Robust Inference for Dynamic Semiparametric Factor Models

J. Bodelet[1*] and D. La Vecchia[1]

[1] *Research center in statistics and Geneva School of Economics and Management, University of Geneva, Blv. Pont d'Arve 40CH-1211, Geneva, Switzerland; Julien.Bodelet@unige.ch, Davide.LaVecchia@unige.ch.*
[*] *Presenting author*

Keywords. *Robust statistics; Dynamic factor model; Functional Magnetic Resonance Imaging; M-estimators on sieves; Outliers; Semiparametric inference.*

The aim of this paper is to define a robust statistical methodology to conduct inference on high dimensional data, collected in a dynamic context and observed at changing locations. A typical example of this kind of data arises in neuroscience, where the functional Magnetic Resonance Imaging (fMRI) datasets convey information about the brain's response (the so called, blood-oxygen-level-dependent, BOLD, signals) to certain stimuli, over time and over different activation areas (the so called, voxels, namely volumetric pixels).

To control for the high-dimensionality of the data, Dynamic Semiparametric Factor Models (DSFM) have been introduced by Park et al. [2009], where the aim is to isolate a limited number L of (latent) factors, explaining the behavior of the response variable (say, Y, representing e.g. the BOLD signals). Specifically, we observe $(X_{t,i}, Y_{t,i})$ for $i = 1, ..., n$ (cross-sectional dimension) and $t = 1, ..., T$ (time series dimension) and we set the model:

$$Y_{t,i} \quad = \quad m_0(X_{t,i}) + \sum_{l=1}^{L} Z_{t,l} m_l(X_{t,i}) + \epsilon_{t,i}, \tag{1}$$

where $(Z_{t,1}, ..., Z_{t,L})$ is an unobservable L-dimensional process (representing the $L \ll n$ latent factors). The $(L+1)$-functions $(m_0, ..., m_L)$ are unknown real-valued functions, defined on a (compact) subset of $\mathbb{R}^d$. The variables $X_{1,1}; ...; X_{T,n}; \epsilon_{1,1}; ...; \epsilon_{T,n}$ are independent, while the errors have zero mean and finite variance. Estimation methods for the model in (1) have been developed (and implemented) in Fengler, M. R., Härdle, W. K. and Mammen, E. [2007], Park et al. [2009] and Härdle, W. K. & Majer, P. [2014]. The extant inferential procedures rely on the least squares method and make use of nonparametric techniques (e.g., kernels or splines), to estimate both the latent factors and the functions m_l, for $l = 0, 1, ..., L$. A challenging open problem related to the inference about DSFM is the treatment of outliers.

Indeed, our simulation exercises emphasize that even a small number of anomalous records can largely distort the inference about the underlying latent factors. This problem is particularly relevant in the neurological applications, where diagnoses are obtained from the statistical analysis of fMRI datasets typically contaminated by motion artifacts and/or recording device failures; see Muschelli et al. [2014] and Power et al. [2014].

To cope with this statistical issue, we consider the class of $M-$estimators on sieves (see Chen, X. & Shen, X. [1997]) for DSFM and we define a class of robust inferential procedures which, by design, mitigates the impact of extreme observations. Since in the considered context a formal characterization of a robustness principle is missing, we first need to introduce a criterion which identifies the class of robust $M-$estimators for the DSFM. Then, we: (i) investigate the asymptotic properties of our new estimators, following [van der Geer, S. A. , 2000]; (ii) define an algorithm to implement our procedure, following Park et al. [2009]; (iii) develop a methodology for selecting the constant which controls the degree of robustness of our $M-$estimates, following La Vecchia, D., Camponovo, L. & Ferrari, D. [2015]. Monte-Carlo simulations provide numerical evidence of the good performance of our methodology, under different contamination settings. An application to fMRI data analysis concludes the paper.

References

Chen, X. & Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, 289–314.

Fengler, M. R., Härdle, W. K. & Mammen, E. (2007). A semiparametric factor model for implied volatility surface dynamics. *Journal of Financial Econometrics*, **5**, 189–218.

Härdle, W. K. & Majer, P. (2014). Yield curve modeling and forecasting using semiparametric factor dynamics. *The European Journal of Finance*, 1–21.

La Vecchia, D., Camponovo, L. & Ferrari, D. (2015). Robust heart rate variability analysis by generalized entropy minimization. *Computational Statistics & Data Analysis*, **82**, 137–151.

Muschelli, J., Nebel, M. B., Caffo, B. S, Barber, A. D., Pekar, J. J. & Mostofsky, S. H. (2014). Reduction of motion-related artifacts in resting state fMRI using aCompCor. *Neuroimage*, **96**, 22–35.

Park, B. U., Mammen, E., Härdle, W. & Borak,S. (2009). Time series modelling with semiparametric factor dynamics. *Journal of the American Statistical Association*, **104**, 284–298.

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L. & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage*, **84**, 320–341.

van der Geer, S. A. (2000). Empirical Processes in $M-$estimation. Cambridge university press.