

Using CVA for the Estimation of Approximate Dynamic Factor Models

Berlin Müggelheim,
20.3.2025

Dank an die DFG für die Finanzierung des Projekts GLASS No. 469278259.

Dietmar Bauer

Introduction

- Often in the analysis of, for example, macro-data or financial data one models a large number of variables jointly.
- Examples include the vectorization of matrix valued time series (MaTS), wherein a number of variables is observed in a number of regions.
- The number of variables typically is smaller but similar to the number of time points.
- To mention just one example, Bai and Ng (2019) use an example with $N = 128$ variables and $T = 676$ (monthly) observations from the FRED-MD data base.
- Joint models for such number of series necessarily contain many parameters and pose problems for specification and estimation.
- In order to cope with the high dimensionality, factor models and generalisations thereof are considered to reduce the dimensionality.

Motivation

- Estimation of such models often involves dynamic modelling of the principal components (PC), see the recent survey of Barigozzi et al. (2024).
- VAR modelling for PCs is straightforward, if the factor process is non-singular (see Stock and Watson, 2011, e.g.).
- This includes integrated settings (e.g., Bai and Ng, 2004).
- State space models have also been used: Doz et al. (2011) discuss a two step approach for estimation using an EM type algorithm.
- Subspace methods have been applied to model combined dynamics of factors and idiosyncraties (Kapetanios and Marcellino, 2009).
- Empirically part of the literature argues for singularity of the factor processes (fewer shocks than components), which complicates estimation.
- Manfred Deistler and others developed structure theory and first ideas on estimation in the singular situation (see Lippi, Deistler, Anderson, 2023).

Question investigated in this paper: **How robust are estimates obtained using the CVA type subspace method with respect to singularity and (co-)integration?**

Outline

- **Factor models (FM)** reduce the dimensionality by focusing on *common* features that appear in many time series first.
- **Dynamics** of the factors can be modelled using state space models.
- State space models can be **estimated** using different methods.
- **Subspace** procedures like CVA are one alternative.
- **Consistency results** confirm that they provide useful information in the aDFM situation: stationary and $I(1)$ case.
- Appendix: (**Simulations**: specification of integer valued parameters.)

This is a methodological talk, I will not show real world data or modelling!

Disclaimer: This talk is based on a talk given at the CFE in Dec. 2023 and reuses many of the slides.

Specification of Factor Models

Approximate Dynamic Factor Models: aDFMs

$$y_{it} = \chi_{it} + \xi_{it} = \underbrace{\lambda_i' F_t}_{\text{factor component}} + \underbrace{\xi_{it}}_{\text{idiosyncratic component}}, i = 1, \dots, N.$$

$$Y_t^N = \Lambda_N F_t + \Xi_t^N \in \mathbb{R}^N$$

- $Y_t^N = [y_{1t}, \dots, y_{Nt}] \in \mathbb{R}^N, t = 1, \dots, T$
- $F_t \in \mathbb{R}^r$... static factors, where $r \ll N$.

Assumption (Independence)

The factors F_t and the idiosyncratic component ξ_{is} are independent for all variables i and all times t, s .

Assumption (Factor Loadings)

The factor loadings λ_i are assumed deterministic such that

$$N^{-1} \sum_{i=1}^N \lambda_i \lambda_i' = N^{-1} \Lambda_N' \Lambda_N \rightarrow M_\Lambda > 0.$$

$$\sup_N \max_i \|\lambda_i\| \leq M_\lambda.$$

Stationarity

- In this talk we first restrict attention to stationary processes.
- This may involve the need to transform some variables for example via taking temporal differences. Integrated processes dealt with at the end.

Assumption (Stationarity)

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.

For each of the processes F_t, Ξ_t^N, y_t^N we have

$$\max_{0 \leq k \leq H_T} \max_{i,j} \left\| T^{-1} \sum_{t=1+K}^T x_{t,i} x_{t-k,j} - \mathbb{E} x_{t,i} x_{t-k,j} \right\| = O(Q_T)$$

where $Q_T := \sqrt{(\log \log T / T)}$ and $H_T = (\log T)^a$ for some integer $a > 1$.

These high level assumptions differ from the literature. Often assumptions on the underlying processes are stated such that the covariance estimates fulfill similar assumptions.

Covariance of observations:

$$\Gamma_{y,N} = \mathbb{E} y_t^N (y_t^N)' = \Lambda_N (\mathbb{E} F_t F_t') \Lambda_N' + \Gamma_{\Xi,N}.$$

Two issues for identification:

- (I) separate common and idiosyncratic parts
- (II) identify common factors and loadings from the product $\Lambda_N F_t$.

Assumption (Identification)

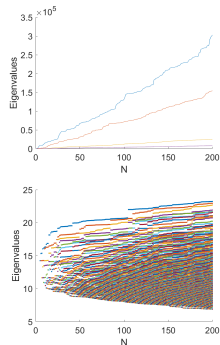
- $\Gamma_{\Xi,N} = \mathbb{E} \Xi_t^N (\Xi_t^N)'$: $\sup_N \lambda_{\max}(\Gamma_{\Xi,N}) \leq M_{\Xi}$ ('weak correlation').
- $\Gamma_F = \mathbb{E} F_t F_t' = I_r$.
- Λ_N is positive lower triangular (for example, its heading $r \times r$ submatrix is lower triangular matrix with positive diagonal entries).

Other conditions are possible, such as a submatrix of Λ_N being the identity matrix.

REMARK: Normalizations using $\Lambda_N' \Lambda_N / N = I_r$ or diagonal do not share the 'nesting' property that Λ_{N_0} is a submatrix of Λ_N for $N_0 < N$.

It follows that

- $\mathbb{E} \chi_t^N (\chi_t^N)' = \Lambda_N (\mathbb{E} F_t F_t') \Lambda_N' = \Lambda_N \Lambda_N'$
- $\Lambda_N' \Lambda_N / N \rightarrow M_\Lambda > 0$: all eigenvalues of $\Lambda_N \Lambda_N'$ grow essentially linearly as a function of N .
- In $\Gamma_{y,N} = \Lambda_N \Lambda_N' + \Gamma_{\varepsilon,N}$ the first r eigenvalues are asymptotically (in N) proportional to N .
- The remaining $N - r$ ones remain bounded.
- This suggest to estimate F_t using PCA.



Asymptotic identification in the sense of Chamberlain and Rothschild (1983) \Rightarrow approximate DFM (aDFM).

State Space Modelling

Modelling Dynamics

- Factors typically are not uncorrelated in time.
- In the literature often autoregressive processes are used:

$$F_t = A_1 F_{t-1} + A_2 F_{t-2} + \dots + A_p F_{t-p} + v_t,$$
$$a(z)F_t = v_t, \quad a(z) = I - A_1 z - A_2 z^2 - \dots - A_p z^p.$$

- More general: state space system in innovation form:

$$y_t = Cx_t + v_t$$
$$x_{t+1} = Ax_t + Bv_t$$

- If the matrix A is stable, stationary solutions are given by

$$F_t = v_t + \sum_{j=0}^{\infty} CA^j B v_{t-j-1} = k(L)v_t.$$

Assumption (Rationality; cf. Lippi, Deistler, Anderson (2023))

The static factor process $(F_t)_{t \in \mathbb{Z}}$, $F_t \in \mathbb{R}^r$, does not depend on N and has a minimal state space representation as

$$F_t = Cx_t + Du_t, \quad x_{t+1} = Ax_t + Bu_t, \quad u_t \in \mathbb{R}^q.$$

where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times q}$, $C \in \mathbb{R}^{r \times n}$, $D \in \mathbb{R}^{r \times q}$ and where $(u_t)_{t \in \mathbb{Z}}$ is a stationary, ergodic martingale difference sequence and has expectation zero and variance matrix $\Omega = I_q$.

The transfer function $k(z) = D + zC(I_n - zA)^{-1}B$ has no zeros* or poles inside the unit circle.

- $(u_t)_{t \in \mathbb{Z}}$: white noise process of *dynamic factors*.
- If $q < r$ we get a 'tall' transfer function. In that case the zeros are defined differently from the square case. See Anderson and Deistler (2008).*
- Lippi, Deistler, Anderson (2023) use different (but related) state space representation.

Assumption (factor dynamics)

$k(z) = D + zC(I_n - zA)^{-1}B$ where

- The matrix $D = k(0) \in \mathbb{R}^{r \times q}$ has full column rank q such that the heading $q \times q$ submatrix is lower triangular with positive entries on the diagonal.
- There exists a pseudo inverse D^\dagger such that $\underline{A} = A - BD^\dagger C$ is stable. Then the transfer function $c(z) = \sum_{j=0}^{\infty} c_j z^j \in \mathbb{R}^{q \times r}$, $c_j = C \underline{A}^{j-1} B D^\dagger$ is a pseudo left-inverse such that $c(z)k(z) = I_q$.
- There exists a real value $\rho_0 < 1$ such that $\|c_j\| \leq \rho_0^j \mu, \forall j \in \mathbb{N}$ for $0 < \mu < \infty$.
- If $c(z)$ is a polynomial $\rho_0 = 0$. The degree of $c(z)$ then is denoted as p_0 .

- We only assume the existence, not the uniqueness.
- Du_t are the innovations for the process $(F_t)_{t \in \mathbb{Z}}$.
- Identification uses recursive definition: $[I_q, 0]D$ is the lower triangular Cholesky factor of the innovations in the first q static factors, which in turn multiply the loadings Λ_N whose heading submatrix also is assumed to be lower triangular. This aids interpretation, but requires knowledge.


Estimation of State Space Models

Autoregressive Static Factor Processes F_t

Estimation based on $\hat{F}_t = \hat{\Lambda}_N^\dagger y_t^N$:

- In the stationary non-singular case $q = r$: OLS is consistent (under our assumptions below).
- For fast enough growing cross section ($T/N^2 \rightarrow 0$) inference remains the same (same asymptotic distribution of coefficient estimators) as if the true factors would be known.
- Similar results hold for the (co-)integrated case for $q = r$.
- For $q < r$: OLS has problems due to near multi-collinearity of regressors.
- Regularization helps, but introduces unnecessary biases in the estimated coefficients.
- Deistler et al. (2010) provide structure theory to select the relevant regressors. This requires the estimation of integer valued structural parameters.

State Space Processes

- Parameterization of state space systems is more complex than for VARs: identifiability issues due to latency of states.
- Different parameterization is needed for stationary, cointegrated as well as singular ($q < r$) cases.
- Once parameterization is obtained: Gaussian likelihood can be formulated as a function of parameters, when modelling the idiosyncracics as white noise with known variance matrix (misspecification).
- Jungbacker and Koopman (2015) show that Kalman filter only is required for 'small' dimensional factor process F_t , not y_t^N 
- Doz et al. (2011) propose to use an EM algorithm instead of Gauss-Newton type procedures.
- Asymptotic properties of the misspecified qMLE are not known in all cases. Most results focus on $q = r$ and VAR models.

Irrespective of usage of EM or qMLE, a good initialization algorithm is needed!

Canonical Variate Analysis (CVA)

5. Canonical Variate Analysis (CVA)



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

⇒ estimation would be simple, if the state was known!

CVA is based on two facts:

1. The state x_t can be approximated by past observations:

$$x_t(p) = \sum_{j=1}^p \mathcal{K}_j(p) F_{t-j} = \mathcal{K}_p F_t^{-} \xrightarrow{p \rightarrow \infty} x_t$$

2. Predictions of F_{t+h} , $h \geq 0$, based on the past of F_t are a function of the state:

$$F_{t+h} = \underbrace{CA^h x_t}_{F_{t+h|t-1}} + \underbrace{Du_{t+h} + \sum_{j=0}^{h-1} CA^j Bu_{t+h-j-1}}_{v_{t+h|t-1}}$$

This holds for $h = 0, 1, \dots, f-1$ and can be seen as a multi-step long VAR approximation.

Jointly this implies using $x_t \approx x_t(p) = \mathcal{K}_p F_t^-$

$$\underbrace{\begin{pmatrix} F_t \\ \vdots \\ F_{t+f-1} \end{pmatrix}}_{F_t^+} = \mathcal{O}_f \mathcal{K}_p \underbrace{\begin{pmatrix} F_{t-1} \\ \vdots \\ F_{t-p} \end{pmatrix}}_{F_t^-} + V_t(p), \quad (*)$$

$$\mathcal{O}_f = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{f-1} \end{pmatrix}, \mathcal{K}_p = (\mathcal{K}_1(p) \quad \mathcal{K}_2(p) \quad \dots \quad \mathcal{K}_p(p)),$$

$$V_t(p) = \begin{pmatrix} v_{t|t-1} \\ \vdots \\ v_{t+f-1|t-1} \end{pmatrix} + \mathcal{O}_f(x_t - x_t(p)).$$

(*) is a regression equation, where the matrix $\mathcal{O}_f \mathcal{K}_p$ has low rank $n \ll fr$.

CVA algorithm in the potentially singular (tall TF) case

1. Choose f, p .
2. Perform a rank restricted regression (RRR) of F_t^+ onto F_t^- . In this step the order n needs to be specified.
3. Use the estimate \hat{K}_p from the last step to estimate the state $\hat{x}_t(p) = \hat{K}_p F_t^-$.
4. Estimate C by regressing F_t onto $\hat{x}_t(p)$. This step provides $\hat{v}_t = F_t - \hat{C}\hat{x}_t(p)$.
5. Obtain \hat{D} from a truncated SVD of

$$\hat{\Sigma}_T = \langle \hat{v}_t, \hat{v}_t \rangle = \hat{D}\hat{D}' + \hat{R}_q, \quad (\langle a_t, b_t \rangle = T^{-1} \sum_{t=p+1}^T a_t b_t')$$

using the q largest singular values, where \hat{D} is p.l.t.

6. Regress $\hat{x}_{t+1}(p)$ onto $\hat{x}_t(p)$ and $\hat{u}_t = \hat{D}^\dagger \hat{v}_t$ to obtain the estimates \hat{A} and \hat{B} .
7. Convert $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ to an overlapping form, for example echelon forms.

CVA has been proposed by W. Larimore (1983).

$$F_t^+ = \mathcal{H}_{f,p} F_t^- + V_t(p)$$

- Unrestricted estimate: $\hat{\mathcal{H}}_{f,p} = \langle F_t^+, F_t^- \rangle \langle F_t^-, F_t^- \rangle^{-1}$.
- To obtain a rank n estimator we can use the SVD of:

$$W_f \langle F_t^+, F_t^- \rangle \langle F_t^-, F_t^- \rangle^{-1} \langle F_t^-, F_t^+ \rangle W_f' = \hat{U} \hat{S} \hat{U}' = \hat{U}_n \hat{S}_n \hat{U}_n' + \hat{R}_n.$$

- Different weights correspond to different estimators:
 - OLS: $W_f = I_{ff}$
 - qMLE: $W_f = \langle F_T^+, F_T^+ \rangle^{-1/2}$.
- Consistency for the estimator typically holds as long as W_f is non-singular.
- The rank of $\mathcal{H}_{f,p}$ is seen in the singular values estimated as the diagonal entries of \hat{S} .

Issues

- F_t is not known \Rightarrow use the largest r principal components \hat{F}_t such that $\langle \hat{F}_t, \hat{F}_t \rangle = I_r$ and $\hat{\Lambda}_N$ is p.l.t. (or other normalization).
- Classical CVA uses $W_f = \langle \hat{F}_t^+, \hat{F}_t^+ \rangle^{-1/2}$ to achieve optimal variance (see Bauer, 2005).

Special features for $q < r$ (singular case) and p large:

- variance of F_t^- has rank $n + pq < rp \Rightarrow$ regularisation is needed:

$$\begin{aligned}\hat{U}\hat{S}\hat{U}' &= W_f \langle \hat{F}_t^+, \hat{F}_t^- \rangle \langle \hat{F}_t^-, \hat{F}_t^- \rangle^\dagger \langle \hat{F}_t^-, \hat{F}_t^+ \rangle W_f' \\ &= \hat{U}_n \hat{S}_n \hat{U}_n' + \hat{R}_n \Rightarrow \hat{K}_p = \hat{U}_n' W_f^\dagger \langle \hat{F}_t^+, \hat{F}_t^- \rangle \langle \hat{F}_t^-, \hat{F}_t^- \rangle^\dagger\end{aligned}$$

- $\langle \hat{F}_t^-, \hat{F}_t^- \rangle^\dagger$: eigenvalues smaller than $\epsilon > 0$ are replaced by ϵ .
- Typical values: $\epsilon = 10^{-6}$.
- Projection $\hat{x}_t(p) = \hat{K}_p \hat{F}_t^-$: exists and can be calculated (unbiasedly) also for (near) singular \hat{F}_t^- .

Issues

Case $q < r$:

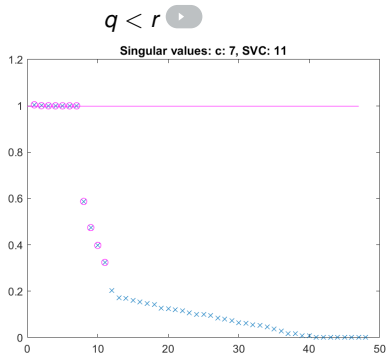
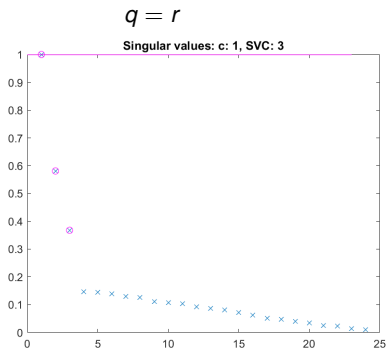
- Typical CVA weighting W_f will be asymptotically singular: If $L'_{f,\perp} F_t^+ = 0$ then $L'_{f,\perp} \hat{F}_t^+$ contains terms $\hat{\Lambda}_N^+ \Xi_t^N$ of variance $O(1/N)$.
- Canonical correlations do not depend on scaling. Hence the canonical correlations between these two pairs of r.v.s are identical:

$$(A) \quad \begin{pmatrix} y_t \\ w_t/\sqrt{N} \end{pmatrix} \text{ and } \begin{pmatrix} x_t \\ z_t/\sqrt{N} \end{pmatrix}, \quad (B) \quad \begin{pmatrix} y_t \\ w_t \end{pmatrix} \text{ and } \begin{pmatrix} x_t \\ z_t \end{pmatrix}$$

- CVA weighting amplifies for $q < r$ asymptotically singular directions of \hat{F}_t^+ and $\hat{F}_t^- \Rightarrow$ may be problematic.

Deciding on the order of the system

- Singular values include information on order n .
- Differences between $q = r$ and $q < r$.
- For $q = r$: use SVC.
- For $q < r$: count singular values close to 1 (unclear, how exactly)!



Asymptotic Results in the Stationary Case

Extraction of the static factors F_t

- First step: static PCA on $Y_t^N \in \mathbb{R}^N$ using $\hat{\Sigma}_{N,T} = (NT)^{-1} \sum_{t=1}^T Y_t^N (Y_t^N)'$.

$$\hat{\Sigma}_{N,T} = \frac{\Lambda_N}{\sqrt{N}} \hat{\Gamma}_F \frac{\Lambda_N'}{\sqrt{N}} + \frac{\hat{\Gamma}_{\Xi}}{N} + \text{cross terms} = \hat{U}_{N,r} \hat{S}_{N,r} \hat{U}_{N,r}' + \hat{R}_N$$

- We get $\hat{\Lambda}_N = \hat{U}_{N,r} \hat{S}_{N,r}^{1/2} \hat{L}_N$ (\hat{L}_N is introduced to fulfill the identification restrictions)
- Then use $(\hat{\Lambda}_N^\dagger = (\hat{\Lambda}_N' \hat{\Lambda}_N)^{-1} \hat{\Lambda}_N')$

$$\hat{F}_t = \hat{\Lambda}_N^\dagger Y_t^N = \underbrace{(\hat{\Lambda}_N^\dagger \Lambda_N)}_{\Delta_T} F_t + \hat{\Lambda}_N^\dagger \Xi_t^N.$$

- We obtain

$$T^{-1} \sum_{t=1}^T (\hat{F}_t \hat{F}_t' - \Delta_T F_t F_t' \Delta_T') = O(Q_T + 1/N).$$

- Furthermore $\|\hat{\Lambda}_N/\sqrt{N} - \Lambda_N/\sqrt{N}\| = O(Q_T + 1/N).$

Under the assumption $T/N^2 \rightarrow 0$ we have $O(Q_T + 1/N) = O(Q_T)$.

Theorem (Consistency)

- Let the data be generated according to the before mentioned Assumptions.
- Assume that $T/N^2 \rightarrow 0, T \rightarrow \infty$.
- Let $p(T) \leq H_T, f \geq n, p = p(T) \geq -\frac{e \log T}{2 \log \rho_0}, e > 1$, for $\rho_0 > 0$ and $p \geq \rho_0$ for $\rho_0 = 0$.
- Further let $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ denote the CVA estimators based on the PCA estimates \hat{F}_t for given order n using $W_f = I_f$ and appropriate regularization converted to an appropriate overlapping form.

Then

$$\max\{\|\hat{A} - A\|, \|\hat{B} - B\|, \|\hat{C} - C\|, \|\hat{D} - D\|\} = O(Q_T).$$

Hence the transfer function is estimated consistently.

For $q = r$ the same result holds using the CVA weighting $W_f = \langle \hat{F}_t^+, \hat{F}_t^+ \rangle^{-1/2}$.

The $I(1)$ case

Co-integrated state space processes

- In state space processes, integration occurs for unit eigenvalues of A .
- A canonical form that stresses this has the form (Bauer and Wagner (2012)):

$$F_t = (C_1 \quad C_\bullet) x_t + Du_t,$$

$$\begin{pmatrix} x_{t+1,1} \\ x_{t+1,\bullet} \end{pmatrix} = \begin{pmatrix} I_c & 0 \\ 0 & A_\bullet \end{pmatrix} x_t + \begin{pmatrix} B_1 \\ B_\bullet \end{pmatrix} u_t.$$

- This generates c common trends $(x_{t,1})_{t \in \mathbb{Z}}$, $x_{t,1} = \sum_{j=1}^{t-1} B_1 u_j + x_{0,1}$.
- The idiosyncratic terms are kept stationary (this is a restriction, that can be dealt with by letting the cross sectional dimension increase faster).

Assumption (Identification, I(1) case)

- There exists a selector matrix $\tilde{I}_r \in \mathbb{R}^{N_0 \times r}$ such that $[\tilde{I}_r', 0] \Lambda_N = I_r$ for $N \geq N_0$.
- D is p.l.t.

Assumptions on the common factors

Assumption (factor dynamics, I(1))

$$F_t = \underbrace{(C_1 \quad C_\bullet)}_C x_t + D u_t, x_{t+1} = \underbrace{\begin{pmatrix} I_c & 0 \\ 0 & A_\bullet \end{pmatrix}}_A x_t + \underbrace{\begin{pmatrix} B_1 \\ B_\bullet \end{pmatrix}}_B u_t$$

where

- $(u_t)_{t \in \mathbb{Z}}$ fulfills the noise assumptions in (rationality),
- $W_t = F_t - C_1 C_1' F_{t-1}$ fulfills the assumptions (Stationarity),
- A_\bullet is stable, $(A_\bullet, B_\bullet, C_\bullet)$ is in echelon overlapping form,
- there exists a stable left pseudo-inverse to $k(z) = D + zC(I_n - zA)^{-1}B$,
- finally

$$\sup_{N \in \mathbb{N}} \max_{i=1, \dots, N} \|\langle C_1' F_t, \xi_{it} \rangle\| = O(\log T). \quad (1)$$

PCA with integrated variables

- Consider a bivariate process $F_{t,1} = F_{t-1,1} + u_{t,1}, F_{t,2} = u_{t,2}$.
- Scaled empirical covariance matrix:

$$\frac{1}{T^2} \sum_{t=1}^T F_t F_t' \xrightarrow{d} \begin{pmatrix} Z & 0 \\ 0 & 0 \end{pmatrix}.$$

- First PC: proportional to $F_{t,1}$, but with $\langle \hat{F}_{t,1}, \hat{F}_{t,1} \rangle = 1$.
- Achieved by $\hat{F}_{t,1} \approx F_{t,1} / \sqrt{\langle F_{t,1}, F_{t,1} \rangle}$.
- Randomly weighted common trend. Weights are of order $O(1/\sqrt{T})$.
- Second PC: approx. $F_{t,2}$

PCA with integrated variables in aDFM

- Principal components are obtained from the sample covariance matrix:

$$\frac{1}{NT} \sum_{t=1}^T y_t^N (y_t^N)' = \frac{\Lambda_N}{\sqrt{N}} \left(\frac{1}{T} \sum_{t=1}^T F_t F_t' \right) \frac{\Lambda_N'}{\sqrt{N}} + \underbrace{\frac{1}{NT} \sum_{t=1}^T \Xi_t^N (\Xi_t^N)'}_{O(1/N)} + \text{cross terms}$$

- For co-integrated process F_t in $\frac{1}{T} \sum_{t=1}^T F_t F_t'$ the c common trends lead to c eigenvalues tending to infinity as $O(T)$. $r - c$ tend to their finite limits.
- The first c principal components then essentially are due to the common trends.
- Typically one uses a normalisation such that $\langle \hat{F}_t, \hat{F}_t \rangle = I_r$.
- For common trends this implies a weighting with dominant term $\langle F_{t,1}, F_{t,1} \rangle^{-1/2}$ which is of order $O(1/\sqrt{T})$.

Applying CVA then uses regression $\hat{F}_t^+ = \mathcal{O}_f \mathcal{K}_p \hat{F}_T^- + V_t(p)$.

Sidestep: Asymptotics in a bivariate setting

- Consider a univariate stationary process z_t and an $I(1)$ process $y_t = y_{t-1} + \alpha z_t + \varepsilon_t$.
- Estimating the regression $y_t = \rho y_{t-1} + \alpha z_{t-1} + \varepsilon_t$ we can consistently estimate ρ and α .
- Replacing y_t with $\tilde{y}_t = y_t/\sqrt{T}$ we get the equation:
$$\tilde{y}_t = \rho \tilde{y}_{t-1} + \frac{\alpha}{\sqrt{T}} z_{t-1} + \varepsilon_t/\sqrt{T}.$$
- This leaves the OLS estimator for ρ consistent, but $\langle \tilde{y}_t, z_{t-1} \rangle \langle z_{t-1}, z_{t-1} \rangle^{-1} \rightarrow 0$.
- Thus in the rescaled equation the information on α diminishes.

CVA for I(1) processes

$$\begin{pmatrix} F_{t,c} \\ F_{t,\bullet} \end{pmatrix} = \begin{pmatrix} I_c & C_{1,\bullet} \\ 0 & C_\bullet \end{pmatrix} \begin{pmatrix} x_{t,c} \\ x_{t,\bullet} \end{pmatrix} + \begin{pmatrix} D_1 \\ D_\bullet \end{pmatrix} u_t,$$

$$\begin{pmatrix} x_{t+1,c} \\ x_{t,\bullet} \end{pmatrix} = \begin{pmatrix} I_c & 0 \\ 0 & A_\bullet \end{pmatrix} \begin{pmatrix} x_{t,c} \\ x_{t,\bullet} \end{pmatrix} + \begin{pmatrix} B_1 \\ B_\bullet \end{pmatrix} u_t,$$

$$\begin{pmatrix} F_{t,1} \\ F_{t,\bullet} \\ \Delta F_{t+1,1} \\ F_{t+1,\bullet} \\ \vdots \\ \Delta F_{t+f-1,1} \\ F_{t+f-1,\bullet} \end{pmatrix} = \begin{pmatrix} I_c & C_{1,\bullet} \\ 0 & C_\bullet \\ 0 & C_{1,\bullet}(A_\bullet - I) \\ 0 & C_\bullet A_\bullet \\ \vdots & \vdots \\ 0 & C_{1,\bullet}(A_\bullet - I)A_\bullet^{f-2} \\ 0 & C_\bullet A_\bullet^{f-1} \end{pmatrix} \begin{pmatrix} x_{t,c} \\ x_{t,\bullet} \end{pmatrix} + \tilde{N}_t^+$$

- Thus $x_{t,\bullet}$ may be identified from the second block rows.
- But the information may be treatly reduced.

CVA for I(1) processes

$$\begin{pmatrix} F_{t,c}/\sqrt{T} \\ F_{t,\bullet} \end{pmatrix} = \begin{pmatrix} I_c & C_{1,\bullet}/\sqrt{T} \\ 0 & C_{\bullet} \end{pmatrix} \begin{pmatrix} x_{t,c}/\sqrt{T} \\ x_{t,\bullet} \end{pmatrix} + \begin{pmatrix} D_1/\sqrt{T} \\ D_{\bullet} \end{pmatrix} u_t,$$

$$\begin{pmatrix} x_{t+1,c}/\sqrt{T} \\ x_{t,\bullet} \end{pmatrix} = \begin{pmatrix} I_c & 0 \\ 0 & A_{\bullet} \end{pmatrix} \begin{pmatrix} x_{t,c}/\sqrt{T} \\ x_{t,\bullet} \end{pmatrix} + \begin{pmatrix} B_1/\sqrt{T} \\ B_{\bullet} \end{pmatrix} u_t,$$

$$\begin{pmatrix} F_{t,1}/\sqrt{T} \\ F_{t,\bullet} \\ \hline \Delta F_{t+1,1}/\sqrt{T} \\ F_{t+1,\bullet} \\ \hline \vdots \\ \hline \Delta F_{t+f-1,1}/\sqrt{T} \\ F_{t+f-1,\bullet} \end{pmatrix} = \begin{pmatrix} I_c & C_{1,\bullet}/\sqrt{T} \\ 0 & C_{\bullet} \\ \hline 0 & C_{1,\bullet}(A_{\bullet} - I)/\sqrt{T} \\ 0 & C_{\bullet}A_{\bullet} \\ \hline \vdots & \vdots \\ \hline 0 & C_{1,\bullet}(A_{\bullet} - I)A_{\bullet}^{f-2}/\sqrt{T} \\ 0 & C_{\bullet}A_{\bullet}^{f-1} \end{pmatrix} \begin{pmatrix} x_{t,c}/\sqrt{T} \\ x_{t,\bullet} \end{pmatrix} + \tilde{N}_t^+$$

- Thus $x_{t,\bullet}$ may be identified from the second block rows.
- But the information may be treatly reduced.

How to avoid dropping $C_{1,\bullet}$?

1. Use the CVA weights $\langle \hat{F}_t^+, \hat{F}_t^+ \rangle^{-1/2}$: works for $q = r$, does not work for $q < r$.
2. Use a different normalization that allows $\langle \hat{F}_t, \hat{F}_t \rangle$ to grow with T : $\tilde{I}'_N \hat{\Lambda}_N = I_r$ for the calculation of the static factors.
3. Define a weight that counteracts the $\langle F_{t,1}, F_{t,1} \rangle^{-1/2}$ term:

$$\begin{aligned}\hat{F}_t &= \hat{\Lambda}_N^\dagger y_t^N = \hat{\Lambda}_N^\dagger \Lambda_N F_t + \delta F_t, \\ I_r &= \langle \hat{F}_t, \hat{F}_t \rangle \approx \hat{\Lambda}_N^\dagger \Lambda_N \langle F_t, F_t \rangle (\hat{\Lambda}_N^\dagger \Lambda_N)' \Rightarrow \hat{\Lambda}_N^\dagger \Lambda_N \approx \langle F_t, F_t \rangle^{-1/2}, \\ \hat{F}_t - \hat{F}_{t-1} &= \hat{\Lambda}_N^\dagger \Lambda_N \Delta F_t + \Delta \delta F_t \\ \Rightarrow \langle \Delta \hat{F}_t, \Delta \hat{F}_t \rangle^{1/2} &\approx \hat{\Lambda}_N^\dagger \Lambda_N \langle \Delta F_t, \Delta F_t \rangle^{1/2}.\end{aligned}$$

Theorem (Consistency, $I(1)$ case)

Let the data be generated according to the Assumptions above. Let $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ be calculated using CVA with

- $f \geq n_O$, the observability index,
- $p = p(T) \geq -(1 + \delta) \log T / (\log \rho_o) \rightarrow \infty$, $p = O(H_T)$ where $\delta > 0$, $H_T = (\log T)^a$ for $\rho_o > 0$ and $p > p_o$ else (where p_o denotes the lag length of an autoregressive pseudo left-inverse of b_{red}),
- $W_f = (I_f \otimes \langle \Delta \hat{F}_t, \Delta \hat{F}_t \rangle^{-1/2})$ or for $r = q$ alternatively $W_f = \langle \hat{F}_t^+, \hat{F}_t^+ \rangle^{-1/2}$.

Then for all $0.5 < \gamma < 1$ for $T \rightarrow \infty$, $T/N^2 \rightarrow 0$

$$\max\{\|\hat{A} - A\|, \|\hat{B} - B\|, \|\hat{C} - C\|, \|\hat{D} - D\|\} = O(T^{1/2-\gamma}).$$

Conclusions

- CVA can be used to obtain consistent estimates in the aDFM setting.
- The main assumption beside the aDFM model structure is that $T/N^2 \rightarrow 0$ such that the cross-sectional dimension grows fast enough compared to the time dimension.
- The procedure is simple being based on static PCA and regression methods.
- The method is robust with respect to the relation between q and r (provided appropriate regularization is used) and with respect to integration in the static factors (for appropriate choice of the weights).
- The weighting matrix W_f^+ should be adapted compared to the classical case, if $q < r$ is possible.
- The method provides information on all required integer parameters, which in all cases considered leads to consistent estimation.

Current research

- CVA does not directly lead to inference, as no asymptotic distributions are available in all cases (hypothesis tests).
- Misspecified qMLE based on Gaussian likelihood modelling the idiosyncratic terms as white noise terms may help in this respect.
- Similar to above for $r = q$ the Gaussian likelihood for known $\hat{\Lambda}_N$ leads to the lower dimensional likelihood for modelling $(\hat{F}_t)_{t \in \mathbb{Z}}$.
- This may lead to inferential procedures in the stationary and in the $I(1)$ case.

- Anderson, B.D.O., Deistler, M. (2008). Properties of Zero-Free Transfer Function Matrices. *SICE Journal of Control, Measurement, and System Integration* 1 (4). 284–292.726
- Bai, J.; Ng, S. (2002): Determining the Number of Factors in Approximate Factor Models. In: *Econometrica* 70 (1), S. 191–221.
- Bai, J.; Ng, S. (2019): Rank regularized estimation of approximate factor models. In: *Journal of Econometrics* 212 (1), S. 78–96.
- Barigozzi, M., Hallin M., Luciani, M., Zaffaroni, P. (2024). Inferential theory for generalized dynamic factor models, *Journal of Econometrics*, 239, 2, 105422.
- Bauer, D. (2005): Estimating linear dynamical systems using subspace methods. In: *Econ. Theory* 21 (01). DOI: 10.1017/S0266466605050127.
- Bauer, D., & Wagner, M. (2012). A state space canonical form for unit root processes. *Econometric Theory*, 28(6), 1313-1349.
- 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 mean-variance analysis on large asset markets. *Econometrica* 51 (5), 1281–1304.
- Deistler, M.; Anderson, B. D.O.; Filler, A.; Zinner, Ch.; Chen, W. (2010): Generalized Linear Dynamic Factor Models: An Approach via Singular Autoregressions. In: *European Journal of Control* 16 (3), S. 211–224. DOI: 10.3166/ejc.16.211-224.
- 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.
- Jungbacker, B., Koopman, S. J. (2015). Likelihood-based dynamic factor analysis for measurement and forecasting. *The Econometrics Journal*, C1-C21.
- Kapetanios, G.; Marcellino, M. (2009). A parametric estimation method for dynamic factor models of large dimensions. *Journal of Time Series Analysis*, 30(2), 208-238.
- Larimore, W. E. (1983). System identification, reduced-order filtering and modeling via canonical variate analysis. In 1983 American Control Conference (pp. 445-451). IEEE.
- Lippi, M.; Deistler, M.; Anderson, B.D.O. (2023): High-Dimensional Dynamic Factor Models: A Selective Survey and Lines of Future Research. In: *Econometrics and Statistics* 26, S. 3–16. DOI: 10.1016/j.ecosta.2022.03.008.
- Stock, J. H., Watson, M. W. (2011) Dynamic Factor Models. In M. Clements, and D. F. Hendry (eds), *The Oxford Handbook of Economic Forecasting*, Oxford Handbooks.



Thank you for your attention!

Tall transfer functions

$k(z) = D + zC(I_n - zA)^{-1}B \in \mathbb{C}^{r \times q}$ where $q < r$ typically: 'tall'.

- Dimension reduction: $Y_t^N \in \mathbb{R}^N \rightarrow F_t \in \mathbb{R}^r \rightarrow u_t \in \mathbb{R}^q$.
- Deistler et al. (2010): generically (in a specific sense) there exists polynomial left pseudo-inverses $k^\dagger(z)$ to $k(z)$.
- In these cases we obtain a *singular* autoregressive representation

$$k^\dagger(L)F_t = v_t = \begin{pmatrix} I \\ 0 \end{pmatrix} u_t.$$
- From this we see that u_t can be written as a linear combination of a finite number of past F_t 's.
- The pseudo-inverse $k^\dagger(z)$ is not unique. Not even the lag order. Not even the system for smallest lag order. This complicates estimation.

Definition of zeros of tall transfer functions (Anderson, Deistler, 2008) :

The rank of the matrix $M(z) := \begin{pmatrix} zI - A & -B \\ C & D \end{pmatrix}$ falls below its usual rank at z_0 .

Consequences of singularity

From state space system we get:

$$\begin{pmatrix} F_t \\ F_{t+1} \\ \vdots \\ F_{t+f-1} \end{pmatrix} = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{f-1} \end{pmatrix} x_t + \begin{pmatrix} D & 0 & \dots & 0 \\ CB & D & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ CA^{f-2}B & \dots & CB & D \end{pmatrix} \begin{pmatrix} u_t \\ u_{t+1} \\ \vdots \\ u_{t+f-1} \end{pmatrix},$$

$$F_t^+ = \mathcal{O}_f x_t + \mathcal{U}_f U_t^+ = \underbrace{(\mathcal{O}_f \quad \mathcal{U}_f)}_{L_f} \begin{pmatrix} x_t \\ U_t^+ \end{pmatrix} \quad (L_f \in \mathbb{R}^{fr \times (n+fq)})$$

$$= \mathcal{O}_f \mathcal{K}_p F_t^- + N_{t,f,p}$$

- Rank of variance of F_t^+ is at most $\min(fr, n + fq)$.
- F_t^+ is singular for $q < r$ and so is F_t^- for f, p large enough.

Consequences of singularity (II)

- If L_f has full column rank, there exists a (non-unique) left pseudo-inverse L_f^\dagger :

$$L_f^\dagger F_t^+ = L_f^\dagger L_f \begin{pmatrix} x_t \\ U_t^+ \end{pmatrix} = \begin{pmatrix} x_t \\ U_t^+ \end{pmatrix}.$$

- Then (I): $[I_n, 0] L_f^\dagger F_t^+ = x_t$.

- Also (II)

$$x_{t+f} = A^f x_t + \begin{pmatrix} A^{f-1}B & A^{f-2}B & \dots & B \end{pmatrix} U_t^+ = \tilde{\mathcal{K}}_f F_t^+.$$

- Hence from (I) and (II): $x_t = \check{\mathcal{K}}_f F_t^- = \mathcal{C}_f F_t^+.$
- In this case we obtain n canonical correlations between F_t^+ and F_t^- that are equal to 1, cf. Breitung and Pigorsch (2013).

Two different state space representations ◀

$$F_t = Cx_t + Du_t, x_{t+1} = Ax_t + Bu_t,$$

$$\tilde{x}_t = \begin{pmatrix} x_t \\ u_t \end{pmatrix} : \quad \tilde{x}_{t+1} = \begin{pmatrix} A & B \\ 0 & 0 \end{pmatrix} \tilde{x}_t + \begin{pmatrix} 0 \\ I \end{pmatrix} u_{t+1}$$

$$F_t = \underbrace{[C, D]}_{\tilde{C}} \tilde{x}_t + \mathbf{0}, \tilde{x}_{t+1} = \tilde{A}\tilde{x}_t + \tilde{B}u_{t+1}.$$

Gaussian likelihood in aDFM case

$$Y_t^N = \Lambda_N F_t + \Xi_t^N = U_N \sqrt{N} F_t + \Xi_t^N = \sqrt{N} (U_N F_t + \Xi_t^N / \sqrt{N})$$

- where Ξ_t^N is modelled as white noise with variance $\sigma^2 I_N$. Normalization $\Lambda_N = U_N \sqrt{N}$ is used where $U_N' U_N = I_r$.
- $\mathbb{E} Y_t^N = 0$, $\text{Var}(Y_t^N) = \Lambda_N \text{Var}(F_t) \Lambda_N' + \sigma^2 I_N$.
- Let $U = [U_N, U_\perp]$ be an orthonormal matrix:

$$\text{Var}(U' Y_t^N / \sqrt{N}) = \begin{pmatrix} \text{Var}(F_t) + \frac{\sigma^2}{N} I_r & 0 \\ 0 & \frac{\sigma^2}{N} I_{N-r} \end{pmatrix}.$$

- Thus the Gaussian likelihood for $Y_{T,N} = [(y_1^N)', \dots, (y_T^N)']' / \sqrt{N}$ can be decomposed into two separate components:
 - the Gaussian likelihood for the process $(F_t)_{t \in \mathbb{Z}}$ of dimension r plus white noise with variance $\frac{\sigma^2}{N}$.
 - the Gaussian likelihood for the process $U_\perp' \Xi_t^N / \sqrt{N}$ of variance $U_\perp' U_\perp \frac{\sigma^2}{N}$.
- The second component can be written as $U_\perp U_\perp' \Xi_t^N = (I - U_N U_N') \Xi_t^N = (I - U_N U_N') y_t^N$. Hence U_\perp does not need to be calculated.

Regularization in regression with noise

$$y_t + \tilde{w}_t = \beta_x x_t + \beta_z \tilde{z}_t + u_t, \tilde{w}_t = \frac{1}{N} w_t, \tilde{z}_t = \frac{1}{N} z_t.$$

- Assume that $X'Z = 0$.

- Without regularization:

$$\begin{aligned} (\hat{\beta}_x \quad \hat{\beta}_z) &= ((Y'X + W'X/N)(X'X)^{-1} \quad (Y'Z + W'Z/N)(Z'Z/N^2)^{-1}) \\ \hat{Y} &= (Y'X + W'X/N)(X'X)^{-1}X' + (Y'Z + W'Z/N)(Z'Z)^{-1}Z'/N \\ &= Y'X(X'X)^{-1}X' + Y'Z(Z'Z)^{-1}Z' + \frac{1}{N}(W'X(X'X)^{-1}X' + W'Z(Z'Z)^{-1}Z'). \end{aligned}$$

- With regularization: replace $Z'Z/N^2$ by $Z'Z/m$:

$$\begin{aligned} (\tilde{\beta}_x \quad \tilde{\beta}_z) &= ((Y'X + W'X/N)(X'X)^{-1} \quad (Y'Z + W'Z/N)(Z'Z/m)^{-1}) \\ \tilde{Y} &= (Y'X + W'X/N)(X'X)^{-1}X' + (Y'Z + W'Z/N)(Z'Z)^{-1}Z'/N \\ &= Y'X(X'X)^{-1}X' + Y'Z(Z'Z)^{-1}Z' \frac{m}{N} \\ &\quad + \frac{1}{N}(W'X(X'X)^{-1}X' + \frac{m}{N}W'Z(Z'Z)^{-1}Z'). \end{aligned}$$

Specification of Integers

Illustration

Simulation system:

- $N = 200, T = 800$
- $r = 5, q = 2, n = 3$.
- Factor dynamics: $A = \text{diag}(0.8, -0.8, 0.4)$, all other matrices chosen randomly.
- Idiosyncratic terms: each individual series follows an AR(1) with randomly chosen $|\rho_i| \leq 0.7, i = 1, \dots, N$ and noise with variance $\sigma^2 = 0.25$.
- $M = 1000$ replications.

Dimension of the static factor r

- There are many methods to choose r .
- To mention just one example, Bai and Ng (2002) suggest to use an information type criterion to select the number of static factors:

$$\widehat{IC}_2(k) = \log SSR_k + k \frac{N+T}{NT} \log(NT/(N+T)), \quad SSR_k = \|Y - \hat{\Lambda}_{N,k} \hat{F}'_k\|_{Fr}^2$$

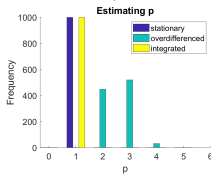
- This penalizes the fit (in terms of variation not explained by the common component) against complexity (modelled as linear in number of factors).
- \hat{r} minimizes the criterion.

f, p

- f needs to be large enough such that the observability matrix \mathcal{O}_f is of full rank.
- In the aDFM case, a small value typically is sufficient, as $r \geq n$ is reasonable, where hence $f = 1$ would suffice.
- p should be chosen such that the approximation of x_t by $x_t(p)$ is accurate.
- This is related to the lag selection in a long VAR representation for \hat{F}_t .
- F_t for $q < r$ is a singular process, the information criterion needs to be adapted.
- A simple fix is: $\hat{\Sigma}_T(p)$ denoting the innovation variance estimate for a lag p AR approximation of F_t :

$$IC(p; C_T) = \underbrace{\text{tr}}_{\log \det} \left[\hat{\Sigma}_T(p) \right] + \frac{2r^2 p C_T}{T}.$$

$T = 800, N = 200$



T	200	400	800	1600
stat.	1.00	1.00	1.00	1.00
overdiff.	1.27	2.00	2.58	4.16
integr.	1.00	1.00	1.00	1.00

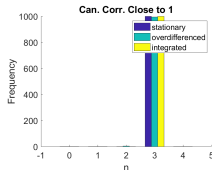
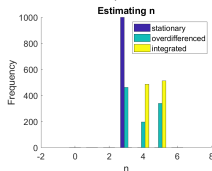
System order n

- Within CVA: singular values in the RRR step used for estimation.
- Selection can be done using

$$SVC(n) = \sum_{j=1}^n \log \hat{\sigma}_j^2 + \frac{2rnC_T}{T}$$

- C_T sufficiently large: consistent estimation of n for $r = q$ based on F_T .
- If $N^2/T \rightarrow \infty$: sampling error $O(Q_T)$ dominates. Thus for $C_T/T \rightarrow 0$, $C_T/(fp \log \log T) \rightarrow \infty$ sufficient for consistency for $r = q$.
- $q < r$: Breitung and Pigorsch (2013) show, the canonical correlations are close to one.

$T = 800, N = 200$



$N = 50$, percent $\hat{n} = n$

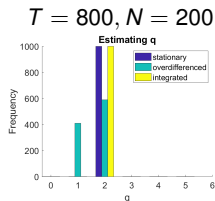
T	200	400	800	1600
stat.	1.00	1.00	1.00	0.95
overdiff.	0.00	0.00	0.05	0.84
integr.	0.00	0.05	0.45	0.99

Dynamic factor dimension q

- Typically this would be determined based on dynamic PCA.
- An alternative is to use the estimated innovation variance $\hat{\Sigma}_T \in \mathbb{R}^{r \times r}$.
- The largest q eigenvalues will tend to their non-zero limits, the remaining ones tending to zero.
- Consequently also here information criteria can be used:

$$IC(q; C_T) = \sum_{j=1}^q \mu_j(\hat{\Sigma}_T) + \frac{rqC_T}{T}.$$

where $\mu_j(\hat{\Sigma}_T)$ denotes the j -th largest eigenvalue.



diff., percent $\hat{q} = q$

T	200	400	800	1600
$N = 50$	0.00	0.02	0.18	1.00
$N = 100$	0.00	0.00	0.46	1.00
$N = 150$	0.00	0.01	0.56	1.00
$N = 200$	0.00	0.00	0.59	1.00

Estimation Accuracy

- Accuracy measured as

$$\|\hat{D}\hat{D}' - DD'\| + \sum_{j=0}^{10} \|\hat{C}\hat{A}^j\hat{K}(\hat{C}\hat{A}^j\hat{K})' - CA^jK(CA^jK)'\|$$

- Averaged over 1000 replications for $N = 50$.
- Convergence is clearly visible.
- Overdifferenced case leads to worst results.

