UNIVERSITÀ DEGLI STUDI DI PALERMO

dSEAS dipartimento scienze economiche aziendali e statistiche

CRoNoS

Workshop on Multivariate Data Analysis and Software

A new method for curves clustering in general dependence models

Gianluca SOTTILE and Giada ADELFIO

Contents

Contents

Literature review

The problem of curves clustering is very complex and has been recently addressed in several fields:

- *structural averaging* in the context of computing an average (Kneip and Gasser, 1992);
- *curves registrastion* in statistics (Silverman, 1995; Ramsay and Li, 1998);
- *time warping* in engineering (Wang and Gasser, 1997)

Literature review

In statistics:

- Silverman, (1995) proposed a general approach, in which a target curve must satisfy a predefined criterion;
- Ramsay and Li, (1998) used a Procrustes fitting procedure (Gower, 1975) to provide maximal alignment to the target function;
- James, (2007) introduced a method for finding similarities between functions by equating the moments among all curves;
- Garcia-Escudero and Gordaliza, (2005) proposed a new approach based on the trimmed *k*-means Robust Curve Clustering;
- Adelfio et al., (2012) introduced a procedure to identify clusters of multivariate waveforms;
- Adelfio et al., (2016) focused on finding clusters of multidimensional curves with spatio-temporal structure.

WHAT ABOUT CURVES CLUSTERING IN GENERAL DEPENDENCE MODELS?

Contents

Contents

QR Koenker and Bassett Jr, (1978) and Koenker, (2005)

Let be y a response variable, and $\boldsymbol{x}$ a q-dimensional vector of covariates. We assume that $Q(p \mid \boldsymbol{x}) = \boldsymbol{x}^T\beta(p)$ is the p-th quantile of y, given $\boldsymbol{x}$. The vector of quantile regression coefficients, $\beta(p)$, can be estimated by

$$\hat{\beta}(p) = \arg\min_{\boldsymbol{\beta}\in\mathcal{R}^q} \sum_{i=1}^{n} \omega_{p,i}(y_i - \boldsymbol{x}_i^T\beta)$$

where $\omega_{p,i} = \mathcal{I}(y_i \leq \boldsymbol{x}_i^T\beta)$ and $\mathcal{I}(\cdot)$ is the indicator function.

QR Koenker and Bassett Jr, (1978) and Koenker, (2005)

Let be y a response variable, and $\boldsymbol{x}$ a q-dimensional vector of covariates. We assume that $Q(p \mid \boldsymbol{x}) = \boldsymbol{x}^T \beta(p)$ is the p-th quantile of y, given $\boldsymbol{x}$. The vector of quantile regression coefficients, $\beta(p)$, can be estimated by

$$\hat{\beta}(p) = \arg \min_{\beta \in \mathcal{R}^q} \sum_{i=1}^{n} \omega_{p,i}(y_i - \boldsymbol{x}_i^T \beta)$$

where $\omega_{p,i} = \mathcal{I}(y_i \leq \boldsymbol{x}_i^T \beta)$ and $\mathcal{I}(\cdot)$ is the indicator function.

Issues

- quantiles are estimated one at the time
- the estimated coefficients are generally non-smooth functions of p

QR Koenker and Bassett Jr, (1978) and Koenker, (2005)

Le

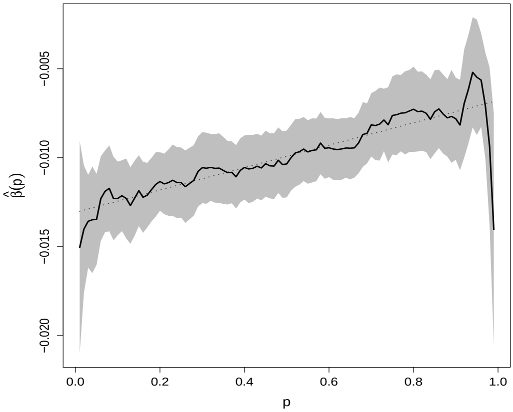
co

y,

est

wh

Iss



Estimated quantile regression coefficient and 95% pointwise confidence intervals (shaded area). The dotted line suggests a possible linear trend.

Contents

QRCM Frumento and Bottai, (2016)

A parametric approach to model the quantile function estimating the coefficients as functions of the order of the quantile $p \in (0, 1)$

$$Q(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}^T \beta(p \mid \boldsymbol{\theta}),$$

QRCM Frumento and Bottai, (2016)

A parametric approach to model the quantile function estimating the coefficients as functions of the order of the quantile $p \in (0, 1)$

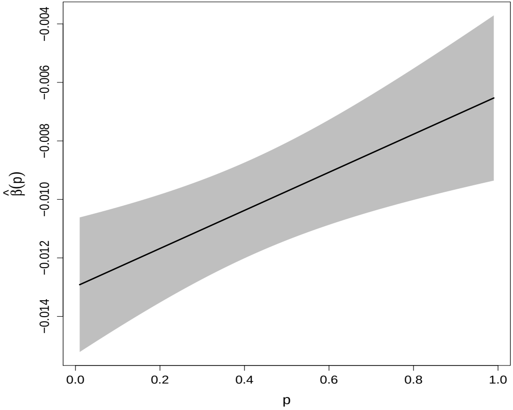
$$Q(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}^T \beta(p \mid \boldsymbol{\theta}),$$

- $\boldsymbol{x}$ is the model matrix ($N \times q$)
- $\beta(p \mid \boldsymbol{\theta}) = \boldsymbol{\theta}\boldsymbol{b}(p)$
- $\boldsymbol{b}(p) = [1, b_1(p), \ldots, b_k(p)]^T$ is a set of $(k+1)$ known functions
- $\boldsymbol{\theta}$ is the unknown parameter matrix

QRCM Frumento and Bottai, (2016)

A p
co

For
$\beta(p$



Estimated quantile regression coefficient and 95% pointwise confidence intervals (shaded area).

Contents

CEC method Sottile and Adelfio

Aims

Our goal is to use the QRCM framework to answer two different questions:

- **Univariate case.** Given one response variable we estimate $\beta_1(p \mid \boldsymbol{\theta}), \ldots, \beta_q(p \mid \boldsymbol{\theta}) \Rightarrow$ the **aim** is to assess if these q curves, can be clustered based on similarities of effects

CEC method Sottile and Adelfio

Aims

Our goal is to use the QRCM framework to answer two different questions:

- **Univariate case.** Given one response variable we estimate $\beta_1(p \mid \boldsymbol{\theta}), \ldots, \beta_q(p \mid \boldsymbol{\theta}) \Rightarrow$ the **aim** is to assess if these q curves, can be clustered based on similarities of effects
- **Multivariate case.** Given m response variables we estimate $\beta_{11}(p \mid \boldsymbol{\theta}), \ldots, \beta_{mq}(p \mid \boldsymbol{\theta}) \Rightarrow$ the **aim** is to assess if there are similar responses given covariates

Our proposal

A new dissimilarity measure, that accounts both for the **shape** and for the **distance**:

Our proposal

A new dissimilarity measure, that accounts both for the **shape** and for the **distance**:

- the *shape* evaluated using its second derivative. Moreover, two different curves are similar in shape if, at any given point, the signs of the second derivatives are concordant;

$$d^{ii'}_{\text{shape}}(\boldsymbol{p}) = \mathcal{I}(\text{sign}(\beta''_i(\boldsymbol{p} \mid \boldsymbol{\theta})) \times \text{sign}(\beta''_{i'}(\boldsymbol{p} \mid \boldsymbol{\theta})) = \mathbf{1})$$

Our proposal

A new dissimilarity measure, that accounts both for the **shape** and for the **distance**:

- the *shape* evaluated using its second derivative. Moreover, two different curves are similar in shape if, at any given point, the signs of the second derivatives are concordant;

$$d^{ii'}_{\text{shape}}(\boldsymbol{p}) = \mathcal{I}(\text{sign}(\beta''_i(\boldsymbol{p} \mid \boldsymbol{\theta})) \times \text{sign}(\beta''_{i'}(\boldsymbol{p} \mid \boldsymbol{\theta})) = \mathbf{1})$$

- the *distance* between two curves evaluated as their differences with respect to other curves. Two curves are said close if their distance at any given point is lower than a fixed value;

$$d^{ii'}_{\text{distance}}(\boldsymbol{p}) = \mathcal{I}(|\beta_i(\boldsymbol{p} \mid \boldsymbol{\theta}) - \beta_{i'}(\boldsymbol{p} \mid \boldsymbol{\theta})| \leq f(\alpha, \text{dist}(\boldsymbol{p})))$$

The cut-off function

The $f(\cdot,\cdot)$ function depends on a probability value α, and on dist(p) that is, for each percentile, the distribution of all possible distances among curves. Therefore, the cut-off function selects the α-th percentile of dist(p)

The cut-off function

The $f(\cdot,\cdot)$ function depends on a probability value α, and on dist(p) that is, for each percentile, the distribution of all possible distances among curves. Therefore, the cut-off function selects the α-th percentile of dist(p)

The role of α

α is the probability value, and it has a central role for finding hoogeneous clusters. Its choice depends on the goal of the analysis and has to be fixed by the researcher.

The new dissimilarity measure

$$d(i,i') = 1 - \int_0^1 \left[d^{ii'}_{\text{shape}}(p) \cdot d^{ii'}_{\text{distance}}(p) \right] dp$$

The product of the two measures is computed, to account for their concordance at each point.

The new dissimilarity measure

$$d(i,i') = 1 - \int_0^1 \left[d^{ii'}_{\text{shape}}(p) \cdot d^{ii'}_{\text{distance}}(p) \right] dp$$

The product of the two measures is computed, to account for their concordance at each point.

Optimization

We implemented the above measure in the `clustEff` R package.

Contents

Computation

Pseudo code of the algorithm implemented in the `clustEff` package

Step	Algorithm
1	fix the α-level and calculate dist($\boldsymbol{p}$) for each $p \in (0, 1)$
2	the cut-off function $f(\cdot, \cdot)$ selects the percentiles of the distribution of dist($\boldsymbol{p}$) used in $d^{ii'}_{\text{distance}}(\boldsymbol{p})$
3	compute $d^{ii'}_{\text{shape}}(\boldsymbol{p})$, $d^{ii'}_{\text{distance}}(\boldsymbol{p})$, and hence $d(i, i')$
4	apply a hierarchical clustering algorithm to the dissimilarity matrix in order to obtain the dendrogram
5	select the optimal number of clusters l^*, unless it is known in advance
6	calculate goodness-of-fit measures

Choice of the number of clusters

Effects curves

$$\pi^l = l \sum_{j=1}^{l} q_j^{-1} \sum_{i=1}^{q_j} \left\{ \int_0^1 \mathcal{I}\left(\overline{\text{LB}}_j(p) \leq \beta_j^i(p \mid \boldsymbol{\theta}) \leq \overline{\text{UB}}_j(p) \right) dp \right\},$$

The value l^* is identified by that partition for which $\pi^l - \pi^{l+1}$ is minimized

Choice of the number of clusters

Effects curves

$$\pi^l = l \sum_{j=1}^{l} q_j^{-1} \sum_{i=1}^{q_j} \left\{ \int_0^1 \mathcal{I}\left(\overline{\text{LB}}_j(p) \leq \beta_j^i(p \mid \boldsymbol{\theta}) \leq \overline{\text{UB}}_j(p) \right) dp \right\},$$

The value l^* is identified by that partition for which $\pi^l - \pi^{l+1}$ is minimized

General curves

$$\text{dist}_{\text{rel}}^l = l \sup_{j \in \{1,\dots,l\}} \left\{ q_j^{-1} \sum_{i=1}^{q_j} \int_0^1 |\overline{\beta}_j(p) - \beta_j^i(p \mid \boldsymbol{\theta})| \, dp \right\}.$$

The value l^* is identified by that partition for which $\text{dist}_{\text{rel}}^l - \text{dist}_{\text{rel}}^{l+1}$ is minimized

Contents

Simulation scenario 1

Clusters of effects

We considered a multivariate scenario in which the general quantile function was defined by

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

where $x \sim \mathcal{U}(0, 5)$.

Simulation scenario 1

Clusters of effects

We considered a multivariate scenario in which the general quantile function was defined by

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

where $x \sim \mathcal{U}(0, 5)$.

We defined three quantile functions and generated 30 response variables, 10 for each of them, using polynomial trends, i.e.,

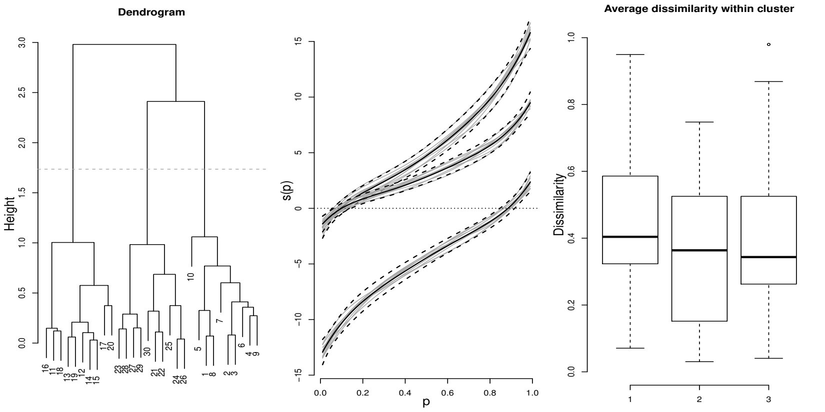
1. $Q_1(p \mid x, \boldsymbol{\theta}_1) = (1 + \phi(p)) + (.5 + .5p + p^2 + 2p^3)x$
2. $Q_2(p \mid x, \boldsymbol{\theta}_2) = (1 + \phi(p)) + (-3 + .5p + p^2 + .5p^3)x$
3. $Q_3(p \mid x, \boldsymbol{\theta}_3) = (1 + \phi(p)) + (.3 - .5p - p^2 + 2p^3)x$

Simulation scenario 1

Clu
We
fur

wh

We
var
1
2
3



Output of the proposed algorithm for one replicate. The left panel shows the dendrogram; the middle panel shows the 30 curves clustered in 3 groups; the right panel shows the boxplot of the average dissimilarity within each cluster.

Simulation scenario 2

Waveform clustering

We simulated 30 harmonic functions evaluated at a grid of size 1000

1. $f(t) = \sin(3\pi t)$ ($\times 10$)
2. $g(t) = \cos(3\pi t)$ ($\times 13$)
3. $h(t) = \sin(3\pi t)\cos(\pi t)$ ($\times 5$)
4. $l(t) = 0$ ($\times 2$)

Simulation scenario 2

Waveform clustering

We simulated 30 harmonic functions evaluated at a grid of size 1000

1. $f(t) = \sin(3\pi t)$ ($\times 10$)
2. $g(t) = \cos(3\pi t)$ ($\times 13$)
3. $h(t) = \sin(3\pi t)\cos(\pi t)$ ($\times 5$)
4. $l(t) = 0$ ($\times 2$)

A random error $\epsilon_t \sim \mathcal{N}(0, \sigma_t)$ was added to each curve to define a segmented relation with multiple change-points, such as

$$\sigma_t = 4\max(t - 0.2, 0) - 8\max(t - 0.5, 0) + 4\max(t - 0.8, 0)$$

Simulation scenario 2

Wa

We

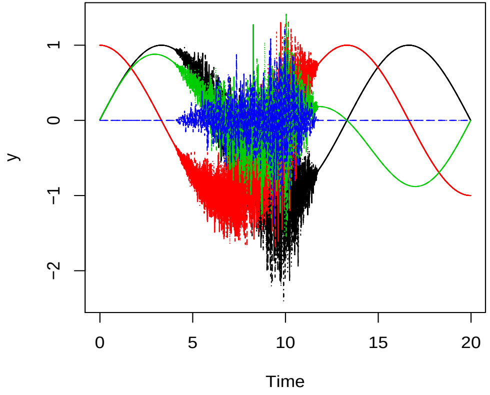
1

2

3

4

A r
seg



One simulated dataset

Simulation scenario 2

Wa

We

1

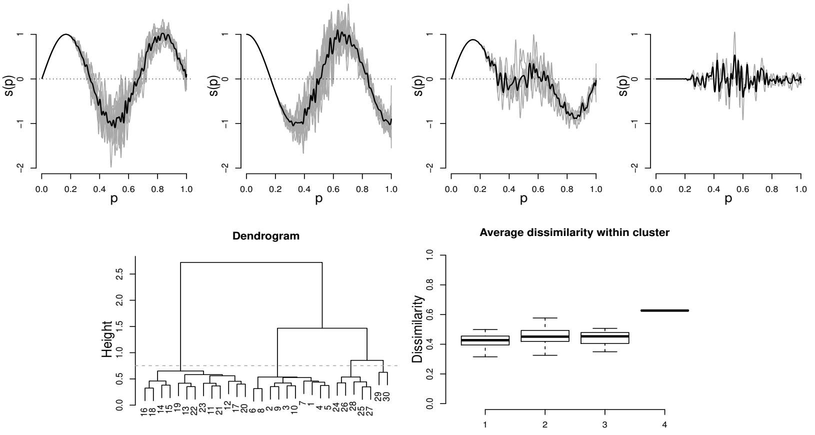
2

3

4

A r

seg



Output of the proposed algorithm for one replicate. Upper panels show the 4 clusters; bottom-left panel shows the dendrogram; bottom-right panel shows the boxplot of the average dissimilarity within each cluster.

Simulations

Methods

- `funFem`: a functional mixture model (Bouveyron and Brunet-Saumard, 2014)
- `FPCA`: a k-means algorithm based on the principal component rotation of data (Adelfio et al., 2011)

Simulations

Methods

- `funFem`: a functional mixture model (Bouveyron and Brunet-Saumard, 2014)
- `FPCA`: a *k*-means algorithm based on the principal component rotation of data (Adelfio et al., 2011)

Mesasures

- $\text{Area}(l^*) = l^{*-1} \sum_{j=1}^{l^*} \left\{ q_j^{-1} \sum_{i=1}^{q_j} \int_0^1 \left(\mid \overline{\beta}_j(p) - \beta_j^i(p \mid \boldsymbol{\theta}) \mid \right) dp \right.$
- $\rho_{\text{dist}}(l^*) = l^{*-1} \sum_{j=1}^{l^*} \left\{ 1 - \left[2 \Big(q_j(q_j - 1) \Big)^{-1} \sum_{i=1}^{q_j - 1} \sum_{z>i}^{q_j} \rho_{iz} \right]^2 \right\}$
- the average number of clusters l^*

Simulations

Average area, average distance based on correlation (ρ_{dist}) and average of the optimal number of discovered clusters (I^*) using the three different algorithms (clustEff, funFEM and FPCA) and as benchmark measure the true partition of curves in the 100 runs. SD in brackets.

		True	clustEff	funFEM	FPCA
Sim 1	I^*	3.00(0.00)	3.51(1.63)	5.06(0.98)	3.35(0.63)
	Area	0.216(0.091)	0.205(0.091)	0.178(0.079)	0.206(0.090)
	ρ_{dist}	0.010(0.015)	0.010(0.015)	0.008(0.008)	0.010(0.015)
Sim 2	I^*	4.00(0.00)	4.23(0.95)	3.44(1.05)	3.83(0.38)
	Area	0.133(0.102)	0.130(0.099)	0.177(0.146)	0.142(0.116)
	ρ_{dist}	0.441(0.177)	0.426(0.175)	0.436(0.203)	0.437(0.171)

Simulations

Average area, average distance based on correlation (ρ_{dist}) and average of the optimal number of discovered clusters (l^*) using the three different algorithms (clustEff, funFEM and FPCA) and as benchmark measure the true partition of curves in the 100 runs. SD in brackets.

		True	clustEff	funFEM	FPCA
Sim 1	l^*	3.00(0.00)	**3.51(1.63)**	5.06(0.98)	**3.35(0.63)**
	Area	0.216(0.091)	**0.205(0.091)**	0.178(0.079)	**0.206(0.090)**
	ρ_{dist}	0.010(0.015)	**0.010(0.015)**	0.008(0.008)	**0.010(0.015)**
Sim 2	l^*	4.00(0.00)	**4.23(0.95)**	3.44(1.05)	**3.83(0.38)**
	Area	0.133(0.102)	**0.130(0.099)**	0.177(0.146)	**0.142(0.116)**
	ρ_{dist}	0.441(0.177)	**0.426(0.175)**	0.436(0.203)	**0.437(0.171)**

Contents

Inspiratory capacity data

A study carried out in 1988-1991 in Northern Italy

- $N = 2,045$ subjects (51% Male and 49% Female)
- $q = 9$ (age, height, body mass index (BMI), sex, current smoking status, occupational exposure, cough, wheezing and asthma)

Inspiratory capacity data

A study carried out in 1988-1991 in Northern Italy

- $N = 2,045$ subjects (51% Male and 49% Female)
- $q = 9$ (age, height, body mass index (BMI), sex, current smoking status, occupational exposure, cough, wheezing and asthma)

The model basis

- Intercept: $\boldsymbol{b}(p) = [1, \log(p), \log(1-p)]^T$
- Covariates: a shifted Legendre polynomials up to a 5$^{\text{th}}$ degree (Abramowitz and Stegun, 1964)

Inspiratory capacity data

A s

Th

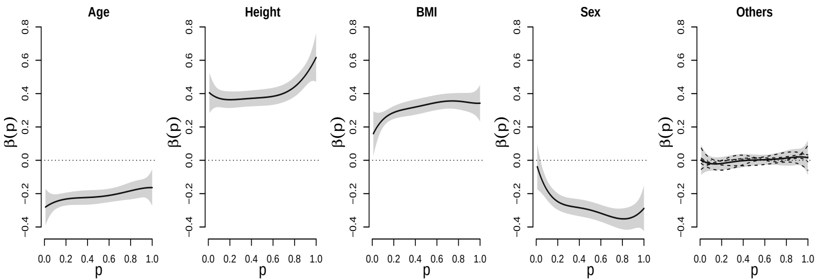
Table 1: Average area, average correlation (ρ_{dist}) and average of the optimal number of discovered clusters (l^*) are compared across the three algorithm (clustEff, funFEM and FPCA).

	clustEff	funFEM	FPCA
l^*	5	3	3
Area	0.004	0.039	0.039
ρ_{dist}	0.197	0.710	0.710

Inspiratory capacity data

A s

Th



The five clusters obtained applying the clustEff algorithm on the estimated quantile regression coefficients of inspiratory capacity dataset. Black solid lines are the mean curves; the dashed lines are the effects curves; the shaded areas are identified by the mean lower and upper bands within each cluster. The dotted line indicates the zero.

Contents

To sum-up

- We proposed a new dissimilarity measure based on similarities in shape and distance among general curves, both effects curves (in QRCM) or waveform curves;

To sum-up

- We proposed a new dissimilarity measure based on similarities in shape and distance among general curves, both effects curves (in QRCM) or waveform curves;
- We developed the `clustEff` R packages implementing the proposed algorithm;

To sum-up

- We proposed a new dissimilarity measure based on similarities in shape and distance among general curves, both effects curves (in `QRCM`) or waveform curves;
- We developed the `clustEff` R packages implementing the proposed algorithm;
- Results of two different simulation scenarios showed good performance of our proposal with respect of two competitors funFEM and FPCA;

To sum-up

- We proposed a new dissimilarity measure based on similarities in shape and distance among general curves, both effects curves (in `QRCM`) or waveform curves;
- We developed the `clustEff` R packages implementing the proposed algorithm;
- Results of two different simulation scenarios showed good performance of our proposal with respect of two competitors funFEM and FPCA;
- Results on the Inspiratory Capacity data showed a variable selection perspective.

Thanks for the attention!!!

Abramowitz, M. and I.A. Stegun (1964). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Vol. 55. Courier Corporation.

Adelfio, G., M. Chiodi, A. D'Alessandro, and D. Luzio (2011). "FPCA algorithm for waveform clustering". In: *Journal of Communication and Computer* 8.6, pp. 494–502.

Adelfio, G., M. Chiodi, A. D'Alessandro, D. Luzio, G. D'Anna, and G. Mangano (2012). "Simultaneous seismic wave clustering and registration". In: *Computers & geosciences* 44, pp. 60–69.

Adelfio, G., F. Di Salvo, and M. Chiodi (2016). "Space-time FPCA Algorithm for clustering of multidimensional curves." In: *Proceeding of the 48th Scientific Meeting of the Italian Statistical Society, Salerno*.

Bouveyron, C. and C. Brunet-Saumard (2014). "Model-based clustering of high-dimensional data: A review". In: *Computational Statistics & Data Analysis* 71, pp. 52–78.

Frumento, P. and M. Bottai (2016). "Parametric modeling of quantile regression coefficient functions". In: *Biometrics* 72.1, pp. 74–84.

Garcia-Escudero, L.A. and A. Gordaliza (2005). "A proposal for robust curve clustering". In: *Journal of classification* 22.2, pp. 185–201.

Gower, J.C. (1975). "Generalized procrustes analysis". In: *Psychometrika* 40.1, pp. 33–51.

James, G.M. (2007). "Curve alignment by moments". In: *The Annals of Applied Statistics*, pp. 480–501.

Kneip, A. and T. Gasser (1992). "Statistical tools to analyze data representing a sample of curves". In: *The Annals of Statistics*, pp. 1266–1305.

Koenker, R. (2005). *Quantile regression*. 38.

Koenker, R. and G. Bassett Jr (1978). "Regression quantiles". In: *Econometrica: journal of the Econometric Society*, pp. 33–50.

Ramsay, J.O. and X. Li (1998). "Curve registration". In: *Journal of the Royal Statistical Society: Series B* 60.2, pp. 351–363.

Silverman, B.W. (1995). "Incorporating parametric effects into functional principal components analysis". In: *Journal of the Royal Statistical Society. Series B*, pp. 673–689.

Wang, K. and T. Gasser (1997). "Alignment of curves by dynamic time warping". In: *The annals of Statistics* 25.3, pp. 1251–1276.