SUBSPACE ALGORITHMS

Dietmar Bauer *,1

* *Institute for Econometrics, Operations Research and System Theory, TU Wien, Argentinierstr. 8, A-1040 Vienna, Austria*

Abstract: Subspace algorithms have been established in the last decades as an alternative to prediction error methods for the estimation of linear dynamical systems. Conceptual simplicity and numerical feasability have been the main arguments in favor of the approach. This article gives a presentation of the mainstream approach and tries to convince the reader, that this class of algorithms has its virtues. Strengths and weaknesses of the approach are discussed.

Keywords: subspace algorithms, estimation, linear dynamical systems

1. INTRODUCTION

'Subspace algorithms' is a technical term, which is both, too broad and misleading. Too broad, since the term is used in many different contexts in totally different meanings. Misleading, because even if one adds the context the term is not connected to a particular algorithm or class of algorithms, but rather to a general idea. Subspace algorithms have their origins in the algorithms of Zeiger and McEwen (1974) and Ho and Kalman (1966). As such, they bear elements of realization algorithms. However, the main idea centers around the concept of the state, as being an interface – in a sense to be made more clear below – between the past and the future, stated loosely. These early ideas have been developed further leading to the three most well known algorithms:

- `N4SID` (numerical algorithms for subspace state space system identification) proposed by Van Overschee and DeMoor (1994)
- `MOESP` (multivariable output error state space) system identification procedure proposed by Verhaegen (1994)
- `CCA` (canonical correlation analysis) proposed as `CVA` (canonical variate analysis) by Larimore (1983).

All three of them are used in the context of linear dynamical systems operating in open loop. Following the suggestion of the algorithms in parallel the analysis of the properties of these algorithms and the adaptation to different model classes occured. The general idea of the method has been adapted to lead to algorithms for the closed loop case (Chou and Verhaegen, 1999; Verhaegen, 1993; Ljung and McKelvey, 1996*b*), frequency domain data (McKelvey, 1995), bilinear models (Favoreel, 1999; Chen and Maciejowksi, 2000; Chou, 1994), piecewise linear models (Babuska *et al.*, 1997), time-varying parameters (Gustafsson, 1999; Verdult and Verhaegen, 2002; Oku and Kimura, 2002), Hammerstein models (Gomez and Baeyens, 2002), continuous-time models (Haverkamp *et al.*, 1997; Ohsumi and Kawano, 2002), errors-in-variables problems (Chou and Verhaegen, 1997), integrated processes (Bauer and Wagner, 2002), hidden markov chains (Andersson, 2002). Here we will only discuss the case of stationary, linear, discrete time, time invariant systems.

The aim of this paper is to present the concept of subspace algorithms in a unified way in order to highlight

[1] Support by the Austrian FWF under the project number P14438-INF is gratefully acknowledged.

the similarities between the various algorithms. The discussion will present the algorithms in much detail, while trying to keep the exposition self contained in order to allow also readers from related areas to follow. At many places the comparison to prediction error methods will be considered, since the subspace methods are an alternative to these methods. It is the purpose of this paper to point out situations, where there are advantages of the subspace approach over the prediction error approach.

2. STATE SPACE MODELS

In this paper the model class considered will always be the class of linear, discrete time, finite dimensional, time invariant state space models, given by

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + K\varepsilon_t \\ y_t &= Cx_t + Du_t + \varepsilon_t \end{aligned} \tag{1}$$

Here $(y_t)_{t\in\mathbb{Z}}$ denotes the s-dimensional output process, observed for $t = 1,\dots,T$, $(\varepsilon_t)_{t\in\mathbb{Z}}$ denotes the s-dimensional innovations, which for simplicity are assumed to be i.i.d. Gaussian random variables with zero mean and variance matrix $\Omega > 0$. As usual the noise is assumed to be unobserved. Furthermore $(u_t)_{t\in\mathbb{Z}}$ denotes the m-dimensional input process, observed for $t = 1,\dots,T$. The n dimensional state process $(x_t)_{t\in\mathbb{Z}}$ is also not observed. The system matrices $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{s\times n}, D \in \mathbb{R}^{s\times m}$ and $K \in \mathbb{R}^{n\times s}$ are to be estimated.

In the following some important properties of state space systems are discussed. Since it is assumed, that these concepts are known to all readers, the discussion is very brief. For a more detailed discussion we refer to (Hannan and Deistler, 1988, Chapter 1). Central to the definition of the model is the concept of a state: x_t is introduced in order to describe all the dynamics present in the model, as the observation equation is a static one. For so called *white box* models derived on the basis of physical principles, the state possesses a specific interpretation. In this talk only the so called *black box* modelling approach will be considered. Here only the input/output map is of interest and the state does not have any physical meaning, but is only a mathematical object to conveniently decsribe the dynamics of the system. As such, the state of a system is not unique: Any change from x_t to $z_t = Tx_t$ using a nonsingular matrix $T \in \mathbb{R}^{n\times n}$ results in a different model $(TAT^{-1}, TB, CT^{-1}, D, TK)$, which represents the same input/output map. In this case, (A,B,C,D,K) and $(TAT^{-1}, TB, CT^{-1}, D, TK)$ are called *observationally equivalent*. A state space representation of an input/output map is called *minimal*, if there exists no state space representation of the same input/output map with lower state dimension. In that case the integer n is called the *order* of the system.

The significance of the state in the state space models comes from the fact, that it summarizes all the dynamics in the model: Given the state trajectory, the ouput is obtained from the static observation equation. As a note we remark, that state space systems hence can be seen as very special hidden markov models. The state is not observed, but given the model and trajectories of the input u_t and the output y_t for $t = 1,\dots,T$, the state x_{T+1} can be estimated. Assume, that the system is stable and strictly minimum-phase, i.e. that $|\lambda_{max}(A)| < 1$ and $|\lambda_{max}(A-KC)| < 1$ hold. Here $\lambda_{max}(.)$ denotes an eigenvalue of maximal modulus of a matrix. The best estimate of the state in the mean square sense (assuming the input to be a covariance stationary stochastic process) is calculated by the Kalman filter. If the input/output data is available for $t = T, T-1, T-2, \dots$, then the steady state Kalman filter estimate of the state coincides with the state, since

$$x_{T+1} = \sum_{j=0}^{\infty} (A-KC)^j \left[Ky_{T-j} + (B-KD)u_{T-j}\right]$$

considering time T to be the 'present'. Therefore, the state can be recovered from the knowledge of the history of the input/output data. On the other hand, the prediction of the output is one of the main goals for identification. Consider (for $f \geq 0$)

$$\begin{aligned} y_{T+f} &= Cx_{T+f} + Du_{T+f} + \varepsilon_{T+f} \\ &= C(Ax_{T+f-1} + Bu_{T+f-1} + K\varepsilon_{T+f-1}) \\ &\quad + Du_{T+f} + \varepsilon_{T+f} \\ &= \cdots \\ &= CA^f x_T + Du_{T+f} + \varepsilon_{T+f} \\ &\quad + \sum_{j=0}^{f-1} \left(CA^j Bu_{T+f-j-1} + CA^j K\varepsilon_{T+f-j-1}\right) \\ &= CA^f x_T + \sum_{j=0}^{f} L_j u_{T+f-j} + \sum_{j=0}^{f} K_j \varepsilon_{T+f-j} \end{aligned} \tag{2}$$

where the last equation defines the impulse response sequences $L_j \in \mathbb{R}^{s\times m}, K_j \in \mathbb{R}^{s\times s}, j \geq 0$, i.e. $L_0 = D, K_0 = I, L_j = CA^{j-1}B, K_j = CA^{j-1}K, j > 0$. This equation decomposes the output y_{T+f} into three components: $CA^f x_T$ gives the contribution of the state at initial time T, $\sum_{j=0}^{f} L_j u_{T+f-j}$ the contribution of the future and the present of the input and $\sum_{j=0}^{f} K_j \varepsilon_{T+f-j}$ the contribution of the future and present of the noise. If one assumes open loop operation, then the input is uncorrelated with the noise. The state, being a function of the past input/output data also is uncorrelated with the noise. Therefore the best linear mean square prediction of y_{T+f} based on the whole input sequence $u_t, t \in \mathbb{Z}$ and the past of the output $y_s, s < T$, say $y(T+f|T)$, equals

$$y(T+f|T) = CA^f x_T + \sum_{j=0}^{f} L_j u_{T+f-j}$$

The following two facts constitute the role of the state in state space models:

(1) The state is a function of the past input/output data.
(2) The state summarizes all information contained in the past input/output measurements that is relevant for the prediction of the future output.

In this sense, the state is the interface between the past and the future. This basic fact lies at the heart of all subspace algorithms.
Choosing two integers f and p, the following vectors can be defined for arbitrary time instant t:

$$Y_{t,f}^{+} = \begin{bmatrix} y_t \\ y_{t+1} \\ \vdots \\ y_{t+f-1} \end{bmatrix} \in \mathbb{R}^{fs}, Z_{t,p}^{-} = \begin{bmatrix} y_{t-1} \\ u_{t-1} \\ \vdots \\ y_{t-p} \\ u_{t-p} \end{bmatrix} \in \mathbb{R}^{p(m+s)}$$

Additionally $U_{t,f}^{+}$ is defined using u_t analogously to $Y_{t,f}^{+}$ and $E_{t,f}^{+}$ is defined using the innovations ε_t. Let $\mathscr{K}_p$ denote the matrix corresponding to the finite Kalman filter, such that

$$n_t = x_t - \mathscr{K}_p Z_{t,p}^{-}$$

is orthogonal to $Z_{t,p}^{-}$, i.e. uncorrelated. From projection arguments in combination with ($\bar{A} = A - KC$)

$$x_t = \bar{A}^p x_{t-p} + \sum_{j=0}^{p-1} \bar{A}^j (K y_{t-j-1} + (B - KD) u_{t-j-1})$$

it follows that $\|n_t\| \leq \|(A-KC)^p\| \|x_{t-p}\|$ and therefore for p large the strict minimum-phase assumption implies that the error term n_t is small. Combining the equations (2) for $y_{t+j}, j = 0, \ldots, f-1$ one obtains the following central equation:

$$\begin{aligned} Y_{t,f}^{+} &= \mathscr{O}_f x_t + \mathscr{U}_f U_{t,f}^{+} + \mathscr{E}_f E_{t,f}^{+} \\ &= \mathscr{O}_f \mathscr{K}_p Z_{t,p}^{-} + \mathscr{U}_f U_{t,f}^{+} + \mathscr{E}_f E_{t,f}^{+} + \mathscr{O}_f n_t \quad (3) \\ &= \mathscr{O}_f \mathscr{K}_p Z_{t,p}^{-} + \left(\mathscr{U}_f + \mathscr{O}_f \mathscr{N}_{f,p}\right) U_{t,f}^{+} + N_t^{\perp} \end{aligned}$$

where $N_t^{\perp} = \mathscr{E}_f E_{t,f}^{+} + \mathscr{O}_f (n_t - \mathscr{N}_{f,p} U_{t,f}^{+})$, denoting the projection in mean square sense of n_t onto $U_{t,f}^{+}$ by $\mathscr{N}_{f,p} U_{t,f}^{+}$. Here $\mathscr{O}_f = [C', A'C', \cdots, (A^{f-1})'C']'$ denotes the truncated observability matrix, $\mathscr{U}_f = [L_{i-j}]_{i,j=1,\ldots,f}$ the Toeplitz matrix of the impulse responses L_j, where $L_j = 0, j < 0$ is used. $\mathscr{E}_f = [K_{i-j}]_{i,j=1,\ldots,f}, K_j = 0, j < 0$. This equation is a vector equation for $t \in \mathbb{Z}$. Often this equation is written as a matrix equation having the above equation (3) for $t = p+1, p+2, \ldots, T-f$ as its columns. The structure of the matrices containing the data caused the term 'data Hankel matrices'. We will put forward a different view of the equation.
The central equation (3) decomposes the vector $Y_{t,f}^{+}$ into three components: $\mathscr{O}_f \mathscr{K}_p Z_{t,p}^{-}, (\mathscr{U}_f + \mathscr{O}_f \mathscr{N}_{f,p}) U_{t,f}^{+}$ and $N_t^{\perp}$. For $t = p+1, \ldots, T-f$ the vectors $Y_{t,f}^{+}, U_{t,f}^{+}$ and $Z_{t,p}^{-}$ can be built using input/output data $y_t, u_t, t = 1, \ldots, T$. Hence the equation

$$Y_{t,f}^{+} = \beta_z Z_{t,p}^{-} + \beta_u U_{t,f}^{+} + N_t^{\perp}, t = p+1, \ldots, T-f$$

has the following interesting features:

- Under the assumption of open loop operation, the vector $N_t^{\perp}$ is uncorrelated with the remaining terms on the right hand side of the equation. Under the closed loop assumption, $U_{t,f}^{+}$ and $N_t^{\perp}$ are correlated, but $Z_{t,p}^{-}$ and $N_t^{\perp}$ remain uncorrelated.
- The matrix β_z has rank n, the system order.
- $\beta_u = \mathscr{U}_f + \mathscr{O}_f \mathscr{N}_{f,p} \to \mathscr{U}_f$ for $p \to \infty$.
- $N_t^{\perp} \to \mathscr{E}_f E_{t,f}^{+}$ for $p \to \infty$.
- $\mathscr{E}_f E_{t,f}^{+}$ is an MA(f) process.

These observations build the basis for the subspace algorithms.

3. DESCRIPTION OF THE ALGORITHMS

Most subspace algorithms share a common outline. They can be decomposed into three main steps [2]:

(1) Use the central equation to estimate β_z, β_u by regressing $Y_{t,f}^{+}$ onto $Z_{t,p}^{-}$ and $U_{t,f}^{+}$ for the open loop case. In the closed loop case, given an estimate $\hat{\beta}_u$ of $\mathscr{U}_f$, an estimate of β_z is obtained using regression of $Y_{t,f}^{+} - \hat{\beta}_u U_{t,f}^{+}$ onto $Z_{t,p}^{-}$. This leads to estimates $[\hat{\beta}_z, \hat{\beta}_u]$.
(2) The estimate $\hat{\beta}_z$ will typically be of full rank, whereas $\mathscr{O}_f \mathscr{K}_p$ is of rank n. Hence a rank n approximation $\hat{\mathscr{O}}_f \hat{\mathscr{K}}_p$ of $\hat{\beta}_z$ is obtained.
(3) Based on the estimates $\hat{\mathscr{O}}_f, \hat{\mathscr{K}}_p$ and $\hat{\beta}_u$, estimates of the system matrices are obtained.

This outline is shared by most of the commonly used subspace procedures. In particular `MOESP` and `CCA` fit into this framework, whereas `N4SID` uses a slightly different third step while using the same first two steps. Note, that the description was given for the open loop case and the closed loop case, whereas most of the literature only considers the open loop case.
In the following, we will describe the various approaches to the three steps in more detail, where the emphasis will be on a discussion with respect to applicability to real world data sets, numerical aspects and also asymptotic properties, above all consistency issues and asymptotic variance considerations.

3.1 Step 1: Regression

The first step in the procedure is a regression. Least squares regression is maybe the best understood statistical method. Efficient numerical procedures exist. The pitfalls are understood to a large extent. Using recursive regression methods, one immediately obtains recursive subspace methods (cf. e.g. Oku and Kimura, 2002). In particular, the regression faced in subspace methods has

- lagged output variables as regressors
- residuals, which are not white.
- reduced rank coefficient matrices
- for $p = \infty$ the matrix $\mathscr{U}_f$ has a rich structure, i.e. it is block Toeplitz.

[2] This decomposition has been given in (Peternell *et al.*, 1996). A similar view of the algorithms and in particular the use of the regression interpretation is independently given in (Shi, 2002).

- potential problems with illconditioning due to multicollinearity.

The consequences of these facts are discussed next.

3.1.1. Lagged Output Variables Different possible choices with respect to the initial and end conditions are possible. The regression equation was written for $t = p+1,\ldots,T-f$. As for ARX systems, setting the initial and end conditions to be equal to zero, the regressions can be calculated using the estimates of the covariance sequence, since in $Y_{t,f}^{+}, U_{t,f}^{+}$ and $Z_{t,p}^{-}$ lagged versions of the two processes y_t and u_t appear. This is equivalent to extending the regression equation to $t = 1,\ldots,p$ and $t = T-f+1,\ldots,T$, while replacing missing values with zeros. It will be clear from the following, that throughout the algorithms not the observed processes themselves are needed, but the estimates of the first $f+p-1$ covariances are sufficient. A different approach is to discard the time instants, where some variables are not observed and to use the regression equation only for $t = p+1,\ldots,T-f$. Arguments paralleling the autoregressive case could be made keeping in mind, that the obtained estimate is $\hat{\beta}_z$ rather than the system matrices themselves: Setting initial and end conditions to zero definitely leads to serious distortions for f and p relatively large in comparison to T. Additionally a different argument has been put forward in favor of not using covariance estimates: For linear equations it is known, that solving the least squares problem using the QR decomposition is numerically favorable to solving the normal equations. This is the reason for using the QR decomposition in the original versions of `MOESP` and `N4SID`. It is the belief of the author that the contribution of the numerical errors to the total error is minor. Therefore the decision, how to choose the initial and end conditions in the regression should be based on statistical grounds rather than for numerical reasons.
There are cases, where the idea of viewing subspace algorithms as being a nonlinear function of the estimated sample covariances is beneficial. First of all, this view is very convenient for the derivation of asymptotic properties of the estimators obtained from using the subspace approach. Basically this idea underlies all results proving consistency and asymptotic normality (with the exception of the case that there is an integrator present in the data generating process). But secondly, and more important for the practitioner, basing the estimation on estimated covariances brings many convenient features:

- Huge sample sizes can be dealt with: Calculating the sample covariances can be done even for very large data sets in a few seconds. On the contrary, the regression matrix can become huge even for moderate sample sizes: Choosing $f = p = 30$ e.g. for a three dimensional input and three dimensional output observed for 5000 time instants results in $Z_{t,p}^{-}$ being of dimension 180 and with t varying between 31 and 4970 the corresponding regression matrix would be of dimension 180 times 4940 containing approx. 890.000 entries. The QR decomposition of the original `MOESP` algorithm, which calculates in effect the regression, would thus have to be performed on a matrix of size 360 times 4940 having more than 1.6 million entries.
- Missing values: Due to the dynamic structure of the regression single irregularily missing values might reduce the effective sample size substantially. The estimated covariances, however, do not suffer such a loss in accuracy.
- Outliers: Covariance estimators exist, which are robust with respect to outlying data points. These might be easier to apply than robustifications for the regression itself.
- Time varying parameters: Recursive covariance estimators can be used in order to cope with time varying parameters, although this might not be preferable.

3.1.2. Nonwhite residuals Recall that the residuals are equal to $\mathscr{E}_f E_{t,f}^{+} + \mathscr{O}_f(n_t - \mathscr{N}_{f,p} U_{t,f}^{+})$. These are nonwhite, since $E_{t,f}^{+}$ and $E_{t-1,f}^{+}$ have a considerable overlap and because $n_t - \mathscr{N}_{f,p} U_{t,f}^{+}$ is nonwhite. Choosing p large, the second problem can be made negligible, whereas the first problem remains. The usual solution to the problem of correlated errors is to use the GLS estimator rather than the OLS estimator with an estimate of the covariance matrix of the residuals. However, there exist cross restrictions between the regression parameters and the noise covariance matrix, which is a problem for GLS estimation. Furthermore it is noted, that $\hat{\beta}_z$ is only an intermediate estimate, therefore it might not prove essential to use an optimal estimator in this stage.

3.1.3. Reduced rank regression It has been noted, that $\mathscr{O}_f \mathscr{K}_p$ has rank equal to the system order n, whereas the estimate $\hat{\beta}_z$ typically has full rank (for f and p sufficiently large). A natural idea would be to incorporate the reduced rank property already in the regression. Consider a regression problem

$$y_t = \beta_z z_t + \beta_u u_t + n_t$$

where y_t, u_t and z_t are observed for $t = 1,\ldots,T$ and the rank of $\beta_z = \alpha\beta'$ is restricted to be equal to n. In order to simplify notation, let $\langle a_t, b_t\rangle = \sum_{t=1}^{T} a_t b_t'$, where a_t and b_t here stand for any of the processes y_t, z_t or u_t. First consider the criterion function ($\hat{n}_t(\alpha,\beta,\beta_u) = y_t - \alpha\beta' z_t - \beta_u u_t$)

$$L_T(\alpha,\beta,\beta_u) = \mathrm{tr}[W_f \langle \hat{n}_t(\alpha,\beta,\beta_u), \hat{n}_t(\alpha,\beta,\beta_u)\rangle]$$

for some positive definite weighting matrix $W_f = W_f'$. The solution to this problem can be found e.g. in (Reinsel, 1998) and is given using the SVD of [3]

[3] Here the symmetric square root of a matrix is used.

$$W_f^{1/2}\langle y_t^\perp, z_t^\perp\rangle \hat{W_p^-} = \hat{U}\hat{\Sigma}\hat{V}' = \hat{U}_n\hat{\Sigma}_n\hat{V}_n' + \hat{R}_n \quad (4)$$

where $\hat{U} \in \mathbb{R}^{fs\times fs}$ denotes the matrix of left singular vectors and $\hat{U}_n$ is the principal submatrix constituted of the first n columns, $\hat{V}$ and $\hat{V}_n$ are the corresponding quantities corresponding the right singular vectors and $\hat{\Sigma}_n = \mathrm{diag}(\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_n)$, where $\hat{\sigma}_1 \geq \ldots \geq \hat{\sigma}_n > \hat{\sigma}_{n+1} \geq 0$ are the estimated singular values ordered decreasing in size. The residuals from a regression of y_t onto u_t are denoted by $y_t^\perp = y_t - \langle y_t, u_t\rangle\langle u_t, u_t\rangle^{-1}u_t$. Similarily $z_t^\perp$ is defined. Further $\hat{W_p^-} = \langle z_t^\perp, z_t^\perp\rangle^{-1/2}$. This leads to a minimum of $\hat{\beta} = (\hat{W_p^-})^{-1}\hat{V}_n$ and $\hat{\alpha} = (W_f)^{-1/2}\hat{U}_n\hat{\Sigma}_n$. Clearly the minimum is not unique.
Alternatively, pseudo maximum likelihood estimation, i.e. estimation based on the Gaussian likelihood for i.i.d. white noise n_t as the criterion function, can be used. Note, that this criterion function leads to reasonable estimators, even if n_t is not Gaussian white noise. It follows from similar arguments to the ones given above, that the solution in this case is identical to the one given above for $W_f = \langle y_t^\perp, y_t^\perp\rangle^{-1}$.
This procedure applied to the regression used in the subspace approach leads to the SVD

$$W_f^{1/2}\langle Y_{t,f}^{+,\perp}, Z_{t,p}^{-,\perp}\rangle \hat{W_p^-}$$

where $\hat{W_p^-} = \langle Z_{t,p}^{-,\perp}, Z_{t,p}^{-,\perp}\rangle^{-1/2}$ and $Z_{t,f}^{+,\perp}, Z_{t,p}^{-,\perp}$ denote the residuals from regression onto $U_{t,f}^+$. It will be seen below, that these choices are used in some procedures.

3.1.4. Structure in $\mathscr{U}_f$ It has been suggested to use the block Toeplitz structure in $\mathscr{U}_f$ in the regression in order to obtain better estimates of β_z in (Peternell *et al.*, 1996). In that paper $p \to \infty$ has been used as a justification for neglecting $\mathscr{O}_f(n_t - \mathscr{N}_{f,p}U_{t,p}^+)$. In some simulation examples also advantages in the accuracy have been shown. There is no general result backing the intuition of better estimates obtained by using the structure. In the case of white noise inputs, moreover, the restricted regression approach does not lead to more accurate estimates, as is shown in (Bauer, 1998). A disadvantage of the restricted regression method is the significant increase in the computational complexity.

3.1.5. Multicollinearity There are two different kinds of multicollinearity problems, and each has to be dealt with differently. The first kind is the obvious problem of perfectly correlated regressors. This occurs e.g. if certain deterministic terms such as the constant are included as inputs. The solution in this case is simply to omit the corresponding variables and this multicollinearity does not introduce any serious problems. The second problem is concerned with almost perfect collinearities. Again, the viewpoint of a regression analysis is helpful in this respect: Ridge regression techniqes can be used in this case and in fact have been proposed (Shi, 2002; Gustafsson, 1999).

3.2 Step 2: Rank n approximation

In the second step of the subspace algorithms the estimate $\hat{\beta}_z$ is approximated by a rank n matrix. This is usually accomplished using a weighted singular value decomposition: Let $\hat{W_f^+} \in \mathbb{R}^{fs\times fs}$ and $\hat{W_p^-} \in \mathbb{R}^{p(s+m)\times p(s+m)}$ be two symmetric positive definite matrices. Then consider the SVD

$$\hat{W_f^+}\hat{\beta}_z\hat{W_p^-} = \hat{U}\hat{\Sigma}\hat{V}' = \hat{U}_n\hat{\Sigma}_n\hat{V}_n' + \hat{R}_n$$

Note, that this SVD is totally analogous to the SVD in (4), which uses a special choice of $\hat{W_p^-}$. Therefore the reduced rank regression approach leads to the same result as the unrestricted regression approach combined with a weighted rank n approximation. `N4SID` does not have this interpretation, since it uses a different weighting $\hat{W_p^-}$, but both `MOESP` and `CCA` fall into this category.
In this step, (almost) all the user choices to be taken in subspace algorithms are of crucial importance. The integers f and p define the dimensions of the matrix on which the SVD is performed. A lower bound on these integers has to be imposed, in order to make sure, that the essential dynamics of the system can be estimated. The weighting matrices $\hat{W_f^+}$ and $\hat{W_p^-}$ have to be chosen, which influence the approximation quality. And finally the order of the estimated system, n say, has to be prescribed in this step.
Order estimation is a - in my opinion - neglected topic. There have been two different approaches proposed: Estimating the order using criterion minimization and alternatively statistical rank testing (for a discussion see e.g. Camba-Mendez and Kapetanios, 2001). Estimating the order has been based mostly on criterion functions comparing the norm of the neglected part of the SVD, i.e. $\hat{R}_n$, to a penalty function:

$$IC(n) = \|\hat{R}_n\|^2 + \frac{C(T)d(n)}{T}$$

where $d(n) = ns + n(s+m) + sm$ denotes the number of parameters needed to parametrize the state space systems of the form (1). $C(T) > 0, C(T)/T \to 0$ is a term penalizing large models. The choice of $C(T)$ determines the properties of the estimates of the order, obtained as the minimizing argument of the criterion. With respect to the norm, the two norm ($\|\hat{R}_n\|^2 = \hat{\sigma}_{n+1}^2$, `SVC`, Bauer (2001)) and the Frobenius norm ($\|\hat{R}_n\|^2 = \sum_{j=n+1}^M \hat{\sigma}_j^2, M = \min\{fs, p(s+m)\}$, `NIC`, Peternell (1995)) have been proposed. Based on different grounds also $\|\hat{R}_n\|^2 = 1 - \sum_{j=n+1}^M \log(1 - \hat{\sigma}_j^2)$ has been proposed (Camba-Mendez and Kapetanios, 2001).
The statistical testing approach is based on a series of tests on the rank of a matrix according to the ideas of Gragg and Donald (1997): The series is started at the hypothesis of the order of the system being equal to null. If the null hypothesis is rejected, the null hypothesis is adapted, now saying that the order is equal to one. This procedure is continued as long as the null is rejected.

For both methods the asymptotical properties have been derived, mainly proving consistency. Small simulation studies compare the various approaches, but to the best of my knowledge no procedure has been found to be superior. Also the motivation for the estimation methods is relatively weak, basically only hinging on consistency. But of course, many consistent procedures can be defined.

3.3 Step 3: Estimation of system matrices

From the previous steps, the estimates

$$\hat{\mathscr{O}}_f = (\hat{W}_f^+)^{-1/2}\hat{U}_n\hat{\Sigma}_n, \hat{\mathscr{K}}_p = ((\hat{W_p^-})^{-1}\hat{V}_n)'$$

and $\hat{\beta}_u$ have been obtained. Up to now, the discussion did not distinguish between the various different approaches, except for pointing to different choices of weighting matrices, which however only apply for the default algorithms. There is no difficulty in applying, say, `CCA` using the weighting scheme put forward in `MOESP`. The estimation of the system matrices, however, is where the differences in the algorithms show up. Hence this section is divided into two subsections, the first one dealing with the `MOESP` type of methods, whereas the second one deals with state based approaches.

3.3.1. `MOESP` *type of methods* The distinctive feature of this type of algorithms is the usage of the matrix $\hat{\beta}_u$ in the estimation. The estimation hinges on the estimates $\hat{\mathscr{O}}_f$ and $\hat{\beta}_u$ and most algorithms in this class only estimate (A,B,C,D), the subsystem describing the effects of the input on the output. The noise model is included in the estimation, since it is hoped that by doing this the estimation accuracy is increased. It is debatable, whether this is really true. Chiuso and Picci (2002) find examples, where the joint modelling approach leads to worse estimates, than seperately modelling the systems (A,B,C,D) and (A,K,C).
In the first part of step 3, the `MOESP` type of approach uses the structure of the matrix $\mathscr{O}_f$: Define $\overline{\mathscr{O}_f}$ as the submatrix of $\mathscr{O}_f$, which is obtained by omitting the first block row. Then obviously

$$\overline{\mathscr{O}_f} = \mathscr{O}_{f-1}A$$

Letting $\underline{\hat{\mathscr{O}}_f}$ be defined as the first $f-1$ block rows of $\hat{\mathscr{O}}_f$ this equation can be used in order to obtain an estimate $\hat{A}$ of A as the least squares solution to

$$\overline{\hat{\mathscr{O}}_f} = \underline{\hat{\mathscr{O}}_f}A + r$$

The estimate $\hat{C}$ is defined as the first block row of $\hat{\mathscr{O}}_f$. This procedure is usually termed 'shift invariance approach'.
Given these two estimates, a number of different procedures for the estimation of B and D have been proposed and it does not seem to be clear, which approach is to be favored. (Ljung and McKelvey, 1996*a*) propose to use the representation $y_t = \sum_{j=0}^{\infty} L_j u_{t-j} + v_t$ as the basis for the estimation of B and D, as $L_0 = D, L_j = CA^{j-1}B, j > 0$ are linear in B and D and hence given estimates of A and C the estimates of B and D are obtained as

$$(\hat{B},\hat{D}) = \arg\min \sum_{t=1}^{T} \|y_t - L(u_t,B,D)\|^2$$

where $L(u_t,B,D) = \sum_{j=0}^{t-1} L_j u_{t-j}$ is linear in B and D. Closed form expressions for the solution exist.
Alternatively the structure of $\beta_u = \mathscr{U}_f + \mathscr{O}_f\mathscr{N}_{f,p}$ can be used to construct estimates of B and D. This is in fact done in the original `MOESP` procedure. Note, that $\mathscr{U}_f$, being a matrix whose entries are L_j, is linear in B and D. Let $\mathscr{O}_f^{\perp} \in \mathbb{R}^{fs\times(fs-n)}$ denote a matrix, such that $\mathscr{O}_f'\mathscr{O}_f^{\perp} = 0$, while $\mathscr{O}_f^{\perp}$ is of full column rank. Further let $\mathscr{U}_f = L(A,B,C,D)$. Then

$$(\mathscr{O}_f^{\perp})'\beta_u = (\mathscr{O}_f^{\perp})'\mathscr{U}_f = (\mathscr{O}_f^{\perp})'L(A,B,C,D)$$

If estimated quantitites replace true quantities, the equation only holds approximately and B and D can be determined using least squares fitting on the vectorized equations:

$$\mathrm{vec}(\widehat{(\mathscr{O}_f^{\perp})}'\hat{\beta}_u) = \mathrm{vec}(\widehat{(\mathscr{O}_f^{\perp})}'L(\hat{A},B,\hat{C},D)) + r$$

There are two issues related to this equation: The first issue is the choice of the estimate $\widehat{(\mathscr{O}_f^{\perp})}$, which is based on an estimate of $\mathscr{O}_f$. Two possible choices are $\hat{\mathscr{O}}_f$ and $[\hat{C}',\hat{A}'\hat{C}',\ldots,(\hat{A}^{f-1})'\hat{C}]'$. The second issue is the distribution of r. Given (A,C), Chiuso and Picci (2002) give the variance of r and find an estimate of the variance in order to obtain GLS estimates.
For all these procedures it is unclear, which is the preferable one. Except for a few simulation examples, no evidence exists. Moreover, the estimation of the system matrices corresponding to the noise characteristics usually is neglected. There exist procedures to estimate the matrix K, however, we will not present them. For the case of no observed inputs present only realization methods can be seen to fall into the `MOESP` type of algorithms.

3.3.2. The state approach Contrary to the `MOESP` type of approach, the state approach uses the estimate $\hat{\mathscr{K}}_p$ and neglects $\hat{\mathscr{O}}_f$ and $\hat{\beta}_u$. Recalling that the state is equal to

$$x_t = \mathscr{K}_p Z_{t,p}^- + n_t$$

an estimate of the state can be given as $\hat{x}_t = \hat{\mathscr{K}}_p Z_{t,p}^-, t = p+1,\ldots,T$. This estimate can be used in the observation equation in place of the true state x_t in order to obtain an estimate of C and D from

$$y_t = \hat{C}\hat{x}_t + \hat{D}u_t + \hat{\varepsilon}_t$$

This also defines an estimate $\hat{\varepsilon}_t$ of the innovations. Secondly, if an estimate $\tilde{x}_{t+1}, t = p+1,\ldots,T$ is available, the state equation could be used in order to obtain estimates of A,B and K using the regression equation

$$\tilde{x}_{t+1} = \hat{A}\hat{x}_t + \hat{B}u_t + \hat{K}\hat{\varepsilon}_t + r_t$$

One obvious estimate is $\tilde{x}_{t+1} = \hat{x}_{t+1}, \tilde{x}_{T+1} = 0$ using the shifted estimated state sequence. Alternative estimates for $\tilde{x}_{t+1}$ have been proposed in the original N4ISD procedure and recently by Chiuso and Picci (2002). In the case of no observed inputs, the formulae are valid without a change, setting $m = 0$.

4. SOME ASYMPTOTIC RESULTS

After having described the algorithm in detail, in this section the main theoretical results are cited, which are important in order to obtain an understanding of the possible applications of the algorithms. We will not state the results in full technical detail, but rather refer to the original sources for the interested reader.

4.1 MOESP *type of procedures*

Corresponding to this class of procedures, there only exists a limited set of results on the asymptotic properties. The effects of the user choices (f, p, the weighting matrices) on the asymptotic properties are not well understood. For all procedures, consistency has been fairly well investigated and cases, where the algorithms are not consistent, have been singled out (Jansson and Wahlberg, 1998). Also asymptotic normality has been proved (Bauer and Jansson, 2000) and the calculations of the asymptotic variance described in detail (Jansson, 2000). It is known, that the choice of $\hat{W_f^+}$ does not influence the accuracy of the estimated poles of the system, i.e. the eigenvalues of A (Jansson, 1997).
Corresponding to the effects of the weighting matrices on the asymptotic bias in the case of underspecification of the order, there exist a number of examples, which show that in some cases, the bias can be affected to the favor of the modeller. However, there do not exist any results making this observation more concrete than rules of thumb based on a couple of observations. Also the effects of the choices of the weightings on the asymptotic variance are not well understood. The expressions derived so far do not provide good insights. With respect to the effects of the choice of f and p almost no advice exists to the best of the knowledge of the author. An invited session at this conferences is dedicated to these topics, which are an area of ongoing research, which also needs impact from applications.

4.2 State approach

For the state approach, there is much more knowledge present on the effects of the various user choices. Two different cases are distinguished: For the case of no observed inputs or white noise inputs, the asymptotic properties of subspace algorithms are well understood. For the case of coloured input, the situation ressembles much the situation for the MOESP type of procedures.

4.2.1. No observed inputs or white noise inputs

In this case, the procedure based on $\tilde{x}_{t+1} = \hat{x}_{t+1}$ is understood quite completely with respect to the asymptotic properties. A necessary condition in order to achieve consistency in this setting is to let p tend to infinity as a function of the sample size (cf. Deistler *et al.*, 1995). If additionally asymptotic normality of the estimators is to be ensured, then $p \geq -d \log T / (2 \log |\lambda_{max}(A - KC)|)$ for some arbitrary $d > 1$ is assumed in the proofs (cf. Bauer *et al.*, 1999). This bound depends on unknown system quantities. However, it can be shown (cf. e.g. Hannan and Deistler, 1988, Theorem 6.6.3) that $2\hat{p}_{AIC}$ fulfills this bound almost surely, where $\hat{p}_{AIC}$ is the order estimated using AIC in an autoregressive approximation of the output process y_t.
In the case of no observed inputs there exist expressions for the asymptotic bias term for underspecification of the order, which also include some results on the dependence of the bias distribution over frequency on the choice of the weighting matrix (cf. Bauer, 1998, Chapter 2). However, these results are not very sharp and particularly not of much use in practice. For the case of correctly specified order, (Bauer and Ljung, 2002) provide very transparent expressions for the asymptotic variance, which reveal the influence of the weighting $\hat{W_f^+}$ and the choice of f on the asymptotic accuracy. $\hat{W_p^-}$ has been shown to be of no concern in this case in (Bauer *et al.*, 2000). The bottom line of these results is that the CCA choice of the weighting matrices is optimal for each fixed f. Furthermore, the asymptotic accuracy for the CCA estimates increases monotonically with f. This implies, that an optimal procedure has to use $f \to \infty$. (Bauer, 2000) finally shows, that in the case of no observed inputs, CCA together with $f = p = 2\hat{p}_{AIC}$ leads to a procedure, which asymptotically is equivalent to prediction error methods and hence in the case of Gaussian innovations achieves the optimal accuracy given by the Cramer Rao lower bound.
(Dahlen and Scherrer, 2001) show, that CCA is asymptotically equivalent to a procedure, which performs balanced model reduction on a preliminary AR estimate in the sense, that the difference of the obtained estimates tends to zero faster than $1/\sqrt{T}$. This provides an alternative interpretation of the procedure, giving also some motivation to the choice of p as suggested above.
In the case of white noise observed inputs, the variance expressions given in (Bauer and Ljung, 2002) are still valid. Asymptotic equivalence to prediction error methods also has been shown. Therefore in these cases, CCA can be seen as an equivalent of (pseudo) maximum likelihood methods *under the assumption of correctly specified order*.

4.2.2. Coloured inputs In the case of coloured observed inputs some examples have been given, which show, that in this case CCA does not achieve optimal accuracy. The knowledge about the asymptotic properties is limited to the basic results of consistency (Peternell *et al.*, 1996) and asymptotic normality (Bauer, 1998). Expressions for the asymptotic variance exist, but are computationally demanding.

5. APPLICATION TO STOCK RETURN DATA

In order to illustrate the advantages of subspace based state space modeling we use a data set of so called 'high frequency' stock returns, provided by Tim Bollerslev. The data set is further described in Bollerslev and Zhang (2003). The data set consists of five-minute returns on a value weighted market portfolio consisting of more than 6000 of the largest issues traded on the NYSE, NASDAQ and AMEX stock exchanges. The sample consists of 1761 trading days from January 2, 1993 through December 1, 1999. For each trading day, 79 five minute returns are given, resulting in a total of 139119 observations. The data set hence covers an extended period in time and is also quite demanding with respect to its size.
One commonly used hypothesis in financial econometrics is the so called 'efficient market hypothesis', which basically means that the knowledge of past return data cannot be used to obtain a forecast of future returns, which beats the no change prediction systematically. For daily return data this hypothesis seems to be rather accurate and modelling the mean return is not a rewarding task. For five minute returns by contrast, even a simple AR(1) model already beats the no change prediction on average. Hence, fitting an ARMA model to the five minute return series seems to make sense.
Financial data sets share a number of commonly found characteristics:

- Heteroskedasticity: Conditional on the past, the variance of the innovations varies.
- 'Fat tails': The amount of 'large' innovations is higher than would be expected from a normal density, i.e. the distribution of the innovations is leptokurtic.

Subspace algorithms are known to be robust with respect to certain forms of heteroskedasticity including the commonly used GARCH models (Bauer, 2002). The leptokurtosis can be dealt with using outlier robust covariance estimators. Usually the analysis of financial data is done in two steps: First a model for the conditional mean is derived in order to obtain an estimate of the innovations. Secondly a model for the conditional variance (or respectively a model for the squared innovations) is derived based on the estimated innovations. For efficiency of estimation the final model is estimated jointly.
Therefore, consider a state space model estimated for

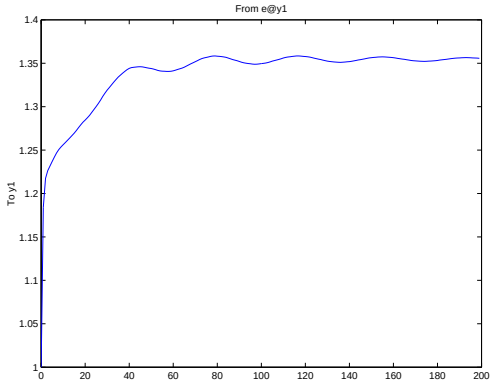

Fig. 1. Step function for the estimated 11-th order transfer function.

the whole data set: As the outlier robust estimator of the covariance sequence we used the trimmed mean, where we neglect 5% percent of the data. These covariance estimates are then inserted into the subspace algorithm and a model of order 11 (estimated from the data) is estimated for the full data set. In this step the integers f and p have been specified externally. A plot of the estimated step function can be seen in picture 1. The plot shows that the immediate effects dominate, while there is some influence on the first halfday. The step response levels out at a value of approximately 1.35 after half a day, although there is some fluctuation at the period length one half day.
Due to the extended time span (7 years), time constancy of the system is highly questionable. As an alternative to the constant parameter model, the data set has been partitioned into blocks of 10 consecutive trading days, resulting in 176 data sets of 790 data points each. For each data block, a seperate model has been specified and estimated. A number of different techniques for the specification have been tried out, including the estimation of robust estimators for the covariances, fully automated model selection procedures, and robust estimation for fixed model structure (eleventh order model). As a criterion to compare the time varying models to the constant parameter model, the one step ahead prediction error on the following ten trading days for the time varying models is compared to the prediction error on the same data set for the constant parameter model. This comparison is friendly to the constant model, since there all the data was used for estimation. Nevertheless, the time varying models perform better, as can be seen in figure 2, showing the quotient of the standard deviation of the one step ahead prediction error on the 10 consecutive trading days to the standard deviation of the one step ahead prediction using the constant parameter model. The plot shows, that the time varying models outperform the constant parameter model in the first half of the observation period, while it is somewhat worse on the second half. Overall the difference is negligible. Thus on these grounds the constant parameter hypothesis is rejected and the model for the mean is given by the time varying parameter models. These models

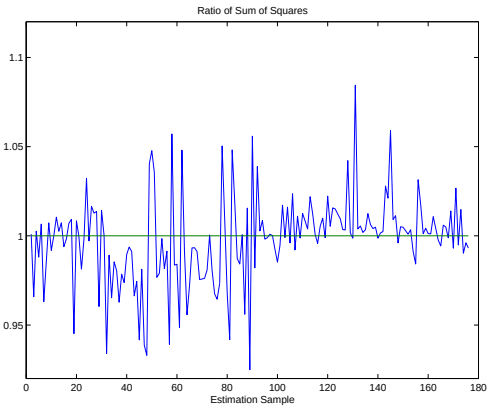

Fig. 2. Ratio of standard deviation of prediction error corresponding to the time varying parameter models against the standard deviation of the prediction error corresponding to the constant parameter model.

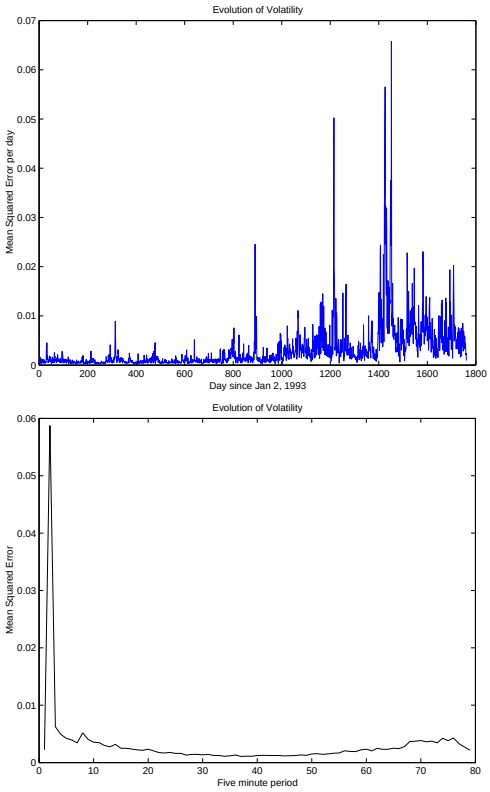

Fig. 3. Mean of squared residuals for the various days (upper plot) and for different five minute intervals (lower plot).

result in an overall R^2 of 0.11.

Figure 3 shows, that the variability as measured by the daily means of the squared errors is not constant over the various days. Also the variability is not constant for the various five minute intervals, as documented by the lower plot. It can be clearly seen, that the most variability in the stock returns occurs five minutes after the opening of the market. One commonly used model for this sort of data is the so called GARCH model (Bollerslev, 1986). Squared GARCH

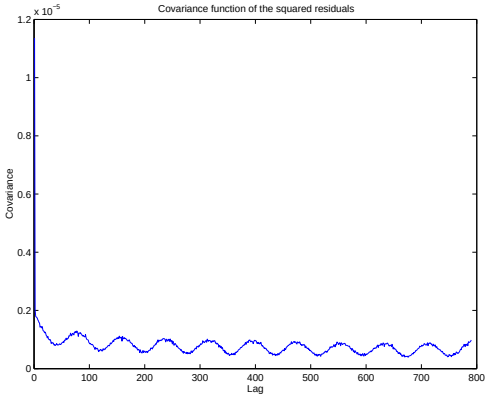

Fig. 4. Estimated covariance sequence of the squared innovations adjusted for intra day means.

processes have been interpreted as a linearily filtered heteroskedastic white noise including a nonzero intercept. Therefore the usual modeling techniques can be used. Clearly in our case, the deterministic term depends on the particular five minute interval and cannot be chosen as constant. Two alternatives are the introduction of one dummy for each interval or a parametric model for the distribution over time. We will here only deal with the dummy approach, implemented by subtracting the trimmed mean (10%) of the squared innovations for each five minute time period. The (trimmed, 10 %) estimated covariance sequence can be seen in figure 4. The plot shows some of the features, that are often found in these data sets: The covariance at lag one is already rather small compared to the covariance at lag 0 (correlation of about 0.18). The remaining covariances show a slow decay and a cyclical behaviour at the daily frequency.

In principle there are a number of different strategies to model the conditional variability: These include one model for all instances, a different model for each five minute interval or a multivariate model for the vector of all 79 five minute returns jointly. Each of these has its drawbacks: Building one univariate model is the most restrictive model. On the other hand, one model for each five minute interval leads to a large amount of work due to the necessity to specify 79 models. The multivariate model also has a number of drawbacks: First of all, the information set is different: Whereas in the univariate models, the prediction is performed on the basis of all returns up to time t, the multivariate model predicts on the basis of all data up to the last day and hence does not take into account the returns of the current day. Secondly, the properties of multivariate GARCH models are largely unknown. Especially the positivity constraint is problematic. Hence logarithms are taken, which also leviates the problems due to the leptokurtosis.

Using subspace methods, multivariate models can be estimated and specified for output dimension 79 without big problems. The specification step is numerically feasible, as the main computational load in this case lies in the calculation of the covariance sequence, which can be done prior to model estimation. The

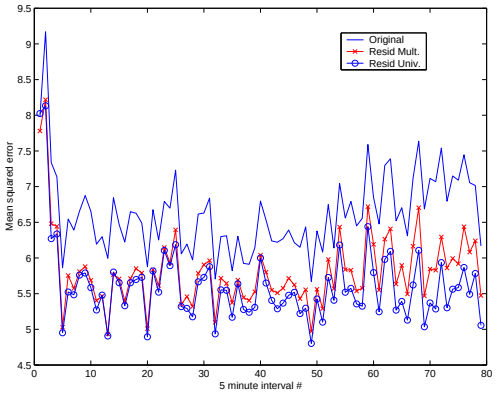

Fig. 5. Mean squared errors for the squared adjusted residuals, the residuals from the multivariate model and the residuals from the univariate model.

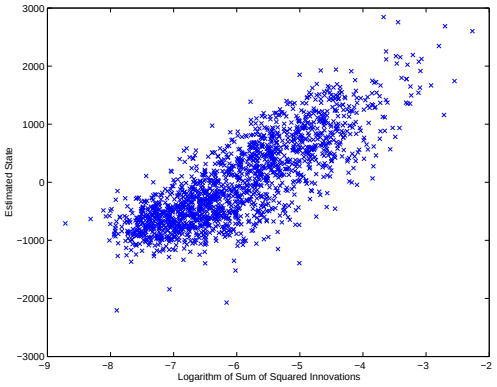

Fig. 6. Logarithm of the sum of the squared innovations plotted against the estimates of the state in the multivariate model.

automatic procedure estimates the order to equal 1. Plot 5 shows the mean squared errors of the logarithm of the squared residuals, which have been corrected for the mean for given five minute interval, the mean squared errors for the multivariate model and the mean squared error for the univariate model. It is observed, that both models substantially reduce the mean squared error over the naive model of constant volatility. The univariate model performs a bit worse for the early returns, but better for the late returns. The explanation for this lies in the different amount of information on which the prediction is based. Nevertheless, if one considers the estimated logarithm of the determinant of the residual variance, the multivariate model results in 133.01, whereas the univariate model only achieves 133.48. The number of parameters is $2 * 79 = 158$ for the multivariate model and 166 for the univariate model. The one dimensional state can be interpreted as an estimate of the logarithm of the mean squared errors of the various days, as can also be seen in a scatter plot (cf. Figure 6).

6. COMPARISON TO PREDICTION ERROR METHODS

Subspace methods are an alternative to prediction error methods in the sense, that they can be used to fit linear dynamical models to input/output data. It has been cited, that for the case of no observed inputs one particular method, namely `CCA` achieves the same asymptotic accuracy as the prediction error method, while being much more computationally efficient (especially for large sample sizes). For the case of coloured observed inputs, no subspace method has been proven up to now to provide asymptotically efficient estimates. All the results given above only correspond to asymptotic reasoning. The question remains, what place subspace methods should take in the toolbox of the model builder? My personal views, which are definitely not shared by all people in the community, are the following: At the very least, subspace methods for properly chosen user parameters lead to good initial estimates, which can then be used in gradient based optimization methods. This view is implementated in the fully automatic `pem` procedure in the system identification toolbox of `MATLAB`. Based on theoretical arguments, however, for the case of no observed inputs, there is no reason for rejecting the estimates obtained using `CCA` in favor of estimates obtained from numerical optimization of criterion functions.
Additionally there are a number of situations, where subspace methods can be a useful alternative:

- Model specification: Subspace algorithms provide an additional possibility for estimation of the order.
- Systems with moderate output dimension. For output dimension say up to $s = 5$ prediction error methods are probably still computationally feasible, while subspace algorithms provide a quick second look at the data.
- Automatic modelling: Providing rules of thumb for the choice of f and p and the weighting matrices, combined with order estimation procedures, immediately renders the subspace methods into an automatic modelling method: data in, estimated system out (cf. Bauer and de Waele, 2003).

In a number of cases, subspace algorithms seem to be the only choice, since prediction error methods for state space models are not feasible:

- Very large data sets. In the example the number of sampling instants was equal to 139119. This data set can still be dealt with using standard prediction error methods. Order estimation on this data set using prediction error methods nevertheless becomes infeasible. Using the subspace approach, much larger data sets can be dealt with.
- Many outputs. In the application example a system for a 79 dimensional output has been es-

timated. Using prediction error methods this would be numerically infeasible. Even autoregressive modelling would not be feasible, since for an AR(1) system, $79^2 = 6241$ parameters would need to be estimated. Hence, reduced rank regression in an autoregressive setting represents the only alternative in this case.

- Many models at a time. Again, the example considered used a model for two consecutive weeks, resulting in a total of 176 models to be estimated. Using prediction error methods, the only choice due to time restrictions would be to use the same model structure for all models. Subspace methods on the contrary allow for specification of each model at a time.
- Input selection. In econometrics (but also in other disciplines) it is common, that there exists only a set of regressors, which are seen as influental, but not necessary each in fact is. Hence the first step usually is input selection, i.e. the specification, which of the potential inputs contributes to the output. Typically in this step a huge number of models has to be estimated and hence only a computationally feasible method is of use.

It should be stressed again, that this list is only my personal belief. It definitely is not complete and some of the points are debatable.

Finally I would like to correct a possible misunderstanding: I do not argue, that subspace methods are superior to prediction error methods. In many cases, prediction error methods are still the better tool. In particular simulations indicate, that the small sample properties of subspace methods are quite poor for very small samples, such as the ones typically found in macroeconometrical applications. In this case, the estimates are of use only as initial estimates. Additionally, it is not possible to include prior information into algorithms easily. Therefore, they are only of interest for black box modelling. And last, but not least, there are still many topics, which are not completely clarified: (Bauer and de Waele, 2003) show, that although nice on a theoretical ground, the automatic modelling approach based on the recommendations of this paper works surprisingly poor in certain test cases. This is an indication, that especially the choice of the user supplied quantities $f, p, \hat{W}_f^+, \hat{W}_p^-$ and the order of the system has to be analyzed in more depth, also from the perspective of finite sample properties, which are relevant for applications.

ACKNOWLEDGEMENT

Even though my name is on the front page of this article, it really is the work of the group in Vienna, that is the root of this contribution. Almost all of my work in the area has been done in cooperation with and under the guidance of Manfred Deistler and Wolfgang Scherrer from the TU Wien. I am very grateful for their support. I am also indebted to Thomas Ribarits for careful proof reading. The data set has been provided by Tim Bollerslev, which is gratefully acknowledged.

REFERENCES

Andersson, S. (2002). Hidden Markov Models - Traffic Modeling and Subspace Methods. PhD thesis. Lund University, Sweden.

Babuska, R., J. Keizer and M. Verhaegen (1997). Identification of nonlinear dynamic systems as a composition of local linear parametric or state space models. In: *Proceedings of the SYSID'97 Conference, Fukuoka, Japan*. pp. 703–708.

Bauer, D. (1998). Some Asymptotic Theory for the Estimation of Linear Systems Using Maximum Likelihood Methods or Subspace Algorithms. PhD thesis. TU Wien, Austria.

Bauer, D. (2000). Asymptotic efficiency of the CCA subspace method in the case of no exogenous inputs. Technical report. Department of Automatic Control, Linköping Universitetet.

Bauer, D. (2001). Order estimation for subspace methods. *Automatica* **37**, 1561–1573.

Bauer, D. (2002). Identification of state space systems with conditionally heteroskedastic innovations. In: *Proceedings of the 15th IFAC World Congress*. pp. T–Mo–M02.

Bauer, D. and L. Ljung (2002). Some facts about the choice of the weighting matrices in Larimore type of subspace algorithms. *Automatica* **38**, 763–773.

Bauer, D. and M . Wagner (2002). Estimating cointegrated systems using subspace algorithms. *Journal of Econometrics* **111**, 47–84.

Bauer, D. and M. Jansson (2000). Analysis of the asymptotic properties of the MOESP type of subspace algorithms. *Automatica* **36**(4), 497–509.

Bauer, D. and St. de Waele (2003). A finite sample comparison of automatic model selection methods. In: *Proceedings of the SYSID'03 conference, August 2003, Rotterdam, The Netherlands*.

Bauer, D., M. Deistler and W. Scherrer (1999). Consistency and asymptotic normality of some subspace algorithms for systems without observed inputs. *Automatica* **35**, 1243–1254.

Bauer, D., M. Deistler and W. Scherrer (2000). On the impact of weighting matrices in subspace algorithms. In: *Proceedings of the SYSID'2000 conference*, Santa Barbara, California.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**, 307–327.

Bollerslev, T. and B. Zhang (2003). Measuring and Modeling Systematic Risk in Factor Pricing Models using High-Frequency Data. *Journal of Empirical Finance*, No. 10, forthcoming.

Camba-Mendez, G. and G. Kapetanios (2001). Testing the rank of the hankel covariance matrix: A statistical approach. *IEEE Transactions on Automatic Control* **46**, 331–336.

Chen, H. and J. Maciejowksi (2000). Subspace identification methods for combined deterministic-stochastic bilinear systems. In: *Proceedings of the SYSID'2000 conference*, Stanta Narbara California.

Chiuso, A. and G. Picci (2002). Asymptotic variances of subspace identification by data orthogonalization and model decoupling. Technical report. University of Padua, Italy.

Chou, C. and M. Verhaegen (1997). Subspace algorithms for the identification of multivariable dynamic errors-in-variables models. *Automatica* **33**, 1857–1869.

Chou, C. T. (1994). Geometry of Linear Systems and Identification. PhD thesis. University of Cambridge.

Chou, C. T. and M. Verhaegen (1999). Closed-loop identification using canonical correlation analysis. In: *Proceedings of the ECC'99 Conference, Karlsruhe, Germany*.

Dahlen, A. and W. Scherrer (2001). The relation of the CCA subspace method to a balanced reduction of an autoregressive model. Submitted to Journal of Econometrics.

Deistler, M., K. Peternell and W. Scherrer (1995). Consistency and Relative Efficiency of Subspace Methods. *Automatica* **31**(12), 1865–1875.

Favoreel, W. (1999). Subspace Methods for Identification and Control of Linear and Bilinear Systems. PhD thesis. Katholieke Universiteit Leuven.

Gomez, J. and E. Baeyens (2002). Subspace identification of multivariable hammerstein and wiener models. In: *Proceedings of the 15th IFAC World Congress*. pp. T–Th–M01.

Gragg, J. and S. Donald (1997). Inferring the rank of a matrix. *Journal of Econometrics* **76**, 223–250.

Gustafsson, T. (1999). Subspace Methods for System Identification and Signal Processing. PhD thesis. Chalmers University, Gothenburg, Sweden.

Hannan, E. J. and M. Deistler (1988). *The Statistical Theory of Linear Systems*. John Wiley. New York.

Haverkamp, B., M. Verhaegen, C. Chou and R. Johansson (1997). Continuous-time subspace model identification method using laguerre filtering. In: *Proceedings of the SYSID'97 Conference, Fukuoka, Japan*. pp. 1143–1148.

Ho, B. and R. E. Kalman (1966). Efficient construction of linear state variable models from input/output functions. *Regelungstechnik* **14**, 545–548.

Jansson, M. (1997). On Subspace Methods in System Identification and Sensor Array Signal Processing. PhD thesis. KTH, Stockholm.

Jansson, M. (2000). Asymptotic variance analysis of subspace identification methods. In: *Proceedings of the SYSID'2000 Conference*. Santa Barbara, California.

Jansson, M. and B. Wahlberg (1998). On consistency of subspace methods for system identification. *Automatica* **34**(12), 1507–1519.

Larimore, W. E. (1983). System identification, reduced order filters and modeling via canonical variate analysis. In: *Proc. 1983 Amer. Control Conference 2*. (H. S. Rao and P. Dorato, Eds.). Piscataway, NJ. pp. 445–451.

Ljung, L. and T. McKelvey (1996*a*). A least squares interpretation of sub-space methods for system identification. In: *Proccedings of the CDC96 Conference*. Kobe, Japan.

Ljung, L. and T. McKelvey (1996*b*). Subspace identification from closed loop data. *Signal Processing, Special Issue on Subspace Methods, Part II: System Identification* **52**(2), 209–216.

McKelvey, T. (1995). Identification of State-Space Models from Time and Frequency Data. PhD thesis. Dept. of Electr. Eng., Linköping.

Ohsumi, A. and T. Kawano (2002). Subspace identification for a class of time-varying continuous-time stochastic systems via distribution-based approach. In: *Proceedings of the 15th IFAC World Congress*. pp. T–Mo–M02.

Oku, H. and H. Kimura (2002). Recursive 4SID algorithms using gradient type subspace tracking. *Automatica* **38**, 1035–1043.

Peternell, K. (1995). Identification of Linear Dynamic Systems by Subspace and Realization-Based Algorithms. PhD thesis. TU Wien.

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

Reinsel, G. (1998). *Multivariate Reduced-Rank Regression*. Springer, New York.

Shi, R. (2002). Subspace Identification Methods for Process Dynamic Modeling. PhD thesis. McMaster University, Canada.

Van Overschee, P. and B. DeMoor (1994). N4sid: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica* **30**, 75–93.

Verdult, V. and M. Verhaegen (2002). Subspace identification of multivariable linear parameter-varying systems. *Automatica* **38**, 805–814.

Verhaegen, M. (1993). Application of a subspace model identification technique to identify lti systems operating in closed loop. *Automatica* **29**(4), 1027–1040.

Verhaegen, M. (1994). Identification of the deterministic part of mimo state space models given in innovations form from input-output data. *Automatica* **30**(1), 61–74.

Zeiger, H. P. and A. J. McEwen (1974). Approximate linear realizations of given dimension via Ho's algorithm. *IEEE Transaction on Automatic Control* **19**, 153.