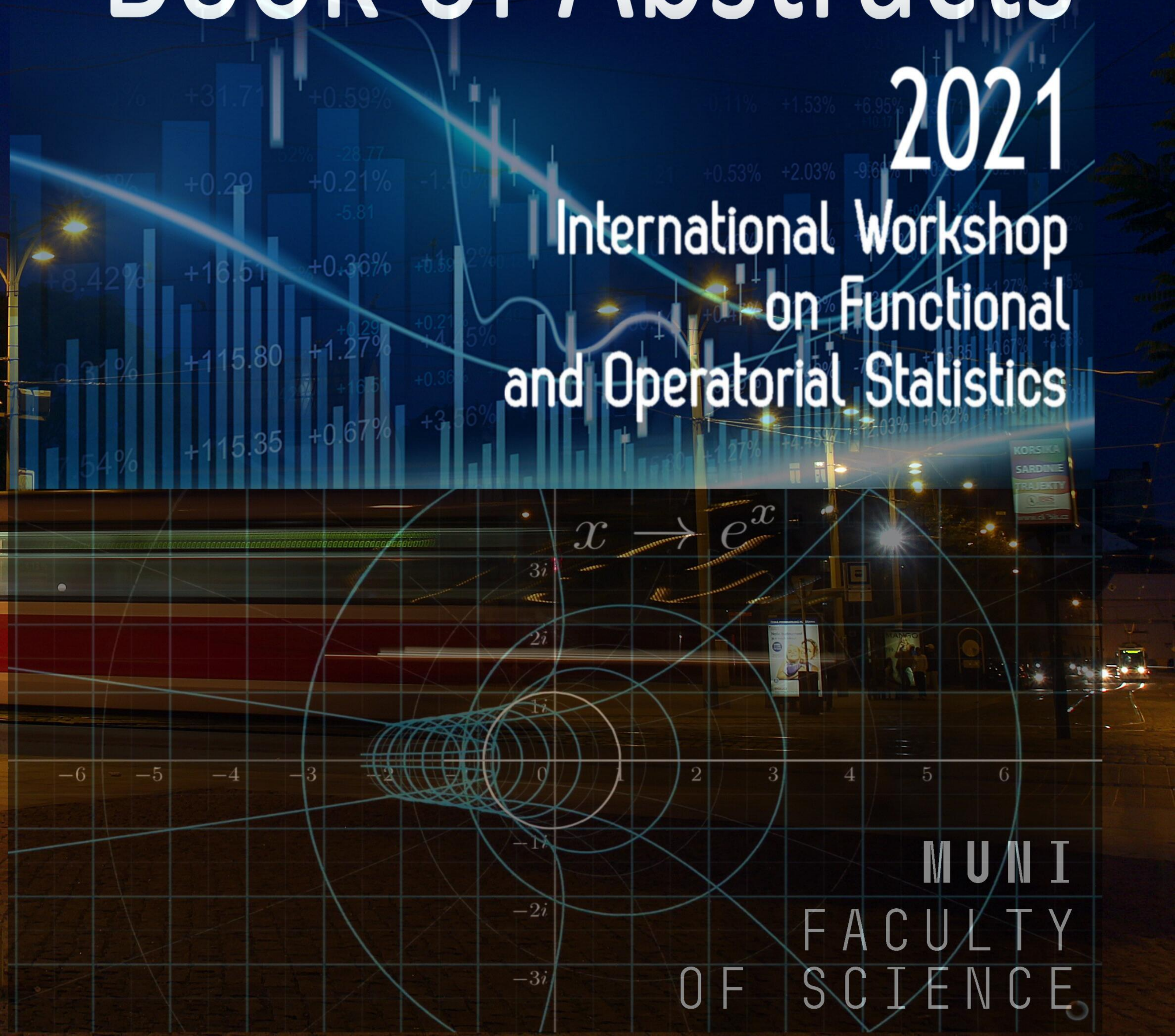
IWFOS

Book of Abstracts

2021

International Workshop on Functional and Operatorial Statistics



MUNI
FACULTY OF SCIENCE

Contents

Different Varimax Rotation Approaches of Functional PCA for the evolution of COVID-19 pandemic in Spain

Christian Acal, Ana M. Aguilera and Manuel Escabias

Abstract

It is well known that Functional Principal Component Analysis (FPCA) provides an approximated linear representation of a sample of curves in terms of a reduced set of ortonormal functions and uncorrelated scalar variables (principal components), whose interpretation allows to explain the main patterns of variability in a functional data set [4]. In this work, FPCA is performed to explain the evolution of the curves of positive cases in the first wave of COVID-19 pandemic in the Spanish autonomous communities. After properly homogenizing and registering the data in a common interval so that the observed curves become comparable, a B-spline basis expansion approach is considered for reconstructing the true functional form of the curves from their daily registration. The inherent problem in this application is that almost all variability falls on the first principal component, which is a straightforward average or size effect. An usual solution to redistribute variability in PCA and to make the interpretation easier consists of applying some kind of rotation on the weights of principal components, with Varimax criterion being the most popular due to its good properties and simpleness [2]. The main drawback of rotation lies in that one of the two essential properties of PCA, orthogonality of the weights or uncorrelatedness of the components, is lost.

Varimax rotation was first extended to FPCA in two different ways: one is based on Varimax rotation of the matrix of values of the weight functions at equally spaced points (R1), and the other on Varimax rotation of the matrix of basis coefficients of the weight functions (R2) (see [4] for a detailed study and interesting examples). Both approaches retain the property of orthogonality but the rotated principal component scores are not uncorrelated anymore. Besides, neither of them are a true rotation of weight functions. More recently, two new functional Varimax procedures have been introduced to solve these disadvantages [1]. Both of them are based in the equivalence between FPCA and multivariate PCA of certain transformation of the matrix of basis coefficients of the sample curves [3]. On the one hand, rotation of the matrix of weight vectors (eigenvectors of the sample covariance matrix) is considered to preserve the orthogonality between the rotated weight functions (eigenfunctions of the sample covariance operator) (R3). On the other hand, rotating the matrix of loadings of the standardised principal component scores is proposed to provide uncorrelated rotated scores (R4). An exhaustive simulation study is developed in this work to compare the four functional Varimax approaches by concluding that the third ensures a more accurate estimation and is more robust with respect to the number of discrete time observations of the sample curves. Finally, the combination of the new approaches R3 and R4 will show its potential to help to distinguish different behaviors in the evolution of positive cases in the Spanish autonomous communities.

References

[1] Acal, C., Aguilera, A.M., Escabias, M.: New Modeling Approaches Based on Varimax Rotation of Functional Principal Components. Mathematics. **8**, 2085 (2020)

[2] Jolliffe, I.: Principal Component Analysis. Springer, Berling (2002)

[3] Ocaña, F.A., Aguilera, A.M., Escabias, M.: Computational considerations in functional principal component analysis. Comput. Stat. **22**, 449–465 (2007)

[4] Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer, Berling (2005)

Christian Acal
Department of Statistics and O.R. and IMAG, University of Granada, Spain, e-mail: chracal@ugr.es

Ana M. Aguilera
Department of Statistics and O.R. and IMAG, University of Granada, Spain, e-mail: aaguiler@ugr.es

Manuel Escabias
Department of Statistics and O.R. and IMAG, University of Granada, Spain, e-mail: escabias@ugr.es

Generalized Functional Partially Linear Single-Index Models

Mohamed Alahiane, Idir Ouassou, Philippe Vieu and Mustapha Rachdi

Abstract

Single-index models are potentially important tools for multivariate nonparametric regression. They generalize linear regression by replacing the linear combination $\alpha_0^\top X$ with a nonparametric component, $\eta_0\left(\alpha_0^\top X\right)$, where $\eta_0(\cdot)$ is an unknown univariate link function. L. Wang and G. Cao (2018) has study generalized partially linear single-index models (GPLSIM) where the systematic component in the model has a flexible semi-parametric form with a general link function. In this paper we generalize these models to have a functional component, replacing the generalized Partially Linear Single-Index Models $\eta_0\left(\alpha_0^\top X\right) + \beta_0^\top Z$ by $\eta_0\left(\alpha_0^\top X\right) + \int_0^1 \beta_0(t)Z(t)dt$ where α is a vector in R^d to be estimated and $\eta_0(\cdot)$ and $\beta_0(\cdot)$ are a unknown functions. We call these generalized functional partially linear single-index models (GFPLSIM). We propose estimates of the unknown parameter α_0 and the unknown functions $\beta_0(\cdot)$ and $\eta_0(\cdot)$ and obtain their asymptotic distributions. Examples illustrate the models and the proposed estimation methodology.

References

[1] Li W. and Guanqun C.: Efficient Estimation for Generalized Partially Linear Single-Index Models. Bernoulli. **24**(2), 1101-1127 (2018)

[2] Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P.: Generalized partially linear single-index models. Journal of the American Statistical Association. **92** 477–489 (1997)

[3] De Boor, C. . A practical guide to splines, revised Edition, **Vol. 27** of Applied Mathematical Sciences. Springer-Verlag, Berlin. (2001)

[4] Wang, L. and Yang, L.: Spline estimation of single-index models. Statistica Sinica **19** 765–783 (2009)

[5] Wang, J. L., Xue, L., Zhu, L. and Chong, Y. S.: Estimation for a partial-linear single-index model. The Annals of Statistics **38** 246–274 (2010)

[6] Wang, J. L., Xue, L., Zhu, L. and Chong, Y. S.: Estimation for a partial-linear single-index model. The Annals of Statistics. **38** 246–274 (2010)

[7] Yu, Y. and Ruppert, D.: Penalized spline estimation for partially linear single-index models. Journal of the American Statistical Association. **97** 1042–1054 (2002)

Mohamed Alahiane
Université Cadi Ayyad, Ecole Nationale des Sciences Appliquées, Marrakech, Moroc. e-mail: alahianemed@gmail.com

Idir Ouassou
Université Cadi Ayyad, Ecole Nationale des Sciences Appliquées, Marrakech, Moroc. e-mail: i.ouassou@uca.ma

Philippe Vieu
Institut de Mathématiques de Toulouse, Université Paul Sabatier, 31062 Toulouse cedex 9, France. e-mail: vieu@math.univ-toulouse.fr

Mustapha Rachdi
Grenoble-Alpes, Laboratoire AGIM FRE 3405 CNRS, Université P. Mendès France (Grenoble 2), UFR SHS, BP. 47, 38040 Grenoble Cedex 09, France.
e-mail: Mustapha.Rachdi@upmf-grenoble.fr

Analysis of Telecom Italia mobile phone data by space-time regression with differential regularization

Eleonora Arnone, Mara S. Bernardi, Laura M. Sangalli and Piercesare Secchi

Abstract

In this work, we apply spatial regression methods with Partial Differential Equation (PDE) regularization [8, 9, 4, 5] to the Telecom Italia mobile phone data. In particular, we consider the Space-Time regression with PDE penalization method (ST-PDE) introduced in [6] and extend it to deal with observations featuring complex spatial dependency. The technique proposed allows to include specific information on the phenomenon under study through a definition of the non-stationary anisotropy characterizing the spatial regularization based on the texture of the domain on which the data are observed.

ST-PDE is a penalized regression method that models separately the spatial and the temporal regularization by considering two roughness penalties, which account separately for the regularity of the field in space and in time by using a tensor product, following the approach used also by [1, 3, 7]; while, in the generalization of the technique proposed by [2], a single roughness penalty is used to jointly model the spatial and temporal dimensions. Therefore, in the ST-PDE model, the field is estimated minimizing a functional composed by three parts: a data-fitting part, a penalization for the spatial regularity, and a penalization for the temporal regularity. In [6], the spatial penalization involves a simple differential operator that imposes smoothness to the solution. Instead, in this work, we consider a spatial penalization involving a more general PDE, that allows to impose non-stationary anisotropy to the solution, thus modeling more complex spatial dependencies. Moreover, the PDE can model problem-specific knowledge on the phenomenon under study. For example, if the PDE governing the physical phenomenon generating the data is available, it can be exploited in the spatial regularization term of the ST-PDE functional, thus driving the estimation towards a physically sound solution. In the context of the analysis of mobile phone data, where no physical knowledge on the phenomenon under study is available, we use the PDE to include in the model information about the texture of the spatial domain; in particular, we here characterize the PDE using the road network, which highly influences the data. This application highlights the high flexibility of the definition of spatial dependence imposed by the ST-PDE model.

References

[1] Aguilera-Morillo M. C., Durbán M., Aguilera A. M.: Prediction of functional data with spatial dependence: a penalized approach. Stochastic Environ Res Risk Assess **31**, 7–22 (2017)

[2] Arnone E., Azzimonti L., Nobile F., Sangalli L. M.: Modeling spatially dependent functional data via regression with differential regularization. Journal of Multivariate Analysis **170**, 275–295 (2019)

[3] Augustin N. H., Trenkel V. M., Wood S. N., Lorance P.: Space-time modelling of blue ling for fisheries stock management. Environmetrics **24**(2) 109–119 (2013)

[4] Azzimonti L., Nobile F., Sangalli L. M., Secchi P.: Mixed Finite Elements for Spatial Regression with PDE Penalization. SIAM/ASA Journal on Uncertainty Quantification **2**(1), 305–335 (2014)

[5] Azzimonti L., Sangalli L. M., Secchi P., Domanin M., Nobile F.: Blood flow velocity field estimation via spatial regression with PDE penalization. Journal of the American Statistical Association **110**(511), 1057–1071 (2015)

[6] Bernardi M. S., Sangalli L. M., Mazza G., Ramsay J. O.: A penalized regression model for spatial functional data with application to the analysis of the production of waste in Venice province. Stochastic Environ Res Risk Assess **31**(1), 23–38 (2017)

[7] Marra G., Miller D. L., Zanin L.: Modelling the spatiotemporal distribution of the incidence of resident foreign population. Statistica Neerlandica **66**(2) 133–160 (2012)

[8] Ramsay, T.: Spline smoothing over difficult regions. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **54**(2), 307–319 (2002)

[9] Sangalli L. M. and Ramsay J. O. and Ramsay T. O.: Spatial spline regression models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **75**(4), 681–703 (2013)

Eleonora Arnone
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy, e-mail: eleonora.arnone@polimi.it

Mara S. Bernardi
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy e-mail: marasabina.bernardi@polimi.it

Laura M. Sangalli
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy e-mail: laura.sangalli@polimi.it

Piercesare Secchi
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy e-mail: piercesare.secchi@polimi.it

Learning with Signatures

Gérard Biau

Abstract

Sequential and temporal data arise in many fields of research, such as quantitative finance, medicine, or computer vision. The present talk is concerned with a novel approach for sequential learning, called the signature method and rooted in rough path theory. Its basic principle is to represent multidimensional paths, i.e., functions from $[0, 1]$ to $\mathbb{R}^d$, by a graded feature set of their iterated integrals, called the signature. After a survey of their basic principles, I will investigate how signatures can be used in machine learning, with illustrations on recent and challenging datasets.

Gérard Biau
Sorbonne University, e-mail: gerard.biau@sorbonne-universite.fr

About the complexity of a Functional Time Series

Enea G. Bongiorno, Lax Chan, Aldo Goia and Philippe Vieu

Abstract

Consider a functional time series taking values in a general topological space and assume that its Small-Ball Probability (SmBP) factorizes into two terms that play the role of a surrogate density and of a volumetric term. The latter is a mean for studying the complexity of the underlying process, since it may reveal some latent feature of it. In some cases, it can be analytically specified in a parametric form: a special situation is given when the process belongs to the monomial family, like in the finite dimensional and fractal case, for which the volumetric term has monomial form depending on the SmBP radius and a parameter named complexity index. This work presents some recent developments concerning the study of a nonparametric estimator for the volumetric term based on a U-statistic. Weak consistency of this estimator is provided in the beta-mixing case. In the particular case of a monomial family, it is possible to estimate the complexity index by minimizing a suitable dissimilarity measure. For this estimator asymptotic Gaussianity is shown, providing a theoretical support to build confidence interval for the complexity index. A Monte Carlo simulation is carried out in order to assess the performance of the methodology for finite sample sizes. Finally, the new method is applied to detect the complexity of a real world dataset.

References

[1] Bongiorno, E.G., Goia, A. Vieu. P.: Evaluating the complexity of some families of functional data. SORT **42** (1), 27–44 (2018)

[2] Bongiorno, E.G., Goia, A. Vieu. P.: Estimating the complexity index of functional data: Some asymptotics. Statistics and Probability Letters **161**, 108731 (2020)

[3] Bongiorno, E.G., Chan, L., Goia, A.: On the use of the Complexity index for detecting the dimensionality of a Functional time series. Submitted (2021)

Enea G. Bongiorno
DISEI, Università del Piemonte Orientale, Novara, Italia, e-mail: enea.bongiorno@uniupo.it

Lax Chan
DISEI, Università del Piemonte Orientale, Novara, Italia, e-mail: lax.chan@uniupo.it

Aldo Goia
DISEI, Università del Piemonte Orientale, Novara, Italia, e-mail: aldo.goia@uniupo.it

Philippe Vieu
Institut de Mathématiques de Toulouse, Université Paul Sabatier, France, e-mail: philippe.vieu@math.univ-toulouse.fr

Frequency Analysis of a Cyclostationary Random Function

Alain Boudou and Sylvie Viguier-Pla

Abstract

We develop and extend the principal components analysis (PCA) of a cyclostationary random function $(X_t)_{t\in\mathbb{R}}$, that is a function such that $\mathrm{cov}(X_t, X_{t'}) = \mathrm{cov}(X_{t+\Delta}, X_{t'+\Delta})$, for any (t, t') of $\mathbb{R}\times\mathbb{R}$. This property of cyclostationarity for random functions, also refered as periodically correlated, is encountered in various phenomena where some statistics present a periodicity. We can find examples of study of this property in Gardner [5] for telecommunications, Randall et al. [9] for mechanic transmission, Weber et al. [11] for radioastronomy, Zakaria [13] for locomotion, or Roussel [10] for medicine. The earliest mention of such processes can be found in Voychishin et al [12], which gives an english translation of articles first published ind 1957 and 1960. The mathematical formulation has first been given by Gladyshev [6], and then largely developped by Hurd [7], Hurd et al. [8], and more recently by Bouleux et al. [4], in a multidimentional context.

In this presentation, from such a random function, we define a series $(Y_n)_{n\in\mathbb{N}}$ of random functions, where $Y_n = (X_{n+t})_{t\in[0;\Delta[}$. This series is multidimentional, it takes values in the Hilbert space $L^2([0;\Delta[)$, and it is stationary. This stationarity lets us proceed to Principal Components Analysis in the frequency domain, which is more powerful than Principal Components Analysis of each of the random vectors Y_n.

In Boudou and Viguier-Pla [3], this method of PCA is exposed for most common cyclostationary functions. In this presentation, we make use of more complex mathematical tools for an extension of the field of work.

We generalize the set of the indexes $\mathbb{R}$ to an abelian group G_1, which can be $\mathbb{Z}$, $\mathbb{Z}/k\mathbb{Z}$, $\mathbb{Z}^k$, $\mathbb{R}^k$, and as for the subgroup $\Delta\mathbb{Z}$ of $\mathbb{R}$, it becomes a subgroup G of G_1. Then we say that a random function $(X_g)_{g\in G_1}$ is cyclostationary when $\mathrm{cov}(X_{g_1}, X_{g_2}) = \mathrm{cov}(X_{g_1+g}, X_{g_2+g})$, for any (g_1, g_2, g) of $G_1 \times G_1 \times G$.

Of course, the type of cyclostationarity is linked with the choice of the sub-group G of G_1. We get, as particular cases, various forms of cyclostationarity more or less known.

In this work, the new skill is based on the fact that with a cyclostationary function we associate a stationary multidimensional random function, and then lets us use the powerful tool of reduction of the data which is PCA in the frequency domain (cf. Brillinger [1] and Boudou and Dauxois [2]). Let us finaly mention a result which interest is only theoretical: with cyclostationary random function we associate, in a biunivoque way, a spectral measure, as well as with a stationary continuous random function, we associate a random measure from which it is the Fourier transform.

References

[1] Brillinger, D.M. (2001). *Time Series Data Analysis and Theory.* Reprint of the 1981 edition. Classics in Applied Mathematics, 36. Society for Industrial Applied Mathematics (SIAM), Philadelphia.

[2] Boudou, A. and Dauxois, J. (1994). Principal Component Analysis for a Stationary Random Function Defined on a Locally Compact Abelian Group. *J. Multivariate Anal.* **51** 1-16.

[3] Boudou, A. and Viguier-Pla, S. (2020). Principal Components Analysis of a Cyclostationary Random Function. In *Functional and High-Dimensional Statistics and Related Fields*, Chapter 6, Contributions to statistics, Springer Nature Switzerland.

[4] Bouleux, G., Dugast, M. and Marcon, F. (2019). Information Topological Characterization of Periodically Correlated Processes by Dilation Operators. *IEEE Trans. on Information Theory, Institute of Electrical and Electronics Engineers, I* **65** 10 6484-6495.

[5] Gardner, G. D. Z. (1994) *Cyclostationarity in communications and signal processing*. IEEE Press.

[6] Gladyshev, E. G. (1961) Periodically correlated random sequences. *Soviet Math.* **2** 385-388.

[7] Hurd, H. L. (1970) *An investigation of periodically correlated processes*. Durham, North Carolina, USA: Ph. D. dissertation of the Duke University.

[8] Hurd, H., Kallianpur, G., Farshidi, J. (2004) Correlation and spectral theory for periodically correlated random fields indexed on $\mathbb{Z}^2$. *J. Multivariate Anal.* **90** 359-383.

[9] Randall, R. B., Antoni, J. and Chobsaard, S. (2001) The Relationship Between Spectral Correlation and Envelope Analysis in the Diagnostics of Bearing Faults and Other Cyclostationary Machine Signals. *Mechanical Systems and Signal Processing*, **15** 5 945-62.

[10] Roussel, J. (2014). *Modélisation cyclostationnaire et séparation de sources des signaux électromyographiques*. PHD Thesis, Université d'Orléans.

[11] Weber, R., Faye, C. (1998). Real Time Detector for Cyclostationary RFI in Radio Astronomy. *EUSIPCO*.

[12] Voychishin, K.S., Dragan, Y.P. (1973) Example of formation of periodically correlated random processes. *Radio Eng. Electron. Phys.* **18** 1426-1429 (English translation of Radiotekh. Elektron. 18, 1957, 1960).

[13] Zakaria, F. (2015). *Analyse de la locomotion humaine: exploitation des propriétés de cyclostationarité des signaux*. PhD thesis, Université Jean Monnet, Saint Étienne, France.

Alain Boudou, Sylvie Viguier-Pla
Equipe de Stat. et Proba., Institut de Mathématiques, UMR5219, Université Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France,
e-mail: boudou@math.univ-toulouse.fr

Sylvie Viguier-Pla
Université de Perpignan via Domitia, LAMPS, 52 av. Paul Alduy, 66860 Perpignan Cedex 9, France, e-mail: viguier@univ-perp.fr

Identification of higher order derivatives of multivariate functions, with statistical applications

Jose E. Chacón and Tarn Duong

Abstract

Identification theorems for the gradient and Hessian of a multivariate function from its first and second order differentials, respectively, are well known. Similar results for higher order derivatives, however, remained largely unexplored. Here we advocate for the use of a vectorized representation of higher order derivatives, which allows us to provide an identification theorem that is valid for any arbitrary derivative order. We illustrate its usefulness with statistical applications, including vector Hermite polynomials, moments and cumulants.

References

[1] Chacón, J.E., Duong, T.: Higher order differential analysis with vectorized derivatives. arXiv preprint 2011.01833 (2020)

Jose E. Chacón
Universidad de Extremadura, e-mail: jechacon@unex.es

Tarn Duong
ecov, Saint-Denis, France, e-mail: tarn.duong@gmail.com

Minimax estimation in the functional regression model with a functional output

Gaëlle Chagny, Anouar Meynaoui and Angelina Roche

Abstract

We address the non-parametric estimation of a linear regression model S, with functional input and output. Mathematically,

$$Y = S(X) + \varepsilon,$$

where the covariates X, Y and the noise ε belong to $\mathbb{L}^2([0,1])$, the set of square-integrable functions on $[0,1]$, with its usual scalar product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$. We assume without a loss of generality that X, Y and ε are centered. Given *i.i.d.* observations $(X_i, Y_i)_{i=1,\dots,n}$ of (X, Y), we aim to provide a non-asymptotically minimax estimator of the model S. To do so, we consider a Hilbert basis $(\psi_j)_{j\geq 1}$ of $\mathbb{L}^2([0,1])$ and the operator collection $E_{m_1,m_2} = \mathrm{Span}\{\psi_k \otimes \psi_j; 1 \leq j \leq m_1, 1 \leq k \leq m_2\}$, where $m_1, m_2 \in \mathbb{N}^*$ and $\psi_k \otimes \psi_j : f \mapsto \langle f, \psi_j\rangle\psi_k$. Let γ_n be the contrast function, defined for all linear operator T by: $\gamma_n(T) = \frac{1}{n}\sum_{i=1}^n \|Y_i - T(X_i)\|^2$. We then introduce the projection estimators S_{m_1,m_2} constructed by minimizing the contrast function over E_{m_1,m_2}:

$$S_{m_1,m_2} = \underset{T\in E_{m_1,m_2}}{\arg\min}\ \gamma_n(T).$$

The explicit form of S_{m_1,m_2} is established in the general case (regardless of the basis choice). Subsequently, we study the particular case where $(\psi_j)_{j\geq 1}$ is the empirical PCA (Principal Component Analysis) basis associated to X. In the sequel, we denote by $\Gamma = \mathbb{E}[X \otimes X]$ the covariance operator of X and $\Gamma_n = \frac{1}{n}\sum_{i=1}^n X_i \otimes X_i$ its empirical version. The eigenelements of Γ and Γ_n are respectively denoted $(\lambda_j, \varphi_j)_{j\geq 1}$ and $(\hat\lambda_j, \hat\varphi_j)_{j\geq 1}$ (where the λ_j's and $\hat\lambda_j$'s are sorted in decreasing order). We also introduce the empirical cross-covariance operator $\Delta_n = \frac{1}{n}\sum_{i=1}^n Y_i \otimes X_i$ and the "pseudo-inverse" of Γ_n as $\Gamma_n^\dagger = \sum_{j=1}^{m_1} \hat\lambda_j^{-1}\hat\varphi_j \otimes \hat\varphi_j$. We then show that the estimator S_{m_1,m_2} is defined as $S_{m_1,m_2} = \hat\Pi_{m_2}\Delta_n\Gamma_n^\dagger$, where $\hat\Pi_{m_2}$ stands for the orthogonal projection onto $\mathrm{Span}\{\hat\varphi_1, \dots, \hat\varphi_{m_2}\}$. Our estimator differs slightly from the one in [2]. To theoretically select the optimal projection dimensions m_1 and m_2, we consider as in [2] the Mean Square Prediction Error (MSPE) of the estimator S_{m_1,m_2} defined as $\mathrm{MSPE}(S_{m_1,m_2}) = \mathbb{E}\|S_{m_1,m_2}(X_{n+1}) - S(X_{n+1})\|^2$, where X_{n+1} is a new observation of X, independent of $(X_i, \varepsilon_i)_{i=1,\dots,n}$. Denoting $\Gamma^{1/2} = \sum_{j\geq 1}\sqrt{\lambda_j}\varphi_j \otimes \varphi_j$, we assume by analogy with [1] that $S\Gamma^{1/2}$ belongs to the regularity space: $\mathcal{W}^R_{\alpha,\beta} = \{T \text{ linear operator: } \sum_{j\geq 1}\sum_{r\geq 1}\eta_\alpha(j)\psi_\beta(r)\langle T(\varphi_j), \varphi_r\rangle^2 \leq R^2\}$, where $\alpha, \beta, R > 0$ and for all $\gamma > 0$, the functions η_γ and ψ_γ are either $u \mapsto u^\gamma$ or $u \mapsto \exp(u^\gamma)$. Under the regularity condition stated below and other classically prescribed assumptions (omitted here), we achieve a sharp upper-bound of the prediction risk as a classical bias-variance trade-off. This technical decomposition requires an extensive use of the perturbation theory, presented in [3]. We show in particular that if ψ_β is polynomial with $\beta > 6$ or exponential, the non-asymptotic minimax prediction risk is upper bounded as

$$\inf_{S_{m_1,m_2}}\ \sup_{S\Gamma^{1/2}\in\mathcal{W}^R_{\alpha,\beta}} \mathrm{MSPE}(S_{m_1,m_2}) \leq \inf_{m_1\in\mathbb{N}^*}\left\{\sigma_\varepsilon^2\frac{m_1}{n} + \frac{3}{\eta_\alpha(m_1)}\right\}, \tag{1}$$

where $\sigma_\varepsilon^2 = \mathbb{E}\|\varepsilon\|^2$. We finally prove that whenever ε is Gaussian, the upper bound of Equation (1) is also the minimax lower bound (up to a positive constant) over the same regularity space. This means that the last estimator is non-asymptotically minimax.

References

[1] Brunel, É., Mas, A., Roche, A.: Non-asymptotic adaptive prediction in functional linear models. Journal of Multivariate Analysis. 143, 208–238 (2016)

[2] Crambes, C., Mas, A.: Asymptotics of prediction in functional linear regression with functional outputs. Bernoulli. 19 (5B), 2627–2651 (2013)

[3] Dunford, N., Schwartz, J. T.: Linear operators, Vols. I & II. New York: Wiley.

Gaëlle Chagny
LMRS, UMR CNRS 6085, Université de Rouen Normandie, e-mail: gaelle.chagny@univ-rouen.fr

Anouar Meynaoui
LMRS, UMR CNRS 6085, Université de Rouen Normandie, e-mail: anouar.meynaoui@gmail.com

Angelina Roche
CEREMADE, UMR CNRS 7534, Université Paris Dauphine, e-mail: roche@ceremade.dauphine.fr

Reconstruction of motion signals with curvature and torsion

Perrine Chassat, Nicolas Brunel and Juhyun Park

Abstract

Among the many fields of exploration of motion capture is the very specific field of sign language involving movements of the body, hands, fingers, face and eyes and achieve a capacity for expression as rich and structured as that offered by speech [5]. These types of movement are interesting for us as they are particularly meaningful but are difficult to characterize without specific knowledge in the field. Our idea in treating the “motion" signals, in collaboration with the company MOCAPLAB [1], specialized in Motion Capture, is to incorporate current mathematical and statistical tools to extract “primitives" specific to the nature of the signals studied.

From the perspective of motor control theories, the axis of the analysis envisaged should be able to exploit the link between speed and motion geometry [4]. The starting point of our methodology consists in the estimation of trajectories resulting from motion capture and the characterization of the geometry and kinematics by appropriate functional representations. This simultaneous analysis is particularly possible in the trajectory of a point particle, using the Frenet-Serret frame [1, 2].

Our approach is to focus on estimating curvature and torsion [6, 7]. Indeed, the geometry of the trajectories of movement have physical significance: curvature and torsion characterize this geometry and can provide insightful summaries of kinetic curves. Discovering and exploiting these relationships require a good estimation of the functional parameters such as curvature and Frenet paths. This is a challenging statistical task as curvature and torsion depend on higher order derivatives and their estimation from real data (even with a low noise) can be very unstable.

Our work is motivated by the new development in [8], which offers a unified framework for the estimation of these functional quantities in the setting of functional data. A motivation of this type of analysis is to exploit the link between the curvilinear speed and the geometry of the trajectories (typically curvatures). Although the previous work is designed for multiple trajectories as an extension of functional data, it is still applicable to single curves and here we adopt the geometric framework for the analysis of single trajectories. Arguably, this is more challenging as the benefit of borrowing information from multiple curves is missing in the single curve setting.

Under this framework, the problem of estimating curvature and torsion can be treated as estimation of an ordinary differential equation in a Lie group. We develop a new algorithm for estimation by considering the perturbed Frenet-Serret ODE and solving the optimal control problem with a computationally fast Kalman filter [3]. The quality of the estimates is evaluated based on the quality of reconstruction of the trajectory by solving the Frenet-Serret ordinary differential equation. In particular, we show that our proposed method dominates the straightforward estimates of curvature and torsion based on the extrinsic formulas.

References

[1] Brunel, N. and Park, J. *Removing phase variability to extract a mean shape for juggling trajectories.* Electronic Journal of Statistics 8.2, 2014.

[2] Brunel, N. and Park, J. *The Frenet-Serret framework for aligning geometric curves.* International Conference on Geometric Science of Information. Springer, Cham, 2019.

[3] Clairon, Q. and Brunel, N. *Tracking for parameter and state estimation in possibly misspecified partially observed linear Ordinary Differential Equations* Journal of Statistical Planning and Inference. 2019.

[4] Flash, Tamar, and al. *Motor compositionality and timing: combined geometrical and optimization approaches.* Biomechanics of Anthropomorphic Systems. Springer, Cham, 2019.

[5] Gibet, S., Lefebvre-Albaret, F., Hamon, L., Brun R. and Turki, A. *Interactive editing in French Sign Language dedicated to virtual signers: requirements and challenges* Universal Access in the Information Society, Springer Verlag, 2016.

[6] Kim, K.-R., P. Kim, J.-Y. Koo, and M. Pierrynowski *Frenet-serret and the estimation of curvature and torsion.* IEEE Journal of Selected Topics in Signal Processing 7(4), 2013.

[7] Lewiner, T., J. Gomes, H. Lopes, and M. Craizer. *Curvature and torsion estimators based on parametric curve fitting.* Computers and Graphics 29(5), 2005.

[8] Park, J. and Brunel, N. *Mean curvature and mean shape for multivariate functional data under Frenet-Serret framework.* arXiv:1910.12049, 2019.

Perrine Chassat
Université Paris Saclay, CNRS, Univ. Evry, Laboratoire de Mathématiques et Modélisation d’Evry, e-mail: perrine.chassat@gmail.com

Nicolas Brunel
Université Paris Saclay, CNRS, ENSIIE, Laboratoire de Mathématiques et Modélisation d’Evry, e-mail: nicolas.brunel@ensiie.fr

Juhyun Park
Université Paris Saclay, CNRS, ENSIIE, Laboratoire de Mathématiques et Modélisation d’Evry, e-mail: juhyun.park@ensiie.fr

[1] https://www.mocaplab.com/fr/

Gaussian Graphical Modeling for Spectrometric Data Analysis

Laura Codazzi, Alessandro Colombi, Matteo Gianella, Raffaele Argiento, Lucia Paci and Alessia Pini

Abstract

Motivated by the analysis of spectrometric data, we introduce a functional graphical model for learning the conditional independence structure of infrared spectra. Infrared spectroscopy is a technique used to study chemical substances through the measurement of their infrared radiation spectra. To make those signals more meaningful it is important to understand which wavelength bands are related to the different components. The dependence structure between the signal at different wavelengths is particularly informative in this sense: if two different bands of the spectrum are dependent, we can conclude that they refer to the same components. Our goal is to investigate the structure of conditional dependence among different portions of an absorbance spectrum.

From a mathematical point of view, infrared spectra are continuous functions of the wavelength over a common domain, and are thus modeled as functional data. We interpret the analysis of dependence structure of spectrometric data as a smoothing problem of functional data analysis, followed by inference on the smoothing coefficients in a Bayesian framework. Several recent works focus on the problem of smoothing functional data. [4] proposed a Bayesian hierarchical model with Gaussian-Wishart processes for simultaneously smoothing multiple functional observations and estimating mean-covariance functions. However, their model suffers serious computational burden when data are observed on high-dimensional grids. To address the computational issue, [5] proposed to approximate the underlying true functional data with basis functions, and derive the induced Bayesian hierarchical model for the associated smoothing coefficients based on a Gaussian-Wishart prior.

Following the latter approach, we use B-spline basis expansion to represent the functional data. The B-spline basis expansion will also help us in the interpretation of the final dependency structure. Indeed, due to the compact support of B-spline basis functions, the conditional dependence between basis coefficients translates into conditional dependence between the corresponding portions of the domain of functional data. As a consequence, we can detect an association between different bands of the spectrum through a non-zero partial correlation between the corresponding smoothing coefficients.

The focus of this work is on the dependence structure of the coefficients. With a Bayesian perspective, this concerns the specification of a structured prior distribution for the smoothing coefficients. To do so, we frame our approach in to the graphical modeling setting, assuming a Gaussian graphical prior for the basis expansion coefficients. Graphical models describe a mapping between a graph and a family of multivariate probability models [2]. In this framework, the structure of the underlying graph is unknown and is estimated on the basis of the available data (strictural learning). This approach allows us to infer the structure of conditional dependence between basis coefficients. To the best of our knowledge, this is the first work where a Gaussian graphical model is assumed as a prior model for the basis expansion coefficients in the analysis of functional data.

On the computational side, we design an efficient sampling strategy to approximate the joint posterior distribution of the graph and model parameters. Specifically, the strategy builds upon the birth and death Markov Chain Monte Carlo algorithm of [3]. We illustrate our method on a real data set, studying the infrared absorbance spectra of strawberry purees. Further details on our proposed method can be found in [1].

References

[1] Codazzi, L., Colombi, A., Gianella, M., Argiento, R., Paci, L., Pini, A.: Functional graphical model for spectrometric data analysis, arXiv:2103.11666 (2021)

[2] Lauritzen, S. L.: *Graphical Models*. Oxford University Press, Oxford (1996)

[3] Mohammadi, A., Wit, E. C.: Bayesian structure learning in sparse Gaussian graphical models, Bayesian Analysis **10** (2015) 109–138 doi: 10.1214/14-BA889

[4] Yang, J., Zhu, H., Choi, T., Cox, D. D.: Smoothing and mean covariance estimation of functional data with a Bayesian hierarchical model. Bayesian Analysis **11** (2016) 649–670 doi: 10.1214/15-BA967

[5] Yang, J., Cox, D. D., Lee, J. S., Ren, P., Choi, T.: Efficient Bayesian hierarchical functional data analysis with basis function approximations using Gaussian Wishart processes. Biometrics **73** (2017) 1082–1091 doi: doi.org/10.1111/biom.12705

Laura Codazzi, Alessandro Colombi, Matteo Gianella
Department of Mathematics, Politecnico di Milano, e-mail: laura.codazzi@mail.polimi.it, e-mail: alessandro3.colombi@mail.polimi.it, e-mail: matteo1.gianella@mail.polimi.it

Raffaele Argiento, Lucia Paci, Alessia Pini
Department of Statistical Sciences, Università Cattolica del Sacro Cuore, e-mail: raffaele.argiento@unicatt.it, e-mail: lucia.paci@unicatt.it, e-mail: alessia.pini@unicatt.it

Probabilistic local clustering of misaligned functional data: analysis of Italian COVID-19 death curves

Marzia A. Cremona, Tobia Boschi and Francesca Chiaromonte

Abstract

On February 19, 2020, the first non-travel-related COVID-19 case in Italy was diagnosed in Codogno (Lombardia). Since then, the number of identified cases has rapidly increased and Italy has become one of the most hardly hit countries in the world by the COVID-19 pandemic. As of mid-May 2021, Italy saw a total of 4.2 million cases and more than 120,000 COVID-attributed deaths, corresponding to a mortality of 206 deaths per 100,000 inhabitants. A striking aspect of COVID-19 pandemic in Italy has been its heterogeneity. Indeed, Italian regions were hit at different times and with different strengths, especially during the first wave. Hence, comparing the evolution of the pandemic across regions can provide important insights on the role of underlying contributors to this heterogeneity.

We consider official COVID-19 death curves, as well as excess mortality curves due to COVID-19 – estimated as the daily difference between 2020 deaths and average deaths in the period 2015-2019 – for the 20 Italian regions. The goal is to cluster these misaligned functional data, in order to assess whether there are regions sharing similar pandemic patterns [2]. Importantly, we are looking for clusters based on a local similarity among curves, since patterns might differ only on a (misaligned) portion of the domain.

We develop probabilistic K-mean with local alignment (probKMA), a new functional data analysis method to locally cluster a set of (possibly multidimensional) curves and discover functional motifs, i.e. typical "shapes" that may recur several times along and across the curves capturing important local characteristics of these curves [1]. This method leverages ideas from functional data analysis (joint clustering and alignment of curves), bioinformatics (local alignment through the extension of high similarity seeds) and fuzzy clustering (curves belonging to more than one cluster, if they contain more than one typical"shape"). It can employ various dissimilarity measures and incorporate derivatives in the discovery process, thus exploiting complex facets of shapes.

Using probKMA as a probabilistic clustering method to group COVID-19 death curves and excess mortality curves based on their local similarity, we find two starkly different first waves of COVID-19 pandemics; an "exponential" one unfolding in Lombardia and the worst-hit areas of the north, and a milder, "flat(tened)" one in the rest of the country – including Veneto, where cases appeared concurrently with Lombardia but aggressive testing was implemented early on. Local alignments of curves provide an indication of the lags between different regions, which can be employed in subsequent analyses to associate patterns of mortality with functional covariates such as mobility and positivity [2].

References

[1] Cremona, M.A., Chiaromonte F.: Probabilistic K-mean with local alignment for clustering and motif discovery in functional data. arXiv 1808.04773 (2020)

[2] Boschi, T., Di Iorio, J., Testa, L., Cremona, M.A., Chiaromonte F.: The shapes of an epidemic: using functional data analysis to characterize COVID-19 in Italy. arXiv 2800.04700 (2020)

Marzia A. Cremona
Department of Operations and Decision Systems, Université Laval, Québec, G1V 0A6, Canada e-mail: marzia.cremona@fsa.ulaval.ca
CHU de Quebec – Université Laval Research Center, Québec, G1V 4G2, Canada

Tobia Boschi
Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA e-mail: tub37@psu.edu

Francesca Chiaromonte
Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA e-mail: fxc11@psu.edu
Institute of Economics and EMbeDS, Sant'Anna School of Advanced Studies, Pisa, 56127, Italy

A Conformal approach for multivariate functional data prediction

Jacopo Diquigiovanni, Matteo Fontana and Simone Vantini

Abstract

A crucial challenge in Functional Data Analysis (FDA) is the issue of uncertainty quantification in prediction. Intuitively, we are interested in creating prediction sets, namely subsets of the sample space including a new functional observation with a certain nominal confidence level $1-\alpha$. Only very recent works in FDA provide some knowledge into this theoretical (but yet full of applied repercussions) issue. These approaches can be classified in two groups: the first one consists of works principally based on parametric bootstrapping techniques [e.g., 4, 2], the second one is characterized by the application of dimensionality reduction techniques to manage the naturally infinite dimensionality [e.g., 5, 1]. Both groups carry obvious drawbacks since they are either based on not easily provable distributional assumptions and/or on asymptotic results. In addition, the first class of approaches is computationally demanding, whereas the second one relies on the approximations induced by basis projection.

Focusing on a general regression framework in which independent and identically distributed regression data consist of a multivariate functional response variable and a set of (not necessarily scalar) covariates, we propose a procedure able to bypass the main methodological shortcomings identified in the previous literature. To do that, the framework we consider is Conformal Prediction [6], a novel method of forecasting firstly developed in the Machine Learning community as a way to define prediction intervals for Support Vector Machines. Specifically, we introduce a new family of nonconformity measures which outputs closed-form finite-sample exact prediction sets - i.e. sets ensuring a coverage equal to the nominal confidence level - for the multivariate functional response variable under no assumptions other than i.i.d. regression data. Moreover, our proposal ensures that the prediction sets obtained are multivariate functional bands, an essential feature in the functional setting that allows the visualization and interpretation of such sets. The procedure does not rely on functional dimension reduction techniques and is scalable, because - conditional on the computational cost required to calculate the regression estimates - the time required to compute the multivariate prediction band increases linearly with the sample size. Since the procedure creates exact prediction bands regardless the (true) regression function and the regression estimator used, it can be applied in a wide range of application scenarios. A criterion that naturally arises in the prediction framework to discriminate between the elements belonging to this family of nonconformity measures is maximization of efficiency, i.e. minimization of the size of prediction sets. The reason of this choice is very intuitive: since prediction bands are, by construction, exact, one is justified in seeking small prediction bands because they include subregions of the sample space where the probability mass is concentrated. Within this family of measures, we therefore propose a specific conformal predictor that modulates the width of the multivariate band over the domains based on the local behavior and magnitude of the functional data and which is able to create prediction bands asymptotically no less efficient than those with constant width. The work extends the method detailed in [3] to multivariate functional data and to a regressive framework.

References

[1] Antoniadis, A., Brossat, X., Cugliari, J., Poggi, J.M.: A prediction interval for a function-valued forecast model: Application to load forecasting. Int. J. Forecast. **32**, 939–947 (2016)

[2] Cao, G., Yang, L., Todem, D.: Simultaneous Inference For The Mean Function Based on Dense Functional Data. J. Nonparametr. Stat. **24**, 359–377 (2012)

[3] Diquigiovanni, J., Fontana, M., Vantini, S.: The Importance of Being a Band: Finite-Sample Exact Distribution-Free Prediction Sets for Functional Data. arXiv preprint arXiv:2102.06746. (2021)

[4] Degras, D.A.: Simultaneous confidence bands for nonparametric regression with functional data. Statist. Sinica. **21**, (2011)

[5] Hyndman, R.J., Shahid Ullah, M.: Robust forecasting of mortality and fertility rates: A functional data approach. Comput. Statist. Data Anal. **51**, 4942–4956 (2007)

[6] Vovk, V., Gammerman, A., Shafer, G.: Algorithmic learning in a random world. Springer Science & Business Media (2005)

Jacopo Diquigiovanni
Department of Statistical Sciences, University of Padova, Italy, e-mail: jacopo.diquigiovanni@phd.unipd.it

Matteo Fontana
Joint Research Centre - European Commission, Ispra (VA), Italy, e-mail: matteo.fontana@ec.europa.eu

Simone Vantini
MOX - Department of Mathematics, Politecnico di Milano, Italy, e-mail: simone.vantini@polimi.it

Plug-in classification procedures for diffusion paths

Eddy Ella-Mintsa, ChristopheDenis, CharlotteDion and Viet-Chi Tran

Abstract

Recent advents in modern technology have generated labeled data, recorded at high frequency, that can be modelled as functional data. This work focuses on multiclass classification problem for functional data modelled by a stochastic differential equation. The drift function depends on the label of the class $Y \in \{1, ..., K\}, K \in \mathbb{N} \setminus \{0, 1\}$. An observation is a solution $X = (X_t)_{t\in[0,1]}$ of the following time-homogeneous stochastic differential equation

$$dX_t = b^*_Y(X_t)dt + \sigma^*(X_t)dW_t, \;\; x_0 = 0, \;\; t \in [0, 1],$$

with unknown drift functions $(b^*_k)_{k=1,...,K}$ and unknown diffusion coefficient σ^*. Furthermore, we assume that the law $p = (p_1, ..., p_K)$ of the label Y is unknown. From a learning sample $D_N = \left(X^{(i)}, Y_i\right)_{i\in[[1,N]]}$ that consists of N independent copies of the pair (X, Y), our aim is to build an implementable *plug-in* nonparametric classification procedure and derive upper bounds of its excess risk over Hölder spaces.

Few works have investigated the classification of functional data in the stochastic differential equation framework. In [3], it is done from a parametric point of view with a known diffusion coefficient. Here, we deal with a more challenging problem : the drift function is nonparametric, plus, the diffusion coefficient and the distribution $p = (p_1, ..., p_K)$ are unknown. Our classification procedure relies on the nonparametric estimation of functions $b^*_k, k = 1, ..., K$ and σ^{*2} minimizing a least-squares contrast over a spline basis (as in [1] for the estimation of the drift function). We establish the consistency of the resulting empirical classifier as a function of N, the size of the learning sample and $n \in \mathbb{N}^*$, the number of discrete observations for each path. We obtain rates of convergence under mild assumptions. These computations rely here on stochastic calculus, in particular fine estimate of the transition densities. Finally, the obtained empirical classifier is implemented and successfully evaluated from simulated data.

References

[1] Denis C., Dion C., Martinez M. : A ridge estimator of the drift from discrete repeted observations of the solutions of a stochastic differential equation, Bernouilli (2020)

[2] Comte F., Genon-Catalot V., Rozenholc Y. : Penalized nonparametric mean square estimation of the coefficients of diffusion processes, Bernouilli Society for Mathematical Statistics and Probability (2007)

[3] Denis C., Dion C., Martinez M.: Consistent procedures for multiclass classification of descrete diffusion paths, Scandinavian Journal of Statistics, 47(2):516-554 (2020)

[4] Gadat S., Gerchinovitz S., Marteau C.: Optimal functional supervised classification with separation condition. Bernouilli, 26(3):1797-1831 (2020)

[5] Cadre B.: Supervised classification of diffusion paths, Mathematical Methods of Statistics, 22(3):213-225, Springer (2013)

Eddy Ella-Mintsa
LAMA UMR 8050, 5 Boulevard Descartes, 77454 Marne-la-Vallée cedex 2, e-mail: eddy-michel.ella-mintsa@univ-eiffel.fr

A Review of Semi-Functional Partial Linear Regression model and their Extensions

Mohammad Fayaz

Abstract

The research article titled "Semi-Functional Partial Linear Regression" by Germán Aneiros-Pérez and Philippe Vieu was published in 2006 in the Statistics and Probability Letters. The main idea of this model is to consider both the functional and non-functional, or mixed and hybrid, covariates for the prediction of the real-valued response. It used the non-parametric method for functional covariate with the weights of the functional version of the Nadaraya–Watson and the parametric method for the non-functional covariate with the linear relation. In the next few years, the authors and some researchers extend this model and recently many other extensions are published. In this study, we search these extensions with Google Scholar from 2006 to 2020 and there is at least 200 result. In the next step, the inclusion criteria are published in the ISI-indexed journals and the exclusion criteria are non-English, preprints, thesis, conference proceedings, and slides. Finally, 85 research articles were selected. The results are summarized in some tables, for example they categorized into the following main topics: time-series, quintile regression, varying coefficient model, statistical testing, robust estimation, Bayesian estimation, multi-functional covariates, variable selection, Confidence bands and prediction intervals, missing data, errors in variable and others. And there are also different applications such as in spectroscopy, in air-pollution and related topics, child growth study, neuroimaging, electricity demand and price, and others. Most of them are published in statistical journals but some of them are published in neuroscience, energy, and mathematics journals. Finally, an application of this model with the fda.usc package is presented.

Mohammad Fayaz
PhD graduate in Biostatistics, Shahid Beheshti University of Medical Sciences, Tehran, Iran. ORCID : 0000000256439763,
e-mail: Mohammad.Fayaz.89@gmail.com

A Functional Data Analysis Approach to the Estimation of Densities over Complex Regions

Federico Ferraccioli, Eleonora Arnone, Livio Finos, James O. Ramsay and Laura M. Sangalli

Abstract

We propose a nonparametric method for density estimation over (possibly complicated) spatial domains. Following a functional data analysis approach, we consider a penalized likelihood estimator, with a roughness penalty based on a differential operator. We demonstrate the good inferential properties of the method. Moreover, we develop an estimation procedure based on advanced numerical techniques, and in particular making use of finite elements. This ensures high computational efficiency and enables great flexibility. The proposed method efficiently deals with data scattered over regions having complicated shapes, featuring complex boundaries, sharp concavities or holes. Moreover, it captures very well complicated signals having multiple modes with different directions and intensities of anisotropy. We show the comparative advantages of the proposed approach over state of the art methods, in simulation studies and in an application to the study of criminality in the city of Portland, Oregon.

Keywords: differential regularization; finite elements; heat diffusion density estimator; functional data analysis.

The talk is based on:
Ferraccioli, F., Arnone, E., Finos, L., Ramsay, J. O., Sangalli, L. M.: Nonparametric density estimation over complicated domains. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83(2), 346-368 (2021)

Federico Ferraccioli
Dipartimento di Scienze Statistiche, Via Cesare Battisti, 241, 35121 Padova, Italy, e-mail: ferraccioli@stat.unipd.it

Eleonora Arnone
MOX-Dipartimento di Matematica, Piazza L. da Vinci, 32, 20133 Milano, Italy, e-mail: eleonora.arnone@polimi.it

Livio Finos
Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Via Venezia, 8, 35131 Padova, e-mail: livio.finos@unipd.it

James O. Ramsay
Department of Psychology, 1205 Dr Penfield Avenue, Montreal, QC, e-mail: ramsay@psych.mcgill.ca

Laura M. Sangalli
MOX-Dipartimento di Matematica, Piazza L. da Vinci, 32, 20133 Milano, Italy, e-mail: laura.sangalli@polimi.it

Permutation Approach to Testing for Effect of Covariates in the Spatial Regression Model with Functional Response

Eva Fišerová, Veronika Římalová, Alessandra Menafoglio and Alessia Pini

Abstract

The aim of this contribution is to introduce an approach to hypotheses testing in a functional linear model for spatial data. The proposed method can deal with the spatial structure of data by building a permutation testing procedure on spatially filtered residuals of a spatial regression model. Indeed, due to the spatial dependence existing among the data, the residuals of the regression model are not exchangeable, breaking the basic assumptions of the Freedman and Lane permutation scheme. Instead, it is proposed here to base the permutation test on approximately exchangeable spatially filtered residuals, i.e. the variance-covariance structure of the residuals is estimated by variography and then the correlation of the residuals is removed by a spatial filtering. To evaluate the performance of the proposed method in terms of empirical size and power, a simulation study, examining its behaviour under different covariance settings, is conducted. It will be shown that neglecting the residuals spatial structure in the permutation scheme (thus permuting the correlated residuals directly) yields a very liberal testing procedures, whereas the proposed procedure based on spatially filtered residuals is close to the nominal size of the test. The methodology will be demonstrated on a real world data set on the amount of waste production in the Venice province of Italy.

References

[1] Menafoglio, A., Secchi, P.: Statistical analysis of complex and spatially dependent data: a review of object oriented spatial statistics. European journal of operational research, **258**(2), 401–410 (2017)

[2] Ramsay, J.O., Silverman, B.W.: Functional data analysis. New York, NY, Springer (2013)

[3] Římalová, V., Fišerová, E., Menafoglio, A., Pini, A.: Inference for Spatial Regression Models with Functional Response using a Permutational Approach. Journal of Multivariate Analysis (*under review*)

Eva Fišerová
Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc, Olomouc, Czech Republic, e-mail: eva.fiserova@upol.cz

Veronika Římalová
Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc, Olomouc, Czech Republic, e-mail: veronika.rimalova01@upol.cz

Alessandra Menafoglio
MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy, e-mail: alessandra. menafoglio@polimi.it

Alessia Pinni
Department of statistical sciences, Università Cattolica del Sacro Cuore, Milan, Italy, e-mail: alessia.pini@unicatt.it

A nonparametric bootstrap method for functional data

Miguel Flores, Rubén Fernández-Casal and Sergio Castillo-Páez

Abstract

In this work, a nonparametric bootstrap method, which can be used to approximate characteristics of the distribution of some statistic derived from functional data, is described. This approach is an adaptation of the method proposed in [2] for inference on spatial data. In this case, the observed data consists of repeated (and independent) realizations of a continuous one-dimensional process.

Let us assume that $\mathcal{S}_n = \{Y_i(t)\}_{i=1}^n$, for $t \in [a, b] \subset \mathbb{R}$, is a set of n independent observations of a functional variable $Y(t)$ defined over $\mathbb{R}$, verifying:

$$Y_i(t) = \mu(t) + \varepsilon_i(t),$$

being $\mu(t)$ the deterministic trend function y $\varepsilon_i(t)$ a random error process with zero mean and covariances $Cov\,(\varepsilon_i(t), \varepsilon_{i'}(t')) = \delta_{ii'} C\,(\|t - t'\|)$, for $1 \leq i, i' \leq n$ and $a \leq t, t' \leq b$, where $\delta_{ii'} = 1$ if $i = i'$, $\delta_{ii'} = 0$ if $i \neq i'$ and $C(\cdot)$ is the covariogram function.

In practice, each $Y_i(t)$ is observed in a discrete set of points $t_j \in [a, b] \subset \mathbb{R}$, with $j = 1, \ldots, p$. Then, these set of observations can be expresed as a matrix $\mathbf{Y}$ of order $n \times p$, with $\mathbf{Y}_{ij} = Y_i(t_j)$. Furthermore, if $\mathbf{y}_i = \left(Y_i(t_1), \ldots, Y_i(t_p)\right)^\top$ is the vector corresponding to the i-th row of $\mathbf{Y}$, its covariance matrix $Cov(\mathbf{y}_i) = \Sigma_0$ (within-curve covariance matrix) has $(\Sigma_0)_{jj'} = C\left(\left\|t_j - t_{j'}\right\|\right)$, for $i = 1, \ldots, n$.

The proposed procedure starts with the nonparametric estimation of the trend and the dependence, following an iterative algorithm similar to that described in [2]. The local linear estimator is used to estimate the trend, using a bandwidth that takes the temporal dependence into account, and a Shapiro-Botha variogram model is fitted to pilot variograma estimates, from which a valid covariogram estimate can be derived. Then, the following bootstrap algorithm is applied:

1. Compute the residuals matrix $\mathbf{R}$, with $\mathbf{r}_i = \mathbf{y}_i - \hat{\boldsymbol{\mu}}$, where $\hat{\boldsymbol{\mu}} = \left(\hat{\mu}(t_1), \ldots, \hat{\mu}(t_p)\right)^\top$ are the local linear trend estimates at the discretization points.
2. Obtain an estimate $\hat{\boldsymbol{\Sigma}}_0$ of the within-curve covariance matrix and its Cholesky decomposition $\hat{\boldsymbol{\Sigma}}_0 = \mathbf{U}^\top\mathbf{U}$.
3. Compute the uncorrelated data $\mathbf{E} = \mathbf{R}\mathbf{U}^{-1}$ and center them (by subtracting from them the overall sample mean).
4. Use the centered values to derive an independent bootstrap sample $\mathbf{E}^*$, by resampling the rows and columns of $\mathbf{E}$.
5. Compute the bootstrap errors $\boldsymbol{\varepsilon}^* = \mathbf{E}^*\mathbf{U}$.
6. Obtain the bootstrap sample $\mathbf{Y}^*$, with $\mathbf{y}_i^* = \hat{\boldsymbol{\mu}} + \boldsymbol{\varepsilon}_i^*$, for $i = 1, \ldots, n$.
7. Repeat B times steps 4-6 to obtain the B bootstrap replicates $\left\{\mathbf{Y}_1^*, \ldots, \mathbf{Y}_B^*\right\}$.

The replicates derived from this algorithm can be used to approximate characteristics of the distribution of a statistic under study. For instance they can be used for approximating the standard error and bias of an estimator.

Numerical studies were carried out to study the behavior of this bootstrap procedure under different scenarios. Specifically, we checked its performance to approximate the bias and variance of two trend estimators, the sample mean and the local linear trend estimator. The results were compared with those obtained by using the smoothed bootstrap proposed in [1].

References

[1] A. Cuevas, M. Febrero, R. Fraiman, On the use of the bootstrap for estimating functions with functional data, Computational Statistics & Data Analysis **51** 1063-1074(2006)

[2] S. Castillo-Páez, R. Fernández-Casal, P. García-Soidán, A nonparametric bootstrap method for spatial data, Computational Statistics & Data Analysis **137** 1-15 (2019)

Miguel Flores
MODES, SIGTIG, Dep. de Matemática, Escuela Politécnica Nacional, Ecuador e-mail: miguel.flores@epn.edu.ec

Rubén Fernández-Casal
Dep. de Matemáticas, Universidade da Coruña, Spain e-mail: ruben.fcasal@udc.es

Sergio Castillo-Páez
Dep. de Ciencias Exactas, Universidad de las Fuerzas Armadas ESPE, Ecuador e-mail: sacastillon@espe.edu.ec

Factor-augmented Model for Functional Data

Yuan Gao, Han Lin Shang and Yanrong Yang

Abstract

We propose modeling raw functional data as a mixture of a smooth function and a highdimensional factor component. The conventional approach to retrieving the smooth function from the raw data is through various smoothing techniques. However, the smoothing model is not adequate to recover the smooth curve or capture the data variation in some situations. These include cases where there is a large amount of measurement error, the smoothing basis functions are incorrectly identified, or the step jumps in the functional mean levels are neglected. To address these challenges, a factor-augmented smoothing model is proposed, and an iterative numerical estimation approach is implemented in practice. Including the factor model component in the proposed method solves the aforementioned problems since a few common factors often drive the variation that cannot be captured by the smoothing model. Asymptotic theorems are also established to demonstrate the effects of including factor structures on the smoothing results. Specifically, we show that the smoothing coefficients projected on the complement space of the factor loading matrix is asymptotically normal. As a byproduct of independent interest, an estimator for the population covariance matrix of the raw data is presented based on the proposed model. Extensive simulation studies illustrate that these factor adjustments are essential in improving estimation accuracy and avoiding the curse of dimensionality. The superiority of our model is also shown in modeling Australian temperature data and Canadian weather data.

Yuan Gao
Australian National University, Acton 2601, Australia, e-mail: yuan.gao@anu.edu.au

Han Lin Shang
Macquarie University, Macquarie Park NSW 2109, Australia, e-mail: hanlin.shang@mq.edu.au

Yanrong Yang
Australian National University, Acton 2601, Australia, e-mail: yanrong.yang@anu.edu.au

Goodness-of-fit tests for functional linear models based on integrated projections

Eduardo García-Portugués, Javier Álvarez-Liébana,
Gonzalo Álvarez-Pérez, and Wenceslao González-Manteiga

Abstract

Functional linear models are one of the most fundamental tools to asses the relation between two random variables of a functional or scalar nature. In this talk, we present a goodness-of-fit test for the functional linear model with functional response that neatly adapts to functional/scalar responses/predictors. In particular, the new goodness-of-fit test extends a previous proposal for scalar response. The test statistic can be seen as a weighted quadratic norm of the functional residuals, is based on a convenient regularized estimator, is easy to compute, and is calibrated through an efficient bootstrap resampling. Comparative simulations for the simple hypothesis of no effect and the composite hypothesis empirically show the appropriateness of the advocated test. A graphical diagnostic tool, useful to visualize the possible departures from the functional linear model, is introduced and illustrated with a couple of data applications. The companion R package `goffda` implements the proposed methods and allows for the reproducibility of the data applications.

The talk is based on the paper [2] and package [1].

References

[1] García-Portugués, E., Álvarez-Liébana, J.: goffda: Goodness-of-fit tests for functional data (2020). https://CRAN.R-project.org/package=goffda. R package version 0.0.7

[2] García-Portugués, E., Álvarez-Liébana, J., Álvarez-Pérez, G., González-Manteiga, W.: A goodness-of-fit test for the functional linear model with functional response. Scand. J. Stat., to appear.

Eduardo García-Portugués
Department of Statistics, Carlos III University of Madrid, Avda. Universidad 30, 28911 Leganés (Spain) e-mail: edgarcia@est-econ.uc3m.es

Javier Álvarez-Liébana
Department of Statistics and Operations Research and Mathematics Didactics, University of Oviedo, C/ Federico García Lorca, 18, 33007 Oviedo (Spain)
e-mail: alvarezljavier@uniovi.es

Gonzalo Álvarez-Pérez
Department of Physics, University of Oviedo, C/ Federico García Lorca, 18, 33007 Oviedo (Spain) e-mail: gonzaloalvarez@uniovi.es

Wenceslao González-Manteiga
Department of Statistics, Mathematical Analysis and Optimization, University of Santiago de Compostela, Rúa Lope Gómez de Marzoa s/n, 15782 Santiago de Compostela (Spain) e-mail: wenceslao.gonzalez@usc.es

Constructing a T^2 Hotelling Control Chart Using Functional Data

Priscila Guayasamín, Miguel Flores, Rubén Fernández-Casal, Salvador Naya and Javier Tarrío-Saavedra

Abstract

In this work a non-parametric control chart for functional data is proposed. We assume that the process is under control, being $\mathbf{X}_1, ..., \mathbf{X}_n$ the reference sample (n i.d.d. observations of a functional random variable), and we want to check if a new observation $\mathbf{X}_0$ deviates from de distribution of the reference sample (see e.g. [3]). For this purpose, we use the generalized Hotelling's T^2 proposed in [1], for a separable Hilbert space $\mathbb{H}$, and we propose a bootstrap procedure to approximate its distribution in phase II.

For simplicity, we hereby describe the procedure for the case in which $\mathbb{H} = \mathbb{R}^p$. For $p < n$, the generalized Hotelling's T^2 statistic coincides with the traditional expression $T^2 = n\left(\bar{\mathbf{X}} - \mathbf{m}\right)^\top S^{-1}\left(\bar{\mathbf{X}} - \mathbf{m}\right)$, where $\mathbf{m}$ is the theoretical mean, $\bar{\mathbf{X}}$ the sample mean and S the sample covariance. When $p > n$, [2] suggests the generalization $T^2 = n\left(\bar{\mathbf{X}} - \mathbf{m}\right)^\top S^{+}\left(\bar{\mathbf{X}} - \mathbf{m}\right)$, where S^+ is the Moore-Penrose inverse of the sample covariance matrix S.

The **general procedure** to monitor a new observation $\mathbf{X}_0$ (monitoring observation) in phase II is as follows:

1. Compute the mean $\bar{\mathbf{X}}$, and the Moore-Penrose inverse S^+ of the sample covariance matrix of the reference sample.
2. Obtain the observed value of the statistic $T^2 = \left(\mathbf{X}_0 - \bar{\mathbf{X}}\right)^\top S^+\left(\mathbf{X}_0 - \bar{\mathbf{X}}\right)$
3. Assuming that the process is under control, for a fixed significance level α, approximate the upper control limit UCL using the bootstrap algorithm described below.
4. Generate the control chart to monitor the new observation. The process is considered out-of-control if $T^2 \geq UCL$.

The **bootstrap algorithm** to approximate the quantile of the distribution of the statistic is as follows:

1. Generate a boostrap sample $\mathbf{X}^* = \{\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*\}$ from de reference sample, by generating a set of random indices with replacement from $\{1, \ldots, n\}$.
2. Compute the mean $\bar{\mathbf{X}}^*$, and the Moore-Penrose inverse S^{+*} of covariance of the boostrap sample.
3. Select a random index $i_0 \in I_0$, where I_0 is the set of indices not included in the bootstrap sample, and set $\mathbf{X}_0^* = \mathbf{X}_{i_0}$.
4. Compute the bootstrap version of the statistic $T^{2*} = \left(\mathbf{X}_0^* - \bar{\mathbf{X}}^*\right)^\top S^{+*}\left(\mathbf{X}_0^* - \bar{\mathbf{X}}^*\right)$
5. Repeat steps 1-4 a large number of times to obtain B replicates of the statistic (steps 3 and 4 can be repeated many times).
6. Compute UCL as the $1 - \alpha$ empirical quantile of the bootstrap replicates, $UCL = T^{2*}_{1-\alpha}$.

The behavior the proposed method was analyzed by several simulation studies with different sample sizes and number of discretization points. Considering changes in the mean of the process in terms of the magnitude and shape, a greater power was observed when the number of discretization points is close to the sample size. Additionally, this control chart seems to be more sensitive to changes in the shape of the mean.

References

[1] A. Pini , A. Stamm , S. Vantini, Hotelling's T^2 in separable Hilbert spaces. Journal of Multivariate Analysis, **167**, 284-305, (2018).

[2] P. Secchi, A. Stamm, S. Vantini. Inference for the mean of large p small n data: A finite-sample high-dimensional generalization of Hotelling's theorem. Electronic journal of statistics, **7**, 2005-2031, (2013).

[3] M. Flores, S. Naya, R. Fernández-Casal, S. Zaragoza, P. Raña, J. Tarrío-Saavedra, Constructing a control chart using functional data. Mathematics, **8**, 58, (2020).

Priscila Guayasamín
Dep. de Matemática, Escuela Politécnica Nacional, Ecuador, e-mail: priscila.guayasamin@epn.edu.ec

Miguel Flores
MODES,SIGTIG, Dep. de Matemática, Escuela Politécnica Nacional, Ecuador e-mail: miguel.flores@epn.edu.ec e-mail: miguel.flores@epn.edu.ec

Rubén Fernández-Casal
Dep. de Matemáticas, Universidade da Coruña, Spain, e-mail: ruben.fcasal@udc.es

Salvador Naya
MODES, CITIC, ITMATI, Universidade da Coruña, Escola Politécnica Superior, Mendizábal s/n, Ferrol, Spain , e-mail: salva@udc.es

Javier Tarrío-Saavedra
MODES, CITIC, Universidade da Coruña, Escola Politécnica Superior, Mendizábal s/n, Ferrol, Spain , e-mail: javier.tarrio@udc.es

An Application of Multivariate Functional Data Analysis in Sports Biomechanics

Edward Gunning, Liwen Zhang, Drew Harrison and Norma Bargary

Abstract

In sports biomechanics, functional data analysis (FDA) allows the streams of data that are measured continuously during a movement, such as joint angles or forces, to be modelled as smooth, time-varying functions. When multiple streams of biomechanical data are observed concurrently (e.g., the hip and the knee angle), they can be considered jointly as multivariate functional data. This is useful for studying coordination, as it considers the cross-covariance between different joints, allowing their complex interactions to be examined during during dynamic sporting tasks. Our application of multivariate FDA concerns a study of hip-knee coordination of the kicking leg in soccer. Kicking is the most common and complex movement in soccer, where inter-joint coordination plays a crucial role. In our data set, soccer players performed two kicking tasks, instep kicking (the ball is kicked from the ground) and punt kicking (the ball is kicked in the air after leaving the hands, i.e. a goalkeeper's kick). The data also had a repeated-measures structure due to the soccer players performing multiple replicates. A mixed-effects modelling approach was adopted, which allowed task-specific effects (fixed effects) and individual-level variation (random effects) to be modelled. Initially, we applied bivariate functional principal components anlaysis (Ramsay and Silverman, 2005) (BFPCA) to the combined hip and knee angle data and then modelled the BFPCA scores post-hoc using a *scalar* linear mixed effects model. More recently, we modelled the hip-knee angle data directly in a bivariate *functional* linear mixed effects model, using the multivariate functional additive mixed model (multiFAMM) methodology of Volkmann et al. (2021). This approach is useful for studying coordination - it preserves the time-dependent structure of biomechanical data (*functional*), considers the cross-covariance between joints (*multivariate*), and allows task-specific constraints (*fixed effects*) and individual-level differences (*random effects*) to be modelled. We will mention some considerations and possibile extensions for the application of these models in sports biomechanics.

References

[1] Ramsay, J. O., Silverman, B. W.: Functional Data Analysis. Springer, New York (2005)

[2] Volkmann, Alexander, et al. : Multivariate Functional Additive Mixed Models. arXiv preprint arXiv:2103.06606 (2021).

Edward Gunning
University of Limerick, Ireland. e-mail: Edward.Gunning@ul.ie

Liwen Zhang
Beijing Sport University, China. e-mail: Liwen.Zhang@ul.ie

Drew Harrison
University of Limerick, Ireland. e-mail: Drew.Harrison@ul.ie

Norma Bargary
University of Limerick, Ireland. e-mail: Norma.Bargary@ul.ie

Mode regression for functional data

Chaima Hebchi

Abstract

In nonparametric statistics, many studies are carried out to give powerful tools to model and study the relationship between the response variable. The nonparametric mode regression function has long been a question of great interest in a wide range of fields for instance in econometrics, biology, astronomy...

The aim of this abstract is to join the advantages of mode with regression function by using local linear method. When the high dimensional causes many problems in nonparametric mode regression, we attempt to project the explanation of Y given X on one functional direction.

In what follow, we study the uniform almost complete convergence of mode regression estimator then we establish the uniform almost complete convergence of our estimator in the single functional index modelling.

uniform almost complete convergence of mode regression

Under some conditions also the following notations, $\hat{M}r$: is the mode regression estimator of Mr, $Mr(x) = \sup_{x\in\mathcal{F}} m(x)$, with : $\mathcal{F}$: semi metric and $m(.)$ is regression function. For $S_{\mathcal{F}} \subset \cup_{k=1}^{d_n} B(x_k, r_n)$ where : r_n (resp d_n) is a sequence of positive real numbers, $B(x_k, r_n)$ the ball centered at x_k with radius r_n. we can get the next result

Theorem 1.[1]

$$\sup_{x\in S_{\mathcal{F}}} |\hat{M}r(x) - Mr(x)| = O(h^b) + O_{a.co.}\left(\sqrt{\frac{\ln d_n}{n\phi_x(h)}}\right)$$

with : $\phi_x(h) = \mathbb{P}[X \in B(x_k, h)]$, h is chosen as a sequence of positive real numbers and convergences to 0 when $n \longrightarrow \infty$.

mode regression estimator in the single functional index modelling

At this stage, we observe n pairs (X_i, Y_i) for $i = 1, ..., n$ identically distributed as (X, Y), this last is valude in $\mathcal{F} \times \mathbb{R}$, where $\mathcal{F}$ is a Hilbertian space and there exists a $\theta \in \Theta_{\mathcal{F}} \subset \mathcal{F}$ such that : $\mathbb{E}[Y|X] = \mathbb{E}[Y| < \theta, X >]$ in order to establish the uniform almost complete convergence we have to take the following notions $S_{\mathcal{F}} \subset \cup_{k=1}^{d_n^{S_{\mathcal{F}}}} B(x_k, r_n)$, $\Theta_{\mathcal{F}} \subset \cup_{j=1}^{d_n^{\Theta_{\mathcal{F}}}} B(t_j, r_n)$ with, $k(x) = \arg \min_{k\in\{1,...,d_n^{S_{\mathcal{F}}}\}} ||x - x_k||$, $k'(\theta) = \arg \min_{k'\in\{1,...,d_n^{\Theta_{\mathcal{F}}}\}} ||\theta - t_{k'}||$, with $(x_k, t_{k'}) \in \mathcal{F}^2$ and $r_n, d_n^{S_{\mathcal{F}}}, d_n^{\Theta_{\mathcal{F}}}$ are a sequences of positive real numbers and $B(x_k, r_n)$ (res, $B(t_j, r_n)$) the ball centered at x_k (res, t_j) with radius r_n.

Theorem 2.

$$\sup_{\theta\in\Theta_{\mathcal{F}}} \sup_{x\in S_{\mathcal{F}}} |\hat{M}r_\theta(x) - Mr_\theta(x)| = O(h^b) + O_{a.co.}\left(\sqrt{\frac{\ln d_n^{S_{\mathcal{F}}} + \ln d_n^{\Theta_{\mathcal{F}}}}{n\phi_x(h)}}\right)$$

References

[1] Hebchi, C.: Uniform almost complete convergence of local linear mode regression. IJSE. **21**(1), 54–62 (2020)

Chaima Hebchi
Laboratoire de Statistique et Processus Stochastiques, (LSPS). Université Djillali liabès. BP 89, Sidi bel Abbès 22000, Algeria, e-mail: chaimahabchi@yahoo.fr

Probabilistic approximations to discrete optimal transport

Florian Heinemann, Axel Munk and Yoav Zemel

Abstract

Optimal transport is now a popular tool in statistics, machine learning, and data science. A major challenge in applying optimal transport to large-scale problems is its excessive computational cost. We propose a simple resampling scheme for fast randomized approximate computation of optimal transport distances on finite spaces. This scheme operates on a random subset of the full data and can use any exact algorithm as a black-box back-end, including state-of-the-art solvers and entropically penalized versions. We give non-asymptotic bounds for the expected approximation error. Remarkably, in many important instances such as images (2D-histograms), the bounds are independent of the size of the full problem. Our resampling scheme can also be employed for the barycentre problem, namely computing Frećhet means with respect to the optimal transport metric. We present numerical experiments demonstrating very good approximations can be obtained while decreasing the computation time by several orders of magnitude.

(based on joint work with Florian Heinemann and Axel Munk [1])

References

[1] Heinemann, F., Munk, A., Zemel, Y.: Randomised Wasserstein barycenter computation: Resampling with statistical guarantees. arXiv:2012.06397 (2020)

Florian Heinemann
Georg–August–Universität Göttingen, Göttingen, Germany, e-mail: florian.heinemann@uni-goettingen.de

Victor M. Panaretos
Georg–August–Universität Göttingen, Göttingen, Germany, e-mail: amunk1@gwdg.de

Yoav Zemel
University of Cambridge, Cambridge, United Kingdom, e-mail: zemel@statslab.cam.ac.uk

From High-Dimensional to Functional Data: Stringing via Manifold Learning

Harold A. Hernández-Roig, M. Carmen Aguilera-Morillo and Rosa E. Lillo.

Abstract

The study of high-dimensional data is becoming a common trend in modern research. Recently, *stringing* [1] emerged as a methodology to treat high-dimensional sample vectors as realizations of smooth stochastic processes. Under the hypothesis of noisy and order-perturbed measurements, stringing introduces smooth transitions between predictors and takes advantage of Functional Data Analysis to study the data. Once a functional representation is achieved, it is possible to visualize intrinsic patterns, or fit functional regression models. We propose *stringing via Manifold Learning* [2] as an alternative to the reordering step based on Multidimensional Scaling. In a simulation study we show that our proposal achieves smaller relative order errors, and that it can recover more complex relations between predictors.

References

[1] Chen, K., Müller, H.-G., Wang, J.: Stringing High-Dimensional Data for Functional Analysis. Journal of the American Statistical Association. **106(493)**, 275–284 (2011)

[2] Hernández-Roig, H. A., Aguilera-Morillo, M. C., Lillo, R. E.: Functional Modeling of High-Dimensional Data: A Manifold Learning Approach. Mathematics, **9(4)**, 406 (2021).

Harold A. Hernández-Roig
Universidad Carlos III de Madrid and uc3m-Santander Big Data Institute, Spain.
e-mail: haroldantonio.hernandez@uc3m.es

M. Carmen Aguilera-Morillo
Universitat Politècnica de València and uc3m-Santander Big Data Institute, Spain.
e-mail: mdagumor@eio.upv.es

Rosa E. Lillo
Universidad Carlos III de Madrid and uc3m-Santander Big Data Institute, Spain.
e-mail: rosaelvira.lillo@uc3m.es

Fourier-type tests of mutual independence between functional time series

Zdeněk Hlávka, Marie Hušková and Simos G. Meintanis

Abstract

For random variables X and Y, the null hypothesis of independence may be conveniently stated in terms of *characteristic functions* (CFs) as: $\mathcal{H}_0 : \Phi_{X,Y}(u,v) = \varphi_X(u)\varphi_Y(v),\ \forall(u,v) \in \mathbb{R} \times \mathbb{R}$, where $\Phi_{X,Y}(u,v) := \mathbb{E}\left[e^{\mathrm{i}(uX+vY)}\right]$ is the joint CF, and $\varphi_X(u) := \Phi_{X,Y}(u,0)$ and $\varphi_Y(v) := \Phi_{X,Y}(0,v)$, are the marginal CFs of X and Y, respectively. In order to generalize this technique to random functions, we restrict our attention to the space $L^2[0,1]$ of square integrable functions in which case the *characteristic functional* (CFL) of the random curve $X(t)$, $t \in [0,1]$, is defined by: $\varphi_X(u) = \mathbb{E}\left[e^{\mathrm{i}\int_0^1 u(t)X(t)dt}\right]$, $u \in L^2[0,1]$, see, e.g., [3].

Tests of independence between functional time series were proposed, e.g., in [2] but these procedures are typically based on the covariance operator and the results thus may not be easily interpretable without additional assumptions of Gaussianity, i.e., with heavy tailed observations.

Consider a pair of stationary random curves $\{X_1(t), Y_1(t)\}, \ldots, \{X_n(t), Y_n(t)\}, \ldots,\ t \in [0,1]$ and suppose we wish to test the null hypothesis:

$$\Upsilon_0 : \{X_j\}_{j=1}^{\infty} \text{ and } \{Y_j\}_{j=1}^{\infty} \text{ are independent.}$$

We show that our tests provide good power against the alternative:

$$\Upsilon_H : \exists |h_0| \le H, \text{ such that } \int_0^1 \int_0^1 \int_{\mathbb{R}} \int_{\mathbb{R}} \left|\varphi_{X,Y,h_0}(t_1,t_2;u_1,u_2) - \varphi_X(t_1;u_1)\varphi_{Y,h_0}(t_2;u_2)\right|^2 w(u_1)w(u_2)\mathrm{d}u_1\mathrm{d}u_2\mathrm{d}t_1\mathrm{d}t_2 > 0,$$

where H is an a priori chosen fixed integer $0 \le H < \infty$. Here $\varphi_{X,Y,h}(t_1,t_2;u_1,u_2)$ denotes the joint CF of $X_1(t_1)$ and $Y_{1+h}(t_2)$, and $\varphi_X(t_1;u_1)$ and $\varphi_{Y,h}(t_2;u_2)$ the corresponding marginal CFs, computed at fixed "time"-pair (t_1,t_2) and argument (u_1,u_2).

Proceeding similarly as [1], we reject the null hypothesis Υ_0 versus Υ_H for large values of the test statistic:

$$T_{n,H}^{(w)} = \sum_{h=-H}^{H} (n - |h|) \int_0^1 \int_0^1 \Delta_{n,h}^{(w)}(t,s)\mathrm{d}s\mathrm{d}t,$$

where

$$\Delta_{n,h}^{(w)}(t,s) = \int_{\mathbb{R}} \int_{\mathbb{R}} \left|\widehat{\phi}_{X,Y,n,h}(t,s;u,v) - \widehat{\phi}_{X,n}(t;u)\widehat{\phi}_{Y,n,h}(s;v)\right|^2 w(u)w(v)\mathrm{d}u\mathrm{d}v,$$

with $\widehat{\phi}_{X,Y,n,h}(t,s;u,v) = \frac{1}{n-|h|}\sum_{j=\max(1,1-h)}^{\min(n-h,n)} e^{\mathrm{i}\left(uX_j(t)+vY_{j+h}(s)\right)}$ being the empirical joint CF and $\widehat{\phi}_{X,n}(t;u)$ and $\widehat{\phi}_{Y,n,h}(t;u)$ the corresponding empirical marginal CFs, respectively.

The limit distribution of the new test statistic is obtained under the null hypothesis, while under alternatives it is shown that the same test statistic almost surely diverges as the sample size increases. Since the limit null distribution is complicated, a bootstrap version of the test is suggested to assess the test's performance in finite samples. Also, an application illustrates the use of the method with real data from financial markets. Extension to tests of mutual independence for multiple time series is also considered.

Acknowledgement: The research was supported by Czech Science Foundation, grant GAČR 18-08888S.

References

[1] Hlávka, Z., Hušková, M., Meintanis, S. G.: Testing serial independence with functional data. TEST (2020)
[2] Horváth, L., Rice, G.: Testing for independence between functional time series. J. Econom. **189**, 371–382 (2015)
[3] Prohorov, Y.V.: The method of characteristic functionals. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2, pp. 403–419. University of California Berkeley (1961)

Zdeněk Hlávka
Charles University, Department of Statistics, Sokolovská 83, Prague, Czech Republic, e-mail: hlavka@karlin.mff.cuni.cz

Marie Hušková
Charles University, Department of Statistics, Sokolovská 83, Prague, Czech Republic, e-mail: huskova@karlin.mff.cuni.cz

Simos G. Meintanis
Unit for Business Mathematics and Informatics, North–West University, Potchefstroom, South Africa, on sabbatical leave from the University of Athens, e-mail: simosmei@econ.uoa.gr

Functional ANOVA based on empirical characteristic functionals

Zdeněk Hlávka and Daniel Hlubinka and Kateřina Koňasová

Abstract

Functional two-sample tests based on empirical characteristic functionals are studied. We consider a test statistic of Cramér–von Mises type with integration over a preselected family of probability measures, say Q, leading to a computationally feasible and powerful test statistic. Small sample properties of the resulting two- and k-sample functional tests are investigated in a simulation study. In particular, we show that the resulting tests are much stronger than the previously proposed F-statistic-based tests. Moreover, a proper choice of the probability measure Q gives very good power in detecting shift and scale alternatives.

References

[1] Hlávka,Z., Hlubinka, D., Koňasová, K.: Functional ANOVA based on empirical characteristic functionals. To appear. (2020)

[2] Hlávka, Z., Hlubinka, D.: Functional Two-sample Tests Based on Empirical Characteristic Functionals. In: Functional and High-Dimensional Statistics and Related Fields, Aneiros, G., Horová, I., Hušková, M., Vieu, P. (Eds.) pp. 123–130 (2020)

Daniel Hlubinka
Univerzita Karlova, Department of Probability and Statistics, e-mail: hlubinka@karlin.mff.cuni.cz

Preprocessing functional data by a factor model approach

Siegfried Hörmann and Fatima Jammoul

Abstract

We consider noisy functional data $Y_t(s_i) = X_t(s_i) + \varepsilon_{ti}$ that has been recorded at a discrete set of observation points. Naturally, the goal is to recover the underlying signal X_t. Commonly, this is done by non-parametric smoothing approaches, e.g. kernel smoothing or spline fitting. These methods act function by function and do not take the overall presented information into consideration. We argue that it is often more accurate to take the entire data set into account, which can help recover systematic properties of the underlying signal. Other approaches using functional principal components do just that, but require strong assumptions on the smoothness of the underlying signal. We show that under very mild assumptions, the signal may be viewed as the common components of a factor model. Using this discovery, we develop a PCA driven approach to recover the signal and show consistency. Our theoretical results hold under rather mild conditions, in particular we do not require specific smoothness assumptions for the underlying curves and allow for a certain degree of autocorrelation in the noise. We demonstrate the applicability of our approach with simulation experiments and real life data analysis.

Siegfried Hörmann
Institute of Statistics, Graz University of Technology, Graz, Austria, e-mail: shoermann@tugraz.at

Fatima Jammoul
Institute of Statistics, Graz University of Technology, Graz, Austria, e-mail: f.jammoul@tugraz.at

Estimating the conditional distribution in functional regression problems

Siegfried Hörmann, Thomas Kuenzer and Gregory Rice

Abstract

We consider the problem of consistently estimating the conditional distribution $P(Y \in A|X)$ of a functional data object $Y = (Y(t) : t \in [0, 1])$ given covariates X in a general space, assuming that Y and X are related by a functional linear regression model. Two natural estimation methods for this problem are proposed, based on either bootstrapping the estimated model residuals, or fitting functional parametric models to the model residuals and estimating $P(Y \in A|X)$ via simulation. We show that under general consistency conditions on the regression operator estimator, which hold for certain functional principal component based estimators, consistent estimation of the conditional distribution can be achieved, both when Y is an element of a separable Hilbert space, and when Y is an element of the Banach space of continuous functions on the unit interval. The latter results imply that sets A that specify path properties of Y that are of interest in applications can be considered, such as the maximum of the curve. Our methods have numerous applications in the context of constructing prediction sets, quantile regression and VaR estimation. Compared to direct modelling these curve properties using scalar-on-function regression, modelling the whole response distribution and extracting the curve properties in a second step allows us to harness the full information contained in the functional data to fit the regression model and achieve better results. We study the proposed methods in several simulation experiments and real data analysis of electricity price curves and show that they outperform both the non-parametric kernel estimator and functional binary regression.

Siegfried Hörmann
Institute of Statistics, Graz University of Technology, Austria, e-mail: shoermann@tugraz.at

Thomas Kuenzer
Institute of Statistics, Graz University of Technology, Austria, e-mail: kuenzer@tugraz.at

Gregory Rice
Department of Statistics and Actuarial Science, University of Waterloo, Canada, e-mail: grice@uwaterloo.ca

Inference in dynamic Nelson–Siegel models

Lajos Horváth, Piotr Kokoszka, Jeremy VanderDoes and Shixuan Wang

Abstract

We consider functional observations $X_1(t), X_2(t), \ldots, X_N(t)$ defined on the interval $\mathcal{T}$. It is often assumed that these observations follow a semi parametric model

$$X_i(t) = \sum_{\ell=1}^{K} b_{i,\ell,0} f_\ell(t; \lambda_0) + \epsilon_i(t), \quad \text{with} \quad E\epsilon_i(t) = 0, \quad t \in \mathcal{T}, \quad 1 \le i \le N,$$

where the random coefficients satisfy

$$b_{i,\ell,0} = c_{\ell,0} + e_{i,\ell} \quad \text{with} \quad Ee_{i,\ell} = 0, \quad t \in \mathcal{T}, \qquad 1 \le \ell \le K \quad \text{and} \quad 1 \le i \le N.$$

Hence in this semi parametric model

$$EX_i(t) = \sum_{\ell=1}^{K} c_{\ell,0} f_\ell(t; \lambda_0), \quad t \in \mathcal{T} \quad \text{and} \quad 1 \le i \le N,$$

i.e. the mean of the observations can be written as a linear combination of the functions $f_1(t; \lambda_0), f_2(t; \lambda_0), \ldots, f_K(t; \lambda_0)$, where the functions $f_1, f_2, \ldots, f_K$ are known and $\lambda_0 \in R^d$ is the true value of an unknown parameter.

We discuss the estimation of the parameters $c_{\ell,0}$, $1 \le \ell \le K$ and λ_0. We show that the estimators are asymptotically consistent and multivariate normal under minor regularity conditions. We also investigate the question if the functional observations can be written in a semi parametric form.

Lajos Horváth
University of Utah, Salt Lake City, UT 841120090, USA, e-mail: horvath@math.utah.edu

Testing stability in event observations with applications to IPO

Lajos Horváth, Zhenya Liu, Gregory Rice, Shixuan Wang and Yaosong Zhan

Abstract

Event study is a widely used methodology in the fields of economics, finance, accounting, operational management, marketing, and political science. The application of event studies enjoys an abundant and expanding literature, with the pioneering research dated back to Dolley (1933). MacKinlay (1997) established a paradigm for conducting event studies in the fields of economics and finance by testing whether cumulative abnormal returns are zero. See also Campbell et al. (1997), Kothari and Warner (2007) and Linton (2019) for reviews of conventional econometrics of event studies.

The available literature on event studies focuses primarily on whether a particularly type of event has an significant effect on the variable of interest. An implicit assumption in such framework is that the event effect stays the same over the sample period. A less explored consideration which offers new insight into event studies is to regard the effects from individual events as random functions which are subject to changes. This consideration is well motivated by our empirical data of IPO premium that policy changes may have a significant impact on the event effect over time. To illustrate, suppose that we have N events occurred in the date sequence of $\{t_1 \leqslant t_2 \leqslant ... \leqslant t_i \leqslant t_{i+1} ... \leqslant t_N\}$, $i = 1, 2, ...N$ in the sample period. The event effect is measured by the variable of interest (such as cumulative abnormal return) during a period called event window. The variable of interest during the event window is continuous in nature and can be formulated as functional observations denoted as $\left\{Y_{t_1}(\tau), Y_{t_2}(\tau), ..., Y_{t_N}(\tau)\right\}, \tau \in [0, 1]$, and they are assumed to satisfy $Y_{t_i}(\tau) = \mu_{t_i}(\tau) + \eta_{t_i}(\tau)$, where $Y_{t_i}(\tau)$ is the impact from the i^{th} event occurred, $\mu_{t_i}(\tau)$ is the common mean function of the event effect, $\eta_{t_i}(\tau)$ is a random error function. One noticeable feature is that it is common in event studies that events may come in irregular frequency. Under the setting in the context of event studies, we develop a new procedure to test for changes in the event effect that $\mu_{t_1}(\tau) = \mu_{t_2}(\tau) = ... = \mu_{t_N}(\tau)$ under the no-change null hypothesis against the alternative of at-least-one change at unknown time. The limit distributions of our test statistics are derived under different assumptions on the time series dependence of the error functions. Monte Carlo simulation illustrates that our test procedure has good size control and high power in relatively small samples. In the empirical illustration, we apply the developed test on a comprehensive data set containing 1,297 functional observations of IPO premium in China over the five-year period from December 2015 to September 2020. We find strong evidence in the change of IPO premium curves around the time of policy changes.

The developed technique of detecting changes in event studies is potentially useful for the researchers who are interested in examining whether the event effect is constant over time.

References

[1] Campbell, J. Y., Lo, A. W., & MacKinlay, A. C.: The econometrics of financial markets. Princeton University press, New Jersey (1997)

[2] Dolley, J. C.: Characteristics and procedure of common stock split-ups. Harvard Business Review 11, **3**, 316-326 (1933)

[3] Kothari, S. P., & Warner, J. B.: Econometrics of event studies. In ECKBO, B. E. (eds.) Handbook of empirical corporate finance, pp. 3-36. Elsevier, Amsterdam (2007)

[4] Linton, O.: Financial econometrics models and methods. Cambridge University press, Cambridge (2019)

[5] MacKinlay, A.C.: Event studies in economics and finance. Journal of economic literature 35, **1**, 13-39, (1997)

Lajos Horváth
University of Utah, Salt Lake City, UT 841120090, USA, e-mail: horvath@math.utah.edu

Zhenya Liu
Renmin University of China, Beijing, 100872, China, e-mail: zhenya.liu@ruc.edu.cn

Gregory Rice
University of Waterloo, Waterloo, N2L 3G1, Canada, e-mail: grice@uwaterloo.ca

Shixuan Wang
Univeristy of Reading, Reading, RG6 6AA, UK e-mail: shixuan.wang@reading.ac.uk

Yaosong Zhan
Renmin University of China, Beijing, 100872, China e-mail: zys1994zys@ruc.edu.cn

Orthogonal Decomposition of Bivariate Densities Using Bayes Spaces

Karel Hron, Jitka Machalová and Alessandra Menafoglio

Abstract

Bivariate probability densities capture relationships within and between two continuous random variables. As such, they carry essentially relative information and follow the scale invariance property which is widely recognized in Bayesian statistics (e.g., when normalizing constant are neglected from computations). Both these properties are captured by the Bayes spaces [3, 11] equipped with the Hilbert space structure which were developed as a generalization of the logratio methodology of compositional data [1, 5] and form a natural sample space for "scale invariant" measures and their respective Radon-Nikodym derivatives (i.e., their densities). The Bayes space methodology enables for a flexible choice of the reference measure (in addition to the default choice, the Lebesgue measure) with weighting effects on the domain of densities and shrinkage/expansion of the space according to the scale of the reference measure. Since bivariate densities as objects in the Bayes space can be also considered as continuous counterparts of compositional tables [2, 4], it is possible to decompose them *orthogonally* into independent and interactive parts, the former being product of revised definitions of marginal densities and the latter capturing the relationships between the random variables [7]. This has several important consequences when probability reference measures are considered. Among them, we mention the so called *marginal invariance* [12], i.e., when the bivariate density is shifted (in the Bayes space sense) by marginal densities, the interaction density is not changed. In addition, the orthogonality of the decomposition opens totally new and groundbreaking perspectives to the theory of copulas [9]. Finally, the centred logratio transformation [11] of bivariate densities enables to move them from the Bayes space to the standard L^2 space where popular methods of functional data analysis [10] can be applied. The novel theoretical framework here proposed has thus clear potential on the application side, allowing to analyse samples of densities arising, for example, as a result of aggregation of massive data coming from large-scale studies or automated collection of data (see, e.g., [6, 8]). The theoretical developments will be applied to both simulated and real-world data sets, the latter containing densities of body height and weight in different age groups from an anthropometric study.

References

[1] Aitchison, J.: The Statistical Analysis of Compositional Data. Chapman and Hall, London (1986)

[2] Egozcue, J.J., Díaz-Barrero, J.L., Pawlowsky-Glahn, V.: Compositional analysis of bivariate discrete probabilities. In: Daunis-i-Estadella, J., Martín-Fernández, J.A. (eds.), The 3rd Compositional Data Analysis Workshop: Proceedings of CODAWORK'08. University of Girona, Girona (2008)

[3] Egozcue, J.J., Díaz-Barrero, J.L., Pawlowsky-Glahn, V.: Hilbert space of probability density functions based on Aitchison geometry. Acta Mathematica Sinica, English Series **22**, 1175–1182 (2006)

[4] Egozcue, J.J., Pawlowsky-Glahn, V., Templ, M., Hron, K.: Independence in contingency tables using simplicial geometry. Communications in Statistics - Theory and Methods **44**, 3978-–3996 (2015)

[5] Filzmoser, P., Hron, K., Templ, M.: Applied Compositional Data Analysis. Springer, Cham (2018)

[6] Hron, K., Menafoglio, A., Templ, M., Hrůzová, K., Filzmoser, P.: Simplicial principal component analysis for density functions in Bayes spaces. Computational Statistics and Data Analysis 94, 330–350 (2016)

[7] Hron, K., Machalová, J., Menafoglio, A.: Bivariate densities in Bayes spaces: Orthogonal decomposition and spline representation. arXiv:2012.12948 (2020)

[8] Menafoglio, A., Grasso, M., Secchi, P., Colosimo B.M.: Profile monitoring of probability density functions via simplicial functional PCA with application to image data. Technometrics **60**, 497–510 (2018)

[9] Nelsen, R.B.: An Introduction to Copulas, second edition. Springer, New York (2006)

[10] Ramsay, J., Silverman, B.W.: Functional Data Analysis, second edition. Springer, New York (2005)

[11] van den Boogaart, K.G., Egozcue, J.J., Pawlowsky-Glahn, V.: Hilbert Bayes spaces. Australian & New Zealand Journal of Statistics **54**, 171–194 (2014)

[12] Yule, G.U.: On the methods of measuring association between two attributes. Journal of the Royal Statistical Society **75**, 579–642 (1912)

Karel Hron
Department of Mathematical Analysis and Applications of Mathematics, Palacký University, 17. listopadu 12, 77146 Olomouc, Czech Republic,
e-mail: karel.hron@upol.cz

Jitka Machalová
Department of Mathematical Analysis and Applications of Mathematics, Palacký University, 17. listopadu 12, 77146 Olomouc, Czech Republic,
e-mail: jitka.machalova@upol.cz

Alessandra Menafoglio
MOX - Department of Mathematics, Politecnico di Milano Piazza Leonardo da Vinci 32, 20133 Milano, Italy, e-mail: alessandra.menafoglio@polimi.it

Invariant Tests for Functional Data with Application to an Earthquake Impact Study

Wei-Hsueh Huang, Li-Shan Huang and Cheng-Tao Yang

Abstract

Motivated by an earthquake impact study, this paper develops new tests with several invariant properties for functional data. For multi-sample functional ANOVA (mfANOVA), based on local polynomial regression, exact local and global ANOVA decompositions into within- and between-group of variations are obtained. A local mfANOVA test is developed to examine differences in a local neighborhood, and combining local quantities, a global mfANOVA test statistic is formed. We show that both the local and global mfANOVA test statistics are location, scale, and translation invariant, enjoy the Wilks phenomenon, allow interchanging the order of smoothing and ANOVA projection, and have asymptotic F-distributions under the Gaussian assumption. This paper contributes to the literature by being the first, to our knowledge, to study mfANOVA tests with several invariant properties. Simulation studies are presented to compare the proposed global mfANOVA test with some existing procedures. Application to an earthquake impact study in Taiwan reveals that when an earthquake in 2016 resulted in closed highways, the patterns of traffic flows were significantly different between three time periods, before the earthquake, during, and after the repair period. The information could be useful in planning for disaster preparedness.

Wei-Hsueh Huang
Institute of Statistics, National Tsing Hua University, TAIWAN, e-mail: sheynehuang@gapp.nthu.edu.tw

Li-Shan Huang
Institute of Statistics, National Tsing Hua University, TAIWAN, e-mail: lhuang@stat.nthu.edu.tw

Cheng-Tao Yang
National Center for Research on Earthquake Engineering, TAIWAN, e-mail: ctyang@ncree.narl.org.tw

Modeling the effect of recurrent events on time-to-event processes by means of Functional Data

Francesca Ieva, Marta Spreafico and Davide Burba

Abstract

In this paper we propose a methodological framework for modeling information carried out by a longitudinal process by means of functional data, within a survival framework targeting the time-to-event process of interest. In particular, the longitudinal process is represented by the compensator of a marked point process the recurrent events are supposed to derive from. By means of Functional Principal Component Analysis (FPCA), a suitable dimensional reduction of these objects is carried out in order to plug them into a survival Cox regression model. In doing so, we enrich the information available for modeling survival with relevant dynamic features, whose time-varying nature is properly taken into account. Such methodology is applied to data provided by the healthcare division of Lombardia regional district in Italy, related to patients hospitalized for Heart Failure (HF) between 2000 and 2012, who assume multiple drugs over time. The model enables personalized predictions, quantifying the effect of personal behaviors and therapeutic patterns on long-term survival.

Francesca Ieva, Marta Spreafico, Davide Burba
MOX, Department of Mathematics, Politecnico di Milano, P.zza Leonardo da Vinci 32, 20133 Milano (IT)
e-mail: francesca.ieva@polimi.it e-mail: marta.spreafico@polimi.it e-mail: davide.burba@mail.polimi.it

Process of R-estimators of slope vector in linear model

Jana Jurečková

Abstract

We consider the linear regression model

$$Y_{ni} = \beta_0 + \mathbf{x}_{ni}^{\top}\boldsymbol{\beta} + e_{ni}, \quad i = 1, \ldots, n.$$

While only the intercept part of the α regression quantile of this model reflects the quantile $F^{-1}(\alpha)$ of the model errors, the slope components are not generally monotone and have an undetermined shape. We study the process of R-estimators of the slope parameters, especially generated by the Hájek rank scores, running over $\alpha \in [0, 1]$. Under some conditions on the tails of the basic distribution and on the covariates, the process of R-estimators of slopes converges, after a pertinent standardization, to the vector of independent Brownian bridges.

Jana Jurečková
The Czech Academy of Sciences,
Institute of Information Theory and Automation,
Charles University, Czech Republic,e-mail: jurecko@karlin.mff.cuni.cz

On Robust Training of Regression Neural Networks

Jan Kalina and Petra Vidnerová

Abstract

Estimation, prediction or smoothing of curves represents a fundamental task of functional data analysis [1, 3]. Nonlinear regression methods allow to search for the best-fit curves explaining the dependence of a response variable on available independent variables. Neural networks, commonly used for the task of nonlinear regression, are however highly vulnerable to the presence of outlying measurements in the data [2]. New robust versions of common types of neural networks, namely multilayer perceptrons and radial basis function networks, are proposed here based on nonlinear regression quantiles or highly robust loss functions. Three datasets are analyzed to illustrate the performance of the novel robust approaches, which turn out to outperform standard neural networks or other competing regression tools over contaminated data.

The research was supported by the projects GA19-05704S and GA18-23827S of the Czech Science Foundation.

References

[1] Ramsay, J., Silverman, B.W.: Functional data analysis. 2nd edn. Springer, New York (2005)

[2] Rusiecki, A., Kordos, M., Kamiński, T., Greń, K.: Training neural networks on noisy data. Lecture Notes in Artificial Intelligence **8467**, 131–142 (2014)

[3] Ullah, S., Finch, C.F.: Applications of functional data analysis: A systematic review. BMC Medical Research Methodology **13**, Article 43 (2013)

Jan Kalina
The Czech Academy of Sciences, Institute of Computer Science, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic e-mail: kalina@cs.cas.cz

Petra Vidnerová
The Czech Academy of Sciences, Institute of Computer Science, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic e-mail: petra@cs.cas.cz

Lagged Covariance and Cross-Covariance Operators of Processes in Cartesian Products of Abstract Hilbert Spaces

Dr. Sebastian Kühnert

Abstract

A major task in Functional Time Series Analysis is measuring the dependence within and between processes, for which lagged covariance and cross-covariance operators have proven to be a practical tool. Probabilistic features of and estimators for lagged covariance operators of stationary processes with values in $L^2[0,1]$, the space of measurable, square-Lebesgue integrable real valued functions with domain $[0,1]$, are widely studied for fixed lag h, see, e.g., [3], [5], [7], [13], [9], and under several limitations also in the space of continuous functions $C[0,1]$, in tensor product Sobolev-Hilbert spaces, for continuous surfaces, and for arbitrary separable Hilbert spaces, see [16], [17], [12] resp. [4], [1]. Further, lagged cross-covariance operators of stationary $L^2[0,1]$-valued processes were comprehensively studied in Rice & Shum [14], and Aue & Klepsch [2], who esimtated operators of linear, invertible processes in $L^2[0,1]$, had to estimate lagged cross-covariance operators of processes with values in Cartesian products of $L^2[0,1]$ in order derive their main results.

This work deduces, based on ideas in [14], [2] and [11], and the notion for weak dependence developed by Hörmann and Kokoszka [6], estimators and asymptotic upper bounds of the estimation errors for lagged covariance and cross-covariance operators of processes in Cartesian products of abstract Hilbert spaces for fixed and increasing lag and Cartesian powers. We allow the processes to be non-centered, and to have values in different spaces when investigating the dependence between processes. Also, we discuss features of estimators for the principle components of our covariance operators.

References

[1] Allam, A., Mourid, T.: Optimal rate for covariance operator estimators of functional autoregressive processes with random coefficients. J. Multivariate Anal. **169**, 130–137 (2019)

[2] Aue, A., Klepsch, J.: Estimating functional time series by moving average model fitting. arXiv: 1701.00770v1 (2017)

[3] Bosq, D.: Linear Processes in Function Spaces. Lecture Notes in Statistics, 149. Springer, New York (2000)

[4] Hashemi, M., Zamani, A., Haghbin, H.: Rates of convergence of autocorrelation estimates for periodically correlated autoregressive Hilbertian processes. Statistics **53** (2), 283–300 (2019)

[5] Horváth, L., Kokoszka, P.: Inference for Functional Data with Applications. Springer, New York (2012)

[6] Hörmann, S., Kokoszka, P.: Weakly dependent functional data. Ann. Statist. **38**, 1845–1884 (2010)

[7] Hsing, T., Eubank, R.: Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. Wiley, West Sussex (2015)

[8] Kokoszka, P., Reimherr, M.: Asymptotic normality of the principal components of functional time series. Stoch. Process. Appl. **123** (5), 1546–1562 (2013)

[9] Kokoszka, P., Rice, G., Shang, H.L.: Inference for the autocovariance of a functional time series under conditional heteroscedasticity. J. Multivariate Anal. **162**, 32–50 (2017)

[10] Kokoszka, P., Stoev, S., Xiong, Q.: Principal components analysis of regularly varying functions. Bernoulli **25** (4B), 3864–3882 (2019)

[11] Kühnert, S.: Functional ARCH and GARCH models: A Yule-Walker approach. Electron. J. Statist. **14** (2), 4321–4360 (2020)

[12] Martínez-Hernández, I., Genton, M.G.: Recent developments in complex and spatially correlated functional data. Braz. J. Probab. Stat. **34** (2), 204–229 (2020)

[13] Mas, A.: Weak convergence in the functional autoregressive model. J. Multivariate Anal. **98** (6), 1231–1261 (2007)

[14] Rice, G., Shum, M.: Inference for the lagged cross-covariance operator between functional time series. J. Time Series Anal. **40** (5), 665–692 (2019)

[15] Rice, G., Wirjanto, T., Zhao, Y.: Forecasting value at risk with intra-day return curves. Int. J. Forecast., Elsevier **36** (3), 1023–1038 (2020)

[16] Ruiz-Medina, M.D., Álvarez-Liébana, J: Strongly consistent autoregressive predictors in abstract Banach spaces. J. Multivariate Anal. **170**, 186–201 (2019)

[17] Wong, R.K.W., Zhang, X.: Nonparametric operator-regularized covariance function estimation for functional data. Comput. Statist. Data Anal. **131**, 131–144 (2019)

Dr. Sebastian Kühnert
University of Rostock (formerly), e-mail: s.kuehnert_math@gmx.de

Decomposing the Asset Pricing Anomalies by Functional PCA

Bo Li, Zhenya Liu and Yifan Zhang

Abstract

This paper applies a functional principal component analysis (FPCA) to decompose China's A-share portfolio returns on time-series and cross-section simultaneously.

First we obtain a sequence of functional observations $\{r_t(u),\ 1 \le t \le T,\ 0 \le u \le 1\}$ smoothing the discretely observed vectors $\boldsymbol{r}_t = (r_{t,1}, \ldots, r_{t,N}),\ 1 \le t \le T$ by B-spline basis. We assume that $\{r_t(u),\ 0 \le u \le 1\}$ is a stationary sequence with sample paths in $L^2[0,1]$, and the moment condition $E||r_t(u)||^4 < \infty$ is satisfied. The covariance function $c(u,v)$ of $\{r_t(u),\ 0 \le u \le 1\}$ is estimated by (cf. Cao et al., 2019):

$$\hat{c}_T(u,v) = \frac{1}{T}\sum_{t=1}^{T}\big(r_t(u) - \bar{r}_T(u)\big)\big(r_t(v) - \bar{r}_T(v)\big),$$

where $\bar{r}_T(u) = \frac{1}{T}\sum_{t=1}^{T} r_t(u)$. The eigenvalues and eigenfunctions of the $c(u,v)$ are estimated by $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \cdots \ge 0$ and $\hat{\phi}_1(u), \hat{\phi}_2(u), \cdots$ satisfying

$$\hat{\lambda}_m \hat{\phi}_m(u) = \int_0^1 \hat{c}_T(u,v)\hat{\phi}_m(v)dv, \quad m = 1, 2, \ldots.$$

Thus the functional observations $r_t(u)$ can be projected into the space spanned by the eigenfunctions $\big(\hat{\phi}_1(u), \hat{\phi}_2(u), \ldots, \hat{\phi}_M(u)\big)$:

$$r_t(u) \approx \hat{r}_t(u) = \sum_{m=1}^{M} \hat{\xi}_{t,m}\hat{\phi}_m(u),$$

where $\hat{\xi}_{t,m} = \langle r_t(u), \hat{\phi}_m(u)\rangle = \int_0^1 r_t(u)\hat{\phi}_m(u)du$. We refer Horváth and Kokoszka (2012) for more details. In this case, the variation of $r_t(u)$ along sorted groups are decomposed by a finite number of estimated eigenfunctions $\hat{\phi}_m$, and their variance contributions are captured by their corresponding eigenvalues $\hat{\lambda}_m$.

The empirical analysis is based on the following statistics:

1) The expectation of m-th component on percentile u stands for the estimated variation pattern along with time t, referred as $\mathbb{E}_u[\xi_{t,m}\phi_m(u)]$, which is estimated by $\mathbb{E}_u[\hat{\xi}_{t,m}\hat{\phi}_m(u)]$;
2) The expectation on time t stands for the estimated long-run variation pattern on the cross-section, referred as $\mathbb{E}_t[\xi_{t,m}\phi_m(u)]$, which is estimated by $\mathbb{E}_t[\hat{\xi}_{t,m}\hat{\phi}_m(u)]$;
3) $\sum_{m=1}^{M}\mathbb{E}_t[\xi_{t,m}\phi_m(u)]$, which is estimated by $\sum_{m=1}^{M}\mathbb{E}_t[\xi_{t,m}\phi_m(u)]$, represents the overall long-run variation pattern on the cross-section.

The results show that the first empirical functional principal component (EFPC) stands for the market factor and the others for an anomaly. The second and third ones reveal the cross-sectional linear and convex patterns, and the joint of them dominates the asset pricing anomalies. Furthermore, the EFPCs illustrate much more information than the portfolio-based approach, and can be used to explain the debates about some anomalies.

References

[1] Horváth, L. and Kokoszka, P.: Inference for Functional Data with Applications. Springer, New York (2012)

[2] Cao, R.M., Horváth, L., Liu, Z.Y., Zhao,Y.Q.: A study of data-driven momentum and disposition effects in the Chinese stock market by functional data analysis. Review of Quantitative Finance and Accounting. **52(1)**: 1–24 (2019)

Bo Li
Business School, Beijing International Studies University, China, e-mail: libo@bisu.edu.cn.

Zhenya Liu
School of Finance, Renmin University of China, China, CERGAM, Aix-Marseille University, France, e-mail: zhenya_liu@hotmail.com

Yifan Zhang
School of Finance, Renmin University of China, China, e-mail: zhangyifan@ruc.edu.cn

Nonstationary Fractionally Integrated Functional Time Series

Degui Li, Peter M. Robinson and Han Lin Shang

Abstract

We study a functional version of nonstationary fractionally integrated time series, covering the functional unit root as a special case. The time series taking values in an infinite-dimensional separable Hilbert space are projected onto a finite number of sub-spaces, the level of nonstationarity allowed to vary over them. Under regularity conditions, we derive a weak convergence result for the projection of the fractionally integrated functional process onto the asymptotically dominant sub-space, which retains most of the sample information carried by the original functional time series. Through the classic functional principal component analysis of the sample variance operator, we obtain the eigenvalues and eigenfunctions which span a sample version of the dominant sub-space. Furthermore, we introduce a simple ratio criterion to consistently estimate the dimension of the dominant sub-space, and use a semiparametric local Whittle method to estimate the memory parameter. Monte-Carlo simulation studies and empirical applications are given to examine the finite-sample performance of the developed techniques.

Degui Li
Department of Mathematics, University of York, York, YO10 5DD, UK, e-mail: degui.li@york.ac.uk

Peter M. Robinson
Department of Economics, London School of Economics, London WC2A 2AE, UK, e-mail: P.M.Robinson@lse.ac.uk

Han Lin Shang
Department of Actuarial Studies and Business Analytics, Macquarie University, NSW 2109, Australia, e-mail: hanlin.shang@mq.edu.au

Simultaneous Inference for Function-Valued Parameters: A Fast and Fair Approach

Dominik Liebl and Matthew Reimherr

Abstract

This work presents a new approach for constructing simultaneouse confidence bands for function-valued parameters. The bands are fast to compute as they are based on nearly closed-form expressions and, therefore, do not require computationally expensive resampling based methods. The shape of the bands can be constructed according to a desired criteria specified by the user. A particularly interesting criteria is the proposed concept of "fair" or equitable bands which leads to simultaneous confidence bands that have an adaptive width reflecting the local multiple testing problem. The theoretical foundations of our simultanouse confidence bands are presented in [9]. In this short paper, we deviate from [9] and consider the practically important special case of the linear function-on-scalar regression model. Moreover, we present a novel application on testing for differences in yield curves of A and B-type rated countries.

References

[1] Abramowicz, K., Häger C. K., Pini A., Schelin L., Sjöstedt de Luna S., and Vantini S. (2018). Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament. *Scandinavian Journal of Statistics 45*(4), 1036–1061.

[2] Adler, R. J. and Taylor J. E. (2007). *Random Fields and Geometry*, 1st ed. Springer.

[3] Cao, G., Yang, L. and Todem, D. (2012). Simultaneous inference for the mean function based on dense functional data. *Journal of Nonparametric Statistics 24*(2), 359–377.

[4] Chang, C. and Ogden, R. T. (2009). Bootstrapping sums of independent but not identically distributed continuous processes with applications to functional data. *Journal of Multivariate Analysis*, *100*(6), 1291–1303.

[5] Degras, D. A. (2011). Simultaneous confidence bands for nonparametric regression with functional data. *Statistica Sinica 21*(4), 1735–1765.

[6] Dette, H., Kokot, K. and Aue, A. (2020). Functional data analysis in the Banach space of continuous functions. *The Annals of Statistics (forthcoming).*

[7] Kac, M. (1943). On the average number of real roots of a random algebraic equation. *Bulletin of the American Mathematical Society 49*(4), 314–320.

[8] Kokoszka, P. and Reimherr, M. (2017). Introduction to Functional Data Analysis, 1st ed. Chapman & Hall/CRC

[9] Liebl, D. and Reimherr, M. (2020). Fast and Fair Simultaneous Confidence Bands for Functional Parameters. arXiv preprint arXiv:1910.00131.

[10] Lopes, M. E., Lin, Z. and Müller, H. G. (2020). Bootstrapping max statistics in high dimensions: Near-parametric rates under weak variance decay and application to functional data analysis. *The Annals of Statistics (forthcoming).*

[11] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer.

[12] Rice, S. O. (1945). Mathematical analysis of random noise. *Bell System Technical Journal 24*, 46–156.

[13] Taylor, J., Takemura, A. and Adler, R. J. (2005). Validity of the expected euler characteristic heuristic. *The Annals of Probability 33*(4), 1362–1396.

[14] Telschow, F. J. E. and Schwartzman, A. (2020) Simultaneous confidence bands for functional data using the gaussian kinematic formula. arXiv preprint arXiv:1901.06386.

[15] Wang, J., Cao, G., Wang, L. and Yang, L. (2020). Simultaneous Confidence Band for Stationary Covariance Function of Dense Functional Data. *Journal of Multivariate Analysis 176*, 104584.

[16] Wang, Y., Wang, G., Wang, L. and Ogden, R. T. (2020). Simultaneous confidence corridors for mean functions in functional data analysis of imaging data. *Biometrics (forthcoming).*

Dominik Liebl
Institute of Finance and Statistics and Hausdorff Center for Mathematics, University of Bonn, Adenauerallee 24-26, 53113 Bonn, e-mail: dliebl@uni-bonn.de

Matthew Reimherr
Department of Statistics, Penn State University, 411 Thomas Building University Park, PA 16802 e-mail: mreimherr@psu.edu

Single functional index model under responses MAR and dependent observations

Nengxiang Ling, Lilei Cheng and Philippe Vieu

Abstract

This contribution deals with the estimation of the functional single index regression model (FSIRM) with responses missing at random (MAR) for strong mixing time series data. Some asymptotic properties such as the uniform almost complete convergence rate and asymptotic normality of the estimator are obtained respectively under some general conditions.

Nengxiang Ling
School of Mathematics, Hefei University of Technology, China, e-mail: hfut.lnx@163.com

Lilei Cheng
School of Mathematics, Hefei University of Technology, China, e-mail: hfutcll@163.com

Philippe Vieu
Institut de Mathématiques, Université Paul Sabatier, Toulouse, France, e-mail: philippe.vieu@math.univ-toulouse.fr

Outlier detection for multivariate time series: a functional data approach

Ángel López Oriona and José Antonio Vilar

Abstract

A method for detecting outlier samples in a multivariate time series dataset is proposed. It is assumed that an outlying series is characterized by having been generated from a different process than those of the rest of the series. Each multivariate time series is described by means of an estimator of its quantile cross-spectral density, which is treated as a multivariate functional datum. Then an outlier score is assigned to each series by using functional depths. A broad simulation study shows that the proposed approach is superior to the alternatives suggested in the literature and demonstrates that the consideration of functional data constitutes a critical step. The procedure runs in linear time with respect to both the series length and the number of series, and in quadratic time regarding the number of components. Two applications concerning financial series and ECG signals highlight the usefulness of the technique.

Ángel López Oriona
University of A Coruña, Spain, e-mail: oriona38@hotmail.com

José Antonio Vilar
University of A Coruña, Spain, e-mail: jose.vilarf@udc.es

Compositional splines for representation of density functions

Jitka Machalová and Karel Hron

Abstract

Similar as in functional data analysis (FDA) [4], also in statistical processing of density functions which relies on the Bayes space methodology [1] proper preprocessing of the input discrete data obtained by sampling a (hypothetical) density function is crucial for a reliable output [3]. Spline functions are extensively used in FDA for approximation of non-periodical functions as they are flexible enough to cover a wide range of their specific behavior, hence they are also a natural choice for construction of density functions. There are several different basis systems for constructing spline functions; we will focus on B-spline functions which have minimal support with respect to a given degree, smoothness and domain partition. Any spline function of given degree can be expressed as a linear combination of B-spline functions of that degree. Here, we restrict to a bounded support $I = [a, b] \subset \mathbb{R}$ and density functions are considered with respect to the Lebesgue reference measure using the Bayes space $\mathcal{B}^2(I)$ of functions with square-integrable logarithm. In order to express the density functions in the standard L^2 space, $L^2(I)$, *centered log-ratio* (clr) transformation is taken, which induces, however, an additional zero-integral constraint. As the clr space is clearly a subspace of $L^2(I)$, hereafter it is denoted as $L_0^2(I)$.

Using the standard B-spline basis system for approximation of density functions in $L_0^2(I)$ is unfortunate, because it does not belong to this space. Therefore so called *compositional splines* were proposed which honor the zero integral constraint in $L_0^2(I)$ for both the B-spline basis functions and the resulting spline function. Consequently, the compositional splines can be implemented instead of the standard ones into FDA methods for statistical processing of density functions like simplicial functional principal component analysis (SFPCA) [2] or regression analysis with functional response [5]. In this contribution, the case of SFPCA is considered for dimension reduction of density functions of anthropological data. Moreover, the presented theory of univariate spline functions can be extended to more than one variable, i.e., for approximation of multivariate density functions. The easiest and flexible way to do that is to use tensor product splines which are currently intensively studied. For the purpose of simplicity only bivariate splines will be considered and basic ideas of this extension will be presented.

References

[1] van den Boogaart, K.G., Egozcue, J.J., Pawlowsky-Glahn, V.: Bayes Hilbert spaces. Australian & New Zealand Journal of Statistics, **56**,(2), pp. 171–194 (2014)

[2] Hron, K., Menafoglio, A., Templ, M., Hrůzová, K., Filzmoser, P.: Simplicial principal component analysis for density functions in Bayes spaces, Computational Statistics and Data Analysis, **94**, pp. 330–350 (2016)

[3] Machalová, J., Hron, K., Monti, G. S.: Preprocessing of centred logratio transformed density functions using smoothing splines, Journal of Applied Statistics, **43**, (8), pp. 1419–1435 (2016)

[4] Ramsay, J. Silverman, B.W.: Functional Data Analysis, Springer, Heidelberg (2005)

[5] Talská R., Menafoglio, A., Machalová, J., Hron, K., Fišerová, E.: Compositional regression with functional response, Computational Statistics and Data Analysis, **123**, pp. 66–85 (2018)

Jitka Machalová
Faculty of Science, Palacký University Olomouc, 17. listopadu 1192/12, 771 46 Olomouc, e-mail: jitka.machalova@upol.cz

Karel Hron
Faculty of Science, Palacký University Olomouc, 17. listopadu 1192/12, 771 46 Olomouc, e-mail: karel.hron@upol.cz

O2S2 for the geodata deluge

Alessandra Menafoglio, Davide Pigoli and Piercesare Secchi

Abstract

We illustrate a few recent ideas of Object Oriented Spatial Statistics (O2S2), focusing on the problem of kriging prediction in situations where a global second order stationarity assumption for the random field generating the data is not justifiable or the space domain of the field is complex. By localizing the analysis through the Random Domain Decomposition algorithm, we build ensembles of local predictors eventually aggregated in an ultimate one. The localization allowed by the algorithm is also effective for dealing with data which are mildly non-Euclidean and can be locally linearized, as it happens for data embedded in a Riemannian manifold.

References

[1] Menafoglio, A., Pigoli, D., Secchi, P: O2S2 for the geodata deluge. In: Aneiros G., Horová I., Hušková M., Vieu P. (eds) Functional and High-Dimensional Statistics and Related Fields. IWFOS 2020. Contributions to Statistics. Springer, Cham. (2020) https://doi.org/10.1007/978-3-030-47756-1_23

Alessandra Menafoglio
MOX - Department of Mathematics, Politecnico di Milano, Milano, Italy e-mail: alessandra.menafoglio@polimi.it

Davide Pigoli
Department of Mathematics, King's College London, London, United Kingdom e-mail: davide.pigoli@klc.ac.uk

Piercesare Secchi
MOX - Department of Mathematics, Politecnico di Milano, Milano, Italy and Center for Analysis Decisions and Society, Human Technopole, Milano, Italy
e-mail: piercesare.secchi@polimi.it

Entropic regularization of Wasserstein distance between infinite-dimensional Gaussian measures and Gaussian processes

Hà Quang Minh

Abstract

Let (X, d) be a complete separable metric space, $c : X \times X \to \mathbb{R}_{\geq 0}$ a lower semi-continuous *cost function*. Let $\mathcal{P}(X)$ denote the set of all probability measures on X. The *optimal transport* (OT) problem between two probability measures $\nu_0, \nu_1 \in \mathcal{P}(X)$ is $\mathrm{OT}(\nu_0, \nu_1) = \min_{\gamma \in \mathrm{Joint}(\nu_0,\nu_1)} \mathbb{E}_\gamma[c] = \min_{\gamma \in \mathrm{Joint}(\nu_0,\nu_1)} \int_{X \times X} c(x, y) d\gamma(x, y)$ where $\mathrm{Joint}(\nu_0, \nu_1)$ is the set of joint probabilities with marginals ν_0 and ν_1. The OT problem is often computationally challenging and it is more numerically efficient to solve the following *entropic regularized* optimization problem $\mathrm{OT}^\epsilon_c(\mu, \nu) = \min_{\gamma \in \mathrm{Joint}(\mu,\nu)} \left\{\mathbb{E}_\gamma[c] + \epsilon \mathrm{KL}(\gamma || \mu \otimes \nu)\right\}, \epsilon > 0$, with KL denoting the Kullback-Leibler divergence. The KL term acts as a bias and in general $\mathrm{OT}^\epsilon_c(\mu, \mu) \neq 0$. This bias is removed in the *p-Sinkhorn divergence* $S^\epsilon_p(\mu, \nu) = \mathrm{OT}^\epsilon_{d^p}(\mu, \nu) - \frac{1}{2}(\mathrm{OT}^\epsilon_{d^p}(\mu, \mu) + \mathrm{OT}^\epsilon_{d^p}(\nu, \nu))$. If $X = \mathcal{H}$ is a separable Hilbert space and μ, ν are Gaussian measures on $\mathcal{H}$, both the entropic Wasserstein distance $\mathrm{OT}^\epsilon_{d^2}$ and S^ϵ_2 admit closed form expressions, as follows.

Theorem 1 (Entropic Wasserstein distance and Sinkhorn divergence between Gaussian measures on Hilbert space [1]) *Let $\mu_0 = \mathcal{N}(m_0, C_0)$, $\mu_1 = \mathcal{N}(m_1, C_1)$ be two Gaussian measures on a separable Hilbert space $\mathcal{H}$. For each fixed $\epsilon > 0$,*

$$\mathrm{OT}^\epsilon_{d^2}(\mu_0, \mu_1) = ||m_0 - m_1||^2 + \mathrm{Tr}(C_0) + \mathrm{Tr}(C_1) - \frac{\epsilon}{2}\mathrm{Tr}(M^\epsilon_{01}) + \frac{\epsilon}{2} \log\det\left(I + \frac{1}{2}M^\epsilon_{01}\right). \tag{1}$$

$$\mathrm{S}^\epsilon_2(\mu_0, \mu_1) = ||m_0 - m_1||^2 + \frac{\epsilon}{4}\mathrm{Tr}\left[M^\epsilon_{00} - 2M^\epsilon_{01} + M^\epsilon_{11}\right] + \frac{\epsilon}{4}\log\det\left[\frac{\left(I + \frac{1}{2}M^\epsilon_{01}\right)^2}{\left(I + \frac{1}{2}M^\epsilon_{00}\right)\left(I + \frac{1}{2}M^\epsilon_{11}\right)}\right]. \tag{2}$$

Here det *is the Fredholm determinant, $M^\epsilon_{ij} = -I + \left(I + \frac{16}{\epsilon^2}C_i^{1/2}C_jC_i^{1/2}\right)^{1/2}$, $i, j = 0, 1$, $\lim_{\epsilon \to 0} S^\epsilon_2(\mu_0, \mu_1) = \mathrm{OT}_{d^2}(\mu_0, \mu_1)$.*

Theorem 2 (Convergence in Hilbert-Schmidt norm [2]) *Let $A, \{A_n\}_{n \in \mathbb{N}} \in \mathrm{Sym}^+(\mathcal{H}) \cap \mathrm{Tr}(\mathcal{H})$. $\forall \epsilon > 0$,*

$$\mathrm{S}^\epsilon_2[\mathcal{N}(0, A_n), \mathcal{N}(0, A)] \leq \frac{3}{\epsilon}[||A_n||_{\mathrm{HS}} + ||A||_{\mathrm{HS}}]||A_n - A||_{\mathrm{HS}}. \tag{3}$$

This is *weaker* than the trace class norm convergence of the squared Wasserstein distance OT_{d^2} when $\dim(\mathcal{H}) = \infty$. The Hilbert-Schmidt convergence allows application of concentration results for Hilbert space-valued random variables and gives the following *dimension-independent* estimate for Sinkhorn divergence between Gaussian processes.

Theorem 3 (Estimation of Sinkhorn divergence between Gaussian processes [3]) *Let T be a σ-compact metric space, ν a non-degenerate Borel probability measure on T. Let $\mathrm{GP}(0, K^i)$, $i = 1, 2$, be centered Gaussian processes with sample paths in $\mathcal{L}^2(T, \nu)$, with continuous covariance functions K^i satisfying $\sup_{t \in T} K^i(t, t) \leq \kappa_i^2 < \infty$. Let $C_{K^i} : \mathcal{L}^2(T, \nu) \to \mathcal{L}^2(T, \nu)$ be the corresponding covariance operators, given by $(C_{K^i}f)(x) = \int_T K^i(x, t)f(t)d\nu(t)$. Let $\mathbf{X} = (x_i)_{i=1}^m$ be independently sampled from (T, ν), $K^i[\mathbf{X}]$ be the $m \times m$ Gram matrix defined by $(K^i[\mathbf{X}])_{jk} = K^i(x_j, x_k)$. For any $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\Bigg|\mathrm{S}^\epsilon_2\Bigg[\mathcal{N}\left(0, \frac{1}{m}K^1[\mathbf{X}]\right), \mathcal{N}\left(0, \frac{1}{m}K^2[\mathbf{X}]\right) - \mathrm{S}^\epsilon_2\left[\mathcal{N}(0, C_{K^1}), \mathcal{N}(0, C_{K^2})\right]\Bigg]\Bigg| \leq \frac{6}{\epsilon}(\kappa_1^2 + \kappa_2^2)^2\left[\frac{2\log\frac{6}{\delta}}{m} + \sqrt{\frac{2\log\frac{6}{\delta}}{m}}\right]. \tag{4}$$

References

[1] Hà Quang Minh. Entropic regularization of Wasserstein distance between infinite-dimensional Gaussian measures and Gaussian processes, preprint, 2020, https://arxiv.org/abs/2011.07489.

[2] Hà Quang Minh. Convergence and finite sample approximations of entropic regularized Wasserstein distances in Gaussian and RKHS settings, preprint, 2021, https://arxiv.org/abs/2101.01429.

[3] Hà Quang Minh. Finite sample approximations of exact and entropic Wasserstein distances between covariance operators and Gaussian processes, preprint, 2021, https://arxiv.org/abs/2104.12368.

Hà Quang Minh
RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, 15F, Chuo-ku, Tokyo, Japan, e-mail: minh.haquang@riken.jp

Combined functional ordering in view of box plot and clustering

Tomáš Mrkvička

Abstract

The functions are complex objects and when the aim is to rank the functions from the most extreme to the least extreme, different criteria of extremeness can be of interest. The raw data are suitable for determining the magnitude outliers, but a transformation such as derivative or normalized centered functions can give more reliable information about the shape of the functions [1]. So, the choice of the functional ordering and transformation determines the ranking of the functions and especially which kind of the extremeness it will be sensitive on. In case when one universal ordering is of interest, the combined ordering, which joins different functional orderings or different functional transformations, is of interest.

We will show the extreme rank length depth [2] (or extremal depth [3]) which can be used for joining several functional depths in equal way by employing pointwise ranks. This functional depth has graphical interpretation, such that if a function is not completely contained in a central region then its depth is smaller than depth of all functions forming the central region. And further we will show how this joint depth can be used in order to define functional box plot, which is sensitive both on magnitude outliers and shape outliers. Also, we will show how this joint depth can be used in order to define functional clustering, which is sensitive both on magnitude differences and shape differences.

References

[1] Dai, W., Mrkvička, T., Sun, Y., Genton, M.: Functional Outlier Detection and Taxonomy by Sequential Transformations. https://arxiv.org/abs/1808.05414 (2018)

[2] Myllymäk,i M., Mrkvička, T., Grabarnik, P., Seijo, H., Hahn, U..:lobal Envelope Tests forSpatial Processes. Journal of the Royal Statistical Society, Series B (Statistical Methodol-ogy) **79**, 381–404 (2017)

[3] Narisetty, N. N., Nair, V. J .: Extremal Depth for Functional Data and Applications.our-nal of the American Statistical Association.**111**, 1705–1714 (2016)

Tomáš Mrkvička
University of South Bohemia, Faculty of Economics, University of South Bohemia, Studentská 13, 370 05 České Budějovice, Czech Republic,
e-mail: mrkvicka.toma@gmail.coms

New methods for multiple testing in permutation inference for the general linear model

Mari Myllymäki

Abstract

Permutation methods are often used to test significance of regressors of interest in general linear models (GLMs) for functional (image) data sets, in particular for neuroimaging applications as they rely on mild assumptions. Permutation inference for GLMs typically consists of three parts: choosing a relevant test statistic, computing pointwise permutation tests and applying a multiple testing correction. This talk discusses new multiple testing methods as an alternative to the commonly used maximum value of test statistics across the image. The proposed methods increase power and robustness against inhomogeneity of the distribution of the test statistic across its domain and also allow to identify the regions of potential rejection via a graphical output [1, 2]. The methods rely on sorting the permuted functional test statistics based on pointwise rank measures. The methods are implemented in the R package GET [3].

References

[1] Mrkvička, T., Myllymäki, M., Narisetty, N. N.: New methods for multiple testing in permutation inference for the general linear model. arXiv:1906.09004 [stat.ME] (2019)

[2] Mrkvička, T., Roskovec, T., Rost, M.: A nonparametric graphical tests of significance in functional GLM. arXiv:1902.04926 [stat.ME] (2019) Methodology and Computing in Applied Probability. https://doi.org/10.1007/s11009-019-09756-y

[3] Myllymäki, M., Mrkvička, T.: GET: Global envelopes in R. arXiv:1911.06583 [stat.ME] (2020)

Mari Myllymäki
Natural Resources Institute Finland (Luke), Latokartanonkaari 9, FI-00790 Helsinki, Finland e-mail: mari.myllymaki@luke.fi

Statistical Depth for Functional Data

Stanislav Nagy

Abstract

Statistical depth is a tool of nonparametric analysis applicable to multivariate and non-Euclidean datasets living in general spaces $\mathcal{X}$. To a point $x \in \mathcal{X}$ and a Borel probability measure P on $\mathcal{X}$ the depth attaches a number $D(x; P)$ that quantifies how much “centrally positioned" x is with respect to the geometry of the distribution P [5]. In the standard case of multivariate observations and $\mathcal{X} = \mathbb{R}^d$ with $d \geq 1$, many useful depth functions are available in the literature [6]. The depth as a general concept is already fairly well established in $\mathbb{R}^d$, and several sets of desiderata demanded from a proper statistical depth function are available. Prominent examples of depths in $\mathbb{R}^d$ are, e.g., the halfspace, zonoid, spatial, or projection depth.

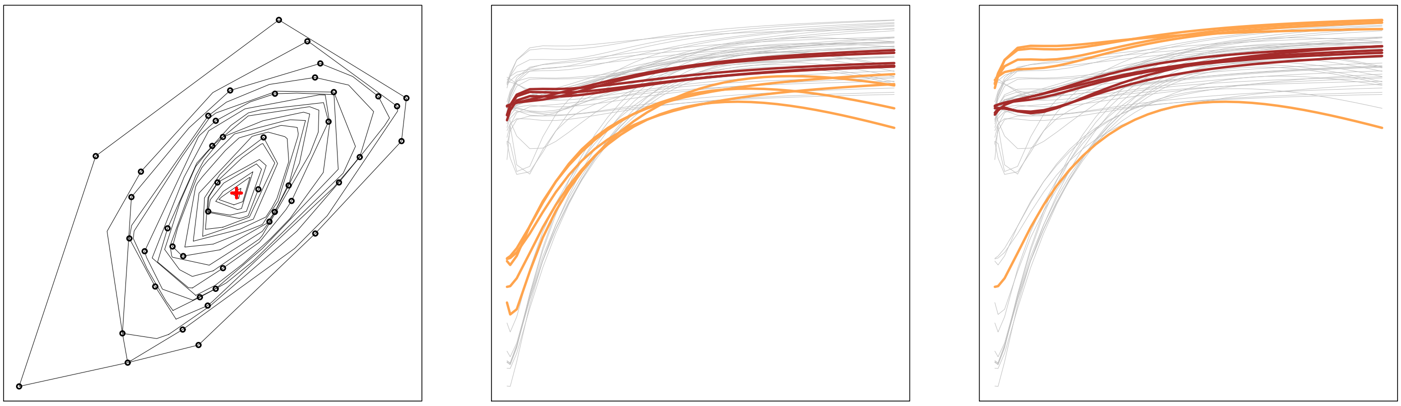


Fig. 1 Left panel: A random sample of bivariate data (black points) with the contours of the corresponding halfspace depth (convex polygons). The most centrally located point (thick red cross) is an analogue of the median for multivariate data. Middle and right panel: A random sample of functional data $[0, 1] \to \mathbb{R}$ with several deepest (thick dark red lines) and least deep (thick light brown lines) sample functions evaluated using an h-depth (middle panel) and an integrated depth (right panel). Different depth functionals capture different traits of the functional data.

We deal with function spaces $\mathcal{X}$ and discuss the recent progress made into the definitions and the properties of various depths proposed for functional observations [1, 2, 3, 4]. We will show that the infinite-dimensional nature of the space $\mathcal{X}$, as well as other peculiarities of functional data, do not enable direct extensions of the theory developed in $\mathbb{R}^d$ toward function spaces $\mathcal{X}$. Instead, fundamentally different approaches need to be developed. In the talk, we review a number of depths for functional data that have been proposed in the literature. We point to their similarities and differences and outline their theoretical properties. We will see that many open problems and partially resolved issues stimulate research in this fast-growing field of functional data analysis.

References

[1] Cuevas, A. and Fraiman, R.: On depth measures and dual statistics. A methodology for dealing with general data. *J. Multivariate Anal.*, 100(4):753–766 (2009).

[2] Fraiman, R. and Muniz, G.: Trimmed means for functional data. *TEST*, 10(2):419–440 (2001).

[3] López-Pintado, S. and Romo, J.: On the concept of depth for functional data. *J. Amer. Statist. Assoc.*, 104(486):718–734 (2009).

[4] Mosler, K.: Depth statistics. In Becker, C., Fried, R., and Kuhnt, S. (eds.) *Robustness and complex data structures*, Springer, Heidelberg, pages 17–34 (2013).

[5] Tukey, J. W.: Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 2*, pages 523–531 (1975). Canad. Math. Congress, Montreal, Que.

[6] Zuo, Y., Serfling, R.: General notions of statistical depth function. *Ann. Statist.*, 28(2):461–482 (2000). 2000.

Stanislav Nagy
Charles University, Department of Probability and Mathematical Statistics, Sokolovská 83, 186 75 Praha 8, Czech Republic,
e-mail: nagy@karlin.mff.cuni.cz

Variable selection in semiparametric bi-functional models

Silvia Novo, Germán Aneiros and Philippe Vieu

Abstract

Functional variables are more and more common in practical situations and developing techniques with high level of flexibility and interpretability has became a target in current statistical researches. To be adapted to these practical requirements, models and procedures able to reduce dimensionality are of first necessity (see [4]) and both semiparametric and sparse ideas are of great interest for reaching this purpose.

Accordingly, a new sparse semiparametric functional model is proposed, which tries to incorporate the influence of two functional variables in a scalar response in a flexible way, but involving interpretable parameters. One of the functional variables is included trough a single-index structure and the other one linearly, but trough the high-dimensional vector formed by its discretized observations. Due to the sparse nature of the linear component, variable selection is needed in the estimation task. The problem is that standard variable selection methods, coming from an adaptation of the multivariate methodology, such as the proposed in [3], can provide inadequate results. On the one hand, these procedures are affected by the strong dependence between variables, which in this case is directly derived from its functional origin. On the other hand, the great quantity of observations makes difficult obtaining results in reasonable amount of time. As a consequence, a new algorithm for variable selection in the linear part is proposed. This procedure takes advantage of the functional origin of the scalar covariates with linear effect, as the proposed in [1, 2]. Some asymptotic results will ensure the good performance of the method. Finally, Tecator's data will illustrate the great applicability of the presented methodology: good predictive power together with interpretability of the outputs.

Acknowledgements This work was supported in part by the Spanish Ministerio de Economía y Competitividad under Grant MTM2014-52876-R and Grant MTM2017-82724-R, in part by the Xunta de Galicia through Centro Singular de Investigación de Galicia accreditation under Grant ED431G/01 2019 and through the Grupos de Referencia Competitiva under Grant ED431C2020-014, and in part by the European Union (European Regional Development Fund - ERDF). The first author also thanks the financial support from the Xunta de Galicia and the European Union (European Social Fund - ESF), the reference of which is ED481A-2018/191.

References

[1] Aneiros, G., Vieu, P.: Variable selection in infinite-dimensional problems. Stat. Probab. Lett. **94**, 12–20 (2014)

[2] Aneiros, G., Vieu, P.: Partial linear modelling with multi-functional covariates. Comput. Stat. **30**(3), 647–671 (2015)

[3] Novo, S., Aneiros, G., Vieu, P.: Sparse semiparametric regression when predictors are mixture of functional and high-dimensional variables. TEST, https://doi.org/10.1007/s11749-020-00728-w (2020)

[4] Vieu, P.: On dimension reduction models for functional data. Statist. Probab. Lett. **136**, 134–138 (2018)

Silvia Novo
Universidade da Coruña, CITIC, e-mail: s.novo@udc.es

Germán Aneiros
Universidade da Coruña, CITIC, ITMATI, e-mail: german.aneiros@udc.es

Philippe Vieu
Université Paul Sabatier, Institut de Mathématiques de Toulouse, e-mail: philippe.vieu@math.univ-toulouse.fr

Local inference for functional data controlling the functional false discovery rate

Niels Lundtorp Olsen, Alessia Pini and Simone Vantini

Abstract

In functional data analysis (FDA), the object of statistical methods are functions, which are typically modeled as random elements of a Hilbert space [3]. In this framework inference is particularly challenging, since it deals with elements of infinite dimensional spaces.

A topic which is becoming more and more popular in Functional Data Analysis is local inference, i.e., the continuous statistical testing of a null hypothesis along a domain of interest. Iinference is performed locally on the domain, and the identification of the areas of the domain responsible for the rejection of the null hypothesis is provided. The principal issue in this topic is the infinite amount of tested hypotheses, which can be seen as an extreme case of multiple comparisons problem. Typically, local inferential techniques are either based on simultaneous confidence bands, which are provided with a fixed coverage probability [2, 4], or on the definition of a p-value function, which provides a p-value at each point of the domain, guaranteeing a control of a quantity related with the error rate on the whole domain. Focusing on this second line of research, depending on the quantity that is controlled, different methods can be defined. Many papers deal with the extension of the control of the family-wise error rate (FWER) - a well known quantity defined for multivariate data - to the case of functional data [7, 6].

In this work we focus instead on the false discovery rate (FDR), which is the expected proportion of false discoveries (rejected null hypotheses) among all discoveries, and was first introduced in the seminal paper by Benjamini and Hochberg [1]. We define FDR in the setting of functional data defined on a compact set of $\mathbb{R}^d$, and we further generalize this definition to functional data defined on a manifold. Furthermore, we introduce the functional Benjamini-Hochberg (fBH) procedure: a procedure able to control the previously defined functional FDR. We state some general conditions under which the fBH procedure provides control of FDR., and show how the procedure can be plugged-in with every parametric or nonparametric pointwise test, given that such test is exact. All details about the fBH procedure, as well as the proofs of all results described here are reported in [5].

Finally, to show the practical usefulness of our procedure, the proposed method - plugged-in with a nonparametric test - is applied to the analysis of a data set of daily temperatures on the Earth to identify the regions where the temperature has significantly increased over the last decades.

References

[1] Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B, **57** (1) 289–300 (1995) doi: 10.1111/j.2517-6161.1995.tb02031.x

[2] Choi, H., Reimherr, M.: A geometric approach to confidence regions and bands for functional parameters. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(1), 239-260 (2018) doi: 10.1111/rssb.12239

[3] Hsing, T., Eubank, R.: Theoretical foundations of functional data analysis, with an introduction to linear operators. John Wiley & Sons (2015)

[4] Liebl, D., Reimherr, M.: Fast and Fair Simultaneous Confidence Bands for Functional Parameters. arXiv:1910.00131 (2019)

[5] Lundtorp Olsen, N., Pini, A., Vantini, S.: False discovery rate for functional data. Test (2021)

[6] Pini, A., Vantini, S.: Interval-wise testing for functional data. Journal of Nonparametric Statistics **29** (2): 407–424 (2017) doi: 10.1080/10485252.2017.1306627

[7] Vsevolozhskaya, O., Greenwood, M., Holodov, D.: Pairwise comparison of treatment levels in functional analysis of variance with application to erythrocyte hemolysis. Ann. Appl. Stat., **8** (2), 905–925 (2014) doi:10.1214/14-AOAS723

Niels Lundtorp Olsen
University of Copenhaguen, e-mail: niels.olsen@math.ku.dk,

Alessia Pini
Università Cattolica del Sacro Cuore, e-mail: alessia.pini@unicatt.it

Simone Vantini
Politecnico di Milano e-mail: simone.vantini@polimi.it

Non-Parametric Functional Partially Linear Single-Index Models

Idir Ouassou

Abstract

Generalized linear models (GLM) provide a unified framework of likelihood for parametric regression analysis and are also an extension of linear models. Indeed, they allow to model, in a parametric way, the relation between a transformation of the average response and some covariates.

In this paper, we consider the estimation of generalized functional non-parametric partially linear single-index modeles of the form

$$g\left(\mu\left(X,Z\right)\right)=\eta\left(\alpha^{\top}X\right)+r\left(Z\right)\quad\text{where}\quad\mu\left(x,z\right)=E\left[Y\mid X=x,Z=z\right]$$

where $\eta(\cdot)$ is a unkown and smooth function (single index link function), α is a single index coefficient vector to be estimated, $r(\cdot)$ is the kernel estimator of the regression operator and g is a known link function.

These models would be called " Non Parametric Functional Partially Linear Single-Index Models (FNPPLSIM), where the systematic component in the model has a flexible semi-parametric form with a general link function. We propose an efficient and practical approach to (i) estimate the single-index link function, (ii) estimate single-index coefficients as well as (iii) the non-parametric component $r(\cdot)$ of the model. The estimation procedure is developed by applying the quasi-likelihood estimation. After constructing the estimators of the function and the coefficient described above, we present a wide range of properties of these estimators and we will give the rate of convergence of the constructed estimators. Moreover, by making use of normalized cubic B-splines basis approximation and the Fisher score iterations, we show the efficiency and the practicality of our estimation methodology on some simulated data.

References

[1] Ping y., Jiang D. and Zhongzhan, Z.: Single-Index partially functional linear regression model. Statistical Papers, DOI: 10.1007/s00362-018-0980-6, (2018)

[2] Li, W. and Guanqun, C.: Efficient estimation for generalized partially linear single-index models. Bernoulli, **24**(2), 1101–1127 (2018)

[3] Carroll, R.J., Fan, J., Gijbels, I. and Wand, M. P.: Generalized partially linear single-index models. Journal of the American Statistical Association, **92**, 477–489 (1997)

[4] De Boor, C.: A practical guide to splines, Revised Edition, vol. **27** of Applied Mathematical Sciences. Springer-Verlag, Berlin (2001)

[5] Wang, L. and Yang, L.: Spline estimation of single-index models. Statistica Sinica, **19**, 765–783 (2009)

[6] Chen, J., Li, D., Liang, H. and Wang, S.: Semiparametric GEE analysis in partially linear single-index models for longitudinal data. Annals of Statistics, **43**, 1682–1715 (2015)

[7] Ma, S., Liang, H. and Tsai, C. L.: Partially linear single index models for repeated measurements. Journal of Multivariate Analysis, **130**, 354–375 (2014)

[8] Wang, J. L., Xue, L., Zhu, L. and Chong, Y. S.: Estimation for a partial-linear single-index model. Annals of Statistics, **38**, 246–274 (2010)

Idir Ouassou
Ecole Nationale des Sciences Appliquées, Marrakech, Morocco, e-mail: i.ouassou@uca.ma

Quantifying brain age prediction uncertainty from imaging using scalar-on-image quantile regression

Marco Palma, Shahin Tavakoli, Julia Brettschneide and Thomas E. Nichols

Abstract

Prediction of subject age from brain anatomical MRI has the potential to provide a sensitive summary of brain changes, indicative of different neurodegenerative diseases. However, existing studies typically neglect the uncertainty of these predictions. In this work we take into account this uncertainty by applying methods of functional data analysis. We propose a penalised functional quantile regression model of age on brain structure with cognitively normal (CN) subjects in the Alzheimer's Disease Neuroimaging Initiative (ADNI) and use it to predict brain age in Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD) subjects. Unlike the machine learning approaches available in the literature of brain age prediction, which provide only point predictions, the outcome of our model is a prediction interval for each subject. The prediction accuracy obtained with this model is similar to more sophisticated approaches, while being also more principled and interpretable. The gap between predicted and chronological age correlates with cognitive decline.

References

[1] Palma, M., Tavakoli, S., Brettschneider, J., Nichols, T.E. for the Alzheimer's Disease Neuroimaging Initiative: Quantifying uncertainty in brain-predicted age using scalar-on-image quantile regression. NeuroImage, **219**, 116938 (2020)

Marco Palma
Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom,
e-mail: M.Palma@warwick.ac.uk

Shahin Tavakoli
Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom
e-mail: s.tavakoli@warwick.ac.uk

Julia Brettschneider
Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom
The Alan Turing Institute, London, NW1 2DB, United Kingdom
e-mail: julia.brettschneider@warwick.ac.uk

Thomas E. Nichols
Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Oxford, OX3 7LF, United Kingdom,
Wellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, OX3 9DU, United Kingdom
e-mail: thomas.nichols@bdi.ox.ac.uk

Analysis on Stratified Spaces and an RNA Based Investigation of the SARS-CoV-2 Hypotheses

Vic Patrangenaru and Roland Moore

Abstract

After an introduction to data analysis on stratified spaces, with a special emphasis on tree spaces (Billera et. al. (2001)), one considers phylogenetic tree data to investigate RNA sequences of the SARS-CoV-2 virus (Shen(2021)), and two of its possible origins, formulated so far: the bat SARS origin, and the Wuhan 'lab leak' hypothesis(Bloom et al.(2021). In particular the stickiness of the intrinsic sample mean (Hotz et al(2011)) applied to phylogenetic trees from United States SARS-CoV-2 RNA sequences, is used in testing such hypotheses.

References

[1] L. J. Billera, S. P. Holmes and K. Vogtmann: Geometry of the space of phylogenetic trees, *Adv. Appl. Math.* **27** , 733—767 (2001).

[2] 2. J. D. Bloom, Y. A. Chan, R. S. Baric, P. J. Bjorkman, S. Cobey, B. E. Deverman, D. N. Fisman, R. Gupta, A. Iwasaki, M. Lipsitch, R. Medzhitov, R. A. Neher, R. Nielsen, N. Patterson, T. Stearns, E. van Nimwegen, M. Worobey, D. A. Relman: Investigate the origins of COVID-19. SCIENCE, **372**, Issue 6453, p. 694 (2021).

[3] C. Shen. *Topological Data Analysis for Medical Imaging and RNA Data Analysis on Tree Spaces*, PhD dissertation, Florida State University (2021).

[4] T. Hotz, S. Huckemann, H. Le, J. S. Marron, J. C. Mattingly, E. Miller, J. Nolen, M. Owen, V. Patrangenaru. S. Skwerer: Sticky Central Limit Theorems on Open Books. *Annals of Applied Probability,* **23**, 2238–2258 (2013).

Vic Patrangenaru
Florida State University, Department of Statistics, Tallahassee, FL 32306-4330, USA, e-mail: vic@stat.fsu.edu

Roland Moore
Florida State University, Department of Statistics, Tallahassee, FL 32306-4330, USA, e-mail: rm16n@my.fsu.edu

A functional approach to linear discriminant analysis: Classification of probability density functions using the Bayes space methodology

Ivana Pavlů, Karel Hron, Alessandra Menafoglio and Peter Filzmoser

Abstract

Classification of observations into one of pre-existing classes is one of the basic tasks in both multivariate and functional data analysis. While considering probability density functions (PDFs) as the objects of classification, standard functional methods cannot be used as they are usually implemented for functions from the standard L^2 space [4]. Some specific properties of PDFs (scale invariance, relative scale, unit integral), however, cause the L^2 space to fail. To solve this issue, the idea of Bayes spaces [5] is used for the representation of PDFs as this way their key features are maintained. The centred log-ratio transformation, a common tool of compositional data analysis [1], is frequently used to represent the original data (from Bayes space) into the L^2 space. This way, standard functional procedures can be applied onto the transformed data for further analysis.

Here, a functional linear discriminant analysis model [2] is modified for classification of PDFs. As the raw data often come in discretized rather than continuous form, their functional is addressed by using the so-called compositional smoothing splines [3]. To filter out possible noise, a projection into a reduced discriminant space is performed on functional observations as well as the class representatives (centroids) of the predefined classes. The classification criterion itself is based on the generalized Bayes' formula – it minimizes the distance between the linear projections of both class centroids and examined observations. The introduced method is then used on geological data consisting of particle size distribution of 250 soil samples from Moravia region, Czech Republic. The measuring took place at four different sites, providing natural classes for the performed classification. 5-fold cross-validation was used for estimating the quality of classification, showing significant differences between localities.

References

[1] Aitchison, J.: The Statistical Analysis of Compositional Data. Chapman & Hall, London (1986)

[2] James, G. M., Hastie, T. J.: Functional linear discriminant analysis for irregularly sampled curves. Journal of the Royal Statistical Society: Series B (Statistical Methodology), **63**, 533-550 (2001)

[3] Machalová, J., Talská, R., Hron, K., Gába, A.: Compositional splines for representation of density functions. Computational Statistics, DOI: 10.1007/s00180-020-01042-7 (2020)

[4] Ramsay, J., Silverman, B. W.: Functional Data Analysis, 2nd edition. Springer, New York (2005)

[5] van den Boogaart, K. G., Egozcue, J. J., Pawlowsky-Glahn, V.: Bayes Hilbert Spaces. Australian & New Zealand Journal of Statistics 56(2):171-194 (2014)

Ivana Pavlů
Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, 17. listopadu 12, CZ-77146 Olomouc, Czech Republic, e-mail: ivana.pavlu@upol.cz

Karel Hron
Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, 17. listopadu 12, CZ-77146 Olomouc, Czech Republic, e-mail: karel.hron@upol.cz

Alessandra Menafoglio
MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy, e-mail: alessandra.menafoglio@polimi.it

Peter Filzmoser
Institute of Statistics and Mathematical Methods in Economics, TU Wien, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria,
e-mail: peter.filzmoser@tuwien.ac.at

Functional clustering via multivariate clustering

Belén Pulido, Alba M. Franco-Pereira and Rosa E. Lillo

Abstract

Clustering techniques applied to multivariate data are popular nowadays and, thus, have been fully studied in the literature. It is remarkable that despite widely used multivariate techniques are adapted to functional data, clustering techniques are less well known when dealing with this data type. In this work, we address this problem making use of the epighaph and the hypograph indexes, as well as their modified versions that were firstly defined in [3]. The combination of these indexes has been previously considered, for example, for outliers detection [1], for defining a functional boxplot [4] and for testing homogeneity between two samples [2]. The functional clustering technique that we propose consist of applying these indexes to a given functional data set and thereby, converting it from a functional data problem into a multivariate problem, where the multivariate clustering techniques can be applied. Since these indexes are able to reflect the shape of the curves, their first and second derivatives are also taken into account. Then, several multivariate clustering techniques are considered and compared in terms of different metrics.

Our procedure is applied to different data sets, both simulated and real ones. Moreover, our methodology is also compared to some clustering techniques originally designed for functional data. Two recent results concerning functional clustering techniques can be found in [5] and [6]. In view of the results, we conclude that the proposed methodology is competitive in terms of computational time and performance.

References

[1] Arribas-Gil, A., Romo, J.: Shape outlier detection and visualization for functional data: the outliergram. Biostatistics, **15(4)**, 603-619 (2014)

[2] Franco-Pereira, A.M., Lillo, R.E.: Rank tests for functional data based on the epigraph, the hypograph and associated graphical representations. Advances in Data Analysis and Classification, **14(3)**, 651-676 (2020)

[3] Franco-Pereira, A. M., Lillo, R. E., Romo, J.: Extremality for functional data. In: Recent advances in functional data analysis and related topics 131–134. Physica-Verlag HD (2011)

[4] Martín-Barragán, B., Lillo, R.E., Romo, J.: Functional boxplots based on half-regions. Journal of Applied Statistics, 1088-1103 (2018)

[5] Martino, A., Ghiglietti, A., Ieva, F., Paganoni, A. M.: A k-means procedure based on a Mahalanobis type distance for clustering multivariate functional data. Statistical Methods and Applications, **28(2)**, 301-322 (2019)

[6] Zambom, A. Z., Collazos, J. A., & Dias, R.: Functional data clustering via hypothesis testing k-means. Computational Statistics, **34(2)**, 527-549 (2019)

Belén Pulido
uc3m-Santander Big Data Institute, Universidad Carlos III de Madrid, Getafe, Madrid, Spain, e-mail: belenpulidobravo@gmail.com

Alba M. Franco-Pereira
Department of Statistics and O.R., Universidad Complutense de Madrid, Madrid, Spain,
uc3m-Santander Big Data Institute, Universidad Carlos III de Madrid, Getafe, Madrid, Spain, e-mail: albfranc@ucm.es

Rosa E. Lillo
Department of Statistics, Universidad Carlos III de Madrid, Getafe, Madrid, Spain,
uc3m-Santander Big Data Institute, Universidad Carlos III de Madrid, Getafe, Madrid, Spain, e-mail: rosaelvira.lillo@uc3m.es

scikit-fda: A Python package for Functional Data Analysis

Carlos Ramos-Carreño, José Luis Torrecilla and Alberto Suárez

Abstract

We present *scikit-fda* (https://github.com/GAA-UAM/scikit-fda), a Python package for the statistical analysis of functional data. The library is fully integrated in the SciPy ecosystem, which is extensively used by machine learning researchers and practitioners. The package *scikit-fda* provides structures for the representation and manipulation of functional data both in the discretized and basis expansion forms. It includes methods for preprocessing, such as smoothing and registration, exploratory analysis, such as visualization and outlier detection, and for machine learning, such as clustering, classification, and regression. Special care has been taken in the design of these procedures so that they follow the existing practices in the SciPy ecosystem. Specifically, they conform to the application programming interface (API) of *scikit-learn*, a powerful machine learning library in Python. This design makes it possible to directly apply the *scikit-learn* methods to multivariate representations of the functional data. Furthermore, *scikit-learn* utilities, such as those for hyperparameter tuning and model selection, can be combined with the purely functional machine learning methods implemented in *scikit-fda* in a seamless manner.

References

[1] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G.: API design for machine learning software: experiences from the scikit-learn project In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 108–122 (2013)

[2] Febrero-Bande, M., Oviedo de la Fuente, M.: Statistical computing in functional data analysis: the R package fda.usc. JSS. **51**(4), 1–28 (2012)

[3] Millman, K. J., Aivazis, M.: Python for Scientists and Engineers. CiSE, **13**, 9–12 (2011)

[4] Oliphant, T. E.: Python for Scientific Computing. CiSE, **9**, 10–20 (2007)

[5] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. JMLR, **12**, 2825–2830 (2011)

[6] Ramsay, J., Wickham, H., Graves, S., Hooker, G.: fda: Functional Data Analysis. R package version 2.4.8. (2018)

Carlos Ramos-Carreño
Universidad Autónoma de Madrid, Spain, e-mail: carlos.ramos@uam.es

José Luis Torrecilla
Universidad Autónoma de Madrid, Spain, e-mail: joseluis.torrecilla@uam.es

Alberto Suárez
Universidad Autónoma de Madrid, Spain, e-mail: alberto.suarez@uam.es

Variable selection in multivariate functional linear model via the Lasso

Angelina Roche

Abstract

In more and more applications, a quantity of interest may depend on several covariates, with at least one of them infinite-dimensional (e.g. a curve). The aim of this contribution is to study the link between a real response Y and a vector of covariates $\mathbf{X} = (X^1, ..., X^p)$. For that, we suppose we observe $\{(Y_i, \mathbf{X}_i), i = 1, ..., n\}$ where $\mathbf{X}_i = (X_i^1, ..., X_i^p)$ is a vector of covariates which can be of different nature (curves or vectors). More precisely, we suppose that, for all $j = 1, ..., p$, $i = 1, ..., n$, $X_i^j \in \mathbb{H}_j$ where $(\mathbb{H}_j, \|\cdot\|_j, \langle\cdot,\cdot\rangle_j)$ is a separable Hilbert space. Our covariate $\{\mathbf{X}_i\}_{1\le i\le n}$ then lies in the space $\mathbf{H} = \mathbb{H}_1 \times ... \times \mathbb{H}_p$, which is also a separable Hilbert space with its natural scalar product

$$\langle \mathbf{f}, \mathbf{g}\rangle = \sum_{j=1}^{p} \langle f_j, g_j\rangle_j \text{ for all } \mathbf{f} = (f_1, ..., f_p), \mathbf{g} = (g_1, ..., g_p) \in \mathbf{H}$$

and usual norm $\|\mathbf{f}\| = \sqrt{\langle \mathbf{f}, \mathbf{f}\rangle}$. We consider in this context the *multivariate functional linear model*,

$$Y_i = \sum_{j=1}^{p} \langle \beta_j^*, X_i^j\rangle_j + \varepsilon_i = \langle \boldsymbol{\beta}^*, \mathbf{X}_i\rangle + \varepsilon_i,$$

where, $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^*, ..., \boldsymbol{\beta}_p^*) \in \mathbf{H}$ is unknown and $\{\varepsilon_i\}_{1\le i\le n} \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$. We suppose that $\{\mathbf{X}_i\}_{1\le i\le n}$ can be either fixed elements of $\mathbf{H}$ (fixed design) or i.i.d centered random variables in $\mathbf{H}$ (random design) and that it is independent of $\{\varepsilon_i\}_{1\le i\le n}$. To select the relevant covariates in this context, we propose an adaptation of the Lasso method. Two estimation methods are defined. The first one consists in the minimisation of a criterion inspired by classical Lasso inference under group sparsity ([9, 8]) on the whole multivariate functional space $\mathbf{H}$:

$$\widehat{\boldsymbol{\beta}}_\lambda \in \arg\min_{\boldsymbol{\beta}=(\beta_1,...,\beta_p)\in\mathbf{H}} \left\{ \frac{1}{n}\sum_{i=1}^{n} (Y_i - \langle \boldsymbol{\beta}, \mathbf{X}_i\rangle)^2 + 2\sum_{j=1}^{p} \lambda_j \|\beta_j\|_j \right\}.$$

The second one minimises the same criterion but on a finite-dimensional subspace. A data-driven dimension selection criterion is proposed to select the dimension m, inspired by the works of [1] and their adaptation to the functional linear model [5, 6, 4, 3]

$$\widehat{m} \in \arg\min_m \left\{ \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \langle \widehat{\boldsymbol{\beta}}_{\lambda,m}, \mathbf{X}_i\rangle \right)^2 + \kappa\sigma^2 \frac{m\log(n)}{n} \right\},$$

where $\kappa > 0$ is a constant which can be calibrated by a simulation study or selected from the data by methods stemmed from slope heuristics (see e.g. [2]). Sparsity-oracle inequalities are proven in case of fixed or random design in our infinite-dimensional context. To calculate the solutions of both criteria, we propose a coordinate-wise descent algorithm, inspired by the *glmnet* algorithm ([7]). A numerical study on simulated and experimental datasets illustrates the behavior of the estimators.

References

[1] Barron, A., Birgé, A., Massart, P.: Risk bounds for model selection via penalization. Probab. Theory Relat. Fields, **113** (3), 301–413 (1999).

[2] Baudry, J.-P., Maugis, C., Michel, B.: Slope heuristics: overview and implementation. Stat. Comput., **22** (2), 455–470 (2012).

[3] Brunel, E., Mas, A., Roche, A.: Non-asymptotic adaptive prediction in functional linear models. J. Multivariate Anal., **143**, 208–232 (2016).

[4] Brunel, E., Roche, A.: Penalized contrast estimation in functional linear models with circular data. Statistics, **49** (6), 1298–1321 (2015).

[5] Comte, F., Johannes, J.: Adaptive estimation in circular functional linear models. Math. Methods Statist., **19** (1), 42–63 (2010).

[6] Comte, F., Johannes, J.: Adaptive functional linear regression. Ann. Statist., **40** (6), 2765–2797 (2012).

[7] Friedman, J., Hastie, T., Höfling, H., Tibshirani, R.: Pathwise coordinate optimization. Ann. Appl. Stat., **1** (2), 302–332 (2007).

[8] Lounici, K., Pontil, M., van de Geer, S., Tsybakov, A. B.: Oracle inequalities and optimal inference under group sparsity. Ann. Statist., **39**(4), 2164–2204 (2011).

[9] Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B Stat. Methodol., **68** (1),49–67 (2006).

Angelina Roche
Université Paris-Dauphine, CNRS, UMR 7534, CEREMADE, 75016 Paris, FRANCE, e-mail: roche@ceremade.dauphine.fr

Factor analysis for high-dimensional functional time series

Chen Tang, Han Lin Shang and Yanrong Yang

Abstract

The needs in handling increasing volumes of data with complicated structures give rises to modeling large sets of functional time series, that is, high-dimensional functional time series. In high-dimensional functional time series, functions of each set (cross-section) contain serial dependence while functions from different cross-sections may also be correlated. In high-dimensional functional time series, the ’curse of dimensionality’ is twofold: the infinite dimensionality of functional data and the high-dimensionality of the number of cross-sections. To address this issue, our paper proposes a factor model for high-dimensional functional time series. By isolating the heterogeneity along cross-sections, the high-dimensional functional time series can be reduced to functional time series with lower dimensions. Through a functional dynamic factor model, the dimension-reduced functional time series are further reduced into low-dimensional scalar factor matrices such that all the temporal dynamics contained in the original high-dimensional functional time series are extracted to facilitate forecasting. Through a Monte Carlo simulation, we demonstrate the performance of the proposed method in model fitting. In a empirical study of the Japanese subnational age-specific mortality rates, the proposed model produces more accurate forecasts than several existing methods in joint modeling the mortality rates of different prefectures, the dimensional reduced factor matrices can convey the useful information in the original high-dimensional functional time series reasonably well.

Chen Tang
Australian National University, e-mail: chen.tang@anu.edu.au

Han Lin Shang
Macquarie University, e-mail: hanlin.shang@mq.edu.au

Yanrong Yang
Australian National University, e-mail: yanrong.yang@anu.edu.au

A Kernel Nonlinear Principal Component Analysis to build a spatial inequality social indicator

Jared Abigail Valencia

Abstract

Kernel-based principal components analysis (KPCA) method is a generalization of the nonlinear PCA methods using positive definite kernels. The objective is to find projected variables of the feature space into a induced kernel, with maximum variance. Moreover, if we use a distance decay function as a kernel, the KPCA will handle spatial interaction in our data.

KPCA has been used for urban growth simulation, pattern recognition ,facial recognition, nonlinear dynamic process monitoring, time series prediction, emotion learning and recognition, among others. Here, in this paper, we are interested in social inequality problems.

Social inequality occurs when a group of people receive different treatment as a consequence of their social position, economic situation, gender, the religion they profess and other aspects. All of the factors mentioned before intersect in territories, that is why a lot of social studies suggest that the place where a person is born or lives influences in their societal development, as well as access and distribution of social and economic opportunities, resulting in territorial inequalities.

Therefore, our goal is to analyze the variables of the Survey of Living Conditions 2014, Sixth Round, for the construction of an indicator of social inequality that considers the spatial and geographical influence on the data in order to measure social inequality in Ecuador, specifically in Quito's Metropolitan District, using KPCA with a distance decay function as a kernel.

As a result, there were a large number of regions in Quito, in particular peripherical zones, that presented high amounts of social inequality. In fact, this was not new because this trend of fewer opportunities in peripheral areas occurs worldwide.

An interesting observation was that around 75% of analyzed regions presented more than 70% of social inequality and 7% presented less than 14% of social inequality. We compared the results with a study of unsatisfied basic needs carried out by the Municipality of Quito's Metropolitan District and our results matched.

To conclude, the KPCA method gave us a good representation of social reality in Quito taking into consideration spatial interaction. Indeed, we reinforce the fact that social inequality in Quito is an unsolved problem, and authorities should propose new development and inclusion policies in order to enhance the quality of life of Quito's citizens.

References

[1] CEPAL:La hora de la igualdad: brechas por cerrar, caminos por abrir. CEPAL, 2010.

[2] CEPAL:La matriz de la desigualdad social en America Latina. CEPAL, 2016.

[3] CEPAL:Desarrollo y migracion: Desafıos y oportunidades en los paıses del norte de Centroamerica CEPAL, 2019.

[4] Chen, Qisong, Xiaowei Chen, and Yun Wu:Optimization algorithm with kernel pca to supportvector machines fortime series prediction. ISSN 1796-203X, 5(3):380, 2010.

[5] Choi, Sang Wook and In Beum Lee:Nonlinear dynamic process monitoring based on dynamickernel pca. Chemical engineering science, 59(24):5897–5908, 2004.1

[6] Ebied, Hala M:Feature extraction using pca and kernel-pca for face recognition. In 2012,8th International Conference on Informatics and Systems (INFOS), pages MM–72. IEEE, 2012.

[7] Estupinan Velasco, John Alexanderet al.:Aprendizaje y reconocimiento de emociones en imagenes empleando kernel pca. B.S. thesis, Uniandes, 2012.

[8] Feng, Yongjiu and Yan Liu:A cellular automata model based on nonlinear kernel principal component analysis for urban growth simulation. Environment and Planning B: Planning and Design,40(1):117–134, 2013.

[9] Jayasinghe, Maneka and Christine Smith:Poverty implications of household headship and foodconsumption economies of scales: A case study from sri lanka. Social Indicators Research, pages1–29, 2021.

[10] Naveed, Tanveer Ahmed, David Gordon, Sami Ullah, and Mary Zhang:The construction of anasset index at household level and measurement of economic disparities in punjab (pakistan) byusing mics-micro data. Social Indicators Research, pages 1–23, 2021.

[11] Samaniego Ponce, Pablo:Evolucion de la pobreza y la desigualdad en Quito. Questiones Urbano Regionales.

[12] Sholkof, B.:Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation,10:1299 – 1319, 1998.

Jared Valencia
Escuela Politécnica Nacional, Quito - Ecuador e-mail: jared.valencia.salvador@gmail.com

Independent Component Analysis Techniques for Functional Data

Marc Vidal and Ana M. Aguilera

Abstract

The concept of statistical independence in infinite-dimensional Hilbert spaces has attracted the interest of many researchers. As no distributional assumption is considered for a set of random functions, the extension of independent component analysis (ICA) to functional data faces several problems for its generalization. To accommodate this scenario, the definition of independence is constrained under the umbrella of an orthogonal projection. Thus, numerous solutions may arise to solve the ICA problem in functional spaces. A natural basis choice to represent a random variable in L^2 is given by the covariance kernel eigenfunctions. However, too frequently, there is a need for regularizing or smoothing the estimated covariance functions. In this sense, the functional independent measure proposed is derived from the kurtosis of a smoothed principal components expansion. We develop a generalization of Silverman's method to obtain a smoothed orthonormal basis in the L^2 sense, for which the functional independent component model is well suited. Therefore, our modelling strategy can be seen as a bi-smoothed procedure that simultaneously regularizes the independent components in terms of the principal components eigendirections, but also with respect to a roughness penalty into the orthonormality constraint of the aforementioned functions. A kurtosis operator is defined to capture intrinsic patterns aimed at maximizing the independence of the components over the smoothed observations.

With the development of brain-computer interfaces and technologies that depend on real-time information from brain activity (e.g., via electroencephalography, EEG), there is a compelling need to reduce the signal with minimum brain potential loss. Moreover, the processing of such neural information into qualitative biofeedback requires flexible techniques able to extract highly interpretable patterns from the observed data. Because of the transient nature and the complex morphology of EEG data, B-splines provide a good alternative to represent the non-sinusoidal behaviour of the neural oscillations due to their well-behaved local smoothing. The proposed functional ICA approach is based on a P-spline penalty computed at a relatively low cost to enhance the performance of the ICA estimators. As a motivating example, a novel functional data framework is proposed for extracting and removing smoothed artifacts in order to estimate genuine brain sources from EEG functional representations generated on large and densely sampled grids. We incorporate shrinkage in the functional principal components estimation process to avoid numerical instabilities. An ad-hoc cross-validation method is derived using a backward shrinkage transformation to select the penalty parameter. Although our results provide evidence of the effectiveness and flexibility of our methods, a potential drawback is the emerging non-linearities in the EEG signal, suggesting the extension of functional ICA to non-parametric realms.

References

[1] Aguilera, A.M., Aguilera-Morillo, M.C. Penalized PCA Approaches for B-Spline Expansions of Smooth Functional Data. *Appl. Math. Comput.* **219**, 7805–7819 (2013).

[2] Ocaña, F.A., Aguilera, A.M., Valderrama, M.J. Functional Principal Component Analysis by Choice of Norm. *J. Multivar. Anal.* **71**, 262–276 (1999).

[3] Silverman, B.W. Smoothed Functional Principal Components Analysis by Choice of Norm. *Ann. Stat.* **24**, 1-24 (1996).

[4] Vidal, M., Rosso, M., Aguilera, A.M. Bi-Smoothed Functional Independent Component Analysis for EEG Artifact Removal. *Mathematics*, **9**, 1243 (2021).

Marc Vidal
Ghent University, e-mail: marc.vidalbadia@ugent.be

Ana M. Aguilera
Granada University, e-mail: aaguiler@ugr.es

Kernel Mean Embeddings for Functional Data Analysis

George Wynne, Andrew B. Duncan, Stanislav Nagy and Mikołaj Kasprzak

Abstract

Kernel mean embeddings (KMEs) [1] have enjoyed wide success in statistical machine learning over the past fifteen years. They offer a non-parametric method of reasoning with probability measures by mapping measures into a reproducing kernel Hilbert space (RKHS). The RKHS facilitates easy to compute, closed form expressions which makes the methodology practical and amenable to statistical analysis. Much of the existing theory and practice has revolved around Euclidean data whereas functional data has received very little investigation. Likewise, in functional data analysis (FDA) the technique of KMEs has not been explored.

In this talk I will describe work which aims to bridge this gap [2] by defining kernels which take functional inputs. In this context, KMEs offer an alternative paradigm than the common practice of projecting data to finite dimensions then employing classical finite dimensional statistical procedures. Indeed, the KME framework can handle infinite dimensional input spaces, offers an elegant theory and leverages the spectral structure of functional data. Applications include two-sample testing, goodness-of-fit testing and functional depth.

References

[1] Muandet, K., Fukumizu, K., sriperumbudur, B, Schölkopf, B.: Kernel Mean Embedding of Distributions: A Review and Beyond. Foundations and Trends® in Machine Learning (2017)

[2] Wynne, G, Duncan, A.D.: A Kernel Two-Sample Test for Functional Data. *arXiv:2008.11095* (2020)

George Wynne
Imperial College London, e-mail: g.wynne18@imperial.ac.uk

Andrew B. Duncan
Imperial College London, e-mail: a.duncan@imperial.ac.uk

Stanislav Nagy
Charles University, e-mail: nagy@karlin.mff.cuni.cz

Mikołaj Kasprzak
University of Luxembourg, e-mail: mikolaj.kasprzak@uni.lu

Temperature Forecasting: A Spatial Functional Time Series Approach

Ruofan Xu, Han Lin Shang and Yanrong Yang

Abstract

The literature of econometrics has recently seen a growing interest on modelling the climate change. Motivated by the nature of climate data and the associated modelling challenges, this paper considers a spatial functional time series (SFTS) framework using a functional principal component analysis (FPCA) approach. The newly proposed methodology utilizes the dependence over spatio-temporal units through a long-run covariance estimator, and is capable of extracting the core features in a data rich environment. The asymptotic properties of the proposed estimator are established accordingly. In the empirical study, our analysis shows significant improvement in both fitting and forecasting accuracy compared with some existing methods in the literature.

Ruofan Xu
Monash University, Melbourne VIC Australia, e-mail: ruofan.xu@monash.edu

Han Lin Shang
Macquarie University, Sydney NSW Australia, e-mail: hanlin.shang@mq.edu.au

Yanrong Yang
Australian National University, Canberra ACT Australia, e-mail: yanrong.yang@anu.edu.au

A Basis Approach to Surface Clustering

Adriano Zanin Zambom and Qing Wang and Ronaldo Dias

Abstract

We introduce a novel method for clustering surfaces. The proposal involves first using basis functions in a tensor product to smooth the data and thus reduce the dimension to a finite number of coefficients, and then using these estimated coefficients to cluster the surfaces via the k-means algorithm. An extension of the algorithm to clustering tensors is also discussed. We show that the proposed algorithm exhibits the property of strong consistency, with or without measurement errors, in correctly clustering the data as the sample size increases. Simulation studies suggest that the proposed method outperforms the benchmark k-means algorithm which uses the original vectorized data. In addition, an EGG real data example is considered to illustrate the practical application of the proposal.

Adriano Zanin Zambom
California State University Northridge, e-mail: adriano.zambom@csun.edu

Qing Wang
Wellesley College, e-mail: qwang@wellesley.edu

Ronaldo Dias
State University of Campinas, e-mail: dias@ime.unicamp.br