A new approach for clustering of effects in quantile regression

UNIVERSITÀ DEGLI STUDI DI PALERMO

Gianluca Sottile & Giada Adelfio

dSEAS dipartimento scienze economiche aziendali e statistiche

Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università di Palermo, ITALY

gianluca.sottile@unipa.it

Background and Aims

In quantile regression coefficients modeling (QRCM), Frumento and Bottai (2015) suggest to adopt a parametric approach to model the quantile function (Q). Assume that for any $p \in (0,1)$ there exists a q-dimensional vector $\boldsymbol{\beta}(p)$ such that $Q(p \mid \boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}(p)$, where $\boldsymbol{x}$ is the model matrix of dimension $n \times q$ and $\boldsymbol{\beta}(p)$ is a function of p that depends linearly on a finite dimensional parameter $\boldsymbol{\theta}$, that is $\boldsymbol{\beta}(p \mid \boldsymbol{\theta}) = \boldsymbol{\theta}\boldsymbol{b}(p)$. Moreover, $\boldsymbol{b}(p) = [b_1(p), \ldots, b_k(p)]^T$ is a set of k known functions of p and $\boldsymbol{\theta}$ is a $q \times k$ matrix with entries θ_{jh} associated to the j-th covariate and the h-th function, $j = 1, \ldots, q$ and $h = 1, \ldots, k$.

- In a univariate framework, let us consider a response variable y of length n and a model matrix $\boldsymbol{x}$; applying the QRCM on y, we estimate the regression coefficients functions $\beta_1(p \mid \boldsymbol{\theta}), \ldots, \beta_q(p \mid \boldsymbol{\theta})$, namely **effect curves**

  $\Rightarrow$ the **aim** is to assess if these q curves, that describe the effects of each covariate on the response, can be clustered based on similarities of effects.

- In a multivariate framework, let us consider a set of m response variables $\boldsymbol{y} = [y_1, \ldots, y_t, \ldots, y_m]$, each of length n, and the model matrix $\boldsymbol{x}$; applying the QRCM on each y_t, we estimate the $m \times q$ effect curves $\beta_{11}(p \mid \boldsymbol{\theta}), \ldots, \beta_{mq}(p \mid \boldsymbol{\theta})$

  $\Rightarrow$ the **aim** is to assess if any response is related by similar effect to a given covariate.

Methods

The proposed clustering approach is based on a new dissimilarity measure, that accounts both for the **shape** and for the **distance**. Let us consider N percentiles and two different curves i and i'.

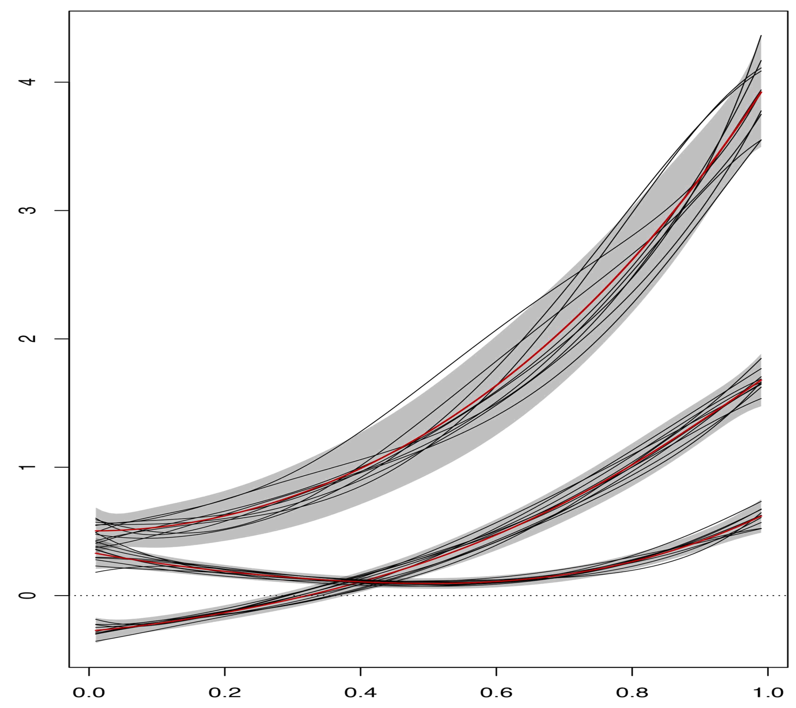
1. The shape of a curve is evaluated using its curvature point by point. Let be $s_i(p)$ a spline approximation of $\beta_i(p \mid \boldsymbol{\theta})$, then
$$d^{ii'}_{\text{shape}}(p) = I(\text{sign}(s''_i(p)) \times \text{sign}(s''_{i'}(p)) = 1)$$
where $s''_i(p)$ is the approximation of the second derivative of $\beta_i(p \mid \boldsymbol{\theta})$.
2. The distance between two curves is evaluated as the distance point by point of them, then
$$d^{ii'}_{\text{distance}}(p) = I(|\beta_i(p \mid \boldsymbol{\theta}) - \beta_{i'}(p \mid \boldsymbol{\theta})| \le f(\alpha, \text{dist}(p)))$$
$f(\cdot,\cdot)$ is a cut-off function, that depends on a probability value α, and on dist(p), that is the vector of the distances between all the pairs of curves for each percentile. The cut-off function selects the α-th percentile vector of dist(p).

Therefore, the **proposed measure** is defined as:
$$d^{ii'} = 1 - \frac{1}{N}\sum_{l=1}^{N}\left[d^{ii'}_{\text{shape}}(l) \cdot d^{ii'}_{\text{distance}}(l)\right].$$
We used the product of the two measures to take into account the concordance of both, in each point, to keep its general applicability. We implemented this approach in the forthcoming `clustEff` package in R.

Simulations

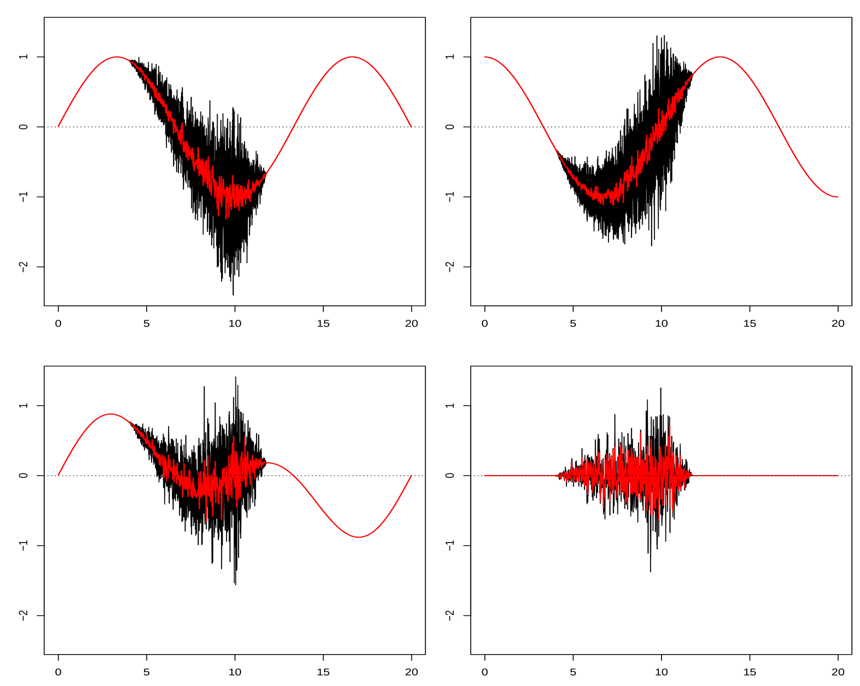
In order to briefly show the performance of our proposal, we report simulated results in clustering of curves, both referring to curves of effects in a QRCM and to general waveform framework.

- **Curve effects framework.** Let us consider a multivariate scenario in which the quantile functions are simulated as:
$$Q_1(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = (1 + \phi(p)) + (.5 + .5p + p^2 + 2p^3)x$$
$$Q_2(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = (1 + \phi(p)) + (-3 + .5p + p^2 + .5p^3)x$$
$$Q_3(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = (1 + \phi(p)) + (.3 - .5p - p^2 + 2p^3)x$$
where the intercept is modelled as a quantile normal distribution function (ϕ) and the covariate $x \in \mathbb{U}(0,5)$ is modelled as a third degree polynomial. Ten response variables are generated for each quantile function (Q_1, Q_2, Q_3), and applying the QRCM approach, we obtained $m = 30$ effect curves and the lower and upper bounds, useful to select the optimal number of clusters.



The 30 curves clustered in 3 groups. Red solid line is the mean curve and shaded area highlights the mean lower and upper bands within each cluster.

- **Waveform framework.** 30 curves are generated, 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 $h(x) = \sin(3\pi x)\cos(\pi x)$ and 2 outlying curves from $l(x) = 0$ 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.



The 30 curves clustered in 4 groups. Red solid line is the mean curve.

Conclusions

- The proposed approach can be seen as a new dimensionality reduction tool for dependence model, in a quantile regression.
- The **new** dissimilarity measure, accounting both for shape and distance among curves, allows to highlight i) similarities among curves that represent the effect of covariates on response(s) and ii) similarities in a general waveform framework.
- We proposed a new R package (*coming soon!*), that results flexible, computationally fast and user-friendly.

Acknowledgements

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

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