Using Subspace Algorithms for the Estimation of Linear State Space Models in the Context of Approximate Dynamic Factor Models

Dietmar Bauer[1]

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

Abstract

Approximate Dynamic Factor Models (aDFM) are a popular tool for modelling a large number of time series jointly. aDFMs decompose the observations into a common component, containing the most valuable information, and idiosyncratic components that are typically seen as additive noise (at least in the first step of modelling). Identification for aDFMs is achieved for number of variables tending to infinity assuming that information on the common components accumulates while the idiosyncratic component is only weakly correlated across variables.

In this setting the common part is often estimated using principal components in the first step. In the second step then a linear dynamic model for the static common factor process is estimated explaining the evolution of the static factors by an underlying latent dynamic factor process modeled as white noise.

In this paper we show, that the canonical variate analysis (CVA) type of subspace methods can be used in order to obtain consistent estimates of the transfer function relating the dynamic factors to the static factors. Our results cover integrated processes as well as stationary processes obtained, for

[1]Unfinished paper draft.

example, by differencing the integrated process, even if the differentiation leads to spectral zeros. In that case, the convergence rate decreases considerably.

Furthermore, we discuss the differences that arise if there happen to be less dynamic factors than static factors.

Keywords: approximate DFMs, linear state space systems, subspace algorithms, spectral zeros

1. Introduction

For macro-economic and financial data the usage of so called approximate dynamic factor models (aDFMs) has become popular. Thereby, the multivariate time series $(y_{i,t})_{t\in\mathbb{Z}}, y_{i,t} \in \mathbb{R}, i = 1, 2, ..., N, t \in \mathbb{Z}$, is modeled jointly as a vector process. The cross sectional dimension, N, is typically large in such applications, ranging in the hundreds and therefore is similar to the number of time points (to cite just one example, Bai and Ng , 2019, use $N = 128$ for $T = 676$ monthly observations). Such data sets arise, for example, for matrix valued time series, wherein at each time point several variables are observed in different regions. Thus, in this case the observations at a given time point can be arranged into a matrix with rows corresponding to the same variables and columns to variables from within one region or country. Vectorizing the matrix, we obtain a vector valued time series with a typically large dimensionality.

In such situations, providing a joint unrestricted model of the vector autoregressive (VAR) type uses too many parameters (N^2 for each lag) to allow for accurate inference. Often the general modeling idea then is to assume a smaller number of factors, say $r \ll N$, that influence most of the variables. To be more concrete, we in this paper use the following model (here L denotes the backward-shift operator):[2]

$$y_{i,t} = \chi_{i,t} + \xi_{i,t} = \lambda_i' F_t + \xi_{i,t}, \tag{1}$$

$$F_t = b(L)u_t \in \mathbb{R}^r, \quad u_t \in \mathbb{R}^q \tag{2}$$

where $b(z)$ is a rational transfer function. Consequently $(F_t)_{t\in\mathbb{Z}}$ has a minimal state space representation:

$$F_t = Cx_t + Du_t, \quad x_{t+1} = Ax_t + Bu_t \tag{3}$$

where $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times q}, C \in \mathbb{R}^{r\times n}, D \in \mathbb{R}^{r\times q}, q \leq r$. Such a model structure is advocated by Lippi (2024) arising from dynamic stochastic generalized equilibrium models (DSGE). It has been used often in empirical macro-economic studies, see e.g. (Stock and Watson , 2011; Forni et al. , 2000; Barigozzi et al , 2021) and the references therein. Sometimes this model is extended by replacing $\lambda_i' b(L)$ by a rational transfer function $b_i(L)$ (see, for example, Barigozzi et al , 2024, and the references therein). We will not consider this potentially larger model class.

Below we will also use the notation $y_t^N = (y_{i,t})_{i=1,\dots,N} \in \mathbb{R}^N, \xi_t^N \in \mathbb{R}^N$ and $\Lambda_N = (\lambda_1, ..., \lambda_N)' \in \mathbb{R}^{N\times r}$, stressing the dependence on the number of variables N. With this notation we obtain $y_t^N = \Lambda_N F_t + \xi_t^N$. Note, however, that we assume that the static factor process $(F_t)_{t\in\mathbb{Z}}$ does not depend on N.

In order to identify the common and the idiosyncratic component we use the framework of asymptotic (in the dimension N) identification of Cham-

[2]The presentation here follows closely the survey Lippi, Deistler and Anderson (2023).

berlain and Rothschild (1983) (see below for details). As a consequence, the static factors often are approximated using the first r principal components of the vector process y_t^N. Given the estimate of the static common factors then a dynamic model is estimated.

Since F_t is of dimension r and driven by a q dimensional noise u_t, it is a singular process for $r > q$ as is empirically often found to fit data sets well (compare Lippi, Deistler and Anderson , 2023). For $r > q$ the tall transfer function $b(z) = D + zC(I - zA)^{-1}B \in \mathbb{C}^{r\times q}$ does not have a unique left pseudo-inverse. The tall transfer function is called *zero-less*, if the rank of $b(z)$ equals q for all ω. For a square transfer function this is only fulfilled for uni-modular matrices, whereas tall transfer functions generically (in a certain sense) are zero-less, as Anderson and Deistler (2008) show. This has profound consequences for estimation as we will discuss below.

In such a zero-less case, Anderson and Deistler (2008) show that there exists a vector autoregressive (VAR) representation of the static factors F_t such that $A(L)F_t = D_0 u_t, A(z) \in \mathbb{C}^{r\times r}, D_0 \in \mathbb{R}^{r\times q}$. Contrary to the square case, the representation here is not unique even for fixed polynomial degree. Deistler et al. (2012) provide ways to restrict the model set in order to obtain a unique representation involving the specification of a multi-index. This structure theory has led to the suggestion of methods to specify and estimate the corresponding models subject to the assumption of zero-less-ness of the tall transfer function, see for example Barigozzi et al (2024) and the references therein.

In the context of integrated and cointegrated processes, that are ubiquitous for macro-economic or finance datasets, this assumption of zero-less-ness may be doubted: If $b(1)$ for the differenced process $\Delta(L)F_t = b(L)u_t$ has full column rank q, this implies that all q components of u_t have a long-run impact on F_t and hence on y_t^N. This might not be wanted, as persistent shocks have been related to supply shocks, while demand shocks typically are assumed to have only short-term non-persistent impact (see, for instance, Forni et al. , 2023). Therefore, if demand shocks are present, $b(1)$ for the differenced process will not have full column rank. Analogously, seasonal adjustment by including a seasonal differencing filter may lead to spectral zeros in the seasonally differenced series.

If the transfer function is not zero-less, then there does not exist a VAR representation of the singular process, but only state space representations. Thus, it appears to be of interest to develop methods that also work in the case of zeros in the tall transfer function.

The recent paper Forni and Lippi (2023) points out that, while the VAR representation of F_t is not unique, the projections of F_t onto the space spanned by the past of the process is. Consequently, also the innovations are uniquely defined based on these projections. This observation builds the basis for this paper: The main contribution is to point out that subspace procedures like canonical variate analysis (CVA) (see Larimore, 1983), being based upon such projections of the future of a process onto its past, can be used to obtain consistent estimates of the tall transfer function linking dynamic to static factors based upon the estimated first r principal components.

This consistency result is very robust, as it holds both for $q = r$ and $q < r$ cases. It even holds when spectral zeros exist, in which case, however, the convergence rate drops. Moreover, consistency also holds for certain integrated factor processes, while the idiosyncratic terms will be assumed to be stationary throughout.

The outline of this paper is as follows: First we provide the model set up in the next section. In section 3 we discuss the state space representation for tall transfer functions. Subsequently we present the CVA subspace algorithm and point out the differences for singular processes in section 4, where also the main result in the stationary case is presented. The non-stationary case is examined in section 5. Section 6 provides a discussion on how to choose the integer values required by the CVA procedure. Finally section 7 concludes the paper. The proofs of the theorems are collected in the appendix.

2. Model set

In this paper we assume that $y_t^N = \Lambda_N F_t + \xi_t^n$. Note, that the process $(F_t)_{t\in\mathbb{Z}}$ hence does not depend on the cross sectional dimension, which is a restriction that is followed in big parts of the literature.

Identification between the idiosyncratic part and the common factor part will use a growing number of variables N in the sense of Chamberlain and Rothschild (1983): The idiosyncratic components $\xi_{i,t}$ and the common components $\Lambda_N F_t$ are assumed to be independent. The idiosyncratic component contains weakly correlated variables such that $\xi_t^N \in \mathbb{R}^N$ has a covariance matrix subject to a uniform (in N) upper bound for its norm (see below for explicit assumptions). For the common components $\chi_{i,t}$ on the other hand we use the following identifying assumptions:

Assumption 1 (Identification of loadings and factors)**.** *For a selector matrix*[3] $\tilde{I}_{N_0} \in \mathbb{R}^{N_0 \times r}$ *we have* $\tilde{I}'_{N_0}\Lambda_{N_0} = I_r$.

Without restriction of generality we may assume that $\tilde{I}_{N_0} = [I_r, 0]'$ such that the loading matrix in the first r rows equals the identity matrix. Clearly this identifies the static factors $F_t = [\tilde{I}'_{N_0}, 0](y_t^N - \xi_t^N)$. These restrictions imply that the representation does not change for different values of N and Λ_{N_0} is a submatrix of Λ_N for $N_0 < N$.

Note, that this restriction requires knowledge on the impact of the static factors. Imposing the identity matrix appears a strong assumption on first sight. However, it is equivalent to the assumption that $\tilde{I}'_{N_0}\Lambda_{N_0}$ is non-singular.

In that case all eigenvalues of the covariance matrix of $\Lambda_N F_t$ grow essentially linearly as a function of N. Thus the common components are assumed to correspond to the r dominant directions in the variance matrix of $y_t^N = (y_{i,t})_{i=1,\dots,N} \in \mathbb{R}^N$ identifying the column spaces spanned by the columns of Λ_N.

Note, that the identification here is obtained from uniquely factoring the variance $\Lambda_N \mathbb{E} F_t F_t' \Lambda_N'$. There are a number of alternatives (cf. Bai and Ng , 2013):

- $\Lambda_N' \Lambda_N / N = I_r$ such that Λ_N is an orthonormal column (up to scaling by $\sqrt{N}$) and $\mathbb{E} F_{t,N} F_{t,N}'$ is diagonal. This implies that $F_{t,N}$ depends on N via the choice of the basis.

[3] A selector matrix has only entries 0 and 1, where exactly one entry per column equals 1 and the matrix has full column rank.

- $\Lambda_N'\Lambda_N/N$ is diagonal with decreasingly ordered diagonal entries and $\frac{1}{T}\sum_{t=1}^T F_{t,N}F_{t,N}' = I_r$: again $F_{t,N}$ depends on N and additionally its basis is stochastic as the normalisation depends on the sample size T (PC1 of (cf. Bai and Ng , 2013)).

- $\frac{1}{T}\sum_{t=1}^T F_{t,N}F_{t,N}' = I_r$, $[I_r, 0]\Lambda_N$ is positive lower triangular (PC2 of (cf. Bai and Ng , 2013)).

Our restriction is called PC3 in Bai and Ng (2013). The restrictions trade off knowledge on the impacts of static factors with closeness to the PCA: Restrictions like PC1 fit nicely with principal component analysis (PCA) since principal components are obtained from the eigenvalue decomposition of the matrix $\hat{\Sigma}_{N,T} := (NT)^{-1}\sum_{t=1}^T y_t^N(y_t^N)'$ containing the component $\Lambda_N\left(T^{-1}\sum_{t=1}^T F_{t,N}F_{t,N}'\right)\Lambda_N'/N$. By assuming $T^{-1}\sum_{t=1}^T F_{t,N}F_{t,N}' = I_r$ then $\Lambda_N Q$ for orthonormal $Q'Q = I_r, Q \in \mathbb{R}^{r\times r}$ describes the equivalence class. A unique Λ_N may be chosen by (A) restricting the heading matrix to be positive lower triangular or (B) $\Lambda_N'\Lambda_N$ diagonal. Option (A) requires the knowledge that the heading matrix is non-singular. In this situation hence $[I_r, 0]\Lambda_N = I_r$ may as well be assumed.

Option (B) corresponds to PC1 and is attractive, if one does not want to impose knowledge of a set of variables such that the corresponding loading matrix has full column rank.

In this paper we will use the normalization of a submatrix of Λ_N to equal the identity matrix. If no sufficient knowledge exists, then this normalization is infeasible. If instead a different identification scheme is used, the results can be transferred from the results for the infeasible normalization as the two different identification schemes are related by a unique invertible transformation matrix $\hat{G}_T$. Depending on the identification used additionally error bounds or convergence results for $\hat{G}_T$ need to be provided. For the identification schemes described above the literature contains many results in this respect, see in particular the discussion in Bai and Ng (2013).

3. State Space Representations of Tall Transfer Functions

The state space representation of tall transfer functions, that is the case $q < r$ such that $b(z)\mathbb{C}^{r\times q}$, shows some subtle differences compared to the square case of $q = r$, as has been pointed out in a series of articles (such as Anderson and Deistler , 2008) by Manfred Deistler and coworkers. They use a slightly different state space form:

$$F_t = \tilde{C}\tilde{x}_t, \quad \tilde{x}_{t+1} = \tilde{A}\tilde{x}_t + \tilde{B}u_{t+1}. \tag{4}$$

Note that – compared to the innovations form (3) – there is a time shift in the state equation, with the innovations entering at the same time point as the state on the left hand side. Letting $\tilde{x}_t = [x_t', u_t']'$ for a system in innovation form, we see that $\tilde{C} = [C, D], \tilde{B} = [0, I]', \tilde{A} = \begin{bmatrix} A & B \\ 0 & 0 \end{bmatrix}$ is a potentially non-minimal state space representation of order $q + n$ of the form (4). Letting $\tilde{n}$ denote the order of a corresponding minimal representation, we obtain $\tilde{n} \geq r \geq q$.

The tall transfer function corresponding to the factor process thus equals

$$b(L) = D + zC(I_n - zA)^{-1}B = \tilde{C}(I_{\tilde{n}} - z\tilde{A})^{-1}\tilde{B}. \tag{5}$$

It follows that the column rank of $\tilde{C}$ equals r. If $\tilde{C}$ is of full column rank (which is in a certain sense generic), such that $r = \tilde{n}$ and the transfer function has a zero at z_0 (defined as a point where the column rank of $b(z)$ drops), then there exists a vector $x \in \mathbb{R}^q$ such that

$$0 = b(z_0)x = \tilde{C}(I_{\tilde{n}} - z_0\tilde{A})^{-1}\tilde{B}x \Rightarrow (I_{\tilde{n}} - z_0\tilde{A})^{-1}\tilde{B}x = 0 \Rightarrow \tilde{B}x = 0. \qquad (6)$$

In this case $b(z)x = 0$ for all z and q is mis-specified. Thus in the generic case of full column rank of $\tilde{C}$ the transfer function is zero-less.

If $\tilde{C}$ is not of full column rank, then it is possible that the rank of $\tilde{C}(I_{\tilde{n}} - z_0\tilde{A})^{-1}\tilde{B}$ drops just at some values $z_0 \in \mathbb{C}$. As an example consider $F_t = (\Delta u_t, \Delta^2 u_t)'$. Clearly then

$$\begin{aligned} b(z) &= \begin{pmatrix} 1 \\ (1-z) \end{pmatrix}(1-z) = \begin{pmatrix} 1 & -1 & 0 \\ 1 & -2 & 1 \end{pmatrix}(I_3 - z\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix})^{-1}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 1 \\ 1 \end{pmatrix} + z\begin{pmatrix} -1 & 0 \\ -2 & 1 \end{pmatrix}\left(I_2 - z\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}\right)^{-1}\begin{pmatrix} 1 \\ 0 \end{pmatrix}. \end{aligned} \qquad (7)$$

Here $\tilde{n} = 3 > r = 2 > q = 1$ since the state space system on the right is minimal. Anderson and Deistler (2008) define for the case of tall transfer functions in their Definition 1 on p. 285 that the transfer function realised by the tuple (A, B, C, D) has a zero at the finite complex value z_0 if the matrix

$$\begin{pmatrix} z_0 I - A & -B \\ C & D \end{pmatrix} \qquad (8)$$

falls below its normal rank. Then they show that every zero of $b(z)$ in this definition leads to an increase of the gap between $\tilde{n}$ and r.

In order to understand the dynamical properties of the process $(F_t)_{t\in\mathbb{Z}}$, the innovation representation $F_t = Cx_t + Du_t, x_{t+1} = Ax_t + Bu_t$ is more revealing. To this end consider for given integer f the stacked processes $F_t^+ = (F_t', F_{t+1}', ..., F_{t+f-1}')' \in \mathbb{R}^{fr}, U_t^+ = (u_t', u_{t+1}', ..., u_{t+f-1}')' \in \mathbb{R}^{fq}$:

$$F_t^+ = \underbrace{\begin{pmatrix} C \\ CA \\ \vdots \\ CA^{f-1} \end{pmatrix}}_{\mathcal{O}_f} x_t + \underbrace{\begin{pmatrix} D & 0 & \dots & 0 \\ CB & D & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ CA^{f-1}B & \dots & CB & D \end{pmatrix}}_{\mathcal{U}_f} U_t^+. \tag{9}$$

Note that $\mathcal{O}_f \in \mathbb{R}^{fr\times n}, \mathcal{U}_f \in \mathbb{R}^{fr\times fq}$ and thus[4] $L_f = [\mathcal{O}_f, \mathcal{U}_f] \in \mathbb{R}^{fr\times(fq+n)}$. For $r > q$ it follows that L_f is tall for large enough f, admitting a left kernel of dimension at least $f(r-q) - n > 0$.

Consider left multiplication with a matrix $A_- = [A_{f-1}, ..., A_1, I_r] \in \mathbb{R}^{r\times fs}$. We obtain

$$F_{t+f-1} + \sum_{j=1}^{f-1} A_j F_{t+f-1-j} = A_- F_t^+ = A_- \mathcal{O}_f x_t + A_- \mathcal{U}_f U_t^+. \tag{10}$$

It follows that each such matrix fulfilling $A_- \mathcal{O}_f = 0, A_- \mathcal{U}_f = [0, ..., 0, D]$ results in a singular VAR representation for $(F_t)_{t\in\mathbb{Z}}$. Clearly this representation is not unique due to the existence of the left kernel for L_f.

Next consider the role of the state:

[4]The notation L_f is used in Anderson and Deistler (2008).

$$x_t = Ax_{t-1} + Bu_{t-1} = A^p x_{t-p} + \sum_{j=1}^{p} A^{j-1} B u_{t-j}. \tag{11}$$

We can assume without restriction of generality that D is of full column rank. If this would not hold, then we can shift time in one direction leading to an adjustment of the transfer function. Then it follows that

$$u_t = D^\dagger (F_t - Cx_t) \tag{12}$$

as in the square case, where $D^\dagger$ denotes the Moore-Penrose pseudo-inverse of D. This defines a square system:

$$D^\dagger F_t = D^\dagger C x_t + u_t. \tag{13}$$

Now assume that there exists a pseudo-inverse $D^\dagger$ such that $D^\dagger b(z)$ can be stably inverted, that is such that the matrix $\underline{A} = A - BD^\dagger C$ has only eigenvalues of modulus smaller than one. In that case we obtain

$$x_t = \underline{A} x_{t-1} + BD^\dagger F_{t-1} = \underline{A}^p x_{t-p} + \sum_{j=1}^{p} \underline{A}^{j-1} BD^\dagger F_{t-j}. \tag{14}$$

Thus x_t can be approximated using the finite past of F_t since $\underline{A}^p \to 0, p \to \infty$. Even without this assumption, we may assume without restriction of generality that x_t can be approximated using the finite past of F_t as it is a linearly regular (in the sense of Theorem 1.3.1 of Hannan and Deistler, 1988) stationary process and the past spaces of F_t and u_t coincide. Thus, if $x_t(p) = \mathcal{K}_p F_t^-(p)$ denotes the best (in mean square sense) approximation of x_t using $F_t^- = [F'_{t-1}, ..., F'_{t-p}]'$ then we have $\mathbb{E}\|x_t(p) - x_t\|^2 \to 0$ for $p \to \infty$.

For non-invertible cases (due to spectral zeros) the convergence may be slower than exponential, see the example in Bauer (2025).

Taking the two equations (14) and (9) together we obtain

$$F_t^+ = \mathcal{O}_f x_t(p) + \mathcal{U}_f U_t^+ + \mathcal{O}_f(x_t - x_t(p)) = \mathcal{O}_f \mathcal{K}_p F_t^-(p) + N_t^+ \tag{15}$$

where N_t^+ (neglecting its dependence on f, p in the notation) is orthogonal to $F_{t-j}, 0 < j \leq p$. This is the main ingredient for the CVA approach to subspace methods due to Larimore (1983).

Now consider the zero-less case more closely. In this case there is an autoregressive representation of F_t, which implies that $x_t(p) = x_t = \mathcal{K}_p F_t^-(p)$ for p large enough. Then Anderson and Deistler (2008) show that the state x_t can be recovered from F_t^+, the future of F_t. Equation (9) shows that this may happen, if $L_f = [\mathcal{O}_f, \mathcal{U}_f]$ has full column rank: Denoting with $L_f^\dagger$ the Moore-Penrose pseudo-inverse we get

$$L_f^\dagger F_t^+ = \begin{pmatrix} x_t \\ U_t^+ \end{pmatrix}, x_t = \mathcal{K}_p F_t^-(p). \tag{16}$$

In that situation the state is a linear combination of $F_t^-(p)$ for p large enough. But also a linear combination of F_t^+ (using the matrix of the first n rows of $L_f^\dagger$). It follows that in this situation there are n unit canonical correlations between F_t^+ and $F_t^-(p)$, compare Breitung and Pigorsch (2012).

For $r = q$ as well as for zeros in $b(z)$ the canonical correlations are strictly smaller than 1 as in these cases $x_t(p) \neq x_t$ and x_t is not contained in the space spanned by the components of F_t^+.

4. Canonical variate analysis

CVA is a numerically cheap algorithm to estimate a linear dynamic time invariant model (3) to vector valued time series data. It consists of a series of regressions. It is asymptotically equivalent to quasi maximum likelihood estimation (using the Gaussian likelihood) for non-singular, invertible stationary processes and robust to the existence of simple unit roots (see Bauer, 2005a, for a survey).

The CVA method adapted to the estimation for singular processes on the basis of the data $F_t \in \mathbb{R}^r, t = 1, .., T$ is proposed to be performed in four steps, using two integers f, p ('future' and 'past') and information of the system order n as well as the number of dynamic factors q (compare Bauer, 2005a):

1. Obtain an estimate $\hat{x}_t$ of the state x_t for $t = p + 1, ..., T + 1$.
2. Estimate C by regressing F_t onto $\hat{x}_t$. This step provides residuals $\hat{\varepsilon}_t = F_t - \hat{C}\hat{x}_t, t = p + 1, ..., T$.
3. Perform a lower triangular Cholesky decomposition of the Frobenius norm optimal rank q approximation to
$$\hat{\Omega} = (T - p)^{-1} \sum_{t=p+1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t'$$
as $\hat{\Omega} = \hat{D}_q \hat{D}_q' + \hat{R}_q$ where $\hat{D}_q \in \mathbb{R}^{r \times q}$ denotes the positive lower triangular matrix square root. Obtain $\hat{u}_t = \hat{D}_q^{\dagger} \hat{\varepsilon}_t$.
4. Estimate A and B by regressing $\hat{x}_{t+1}$ onto $\hat{x}_t$ and $\hat{u}_t, t = p + 1, ..., T$.
5. Convert the estimated system to an appropriate echelon overlapping form resulting in the estimates $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$.

The essential idea of CVA lies in the estimation of x_t which uses (15): The components of F_t^+ projected onto $F_t^-(p)$ are all linear functions of the projected state at time t.

This implies that regressing F_t^+ onto $F_t^-(p)$ results in an estimate of $\mathcal{O}_f x_t(p)$. Taking the first n principal components of this vector then provides an estimate of $\hat{\mathcal{T}} x_t$ (the state in a possibly randomly transformed basis). In more detail we can estimate (9) using OLS to obtain the typically full rank matrix $\hat{\beta}_{f,p} := \hat{\mathcal{H}_{f,p}} \hat{\Gamma}_p^{-1}$ where $\hat{\mathcal{H}_{f,p}} = T^{-1} \sum_{t=p+1}^{T-f+1} F_t^+ (F_t^-(p))', \hat{\Gamma}_p = T^{-1} \sum_{t=p+1}^{T-f+1} F_t^-(p)(F_t^-(p))'$ with predictions $\hat{\beta}_{f,p} F_t^-(p)$. The sample variance of the predictions hence equal

$$\hat{\beta}_{f,p} \hat{\Gamma}_p \hat{\beta}_{f,p}' = \hat{\mathcal{H}_{f,p}} \hat{\Gamma}_p^{-1} \hat{\mathcal{H}_{f,p}}' = \hat{U} \hat{S} \hat{U}' = \hat{U}_n \hat{S}_n \hat{U}_n' + \hat{R}_n. \tag{17}$$

Here $\hat{U} \hat{S} \hat{U}'$ denotes the singular value decomposition (SVD) for calculating the principal components. $\hat{U}_n \in \mathbb{R}^{fr \times n}$ denotes the submatrix of the first n columns providing the loadings of the first n principal components.

Note that this is different from the usual CVA approach, which uses a SVD of

$$W_f^+ \hat{\beta}_{f,p} W_p^- (W_p^-)' \hat{\beta}_{f,p}' (W_f^+)' \tag{18}$$

to estimate the state sequence. W_p^- is chosen typically as $\hat{\Gamma}_p^{1/2}$ as above, while $W_f^+ = (\hat{\Gamma}_f^+)^{-1/2}$ (where $\hat{\Gamma}_f^+ = T^{-1} \sum_{t=p+1}^{T-f+1} F_t^+ (F_t^+)'$) leads to canonical variates. We advocate for a different weighting matrix $W_f^+ = I_{fr}$ below for the case $q < r$ as in this situation $\hat{\Gamma}_f^+$ will be asymptotically singular.

Compared to the usual case of a non-singular invertible process two changes occur for the tall transfer function case corresponding to singular processes considered here:

- Singularity of the process F_t implies that the calculation of the projection needs to be done more carefully numerically in order to take the singularity of the covariance of $F_t^-(p)$ into account.

- The innovation estimates $\hat{\varepsilon}_t$ will tend to a process with singular variance. Hence in the equation $x_{t+1} = Ax_t + Bu_t$ the noise u_t is not simply the error of the observation equation.

Therefore, we use regularization in the procedure above by replacing $\hat{\Gamma}_p$ by $\tilde{\Gamma}_p$ wherein all eigenvalues smaller than $\epsilon := 10^{-6}$ are changed to ϵ:

$$\hat{\Gamma}_p = \hat{V}\hat{\phi}\hat{V}', \hat{\phi} = \mathrm{diag}(\hat{\phi}_1, ..., \hat{\phi}_{rp}), \tag{19}$$

$$\tilde{\Gamma}_p(\epsilon) = \hat{V}\tilde{\phi}\hat{V}', \tilde{\phi} = \mathrm{diag}(\max(\hat{\phi}_1, \epsilon)..., \max(\hat{\phi}_{rp}, \epsilon)). \tag{20}$$

Of course the choice of ϵ is heuristic. Below we will see that it is immaterial for the application in aDFMs under the assumption that it is chosen small enough.

Having stated the algorithm, the next step is to derive its asmyptotic properties. In order to do so, in this paper we use the following assumption on the data generating process in the stationary case:

Assumption 2. *The stationary process $(F_t)_{t\in\mathbb{Z}}, F_t \in \mathbb{R}^r$, is generated as $F_t = b_\circ(L)u_t$ (L denoting the backward-shift operator) where*

$$b_\circ(z) = D_\circ + \sum_{j=1}^{\infty} C_\circ A_\circ^{j-1} B_\circ z^j \in \mathbb{C}^{r\times q} \tag{21}$$

for $q \leq r$ where $\lambda_{|max|}(A_\circ) < 1$ (which hence has all its poles outside the closed unit circle) and where $D_\circ \in \mathbb{R}^{r \times q}$ has full column rank and is positive lower triangular such that its heading submatrix is non-singular with positive entries on the main diagonal.

The minimal system $(A_\circ, B_\circ, C_\circ, D_\circ)$ is assumed to be in an interior point of an appropriate echelon overlapping form.

Additionally, we assume that there is a pseudo-inverse $D_\circ^\dagger$ such that $D_\circ^\dagger D_\circ = I_q$ where $\underline{A}_\circ = A_\circ - B_\circ D_\circ^\dagger C_\circ$ is stable such that $c_\circ(z) = D_\circ^\dagger(I_r - zC_\circ(I_n - z\underline{A}_\circ)^{-1} B_\circ D_\circ^\dagger)$ is a stable left pseudo-inverse for $b_\circ(z)$ such that $\|\underline{A}_\circ^p\| \leq M\rho_\circ^p$ for some $M < \infty$ and $0 \leq \rho_\circ < 1$ (where $\rho_\circ = 0$ is defined for nilpotent matrices $\underline{A}_\circ$).

Here $(u_t)_{t \in \mathbb{Z}}, u_t \in \mathbb{R}^q$, 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 u_t fulfilling

$$\mathbb{E}(u_t|\mathcal{F}_{t-1}) = 0 \quad , \quad \mathbb{E}(u_t u_t'|\mathcal{F}_{t-1}) = \mathbb{E}(u_t u_t') = I_r. \tag{22}$$

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

We use the same noise assumptions as Hannan and Deistler (1988). Clearly such processes have a spectral density of rank q (which is hence singular for $q < r$).

Note that we assume that the heading $q \times q$ submatrix of $D_\circ$ is non-singular such that the innovations of the first q static factors are influenced by all q dynamic factors. Jointly with assumptions 1 this implies that the dynamic factors can be given the interpretation of shocks to certain variables.

Again, this assumption can be seen as a technical assumption potentially leading to the definition of an infeasible estimator, for which asymptotics can be derived. If one does not use these assumptions then an additional step requires the handling of the corresponding transformation matrix $\hat{G}_T$ relating the infeasible and the feasible normalization.

Finally, note that we use an overlapping echelon form to represent both the true and the estimated system. There always exists an overlapping form such that the true system is an interior point and in particular $\hat{b}(z) \to b_\circ(z)$ is equivalent to the state space matrices converging, see Chapter 2 of (Hannan and Deistler, 1988).

Theorem 1. *Let the process $(F_t)_{t\in\mathbb{Z}}$ be generated according to Assumptions 2 where $q < r$. Let the CVA procedure be applied to the process $(F_t)_{t\in\mathbb{Z}}$ with $f \geq n_O$, the observability index,[5] not depending on T and $p = p(T) \to \infty$ for $T \to \infty$ such that $p(T) \geq -(1+\delta)\log T/(2\log\rho_\circ)$ for $\delta > 0$ for $0 < \rho_\circ$ and $p \geq p_\circ$ else. Additionally $p(T) = O((\log T)^a), a > 0$. Here the weights are chosen as $W_f^+ = I_{fr}, W_p^- = \tilde{\Gamma}_p(\epsilon)^{1/2}$, where ϵ is chosen smaller than the smallest nonzero eigenvalue of $\mathbb{E}F_t(p)^-(F_t(p)^-)'$.*

Denote the corresponding CVA estimate as $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$. Then:

$$\max\{\|\hat{A} - A_\circ\|, \|\hat{B} - B_\circ\|, \|\hat{C} - C_\circ\|, \|\hat{D} - D_\circ\|\} = O(Q_T). \qquad (23)$$

If $q = r$ then no regularization of $\hat{\Gamma}_p$ is needed and the choice $W_f^+ = (\hat{\Gamma}_f^+)^{-1/2}$ leads to the same consistency with order $O(Q_T)$.

The proof is given in Appendix B. The theorem extends results for the

[5]The observability index is the smallest integer f such that $\mathcal{O}_f$ has full column rank.

stationary non-singular case to include singular processes.[6]

Theorem 1 builds the basis for the main result of this paper, but applies to the unrealistic situation that the static factor process F_t is directly observed. In aDFMs the common factor process is observed within a larger set of variables $N \gg r$ subject to additional idiosyncratic components ξ_t^N. In a first step an estimate $\hat{F}_t$ of the static common factors is achieved using the first r leading principal components of the process $(y_t^N)_{t\in\mathbb{Z}}$ as $\hat{\Lambda}_N^\dagger y_t^N$ for some matrix $\hat{\Lambda}_N \in \mathbb{R}^{N\times r}$ depending on the identification assumption.

This has two consequences: first, the input to CVA is changed from $\hat{\gamma}_j^F := \langle F_t, F_{t-j}\rangle$ to $\hat{\gamma}_j^{\hat{F}} := \langle \hat{F}_t, \hat{F}_{t-j}\rangle$ using the notation

$$\langle a_t, b_{t-j}\rangle := T^{-1}\sum_{t=p+1}^{T-f+1} a_t b_{t-j}' \tag{24}$$

for vector valued sequences $a_t, b_t, t = 1, ..., T$. In the aDFM literature often assumptions are used such that the difference $\hat{\gamma}_j^F - \hat{\gamma}_j^{\hat{F}} = O_P(N^{-1}) + O_P(T^{-1/2})$. The second term here is dominated by the LIL rate Q_T and the same is true for the first term if $T/N^2 \to 0$. Under this assumption it has been shown in (Stock and Watson , 2011; Doz et al. , 2011) for the $q = r$ case that VAR(p) model estimators based on $\hat{F}_t$ have the same asymptotic behavior as the ones using F_t.

Second, while the variance of $\Lambda_N F_t$ is singular, the variance of y_t^N typically is non-singular. This has an impact on the regularization: Doz et al. (2011)

[6] An earlier version of this working paper als included cases with spectral zeros. These also lead to consistency of the estimated system, but the convergence rate changes. See Bauer (2025).

impose the assumption that the smallest eigenvalue of $\mathbb{E}\xi_t^N(\xi_t^N)'$ is bounded away from zero uniformly. This carries over to y_t^N and hence also to its principal components. Hence with proper normalization, no regularization is necessary for principal components and typical values for N.

Here we use the following setting: Let $\hat{\Sigma}_{N,T} = \langle y_t^N, y_t^N \rangle / N$ denote the scaled sample variance of y_t^N (assuming zero mean). Then the principal components are obtained from an SVD of this matrix:

$$\hat{\Sigma}_{N,T} = \hat{U}_r \hat{S}_r \hat{U}_r' + \hat{R}_r \in \mathbb{R}^{N\times N}, \hat{U}_r \in \mathbb{R}^{N\times r}, \hat{U}_r'\hat{U}_r = I_r. \tag{25}$$

Above we used the identification restrictions $\tilde{I}_N'\Lambda_N = I_r$ where $\Lambda_N'\Lambda_N/N \to M_\circ > 0$. This implies $\hat{\Lambda}_N := \sqrt{N}\hat{U}_r(\tilde{I}_N'\sqrt{N}\hat{U}_r)^{-1}$. Then we use $\hat{\Lambda}_N^\dagger = (\hat{\Lambda}_N'\hat{\Lambda}_N)^{-1}\hat{\Lambda}_N'$ such that

$$\hat{F}_t = \hat{\Lambda}_N^\dagger y_t^N = \hat{\Lambda}_N^\dagger \Lambda_N F_t + \hat{\Lambda}_N^\dagger \xi_t^N. \tag{26}$$

Doz et al. (2011), for example, provide sufficient conditions such that $\hat{\Lambda}_N^\dagger \Lambda_N \to I_r$ and

$$\hat{\Lambda}_N^\dagger \mathbb{E}\xi_t^N(\xi_t^N)'(\hat{\Lambda}_N^\dagger)' \leq \hat{\Lambda}_N^\dagger(\hat{\Lambda}_N^\dagger)'\Psi = (\hat{\Lambda}_N'\hat{\Lambda}_N)^{-1}\Psi \to 0. \tag{27}$$

Here the inequality follows from $\lambda_{max}(\mathbb{E}\xi_t^N(\xi_t^N)') \leq \Psi$, that is, weak dependence. In this paper the following high level assumptions are used:

Assumption 3. *(I) The processes $(F_t)_{t\in\mathbb{Z}}, (\xi_t^N)_{t\in\mathbb{Z}}$ are jointly wide sense stationary with zero expected value for all N and possess spectral densities. The factor process $(F_t)_{t\in\mathbb{Z}}$ and the idiosyncratic process $(\xi_t^N)_{t\in\mathbb{Z}}$ are assumed* *to be independent.*

For each of the processes F_t, ξ_t^N, y_t^N we have uniformly in $N \in \mathbb{N}$

$$\max_{0 \le k \le H_T} \max_{i,j} \|\langle x_{t,i}, z_{t-k,j}\rangle - \mathbb{E}x_{t,i}z_{t-k,j}\| = O(Q_T) \tag{28}$$

where $Q_T := \sqrt{(\log\log T/T)}$ and $H_T = (\log T)^a$ for some integer $a > 1$ and x_t and z_t here stand for any of the processes y_t^N, F_t, ξ_t^N.

(II) The idiosyncratic process is weakly dependent such that $\lambda_{max}(\mathbb{E}\xi_t^N(\xi_t^N)') \le \Psi < \infty$ uniformly in N.

These assumptions hold for a finite dimensional stationary process with rational spectral density for martingale difference sequence assumptions on the noise, see, for example, Hannan and Deistler (1988) Theorem 5.3.2. The upper bound on the lags can be traded against a slightly larger order of convergence using Theorem 7.4.3. of Hannan and Deistler (1988). Additionally here we assume uniformity in N. This is related to the assumption that the idiosyncratic components do not have strong links. With these assumptions we can show that the error introduced by replacing static factors by its estimates are of the same order as the difference to the expectations (for the proof see the Appendix):

Theorem 2. *Let the process be generated according to Assumptions 2 and 3, where $T/N^2 \to 0$. Then*

$$\sup_{0 \le k \le H_T} \|\langle \hat{F}_t, \hat{F}_{t-k}\rangle - \langle F_t, F_{t-k}\rangle\| = O(Q_T). \tag{29}$$

We note that under the assumptions of Doz et al. (2011) the error term is $O_P(T^{-1/2})$ and hence similar. Also Barigozzi et al (2024) contains very similar results derived from sufficient conditions for the process, see (28) on p. 9 or footnote 10 on p. 12. We prefer the a.s. bound as they make some of the calculations easier.

This leads to the following result (which is proved in Appendix B):

Theorem 3. *Under the assumptions of Theorem 2 the results of Theorem 1 remain true, if $\hat{F}_t$ is used in CVA instead of F_t.*

Remark 1. *Note, that here the weight $W_f^+ = I_{fr}$ is suggested which differs from the usual CVA choice $W_f^+ = \langle \hat{F}_t^+, \hat{F}_t^+ \rangle^{-1/2}$. This is necessary due to the singularity of the process $(F_t)_{t \in \mathbb{Z}}$ which implies that $\langle F_t^+, F_t^+ \rangle$ and $\langle F_t^-(p), F_t^-(p) \rangle$ both are singular. For $\hat{F}_t^+$ and $\hat{F}_t^-(p)$ we obtain the same limits for $T, N \to \infty$. Fixing N and the integers f, p and letting $T \to \infty$ one notices that because of the idiosyncratic terms the limit of $\langle \hat{F}_t^+, \hat{F}_t^+ \rangle$ in general is non-singular, even if the idiosyncratic variables only contribute a small variability due to the factor $1/N$ involved. However, since CVA with the weight $W_f^+ = \langle \hat{F}_t^+, \hat{F}_t^+ \rangle^{-1/2}$ calculates the canonical correlations, the small variance is compensated such that the idiosyncratic terms show up in the estimation.*

Another way to put this is that the correlations between the variables $z_{t,-}$ and $z_{t,+}$ are identical to the correlations for the variables $z_{t,-}/N$ and $z_{t,+}/N$ for every value of N and the same holds for the correlation matrix between the vectors $[y_{t-1}', z_{t,-}/N]'$ and $[y_t', z_{t,+}/N]'$. Clearly the corresponding variance matrices tend to singular limits for $N \to \infty$, while the correlations do not depend on N. For our purposes this is critical as it implies that the canonical correlations between some aspects of the idiosyncratic components would influence the CVA estimates when using the weighting matrix W_f^+

leading to canonical correlations.

The choice $W_f^+ = I_{fs}$ avoids this. For the regressors $\hat{F}_t^-(p)$ the weighting magnifies the idiosyncratic variables in the kernel of $\langle F_t^-(p), F_t^-(p)\rangle$, but the corresponding covariance in $\langle \hat{F}_t^+, \hat{F}_t^-(p)\rangle$ is small eliminating these terms in the limit of (18). Thus the choice $W_f^+ = I_{fr}$ is robust to whether $q = r$ or not, while the CVA choice $W_f^+ = (\hat{\Gamma}_f^+)^{-1/2}$ is not and only provides consistent results for $q = r$.

5. Integrated static factors

Many economic series show a strong persistence that often is modeled using integrated processes. Analyzing such series in first differences, as is often done, carries the risk of introducing spectral zeros. Bauer (2025) shows that this involves a penalty as methods such as CVA relying on autoregressive approximations have difficulties with spectral zeros.

As an alternative one may work with the original series. In order to extract the static factors two different approaches are obvious: One may differentiate y_t^N and calculate the principal components for the differenced series. Once the matrix $\hat{\Lambda}_N^\dagger$ is estimated, it may be applied to the original series. Such an approach is used for example in Bai and Ng (2004).

While such an approach provides useful results, it may lack efficiency as differencing reduces variation in the series and hence distorts the principal components. Below, we only deal with estimates of the principal components of the original series y_t^N.

In the integrated case we use different assumptions on the data generating process:

Assumption 4 (dgp, I(1) case)**.** *The process* $(F_t)_{t\in\mathbb{Z}}, F_t \in \mathbb{R}^r$, *has a state space representation for some* $0 < c \le q$*:*

$$F_t = \underbrace{\begin{pmatrix} C_1 & C_\bullet \end{pmatrix}}_{C_\circ} x_t + D_\circ u_t, x_{t+1} = \underbrace{\begin{pmatrix} I_c & 0 \\ 0 & A_\bullet \end{pmatrix}}_{A_\circ} x_t + \underbrace{\begin{pmatrix} B_1 \\ B_\bullet \end{pmatrix}}_{B_\circ} u_t \tag{30}$$

where $\lambda_{|max|}(A_\bullet) < 1$, $C_1'C_1 = I_c, B_1$ *is p.u.t. and the subsystem* $(A_\bullet, B_\bullet, C_\bullet)$ *is in echelon overlapping form. Further* $[I_q, 0]D_{red}$ *is non-singular and p.l.t.*

Furthermore we assume that there is a pseudo-inverse $D_\circ^\dagger$ *such that* $D_\circ^\dagger D_\circ = I_q$ *where* $\underline{A}_\circ = A_\circ - B_\circ D_\circ^\dagger C_\circ$ *is stable such that* $c_\circ(z) = D_\circ^\dagger(I_r - zC_\circ(I_n - z\underline{A}_\circ)^{-1}B_\circ D_\circ^\dagger))$ *is a stable left pseudo-inverse for* $b_\circ(z) = D_\circ + zC_\circ(I_n - zA_\circ)^{-1}B_\circ$ *such that* $\|\underline{A}_\circ^p\| \le M\rho_\circ^p$ *for some* $M < \infty$ *and* $0 \le \rho_\circ < 1$ *(where* $\rho_\circ = 0$ *is defined for nilpotent matrices* $\underline{A}_\circ$*).*

Here $(u_t)_{t\in\mathbb{Z}}, u_t \in \mathbb{R}^q$, *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* u_t *fulfilling*

$$\mathbb{E}(u_t|\mathcal{F}_{t-1}) = 0 \quad , \quad \mathbb{E}(u_t u_t'|\mathcal{F}_{t-1}) = \mathbb{E}(u_t u_t') = I_q. \tag{31}$$

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

Finally Assumption 2 holds for $(W_t)_{t\in\mathbb{Z}}$ *replacing* $(F_t)_{t\in\mathbb{Z}}$ *where*

$$P_C = \begin{pmatrix} C_1 & C_{1,\perp} \end{pmatrix}, P_C'P_C = I_r, W_t = \begin{pmatrix} \Delta & 0 \\ 0 & I \end{pmatrix} P_C'F_t \tag{32}$$

and

$$\sup_{N\in\mathbb{N}} \max_{i=1,\dots,N} \|\langle [I_c, 0]F_t, \xi_{i,t}\rangle\| = O(\log T). \tag{33}$$

The process $(F_t)_{t\in\mathbb{Z}}$ hence is assumed to be generated by a state space system in the canonical form for $I(1)$ systems proposed in Bauer and Wagner (2002). Consequently the common trends in $(F_t)_{t\in\mathbb{Z}}$ equal $C_1 B_1 \sum_{j=1}^{t-1} u_j$. These trends dominate the variance matrix $\hat{\Sigma}_{N,T}$ for large T. Hence it is useful to consider the transformed process $P_C F_t = (F'_{t,c}, F_{t,\bullet})'$ which contains the common trends $B_1 \sum_{j=1}^{t-1} u_j$ plus a stationary process as the first c components.

With this assumption we can show the approximation quality using the principal components:

Theorem 4. *Under Assumptions 4 with the matrix* $\hat{\Lambda}_N^\dagger = \hat{S}_r^{-1/2}\hat{U}_r'/\sqrt{N}$ *obtained in the PCA step using the SVD of (25) let* $\hat{F}_t = \hat{\Lambda}_N^\dagger y_t^N$ *normalized such that* $\langle \hat{F}_t, \hat{F}_t\rangle = I_r$. *Consider the normalization*[7]

$$\tilde{F}_t = \begin{pmatrix} \langle F_{t,c}, F_{t,c}\rangle^{-1/2} F_{t,c} \\ F_{t,\bullet} \end{pmatrix}. \tag{34}$$

Then there exists a sequence of random matrices $H_{T,N}$ *such that for each* $0.5 < \gamma < 1$ *for* $T/N^2 \to 0, T \to \infty$ *we have*

$$\|\langle \tilde{F}_t, \tilde{F}_{t-k}\rangle - H_{T,N}\langle \hat{F}_t, \hat{F}_{t-k}\rangle H'_{T,N}\| = O(T^{1/2-\gamma}). \tag{35}$$

This result is slightly different to the result for the stationary case. It implies that taking the principal components leads to a self-normalization

[7] The upper triangular Cholesky factor is used as the matrix square root.

due to the attempt to achieve unit variance. The self-normalization also implies that the off-diagonal elements (cross products between stationary components of the static factors and normalized integrated components) tend to zero with rate slightly slower than $T^{-1/2}$. Contrary to the stationary case here we did not find a way to avoid the random matrices $H_{T,N}$, as the normalization depends on the inner product of the integrated components. The theorem is proved in Appendix B.

Having established that the empirical second moments of the estimated principal components correspond to the ones of the appropriately normalized latent static factors, the last step is to investigate the properties of the CVA estimates for the integrated case. For this we use the notation $W_t = \text{diag}(\Delta I_c, I_{r-c})P_C' F_t$ such that $(W_t)_{t\in\mathbb{Z}} = b_{red}(L)(u_t)_{t\in\mathbb{Z}}$ is a singular stationary process where $b_{red}(z) = D_{red} + zC_{red}(I - A_{red}z)^{-1}B_{red}$ denotes a minimal representation of the transfer function with stable left pseudo inverse according to Assumption 4. We also use the partitioning $W_t = [W_{t,c}', W_{t,\bullet}]', W_{t,c} \in \mathbb{R}^c$.

Then the central equation (15) allows to identify the state as the projection of the future space onto the past space. To examine this equation, the vectors F_t^+ and $F_t^-(p)$ are transformed with a non-singular matrix. For $F_t^-(p)$, note, that choosing the basis for the space onto which we project, does not change the projection. For F_t^+ we apply a transformation matrix $\mathcal{T}_W$ in order to separate the stationary from the non-stationary components.

To this end we consider

$$
\underbrace{\begin{pmatrix}
I_c & 0 & \dots & \dots & \dots & 0 \\
0 & I_{r-c} & \ddots & & & \vdots \\
-I_c & 0 & I_c & \ddots & & \vdots \\
0 & 0 & 0 & I_{r-c} & \ddots & \vdots \\
 & & & \ddots & \ddots & 0 \\
 & & & & 0 & I_{r-c}
\end{pmatrix}}_{\mathcal{T}_W^+} F_t^+ = \begin{pmatrix} P_C' F_t \\ W_{t+1} \\ \vdots \\ W_{t+f-1} \end{pmatrix} = W_t^+ + \begin{pmatrix} I_c \\ 0 \\ \vdots \\ 0 \end{pmatrix} F_{t-1,c},
$$

$$
\underbrace{\begin{pmatrix}
I_c & 0 & \dots & \dots & \dots & 0 \\
0 & I_{r-c} & \ddots & & & \vdots \\
I_c & 0 & -I_c & \ddots & & \vdots \\
0 & 0 & 0 & I_{r-c} & \ddots & \vdots \\
 & & & \ddots & \ddots & 0 \\
 & & & & 0 & I_{r-c}
\end{pmatrix}}_{\mathcal{T}_W^-} F_t^-(p) = \begin{pmatrix} F_{t-1,c} \\ W_{t-1,\bullet} \\ W_{t-1,c} \\ W_{t-2,\bullet} \\ \vdots \\ W_{t-p+1,c} \\ W_{t-p,\bullet} \end{pmatrix}. \tag{36}
$$

With these transformations (15) reads

$$
\mathcal{T}_W^+ F_t^+ - \begin{pmatrix} I_c \\ 0 \\ \vdots \\ 0 \end{pmatrix} F_{t-1,c} = W_t^+ = \mathcal{O}_f \mathcal{K}_p W_t^-(p) + N_t^+ \tag{37}
$$

where $\mathcal{O}_f \mathcal{K}_p$ corresponds to $(A_{red}, B_{red}, C_{red}; D_{red})$ realizing $b_{red}(z)$. Note that $\mathcal{T}_W^- F_t^-(p)$ contains $W_t^-(p)$ except for $W_{t-p,c}$ but including $F_{t-1,c}$. Further note that (all orders hold uniformly in p element-wise)

$$
\begin{aligned}
\langle F_{t-1,c}, W_t^-(p)\rangle &= O(\log T), \\
\langle F_{t-1,c}, W_t^+\rangle &= O(\log T), \\
\langle W_t^-(p), W_t^-(p)\rangle &= \mathbb{E}W_t^-(p)(W_t^-(p))' + O(Q_T), \\
\langle W_t^+, W_t^-(p)\rangle &= \mathbb{E}W_t^+ W_t^-(p)' + O(Q_T), \\
\langle F_{t-1,c}, F_{t-1,c}\rangle / T &= O(\log T).
\end{aligned} \tag{38}
$$

These are the main results needed for the consistency proof, for which one more hurdle exists: The PCA estimate $\hat{F}_t$ is related to the self-normalized vector $\tilde{F}_t$ rather than F_t. For $F_{t-1,c}$ this is an advantage as $\langle \tilde{F}_{t,c}, \tilde{F}_{t,c}\rangle = I_c$ and $\langle \tilde{F}_{t,c}, \tilde{F}_{t,\bullet}\rangle = O((\log T)/\sqrt{T})$. However, this also implies $\langle \Delta\tilde{F}_{t,c}, \Delta\tilde{F}_{t,c}\rangle \to 0$ implying that in $\mathcal{T}_W^+ \tilde{F}_t^+$ the components corresponding to $W_{t+j,c}, j > 0$, converge to zero. While this would be dealt with by the usual CVA weight $\langle \tilde{F}_t^+, \tilde{F}_t^+\rangle^{-1/2}$ that would undo the self-normalisation, this choice of the weight for $q < r$ also introduces dynamics due to the idiosyncratic components (see the discussion in Remark 1 above). Ignoring this, the system still may be identifiable if $(I_f \otimes \begin{pmatrix} 0 & 0 \\ 0 & I_{r-c} \end{pmatrix})\mathcal{O}_f$ has rank n. As an alternative we use a different weight to show our second main result (for the proof see Appendix B):

Theorem 5 (Consistency in the I(1) case)**.** *Let the data be generated according to Assumption 4 with identification restrictions listed in Assumption 1. Let $\hat{F}_t = \hat{\Lambda}_N^\dagger y_t^N$ be obtained from the largest r principal components of the process y_t^N normalised such that $\langle \hat{F}_t, \hat{F}_t\rangle = I_r$.*

Let $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ be calculated using CVA with

- $f \geq n_O$, *the observability index,*

- $p = p(T) \geq -(1+\delta) \log T / (\log \rho_\circ) \to \infty, p = O(H_T)$ *where* $\delta > 0, H_T = (\log T)^a$ *for* $\rho_\circ > 0$ *and* $p > p_\circ$ *else (where* $p_\circ$ *denotes the lag length of an autoregressive pseudo left-inverse of* $b_\circ$*),*

- $W_f^+ = (I_f \otimes \langle \Delta \hat{F}_t, \Delta \hat{F}_t \rangle^{-1/2}), W_p^- = \langle \hat{F}_t^-(p), \hat{F}_t^-(p) \rangle^{1/2}$.

- *for* $T/N^2 \to 0, T \to \infty$

Then there exists a sequence of random matrices $H_{T,N}$ *such that the transfer function* $\hat{k}(z) := H_{T,N}(\hat{D} + z\hat{C}(I_n - z\hat{A})^{-1}\hat{B})$ *is consistent for* $b_\circ(z)$ *such that the impulse response sequence has an error of order* $O(T^{1/2-\gamma})$ *for all* $0.5 < \gamma < 1$.

Note, that the theorem does not assume any knowledge on the integer c. The logic behind the special choice of the weighting is the following: Consider $\tilde{F}_{t,c} = \langle F_{t,c}, F_{t,c} \rangle^{-1/2} F_{t,c}$. Taking the first difference we get:

$$\tilde{F}_{t,c} - \tilde{F}_{t-1,c} = \langle F_{t,c}, F_{t,c} \rangle^{-1/2}(F_{t,c} - F_{t-1,c}). \tag{39}$$

Here $\langle \Delta F_{t,c}, \Delta F_{t,c} \rangle \to \mathbb{E} \Delta F_{t,c} (\Delta F_{t,c})' > 0$, but $T^{-1} \langle F_{t,c}, F_{t,c} \rangle \xrightarrow{d} Z$ for some random matrix Z.

Therefore $\langle \Delta \tilde{F}_{t,c}, \Delta \tilde{F}_{t,c} \rangle$ tends to zero. Multiplying with its inverse undoes the normalisation of the integrated variables and hence avoids the nulling of the rows corresponding to the integrated variables in F_t^+. As it operates on the blocks, it does not upweigh the contribution of the idiosyncratic components, however. The weighting can also be applied in the stationary case as it then converges to a non-singular matrix. Hence it can be used in situations where the persistence of some of the variables is unclear.

Alternatively the normalization condition $\tilde{I}_N'\hat{\Lambda}_N = I_r$ can be applied leading to the same result of undoing the self-normalization.

Note, that as usual for CVA the lower bound on $p(T)$ is chosen such that $\underline{A}^{p(T)} = o(T^{-1})$. For consistency smaller values would be possible, but then the convergence rate may depend on the choice.

6. Choice of the integers

In order to apply the above procedure based on CVA a number of integers have to be supplied: r, f, p, n, q. In this section we compile a number of ideas on how to specify them. Some of these ideas are heuristic and hence subject to additional research, while for some there are many established procedures.

6.1. Choosing r

The choice of the number of common components has been investigated a lot in the literature starting with the PANIC approach of Bai and Ng (2004). See also Barigozzi et al (2024) for a list of procedures. This paper adds nothing in this respect.

6.2. Choosing f and p

In CVA the choice of f often is tied to the choice of p. For our results f must be chosen large enough such that $\mathcal{O}_f$ is of full column rank. This implies that f must be at least equal to the observability index n_O (the smallest integer such that $\mathcal{O}_f$ has full column rank). A simple sufficient condition is the choice $f \geq n$. Generically (in the set of all systems of order n) the observability index equals $\lceil n/r \rceil$ (see Hannan and Deistler, 1988, Theorem 2.5.3.). Note that this may equal $n_O = 1$ if $n \leq r$. Hence f much smaller than n often can be used.

In (Bauer and Ljung , 2002) it is shown that $f \to \infty$ is favourable with respect to the asymptotic variance for the choice of the weight W_f^+ related to canonical correlations. This choice also is a sufficient condition for the asymptotic equivalence to Gaussian quasi likelihood estimation (see Bauer, 2005b).

For our suggestion $W_f^+ = I_{fr}$ one example in Bauer and Ljung (2002) shows the best accuracy for small values of f. It is unclear, whether this is an exception.

With respect to the choice of p in the literature the recommendation to use the lag order of a long VAR approximation of F_t using information criteria can be found, cf. eg. Bauer (2005a). This is rooted in the idea that (15) has similarities to such an approximation, but includes more prediction horizons simultaneously. In the integrated case a doubling of the integer has been proposed in Bauer and Wagner (2002) to achieve $\underline{A}^p = o(T^{-1})$ rather than only $\underline{A}^p = o(T^{-1/2})$.

In this respect a complication arises since typically information criteria are formulated as

$$IC(h; C_T) = \log \det \hat{\Omega}_{T,h} + C_T h r^2 / T \tag{40}$$

for an autoregressive approximation using h lags, where $\hat{\Omega}_{T,h}$ denotes an estimate of the variance matrix for the error term. Since the limit Ω_h of $\hat{\Omega}_{T,h}$ is singular for a singular process at least in the limit for $h \to \infty$, the determinant tends to zero and thus the logarithm to $-\infty$.

For singular processes the criterion, therefore, should be adapted, for example to

$$\widetilde{IC}(h; C_T) = \text{trace}[\hat{\Omega}_{T,h}] + C_T h r^2/T \tag{41}$$

It is not too hard to see that minimizing this criterion leads to similar asymptotic properties as AIC or BIC also for singular processes: Theorem 7.4.7 of Hannan and Deistler (1988) states for the strict minimum-phase case and square, non-singular transfer functions that

$$\hat{\Omega}_{T,h} = \dot{\Omega}_T + (\Omega_h - \Omega)(1 + o(1)) + O(\frac{h \log T}{T}), \quad \dot{\Omega}_T = T^{-1} \sum_{t=1}^{T} \varepsilon_t \varepsilon_t'. \tag{42}$$

It is not too hard to see that this carries over to the singular case. This implies that the estimated innovation variance equals the estimated variance based on the true innovations $\dot{\Omega}_T$ (that does not depend on h), the expected loss in accuracy from using a lag h approximation $(\Omega_h - \Omega)$ and an error term that is of order $h(\log T)/T$. If $C_T/\log T \to \infty$ the penalty term dominates the error leading to a consistent order selection procedure. The factor $\log T$ can be eliminated providing (weak) consistency for BIC (see Theorem 7.4.6 and 7.4.7 of Hannan and Deistler, 1988).

Taking the trace rather than the logarithm of the determinant implies that the scale is not eliminated from the criterion. Thus, even if it is simple to show consistency for lag length selection using $\widetilde{IC}$ and large enough penalty term, it is clear that this rough idea is not optimal in any sense.

6.3. Choosing n

The system order n in subspace methods is typically found by detecting non-zero canonical correlations between the past and the future, compare Bauer (2001). This is obtained from a SVD of the weighted projection matrix using the weight $W_f^+ = \langle F_t^+, F_t^+ \rangle^{-1/2}$.
It has been noted by Breitung and Pigorsch (2012) that for singular processes unit canonical correlations may arise. As in section 3 we see that for zero-free tall transfer functions we have

$$[I_n, 0] L_f^{\dagger} F_t^+ = x_t = \mathcal{K}_p F_t^-(p). = x_t. \tag{43}$$

In this case, thus, there are exactly n linear combinations of the space spanned by the future observations $F_t, F_{t+1}, ..., F_{t+f-1}$ that are also contained in the space of the past observations. Correspondingly, there must be exactly n canonical correlations equal to 1. This also holds in the $I(1)$ case.
Using $\hat{F}_t$ in place of F_t implies that the unit canonical correlations are not exact but only approximate due to the noise added via the idiosyncratic components. Additionally, since the idiosyncratic terms act as noise, the rest of the canonical correlations (after the first n) may be non-zero even in the limit for $q < r$ (compare Remark 1; this is different from the square, non-singular case, where only the first n canonical correlations are non-zero asymptotically). This implies that using the weight $W_f^+ = \langle \hat{F}_t^+, \hat{F}_t^+ \rangle^{-1/2}$ we – contrary to the non-singular case – obtain information on the system order in some situations by detecting the number of canonical correlations close to 1, while additional non-zero canonical correlations come from the idiosyncratic part and hence in this setting do not correspond to the signal part, but to the

noise. Note, however, that zeros in $b_\circ(z)$ will lead to canonical correlations for F_t smaller than 1 even in the limit. Hence it is not clear that detecting the number of unit canonical correlations is a reliable procedure.

6.4. Choosing q

There are two sources of information for q:

- The spectrum of F_t has rank q everywhere under our assumptions.
- The innovation variance $\Omega = D_\circ D_\circ'$ where $D_\circ \in \mathbb{R}^{r\times q}$ and hence has rank q.

Procedures to infer the rank from the spectrum can be taken from the literature, see, for example, the references in section 4 of Barigozzi et al (2016). Note that in aDFMs the identification method requires $N \to \infty$ in order to infer the number of common trends.

Secondly the innovation variance Ω has q non-zero eigenvalues. Thus inferring the rank of an estimate for this matrix provides cues. This can be done for example using simple thresh-holding methods using the error bound $O(Q_T)$. Such crude methods in first simulations work surprisingly well.

7. Conclusions

In this paper we show that the CVA subspace procedure can be used in order to obtain consistent estimates of the (potentially tall) transfer function linking the dynamic to the static factors in the aDFM representation. This can be based on consistent estimates of the static factors that can be, for example, obtained using the first r principal components.

Consistency for the estimation of the dynamic factor part of the model holds for stationary processes as well as for cointegrated static common factor processes.

The method requires the specification of a number of integer values. While we give some hints on how to choose them, the suggestions are not fully satisfactory and more refined procedures need to be researched.

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

References

Anderson, B.D.O., Deistler, M. (2008). Properties of Zero-Free Transfer Function Matrices. SICE Journal of Control, Measurement, and System Integration 1 (4). 284–292.

Bai, J., Ng, S. (2004). A PANIC attack on unit roots and cointegration. Econometrica 72, 4 pp. 1127–1177.

Bai, J., Ng, S. (2013). Principal components estimation and identification of static factors. Journal of Econometrics 176, 1 pp. 18–29.

Bai, J., Ng, S. (2019). Rank regularized estimation of approximate factor models. Journal of Econometrics 212, 1 pp. 78—96

Barigozzi, M, Hallin M., Luciani, M., Zaffaroni, P. (2024). Inferential theory for generalized dynamic factor models, Journal of Econometrics, 239, 2, 105422, Barigozzi, M, Lippi, M., Luciani, M. (2021). Large-dimensional dynamic factor models: Estimation of impulse–response functions with I(1) cointegrated factors, Journal of Econometrics, 221, 2, 455–482

Barigozzi, M, Lippi, M., Luciani, M. (2016). Non-stationary dynamic factor models for large datasets, Available at SSRN 2741739

Bauer, D. (2001). Order estimation for subspace methods, Automatica, 37, 10, 1561-1573

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

Bauer, D. (2005b). Comparing the CCA subspace method to pseudo maximum likelihood methods in the case of no exogenous inputs, Journal of Time Series Analysis, 26, 5, 631-668

Bauer, D. (2009). Almost sure bounds on the estimation error for OLS estimators when the regressors include certain MFI (1) processes, Econometric Theory 25, 2, 571–582

Bauer, D. (2025). Using Subspace Algorithms for the Estimation of Linear State Space Models for Over-Differenced Processes., Available at arxiv, no. 2502.09987.

Bauer, D., Buschmeier, R. (2021). Asymptotic properties of estimators for seasonally cointegrated state space models obtained using the CVA subspace method. Entropy 23, 4, 436

Bauer, D., Ljung, L. (2002). Some facts about the choice of the weighting matrices in Larimore type of subspace algorithms, Automatica, 38, 5, 763-773

Bauer, D., Wagner, M. (2002). Estimating cointegrated systems using subspace algorithms, Journal of Econometrics, 111, 1, 47–84

Breitung, J., Pigorsch, U. (2013). A canonical correlation approach for selecting the number of dynamic factors., Oxford Bulletin of Economics and Statistics, 75, 1, 23–36

Chamberlain, G., Rothschild, M. (1983). Arbitrage, factor structure and meanvariance analysis in large asset markets, Econometrica, 51, 1305—1324

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

Deistler, M., Anderson, B.D.O., Chen, W., Filler, A., (2012). Autoregressive model of singular spectral matrices. Automatica 48. pp. 2843–2849.

Doz, C., Giannone, D., Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering, Journal of Econometrics, 164, 1, 188–205,

Forni, M., Gambetti, L., Granese, A., Sala, L. , Soccorsi, S. (2023). An American Macroeconomic Picture. Supply and Demand Shocks in the Frequency Domain, CEPR Discussion Papers, 18070

Forni, M., Hallin, M., Lippi, M., Reichlin, L. (2004). The generalized dynamic factor model: consistency and rates, Journal of Econometrics, 119, 231–255

Forni, M., Lippi, M. (2023). Estimating Singular VARs by Means of Non-Singular Approximations. EIEF Working Paper 23/1, October 2023. .

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

Larimore, W.E. (1983). System Identification, reduced order filters and modeling via canonical variate analysis, In proceedings of the 1983 American Control Conference, Piscataway, NJ, 445–451

Lippi, M. (2024). Validating DSGE models with SVARs and high-dimensional dynamic factor models, Econometric Theory, 39, 6, 1273–1291

Lippi, M., Deistler, M. and Anderson, B.D.O. (2023). High-dimensional dynamic factor models: A selective survey and lines of future research, Econometrics and Statistics, 26, 3–16,

Stock, J. H., Watson, M. W. (2011). Dynamic Factor Models. In M. P. Clements, and D. F. Hendry (eds), The Oxford Handbook of Economic Forecasting, Oxford Handbooks

Appendix A. Preliminary Lemmas

Lemma 1. *Let* $(F_t)_{t\in\mathbb{Z}}$ *be a process generated according to Assumption 2. Then the following holds for the covariance matrix* Γ_f *of the vector* $F_t^+ = [F_t', F_{t+1}', ..., F_{t+f-1}']' \in \mathbb{R}^{rf}$ *(*M *denotes a generic constant, not necessarily the same everywhere):*

(I) rank$(\Gamma_f) = n_\circ + fq = n(f)$ *where* $0 \le n_\circ \le n$ *equals* n *minus the number of zeros of* $b_\circ(z)$ *(including their multiplicity) defined as in Definition 1 of Anderson and Deistler (2008), see (8).*

(II) $\sup_{f\in\mathbb{N}} \|\Gamma_f\|_2 < M, \sup_{f\in\mathbb{N}} \|\Gamma_f\|_\infty < M$.

(III) If $\tilde\Gamma_f$ *is defined as* $\tilde\Gamma_f = U_f \tilde S_f U_f'$ *where* $\Gamma_f = U_f S_f U_f'$ *with* $U_f' U_f = I_{rf}, U_f \in \mathbb{R}^{rf\times rf}$ *and* $S_f = diag(s_1, ..., s_{n(f)}, 0) \in \mathbb{R}^{rf\times rf}$ *as well as* $\tilde S_f = diag(\tilde s_1, ..., \tilde s_{n(f)}, \epsilon I_{rf-n(f)})$ *where* $\tilde s_j = \max(s_j, \epsilon)$ *for* $\epsilon > 0$*. Then* $\sup_{f\in\mathbb{N}} \|\tilde\Gamma_f^{-1}\|_2 < M$.

Proof. (I) The representation $F_t^+ = \mathcal{O}_f x_t + \mathcal{U}_f \tilde U_t^+ = L_f \begin{pmatrix} x_t \\ \tilde U_t^+ \end{pmatrix}$ implies that the rank of Γ_f equals the column rank of $L_f \in \mathbb{R}^{rf\times(n+fq)}$. Here $\tilde U_t^+$ denotes U_t^+ built using u_t. Theorem 5 of Anderson and Deistler (2008) states that L_f for large enough f ($f \ge n$ is sufficient) does not have full column rank for $D_\circ$ of full column rank, if $b(z)$ has a zero. For each zero a different vector in the kernel is constructed, such that the rank is reduced by one. The matrix $\mathcal{U}_f$ has full column rank for D being of full column rank.

(II) The stability of $A_\circ$ implies $\sup_{f\in\mathbb{N}} \|\Gamma_f\|_2 < M$ as in the square case since $x'\Gamma_f x = \int_{-\pi}^{\pi} x(\omega)^* f_F(\omega) x(\omega) d\omega$ where $f_F(\omega)$ denotes the spectrum of the process $(F_t)_{t\in\mathbb{Z}}$ and $x(\omega) = \sum_{j=1}^f x_j e^{-ij\omega} \in \mathbb{C}^r$.

For the infinity norm, note that the $(i,j) - th, i < j$, block of Γ_f equals (where $P = \mathbb{E} x_t x_t'$ has finite norm)

$$C_\circ A_\circ^{i-1} P (C_\circ A_\circ^{j-1})' + \begin{pmatrix} C_\circ A_\circ^{i-2} B_\circ & \dots & C_\circ B_\circ & D_\circ \end{pmatrix} \begin{pmatrix} (C_\circ A_\circ^{j-2} B_\circ)' \\ \vdots \\ (C_\circ A_\circ^{j-2-i+1} B_\circ)' \end{pmatrix} \tag{A.1}$$

since $\mathbb{E} u_t u_s' = I_q \mathbb{I}(t = s)$ by assumption. It follows that the first term tends to zero exponentially with i, j growing. The second term is of the form $X_i (A_\circ^{j-i-1})' C_\circ'$ where $\|A_\circ^{j-i-1}\| \to 0$ for $j - i \to \infty$ and $\sup_{i \in \mathbb{N}} \|X_i\| < M$.

(III) This directly follows from regularization.

(IV) Note that with the left pseudo-inverse of Assumption 2 we get for

$$\begin{aligned} P_j &= \begin{pmatrix} -C_\circ \underline{A}_\circ^{j-2} B_\circ D_\circ^\dagger & \dots & -C_\circ B_\circ D_\circ^\dagger & I_r \end{pmatrix} \in \mathbb{R}^{r \times rj}, \\ \Rightarrow P_j & \begin{pmatrix} C_\circ & D_\circ & 0 & \dots & 0 \\ C_\circ A_\circ & C_\circ B_\circ & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ C_\circ A_\circ^{j-1} & C_\circ A_\circ^{j-2} B_\circ & \dots & C_\circ B_\circ & D_\circ \end{pmatrix} = \begin{pmatrix} C_\circ \underline{A}_\circ^{j-1} & 0 & \dots & 0 & D_\circ \end{pmatrix} \end{aligned} \tag{A.2}$$

such that using $\tilde{P}_j = [P_j, 0] \in \mathbb{R}^{r \times rf}$

$$\begin{pmatrix} \tilde{P}_1 \\ \tilde{P}_2 \\ \vdots \\ \tilde{P}_f \end{pmatrix} L_f = \begin{pmatrix} C_\circ & D_\circ & 0 & \dots & 0 \\ C_\circ \underline{A}_\circ & 0 & D_\circ & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ C_\circ \underline{A}_\circ^{f-1} & 0 & \dots & 0 & D_\circ \end{pmatrix}. \tag{A.3}$$

Pre-multiplying with $(I_f \otimes D_\circ^\dagger)$ results in a matrix of rank qf. Furthermore for $D_\perp' D_\circ = 0, D_\perp' D_\perp = I$ $(D_\perp \in \mathbb{R}^{r \times (r-q)})$ we get $D_\perp' \tilde{P}_j \mathcal{U}_f = 0$ and hence

$$(I_f \otimes D'_\perp) \begin{pmatrix} \tilde{P}_1 \\ \tilde{P}_2 \\ \vdots \\ \tilde{P}_f \end{pmatrix} F_t^+ = \underbrace{\begin{pmatrix} D'_\perp C_\circ \\ D'_\perp C_\circ \underline{A}_\circ \\ \vdots \\ D'_\perp C_\circ \underline{A}_\circ^{f-1} \end{pmatrix}}_{\underline{\mathcal{O}_f}^\perp} x_t. \tag{A.4}$$

The matrix on the right hand side achieves its maximal column rank $n_\circ$, say, for $f = n$ (taking f larger does not change the rank). Accounting for this we obtain a matrix $P_\perp \in \mathbb{R}^{[f(r-q)-n_\circ]\times fr}$ such that $P_\perp \Gamma_f = 0$. Note that for each row the sum of the absolute values of the entries of $\tilde{P}_j$ is uniformly bounded. The same then holds for appropriate choice of $P_\perp$, the left kernel of Γ_f. This can be achieved, for example by the following construction:

1. Choose $O_{\underline{n},\perp}$ as an orthonormal column orthogonal to the first $\underline{n}$ block rows of $\underline{\mathcal{O}_f}^\perp$.
2. Let $P_{\underline{\mathcal{O}_f}^\perp}$ denote the projection onto the column space of $\underline{\mathcal{O}_f}^\perp$. then $[O'_{\underline{n},\perp}, 0]P_{\underline{\mathcal{O}_f}^\perp} = 0$.
3. The j-th block row (for $j \in \{\underline{n}+1, ..., f\}$) of the matrix $P_{\underline{\mathcal{O}_f}^\perp}$ then is orthogonal to this column. The norm of these block rows is bounded by a constant times $\|\underline{A}_\circ^j\|$ which tends to zero for $j \to \infty$. Hence for $\underline{n}$ large enough the j-th block row of $I - P_{\underline{\mathcal{O}_f}^\perp}$ is of full row rank and also orthogonal to $[O'_{\underline{n},\perp}, 0]$.
4. Orthogonalising each block row of $I - P_{\underline{\mathcal{O}_f}^\perp}$ with respect to the previous rows involves a matrix of uniformly bounded infinity norm.
5. Dividing each block row by the square root of its norm leads to an orthonormal matrix $P_\perp$.

Furthermore we obtain

$$\underbrace{(I \otimes D_\circ^\dagger)\begin{pmatrix}\tilde{P}_1\\ \tilde{P}_2\\ \vdots\\ \tilde{P}_f\end{pmatrix}}_{P_{1,f,D}} F_t^+ = \begin{pmatrix} D_\circ^\dagger C_\circ \\ D_\circ^\dagger C_\circ \underline{A}_\circ \\ \vdots \\ D_\circ^\dagger C_\circ \underline{A}_\circ^{f-1}\end{pmatrix} x_t + \begin{pmatrix} u_t \\ u_{t+1} \\ \vdots \\ u_{t+f}\end{pmatrix} \tag{A.5}$$

where $P_{1,f,D} \in \mathbb{R}^{qf \times rf}$ such that $P_{1,f,D}\Gamma_f P'_{1,f,D} = I_{qf} + \underline{\mathcal{O}}_f V_x \underline{\mathcal{O}}_f{}'$. Adding $n_\circ + nq$ rows corresponding to $\tilde{\Gamma}_n$, the non-null part of Γ_n, to $P_{n+1,f,D}$ we obtain a matrix $\tilde{P}_D \in \mathbb{R}^{(fq+n_\circ)\times fr}$ (with $\sup_f \|\tilde{P}_D\|_\infty < \infty$) such that

$$\tilde{P}_D \Gamma_f \tilde{P}_D' = \begin{pmatrix} \tilde{\Gamma}_n & \tilde{M}_n' \underline{\mathcal{O}}_f' \\ \underline{\mathcal{O}}_f \tilde{M}_n & \tilde{I}_{q(f-n)} + \underline{\mathcal{O}}_f V_x \underline{\mathcal{O}}_f' \end{pmatrix}. \tag{A.6}$$

Orthogonalization with respect to the first block leads to

$$\underbrace{\begin{pmatrix} I & 0 \\ -\underline{\mathcal{O}}_f \tilde{M}_n \tilde{\Gamma}_n^{-1} & I \end{pmatrix} \tilde{P}_D}_{P_D} \Gamma_f \tilde{P}_D' \begin{pmatrix} I & 0 \\ -\underline{\mathcal{O}}_f \tilde{M}_n \tilde{\Gamma}_n^{-1} & I \end{pmatrix}' = \begin{pmatrix} \tilde{\Gamma}_n & 0 \\ 0 & I_{q(f-n)} + \underline{\mathcal{O}}_f \tilde{V}_x \underline{\mathcal{O}}_f' \end{pmatrix} \tag{A.7}$$

where $\tilde{V}_x = V_x - \tilde{M}_n \tilde{\Gamma}_n^{-1} \tilde{M}_n' \geq 0$. Stacking P_D and $P_\perp$ we obtain a matrix $P \in \mathbb{R}^{fr \times fr}$ such that

$$P \Gamma_f P' = \begin{pmatrix} \tilde{\Gamma}_n & 0 & 0 \\ 0 & I_{q(f-n)} + \underline{\mathcal{O}}_f \tilde{V}_x \underline{\mathcal{O}}_f' & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{A.8}$$

Further $\|P\|_\infty, \|P\|_1$ are bounded uniformly. Note that the rows of P_D

and $P_\perp$ are not necessarily orthogonal. However, using an appropriate transformation this can be achieved:

$$P_{D\perp} = P_D - P_D P_\perp' P_\perp. \tag{A.9}$$

The structure of P_D and $P_\perp$ shows that the one-norm of the rows of this matrix are uniformly bounded.

Now turn to the inverses. The smallest eigenvalue of $\tilde{\Gamma}_f$ is bounded away from zero by construction. Thus $\lambda_{min}(\tilde{\Gamma}_f) \geq \epsilon$ which implies $\sup \|\tilde{\Gamma}_f^{-1}\|_2 \leq 1/\epsilon$. From (A.8) we see that the regularization is required in the row space of $P_\perp$ where the zero eigenvalue is changed to ϵ.

The rest of the proof follows the proof of Theorem 6.6.11 of Hannan and Deistler (1988), in the following denoted as HD. From above we obtain after regularization that

$$P\tilde{\Gamma}_f P' = \begin{pmatrix} \tilde{\Gamma}_n & 0 & 0 \\ 0 & I_{q(f-n)} + \underline{\mathcal{O}}_f \tilde{V}_x \underline{\mathcal{O}}_f' & 0 \\ 0 & 0 & \epsilon I \end{pmatrix} \tag{A.10}$$

This implies as in the proof in HD that $\|(P\tilde{\Gamma}_f P')^{-1}\|_\infty$ is uniformly bounded. The result then follows from $\|P\|_\infty$ being uniformly bounded as derived above. ■

Lemma 2. *Let $(F_t)_{t\in\mathbb{Z}}$ be a process generated according to Assumption 2. Let $D_\otimes = I_p \otimes D^\dagger$ and let $F_t^\dagger(p) = D_\otimes F_t^-(p)$. Then*

$$
\begin{aligned}
\hat{\Phi}_{\otimes}(p) &= \langle F_t, F_t^{\dagger}(p)\rangle \langle F_t^{\dagger}(p), F_t^{\dagger}(p)\rangle^{-1}, \\
\Phi_{\otimes}(p) &= \mathbb{E}F_t(F_t^{\dagger}(p)')(\mathbb{E}F_t^{\dagger}(p)(F_t^{\dagger}(p))')^{-1}, \qquad \text{(A.11)} \\
\hat{\Phi}_{\otimes}(p) - \Phi_{\otimes}(p) &= O(Q_T).
\end{aligned}
$$

Here $\Phi_{\otimes}(p) = [\Phi_{1,\otimes}(p), ..., \Phi_{p,\otimes}(p)]$ *such that* $\Phi_{j,\otimes}(p) \to \Phi_{j,\otimes}$ *for* $p \to \infty$. *Additionally* $\Phi_{j,\otimes}(p) - \Phi_{j,\otimes} = O(\rho_0^p)$.

Proof. The first statement follows from Lemma 1 jointly with the uniform rate of convergence for covariances.

For the second statement note that the existence of the stable left pseudo-inverse $D_\circ^{\dagger} - zD_\circ^{\dagger}C_\circ(I_n - z\underline{A}_\circ)^{-1}BD_\circ^{\dagger}$ implies that

$$
u_t = D_\circ^{\dagger}F_t - D_\circ^{\dagger}C_\circ x_t \Rightarrow x_{t+1} = \underline{A}_\circ x_t + B_\circ D_\circ^{\dagger}F_t. \qquad \text{(A.12)}
$$

Consequently

$$
\begin{aligned}
F_t = C_\circ x_t + D_\circ u_t &= C_\circ(\underline{A}_\circ x_{t-1} + B_\circ D_\circ^{\dagger}F_{t-1}) + D_\circ u_t \\
&= D_\circ u_t + \sum_{j=1}^{p} C_\circ \underline{A}_\circ^{j-1} B_\circ D_\circ^{\dagger} F_{t-j} + C_\circ \underline{A}_\circ^{p} x_{t-p} \\
&= D_\circ u_t + [\Phi_{j,\otimes}]_{j=1,..,p} F_t^{\dagger}(p) + C_\circ \underline{A}_\circ^{p} x_{t-p} \qquad \text{(A.13)}
\end{aligned}
$$

where $\Phi_{j,\otimes} = C_\circ \underline{A}_\circ^{j-1} B_\circ D_\circ^{\dagger}$. From this equation it follows that

$$
\Phi_{\otimes}(p) = [\Phi_{j,\otimes}]_{j=1,..,p} + C_\circ \underline{A}_\circ^{p} \mathbb{E}x_{t-p} F_t^{-}(p)'(\mathbb{E}F_t^{-}(p)F_t^{-}(p)')^{\dagger}. \qquad \text{(A.14)}
$$

Now the uniform bounds on the two and the infinity norm (see Lemma 1 above) shows the result. ■

Note, that here the approximation quality of the AR(p) model $\Phi_{\otimes}(p)$ is only given as an upper bound only using the non-singular process $D_\circ^\dagger F_{t-j}$ for the approximation. The inclusion of $D_\perp' F_{t-j}$ can aid the prediction. In fact in the zero-less case we can reconstruct x_{t-p} from $F_t^-(p)$ for p large enough such that the prediction is perfect already for finite p, as in this case there exist autoregressive left pseudo-inverses.

Next let $F_{t|t-1}(p) = \Phi(p)F_t^-(p)$ denote the best approximation of F_t based on $F_t^-(p)$, whereas $F_{t|t-1} = C_\circ x_t$ denotes its limit (shown to exist below) for $p \to \infty$. Then from above we get:

$$F_t = D_\circ u_t + C_\circ x_t = D_\circ u_t + F_{t|t-1}(p) + (F_{t|t-1} - F_{t|t-1}(p)), \tag{A.15}$$

$$\mathbb{E}(F_{t|t-1} - F_{t|t-1}(p))(F_{t|t-1} - F_{t|t-1}(p))' \leq C_\circ \underline{A}_\circ^p \left(\mathbb{E}x_{t-p}x_{t-p}'\right)(\underline{A}_\circ^p)' C_\circ', \tag{A.16}$$

$$\mathbb{E}(F_t - F_{t|t-1}(p))(F_t - F_{t|t-1}(p))' \leq D_\circ D_\circ' + C_\circ \underline{A}_\circ^p \left(\mathbb{E}x_{t-p}x_{t-p}'\right)(\underline{A}_\circ^p)' C_\circ'. \tag{A.17}$$

Using these equations we obtain:

Lemma 3. *Let $(F_t)_{t\in\mathbb{Z}}$ be a process generated according to Assumption 2. Then for small enough ϵ we have*

$$\hat{\Phi}(p) = \langle F_t, F_t^-(p)\rangle\langle F_t^-(p), F_t^-(p)\rangle^\dagger - \Phi(p) = O(Q_T) \tag{A.18}$$

uniformly for $1 \leq p \leq H_T, H_T = O((\log T)^a), a > 1$, where $\langle F_t^-(p), F_t^-(p)\rangle^\dagger$

denotes $\tilde{\Gamma}_p^{-1}$.

Furthermore letting $\hat{F}_{t|t-1}(p) = \hat{\Phi}(p)F_t^-(p)$ *and* $F_{t|t-1}(p) = \Phi(p)F_t^-$ *we have*

$$(I)\langle \hat{F}_{t|t-1}(p) - F_{t,t-1}(p), \hat{F}_{t|t-1}(p) - F_{t|t-1}(p)\rangle = O(Q_T), \tag{A.19}$$

$$(II)\langle \hat{F}_{t|t-1}(p) - F_{t|t-1}(p), \hat{F}_{t|t-1}(p)\rangle = O(Q_T), \tag{A.20}$$

$$(III)\langle \hat{F}_{t|t-1}(p), \hat{F}_{t|t-1}(p)\rangle = \langle F_{t|t-1}(p), F_{t|t-1}(p)\rangle + O(Q_T), \tag{A.21}$$

$$(IV)\langle F_t, \hat{F}_{t|t-1}(p)\rangle = \langle F_t, F_{t|t-1}(p)\rangle + O(Q_T) \tag{A.22}$$

where $\langle a_t, b_t\rangle = T^{-1}\sum_{t=p+1}^{T} a_t b_t'$ *for processes* $(a_t)_{t\in\mathbb{Z}}$ *and* $(b_t)_{t\in\mathbb{Z}}$.

Proof. Since $\langle F_t, F_{t-j}\rangle - \mathbb{E}F_tF_{t-j}' = O(Q_T)$ (see, e.g., HD, Theorem 5.3.2) the results follow from Lemma 2 above in conjunction with the bounds on the 2- and infinity norms of $\langle F_t^-(p), F_t^-(p)\rangle^\dagger$. Choosing ϵ small enough, the regularization is only active in the kernel of Γ_p, the effects of which are canceled by $\mathbb{E}F_tF_t^-(p)'$.

Therefore all the limits for the estimated quantities equal the ones for the true quantities up to an error of order $O(Q_T)$. ■

Lemma 4. *Under the Assumptions 2 where* $p \to \infty$ *we have* $F_{t|t-1}(p) - F_{t|t-1} \to 0$ *where* $F_{t|t-1} = C_\circ x_t$.

Proof. The existence of a stable left pseudo-inverse to $b_\circ(z)$ implies that

$$b_\circ^\dagger(L)F_t = u_t. \tag{A.23}$$

Hence analogously to the proof of Lemma 2 we get

$$F_t = D_\circ u_t + \sum_{j=1}^{p} \tilde{\Phi}_j(p)F_{t-j} - C_\circ \underline{A}_\circ^p \tilde{x}_{t-p} \tag{A.24}$$

where $\tilde{x}_{t+1} = A_\circ \tilde{x}_t + B_\circ u_t$. This shows that for $p \to \infty$ we obtain a perfect reconstruction of u_t based on the past of F_t where the approximation error is of order $\underline{A}_\circ^p$ and hence tending to zero exponentially fast. This shows the lemma. ■

The last preliminary lemma deals with the calculation of eigendecompositions with three groups of eigenvalues: the ones contained in the diagonal matrix $\hat{S}_1$ tend to infinity, the one in $\hat{S}_2$ tend to their non-zero finite values and the ones in $\hat{S}_3$ tend to zero. Such situations occur for $\langle F_t^+, F_t^+ \rangle$ in the integrated case, where due to the common trends some eigenvalues tend to infinity at rate T, the ones corresponding to the stationary components towards their finite limits, while some eigenvalues are zero for the singular process.

Lemma 5. *(I) Let $\hat{A}_T$ and $A_{\circ,T}$ be square, symmetric matrices with eigenvalue decompositions given as*

$$\hat{A}_T = \hat{U}_1 \hat{S}_1 \hat{U}_1' + \hat{U}_2 \hat{S}_2 \hat{U}_2' + \hat{U}_3 \hat{S}_3 \hat{U}_3', \tag{A.25}$$

$$A_{\circ,T} = U_1 S_{1,T} U_1' + U_2 S_2 U_2' \tag{A.26}$$

where $[U_1, U_2, U_3]'[U_1, U_2, U_3] = I, [\hat{U}_1, \hat{U}_2, \hat{U}_3]'[\hat{U}_1, \hat{U}_2, \hat{U}_3] = diag(\tilde{I}_1, \tilde{I}_2, \tilde{I}_3)$, $\hat{U}_i$ is normalized such that $\hat{U}_i' U_i = I$.

Define the function $f_M(x) = \frac{\exp(x/M)-1}{\exp(x/M)+1}$ for $0 < M < \infty$. Assume that $f_{T^\gamma}(\hat{A}_T) - f_{T^\gamma}(A_{\circ,T}) = O(T^{-\gamma})$ where $\gamma > 0$ is such that $\min_j(S_{1,T,jj})/T^\gamma \to \infty$ a.s. for γ large enough, such that $f_{T^\gamma}(S_{1,T,jj}) \to 1, \forall j$. Further $\hat{S}_2/T^\gamma \to 0$ is assumed such that $f_{T^\gamma}(S_{2,T,jj}) \to 0$.

Furthermore let $[U_2, U_3]' \hat{A}_T [U_2, U_3] = U_2 S_2 U_2' + O(Q_T)$.

Then

$$\begin{aligned}
\hat{U}_1 &= U_1 - [U_2, U_3][U_2, U_3]' H_T U_1 + o(T^{-\gamma}), \|H_T\| = O(T^{-\gamma}), && \text{(A.27)}\\
\hat{U}_2 &= U_2 + U_1 U_1' H_T U_2 - U_3 U_3' \tilde{H}_T \Sigma U_2 + O(Q_T^2 + T^{-\gamma}), && \text{(A.28)}\\
\hat{U}_3 &= U_3 + \Sigma \tilde{H}_T U_3 + \Sigma \tilde{H}_T \Sigma \tilde{H}_T U_3 + o(Q_T^2), \|\tilde{H}_T\| = O(Q_T), && \text{(A.29)}
\end{aligned}$$

where $\Sigma = [U_1, U_2] diag(I, f_M(S_2)^{-1})[U_1, U_2]'$ *for some large* $0 < M < \infty$.

(II) Let $\hat{A}_T = A + \tilde{A}_T$ *where all matrices are symmetric positive semidefinite. Let* $U = [U_1, U_2]$ *denote an orthonormal matrix such that*

$$U' A U = \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix}. \tag{A.30}$$

Assume that B_1 *and* B_2 *do not have joint eigenvalues and define the numbers* $\phi = 1/\delta, s = \|U_1' \hat{A}_T U_2\|_2$ *where* δ *is the smallest distance between the eigenvalues of* B_1 *and* B_2*. Then if*

$$\|\tilde{A}_T\|_2 < \frac{1}{4}[\phi(1 + \phi s)]^{-1} \tag{A.31}$$

there exists a matrix $\hat{U}_1$ *such that* $\hat{A}_T \hat{U}_1 = \hat{U}_1 \hat{B}_1$ *normalized such that* $\hat{U}_1' U_1 = I_r$ *such that*

$$\|\hat{U}_1 - U_1\|_2 \leq 2\phi \|\tilde{A}_T\|_2, \|\hat{B}_1 - B_1\|_2 = (2\phi s + 1)\|\tilde{A}_T U_1\|_2. \tag{A.32}$$

Proof. The lemma summarizes results contained in the book Chatelin (1993). There the Rayleigh-Schrödinger decompositions are derived.

(I) In the first stage we use $f_{T^\gamma}(\hat{A}_T)$ and $f_{T^\gamma}(A_{\circ,T}) \to U_1 U_1'$ according to the assumptions such that in the limit all eigenvalues are either one or zero.

Letting $H_T = f_{T^\gamma}(\hat{A}_T) - f_{T^\gamma}(A_{\circ,T}) = O(T^{-\gamma})$ the first order expressions for $\hat{U}_1$ are given as:

$$\hat{U}_1 = U_1 - [U_2, U_3][U_2, U_3]' H_T U_1 + o(T^{-\gamma}). \tag{A.33}$$

For $[\tilde{U}_2, \tilde{U}_3]$ normalized such that $[\tilde{U}_2, \tilde{U}_3]'[U_2, U_3] = I$ we get:

$$[\tilde{U}_2, \tilde{U}_3] = [U_2, U_3] + U_1 U_1' H_T [U_2, U_3] + o(T^{-\gamma}). \tag{A.34}$$

In the second step the eigenvalue zero in the matrix above is isolated. In the current situation using M larger than the largest eigenvalue contained in S_2 we get

$$f_M(\hat{A}_T) = \hat{U}_1 f_M(\hat{S}_1)\hat{U}_1' + \hat{U}_2 f_M(\hat{S}_2)\hat{U}_2' + \hat{U}_3 f_M(\hat{S}_3)\hat{U}_3', \tag{A.35}$$

$$f_M(A_{\circ,T}) = U_1 f_M(S_{1,T})U_1' + U_2 f_M(S_2)U_2'. \tag{A.36}$$

such that using $\hat{U}_1 - U_1 = O(T^{-\gamma})$ we obtain

$$f_M(\hat{A}_T) - f_M(A_{\circ,T}) = \hat{U}_2 f_M(\hat{S}_2)\hat{U}_2' + \hat{U}_3 f_M(\hat{S}_3)\hat{U}_3' - U_2 f_M(S_2)U_2' + O(T^{-\gamma}). \tag{A.37}$$

From $[\hat{U}_2, \hat{U}_3] = [\tilde{U}_2, \tilde{U}_3]\mathcal{T}_U$ we have $U_1'[\hat{U}_2, \hat{U}_3] = O(T^{-\gamma})$ such that $U_1'(f_M(\hat{A}_T) - f_M(A_{\circ,T})) = O(T^{-\gamma})$. Since $[U_2, U_3]'(\hat{A}_T - A_{\circ,T})[U_2, U_3] = O(Q_T)$ differentiability of f_M implies that $[U_2, U_3]'(f_M(\hat{A}_T) - f_M(A_{T,\circ}))[U_2, U_3] = O(Q_T)$. Therefore $\tilde{H}_T := f_M(\hat{A}_T) - f_M(A_{\circ,T}) = O(Q_T)$.

The matrix Σ in the Rayleigh-Schrödinger expansions equals

$$\Sigma = [U_1, U_2]\mathrm{diag}(I, f_M(S_2)^{-1})[U_1, U_2]'. \tag{A.38}$$

In this situation the second order approximation of $\hat{U}_3$ is obtained as

$$\hat{U}_3 = U_3 + \Sigma\tilde{H}_T U_3 + \Sigma\tilde{H}_T\Sigma\tilde{H}_T U_3 + o(Q_T^2). \tag{A.39}$$

From above we also get $\tilde{U}_2 = U_2 + U_1 U_1' H_T U_2 + o(T^{-\gamma})$.
The result for $\hat{U}_2 - U_2$ then follows from the orthogonality restrictions, projecting this onto $I - \hat{U}_3\hat{U}_3'$.
(II) is a direct consequence of Corollary 4.4.6 of Chatelin (1993), p. 176. ■

Appendix B. Proof of the Theorems

Appendix B.1. Proof of Theorem 1

In section Appendix A a number of preliminary lemmas are derived. The essence is that the best approximation $F_{t|t-1}(p)$ of F_t based on the finite past $F_t^-(p)$ converges to $C_\circ x_t$ and the finite sample version of it provides consistent estimates of second moments.
In order to use this in the proof of Theorem 1 it is necessary to extend these results from approximating F_t to approximating $F_t^+ = (F_t', F_{t+1}', ..., F_{t+f-1}')'$. This extension is straightforward for finite f. Consequently we obtain

$$F_{t|t-1}^+(p) - \mathcal{O}_f x_t \to 0. \tag{B.1}$$

Furthermore we obtain

$$\langle \hat{F}_{t|t-1}^+(p), \hat{F}_{t|t-1}^+(p)\rangle = \langle F_{t|t-1}^+(p), F_{t|t-1}^+(p)\rangle + O(Q_T). \tag{B.2}$$

Additionally

$$\langle F^+_{t|t-1}(p), F^+_{t|t-1}(p)\rangle \to \mathbb{E}F^+_{t|t-1}(p)(F^+_{t|t-1}(p))' \tag{B.3}$$

for $T \to \infty$. Letting $p \to \infty$ we obtain $\mathbb{E}F^+_{t|t-1}(p)(F^+_{t|t-1}(p))' \to \mathcal{O}_f \mathbb{E}x_t x_t' \mathcal{O}_f'$. Thus the limit for $T \to \infty$ where $p = p(T) \to \infty$ is a rank n positive semi-definite matrix. In CVA the SVD of this matrix is truncated:

$$\langle \hat{F}^+_{t|t-1}(p), \hat{F}^+_{t|t-1}(p)\rangle = \hat{U}_n \hat{S}_n \hat{U}_n' + \hat{R}_n. \tag{B.4}$$

We obtain $\hat{\mathcal{O}}_f = \hat{U}_n \hat{S}_n^{1/2} \hat{\mathcal{T}}_n$, where $\hat{\mathcal{T}}_n$ denotes a transformation matrix choosing the state basis. From matrix perturbation theory (see Lemma 5 (II)) one obtains $\hat{\mathcal{O}}_f \to \mathcal{O}_f$ using the (infeasible) normalization $\mathcal{O}_f^\dagger \hat{\mathcal{O}}_f = I_n$. The error $\hat{\mathcal{O}}_f - \mathcal{O}_f$ carries over from the matrices which are decomposed in the SVD. Letting $\mathcal{O}_f(p)$ denote the left factor obtained from $\mathbb{E}F^+_{t|t-1}(p)(F^+_{t|t-1}(p))'$ we then get $\hat{\mathcal{O}}_f - \mathcal{O}_f(p) = O(Q_T)$.

This approximation error implies for example (using $\hat{x}_t = \hat{\mathcal{O}}_f^{\dagger} \hat{F}^+_{t|t-1}(p), x_t(p) = \mathcal{O}_f(p)^\dagger F^+_{t|t-1}(p)$)

$$\langle \hat{x}_t, \hat{x}_t\rangle = \langle x_t(p), x_t(p)\rangle + O(Q_T) = \langle x_t, x_t\rangle + O(Q_T + \|\underline{A}_\circ\|^p). \tag{B.5}$$

Here $\|\underline{A}_\circ\|^p$ is due to the approximation error. Analogously we get for the regression estimates of $C_\circ$ and $A_\circ$ that $\hat{A} - A_\circ = O(Q_T), \hat{C} - C_\circ = O(Q_T)$ where

$$\hat{C} = \langle F_t, \hat{x}_t\rangle \langle \hat{x}_t, \hat{x}_t\rangle^{-1}, \hat{A} = \langle \hat{x}_{t+1}, \hat{x}_t\rangle \langle \hat{x}_t, \hat{x}_t\rangle^{-1}. \tag{B.6}$$

It follows that

$$\langle F_t - \hat{C}\hat{x}_t, F_t - \hat{C}\hat{x}_t\rangle \to D_\circ D_\circ'. \tag{B.7}$$

The estimation error again consists of the approximation error due to finite p and the sampling error $O(Q_T)$.

Differentiability of the calculation of Cholesky factors establishes $\hat{D} - D_\circ = O(Q_T)$. Here the identification restrictions are essential.

Finally the estimation of $B_\circ$ can be dealt with identically using the residuals $\hat{u}_t = \hat{D}^{\dagger}{}_{\circ}(F_t - \hat{C}\hat{x}_t)$. This concludes the proof.

Appendix B.2. Proof of Theorem 3

The theorem mirrors Proposition 3 in Doz et al. (2011) with the difference that almost sure bounds are derived under stronger assumptions on the processes. The proof of Doz et al. (2011) can be applied also here almost unchanged: In their Lemma 2 the arguments from Brockwell and Davis are replaced by the a.s. convergences according to Assumptions 3. This changes the error term from $O_P(T^{-1/2})$ to $O(Q_T)$ and from $O_P(N^{-1})$ to $O(N^{-1})$ which is of lower order and hence dominated by $O(Q_T)$ under the assumption $\sqrt{T}/N \to 0$. The rest of the proof then follows straightforwardly.

For future reference we provide a different version of the proof that can be generalized to the integrated case. To this end note that the principal components are obtained using eigenvectors to the largest r eigenvalues of the matrix

$$\hat{\Sigma}_{N,T} := \frac{1}{NT}\sum_{t=1}^{T} y_t^N (y_t^N)' = \frac{1}{N}\Lambda_N \frac{1}{T}\sum_{t=1}^{T} F_t F_t' \Lambda_N' + \frac{1}{NT}\sum_{t=1}^{T} \xi_t^N (\xi_t^N)'$$
$$+ \frac{1}{NT}\sum_{t=1}^{T} \xi_t^N F_t' \Lambda_N' + \frac{1}{NT}\sum_{t=1}^{T} \Lambda_N F_t (\xi_t^N)' \quad \text{(B.8)}$$

Since $N^{-1}\Lambda_N'\Lambda_N \to M_\circ \in \mathbb{R}^{r\times r}$ is assumed, we can use the renormalization $\tilde{\Lambda}_N = \Lambda_N/\sqrt{N}$. The entries of $T^{-1}\sum_{t=1}^{T} \xi_t^N(\xi_t^N)', T^{-1}\sum_{t=1}^{T}\xi_t^N F_t'$ and $T^{-1}\sum_{t=1}^{T} F_t F_t'$ all deviate from their expectation by a maximal order $O(Q_T)$ (uniformly elementwise). Therefore

$$\|(NT)^{-1}\sum_{t=1}^{T}\left(\xi_t^N(\xi_t^N)' - \mathbb{E}(\xi_t^N(\xi_t^N)')\right)\|_2^2 \le \sum_{a,b=1}^{N} N^{-2} O(Q_T^2) = O(Q_T^2) \quad \text{(B.9)}$$

and $\|\mathbb{E}\xi_t^N(\xi_t^N)'\|_2 = O(1)$ such that $\|(NT)^{-1}\sum_{t=1}^{T}\xi_t^N(\xi_t^N)'\|_2 = O(N^{-1}) + O(Q_T)$. Similarly

$$\|N^{-1/2}T^{-1}\sum_{t=1}^{T}\xi_t^N F_t'\|_2 = O(Q_T), \quad \text{(B.10)}$$
$$\|T^{-1}\sum_{t=1}^{T} F_t F_t' - \mathbb{E}F_t F_t'\|_2 = O(Q_T). \quad \text{(B.11)}$$

Consequently we obtain $\|\frac{1}{NT}\sum_{t=1}^{T} y_t^N(y_t^N)' - \Lambda_N(\mathbb{E}F_tF_t')\Lambda_N'/N\|_2 = O(Q_T)$ uniformly in N since $O(N^{-1})$ is assumed to be negligible.
This error bound can be used in Theorem 4.4.5 of Chatelin (1993) which has been restated for our case in Lemma 5 (II).

In the situation above we obtain δ as the smallest eigenvalue of $\frac{\Lambda_N'\Lambda_N}{N}(\mathbb{E}F_tF_t') \to M_\circ(\mathbb{E}F_tF_t')$ and $\|\tilde{A}_T\|_2 = O(Q_T)$. Note that the bound can be made independent of N as for all involved quantities we have obtained uniform bounds.

Recall the normalization used: $\tilde{I}_N'\Lambda_N = I_r$. This implies that $\Lambda_N = U_N(\tilde{I}_N'U_N)^{-1}$ for $U_N'U_N = I_r$. Since we assume that $\Lambda_N'\Lambda_N/N \to M_\circ$ we obtain $(\tilde{I}_N'U_N)^{-1}(\tilde{I}_N'U_N)^{-T}/N \to M_\circ$. Choosing U_N such that $\tilde{I}_N'U_N$ is p.l.t. we get that $M_N^{1/2} = (\tilde{I}_N'U_N)^{-1}/\sqrt{N} \to M_\circ^{1/2}$ with $M_\circ^{1/2}$ denoting the lower triangular Cholesky factor.

It follows that $\|\hat{U}_N - U_N\|_2 = O(Q_T)$ (with the normalisation $\hat{U}_N'U_N = I_r$) where $\Lambda_N(\mathbb{E}F_tF_t')\Lambda_N'/N = U_NS_NU_N'$ such that $\Lambda_N^\dagger = (\tilde{I}_N'U_N)U_N'$.

Consequently

$$\begin{aligned}
\hat{U}_N'y_t^N/N^{1/2} &= \hat{U}_N'\Lambda_N F_t/N^{1/2} + \hat{U}_N'\xi_t^N/N^{1/2} \\
&= \hat{U}_N'U_N\frac{(\tilde{I}_N'U_N)^{-1}}{\sqrt{N}}F_t + \hat{U}_N'\xi_t^N/N^{1/2} \\
&= M_N^{1/2}F_t + \underbrace{\hat{U}_N'\xi_t^N/N^{1/2}}_{\delta F_t}.
\end{aligned} \tag{B.12}$$

For δF_t we have from above $\|T^{-1}\sum_{t=1}^T \delta F_t\delta F_t'\|_2 = O(Q_T)$ as well as $\|T^{-1}\sum_{t=1}^T \delta F_tF_t'\|_2 = O(Q_T)$. Analogously this holds for all k for the covariance sequence of these terms. Finally letting $\hat{M}_N = \hat{U}_N'\hat{\Sigma}_{N,T}\hat{U}_N$ we get

$$\|\hat{M}_N - M_N\|_2 = O(Q_T) \Rightarrow \|\hat{M}_N^{1/2} - M_N^{1/2}\|_2 = O(Q_T) \tag{B.13}$$

using the same Cholesky factors. The identification restrictions imply that this choice results in differentiability of the corresponding mapping. Therefore, letting $\hat{F}_t = \frac{\hat{M}_N^{-1/2}\hat{U}_N'y_t^N}{N^{1/2}}$, we get

$$\hat{F}_t - F_t = \hat{M}_N^{-1/2}\hat{U}_N'y_t^N/N^{1/2} - F_t = \underbrace{(\hat{M}_N^{-1/2}M_N^{1/2} - I_r)}_{\delta M_N}F_t + \delta\tilde{F}_t \tag{B.14}$$

where $\delta\tilde{F}_t = \hat{M}_N^{-1/2}\delta F_t$ and where $\delta M_N = \hat{M}_N^{-1/2}M_N^{1/2} - I_r = O(Q_T)$. Consequently we obtain

$$\begin{aligned}\langle \hat{F}_t, \hat{F}_{t-k}\rangle &= \langle F_t + \delta M_N F_t + \delta\tilde{F}_t, F_{t-k} + \delta M_N F_{t-k} + \delta\tilde{F}_{t-k}\rangle \\ &= \langle F_t, F_{t-k}\rangle(I + \delta M_N') + \delta M_N\langle F_t, F_{t-k}\rangle(I + \delta M_N') + O(Q_T) \\ &= \langle F_t, F_{t-k}\rangle + O(Q_T). \qquad \text{(B.15)}\end{aligned}$$

Appendix B.3. Proof of Theorem 3

The proof of the theorem uses the fact that the distance between the sample covariances of $\hat{F}_t$ and the ones of F_t is of the same order as the difference between the sample covariances of F_t and its expected values. This follows directly from Theorem 2.

The only difference is that the (sample) covariance of $\hat{F}_t^-(p)$ typically is not singular, but its smallest eigenvalues tend to zero of order $1/N$ (the terms due to the idiosyncratic terms average out in this order). The introduced regularization increases these eigenvalues to $\epsilon > 0$. Note, however, that the regularization is of no importance, as in $\langle \hat{F}^+_{t|t-1}(p), \hat{F}^+_{t|t-1}(p)\rangle$ the regularized directions are filtered out since the variance of $\hat{F}_t^-(p)$ tends to zero in these directions.

This shows the theorem.

Appendix B.4. Proof of Theorem 4

The proof follows closely the proof in Appendix B.2. There are two complications: First, the order of convergence is different, since in the integrated case $\langle F_t, F_t\rangle$ diverges to infinity. Secondly, the limit of $T^{-1}\langle F_t, F_t\rangle$ is not deterministic but random.

Consequently in this section we use a different normalization tha better fits with the PCA: $\Lambda_N = [\Lambda_{N,c}, \Lambda_{N,\bullet}], \Lambda_{N,c} \in \mathbb{R}^{N\times c}$. Furthermore $\Lambda_{N,c}$ (corresponding to the common trends) and $\Lambda_{N,\bullet}$ both are assumed to be positive upper triangular. For some N_0 we assume that $\Lambda'_{N_0,c}\Lambda_{N_0,\bullet} = 0$. $\Lambda'_{N_0,c}\Lambda_{N_0,c}/N_0 = I_c$ identifies $F_{t,c}$, while $\mathbb{E}F_{t,\bullet}F'_{t,\bullet} = I_{r-c}$ does the same for $F_{t,\bullet}$. This leads to an identifiable representation. This representation is infeasible as it requires knowledge of the integer c. It is a technical device in the proof. The normalization $\tilde{I}'_N\Lambda_N = I_r$ and the one used in this section are related via a static matrix describing the basis change. This is included in the matrix $H_{T,N}$ in the formulation of the theorem.

The consequences for PCA are contained in the next lemma:

Lemma 6. *Under the assumptions of Theorem 4 let $\hat{\Sigma}_{NT} = \langle y_t^N, y_t^N\rangle/N$ be used for the PCA where $\hat{U}_{N,c} \in \mathbb{R}^{N\times c}$ denotes the matrix of eigenvectors to all eigenvalues tending to infinity and let $\hat{U}_{N,\bullet}$ denote the ones corresponding to the eigenvalues converging to non-zero finite limits. Here the normalization $U'_{N,c}\hat{U}_{N,c} = I, U'_{N,\bullet}\hat{U}_{N,\bullet} = I$ is used.*

(I) Then $\hat{U}_{N,c} - U_{N,c} = O(T^{-\gamma}), \hat{U}_{N,\bullet} - U_{N,\bullet} = O(Q_T)$.

(II) Furthermore let $U_N = [U_{N,c}, U_{N,\bullet}]$ and $\hat{U}_N = [\hat{U}_{N,c}, \hat{U}_{N,\bullet}]$ such that (using the upper triangular Cholesky factor as the matrix square root here)

$$\Lambda_N/\sqrt{N} = U_N S_N^{1/2}, S_N^{1/2} = \begin{pmatrix} S_{N,c}^{1/2} & T_N \\ 0 & S_{N,\bullet}^{1/2} \end{pmatrix}. \tag{B.16}$$

Here $T_{N_0} = 0$ due to the identification assumption, but $T_N \neq 0$ in general. Define $\hat{S}_N = \hat{U}'_N\hat{\Sigma}_{NT}\hat{U}_N$. Then with $\tilde{D}_T = diag(I_cT^{-1/2}, I_{r-c})$ it follows that

$$\tilde{D}_T \hat{S}_N \tilde{D}_T = \begin{pmatrix} T^{-1} S_{N,c} \langle F_{t,c}, F_{t,c} \rangle S_{N,c} + O(T^{-\gamma}) & O(T^{1/2-\gamma}) \\ O(T^{1/2-\gamma}) & S_{N,\bullet} + O(Q_T) \end{pmatrix}. \tag{B.17}$$

(III) Let $\hat{\Psi}_N$ denote the upper triangular Cholesky factor of $\tilde{D}_T \hat{S}_N \tilde{D}_T$ and let Ψ_N denote the one of diag$(T^{-1} S_{N,c} \langle F_{t,c}, F_{t,c} \rangle S_{N,c}, S_{N,\bullet})$. Then

$$\hat{\Psi}_N - \Psi_N = \begin{pmatrix} O((\log T) T^{-\gamma}) & O((\log T) T^{1/2-\gamma}) \\ 0 & O(Q_T) \end{pmatrix}. \tag{B.18}$$

Proof. (I) Consider $\hat{\Sigma}_{NT}$:

$$\begin{aligned}
\hat{\Sigma}_{NT} &= \Lambda_N \langle F_t, F_t \rangle \Lambda_N'/N + \Lambda_N \langle F_t, \xi_t^N \rangle/N + \langle \xi_t^N, F_t \rangle \Lambda_N'/N + \langle \xi_t^N, \xi_t^N \rangle/N \\
&= \Lambda_N \langle F_t, F_t \rangle \Lambda_N'/N + \Lambda_{N,c} \langle F_{t,c}, \xi_t^N \rangle/N + \Lambda_{N,\bullet} \langle F_{t,\bullet}, \xi_t^N \rangle/N \\
&+ \langle \xi_t^N, F_{t,c} \rangle \Lambda_{N,c}'/N + \langle \xi_t^N, F_{t,\bullet} \rangle \Lambda_{N,\bullet}'/N + O(Q_T) \\
&= \Lambda_{N,c} \langle F_{t,c}, F_{t,c} \rangle \Lambda_{N,c}'/N + \Lambda_{N,c} \langle F_{t,c}, F_{t,\bullet} \rangle \Lambda_{N,\bullet}'/N + \Lambda_{N,\bullet} \langle F_{t,\bullet}, F_{t,c} \rangle \Lambda_{N,c}'/N \\
&+ \Lambda_{N,\bullet} \langle F_{t,\bullet}, F_{t,\bullet} \rangle \Lambda_{N,\bullet}'/N + \langle \xi_t^N, F_{t,c} \rangle \Lambda_{N,c}'/N + \Lambda_{N,c} \langle F_{t,c}, \xi_t^N \rangle/N + O(Q_T)
\end{aligned} \tag{B.19}$$

where the order $O(Q_T)$ holds for the 2-norm of the matrix as has been shown in the proof of Theorem 3.

Here $\|\langle \xi_t^N, F_{t,c} \rangle\|_2 = O((\log T) N^{1/2})$, being a matrix of size $N \times c$ where each entry is $O(\log T)$. Thus the two last terms are of order $O(\log T / N^{1/2})$ elementwise. Analogously the terms involving $\langle F_{t,c}, F_{t,\bullet} \rangle$ are of the same order. Further $(\log \log T)/T^2 \sum_{t=1}^T F_{t,c} F_{t,c}' > 0$ almost surely and (see, e.g., Bauer, 2009)

$$1/(T^2(\log\log T))\sum_{t=1}^{T} F_{t,c}F_{t,c}' < M a.s. \tag{B.20}$$

This clarifies the rates to be considered. With respect to the calculation of the principal components note, that this is exactly the situation of Lemma 5 (I) where $0.5 < \gamma < 1$ can be chosen arbitrarily close to one. This clarifies the rates of convergence for $\hat{U}_N$.

(II) Recall that

$$y_t^N/\sqrt{N} = \Lambda_N/\sqrt{N}F_t + \xi_t^N/\sqrt{N} = U_N S_N^{1/2} F_t + \xi_t^N/\sqrt{N}. \tag{B.21}$$

As above $\langle \xi_t^N, \xi_t^N \rangle/N = O(Q_T)$ and $\hat{U}_N' U_N = I_r + O(T^{-\gamma})$. Therefore

$$\tilde{D}_T \hat{U}_N' U_N \tilde{D}_T^{-1} = \begin{pmatrix} I_c + O(T^{-\gamma}) & O(T^{-\gamma-1/2}) \\ O(T^{1/2-\gamma}) & I_{r-c} + O(T^{-\gamma}) \end{pmatrix}. \tag{B.22}$$

Moreover

$$\tilde{D}_T S_N^{1/2} \langle F_t, F_t \rangle S_N^{T/2} \tilde{D}_T = \begin{pmatrix} S_{N,c}^{1/2} Z_T S_{N,c}^{T/2} + O((\log T)T^{-1}) & O((\log T)T^{-1/2}) \\ O((\log T)T^{-1/2}) & S_{N,\bullet} + O(Q_T) \end{pmatrix} \tag{B.23}$$

where $Z_T := T^{-1}\langle F_{t,c}, F_{t,c} \rangle = O(\log T)$.

Since all terms involving ξ_t^N are of lower order it follows that

$$\tilde{D}_T \hat{S}_N \tilde{D}_T = \begin{pmatrix} S_{N,c}^{1/2} Z_T S_{N,c}^{T/2} + O((\log T)T^{-\gamma}) & O(T^{1/2-\gamma}(\log T)) \\ O(T^{1/2-\gamma}(\log T)) & S_{N,\bullet} + O(Q_T) \end{pmatrix}. \tag{B.24}$$

(III) This follows from the differentiability of the Cholesky factor where the (2,1) block is zero by definition. ■

For calculating the principal components note that the normalization chosen equals $\langle \tilde{F}_t, \tilde{F}_t \rangle = I_r$. We achieve this via

$$\hat{F}_t = \hat{S}_N^{-1/2} \hat{U}_N' y_t^N = (\tilde{D}_T \hat{S}_N \tilde{D}_T)^{-1/2} \tilde{D}_T \hat{U}_N' y_t^N \tag{B.25}$$

where $\hat{S}_N = \hat{U}_N' \langle y_t^N, y_t^N \rangle \hat{U}_N$. The choice of the matrix square root determines the basis for the principal components space. In Lemma 5 this basis is chosen with knowledge of U_N. The identification normalization used in this section uses knowledge of c, the number of common trends. In the theorem the choice is taken care of by introducing the random matrix $H_{T,N}$. Thus in the proof we may use the basis choice implied by Lemma 5 w.r.o.g.

We get $\hat{F}_t = N^{-1/2} \hat{\Psi}_N^{-1} \tilde{D}_T \hat{U}_N' y_t^N$ and $\tilde{F}_t = N^{-1/2} \Psi_N^{-1} \tilde{D}_T U_N' \Lambda_N F_t$ such that ($\hat{\xi}_t^N = \hat{\Psi}_N^{-1} \tilde{D}_T \hat{U}_N' \xi_t^N$)

$$\begin{aligned}
\hat{F}_t - \tilde{F}_t &= N^{-1/2} (\hat{\Psi}_N^{-1} \tilde{D}_T \hat{U}_N' \Lambda_N F_t - \Psi_N^{-1} \tilde{D}_T U_N' \Lambda_N F_t + \hat{\xi}_t^N) \\
&= N^{-1/2} \left((\hat{\Psi}_N^{-1} - \Psi_N^{-1}) \tilde{D}_T \hat{U}_N' \Lambda_N F_t + \Psi_N^{-1} \tilde{D}_T (\hat{U}_N - U_N)' \Lambda_N F_t + \hat{\xi}_t^N \right) \\
&= (\hat{\Psi}_N^{-1} - \Psi_N^{-1}) \tilde{D}_T \hat{U}_N' U_N S_N^{1/2} F_t + \Psi_N^{-1} \tilde{D}_T (\hat{U}_N - U_N)' U_N S_N^{1/2} F_t + \frac{\hat{\xi}_t^N}{N^{1/2}} \\
&= \begin{pmatrix} O((\log T) T^{-1/2-\gamma}) & O((\log T) T^{1/2-\gamma}) \\ O((\log T) T^{-1/2-\gamma}) & O(Q_T) \end{pmatrix} S_N^{1/2} F_t + \frac{\hat{\xi}_t^N}{N^{1/2}}.
\end{aligned} \tag{B.26}$$

Thus, the integrated components of F_t are multiplied with terms of the order $(\log T) T^{-1/2-\gamma}$ which for γ close to 1 is close to -1.5. The stationary terms are multiplied with terms roughly of order $T^{-1/2}$. Therefore we obtain

$$
\begin{aligned}
\langle \hat{F}_t - \tilde{F}_t, \tilde{F}_{t-k}\rangle &= \begin{pmatrix} O((\log T)T^{-1/2-\gamma}) & O((\log T)T^{1/2-\gamma}) \\ O((\log T)T^{-1/2-\gamma}) & O(Q_T) \end{pmatrix} S_N^{1/2}\langle F_t, \tilde{F}_{t-k}\rangle \\
&\quad + \hat{\Psi}_N^{-1}\tilde{D}_T\hat{U}_N'\langle \xi_t^N/N^{1/2}, \tilde{F}_{t-k}\rangle \\
&= \begin{pmatrix} O((\log T)^2 T^{-\gamma}) & O((\log T)T^{1/2-\gamma}) \\ O(Q_T) & O(Q_T) \end{pmatrix} + O(Q_T) \qquad \text{(B.27)}
\end{aligned}
$$

From this it also follows that the difference $\langle \hat{F}_t, \hat{F}_{t-k}\rangle - \langle \tilde{F}_t, \tilde{F}_{t-k}\rangle$ is of the same order. This concludes the proof of the theorem since $Q_T = o((\log T)T^{1/2-\gamma})$ for $\gamma < 1$ and $O((\log T)^2 T^{-\gamma}) = O(T^{-\gamma+\epsilon})$ for small ϵ such that $0.5 < \gamma - \epsilon < 1$.

Appendix B.5. Proof of Theorem 5

The proof again proceeds in two steps: First we show consistency for $\tilde{F}_t$ as the data. Then we use the approximation results in Theorem 4 to deduce the result for $\hat{F}_t$.

First investigate

$$
\Delta\tilde{F}_t = \tilde{F}_t - \tilde{F}_{t-1} = \begin{pmatrix} \langle F_{t,c}, F_{t,c}\rangle^{-1/2} & 0 \\ 0 & I_{r-c} \end{pmatrix} (F_t - F_{t-1}) \qquad \text{(B.28)}
$$

such that

$$
\langle \Delta\tilde{F}_t, \Delta\tilde{F}_t\rangle^{1/2} = \begin{pmatrix} \langle F_{t,c}, F_{t,c}\rangle^{-1/2} & 0 \\ 0 & I_{r-c} \end{pmatrix} \langle \Delta F_t, \Delta F_t\rangle^{1/2} \qquad \text{(B.29)}
$$

due to the use of the lower triangular Cholesky factor as the matrix square root. It follows that

$$(I_f \otimes \langle \Delta\tilde{F}_t, \Delta\tilde{F}_t\rangle^{-1/2})\tilde{F}_t^+ = (I_f \otimes \langle \Delta F_t, \Delta F_t\rangle^{-1/2})F_t^+ = I_\otimes F_t^+ \qquad \text{(B.30)}$$

where the last equation defines $I_\otimes$. We also obtain

$$F_t^-(p) = (\mathcal{T}_W^-)^{-1}[I_{pr}, 0]\begin{pmatrix} F_{t-1,c} \\ W_t^-(p) \end{pmatrix}. \qquad \text{(B.31)}$$

Hence we may examine the regression of F_t^+ onto the regressor vector $Z_t := [F'_{t-1,c}, ([I_{pr-c}, 0]W_t^-(p))']'$. The matrix that is decomposed in CVA equals (redefining $\tilde{D}_T = \text{diag}(T^{-1/2}I_c, I_{pr-c})$)

$$\begin{aligned}\hat{A}_T &:= I_\otimes \langle F_t^+, Z_t\rangle\langle Z_t, Z_t\rangle^{-1}\langle Z_t, F_t^+\rangle I'_\otimes \\ &= I_\otimes \langle F_t^+, Z_t\rangle \tilde{D}_T(\tilde{D}_T\langle Z_t, Z_t\rangle\tilde{D}_T)^{-1}\tilde{D}_T\langle Z_t, F_t^+\rangle I'_\otimes. \end{aligned} \qquad \text{(B.32)}$$

Since $F_t^+ = \mathcal{O}_f x_t + \mathcal{U}_f E_t^+ = \mathcal{O}_{f,c}x_{t,c} + \mathcal{O}_{f,\bullet}x_{t,\bullet} + \mathcal{U}_f E_t^+$, where Z_t and E_t^+ are uncorrelated, it follows that for the limit only x_t projected onto Z_t is of interest. In x_t there are c integrated components, $n-c$ stationary components. The remaining eigenvalues in $\hat{A}_T$ correspond to limiting zero eigenvalues.

Since we can choose the state space basis freely (the result is formulated in terms of the transfer function not a specific realization), $\mathcal{O}_{f,c} = U_1$ where $U_1'U_1 = I_c$ is assumed. Furthermore $\mathcal{O}_{f,\bullet} = U_2 + U_1 M_\bullet$ such that

$$\begin{pmatrix} U_1' \\ U_2' \end{pmatrix}\mathcal{O}_f = \begin{pmatrix} I_c & M_\bullet \\ 0 & I_{n-c} \end{pmatrix}. \qquad \text{(B.33)}$$

It then follows that

$$\langle F_t^+, F_{t-1,c}\rangle = U_1\langle x_{t,c}, F_{t-1,c}\rangle + (U_2 + U_1 M_\bullet)\langle x_{t,\bullet}, F_{t-1,c}\rangle + \langle \mathcal{U}_f E_t^+, F_{t-1,c}\rangle. \tag{B.34}$$

Here the second and the third term are of order $O(\log T)$, while the first diverges such that $T^{-1}\langle x_{t,c}, F_{t-1,c}\rangle$ converges in distribution.

For

$$\langle F_t^+, W_t^-(p)\rangle = U_1\langle x_{t,c}, W_t^-(p)\rangle + (U_2 + U_1 M_\bullet)\langle x_{t,\bullet}, W_t^-(p)\rangle + \langle \mathcal{U}_f E_t^+, W_t^-(p)\rangle. \tag{B.35}$$

we have that the third term is of order $O(Q_T)$, the second converges a.s. and the first is of order $O(\log T)$. Therefore the matrix $\hat{A}_T$ fulfills the assumptions of Lemma 5. Consequently the representation for $\hat{U}_1 = U_1 + O(T^{-\gamma})$ and $\hat{U}_2 = U_2 + O(Q_T)$ hold. Then we obtain for the estimation of the state

$$\hat{x}_t = \begin{pmatrix} \hat{U}_1' \\ \hat{U}_2' \end{pmatrix} \langle F_t^+, Z_t\rangle\langle Z_t, Z_t\rangle^{-1} Z_t = \begin{pmatrix} \hat{U}_1' \\ \hat{U}_2' \end{pmatrix} (\mathcal{O}_f \mathcal{K}_p + \langle N_t^+, Z_t\rangle\langle Z_t, Z_t\rangle^{-1}) Z_t. \tag{B.36}$$

The state corresponding to this choice of the state basis equals (for p large enough, which is ascertained by the assumptions)

$$x_t = \begin{pmatrix} U_1' \\ U_2' \end{pmatrix} \mathcal{O}_f \mathcal{K}_p Z_t + o(T^{-1}). \tag{B.37}$$

Here

$$\begin{pmatrix} \hat{U}_1' - U_1' \\ \hat{U}_2' - U_2' \end{pmatrix} \mathcal{O}_f = \begin{pmatrix} \hat{U}_1' - U_1' \\ \hat{U}_2' - U_2' \end{pmatrix} [U_1, U_2 + U_1 M_\bullet]$$

$$= \begin{pmatrix} 0 & (\hat{U}_1' - U_1')U_2 \\ (\hat{U}_2' - U_2')U_1 & (\hat{U}_2' - U_2')U_1 M_\bullet \end{pmatrix} \tag{B.38}$$

According to the representation in Lemma 5 all blocks are of order $O(T^{-\gamma})$. Then consider the proof of Lemma A6 of Bauer and Buschmeier (2021) showing consistency for CVA in the square non-singular case. There the following terms are seen as relevant:

$$\langle \varepsilon_t, \hat{x}_t - x_t \rangle = O(\frac{(\log T)^a p}{T}) \quad , \quad \tilde{D}_T \langle Z_t, \hat{x}_t - x_t \rangle = O(\frac{p(\log T)^a}{T^{1/2}}), \tag{B.39}$$

$$\tilde{D}_T \langle Z_{t+1}, \hat{x}_t - x_t \rangle = O(\frac{p(\log T)^a}{T^{1/2}}) \quad , \quad \tilde{D}_x \langle x_t, \hat{x}_t - x_t \rangle = O(\frac{p(\log T)^a}{T^{1/2}}), \tag{B.40}$$

$$\langle \hat{x}_t - x_t, \hat{x}_t - x_t \rangle = O(\frac{p(\log T)^a}{T^{1/2}}) \quad . \tag{B.41}$$

Now $\hat{x}_t - x_t$ has two components: The first one equals $\mathcal{K}_p Z_t$ multiplied with a matrix of order $O(T^{-\gamma})$. For this term all results hold replacing T^{-1} by $T^{-\gamma}$ and $T^{-1/2}$ by $T^{1/2-\gamma}$.

The second one equals a matrix times $\langle N_t^+, Z_t \rangle \langle Z_t, Z_t \rangle^{-1} Z_t$. All evaluations to show the results for this term are straightforward and hence omitted.

With these orders of convergence the consistency is shown in Lemma A.6 of Bauer and Buschmeier (2021). This shows consistency of the order given in the theorem for the normalized static factors $\tilde{F}_t$ as data.

For the estimated factors we obtained the result

$$\sup_{0 \le k \le H_T} \| \langle \tilde{F}_t, \tilde{F}_{t-k} \rangle - \langle \hat{F}_t, \hat{F}_{t-k} \rangle \| = O(T^{1/2-\gamma}). \tag{B.42}$$

Consequently

$$\langle \Delta\hat{F}_t, \Delta\hat{F}_t\rangle - \langle \Delta\tilde{F}_t, \Delta\tilde{F}_t\rangle = O(T^{1/2-\gamma}). \tag{B.43}$$

Recall that $\hat{F}_t = \hat{S}_N^{-1/2}\hat{U}_N' y_t^N / N^{1/2}$ and let $\check{F}_t = S_N^{-1/2}\hat{U}_N' y_t^N / N^{1/2}$. It follows that

$$\langle \Delta\hat{F}_t, \Delta\hat{F}_t\rangle = \hat{S}_N^{-1/2} S_N^{1/2} \langle \Delta\check{F}_t, \Delta\check{F}_t\rangle S_N^{T/2} \hat{S}_N^{-T/2} \tag{B.44}$$

such that

$$\langle \Delta\hat{F}_t, \Delta\hat{F}_t\rangle^{-1/2} = \langle \Delta\check{F}_t, \Delta\check{F}_t\rangle^{-1/2} S_N^{-1/2} \hat{S}_N^{1/2}. \tag{B.45}$$

From this we get

$$\langle \Delta\hat{F}_t, \Delta\hat{F}_t\rangle^{-1/2} \hat{F}_t = \langle \Delta\check{F}_t, \Delta\check{F}_t\rangle^{-1/2} \check{F}_t. \tag{B.46}$$

Clearly $\langle \Delta\check{F}_t, \Delta\check{F}_t\rangle^{-1/2} - \langle \Delta F_t, \Delta F_t\rangle^{-1/2} = O(Q_T)$, where both are non-singular a.s. for large enough T.

Therefore, in order to analyze the CVA estimates we may use $\check{F}_t$ in place of F_t. Recall that

$$\check{F}_t = S_N^{-1/2}\hat{U}_N' y_t^N / N^{1/2} = S_N^{-1/2}\hat{U}_N' U_N S_N^{1/2} F_t + S_N^{-1/2}\hat{U}_N' \xi_t^N / N^{1/2} \tag{B.47}$$

where $\hat{U}_N' U_N = I_r + O(T^{-\gamma})$ and $S_N^{1/2}$ is upper triangular. It follows that the properties of $\check{F}_t$ equal the ones of F_t insofar as the first c components are $I(1)$ essentially equal to $F_{t,c}$. The remaining $r - c$ components are stationary plus $F_{t,c}$ multiplied with the matrix $\hat{U}_{N,\bullet}' U_{N,c} = O(T^{-\gamma})$. Consequently the order of convergence for the cross terms changes according

$\langle \check{F}_{t,c}, \check{F}_{t,\bullet} \rangle = O(T^{1-\gamma})$, all other terms remain of the same order. This shows that for the arguments in Bauer and Buschmeier (2021) the term $\log T$ needs to be replaced by $T^{1-\gamma}$ in each occurrence. Since γ can be chosen arbitrarily close to 1 the change is minor.

It follows that the same transformations $\mathcal{T}_W^+$ and $\mathcal{T}_W^-$ can be used for $\check{F}_t$ as for F_t in order to (approximately) separate stationary and non-stationary directions.

The rest of the proof then follows exactly the same lines as above, where only the error bounds for matrices like $\langle W_t^+, W_t^+ \rangle$ are required. This shows the result.