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 (RMR) 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 RMR and RRTS. It covers model formulation, model fitting and asymptotics. Some of the work on RRR has been described by a VAR model of order $d$. The McMillan degree may be much less than $d$ 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

$$y_t = \Sigma u_{t-1} + \epsilon_t, t = 1, \ldots, \bar{T}$$

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

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 $\sim 3700$ 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

$$y_t = \Sigma u_{t-1} + \epsilon_t, t = 1, \ldots, \bar{T}$$

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

II. Review of Reduced Rank Time Series Models

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

$$y_t = \Sigma u_{t-1} + \epsilon_t, t = 1, \ldots, \bar{T}$$

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

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 $\sim 3700$ 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

$$y_t = \Sigma u_{t-1} + \epsilon_t, t = 1, \ldots, \bar{T}$$

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

II. Review of Reduced Rank Time Series Models

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

$$y_t = \Sigma u_{t-1} + \epsilon_t, t = 1, \ldots, \bar{T}$$

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

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 $\sim 3700$ 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

$$y_t = \Sigma u_{t-1} + \epsilon_t, t = 1, \ldots, \bar{T}$$

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

II. Review of Reduced Rank Time Series Models

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

$$y_t = \Sigma u_{t-1} + \epsilon_t, t = 1, \ldots, \bar{T}$$

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

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 $\sim 3700$ 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

$$y_t = \Sigma u_{t-1} + \epsilon_t, t = 1, \ldots, \bar{T}$$

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

II. Review of Reduced Rank Time Series Models

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

$$y_t = \Sigma u_{t-1} + \epsilon_t, t = 1, \ldots, \bar{T}$$

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

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 $\sim 3700$ 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

$$y_t = \Sigma u_{t-1} + \epsilon_t, t = 1, \ldots, \bar{T}$$

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

II. Review of Reduced Rank Time Series Models

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

$$y_t = \Sigma u_{t-1} + \epsilon_t, t = 1, \ldots, \bar{T}$$

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

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 $\sim 3700$ 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

$$y_t = \Sigma u_{t-1} + \epsilon_t, t = 1, \ldots, \bar{T}$$

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

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

B. Methods in Multivariate Reduced Rank Regressions (RRR)

- Methods in multivariate reduced rank regression (RMR) go back to [5] 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 RMR 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 connection 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 RRR 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.
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 possesses 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,
\[
y_t = F D^{-1}(z^{-1}) N(z^{-1}) y_t + \epsilon_t
\]
where, \(F_{d \times k}\) is of rank \(k < \min(d,n)\), \(N = \text{McMillan degree.} \); \(D_{k \times k}(z), N_{k \times d}(z^{-1})\) 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 \((A_s, B, C, F)\) if firstly the parameters satisfy the following conditions:
\(W_s: A_s\) is a stability matrix.
\(W_1: \ F_{d \times k}\) is of rank \(k < \min(d,n)\).
\(W_2: A_{s \times n}, C_{k \times n}\) is observable; \(A_s, B_{n \times d}\) is controllable.
Of course this means \(A_s, B, C\) is minimal.

Secondly we are given a white noise sequence \(\{\epsilon_t\}\) i.e. the sequence is uncorrelated, has zero mean and covariance \(\Sigma\). And are also given an initial condition \(\xi_0\) for a \(n-\) dimensional state \(\xi_t\). Then we generate \((y_t, \xi_{t+1})\) recursively as follows,
\[
y_t = FC\xi_t + \epsilon_t, t = 0, 1, \cdots
\]
\[
\xi_{t+1} = A_s\xi_t + By_t, t = 0, 1, \cdots
\]

Formally we then have
\[
y_t = FC(z I - A_s)^{-1} By_t + \epsilon_t
\]

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:
\(W_0, W_1, W_3\).
\(W_0: A_s = A - BFC\) 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 \(\epsilon_t\) of covariance \(\Sigma\) and an initial condition \(\xi_0\) of a \(n-\) dimensional state \(\xi_t\). Then \((y_t, \xi_t)\) are generated recursively by the innovations state space model
\[
y_t = FC\xi_t + \epsilon_t, t = 0, 1, \cdots
\]
\[
\xi_{t+1} = A\xi_t + B\epsilon_t, t = 0, 1, \cdots
\]

Now we have,

**Theorem 1.** \(y_t\) is generated by a WN-SS model with parameters \((A_s, B, C, F)\) 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,
\[
\xi_{t+1} = A_s\xi_t + By_t
\]
\[
\xi_{t+1} = A_s\xi_t + B(FC\xi_t + \epsilon_t)
\]
\[
\xi_{t+1} = (A_s + BFC)\xi_t + B\epsilon_t
\]
\[
\xi_{t+1} = A\xi_t + B\epsilon_t
\]
as required. We now show \((A_s, B, C, F)\) is minimal. Suppose \((A, B)\) is not controllable then there exists a left eigenvector \(q\) of \(A\) with corresponding eigenvalue \(\lambda\) such that \(q^T A = \lambda q, q^T B = 0\). But then \(\lambda q^T = q^T (A_s + BFC) = q^T A_s\) which now contradicts \(W_2\). Similarly we can establish that \((A, C)\) is observable.

For the converse we simply argue in reverse
\[
\xi_{t+1} = A\xi_t + B\epsilon_t
\]
\[
\xi_{t+1} = A\xi_t + B(y_t - FC\xi_t)
\]
\[
\xi_{t+1} = (A - BFC)\xi_t + By_t
\]
\[
\xi_{t+1} = A\xi_t + B\epsilon_t
\]
as required. Also we have \(W_0 \Rightarrow W_s\). And by the same arguments in reverse \((A_s, B, C)\) 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:

$D_0$: $A_*$ is a stability matrix.

$D_1$: $G_{d \times k}$ is of rank $k < \min(d, n)$.

$D_2$: $A_{n \times n}, C_{d \times n}$ are observable; $A_*, B_{n \times k}$ is controllable. Again this means $A_*, B, C$ is minimal.

Secondly we are given a WN sequence $e_t$ of covariance $\Sigma$ and an initial condition $\xi_0$ of a $n-$ dimensional state $\xi_t$. Then $(y_t, \xi_t)$ are generated recursively by

$$
\begin{align*}
    y_t &= C\xi_t + \epsilon_t, t = 0, 1, \ldots \\
    \xi_{t+1} &= A_\ast \xi_t + B G^T y_t, t = 0, 1, \ldots
\end{align*}
$$

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

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

The stability of $A_\ast$ ensures the filtering on the LHS is well defined. According to results of [14], from this structure, and assuming stationarity, we can conclude that $G^T y_t$ 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:

$D_0$, $D_1$, $D_2$.

$D_3$: $(A, C)$ is observable; $(A, B)$ is controllable.

Note that $D_0$ is consistent with $D_3$.

And secondly we are given a WN sequence $e_t$ of covariance $\Sigma$ and an initial condition $\xi_0$ of a $n-$ dimensional state $\xi_t$. Then $(y_t, \xi_t)$ are generated recursively by the innovations state space model

$$
\begin{align*}
    y_t &= C\xi_t + \epsilon_t, t = 0, 1, \ldots \\
    \xi_{t+1} &= A_\ast \xi_t + B G^T \epsilon_t, t = 0, 1, \ldots
\end{align*}
$$

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

$$
\begin{align*}
    \xi_{t+1} &= A_\ast \xi_t + B G^T y_t \\
    &= A_\ast \xi_t + B G^T (C \xi_t + \epsilon_t) \\
    &= (A_\ast + B G^T C) \xi_t + B G^T \epsilon_t \\
    &= A_\ast \xi_t + B G^T \epsilon_t
\end{align*}
$$

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 $\lambda$ with $q^T A = \lambda q^T, q^T B = 0$. But then $\lambda q^T = q^T (A_\ast + B G^T C) = q^T A_\ast$ 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 $y_t, t = 1, \ldots, T$ choose a lag $m$ and set $N = m d, T = T - 2m + 1$, then form the ‘past’ and ‘future’ matrices

$$
\begin{align*}
    Y_- &= \begin{pmatrix}
        y_m & y_{m+1} & \cdots & y_{T-m} \\
        y_{m-1} & \ddots & \ddots & \ddots \\
        \vdots & \ddots & \ddots & \ddots \\
        y_1 & \ddots & \ddots & \ddots \\
        y_1 & \ddots & \ddots & \ddots \\
        y_{2m} & \ddots & \ddots & \ddots
    \end{pmatrix} \\
    Y_+ &= \begin{pmatrix}
        y_1 & y_{m+2} & \cdots & y_{T-m+1} \\
        y_{m+1} & \ddots & \ddots & \ddots \\
        \vdots & \ddots & \ddots & \ddots \\
        y_1 & \ddots & \ddots & \ddots \\
        y_{m+2} & \ddots & \ddots & \ddots \\
        y_{2m} & \ddots & \ddots & \ddots
    \end{pmatrix} \\
    N \times T
\end{align*}
$$

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 $\Lambda_*$ 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]

$$FV_r = -(T-1)\Sigma^N_{r+1} \ln(1 - \Lambda^2) - 2(N-r)^2$$

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

$$K = \Lambda^T_2 V^T_* L_\ast$$

Then we construct a state estimator as

$$\dot{\xi}_t = K y_{-t}, t = m + 1, \cdots, T + m$$

$$y_{-t} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-m} \end{bmatrix}, t = m + 1, \cdots, T + m + 1$$

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

$$S_{oo} = \frac{1}{T} \Sigma^T_1 \xi_\ast \xi_\ast^T = \Lambda_\ast$$ (4.1)

B. RRO Model

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

$$\min_{F,C} \Sigma^T_1 \| y_t - F_{d \times k} C_{k \times n} \dot{\xi}_t \|^2$$

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

$$S_{yy} = \frac{1}{T} \Sigma^T_1 y_t y_t^T$$

$$S_{yo} = \frac{1}{T} \Sigma^T_1 y_t \xi_\ast^T$$

$$S_{oo} = \frac{1}{T} \Sigma^T_1 \xi_\ast \xi_\ast^T$$

Next we carry out an SVD

$$\tilde{M}_{d \times n} = S_{yy} \tilde{S}_{yo} S_{oo}^{-\frac{1}{2}} = \tilde{P}_d \tilde{Q}_n \tilde{Q}_n^T$$

$$= \tilde{P}_d \tilde{D}_r \tilde{Q}_n^T$$

Where $\tilde{Q}, \tilde{P}$ are orthogonal matrices and $\tilde{D}$ 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 $\tilde{P}_{\ast, d \times k}$ of $\tilde{P}$; the corresponding k largest singular values in $\tilde{D}_{\ast, k \times k}$; and the corresponding first k rows $\tilde{Q}_{\ast, k \times n}$ of $\tilde{Q}$. And $\tilde{P}_\perp, \tilde{D}_\perp, \tilde{Q}_\perp$ denote the remaining singular quantities.

Choice of k.

We propose again the criterion of [16] which is now,

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

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 $\ln(T-1)$.

Now we form the estimators

$$\hat{F} = S_{yy}^\frac{1}{2} \tilde{P}_\ast \tilde{C} = \tilde{D}_\ast \tilde{Q}_n^T \tilde{S}_{oo}^{-\frac{1}{2}}$$ (4.2)

Note for future use that

$$\hat{F} \tilde{C} = S_{yy}^\frac{1}{2} \tilde{P}_\ast \tilde{Q}_n^T S_{yy}^{-\frac{1}{2}}$$ (4.3)

$$\Rightarrow \hat{F} \tilde{C} S_{yo}^T = S_{yy}^\frac{1}{2} \tilde{P}_\ast \tilde{Q}_n^T \tilde{M}^T S_{yy}^\frac{1}{2}$$

$$= S_{yy}^\frac{1}{2} \tilde{P}_\ast \tilde{Q}_n^T \tilde{Q}_\ast \tilde{D}_\perp \tilde{P}_\perp^T S_{yy}^\frac{1}{2}$$

$$= S_{yy}^\frac{1}{2} \tilde{P}_\ast \tilde{D}_\ast^2 \tilde{P}_\ast^T S_{yy}^\frac{1}{2}$$ (4.4)

$\hat{S}_\ast$

Continuing we introduce the residual or error signal

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

The error covariance is

$$\hat{S} = S_e = \frac{1}{T} \Sigma^T_1 e_t e_t^T$$

$$= \frac{1}{T} \Sigma^T_1 (y_t - \hat{F} \tilde{C} \dot{\xi}_t)(y_t - \hat{F} \tilde{C} \dot{\xi}_t)^T$$

$$= S_{yy} - S_{yy} \tilde{C} \tilde{C}^T \hat{F} \tilde{C} - \tilde{F} \tilde{C} S_{yo} \tilde{C}^T \hat{F} \tilde{C}^T$$

$$= S_{yy} - S_{yy} \tilde{P}_\ast \tilde{D}_\ast^2 \tilde{P}_\ast^T S_{yy}$$ (4.5)

Note that since the singular values are less than one this matrix has full rank.

$$[\hat{A}, \hat{B}]$$

Now we estimate $A,B$ by least squares

$$\min_{A,B} \Sigma^T_1 \| \xi_{t+1} - (A,B) (\dot{\xi}_t, e_t) \|^2$$

A perturbation argument leads to the Euler equations

$$0 = \Sigma^T_1 (\dot{\xi}_{t+1} - (\hat{A}, \hat{B}) \left( \dot{\xi}_t, e_t \right)) (\dot{\xi}_T, e_T)$$ (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

$$e_{ot} = y_t - \hat{C}_o \dot{\xi}_t$$

$$= y_t - S_{yo} S_{oo}^{-1} \dot{\xi}_t$$

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

$$\Sigma^T_1 e_{ot} \dot{\xi}_t = 0$$
Further $\hat{A}_o, \hat{B}_o$ obey the Euler equations

$$
0 = \Sigma_1^T (\xi_{t+1} - (\hat{A}_o, \hat{B}_o) \left( \begin{array}{c} \xi_t \\ e_{ot} \end{array} \right)) (\xi_{t}^T, e_{ot}^T) \quad (4.7)
$$

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

$$
S_{yo} = \frac{1}{T} \Sigma_1^T \xi_t, \xi_t
$$

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

Now observe that

$$
\begin{align*}
e_{ot} &= y_t - S_{yo}^{-1} \xi_t \\
&= y_t - S_{yo}^{-1} M S_{oo}^{-1} \xi_t \\
&= y_t - S_{yo}^{-1} (P_\perp D_\perp Q_\perp^T + P_\perp D_\perp Q_\perp^T) S_{oo}^{-1} \xi_t \\
&= y_t - H \xi_t - \hat{F} \hat{C} \xi_t, \text{ by (4.3)} \\
&= e_t - H \xi_t \\
H &= S_{yo}^{-1} P_\perp D_\perp Q_\perp^T S_{oo}^{-1} \\
&= S_{yo}^{-1} (M - P_\perp D_\perp Q_\perp^T) S_{oo}^{-1} \\
&= S_{yo}^{-1} I - \hat{F} \hat{C}
\end{align*}
$$

Thus we have

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

Multiplying (4.7) on the RHS by $\left( \begin{array}{c} I & H \end{array} \right)$ and substituting (4.9) inside the bracket yields

$$
0 = \Sigma_1^T (\xi_{t+1} + (A_t, B_t) \left( \begin{array}{c} I \\
-H \\
I
\end{array} \right) \xi_t (\xi_t^T, e_{ot}^T)
$$

$$
\Rightarrow 0 = \Sigma_1^T (\xi_{t+1} - (\hat{A}_o - \hat{B}_o H, \hat{B}_o) \xi_t) (\xi_t^T, e_{ot}^T)
$$

From (4.6) we can thus read off

$$
\hat{A} = \hat{A}_o - \hat{B}_o H, \hat{B} = \hat{B}_o
$$

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

$$
\hat{B}_o = A_T^+ V_s^+ L - \begin{pmatrix} I_d & 0 \\
0 & 0
\end{pmatrix}
$$

A proof is omitted due to its length.

C. RRG Model

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

$$
\begin{pmatrix}
\hat{C} \\
\hat{\Sigma}
\end{pmatrix}
$$

We estimate $\hat{C}, \hat{\Sigma}$ in the standard way by minimising $\Sigma_1^T \| y_t - C \xi_t \|^2$ leading to the Euler equation

$$
0 = \Sigma_1^T (y_t - \hat{C} \xi_t) \xi_t^T \\
\Rightarrow \hat{C} = \hat{C}_o = S_{yo}^{-1}
$$

In the appendix we show for completeness the known result

$$
\hat{C}_o = \Gamma_d \times N \times V_s^T A_s^+ \xi_t \quad (4.13)
$$

\[
\Gamma = \text{first } d \times N \text{ block row of } H
\]

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

$$
\Sigma_1^T e_{ot} e_{ot}^T = 0
$$

From this we obtain the estimator of the noise variance as

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

Now we estimate $A, B, G$ by a classical RRR by solving

$$
\min_{A,B,G} \Sigma_1^T (\xi_{t+1} - \hat{A} \xi_t - B G e_{ot}) \xi_t^T = 0
$$

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

Optimizing over $A$ leads to

$$
\Sigma_1^T (\xi_{t+1} - \hat{A} \xi_t - B G e_{ot}) \xi_t^T = 0
$$

and orthogonality yields

$$
0 = \Sigma_1^T (\xi_{t+1} - \hat{A} \xi_t) \xi_t^T \\
\Rightarrow \hat{A} = \hat{A}_o = S_{yo}^{-1}
$$

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

$$
u_t = \xi_{t+1} - \hat{A}_o \xi_t
$$

So the problem becomes

$$
\min_{B,G} \Sigma_1^T \| \nu_t - B G e_{ot} \|^2
$$

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

$$
S_{uu} = \frac{1}{T} \Sigma_1^T u_t u_t^T
$$

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

and together with $S_{eeo}$ carry out the SVD

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

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 $Q_{* \times k}$ of $Q$; the corresponding diagonalized singular values $D_{* \times k}$ of $D$; and the corresponding first $k$ rows $P_{* \times k}$ of $P^T$. 

3279
Choice of $k$
As before we propose to use the criterion of [16]

$$F V_r = -(T - 1)\sum_{r+1} q(1 - D_j^2) - 2(d - r)(n - r)$$

The idea again is simply to plot $F V_r$ 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

$$\hat{B} = S_{au}Q_x, \hat{G}^T = D_xP_x^TS_{cc0}^{-\frac{1}{2}} \quad (4.16)$$

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

$$S_{uu} = \frac{1}{T} \sum_{t=1}^T (\xi_{t+1} + \hat{A}_o \hat{\xi}_t)^Tc_{ot}$$
$$= S_{11} - \hat{A}_o S_{1o}$$
$$S_{11} = \frac{1}{T} \sum_{t=1}^T \hat{\xi}_{t+1} \hat{\xi}_t^T$$

Continuing

$$S_{ue} = \frac{1}{T} \sum_{t=1}^T (\xi_{t+1} + \hat{A}_o \hat{\xi}_t)c_{ot}$$
$$= \frac{1}{T} \sum_{t=1}^T \hat{\xi}_{t+1}^Tc_{ot}$$
$$= \frac{1}{T} \sum_{t=1}^T (y_t - \hat{C}_o \hat{\xi}_t)^T \hat{\xi}_t$$
$$= S_{1y} - S_{1o}S_{1o}^{-1}S_{yo}$$

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 associated subspace 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).

$$S_{oo} = \frac{1}{T} \sum_{t=1}^T (\xi_{t+1} + \hat{A}_o \hat{\xi}_t)^T \hat{\xi}_t$$
$$= \frac{1}{T} \sum_{t=1}^T (\xi_{t+1} + \hat{A}_o \hat{\xi}_t)^T \hat{\xi}_t$$
$$= \frac{1}{T} \sum_{t=1}^T y_t \hat{y}_t^T V^T \Lambda_s$$

$$\Rightarrow S_{oo} = \Lambda_s$$

Proof of (4.13).

$$\hat{C}_o = \frac{1}{T} \sum_{t=1}^T y_t \hat{y}_t^T V^T \Lambda_s$$

And we now observe that this is (4.13)

References