UNIVERSITÀ DEGLI STUDI DI PALERMO

Dottorato in Scienze Economiche, Statistiche, Psicologiche e Sociali
Dipartimento di Scienze Economiche, Aziendali e Statistiche
SECS-S/01 – Statistica

# Penalized regression and clustering in high-dimensional data

IL DOTTORE
**Gianluca Sottile**

IL COORDINATORE
**Prof. Vito M.R. Muggeo**

IL TUTOR
**Prof. Marcello Chiodi**

CICLO XXX
ANNO CONSEGUIMENTO TITOLO 2018

*"To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of."*

*"Natural selection is a mechanism for generating an exceedingly high degree of improbability."*

Ronald Fisher

Università degli Studi di Palermo

# *Abstract*

**Penalized regression and clustering in high-dimensional data**

by Gianluca SOTTILE

The main goal of this Thesis is to describe numerous statistical techniques that deal with high-dimensional genomic data.

The Thesis begins with a review of the literature on penalized regression models, with particular attention to least absolute shrinkage and selection operator (LASSO) or $L_1$-penalty methods. $L_1$ logistic/multinomial regression models are used for variable selection and discriminant analysis with a binary/categorical response variable.

The Thesis discusses and compares several methods that are commonly utilized in genetics, and introduces new strategies to select markers according to their informative content and to discriminate clusters by offering reduced panels for population genetic analysis.

After having accomplished its main objective, the thesis addresses the issue of tuning parameter selection in LASSO models, studying consistency with high-dimensional data. The tuning parameter balances the trade-off between model fit and variance reduction in sparse models and its value is crucial in all the lasso-type regression.

Finally, this Thesis introduces a LASSO method that can be applied to quantile regression coefficients modeling (QRCM), an approach that permits describing the coefficients of a quantile regression model as parametric functions of the order of the quantile. Compared with standard quantile regression, QRCM facilitates estimation, inference, and interpretation of the results, and generally yields a gain in efficiency. However, since each predictor has multiple associated coefficients, the total number of parameters escalates quickly with the size of the model matrix, causing numerical problems and large standard errors. Using the $L_1$-penalty in this framework permits keeping a parsimonious set of parameters and performing variable selection in an efficient way.

# *Acknowledgements*

First and foremost I want to thank my supervisor Prof. Marcello Chiodi. It has been an honor to be his Ph.D. student. He has taught me, both consciously and unconsciously, how good Statistics is done. I appreciate all his contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating, even during tough times in the Ph.D program.

The members of the Statistical group have contributed immensely to my personal and professional time at DSEAS. The group has been a source of friendships as well as good advice and collaboration. I am especially grateful to the professors Giada Adelfio and Vito Muggeo who allowed me to further improve my research and grow as a researcher. I would also like to thank Dr. Luigi Augugliaro for the coffee breaks and for inspiring me in programming; the doctors Antonino Abbruzzo, Giovanni Boscaino and Mariangela Sciandra for the beautiful moments spent together and for the inspiring discussions; Prof. Massimo Attanasio, Dr. Francesca Di Salvo, and the professors Gianfranco Lovison and Antonella Plaia for the long talks and trust they have placed in me on many occasions. A thank also goes to my colleagues for this adventure, in particular Fabio, Marco, Marianna, Ola, Pietro, Roberto and Simona (aka Lime).

Moreover, I would also like to thank Prof. Baldassare Portolano and the members of his group, the doctors Rosalia Di Gerlando, Salvatore Mastrangelo, Marco Tolone, Maria Teresa Sardina (aka Ciupy) and Anna Maria Sutera, for having had the opportunity to collaborate and develop my research. It has been and will be a real pleasure to continue to work together.

I would also like to thank the first people who have believed in me from an academic point of view, Prof. Giuseppe Alonzo and his wife Marika.

Furthermore, during my visiting period at Karolinska Institutet in Stockholm, I had the pleasure of meeting amazing people and magical places. For this opportunity, I owe thanks to Prof. Matteo Bottai, head of the Unit of Biostatistics, and to Dr. Paolo Frumento for the several suggestions.

Last but not the least, a special thank goes to my family and to my large family acquired, for all their love and encouragement. Thanks to my parents who have always believed in me and my potential, and to my brother who has always been close to me. Finally, an infinite thank you goes to my wonderful wife Jenny, to support me in every decision,

for giving me some of the best moments of my life and for those who will come in the future.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 The high-dimensional problem

In fields such as as biology and genomics usually many information are avaliable for a limited number of observations. As a result statistical uncertainty can be high and common models cannot be applied. The combination of small sample size $N$ and large number of variables $p$ configures the high dimensional framework. Intuitively, in high-dimensional setting, not all covariates are equally relevant, and the concept of sparsity plays an important role. To shrink to zero most of the covariates' coefficients, if irrelevants, leads to identifiability avoiding the overparametrization problem.

With such data, regularized or penalized methods are needed to fit the model and variable selection is often the most important aspect of the analysis. The Least Absolute Shrinkage and Selection Operator (LASSO) introduced by Tibshirani, (1996) is a penalized method similar to the ridge regression (Tikhonov, 1943; Hoerl and Kennard, 1970) but uses the $L_1$-penalty $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ instead of the $L_2$-penalty $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} \beta_j^2$. Figure 1.1 shows the geometric interpretation of the two penalty constraints.

An important feature of the LASSO is that it can be used for variable selection. Compared to classical variable selection methods, the LASSO has two advantages. First, the selection process is based on continuous trajectories of regression coefficients as function of the penalty level and is hence more stable than subset selection methods. Second, the LASSO is computationally feasible for high-dimensional data (Osborne, Presnell, and Turlach, 2000a; Osborne, Presnell, and Turlach, 2000b; Efron et al., 2004). Several authors have studied the model-selection consistency of the LASSO in the sense of selecting exactly the set of variable with nonzero coefficients, that is, identifying the subset $\{j : \beta_j \neq 0\}$ of $\{1, \ldots, p\}$. In the low dimensional setting with fixed $p$, Knight and Fu,

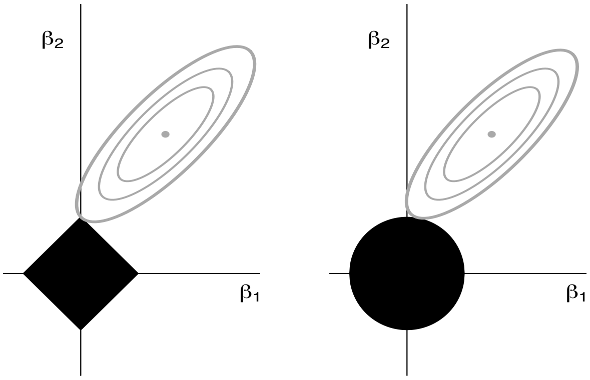


FIGURE 1.1: Left: Contour lines of residual sum of squares, with $\hat{\beta}$ being the least squares estimator, and $L_1$-ball corresponding to the lasso problem. Right: Analogous to left panel but with $L_2$-ball corresponding to ridge regression.

(2000) showed that, under appropriate conditions, the LASSO is consistent for estimating the regression parameters $\beta_j$. However, the LASSO is not variable-selection consistent without proper assumptions (Bühlmann and Van De Geer, 2011).

Penalized regression models have gained popularity and attractiveness to perform selection of variables, and while several extensions have been discussed, such as elastic net (Zou and Hastie, 2005), adaptive lasso (Zou, 2006), fused lasso (Tibshirani et al., 2005) and grouped lasso (Yuan and Lin, 2006), the *'naive'* lasso still represents a valuable tool, in both theory and applications (Tibshirani, 2011).

## 1.2 The LASSO

LASSO was originally introduced in the context of least squares and was later extended to a wider variety of statistical models including generalized linear models, generalized estimating equations, proportional hazards models and quantile regression. Consider a sample consisting of $N$ observations, each of which consists of $p$ covariates and a single outcome. Let $\boldsymbol{y}$ be the outcome and $\boldsymbol{X}$ be the model matrix of dimension $N \times p$. The objective of LASSO is to solve the Lagrangian form

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{N^{-1}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1\right\}, \tag{1.1}$$

where $\lambda$ is the tuning parameter. The tuning parameter balances the trade-off between model fit and model sparsity, and selecting an appropriate value is the key point of LASSO regression. If $\lambda$ is 0, then the problem becomes unconstrained and the coefficients $\hat{\boldsymbol{\beta}}$ are simply obtained by ordinary least squares. Vice versa, all coefficients $\hat{\boldsymbol{\beta}}$ shrink to 0 when $\lambda$ goes to infinity. Figure 1.2 shows an example path of regression coefficients as function of $\log(\lambda)$.

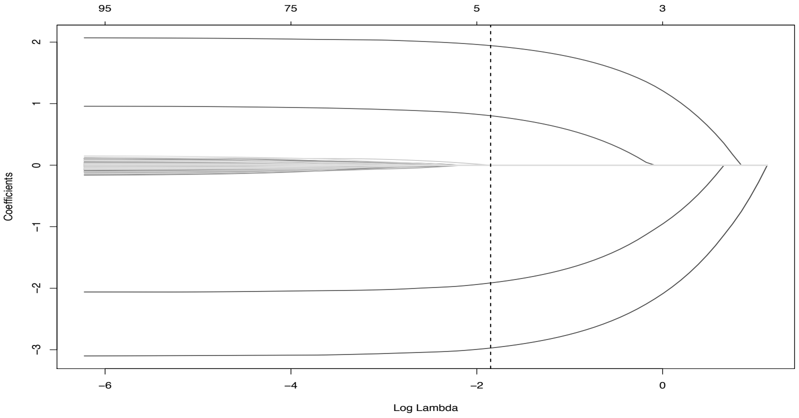


FIGURE 1.2: The grey lines are the paths of regression coefficients shrinking towards zero as $\lambda$ increases. If we draw a vertical line in the figure, it will give a set of regression coefficients corresponding to a fixed $\lambda$. The x-axis shows the logarithm of shrinkage instead of $\lambda$.

## 1.3 The role of simulation studies in statistical methodology research

In many applications, such as studies involving microarray data, theoretical assumptions are quite hard or impossible in practice to meet, since most of them are unverifiable or simply violated. Planning simulation studies can help to assess the gap between theory and realistic expectations. Different simulation scenarios can be created. For example, by building a scenario in which all assumptions are fulfilled, optimal performance is expected; conversely, when assumptions are not met, performance is generally lower. One important drawback of simulation studies is that they are artificial and far from real data. We can combine simulated data and real data to better mimic realistic situations, in order to give a better idea of what has to be expected from the real data.

## 1.4 Outline of thesis

The contribution of this thesis includes several aspects of penalized models and clustering and is summarized in the following.

- **Penalized Multinomial Regression and Stability Selection.** In Chapter 2, we propose a mixed strategy to do variable selection and discriminant analysis. Starting from available high-throughput SNP data, Penalized Multinomial Regression and Stability Selection are applied to identify the optimal set of informative SNPs useful to discriminate among Sicilian dairy sheep breeds. This methodology, as proposed in this thesis, could be considered as a high level strategy to select markers in high-throughput genotyping.

  This Chapter has been published as Sottile et al., (2017). Different sections of the paper were written by different co-authors according to their expertise.

- **Tuning parameter selector.** In Chapter 3, a new criterion to select the tuning parameter $\lambda$ is proposed. The criterion is quite simple to compute and can be interpreted as the maximization of the signal-to-noise ratio. This tuning parameter selector enables to identify the true model consistently when the true model is among a set of candidate models. This methodology could be used as a valid alternative to classical criteria.

  This Chapter is an extension of the poster presented at the Internation Workshop on Statistical Modeling (Sottile and Muggeo, 2016).

- **Penalized qunantile regression coefficients modeling.** In Chapter 4, applying the $L_1$-penalty to the integrated loss function described by Frumento and Bottai, (2016) is proposed. This methodology allows to select variables in a new parametric approach to model the quantile function, in a quantile regression framework.

  This Chapter is based on a paper currently under review in a journal as Sottile et al.

- **Clusters of effects curves.** In Appendix A, a new dissimilarity measure based on both the shape of curves and their distances is proposed. This measure, useful for the application of any hiearchical clustering method, can be used to cluster curves computed in a quantile regression framework, namely effects curves, or waveform curves as in functional data analysis.

  This Appendix is an extension of the poster presented at the Internation Workshop on Statistical Modeling (Sottile and Adelfio, 2017). It is currently under review in a journal.

- **R packages.** In Appendix B, I provide a description of four different R packages (**asnr**, **qrcmNP**, **clustEff**, **islasso**) that I implemented, is provided.

  Codes for the R packages were written mainly by me and inspired, in some cases, by original codes of V. Muggeo and P. Frumento.

# Chapter 2

# Selection of markers from high-throughput genotyping: Application in Sheep Breeds

## 2.1 Introduction

Assignment tests using genetic information to establish population membership of individuals provide the most direct methods to determine population of origin of unknown individuals (Negrini et al., 2009). The identification of individuals' breed/population of origin has several practical applications in livestock and is useful in different biological contexts, such as management of livestock genetic resources for breed confirmation, estimation of hybridization level and authentication of brand products that are produced using only a few particular breeds or populations (Wilkinson et al., 2011; Bertolini et al., 2015). Moreover, assignment of individuals to a specific breed is very important both for biodiversity purposes and products traceability, especially when the phenotypic differentiation among breeds is difficult (Tolone et al., 2012).

Recently developed genomic technologies, such as medium and high-density SNP arrays, are important tools that can be used for these purposes. Dense genome-wide data is valuable but is relatively costly and time-consuming or computationally expensive to analyze. However, some methods are tractable and capable to efficiently predict breed composition using breed frequencies of thousands of markers (Kuehn et al., 2011). Therefore, it is often desirable to reduce the number of markers according to their information content, in order to create reduced panels for population genetic analysis (Paschou et al., 2007). Many clustering algorithms have been developed employing population genetic data to assign individuals to clusters (Jakobsson and Rosenberg, 2007). Several statistical methods were used to determine which genetic markers contain the most information to discriminate among populations (Wilkinson et al., 2011; Rosenberg, 2005), such as the combined approach of Principal Component Analysis (PCA) and Random Forest (RF) (Bertolini et al., 2015), multivariate canonical discriminant analysis (Dimauro et al., 2013), the statistic delta (Shriver et al., 1997), and Wright's $F_{\text{st}}$ (Bowcock et al., 1994). While all these methodologies yielded reduced marker panels useful for breed identification, the power of assignment depended on the utilized method.

## 2.2 Materials and methods

### 2.2.1 Data

A total of 236 animals, randomly collected from several farms in different areas of Sicily, were used for the analysis. Samples consisted of 30 Barbaresca (Bar), 51 Comisana (Com), 77 Pinzirita (Pin), 30 Sarda (Sar) and 48 Valle del Belice (VdB) individuals. The procedures involving animal samples collection followed the recommendation of Directive 2010/63/EU. All animals were genotyped for 54,241 SNPs using the Illumina OvineSNP50K Genotyping BeadChip. Genotyping was performed by Dipartimento Scienze Agrarie e Forestali, University of Palermo. Input data were genotyping data of 54 241 SNPs, i.e. GType data in Illumina AB format exported from GenomeStudio v1.0 (Illumina, Inc.). We excluded all SNPs not assigned to chromosome (OAR) or assigned to X and Y chromosomes. Markers were filtered according to the following quality criteria: i) call frequency ( 95%), ii) Minor Allele Frequency (MAF 0.01). SNPs that did not satisfy these quality criteria were excluded. A total of 48 068 SNPs were retained for subsequent analyses. We transformed the genotyping data to numeric values, without any loss of information, in order to apply into PMR. The initial data table $\boldsymbol{X}$ consisted of $N$ rows, one per animal, and $p$ columns, one per SNP. Each entry of $\boldsymbol{X}$, AA, AB and BB, was scored as -1, 0, 1, or empty. SNPs with missing genotypes were randomly imputed within each breed according to the corresponding genotype frequency.

### 2.2.2 Statistical analysis of Single Nucleotide Polymorphisms and variable reduction

Each sheep breed was divided into a test sample and a validation sample. The validation sample, generated by randomly sampling 15% of animals within each breed, was used for the final validation procedure of breed assignment. The test sample consisted of the remaining animals.

Suppose to have a set of $N$ individuals and $p$ SNPs, with $p \gg N$, partitioned into $K$ groups. The main goal is to select a limited number of SNPs from an initially large set, to be used to predict group membership with high discrimination power. To achieve this aim some authors proposed the LASSO method (Tibshirani, 1996) or $L_1$-penalty, a shrinkage and selection method for linear regression. In statistics, LASSO (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model. The idea behind LASSO is to shrink towards zero the coefficients that are less than a certain fixed value. LASSO was originally introduced in the context of least squares. Let $y_i$ be the outcome and $\boldsymbol{x}_i = (x_1, x_2, \ldots, x_p)^T$ be the covariate vector for the $i^{\text{th}}$ observation, $i = 1, \ldots, n$. LASSO linear regression solves

$$\min_{\beta_0, \boldsymbol{\beta}} \left\{ N^{-1} \sum_{i=1}^{N} (y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t$$

where $t$ is a prespecified free parameter that determines the amount of regularization, $\beta_0$ is the intercept of the model and $\boldsymbol{\beta}$ is the $p$-variate vector of regression coefficients. Letting $\boldsymbol{X}$ be the covariate matrix, so that $\boldsymbol{X}_{ij} = (\boldsymbol{x}_i)_j$ and $\boldsymbol{x}_i^T$ is the $i^{\text{th}}$ row of $\boldsymbol{X}$ we can write this in the so-called Lagrangian form

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ N^{-1} \|y - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

where the relationship between $\lambda$ and $t$ is $t \approx \lambda^{-1}$, i.e., it is approximately the multiplicative inverse of $\lambda$. The $L_1$-norm of $\boldsymbol{\beta}$ is a constraint on the regression coefficients that strictly depends on the tuning parameter $\lambda$.

Although LASSO regression was originally defined for least squares, it is easily extended to a wide variety of statistical models including generalized linear models. In our framework, given the nature of the outcome $y$ that is a categorical variable with $K > 2$ levels, we used a penalized multinomial regression. Here, we model the probability to belong to breed $k$ given the SNPs' matrix $X$ of dimension $N \times p$,

$$\Pr(y = k | \boldsymbol{X}) = \frac{e^{\boldsymbol{X}\boldsymbol{\beta}_k}}{\sum_{l=1}^{K} e^{\boldsymbol{X}\boldsymbol{\beta}_l}}, k = 1, \ldots, K.$$

Let the outcome $y$ be the $N \times K$ indicator response matrix, with elements $y_{il} = I(y_i = l), l = 1, \dots, K$ and $i = 1, \dots, N$. Then the regression coefficients are obtained as the solution of the following optimization problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{k(p+1)}} \left\{ -N^{-1} \sum_{i=1}^{N} \left( \sum_{l=1}^{K} y_{il} \boldsymbol{X} \boldsymbol{\beta}_l - \log \left( \sum_{l=1}^{K} e^{\boldsymbol{X} \boldsymbol{\beta}_l} \right) \right) + \lambda \sum_{j=1}^{p} \sum_{l=1}^{K} |\boldsymbol{\beta}_{jl}| \right\},$$

where $\boldsymbol{\beta}$ is a $p \times K$ matrix coefficients, $\boldsymbol{\beta}_k$ refers to the $k$-th column (for outcome breed $k$), and $\boldsymbol{\beta}_j$ the $j$-th row (vector of $K$ coefficients for variable $j$).

The analysis was conducted using the `glmnet` R package (Friedman, Hastie, and Tibshirani, 2010) All the following analysis was computed using `glmnet` package of the R software 3.3.0 (2017). The Stability Selection method (Meinshausen and Bühlmann, 2010) was used to discover the most stable subset of variables that have nonzero weight in the model. Assume to have a generic structure estimation algorithm (i.e., the LASSO) that takes a dataset $\boldsymbol{X}$ and a regularization parameter $\lambda$, and returns a selection set $S^\lambda$. The $j$-th covariate belongs to $S^\lambda$ if the regression coefficient $\boldsymbol{\beta}_j \neq 0$. The SS algorithm runs as follows:

1. Define a candidate set of dimension $m$ of regularization parameters $\Lambda = \{\lambda_1, \dots, \lambda_m\}$ and a number $n$ of subsample.

2. For each $\lambda \in \Lambda$, do:

   a. Start with the full dataset $\boldsymbol{X}$

   b. For each $i = 1, \dots, n$:

      i. Subsample from $\boldsymbol{X}$ without replacement to generate a smaller dataset of size $N/2$, namely $\boldsymbol{Z}_i$.

      ii. Run the selection algorithm on dataset $\boldsymbol{Z}_i$ with parameter $\lambda$ to obtain a selection set $S^\lambda$.

   c. Given the $n$ selection sets from each subsample, calculate the empirical selection probability for each covariate:

   $$\Pi_j^\lambda = \mathrm{P}\{j \in S^\lambda\} = n^{-1} \sum_{i=1}^{n} I(j \in S^\lambda), j = 1, \dots, p.$$

   The selection probability for covariate $j$ is its probability of being selected by the algorithm.

3. Given the selection probabilities for each covariate and for each value $\lambda \in \Lambda$, construct the stable set according to the following

definition:

$$\hat{S}^{\text{stable}} = \left\{ j : \max_{\lambda \in \Lambda} \Pi_j^{\lambda} \geq \pi_{\text{thr}} \right\}$$

where $\pi_{\text{thr}}$ is a predefined threshold.

In our study, fixing a sequence of 100 values of $\lambda$, step 2.(b) was repeated $B$ times by randomly splitting, within each breed, the test sample. After calculating the empirical selection probability for each SNP and fixing the threshold value, a final set $\hat{S}^{\text{stable}}$ of $p_1$ SNPs was selected. For this reduced panel, a new multinomial regression model was then fitted. To assess the classification performance of this set of $p_1$ SNPs we tested the discrimination rule using the validation sample which is considered to be an independent subset of samples.

### 2.2.3 Other Statistical and Genetic Methods

To better understand the potential of our strategy and the strength of a reduced panel of SNPs we decided to use the *k*-means approach, which is an unsupervised technique. In particular, we used all the principal components up to 70% of explained variance of the model matrix $\boldsymbol{X}$. We applied this technique once to the whole set of SNPs and once to the reduced set $p_1$. The efficiency of the selected markers to cluster individuals was also tested using model-based clustering algorithm implemented in the Admixture software 1.3.0 (Alexander and Lange, 2011) which used unsupervised classification approaches and Genepop software 4.1.4 (Rousset, 2008) to calculate $F_{\text{st}}$ (Bowcock et al., 1994). The most probable number of populations in the dataset ($K$) was estimated using the default (5-fold) Admixture's cross-validation procedure, by which estimated prediction errors are obtained, for each $K$ value, by adopting a kind of 'leave-one-out' approach through which an estimation of prediction errors can be assumed to be the most suitable one. Genepop was also used to estimate population relatedness using pair-wise estimates of $F_{\text{st}}$ among breeds. The reduced panel was analyzed using SNPchiMp (Nicolazzi et al., 2015) to obtain information on the genomic distribution of SNPs.

In order to compare our approach to those previously reported, another mixed strategy was considered (Bertolini et al., 2015). In particular, PCA and RF was used to discover a new SNP panel able to discriminate among the breeds. For each autosome, the top 20 SNPs were selected and merged together, leading to a final panel of 520 markers. RF based on the selected 520 SNPs were built on the test sample. The Mean Decrease in the Gini Index (MDGI) or the Mean Accuracy Decrease (MAD) were used in order to select the most discriminant SNPs. Four different SNP panels were created selecting the first 48 and 96 SNPs from the MDGI and the first 48 and 96 from the MAD, respectively. This SNP panels' size was chosen considering the practical possibilities to develop multiplex SNP panels containing a reduced number of SNPs for field applications (Bertolini et al., 2015). For each of the four reduced panels, a new RF was fitted and the corresponding out-of-bag (OOB) error rate was calculated. Classification performance of these four RFs was assessed also using the validation sample.

A simulation study has been done to compare the performance of our proposed strategy and the PCA-RF strategy. A group of genetic variants has been randomly generated by using the real dataset. In general, we sampled with replacement $N$ observations of the real dataset, so to maintain the same structure and association between SNPs. Moreover, we built $\tilde{X}_{\text{test}}$ which is the simulated test sample and $\tilde{X}_{\text{val}}$ which is the simulated validation sample (15% of the sample size). The response variable $\tilde{y}$ was the label vector of length $n$, indicating the membership of each animal to their own breed in the simulated data. $\tilde{X}_{\text{test}}$ and $\tilde{X}_{\text{val}}$ were used to evaluate the out-of-bag error and misclassification error rate for both strategies.

## 2.3 Results

### 2.3.1 Penalized Multinomial Regression and Stability Selection

Out of a total of 54,241 genotyped SNPs, 378 were unmapped and 1,450 were located on sex chromosomes. Thus, 52,413 SNPs mapped onto 26 sheep autosomes were used, and after filtering (see 2.2), the final number of common SNPs was 48,068. On these SNPs, PMR and SS procedure, with $B = 500$, were performed to select the most informative markers obtaining a final small set of 48 SNPs.

Figure 2.1 shows the 201 animals of the test sample in the subspace defined by the first three principal components calculated on these 48 SNPs. This allowed an assessment of whether the reduced SNP panel leads to loss of important genetic information which is relevant to explain the differences across breeds. Figure 1 shows partial overlaps of historically- and phylogenetically-related breeds (Tolone et al., 2012; Mastrangelo et al., 2012; Mastrangelo et al., 2014) and the difficulty in separating them. To analyze the structure of each cluster, we used the standard deviations (SD) as a measure of spread within each breed in the first three principal components. We observed a standard deviation average of about 0.65 for each principal component.

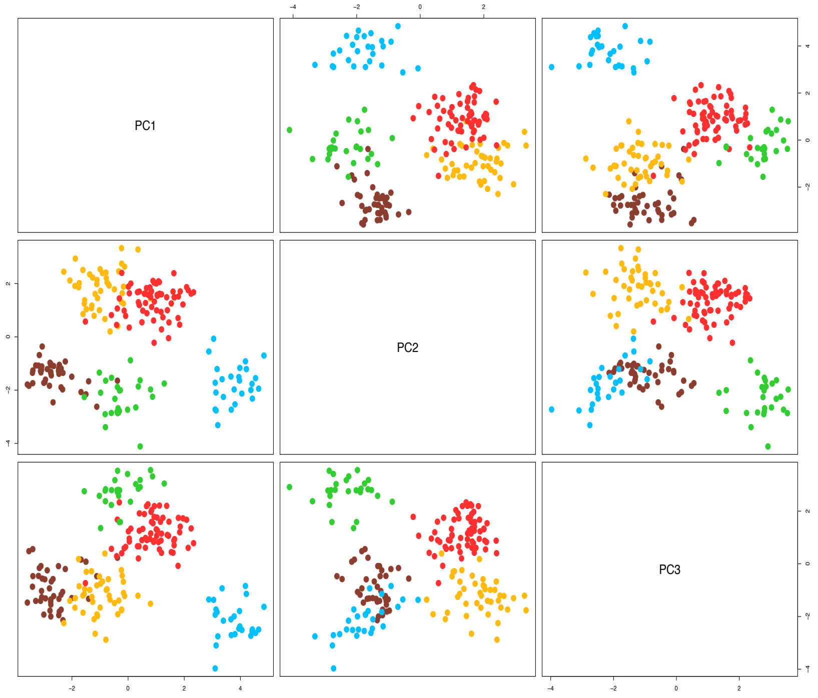


FIGURE 2.1: Plot of the first three principal components obtained using the panel of 48 Single Nucleotide Polymorphisms (coded as genotype), selected after the first step with the Penalized Multinomial Regression and Stability Selection procedure. ● = Valle del Belice (VdB), ● = Comisana (Com), ● = Pinzirita (Pin), ● = Barbaresca (Bar), ● = Sarda (Sar).

Using this SNP panel, the corresponding misclassification error rate both for test and validation sample was equal to 0%. The unsupervised strategy which consists of the combination of PCA and *k*-means also provided excellent results. Using the first 15 principal components, which are highly correlated ($> 0.50$) with 31 SNPs and explain 70% of total variability, we only missed one individual to perfectly discriminate the five breeds involved in the study.

Moreover, using the whole set of SNPs and applying the same unsupervised strategy, we still missed the same individual. In this case, 112 principal components, which are highly correlated ($> 0.50$) with 608 SNPs, are used in the $k$-means step.

### 2.3.2 Random Sampling of Single Nucleotide Polymorphisms

In order to assess the ability of the 48 selected SNPs to efficiently discriminate sheep breeds, a simulation was performed. Another sets of 48 SNPs were randomly sampled 500 times from the whole set of SNPs and the average classification accuracy in the validation sample was about 60%. Repeating this procedure, sampling different numbers of SNPs (i.e. 50, 100, 200, 400, 800, 1 600, 3 200, 6 400), the classification accuracy was tested using Kruskal-Wallis, also rank-sum test (Kruskal and Wallis, 1952), to evaluate if any increment in accuracy was significant. Results are shown in Table 2.1.

TABLE 2.1: Accuracy of classification in the test and validation sample, when different numbers ($p$) of SNPs are sampled from the whole set. Results are based on 500 simulated datasets. The last two columns report the Kruskal-Wallis rank-sum statistic and its p-value, that compare each pair of consecutive values of accuracy.

| p | Avg % test | Avg % val | Kruskal-Wallis test | p-values |
|---:|---:|---:|---:|---:|
| 48 | 99.9% | 60.6% | - | - |
| 50 | 99.9% | 63.3% | 27.66 | <.0001 |
| 100 | 100% | 75.2% | 376.50 | <.0001 |
| 200 | 100% | 83.1% | 254.08 | <.0001 |
| 400 | 100% | 87.7% | 126.38 | <.0001 |
| 800 | 100% | 90.8% | 82.10 | <.0001 |
| 1 600 | 100% | 92.7% | 41.67 | <.0001 |
| 3 200 | 100% | 94.5% | 265.60 | <.0001 |
| 6 400 | 100% | 95.2% | 16.63 | <.0001 |

Figure 2.2 shows the strength of the selected panel of 48 SNPs to discriminate across all the breeds, and the difficulty to perfectly discriminate among them using a large set of SNPs. Moreover, the Kruskal-Wallis test results were significant for each increment after sampling even more SNPs (Table 2.1).

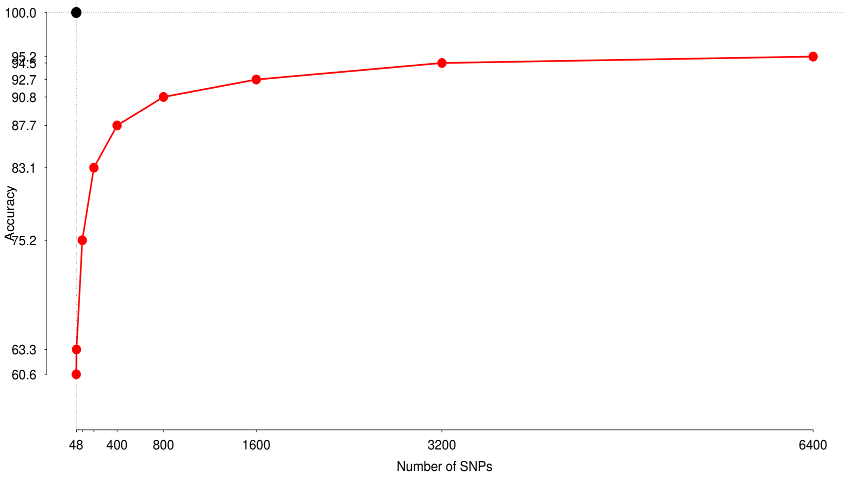


FIGURE 2.2: Plot of the mean accuracy to classify among the five breeds after repeating, for different number of Single Nucleotide Polymorphisms (SNPs), a random sampling procedure from the whole set of available SNPs. Black dot is the accuracy level of the selected 48 SNP panel.

### 2.3.3 Penalized Multinomial Regression and Stability Selection versus Principal Component Analysis and Random Forest

PMR-SS procedure is a new strategy used for assigning animals to a breed. In order to compare our approach with other previously reported strategies and to test its efficiency in assigning individuals, PCA and RF strategy (Bertolini et al., 2015) were also used with the real data. With respect to the two first ranking of SNP panels (MDGI and MAD for 48 and 96 SNPs), the OOB errors in the test sample were 4.09% and 2.03%, respectively, whilst the misclassification error rates for the validation sample were both 2.86%. In the second ranking, the OOB errors for the test sample were 2.55% for the 48 SNP panel and 2.55% for the 96 SNP panel, whilst the misclassification error rates are 5.71% and 2.86%, respectively. These results are summarized in Table 2.2.

TABLE 2.2: Out-of-bag (OOB) errors on the test sample and misclassification error rates on the validation sample for the two Single Nucleotide Polymorphism panels and the two rankings, Mean Decrease in the Gini Index (MDGI) and Mean Accuracy Decrease (MAD) and for the mixed strategy Penalized Multinomial Regression and Stability Selection (PMR-SS).

| **Rankings** | **n. of SNPs** | **OOB** | **Misclassification** |
|---|---|---|---|
| MDGI | 48 | 8.21/201 | 1/35 |
| | 96 | 4.08/201 | 1/35 |
| MAD | 48 | 5.12/201 | 2/35 |
| | 96 | 5.12/201 | 1/35 |
| PMR-SS | 48 | 0/201 | 0/35 |

Figure 2.3 shows the distribution of the 48 selected SNPs across the 26 chromosomes, and the four SNPs panels obtained through PCA and RF procedure. 15 and 13 SNPs out of 48 are the same as in two 48 rankings MDGI and MAD, respectively.

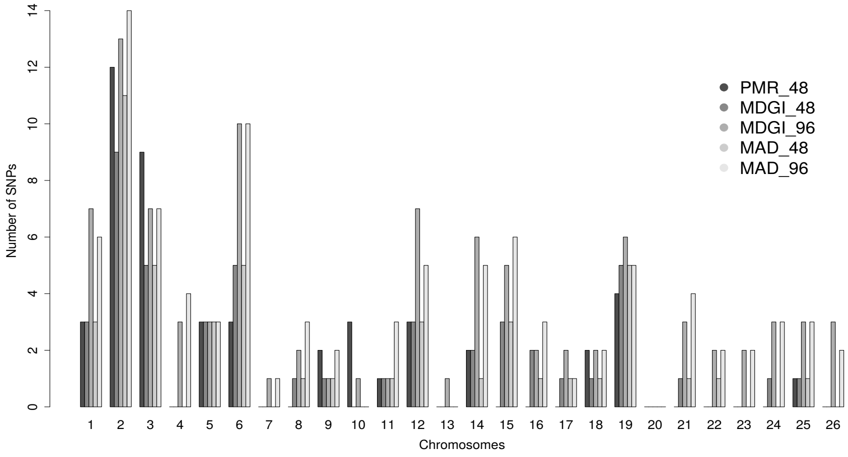


FIGURE 2.3: Chromosome distribution of the Single Nucleotide Polymorphisms (SNPs) selected according to the proposed strategy, Penalized Multinomial Regression and Stability Selection (PMR-SS), and based on the two panels Mean Decrease in the Gini Index (MDGI) and the two panels Mean Accuracy Decrease (MAD).

To compare PMR-SS and PCA-RF strategies in more depth, we performed a simulation study. We artificially built, 300 times, test ($\tilde{X}_{\text{test}}$) and validation ($\tilde{X}_{\text{val}}$) samples by sampling with replacement observation from the real dataset ($\boldsymbol{X}$). For each replicate, OOB and misclassification error rates were calculated according to a new reselected SNP panel. Results are summarized in Table 2.3.

TABLE 2.3: Out-of-bag (OOB%) error and misclassification error rate (MER%) on the test and validation sample for both strategies, Penalized Multinomial Regression-Stability Selection and Principal Component Analysis-Random Forest. Standard deviations in brackets. Results are based on 300 simulation runs.

| | **MDGI 48** | **MDGI 96** | **MAD 48** | **MAD 96** | **PMR-SS 48** |
|---|---|---|---|---|---|
| OOB% | 1.63 (0.81) | 1.29 (0.80) | 1.49 (0.77) | 1.25 (0.83) | 0.00 (0.00) |
| MER% | 2.11 (2.60) | 1.49 (1.89) | 1.91 (2.49) | 1.49 (1.97) | 1.46 (1.82) |

MDGI = mean decrease in the Gini index; MAD = mean accuracy decrease; PMR-SS = Penalized Multinomial Regression-Stability Selection

### 2.3.4 Breed assignment

The performance of the selected informative SNP markers in individual assignment test was evaluated using traditional genetic statistics such as model-based clustering algorithm and Wright's fixation index. These analyses were conducted using the whole set of SNPs (48,068) and the final number of selected SNPs (48). Results from within-population substructure, using admixture analysis and considering a range of 2 through 10 potential clusters ($K$), indicated that the most probable number of inferred populations was $K = 5$. A graphic representation of the estimated membership coefficients, using the whole set of SNPs and the final number of selected SNPs is shown in Figure 2.4, where model-based clustering partitioned the genome of each sample into a predefined number of components. Some breeds tend to have their own distinct cluster (Bar, Sar and VdB), whereas other breeds, such as Pin and Com, showed a complex admixture-like pattern. These results support the findings on the basis of PCA.

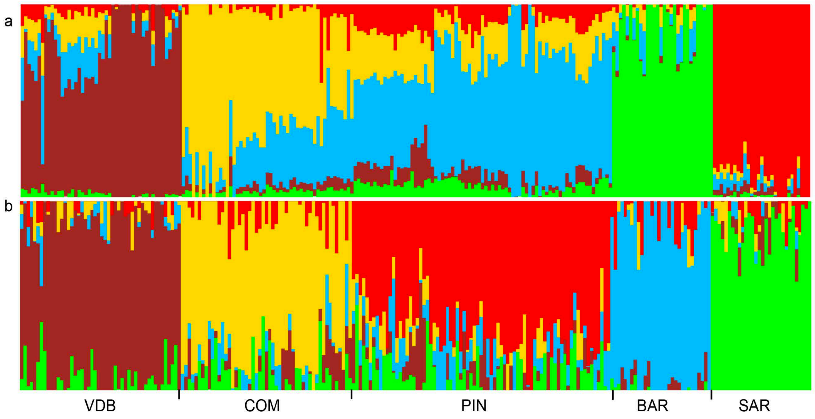


FIGURE 2.4: Model-based clustering of the five sheep breeds analyzed for the most likely clusters ($K$=5), using (a) the whole set of Single Nucleotide Polymorphisms (SNPs; 48,068); and (b) the final number of selected SNPs (48).

The degree of genetic differentiation between pairs of breeds is reported in Table 2.4. The highest $F_{st}$ value, for both SNPs panels, is seen between Bar and Sar and the lowest value was for Com versus Pin. Based upon the reference population, the average pairwise breeds $F_{st}$ showed a higher value using the 48 SNPs, confirming the ability of this method to select discriminating markers.

TABLE 2.4: Population genetic differentiation (statistic) across the five sheep breeds using the whole set of Single Nucleotide Polymorphisms (SNPs; 48,068) (top triangular) and the final number of selected SNPs (48) (bottom triangular).

| | **VdB** | **Com** | **Pin** | **Bar** | **Sar** |
|---|---|---|---|---|---|
| **VdB** | 0 | 0.05 | 0.04 | 0.10 | 0.07 |
| **Com** | 0.24 | 0 | 0.02 | 0.08 | 0.06 |
| **Pin** | 0.26 | 0.18 | 0 | 0.07 | 0.04 |
| **Bar** | 0.43 | 0.37 | 0.30 | 0 | 0.11 |
| **Sar** | 0.31 | 0.31 | 0.25 | 0.42 | 0 |

VdB = Valle del Belice; Com = Comisana; Pin = Pinzirita;
Bar = Barbaresca; Sar = Sarda

## 2.4 Discussion

The aim of this study was to apply a new strategy to identify the minimum number of informative SNPs from high-throughput genotyping data in sheep breeds reared in Sicily and to investigate their usefulness for breed assignment purposes. Generally, the selection of genetic markers useful for these purposes is based on two approaches: a deterministic one, in which markers with different allelic variants fixed in the compared breeds are used, and the probabilistic one, in which selected markers present typical allelic frequencies in different breeds (Negrini et al., 2009).

Several strategies have been already proposed to identify breed informative SNPs derived from high-throughput genotyping platforms. These systems usually include a first step in which SNPs are preselected and a second step in which different assignment methods are applied (Bertolini et al., 2015). For example, Allen et al., (2010) in a study on Irish cattle, reported a set of 43 SNPs for breed identification on the basis of allele frequency. Heaton et al., (2014) identified 163 SNPs for use in parentage testing and traceability in sheep, using the minor allele frequency ($>$ 0.3). In Mastrangelo et al., (2014), a subset of 119 SNPs was tested to evaluate their ability to assign individuals to the same groups that have been used in the present study. These SNPs were selected according to their informativeness in breed pair comparison, meaning that SNPs with the largest allele frequency differences between pairs of breeds were chosen (fixed alleles in one breed and MAF $>$ 0.25 in the others). Principal Component Analysis and k-means using this subset of SNPs showed a lack of ability to discriminate among the breeds and the presence of overlapped areas. Recently, Dimauro et al., (2015) used three complementary multivariate statistical techniques (stepwise discriminant analysis, canonical discriminant analysis and discriminant analysis) and two reduced pools of 110 and 108 SNPs, respectively, to discriminate between divergent sheep breeds.

In this Chapter, supervised approaches, Penalized Multinomial Regression and Stability Selection procedures were applied to identify the minimum number of informative SNPs from high-throughput genotyping data. These were used as a classification method for unknown samples. The method proposed in the present work differs from other studies due to the statistical technique used to reduce the number of SNPs. The main result was the selection of 48 SNPs from a whole set of 48,068. These contained enough genetic information to produce sufficient power for individuals' breed assignment, using a relatively low number of individuals for breed and closely related breeds. The majority of the SNPs are in non-coding/intergenic regions of the sheep genome, which is ideal for identification and assignment purpose since these regions/SNPs should be less influenced by natural or artificial selection (Allen et al., 2010).

The study proved that the combination of these methods allowed efficient discrimination between individuals of the studied breeds. Of course, the 48 identified SNPs that were useful to discriminate among the sheep breeds under study are probably not useful to discriminate other sheep breeds. However, the same strategy could be applied to other breeds, using different markers. Wilkinson et al., (2011) reported poor assignment power for breeds with low sample size and closely related individuals, showing that closely related breeds require about 200 markers to achieve 95% assignment success. Bertolini et al., (2017) in a study on cosmopolitan and autochthonous cattle breeds, showed that a 96-SNP panel was generally sufficient to discriminate all breeds. For the 48-SNP panel, the error rate was larger for autochthonous breeds, probably as a consequence of their admixed origin, lower selection pressure and due to bias in the construction of the SNP chip. In fact, where there is sufficient genetic heterogeneity across populations, few genetic markers can be easily used to identify and verify the origin of individuals. This becomes more complicated for population with low genetic differentiation, such as the sheep breeds involved in this study (Mastrangelo et al., 2012; Tolone et al., 2012). It is well known that a high number of genotyped animals can capture the whole within-population variability reducing the possibility that some individuals would not be assigned correctly due to atypical genotypes (Hulsegge et al., 2013). Considering the high level of admixture among these sheep breeds (Mastrangelo et al., 2017), and the relative low number of analyzed individuals, our study produced relevant results. A good separation among breeds was obtained with high percentages of correct assignment. The applicability of reduced SNP panels with low classification error rate is therefore still possible also for local breeds in which the total or partial lack of selection programs have not shaped the genome as it might be the case for cosmopolite breeds.

The combined use of Principal Component Analysis and Random Forest proposed by Bertolini et al., (2015), applied to our sheep breeds, appeared to perform poorly, even using larger panels of 96 SNPs. Simulation results showed that the proposed strategy performed slightly better than PCA-RF, and similar results were found in the application to real data. Therefore, the proposed strategy could provide a new tool that overcomes the limitation of the existing approaches when breeds are phylogenetically close.

The results reported using independent analyses, such as the model-based clustering algorithm implemented in Admixture software (Alexander and Lange, 2011) and $F_{\mathrm{st}}$ confirmed the ability of this method in selecting discriminating markers. The reduced SNP panel captured a large proportion of genetic variation between the dairy sheep breeds with estimates of $F_{\mathrm{st}}$ exceeding those previously reported using microsatellites (Tolone et al., 2012) and SNPs (Mastrangelo et al., 2014). Moreover, a previous study on Sicilian sheep breeds (Tolone et al., 2012), using a set of 20 microsatellites, reported that the Bayesian assignment test showed a low assignment value for these breeds, and the low robustness of the assignment test prevented its use for traceability purposes.

Validation analyses will be conducted on the identified SNPs using a wider sample of individuals and other laboratory assay, e.g. Sanger sequencing. Finally, a multiplexed genotyping-by-sequence assay will be developed highlighting the economic advantage on of reduced SNP panels, compared to dense genome-wide assay, for routine use in the management of local populations.

## 2.5 Conclusion

Results for assignment test using the mixed strategy were interesting, because 100% of the individuals were correctly assigned to their breeds of origin. Using genotypic data, a small set of SNPs was identified. The results laid the basis to improve the existing strategies. Potential uses of the described approach include breed assignment, and tracing the origin of animal products in an industrial setting.

# Chapter 3

# Selecting tuning parameter in lasso regression: a new proposal

## 3.1 Introduction

In the context of high-dimensional data, typically only a small number of variables are truly informative whilst most of them are redundant. Selecting the appropriate variables is a crucial step of the data analysis process. In fact underfitted models excluding truly informative covariates may lead to severe bias of the estimators. On the other hand, overfitting may hinder interpretation and cause large standard errors (Fan and Li, 2001).

Penalized regression methods have gained popularity as a tool to perform variable selection. This approach requires defining a tuning parameter that affects the degree of shrinking to be applied to the model coefficients. Among the different penalized procedures, the least absolute shrinkage and selection operator (lasso, Tibshirani, 1996) appears to be the most widely utilized approach.

The tuning parameter balances the trade-off between model fit and model sparsity, and selecting an appropriate value is fundamental of lasso regression. Traditional selection criteria include simple and generalized cross-validation (respectively CV and GCV, Craven and Wahba, 1979), Akaike information criterion (AIC, Akaike, (1974)), Bayesian information criterion (BIC, Schwarz, 1978), and its extended version (EBIC, Chen and Chen, 2008), the more recent Generalized information criterion (GIC, Zhang, Li, and Tsai, 2010) and stability selection (Meinshausen and Bühlmann, (2010)).

Unfortunately, no method appears to perform systematically better than others. Broadly speaking, it is recognized that AIC is similar to GCV, and both of them tend to select too complex models (Wang, Li, and Tsai, 2007b); the BIC and EBIC are able to identify consistently the true model but in finite samples they typically leave out important covariates. However, many authors proposed to select the tuning parameters through $k$-fold CV, which is also the default option in several R packages.

We propose a new criterion to select the tuning parameter in lasso regression. The criterion is quite simple to compute and can be interpreted as maximization of the signal-to-noise ratio. We show that our tuning parameter selector enables to consistently identify the true model when the true model is among a set of candidate models.

## 3.2 New information/selection criterion

### 3.2.1 Penalized estimators and penalty condition

Consider data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$ where $y_1, \ldots, y_N$ are independent given $\boldsymbol{x}$. $y_i$ is the response from the $i$-th subject, and $\boldsymbol{x}_i$ the associated $p$-dimensional covariate vector with corresponding parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. Assume that $\mathbb{E}(y_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$. Let $\ell(\boldsymbol{\beta}, \phi)$ be the model log-likelihood function depending on the $p$ dimensional regression parameter $\boldsymbol{\beta}$ and the dispersion parameter $\phi$.

In lasso regression, the objective is to minimize the penalized log likelihood for a *given* value of the tuning parameter $\lambda$, that is:

$$\ell_\lambda(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}, \phi) - \lambda \sum_{j=1}^{p} |\beta_j|.$$

In a more general setting $\lambda|\beta_j| = p_\lambda(|\beta_j|)$, where $p_\lambda(\cdot)$ is a function of the penalization term. In the next subsection we propose a new criterion to select the value of the tuning parameter.

### 3.2.2 The proposed criterion

We suggest to select $\lambda$ as the maximizer of average signal to noise ratio (ASNR)

$$\text{ASNR}(\lambda) = \frac{\sum_{j=1}^{p} |\hat{\beta}_{j\lambda}| / \mathrm{d}_\lambda}{\hat{\phi}_\lambda^{1/2}},$$

where $\mathrm{d}_\lambda$ is the model degrees of freedom, namely the cardinality of the active set and $\hat{\phi}_\lambda^{1/2}$ is the square root of the estimated dispersion parameter.

### 3.2.3 Consistency

A penalty term is said to be selection consistent if the probability that the fitted regression model includes only the truly informative variables tends to one as $N$ tends to infinity, and $\lambda$ is replaced by $\lambda_N$ to emphasize its dependence on $N$ in quantifying the asymptotic behaviors. In particular, Zhao and Yu, (2006) showed that the lasso regression is selection consistent under the irrepresentable condition when $\sqrt{N}\lambda_N \to \infty$ and $\lambda_N \to 0$. Although the asymptotic order of $\lambda_N = O(N^{-1/2})$ is known to guarantee selection consistency, it remains unclear how to select $\lambda_N$ in finite sample. Several tuning parameter selection criteria can be employed.

**Definition 1** (Candidate model)**.** *We define $S_\lambda$ candidate model, i.e., a subset of the full model $S = \{1, \dots, p\}$. We denote the size of the model $S_\lambda$, i.e., the number of nonzero parameters in $S_\lambda$, by $d_\lambda$ and the corresponding parameter estimate by $\hat{\boldsymbol{\beta}}_\lambda$. Moreover, we denote the collection of all candidate models by $\mathcal{A}$.*

We assume that the set of candidate models contains the unique true model, and that the number of parameters in the full model is finite. Under this assumption, we are able to study the asymptotic consistency of ASNR.

**Definition 2** (Underfitted and Overfitted Models)**.** *We assume that there is a unique true model $S_0$ in $\mathcal{A}$, whose corresponding coefficients are nonzero. Therefore, any candidate model $S_\lambda \not\supset S_0$ is referred to an underfitted model, while any $S_\lambda \supset S_0$ other than $S_0$ itself is referred to as an overfitted model.*

Based on the above definitions, we partition the tuning parameter interval $[\lambda_{\min}, \lambda_{\max}]$ into the underfitted, true and overfitted subset, respectively, so that:

$$\begin{aligned}
\Lambda_- &= \{\lambda : S_\lambda \not\supset S_0\} \Rightarrow \lambda \in [\lambda_{\min}, \lambda_0), \\
\Lambda_0 &= \{\lambda : S_\lambda = S_0\} \Rightarrow \lambda = \lambda_0, \\
\Lambda_+ &= \{\lambda : S_\lambda \supset S_0, \text{ and } S_\lambda \neq S_0\} \Rightarrow \lambda \in (\lambda_0, \lambda_{\max}].
\end{aligned}$$

This partition allows us to assess the performance of regularization parameter selections.

To investigate the asymptotic properties of the regularization parameter selectors, the two linear sparsity conditions are needed (Wainwright, 2009a; Wainwright, 2009b; Reeves and Gastpar, 2013; Su, Bogdan, and Candes, 2015). The first, concerns the effect sizes $\mathrm{d}_0$, that is the degree of sparsity. The second concerns the effect sizes of all coefficients of the true model $S_0$. In particular:

A. suppose that the ration $N/p \to \delta > 0$, then $\mathrm{d}_0 < N/(2\log p)$;

B. the *beta-min* condition, $\mid \beta^0 \mid_{\min} = \min_{j \in S_0} \mid \beta_j^0 \mid \geq c \cdot \sigma\sqrt{2\log p}$, where $c$ is an unkown numerical constant (which would have to exceed one).

We show that, for any $\lambda$ which can not identify the true model, the resulting ASNR($\lambda_0$) is consistently larger than ASNR($\lambda$). To this end we consider two cases, underfitting and overfitting.

**Conjecture 1.** *Suppose all regularity and technical conditions hold. We have to prove that, if $\exists\ \lambda_0 \in \Lambda_0$ which identify the true model $S_0$, then:*

*(1) if $\lambda \in \Lambda_- \Rightarrow P\{\inf_{\lambda \in \Lambda_-} \text{ASNR}(\lambda_0) > \text{ASNR}(\lambda)\} \to 1$;*

*(2) if $\lambda \in \Lambda_+ \Rightarrow P\{\inf_{\lambda \in \Lambda_+} \text{ASNR}(\lambda_0) > \text{ASNR}(\lambda)\} \to 1$.*

Conjecture 1 provides guidance on the choice of the regularization parameter. Conjecture 1(A) and 1(B) imply that the ASNR selector, if all conditions are fulfilled, identify the true model consistently.

**Remark 1.** *Only the simulations evidence suggest the validity of Conjecture 1.*

## 3.3 Numerical Studies

In this section, we present a simulation study which incorporates a variety of scenarios. All scenarios consider a setting in which the true model is in the set of candidate models with different noise levels, both in low and high dimensionality.

We consider a linear regression model $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$, where $\boldsymbol{x}_i$ are identically and independently distributed multivariate normal random variables. The sample sizes are $N = 50, 100, 200$, the ratios $p/N = 0.5, 1.2, 2$ and $\epsilon_i$ are identically and independently distributed $\mathcal{N}(0, \sigma^2)$ with $\sigma = 1, 3$. The entire simulation is repeated under all combinations of $N$, $p/N$, and $\sigma$, and two different scenarios: in the first, covariates are not correlated with each other, while in the second, the Toepliz matrix is used to define the correlation between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is $0.5^{|i-j|}$ with $i, j = 1, \ldots, p$ and $i \neq j$.

To investigate the performance of the proposed method we fix the parameter structure $\beta_0 = (3, -3.1, 4, 3.5, -5, 0, \ldots, 0)^T$. In both scenarios when $\sigma = 1$ the conditions (1) and (2) are fulfilled, whilst they are not when $\sigma = 3$.

We simulate 500 data sets, keeping fixed the model matrix $\boldsymbol{X}$, and $\lambda$ is chosen by ASNR, AIC, BIC, EBIC, GCV, GICand 5-fold CV. Their formulations are given as follows,

$$\begin{aligned}
\text{AIC} &= \log(\hat{\phi}) + 2\mathrm{d}_\lambda N^{-1} \\
\text{BIC} &= \log(\hat{\phi}) + \log(n)\mathrm{d}_\lambda N^{-1} \\
\text{EBIC} &= \log(\hat{\phi}) + \left(\log(N) + 2\gamma\log(p)\right)\mathrm{d}_\lambda N^{-1} \\
\text{GCV} &= \hat{\phi}/\left((1-\mathrm{d}_\lambda N^{-1})^2\right) \\
\text{GIC} &= \log(\hat{\phi}) + \mathrm{c}_N \log(p)\mathrm{d}_\lambda N^{-1} \\
\text{CV} &= \sum_{s=1}^{5} \sum_{(y_k,\boldsymbol{X}_k)\in T^{-s}} \left(y_k - \boldsymbol{X}_k^T\hat{\boldsymbol{\beta}}^{(s)}(\lambda)\right)^2
\end{aligned}$$

where $\hat{\phi}$ is the estimated dispersion parameter, $\gamma$ is a non negative parameter and $\mathrm{c}_N$ is a parameter which depends on $N$. In 5-fold CV, $T^s$ and $T^{-s}$ are the training and validation sets, and $\hat{\boldsymbol{\beta}}^{(s)}(\lambda)$ is the estimated vector of regression coefficients using the training set $T^s$ and the tuning parameter $\lambda$. In our simulation both parameters are fixed to $0.5$ and $\log(\log(N))$, respectively.

The performance of AIC and GCV was very similar and results are not reported. Also, only the EBIC (and not the simple BIC) is reported. We report the average number of non-zero coefficients and the true and false positive rates (TPR and FPR) as in Su, Bogdan, and Candes, (2015).

Let be V and T the number of lasso false and true discoveries, respectively. We denote with $\mathrm{d}_0 = |\{j : \beta_j \neq 0\}|$ the number of true non-zero coefficients, with $V(\lambda) = |\{j : \hat{\beta}_j(\lambda) \neq 0 \text{ and } \beta_j = 0\}|$ and with $T(\lambda) = |\{j : \hat{\beta}_j(\lambda) \neq 0 \text{ and } \beta_j \neq 0\}|$. Finally, we define

$$\text{FPR}(\lambda) = \frac{V(\lambda)}{|\{j : \hat{\beta}_j(\lambda) \neq 0\}| \vee 1}, \qquad \text{TPR}(\lambda) = \frac{T(\lambda)}{\mathrm{d}_0 \vee 1},$$

where $a \vee b = \max\{a, b\}$. Moreover, we could be interested in the false discovery and false negative retes. The FPR is a natural measure of type I error while 1-TPR (namely the false negative rate, FNR) is the fraction of missed coefficients, that is a natural notion of type II error.

Table 3.1 shows a comparison between our proposal and the three main competitor, EBIC, GIC and 5-fold CV, in selecting the tuning parameter in a scenario in which no correlation structure is considered between the covariates and in which the *beta-min* condition is fulfilled ($\sigma = 1$) and not fulfilled ($\sigma = 3$). In this scenario all criteria are able to identify all the non-zero coefficients, i.e., the TPR is about 1. When all regularity conditions are fulfilled our proposal shows excellent results. It correctly selects the true active set already with very low sample size, i.e., $N = 50$. Furthermore, when at least one regularity condition is missing, our method requires large sample size to reach a good performance. Indeed, when the sample size is very low our criterion tends to select more parameters than the other criteria do. In general, our proposal ASNR shows good results in this setting committing a negligible type I error and a null type II error also in a diverging scenario, selecting the correct number of non-zero coefficients already with $N = 50$.

Table 3.2 shows a comparison between our proposal and the three main competitor, EBIC, GIC and 5-fold CV, in selecting the tuning parameter in a scenario in which a Toeplitz correlation structure is considered between the covariates and in which once again the *beta-min* condition is fulfilled ($\sigma = 1$) and not fulfilled ($\sigma = 3$). In this scenario, when the regularity conditions are fulfilled, the same considerations hold for all criteria. However, when the magnitude of the betas is lower than the noise level with correlated covariates some difficulties arise in identifying the true non-zero parameters. Only the CV criterion is able to include the true active set in a larger set of non-null coefficients, at the expense of a large number of degrees of freedom. The other criteria, especially when $N = 50, 100$, have difficulties in identifying all the non-zero coefficients committing a large FNR, that is a large type II error. Our proposal seems more conservative in selecting parameters than the others, indeed the other criteria tends to select overparametrized model. Finally, when $N = 200$, our proposal reaches good performance also in a hard setting.

Concluding, the increment in the noise level given by $\sigma$ in small sample size, shows difficulties in the identification of the true non-zero coefficients and in keeping under control the type I and II error measures. However, as the sample size increases our criterion provides excellent results in terms of both measures FPR and FNR, and also in number of non-zero coefficients identified. All the results in both scenarios, with and without correlation between the covariates, are satisfactory. We remark that, if a correlation structure is present, our criterion fails to identify the true non-zero coefficients. Instead, a high power to identify the zero coefficients is maintained. In terms of model specification our method tends to be more parsimonious than the competitors, which tends to create overparametrized models..

TABLE 3.1: Tuning parameter selection comparison between different criteria, ASNR, EBIC, GIC and 5-fold CV. In this simulation for each $N$ and $p/N$ ratio the averages of false positive ratio (FPR), the true positive ratio (TPR) and the number of non-zero coefficients ($\mathrm{d}_\lambda$) measures are reported. We also reported in the two block columns the same measures for two $\sigma$ values, in a scenario in which no correlation structure between covariates is considered.

| $N$ | $p$ | | $\sigma = 1$ | | | | $\sigma = 3$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | ASNR | EBIC | GIC | CV | ASNR | EBIC | GIC | CV |
| 50 | 25 | FPR | 0.124 | 0.201 | 0.299 | 0.463 | 0.209 | 0.189 | 0.296 | 0.581 |
| | | TPR | 1.000 | 1.000 | 1.000 | 1.000 | 0.990 | 0.993 | 0.999 | 1.000 |
| | | $\mathrm{d}_\lambda$ | 5.86 | 6.49 | 7.53 | 10.37 | 6.76 | 6.34 | 7.50 | 12.89 |
| | 40 | FPR | 0.054 | 0.169 | 0.246 | 0.524 | 0.255 | 0.167 | 0.244 | 0.651 |
| | | TPR | 1.000 | 1.000 | 1.000 | 1.000 | 0.998 | 0.993 | 1.000 | 1.000 |
| | | $\mathrm{d}_\lambda$ | 5.34 | 6.20 | 6.98 | 12.02 | 9.30 | 6.15 | 6.97 | 16.56 |
| | 100 | FPR | 0.111 | 0.169 | 0.243 | 0.607 | 0.413 | 0.161 | 0.239 | 0.719 |
| | | TPR | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.984 | 0.999 | 1.000 |
| | | $\mathrm{d}_\lambda$ | 5.78 | 6.23 | 6.98 | 14.36 | 16.75 | 6.07 | 6.94 | 20.32 |
| 100 | 50 | FPR | 0.036 | 0.195 | 0.260 | 0.667 | 0.096 | 0.196 | 0.257 | 0.673 |
| | | TPR | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | | $\mathrm{d}_\lambda$ | 5.23 | 6.46 | 7.16 | 16.95 | 5.64 | 6.46 | 7.12 | 17.30 |
| | 120 | FPR | 0.108 | 0.241 | 0.287 | 0.749 | 0.178 | 0.239 | 0.291 | 0.755 |
| | | TPR | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | | $\mathrm{d}_\lambda$ | 5.71 | 6.84 | 7.39 | 22.94 | 6.32 | 6.83 | 7.43 | 23.80 |
| | 200 | FPR | 0.068 | 0.219 | 0.267 | 0.806 | 0.182 | 0.216 | 0.259 | 0.812 |
| | | TPR | 1.000 | 1.000 | 1.000 | 1.000 | 0.999 | 0.999 | 1.000 | 1.000 |
| | | $\mathrm{d}_\lambda$ | 5.45 | 6.68 | 7.20 | 29.16 | 6.57 | 6.66 | 7.11 | 30.77 |
| 200 | 100 | FPR | 0.003 | 0.116 | 0.153 | 0.701 | 0.010 | 0.113 | 0.149 | 0.700 |
| | | TPR | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | | $\mathrm{d}_\lambda$ | 5.02 | 5.79 | 6.10 | 19.61 | 5.06 | 5.77 | 6.07 | 19.55 |
| | 240 | FPR | 0.005 | 0.105 | 0.126 | 0.755 | 0.016 | 0.108 | 0.130 | 0.755 |
| | | TPR | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | | $\mathrm{d}_\lambda$ | 5.03 | 5.72 | 5.88 | 25.74 | 5.10 | 5.74 | 5.91 | 25.89 |
| | 400 | FPR | 0.003 | 0.092 | 0.106 | 0.769 | 0.010 | 0.093 | 0.107 | 0.771 |
| | | TPR | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | | $\mathrm{d}_\lambda$ | 5.02 | 5.60 | 5.71 | 28.76 | 5.06 | 5.61 | 5.71 | 29.09 |

TABLE 3.2: It is identical to Table 3.1, but there is a non-zero correlation structure between covariates

| $N$ | $p$ | | $\sigma = 1$ | | | | $\sigma = 3$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | ASNR | EBIC | GIC | CV | ASNR | EBIC | GIC | CV |
| 50 | 25 | FPR | 0.239 | 0.300 | 0.398 | 0.575 | 0.451 | 0.267 | 0.392 | 0.601 |
| | | TPR | 1.000 | 1.000 | 1.000 | 1.000 | 0.999 | 0.928 | 0.997 | 1.000 |
| | | $d_\lambda$ | 6.95 | 7.48 | 8.85 | 12.59 | 11.03 | 6.79 | 8.75 | 13.38 |
| | 40 | FPR | 0.352 | 0.370 | 0.445 | 0.691 | 0.579 | 0.123 | 0.269 | 0.745 |
| | | TPR | 1.000 | 1.000 | 1.000 | 1.000 | 0.895 | 0.615 | 0.780 | 1.000 |
| | | $d_\lambda$ | 8.50 | 8.35 | 9.69 | 17.36 | 22.33 | 3.83 | 6.23 | 21.42 |
| | 100 | FPR | 0.286 | 0.307 | 0.370 | 0.663 | 0.618 | 0.231 | 0.344 | 0.733 |
| | | TPR | 1.000 | 1.000 | 1.000 | 1.000 | 0.987 | 0.856 | 0.956 | 1.000 |
| | | $d_\lambda$ | 7.53 | 7.57 | 8.42 | 16.36 | 24.55 | 6.04 | 7.86 | 20.85 |
| 100 | 50 | FPR | 0.213 | 0.353 | 0.425 | 0.743 | 0.154 | 0.238 | 0.369 | 0.743 |
| | | TPR | 1.000 | 1.000 | 1.000 | 1.000 | 0.734 | 0.853 | 0.946 | 1.000 |
| | | $d_\lambda$ | 6.60 | 8.16 | 9.25 | 20.81 | 5.07 | 6.16 | 8.28 | 20.81 |
| | 120 | FPR | 0.302 | 0.406 | 0.460 | 0.830 | 0.073 | 0.113 | 0.158 | 0.817 |
| | | TPR | 1.000 | 1.000 | 1.000 | 1.000 | 0.797 | 0.800 | 0.805 | 0.985 |
| | | $d_\lambda$ | 7.58 | 8.88 | 9.89 | 31.45 | 4.81 | 4.65 | 5.01 | 30.85 |
| | 200 | FPR | 0.193 | 0.315 | 0.362 | 0.823 | 0.172 | 0.198 | 0.267 | 0.826 |
| | | TPR | 1.000 | 1.000 | 1.000 | 1.000 | 0.794 | 0.840 | 0.892 | 1.000 |
| | | $d_\lambda$ | 6.43 | 7.65 | 8.28 | 30.97 | 7.25 | 5.66 | 6.64 | 32.10 |
| 200 | 100 | FPR | 0.031 | 0.227 | 0.297 | 0.785 | 0.087 | 0.226 | 0.303 | 0.785 |
| | | TPR | 1.000 | 1.000 | 1.000 | 1.000 | 0.997 | 1.000 | 1.000 | 1.000 |
| | | $d_\lambda$ | 5.19 | 6.77 | 7.53 | 25.78 | 5.56 | 6.76 | 7.59 | 25.81 |
| | 240 | FPR | 0.036 | 0.233 | 0.271 | 0.839 | 0.072 | 0.232 | 0.269 | 0.838 |
| | | TPR | 1.000 | 1.000 | 1.000 | 1.000 | 0.959 | 0.998 | 0.999 | 1.000 |
| | | $d_\lambda$ | 5.23 | 6.79 | 7.23 | 35.16 | 5.27 | 6.79 | 7.21 | 35.24 |
| | 400 | FPR | 0.075 | 0.300 | 0.329 | 0.874 | 0.118 | 0.288 | 0.322 | 0.874 |
| | | TPR | 1.000 | 1.000 | 1.000 | 1.000 | 0.934 | 0.992 | 0.996 | 1.000 |
| | | $d_\lambda$ | 5.50 | 7.49 | 7.88 | 44.74 | 5.52 | 7.36 | 7.79 | 44.79 |

## 3.4 Real Data Studies

### 3.4.1 Riboflavin data set

The first real example is the Riboflavin data by Bacillus subtilis (Bühlmann, Kalisch, and Meier, 2014). Riboflavin is an essential micronutrient in the human diet. The data set contains $N = 71$ observations and $p = 4088$ covariates (logarithm of the gene expression level) and a one-dimensional response, which is log-transformed riboflavin production rate (q_RIBFLV).

To assess the performance of all regularization parameter selectors we decided to split the data set in a training set (ts) and a validation set (vs). The training set, containing $N_{\text{ts}} = 53$, that is 75% of the units, is used to estimate the coefficients. The validation set, containing the remaining units ($N_{\text{vs}} = 18$) is not used in estimating the coefficients, and is used to calculate the mean prediction error ($\bar{\text{PE}} = (y_{\text{vs}} - \boldsymbol{x}_{\text{vs}}^T \hat{\boldsymbol{\beta}}_{\text{ts}})^2 / N_{\text{vs}}$). This measure is obtained using a linear model in which only the intercept and the selected covariates are included.

Table 3.3 reports the number of nonzero coefficients, the tuning parameter selected and the mean prediction error for each criteria.

TABLE 3.3: Tuning parameter selection of the riboflavin data set. Different criteria are used ASNR, EBIC, GIC and 5-fold CV. The number of nonzero coefficients, the tuning parameter selected ($\lambda$) and the mean prediction error are reported.

| | # of Nonzero Coefficients | $\lambda$ | Prediction Error |
|---|---|---|---|
| ASNR | 44 | 0.0131 | 0.1719 |
| EBIC | 1 | 0.4917 | 1.3714 |
| GIC | 1 | 0.4917 | 1.3714 |
| CV | 30 | 0.0579 | 0.2217 |

### 3.4.2 Prostate Cancer data set

The second real example is the well-known Prostate Cancer data from a study of Stamey et al. 1989 on prostate cancer, measuring the correlation between the level of a prostate-specific antigen and some covariates. The covariates are $x_1$ = lcavol (log-cancer volume), $x_2$ = lweight (log-prostate weight), $x_3$ = age (age of patient), $x_4$ = lbhp (log-amount of benign hyperplasia), $x_5$ = svi (seminal vesicle invasion), $x_6$ = lcp (log-capsular penetration), $x_7$ = gleason (Gleason Score), $x_8$ = pgg45 (percent of Gleason scores 4 or 5) and response variable $y$ = lpsa (log-psa). The data set consists of $N = 97$ observations and $p = 8$ covariates. To assess the performance of all regularization parameter selectors, the same procedure as for the riboflavin data set is applied ($N_{\text{ts}} = 73$, $N_{\text{vs}} = 24$).

Table 3.4 reports the number of nonzero coefficients, the tuning parameter selected and the mean prediction error measure for each criterion. It is possible to see that ASNR and EBIC selected the same tuning parameter identifying, as already known in literature, the three non-zero covariates lcavol, lweight and svi.

TABLE 3.4: Tuning parameter selection of the Prostate Cancer data set. Different criteria are used ASNR, EBIC, GIC and 5-fold CV. The number of nonzero coefficients, the tuning parameter selected ($\lambda$) and the mean prediction error are reported. An intercept is added to the model.

| | **# of Nonzero Coefficients** | $\boldsymbol{\lambda}$ | **Prediction Error** |
|---|---|---|---|
| ASNR | 3 | 0.1096 | 0.4765 |
| EBIC | 3 | 0.1096 | 0.4765 |
| GIC | 6 | 0.0171 | 0.5468 |
| CV | 8 | 0.0032 | 0.5367 |

## 3.5 Discussion

In the context of variable selection, we propose a new information criterion to choose regularization parameter. Furthermore, we study the theoretical properties of ASNR. If we believe that the true model is contained in a set of candidate models with the generalized linear model structure, then our selector identifies the true model consistently, while the other criteria tend to overfit. Simulation studies and empirical examples support the performance of the selection criteria.

Even if the theoretical property of ASNR is not yet formally proven, the empirical results suggest the potential of it. Moreover, our proposal can be extended to generalized linear models, e.g., Poisson and logistic regression. Application in very high-dimensional settings ($N \ll p$) that are today one of the most challenging concerns, represents a noteworthy application to be investigated.

# Chapter 4

# Penalized Quantile Regression Coefficients Modeling

## 4.1 Introduction

Conditional quantiles fully describe the conditional distribution of a response variable given covariates. Quantile regression (QR; Koenker and Bassett Jr, 1978) and its generalizations (e.g., Chaudhuri, 1991) are useful tools in modeling quantiles. The conditional quantile function is usually assumed to have a linear specification of the form

$$Q(p \mid \boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta}(p), \tag{4.1}$$

where $\boldsymbol{x}$ is a $q$-dimensional vector of covariates, and $\boldsymbol{\beta}(p)$ is a vector of unknown coefficients describing the relationship between $\boldsymbol{x}$ and $p$-th quantile of the response variable, $p \in (0, 1)$.

Standard quantile regression estimates different quantiles one at the time. When a grid of quantiles is estimated, e.g., $p = 0.01, 0.02, \ldots, 0.99$, results can be summarized graphically. The estimated coefficients are generally non-smooth functions of $p$ and may suffer from high volatility, which can make their interpretation not simple.

Recently, Frumento and Bottai, (2016) suggested modeling the quantile regression coefficient functions, $\boldsymbol{\beta}(p)$, as parametric functions of the order of the quantile:

$$Q(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}^T \boldsymbol{\beta}(p \mid \boldsymbol{\theta}), \tag{4.2}$$

where $\boldsymbol{\theta}$ is a vector of model parameters. This approach is referred to as *quantile regression coefficients modeling* (QRCM) and permits modeling the entire quantile function, while keeping the quantile regression structure expressed by equation (4.1). This modeling approach facilitates estimation, inference, and interpretation of the results, and is generally more efficient than standard quantile regression. Sometimes, it may be useful to assume flexible, high-dimensional models. Consider, for example, describing $\boldsymbol{\beta}(p \mid \boldsymbol{\theta})$ by $k$-th degree polynomial functions:

$$\beta_j(p \mid \boldsymbol{\theta}) = \theta_{j0} + \theta_{j1}p + \ldots + \theta_{jk}p^k, j = 1, \ldots, q.$$

Each covariate has $(k+1)$ associated parameters, for a total of $q \times (k+1)$ model coefficients. When $q$ and $k$ are large, estimation may become difficult and sampling variability large. The model can be simplified by restricting some of the parameters to be equal to zero. For instance, some of the $\beta_j(p \mid \boldsymbol{\theta})$ may be assumed to be linear functions of $p$, or not to depend on $p$ at all; and some covariates may be assumed conditionally independent of the response given the remaining covariates by imposing the constraint $\beta_j(p \mid \boldsymbol{\theta}) = 0$.

Numerous papers (e.g., Belloni and Chernozhukov, 2011; Wang, Wu, and Li, 2012; Wu and Liu, 2009; Zheng, Gallagher, and Kulasekera, 2013) have investigated the estimation of penalized quantile regression models in high-dimensional setting. The penalization is usually given by the $L_1$-norm of the coefficients, denoted by $L_1$-QR (Belloni and Chernozhukov, 2011; Li and Zhu, 2008). These approaches, however, focus on model selection when estimating one quantile at a time. Generally, this is inefficient and makes it difficult to interpret the results. The main advantage of adopting the QRCM framework is that of performing model selection directly on the parameters of the conditional quantile function.

We propose applying the $L_1$-penalty to the integrated loss function described by Frumento and Bottai, (2016), which is minimized to estimate the unknown parameter $\boldsymbol{\theta}$ in model (4.2). We refer to this procedure as *penalized quantile regression coefficients modeling* (QRCMPEN).

## 4.2 Motivating example

The motivation for this methods comes from a study on the association between pulmonary inspiratory capacity, a measure of lung's volume, and the following nine predictors: age, height, body mass index (BMI), a binary indicator of smoking, and indicators for sex, occupation exposure, cough, wheezing and asthma. Extreme quantiles of inspiratory capacity, and in particular very low quantiles, can be used to identify health problems and implement therapy.

We first estimated thirteen percentiles $(0.01, 0.05, 0.10, \ldots, 0.95, 0.99)$ with standard penalized quantile regression, as implemented in the R package rqPen. Each variable was standardized to have zero mean and unit standard deviation. Results are summarized in Table 4.1.

The coefficients associated with strong predictors, like age, height and bmi, were consistently and significantly positive or negative at nearly all quantiles. Other coefficients, however, were only significant at some quantiles, making it difficult to interpret the results. All coefficients showed volatility, for example, the coefficient of smoking was not significant at $p = 0.5$ and $p = 0.7$, but significantly greater than zero at $p = 0.6$ and $p = 0.8$. The observed volatility is probably accounted for by random variation and data sparsity, which represents a main source of sampling error in the tails of the distribution.

The idea behind QRCM is to use a more parsimonious representation of $\boldsymbol{\beta}(p)$. Consider, for example, Figure 4.1, showing the estimated regression coefficients associated with age. The underlying trend appears to be adequately described by a linear function, $\beta_{\text{age}}(p \mid \boldsymbol{\theta}) = \theta_0 + \theta_1 p$.

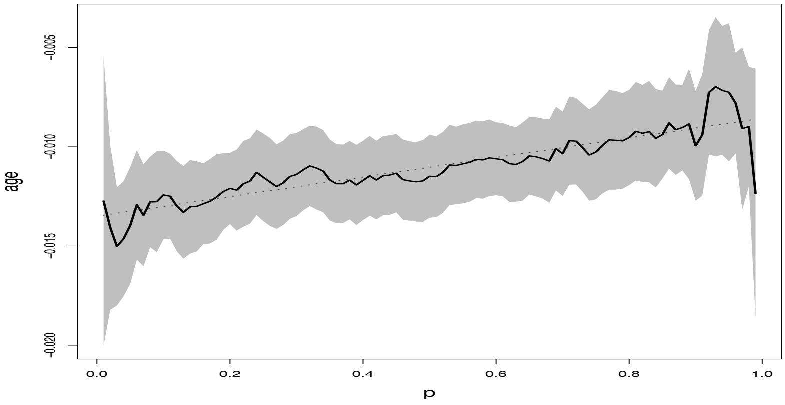


FIGURE 4.1: Estimated quantile regression coefficients associated with age and 95% pointwise confidence intervals (shaded area). The dashed line indicates the linear trend.

Writing $\boldsymbol{\beta}(p) = \boldsymbol{\beta}(p \mid \boldsymbol{\theta})$ and directly estimating $\boldsymbol{\theta}$ permit estimating the entire conditional quantile function. It also enables performing model selection using information on all quantiles simultaneously. As shown later, this has important consequences on the performance of penalized regression methods.

TABLE 4.1: Results of the variable selection using standard penalized quantile regression. The symbol "✓" indicates that the covariate was included in the model. We used a grid of percentiles, namely $p = (0.01, 0.05, 0.10, \ldots, 0.90, 0.95, 0.99)$, and utilized BIC for model selection.

| | **0.01** | **0.05** | **0.10** | **0.20** | **0.30** | **0.40** | **0.50** | **0.60** | **0.70** | **0.80** | **0.90** | **0.95** | **0.99** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Intercept | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Height | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| BMI | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | - | ✓ |
| Male | - | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Smoker | ✓ | ✓ | - | ✓ | - | - | - | ✓ | - | ✓ | ✓ | ✓ | ✓ |
| Non-exposure | ✓ | ✓ | - | - | - | - | - | - | ✓ | - | ✓ | ✓ | ✓ |
| Cough | ✓ | ✓ | - | ✓ | ✓ | ✓ | - | ✓ | - | - | ✓ | ✓ | - |
| Wheezing | - | ✓ | ✓ | ✓ | ✓ | ✓ | - | - | - | - | ✓ | ✓ | ✓ |
| Asthma | ✓ | ✓ | ✓ | - | - | - | ✓ | - | - | - | ✓ | - | ✓ |

## 4.3 The estimator

Throughout, we assume that model (4.2) hold, and adopt the following parametrization:

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

where $\boldsymbol{b}(p) = [b_1(p), \ldots, b_k(p)]^T$ is a set of $k$ known functions of $p$, and $\boldsymbol{\theta}$ is a $q \times k$ matrix with entries $\theta_{jh}$ such that $\beta_j(p \mid \boldsymbol{\theta}) = \theta_{j1}b_1(p) + \ldots + \theta_{jk}b_k(p)$, $j = 1, \ldots, q$. The conditional quantile function is

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

As shown by Frumento and Bottai, (2016), estimation is carried out by minimizing

$$\overline{L}(\boldsymbol{\theta}) = \int_0^1 L(\boldsymbol{\beta}(p \mid \boldsymbol{\theta}))\mathrm{d}p, \tag{4.3}$$

where $L(\boldsymbol{\beta}(p))$ is the loss function of standard quantile regression given by

$$L = \sum_{i=1}^{n} (p - I(y_i \leq \boldsymbol{x}_i^T\boldsymbol{\beta}(p)))(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}(p)).$$

This estimation procedure is referred to as *integrated loss minimization* (ILM), and implemented in the `qrcm` package in R. The model is determined by the choice of $\boldsymbol{b}(p)$. When the model is not known, an intuitive approach is to define $\boldsymbol{b}(p)$ as a sufficiently large collection of functions, e.g., $\boldsymbol{b}(p) = \left[1, p, p^2, p^3, \sqrt{p}, \log(p), \log(1-p), \ldots\right]^T$. This modeling approach is very flexible, and usually provides a good fit of the data. However, it tends to generate large models, causing overparametrization and loss of efficiency.

To implement an automatic procedure for model selection, we propose to modify the loss function (4.3) by introducing a $L_1$-norm penalizing factor:

$$\overline{L}_{\text{PEN}}^{(\lambda)}(\boldsymbol{\theta}) = \int_0^1 L(\boldsymbol{\beta}(p \mid \boldsymbol{\theta})) + \lambda \sum |\boldsymbol{\theta}|\mathrm{d}p. \tag{4.4}$$

We refer to this estimation approach as *penalized integrated loss minimization* (PILM). To minimize $L_{\text{PEN}}^{(\lambda)}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, we use a pathwise coordinate descent algorithm (Friedman, Hastie, and Tibshirani, 2010). The described PILM estimator has been implemented in the `qrcmNP` package in R.

## 4.4 Tuning parameter selection

With a given set of data, the true model is not known. Having adequate criteria for model selection is therefore crucial. In penalized regression, the tuning parameter $\lambda$ balances the trade-off between goodness of fit and efficiency.

We denote by $\hat{\boldsymbol{\theta}}^{(\lambda)}$ the estimator of $\boldsymbol{\theta}$ obtained by minimizing (4.4) at a given value of $\lambda$. AIC-type and BIC-type selectors are grid-search criteria that minimize

$$\text{Dev}^{(\lambda)} + c_n \cdot \text{df}^{(\lambda)},$$

where $\text{Dev}^{(\lambda)}$ is the explained deviance of the model (a measure of goodness-of-fit defined below) corresponding to $\hat{\boldsymbol{\theta}}^{(\lambda)}$, $c_n$ is a constant that could depend on the sample size $n$, and $\text{df}^{(\lambda)}$ reflects the number of nonzero elements of $\hat{\boldsymbol{\theta}}^{(\lambda)}$.

Considering that each of the $\boldsymbol{\beta}_j(p \mid \boldsymbol{\theta})$ has up to $k$ associated parameters, where $k$ is the dimension of $\boldsymbol{b}(p)$, we suggest defining

$$\text{df}^{(\lambda)} = \sum_{j=1}^{q} \text{df}_j^{(\lambda)},$$

where

$$\text{df}_j^{(\lambda)} = k^{-1} \sum_{h=1}^{k} I(\hat{\boldsymbol{\theta}}_{jh}^{(\lambda)} \neq 0),\ j = 1, \ldots, q.$$

Note that $\{\theta_{j1}, \ldots, \theta_{jk}\}$ is the subset of model parameters that contribute to $\beta_j(p \mid \boldsymbol{\theta})$. In particular, $\text{df}_j^{(\lambda)} = 1$ when all elements of $\boldsymbol{b}(p)$ are used to build $\beta_j(p \mid \boldsymbol{\theta})$; and $\text{df}_j^{(\lambda)} = 0$ when all parameters associated with $\beta_j(p \mid \boldsymbol{\theta})$ are shrunk to zero. By using this definition, we attribute one degree of freedom to each quantile regression coefficient $\beta_j(p \mid \boldsymbol{\theta})$, for a total of $q$ (and not $q \times k$) degrees of freedom.

To improve efficiency and computation, we propose standardizing both $\boldsymbol{x}$ and $\boldsymbol{b}(p)$. In our simulation study and data analysis, we followed Lee, Noh, and Park, (2014) and defined

$$\text{Dev}^{(\lambda)} = \log \overline{L}_{\text{PEN}}^{(\lambda)}(\hat{\boldsymbol{\theta}}^{(\lambda)}),$$

the logarithm of the minimized loss function given by (4.4). The AIC and BIC criteria are given by

$$\begin{aligned}\mathrm{AIC}^{(\lambda)} &= \log \overline{L}^{(\lambda)}_{\mathrm{PEN}}(\hat{\boldsymbol{\theta}}^{(\lambda)}) + N^{-1}\mathrm{df}^{(\lambda)},\\ \mathrm{BIC}^{(\lambda)} &= \log \overline{L}^{(\lambda)}_{\mathrm{PEN}}(\hat{\boldsymbol{\theta}}^{(\lambda)}) + (2N)^{-1}\log(N)\mathrm{df}^{(\lambda)}C_N\ .\end{aligned}$$

where $C_n$ is some positive constant, that diverges to infinity as $n$ increase. If $C_n = 1$ it corresponds to the ordinary BIC. Unfortunately, the new criterion ASNR, as discussed in Chapter 3, has not been investigated here.

## 4.5 Simulations

To evaluate empirically the finite-sample properties of the proposed estimator, we considered three different simulation scenarios in which the quantile function was:

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

where $x_1, x_2, \ldots, x_q$ were independent $U(0,5)$ variables. The three simulation scenarios are described below.

*Simulation 1*. We used $q = 3$ covariates and defined

$$Q(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = (1 + \log(p) - .5\log(1-p)) + x_1 + (1 + 1.5p)x_2 + x_3,$$

$$\boldsymbol{b}(p) = [1, \log(p), \log(1-p), p, p^2, p^3]^T.$$

*Simulation 2*. We used $q = 15$ covariates and defined

$$Q(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = (1+\log(p) - .5\log(1-p)) + (1+\sqrt{p})x_1 + (1+1.5p+p^2)x_2 + x_3,$$

$$\boldsymbol{b}(p) = [1, \log(p), \log(1-p), p, p^2, p^3, \sqrt{p}]^T.$$

*Simulation 3*. We used $q = 15$ covariates and defined

$$Q(p \mid \boldsymbol{x}, \boldsymbol{\theta}) = (1+\log(p) - .5\log(1-p)) + (1+1.5\sqrt{p}+\sqrt[3]{p})x_1 + (1+\sqrt{p})x_2 + (1+2\sqrt[4]{p})x_3,$$

$$\boldsymbol{b}(p) = [1, \log(p), \log(1-p), \mathcal{P}_3(p)]^T,$$

where $\mathcal{P}_k(p)$ denotes $k$-th degree shifted Legendre polynomials (e.g., Abramowitz and Stegun, 1964), an orthogonal polynomial in $(0,1)$ that is used as flexible model for $\boldsymbol{\beta}(p \mid \boldsymbol{\theta})$.

In all simulations, we modeled the intercept using $\log(p)$ and $\log(1-p)$, that define the asymmetric Logistic distribution, while the maximal model for $\beta_1(p \mid \boldsymbol{\theta}), \dots, \beta_q(p \mid \boldsymbol{\theta})$ included all the other entries of $\boldsymbol{b}(p)$. In simulation 3, the specification of $\boldsymbol{b}(p)$ did not correspond to the true model. This allowed assessing the performance of the described estimator when the true data-generating process is not known and a flexible parametrization is used to approximate $\boldsymbol{\beta}(p \mid \boldsymbol{\theta})$. For each scenario we generated $B = 500$ simulated quantile functions, keeping fixed the model matrix $\boldsymbol{x}$. We applied the described PILM estimator and selected the model according to AIC and BIC, separately. We also applied the unpenalized ILM estimator, fitting the true model.

We measured bias as follows. Let $\widehat{\boldsymbol{\theta}}_b$ indicate the parameters' estimates in the $b$-th simulated dataset, $b = 1, \dots, B$, and define

$$\bar{\boldsymbol{\theta}} = B^{-1} \sum_{b=1}^{B} \widehat{\boldsymbol{\theta}}_b.$$

This quantity estimates the expected value of $\widehat{\boldsymbol{\theta}}$. For a given value $\boldsymbol{x}_i$ of the covariates, we defined the following measures of bias:

$$\text{bias}_{F_{\boldsymbol{x}_i}} = \sup_{p \in (0,1)} |F(Q(p \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \mid \boldsymbol{x}_i, \bar{\boldsymbol{\theta}}) - p|, \tag{4.5}$$

$$\text{bias}_{Q_{\boldsymbol{x}_i}} = \sup_{p \in (0,1)} |Q(p \mid \boldsymbol{x}_i, \boldsymbol{\theta}) - Q(p \mid \boldsymbol{x}_i, \bar{\boldsymbol{\theta}})| \tag{4.6}$$

where $F(\cdot)$ denotes the cumulative distribution function that corresponds to the inverse of $Q(\cdot)$. Expression (4.5) returns a value in $(0, 1)$ and represents a bias on the scale of $F(\cdot)$, while expression (4.6) is on the same scale as $Q(\cdot)$. Aggregated measures of bias are obtained by averaging over the distribution of $\boldsymbol{x}_i$. Letting the covariates' values be the same across all simulated datasets, this corresponds to calculating the following quantities:

$$\text{bias}_F = N^{-1} \sum_{i=1}^{N} \text{bias}_{F_{\boldsymbol{x}_i}}, \tag{4.7}$$

$$\text{bias}_Q = N^{-1} \sum_{i=1}^{N} \text{bias}_{Q_{\boldsymbol{x}_i}}. \tag{4.8}$$

The results of the simulations are summarized in Table 4.2. In scenario 1, where the model was relatively simple, both AIC and BIC selected the correct model. In scenario 2, with more covariates, both criteria appeared to overestimate the number of nonzero parameters. However, the bias was found to be negligible. In scenario 3, the true model was not known and an orthogonal polynomial was used to form a basis. A small bias was found with both AIC and BIC.

TABLE 4.2: Results of simulations 1-3. For each scenario, we report the measures of bias described in (4.7) and (4.8), the degrees of freedom $\mathrm{df}^{(\lambda)}$ as in Section 4.4, and the median number of nonzero model parameters as a fraction of the parameters of the maximal model, using the PILM estimator with AIC and BIC criteria. In the table, ILM denotes the unpenalized estimator in which the model is correctly specified and only the nonzero parameters are computed.

| | | $\mathrm{bias}_{\boldsymbol{F}}$ | $\mathrm{bias}_{\boldsymbol{Q}}$ | $\mathrm{df}^{(\boldsymbol{\lambda})}$ | **n. of parameters** |
|---|---|---|---|---|---|
| | ILM | 0.001 | 0.021 | 2.000 | 7 |
| Sim 1 | AIC | 0.007 | 0.084 | 2.336 | $\simeq 8/15$ |
| | BIC | 0.012 | 0.156 | 2.082 | $\simeq 7/15$ |
| | ILM | 0.039 | 0.002 | 2.200 | 9 |
| Sim 2 | AIC | 0.014 | 0.389 | 5.832 | $\simeq 27/78$ |
| | BIC | 0.028 | 0.692 | 3.525 | $\simeq 16/78$ |
| | ILM | 0.002 | 0.054 | - | 10 |
| Sim 3 | AIC | 0.016 | 0.893 | 8.093 | $\simeq 31/60$ |
| | BIC | 0.038 | 1.247 | 4.231 | $\simeq 13/60$ |

## 4.6 Computation

The described PILM estimator has been implemented in the `qrcmNP` (penalized quantile regression coefficients modeling) package in R. Although computation is easy, there are important issues:

- for a correct application of the penalization described by (4.4), covariates $\boldsymbol{x}$, response variable $y$ and basis $\boldsymbol{b}(p)$ need to be standardized;

- as shown by Frumento and Bottai, (2016), computing $L(\boldsymbol{\beta}(p \mid \boldsymbol{\theta}))$ requires a numerical evaluation of $\boldsymbol{b}(p)$, its derivative $\boldsymbol{b}'(p)$, and its integral $\boldsymbol{B}(p)$;

- the initial parameters' values must be selected to ensure that the conditional quantile function is well-defined. Starting points are computed based on a preliminary estimate of the conditional distribution, obtained using a flexible parametric model implemented by the package `pch`.

- the sequence of $\lambda$ is obtained according to the penalized integrated gradient fixing the intercept parameters to their current estimates $\hat{\boldsymbol{\theta}}^{(t)}_{1\times k}$ and all the other parameters to $\mathbf{0}_{(q-1)\times k}$. Defining $\tilde{\boldsymbol{\theta}}^{(t)} = (\hat{\boldsymbol{\theta}}^{(t)}_{1\times k}; \mathbf{0}_{(q-1)\times k})$, the initial value of $\lambda$ is given by $\max \mid c_n \mid$, where $c_n = \nabla_{\boldsymbol{\theta}} L^{(\lambda)}_{\text{PEN}}(\tilde{\boldsymbol{\theta}}^{(t)})$.

- standard errors for the nonzero parameters are estimated using standard asymptotic theory of M-estimators as in Frumento and Bottai, (2016).

The `qrcmNP` package contains a main function that implements model fitting using pathwise coordinate descent and quasi Newton-Raphson algorithms (Bottai, Orsini, and Geraci, 2015) to minimize $L^{(\lambda)}_{\text{PEN}}$ at the selected values of the tuning parameter $\lambda$. A variety of summary measures, variable selection procedures, predictions, and graphical tools are available.

## 4.7 Variables selection for inspiratory capacity

We applied the QRCMPEN estimator to a subset (n = 2045) of the data analyzed in Bottai et al., (2011). The data arose in to a study carried out in 1988-1991 in Northern Italy, and included 1053 males and 992 females. The study aimed to estimate percentiles of inspiratory capacity (IC), a measure of lungs function. The following nine predictors were available: age, height, body mass index (BMI), sex, and indicators for current smoking, occupational exposure, cough, wheezing, and asthma.

We used $\boldsymbol{b}(p) = [1, \log(p), \log(1-p)]^T$ to model the intercept, while the coefficients associated with the covariates were described by a shifted Legendre polynomial up to fifth degree, inclusive of an intercept. The maximal model had $3 + 6 \times 9 = 57$ parameters. We used AIC and BIC to assess model fit. Figure 4.2 illustrates the results of our analysis and some graphical diagnostic tools.

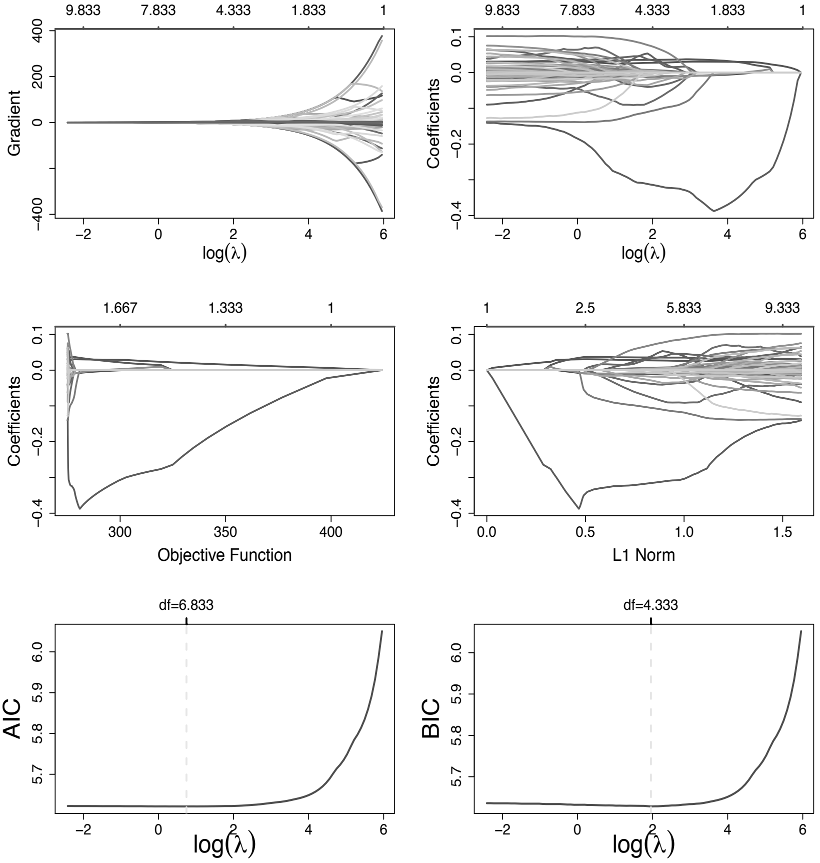


FIGURE 4.2: Gradient plot and coefficient profile plot versus $\log(\lambda)$ (upper panels); coefficients versus objective function plot and $l_1$-norm plot (mid panels), AIC and BIC curves versus $\log(\lambda)$ (bottom panels), for the inspiratory capacity data.

All criteria indicated that the best model included all the considered covariates. However, several model parameters were shrunk to zero. Results are reported in Table 4.3.

TABLE 4.3: Model selection based on different criteria. Degrees of freedom (as defined in Section 4.4), number of parameters (as a fraction of that of the maximal model), minimized loss function, and p-value of a Kolmogorov-Smirnov goodness-of-fit test (Frumento and Bottai, 2016) of the model selected by AIC and BIC. The maximal model had 57 parameters. AIC and BIC criteria selected 6.83 and 4.33 degrees of freedom, respectively, corresponding to 37 and 23 nonzero parameters. The model selected by AIC is summarized in Table 4.4 and represented graphically in Figure 4.3.

| **Criterion** | df$^{(\boldsymbol{\lambda})}$ | **n. of parameter** | **Loss** | **P-value KS** |
|---|---|---|---|---|
| AIC | 6.83 | 37/57 | 275.19 | .43 |
| BIC | 4.33 | 23/57 | 275.49 | .33 |

We used the model selected by AIC and estimated it again using unpenalized QRCM. The model is summarized in Table 4.4, and represented graphically in Figure 4.3. Because we were mostly interested in the low quantiles of IC, in Table 4.5 we only report the estimated quantile regression coefficients, $\widehat{\boldsymbol{\beta}}(p) = \boldsymbol{\beta}(p \mid \widehat{\boldsymbol{\theta}})$, at $p = 0.01$, $p = 0.05$, and $p = 0.50$.

Age, height, BMI and sex were statistically significant. Figure 4.3 shows the regression coefficient functions for all covariates over the interval $p \in (0, 1)$. Age had a negative effect at all quantiles, and the associated coefficient function showed an increasing linear trend. As age increased by one year, IC decreased by about $0.01$ liters. Height had a positive effect, and its regression coefficient function showed a J-shape. As height increased by one centimeter, IC increased by about $0.03$ liters. BMI had a positive effect with a reverse J-shape. As BMI increased by one unit, IC increased by $0.03$ at the first percentile, $0.04$ at the fifth percentile and $0.06$ at the median. The coefficient function associated with male was negative and decreasing. This indicated that, after including all the other covariates, the distribution of IC in males was shifted towards lower values and had larger variability than in females.

## 4.8 Discussion

We described a penalized approach that can be applied to the QRCM framework introduced by Frumento and Bottai, (2016). Modeling the

TABLE 4.4: Summary of the model selected by AIC. ILM estimates of $\boldsymbol{\theta}$ and asymptotic standard errors (in brackets) based on the model selected by AIC. The corresponding graphical representation is proposed in Figure 4.3. The bottom row contains the p-values for the null hypothesis that the corresponding column 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.

| | **1** | **log $(p)$** | **log $(1-p)$** | $\boldsymbol{slp}(\boldsymbol{p},\mathbf{5})[1]$ | $\boldsymbol{slp}(\boldsymbol{p},\mathbf{5})[2]$ | $\boldsymbol{slp}(\boldsymbol{p},\mathbf{5})[3]$ | $\boldsymbol{slp}(\boldsymbol{p},\mathbf{5})[4]$ | $\boldsymbol{slp}(\boldsymbol{p},\mathbf{5})[5]$ | **P-value** |
|---|---|---|---|---|---|---|---|---|---|
| Intercept | 3.482 (.095)* | 0.347 (.047)* | -0.251 (.046)* | - | - | - | - | - | .000* |
| Age | -0.013 (.001)* | - | - | 0.002 (.001)* | - | - | - | - | .000* |
| Height | 0.033 (.004)* | - | - | 0.006 (.002)* | 0.005 (.002)* | - | 0.001 (.002) | -0.001 (.001) | .000* |
| BMI | 0.030 (.007)* | - | - | 0.011 (.003)* | -0.009 (.003)* | - | -0.004 (.003) | - | .000* |
| Male | -0.162 (.065)* | - | - | -0.144 (.039)* | -0.064 (.033)* | -0.034 (.020) | 0.020 (.013)* | - | .000* |
| Smoker | -0.026 (.054) | - | - | 0.064 (.027)* | -0.001 (.025) | -0.002 (.024) | - | -0.012 (.016) | .138 |
| Non-exposure | 0.030 (.055) | - | - | 0.021 (.030) | - | -0.016 (.022) | - | 0.008 (.018) | .568 |
| Cough | 0.020 (.071) | - | - | - | 0.026 (.040) | -0.042 (.044) | - | - | .238 |
| Wheezing | -0.037 (.063) | - | - | 0.107 (.034)* | 0.027 (.037) | - | 0.062 (.030)* | - | .003* |
| Asthma | 0.082 (.120) | - | - | - | -0.078 (.058) | - | - | -0.148 (.069) | .098 |
| P-value | .000* | .000* | .000* | | | .000* | | | |

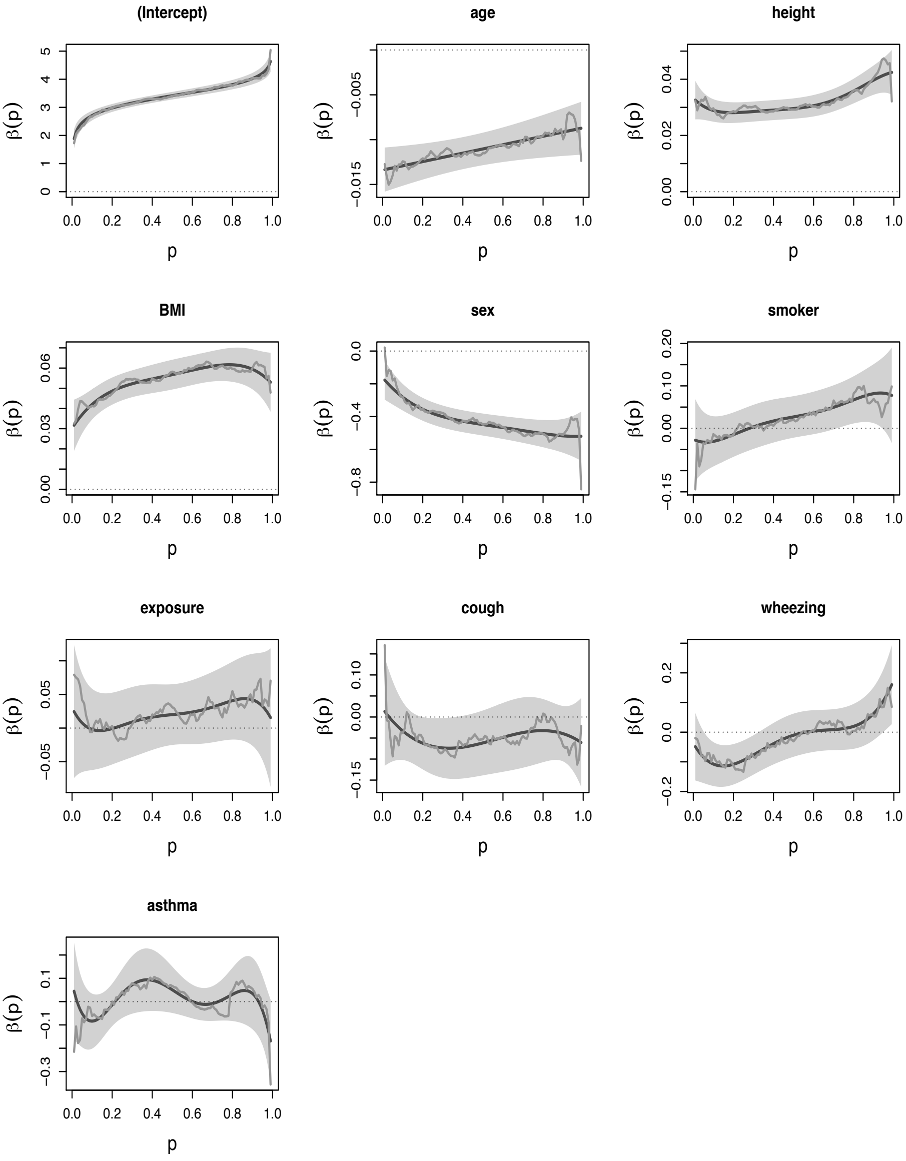


FIGURE 4.3: ILM estimates of $\boldsymbol{\beta}(p)$ under the model selected by AIC (see Table 4.4). Confidence bands are displayed as shaded areas. The broken lines connect the coefficients of ordinary quantile regression estimated at a grid of quantiles. The dashed line indicates the zero.

TABLE 4.5: Estimated quantile regression coefficients at $p = 0.01$, $p = 0.05$ and $p = 0.50$, obtained from the model selected by AIC. Estimated standard errors in brackets. The asterisk (*) denotes significance less than $0.05$.

| | $p = 0.01$ | $p = 0.05$ | $p = 0.50$ |
|---|---|---|---|
| Intercept | 1.885 (.174)* | 2.454 (.110)* | 3.415 (.069)* |
| Age | -0.013 (.001)* | -0.013 (.001)* | -0.011 (.001)* |
| Height | 0.033 (.003)* | 0.030 (.002)* | 0.029 (.002)* |
| BMI | 0.032 (.006)* | 0.037 (.005)* | 0.057 (.004)* |
| Male | -0.177 (.061)* | -0.229 (.050)* | -0.448 (.034)* |
| Smoker | -0.028 (.049) | -0.033 (.037) | 0.027 (.025) |
| Non-exposure | 0.024 (.050) | 0.008 (.037) | 0.020 (.023) |
| Cough | 0.013 (.066) | -0.010 (.049) | -0.062 (.039) |
| Wheezing | -0.049 (.058) | -0.084 (.045) | -0.010 (.037) |
| Asthma | 0.045 (.106) | -0.051 (.074) | 0.051 (.051) |

conditional quantile function parametrically can be more efficient than estimating quantiles one at a time, as in ordinary quantile regression. Moreover, it permits performing model selection directly on the parameters that describe conditional quantiles, instead of proceeding quantile-by-quantile, as the penalized methods for quantile regression proposed so far do. By working with QRCM, it is easy to formulate highly parametrized models, as each covariate has multiple associated parameters.

The QRCMPEN estimator demonstrated to select the correct model with a high probability. A computationally efficient algorithm has been implemented in the `qrcmNP` package in R.

# Appendix A

# Clusters of effects curves in quantile regression models

## A.1 Introduction

General statistical techniques aim to reduce dimensionality aiming to detect the most relevant information for a better interpretation of observed data. In particular, various methods, combining cluster analysis and the search for a lower-dimension representation, have been proposed in a finite-dimensional setting by Vichi and Saporta, (2009). More recently, the use of clustering is considered as a preliminary step for exploring data represented by curves, with additional difficulty associated to the infinite space dimension of data (Jacques and Preda, 2014).

In this Appendix, we focus on a new method to find similarities of curves in a quantile regression coefficient modeling framework, possibly multivariate, in which the effect of covariates on a response variable is represented by curves in the space of percentiles. The proposed approach is very flexible and, as shown, can be also generalized to different contexts, such as clustering of waveform curves (i.e. seismic events or signals).

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 Clogg, Petkova, and Haritou, (1995).

The general issue of curve clustering could be very complex for many reasons, which can be due, for instance, to subjective choices related to the transformation of the observed data. The variability across curves can be distinguished into two components: phase variability (removed after the alignment of the curves) and amplitude variability (Sangalli et al., 2009). 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 Silverman, (1995) and Ramsay and Li, (1998), “time warping” in engineering Wang and Gasser, (1997) and “structural averaging” in the context of computing an average curve Kneip and Gasser, (1992).

Silverman, (1995) proposed a more general approach, in which a target curve is defined to which each other curve must be registered based on some meaningful criterion, such as a local feature of the curve, or the minimization of a distance measure.

Ramsay and Li, (1998) used a Procrustes fitting procedure (Gower, 1975) to provide maximal alignment to the target function, subject to the suitable smoothness of the transformations. Adelfio et al., (2012) introduced a simple procedure to identify clusters of multivariate waveforms based on a simultaneous assignation and alignment procedure. James, (2007) 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 Garcia-Escudero and Gordaliza, (2005) is introduced in Adelfio et al., (2011), considering a functional principal component rotation of data (Ramsay, 2006). This approach has been extended in Adelfio, Di Salvo, and Chiodi, (2016), where the authors focused on finding clusters of multidimensional curves with spatio-temporal structure.

All the aforementioned methods have been defined in a slightly different context with respect to the one we consider here; indeed, none of the above approaches can be suitable for clustering curves of effects in quantile regression. These curves have typically variable trends and polynomial shapes. Therefore, our aim is to find effects (i.e. curves) that are not *significantly* different and then, to identify clusters of *similar* covariates, according to a variable selection perspective.

This Appendix is organized as follows: in Section A.2 we report the usual notation of quantile regression, together with some recent developments that permit describing coefficient functions parametrically. In Section A.3 we introduce the new method for curves clustering, together with the algorithm details. In Section A.4 simulated results are reported both for effects curves in quantile regression coefficient modeling and in general waveform context. Example of applications on real data are reported in Section A.5. Section A.6 is didicated to conclusive remarks.

## A.2 Quantile regression and recent extensions

Dealing with non-normal distributions and outliers, the use of quantile regression (QR; Koenker and Bassett Jr, 1978; Koenker, 2005) to investigate the influence of some covariates on a response is suggested. Indeed, the ordinary least squares (OLS) regression does not take into account the whole shape of distribution of the outcome variable. Conversely, QR provides information on the entire distribution, including for example the tails, and not just its mean. Additionally, quantile regression estimators are generally more robust to outliers than ordinary least squares. 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.

Frumento and Bottai, (2016), suggest adopting a parametric model for the coefficient function of a quantile regression. They refer to this estimation approach as quantile regression coefficients modeling (QRCM) and implemented it in the qrcm R package (Frumento, 2017). Standard quantile regression suffers from the following limitations: (i) quantiles are estimated one at the time, (ii) the estimated coefficients are generally unsmoothed functions of the percentiles ($p$) and may suffer from a high volatility that hinders their interpretability. The QRCM framework overcomes the aforementioned limitations: (i) the entire quantile function is estimated at once; (ii) this modeling approach facilitates estimation, inference, and interpretation of the results, and generally yields a gain in terms of efficiency. More in details, given a response variable $y$ and a model matrix $\boldsymbol{x}$ of dimension $N \times q$, 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 $\boldsymbol{\beta}(p)$ 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 $\theta_{jh}$ associated to the $j$-th covariate and the $h$-th function, $j = 1, \ldots, q$ and $h = 1, \ldots, k$. The authors suggest estimating $\boldsymbol{\theta}$ as the minimizer of the integrated objective function

$$\overline{L}_N(\boldsymbol{\theta}) = \int_0^1 N^{-1} \sum_{i=1}^{N} (p - \omega_{p,i})(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}(p \mid \boldsymbol{\theta}))\mathrm{d}p,$$

where $\omega_{p,i} = I(y_i \leq \boldsymbol{x}_i^T \boldsymbol{\beta}(p \mid \boldsymbol{\theta}))$ and $I(\cdot)$ is the indicator function. With this approach, $\boldsymbol{\beta}(p \mid \boldsymbol{\theta})$ is treated as a finite-dimensional parameter.

In this framework, our proposal tries to answer two different questions:

1. in the univariate case, 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 effects curves. 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, as a variable selection procedure;

2. in the multivariate case, let $\boldsymbol{y} = [y_1, \ldots, y_t, \ldots, y_m]$ be a set of $m$ response variables, each of length $N$; applying the QRCM on each $y_t$, we estimate the $m \times q$ effects curves $\beta_{11}(p \mid \boldsymbol{\theta}), \ldots, \beta_{mq}(p \mid \boldsymbol{\theta})$. The aim is to assess if there are similar responses given covariates, as a preliminary step to describe clustered outcomes most influenced by a given covariate.

## A.3 The proposed clustering method

The proposed clustering approach based on a new measure of dissimilarity that used both the shape of a curve and the distance with respect to other curves:

- the *shape* of a curve is evaluated using its second deravative. Moreover, two different curves are similar in shape if for a fixed point the signs of the second derivatives are concordant;

- the *distance* between two curves is evaluated as their a new measure of dissimilarity that used both the shape of a curve and the distance with respect to other curves. Two curves are said close if their distance at any given point is lower than a fixed value.

Let $i$ and $i'$ be two different curves, $\boldsymbol{p} \in (0, 1)$ the vector of percentiles.

$$d^{ii'}_{\text{shape}}(\boldsymbol{p}) = I(\text{sign}(\beta''_i(\boldsymbol{p} \mid \boldsymbol{\theta})) \times \text{sign}(\beta''_{i'}(\boldsymbol{p} \mid \boldsymbol{\theta})) = \mathbf{1})$$
$$d^{ii'}_{\text{distance}}(\boldsymbol{p}) = I(|\beta_i(\boldsymbol{p} \mid \boldsymbol{\theta}) - \beta_{i'}(\boldsymbol{p} \mid \boldsymbol{\theta})| \leq f(\alpha, \text{dist}(\boldsymbol{p}))),$$

where $f(\cdot, \cdot)$ is a cut-off function, that depends on a probability value $\alpha$, and on dist($\boldsymbol{p}$), the vector of the distances between two curves across all percentiles. The probability value $\alpha$ has a central role for finding homogeneous clusters and its choice depends on the analysis aim and, therefore, has to be fixed by the researcher. Fixing an $\alpha$-level too small or too big could provide inhomogeneous clusters. The cut-off function $f(\cdot, \cdot)$

selects the $\alpha$-th percentile vector of $\text{dist}(\boldsymbol{p})$. In our opinion, the median ($\alpha = 0.50$) is strongly suggested in case of waveforms clustering, while the first quantile ($\alpha = 0.25$) is preferable in the clustering of curves of effects.

Finally, the proposed dissimilarity measure between two curves is defined as

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

In (A.1), the product of the two measures is computed, to account for their concordance at each point. The value in (A.1) defines a metric, as it satisfies the following properties:

1. Nonnegativity: $d(i,i') \geq 0$;
2. Reflexivity: $d(i,i') = 0$ if and only if $i = i'$;
3. Symmetry: $d(i,i') = d(i',i)$;
4. Triangle Inequality: $d(i,i') + d(i',i'') \geq d(i,i'')$.

In the proposed approach, defining a dissimilarity matrix is useful for the application of any hierarchical clustering method. We implemented the proposed procedure in the `clustEff` package in R, that includes several very flexible functions.

### A.3.1 Choice of the number of clusters

The choice of the number of clusters is crucial for most clustering algorithms. We discuss this issue with the goal of offering a classification tool that is sufficiently flexible and, at the same time, can be used in different contexts.

In the case of clustering of effects curves of a QRCM, the optimal number of clusters (say $k^*$) is obtained with a criterion based on the confidence bands of curves. Starting from each partition of curves in $k$ clusters and their estimated confidence bands, we build the average lower and upper bands within cluster ($\overline{\text{LB}}^j(\boldsymbol{p})$ and $\overline{\text{UB}}^j(\boldsymbol{p})$, $j \in \{1,k\}$). Then, we compute the proportion of curves that are outside the average bands. For each $k = 1, ..., K \leq q$, let us define

$$\pi^k_{\text{out}} = k \sum_{j=1}^{k} q_j^{-1} \sum_{i=1}^{q_j} \left\{ \int_0^1 I\left( \overline{\text{LB}}^j(p) \leq \beta_i^j(p \mid \boldsymbol{\theta}) \leq \overline{\text{UB}}^j(p) \right) \mathrm{d}p \right\},$$

where $q_j$ is the number of curves in the $j$-th cluster and $I(\cdot)$ is the indicator function. The value of $k^*$ is identified by that partition for which $\pi^k_{\text{out}} - \pi^{k+1}_{\text{out}}$ is minimized.

The proposed measure, as defined by (A.1), could be also an useful tool for clustering time-dependent signals, usually analyzed in functional data analysis (FDA). The nature of these curves is generally different from that of quantile regression coefficient functions. In FDA, 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, we look for the relative distances between curves belonging to the same cluster and their centroid ($\overline{\beta}^j$, $j \in \{1, k\}$). For each $k = 1, ..., K \leq q$, let us define

$$\text{dist}^k_{\text{rel}} = k \sup_{j\in\{1,\ldots,k\}} \left\{ q_j^{-1} \sum_{i=1}^{q_j} \int_0^1 |\overline{\beta}^j(p) - \beta^j_{q_j}(p \mid \boldsymbol{\theta})| \, \mathrm{d}p \right\}.$$

Then, $k^*$ is identified by that partition for which $\text{dist}^k_{\text{rel}} - \text{dist}^{k+1}_{\text{rel}}$ is minimized.

### A.3.2 Steps of the Algorithm

The main steps of the algorithm are summarized as following:

1. fix the $\alpha$-level and calculate all the possible distances between the pairs of curves for each percentile (i.e., $\text{dist}(\boldsymbol{p})$); then, the cut-off function selects the percentile of the distribution of $\text{dist}(\boldsymbol{p})$ used in $d^{ii'}_{\text{distance}}(\boldsymbol{p})$;

2. after computing $d^{ii\prime}_{\text{shape}}(\boldsymbol{p})$ and $d^{ii'}_{\text{distance}}(\boldsymbol{p})$, the dissimilarity matrix is calculated as in (A.1);

3. apply a hierarchical clustering algorithm in order to obtain the dendrogram;

4. select the optimal number of clusters as in Section A.3.1, unless $k$ is known in advance;

5. after selecting the number of clusters calculate the mean curves within each cluster. Provide goodness-of-fit measures..

The `clustEff` package includes the main function that implements the described algorithm, along with a summary function and graphical tools.

Based on numerous applications and simulation results, that are only partially summarized in the rest of this section, the algorithm seems to be stable and computationally efficient.

## A.4 Simulation study

In this section, we report simulated results to prove the validity of the proposed approach for cluster of curves, both referring to effects curves in a quantile regression coefficients modeling and to general curves of waveforms. Moreover, the proposed clustering is compared with two different algorithms: the model-based clustering algorithm proposed by Bouveyron and Brunet-Saumard, (2014) (funFEM) and the functional principal component analysis algorithm proposed by Adelfio et al., (2011) (FPCA).

The funFEM method is based on a functional mixture model that allows the clustering of the data in a discriminative functional subspace. This model clusters observed curves into $K$ homogeneus groups and assuming that there exists an unobserved random variable $Z = (Z_1, \dots, Z_k) \in \{0, 1\}^K$ indicating the group membership of each curve. However, because the group memeberships of the curves are unknown, the direct maximization of the likelihood associated with the model proposed by Bouveyron and Brunet-Saumard, (2014) is intractable. The EM algorithm is applied to perform optimization.

The FPCA method combines the aim of finding clusters from a set of observed curves with the functional nature of data. It applies a variant of *k*-means algorithm based on the principal component rotation of data. The main idea behind this clustering approach is to find a linear approximation of each curve by a finite-dimensional vector of coefficients defined by the FPCA scores. This method assigns curves to a cluster if the distances between them are less than a fixed threshold in the space of the PCA scores.

In both simulations we generate 100 replicates and in each of them the optimal number of clusters $k^*$ is automatically provided. We report and compare the following statistics:

- the average area between each curve and the mean curve of each cluster, calculated as

$$\text{Area}(k^*) = k^{*-1} \sum_{j=1}^{k^*} \Bigg\{ q_j^{-1} \sum_{i=1}^{q_j} \int_0^1 \Bigg( \mid \overline{\beta}^j(p) - \beta_i^j(p \mid \boldsymbol{\theta})| \Bigg) \mathrm{d}p,$$

where $q_j$ is the number of curves in the $j$-th cluster with $j \in \{1, k^*\}$ and $\overline{\beta}^j(\cdot)$ is the mean effect curve of cluster $j$;

- the average distance based on the correlation among all the curves in each cluster, calculated as

$$\rho_{\text{dist}}(k^*) = k^{*-1} \sum_{j=1}^{k^*} \left\{ 1 - \left[ 2\Big( q_j(q_j - 1) \Big)^{-1} \sum_{i=1}^{q_j - 1} \sum_{z>i}^{q_j} \rho_{iz} \right]^2 \right\},$$

  where $q_j$ is the number of curves in the $j$-th cluster with $j \in \{1, k^*\}$ and $\rho_{iz}$ is the correlation between $i$-th and $z$-th curve;

- the average number of clusters.

### A.4.1 Clusters of effects

We considered a multivariate scenario in which the general quantile function was simulated as

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

where $x \sim \mathbb{U}(0, 5)$ and $p \in (0, 1)$. In the first simulation scenario, the intercept was modeled as the quantile function of a standard normal distribution ($\phi$). Other choices, as suggested in the original paper of Frumento and Bottai, (2016), could be considered. We defined three groups of quantile functions

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

where $\boldsymbol{\theta}_1 = (.5, .5, 1, 2), \boldsymbol{\theta}_2 = (-.3, .5, 1, .5), \boldsymbol{\theta}_3 = (.3, .5, -1, 2)$. For each quantile function $(Q_1, Q_2, Q_3)$ ten response variables were generated according to the parameters $\boldsymbol{\theta}_i + \epsilon_i,\ i = 1, 2, 3$, where $\epsilon_i \sim \mathcal{N}(0, 2)$.

In each replicate we obtained 30 responses $y_1, \ldots, y_{30}$. Applying the QRCM method to these response variables, we could evaluate the effect of the covariate $x$ on each of them. The lower and upper bounds are easily estimated and used within the `clustEff` algorithm to select the optimal number of clusters. Figure A.1 shows the dendrogram, the 30 effect curves clustered and the boxplot of the average dissimilarity within each cluster after applying the `clustEff` algorithm for one replicate.

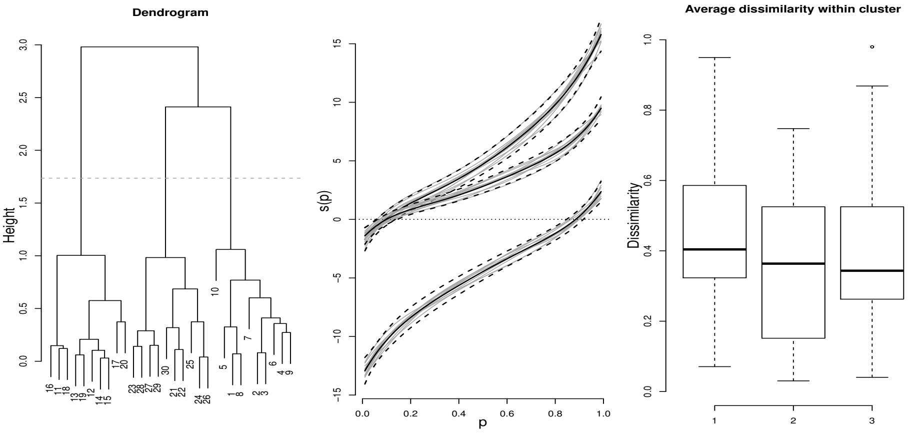


FIGURE A.1: Output of the proposed algorithm for one replicate. Left panel shows the dendrogram; mid panel shows the 30 curves clustered in 3 groups; right panel shows the boxplot of the average dissimilarity within each cluster.

In Table A.1 results are summarized and compared with the true number of clusters and the true partition of curves, used as benchmark measures. Since the average of the area between each curve and the mean curve decreases as the number of clusters increases, in our opinion this is a reasonable statistics for comparing different clustering methods only being equal the number of clusters. Therefore, operatively, we first identified the best approach in terms of the selected number of clusters; then, we assessed the goodness of the partition by looking at the average area and at the average correlation.

TABLE A.1: Average area, average distance based on correlation ($\rho_{\text{dist}}$) and average of the optimal number of discovered clusters ($k^*$) are compared with the three algorithm (clustEff, funFEM and FPCA) using as benchmark measure the true partition of curves in the 100 runs. Standard deviations in brackets.

| | **True** | **clustEff** | **funFEM** | **FPCA** |
|---|---|---|---|---|
| $k^*$ | 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) |
| $\rho_{\text{dist}}$ | 0.010(0.015) | 0.010(0.015) | 0.008(0.008) | 0.010(0.015) |

As shown in the Table A.1, the `clustEff` and the `FPCA` methods are both, on average, more precise than the funFEM in terms of number of clusters. Both algorithms choose, on average, the true number of clusters, even if the clustEff has the largest standard error. The funFEM algorithm overestimates the number of cluster, which makes it difficult to compare results in terms of area and correlation. Moreover, in terms of the distance based on the correlation, the three methods have the same performance. In our opinion, in this framework, the clustEff approach could be preferable, both for its user-friendly nature and since it introduces a new perspective of curves clustering, comparing also the shape of curves. This point is relevant for a clustering method that aims at clustering effects curves. Indeed, the similarity of effects in a quantile regression model can not be based just on the closeness of curves, since also the shape represents an important information, suggesting for instance trends or direction, functions of the percentiles.

### A.4.2 Waveform clustering

For the waveform clustering context, we simulate 30 harmonic functions evaluated on a grid of size 1000 in $t \in [0,1]$, such that 10 are generated from the function $f(t) = \sin(3\pi t)$, 13 from the function $g(t) = \cos(3\pi t)$, 5 from the function $h(t) = \sin(3\pi t)\cos(\pi t)$, and 2 from the function $l(t) = 0$, as outlier curves.

To each curve a random error $\epsilon_t \sim \mathcal{N}(0, \sigma_t)$ is added, where $\sigma_t$ is the square root of the variance function defined by a segmented relation with multiple change-points, such as

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

Figure A.2 shows the 30 curves clustered, the dendrogram and the boxplot of the average dissimilarity within each cluster applying the clustEff algorithm to one replicate.

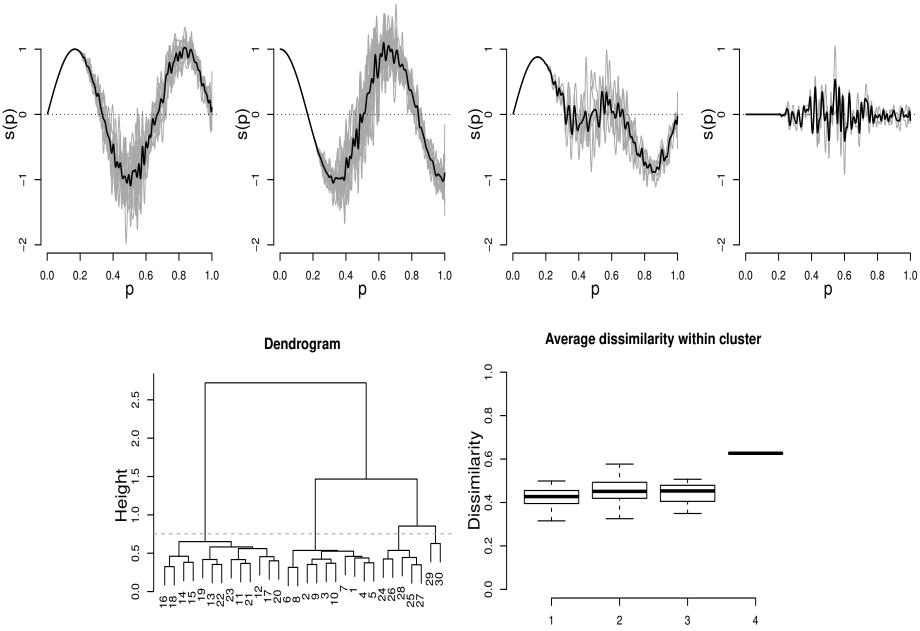


FIGURE A.2: Output of the proposed algorithm for one replicate. Upper panels show the 4 clusters; bottom-left panel shows the dendrogram; bottom-right panel shows the boxplot of the average dissimilarity within each cluster.

The three algorithms, `clustEff`, `funFEM` and `FPCA`, are compared using the same summary statistics used in the simulation Section A.4.1, i.e., the average area between each curve and the mean curve in each cluster, the average correlation betwen curves in each cluster, and the average value of the selected number of clusters ($k^*$). The true number of clusters and the true partition of curves are considered as benchmark measures. As in the previous section, the same considerations about the statistics hold. Results are reported in Table A.2.

In this example, basing on the three summary statistics defined in Section A.4, it is not possible to assess which is the outperforming method. The `clustEff` and the `FPCA` methods perform similarly in terms of the chosen $k^*$, but the `clustEff` is the best one in terms of the average area and the correlation distance, the best one. The `funFEM` algorithm, instead, underestimates the number of clusters.

In clustering of waveforms, where the curves typically have zero-mean with high time-dependent variance, the comparison of shapes, which

TABLE A.2: Average area, average distance based on correlation ($\rho_{\text{dist}}$) and average of the optimal number of discovered clusters ($k^*$) are compared across the three algorithm (clustEff, funFEM and FPCA) using as benchmark measure the true partition of curves in the 100 runs. Standard deviations in brackets.

| | **True** | **clustEff** | **funFEM** | **FPCA** |
|---|---|---|---|---|
| $k^*$ | 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) |
| $\rho_{\text{dist}}$ | 0.441(0.177) | 0.426(0.175) | 0.436(0.203) | 0.437(0.171) |

is one of the two components of the proposed measure in (A.1), should weigh less than the closeness assessment. Nevertheless, these results still confirm the efficiency of the `clustEff` approach, also in a FDA context.

## A.5 Examples of application of the clusteEff algorithm on real datasets

In this section, we apply the proposed clustering algorithm to three different real datasets, in order to show its flexibility and wide spectrum of application. Results are compared with `funFEM` and `FPCA` methods using the same summary statistics as in Section A.4.

### A.5.1 Dataset 1

The first analyzed 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): 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;
- 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 seven dependent variables (mag, errM, depth, lon, lat, errZ, errH) and a set of four independent variables (gap, mDst, rms, nSt).

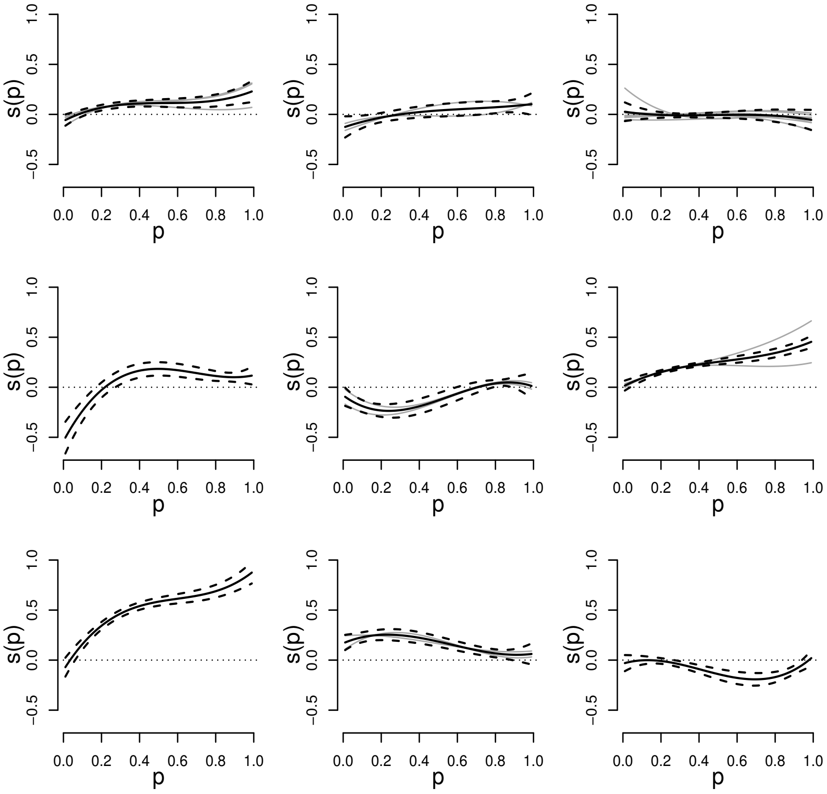
In this example, the main purpose is to describe some kind of relationship among the set of dependent variables and the set of independent variables. In particular, we look for clusters of dependent variables after estimating multiple quantile regressions, one for each response. Clustering of effects on different responses could reflect existing relationships between the responses.

In Table A.3, 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-nSt) and negative correlations (errH-nSt, errZ-nSt, gap-nSt) are shown. Since gap and mDst are higlhy correlated, we decided to exclude from the next analysis the independent variable mDst.

TABLE A.3: 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, mDst=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** | **mDst** | **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 | | | |
| **mDst** | 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 |

Using the QRCM approach, we model the intercept, $\beta_0(p)$, using the quantile function of a standard Normal distribution, and the coefficients associated to the covariates by a shifted Legendre polynomial (Abramowitz and Stegun, 1964) up to the third degree, an orthogonal polynomial in $(0, 1)$ that can be used to define flexible models for $\beta(p)$. We obtain $21$ effects curves, from seven models and three covariates. In Figure A.3 we report the $21$ curves clustered in $9$ groups of size $(3, 2, 7, 1, 2, 2, 1, 2, 1)$ after applying the `clustEff` algorithm.



---

FIGURE A.3: Clusters of the 21 curves from the seven models with the three covariates (gap, rms, nSt). Solid line represents the mean curve. The dotted lines are the mean lower and upper bands; grey solid lines are the effects curves

In details:

- in cluster 1 there are the effects curves of *Gap azimuth* on the responses magnitude and hypocentral uncertainty, and of *RMS* on the response hypocentral uncertainty; these effects are positive for percentiles greater than .08 (Figure A.3, first row on the left);
- in cluster 2, there are the effects curves of *Gap azimuth* on the response magnitude error, and of *RMS* on the response latitude; these effects are positive for percentiles greater than .30 (Figure A.3, first row in the middle);
- cluster 3 consists of the effects curves of *Gap azimuth* on the response depth, of *RMS* on the responses magnitude, magnitude error and

depth, and of *Number of stations* on the responses depth, hypocentral uncertainty and epicentral uncertainty; these effect are almost all negative for percentiles greater than .15 (Figure A.3, first row on the right);

- cluster 4 contains the only curve representing the effect of *Gap azimuth* on the response latitude; it is positive for percentiles greater than .22 (Figure A.3, second row on the left);
- in cluster 5, there are the effects curves of *Gap azimuth* and *RMS* on the response longitude; these curves are negative for percentiles up to .70 (Figure A.3, second row in the middle);
- cluster 6 includes the effects curves of *Gap azimuth* and *RMS* on the response epicentral uncertainty, with positive values for all the percentiles (Figure A.3, second row on the right);
- cluster 7 contains the only effect curve of *Number of stations* on the response magnitude; it is positive for percentiles greater than .04 (Figure A.3, third row on the left);
- in cluster 8, the effects of *Number of stations* on the responses magnitude error and longitude are positive for all the percentiles (Figure A.3, third row in the middle);
- cluster 9 contains the only effect curve of *Number of stations* on the response latitude; it is negative for percentiles up to .97 (Figure A.3, third row on the right).

In this application, we show an interesting usage of the proposed clustering method, identifying clusters of curves in a multivariate context. Though the difficulty in interpreting the results, this approach could represent an useful tool to describe the relationship between variables according to a dependence model. We observe that some covariates are more relevant for some outcomes than for others, and in addition, there are covariates that have the same behavior with respect to a given response variable, as for instance cluster 3, where *Gap azimuth, RMS, Number of stations* have the same effect on the response depth. These results could be useful for addressing some operative choices for the seismic network definition of a given region. The three clustering methods are compared in Table A.4, which has the same format as in Section A.4. The three methods find different value for $k^*$, which does not permit a direct comparison of the area, which is a decreasing function of $k^*$. However, in terms of the average distance based on the average correlation our proposal slightly outperforms the others.

TABLE A.4: Average area, average distance based on correlation ($\rho_{\text{dist}}$) and average value of the optimal number of discovered clusters ($k^*$) are compared across the three algorithm (clustEff, funFEM and FPCA).

| | **clustEff** | **funFEM** | **FPCA** |
|---|---|---|---|
| $k^*$ | 9 | 5 | 3 |
| Area | 0.023 | 0.052 | 0.077 |
| $\rho_{\text{dist}}$ | 0.352 | 0.368 | 0.700 |

### A.5.2 Dataset 2

The second dataset refers to a study carried out in 1988-1991 in the North of Italy, including 1053 males and 992 females. The study aims to assess determinants of the inspiration capability (IC), a measure of lung's function, using the following nine predictors:: age, height, body mass index (bmi), sex, and indicators for current smoking, occupational exposure, cough, wheezing, and asthma.

We adopt the QRCM framework and model the intercept as a linear combination of $\log(p)$ and $\log(1-p)$, that together define the quantile function of the asymmetric Logistic distribution, which is a very flexible model that can be used to describe possibly skewed random varaibles with heavy tails. The coefficients associated with the covariates are modeled by a fifth-degreen shifted Legendre polynomial. The effects curves of the fitted model ($\beta_{\text{age}}, \ldots, \beta_{\text{asthma}}$) are represented in Figure A.4. To discover similarity of effects of covariates, we applied the clustEff algorithm and identified five clusters. Results are summarized in Figure A.5. The clusters can be summarized as follows:

- in cluster $1$, age has a negative effect on IC at all percentiles;
- in cluster $2$, height has a positive effect on IC at all the percentiles;
- in cluster $3$, bmi has a positive effect on IC at all the percentiles;
- in cluster $4$, sex has a negative effect on IC at all the percentiles;
- in cluster $5$, current smoking, occupational exposure, cough, wheezing and asthma have a positive effect on IC at percentiles greater than .45. However, the mean effect of these cluster is almost zero

and is not statistically significant, since the mean lower and upper bound contain the zero (the dotted line in Figure A.5).

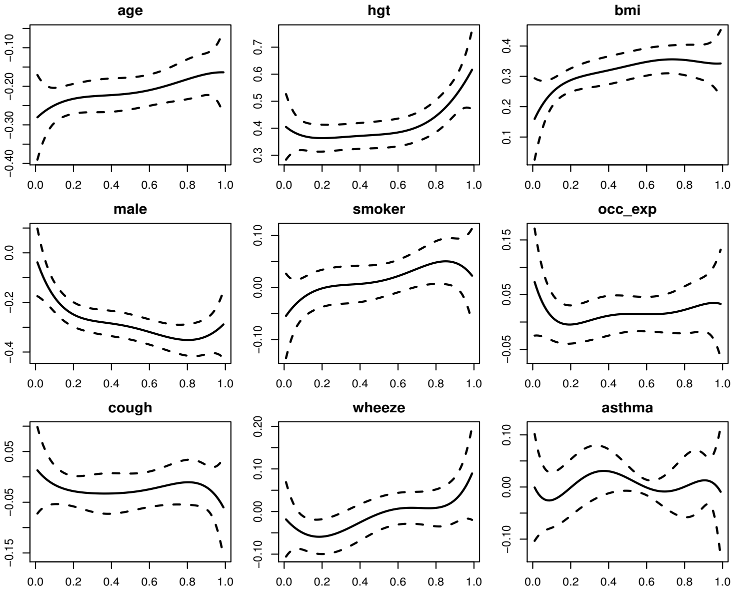


FIGURE A.4: QRCM estimates of $\beta(p \mid \boldsymbol{\theta})$. Confidence bands are displayed as dashed lines

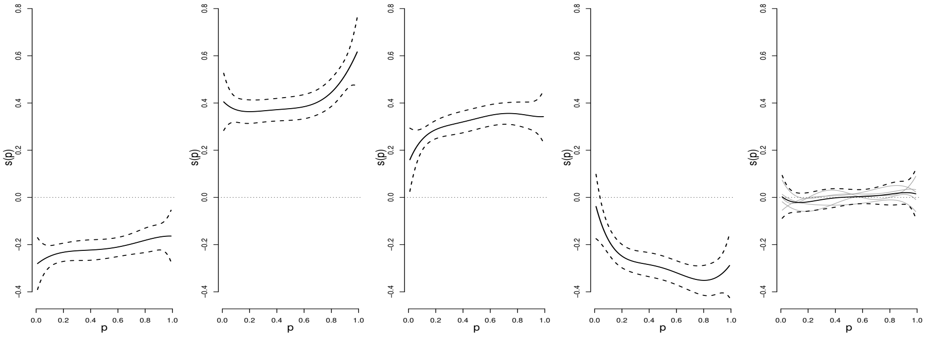


FIGURE A.5: The five clusters obtained applying the clustEff algorithm on the estimated quantile regression coefficients of inspiration capability dataset. Red line is the mean curve; the shaded areas are identified by the mean lower and upper bands within each cluster. The dotted line indicates the zero.

In this application, we focus on a new perspective of variables selection, applied in a quantile regression context. 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. These results could be even more relevant in presence of several regressors. We could observe that the last group is associated with the variables that are not related to the subject characteristics (the "clinical variables"); these variables are not statistically significant. Therefore, in describing the effect of covariates on the response, we interpret the average effect of each cluster, as a proxy for an average "characteristic effect" that is associated to the covariates in that cluster.

In Table A.5 the three methods are compared. Our proposal finds more clusters than the others, which prevents a direct comparison in terms of area.. However, in terms of the average distance based on the correlation our method identified clusters with more similar effects curves.

TABLE A.5: Average area, average correlation ($\rho_{\text{dist}}$) and average of the optimal number of discovered clusters ($k^*$) are compared across the three algorithm (clustEff, funFEM and FPCA).

| | **clustEff** | **funFEM** | **FPCA** |
|---|---|---|---|
| $k^*$ | 5 | 3 | 3 |
| Area | 0.004 | 0.039 | 0.039 |
| $\rho_{\text{dist}}$ | 0.197 | 0.710 | 0.710 |

### A.5.3 Dataset 3

The last dataset concerns waveform clustering, in a functional data analysis context, where curves are waves characterized by high concentrations around zero. The outcome of this dataset is the concentration of one pollutant (PM10) recorded during 2011 at different monitoring stations dislocated along the California state. It consists of 59 monitoring stations (curves) per 365 days observations (Adelfio, Di Salvo, and Chiodi, 2016). Applying our cluster algorithm, the time $t \in [1, 365]$ in days is scaled in $t \in [0, 1]$. Moreover, the 59 curves $(\beta_1, \ldots, \beta_{59})$ are clustered in six groups of size $20, 7, 7, 8, 7, 10$, respectively. The identified clusters, boxplot of the average cluster dissimilarity, and the dendrogram are reported in Figure A.6.

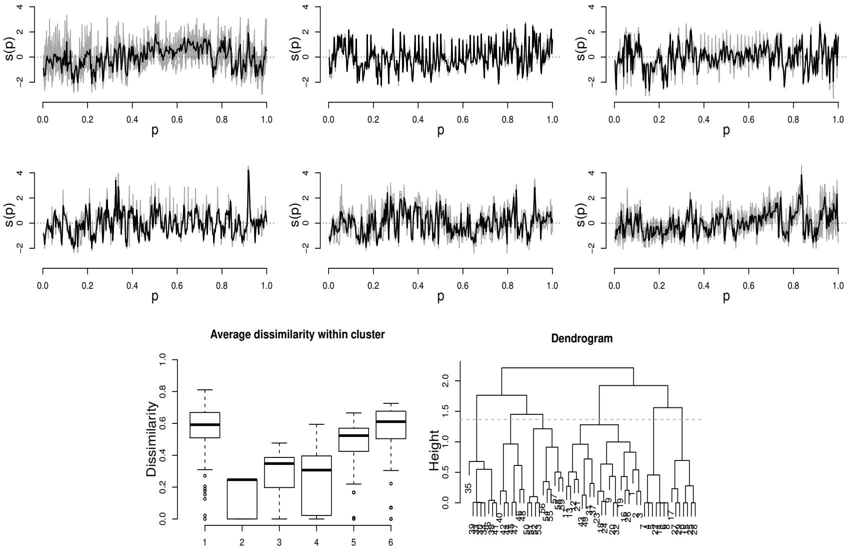


FIGURE A.6: The identified 6 clusters of dataset 3 (on the top): red lines are the mean curves. Boxplot of the average dissimilarity measure within each cluster (on the bottom-left). Dendrogram of the clustering algorithm and height level used to cut the tree (on the bottom-right)

In Table A.6 the three methods are compared. As in the previous datasets, the three methods select a different value for $k^*$, making the average area statistics not properly comparable. However, in terms of the average distance based on the correlation, although the `clustEff` performs slightly worse than the `funFEM`, it selects almost half of the clusters.

TABLE A.6: Average area, average distance based on correlation ($\rho_{\text{dist}}$) and average of the optimal number of discovered clusters ($k^*$) are compared across the three algorithm (clustEff, funFEM and FPCA).

| | **clustEff** | **funFEM** | **FPCA** |
|---|---|---|---|
| $k^*$ | 6 | 11 | 4 |
| Area | 0.286 | 0.187 | 0.406 |
| $\rho_{\text{dist}}$ | 0.362 | 0.227 | 0.539 |

## A.6 Conclusion

The proposed `clustEff` approach is not an 'usual' method for curves clustering, that is an important issue in many areas of science. Indeed, the `clustEff` approach is based on a new dissimilarity measure, that accounts both for the shape of curves and the distance between them; moreover, the new approach looks for similar effects in a quantile regression context, n which curves represent the effect of covariates on quantiles of one or multiple response variables. This method can be used to look for similarity of effects in a variable selection perspective. Simulation results confirm the advantages of the proposed method. Finally, we apply the `clustEff` algorithm to three different real datasets, among which also an application for generic waveforms, in order to show the wide spectrum of application for curves clustering. This approach, developed also in the `clustEff` R package, is very flexible and computationally fast.

# Appendix B

# R packages

## B.1 asnr

This package is inspired by the methodology presented in Chapter 3 and in part by the mixed strategy presented in Chapter 2. It implements a procedure based on the maximization of the average signal-to-noise ratio able to select efficiently the tuning parameter $\lambda$ in a LASSO regression. The main function of the package **asnr** is a function implemented in R which takes as input the model matrix $\boldsymbol{X}$ of dimension $N \times p$, the response vector $y$ of length $N$, the `family` to model the error distribution, and returns in output an object with S3 class `"asnr"`. This object is a list containing the optimal $\lambda$ value, the regression coefficients (only for the selected variables), the degrees of freedom and many other informations.

The main function allows the user to choose which criteria should be used to select the best model, i.e. `type = "asnr"` is the default. Other available options are `"aic", "bic", "ebic", "gcv, "cv", "gic", "stabs"`. The general sintax is displayed below:

```
asnr(x, y, obj = NULL, family = gaussian(),
     intercept = TRUE, standardize = TRUE,
     opt = c("max", "min"), dispersion = NULL,
     method = c("deviance", "pearson"),
     type = c("asnr", "aic", "bic", "ebic",
              "gcv", "cv", "gic", "stabs"),
     cn = c("log(log(n))", "1"), gamma = 0.5,
     nlambda = 100, plot.it = TRUE, set.seed,
     sample = 100, subsample = 0.5, pi = 0.5,
     num.select, ...).
```

This function takes in input, as alternative to the model matrix and the response variable, a fitted model (`obj`) of class `glmnet`; `intercept` and `standardize` if the intercept has to be inserted in the model and if the model matrix has to be standardized; opt, used only if type="asnr", allows to choose if the best value of $\lambda$ has to be select as maximizer of ASNR or minimizer of ASNR$^{-1}$; dispersion has to be a number and if fixed, it is used as dispersion parameter; method allows to choose which estimation method for the dispersion parameter has to be used; cn and gamma are constant for the gic and ebic criteria, respectively; nlamdba is the maximum length of the sequence of $\lambda$; plot.it allows to display the curve to optimize; set.seed, sample, subsample and pi are options only for the stability selection criterion, see Section 2.2.2; finally, num.select allows, in the stability selection criterion, to bypass the option pi, selecting as many variables as indicated by num.select.

A summary, predict and plot S3 functions are also implemented for this class object.

This package is available from the author.

## B.2 qrcmNP

This package is inspired mainly by the methodology presented in Chapter 4. It implements a nonlinear Frumento and Bottai, 2016 method for quantile regression coefficient modeling (QRCM), in which quantile regression coefficients are described by (flexible) parametric functions of the order of the quantile. In the classical qrcm framework the linearity in $\boldsymbol{b}(p)$ and/or in $\boldsymbol{\theta}$ could be relaxed at a cost of more complicated expressions for the ojective and the gradient functions. Here, we propose an efficiently algorithm to use more flexible structures for the regression coefficients. With respect to the most famous function nlrq (quantreg package) our main function niqr implements the integrated quantile regression idea for nonlinear functions. As already known, this practice allows to estimate quantiles all at one time and not one at a time.

The main function in nonlinear QRCM is

```
niqr(fun, fun2, x0, X, y, control = list()).
```

It takes in input fun that is a linear or nonlinear function describing the $\boldsymbol{\beta}(p \mid \boldsymbol{\theta}) = g(p, \boldsymbol{\theta})$; fun2 that is a linear or nonlinear function describing the $Q(p \mid \boldsymbol{\beta}) = h(\boldsymbol{x}, \boldsymbol{\beta})$; X and y that are the model matrix and the response variable; control that is a list of control parameters.

This package also implements a penalized Frumento and Bottai, 2016 method for the qrcm, as proposed in this thesis. This package fits LASSO qrcm using pathwise coordinate descent algorithm. With respect to some other packages which implements the $L_1$-quantile regression (e.g. `rqPen`, `quantreg`) estimating quantiles one at a time our proposal allows to estimate the conditional quantile function parametrically estimating quantiles all at one and to do variable selction in the meanwhile. Here, two proposal to select the tuning parameter ($\lambda$) are implemented. In particular, the function `gof.piqr` allows to select the best tuning parameter minimizing several criteria (e.g. AIC, BIC).

The main function in penalized QRCM is

```
piqr(formula, formula.p = ~slp(p, 1), weights, data, s,
     nl = 70, display = TRUE, tol = 1e-06, maxit = 100).
```

It takes in input `formula` a two-sided formula of the form `y   x1 + x2 + ...`, that is a symbolic description of the quantile regression model; `formula.p` a one-sided formula of the form `  b1(p, ...)  + b2(p, ...)  + ...`, describing how quantile regression coefficients depend on $p$; `weigths`, `data` and `s` that are optionals arguments, i.e. vector of weights, data frame and 0/1 matrix that permits excluding some model coefficients, respectively; `nl` that is the maximum length of the sequence of $\lambda$; `display` allows to print some informations during each iteration; `tol` and `maxit` that are the tolerance for the convergence criterion and the maximum number of iterations, respectively.

A `summary`, `predict` and `plot` S3 functions are also implemented in this package for both the class objects `niqr` and `piqr`.

The package **qrcmNP** is available under the general public license (GPL $\geq$ 2) from the Comprehensive R Archive Network at http://CRAN.R-project.org/package=qrcmNP. Mantainer: Gianluca Sottile.

## B.3 clustEff

This package is inspired by the methodology presented in Appendix A. It implements a general algorithm to cluster coefficient functions (i.e. clusters of effects) obtained from a quantile regression coefficient modeling (QRCM; Frumento and Bottai, 2016). This algorithm is also used for clustering curves observed in time, as in functional data analysis. The objectives of this algorithm vary with the scenario in which it is used. In the univariate case, the goal is to perform variable selection. In the multivariate case, the algorithm can be used to describe relationships between outcomes and covariates. In the case of a functional data analysis the main objective is to cluster waves or any other function of time or space.

The main function is

```
clustEff(Beta, p, alpha, k, ask = FALSE, k.min = 1,
         k.max = min(10, (ncol(Beta) - 1)),
         cluster.effects = TRUE, Beta.lower = NULL,
         Beta.upper = NULL, step = c("both", "shape",
                                      "distance"),
         plot = TRUE, approx.spline = FALSE,
         nbasis = 50, method = "ward.D2").
```

This function takes in input `Beta` that is a matrix of dimension $N \times q$, where $q$ represents the number of curves to cluster and $N$ is either the length of percentiles used in the quantile regression or the length of the time vector; `p` that is the percentiles or the time vector; `alpha` that is the probability value used for computing the dissimilarity matrix; `k.min` and `k.max` that are the minimum and maximum number of clusters in which to look for the best; `cluster.effects` allows to select the framework in which to apply the clustering algorithm; `Beta.lower` and `Beta.upper`, used only if `cluster.effects=TRUE`, allows to select the best number of clusters using the lower and upper bands of each curve; `step` allows to select which measure has to be used to compute the dissimilarity matrix; `plot` allows to display information graphically.; `approx.spline` and `nbasis` allow to approximate curves using smooth splines with a specific number of basis, before computing the clustering alogrithm; finally, `method` allows to choose the agglomeration method to be used.

A `summary` and `plot` S3 functions are also available for this class object.

The package **clustEff** is available under the general public license (GPL $\geq$ 2) from the Comprehensive R Archive Network at http://CRAN.R-project.org/package=clustEff. Mantainer: Gianluca Sottile.

## B.4 islasso

This package is the result of a collaboration with Prof. Vito Muggeo and Dr. Giovanna Cilluffo. **islasso** is an R package that implements the methods proposed in Cilluffo et al., 2016 and in Cilluffo et al., 2017, that focuses on hypothesis testing and confidence interval estimation in lasso regression. This method, called IS-lasso, is based on the recent idea of induced smoothing that allows to obtain appropriate covariance matrix. the core of **islasso** package consist of a main algorithm implemented in

C++ to efficiently compute the Newthon-Raphson algorithm. The package allows to compute easily $p$-values and confidence intervals estimation in lasso regression. Gaussian, Binomial, Poisson and Gamma families with several links have already been implemented in this version of the package.

The main function of the package **islasso** is a wrapper function implemented to handle the formula interface usually used in R to create the $N \times p$-dimensional design matrix $\boldsymbol{X}$ and the $N$-dimensional response vector $y$

```
islasso(formula, family = gaussian(), lambda, data,
   weights, subset, offset, unpenalized, control = list()).
```

The arguments `family, lambda, weights, offset, unpenalized` and `control`, are then passed to the function `islasso.fit()`, the R function that performs the optimization steps of the algorithm

```
islasso.fit(X, y, family = gaussian, lambda,
    intercept = FALSE, weights = NULL, offset = NULL,
    unpenalized = NULL, control = list()).
```

The output of the function is an object of S3 class `"islasso"`. It is presented in a way that is easy to interpret for people familiar with standard `lm()` or `glm()` output. The main R functions are:

```
islasso(), islasso.fit(), print.islasso(),
summary.islasso(), predict.islasso(), plot.islasso(),
coef.islasso(), residuals.islasso(), fitted.islasso(),
deviance.islasso(), logLik.islasso(), AIC.islasso(),
confint.islasso(), model.matrix.islasso(), best.islasso().
```

The package is available from the author.

# Bibliography

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

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*.

Adelfio, G. et al. (2011). "FPCA algorithm for waveform clustering". In: *Journal of Communication and Computer* 8.6, pp. 494–502.

Adelfio, G. et al. (2012). "Simultaneous seismic wave clustering and registration". In: *Computers & geosciences* 44, pp. 60–69.

Akaike, H. (1974). "A new look at the statistical model identification". In: *IEEE Trans. on Automatic Control* 19, pp. 716–723.

Alexander, D.H. and K. Lange (2011). "Enhancements to the ADMIXTURE algorithm for individual ancestry estimation". In: *BMC bioinformatics* 12.1, p. 246.

Allen, A.R. et al. (2010). "Compilation of a panel of informative single nucleotide polymorphisms for bovine identification in the Northern Irish cattle population". In: *BMC genetics* 11.1, p. 5.

Belloni, A. and V. Chernozhukov (2011). "L1-penalized quantile regression in high-dimensional sparse models". In: *The Annals of Statistics* 39.1, pp. 82–130.

Bertolini, F. et al. (2015). "Combined use of principal component analysis and random forests identify population-informative single nucleotide polymorphisms: application in cattle breeds". In: *Journal of Animal Breeding and Genetics* 132.5, pp. 346–356.

Bertolini, F. et al. (2017). "Preselection statistics and Random Forest classification identify population informative single nucleotide polymorphisms in cosmopolitan and autochthonous cattle breeds". In: *Animal*, pp. 1–8.

Bottai, M., N. Orsini, and M. Geraci (2015). "A gradient search maximization algorithm for the asymmetric Laplace likelihood". In: *Journal of Statistical Computation and Simulation* 85.10, pp. 1919–1925.

Bottai, M. et al. (2011). “Percentiles of inspiratory capacity in healthy nonsmokers: a pilot study”. In: *Respiration* 82.3, pp. 254–262.

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.

Bowcock, A.M. et al. (1994). “High resolution of human evolutionary trees with polymorphic microsatellites.” In: *nature* 368.6470, p. 455.

Bühlmann, P., M. Kalisch, and L. Meier (2014). “High-dimensional statistics with a view toward applications in biology”. In: *Annual Review of Statistics and Its Application* 1, pp. 255–278.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*.

Chaudhuri, P. (1991). “Global nonparametric estimation of conditional quantile functions and their derivatives”. In: *Journal of multivariate analysis* 39.2, pp. 246–269.

Chen, J. and Z. Chen (2008). “Extended Bayesian information criteria for model selection with large model spaces”. In: *Biometrika* 95, pp. 759–771.

Cilluffo, G. et al. (2016). “The induced smoothed LASSO”. In: *Proc. 31st International Workshop on Statistical Modelling*. Vol. 1. Rennes, France.

Cilluffo, G. et al. (2017). “Score inference in LASSO regression”. In: *Proc. 32nd International Workshop on Statistical Modelling*. Vol. 1. Groningen, Netherlands.

Clogg, C.C., E. Petkova, and A. Haritou (1995). “Statistical methods for comparing regression coefficients between models”. In: *American Journal of Sociology* 100.5, pp. 1261–1293.

Craven, P. and G. Wahba (1979). “Smoothing noisy data with spline functions”. In: *Numerische Mathematik* 31, pp. 377–403.

Dimauro, C. et al. (2013). “Use of the canonical discriminant analysis to select SNP markers for bovine breed assignment and traceability purposes”. In: *Animal genetics* 44.4, pp. 377–382.

Dimauro, C. et al. (2015). “Selection of discriminant SNP markers for breed and geographic assignment of Italian sheep”. In: *Small Ruminant Research* 128, pp. 27–33.

Efron, B. et al. (2004). “Least angle regression”. In: *The Annals of Statistics* 32, pp. 407–499.

Fan, J. and R. Li (2001). “Variable selection via nonconcave penalized likelihood and its oracle properties”. In: *Journal of the American Statistical Association* 96, pp. 1348–1360.

Friedman, J., T. Hastie, and R. Tibshirani (2010). “Regularization Paths for Generalized Linear Models via Coordinate Descent”. In: *Journal of Statistical Software* 33.1, pp. 1–22.

Frumento, P. (2017). *qrcm: Quantile Regression Coefficients Modeling*. R package version 2.1. URL: https://CRAN.R-project.org/package=qrcm.

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.

Heaton, M.P. et al. (2014). “SNPs for parentage testing and traceability in globally diverse breeds of sheep”. In: *PloS one* 9.4, e94851.

Hoerl, A.E. and R.W. Kennard (1970). “Ridge regression: Biased estimation for nonorthogonal problems”. In: *Technometrics* 12.1, pp. 55–67.

Hulsegge, B. et al. (2013). “Selection of SNP from 50K and 777K arrays to predict breed of origin in cattle”. In: *Journal of animal science* 91.11, pp. 5128–5134.

Jacques, J. and C. Preda (2014). “Functional data clustering: a survey”. In: *Advances in Data Analysis and Classification* 8.3, pp. 231–255.

Jakobsson, M. and N.A. Rosenberg (2007). “CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure”. In: *Bioinformatics* 23.14, pp. 1801–1806.

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.

Knight, K. and W. Fu (2000). “Asymptotics for lasso-type estimators”. In: *Annals of statistics*, pp. 1356–1378.

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.

Kruskal, W.H. and W.A. Wallis (1952). “Use of ranks in one-criterion variance analysis”. In: *Journal of the American statistical Association* 47.260, pp. 583–621.

Kuehn, L.A. et al. (2011). “Predicting breed composition using breed frequencies of 50,000 markers from the US Meat Animal Research Center 2,000 Bull Project”. In: *Journal of animal science* 89.6, pp. 1742–1750.

Lee, E.R., H. Noh, and B.U. Park (2014). "Model selection via bayesian information criterion for quantile regression models." In: *Journal of the American Statistical Association* 109, pp. 216–229.

Li, Y. and J. Zhu (2008). "L1-norm quantile regression". In: *Journal of Computational and Graphical Statistics* 17.1, pp. 163–185.

Mastrangelo, S. et al. (2012). "Study of polymorphisms in the promoter region of ovine $\beta$-lactoglobulin gene and phylogenetic analysis among the Valle del Belice breed and other sheep breeds considered as ancestors". In: *Molecular biology reports* 39.1, pp. 745–751.

Mastrangelo, S. et al. (2014). "Genome wide linkage disequilibrium and genetic structure in Sicilian dairy sheep breeds". In: *BMC genetics* 15.1, p. 108.

Mastrangelo, S. et al. (2017). "Genome-wide analysis in endangered populations: a case study in Barbaresca sheep". In: *Animal*, pp. 1–10.

Meinshausen, N. and P. Bühlmann (2010). "Stability selection". In: *Journal of the Royal Statistical Society: Series B* 72.4, pp. 417–473.

Negrini, R. et al. (2009). "Assessing SNP markers for assigning individuals to cattle populations". In: *Animal genetics* 40.1, pp. 18–26.

Nicolazzi, E.L. et al. (2015). "SNPchiMp v. 3: integrating and standardizing single nucleotide polymorphism data for livestock species". In: *BMC genomics* 16.1, p. 283.

Osborne, M.R., B. Presnell, and B.A. Turlach (2000a). "A new approach to variable selection in least squares problems". In: *IMA journal of numerical analysis* 20.3, pp. 389–403.

— (2000b). "On the lasso and its dual". In: *Journal of Computational and Graphical statistics* 9.2, pp. 319–337.

Paschou, P. et al. (2007). "PCA-correlated SNPs for structure identification in worldwide human populations". In: *PLoS genetics* 3.9, e160.

R Development Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Ramsay, J.O. (2006). *Functional data analysis*. Wiley Online Library.

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

Reeves, G. and M. C. Gastpar (2013). "Approximate sparsity pattern recovery: Information-theoretic lower bounds". In: *IEEE Transactions on Information Theory* 59, pp. 3451–3465.

Rosenberg, N.A. (2005). "Algorithms for selecting informative marker panels for population assignment". In: *Journal of computational biology* 12.9, pp. 1183–1201.

Rousset, F. (2008). "genepop'007: a complete re-implementation of the genepop software for Windows and Linux". In: *Molecular ecology resources* 8.1, pp. 103–106.

Sangalli, L.M. et al. (2009). "A case study in exploratory functional data analysis: geometrical features of the internal carotid artery". In: *Journal of the American Statistical Association* 104.485, pp. 37–48.

Schwarz, G. (1978). "Estimating the dimension of a model". In: *The Annals of Statistics* 6, pp. 461–464.

Shriver, M.D. et al. (1997). "Ethnic-affiliation estimation by use of population-specific DNA markers." In: *American journal of human genetics* 60.4, p. 957.

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

Sottile, G. and G. Adelfio (2017). "A new approach for clustering of effects in quantile regression". In: *Proc. 32nd International Workshop on Statistical Modelling*. Vol. 2. Groningen, Netherlands.

Sottile, G. and V.M.R Muggeo (2016). "Tuning parameter selection in LASSO regression". In: *Proc. 31st International Workshop on Statistical Modelling*. Vol. 2. Rennes, France.

Sottile, G. et al. (2017). "Penalized classification for optimal statistical selection of markers from high-throughput genotyping: application in sheep breeds". In: *animal*, pp. 1–8.

Stamey, T. A et al. (1989). "Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients." In: *The Journal of Urology* 141, pp. 1076–1083.

Su, W., M. Bogdan, and E. Candes (2015). "False discoveries occur early on the lasso path". In: *arXiv preprint arXiv:1511.01957*.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society, Series B* 58, pp. 267–288.

— (2011). "Regression shrinkage and selection via the lasso: a retrospective". In: *Journal of the Royal Statistical Society: Series B* 73.3, pp. 273–282.

Tibshirani, R. et al. (2005). "Sparsity and smoothness via the fused lasso". In: *Journal of the Royal Statistical Society: Series B* 67.1, pp. 91–108.

Tikhonov, Andrey Nikolayevich (1943). "On the stability of inverse problems". In: *Dokl. Akad. Nauk SSSR*. Vol. 39, pp. 195–198.

Tolone, M. et al. (2012). "Genetic diversity and population structure of Sicilian sheep breeds using microsatellite markers". In: *Small Ruminant Research* 102.1, pp. 18–25.

Vichi, M. and G. Saporta (2009). "Clustering and disjoint principal component analysis". In: *Computational Statistics & Data Analysis* 53.8, pp. 3194–3208.

Wainwright, M. J. (2009a). "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting". In: *IEEE Transactions on Information Theory* 55, pp. 5728–5741.

— (2009b). "Sharp thresholds for high-dimensional and noisy sparsity recovery using $l_1$-constrained quadratic programming (Lasso)". In: *IEEE transactions on information theory* 55, pp. 2183–2202.

Wang, H., R. Li, and C.-L. Tsai (2007b). "Tuning parameter selectors for the smoothly clipped absolute deviation method". In: *Biometrika* 94, pp. 553–568.

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

Wang, L., Y. Wu, and R. Li (2012). "Quantile regression for analyzing heterogeneity in ultra-high dimension". In: *Journal of the American Statistical Association* 107.497, pp. 214–222.

Wilkinson, S. et al. (2011). "Evaluation of approaches for identifying population informative markers from high density SNP chips". In: *BMC genetics* 12.1, p. 45.

Wu, Y. and Y. Liu (2009). "Variable selection in quantile regression". In: *Statistica Sinica*, pp. 801–817.

Yuan, M. and Y. Lin (2006). "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society: Series B* 68.1, pp. 49–67.

Zhang, Y., R. Li, and C.-L. Tsai (2010). "Regularization parameter selections via generalized information criterion". In: *Journal of the American Statistical Association* 105.489, pp. 312–323.

Zhao, P. and B. Yu (2006). "On model selection consistency of Lasso". In: *Journal of Machine learning research* 7, pp. 2541–2563.

Zheng, Q., C. Gallagher, and K.B. Kulasekera (2013). "Adaptive penalized quantile regression for high dimensional data". In: *Journal of Statistical Planning and Inference* 143.6, pp. 1029–1038.

Zou, H. (2006). "The adaptive lasso and its oracle properties". In: *Journal of the American Statistical Association* 101, pp. 1418–1429.

Zou, H. and T. Hastie (2005). "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society, Series B* 67, pp. 301–320.