using singular value decomposition type methods that avoid iterative nonlinear parameter optimization. The result is statistically optimal maximum likelihood parameter estimates and likelihood ratio tests of hypotheses. The parameter estimates have optimal Cramer-Rao lower bound accuracy, and the likelihood ratio hypothesis tests on model structure, model change, and process faults produce optimal decisions.

The extension to general nonlinear systems determines optimal nonlinear functions of the past and future using the theory of maximal correlation. This gives the nonlinear canonical variate analysis. New results show that to avoid redundancy and obtain gaussian variables, it is necessary to determine independent canonical variables that are then used in the likelihood function evaluation.

These new results greatly extend the possible applications of subspace ID to closed-loop linear and nonlinear systems for monitoring, fault detection, control design, and robust and adaptive control. Potential applications include system fault detection for control reconfiguration, autonomous system monitoring and learning control, and highly nonlinear processes in emerging fields such as bioinformatics and nano technology. Applications are discussed to identification of vibrating structures under feedback including online adaptive control of aircraft wing flutter, and identification of the chaotic Lorenz attractor.

Index Terms— Subspace system identification, closed-loop, linear, nonlinear, maximum likelihood estimation, optimal estimation.

I. INTRODUCTION

The method of subspace system identification has had some very impressive accomplishments as will become clear in the development. One of these is the use of a singular value decomposition to obtain estimates that in some cases are ‘efficient’, i.e. approach the Cramer-Rao lower bound for variance. Other known efficient methods require nonlinear iterative parameter optimization, whereas using subspace methods the global solution to the problem is immediately obtained by a numerically efficient and stable computation. The method has been widely applied in almost every area requiring systems identification of linear, time-invariant systems with equal spaced data. However, because of lack of a general theory describing the statistical behavior of subspace methods, it has been regarded as an approximate method that may not compete well with maximum likelihood methods in various situations, particularly involving systems with feedback or nonlinear behavior.

The major challenge has been the development of a sound statistical theory that describes the conditions for which it approaches the lower bound accuracy of a maximum likelihood method. Reliance upon subspace system identification in a number of critical applications requires knowledge of the statistical behavior of the subspace algorithm in use. This includes determination of the parameter estimation error or an equivalent model uncertainty description for assessing model accuracy. Such an uncertainty description can be used, for example, in controller robustness analysis or robust control design. For monitoring and fault detection, it is critical that the distribution of test statistics be available for performing tests of hypotheses. This is only generally available for maximum likelihood methods. Later discussions in this tutorial paper as well as some of the other papers in this session give detailed examples of model uncertainty descriptions and distributions of test statistics.

In this paper, we discuss primarily the canonical variate analysis (CVA) method because it results in maximum likelihood accuracy, but comment and compare with the various other subspace methods and algorithms concerning various issues. The paper is organized with the next section giving an overview of the CVA. The following section provides a detailed discussion of CVA for linear systems. Then a discussion of the problem of closed-loop feedback around the plant to be identified is given. The next several sections concern the extension to general nonlinear systems. This is described first as a nonlinear regression problem to focus on the the substantial complexities of the nonlinear aspects of the problem. Then the issues of redundancy and nongaussian that have been major hurdles to solving the problem are discussed, and a method for obtaining a solution is described using optimal normalizing transformations to evaluate the likelihood function. Finally application of these methods to the chaotic Lorentz attractor is presented. The other papers in this tutorial session are Palanthandalam-Madapusi et al. (2005), Lacy et al. (2005), and Juricek et al. (2005).

0-7803-9098-9/05/$25.00 ©2005 AACC

II. OVERVIEW OF CVA CONCEPTS

The basic concept of a canonical variate analysis (CVA) of a time series is discussed, along with some of the characteristics of the solution, particularly for linear systems.

A conceptual starting point for the CVA approach is the description of a process in terms of energy storing quantities referred to as states. The states act as memory and are the central variables in describing the process dynamics. The states include the chemical, thermal, potential or kinetic energy of the process stored in the form of composition, temperature, position or velocity, or rotational energy. The true state vector, or state for brevity, can have very high or even infinite dimension, but at some point in practice a finite dimensional approximation is needed. A major difficulty in previous approaches has been the fitting of a great many possible models of various complexities to the observations, and the determination of adequately fitting models.

In the canonical variate analysis (CVA) approach (Larimore, 1983b), the statistically significant states are first determined directly from the observations by a canonical variate analysis. This determines which of the possible energy storing relationships are statistically significant and provides a mathematical basis for describing them. Then the detailed mathematical description of the system dynamics is directly determined by multivariate regression methods.

The determination of the appropriate model state order is based on an information measure known as the Akaike information criterion (AIC) (Akaike, 1973, 1976). This measure of statistical fit has been justified from the fundamental statistical inference principles of sufficiency and repeated sampling in a predictive inference setting (Larimore, 1983a; Larimore and Mehra, 1985). Once the model state order is determined, a state space model of the process is determined by multivariate regression methods. In the linear case, the computational procedure is based on a singular value decomposition (SVD), which is numerically stable and accurate (Golub, 1969). The computational procedure always gives a meaningful solution to the problem. For the nonlinear case, the computation is more involved and is discussed in following sections.

The CVA method applies to data taken from a very general class of multivariable time-invariant stochastic systems. The assumptions are that:

- The observations are equally spaced in time.
- The system is finite-dimensional, time-invariant, and possibly multivariable.
- The noise disturbances are a finite-dimensional Gaussian processes, i.e. the output of white Gaussian noise exciting a time-invariant finite-dimensional system.
- The observations may include the addition of a time-varying mean function.
- The process may include delayed inputs or internal delayed feedback.
- The measurements may be highly illcondiioned or singular or the process cointegrated.

For the case of linear systems, the CVA method of system identification has been applied to a variety of systems. In the linear case, the theory and computational methods are quite advanced, and has a number of features that are well suited to model estimation and identification on existing microprocessors including:

- General state space model including inputs, and process and measurement noise
- Multi-input, multi-output systems
- Computation using the singular value decomposition
- Numerically stable and accurate – highly reliable
- Automatic determination of the best choice of model state order
- Finite amount of computation – nonrecursive
- No initialization required – accurate on small samples
- Near the maximum likelihood lower bound in accuracy in open- or closed-loop operation as verified in simulations
- Nonminimum phase and time delay systems
- Simultaneous identification of transfer function and noise spectrum of the disturbance process
- Confidence bands giving accuracy on the estimated transfer function
- The required excitation for achieving a given model accuracy is determined using the confidence bands

These properties of the algorithm give it a unique place among the other currently available algorithms. It is reliable, accurate, and yet will handle the very difficult problem of system identification of high-order multivariable systems using short data lengths in the presence of feedback.

The recently developed statistical theory (Larimore, 2004) is quite extensive and shows that for large samples CVA is a maximum likelihood method with optimal accuracy. This has considerable implications for a number of modeling, failure detection and control applications:

- The identification accuracy is optimal and the uncertainty description is accurate for use in stability and performance assessment and robust control design.
- For monitoring process change and failure detection, tests of hypotheses are optimal with minimum probability of false alarm and maximum probability of failure detection
- In reidentifying a system, the required accuracy is achieved with the least amount of data, i.e. control reconfiguration can occur in the shortest time.

These considerable benefits are available only from the use of a maximum likelihood procedure.

While these results for linear CVA are very impressive and have considerable potential for some application, other areas require treatment of nonlinear systems. A number of methods have been developed and applied to nonlinear systems with major advances and impressive results such as Wiener and Volterra methods and neural networks. However most of the methods have major drawbacks:

- They do not utilize the powerful theory of statistical inference and the optimal properties of maximum likelihood estimation and likelihood ratio tests of hypotheses

- They do not produce time domain state space models that are predominantly used in simulation and control system analysis, design, and implementation

On the other hand, the developments in nonlinear modeling in statistics have yielded some major insights and interpretations into seemingly ad hoc procedures such as neural networks. Indeed some recent ‘kernelization’ methods have adopted regularization methods with a statistical interpretation (Hastie, 2001). These developments are related to the maximum likelihood methods of this paper in the discussion below.

In latter sections of this paper, the attractive features of the linear CVA approach are extended to a nonlinear canonical variate analysis (NLCVA) problem. Some of the theory of NLCVA was previously developed in Larimore (1989; 1992b) showing that many of the results apply in a modified form to very general nonlinear systems including chaotic multiple equilibria systems. In this paper, progress is discussed in extending the statistical theory. Such results would provide asymptotic maximum likelihood procedures for model estimation and tests of hypotheses in system with nonlinear dynamics. This would provide optimal estimation accuracy and optimal tests of hypotheses for monitoring system change and failure detection, and accurate uncertainty descriptions for stability and performance assessment.

III. LINEAR CVA OF DYNAMIC PROCESSES

In this section, some of the technical details of the linear CVA method for linear dynamic systems are developed. This provides a basis for further extensions of the CVA method to closed-loop and nonlinear systems. A tutorial paper with additional details on linear CVA is Larimore (1999) that gives a geometric interpretation of CVA and relates it to the other reduced-rank methods of principal component analysis, partial least squares, and instrumental variables.

Canonical correlation and variate analysis was developed by Hotelling (1936) who also developed principal component analysis. Consider the multivariate regression problem

$$f_t = Mp_t + e_t \tag{1}$$

involving the vector f_t of dependent variables and the vector p_t of independent variables with N observations indexed by $t = 1, \ldots, N$. Assume that p and f are jointly distributed as normal random variables with mean zero and covariance matrices Σ_{pp}, Σ_{ff}, Σ_{pf}. These vectors may be high dimensional with considerable redundancy if all of the variables are used. That redundancy can lead to poor models when estimating a large number of parameters if many of them have little effect. For any choice of rank r, canonical variate analysis provides a way to determine transformations of the original variables to new vectors of dimension r defined by variables $c_t = J_r p_t$ and $d_t = L_r f_t$. The new variables c_t and d_t are an optimal choice in terms of maximum likelihood for a specified rank r.

Furthermore, the statistically optimal choice of the rank, i.e., the dimension of c and d, can be determined using maximum likelihood ratio tests. The method of maximum likelihood and likelihood ratio tests have well-known optimal statistical properties in the case of linear regression for the normal distribution (Anderson, 1984). The solution to this problem can be expressed in the form of a generalized singular value decomposition as follows (Larimore, 1990a).

Theorem 1: Canonical Variate Analysis. Let $\Sigma_{pp}(m \times m)$ and $\Sigma_{ff}(n \times n)$ be nonnegative definite (satisfied by covariance matrices). Then there exist matrices $J(\overline{m} \times m)$ and $L(\overline{n} \times n)$ satisfying the generalized singular value decomposition

$$J\Sigma_{pp}J^T = I_{\overline{m}} \; ; \quad L\Sigma_{ff}L^T = I_{\overline{n}} \tag{2}$$

$$J\Sigma_{pf}L^T = \Gamma = diag(\gamma_1 \geq \ldots \geq \gamma_r \geq 0, \ldots, 0), \tag{3}$$

where $\overline{m} = \text{rank}(\Sigma_{pp})$ and $\overline{n} = \text{rank}(\Sigma_{ff})$.

For a specified dimension or rank r to use for c and d, the optimal choice of J_r and L_r is the first r rows of J and L, respectively. Also the maximum of the log likelihood function is simply expressed in terms of the canonical correlations γ_i as

$$\max \log p(Y|X; C, \Sigma_{ee}) = \frac{N}{2} \sum_{i=1}^{r} \log |S_{yy}|^{-1}(1 - \gamma_i^2). \tag{4}$$

Optimal statistical tests on rank involve likelihood ratios. Thus the optimal rank or order selection depends only on the canonical correlations γ_i. A comparison of potential choices of rank can thus be determined from a single GSVD computation on the covariance structure. The above theory applies exactly to the linear regression problem with normally distributed errors that are independent for different samples t.

To extend the CVA concept to time series requires the concept of the past and future of a process. Suppose that data are given consisting of observed outputs y_t and possibly observed inputs u_t at time points $t = 1, \ldots, N$ that are equally spaced in time. Associated with each time t is a past vector p_t consisting of the past outputs and inputs occurring prior to time t as well as a future vector f_t consisting of outputs at time t or later, specifically,

$$p_t = (y_{t-1}^T, u_{t-1}^T, y_{t-2}^T, u_{t-2}^T, \ldots)^T, \quad f_t = (y_t^T, y_{t+1}^T, \ldots)^T \tag{5}$$

Pioneering work by Akaike (1973, 1976) extended the concepts of CVA to the identification of time series models. One of the major issues is that the time series generally violates the assumption of independent errors for different times. Akaike chose the past p_t and future f_t to include the present time, that can result in choosing too high an order for the system state. It was still found to be useful, but the results were substantially less accurate than maximum likelihood and were primarily used to narrow the number of potential model structures that are likely to lead to good models. In a multivariate time series, there can be a very large number of structures (orders of the AR, MA and X components of the model) to sort through. Also, the method used by Akaike was a constructive approach involving the sequential selection of basis elements, and thus the very direct and simple generalized SVD as in the regression case above was not used.

Larimore (1983b) first proposed using the GSVD above directly on the past and future of the time series to determine the rank of the relationship between the past and future. For a process with no inputs u_t, the transformed variable $c_t = J_r p_t$ has all of the information in the past for prediction of the future. This is the definition of the state of a Markov process usually denoted x_t. For a given choice of the state order r, the GSVD defines J_r, that in turn defines a state estimate $x_t = J_r p_t$. A Markov process with state x_t satisfies state equations of the form

$$x_{t+1} = \Phi x_t + G u_t + w_t \tag{6}$$

$$y_t = H x_t + A u_t + B w_t + v_t \tag{7}$$

where w_t and v_t are white noise processes that are independent with covariance matrices Q and R respectively. Since the state estimate is available from $x_t = J_r p_t$, the estimated values can be used in doing a regression of the left-hand variables x_{t+1} and y_t on the right-hand variables x_t and u_t. This simple multivariate regression produces estimates of the matrices Φ, G, H, and A, and similar computations using the error in the regression to produce estimates of B and the covariance matrices Q and R. The entire computations can be implemented using numerically stable and accurate SVD computations.

For time series processes with no inputs, it was found in a particular case (Larimore, Mahmood, and Mehra, 1984) that the accuracy of the model identified using CVA was very close to achieving the Cramer-Rao lower bound so that it was essentially equivalent to maximum likelihood estimation. More detailed simulations followed (Deistler et al, 1995; Larimore, 1996) showing the maximum likelihood behavior of CVA in large samples. This was followed by considerable effort on the asymptotic theory as the sample size becomes large resulting in optimal properties of asymptotic normality and minimum variance (Bauer, 1998; 2005).

To extend the results of the above CVA regression problem to the time series case, the likelihood function is expressed in terms of conditional multistep predictions. To compute, the dimension of the past and future are truncated to a sufficiently large finite number $\zeta = \dim(f_t)$ and $\rho = \dim(p_t)$. Following Akaike (1976), this dimension is determined by autoregressive (ARX) modeling and determining the optimal ARX order using the AIC. The notation $Y_s^t = (y_s, \ldots, y_t)$ is used to denote the observations and similarly for Q_s^t. Suppose that the number of samples N is exactly $N = M\zeta + \rho$ for some integer M. Then by successively conditioning, the log likelihood function of the observations conditional on the initial state $p_{\rho+1}$ at time $\rho + 1$ is

$$\log p(Y_{\rho+1}^{M\zeta+\rho}|p_{\rho+1}, Q, \theta) = \sum_{m=0}^{M-1} \log p(f_{m\zeta+\rho+1}|p_{m\zeta+\rho+1}, Q, \theta) \tag{8}$$

where $Q = Q_1^{M\zeta+\rho}$ so the likelihood function decomposes into the product of M multistep conditional probabilities. Now by shifting the interval of the observations in the above by time s, the likelihood of the observations $Y_{\rho+1+s}^{M\zeta+\rho+s}$ is obtained. Consider the average of these shifted likelihood functions which gives

$$\frac{1}{\zeta}\sum_{s=0}^{\zeta-1} \log p(Y_{\rho+1+s}^{M\zeta+\rho+s}|p_{\rho+1+s}, Q, \theta) \tag{9}$$

$$= \frac{1}{\zeta}\sum_{s=0}^{\zeta-1}\sum_{m=0}^{M-1} \log p(f_{m\zeta+\rho+1+s}|p_{m\zeta+\rho+1+s}, Q, \theta) \tag{10}$$

$$= \frac{1}{\zeta}\sum_{t=\rho+1}^{N} \log p((f_t|q_t)|p_t, \theta)) \tag{11}$$

Now each likelihood function in this average is a likelihood of $N-\rho$ points that differs only on the particular end points included in the time interval. This effect will disappear for large sample size, and even in small sample size will provide a suitable approximate likelihood function. Note that the only difference between the likelihood function for the iid vector case of Section 2 and the time series case here is the normalization $1/\zeta$ involving the dimension of the future and replacing x and y by p_t and $f_t|q_t$ respectively.

In the history of maximum likelihood estimation, major advances were made when a suitable expression for the likelihood function was obtained. The Whittle (1954) likelihood function expressed in the frequency domain allows for ML estimation of random processes in space and time. The Schweppe (1965) likelihood function expressed in terms of the Kalman filter allows for exact ML estimation of time series models. The likelihood function (11) allows for ML estimation using CVA in terms of the past and future as shown below.

CVA is one of the class of subspace methods for fitting state space models from observational data that has been called a 'Larimore type' of subspace algorithm (Bauer and Ljung, 2002). It was noted in Larimore (1983b) that different weightings Λ can be used in place of Σ_{ff} for various purposes. Various weightings were discussed in Larimore (1990a) that correspond to principal component analysis, partial least squares, and instrumental variables, and it was noted that only the weighting Σ_{ff} leads to maximum likelihood results. Bauer and Ljung (2002) show that the CVA weighting is optimal in the possible choice of weighting Λ. The other various weightings cannot be maximum likelihood since they are not invariant to an arbitrary scaling of the data that is a property of maximum likelihood. In particular, the N4SID algorithm of Van Overschee and De Moor (1996) originally used the identity weighting, that was subsequently shown to be much less accurate in some cases. The CVA weighting can be used in the N4SID algorithm, but it still can be much less accurate than CVA (Juricek et al, 2002) due to a number of differences in the details of the two methods.

Van Overschee and De Moor (1994) have shown that the known subspace algorithms are approximately a generalized SVD of the form of CVA with various weightings followed by various procedures for estimating the state space matrices from the GSVD. The CVA method uses regression to estimate the state space matrices. On the other hand, the N4SID algorithm of Van Overschee and De Moor (1996) uses the structure of the observability and controllability matrices to obtain estimates of the state space matrices. N4SID is the only other commercially available subspace system identification software besides the ADAPTx implementation of CVA (Larimore, 1992a).

Dahlen et al (1998) have studied CVA and the three N4SID algorithms given in the book by Van Overschee and De Moor (1996) in terms of the fundamental requirement that any solution must be positive real. Positive real is equivalent to the requirement that a covariance sequence be positive semi-definite to be a meaningful solution. They construct systems that generate data which cause all three algorithms in Van Overschee and De Moor (1996) to fail even when using the CVA weighting. Also the algorithm of Aoki (1987) is shown to fail. This demonstrates that these algorithms are not completely reliable. In the dissertation (Dahlen, 2001), Dahlen also analyzes the CVA algorithm and says the solution is guaranteed to be positive real so that no such failures of the algorithm will occur. Thus there is a major reliability issue with N4SID that is critical in applications such as control of aerospace vehicles that requires a completely automatic and reliable implementation.

Finally, very recent results have been published proving the large sample efficiency of CVA, that shows the parameter estimation error variance achieves the Cramer-Rao lower bound for large samples. Thus no method of estimating such a parametric model will achieve a smaller estimation error. This is shown for the case of no inputs by Bauer (2005), and is outlined for the case of inputs with feedback for the ADAPTx algorithm in Larimore (2004). One of the central methods used in both of these papers is the multistep likelihood function developed in Larimore (1997a) and called a pseudo-likelihood function in Bauer (2005). As discussed in sections below, the multistep likelihood is a key to the interpretation and theory for the nonlinear CVA. Another key concept (Larimore, 2004) is viewing the linear CVA method for the case of inputs with feedback as a sequence of nested models successively projected onto lower dimensional models as discussed in Cox and Hinkley (1974).

IV. CLOSED-LOOP IDENTIFICATION

In this section an outline is given of the large sample efficiency of adaptx for the case of unknown feedback. A more detailed technical development will appear elsewhere. Asymptotic efficiency means the parameter estimation error approaches the minimum variance bound for large sample size.

Over the past two decades, the computational methods, statistical theory, and applications of canonical variate analysis (CVA) have been developed considerably. The basic algorithm (Larimore, 1983) has been significantly improved with model order selection (Larimore, 1990a; 1990b), confidence bands on spectral functions such as frequency response and power spectrum (Larimore, 1993), monitoring and fault detection (Larimore, 1997a; Wang et al, 1997; Juricek et al, 2004; Conner et al, 2004), and delay estimation (Larimore, 2003).

There were early empirical demonstrations of near optimal estimation approaching the Cramer-Rao lower bound (Larimore et. al., 1984), with more detailed simulations to follow (Deistler et al, 1995; Larimore, 1996a, 1996b; Peternell et al, 1996). In the case of no inputs, this was followed by considerable effort on the asymptotic theory, as the sample size becomes large, showing the optimal properties of asymptotic normality and minimum variance (Bauer, 1998; 2005).

A much discussed aspect in the literature has been the behavior of subspace system identification for the case of colored inputs perhaps with feedback. The fundamental problem is the necessity to compute and remove the effects of future inputs on future outputs before the CVA is done to determine the system state. But it appears that the CVA solution itself is required to compute these effects on future outputs. In Larimore (1996a, 1996b), simulation results were presented that strongly suggest such efficiency for that simulation model. The algorithm used in those simulations, and incorporated in the first release of the ADAPTx^{TM} software (Larimore, 1992) as well as all subsequent releases, is as follows:

- Fit ARX. Using conditional maximum likelihood (ML), fit ARX models recursively on order and evaluate the AIC_C statistic to determine the optimal number ℓ of delayed inputs and outputs to use in the CVA computation.
- Remove effects of future inputs q_t on future outputs f_t. Compute the multistep predictor matrix Ω using the ARX model, and compute the corrected future $f_t|q_t = f_t - \Omega q_t$.
- CVA. Do a CVA between the past p_t and corrected future $f_t|q_t$ to determine the states ordered by their associated canonical correlation.
- Select State Order k. Compute the estimated one-step prediction error covariance matrix for each state order from 0 to order $\ell\text{Dim}(y_t)$, compute the associated AIC_C for each order, and select the minimum AIC_C.

- Estimate Model. Compute estimates of the state space matrices and the one-step error covariance in the state equations by regression.
- Alternate Model Forms. Solve the Riccati equation and compute the innovations, overlapping parameterization, and ARMAX models.

It may seem surprising that the use of the ARX model to remove the effects of future inputs from future outputs results in an optimal procedure with asymptotic efficiency. Questions that come to mind are the well known issues:

- High Order ARX. The ARX model can have far more parameters to obtain a reasonable approximation to the process than the state space model especially for a process with moving average terms in the noise requiring a high AR order.
- ARX Model Error. Such a high order ARX model will have modeling error proportional to the number of estimated parameters so the modeling error for the ARX could be much larger than that potentially achievable using a SS model.
- SS Model Error. Thus using the ARX model to remove the effect of future inputs on future outputs could result in additional error in the future outputs, and consequently increase the error in fitting the SS model in subsequent steps.

While these issues are well founded concerns, it will be shown that there is much additional structure to the problem that effectively projects these additional errors to zero.

The adaptx algorithm is discussed below in terms of a number of statistical concepts and how they impact the estimation problem. It has long been noted in the literature (Larimore, 1990a) that the difficulty is the presence of future inputs that introduce errors in the prediction of the future outputs from the past, and this introduces errors in the CVA step to determine the state. The use of the ARX model avoids this problem for a number of reasons that will become more evident in later sections. The basic concept is given in Cox and Hinkley (1974, pp. 307, 321-4) concerning nested models, projection, and sufficiency. The use of the ARX model to remove future inputs from future outputs has the following advantages:

- Linear Computation. Fitting of the ARX model permits the approximate maximum likelihood identification of a model using efficient and non-iterative linear computations that are needed also to determine the number of lags ℓ of the past to use in the CVA calculation.
- Order-recursive Computation. A process can be approximated arbitrarily closely by an ARX process, and recent methods permit the use of an efficient (order ℓ^3 verses ℓ^4 multiplications) order-recursive computation that is highly accurate with no error accumulation (Larimore, 1990b, 2002).
- ML is Immune to Colored Inputs and Feedback. The ARX procedure is asymptotically ML and as such the estimates of the plant model from input and output data do not depend on knowledge of the spectrum of the inputs or feedback system, i.e. there is no bias in the estimates (Larimore, 1997b; Gustavsson et al, 1977)
- Nested Model. The ARX model class contains the state space model that is fitted by regression so that the subspace model is nested in the ARX model. Specifically, the state space model parameters lie in a subspace of the ARX model.
- Projection to Low Dimension. Because of the nested model structure, fitting of the SS model by regression projects the ARX model onto the low dimensional state subspace of the ARX space of delayed inputs and outputs.
- Decomposition of the ARX Model. The ARX model decomposes into two pieces, the low-dimensional SS model and the part of the ARX model orthogonal to the SS model. This orthogonal piece projects to zero, i.e. errors in this part of the ARX model go to zero when projecting on the SS model.
- ARX Model is Sufficient for SS Model. From model nesting, all of the information in the sample for inference about the SS model is contained in the ARX model parameter estimates.
- Multistep Likelihood Function. The equivalence of the onestep and multistep likelihood functions plays a key role in the technical details to demonstrate orthogonality.

While there have been a number of recent papers on new subspace algorithms to handle colored inputs and feedback, there has been very little discussion concerning the asymptotic efficiency of these subspace methods. An exception is Peternell et al (1996) who discuss two algorithms, one imposing a block shift structure on the model involving future inputs, and the other using an iteration to refit the previous model for removing the effects of inputs. By simulation, the first method was shown not to be efficient, and the second appeared to be. But the iterative method appears not to have been pursued, presumably because a major advantage of CVA is the lack of any iteration.

A method was developed by Ljung and McKelvey (1996) using ARX models to remove the effect of future inputs on future outputs. However, the ARX model is used in a completely different way to predict the future outputs that are then used in place of the measurements. A major disadvantage is that such a procedure will lead to biased estimates of the noise covariance matrix. They mention the potential illconditioning in fitting high order ARX models. Illconditioning is avoided in the adaptx algorithm by using the order-recursive factorization algorithm (Larimore 1990b, 2002, 2003) that has been demonstrated to be accurate to machine precision even in the case of highly rank deficient data (Larimore, 2002).

Shi (2001) and Shi and MacGregor (2001) discuss several algorithms and consider the use of the ARX model to remove the effects of future inputs on future outputs and show it gives unbiased estimates in the presence of unknown feedback. There is no discussion of the efficiency of the procedure.

An easy way to see the immunity of ML estimation to feedback is based on simple conditional probability relationships, as shown in Larimore (1997b). The following notation will be used in the development, $Y_1^N = (y_N, \ldots, y_1)$ and similarly for U_1^N. Also let p_t denote the inputs and outputs in the strict past of t. The joint likelihood function of the outputs Y_1^N and the inputs U_1^N conditional on the initial state expressed by the past p_1 at time $t = 1$ and as a function of the unknown parameters θ can be expressed

$$p(Y_1^N, U_1^N | p_1; \theta) = [\prod_{t=1}^{N} p(y_t | u_t, p_t; \theta)][\prod_{t=1}^{N} p(u_t | p_t; \theta)] \tag{12}$$

The probability densities above involve the conditional random variable $y_t|(u_t, p_t)$ that is the usual output innovations process of the plant input-output model. The conditional random variable $u_t|p_t$ is the innovation of the feedback system with a required delay of one time step between y_t and u_t. The joint likelihood function of (Y_1^N, U_1^N) is expressed as the product of two terms that are thus independently distributed. Each of these terms is the product of probabilities of independently distributed innovations processes.

The above factoring of the likelihood function into two terms as in (12) always holds and is the consequence of simple conditional probability rules. The real usefulness comes, however, when the plant and feedback pieces of the system can be parameterized separately. Suppose that the parameter vector can be written as $\theta = (\theta_p, \theta_f)$ where the two subvectors respectively parameterize the plant and feedback parts of the systems. In this case, the maximum of the likelihood function is the product of the maxima of each of the two pieces. Thus under the hypothesis that the process is in a plant-feedback form with the only relationships between them appearing in the plant inputs and outputs, then ML estimation of the plant does not depend upon the presence or absence of feedback. The actual computation of the ML estimates for the ARX model and other details are discussed in the next section.

V. PROJECTION IN ARX AND MARKOV MODELS

The fitting of ARX models using conditional ML and the fitting of state space models using CVA involve the use of regression. Projection is a very useful concept in regression that greatly clarifies some fundamental orthogonality relationships among the identified parameters. The result of this is the elimination of the effect of future inputs on future outputs even in the presence of unknown feedback in the system.

Consider the multivariate ARX model

$$y(t) = \sum_{s=1}^{\ell} \alpha(s) y(t-s) + \sum_{s=0}^{\ell} \beta(s) u(t-s) + e(t) \tag{13}$$

for $t = \ell + 1, \ldots, N$, and where ℓ is the AR and X orders and the error e_t is normally distributed with covariance matrix Σ and independently for different t. The $\alpha(s)$ are the autoregressive (AR) coefficients and the $\beta(s)$ are the exogenous (X) input coefficients.

In fitting the ARX model using least squares (LS), also called conditional maximum likelihood (ML), the equations (13) are used for $t = \ell + 1, \ldots, N$, and are transposed and stacked up to give

$$Y = Z\Theta + E \tag{14}$$

where $Y^T = [y_{\ell+1}, \ldots, y_N]$ with the first ℓ observations of the output not used in the regression so it is conditional on the first ℓ observations. Also denote $\Theta^T = [\alpha_1, \ldots, \alpha_\ell, \beta_0, \ldots, \beta_\ell]$ and

$$Z^T = \begin{bmatrix} y_\ell & \cdots & y_1 & u_{\ell+1} & \cdots & u_1 \\ \vdots & & & & & \vdots \\ y_{N-1} & \cdots & y_{N-\ell} & u_N & \cdots & u_{N-\ell} \end{bmatrix}$$

The linear model (14) applies to much more general processes than ARX models, that will be denoted by Θ_A when needed. The LS and conditional ML estimates are given as

$$\hat{\Theta} = (Z^T Z)^{-1} Z^T Y$$

$$\hat{\Sigma} = Y^T Y - \hat{\Theta}^T Z^T Z \hat{\Theta}$$

The model for y_t is the right hand side of (13) without the noise e_t, which is the conditional expectation of y_t given the past p_t and present input u_t. This is the systematic part of the model for Y. The ML estimates $\hat{\Theta}$ minimize the error $E = Y - \hat{Y}$ with

$$\hat{Y} = Z\hat{\Theta} = Z_1 \hat{\Theta}_{1*} + \cdots + Z_m \hat{\Theta}_{m*}$$

where Z_i is the i-th column of Z and $\hat{\Theta}_{i*}$ is the i-th row of $\hat{\Theta}$.

A subspace projection interpretation clarifies the nesting of parameter spaces. Primarily the univariate case is discussed for conceptual simplicity (see Schaffe, 1959, pp. 43, for a detailed discussion), but it extends to the multivariate case (Anderson, 1984, pp. 295).

In the case that Y is a vector so that $\hat{\Theta}$ is a vector of parameters, then $\hat{\Theta}$ is the linear combination of the columns of Z that gives the model $\hat{Y}$ for Y. Thus the model $Z\hat{\Theta}$ is an $N - \ell$ dimensional vector that lies in the m-dimensional subspace generated by the m columns of Z, denoted $\mathcal{S}(Z)$. Also the parameters $\hat{\Theta}_i$ can be associated with the basis vectors Z_i, respectively, and are coordinates for the subspace. A change of coordinates can be used to define a different parameterization of the subspace. In the multivariate case that Y is a matrix, then the above interpretation applies to each column Y_i of Y using the corresponding column $\hat{\Theta}_{*i}$ of $\hat{\Theta}$ so that the model for the i-th components Y_i of the observations is

$$\hat{Y}_i = Z\hat{\Theta}_{*i} = Z_1\hat{\Theta}_{1i} + \cdots + Z_m\hat{\Theta}_{mi} \tag{15}$$

This has the following projection interpretation. The estimated model $\hat{Y} = Z\hat{\Theta} = Z(Z^TZ)^{-1}Z^TY$ involves the orthogonal projection operator $Z(Z^TZ)^{-1}Z^T$. The error $Y - \hat{Y}$ is orthogonal to the estimate $\hat{Y}$ since substituting the above for $\hat{Y}$ reduces $\hat{Y}^T(Y - \hat{Y})$ to zero. So $\hat{Y}$ is the orthogonal projection of columns of Y onto the subspace $\mathcal{S}(Z)$ span by the columns of Z with the projections defined by the linear combinations (15) specified by the columns of $\hat{\Theta}$.

Substituting $Y = Z\Theta + E$ into $\hat{Y} = Z\hat{\Theta} = Z(Z^TZ)^{-1}Z^TY$ gives

$$\hat{Y} = Z\Theta + Z(Z^TZ)^{-1}Z^TE \tag{16}$$

Thus, under the hypotheses that the true process lies in a lower dimensional subspace, the first observation is that except for the noise, the estimate $\hat{Y}$ is equal to the true noiseless value $Z\Theta$ plus noise. The second observation is that projecting the data on a lower dimensional subspace reduces the degrees of freedom of the noise to the dimension of the subspace. This is a major concept in obtaining asymptotic efficiency.

In the case of static regression where the regressors Z are not random variables but fixed known values, parameter estimates are unbiased since

$$\begin{aligned}\mathcal{E}[\hat{\Theta} - \Theta] &= \mathcal{E}[(Z^TZ)^{-1}Z^T)Y - \Theta] \\ &= \mathcal{E}[(Z^TZ)^{-1}Z^T(Z\Theta + E) - \Theta] = 0\end{aligned}$$

and the parameter estimation error between any two columns $\hat{\Theta}_i$ and $\hat{\Theta}_j$ of $\hat{\Theta}$ is

$$\begin{aligned}\mathrm{Cov}(\hat{\Theta}_i, \hat{\Theta}_j) &= \mathcal{E}(Z^TZ)^{-1}Z^TE_{i.}E_{j.}^TZ(Z^TZ)^{-1} \\ &= (Z^TZ)^{-1}Z^T\sigma_{ij}Z(Z^TZ)^{-1} = \sigma_{ij}(Z^TZ)^{-1}\end{aligned}$$

Suppose the space $\mathcal{S}(Z)$ decomposes into two subspaces that are orthogonal so $Z_a = (Z_1, \ldots, Z_r)$ and $Z_b = (Z_{r+1}, \ldots, Z_m)$ with $Z = (Z_a\ Z_b)$ and the orthogonality condition $Z_a^TZ_b = 0$. Then the corresponding decomposition of the parameters $\hat{\Theta} = (\hat{\Theta}_a; \hat{\Theta}_b)$ have diagonal covariance matrix with

$$\mathrm{Cov}(\hat{\Theta}_i, \hat{\Theta}_j) = \sigma_{ij}\mathrm{diag}((Z_a^TZ_a)^{-1}, (Z_b^TZ_b)^{-1})$$

so parameter estimates $\hat{\Theta}_a$ and $\hat{\Theta}_b$ are uncorrelated. The converse is also true; if $\hat{\Theta}_a$ and $\hat{\Theta}_b$ are uncorrelated, then Z_a and Z_b are orthogonal.

Now given a subspace $\mathcal{S}(Z_S)$ of a larger space $\mathcal{S}(Z_A)$, the orthogonal compliment Z_{A-S} can always be constructed by orthonormalization, that in turn defines orthogonal parameter estimates Θ_S and Θ_{A-S}. The Z_S and Θ_S are said to be nested respectively in Z_A and Θ_A. Denoting the restricted model as $\hat{Y}_S = Z_S\hat{\Theta}_S$ in such a nested structure, the error $\hat{Y}_A - \hat{Y}_S$ is orthogonal to the estimate $\hat{Y}_S$ as illustrated in Fig. 1 and the and parameter estimates $\hat{\Theta}_S$ and $\hat{\Theta}_{A-S}$ are uncorrelated.

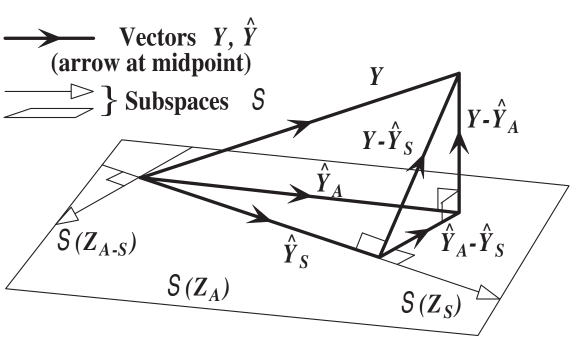

Fig. 1. Nested Subspaces and Orthogonality Relationships.

In the case of estimating an ARX time series with Z random rather than a static regression, the above properties also hold asymptotically for large sample under appropriate assumptions (Lütkepohl, 1993).

Now, consider any finite dimensional multivariable Markov process with vector input u_t and output y_t of the form

$$x_{t+1} = \Phi x_t + Gu_t + w_t \tag{17}$$

$$y_t = Hx_t + Au_t + Bw_t + v_t \tag{18}$$

where x_t is a k-order Markov state and w_t and v_t are white noise processes that are independent with covariance matrices Q and R respectively. An alternative representation is the innovations form where the noise terms w_t and $Bw_t + v_t$ are replaced, respectively, with Kv_t and the output innovation v_t, where K is the Kalman gain obtained from solving the Riccati equation. The state expressed as $x_t = J_k^\infty p_t^\infty$ in terms of the infinite past p_t^∞ is

$$x_t = \sum_{i=1}^{\infty}(\Phi - KH)^{i-1}[(G - KA)u_{t-i} + Ky_{t-i}] \tag{19}$$

that results from recursively substituting (17) for x_t in (17). Eq. (19) is equivalent to (17) provided that $J_k^\infty(\Theta)$ is parameterized as in (17) and (18). By truncating, the approximation $x_t = J_kp_t$ is obtained. The approximation error decreases as $(\Phi - KH)^\ell$ that is exponential in the length ℓ of the past p_t so it can be ignored asymptotically. Since (18) with $x_t = J_kp_t$ is in the ARX form (13) with additional restrictions on the parameters, the Markov model (17) and (18) is nested within the ARX model class, asymptotically.

In the adaptx subspace algorithm, the fitting of the Markov model is done in two steps. First, a reduced-rank regression is done to estimate $\hat{J}_k$ of fixed rank in $x_t = \hat{J}_kp_t$ and with no parametric constraints on $\hat{J}_k$ so it is not parameterized as in (19). The reduced-rank regression is performed using a canonical variate analysis between past and future as developed in Larimore (1997a) for the case of no inputs. The case of inputs with feedback is developed in the next section. In the second step, the constraints are then introduced by regression using (17) and (18) with the state given by $x_t = \hat{J}_k p_t$. In particular, let X^+ denote X with the time index t replaced by $t+1$, and project $(X^+ \; Y)$ on $\mathcal{S}(X \; U)$ to obtain $(\Phi \; G; H \; A)$. Implicit in this regression are additional constraints among the parameters that lead to the various state space canonical forms (Candy et al, 1979).

This can be viewed as a succession of restrictions on ML models starting with the ARX, then the reduced rank regression using CVA, and finally the state space regression using (17) and (18). The latter two involve nonlinear parameterizations, and are developed in detail in Larimore(2004).

VI. NONLINEAR CVA FOR REGRESSION

In this section, an outline is given for the extension of the linear CVA method for linear dynamic processes discussed in the previous section to the static nonlinear regression problem. Nonlinear CVA (NLCVA) can be viewed as a generalization of the linear problem (3) where linear combinations of the observed variables are replaced by nonlinear functions of the variables. This nonlinearity considerably complicates the nature of the problem as well as the methods available for solution. In the following section, this approach will be extended to nonlinear dynamic stochastic systems.

The discussion below of nonlinear CVA is an elementary description of the concepts in terms of the linear CVA, which is much easier to understand. The original development of the method in Larimore (1989) involves a much more detailed theory using functional analysis (Renyi, 1959; Csaki and Fischer, 1960, 1963; Brieman and Friedman, 1985). NLCVA is directly related to recent kernel computational methods that in some cases offer an advantage (Friedman, 2004; Hastie, Tibshirani, and Friedman, 2001; Scholkopf and Smola, 2002: Scholkopf et al., 1998; Kass and Graepel, 2003; Bach and Jordan, 2002).

The nonlinear problem is posed analogous to the original linear formulation by Hotelling (1936) except that nonlinear functions are considered. Early work on the nonlinear problem was done by Renyi (1959; see also Csaki and Fischer, 1960, 1963). Consider random vectors $X = (x_1, \dots, x_m)^T$ and $Y = (y_1, \dots, y_n)^T$ that have a joint probability distribution, and consider the sets $\mathcal{F}_X$ and $\mathcal{F}_Y$ respectively of nonlinear functions $f(X)$ and $g(Y)$ that, without loss of generality, are assumed to be centered and scaled so that they have zero mean and variance 1 (i.e., $E[f(x)]^2 = E[g(y)]^2 = 1$). Then due to the centering and scaling of f and g, the correlation coefficient $\rho(f(X), g(Y))$ equals $E[f(X)g(Y)]$ where $E[\;]$ is expectation or averaging.

The maximal correlation $\rho^*(X,Y)$ of X and Y is defined as the maximum over the functions (f,g) in the respective sets $\mathcal{F}_X$ and $\mathcal{F}_Y$ of the correlation coefficient

$$\rho^*(X,Y) = \max_{(f,g)} \rho(f(X), g(Y)) = \max_{(f,g)} E[f(X)g(Y)] \tag{20}$$

where f and g run over all Borel measurable functions with zero mean and unit variance. Borel measurable functions are very general and allow, for example, jump discontinuities.

The maximal correlation satisfies the following properties:

- Existence. $\rho^*(X,Y)$ is defined for every pair of random vectors X and Y, neither of them being a constant with probability 1.
- Symmetric. $\rho^*(X,Y) = \rho^*(Y,X)$.
- Nonnegative and bounded. $0 \leq \rho^*(X,Y) \leq 1$.
- Stochastic Independence. $\rho^*(X,Y) = 0$ if and only if X and Y are stochastically independent, also called statistically independent.
- Deterministic Dependence. $\rho^*(X,Y) = 1$ 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, so there is deterministic dependence. The converse requires some additional conditions.
- Invariance. Under 1-1 onto Borel-measurable transformations f and g,
  $\rho^*(f(X), g(Y)) = \rho^*(X,Y)$.
- Normal Implies Linear. If the joint distribution of X and Y is normal, then the maximal correlation $\rho^*(X,Y)$ is achieved by considering only linear functions f and g.

The correlation coefficient itself is generally a rather poor measure of relationship between nonlinear functions of random variables. What is remarkable is that the maximal correlation characterizes independence and, under suitable restrictions, strict (or deterministic) dependence. A central concept discussed in this paper is to make full use of these very strong properties of maximal correlation to obtain minimal order descriptions of very general nonlinear processes.

Now consider the nonlinear reduced-rank multivariate regression problem. Extending the above notation, for a given positive integer r consider the sets $\mathcal{F}_X^r$ and $\mathcal{F}_Y^r$ of all Borel measurable r-dimensional vector functions $f(X) = (f_1(X), \dots, f_r(X))^T$ and $g(Y) = (g_1(Y), \dots, g_r(Y))$, where each component is zero mean and unit variance as above. The sets $\mathcal{F}_X^r$ and $\mathcal{F}_Y^r$ are linear vector spaces on which we define the inner product $<\;,\;>$ (also called the dot product)

$$< f, g >= trEf(X)g^T(Y) = E \sum_{i=1}^{r} f_i(X)g_i^T(Y) = tr\Sigma_{fg} \tag{21}$$

where the covariance matrix notation $\Sigma_{fg} = Ef(X)g(Y)^T$ is used and where $tr(\;)$ is the trace, i.e. the sum of the diagonal elements. The pseudonorm is given by

$$\| f \| =< f, f >^{1/2} . \tag{22}$$

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

A number of results can be shown using the theory of operators on Hilbert spaces. In particular, it can be shown that under regularity conditions on the joint probability distribution function $P(X, Y)$, there always exist functions f and g attaining the maximal correlation.

It was shown in Larimore (1989, 1992b) that the generalization of the nonlinear reduced-rank multivariate regression can be phrased as the following minimization problem.

Problem 1: Rank r Nonlinear Prediction. For a given positive integer r, find r-dimensional vector functions $f(X)$ and $g(Y)$ that minimize the relative prediction error

$$\max_{(f,g)} \parallel g(Y) - \hat{g}(f(X)) \parallel_{\Sigma_{gg}^{\dagger}} \tag{23}$$

where $\hat{g} = E(g|f)$ is the conditional expectation of $g(Y)$ given $f(X)$.

The solution to this problem, as shown in Larimore (1989, 1992b), is similar to the solution of the linear CVA problem (3) as follows.

Theorem 2: Nonlinear CVA. For any choice of rank r, there exist multivariable functions f and g of dimension r satisfying $\Sigma_{ff} = \Sigma_{gg} = I_r$ and $\Sigma_{fg} = D = Diag(d_1, \geq \ldots \geq d_k > 0, \ldots, 0)$ that maximize

$$\max_{(f,g)} tr(\Sigma_{gf}\Sigma_{fg}) = \sum_{i=1}^{r} d_i^2 \tag{24}$$

For any choice of rank r, the first r canonical correlations d_i are unique to within a sign change, and the rank r nonlinear prediction problem is solved by the first r components of f and g.

The problem can also be phrased in terms of choosing the components of the canonical functions f and g sequentially and pairwise as follows.

Theorem 3: Sequential Selection. The vector functions f and g giving an optimal solution to the nonlinear prediction problem (23) are obtained sequentially by the following procedure: For each r, find the pair of functions (f_r, g_r) that are uncorrelated with 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

$$d_r = \max_{(f_r, g_r)} \rho(f_r(X), g_r(Y)). \tag{25}$$

VII. REDUNDANCY OF NONLINEAR 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. In this section, some indication of what goes wrong is discussed, and particularly for the case of the normal distribution. A starting point that suggests a much more basic relationship is the following (Lancaster, 1966,1969).

Theorem 4: If X and Y are jointly normal variables, then the maximal correlation occurs for linear transformations $f(X)$ and $g(Y)$, 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 following 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 5: Let X and Y be jointly normal random vectors, and let the nonlinear transformations $f_i(X)$ and $g_j(Y)$ be recursively defined so that $E(f_i(X)f_j(X)) = E(g_i(Y)g_j(Y)) = 0$ for $i < j$. Then f_1 and g_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 f_i and g_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 (Larimore, 1989).

Theorem 6: Let X and Y be jointly normal random vectors of dimensions k and ℓ respectively, and let the nonlinear transformations $f_j(X)$ and $g_j(Y)$ be recursively defined such that $\rho^*(f_i(X), f_j(X)) = \rho^*(g_i(Y), g_j(Y)) = 0$ for $i < j$. Then the functions f_j and g_j have maximal correlation if they are respectively the jth pair of linear canonical variables c_j and d_j. For $\rho_j > 0$, f_j and g_j are strictly linear functions respectively of $c_1, \ldots, c_k$ and $d_1, \ldots, d_\ell$.

With the added requirement of independence among the canonical variables, the univariate result of Theorem 4 is generalized to the multivariate case by Theorem 6. Thus in the case of joint normality, among all possible functions the independent canonical variables involve linear functions, are normally distributed, and the corresponding prediction problem among the canonical variables is linear.

VIII. INDEPENDENT CVA AND NORMALITY

In this section, the construction of an independent canonical variate analysis (ICVA) is discussed.

For a nonlinear process, the number of orthogonal canonical variables with nonzero canonical correlations may not equal the minimal state order as it does for linear processes. The problem is that two canonical variables that are by definition orthogonal may be such that one is a deterministic function of the other. Thus there is no new information in the second that is not available in the first. So there may be considerable redundancy in the canonical variables, i.e. some nonlinear functions of different canonical variables may be highly correlated or functionally dependent.

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 are such that there is no functional redundancy between them in the sense that for any functions e and f, $e(g_1)$ and $f(g_2)$ are uncorrelated. This is equivalent to the statement that the maximal correlation is zero, i.e. $\rho^*(g_1, g_2) = 0$, which from Section VI 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. in the Sequential Selection Theorem 3 replace the orthogonality condition $< g^{(r-1)}, g_r > = 0$ with the mutual independence condition $\rho^*(g^{(r-1)}(Y), g_r(Y)) = 0$

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 remainder of this section, regularity conditions are studied that permit the construction of independent canonical variables. In particular, we wish to show that for any sets W and X of random variables, there exists a set $U(W, X)$ of random variables, where each of the components $u_i(W, X)$ is a function of W, X, and are such that W and U are mutually independent and generate the same space as W and X. In the statement and proof of the theorem, the standard statistical notation is used where upper case W denotes a random variable or random vector and lower case w denotes a particular real value of a random variable or vector. One version of such a theorem is given below after stating the regularity condition that the probability density function must satisfy.

Condition 1: Density Function. The density $p_{W,X}(w, x)$ of the joint distribution $P_{W,X}(w, x)$ with respect to the product $P_W(w)P_X(x)$ of the marginals exists, is continuous and nonzero.

Theorem 7: Independent Generators. Under Condition 1 there exists a transformation $U(W, X)$ such that the map: $(W, X) \rightarrow (W, U)$ is 1-1 and W and U are mutually independent.

Proof: For simplicity, first the case of W possibly a vector and X a scalar is considered. From Condition 1, the probability densities exist and are nonzero. Since $p(w, u) = p(u|w)p(w)$ and w and u are independent if and only if $p(w, u) = p(u)p(w)$, it follows that w and u are independent if and only if $p(u|w) = p(u)$, i.e. if and only if the conditional density is equal to the marginal density. Thus we construct a transformation of x to a variable u such that this is true. Let $F_W(w) = P(W < w)$ be the cumulative distribution function for the scalar random variable W. Consider the conditional random variable $X|w$ with density $p_{X|w}(X|w)$ which can be transformed to the uniform random variable t defined in terms of the cumulative distribution

$$t(w, x) = F_{X|w}(x) \tag{26}$$

and transformed back to the scalar random variable u with the same density as the marginal $p(x)$ by the function

$$u(w, x) = F_X^{-1}(t(w, x)) \tag{27}$$

The cumulative distribution of $U|w$ of U for fixed vector w is

$$F_{U|w}(u) = F_X(u), \tag{28}$$

the marginal density of X which does not depend upon w. The marginal density of U is

$$p_U(u) = \int p_{U,X}(u, w)dw = \int p_{U|x}(u|w)p_W(w)dw \tag{29}$$

$$= \int p_X(u)p_W(w)dw = p_X(u) = p_{U|w}(u|w) \tag{30}$$

By construction, the map: $(W, X) \rightarrow (W, U)$ is 1-1.

Now consider the case of X a vector. Then by induction using the case proven for a scalar X, the transformation $(W, U_1, \ldots, U_i, X_{i+1}) \rightarrow (W, U_1, \ldots, U_i, U_{i+1})$ is constructed that is 1-1 with $W, U_1, \ldots, U_i$ independent of U_{i+1} which proves the theorem. ■

Now, the independent canonical variate analysis is constructed using the sequential selection theorem with stochastic independence in place of orthogonality.

Theorem 8: Independent CVA. Assume Condition 1 on the density of the sets (X, Y) and consider the nonlinear prediction problem (23) with the additional requirement that the components of $f(X)$ are mutually stochastically independent and similarly for $g(Y)$. Then there exist vector functions f and g giving an optimal solution that are obtained sequentially by the following procedure: For each r, find the pair of functions (f_r, g_r) such that $f_r(X)$ is stochastically independent of the previously selected random vector $f^{r-1}(X) = (f_1, \ldots, f_{r-1})^T$, and $g_r(Y)$ is independent of $g^{r-1}(Y) = (g_1(Y), \ldots, g_{r-1}(Y))^T$, and simultaneously maximize the correlation

$$d_r = \max_{(f_r, g_r)} \rho(f_r(X), g_r(Y)). \tag{31}$$

Furthermore, the canonical variables are jointly normally distributed.

Proof: Suppose that the theorem is true for $r - 1$ so that there exist functions $f^{r-1} = (f_1, \ldots, f_{r-1})^T$ and $g^{r-1} = (g_1, \ldots, g_{r-1})^T$ with $f^{r-1}(X)$ mutually independent and $g^{r-1}(Y)$ mutually independent that pairwise maximize the correlation (31). Then by Theorem 7 with $W = f^{r-1}(X)$, there exist independent random variables $U(W, X)$ as functions of W and X that are independent of W so that the transformation $(W, X) \rightarrow (W, U)$ is 1 to 1 and generates the same space of random variables. Similarly, taking $W = g^{r-1}(Y)$, there exist independent random variables $V(W, Y)$ as functions of W and Y that are independent of W so that the transformation $(W, Y) \rightarrow (W, V)$ is 1 to 1 and generates the same space of random variables. Now define the functions f_r and g_r as those maximizing the correlation $\rho(f(U)g(V))$. Since the random variables U are independent of $f^{r-1}(X)$, so is the function $f_r(U) = f_r(U(W, X)) = f_r(U(f^{r-1}(X), X))$ that is a function only of X. Similarly $g_r(V)$ is a function only of Y and is independent of $g^{r-1}(Y)$. Thus the theorem holds for r. By construction in the proof of Theorem 7, it is trivial to construct transformations from each of $f_r(U)$ and $g_r(V)$ to gaussian variables, so that the canonical variables are joint gaussian random variables. Thus the independent canonical variables can be selected as gaussian under the assumed regularity conditions. ■

IX. NONLINEAR MARKOV PROCESSES

In this section, various aspects and representations of Markov processes are developed. The fundamental properties of the state of a Markov process are reviewed. Given a state for a Markov process, the development of the state space innovations representation is immediate.

A fundamental concept in the CVA approach is the past and future of a process. Suppose that data are given consisting of observed outputs y_t and observed inputs u_t at time points labeled $t = 1, \ldots, N$ that are equal spaced in time. Associated with each time t is a past vector p_t consisting of the past outputs and inputs occurring prior to time t as well as a future vector f_t consisting of outputs at time t or later, specifically

$$p_t = (y_{t-1}^T, y_{1-2}^T, \ldots, u_{t-1}^T, u_{1-2}^T, \ldots)^T, \quad f_t = (y_t^T, y_{t+1}^T, \ldots)^T \tag{32}$$

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

$$x_t = C_t(p_t) \tag{33}$$

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

$$P\{f_t|p_t\} = P\{f_t|x_t\} \tag{34}$$

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 q_t denote the future inputs $q_t^T = (u_t^T, u_{t+1}^T, \ldots,)$ and consider the conditional random variable $f_t|q_t$. Then the process is a controlled Markov processes of order k if there exists a k-order state such that the conditional distribution of $f_t|q_t$ given the past p_t is identical to the conditional distribution of $f_t|q_t$ given the state x_t so

$$P\{(f_t|q_t)|p_t\} = P\{(f_t|q_t)|x_t\} \tag{35}$$

This is equivalent of the statement that

$$P\{f_t|(q_t, p_t)\} = P\{f_t|(q_t, x_t)\} \tag{36}$$

Now suppose that the state x_t is given from CVA of the past and future as in the previous sections with $X = p_t$ and $Y = f_t$, 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 ν_t 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

$$\nu_t = y_t - E(y_t|x_t) \tag{37}$$

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 (Larimore, 1988).

Theorem 9: State Space Representation. Suppose that the joint and marginal densities among p_t, f_t, q_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

$$x_{t+1} = \phi(x_t, u_t, \nu_t) \tag{38}$$

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

$$y_t = \mu_t(x_t, u_t) + \nu_t \tag{39}$$

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

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

$$x_{t+1} = E(x_{t+1}|(\nu_t, u_t, x_t)) = \phi_t(\nu_t, u_t, x_t) \tag{40}$$

using the conditional expectation operator $E\{\cdot|\cdot\}$.

For nonlinear processes for which Theorem 8 holds, an independent CVA (ICVA) procedure parallels that of the linear CVA. It is first necessary to remove the effect of future inputs q_t on future outputs f_t. This is done by fitting a nonlinear ARX model involving linear combinations of nonlinear functions of the past. The ICVA then results in a likelihood function given by (11) involving the canonical variables as states and a normally distributed error. It is then a matter of determining the nonlinear functions ϕ of (38) and μ of (39) by a nonlinear regression of x_{t+1} and y_t on x_t and u_t. Future work will involve the investigation of general conditions for existence of the ICVA and computational methods for obtaining the canonical variables.

X. APPLICATIONS OF CVA

A. Adaptive Control of Aircraft Wing Flutter

A particularly impressive application involved a wind tunnel test of on-line adaptive control of unstable aircraft wing flutter using CVA system identification and linear quadratic gaussian (LQG) control design (Larimore and Mehra, 1984; Peloubet et al., 1990). The system had 2 inputs, 6 outputs, and substantial wind tunnel turbulence. This example illustrates the use of a single system identification and control design procedure to successfully identify over 100,000 multivariable systems with up to 30 states for a wide range of system dynamics and structural configurations. There were no failures of the system identification algorithm that was operating completely automatically (Peloubet et al, 1990).

In the wing flutter problems, the dynamic system is a globally nonlinear systems. However, an approximate linear system about the operating point provides a useful model for feedback control. The linearized dynamic model can be determined directly and automatically from the measured input and output data. Then the identified dynamic model can be used to automatically design a feedback control to suppress the vibration. For the wind tunnel test, most of the time during the test the model aircraft was flying beyond the critical flutter speed where the flutter dynamics were unstable. The LQG controller developed from the CVA identified model was being used during the whole experiment to stabilize the model dynamics that were unstable most of the time.

In the final days of the three weeks of windtunnel testing, a particularly unstable configuration of stores (tanks, missiles, etc) were hung from the wing for testing. The testing was conducted as usual to the point were it appeared close to being unstable even with the real-time system identification and on-line control system design and feedback. At one point the system went unstable and broke the wing ending the testing. In post analysis, it was determined that the controller in combination with the identified wing flutter model was stable but only very marginally so. Of course the identified model is only an estimate of the true wing flutter dynamics and has some error associated with it. From comparison of the dispersion of repeated identifications of the wing flutter, it was determined that the model identification error exceeded the robustness of the associated controller. As a result, the controller in combination with the true flutter dynamics was unstable. Measurements indicated that there was amplitude doubling every 4 cycles, or equivalently every 0.25 seconds.

The lesson learned was that it is critical to have an assessment of the accuracy of the identified model, and to use this in determining the robustness of the controller to modeling errors. At the time of the tests, there was no means of doing this. Subsequently, methods for computing confidence bands for the accuracy of the identified model were developed (Larimore, 1993) based on the assumed maximum likelihood accuracy of CVA that was proven in Larimore (2004).

B. Monitoring and Fault Detection

Several monitoring and fault detection methods have been developed based on the ML properties of CVA. Larimore (1997a) developed a likelihood ratio test, ΔAIC, for model change. It was used by Wang et al. (1997)in CSTR simulations, and distribution theory was developed by Conner et al. (2004) showing that it is optimal. Juricek et al. (2004) compared several failure detection methods based on CVA methods to conventional methods.

C. Nonlinear Identification of the Lorenz Attractor

The Markov process considered is the Lorenz attractor (Lorenz, 1963) with process excitation noise. 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

$$
\begin{aligned}
x_{t+1}^{(1)} &= x_t^{(1)} + \Delta t\sigma(x_t^{(2)} - x_t^{(1)}) + n_t^{(1)} && (41)\\
x_{t+1}^{(2)} &= x_t^{(2)} + \Delta t[\rho x_t^{(1)} - x_t^{(2)} - x_t^{(1)}x_t^{(3)}] + n_t^{(2)} && (42)\\
x_{t+1}^{(3)} &= x_t^{(3)} - \Delta t[\beta x_t^{(3)} + x_t^{(1)}x_t^{(2)}] + n_t^{(3)} && (43)
\end{aligned}
$$

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 noise covariance matrix of the white process excitation noise $(n_t^{(1)}, n_t^{(2)}, n_t^{(3)})^T$ used in the simulation is $10^{-2} \times I_3$ 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 are $y_t = x_t^{(1)}$, the first component of the discretized Lorenz process observed at 1000 time points. The presence of the noise on the process was very noticeable. It is show (Larimore, 1989, 1992b) that the entire 3-dimensional dynamics of the process can be reconstructed from the measured first component.

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

$$
f_{i_1,i_2,i_3}(y_{t-1}, y_{t-2}, y_{t-3}) = y_{t-1}^{i_1} y_{t-2}^{i_2} y_{t-3}^{i_3} \quad \text{for } i_1 + i_2 + i_3 \leq 3 \qquad (44)
$$

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

$$f_t = (f_t, \ldots, f_{t+20})^T \tag{45}$$

A canonical variate analysis of sample covariances of the past and future are given in Table I. 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. The canonical state is chosen as the first 5 canonical variables.

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 I

Canonical Correlations for Lorenz Attractor.

D. Other Applications of Nonlinear CVA

Several other research teams have independently applied the methodology of Larimore (1989) including Verhoeven et al (2002), Pilgram et al (2000), and DeCicco and Cinar (2000). Verhoeven et al (2002) and Pilgram et al (2000) apply NLCVA to financial data of the S& P500 and foreign exchange rates respectively. They compare the NLCVA model with a GARCH(1,1) model of Engle (1982) and Bollerslev (1986) for modeling volatility. The GARCH is a model developed specifically for modeling financial data volatility, and has been used with great success. Engle recently received the Nobel prize for his contributions to ARCH and GARCH modeling of volatility in financial systems. What is intriguing is that the NLCVA model that is identified with no knowledge about financial systems or prior information performs significantly better than the GARCH model that was developed and refined over a number of years specifically for that purpose. This suggests at least in some cases that black box nonlinear modeling can outperform detailed models that are constructed with a particular structure based upon somewhat imperfect prior information.

References

[1] 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.

[2] 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.

[3] Anderson, T.W. (1984), An Introduction to Multivariate Statistical Analysis, New York: Wiley.

[4] Aoki, M. (1987). State Space Modeling of Time Series. Springer Verlag

[5] Bach, F.R., and M.I. Jordan (2002), “Kernel Independent Component Analysis”, J. Machine Learning Res., Vol. 3, pp. 1-48.

[6] Bauer, D. (1998). Some asymptotic theory for the estimation of linear systems using maximum likelihood methods or subspace algorithms. Ph.D. thesis, Technischen Universitat Wien, Austria

[7] Bauer, D. and L. Ljung (2002), “ Some facts about the choice of the weighting matrices in Larimore type of subspace algorithms”, Automatica, Vol. 38, pp. 763-773.

[8] Bauer, D. (2005) “Comparing the CCA subspace method to pseudo maximum likelihood methods in the case of no exogenous inputs”, accepted for publication, J. Time Series Analysis.

[9] Bollerslev, T. (1986), “Generalized autoregressive conditional heteroskedasticity”, J. Econometrics, Vol. 31, pp. 307-327.

[10] 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.

[11] Candy, J.V., Bullock, T.E., and Warren, M.E. (1979), “Invariant Description of the Stochastic Realization,” Automatica, Vol. 15, pp. 493-5.

[12] Conner, J.S., D.E. Seborg, and W.E. Larimore (2004), “A Theoretical Analysis of the DeltaAIC Statistic for Optimal Detection of Small Changes”, Proc. 2004 American Control Conference, June 30 - July 2, Boston MA.

[13] Cox, D.R., and D.V. Hinkley (1974), Theoretical Statistics, New York: Chapman and Hall.

[14] 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.

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

[16] Dahlen, A., A. Liindquist, and J. Mari (1998), “Experimental evidence showing that stochastic subspace identification methods may fail,” Systems and Control Letters. Vol. 34, pp 303-312

[17] Dahlen, A. (2001), “Identification of stochastic systems: Subspace methods and covariance extension”, Doctoral Thesis, Royal Institute of Technology, Stockholm, Sweden.

[18] DeCicco, J., and A. Cinar (2000) "Empirical Nonlinear Dynamic Modeling of Processes with Output Multiplicities", Proc. American Control Conference, held June 28-30, Chicago, IL, pp. 2265-2269

[19] Deistler, M., K. Peternell and W. Scherrer (1995), “Consistency and Relative Efficiency of Subspace Methods,” Automatica, Vol. 31, pp. 1865-1875.

[20] Engle, R.F. (1982), “Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation”, Econometrica, Vol. 50, pp. 987-1008.

[21] Friedman, J.H. (2004), “Recent advances in predictive (machine) learning”, SLAC-PUB-10321, Stanford Linear Accelerator Center, Stanford, CA, January 2004. Automatica, Vol. 31, pp. 1317–1324

[22] 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.

[23] Gustavsson, G., L. Ljung, and T. Soderstrom (1977), “Identification of Processes in Closed Loop – Identifiability and Accuracy Aspects,” Auotmatica, Vol. 13, No. 1, pp. 59-75.

[24] Hastie, T., R. Tibshirani, and J. Friedman (2001), The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer, NY.

[25] Hotelling, H. (1936). “Relations Between Two Sets of Variates”, Biometrika, Vol. 28, pp. 321-377.

[26] Juricek, B.C., D.E. Seborg, and W.E. Larimore (2001), “Identification of the Tennessee Eastman Challenge Process with Subspace Methods,” Control Engineering Practice, Vol. 9, pp. 1337-1351.

[27] Juricek, B.C., W.E. Larimore, and D.E. Seborg (2002), "Reduced Rank ARX and Subspace System Identification for Process Control", Ind. & Eng. Chem. Research 41, 2185-2203. Also appeared in Proc. IFAC DYCOPS Sympos., 245-250, Corfu, Greece (1998).

[28] Juricek, B.C., D.E. Seborg, and W.E. Larimore (2004), “Fault Detection Using Canonical Variate Analysis,” Ind. Eng. Chem. Res., Vol. 43, pp. 458-474.

[29] Juricek, B.C., D.E. Seborg, and W.E. Larimore (2005), “ Process Control Applications of Subspace and Regression-Based Identification and Monitoring Methods,” Proc. American Control Conference, Portland, OR.

[30] Kuss, M., and T. Graepel (2003), The geometry of kernel canonical correlation analysis, Technical Report No. 108, Max Plank Institute for Biological Cybernetics, Tubingen, Germany.

[31] Lacy, S.L., V. Babuska, K.N. Schrader, and R. Fuentes (2005), “System Identification of Space Structures,” Proc. American Control Conference, Portland, OR.

[32] Lancaster, H.O. (1966), “Kolmogorov’s Remark on the Hotelling Canonical Correlations,” Biometrika, Vol. 53, pp. 585-588.

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

[34] Larimore, W.E. (1983a). “Predictive Inference, Sufficiency, Entropy, and an Asymptotic Likelihood Principle”, Biometrika, Vol. 70, pp. 175-81.

[35] 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.

[36] 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. Held December 12-14, 1984, Las Vegas, NV. New York: IEEE.

[37] Larimore, W.E. and R.K. Mehra (1984), “Technical Assessment of Adaptive Flutter Suppression Research,” Air Force Wright Aeronautical Lab Report No. AFWAL-TR-84-3052.

[38] Larimore, W.E., W.M. Lebow, and R.K Mehra (1985), “Identification of Parameters and Model Structure for Missile Aerodynamics,” American Control Conference, Eds. Y. Bar-shalom and D. Wormley, Vol. 1, pp. 18-26. Piscataway, N.J.: IEEE Service Center.

[39] Larimore, W.E. and R.K. Mehra (1985), “The Problem of Overfitting Data,” Byte, Vol. 10, pp. 167-80.

[40] Larimore, W.E. (1988). “Generalized Canonical Variate Analysis of Nonlinear Systems”, Proceedings of the 27th IEEE Conference on Decision and Control, Vol. 3, pp. 1720-5, held December 7-9, 1988, Austin, TX.

[41] Larimore, W.E. (1989), “System Identification and Filtering of Nonlinear Controlled Markov Processes By Canonical Variate Analysis, ” Final Report for Air Force Office of Scientific Research, Computational Engineering, Inc, 1989. Summarized in Larimore (1992b).

[42] Larimore, W.E. (1990a), “Canonical Variate Analysis for System Identification, Filtering, and Adaptive Control, ” Proc. 29th IEEE Conference on Decision and Control, Honolulu, Hawaii, December, Vol. 1, pp. 635-9.

[43] Larimore, W.E. (1990b), “Order-Recursive Factorization of the Pseudoinverse of a Covariance Matrix”, IEEE Trans. of Automatic Control, Vol. 35, pp. 1299-1303.

[44] Larimore, W.E. (1992a), ADAPT$_X$ Automated System Identification Software Users Manual, Adaptics, Inc, 40 Fairchild Drive, Reading, MA 01867.

[45] Larimore, W.E. (1992b), “Identification and Filtering of Nonlinear Systems Using Canonical Variate Analysis,” Nonlinear Modeling and Forecasting, Eds. M. Casdagli and S. Eubank, pp. 283-303. Reading, MA: Addison-Wesley.

[46] Larimore, W.E. (1993), “Accuracy Confidence Bands Including the Bias of Model Under-fitting,” Proc. 1993 American Control Conference, San Francisco, CA, June 2-4, 1993, Vol. 2, pp. 1995-9.

[47] Larimore, W.E. (1996a), “Statistical Optimality and Canonical Variate Analysis System Identification,” Signal Processing, Vol. 52, pp. 131-144.

[48] Larimore, W.E. (1996b), “Optimal Order Selection and Efficiency of Canonical Variate Analysis System Identification,” Proc. 13th IFAC World Congress, Vol. I, San Francisco, July 1-5, 1996, Vol. I, pp. 151-156.

[49] Larimore, W.E. (1997a), “Optimal Reduced Rank Modeling, Prediction, Monitoring, and Control Using Canonical Variate Analysis,” IFAC Internat. Symp. on Advanced Control of Chemical Processes, Banff, Canada, June 9-11, 1997.

[50] Larimore, W.E. (1997b), “System Identification of Feedback and ‘Causality’ Structure using Canonical Variate Analysis,” Preprints 11th IFAC Symposium on system Identification, held July 8-11, 1997, Fukuoka, Japan, Vol. 3, pp. 1101-6.

[51] Larimore, W.E. (1999), “Automated Multivariable System Identification and Industrial Applications,” Proc. 1999 American Control Conference, June 24, 1999, San Diego, CA, pp. 1148-1162.

[52] Larimore, W.E. (2002). Reply to ’Comment on ’Order-recursive factorization of the pseudoinverse of a covariance matrix’ ’. IEEE Trans. Automat. Contr., 47, pp. 1953-7.

[53] Larimore, W.E. (2003). "Inferring Multivariable Delay and Seasonal Structure for Subspace Modeling". Preprints 13th IFAC Symposium on System Identification, held August 27-29, 2003, Rotterdam, Netherlands.

[54] Larimore, W.E. (2004), “Large Sample Efficiency for ADAPTx Subspace System Identification with Unknown Feedback”, to be presented at IFAC DYCOPS, to be held July 5-7, Boston, MA.

[55] Ljung, L and T. McKelvey (1996), “Subspace identification from closed loop data”, Signal Processing, Vol. 52, pp. 209-215.

[56] Lorenz, E.N. (1963). “Deterministic Nonperiodic Flow”, J. Atmospheric Sciences, Vol. 20, pp. 130-41.

[57] Lütkepohl, H. (1993), Introduction to Multiple Time Series Analysis, Second Edition, New York: Springer-Verlag.

[58] Palanthandalam-Madapusi, H., S.L. Lacy, J.B. Hoagg, and D.S. Bernstein (2005), “Subspace-Based Identification for Linear and Nonlinear Systems,” Proc. American Control Conference, Portland, OR.

[59] Peloubet, R.P., R.L. Haller, R.M Bolding (1990), “On-line Adaptive Control of Unstable Aircraft Wing Flutter, ” Proc. 29th IEEE Conference on Decision and Control, Honolulu, Hawaii, December, Vol. 1, pp. 643-51.

[60] Peternell, K., W. Scherrer, and M. Deistler (1996), “Statistical analysis of novel subspace identification methods”, Signal Processing, Vol. 52, pp. 161-177.

[61] Pilgram, B., P. Verhoeven, A. Mees, and M. McAleer (2002), “Nonlinear Markov modelling using canonical variate analysis: forecasting exchange rate volatility”, Department of Mathematics and Statistics, University of Western Australia, WA 6907.

[62] Renyi, A. (1959), “On Measures of Dependence”, Acta. Math. Acad. Sci. Hungar., Vol. 10, pp. 441-451.

[63] Schaffé, H. (1959), The Analysis of Variance, New York: Wiley.

[64] Scholkopf, B., A.J. Smola, and K.R. Muller (1998), “Nonlinear component analysis as a kernel eigenvalue problem”. Neural Computation, Vol. 10, pp. 1299-1319.

[65] Scholkopf, B., and A.J. Smola (2002), Learning with Kernels, cambridge, MA: MIT Press.

[66] Schweppe, F.C. (1965), “Evaluation of Likelihood Functions for Gaussian Signals,” IEEE Trans. Inf. Theory Vol 11, pp. 61-70.

[67] Shi, R. (2001), Ph.D thesis, McMaster University, ON, Canada.

[68] Shi, R., and J.F. MacGregor (2001), “A Framework for Subspace Identification Methods,” Proc. American Control Conference, Arlington, VA, June 25-7, pp. 3678-83.

[69] Van Overschee, P., and B. De Moor (1994), “A Unifying Theorem for Three Subspace System Identification Algorithms,” American Control Conf., pp. 1645-1649, June 29-July 1, 1994, Baltimore, MD.

[70] Van Overschee, P., and B. De Moor (1996), Subspace Identification - Theory - Implementation - Applications. Kluwer Academic Publishers.

[71] Verhoeven, P., B. Pilgram, M. McAleer, and A. Mees (2002), “Nonlinear modelling and forecasting of S&P 500 volatility”, Mathematics and Computers in Simulation, Vol. 59, pp. 233-241.

[72] Wang, Y., D.E. Seborg, and W.E. Larimore (1997), "Process Monitoring Using CVA and PCA", Proc. IFAC ADCHEM Sympos., pp. 523-528, Banff, Canada.

[73] Whittle, P. (1954), “On Stationary Processes in the Plane,” Biometrika, Vol. 41, pp. 434-449.