Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

A Subspace Approach to Reduced Rank Time-series Models.

Victor Solo

Abstract— **While reduced rank time-series models go back over 30 years, there is a renewed interest because of the now commonplace occurrence of high dimensional time-series. Here, for the first time, we characterize the two basic reduced rank vector time-series models in state space terms in a surprisingly simple way. This allows us to extend these models from vector AR to vector ARMA and we develop two new associated subspace fitting algorithms.**

I. Introduction

With the advent of large time series data sets in areas such as system identification, econometrics, neuroscience, there is emerging interest in modeling high dimensional time series e.g. [1],[2],[3]. And reduced rank models are thus of import [4].

Methods in multivariate reduced rank regression (RRR) go back to [5] (see [4]) and are closely related to canonical correlation analysis [4]. The development of reduced rank time series (RRTS) methods starts with [6], [7] and has gained a big impetus from econometric work on cointegration [8].

The monograph [4] summarizes the important methods in RRR and RRTS. It covers model formulation, model fitting and asymptotics. Some of the work on RRR has been rediscovered in the signal processing literature but there have also been new algorithmic developments e.g. [9]. However no connexion is made in [4] between RRTS and state space methods and indeed there seems to have been no work in this direction.

However connexions have been made between RRR and subspace methods e.g. [10],[11]. These authors and others show how subspace fitting methods can be viewed as RRR with correlated noise and use this representation to formulate various types of results. In these works it is a certain Hankel matrix that is of reduced rank. This connexion is important and fruitful but is totally different from what we do here. Namely connect RRTS to SS where as will be seen, other matrices will be of reduced rank.

In this work we provide a state space formulation of RRTS for the first time. This allows an extension of RRTS models from vector AR to vector ARMA and leads to new subspace fitting algorithms .

In section 2 we review RRTS models. In section 3 we develop the surprisingly simple state space characterizations of the two main RRTS models. In section 4 we develop new subspace algorithms to fit these models. Conclusions are in section 5.

V Solo is with School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, AUSTRALIA. v.solo@unsw.edu.au

A. Notation and Acronyms

DIM is dynamic index model; **MFD** is matrix fraction description; **MIL** is matrix inversion lemma; **RRR** is reduced rank regression; **RRTS** is reduced rank time series; **RH** is right hand; **LH** is left hand; **RHS** is right hand side; **SS** is state space; **SVD** is singular value decomposition; **VAR** is vector autoregression; **VARMA** is vector autoregressive moving average; **WN** is white noise.

II. Review of Reduced Rank Time Series Models

To date the development of RRTS models has focussed on VARs. Consider a d-dimensional vector time series y_t described by a VAR

$$y_t = \Sigma_1^p H_u y_{t-u} + \epsilon_t, t = 1, \cdots, \bar{T}$$

where $\{\epsilon_t\}$ is a zero mean WN sequence of variance Σ.

If d is large e.g. d=50, then even with order $p = 1$ we have of order $1\frac{1}{2}d^2 + \frac{1}{2}d$ parameters ~ 3700 parameters. The McMillan degree may be much less than $dp = 50$ and this would provide some dimension reduction; but more would be useful. And this is what RRTS can potentially provide.

The two most significant RR-VAR models are

(i) RH model or WN model

$$H_u = FG_u^T, 1 \leq u \leq p$$

(ii) LH model or DIM model

$$H_u = F_u G^T, 1 \leq p$$

where F, G each have rank $k < d$. These representations are not unique since if W is any $k \times k$ matrix of full rank we can e.g. replace F by FW^T and G_u by $G_u W^{-1}$. Any estimation method resolves this by using some kind of normalization. We now discuss each of these models in turn.

WN-model.
This model has the great attraction that model fitting is very simple; it just involves forming some cross variance matrices followed by a SVD [4][section 5.2]. In previous work [12] we have pointed out a hitherto unremarked disadvantage of the WN-model. If $F_\perp$ is the $d \times (d-k)$ matrix of rank $d-k$ orthogonal to F, then $F_\perp^T y_t$ is a WN (this explains the name we have given it). In the context of stationary processes this does not seem likely to occur in practice.

However if we allow random walk components in y_t then this property becomes significant since it leads to the cointegration property of econometrics (i.e. y_t is not stationary but there are linear combinations that are). For this reason, although we do not discuss cointegration, we continue to deal with this model.

978-1-4244-3124-3/08/$25.00 ©2008 IEEE

DIM-model.

This model is much more interesting, partly because it requires nonlinear fitting. If we introduce the reduced dimension process $y_{G,t} = G^T y_t$ then the whole time series is generated from the index time series $y_{G,t}$.

The seminal work on the DIM is due to [13] who develops model fitting and asymptotics. The model was rediscovered by [9] who gave a new model fitting algorithm . Also as discussed in [12] the model posesses a significant feature not previously noted. Namely it induces a natural casual structure; this is discussed further below.

III. State Space View of Reduced Rank Time Series Models

In this section we begin by posing VARMA extensions of the WN and DIM models and then develop the state space versions which admit very simple characterizations.

The rank reduction condition will now be $k < min(d, n)$ where n is the minimal state space dimension.

A. WN Model

Formally the model is a natural VARMA extension of the VAR model,

$$y_t = FD^{-1}(z^{-1})N(z^{-1})y_t + \epsilon_t$$

where, $F_{d\times k}$ is of rank $k < min(d, n), n$ =McMillan degree.;

$D_{k\times k}(z^{-1}), N_{k\times d}(z^{-1})$ are left coprime matrix polynomials; The MFD $D^{-1}(z^{-1})N(z^{-1})$ is required to be strictly proper, which ensures the lead term on the RHS has a time lag of 1; $D(z^{-1})$ is stable i.e. if b is the maximum degree of any polynomial in $D(z^{-1})$ then $|z^b D(z^{-1})|$ has all roots inside the unit circle. The stability ensures the filtering process can be well defined. We say formally because the process y_t could have unbounded variance e.g. the random walk $y_t = y_{t-1} + \epsilon_t$ fits the definition. So the filtering on the RHS has to be initiated at a finite time.

Of course the easiest way to do this is to specify the model in SS terms. And since MFD/SS connexions are well understood there seems little point in labouring through this standard material. Hence we just redefine the WN model in SS form from the start.

WN-SS Model

We say y_t is generated by a WN-SS model with parameters (A_*, B, C, F) if firstly the parameters satisfy the following conditions:

W_*: A_* is a stability matrix.

W_1: $F_{d\times k}$ is of rank $k < min(d, n)$.

W_2: $A_{*,n\times n}, C_{k\times n}$ is observable; $A_*, B_{n\times d}$ is controllable. Of course this means A_*, B, C is minimal.

Secondly we are given a white noise sequence $\{\epsilon_t\}$ i.e. the sequence is uncorrelated, has zero mean and covariance Σ. And are also given an initial condition ξ_0 for a $n-$dimensional state ξ_t. Then we generate (y_t, ξ_{t+1}) recursively as follows,

$$\begin{aligned} y_t &= FC\xi_t + \epsilon_t, t = 0, 1, \cdots \\ \xi_{t+1} &= A_*\xi_t + By_t, t = 0, 1, \cdots \end{aligned}$$

Formally we then have

$$y_t = FC(zI - A_*)^{-1}By_t + \epsilon_t$$

Let us note here that the model has the WN property; i.e. if $F_\perp$ is a $d \times d - k$ matrix of rank $d - k$ which is orthogonal to F then $F_\perp^T y_t = F_\perp^T \epsilon_t$ is a WN.

To establish our first result we introduce an innovations SS model with reduced rank observation (RRO).

RRO Model

We say y_t is generated by an RRO model with parameters (A, B, C, F) if firstly the parameters satisfy the conditions: W_0, W_1, W_3,

W_0: $A_* = A - BFC$ is a stability matrix.

W_3: (A, C) is observable; (A, B) is controllable .

Note that W_0 is consistent with W_3.

And secondly we are given a WN sequence ϵ_t of covariance Σ and an initial condition ξ_0 of a $n-$ dimensional state ξ_t. Then (y_t, ξ_t) are generated recursively by the innovations state space model

$$\begin{aligned} y_t &= FC\xi_t + \epsilon_t, t = 0, 1, \cdots \\ \xi_{t+1} &= A\xi_t + B\epsilon_t, t = 0, 1, \cdots \end{aligned}$$

Now we have,

Theorem 1. y_t is generated by a WN-SS model with parameters (A_*, B, C, F) iff y_t is generated by an RRO model with parameters (A, B, C, F).

Proof.

If the WN-SS model holds we have simply,

$$\begin{aligned} \xi_{t+1} &= A_*\xi_t + By_t \\ &= A_*\xi_t + B(FC\xi_t + \epsilon_t) \\ &= (A_* + BFC)\xi_t + B\epsilon_t \\ &= A\xi_t + B\epsilon_t \end{aligned}$$

as required.

We now show (A, B, C) is minimal. Suppose (A, B) is not controllable then there exists a left eigenvector q of A with corresponding eigenvalue λ such that $q^T A = \lambda q, q^T B = 0$. But then $\lambda q^T = q^T(A_* + BFC) = q^T A_*$ which now contradicts W_2. Similarly we can establish that (A, C) is observable .

For the converse we simply argue in reverse

$$\begin{aligned} \xi_{t+1} &= A\xi_t + B\epsilon_t \\ &= A\xi_t + B(y_t - FC\xi_t) \\ &= (A - BFC)\xi_t + By_t \\ &= A_*\xi_t + B\epsilon_t \end{aligned}$$

as required. Also we have $W_0 \Rightarrow W_*$. And by the same arguments in reverse (A_*, B, C) inherits minimality from (A, B, C). And the result is established.

B. DIM Model

The natural extension of the DIM model to the VARMA case is, formally,

$$y_t = N(z^{-1})D^{-1}(z^{-1})G^T y_t + \epsilon_t$$

where $G_{d\times k}$ is of rank $k < min(d,n)$ and $N(z^{-1}), D(z^{-1})$ are specified as before. Again we proceed immediately to respecify this in SS terms.

DIM-SS Model.
We say y_t is generated by a DIM-SS model with parameters (A_*, B, C, G) if firstly the parameters satisfy the conditions:
D_*: A_* is a stability matrix.
D_1: $G_{d\times k}$ is of rank $k < min(d,n)$.
D_2: $A_{*,n\times n}, C_{d\times n}$ is observable; $A_*, B_{n\times k}$ is controllable. Again this means A_*, B, C is minimal.
Secondly we are given a WN sequence ϵ_t of covariance Σ and an initial condition ξ_0 of a $n-$ dimensional state ξ_t. Then (y_t, ξ_t) are generated recursively by

$$\begin{aligned} y_t &= C\xi_t + \epsilon_t, t = 0, 1, \cdots \\ \xi_{t+1} &= A_*\xi_t + BG^T y_t, t = 0, 1, \cdots \end{aligned}$$

We now see that formally, the model posses the same kind of causality structure pointed out in [12] for the VAR case. If $G_\perp$ is a $d\times d-k$ matrix of rank $d-k$ which is orthogonal to G then multiplying through the defining equation by G and separately by $G_\perp$ and reorganizing yields

$$\begin{pmatrix} I - G^T C(zI - A_*)^{-1}B & 0 \\ -G_\perp C(zI - A_*)^{-1}B & I \end{pmatrix} \begin{pmatrix} G^T y_t \\ G_\perp^T y_t \end{pmatrix} = \begin{pmatrix} G^T \epsilon_t \\ G_\perp^T \epsilon_t \end{pmatrix}$$

The stability of A_* ensures the filtering on the LHS is well defined. According to results of [14], from this structure, and assuming stationarity, we can conclude that $G_\perp^T y_t$ does not weakly Granger cause $G^T y_t$. This again points up an important property of the model namely that it implicitly finds Granger-causal structure if there is any.

To establish our next result we introduce an innovations SS model with reduced rank gain (RRG).

RRG Model.
We say y_t is generated by an RRG model with parameters (A, B, C, G) if firstly the parameters satisfy the conditions: D_0, D_1, D_3,
D_0: $A_* = A - BG^T C$ is a stability matrix.
D_3: (A, C) is observable; (A, B) is controllable .
Note that D_0 is consistent with D_3.
And secondly we are given a WN sequence ϵ_t of covariance Σ and an initial condition ξ_0 of a $n-$ dimensional state ξ_t. Then (y_t, ξ_t) are generated recursively by the innovations state space model

$$\begin{aligned} y_t &= C\xi_t + \epsilon_t, t = 0, 1, \cdots \\ \xi_{t+1} &= A\xi_t + BG^T \epsilon_t, t = 0, 1, \cdots \end{aligned}$$

Now we have,
Theorem 2. y_t is generated by a DIM-SS model with parameters (A_*, B, C, G) iff y_t is generated by an RRG model with parameters (A, B, C, G).
Proof.
The proof is similar to that of Theorem 1. If the DIM-SS model holds then we have

$$\begin{aligned} \xi_{t+1} &= A_*\xi_t + BG^T y_t \\ &= A_*\xi_t + BG^T(C\xi_t + \epsilon_t) \\ &= (A_* + BG^T C)\xi_t + BG^T \epsilon_t \\ &= A\xi_t + BG^T \epsilon_t \end{aligned}$$

as required. We now show (A, B, C) inherits minimality. If (A, B) is not controllable then there exists a left eigenvector q and corresponding eigenvalue λ with $q^T A = \lambda q^T, q^T B = 0$. But then $\lambda q^T = q^T(A_* + BG^T C) = q^T A_*$ which now contradicts D_2. Similarly observability is inherited.

For the converse we just repeat the argument in reverse as we did for Theorem 1.

Remarks.
(i) The stationarity alluded to in the causality discussion above will hold if A is a stability matrix.
(ii) The pair of theorems are quite remarkable in providing very simple and natural interpretations of the two kinds of RRTS models in SS terms.

IV. Reduced Rank Subspace Algorithms

We now construct subspace estimators for the parameters in the RRO,RRG models by modifying standard subspace construction procedures. We use canonical correlations type weighting.

A. Preliminaries

Since our procedures rely on some standard subspace computations we recap some basic material briefly [15]. Given data $y_t, t = 1, \cdots, \bar{T}$ choose a lag m and set $N = md, T = \bar{T} - 2m + 1$. Then form the 'past' and 'future' matrices

$$\underset{N\times T}{Y_-} = \begin{pmatrix} y_m & y_{m+1} & \cdot\cdot & y_{\bar{T}-m} \\ y_{m-1} & \cdot & \cdot\cdot & \cdot \\ \cdot & \cdot & \cdot\cdot & \cdot \\ y_1 & \cdot & \cdot\cdot & y_{\bar{T}-2m+1} \end{pmatrix}$$

$$\underset{N\times T}{Y_+} = \begin{pmatrix} y_{m+1} & y_{m+2} & \cdot\cdot & y_{\bar{T}-m+1} \\ y_{m+2} & \cdot & \cdot\cdot & \cdot \\ \cdot & \cdot & \cdot\cdot & \cdot \\ y_{2m} & \cdot & \cdot\cdot & y_{\bar{T}} \end{pmatrix}$$

Next introduce the near-Hankel matrix

$$\mathcal{H}_{N\times N} = Y_+ Y_-^T \frac{1}{T}$$

and the near block Toeplitz matrices

$$\Sigma_- = \frac{1}{T} Y_- Y_-^T, \Sigma_+ = \frac{1}{T} Y_+ Y_+^T$$

and carry out Cholesky factorizations

$$L_+ \Sigma_+ L_+^T = I, L_- \Sigma_- L_-^T = I$$

Then carry out an SVD

$$L_+ \mathcal{H} L_-^T = U \Lambda V^T$$

and keep the first n columns, U_* of U; the corresponding first n diagonal entries Λ_* of the diagonal matrix of singular

values Λ; the corresponding first n rows V_*^T of V^T.
Choice of n.
We propose the criterion of [16]

$$FV_r = -(T-1)\Sigma_{r+1}^N ln(1-\Lambda_j^2) - 2(N-r)^2$$

The idea is simply to plot FV_r for a minimum in r. We denote the minimizer as n. This is an Akaike type criterion; a Bayes type version is easily obtained by multiplying the penalty term $2(N-r)^2$ by $ln(T-1)$.
Now set

$$K = \Lambda_*^{\frac{1}{2}} V_*^T L_-$$

Then we construct a state estimator as

$$\begin{aligned} \hat{\xi}_t &= K y_{-,t}, t = m+1, \cdots, T+m \\ \underset{N\times 1}{y_{-,t}} &= \begin{pmatrix} y_{t-1} \\ y_{t-2} \\ \cdot \\ y_{t-m} \end{pmatrix}, t = m+1, \cdots, T-m+1 \end{aligned}$$

In the appendix we show for completeness (the known result)

$$S_{oo} = \frac{1}{T}\Sigma_1^T \hat{\xi}_t \hat{\xi}_t^T = \Lambda_* \quad (4.1)$$

B. RRO Model

Given the state estimator $\hat{\xi}_t$. we can estimate FC by a RRR of y_t on $\hat{\xi}_t$

$$min_{F,C} \Sigma_1^T \parallel y_t - F_{d\times k} C_{k\times n} \hat{\xi}_t \parallel^2$$

This is a classical RRR and we can read out the solution from [4] except that we use a more compact SVD representation.
First we form the covariance matrices

$$\begin{aligned} S_{yy} &= \frac{1}{T}\Sigma_1^T y_t y_t^T \\ S_{yo} &= \frac{1}{T}\Sigma_1^T y_t \hat{\xi}_t^T \\ S_{oo} &= \frac{1}{T}\Sigma_1^T \hat{\xi}_t \hat{\xi}_t^T \end{aligned}$$

Next we carry out an SVD

$$\begin{aligned} \bar{M}_{d\times n} &= S_{yy}^{-\frac{1}{2}} S_{yo} S_{oo}^{-\frac{1}{2}} = \bar{P}_{d\times d} \bar{D}_{d\times n} \bar{Q}_{n\times n}^T \\ &= \bar{P}_* \bar{D}_* \bar{Q}_*^T + \bar{P}_\perp \bar{D}_\perp \bar{Q}_\perp^T \end{aligned}$$

Where $\bar{Q}, \bar{P}$ are orthogonal matrices and $\bar{D}$ contains the singular values. The number of these is p =min(d,n). If $d \leq n$, then the left $d\times d$ submatrix is diagonal with the singular values in decreasing order down the diagonal; elsewhere are zeros. If $d > n$ then it is the upper $n\times n$ submatrix that has the singular values down the diagonal. By construction the singular values lie between 0 and 1.

We keep the first k columns $\bar{P}_{*,d\times k}$ of $\bar{P}$; the corresponding k largest singular values in $\bar{D}_{*,k\times k}$; and the corresponding first k rows $\bar{Q}_{*,k\times n}$ of $\bar{Q}$. And $\bar{P}_\perp, \bar{D}_\perp, \bar{Q}_\perp$ denote the remaining singular quantities.
Choice of k.
We propose again the criterion of [16] which is now,

$$FV_r = -(T-1)\Sigma_{r+1}^p ln(1-\bar{D}_j^2) - 2(d-r)(n-r)$$

The idea again to plot FV_r for a minimum in r. We denote the minimizer as k. Again a Bayes type version is easily obtained by multiplying the penalty term $2(d-r)(n-r)$ by $ln(T-1)$.
$\boxed{\hat{F}, \hat{C}}$.
Now we form the estimators

$$\hat{F} = S_{yy}^{\frac{1}{2}} \bar{P}_*, \hat{C} = \bar{D}_* \bar{Q}_*^T S_{oo}^{-\frac{1}{2}} \quad (4.2)$$

Note for future use that

$$\begin{aligned} \hat{F}\hat{C} &= S_{yy}^{\frac{1}{2}} \bar{P}_* \bar{D}_* \bar{Q}_*^T S_{oo}^{-\frac{1}{2}} \quad (4.3) \\ \Rightarrow \hat{F}\hat{C} S_{yo}^T &= S_{yy}^{\frac{1}{2}} \bar{P}_* \bar{D}_* \bar{Q}_*^T \bar{M}^T S_{yy}^{\frac{1}{2}} \\ &= S_{yy}^{\frac{1}{2}} \bar{P}_* \bar{D}_* \bar{Q}_*^T \bar{Q}_* \bar{D}_* \bar{P}_*^T S_{yy}^{\frac{1}{2}} \\ &= S_{yy}^{\frac{1}{2}} \bar{P}_* \bar{D}_*^2 \bar{P}_*^T S_{yy}^{\frac{1}{2}} \quad (4.4) \end{aligned}$$

$\boxed{\hat{\Sigma}}$.
Continuing we introduce the residual or error signal

$$e_t = y_t - \hat{F}\hat{C}\hat{\xi}_t$$

The error covariance is

$$\begin{aligned} \hat{\Sigma} &= S_e = \frac{1}{T}\Sigma_1^T e_t e_t^T \\ &= \frac{1}{T}\Sigma_1^T (y_t - \hat{F}\hat{C}\hat{\xi}_t)(y_t - \hat{F}\hat{C}\hat{\xi}_t)^T \\ &= S_{yy} - S_{yo}\hat{C}^T\hat{F}^T - \hat{F}\hat{C}S_{yo}^T + \hat{F}\hat{C}S_{oo}\hat{C}^T\hat{F}^T \\ &= S_{yy} - S_{yy}^{\frac{1}{2}} \bar{P}_* \bar{D}_*^2 \bar{P}_*^T S_{yy}^{\frac{1}{2}} \\ &= S_{yy}^{\frac{1}{2}} (I - \bar{P}_* \bar{D}_*^2 \bar{P}_*^T) S_{yy}^{\frac{1}{2}} \quad (4.5) \end{aligned}$$

Note that since the singular values are less than one this matrix has full rank.
$\boxed{\hat{A}, \hat{B}}$.
Now we estimate A, B by least squares

$$min_{A,B} \Sigma_1^T \parallel \hat{\xi}_{t+1} - (A,B)\begin{pmatrix} \hat{\xi}_t \\ e_t \end{pmatrix} \parallel^2$$

A perturbation argument leads to the Euler equations

$$0 = \Sigma_1^T (\hat{\xi}_{t+1} - (\hat{A},\hat{B})\begin{pmatrix} \hat{\xi}_t \\ e_t \end{pmatrix})(\hat{\xi}_t^T, e_t^T) \quad (4.6)$$

To solve this system in a compact way we recall the standard subspace estimators. We denote them $\hat{A}_o, \hat{B}_o, \hat{C}_o$ and also introduce the standard error signal

$$\begin{aligned} e_{ot} &= y_t - \hat{C}_o \hat{\xi}_t \\ &= y_t - S_{yo} S_{oo}^{-1} \hat{\xi}_t \end{aligned}$$

Note that $e_{ot}, \hat{\xi}_t$ have an orthogonality property

$$\Sigma_1^T e_{ot} \hat{\xi}_t^T = 0$$

Further $\hat{A}_o, \hat{B}_o$ obey the Euler equations

$$0 = \Sigma_1^T(\hat{\xi}_{t+1} - (\hat{A}_o, \hat{B}_o)\begin{pmatrix} \hat{\xi}_t \\ e_{ot} \end{pmatrix})(\hat{\xi}_t^T, e_{ot}^T) \quad (4.7)$$

$$\Rightarrow \hat{A}_o = S_{1o}S_{oo}^{-1}, \hat{B}_o = S_{ye}S_{eeo}^{-1} \quad (4.8)$$

$$S_{10} = \frac{1}{T}\Sigma_1^T\hat{\xi}_{t+1}\hat{\xi}_t^T$$

$$S_{yeo} = \frac{1}{T}\Sigma_1^T y_t e_{ot}^T$$

$$S_{eeo} = \frac{1}{T}\Sigma_1^T e_{ot}e_{ot}^T$$

Now observe that

$$\begin{aligned} e_{ot} &= y_t - S_{yo}S_{oo}^{-1}\hat{\xi}_t \\ &= y_t - S_{yy}^{\frac{1}{2}}\bar{M}S_{oo}^{-\frac{1}{2}}\hat{\xi}_t \\ &= y_t - S_{yy}^{\frac{1}{2}}(\bar{P}_\perp\bar{D}_\perp\bar{Q}_\perp^T + \bar{P}_*\bar{D}_*\bar{Q}_*^T)S_{oo}^{-\frac{1}{2}}\hat{\xi}_t \\ &= y_t - H\hat{\xi}_t - \hat{F}\hat{C}\hat{\xi}_t, \text{by (4.3)} \\ &= e_t - H\hat{\xi}_t \\ H &= S_{yy}^{\frac{1}{2}}\bar{P}_\perp\bar{D}_\perp\bar{Q}_\perp^T S_{oo}^{-\frac{1}{2}} \\ &= S_{yy}^{\frac{1}{2}}(\bar{M} - \bar{P}_*\bar{D}_*\bar{Q}_*^T)S_{oo}^{-\frac{1}{2}} \\ &= S_{yo}S_{oo}^{-1} - \hat{F}\hat{C} \end{aligned}$$

Thus we have

$$\begin{pmatrix} \hat{\xi}_t \\ e_{ot} \end{pmatrix} = \begin{pmatrix} I & 0 \\ -H & I \end{pmatrix}\begin{pmatrix} \hat{\xi}_t \\ e_t \end{pmatrix} \quad (4.9)$$

Multiplying (4.7) on the RHS by $\begin{pmatrix} I & H^T \\ 0 & I \end{pmatrix}$ and substituting (4.9) inside the bracket yields

$$\begin{aligned} 0 &= \Sigma_1^T(\hat{\xi}_{t+1} - (\hat{A}_o, \hat{B}_o)\begin{pmatrix} I & 0 \\ -H & I \end{pmatrix}\begin{pmatrix} \hat{\xi}_t \\ e_t \end{pmatrix})(\hat{\xi}_t^T, e_t^T) \\ \Rightarrow 0 &= \Sigma_1^T(\hat{\xi}_{t+1} - (\hat{A}_o - \hat{B}_oH, \hat{B}_o)\begin{pmatrix} \hat{\xi}_t \\ e_t \end{pmatrix})(\hat{\xi}_t^T, e_t^T) \end{aligned}$$

From (4.6) we can thus read off

$$\hat{A} = \hat{A}_o - \hat{B}_oH, \hat{B} = \hat{B}_o \quad (4.10)$$

So the RRO equations are: $\hat{F}, \hat{C}$ in (4.2),$\hat{\Sigma}$ in (4.5), $\hat{A}, \hat{B}$ in (4.10); with $\hat{A}_o, \hat{B}_o$ given in (4.8). For completeness we quote the known result

$$\hat{B}_o = \Lambda_*^{\frac{1}{2}}V_*^TL_-\begin{pmatrix} I_d \\ 0 \\ . \\ 0 \end{pmatrix} \quad (4.11)$$

A proof is omitted due to its length.

C. RRG Model

As before we start with the state estimator $\hat{\xi}_t$.

$\boxed{\hat{C}, \hat{\Sigma}}$.

We estimate C, Σ in the standard way by minimising $\Sigma_1^T \parallel y_t - C\hat{\xi}_t \parallel^2$ leading to the Euler equation

$$\begin{aligned} 0 &= \Sigma_1^T(y_t - \hat{C}\hat{\xi}_t)\hat{\xi}_t^T \\ \Rightarrow \hat{C} &= \hat{C}_o = S_{yo}S_{oo}^{-1} \end{aligned} \quad (4.12)$$

In the appendix we show for completeness the known result

$$\begin{aligned} \hat{C}_o &= \Gamma_{d\times N}L_-^TV_*\Lambda_*^{-\frac{1}{2}} \\ \Gamma &= \text{first } d\times N \text{ block row of } \mathcal{H} \end{aligned} \quad (4.13)$$

Continuing, if we introduce the residual as before $e_{ot} = y_t - \hat{C}_o\hat{\xi}_t$ we have as before an orthogonality condition

$$\Sigma_1^T e_{ot}\hat{\xi}_t^T = 0$$

From this we obtain the estimator of the noise variance as

$$\begin{aligned} \hat{\Sigma} &= \hat{\Sigma}_o = S_{eeo} = \frac{1}{T}\Sigma_1^T e_{ot}e_{ot}^T \\ &= \frac{1}{T}\Sigma_1^T e_{ot}y_t^T \\ &= S_{yy} - \hat{C}_oS_{yo}^T \\ &= S_{yy} - S_{yo}S_{oo}^{-1}S_{yo}^T \end{aligned} \quad (4.14)$$

$\boxed{\hat{A}, \hat{B}, \hat{G}}$.

Next we estimate A, B, G by a classical RRR by solving

$$min_{A,B,G}\Sigma_1^T \parallel \hat{\xi}_{t+1} - A\hat{\xi}_t - BG^Te_{ot} \parallel^2$$

Because of the orthogonality between $\hat{\xi}_t, e_{ot}$ we will be able to do the optimization over A and B, G separately.

Optimizing over A leads to

$$\Sigma_1^T(\hat{\xi}_{t+1} - \hat{A}\hat{\xi}_t - \hat{B}\hat{G}^Te_{ot})\hat{\xi}_t^T = 0$$

and orthogonality yields

$$\begin{aligned} 0 &= \Sigma_1^T(\hat{\xi}_{t+1} - \hat{A}\hat{\xi}_t)\hat{\xi}_t^T \\ \Rightarrow \hat{A} = \hat{A}_o &= S_{1o}S_{oo}^{-1} \end{aligned} \quad (4.15)$$

So far our estimators agree with the classical subspace estimators. With $\hat{A}_o$ in hand we can reformulate the remaining RRR problem by introducing

$$u_t = \hat{\xi}_{t+1} - \hat{A}_o\hat{\xi}_t$$

So the problem becomes

$$min_{B,G}\Sigma_1^T \parallel u_t - BG^Te_{ot} \parallel^2$$

Again this is a classical RRR problem and we express the solution compactly as before in terms of SVD.

We first use u_t, e_{ot} to form

$$\begin{aligned} S_{uu} &= \frac{1}{T}\Sigma_1^T u_tu_t^T \\ S_{ue} &= \frac{1}{T}\Sigma_1^T u_te_{ot}^T \end{aligned}$$

and together with S_{eeo} carry out the SVD

$$M_{n\times d}^T = S_{uu}^{-\frac{1}{2}}S_{ue}S_{eeo}^{-\frac{1}{2}} = Q_{n\times n}D_{n\times d}P_{d\times d}^T$$

As before P, Q are orthogonal matrices and D contains the ordered singular values. The number of these is $q = min(d, n)$. We keep the first k columns $Q_{*,n\times k}$ of Q; the corresponding diagonalized singular values $D_{*,k\times k}$ of D; and the corresponding first k rows $P_{*,k\times d}^T$ of P^T.

Choice of k.
As before we propose to use the criterion of [16]

$$FV_r = -(T-1)\Sigma_{r+1}^{q} ln(1-D_j^2) - 2(d-r)(n-r)$$

The idea again is simply to plot FV_r for a minimum in r. We denote the minimizer as k. Again a Bayes type version is easily obtained by multiplying the penalty term $2(d-r)(n-r)$ by $ln(T-1)$.
Then we may take

$$\hat{B} = S_{uu}^{\frac{1}{2}} Q_*, \hat{G}^T = D_* P_*^T S_{eeo}^{-\frac{1}{2}} \quad (4.16)$$

So the estimators are $\hat{C}_o$ in (4.13), $\hat{A}_o$ in (4.15),$\hat{B}, \hat{G}$ in (4.16),$\hat{\Sigma}$ in (4.14).

With a little further calculation we can simplify things as follows. We have

$$\begin{aligned} S_{uu} &= \frac{1}{T}\Sigma_1^T(\hat{\xi}_{t+1} - \hat{A}_o\hat{\xi}_t)\hat{\xi}_{t+1}^T \\ &= S_{11} - \hat{A}_o S_{1o}^T \\ &= S_{11} - S_{1o} S_{oo}^{-1} S_{1o}^T \\ S_{11} &= \frac{1}{T}\Sigma_1^T \hat{\xi}_{t+1}\hat{\xi}_{t+1}^T \end{aligned}$$

Continuing

$$\begin{aligned} S_{ue} &= \frac{1}{T}\Sigma_1^T(\hat{\xi}_{t+1} - \hat{A}_o\hat{\xi}_t)e_{ot}^T \\ &= \frac{1}{T}\Sigma_1^T \hat{\xi}_{t+1} e_{ot}^T \\ &= \frac{1}{T}\Sigma_1^T \hat{\xi}_{t+1}(y_t - \hat{C}_o\hat{\xi}_t)^T \\ &= S_{1y} - S_{1o} S_{oo}^{-1} S_{yo}^T \end{aligned}$$

Remarks.
(i) We have indicated along the way some computational simplifications in (4.1,4.13,4.11).
(ii) It is well known [15] that the Q-R algorithm provides a numerically compact and reliable way to carry out many of the computations associated with subspace methods. Details of this kind will be provided elsewhere.

V. Conclusion

In this paper we have provided, for the first time, a formulation of reduced rank time series models in state space terms. This has allowed an extension of these models from VAR to VARMA. The characterization turns out to be very simple involving either a rank reduced observation matrix (RRO) or a rank reduced gain matrix (RRG). We have developed two new asociated subpsace fitting algorithms including methods for choice of minimal state space dimension n as well as for reduction rank k.

VI. Appendix

Proof of (4.1).

$$\begin{aligned} S_{oo} &= \frac{1}{T}\Sigma_{m+1}^{T+m+1}\hat{\xi}_t\hat{\xi}_t^T \\ &= \Lambda_*^{\frac{1}{2}} V_*^T L_- \frac{1}{T}\Sigma_{m+1}^{T+m+1} y_{-,t} y_{-,t}^T L_-^T V_* \Lambda_*^{\frac{1}{2}} \\ &= \Lambda_*^{\frac{1}{2}} V_*^T L_- \frac{1}{T} Y_- Y_-^T L_-^T V_* \Lambda_*^{\frac{1}{2}} \\ &= \Lambda_*^{\frac{1}{2}} V_*^T L_- \Sigma_- L_-^T V_* \Lambda_*^{\frac{1}{2}} \\ &= \Lambda_*^{\frac{1}{2}} V_*^T V_* \Lambda_*^{\frac{1}{2}} \\ \Rightarrow S_{oo} &= \Lambda_* \end{aligned}$$

Proof of (4.13).

$$\begin{aligned} \hat{C}_o &= \frac{1}{T}\Sigma_{m+1}^{T+m+1} y_t \hat{\xi}_t^T S_{oo}^{-1} \\ &= \frac{1}{T}\Sigma_{m+1}^{T+m+1} y_t y_{-,t}^T K^T \Lambda_*^{-1} \\ &= \frac{1}{T}\Sigma_{m+1}^{T+m+1} y_t y_{-,t}^T L_-^T V_* \Lambda_*^{-\frac{1}{2}} \end{aligned}$$

And we now observe that this is (4.13)

References

[1] JH Stock and MW Watson, "Forecasting using principal components from a large number of predictors", *Journal of the American Statistical Association*, vol. 97, pp. 1167–1179, 2002.
[2] J Bai and S Ng, "Determinig the number of factors in approximate factor models", *Econometrica*, vol. 70, pp. 191–221, 2002.
[3] M Forni, M Hallin, M Lippi, and L Reichlin, "The generalized dynamic factor model consistency and rates", *Journal of Econometrics*, vol. 119, pp. 231–255, 2004.
[4] G Reinsel and P Velu, *Multivariate Reduced-Rank regression*, Springer-Verlag, New York, 1998.
[5] TW Anderson, "Estimating linear restrictions on regression coefficients for multivariate normal distributions", *Ann Math Stat*, vol. 22, pp. 327–351, 1951.
[6] DR Brillinger, "The canonical analysis of stationary time series", in *Multivariate Analysis*. Academic Press, 1969, pp. 331–350.
[7] D.R. Brillinger, *Time Series: data analysis and theory*, Holden-Day, Inc.,, San Francisco, 1981.
[8] J. Hamilton, *Time Series Analysis*, Princeton Univ. Press, Princeton, New Jersey, 1994.
[9] JH Manton and Y Hua, "Convolutive reduced rank wiener filtering", in *Proc IEEE ICASSP Salt Lake City*. IEEE, 2001, pp. –.
[10] B Wahlberg and M Jansson, "4sid linear regression", in *Proc 33rd Conference on Decision and Control Lake Buena Vista, Florida*. IEEE, 1994, p. 4pp.
[11] T Gustafsson and BD Rao, "Statistical analysis of subspace-based estimation of reduced-rank linear regressions", *IEEE Trans Sig Proc*, vol. 50, pp. 151, 2002.
[12] Solo V, "High dimensional point process system identification: Pca and dynamic index models", in *Proc IEEE CDC*. IEEE, 2006, pp. –.
[13] GC Reinsel, "Some results on multivariate autoregressive index models", *Biometrika*, vol. 70, pp. 145–156, 1983.
[14] P E Caines and W W Chan, "Feedback between stationary stochastic processes", *IEEE Trans Autom Contr*, vol. 20, pp. 498–508, 1975.
[15] P Van Overschee and B De Moor, *Subspace Identification for linear systems: Theory, Implementation and Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.
[16] Y Fujikoshi and LG Veitch, "Estimation of dimensionality in canonical correlation analysis", *Biometrika*, vol. 66, pp. 345–351, 1979.