LARGE SAMPLE EFFICIENCY FOR ADAPTX SUBSPACE SYSTEM IDENTIFICATION WITH UNKNOWN FEEDBACK

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. Over the last two decades, the canonical variate analysis method for subspace system identification has been widely applied. A number of these applications have demonstrated near maximum likelihood accuracy of the `adaptx` CVA subspace algorithm in large samples with unknown feedback. The critical step in the algorithm is the use of an ARX model estimated by conditional maximum likelihood to remove the effects of future inputs on future outputs. It is shown that the subspace estimates can be considered as restrictions on the ML ARX estimates to a subspace of the parameters obtained by projection methods. As a result, the errors between the models are orthogonal to the subspace model, and the subspace parameter estimates are asymptotically ML. A critical step in showing this orthogonality is use of the multistep form of the likelihood function.

Key Words: Large sample, Subspace identification, Canonical variate analysis, Closed-loop, Feedback, Multistep likelihood.

1. OVERVIEW

The purpose of this paper is to give an outline 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; 2004).

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 $\ell\mathrm{Dim}(y_t)$, 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.

2. ML ESTIMATION UNDER FEEDBACK

In this section, the maximum likelihood estimation of input-output models under the effect of unknown feedback is discussed. ML estimation of models has the considerable advantage of being immune to the presence of colored inputs or feedback.

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, $Y_1^N = (y_N, \dots, y_1)$ 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

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

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 (1) 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.

3. 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

$$y(t) = \sum_{s=1}^{\ell} \alpha(s) y(t-s) + \sum_{s=0}^{\ell} \beta(s) u(t-s) + e(t) \tag{2}$$

for $t = \ell + 1, \dots, N$, 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 (2) are used for $t = \ell + 1, \dots, N$, and are transposed and stacked up to give

$$Y = Z\Theta + E \tag{3}$$

where $Y^T = [y_{\ell+1}, \dots, y_N]$ with the first ℓ observations of the output not used in the regression so it is conditional on the first ℓ observations. Also denote $\Theta^T = [\alpha_1, \dots, \alpha_\ell, \beta_0, \dots, \beta_\ell]$ 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 (3) 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 (2) 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

$$\hat{Y}_i = Z\hat{\Theta}_{*i} = Z_1\hat{\Theta}_{1i} + \cdots + Z_m\hat{\Theta}_{mi} \tag{4}$$

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 (4) 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

$$\hat{Y} = Z\Theta + Z(Z^TZ)^{-1}Z^TE \tag{5}$$

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 parameter estimates $\hat{\Theta}_S$ and $\hat{\Theta}_{A-S}$ are uncorrelated.

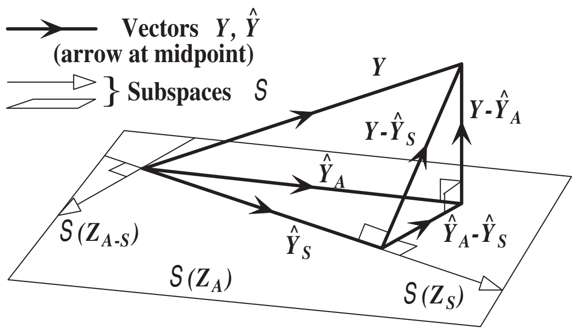

Figure 1: Nested Subspaces and Orthogonality Relationships.

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

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

$$y_t = Hx_t + Au_t + Bw_t + v_t \tag{7}$$

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

$$x_t = \sum_{i=1}^{\infty}(\Phi - KH)^{i-1}[(G - KA)u_{t-i} + Ky_{t-i}] \tag{8}$$

that results from recursively substituting (6) for x_t in (6). Eq. (8) is equivalent to (6) provided that $J_k^\infty(\Theta)$ is parameterized as in (6) and (7). 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 (7) with $x_t = J_kp_t$ is in the ARX form (2) with additional restrictions on the parameters, the Markov model (6) and (7) 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 (8). 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 (6) and (7) with the state given by $x_t = \hat{J}_kp_t$. In particular, let X^+ denote X with the time index t replaced by $t + 1$, and project $(X^+\ Y)$ on

$S(X\ U)$ 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 (6) and (7). The latter two involve nonlinear parameterizations, and the projection methods developed above are generalized in the next section to such nonlinear parameterizations.

4. SUBSPACE DETERMINATION BY CVA

In this section, the CVA procedure for the determination of the state space subspace is developed. The main difficulty in the case of feedback is to remove the effects of future inputs on future outputs. This is done using the estimated ARX model and a multistep form of the likelihood function.

The future $f_t = (y_t^T, \ldots, y_{t+\ell}^T)^T$ of the process is related to the past p_t through the state x_t and the future inputs $q_t = (u_t^T, \ldots, u_{t+\ell}^T)^T$ in the form

$$f_t = \Psi^T x_t + \Omega^T q_t + e_t \qquad (9)$$

where x_t lies in some fixed subspace of p_t, $\Psi^T = (H; H\Phi; \ldots; H\Phi^{\ell-1})$ and the i, j-th block of Ω is $H\Phi^{j-i}G$. The presence of the future inputs q_t creates a major problem in determining the state space subspace from the observed past and future. If the term $\Omega^T q_t$ could be removed from the above equation, then the state space subspace could be estimated accurately. The method used in the `adaptx` algorithm is to fit an ARX model and compute an estimate $\hat{\Psi}$ of Ψ based on the estimated ARX parameters. Note that an ARX process can be expressed in state space form with state $x_t = p_t$ so that the above expressions for Ω and Ψ in terms of the state space model can be used as well for the ARX model. Then the ARX state space parameters (Φ, G, H, A) and Ψ and Ω are themselves functions of the ARX model parameters $\hat{\Theta}_A$.

In Larimore (1997a), the determination of the state space subspace by CVA was developed from basic principles and the maximum likelihood method with a rank constraint. This leads to a CVA between the past p_t and $f_t - \Omega^T q_t$, the future outputs with the effects of future inputs removed. To justify the CVA for a correlated time series, a multistep ahead form of the likelihood function was derived. The log likelihood function of the outputs $Y_{\ell+1}^N$ conditional on the inputs U_1^N and the past $p_{\ell+1}$ at time $\ell + 1$ is of the form

$$\begin{aligned} &\log p(Y_{\ell+1}^N | p_{\ell+1}, Q, \Theta, \Sigma) \\ &= \frac{1}{\ell} \sum_{t=\ell+1}^{N} \log p((f_t - \Omega^T(\Theta) q_t) | p_t, \Theta, \Sigma)) \end{aligned} \qquad (10)$$

To be exact, a maximized likelihood function should be satisfied by the ML estimates $\hat{\Theta}_S$ for the SS model. But of course these estimates are not available for computing $\Omega(\hat{\Theta}_S)$ at the point of trying to determine the state space subspace, so the corrected future cannot be computed.

Now consider removing the effect of future inputs from future outputs using the ARX parameter estimates $\hat{\Theta}_A$. The first point is that the multistep likelihood function (10) is asymptotically equivalent to the one-step prediction likelihood (see Bauer, 2004). Thus parameter estimates from the two are asymptotically equivalent and can be used interchangeably. The estimates of the ARX parameters Θ_A do not involve the estimates of the covariance matrix Σ_A so Θ_A can be estimated separately.

Consider the model (9) of the multistep output f_t with $\Psi(\Theta)$ and $\Omega(\Theta)$ nonlinear in the parameters Θ that appear in the multistep likelihood function (10). The notation $F = (f_{\ell+1}, \ldots, f_{N-\ell})^T$ is used where each column of F is one component of the future vector f_t. Then for an estimate $\hat{\Theta}$, the model for F is

$$\hat{F} = X\Psi(\hat{\Theta}) + Q\Omega(\hat{\Theta}) = W\Pi(\hat{\Theta}) \qquad (11)$$

with $\Pi(\hat{\Theta}) = (\Psi(\hat{\Theta})\ \Omega(\hat{\Theta}))$. As in the linear regression case, $\hat{F}$ lies in the subspace span by the columns of $W = (X\ Q)$.

Suppose the Θ_A and Θ_S are two nested models with $\Theta_S \subset \Theta_A$. As in the linear regression case, consider the refinement of the ML parameter estimate $\hat{\Theta}_A$ by restriction to the model $\hat{\Theta}_S$ that is assumed to be true. The orthogonality condition

$$F(\hat{\Theta}_S)[F(\hat{\Theta}_A) - F(\hat{\Theta}_S)] = 0$$

holds asymptotically for large sample in the nonlinear case. This is proven using results from Magnus and Neudecker (1988) where it is shown that the first differential of the likelihood function leads to the above orthogonality condition in the case of univariate nonlinear regression, and is easily extended to multivariate nonlinear regression. Thus as in the linear regression case, the error $F(\hat{\Theta}_A) - F(\hat{\Theta}_S)$ in the estimate is orthogonal to the estimate $F(\hat{\Theta}_S)$ asymptotically for large sample. For any two columns $\hat{\Pi}_i = \Pi_i(\hat{\Theta})$ and $\hat{\Pi}_j = \Pi_j(\hat{\Theta})$ and where σ_{ij} applies to e_t of (9), it can be shown that asymptotically

$$\sigma_{ij}(W^T W)^{-1} = \text{Cov}(\hat{\Pi}_i, \hat{\Pi}_j)) = \frac{\partial \Pi_i}{\partial \Theta^T} \text{Cov}(\hat{\Theta}, \hat{\Theta}) \frac{\partial \Pi_j^T}{\partial \Theta} \qquad (12)$$

In the last section, the parameter space Θ_A is reparameterized by Θ_S and the difference Θ_{A-S} that are uncorrelated. Let $\Pi(\Theta_S)$ or $\Pi(\Theta_{A-S})$ denote Π as a function with Θ restricted to the respective subspaces Θ_S and Θ_{A-S}, and let W_S and W_{A-S} be the corresponding subspaces of W. Using (12), it can be shown that asymptotically W_S and W_{A-S} are orthogonal if and only if $\hat{\Theta}_A$ and $\hat{\Theta}_{A-S}$ are uncorrelated.

If $\Omega(\hat{\Theta}_S)$ was known, then a canonical variate analysis between $f_t - \Omega(\hat{\Theta}_S)^T q_t$ and the past p_t would estimate the state space subspace. If the ARX estimate $\hat{\Theta}_A$ is used instead of $\hat{\Theta}_S$, then from (9)

$$f_t - \Omega(\hat{\Theta}_A)^T q_t = \Psi(\hat{\Theta}_S)^T x_t - [\Omega(\hat{\Theta}_A) - \Omega(\hat{\Theta}_S)]^T q_t$$
$$+[\Psi(\Theta_S) - \Psi(\hat{\Theta}_S)]^T x_t + [\Omega(\Theta_S) - \Omega(\hat{\Theta}_S)]^T q_t + \upsilon_t$$

where the first term is the projection on the ML estimated state $\hat{x}_t$ and the second is error between the ML ARX and SS estimates in removing future input effects, with the first term orthogonal to the second. Remaining terms are due to ML estimation errors that are minimum and the error υ_t. From the above, it can be shown that the CVA using $f_t - \Omega(\hat{\Theta}_A)^T q_t$, the future corrected with the ARX rather than the SS ML parameter estimates, and the past p_t to determine the state space subspace introduces no bias or added variance asymptotically. This reduces the case of inputs and feedback to the no inputs case, that achieves the ML lower bound asymptotically (Bauer, 2004).

REFERENCES

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

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

Bauer, D. (2004) “Comparing the CCA subspace method to pseudo maximum likelihood methods in the case of no exogenous inputs”, submitted to *J. Time Series Analysis*.

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

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

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

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

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

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.

Larimore, W.E. (1983). “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., 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.

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.

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

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

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.

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

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.

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*, held June 9-11, Banff, Canada, pp. 61-6.

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.

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

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.

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

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

Magnus, J.R., and H. Neudecker (1988), *Matrix Differential Calculus with Applications in Statistics and Econometrics*, New York: Wiley.

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

Schaffé, H. (1959), *The Analysis of Variance*, New York: Wiley.

Shi, R., Ph.D thesis, McMaster University, ON, Canada, 2001

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.

Wang, Y., D.E. Seborg and Larimore (1996), “Process Monitoring based on Canonical Variate Analysis,” *IFAC 1997 Int. Symp. on Advanced Control of Chemical Processes* held June 9-11, 1997, Banff, Canada, pp. 523-8.