Clusters of effects in quantile regression models

Gianluca Sottile[a], Giada Adelfio[a,b,*]

[a] *Dipartimento di Scienze Economiche Aziendali e Statistiche, University of Palermo, Viale delle Scienze ed. 13, 90128, Palermo, Italy*
[b] *Istituto Nazionale di Geofisica e Vulcanologia, Via Ugo La Malfa, 153 90146, Palermo, Italy*

Abstract

In this paper we propose a new method for finding similarity of effects in a multivariate regression context. Using quantile regression, the effect of each covariate on a response variable is represented as a function of percentiles. Collecting all these curves, describing the effects of each covariate on the response, we could assess if there are covariates with similar effects. Moreover, we provide a flexible algorithm which could be used not only for clustering the coefficient effects of a quantile regression framework, but also for finding clusters of generic curves. We provide also some simulated results and applications on real data, highlighting the flexibility of the proposed approach in several research fields.
Keywords: quantile regression coefficients modelling, multivariate analysis, functional data, curves clustering

1. Introduction

In this paper we focus on a new method for classifying effects in general dependence models. Indeed, a first interest of research could be the comparison among explanations of different models, that is, if the coefficients associated to a set of covariates with different responses are different. Another interest could be to check if there are covariates with similar effects with respect to the same

*Corresponding author
Email addresses: `gianluca.sottile@unipa.it` (Gianluca Sottile), `giada.adelfio@unipa.it` (Giada Adelfio)

response. Simple t-tests following the ANOVA theory are usually considered to compare coefficients effects for pooled data, that is, accounting also for some grouping variable. Extended procedures used to compare regression coefficients across models (both linear and generalized linear models) are proposed in [5].

The novelty of the proposed approach is related to a new perspective of comparison, focusing not only on single coefficient effect, but on curves effects, result of a quantile regression fitting.

Looking for curve similarity could be a complex issue characterized by subjective choices related to the continuous transformation of observed discrete data. Here, this problem is handled with the introduction of a new, simple and efficient procedure, based on a similarity measure between curves. The variability among curves can be distinguished in two components: phase variability (removed after the alignment of the curves) and amplitude variability [19].

The complex problem of curves clustering is strictly related to the idea of curves alignment, that is studied in different fields: this is referred to as "curve registration" in statistics [20, 16], "time warping" in engineering [22] and "structural averaging" in the context of computing an average curve [12]. A more general approach is based on the alignment of curves using a target function to which each one has to be registered with respect to some local features or based on the minimization of some measure like the average squared distance between each curve and the target function [20]. [16] used a Procrustes fitting procedure [9] to provide maximal alignment to the target function, subject to the suitable smoothness of the transformations. [3] introduced a simple procedure to identify clusters of multivariate waveforms based on a simultaneous assignation and alignment procedure. More general methods for curves clustering have been proposed in the literature. [11] introduced a method for finding similarities among functions by equating the moments between all curves. This problem can be crucial in several contexts. A new approach based on the trimmed K-means Robust Curve Clustering proposed by [8] is introduced in [2], considering a functional principal component rotation of data [17]. This approach has been extended in [4], where the authors focused on finding clusters of multidimensional curves with spatio-temporal structure.

All the above mentioned methods have been defined in a slight different context with respect to the one we consider here. Indeed, the proposed approach looks for similarities among curves of effects in a quantile regression. These curves have typically variable trends and different shapes, and the main purpose is to find effects that are not *significantly* different and could be associated to covariates belonging to the same cluster, according to a dimensionality reduction perspective.

In general, statistical techniques, aimed at the reduction of huge amounts of information, are relevant in statistics and synthesis (of objects and variables) approaches aim to detect the most relevant information for an appropriate interpretation of data.

Various methods, combining cluster analysis and the search for a lower-dimension representation, have been also proposed in the finite dimensional setting [21]. More recently, the use of clustering is considered as a preliminary step for exploring data represented by curves, with a further difficulty associated to the infinite space dimension of data [10].

The paper is organized as it follows: in Section 2 we report the usual notation of Quantile Regression, together with some recent developments referred to a parametric approach for coefficient functions. In Section 3 we introduce the new method for curves clustering starting from a quantile regression model, together with the algorithm details. In Section 4 simulated results are reported both for curves of effects in quantile regression and in general waveform context. Example of applications on real data are reported in Section 5. Section 6 is devoted to conclusive remarks.

2. Quantile regression and recent extensions

The non-normality of the distribution and the presence of outliers suggest the use of Quantile Regression (QR) approach [13, 14] to investigate the influence of some covariates on the response. Indeed, although the Ordinary Least Squares

(OLS) regression allows to model the average as a measure of synthesis, it does not take into account the whole shape of distribution of the outcome variable. This issue is overcame by the QR approach: it aims at estimating the fixed quantiles of the response variable, using different measures of central tendency (and statistical dispersion), in order to obtain a more comprehensive analysis of the relationship between variables. In the specific context, the QR analysis allows to interpret results also for the tails of the distribution, instead of focusing just on the "average response". QR deals with the estimation of conditional quantile functions for models in which quantiles of the conditional distribution of the response variable are expressed as functions of observed covariates, and with respect to the usual OLS, QR also provides more robust estimates. Unlike the ordinary linear regression, the QR parameter measures the change in a specified quantile of the response variable produced by one unit change in the predictor variable. This allows to compare how some percentiles of the variable of interest may be more affected by certain subject characteristics than other percentiles.

In [6], the authors suggest to adopt a parametric model for the coefficient function of a quantile regression. They refer to this estimation approach as quantile regression coefficients modelling (QRCM). The QRCM method has been also implemented in the R package `qrcm` [15, 7].

Conversely to standard quantile regression which works in a quantile-by-quantile fashion, in the QRCM framework different quantiles are estimated one at the time. This modelling approach facilitates estimation, inference, and interpretation of the results, and generally yields a gain in terms of efficiency. More in the detail, given a response variable y and a set of q-covariates $\boldsymbol{x}$, the coefficients $\boldsymbol{\beta}(p)$ are defined as functions of $p \in (0,1)$ (that is the vector of percentiles), depending on a finite-dimensional parameter $\boldsymbol{\theta}$,

$$\boldsymbol{\beta}(p \mid \boldsymbol{\theta}) = \boldsymbol{\theta}\boldsymbol{b}(p),$$

where $\boldsymbol{b}(p) = [b_1(p), \ldots, b_r(p)]^T$ is a set of r known functions of p [6]. With this approach, $\boldsymbol{\beta}(p)$ is treated as an infinite-dimensional parameter, while the estimated coefficients in a standard quantile regression are generally non-smooth functions of p and may suffer from a high volatility that hinders their interpretability.

In a multivariate framework, let $\boldsymbol{y} = [y_1, \ldots, y_j, \ldots, y_m]$ to be a set of m response variables, correlated or not, and $\boldsymbol{x}$ to be a set of q covariates. Applying the QRCM on each response variable, we estimate the coefficients functions $\beta_{1j}(p, \boldsymbol{\theta}), \ldots, \beta_{qj}(p, \boldsymbol{\theta})$ over the percentiles. In this paper, starting from the QRCM estimation of curve effects, we propose a new algorithm to identify those covariates with the same effect on a single response, or, similarly, to identify the responses that are related by similar effect of a given covariate. In a generic framework, we investigate the similarities among n general curves, parametrized by $\beta_i(p)$, $i = 1, ..., n$.

3. The proposed clustering method

The clustering approach proposed in this paper is based on a new dissimilarity measures based both on shape and distance. More in the detail, we define a new dissimilarity measure, based on two measures accounting both for the shape and for the distance.

Let $\beta_i(p)$ be the coefficient function approximated by a spline function $s_i(p)$, for $p = 1, ..., N_p$, $i = 1, ..., n$. Considering two different curves $\beta_i(p)$ and $\beta_{i'}(p)$ with $i \neq i'$, we define

$$\begin{aligned} d^{ii'}_{\text{shape}}(p) &= I(\text{sign}(s''_i(p)) \times \text{sign}(s''_{i'}(p)) = 1) \\ d^{ii'}_{\text{distance}}(p) &= I(|\beta_i(p) - \beta_{i'}(p)| \leq f(\alpha, \text{dist}(p))), \end{aligned}$$

where $s''_i(\cdot)$ is the second derivative of $\beta_i(\cdot)$ and $f(\cdot, \cdot)$ is a cut-off function, that depends on α, a probability value, and $\text{dist}(p)$, that is the vector of the distances between all the pairs of curves for each value of p. Therefore, computed the distribution of $\text{dist}(p)$ for each value of p, the cut-off function selects the corresponding $\alpha-$th percentile vector.

Therefore, the proposed dissimilarity measure between two curves is defined as:

$$d^{ii'} = 1 - \frac{1}{N_p} \sum_{1=1}^{N_p} \left[d^{ii'}_{\text{shape}}(p) \cdot d^{ii'}_{\text{distance}}(p) \right] \tag{1}$$

In the proposed approach, the new dissimilarity measure is used to define a dissimilarity matrix, useful for the application of a hierarchical clustering method. The proposed procedure has been implemented in the forthcoming R package `clustEff` that develops some very flexible functions, that allow the user to make some starting choices. For instance, the α-level has a central role for finding homogeneous clusters and its choice can depend on the aim of the analysis. Fixing an α-level too small or too big could provide inhomogeneous clusters. The median is strongly suggested in waveform clustering, while the first quantile is preferable in clustering of effects. This, of course, could influence results, but at the same while has the advantage of making the user free to fix starting conditions according to his/her analysis purpose.

3.1. Choice of the number of clusters

In any clustering algorithm, one of the key aspects is the choice of the number of clusters. In our approach, we deal with this point according to the reference framework, to provide a classification tool that is both very flexible and could be used also in different contexts.

In a quantile regression context, where the purpose could be to find clusters of curve effects, we choose the optimal number of clusters (say k^*) basing on the confidence bands of curve. In particular, starting form each partition of curves in k clusters and their estimated confidence bands, we build the average band. Then, we compute the proportion of curves that are outside the average band (say π^k_{out}, $k = 1, ..., K \leq n$). The value of k^* is identified by that partition for which π^k_{out} is minimized.

Anyway, the proposed approach, based on the dissimilarity measure defined in (1), could be also an useful tool for clustering of time-dependent signals, usually analysed in functional data analysis (FDA). The nature of these curves are different from the one of the effects in a QR. Indeed, in FDA clustering, signals are often zero mean, and with high time-dependent variance. Therefore, the criterion for the choice of the optimal k^* can not be the same. In particular, in waveform clustering framework, we look for the relative distances (say $\text{dist}_{\text{rel}}^k$, $k = 1, ..., K$) between curves belonging to the same cluster and their centroid. Then, k^* is identified by that partition for which the average distance $\text{dist}_{\text{rel}}^k$ is minimized.

3.2. Steps of the Algorithm

The main steps of the algorithm are summarized as following:

Step 1. fixed the α-level and calculated all the possible distances between the pairs of curves for each percentile (i.e. $\text{dist}(p)$), the cut-off function selects the percentile of the distribution of $\text{dist}(p)$ used in $d_{\text{distance}}^{ii'}(p)$;

Step 2. according to the measure in (1), the dissimilarity matrix is calculated;

Step 3. applying a hierarchical clustering algorithm a dendrogram is obtained;

Step 4. if the number of clusters is not fixed, the optimal number is obtained as in Section 3.1;

Step 5. after selecting the number of clusters, the mean curves are calculated within each cluster.

The `clustEff` package provides not only the main function that performs the proposed algorithm, but also a summary and different graphical tools.

On the basis of several applications and simulated results, partially here reported, we can conclude that the algorithm seems to be very stable and fast in the computation.

4. Simulation study

In this section, we report simulated results for proving the validity of the proposed approach for cluster of curves, both referring to curves of effects in a quantile regression and to general waveforms.

4.1. Clusters of effects

Let us consider a multivariate scenario in which the quantile function is simulated as

$$Q(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = \beta_0(p \mid \boldsymbol{\theta}) + \beta_1(p \mid \boldsymbol{\theta})x_1 + \cdots + \beta_q(p \mid \boldsymbol{\theta})x_q,$$

where $x_1, x_2, \ldots, x_q$ are independent $\mathbb{U}(0,5)$ variables and $p \in \mathbb{U}(0,1)$. In the first simulation scenario, the intercept is modelled as a quantile normal distribution function (ϕ) for its flexibility. Other choices, as suggested in the original paper of [6], could be also considered. We use $q = 2$ covariates and define three groups of quantile functions

$$Q_1(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = (1 + \phi(p)) + (.5 + .5p + p^2 + 2p^3)x_1 + (.5 + 2p + p^2 + .5p^3)x_2,$$
$$Q_2(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = (1 + \phi(p)) + (-3 + .5p + p^2 + .5p^3)x_1 + (-1.5 - p - .5p^2 + p^3)x_2,$$
$$Q_3(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = (1 + \phi(p)) + (.3 - .5p - p^2 + 2p^3)x_1 + (-.5 + p - .5p^2 - p^3)x_2,$$

Ten response variables are generated for each quantile function (Q_1, Q_2, Q_3). Applying the QRCM method to these response variables, we obtained the 30 coefficients curves, namely curves effect, and their lower and upper bounds, useful to select the optimal number of clusters, for both covariates.

The `clustEff` algorithm is able to select the correct number of clusters and to discriminate the 30 curves effect. In Fig. 1, the curves for both covariates are represented in the three clusters, and in Table 1 results are summarized. In Table 1 average cluster distances and silhouette widths within clusters are reported. The first measure highlights the closeness of curves with respect to the mean curve of each cluster, in particular the smaller is this value the closer are the curves. The silhouette value is a measure to assess the cohesion of each curve to its own cluster compared to other clusters [18]. In particular, observations with a large silhouette (almost 1) are very well clustered; a small silhouette (around 0) means that there would be some observation that lies between two clusters, and negative silhouette means that there are observations probably placed in the wrong cluster.

Starting form the simulation here reported, results show a valid clustering of curves, since silhouette widths are all greater than 0, and in particular, for

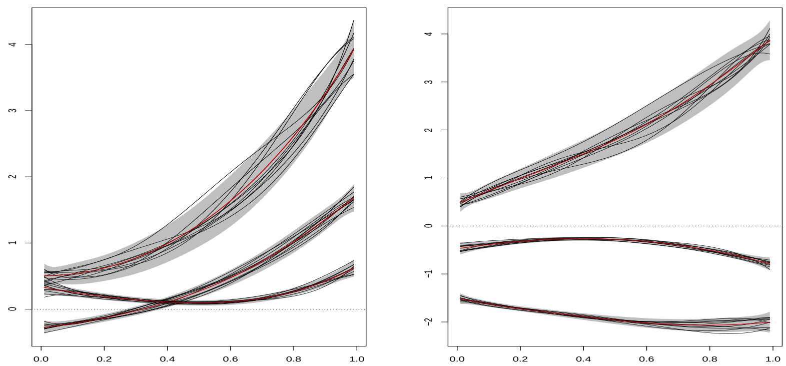

Figure 1: Left and Right panels show the 30 curves clustered in 3 clusters for the first and the second covariate, respectively, after applying the proposed algorithm. Red solid line is the mean curve and dashed red lines are the mean lower and upper bands within each cluster.

Table 1: Results of clustering in correspondence of the two covariates, summarized in terms of average cluster distance (ACD) and silhouette width (SW) within clusters.

	x_1		x_2	
	ACD	SW	ACD	SW
Cluster 1	.41	.39	.50	.27
Cluster 2	.34	.63	.33	.59
Cluster 3	.45	.77	.27	.86

clusters 2 and 3 these values are greater than 0.5. Moreover, all the average cluster distances are lower than or equal to 0.5.

4.2. Curves clustering

Fig. 2 shows 30 curves where 10 of them are obtained from the function $f(x) = \sin(3\pi x)$, 13 from $g(x) = \cos(3\pi x)$ and 5 from $h(x) = \sin(3\pi x)\cos(\pi x)$ evaluated in a grid of size 1000; a $\mathbb{N}(0, \sigma_t^2)$-distributed error is added, with σ_t^2 a variance function defined by segmented relations with multiple change-points. Two outlying curves from $l(x) = 0$ are added, such that they are not pointwise outlier at any coordinate. The proposed clustering methods is applied to the 30

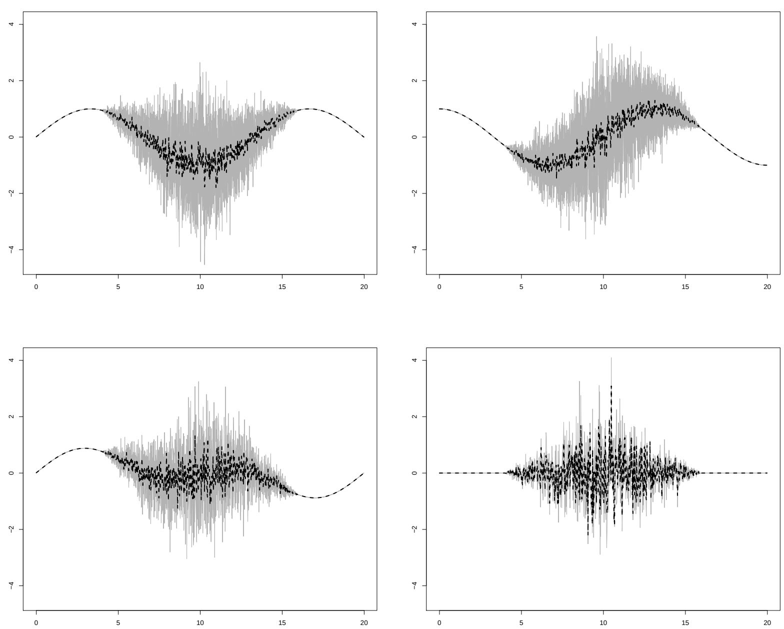

Figure 2: The 30 curves divided in the 4 groups.

curves. The applied procedure finds the three clusters $f(x)$, $g(x)$ and $h(x)$ and also identifies the two outlying curves as a fourth cluster, as reported in Fig. 3.

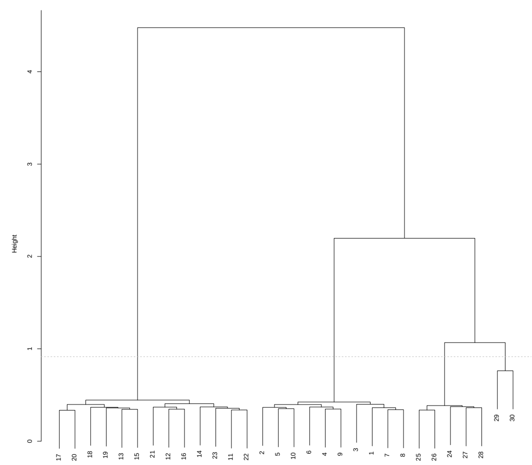

Figure 3: Dendrogram of the clustering algorithm applied in a functional data framework.

The average distances within the first two clusters is approximately 0.04 and the individual silhouette width is around 0.41. These results confirm the good performance of the proposed method in terms of homogeneity of the found clusters and proximity between curves.

5. Examples of application of the clusteEff algorithm on real data

In this section, we apply the proposed clustering algorithm to three different real data, in order to show the flexibility of the proposed method and its wide spectrum of application.

5.1. Dataset 1

The first analysed dataset consists of 2372 earthquakes located in Italy by the INGV (Istituto Nazionale di Geofisica e Vulcanologia) seismic network from

2012 to 2016, with local magnitude greater than 2.5. The selected time interval, as well as the minimum magnitude, have been chosen in order to have a catalogue as homogeneous as possible. Each seismic event is uniquely identified with a sequential numeric (ID). For each event Latitude (lat), Longitude (lon) and Hypocentral Depth (depth), uniquely define the hypocenter position in space.

The precision and accuracy of their estimates is strongly influenced by the quality of the data and the geometry of the stations that recorded the event. In this application, the following variables are further considered:

- Magnitude (mag): measure of the magnitude of the earthquake;
- Magnitude uncertainty (errM): uncertainty about the magnitude of the earthquake;
- Hypocentral uncertainty (errZ): uncertainty about the depth hypocenter;
- Epicentral uncertainty (errH): uncertainty about the depth epicentre;
- Gap azimuth (gap): a synthetic parameter of the geometry of the stations in relation to the epicentre; it expresses the maximum angle between two consecutive stations placing the epicentre to the vertex of the angle. High values of the azimuthal gap, severely affect the quality of the hypocenter location. For values higher than $180°$, i.e. external seismic event from the monitoring network, the localization errors can be very high or the event can not be allocable;
- Distance from the nearest station (mDst): is the minimum distance between the epicentre and stations. In particular for shallow earthquakes, this distance should be sufficiently small. If there is not at least one station close enough to the epicentre, the determination of depth hypocenter can be extremely difficult or even impossible. In Figure 4 minimum distances between epicentres and stations are shown;
- Root Mean Square (rms): the standard deviation between the arrival times of seismic waves estimated automatically or manually (experimental) and

theoretical ones determined on the basis of a velocity model of wave propagation. This variable is therefore a measure of the quality of the location;

- Number of stations that recorded the event (nSt): it is the number of stations used in the localization process. This number is heavily influenced by the magnitude of the event and strongly influences the accuracy of the location.

Starting from all these variables, we could identify a set of dependent variables (mag, errM, depth, lon, lat, errZ, errH) and a set of independent variables (gap, mDst, rms, nSt).

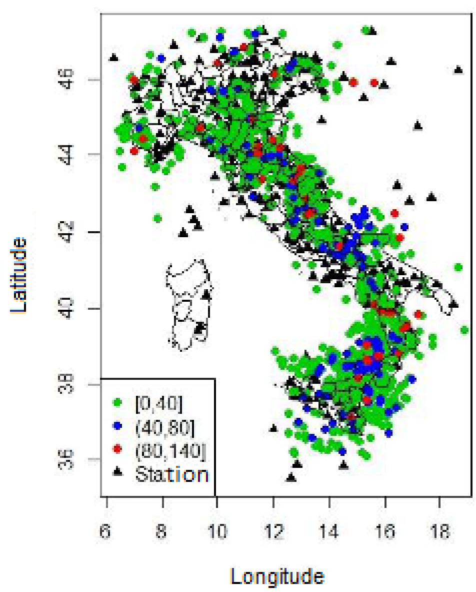

Figure 4: Min. distances (in km) between epicentres and stations (black triangles)

One of the main purposes of this analysis is to find some kind of dependence among these sets of variables and, particularly, to identify some clusters of response variables, conditioning to covariates. Indeed, we look for clusters of dependent variables after estimating different multiple quantile regressions. Clustering of effects on different responses, for a fixed covariate, could identify latent relationships among dependent variables. In table 2, we report the correlation matrix between pairs of variables. As expected, some well known positive correlations (errH-gap, errZ-mDst, errH-rms, errZ-rms, errM-rms) and negative correlations (errH-nSt, errZ-nSt, gap-nSt) are shown.

Table 2: Correlation matrix between dependent and independent variables. The independent variables are: mag=magnitude, errM=magnitude error, depth, lon=longitude, lat=latitude, errZ=hypocentral uncertainty, errH=Epicentral uncertainty. The dependent variables are: gap=Gap azimut, dst=distance of the epicentre from the nearest station, rms, nSt=number of stations that recorded the earthquake.

	mag	errM	depth	lon	lat	errZ	errH	gap	dst	rms
errM	0.03									
depth	0.13	-0.00								
lon	0.01	-0.08	0.33							
lat	-0.01	0.09	-0.38	-0.79						
errZ	0.03	-0.05	0.38	0.19	-0.36					
errH	0.04	-0.12	0.63	0.28	-0.41	0.52				
gap	-0.00	-0.10	0.18	0.20	-0.33	0.32	0.59			
dst	0.15	-0.10	0.31	0.21	-0.38	0.40	0.56	0.61		
rms	0.03	0.01	-0.01	0.03	-0.09	0.15	0.28	0.08	0.11	
nSt	0.52	0.19	0.02	-0.14	0.24	-0.13	-0.21	-0.30	-0.09	0.06

We model the intercept using $\phi(p)$, the quantile normal distribution function, while the coefficients associated to the covariates are described by a shifted Legendre polynomial up to the third degree (e.g., 1), that is, an orthogonal polynomial in $(0, 1)$ used to define a flexible model. In Fig. 5 we report the clusters of the dependent variables conditioned to the variables Gap azimuth (on the top), RMS (in the middle) and Number of Stations (on the bottom).

Conditioning on the Gap azimuth, three clusters of responses are selected:

1. Magnitude, Magnitude error, Latitude and Hypocentral uncertainty;
2. Depth and Longitude;
3. Epicentral uncertainty.

In the first cluster, the covariate Gap azimuth has a positive effect on the

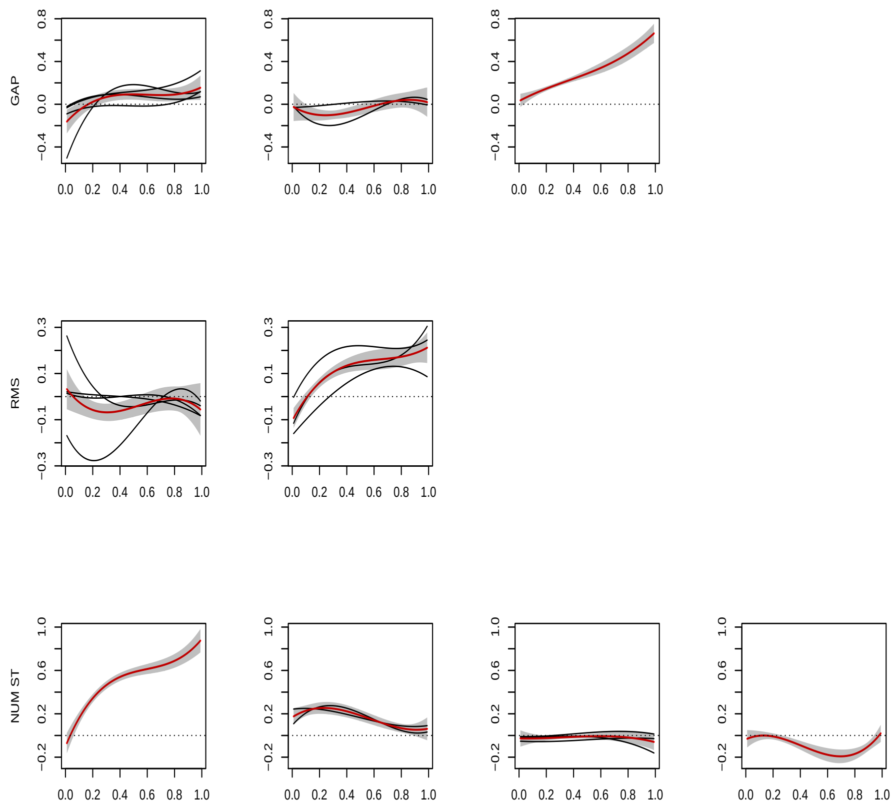

Figure 5: Clusters of responses conditioned to the three independent variables (gap, rms and nSt) for dataset 1. Red solid line is the mean curve; the shaded areas are identified by the mean lower and upper bands within each cluster.

set of responses just for the percentiles greater than .25, that is, the higher the Gap the higher both the Magnitude and the Magnitude error. This could be ascribed to different reasons, i.e. by the distribution of the occurred earthquakes with respect to the stations. Indeed, the maximum magnitude registered in the considered catalogue has been the earthquake occurred in June, 2013 with magnitude 5.9, in the North of Italy (Emilia). For this event, the gap was very high (213° vs the average almost 150°), as well its minimum distance (45 km). The Latitude is also related to the geographical distribution of events. Indeed, the big sequence of earthquakes occurred after the Emilia event has a big influence on the estimates. For these events the gaps are high, because in the North of Italy, and in particular, in the Emilia region, historically considered as a low seismic area, the network station is less dense. On the other hand the estimated effect of Gap on the magnitude error and Hypocentral uncertainty, confirms our previous knowledge.

In the second cluster, we could observe a negative effect on the responses for percentiles lower than .5, and positive otherwise. This reflects the distribution of station again, and the occurrence features of events. For instance, the area of the Ionian slab (greater Longitude), where events are deep, the network is denser (and then the Gap was lower).

In the last cluster, an increasing positive effect of Gap azimuth on Epicentral uncertainty has been estimated, confirming again our previous knowledge.

Conditioning on the rms, the responses are clustered in two sets:

1. Magnitude, Magnitude error, Depth and Longitude;
2. Latitude, Hypocentral and Epicentral uncertainty.

In the first cluster, the effect of the covariate rms on the four responses is constant and negative for percentiles of the distribution between .15 and .51.

In the second cluster, rms has a positive effect on Latitude, Hypocentral and Epicentral uncertainty, conditioned to the percentiles greater than .13.

Finally, conditioning on the covariate nSt, we find four clusters of responses as follows:

1. Magnitude;
2. Magnitude error and Longitude;
3. Depth, Hypocentral and Epicentral uncertainty;
4. Latitude.

In the first cluster, the number of stations has an increasing positive effect on Magnitude for almost all the distribution.

In the second cluster, the number of stations has a decreasing positive effect on Magnitude error and Longitude.

In the third cluster, the covariate number of stations has a low negative effect on Depth, Hypocentral and Epicentral uncertainty for percentiles between .2 and .36.

In the last cluster, the number of stations has almost always a negative effect on Latitude, reflecting what we observed with respect to the covariate Gap.

In this application, we show an interesting usage of the proposed method, identifying clusters of dependent variables on the basis of the estimated curves effect for a better characterization of these dependencies.

5.2. Dataset 2

The data refer to a study carried out in 1988-1991 in the North of Italy, including 1053 males and 992 females. The study aims at assessing determinants of the Inspiration Capacity (IC), a measure of lung's function, among the following nine predictors: age, height, body mass index (bmi), sex, and indicators for current smoking, occupational exposure, cough, wheezing, and asthma.

We model the intercept using $\log(p)$ and $\log(1-p)$, that defines the asymmetric Logistic distribution used for its flexibility, while the coefficients associated to the covariates are described by a fifth degree shifted Legendre polynomial. The estimated model is summarized in Table 3, and curves of effects are represented in Figure 6.

We apply the `clustEff` algorithm to the curves of significant coefficients associated to the predictors in order to look for similar effects of covariates with

Table 3: Summary of the model selected by AIC, for the Inspiratory Capacity data with respect to the nine listed covariates

	1	$\log(p)$	$\log(1-p)$	$slp(p,5)[1]$	$slp(p,5)[2]$	$slp(p,5)[3]$	$slp(p,5)[4]$	$slp(p,5)[5]$	P-value
(Intercept)	-0.050	0.353	-0.403	-	-	-	-	-	.000*
age	-0.285	-	-	0.049	0.002	0.012	-0.013	-0.000	.000*
hgt	0.411	-	-	0.085	0.086	0.031	0.024	-0.006	.000*
bmi	0.146	-	-	0.069	-0.052	0.016	-0.019	0.015	.000*
sex	-0.016	-	-	-0.101	0.077	-0.022	0.055	-0.010	.000*
smoker	-0.060	-	-	0.041	-0.012	0.000	-0.023	-0.002	.208
occ_exp	0.085	-	-	0.008	0.020	-0.018	0.021	-0.016	.594
cough	0.018	-	-	-0.005	0.007	-0.028	-0.009	-0.009	.306
wheeze	-0.013	-	-	0.054	0.017	-0.007	0.037	0.010	.048*
asthma	0.007	-	-	0.004	-0.018	0.020	0.010	-0.035	.380
P-value	.000*	.000*	.000*			.000*			

QRCM estimates of $\boldsymbol{\theta}$. The corresponding graphical representation is proposed in Figure 6. The last row reports the p-values for the null hypothesis that the coefficient of $\boldsymbol{\theta}$ is $\mathbf{0}$ and represents the significance of the components of $\boldsymbol{b}(p)$. The last column is defined analogously and can be interpreted as the significance of a test for a null effect of covariates. The asterisk (*) denotes significance less than 0.05.

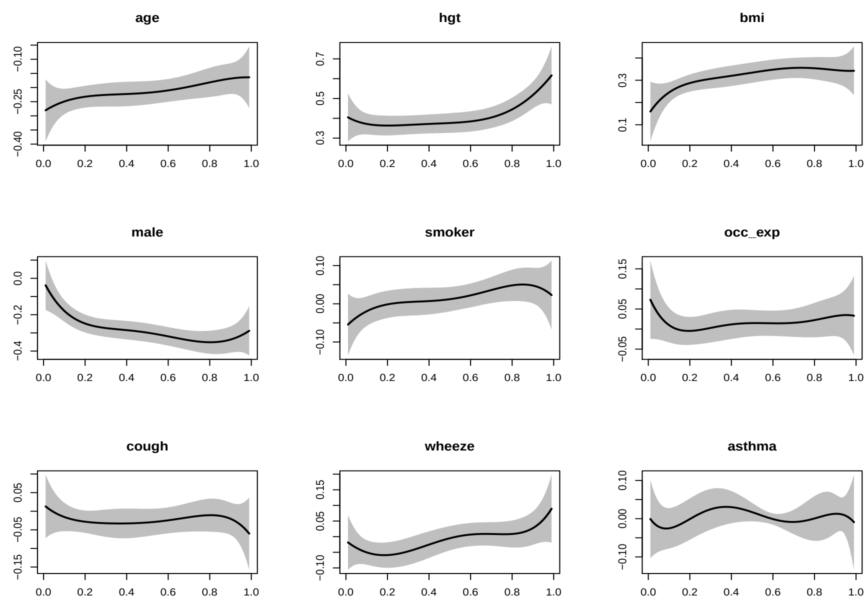

Figure 6: QRCM estimates of $\boldsymbol{b}(p)$ (see Table 3). Confidence bands are displayed as shades.

respect to the Inspiration Capacity response. The application of the proposed algorithm provides three clusters.

The first cluster consist of age and sex with an average negative effect, significant for all the percentiles.

The second cluster is identified by bmi and height, with a positive effect, significant for all the percentiles.

Finally, in the third cluster we find the variables smoker, occupational exposure, cough, wheezing, and asthma with not significant average effect.

These results are summarized in Figure 7.

In this application, we focus on a new perspective of reduction of dimensionality, applied in a quantile regression context. Indeed, we propose the use of the `clustEff` method for finding the main determinants of a quantitative response, assuming that we are interested in looking for dependence structures. Of course, these results could be more relevant in presence of several regressors,

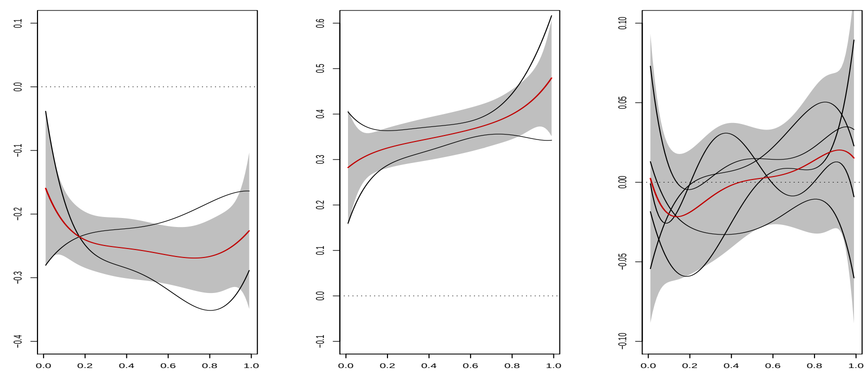

Figure 7: The three clusters obtained applying the clustEff algorithm on the estimated quantile regression coefficients of dataset 2. Red line is the mean curve; the shaded areas are identified by the mean lower and upper bands within each cluster. Black lines are the covariates; in the first cluster there are the variables age and sex, in the second cluster bmi and height and in the third cluster smoker, occupational exposure, cough, wheezing and asthma.

but we showed this example just for its simplicity of interpretation. Indeed, we could observe that covariates are classified in three main groups: the first relative to the subject characteristics, the second relative to body features and the third that associates the clinical aspects. Therefore, in describing the effect of covariates to the response, we interpret the average effect of each cluster, as a proxy of a latent characteristic effect that is associated to the covariates of that cluster. As drawback, this procedure could have some limitations in terms of loss of interpretation, as usual in dimensionality reduction problems.

5.3. Dataset 3

Also this application is reported to show the flexibility of the proposed algorithm. Indeed, we used the clustEff method for waveform clustering, that may be considered as an issue of clustering of functional data.

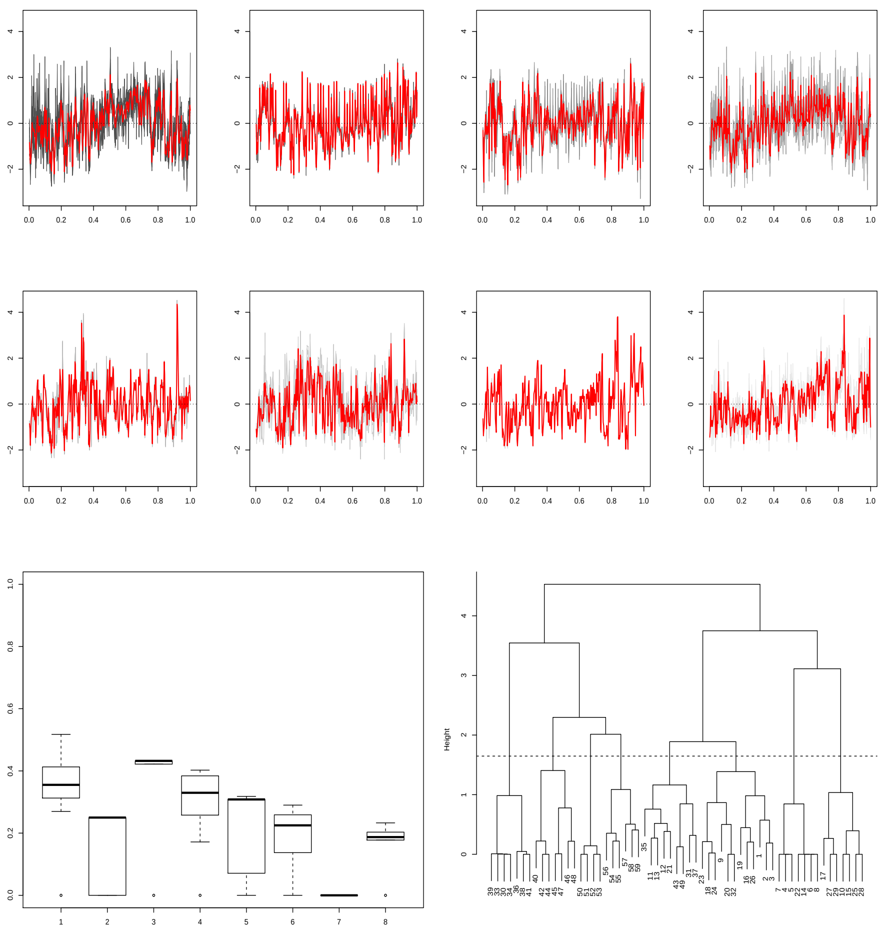

Figure 8: The identified 8 clusters of dataset 3 (on the top): red lines are the mean curves. Boxplot of the average mean distance within each cluster (on the left-bottom). Dendrogram of the clustering algorithm and height level used to cut the tree (on the right-bottom).

6. Conclusion

The proposed approach is not just a method for clustering of curves, that is an important problem in many areas of science, but it can be seen as a new tool for reduction of dimensionality in dependence model, in particular in a quantile regression context. Indeed, the proposed approach, based on a new dissimilarity measure, that accounts both for shape of curves and distance among them, allows to find similarities among curves that represent the effect of covariates on (also multivariate) response. The clustering of these curves, extends the idea of looking for similar effects and, therefore, of covariates in general dependence models, aimed to a selection perspective.

This approach, developed also in a forthcoming R Package, is a very flexible method, that is also very fast in te of computation and user-friendly for general applications. We applied the proposed algorithm to three different real data, included an application for generic waveforms in order to provide a wider spectrum of applications for curves clustering.

Acknowledgements

This paper has been partially supported by the national grant of the Italian Ministry of Education University and Research (MIUR) for the PRIN-2015 program (Progetti di ricerca di Rilevante Interesse Nazionale), "Prot. 20157PRZC4 - Research Project Title Complex space-time modelling and functional analysis for probabilistic forecast of seismic events. PI: Giada Adelfio"

References

[1] Abramowitz, M. and Stegun, I.A. (1964). Handbook of mathematical functions with formulas, graphs, and mathematical tables. Courier Corporation, vol. 55.

[2] Adelfio, G., Chiodi, M., D'Alessandro, A. and Luzio, D. (2011). FPCA algorithm for waveform clustering. Journal of Communication and Computer, vol. 8(6):494-502. ISSN 1548-7709

[3] Adelfio, G., Chiodi, M., D'Alessandro, A. and Luzio, D., D'Anna, G., Mangano, G. (2012) Simultaneous seismic wave clustering and registration. Computers & Geosciences, 8(44), 6069.

[4] Adelfio, G, Di Salvo, F, Chiodi, M. (2016). Space-time FPCA Algorithm for clustering of multidimensional curves. Proceeding of the 48th Scientific Meeting of the Italian Statistical Society, Salerno, June 8th June 10th, 2016. Editors: Monica Pratesi and Cira Pena. ISBN: 9788861970618

[5] Clogg, C, Petkova, E, Haritou, A. (1995). Statistical Methods for Comparing Regression Coefficients Between Models. The American Journal of Sociology, 100, 5: 1261-1293

[6] Frumento, P. and Bottai, M., (2015). Parametric modeling of quantile regression coefficient functions, Biometrics, 72:74-84.

[7] Frumento, P. (2017). qrcm: Quantile Regression Coefficients Modeling. R package version 2.0, https://CRAN.R-project.org/package=qrcm.

[8] Garca-Escudero, L. A. and Gordaliza, A. (2005). A proposal for robust curve clustering, Journal of classification, 22, 185-201.

[9] Gower, J., 1975. Generalized Procrustes analysis. Psychometrika 40(1):33–51.

[10] Jacques J. and Preda C. (2014). Functional data clustering: a survey. Advances in Data Analysis and Classification, Springer Verlag, 8:231-255

[11] James, G.M., (2007). Curve alignment by moments. The Annals of Applied Statistics 1:480–501.

[12] Kneip, A. and Gasser, T., (1992). Statistical tools to analyze data representing a sample of curves. Annals of Statistics 20:1266–1305.

[13] Koenker, R. and Bassett, G., Jr., (1978). Regression quantiles, Econometrica, 46:33-50.

[14] Koenker, R. (2005). Quantile Regression, Cambridge University Pres.

[15] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[16] Ramsay, J.O. and Li, X., (1998). Curve registration. Journal of the Royal Statistical Society, Sec. B, 60:351-363.

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

[18] Rousseeuw, P. J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics. 20: 5365.

[19] Sangalli, L.M., Secchi, P., Vantini, S. and Veneziani, A., (2009). A Case Study in Explorative Functional Data Analysis: Geometrical Features of the Internal Carotid Artery. Journal of the American Statistical Association 104(485):37–48.

[20] Silverman, B.W., (1995). Incorporating parametric effects into functional principal components analysis. Journal of the Royal Statistical Society, Sec. B 57:673-689.

[21] Vichi, M., Saporta, G. (2009). Clustering and Disjoint Principal Component Analysis. Computational Statistics and Data Analysis, 53:3194-3208.

[22] Wang, K. and Grasser, T., (1997). Alignment of curves by dynamic time warping. The Annals of Statistics 25:1251-1276.