Using Subspace Algorithms for the Estimation of Linear State Space Models for Over-Differenced Processes

Dietmar Bauer

[a]*Econometrics, Bielefeld University, Universitätsstrasse 25, Bielefeld, 33615, NRW, Germany*

Abstract

Subspace algorithms like *canonical variate analysis (CVA)* are regression based methods for the estimation of linear dynamic state space models. They have been shown to deliver accurate (consistent and asymptotically equivalent to quasi maximum likelihood estimation using the Gaussian likelihood) estimators for invertible stationary autoregressive moving average (ARMA) processes.
These results use the assumption that there are no zeros of the spectral density on the unit circle corresponding to the state space system.
In this note we consider vector processes made stationary by applying differencing to all variables ignoring potential cointegrating relations. We show consistency for the CVA estimators closing a gap in the literature.

Keywords: Over-differencing, linear state space systems, subspace algorithms

1. Introduction

Subspace algorithms such as the *Canonical Variate Analysis (CVA)* (Larimore, 1983) are used for the estimation of linear dynamical state space systems for time series. CVA is popular since it is numerically cheap consisting of a series of regressions, asymptotically equivalent to quasi maximum likelihood estimation (using the Gaussian likelihood) for stationary processes and robust to the existence of simple unit roots (see Bauer (2005) for a survey).

The algorithm fits a state space system in innovation form

$$y_t = Cx_t + \varepsilon_t, \quad x_{t+1} = Ax_t + B\varepsilon_t, \quad t \in \mathbb{Z}, \tag{1}$$

to an observed time series $y_t \in \mathbb{R}^s, t = 1, ..., T$. Here $A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times s}, C \in \mathbb{R}^{s \times n}$ define the state space system with system order n. We will always assume that the system is minimal (the state dimension cannot be reduced, see (Hannan and Deistler, 1988), Chapter 1, for details) and stable (such that all eigenvalues of A are smaller than one in modulus).

The innovation representation corresponds to the Wold representation of the process $(y_t)_{t \in \mathbb{Z}}$, if and only if the eigenvalues of the matrix $\underline{A} = A - BC$ are inside or on the unit circle: In this case $|\lambda_{max}(\underline{A})| \leq 1$ where $\lambda_{max}(M)$ denotes a maximum modulus eigenvalue of the matrix M.

The asymptotic properties for CVA when the data are generated from a state space system documented in the literature are restricted to the case of invertible processes where the strict inequality $|\lambda_{max}(\underline{A})| < 1$ holds. However, this restriction may be violated in particular for economic data, if the data are transformed to stationarity by temporal differencing without taking possible co-integrating relations into account. If co-integrating relations between the component variables exist and the whole time series is differenced, this leads to over-differencing in some directions introducing spectral zeros at frequency $\omega = 0$. Similar effects occur for seasonal differencing or many other forms of seasonal adjustment methods.

In such a situation the properties of the subspace estimators currently are unknown. This note closes this gap using results of Poskitt (2006) related to the autoregressive approximation of non-invertible processes. We show that CVA provides consistent estimators for the impulse response sequence also in the non-invertible case of some spectral zeros at $\omega = 0$. From the proof it is clear that analogous results also hold for zeros at different frequencies. Consistency is obtained for the integer parameter p of CVA (corresponding to the lag length of an autoregressive approximation) tending to infinity at a certain rate. We investigate the asymptotic bias arising for finite lag lengths and show that for typical choices it is not negligible as it tends to zero slower than $1/\sqrt{T}$, the typical convergence rate involved in asymptotic normality.

2. Canonical variate analysis

The CVA method of estimation is performed in three steps and uses two integers f, p ('future' and 'past') and information of the system order n (compare Bauer (2005)):

1. Obtain an estimate $\hat{x}_t$ of the state x_t for $t = p + 1, ..., T$.

2. Estimate C by regressing y_t onto $\hat{x}_t$. This step provides residuals $\hat{\varepsilon}_t = y_t - \hat{C}\hat{x}_t, t = p+1, ..., T$.
3. Estimate A and B by regressing $\hat{x}_{t+1}$ onto $\hat{x}_t$ and $\hat{\varepsilon}_t, t = p+1, ..., T-1$.

The essential idea of CVA lies in the estimation of x_t which uses the representation of the joint vector $Y_t^+ = (y_t', y_{t+1}', ..., y_{t+f-1}')'$ for some integer $f \geq n$ as the state space system implies

$$Y_t^+ = \mathcal{O}_f x_t + \mathcal{E}_f E_t^+, \quad x_t = \mathcal{K}_p Y_t^- + \delta x_t(p), \tag{2}$$

where $\mathcal{E}_f \in \mathbb{R}^{fs \times fs}$ contains the impulse response coefficients and $\mathcal{O}_f = (C', A'C', ..., (A^{f-1})'C')' \in \mathbb{R}^{fs \times n}$ denotes the observability matrix which has full column rank due to minimality. Further

$$\mathcal{K}_p = \mathbb{E}x_t(Y_t^-)'(\mathbb{E}Y_t^-(Y_t^-)')^{-1} \in \mathbb{R}^{n \times ps}$$

denotes the regression coefficient for explaining x_t by $Y_t^- = (y_{t-1}', ..., y_{t-p}')' \in \mathbb{R}^{ps}$ for integer $p \geq n$ leading to the approximation $x_t(p) = \mathcal{K}_p Y_t^-$. Then $\delta x_t(p) = x_t - x_t(p)$ denotes the approximation error.

Under the strict minimum-phase assumption implying $\underline{A}^p \to 0$ for $p \to \infty$ we may use[1] $\mathcal{K}_p = [B, \underline{A}B, \underline{A}^2B, ..., \underline{A}^{p-1}B]$ leading to

$$x_t = \mathcal{K}_p Y_t^- + \underline{A}^p x_{t-p}$$

to infer that the variance of the approximation error $\delta x_t(p)$ can be bounded by $\underline{A}^p(\mathbb{E}x_{t-p}x_{t-p}')(\underline{A}^p)'$ such that it is of order $O(\rho_0^{2p})$ where $1 > \rho_0 > |\lambda_{max}(\underline{A})|$: If in that case $p = p(T) = -c(\log T)/(2\log\rho_0)$ is used, we obtain

$$\rho_0^p = exp(-c(\log T)/(2\log\rho_0)\log\rho_0) = exp(-c(\log T)/2) = T^{-c/2}$$

such that the variance of the approximation error is of order T^{-c}. If $c > 1$ this implies that the approximation error is negligible in the usual $\sqrt{T}$ asymptotics.

For $\rho_0 = 1$ this argument does not work any more. Poskitt (2006) shows that also in this non-invertible case the approximation error decreases to zero albeit not at the same speed.

[1]This deviates from the definition above and is only used to motivate the size of the approximation error.

Example 1. Consider $y_t = \varepsilon_t - \varepsilon_{t-1} \in \mathbb{R}^s$ for independent identically distributed white noise $(\varepsilon_t)_{t\in\mathbb{Z}}$ with expectation zero and variance $\Omega > 0$. This can be represented in state space form as

$$y_t = I_s x_t + \varepsilon_t, \quad x_{t+1} = 0_{s\times s} x_t - I_s \varepsilon_t$$

and hence $x_t = -\varepsilon_{t-1}$ and $(A, B, C) = (0_{s\times s}, -I_s, I_s)$. Following Poskitt (2006) we see that $\mathcal{K}_p = -[\frac{p}{p+1} I_s, \frac{p-1}{p+1} I_s, ..., \frac{1}{p+1} I_s]$ implying that

$$\begin{aligned} x_t(p) &= \mathcal{K}_p Y_t^- = \frac{-1}{p+1} \sum_{j=1}^{p} (p+1-j)(\varepsilon_{t-j} - \varepsilon_{t-j-1}) \\ &= \frac{-1}{p+1} \sum_{j=1}^{p} (p+1-j)\varepsilon_{t-j} - \frac{-1}{p+1} \sum_{j=2}^{p+1} (p+2-j)\varepsilon_{t-j} \\ &= \frac{-1}{p+1} \left(p\varepsilon_{t-1} - \sum_{j=2}^{p} \varepsilon_{t-j} - \varepsilon_{t-p-1} \right) = -\varepsilon_{t-1} + \frac{1}{p+1} \sum_{j=1}^{p+1} \varepsilon_{t-j}. \end{aligned}$$

Denoting $\overline{\varepsilon}_{t-1}(p) = \sum_{j=1}^{p+1} \varepsilon_{t-j}/(p+1)$ we obtain $x_t = x_t(p) - \overline{\varepsilon}_{t-1}(p), \varepsilon_t(p) = y_t - x_t(p) = \varepsilon_t - \overline{\varepsilon}_{t-1}(p)$ such that the approximation error $\delta x_t(p) = -\overline{\varepsilon}_{t-1}(p)$. It follows that $\mathbb{E}\delta x_t(p)\delta x_t(p)' = \frac{1}{p+1}\Omega$. Thus the approximation error tends to zero in mean square, but the variance is of order $1/p$ and not ρ_0^{2p}. □

This example is typical. The same arguments show that the variance of the approximation error for $y_t = \Delta u_t = u_t - u_{t-1}$ for stationary process $(u_t)_{t\in\mathbb{Z}}$ with non-singular spectral density at $\omega = 0$ (not necessarily white noise) is at most of order p^{-1}.

3. Result

Poskitt (2006) derives results for the estimation accuracy for the autoregressive approximation coefficients: In his Theorem 5 he states that uniformly in $0 < p \leq H_T$ for some upper bound $H_T = O(\sqrt{T/\log T})$ and using $Q_T^2 = (\log T)/T$ we have

$$\sum_{j=1}^{p} |\hat{\alpha}_p(j) - \alpha_p(j)|^2 = O(\frac{p}{\lambda_{min}(\Gamma_p)^2} Q_T^2)$$

where $O(.)$ denotes almost sure convergence at the given rate. Here $\alpha_p(j)$ denote the autoregressive coefficients in a lag p approximation for $(y_t)_{t\in\mathbb{Z}}$ obtained from

$$[\alpha_p(1), ..., \alpha_p(p)] = \mathbb{E}y_t(Y_t^-)'\Gamma_p^{-1}, \quad \Gamma_p = (\mathbb{E}Y_t^-(Y_t^-)')^{-1}$$

and $\hat{\alpha}_p(j)$ are the corresponding least squares estimates. Poskitt (2006) uses a univariate setting, however, the extension to multivariate time series in our framework is obvious.

In this note we do not investigate autoregressive processes but state space processes with spectral zeros. We focus on the case of simple spectral zeros obtained by one time over-differencing:

Assumption 1. *The stationary process $(y_t)_{t\in\mathbb{Z}}, y_t \in \mathbb{R}^s$, is generated using a rational, stable and invertible transfer function $k(z) = I_s + \sum_{j=1}^{\infty} K_j z^j, K_j = C_\circ A_\circ^{j-1} B_\circ$ (which hence has all its zeros and poles outside the unit circle) and an orthonormal matrix $M = [M_c, M_{s-c}] \in \mathbb{R}^{s\times s}, M'M = I_s, M_c \in \mathbb{R}^{s\times c}$, where $0 < c \leq s$ is an integer, as (L denoting the backward-shift operator and $\Delta = (1 - L)$)*

$$(y_t)_{t\in\mathbb{Z}} = M \begin{bmatrix} \Delta I_c & 0 \\ 0 & I_{s-c} \end{bmatrix} M'k(L)(\varepsilon_t)_{t\in\mathbb{Z}}.$$

Here $(\varepsilon_t)_{t\in\mathbb{Z}}$ denotes a zero mean ergodic, stationary, martingale difference sequence with respect to the sequence $\mathcal{F}_t$ of sigma-fields spanned by the past of ε_t fulfilling

$$\mathbb{E}(\varepsilon_t|\mathcal{F}_{t-1}) = 0 \quad , \quad \mathbb{E}(\varepsilon_t\varepsilon_t'|\mathcal{F}_{t-1}) = \mathbb{E}(\varepsilon_t\varepsilon_t') = \Omega.$$

Furthermore $\mathbb{E}\varepsilon_{t,j}^4 < \infty, j = 1, ..., s$.

We use the same noise assumptions as Poskitt (2006) and Hannan and Deistler (1988). Clearly such processes have a spectral density of rank $s - c$ (which is hence singular) for $\omega = 0$ due to the differencing. At all other frequencies the rank equals s since $k(L)$ is assumed to be invertible.

These assumptions would be fulfilled when examining first differences of an I(1) process without knowing the number of cointegrating relations. This is reasonable in systems with many variables where inference on the cointegrating rank is difficult. But even in smaller systems the decision on

the number of cointegrating relations is sometimes not simple as documented in examples in (Johansen, 1995). If in such situations the cointegrating rank is specified too small, spectral zeros result.

To apply this result in our setting note that the multivariate extension to Theorem 2 of Palma and Bondon (2003) implies that $\lambda_{min}(\mathbb{E}Y_t^-(Y_t^-)')$ is bounded from below by $\underline{c}p^{-2}$ for $(y_t)_{t\in\mathbb{Z}}$ according to Assumption 1.

This implies that the bound above amounts to $p^5 \log T/T$ which tends to zero, if $p = c\lfloor T^\delta \rfloor$ for $0 < \delta < 0.2$. Note, however, that for this rate of increase the approximation error (with variance of order p^{-1}) is larger than $O(1/\sqrt{T})$ and hence dominates the asymptotic distribution of terms like $\sqrt{T}(\hat{A} - A_\circ)$.

The results from the autoregressive setting can be used here almost immediately if $f \geq n$ fixed and $p = f\tilde{p}$ where $\tilde{p} = \tilde{p}(T) = o(T^\delta)$ depends on the sample size. This implies that for the approximation of x_t the unrestricted estimate $\hat{\beta}_{f,p} = \langle Y_t^+, Y_t^- \rangle \langle Y_t^-, Y_t^- \rangle^{-1}$ equals an autoregressive model for Y_t^+. Here and below we use the notation $\langle a_t, b_t \rangle = T^{-1} \sum_{t=p+1}^{T} a_t b_t'$ for two processes $(a_t)_{t\in\mathbb{Z}}$ and $(b_t)_{t\in\mathbb{Z}}$. This matrix – which in the limit has rank n – then is low rank approximated leading to the estimate $\hat{\mathcal{O}}_f \hat{\mathcal{K}}_p$ of $\mathcal{O}_f \mathcal{K}_p$.

In order to identify the factors $\hat{\mathcal{O}}_f$ and $\hat{\mathcal{K}}_p$ from the product we use a selector matrix $S_f \in \mathbb{R}^{s \times fs}$ such that $S_f \mathcal{O}_f = I_n$. Such a matrix always exists (cf. for example the overlapping echelon forms, section 2.6 of Hannan and Deistler (1988)). Since the results below correspond to estimates of the impulse response coefficients (which are invariant in this respect) this choice of the state basis can be assumed without restriction of generality.

The second and third step of CVA then amount to least squares using the estimate $\hat{x}_t = \hat{\mathcal{K}}_p Y_t^-$ of the state. If instead we had access to the state approximation $x_t(p) = \mathcal{K}_p Y_t^-$ as well as population instead of sample moments we would obtain the following matrices:

$$\begin{aligned} A_p &= \mathbb{E}x_{t+1}(p)x_t(p)'(\mathbb{E}x_t(p)x_t(p)')^{-1} \\ B_p &= \mathbb{E}x_{t+1}(p)\varepsilon_t(p)'(\mathbb{E}\varepsilon_t(p)\varepsilon_t(p)')^{-1} \\ C_p &= \mathbb{E}y_t x_t(p)'(\mathbb{E}x_t(p)x_t(p)')^{-1}. \end{aligned}$$

If the approximation errors tend to zero and the convergence of sample covariances to population quantities is uniform in p then consistency for $p \to \infty$ follows (for the proof see the appendix):

Theorem 1. *Let the process $(y_t)_{t\in\mathbb{Z}}$ be generated according to Assumptions 1. Let the CVA procedure be applied with $f \geq n$ not depending on T and $p = p(T) \to \infty$ for $T \to \infty$ such that $p(T) = o(T^\delta), 0 < \delta < 0.2$.*

Then:

$$\begin{aligned} \max\{\|\hat{A} - A_p\|, \|\hat{B} - B_p\|, \|\hat{C} - C_p\|\} &= O(\sqrt{p^5/T}), \\ \max\{\|A_\circ - A_p\|, \|B_\circ - B_p\|, \|C_\circ - C_p\|\} &\to 0 \end{aligned}$$

for $p = p(T) \to \infty$ as $T \to \infty$. Consequently $\hat{C}\hat{A}^j\hat{B} \to C_\circ A_\circ^j B_\circ, j = 0, 1, 2, ...$ almost surely in that case.

Note that these two error bounds are differently influenced by the integer p: large p reduces the approximation errors such as $A_p - A_\circ$ but increases the sampling error $\hat{A} - A_p$. It is the first one that tends to zero slower than in the invertible case:

Example 2. Consider again $y_t = \Delta\varepsilon_t$ for white noise $(\varepsilon_t)_{t\in\mathbb{Z}}$. Then $x_t(p) = -\varepsilon_{t-1} + \overline{\varepsilon}_{t-1}(p)$ and $\varepsilon_t(p) = \varepsilon_t - \overline{\varepsilon}_{t-1}(p)$. It follows that $\mathbb{E}x_t(p)\varepsilon_t(p)' = 0$,

$$\begin{aligned} \mathbb{E}\varepsilon_t(p)\varepsilon_t(p)' = \frac{p+2}{p+1}\Omega \quad &, \quad \mathbb{E}x_t(p)x_t(p)' = \frac{p}{p+1}\Omega, \\ \mathbb{E}x_{t+1}(p)\varepsilon_t(p)' = -(1 - \frac{1}{(p+1)^2})\Omega \quad &, \quad \mathbb{E}x_{t+1}(p)x_t(p)' = -\frac{1}{(p+1)^2}\Omega. \end{aligned}$$

Thus $A_p = A_\circ - I_s\frac{1}{p(p+1)}, B_p = B_\circ + \frac{1}{p+1}I_s, C_p = C_\circ$.

This shows for the special case that the system (A_p, B_p, C_p) for fixed p is a biased estimate of the true system $(0, -I_s, I_s)$. The bias is of order p^{-1}. In order for this bias to be asymptotically negligible p has to grow faster than $T^{1/2}$. This is faster than the upper bound $H_T = \sqrt{T/\log T}$ used above, such that with our methods we cannot derive results for the asymptotic distribution of the system estimates.

Additionally note that typically the upper bound for selecting the lag length is $H_T = c\lfloor T^{1/4}\rfloor$ such that the bias derived above will show in the asymptotics.

Similar biases are expected in the general case, as it is the approximation of the inverse of Δ that introduces the issues.

4. Conclusions

In this note we show that working with first differences does not invalidate consistency for CVA. This is a relief in situations where one is not sure about the existence of cointegrating relations.

Inference, on the other hand, gets more complicated as the asymptotic distribution in a situation where some of the variables are over-differenced is not known contrary to the case of no over-differencing.

The results imply that also higher order of differencing as well as spectral zeros at other frequencies introduced for example from yearly differencing can be dealt with using exactly the same methods. In such situations consistency of CVA estimators of the impulse response sequence again follows for p increasing sufficiently slow.

Funding: This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation - Projektnummer 469278259) which is gratefully acknowledged.

Proof of Theorem 1

Note that $\hat{\alpha}(p) = \langle Y_t^+, Y_t^- \rangle \langle Y_t^-, Y_t^- \rangle^{-1}$ is an autoregressive approximation of Y_t^+ by Y_t^- if $p = f\tilde{p}$ for some integer $\tilde{p}$:

$$Y_t^+ = \alpha(p) Y_t^- + U_t^+.$$

Poskitt (2006) Theorem 5 then implies that $\|\hat{\alpha}(p) - \alpha(p)\|_2^2 = O(p^5 Q_T^2)$ using $(\lambda_{min}(\Gamma_p))^{-1} = O(p^2)$ in that case. The proof of this result in Poskitt (2006) can be easily extended to the case of general p in our setting.

Clearly $\alpha(p) = \mathcal{O}_f \mathcal{K}_p$ is of rank n as $\mathcal{E}_f E_t^+$ is orthogonal to Y_t^-. CVA then uses a SVD of $\hat{\Xi_f} \hat{\alpha}(p) \hat{\Xi_p^-}$ or equivalently the SVD of

$$\hat{\Xi_f} \hat{\alpha}(p) \langle Y_t^-, Y_t^- \rangle \hat{\alpha}(p)' (\hat{\Xi_f})'$$

to obtain a rank n approximation where $\hat{\Xi_f} = \langle Y_t^+, Y_t^+ \rangle^{-1/2}$ (the square root denotes the Cholesky decomposition). Due to the uniform convergence of the sample covariances we obtain $\hat{\Xi_f} - \Xi_f = O(Q_T)$ for fixed f since the Cholesky factorization is differentiable for positive definite matrices.

Now $\|\alpha(p)\|_\infty = O(p)$ ($\|.\|_\infty$ denoting the row-sum norm) as can be seen, for example, from the Levinson-Whittle algorithm (see Hannan and Deistler (1988), p. 218). It follows that $\|\alpha(p)\|_1 = O(1)$ (column-sum norm, here

equivalent to maximum entry due to finite f), $\alpha(p)\mathbb{E}Y_t^-(Y_t^-)' = \mathbb{E}Y_t^+(Y_t^-)'$ and $\|\alpha(p)\|_2 = O(p)$. Consequently

$$\hat{\alpha}(p)\langle Y_t^-, Y_t^-\rangle\hat{\alpha}(p)' - \alpha(p)\mathbb{E}Y_t^-(Y_t^-)'\alpha(p)' = O(p^{5/2}Q_T).$$

The properties of the SVD then imply $\|\hat{\mathcal{O}}_f - \mathcal{O}_f\|_2 = O(p^{5/2}Q_T)$ which in turn leads to $\|\hat{\mathcal{K}}_p - \mathcal{K}_p\|_2 = O(p^{5/2}Q_T)$: Key here is the differentiable dependence of the eigenspace to an eigenvalue on the matrix, see Chatelin (1993). This applies here as $\mathcal{O}_f$ spans the orthocomplement of the eigenspace to eigenvalue zero. The convergence for $\hat{\mathcal{O}}_f$ then requires fixing a basis of this space which is achieved by $S_f\mathcal{O}_f = I_n$. We then use the same normalisation for $\hat{\mathcal{O}}_f$ such that $S_f\hat{\mathcal{O}}_f = I_n$ to obtain $\|\hat{\mathcal{O}}_f - \mathcal{O}_f\|_2 = O(p^{5/2}Q_T)$. As $\mathcal{O}_f'\mathcal{O}_f \geq I_s$ we have with $\mathcal{K}_p = (\mathcal{O}_f'\mathcal{O}_f)^{-1}\mathcal{O}_f'\alpha(p)$ and $\hat{\mathcal{K}}_p = (\hat{\mathcal{O}}_f{}'\hat{\mathcal{O}}_f)^{-1}\hat{\mathcal{O}}_f{}'\hat{\alpha}(p)$ that $\|\hat{\mathcal{K}}_p - \mathcal{K}_p\|_2 = O(p^{5/2}Q_T)$.

The remainder of the proof then follows from providing error bounds for terms involving

$$\hat{x}_t(p) - x_t(p) = (\hat{\mathcal{K}}_p - \mathcal{K}_p)Y_t^-.$$

For example,

$$\begin{aligned}
\langle \hat{x}_t, \hat{x}_t\rangle &= \langle \hat{x}_t - x_t(p), \hat{x}_t\rangle + \langle x_t(p), \hat{x}_t - x_t(p)\rangle + \langle x_t(p), x_t(p)\rangle \\
&= (\hat{\mathcal{K}}_p - \mathcal{K}_p)\langle Y_t^-, Y_t^-\rangle\hat{\mathcal{K}}_p{}' + \mathcal{K}_p\langle Y_t^-, Y_t^-\rangle(\hat{\mathcal{K}}_p - \mathcal{K}_p)' + \langle x_t(p), x_t(p)\rangle \\
&= (\hat{\mathcal{K}}_p - \mathcal{K}_p)\mathbb{E}Y_t^-x_t(p)' + \mathbb{E}x_t(p)(Y_t^-)'(\hat{\mathcal{K}}_p - \mathcal{K}_p)' + \langle x_t(p), x_t(p)\rangle + o(p^{5/2}Q_T) \\
&= \langle x_t(p), x_t(p)\rangle + O(p^{5/2}Q_T)
\end{aligned}$$

where the next to last error bound follows from replacing estimates with limits. All evaluations are simple and hence omitted.

These arguments show that uniformly for $0 < p \leq H_T$ the difference between the estimates using $\hat{x}_t$ and using $x_t(p)$ is of order $O(p^{5/2}Q_T)$.

Considering $\langle x_t(p), x_t(p)\rangle - \mathbb{E}x_t(p)x_t(p)'$ we see that

$$\alpha(p)(\langle Y_t^-, Y_t^-\rangle - \Gamma_p)\alpha(p)' = O(p^2Q_T)$$

since $\|\alpha(p)\|_1 = O(p)$. This holds uniformly in $p < H_T$. Similar results show that for p large enough this error rate carries over to the difference in the estimators such that

$$\max\{\|\hat{A} - A_p\|_2, \|\hat{B} - B_p\|_2, \|\hat{C} - C_p\|_2\} = O(p^{5/2} Q_T) = o(1)$$

for $p \leq H_T$.

Next to investigate $A_p - A_\circ$, for example, the difference of the second moments such as $\mathbb{E}x_t(p)x_t(p)' - \mathbb{E}x_t x_t'$ is essential: For these convergence to zero follows since $\mathbb{E}\delta x_t(p)(\delta x_t(p))' \to 0$, as the approximation error converges to zero, compare Lemma 1 of (Poskitt, 2006). This finishes the proof.

References

Bauer, D., 2005. Estimating linear dynamical systems using subspace methods. Econometric Theory 21. pp. 181–211.

Chatelin, F., 1993. Eigenvalues of Matrices. John Wiley & Sons.

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

Johansen, S., 1995. Likelihood-Based Inference in Cointegrated Vector Auto-Regressive Models. Oxford University Press, Oxford, UK.

Larimore, W.E., 1983. System Identification, reduced order filters and modeling via canonical variate analysis, Piscataway, NJ. pp. 445–451.

Palma, W., Bondon, P., 2003. On the Eigenstructure of Generalized Fractional Processes. Statistics and Probability Letters 65, 93–101.

Poskitt, D.S., 2006. Autoregressive approximation in nonstandard situations: the fractionally integrated and non-nivertible case. Annals of Institute of Statistical Mathematics 59, 697–725.