Robust modeling of large dimensional time series with cluster structure

Pedro Galeano

Departamento de Estadística and UC3M-Santander Big Data Institute

Universidad Carlos III de Madrid

pedro.galeano@uc3m.es

Joint work with Andrés M. Alonso (UC3M) and Daniel Peña (UC3M)

Financial support from Agencia Estatal de Investigación
PID2019-108311GBI00/AEI/10.13039/501100011033

New Bridges between Mathematics and Data Science

uc3m | Universidad Carlos III de Madrid

Introduction

- The data science environment:
  - Main characteristic: There are abundant, accessible and low-cost data.
  - In particular: Many problems in many areas including finance, genomic, climatology, environment, etc. involve the analysis of large sets of time series.
  - Common feature: Large dimensional time series are heterogeneous, including the presence of outliers and clusters.
  - Therefore: There is a need for methods that allow modeling time series with these characteristics.
  - Question: What kind of models can we use?

Introduction

- Dynamic Factor Models (DFMs):
  - What are DFMs?: General representations of large sets of time series with very weak assumptions on the data generating process.
  - Idea: The common evolution of the time series is driven by a few latent dynamic factors.
  - Some references: Engle and Watson (1981), Peña and Box (1987), Stock and Watson (1988, 2002), Forni et al. (2000, 2015, 2017), Bai and Ng (2002), Hallin and Lippi (2013), and Gao and Tsay (2019, 2021), among many others.
  - However: The standard DFMs do not consider the presence of clusters and the inference does not usually consider the presence of outliers.

Introduction

- Dynamic Factor Models with Cluster Structure (DFMCSs):
  - What are DFMCSs?: DFMs that take into account that the dynamic evolution of the time series may be affected by global and specific (cluster-dependent) factors.
  - Methods to fit DFMCSs:
    - Wang (2010) and Hallin and Liška (2011) proposed methods for identifying and estimating global and specific factors assuming that the clusters are known.
    - Ando and Bai (2017) proposed a method with the same purposes for financial returns but assuming that the clusters are unknown.
  - Previous proposals: Do not take into account the presence of outliers.
  - In this work: We propose a robust procedure to fit DFMCSs.
  - The proposed method: It is more general, more robust and has better performance than the procedure proposed by Ando and Bai (2017).

Dynamic factor model with cluster structure

- DFMCS:
  - General expression:

$$x_t = P_0 f_{0t} + \sum_{i=1}^{k} P_i f_{it} + n_t = c_t + n_t$$

  - ★ $x_t = (x_{1t}, \ldots, x_{mt})'$ is a vector of m stationary time series with mean 0_m.
  - ★ $f_{0t} = (f_{01t}, \ldots, f_{0r_0t})'$ is a vector of r_0 global factors with mean 0_{r_0}.
  - ★ $P_0 = \left[P'_{0,1}|\cdots|P'_{0,k}\right]'$ is the loading matrix for the r_0 global factors.
  - ★ k is the number of clusters.
  - ★ $f_{it} = \left(f_{i1t}, \ldots, f_{ir_it}\right)'$ is a vector of r_i specific factors in the i-th cluster with mean 0_{r_i}.
  - ★ $P_i = \left[0'_{i,1}|\cdots|P'_{i,i}|\cdots|0'_{i,k}\right]'$ is the loading matrix for the r_i specific factors in the i-th cluster.
  - ★ $n_t = (n_{1t}, \ldots, n_{mt})'$ is a vector of m idiosyncratic noises with mean 0_m and such that $E\left[c_t n'_{t-h}\right] = 0_{m\times m}$, for $h = 0, \pm1, \pm2, \ldots$
  - ★ c_t are called the common components.

Dynamic factor model with cluster structure

- Some remarks:
  - The series are ordered: Wlog, the first m_1 series correspond to the first cluster and the last m_k to the last cluster, such that $\sum_{i=1}^{k} m_i = m$.
  - DFMCS written as a DFM:

$$x_t = Pf_t + n_t = c_t + n_t$$

    - $f_t = \left(f_{0t}', f_{1t}', \ldots, f_{kt}'\right)'$ is the vector of $r = \sum_{j=0}^{k} r_j$ global and specific factors.
    - $P = [P_0|P_1|\cdots|P_k]$ is the loading matrix (with blocks of zeros) for the r global and specific factors.
  - Identifiability restrictions on the loading matrices:
    1. $P_0'P_0 = I_{r_0}$.
    2. $P_i'P_i = I_{r_i}$, for $i = 1, \ldots, k$.
    3. $P_0'P_i = 0_{r_0 \times r_i}$.
    4. $P_i'P_j = 0_{r_i \times r_j}$, for $i \neq j \in \{1, \ldots, k\}$.

Dynamic factor model with cluster structure

- Unknown objects to estimate when fitting the DFMCS:
  1. Number of clusters: k.
  2. Clusters: The subset of time series in each cluster.
  3. Number of global and specific factors in each cluster: r_0 and $r_1, \ldots, r_k$, respectively.
  4. Loading matrices for global and specific factors: P_0 and $P_1, \ldots, P_k$, respectively.
  5. Global and specific factors: f_{0t} and $f_{1t}, \ldots, f_{kt}$, respectively.

Robust procedure to fit a DFMCS

- Step 1 - Clean the observed set of time series:
  - Idea: Use the procedure proposed in Galeano, Peña and Tsay (2022) for detecting outliers in DFMs.
  - The procedure: Detects multivariate additive outliers (MAOs) and multivariate level shifts (MLSs) as well as univariate additive outliers and level shifts.
  - Main idea of the procedure: AOs and LSs in the dynamic factors generates MAOs and MLSs in the observed time series, while AOs in the idiosyncractic noise generates AOs in the observed time series.
  - Then: Detect and clean AOs and LSs in the dynamic factors and AOs in the idiosyncractic noise.
  - Output after cleaning: $x_1^*, \ldots, x_T^*$.

Robust procedure to fit a DFMCS

- Step 2 - Initial estimation of dynamic factors (global and/or specific):
  1. Idea: Fit a standard DFM.
  2. Estimate P: $\widehat{P} = \left[\widehat{P}_1 | \cdots | \widehat{P}_{r_c}\right]$:
     - $\widehat{P}_1, \ldots, \widehat{P}_{r_c}$ are the eigenvectors linked to the r_c largest eigenvalues of $S_x = \frac{1}{T}\sum_{t=1}^{T} x_t^* x_t^{*\prime}$, i.e., the sample covariance matrix of $x_1^*, \ldots, x_T^*$.
  3. Determine r_c: Use the test proposed by Ahn and Horestein (2013).
  4. Estimate f_t: $\widehat{f}_t = \widehat{P}' x_t^*$.
  5. Estimate c_t: $\widehat{c}_t = \widehat{P}\widehat{f}_t$.

Robust procedure to fit a DFMCS

- Step 3 - Clustering:
  - Find the clusters: Use the clustering algorithm proposed by Alonso and Peña (2019) with $\widehat{c}_t$.
  - Algorithm: Hierarchical clustering where the dissimilarity between $\widehat{c}_{it}$ and $\widehat{c}_{jt}$ is given by:

$$d(\widehat{c}_{it}, \widehat{c}_{jt}) = \left(\frac{|R_{ij,p}|}{|R_{ii,p}|\,|R_{jj,p}|} \right)^{1/(p+1)}$$

    1. $R_{ij,p}$ is the sample correlation matrix of $\left(\widehat{c}_{it}, \ldots, \widehat{c}_{it-p}, \widehat{c}_{jt}, \ldots, \widehat{c}_{jt-p}\right)'$.
    2. $R_{ii,p}$ and $R_{jj,p}$ are the sample correlation matrices of $\left(\widehat{c}_{it}, \ldots, \widehat{c}_{it-p}\right)'$ and $\left(\widehat{c}_{jt}, \ldots, \widehat{c}_{jt-p}\right)'$, respectively.
  - $d(\widehat{c}_{it}, \widehat{c}_{jt})$: Equal to 0 if one of the common components is a linear combination of its past and the values of the other common component and it is positive otherwise.
  - Number of clusters, k: Using the Silhouette algorithm proposed by Rousseeuw (1987).

Robust procedure to fit a DFMCS

- Step 4 - Factor classification:
  - Estimate factors in the clusters: Proceed as in step 2. to obtain r_i^g factors, denoted by $\widehat{f}_{i1t}, \ldots, \widehat{f}_{ir_i^g t}$, in cluster i.
  - Compare initial and new estimated factors: For each $j = 1, \ldots, r_c$, compute the first canonical correlations between $\widehat{f}_{jt}$ and $\widehat{f}_{i1t}, \ldots, \widehat{f}_{ir_i^g t}$, for $i = 1, \ldots, k$.
  - Classify each $\widehat{f}_{jt}$ as global or specific:
    1. If several canonical correlations for $\widehat{f}_{jt}$ are larger than a threshold (e.g. 0.9), then $\widehat{f}_{jt}$ is a global factor.
    2. If only one of the canonical correlations for $\widehat{f}_{jt}$ is larger than the threshold value, then $\widehat{f}_{jt}$ is a specific factor in the corresponding cluster.
    3. If all the canonical correlations for $\widehat{f}_{jt}$ are smaller than the threshold, then $\widehat{f}_{jt}$ is a global factor.
  - Add additional specific factors: Label as specific factors those in $\widehat{f}_{i1t}, \ldots, \widehat{f}_{ir_i^g t}$, for $i = 1, \ldots, k$, not correlated with the global factors.

Robust procedure to fit a DFMCS

- Step 5 - Re-estimate specific factors and check that the clusters obtained are due to different specific factors:
  - Compute: $\widehat{v}_t = x_t^* - \widehat{P}_0 \widehat{f}_{0t}$:
    - $\widehat{f}_{0t}$ is the vector of estimated global factors obtained in step 4.
    - $\widehat{P}_0$ is the loading matrix corresponding to these factors.
  - Re-estimate specific factors: As in step 2. but using $\widehat{v}_t$.
  - Check that the clusters obtained are due to different specific factors:
    1. All the k clusters found include at least one specific factor, and we conclude that we have a DFMCS with k clusters.
    2. k_1 clusters ($1 \leq k_1 < k$) contain specific factors, and $k_2 = k - k_1$ clusters only contain global factors, then we have a DFMCS with $k_1 + 1$ clusters, i.e., there is one cluster with no specific factors.
    3. All the clusters only contain global factors, then we have the standard DFM.

Simulations

- Data Generating Processes (DGPs):
  - Number of clusters: $k = 2$.
  - Number of global and specific factors: $r_0 = 2$, $r_1 = 1$ and $r_2 = 3$.
  - Number of series in each cluster: $m_1 = m/3$ and $m_2 = 2m/3$, where $m = 300$ and $m = 600$.
  - Sample size: $T = 200$ and $T = 400$.
  - Dynamic (global and specific) factors: AR(1) with $\phi = 0.75$.
  - Idiosyncratic noise: i.i.d. (DGP1), heteroscedastic with cross-sectional dependence (DGP2) and serial and cross-sectional dependence (DGP3).
  - Signal to noise ratio: Large, medium and small.
  - Outliers: Four MAOs at locations $a = \left[\frac{T}{5}\right]$, $\left[\frac{2T}{5}\right]$, $\left[\frac{3T}{5}\right]$, and $\left[\frac{4T}{5}\right]$, and two MLSs at locations $l = \left[\frac{T}{3}\right]$ and $\left[\frac{2T}{3}\right]$.
  - Number of generated data set: 100.

Simulations

- Mean number of clusters selected (true $k = 2$):

		Large SNR			Medium SNR			Small SNR		
T	m	DGP1	DGP2	DGP3	DGP1	DGP2	DGP3	DGP1	DGP2	DGP3
200	300	2.01	2.00	2.02	2.00	2.02	2.04	1.24	1.25	1.28
400	300	2.01	2.01	2.00	2.00	2.00	2.00	2.00	2.03	2.05
200	600	2.00	2.00	2.00	2.00	2.00	2.01	2.00	2.01	2.03
400	600	2.00	2.00	2.00	2.00	2.00	2.00	2.01	2.00	2.00

Simulations

- Mean number of factors selected (true $r_0 = 2$, $r_1 = 1$ and $r_2 = 3$):

			Large SNR			Medium SNR			Small SNR		
Factor	T	m	DGP1	DGP2	DGP3	DGP1	DGP2	DGP3	DGP1	DGP2	DGP3
Global	200	300	1.98	2.01	2.00	1.81	1.89	1.67	1.01	1.01	0.46
Spec. 1			0.98	1.00	0.98	0.98	0.98	0.96	0.94	0.96	0.91
Spec. 2			3.02	2.99	3.01	3.06	3.02	3.02	3.07	3.00	3.19
Global	400	300	2.00	2.00	2.00	2.00	2.00	1.95	1.89	1.89	1.47
Spec. 1			1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.98	0.95
Spec. 2			3.00	3.00	3.00	3.00	3.00	3.02	2.99	3.03	3.09
Global	200	600	2.00	2.00	2.00	1.94	2.00	1.79	1.84	1.79	1.35
Spec. 1			1.00	1.00	1.00	0.98	1.00	0.95	0.97	0.96	0.88
Spec. 2			3.00	3.00	3.00	3.06	3.00	3.12	3.06	3.09	3.23
Global	400	600	2.00	2.00	2.00	2.00	2.00	2.00	1.97	2.00	1.90
Spec. 1			1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.99
Spec. 2			3.00	3.00	3.00	3.00	3.00	3.00	3.03	3.00	3.03

Simulations

- Loading estimates evaluation with the discrepancy measure in Gao and Tsay (2019) (the closer to 0, the better the estimation):

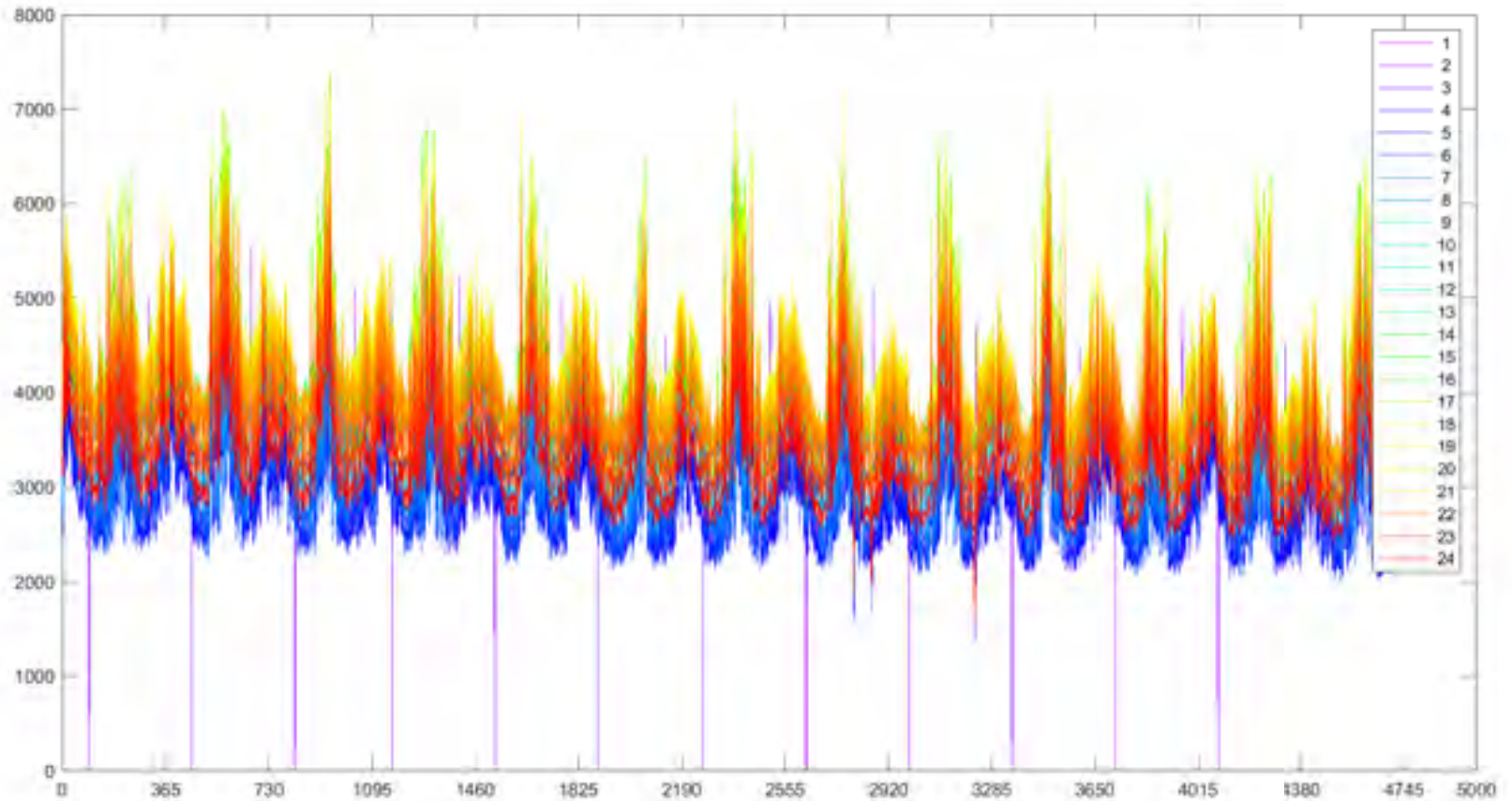
				Large SNR			Medium SNR			Small SNR		
Method	Factor	T	m	DGP1	DGP2	DGP3	DGP1	DGP2	DGP3	DGP1	DGP2	DGP3
A&B	Glob.	200	300	.871	.855	.877	.883	.861	.874	.894	.898	.894
	Spec. 1			.392	.325	.406	.449	.333	.415	.542	.428	.489
	Spec. 2			.598	.556	.667	.683	.624	.627	.635	.706	.689
A,G&P	Glob.			.303	.265	.320	.361	.310	.408	.433	.355	.490
	Spec. 1			.165	.138	.187	.213	.171	.253	.256	.230	.291
	Spec. 2			.145	.126	.149	.171	.146	.193	.191	.161	.209
A&B	Glob.	400	300	.879	.873	.869	.863	.877	.891	.893	.874	.894
	Spec. 1			.310	.355	.314	.398	.298	.372	.722	.299	.403
	Spec. 2			.595	.551	.612	.576	.624	.696	.721	.609	.644
A,G&P	Glob.			.255	.229	.267	.289	.258	.323	.335	.293	.372
	Spec. 1			.141	.116	.144	.162	.131	.178	.192	.154	.225
	Spec. 2			.119	.110	.126	.133	.123	.154	.155	.137	.169
A&B	Glob.	200	600	.870	.859	.863	.879	.869	.892	.883	.873	.898
	Spec. 1			.348	.334	.294	.422	.349	.468	.461	.324	.525
	Spec. 2			.562	.501	.578	.606	.548	.631	.652	.618	.677
A,G&P	Glob.			.261	.223	.288	.334	.275	.372	.392	.324	.451
	Spec. 1			.149	.121	.171	.196	.156	.226	.247	.189	.292
	Spec. 2			.119	.100	.132	.153	.121	.172	.181	.143	.207
A&B	Glob.	400	600	.862	.848	.867	.879	.871	.883	.882	.876	.885
	Spec. 1			.293	.257	.251	.357	.310	.379	.328	.314	.367
	Spec. 2			.495	.433	.529	.612	.564	.649	.634	.609	.658
A,G&P	Glob.			.206	.181	.226	.255	.214	.285	.296	.247	.335
	Spec. 1			.115	.098	.130	.145	.118	.164	.171	.140	.196
	Spec. 2			.095	.083	.104	.115	.094	.128	.136	.110	.151

Real data example: Electricity demands in New England

- Summary:
  - Data set: Hourly day-ahead demand for the ISO New England electricity market.
  - Dates: January 2004 to December 2016.
  - Eight load zones: Connecticut (CT), Maine (ME), New Hampshire (NH), Rhode Island (RI), Vermont (VT), Northeastern Massachusetts and Boston (NEMA), Southeastern Massachusetts (SEMA) and Western/Central Massachusetts (WCMA).
  - $D_{t,i}$: Demand of electricity in one of the eight regions at one of the 24 hours in a given day, that is, $1 \leq i \leq 192$ and $1 \leq t \leq 4749$.
  - $X_{t,i} = \nabla_7 \log D_{t,i}$: The series require a weekly seasonal difference and a logarithm transformation to become stationary.

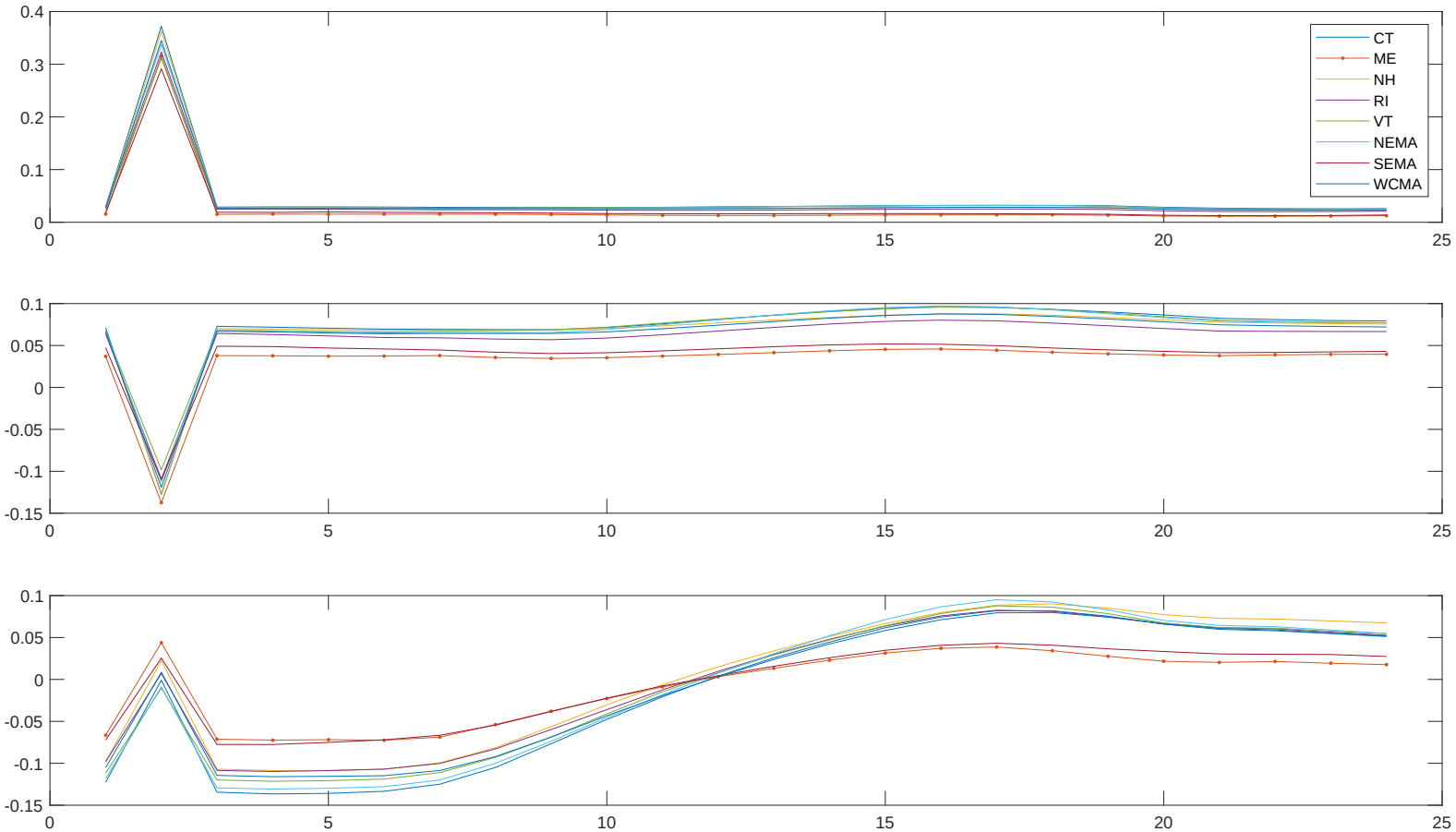
Real data example: Electricity demands in New England

- Demands at 01:00 – 24:00 for Connecticut (January 2004 – December 2016):



Real data example: Electricity demands in New England

- Estimated loadings for a standard DFM: Three factors highly affected by the time change at 2:00 AM:



Real data example: Electricity demands in New England

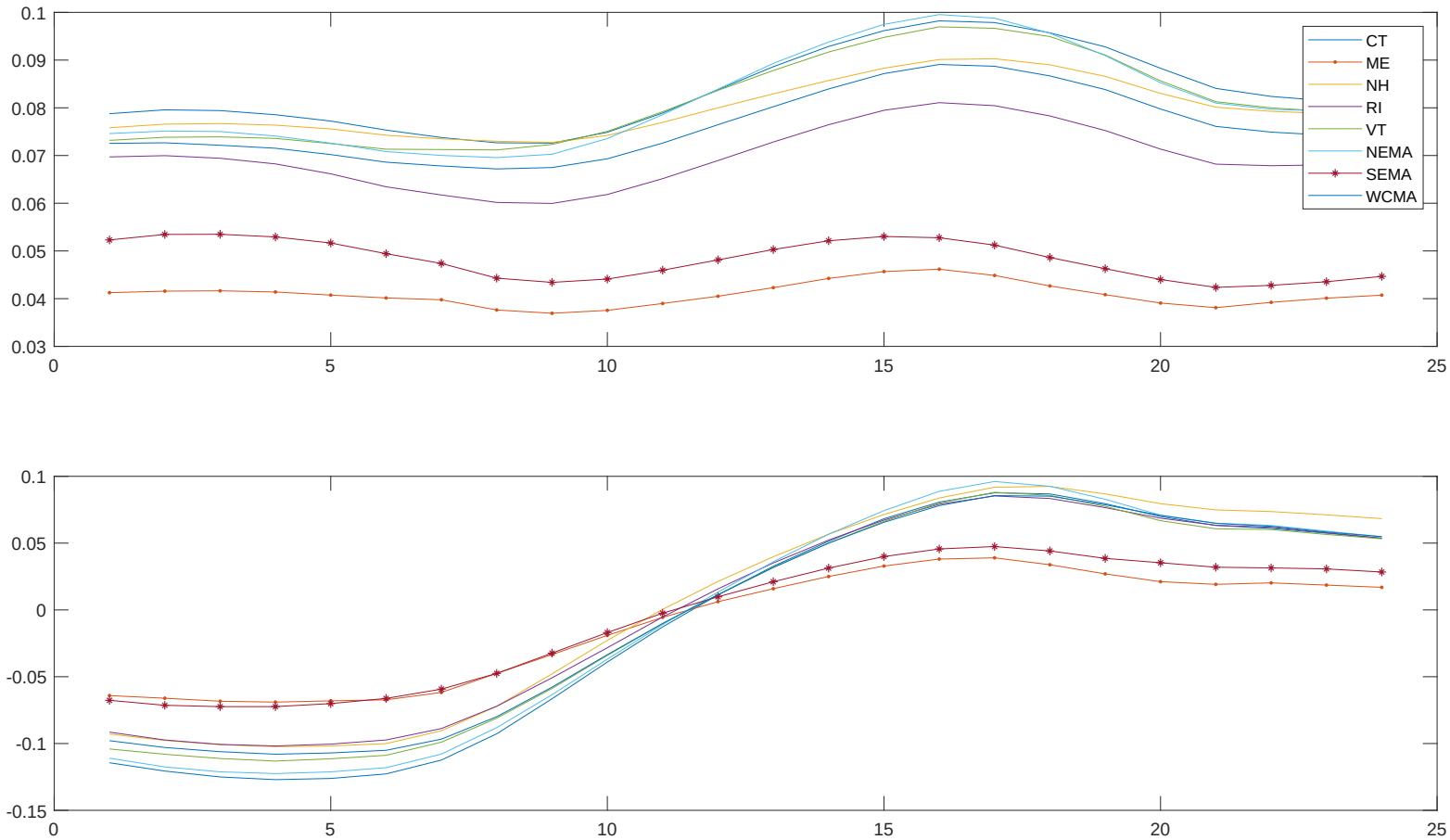
- Step 1: Number of outliers detected by day of the week:

	Day of the week							
Outlier type	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	Total
MAO	22	10	5	3	9	8	56	113
AO	14	9	7	11	10	7	4	59

- Steps 2 to 5: Leads to $k = 2$ clusters with 2 global factors and 6 and 5 specific factors, respectively, in the first and second clusters.

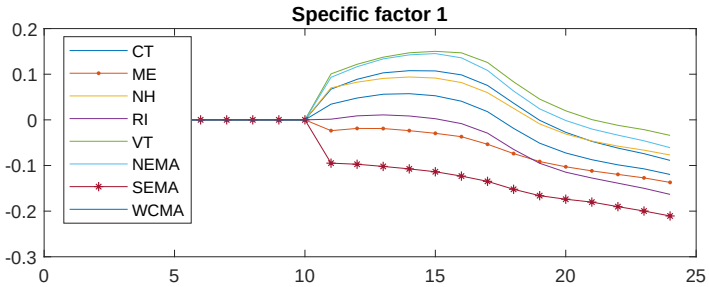
Real data example: Electricity demands in New England

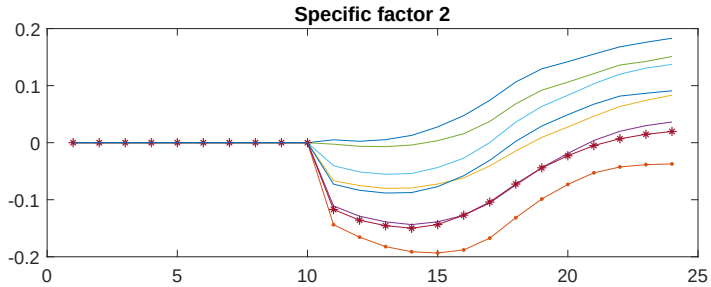
- Estimated loadings with the DFMCS: Two global factors

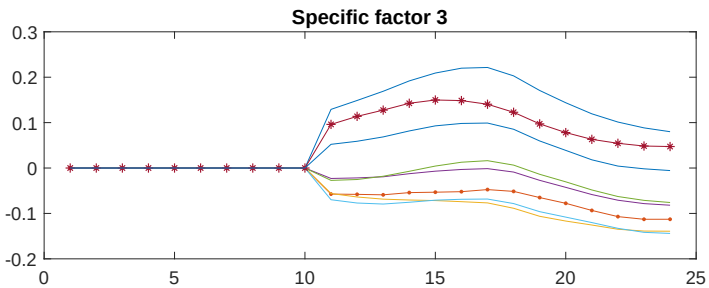


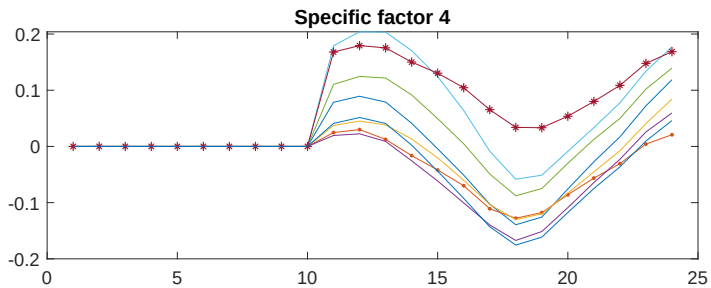
Real data example: Electricity demands in New England

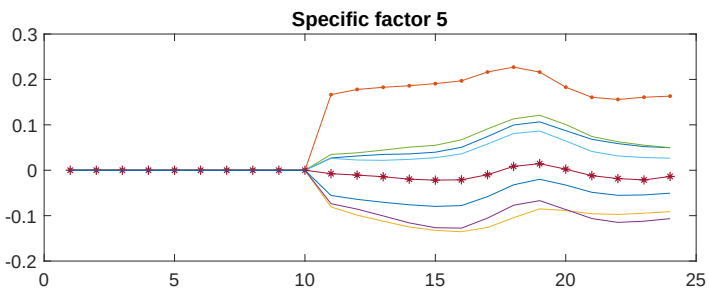
- Estimated loadings with the DFMCS: Six specific factors in the first cluster

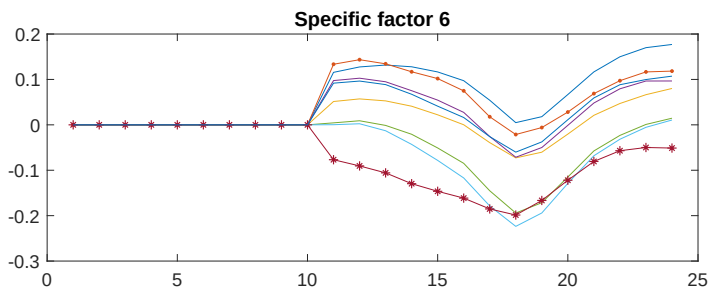






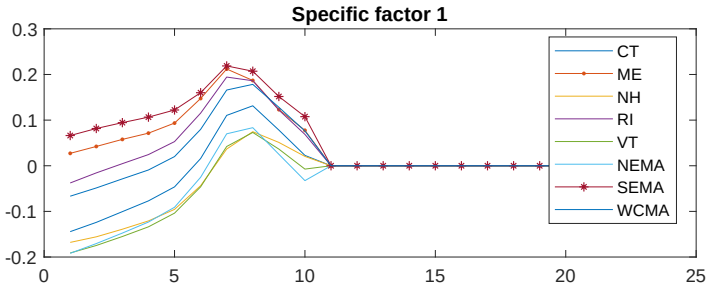


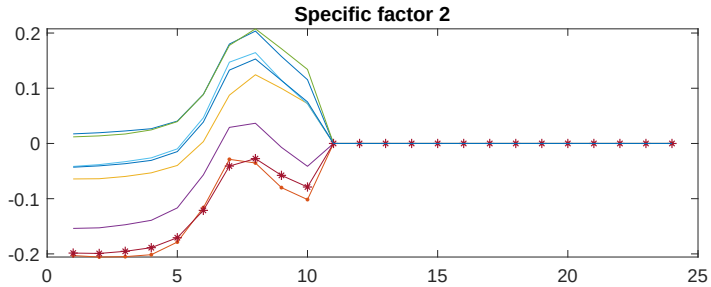


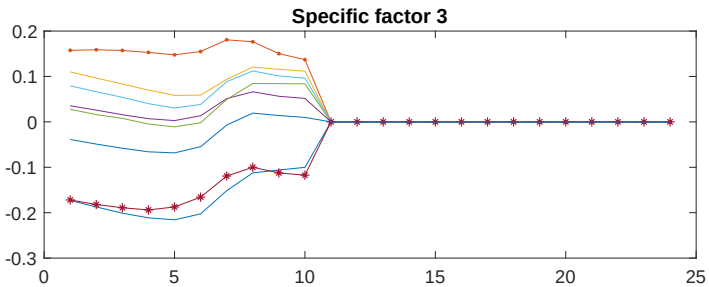


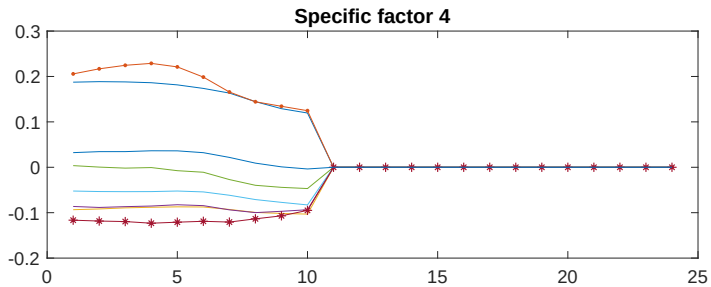
Real data example: Electricity demands in New England

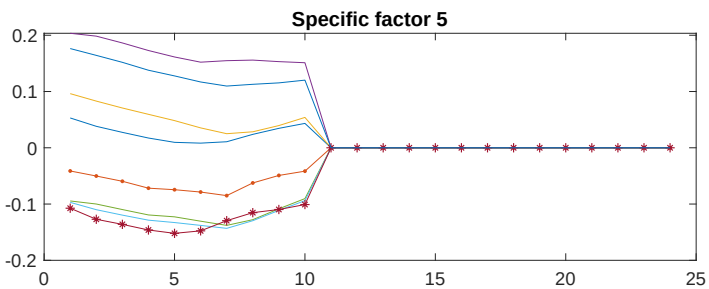
- Estimated loadings with the DFMCS: Five specific factors in the second cluster











Real data example: Electricity demands in New England

- Out of sample prediction exercise:
  - Two fitted models (to the outliers free series): M1 is the fitted DFM with two global factors and M2 is the fitted DFMCS model with two global factors and eleven specific factors.
  - Estimating and forecasting periods: The first ten years of data (3654 days) were used for model fitting and the last three years (1095 days) for forecasting.
  - Perform: A one-day ahead prediction exercise using a rolling windows across the forecasting period.
  - ARIMA models: Fitted to the factors in models M1 and M2.
  - MAE: The forecasting performance of M2 is a 7.41% better than M1.
  - RMSE: The forecasting performance of M2 is a 2.74% better than M1.

References

- Alonso, A. M., Galeano, P. and Peña, D. (2020). A robust procedure to build dynamic factor models with cluster structure. *Journal of Econometrics*, 216, 35-52.
- Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81, 1203–1227.
- Alonso, A. M. and Peña, D. (2019). Clustering time series by linear dependency. *Statistics and Computing*, 29, 655–676.
- Ando, T. and Bai J. (2017). Clustering huge number of financial time series: A panel data approach with high-dimensional predictor and factor structures. *Journal of the American Statistical Association*, 112, 1182–1198.
- Galeano, P., Peña, D. and Tsay, R. S. (2022). Outlier detection in high dimensional time series. Manuscript.
- Hallin M. and Liška, R. (2011). Dynamic factors in the presence of blocks. *Journal of Econometrics*, 163, 29-41.
- Wang, P. (2010). Large dimensional factor models with a multi-level factor structure. Working paper, Department of Economics, HKUST.

Robust modeling of large dimensional time series with cluster structure

Pedro Galeano

Departamento de Estadística and UC3M-Santander Big Data Institute

Universidad Carlos III de Madrid

pedro.galeano@uc3m.es

Joint work with Andrés M. Alonso (UC3M) and Daniel Peña (UC3M)

Financial support from Agencia Estatal de Investigación
PID2019-108311GBI00/AEI/10.13039/501100011033

New Bridges between Mathematics and Data Science

uc3m | Universidad Carlos III de Madrid