SIS 2017
Statistics and Data Science: new challenges, new generations

28–30 June 2017
Florence (Italy)

Proceedings of the Conference of the Italian Statistical Society

edited by
Alessandra Petrucci
Rosanna Verde

FIRENZE UNIVERSITY PRESS
2017

SIS 2017. Statistics and Data Science: new challenges, new generations : 28-30 June 2017 Florence (Italy) : proceedings of the Conference of the Italian Statistical Society / edited by Alessandra Petrucci, Rosanna Verde. – Firenze : Firenze University Press, 2017.
(Proceedings e report ; 114)

http://digital.casalini.it/9788864535210

ISBN 978-88-6453-521-0 (online)

Peer Review Process
All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Committees of the individual series. The works published in the FUP catalogue are evaluated and approved by the Editorial Board of the publishing house. For a more detailed description of the refereeing process we refer to the official documents published on the website and in the online catalogue of the FUP (www.fupress.com).

Firenze University Press Editorial Board
A. Dolfi (Editor-in-Chief), M. Boddi, A. Bucelli, R. Casalbuoni, M. Garzaniti, M.C. Grisolia, P. Guarnieri, R. Lanfredini, A. Lenzi, P. Lo Nostro, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, G. Nigro, A. Perulli, M.C. Torricelli.

Università degli Studi di Firenze
Firenze University Press
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

Index

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

Functional principal component analysis of quantile curves

Analisi in componenti principali funzionali di curve quantiliche

M. Ruggieri, F. Di Salvo and A. Plaia

Abstract Literature on functional data analysis is mainly focused on estimation of individuals curves and characterization of average dynamics. The idea underlying this proposal is to focus attention on other particular features of the distribution of the observed data, moving from mean functions towards functional quantiles. The motivating examples are functional data sets that are collections of high frequency data recorded along time. As quantiles provide information on various aspects of a time series, we propose a modelling framework for the joint estimation of functional quantiles, varying along time, and functional principal components, summarizing some common dynamics shared by the functional quantiles.

Abstract *La letteratura sull'analisi di dati funzionali é prevalentemente rivolta alla modellazione e stima delle singole curve aleatorie e alla caratterizzazione del momento primo. L'idea di base di questo lavoro é considerare altri aspetti della distribuzione dei dati osservati, spostando l'attenzione verso i quantili funzionali. La tipologia di dati a cui questa analisi si rivolge é rappresentata da dati ad alta frequenza osservati nel tempo. Poiché i quantili sintetizzano informazioni sulle dinamiche temporali di una serie storica, si propone un approccio per la stima di quantili funzionali, in corrispondenza di diversi valori di probabilitá, e per la derivazione di componenti principali funzionali che ne riassumano le dinamiche comuni.*

Key words: functional data, nonparametric quantile regression, penalized splines, functional principal components

M. Ruggieri e-mail: mariantonietta.ruggieri@unipa.it,
F. Di Salvo, e-mail: francesca.disalvo@unipa.it,
A. Plaia, e-mail: antonella.plaia@unipa.it

Department of Economics Business and Statistics, University of Palermo, viale delle Scienze, Palermo.

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

1 Introduction

Let consider high dimensional data observed at discrete times; although we observe a finite number of measured values, they are often analyzed as if they were defined in continuous time.

Traditional analysis concerns the conditional distribution at each time point, while in a functional data analysis (FDA) approach each time series is considered as a sample generated by a random curve, varying over a continuum; in both cases, more frequently the goal is the centre of the conditional distribution or a mean function describing the pattern of the set of functions.

In the univariate regression setting, quantile regression models the quantiles of the conditional distribution of the response variable; this is a valuable alternative to the conditional mean, when the interest is in the tails of the distribution or in presence of model mis-specification (see [4] and [5]). With the increasing demand of statistical tools for FDA is therefore natural to try to extend the definition and the estimation of quantile functions for infinite-dimensional data. However the extension of quantile function to a multivariate setting is not straightforward, because quantiles are basically defined by ordering values of a random variable. Since there is no natural order for R^n when $n \geq 2$, there is no obvious extension and a number of efforts has been devoted to this problem in the last years.

Our proposal explore the performance of the multi-way functional principal component analysis $(FPCA)$ when functional quantiles of different order are simultaneously considered. There are some previous works motivating our idea and in particular [3] and [1]. An approach on generalized regression quantiles with their synthesis by means of a small number of principal components is proposed in [3] in a FDA framework.

Fraiman et al. [1] define directional quantiles and extend a projection-based definition of quantiles to infinite-dimensional Hilbert and Banach spaces; the authors develop a factor analysis based on principal directions and robust principal directions. The main results in [1] are based on an intuitive definition of directional quantiles, indexed by an order $\alpha \in (0,1)$ and a direction u in the unit sphere; the directional quantile describe the behavior of the probability distribution in finite and infinite-dimensional spaces; principal quantile directions are defined to summarize their information. Moreover, exploiting the idea of statistical depth, they generalize the definition of robust principal components for functional data.

In a previous paper [2] , we estimate multivariate functional data by penalized B-spline; a working covariance matrix is also derived on the basis of coefficients of the splines, accounting for the main temporal effects; FPCA allow us to project data variations, observed in multidimensional space, into few dimensions. Due to dimensionality reduction and applying the Karhunen-Loéve decomposition, this method is also useful for the representation of the random curves in terms of the factor functions.

In the present paper our main purpose is to investigate data by means of FPCA, capturing the tail behaviour of the distributions. The FDA approach is proposed for the simultaneous estimation of the functional regression quantiles; assuming that quan-

tiles, estimated at different values of probability, share some common features, they can be summarized by a small number of functional principal components, identifying the directions along which resuming the most interesting characteristics. The method is applied to air pollution data from a monitoring network.

2 The Methodology

The α-quantile is defined as the inverse of a cumulative distribution function, given a real valued random variable X, with distribution F_x:

$Q_X(\alpha) =: Q(F_X, \alpha)$, $inf\{x \in R : F_X(x) \geq \alpha\}$.

We refer to the situation in which the interest is in the $\alpha - th$ theoretical quantile of the conditional distribution of X at time t:

$$Q_{X|t}(\alpha|t) =: Q(F_{X|t}, \alpha). \tag{1}$$

In this setting the (1) is a time varying function:

$$Q_{X|t}(\alpha|t) = l_\alpha(t), \tag{2}$$

and the estimator is the minimizer of a expected (generally asymmetric) loss function.

Kato [4] is one of the earlier paper studying functional quantile regression; starting with a linear quantile regression, in which the response is scalar while the covariate is a function, and expanding the covariate and the slope function in terms of their principal components, the model is transformed into a quantile regression model with an infinite number of regressors. More recently, in [3] the functional quantiles $l_{\alpha i}(t)$ are estimated nonparametrically; on the basis of the Karhunen-Loéve decomposition, they may be approximately represented by means of an Empirical Orthonormal Basis (EOF):

$$l_{\alpha i}(t) = \mu(t) + \sum_{h=1}^{H} \psi^h(i)\xi_h(t), \tag{3}$$

where $\mu(t)$ is a mean function and $\sum_{h=1}^{H} \psi^h(i)\xi_h^\alpha(t)$ is the reduced rank model obtained fixing the number H of bases; it is linear combination of principal components $\psi^h(i)$ and eigenfunctions $\xi_h(t)$.

The authors perform the analysis combining the representation (3) with the estimation procedure of the quantile function, after choosing a proper loss function.

We generalize this approach to the joint estimation of a collection of quantile functions, defined for a relevant set of probability values, $\alpha = [\alpha_1, \alpha_2, \ldots \alpha_q]$, implementing a three-mode FPCA analysis together with a general smoothing approach.

A functional form is presented by means of multidimensional linear smooth functions:

$$l_\alpha(t) = \sum_{k=1}^{K} \theta_k^\alpha \phi_k(t), \tag{4}$$

where: $\Phi(t) = \{\phi_k(t)\}$ is the set of K basis functions and $\theta_{\mathbf{i}}^\alpha = \left[\theta_{\mathbf{i},\mathbf{1}}^\alpha, \ldots, \theta_{\mathbf{i},\mathbf{K}}^\alpha\right]$ is the vector of the coefficients.

In order to estimate the $K \times N$ matrix Θ^α of parameters, the P-spline (penalized B-spline) approach is here considered, minimizing the penalized loss function:

$$PENRSS_\lambda(y) = \mathbf{w}(\alpha)\mathbf{I}|X - \Phi\Theta^\alpha|^T|X - \Phi\Theta^\alpha| + \lambda\Theta^\alpha\mathbf{H}\Theta^\alpha, \tag{5}$$

where the elements of vector $w(\alpha) = [w_1(\alpha), \ldots, w_n(\alpha)]$: are:

- $w_i(\alpha) = \alpha$, if $X_i > \alpha\Phi_i\Theta^\alpha$;
- $w_i(\alpha) = (1-\alpha)$, if $X_i \leq \alpha\Phi_i\Theta^\alpha$.

For details of penalty term $\lambda\Theta^\alpha\mathbf{H}\Theta^\alpha$ in (5), as well as for the estimation procedure, we refer to [2]. In this framework, three-mode functional principal quantile are derived straightforward by decomposition of the variance function, estimated by a working variance array (referred to N curves, T time units and Q quantiles) defined in terms of the estimated coefficients (see also [2]). An interesting result is the decomposition of a random function into two sets: the set of factor scores, one for each curve, on the basis of all their quantiles, and the set of corresponding factor loadings, defining the mood of variations.

3 The application

We illustrate the proposed method with an example of PM_{10} daily time series registered in one year in $N = 59$ stations of a monitoring network in California.

In Fig. 1 (a) the set of the N observed curves are represented (gray lines); a subset of seven curves are selected (coloured lines) in order to highlight the results of the procedure. In Fig. 1 $(b) - (f)$ the estimated quantile functions for different probability values, from 0.1 to 0.9, synthesize the specific pattern of the respective curves. Fig. 2 show the projections of the quantile functions in the space of the first two principal components with proportion of variance explained 0.793 and 0.073; figures $(a) - (e)$ are the partial scores for each quantile and (f) the total scores. We can observe that the functional principal components retain the most information of the original curves and curves with similar pattern have similar scores.

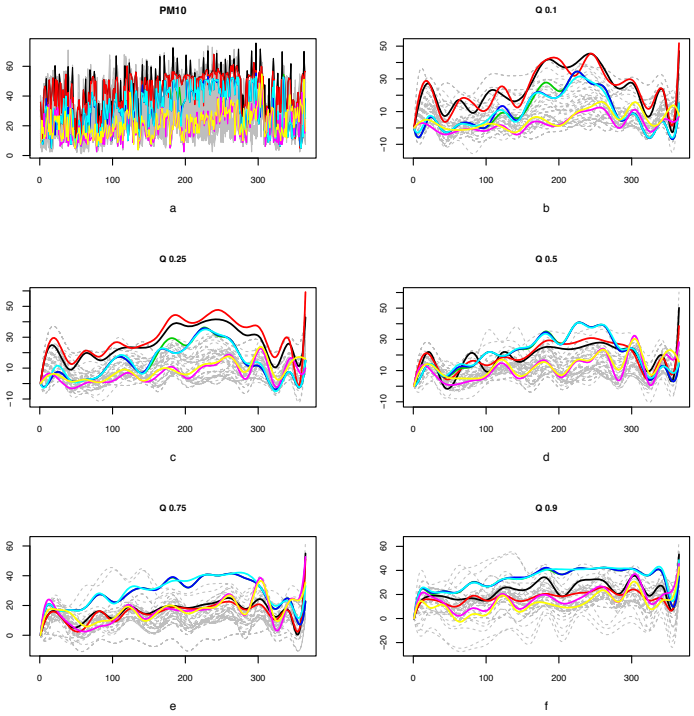

Fig. 1 Observed curves and estimated quantile functionals

4 Conclusion

The FDA approach is proposed for the simultaneous estimation of functional regression quantiles, when the main purpose is capturing the tail behaviour; assuming that quantiles estimated at different values of probability share some common features, they can be summarized by a small number of functional principal components, identifying the directions along which resuming the interesting characteristics. Some implication and appealing intuitions can be borrowed from approaches relied on depth measures, in order to construct basic tools for functional data. The approach has the advantage of further generalization, such as the inclusion of explanatory variables and distributional assumptions. Many consequent applications of the FPCA in quantile regression are motivated by the Karhunen-Loéve theorem, by means of which the random curves find convenient representations in terms of empirical orthogonal functions.

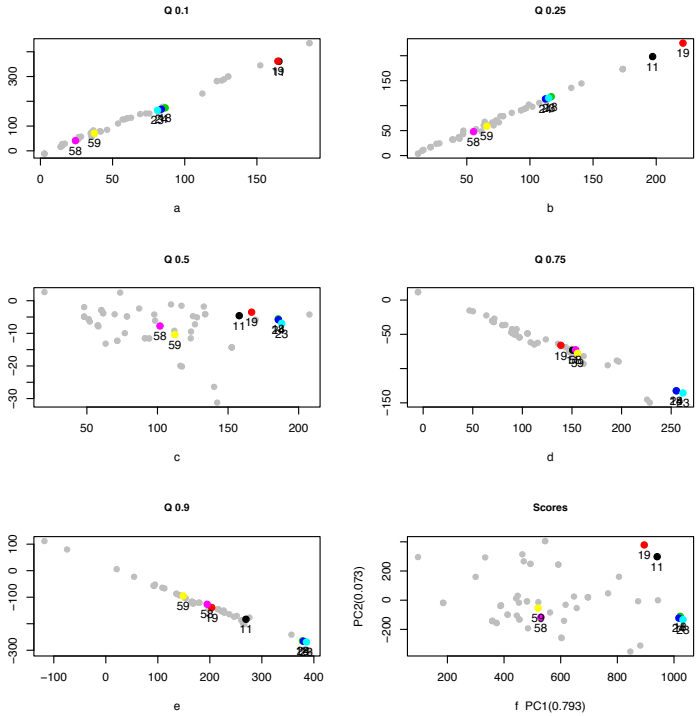

Fig. 2 Projection of the curves in the space of the first two partial (a)-(e) and total (f) principal components

References

1. Fraiman, R., Pateiro-Lopez, B. Functional quantiles, in F. Ferraty and Y. Romain, eds, Recent Advances in Functional Data Analysis and Related Topics. Contributions to Statistics, Physica-Verlag HD, pp. 123129 (2011).
2. Di Salvo, F., Ruggieri, M., Plaia, A., Functional Principal Component analysis for multivariate multidimensional environmental data. Environmental ad Ecological Statistics, 22 (4), 739-757 (2015).
3. Guo, M., Zhou, L., Huang, J.Z., Hardle, W.K., Functional data analysis of generalized regression quantiles. Statistics and Computing, 25 (2), 189-202 (2015). DOI 10.1007/s11222-013-9425-1
4. Kato, K., Estimation in functional linear quantile regression. The Annals of Statistics, 40 (6), 3108-3136 (2012). DOI: 10.1214/12-AOS1066
5. Koenker, R., Quantile Regression. Cambridge University Press, New York (2005).