CANDIDATURA PARA EL PREMIO RAMIRO MELENDRERAS
A NOMBRE DE JAVIER ÁLVAREZ LIÉBANA

Prediction of air pollutants PM_{10} by ARBX(1) processes

J. Álvarez-Liébana · M. D. Ruiz-Medina

1 State of the art

We have formulated an easy and flexible framework for the estimation and prediction of Banach-valued autoregressive processes of order one with exogenous variables (ARBX(1) processes), without conditions over the Banach space involved. These theoretical developments have been applied to the short–term forecasting of daily average concentrations of atmospheric aerosol particles with diameters less than 10 μm, also known as PM_{10} (coarse particles).

The importance of the accurate forecasting of this kind of particles relies on being inhalable atmospheric pollution particles, which impact the public health. Following the suggestions by the World Health Organization, the European Union developed in 2008 (in particular, directive 2008/50/EU) a complete legislative package, establishing health based standards for the levels of PM_{10}: daily mean concentration of PM_{10} should not be greater than 50 $\mu g\ m^{-3}$ more than 35 days per year, neither the annual average of concentration of PM_{10} shall not be greater than 40 $\mu g\ m^{-3}$. However, this limit has been exceed during the last years in heavily industrialized areas, deriving in severe people's health problems. Therefore, PM_{10} forecasting is crucial to adopting efficient public transport policies.

Several approaches have been adopted in the analysis of pollution data (see, e.g., [16], for a comparative study). In [21], the singular value decomposition is applied to identify spatial air pollution index (API) patterns, in relation to meteorological conditions in China. A novel hybrid model combining Multilayer perceptron model and Principal Component Analysis (PCA) is introduced in [14], to improve the air quality prediction accuracy in urban areas. Factor analysis and Box–Jenkins methodology are considered in [11], to examine concentrations of primary air pollutants such as NO, NO_2 , NO_x , PM_{10}, SO_2 and ground level O_3 in the town of Blagoevgrad, Bulgaria. In the recent literature, one can find several modelling approaches for PM_{10} forecasting. Among the most common statistical techniques applied, we mention multiple regression, non–linear state space modelling and artificial neural networks (see, e.g., [12, 17, 20]). Functional Data Analysis (FDA) techniques also play a crucial role in air quality forecasting.

This work was supported in part by projects PGC2018-099549-BI00 and MTM2015–71839–P (co-funded by Feder funds), of the DGI, MINECO, Spain.

J. Álvarez-Liébana
Department of Statistics, O.R. and Didactics of Mathematics, University of Oviedo
E-mail: alvarezljavier@uniovi.es

M. D. Ruiz-Medina
Department of Statistics and O.R., University of Granada E-mail: mruiz@ugr.es

Concerning the methodology, there exists an extensive literature on functional time series prediction, when a Hilbert space is adopted. Since the easily on working in the Hilbertian framework (Hilbert space provides a generalization of an Euclidean space), several proposals arise, in the parametric and nonparametric framework, for the estimation of the autocorrelation operator, and the prediction of the corresponding processes in function spaces. This work will be focused on parametric functional linear time series techniques which have been demonstrated fast and computational low-cost, in contrast with the more flexible nonparametric functional statistical approach, where the so–called curse of dimensionality problem arises (see [9]). Particularly, the approach presented allows a flexible analysis of the local variability of the functional values of the random variables studied, as well as the derivation of strong-consistent functional plug–in predictors, under a state space based framework. From a theoretical point of view, in the autoregressive Hilbertian process framework, the asymptotic properties of componentwise estimators of the autocorrelation operator, and their associated plug-in predictors have been derived in [4], among others. Recently, in [1,18], alternative operator norms for consistency have been investigated. The separable Banach space context has also been adopted in linear functional time series modeling, under a state space based approach. This literature has mainly been focused on the spaces of continuous functions $\mathcal{C}([0,1])$ with the supremum norm (see [5], among others), and on the Skorokhod space of right–continuous functions on $[0,1]$, having limit at the left at each $t \in [0,1]$, equipped with the Skorokhod topology, usually denoted as $\mathcal{J}_1$–topology (see, e.g., [2]). The lack of an inner product structure, in the abstract Banach–valued time series framework, is supplied in [19] by considering suitable embeddings into related Hilbert spaces.

Concerning the inclusion of exogenous information, a first attempt for the inclusion of exogenous information in the functional time series framework can be found in [6,7], where the so-called ARHX(1) processes (Hilbert-valued autoregressive processes of order one with exogenous variables) are introduced. Enhancements were subsequently proposed by [15]. First order conditional autoregressive Hilbertian processes were introduced in [13]. The present paper extends the time series framework in [19] to the case of first-order Banach-valued autoregressive processes with exogenous variables (ARBX(1) processes). Functional parameter estimation and plug–in prediction can be addressed in our ARBX(1) context, from the multivariate infinite–dimensional formulation of the results in [19]. Specifically, a matrix–operator–based formulation of the ARB(1) process (Banach–valued autoregressive process of order one) state equation is considered. The required Hilbert space embeddings, and sufficient conditions for the strong–consistency of the autocorrelation operator estimator (reflecting temporal correlations between endogenous and exogenous variables), and the associated plug–in predictor, were obtained.

2 Summary: motivation and contributions

It is well-known that Functional Data Analysis (FDA) techniques provide a flexible framework for the local analysis of high–dimensional data which are continuous in nature. Computational advances have made possible the implementation of flexible models for random elements in function spaces. One of the main subjects in FDA is the suitable choice of the function space, where the observed data take their values.

In particular, the norm of the selected space should provide an accurate measure of the local variability of the observed endogenous and exogenous variables, that could be crucial in the posterior representation of the possible interactions.

This paper adopts an abstract Banach space (a complete norm space: the notion of orthogonality disappears) framework assuming an autoregressive dynamics in time, for all the functional random variables involved in the model. Most of authors have worked with Banach spaces (such as $\mathcal{C}([0,1])$) well-adapted for measuring the local regularity. This work is focused on exploiting the opposite motivation, considering Banach spaces in which functions are locally singular, and then allowing irregular (non-smoothed), preserving the information included in singular data such as meteorological variables. Particularly, the scale of fractional Besov spaces provides a suitable functional framework, modelling local singularity in an accurate way. Note that the norms in these spaces can be characterized in terms of the wavelet transform. Specifically, wavelet bases provide countable dense systems in Besov spaces, that can be used in the definition of the inner product and associated norms in weighted fractional Sobolev spaces, constructed from the space of square integrable functions on an interval (see [19]).

Then, a Banach-valued autoregressive process of order one with exogenous variables (ARBX(1) process) is considered. Let $X = \{X_n,\ n \in \mathbb{Z}\}$ be a zero-mean ARB(1) process, with $\mathcal{P}\left(X_n \in B\right) = 1$, $n \in \mathbb{Z}$, satisfying

$$X_n = \rho\left(X_{n-1}\right) + \varepsilon_n, \quad n \in \mathbb{Z}, \tag{1}$$

where ρ is the autocorrelation operator, which is assumed to be a bounded linear operator on B, that is, $\rho \in \mathcal{L}(B)$, with $\left(\mathcal{L}(B), \|\cdot\|_{\mathcal{L}(B)}\right)$ denoting the Banach space of continuous operators with the supremum norm. Here, $\varepsilon = \{\varepsilon_n,\ n \in \mathbb{Z}\}$ represents the innovation process, which is assumed to be a B-valued strong white noise, and uncorrelated with the random initial condition. In this work, exogenous information is incorporated to equation (1) in an additive way. Thus, the state space equation of an ARBX(1) process is given by:

$$X_n = \rho\left(X_{n-1}\right) + \sum_{i=1}^{b} a_i\left(Z_{n,i}\right) + \varepsilon_n, \quad n \in \mathbb{Z}, \tag{2}$$

where $\{a_i,\ i = 1, \ldots, b\}$ are bounded linear operators on B. The exogenous functional random variables $Z_i = \{Z_{n,i},\ n \in \mathbb{Z}\}$, $i = 1, \ldots, b$, are assumed to satisfy an ARB(1) model. The endogenous and exogenous information affecting the functional response at a given time is incorporated through a suitable linear model, involving a matrix autocorrelation operator. This operator model reflects possible interactions between all endogenous and exogenous functional random variables at any time. Sufficient conditions are now formulated to ensure the strong consistency with respect to the supremum norm of the formulated componentwise functional parameter estimator. The simulation study and real–data application illustrate the fact that our approach is sufficiently flexible to describing the local behaviour of both, regular and singular functional data. The goal of the simulation study undertaken is to illustrate the flexibility, and the large–sample–size properties of the ARBX(1) parameter estimator, and associated functional plug–in predictor. The effect of the discretization step size is investigated as well. Note that, in the singular case, we can choose a suitable norm that measures the local fluctuations in a precise way. This information is relevant in the analysis of PM_{10} concentrations, as illustrated in the real-data application.

3 Future research lines

The incorporation of spatial interactions in the analysis could be addressed in a multivariate infinite-dimensional spatial framework, and constitutes the subject of a subsequent paper: a spatial functional correlation analysis should be considered. There are several available methods in the current spatial functional statistical literature (see [8]) to address this issue. Particularly, we refer to the frameworks of multivariate functional random field based prediction (see, e.g., [3]); and, spatial functional kriging–based techniques (see, e.g., [10], among others).

References

1. Álvarez-Liébana J, Bosq D, Ruiz-Medina MD (2017) Asymptotic properties of a componentwise ARH(1) plug-in predictor. J. Multivariate Anal. 155:12–34.
2. Blanke D, Bosq D (2016) Detecting and estimating intensity of jumps for discretely observed ARMAD(1,1) processes. J. Multivariate Anal. 146:119–137.
3. Bohorquez M, Giraldo R, Mateu J (2017) Multivariate functional random fields: prediction and optimal sampling. Stoch. Environ. Res. Risk Assess. 31:53–70.
4. Bosq D (2000) Linear Processes in Function Spaces. Springer, New York.
5. Bueno–Larraz B, Klepsch J (2018) Variable selection for the prediction of $C[0,1]$–valued AR processes using RKHS. arXiv:1710.06660.
6. Damon J, Guillas S (2002) The inclusion of exogenous variables in functional autoregressive ozone forecasting. Environmetrics 13:759–774.
7. Damon J, Guillas S. (2005) Estimation and simulation of autoregressie Hilbertian processes with exogenous variables. Stat. Inference Stoch. Process. 8:185–204.
8. Delicado, P, Giraldo R, Comas, C, Mateu, J (2010) Statistics for spatial functional data: some recent contributions. Environmetrics. 21:224–239.
9. Geenens G (2011) Curse of dimensionality and related issues in nonparametric functional regression. Statistics Surveys 5:30–43.
10. Giraldo R, Delicado P, Mateu J (2010) Geostatistics for functional data: an ordinary kriging approach. Environ. Ecol. Stat. 18:411–426.
11. Gocheva–Ilieva S, Ivanov A, Voynikova D, Boyadzhiev D (2014) Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach. Stoch. Environ. Res. Risk Assess. 28:1045–1060.
12. Grivas G, Chaloulakou A (2006) Artificial neural network models for prediction of PM_{10} hourly concentrations in the greater area of Athens, Greece. Atmospheric Environment 40:1216–1229.
13. Guillas S (2002) Doubly stochastic Hilbertian processes. J. Appl. Probab. 39:566–580.
14. He H-D, Lu W-Z, Xue Y (2015) Prediction of particulate matters at urban intersection by using multilayer perceptron model based on principal components. Stoch. Environ. Res. Risk Assess. 29:2107–2114.
15. Marion JM, Pumo B (2004) Comparaison des modéles ARH(1) et ARHD(1) sur des données physiologiques. Ann. I.S.U.P. 48:29–38.
16. Pang W, Christakos G, Wang J-F (2009) Comparative spatiotemporal analysis of fine particulate matter pollution. Environmetrics 21:305–317.
17. Paschalidou AK, Karakitsios S, Kleanthous S, Kassomenos PA (2011) Forecasting hourly PM_{10} concentration in Cyprus through artificial neural networks and multiple regression models: implications to local environmental management. Enviromental Science and Pollution Research 18:316–327.
18. Ruiz-Medina MD, Álvarez-Liébana J (2019) A note on strong-consistency of componentwise ARH(1) predictors. Stat. Probab. Lett. 145:224–248.
19. Ruiz-Medina MD, Álvarez-Liébana J (2019) Strongly consistent autoregressive predictors in abstract Banach spaces. J. Multivariate Anal. DOI: 10.1016/j.jmva.2018.08.001.
20. Stadlober E, Hormann S, Pfeiler B (2008) Quality and performance of a PM_{10} daily forecasting model. Atmospheric Environment 42:1098–1109.
21. Zhang L, Liu Y, Zhao F (2018) Singular value decomposition analysis of spatial relationships between monthly weather and air pollution index in China. Stoch. Environ. Res. Risk Assess. 32:733–748.