UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXIX

Extensions of marginal quantile regression to the analysis of dependent data

Direttore della Scuola: Monica Chiogna

Supervisori: Matteo Bottai

Dottorando: Davide Bossoli

15/11/2016

Acknowledgments

I would like to thank several people for their support during these years.

Matteo Bottai, my supervisor of master and doctoral educations. Thanks for giving me the opportunity to spend part of my PhD at the Karolinska Institutet, for all the things you taught me and for your skillful guidance through this PhD project.

My colleagues from the 29th PhD cycle, **Paolo, Mirko, Lucia, Khanh, Elisa, Claudia**. I am glad to have shared this path with amazing people like you. Thanks for all the support you provided during these years.

My colleagues from University of Padua and Karolinska Institutet, **Erlis, Vera, Clovis, Mauro, Xin, Daniel, Celia, Michele, Andrea, Alessio, Nicola, Jonas, Michael and Ulf.** Thank you for all the interesting discussions we had during this years, along with your enormous support.

A special thanks to **Paolo Frumento**, for always being willing to help me. Words cannot express how grateful I am for what you did during these years. They would have been much thougher without your advices.

My parents, **Sauro and Elisabetta**, and my brother, **Marco**, for being the pillars of my support.

My friends, **Jacopo, Sofia, Erika, Jessica and Massimiliano**, for being always there for me.

Above all, I would like to thank **Lucia**, for her unbounded love and cheerfulness. This journey would not have been the same without you.

Abstract

Dependent data arise frequently in applied research. When quantile regression is the statistical method of choice, several approaches have been proposed that can accommodate dependence among observations. Cluster bootstrap is one of the most popular among them. While practical, this method is generally inefficient and computationally demanding, especially when the number of clusters is large. When the primary interest is on marginal quantiles, estimating equations have been proposed that model the association between the sign of the regression residuals with the Pearson's correlation coefficient. The latter, however, is an inadequate measure of dependence between binary variables because of its range depends on their marginal probabilities. Instead, we propose to model a working association matrix through odds ratios, which are popular measures of association of binary outcomes. Different working structures can be easily estimated by suitable logistic regression models. These structures can be parameterized and may depend on covariates and clusters. Simulations demonstrated that the efficiency of the estimator increases as the working correlation structure approaches the true one. We extend the proposed method to penalized estimating equations, they have increasingly been used to reduce model complexity in several applications. We focus on penalized smoothly clipped absolute deviation models for feature selection and reduced-rank penalized smoothing splines. Simulations showed that the proposed methods potentially improve the performance of the marginal quantile regression estimator. When the correlation structure is correctly specified the estimator's efficiency increases, similarly to what happens in the non-penalized case tackled in the first part of the thesis. We applied the proposed methods to data from a study on cognitive behavior and treatment in patients with obsessive compulsive disorder. To show the full potential of the methods, we modified the original data in some of the analyses.

Abstract

Nella ricerca applicata, i dati dipendenti sono molto frequenti. Nella regressione quantile sono stati proposti diversi approcci per tenere in considerazione la dipendenza tra le osservazioni. Uno dei metodi più utilizzato è il cluster bootstrap, sebbene sia generalmente inefficiente e computazionalmente dispendioso, soprattutto quando il numero di cluster è elevato. Quando l'interesse principale è sui quantili marginali, sono state proposte delle equazioni di stima che modellizzano l'associazione tra i segni dei residui di regressione attraverso il coefficiente di correlazione di Pearson. Tuttavia, questa misura è inadeguata per la dipendenza tra variabili binarie, poichè il suo range dipende dalle loro probabilità marginali. Nella prima parte della tesi viene proposta una matrice di dipendenza definita attraverso gli odds ratios. Le diverse strutture di associazione possono essere stimate attraverso modelli di regressione logistica e possono essere parametrizzate per dipendere da covariate e gruppi. Attraverso uno studio di simulazione viene mostrato che l'efficienza degli stimatori aumenta quando la matrice di associazione è vicina a quella vera. Nella seconda parte della tesi si estende questo metodo ad equazioni di stima penalizzate, che sono utilizzate per ridurre automaticamente la complessità del modello stimato. In quest'ultima parte del lavoro si concentra l'attenzione sui modelli con penalità smoothly clipped absolute deviation per la selezione automatica dei predittori e sulle spline penalizzate tramite riduzione di rango. Attraverso uno studio di simulazione mostriamo che questi metodi hanno performance migliori rispetto a quelli senza penalizzazione. Quando la struttura di associazione è vicina a quella vera l'efficienza dello stimatore aumenta, analogamente al metodo proposto nella prima parte della tesi. I metodi discussi nella tesi sono stati applicati ad un dataset proveniente da uno studio sul comportamento cognitivo in pazienti con disturbi ossessivi-compulsivi; inoltre, per mostrare il massimo potenziale dei metodi penalizzati, si è provveduto a modificare il dataset originale in alcune analisi.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

This thesis focuses on extensions of population-averaged (marginal) quantile regression. The interest in quantile regression has substantially grown in recent years for several reasons. First, it can describe the whole conditional distribution of the response variable. Second, it is much less sensitive to outliers. Third, it is equivariant to monotone transformations. This property has been used effectively in many settings (Bottai et al. (2010), de Luca and Boccuzzo (2014)). Quantiles have been used as an alternative to hazard ratio to describe survival curves of groups of individuals, with a number of applications in econometrics and epidemiology (Koenker and Bilias (2002), Bellavia et al. (2015), Bellavia et al. (2013)). They are also commonly used to study skewed distributions such as income, health care expenditures and unemployment rates (Centeno and Novo (2006), Wang (2011), Buchinsky (1994)).

Dependent data arise frequently in applied research. For example in econometrics, panel data are used to assess time trends and to forecast. A random sample of individuals is selected from a population, and each subject is observed at multiple occasions over time. The multiple observations within each subject may be dependent, while observations from different subjects are independent. This sampling scheme is also widely used in randomized clinical trials to evaluate treatment effects. In genetic studies, siblings may be included in the study to separate heredity effects from environmental effects. The dependence can also be caused by spatial factors. Units that are geographically near may behave similarly, for instance trees in a forest or individuals living in the same area. It is important to take into account the dependence of the data when it is present. Ignoring it may lead to wrong coverage of the confidence intervals and loss of statistical efficiency.

The first contribution of the thesis consists of a novel method to estimate marginal quantile regression where the dependence between observations within the same cluster is modeled through a working odds-ratio matrix. Different working structures can be easily estimated by suitable logistic regression models. These structures can be parametrized to depend on covariates and clusters. When the working correlation is close to the true one, the efficiency of the regression-model estimator increases. We analyzed data from a randomized clinical trial on cognitive behavior. The dataset consisted of 95 families with a child aged 8-12 years with a principal diagnosis of generalised anxiety, panic disorder, separation anxiety, social phobia or specific phobia. Participants were randomised to 10 weeks of internet-based cognitive theraphy (ICBT) with therapist support (70 families), or to a waitlist control condition (25 families). At weekly intervals, the amount of difficulties in emotion regulation scale (DERS) was measured on a 0-100 scale, where 0 indicated no difficulties and 100 extreme difficulties. The maximum number of repeated measurements was 14. The main interest of the study was to assess the relationship between the treatment and DERS.

This method was the basis for the remaining contributions of the thesis, two cases of penalized marginal quantile regression: reduced-rank penalized splines smoothing and regularized models for feature selection. The former is common in various fields. In panel data, complex nonlinear trends might be modeled using certain fixed periodic functions or regression splines. However, fixed functions that are not data-driven might not be flexible enough to capture nonlinear trends. Regression splines are flexible but might lead to overfitting. Instead, penalized splines would provide enough flexibility while prevent overfitting by automatically selecting the optimal degree of smoothness of a regression spline. The spline complexity is controlled by a penalty term chosen through cross-validation or by minimization of a measure that balances goodness of fit and complexity. The penalty term finds the optimal tradeoff between spline complexity and fit of the predicted function. In contrast to regression splines, penalized splines are less influenced by the initial choice of the number of knots and the degree of the chosen spline.

Regularized models for feature selection are used to select the most important predictors in a regression models. The model complexity is controlled by a penalty term chosen through either cross-validation or by minimizing a measure of balance between goodness of fit and model complexity. This term finds the optimal tradeoff between spline complexity and fit of the predicted function. Regularized models are used when the number of predictors is large and classical selection methods such as stepwise regression cannot be used. Analogously to reduced-rank penalized splines smoothing methods, they can be used also to

automatically select the splines complexity in a regression spline model.

We applied the penalized method to the cognitive-behavior study for treatment of obsessive compulsive disorder. To show their full potential, we added a predictor that was nonlinearly associated with the response. We modelled this association through regression splines and we chose the splines complexity through the penalized methods. We also added a large number of uncorrelated predictors to study the behavior of the regularized model.

The outline of the thesis is as follows. Chapter 2 presents some basic notions of quantile regression, generalized estimating equations (GEE), regression splines and regularized models. Chapter 3 introduces a novel approach to estimate unpenalized marginal quantile models using an odds-ratio matrix. Chapter 4 discusses penalized marginal quantile regression splines and regularized marginal quantile models for feature selection in high dimensional data. We conclude the thesis with a discussion on Chapter 5.

1.2 Main contributions of the thesis

The main contributions of the thesis can be summarized as follows.

1. Development of a novel method to estimate marginal quantile regression using a working odds-ratio matrix, with application to a randomized clinical trial study on cognitive behavior.

2. Development of reduced-rank penalized splines smoothing for marginal quantile regression.

3. Development of regularized marginal quantile regression for feature selection.

4. Application of the penalized methods to a randomized clinical trial study on cognitive behavior where we added a nonlinear association along with a set of uncorrelated predictors.

5. Implementation of R functions to estimate the proposed models.

Chapter 2

Background

In this chapter we review some important concepts that will be used in the rest of the thesis. In Section 2.1 we present quantile regression and follow the book of Koenker (2005). In Section 2.2 we review generalized estimating equations, and we use mainly the book of Fitzmaurice et al. (2008). In Section 2.3 and 2.4 we present regression splines and methods for feature selection. These concepts will be the basis of penalized methods of Chapter 4 and are taken from the books of Seber and Wild (2003) and Hastie et al. (2011).

2.1 Quantile Regression

Any real-valued random variable X may be characterized by its distribution function

$$F(x) = P(X \leq x)$$

whereas for any $0 < \tau < 1$,

$$F^{-1}(\tau) = inf\{x : F(x) \geq \tau\}$$

is called the τth quantile of X. The median, $F^{-1}(0.5)$ plays the central role. The quantiles arise from a simple optimization problem. Consider the following problem: a point estimate is required for a random variable with distribution function F. If loss is described by the piecewise linear function illustrated in Figure 2.1

$$\psi(u) = u(\tau - I(u < 0))$$

for some $\tau \in (0, 1)$, find $\hat{x}$ to minimize expected loss.

The aim is to minimize

$$E(\psi(X-\hat{x})) = (\tau-1)\int_{-\infty}^{\hat{x}}(x-\hat{x})dF(x) + \tau\int_{\hat{x}}^{\infty}(x-\hat{x})dF(x).$$

Differentiating with respect to $\hat{x}$, we have

$$0 = (1-\tau)\int_{-\infty}^{\hat{x}}dF(x) - \tau\int_{\hat{x}}^{\infty}dF(x) = F(\hat{x}) - \tau.$$

Because F is monotone, any element of $\{x : F(x) = \tau\}$ minimized expected loss. When the solution is unique, $\hat{x} = F^{-1}(\tau)$; otherwise, we have an "interval of τth quantiles" from which the smallest element must be chosen to adhere to the convention that the empirical quantile function be left-continuous.

It is natural that an optimal point estimator for asymmetric linear loss should lead us to the quantiles. In the symmetric case of absolute value loss it is well known to yield the median. When loss is linear and asymmetric, we prefer a point estimate more likely to leave us on the flatter of two brances of marginal loss. Thus, for example, if an underestimate is marginally three times more costly than an overestimate, we will chose $\hat{x}$ so that $P(X \leq \hat{x})$ is three times greater than $P(X > \hat{x})$ to compensate. That is, we will choose $\hat{x}$ to be the 75th percentile of F. When F is replaced by the empirical distribution function

$$F_n(x) = n^{-1}\sum_{i=1}^{n} I(X_i \leq x),$$

we may still choose $\hat{x}$ to minimize expected loss:

$$\int \psi(x-\hat{x})dF_n(x) = n^{-1}\sum_{i=1}^{n}\psi(x_i-\hat{x})$$

and doing so yields the τth sample quantile. When τn is an integer there is again some ambiguity in the solution, but this generally is of little practical consequence. Much more important is the fact that the problem of finding the τth sample quantile, that might seem inherently tied to the notion of an ordering of the sample observations, has been expressed as the solution to a sample optimization problem. This lead to more general methods of estimating models of conditional quantile functions. Least squares offers a template for this development. Knowing that the sample mean solves the problem

$$\min_{\mu\in R}\sum_{i=1}^{n}(y_i-\mu)^2$$

suggests that, if we are willing to express the conditional mean of y given x as

$\mu(x) = x^T\beta$, then β may be estimated by solving

$$\min_{\beta \in R^p} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2.$$

Similarly, since the τth sample quantile, $\hat{\alpha}(\tau)$, solves

$$\min_{\alpha \in R} \sum_{i=1}^{n} \psi(y_i - \alpha)$$

we are lead to specifying the τth conditional quantile function as $Q_y(\tau|x) = x^T\beta_\tau$, and to consideration of $\hat{\beta}_\tau$ solving

$$\min_{\beta \in R^p} \sum_{i=1}^{n} \psi(y_i - x_i^T \beta).$$

A possible parametric link between the minimization of the sum of the absolute deviates and the maximum likelihood theory is given by the asymmetric Laplace distribution (ALD). We say that a random variable Y is distributed as an ALD with parameters μ, σ and τ and we write it $Y \sim ALD(\mu, \sigma, \tau)$ if the corresponding density is given by

$$f(y|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\psi\left(\frac{y-\mu}{\sigma}\right)\right\},$$

where $\psi(u) = u(\tau - I(u \leq 0))$ is the loss function, $0 < \tau < 1$ is the skewness parameter, $\sigma > 0$ is the scale parameter, and $-\infty < \mu < \infty$ is the location parameter. Set $\mu_i = x_i^T\beta$, and $y = (y_1, \ldots, y_N)$. Assuming that $y_i \sim ALD(\mu_i, \sigma, \tau)$ then the likelihood for N independent observations is, bar a proportionality constant,

$$L(\beta, \sigma, y, \tau) \propto \sigma^{-1} \exp\left\{-\sum_{i=1}^{n} \psi\left(\frac{y_i - \mu_i}{\sigma}\right)\right\}$$

If we consider σ a nuisance parameter, the maximization of the ALD likelihood is equivalent to the minimization of the loss function. The quantile regression estimator can be seen as the quasi-maximum likelihood estimator of an asymmetric Laplace distribution. As we shall see in the next section, this has important connections to generalized estimating equations (GEE).

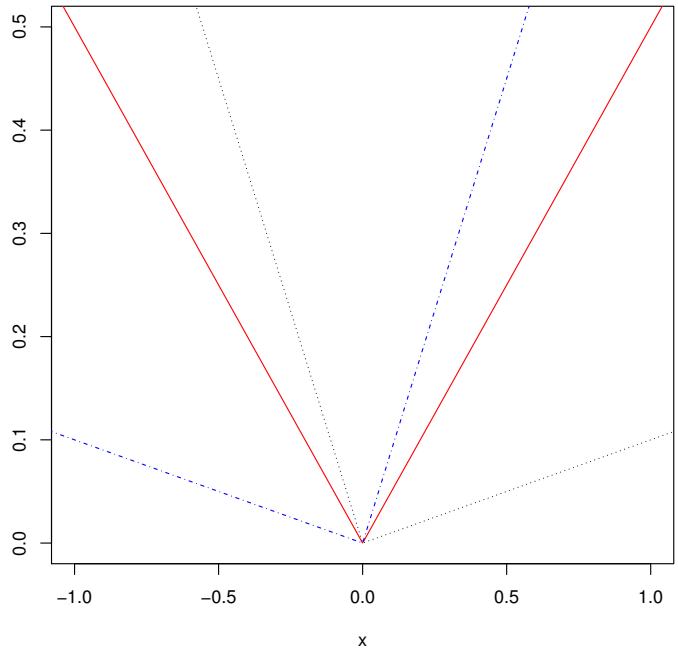

Figure 2.1: Quantile regression loss function for $\tau = 0.5$ (red), $\tau = 0.1$ (blue) and $\tau = 0.9$ (black).

2.2 Generalized Estimating Equations

Generalized estimating equations are a common approach to account for dependence in the data. They provide marginal estimates, where the term marginal is used to emphasize that the model for the mean response at each occasion depends only on the covariates of interest, and does not incorporate dependence on previous responses. This is constrast to classical mixed-effect (conditional) models, where the mean response is modeled not only as a function of covariates but is conditional also on random effects.

Marginal models for dependent data separately model the mean response and the dependence among observations. In a marginal model, the goal is to make inferences about the former, whereas the latter is regarded as a nuisance characteristic of the data that must be taken into account in order to make correct inferences about changes in the population mean response over time.

A marginal model for dependent data has the following three-part specification, with the subscript i and j denoting the cluster and the observation within the cluster, respectively:

- The conditional expectation of each response, $E(Y_{ij}|X_{ij}) = \mu_{ij}$, is assumed to depend on the covariates through a known link function $h^{-1}(\cdot)$,

$$h^{-1}(\mu_{ij}) = \eta_{ij} = X_{ij}^T\beta,$$

  where β is a $p \times 1$ vector of marginal regression parameters.

- The conditional variance of each Y_{ij}, given X_{ij}, is assumed to depend on the mean according to

$$Var(Y_{ij}) = \phi\nu(\mu_{ij}),$$

  where $\nu(\cdot)$ is a known "variance function" and ϕ is a scale parameter that may be fixed and known or may need to be estimated.

- The conditional dependence of the observations is assumed to be a function of an additional vector of association parameters, say α.

A crucial aspect of marginal models is that the mean response and the dependence among observations are modeled separately. This separation has important implications for interpretation of the regression parameters in the model for the mean response. In particular, the regression parameters, β, in the marginal model have so-called population-averaged interpretations. That is, they describe how the mean response in the population is related to the covariates. For example, regression parameters in a marginal model might have interpretation

in terms of contrasts of the changes in the mean responses in subpopulations (e.g., different treatment, intervention or exposure groups).

The three-part marginal model specification does not require full distributional assumptions for the data, only a regression model for the mean response. The avoidance of distributional assumptions leads to a method of estimation known as generalized estimating equations. Thus, the GEE approach can be though of as providing a convenient alternative to ML estimation. The GEE approach is a multivariate generalization of the quasi-likelihood approach for generalized linear models. To better understand this connection to quasi-likelihood estimation, in the following we briefly outline the quasi-likelihood approach for generalized linear models for a univariate response before discussing its extension to multivariate responses.

In the following, we now assume N independent observations of a scalar response variable, Y_i. Associated with the response, Y_i, there are p covariates, $X_{1i}, \ldots, X_{pi}$. We assume that the primary interest is in relating the mean of Y_i, $\mu_i = E(Y_i|X_{1i}, \ldots, X_{pi})$ to the covariates via an appropriate link function,

$$h^{-1}(\mu_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi},$$

where the link function $h^{-1}(\cdot)$ is a known function. The assumption that Y_i has an exponential family distribution has implications for the variance of Y_i. In particular, a feature of exponential family distributions is that the variance of Y_i can be expressed in terms of a known function of the mean and a scale parameter,

$$V(Y_i) = \phi\nu(\mu_i),$$

where the scale parameter $\phi > 0$. The variance function, $\nu(\mu_i)$ describes how the variance of the response is functionally related to the mean of Y_i. Next, we consider estimation of β. Assuming Y_i follows an exponential family density with $Var(Y_i) = \phi\nu(\mu_i)$, the maximum likelihood estimator of β is obtained as the solution to the likelihood score equations,

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta}\right)' \frac{1}{\phi\nu(\mu_i)}\{Y_i - \mu_i(\beta)\} = 0,$$

where $\partial\mu_i/\partial\beta$ is the $1 \times p$ vector of derivatives, $\partial\mu_i/\partial\beta_i$, $i = 1, \ldots, p$. Interestingly, the likelihood equations for generalized linear models depend only on the mean and variance of the response (and the link function). It was suggested using them as "estimating equations" for any choice of link or variance function, even when the particular choice of variance function does not correspond to an exponential family distribution. It was proposed estimating β by solving the

quasi-likelihood equations,

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right)' V_i^{-1} (Y_i - \mu_i(\beta)) = 0.$$

For any choice of weights, V_i, the quasi-likelihood estimator of β, say $\hat{\beta}$, is consistent and asymptotically normal. The choice of weights, $V_i = Var(Y_i)$ yields the estimator with smallest variance among all estimators in this class. In summary, the estimation of β does not require distributional assumptions on the response. Quasi-likelihood estimation only requires correct specification of the model for the mean to yield consistent and asymptotically normal estimators of β. That is, a key property of quasi-likelihood estimators is that they are consistent even when the variance of the response has been misspecified, that is, $V_i \neq Var(Y_i)$. Specifically, it can be shown that the asymptotic distribution of $\hat{\beta}$ satisfies

$$\sqrt{n}(\hat{\beta} - \beta) \to N(0, C_\beta),$$

where

$$C_\beta = \lim_{n \to infty} I_0^{-1} I_1 I_0^{-1},$$

$$I_0 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right)' V_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta} \right)$$

and

$$I_1 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right)' V_i^{-1} \, Var(Y_i) V_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta} \right)$$

Consistent estimators of the asymptotic covariance of the estimated regression parameters can be obtained using the empirical estimator of C_β. The empirical variance estimator is obtained by evaluating $\partial \mu_i / \beta$ at $\hat{\beta}$ and substituting $(Y_i - \hat{\mu_i})^2$ for $Var(Y_i)$; this is widely known as the sandwich variance estimator. It can be shown that the same asymptotic distribution holds when V_i is estimated rather than known, with V_i replaced by estimated weights, say $\hat{V}_i$. The multivariate extension of this quasi-likelihood approach can be obtained by replacing Y_i and μ_i by their corresponding multivariate counterparts, and using a matrix of weights V_i. In the multivariate case, V_i depends not only on β but also on the pairwise associations among the observations in the ith cluster. In general, the assumed covariance matrix among the responses within a cluster V_i can be specified as

$$V_i = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2},$$

where $A_i^{1/2} = diag\{\nu(\mu_{ij})\}$ is a diagonal matrix with diagonal elements $\nu(\mu_{ij})$. For the correlation matrix $R_i(\alpha)$, α represents a vector of parameters associated

with a specified model for $Corr(Y_i)$, with typical element

$$\rho_{ist} = \rho_{ist}(\alpha) = Corr(Y_{is}, Y_{it}; \alpha), \quad s \neq t.$$

In the GEE approach, V_i is usually referred to as a "working covariance", where the term "working" is used to emphasize that V_i is only an approximation to the true covariance matrix. Note that if $R_i(\alpha) = I$, the $n_i \times n_i$ identity matrix, then the GEE reduces to the quasi-likelihood estimating equations for a generalized linear model that assume the repeated measures are independent. Some common examples of models for the correlation are:

- exchangeable, $\rho_{ist} = \alpha \; \forall \; s < t$;
- Toeplitz, $\rho_{ist} = \alpha_{|s-t|}$
- first-order autoregressive (AR-1), $\rho_{ist} = \alpha^{|t-s|}$, where the correlation decreases as the time between measurements increases
- unstructured, $\rho_{ist} = \alpha_{st}$

For the special case where the outcome is binary, an alternative to the correlation as a measure of association between pairs of binary responses is the odds ratio. In Chapter 3, the odds-ratio has many desirable properties and a more straightforward interpretation.

2.3 Regression Spline

A spline function (q-spline) is a piecewise or segmented polynomial of degree q with $q-1$ continuous derivatives at the changepoints. In the spline literature the changepoints are called knots. We consider only functions of a single variable x. Suppose $y = f(x) + \epsilon$, but $f(x)$ is completely unknown. When faced with a well-behaved curved trend in a scatterplot, most statisticians would fit a low-order polynomial in x and go on to make inferences which are conditional upon the truth of the polynomial model. This is done largely because the family of polynomials is a family of simple curves which is flexible enough to approximate a large variety of shapes. More technically, one could appeal to the Stone-Weierstrass theorem (see Royden [1968:174]), from which it follows that any continuous function on $[a, b]$ can be approximated arbitrarily closely by a polynomial. In pratical terms, a better approximation can be obtained by increasing the order of the polynomial. The cost incurred for doing this is the introduction of additional parameters and some oscillation between data points. Spline functions can be viewed as another way of improving a polynomial approximation. Suppose we are prepared to assume that $f(x)$ has $q+1$ continuous

derivatives on $[a, b]$. Then by Taylor's theorem,

$$f(x) = \sum_{j=0}^{q} \phi_j x^j + r(x),$$

where

$$\sum_{j=0}^{q} \phi_j x^j = \sum_{j=0}^{q} \frac{1}{j!} \frac{\partial^j f(a)}{\partial x^j} (x-a)^j,$$

and the remainder is given by

$$\begin{aligned} r(x) &= \frac{1}{q!} \int_a^x \frac{\partial^{q+1} f(t)}{\partial t^{q+1}} (x-t)^q dt \\ &= \frac{1}{q!} \int_a^b \frac{\partial^{q+1} f(t)}{\partial t^{q+1}} (x-t)_+^q dt \end{aligned}$$

where $z_+^q = z^q$ if $z \geq 0$ and 0 otherwise. Use of an order-q polynomial model ignores the remainder term. Instead of improving the polynomial approximation by increasing the order of the polynomial, it can be achieved by approximating the remainder term. Consider a partition $a \equiv \alpha_0 < \alpha_1 < \cdots < \alpha_{D-1} < \alpha_D \equiv b$ and approximate the $(q+1)$th derivative of f by a step function, then

$$r(x) \approx \frac{1}{q!} \sum_{d=1}^{D-1} \xi_d (x - \alpha_d)_+^q.$$

The final term of the summation has been omitted as $(x-b)_+^q = 0$ for $x \leq b$. The function $f(x)$ can now be approximated by

$$f(x) \approx \sum_{j=0}^{q} \phi_j x^j + \frac{1}{q!} \sum_{d=1}^{D-1} \xi_d (x - \alpha_d)_+^q. \tag{2.1}$$

This spline approximation can be improved either by increasing the number of knots, or by allowing the data to determine the position of the knots, thus improving the approximation of the integral in the remainder term $r(x)$. The advantage of using splines over increasing the order of the polynomial, in some situations, is that the oscillatory behavior of high-order polynomials can be avoided with splines.

When the positions of the knots α_d are treated as fixed, model (2.1) is a linear regression model. Testing whether a knot can be removed and the same polynomial equation used to explain two adhacent segments can be test by testing $H_0 : \xi_d = 0$, which is the common t-test statistics.

2.4 Feature selection methods

2.4.1 Subset selection

Subset selection methods are a common technique to perform feature selection. With subset selection, we retain only a subset of the variables, and eliminate the rest from the model. Least squares regression (or quantile regression) is used to estimate the coefficients of the inputs that are retained. There are a number of different strategies for choosing the subset, including best-subset selection, forward and backward stepwise selection and forward-stagewise regression. By retaining a subset of the predictors and discarding the rest, subset selection produces a model that is interpretable and has possibly lower prediction error than the full model. However, because it is a discrete process, in which variables are either retained or discarded, it often exhibits high variance and does not reduce the prediction error of the full model. Regularized methods are more continuous, and don't suffer as much from high variability.

2.4.2 Ridge regression

Ridge regression shrinks the regression coefficients by imposing a penalty on their size. The ridge coefficients minimize a penalized residual sum of least squares,

$$\hat{\beta}^{ridge} = \min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_p^2 \right\}$$

Here $\lambda > 0$ is a tuning parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage. The coefficients are shrunk towards zero. An equivalent way to write the ridge problem is

$$\begin{aligned} \hat{\beta}^{ridge} &= \min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \\ &\quad subject\ to \sum_{j=1}^{p} \beta_j^2 \leq t \end{aligned}$$

which makes explicit the size constraint on the parameters. There is a one-to-one correspondence between the parameters λ and t. When there are many correlated variables in a linear regression model, their coefficients can become poorly determined and exhibit high variance. A wildly large positive coefficient on one variable can be canceled by a similarly large negative coefficients on its correlated cousing. By imposing a size constraint on the coefficients, this problem is alleviated. Note that the intercept β_0 has been left out of the penalty term. Penalization of the intercept would make the procedure depend on the

origin chosen for Y; that is, adding a constant c to each of the targets y_i would not simply result in a shift of the predictions by the same amount c. Writing the ridge criterion in matrix form,

$$RSS(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T(y - \mathbf{X}\beta) + \lambda\beta^T\beta$$

the ridge regression solutions are easily seen to be

$$\hat{\beta}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y},$$

where $\mathbf{I}$ is the $p \times p$ identity matrix. Notice that with the choice of quadratic penalty $\beta^T\beta$, the ridge regression solution is again a linear function of y.

2.4.3 Lasso

The lasso is a shrinkage method like ridge, with subtle but important differences. The lasso estimate is defined by

$$\hat{\beta}^{lasso} = \min_{\beta} \left\{ \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Equivalenty, we can write

$$\begin{aligned} \hat{\beta}^{lasso} \quad = \quad & \min_{\beta} \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 \\ & \textit{subject to} \sum_{j=1}^{p} |\beta_j| \leq t \end{aligned}$$

The L_2 ridge penalty $\sum_1^p \beta_j^2$ is replaced by the L_1 lasso penalty $\sum_1^p |\beta_j|$. This latter constraint makes the solutions nonlinear in the y_i, and there is no closed form expression as in ridge regression. Computing the lasso solution is a quadratic programming problem, although efficient algorithms are available for computing the entire path of solutions as λ is varied, with the same computational cost as for ridge regression. Because of the nature of the constraint, making t sufficiently small will cause some of the coefficients to be exactly zero. Thus the lasso does a kind of continuous subset selection.

2.4.4 Smoothly clipped absolute deviation

The main disadvantage of lasso is that its non-zero coefficients might be biased. The smoothly clipped absolute deviation (SCAD) penalty was shown to

overcome this issue. It is defined as

$$p_{\lambda}^{SCAD}(\beta) = \left\{ \begin{array}{ll} \lambda|\beta|; & \text{if } |\beta| \leq \lambda \\ -\frac{|\beta|^2 - 2a\lambda|\beta| + \lambda^2}{2(a-1)}; & \text{if } \lambda \leq |\beta| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2}; & \text{if } |\beta| > a\lambda \end{array} \right\}$$

where $a > 2$ and $\lambda > 0$. It corresponds to a quadratic spline function with knots at λ and $a\lambda$. The function is continuos and differentiable on $(-\infty, 0) \cap (0, \infty)$, but singular at 0 with its derivatives zero outside the range $[-a\lambda, a\lambda]$. This results in small coefficients being set to zero, a few other coefficients being shrunk towards zero while retaining the large coefficients as they are. Thus, SCAD can produce sparse set of solution and approximately unbiased coefficients for large coefficients.

The solution to the SCAD penalty can be given as

$$\hat{\beta}_j^{SCAD} = \left\{ \begin{array}{ll} (|\hat{\beta}_j| - \lambda)_+ sign(\hat{\beta}_j); & \text{if } |\hat{\beta}_j| \leq 2\lambda \\ \{(a-1)\hat{\beta}_j - sign(\hat{\beta}_j)a\lambda/(a-2)\}; & \text{if } 2\lambda < ||\hat{\beta}_j| \leq a\lambda \\ \hat{\beta}_j; & \text{if } |\beta| > a\lambda \end{array} \right\}$$

This thresholding rule involves two unknown parameters λ and a. Theoretically, the best pair (λ, a) could be obtained using two dimensional grids search using some criteria like cross validation methods. However, such an implementation could be computationally expensive. Fan and Li (2001) suggested $a = 3.7$ is a good choice for variance problems.

Chapter 3

Marginal quantile regression with a working odds-ratio matrix

3.1 Introduction

Longitudinal and clustered data represent two frequent data structures in which observations within clusters may be dependent. In this chapter we analyze data from a longitudinal randomized trial on cognitive behavior therapy for treatment of obsessive compulsive disorder (Vigerland et al. (2016)). Participants were randomised to ten weeks of internet-based cognitive theraphy (ICBT) with therapist support, or to a waitlist control condition. The main interest of the study was to assess the relationship between the treatment and anxiety disorder score.

Several methods have been proposed to account for the dependence induced by the clustering when estimating marginal quantiles of a response variable. Cluster bootstrap is practical but can become computationally slow when the number of clusters is large. Besides, because it does not model the dependence, it may also be inefficient. Generalized estimating equations (GEE), which can estimate population-averaged models, have become a popular alternative to the bootstrap. The dependence between observations within the same cluster is modeled through a covariance matrix, which is usually assumed to be the same for all clusters. In the literature this matrix is called working covariance matrix, because the estimator is asymptotically correct even if the correlation is misspecified. The term "population-averaged" refers to the fact that the method models the average response over the subpopulation that shares a common value

of the predictors as a function of such predictors (Diggle et al. (2002)).

Quantile regression is a distribution-free method (Koenker (2005)) that describes the entire conditional distribution of a response variable. Marginal quantiles were analyzed by Jung (1996), who linked the GEE approach to quantile regression. This method requires the estimation of the density of the residual errors, and is based on estimating equations that are non-smooth with respect to the parameters. More recently, Fu and Wang (2012) provided a smoothed version of Jung's estimating equation by means of induced smoothing (Brown and Wang (2005)). This method does not require specifying the distribution of the residuals and estimates the parameters and their standard errors jointly.

The main difference between mean and quantile GEE is related to the estimation of the working correlation matrix. In the quantile approach, this matrix is estimated from the regression residuals' signs. However, the Pearson's correlation coefficient is not a good measure of dependence between binary variables because it is bounded by their marginal probabilities.

Because correlation is not a natural scale for binary sign variables, modeling on this scale has several disadvantages. First, it does not provide enough flexibility. For instance, parameters cannot depend on covariates. Second, estimation procedures might be complicated because they have to respect the contraints given by the marginal probabilities. Third, the interpretation of the correlation coefficients may not be straightforward.

Fu et al. (2015) proposed to estimate the working correlation matrix using a gaussian pseudolikelihood. Although this may formulation improved the flexibility, the computational and interpretational issues persist.

Instead, we propose to model a working association matrix defined through odds ratios. In the context of marginal logistic regression, this parametrization was analyzed by Lipsitz et al. (1991) and Carey et al. (1993).
Touloumis et al. (2013) studied the multinomial case. In marginal quantile regression, this alternative was mentioned by Yi and He (2009). This chapter explores the odds ratios parametrization in GEE applied to quantile regression. Different working structures can be estimated by specifying appropriate logistic regression models, including multilevel hierarchical data structures. Logistic regression models can be used to select the working dependence structure appropriately.

The rest of the chapter is organized as follows. Section 3.2 presents the smoothed quantile generalized estimating equations and the odds-ratio models. A simulation study is described in Section 3.3. We analyze data from a randomized trial on cognitive in behavior therapy for treatment of obsessive compulsive disorder in Section 3.4. The chapter is concluded with a summary in Section 3.5.

3.2 Method

Let $\{y_{ij}, x_{ij}\}$, $i = 1, \ldots, n$, $j = 1, \ldots, T_i$ be longitudinal data, where $y_{ij} \in \mathbb{R}$ is the response variable and $x_{ij} \in \mathbb{R}^P$ is the covariate vector. For brevity, we assume that the number of observations in a cluster T_i is constant across clusters, $T_i = T$. The general case when $T_i \neq T$ is a straightforward extension. Consider the problem of estimating the conditional τ-th quantile of y given x. A simple solution is to treat the observation as independent and minimize the following objective function:

$$\sum_{i,j} \epsilon_{ij} \, \psi(\epsilon_{ij}),$$

where $\epsilon_{ij} = y_{ij} - x_{ij}^T \beta_\tau$ indicates the residual and $\psi(\epsilon_{ij}) = \tau - I(\epsilon_{ij} < 0)$ is a linear transformation of the residual sign. The parameter vector β_τ can be estimated by solving

$$\sum_{i,j} x_{ij} \psi(\epsilon_{ij}) = 0$$

Ignoring the dependence within clusters may lead to wrong standard errors.

Consider the set of repeated measures on the i-th individual, denoted by $y_i = (y_{11}, \ldots, y_{1T})$, and its design matrix $x_i = (x_{11}, \ldots, x_{1T})$. Each element of the vector $\psi(\epsilon_i) = (\psi(\epsilon_{i1}), \ldots, \psi(\epsilon_{iT}))$ follows a Bernoulli distribution with expectation τ. Therefore, marginal quantile regression can be regarded as a special case of GEE where the mean model is a constant τ and the response variable contains a function of the parameters, $I(\epsilon_{ij} < 0)$. Jung (1996) showed that marginal quantiles can be obtained by

$$U_Q(\beta) = \sum_{i=1}^{n} x_i^T \Gamma_i W_i^{-1} \psi_\tau(\epsilon_i) = 0 \tag{3.1}$$

where $W_i = A_i^{1/2} R_i A_i^{1/2}$, $A_i = diag(\tau(1-\tau), \ldots, \tau(1-\tau))$, R_i is the residuals sign correlation matrix of the i-th individual, $\Gamma_i = diag(f_{i1}(0)), \ldots, f_{i_T}(0))$, and f_{ij} indicates the probability density function of ϵ_{ij}. The latter can be estimated by (Hall and Sheather (1988))

$$\hat{f}_{ij}(0) = 2h_n \left[x_{ij}^T \left\{ \hat{\beta}_{\tau+h_n} - \hat{\beta}_{\tau-h_n} \right\} \right]^{-1}$$

where h_n is a bandwidth parameter such that $h_n \to 0$ for $n \to \infty$, often calculated as $h_n = 1.57 n^{-1/3} \left(1.5\phi^2 \left\{ \Phi^{-1}(\tau) \right\} / \left[2 \left\{ \Phi^{-1}(\tau) \right\}^2 \right] \right)^{2/3}$. The covariance matrix W_i can be parametrized to increase efficiency. To protect against misspecification, a sandwich estimator of the standard errors can be used.

The correlation matrix of the regression residuals and of the regression residuals' sign can be different from each other.
As pointed out by Leng and Zhang (2014), if ϵ_i has an AR-1 covariance matrix with parameter ϕ then $\psi(\epsilon_i)$ depends on ϕ as a function of a computationally untractable two-dimensional integral. Other examples can be found in Fu et al. (2015). In general, the regression residuals and their signs have the same correlation structure only if ϵ_i has an exchangeable or Toeplitz covariance structure.

Fu and Wang (2012) proposed to smooth Jung's estimating equation by means of induced smoothing (Brown and Wang (2005)). They approximated the estimator by adding to the true value β a multivariate standard normal distribution Z and a smoothing parameter Ω, $\hat{\beta} = \beta + \Omega^{1/2}Z$. The smoothed estimating equations are obtained by

$$\tilde{U}_Q(\beta) = E_Z(U_Q(\beta + \Omega^{1/2}Z)) = \sum_{i=1}^{n} x_i^T \Gamma_i W_i^{-1}(\eta)\tilde{\psi}_\tau(\epsilon_i) = 0 \qquad (3.2)$$

where $\tilde{\psi}_\tau(\epsilon_i) = \left(1 - \Phi\left(\frac{y_{i1} - x_{i1}^T\beta}{r_{i1}}\right) - \tau, \ldots, 1 - \Phi\left(\frac{y_{iT} - x_{iT}^T\beta}{r_{iT}}\right) - \tau\right)^T$, $\Phi(\cdot)$ is the standard normal cumulative distribution and $r_{ik} = (x_{ik}^T \Omega x_{ik})^{1/2}$.

The derivative of the smoothed score are given by

$$\tilde{D}(\beta) = \frac{\partial \tilde{U}_Q(\beta)}{\partial \beta} = \sum_{i=1}^{n} X_i^T \Gamma_i W_i^{-1}(\eta)\tilde{\Lambda}_i X_i$$

where $\tilde{\Lambda}_i$ is a diagonal matrix with the kth diagonal element $r_{ik}^{-1}\phi((y_{ik} - x_{ik}^T\beta)/r_{ik})$.

The estimation of $\tilde{\beta}$ and its covariance matrix is obtained through an algorithm, summarized by the following steps:

1. Initialization: obtain $\tilde{\beta}^0$ using ordinary quantile regression; set $\Omega^0 = n^{-1}I_p$ and $K = 0$.

2. Compute the covariance matrix $W_i(\eta)$ using logistic regression on the residuals sign of the current estimate $\tilde{\beta}^K$.

3. Update $\tilde{\beta}^{K+1}$ and $\tilde{\Omega}^{K+1}$ by:

$$\begin{aligned} \tilde{\beta}^{K+1} &= \beta^K + \{-\tilde{D}(\tilde{\beta}^K, \tilde{\Omega}^K)\}^{-1}\tilde{U}_Q(W_i^K, \tilde{\beta}^K, \tilde{\Omega}^K) \\ \tilde{\Omega}^{K+1} &= [\tilde{D}(\tilde{\beta}^{K+1}, \tilde{\Omega}^K)]^{-1} Cov\{\tilde{U}_Q(W_i^K, \tilde{\beta}^{K+1}\}\{[\tilde{D}(\tilde{\beta}^{K+1}, \tilde{\Omega}^K)]^{-1}\}^T, \end{aligned}$$

   where $Cov\{\tilde{U}_G(\beta)\} = \sum_{i=1}^{n} x_i^T \Gamma_i W_i^{-1}(\eta)\tilde{\psi}_\tau(\epsilon_i)\tilde{\psi}_\tau^T(\epsilon_i)W_i^{-1}(\eta)\Gamma_i x_i$.

4. Repeat steps (2) and (3) until convergence.

The final values of $\tilde{\beta}$ and $\tilde{\Omega}$ are the smoothed estimator of β and its covariance matrix. The algorithm is fast and it usually requires few iterations to achieve convergence. Some difficulties may arise when estimating marginal quantiles at extreme quantiles, because $W_i(\eta)$ is less likely to be positive definite.

3.2.1 Estimating the working association matrix

Let $V_{it} = I(y_{it} \leq x_{it}^T \beta_\tau)$ be the residual sign of the i-th individual at time t and $V_t = (V_{1t}, \ldots, V_{nt})$ be the set of residual signs at time t.

A generic element w_{zu} of the working covariance matrix $W_i(\eta)$ can be written as

$$w_{zu} = \begin{cases} \tau(1-\tau), & z = u \\ p_{zu} - \tau^2, & z \neq u \end{cases}$$

where $p_{zu} = E(V_z V_u) = P(V_z = 1, V_u = 1)$. The elements of $W_i(\eta)$ can be computed simultaneously through a second set of estimating equations (Prentice (1988)). These probabilities are bounded by τ, the marginal probability of V_z and V_u, as follows:

$$\max(0, 2\tau - 1) \leq p_{zu} \leq \tau. \tag{3.3}$$

The correlation coefficient is not a good measure of dependence between binary variables, because it is bounded by their marginal frequencies. Instead, we propose to model a working association matrix defined through odds ratios.

Let η_{zu} be the odds ratio between V_z and V_u,

$$\eta_{zu} = \frac{P(V_z = 1, V_u = 1)/P(V_z = 0, V_u = 0)}{P(V_z = 0, V_u = 1)/P(V_z = 1, V_u = 0)}.$$

Let $\mathcal{A} = \{(V_z, V_u)\},\ z = u, \ldots, T,\ u = 1, \ldots, T$, be the set of pairwise vectors of all $\binom{T}{2}$ residual signs corresponding to the odds ratios in the lower triangular part of the working covariance matrix $W_i(\eta)$. Because $W_i(\eta)$ is symmetric, the upper triangular part is the mirror image of the lower part. Consider the new dataset (V_z, V_u, z, u, c). For any working structure of $W_i(\eta)$, the respective set of odds ratios can be estimated simultaneously through an appropriate definition of the linear predictor in a logistic regression of V_z on V_u, $V_z | V_u \sim Be(\mu_{zu})$.

Estimating $W_i(\eta)$ using logistic models has some advantages. First, it makes it easy to specify the form of the working covariance matrix. For example,

- Exchangeable: $\mathit{logit}(\mu_{zu}) = \alpha + \eta V_u$;
- Toeplitz: $\mathit{logit}(\mu_{zu}) = \alpha + \sum_{i=1}^{T-1} \eta_i I_{z-u=i} V_u + \sum_{i=1}^{T-1} I_{z-u=i}$;

- Unstructured: $logit(\mu_{zu}) = \alpha + \sum_{i=1}^{\binom{T}{2}} \eta_i I_{c=i} V_u + \sum_{i=1}^{\binom{T}{2}} I_{c=i}$.
- Nested Exchangeable: $logit(\mu_{zu}) = \alpha + \sum_{i=1}^{K-1} \eta_i V_u I_{c_i=1} + \sum_{i=1}^{K-1} I_{c_i=1}$, where c_i indicates whether V_z and V_u belong to the same i-th cluster.
- Exchangeable dependent on a categorical covariate X: $logit(\mu_{zu}) = \alpha + \sum_{i=1}^{K} \eta_i V_u I_{X=i} + \sum_{i=1}^{K} I_{X=i}$, where K indicates the number of categories of X.

Second, logistic models can be used to easily select the most appropriate structure of the working covariance matrix. Likelihood comparisons and Akaike's information criterion may help identify the best structure. Given the marginal probabilities τ and the odds ratios η_{zu}, the joint probabilities p_{zu} can be obtained by solving the following equation:

$$(\eta_{zu} - 1)p_{zu}^2 + (2\tau(1 - \eta_{zu}) - 1)p_{zu} + \tau^2\eta_{zu} = 0.$$

Only the smaller root of the former equation provides probabilities that respect the constraints in equation (3.3).

3.3 Simulation study

We conducted simulations to assess the performance of the proposed method. The response variable was generated from the following model:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + (1 + |x_{1ij}|)(\epsilon_{ij} - q_\tau) \quad i = 1, \ldots, 250, \quad j = 1, \ldots, T,$$

where q_τ was such that $p(\epsilon_{ij} \leq q_\tau) = \tau$. The size of each cluster was sampled from a binomial distribution $Binom(T, p)$, with $T = (5, 10)$ and $p = 0.8$. The covariate x_{1ij} was sampled from a uniform distribution $U(0, 1)$ and the regression coefficients were set to 1.

We considered the following dependence structures for $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})$:

- exchangeable, with correlation of 0.3 and 0.6;
- independence;
- Toeplitz, with correlation equal to 0.4 for the first two lags and zero otherwise.

The distribution of the error term was multivariate Gaussian with unit variance diagonal.
We conducted a simulation study with 500 independent realizations and estimated three different quantiles, $\tau = (0.10, 0.25, 0.50)$. All computations were

performed using R version 3.23. The multivariate Gaussian random variable were generated using the "mvtnorm" library.

Coverage probabilities were close to their nominal value in all the settings (Table 3.1). The estimator of the 10-th percentile showed undercoverage, with smallest observed value equal to 0.88. Cluster bootstrap had a smaller mean squared error (MSE) than the proposed method when the 10-th percentile was estimated (Table 3.2). This may indicate that the number of clusters for this percentile was not large enough to ensure asymptotic properties. When the 25-th and 50-th percentile were estimated, the proposed method provided smaller mean squared error than cluster bootstrap, regardless of the dependence structure. Similarly to the classic GEE for the expectation of a response variable, selecting the true dependence structure improved the performances of the estimator in terms of mean squared error. When the true dependence structure was Toeplitz, the estimator obtained using the true dependence structure provided the smallest mean squared error in most of the settings. In the independent case, results were very similar across all the working structures. In the exchangeable case, the independence and exchangeable structures had the smallest mean squared error.

Model-based (naive) standard errors were on average smaller than robust standard errors in all settings (Table 3.3), except the independence case. Misspecification of the working correlation structure lead to a decrease in the empirical coverage of the estimators.

Model selection performed through Akaike's information criterion selected the true dependence structure with a probability between 80 and 100 percent for the 25-th and 50-th percentiles and between 40 and 75 percent for the 10-th percentile(Table 3.4). The smallest percentage of correct selection was 40 percent, obtained when the true dependence structure was Toeplitz, $\tau = 0.10$ and $T = 5$. The highest was 100 percent, obtained when the true dependence structure was Toeplitz, $\tau = 0.50$ and $T = 10$. In general, the percentage of correct selection was higher for the median than the 25-th percentiles.

		$\tau = 0.10$		$\tau = 0.25$		$\tau = 0.50$	
	T	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
True structure: Toeplitz							
WI		0.914	0.926	0.934	0.950	0.940	0.938
EX		0.914	0.926	0.930	0.948	0.938	0.934
AR	10	0.912	0.920	0.930	0.946	0.944	0.934
TO		0.924	0.922	0.924	0.948	0.938	0.940
BS		0.944	0.940	0.946	0.956	0.938	0.942
WI		0.910	0.902	0.926	0.946	0.944	0.944
EX		0.912	0.902	0.926	0.948	0.948	0.940
AR	5	0.912	0.896	0.924	0.950	0.944	0.940
TO		0.912	0.906	0.914	0.940	0.938	0.932
BS		0.934	0.926	0.938	0.948	0.958	0.950
True structure: Independence							
WI		0.914	0.928	0.952	0.950	0.942	0.948
EX		0.910	0.928	0.950	0.950	0.942	0.942
AR	10	0.914	0.926	0.950	0.950	0.940	0.948
TO		0.906	0.932	0.948	0.944	0.940	0.946
BS		0.938	0.958	0.966	0.966	0.948	0.954
WI		0.908	0.922	0.942	0.926	0.946	0.956
EX		0.908	0.920	0.946	0.928	0.946	0.954
AR	5	0.910	0.920	0.946	0.926	0.946	0.954
TO		0.912	0.918	0.942	0.932	0.946	0.950
BS		0.930	0.930	0.958	0.944	0.964	0.964
True structure: Exchangeable with $\rho = 30$							
WI		0.926	0.930	0.934	0.954	0.946	0.948
EX		0.934	0.930	0.932	0.962	0.948	0.950
AR	10	0.932	0.932	0.932	0.954	0.950	0.954
TO		0.928	0.934	0.934	0.960	0.946	0.950
BS		0.942	0.950	0.940	0.958	0.964	0.956
WI		0.892	0.906	0.946	0.944	0.922	0.926
EX		0.888	0.900	0.948	0.946	0.924	0.938
AR	5	0.894	0.898	0.950	0.946	0.920	0.938
TO		0.892	0.904	0.948	0.944	0.918	0.938
BS		0.942	0.954	0.950	0.958	0.940	0.938
True structure: Exchangeable with $\rho = 60$							
WI		0.902	0.908	0.954	0.936	0.956	0.942
EX		0.900	0.914	0.944	0.930	0.944	0.944
AR	10	0.908	0.908	0.946	0.926	0.944	0.944
TO		0.904	0.916	0.944	0.930	0.942	0.946
BS		0.932	0.948	0.962	0.942	0.952	0.948
WI		0.898	0.922	0.938	0.946	0.932	0.934
EX		0.906	0.910	0.944	0.946	0.938	0.946
AR	5	0.912	0.908	0.934	0.944	0.934	0.946
TO		0.902	0.908	0.928	0.950	0.934	0.944
BS		0.938	0.946	0.966	0.968	0.956	0.946

Table 3.1: Empirical coverage of the proposed estimator at three quantiles for different working correlation structures, cluster size T and true correlation structures. WI=independence, EX=exchangeable, AR=autoregressive, TO=Toeplitz, BS=cluster bootstrap.

		$\tau = 0.10$		$\tau = 0.25$		$\tau = 0.50$	
	T	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
True structure: Toeplitz							
WI		0.1381	0.6603	0.0903	0.3789	0.0798	0.3414
EX		0.1387	0.6593	0.0907	0.3802	0.0796	0.3430
AR	10	0.1373	0.6555	0.0889	0.3737	0.0790	0.3365
TO		0.1366	0.6522	0.0890	0.3633	0.0783	0.3286
BS		0.1273	0.5966	0.0899	0.3772	0.0820	0.3514
WI		0.2880	1.3025	0.1699	0.7360	0.1410	0.6469
EX		0.2863	1.2933	0.1685	0.7240	0.1373	0.6269
AR	5	0.2825	1.2753	0.1678	0.7216	0.1372	0.6259
TO		0.2818	1.2864	0.1656	0.7259	0.1399	0.6226
BS		0.2536	1.1050	0.1675	0.7340	0.1455	0.6673
True structure: Independence							
WI		0.0992	0.4368	0.0480	0.2068	0.0449	0.1811
EX		0.0992	0.4381	0.0481	0.2074	0.0448	0.1810
AR	10	0.0992	0.4377	0.0481	0.2074	0.0449	0.1810
TO		0.0997	0.4396	0.0487	0.2098	0.0448	0.1820
BS		0.0881	0.3795	0.0482	0.2054	0.0459	0.1847
WI		0.1868	0.8453	0.0984	0.4693	0.0760	0.3392
EX		0.1869	0.8462	0.0984	0.4694	0.0758	0.3387
AR	5	0.1870	0.8485	0.0984	0.4691	0.0758	0.3386
TO		0.1892	0.8507	0.0989	0.4689	0.0758	0.3394
BS		0.1711	0.7612	0.0995	0.4678	0.0808	0.3660
True structure: Exchangeable with $\rho = 30$							
WI		0.1771	0.7963	0.1160	0.4746	0.1106	0.4563
EX		0.1761	0.8009	0.1152	0.4702	0.1105	0.4490
AR	10	0.1772	0.7989	0.1151	0.4697	0.1095	0.4457
TO		0.1772	0.7977	0.1157	0.4739	0.1105	0.4456
BT		0.1667	0.7360	0.1212	0.4990	0.1156	0.4804
WI		0.2905	1.3773	0.1462	0.6456	0.1526	0.6486
EX		0.2886	1.3648	0.1454	0.6326	0.1500	0.6286
AR	5	0.2913	1.3820	0.1456	0.6357	0.1517	0.6311
TO		0.2875	1.3484	0.1461	0.6290	0.1509	0.6283
BS		0.2242	0.9985	0.1492	0.6439	0.1519	0.6523
True structure: Exchangeable with $\rho = 60$							
WI		0.3107	1.3797	0.1883	0.8730	0.1783	0.7632
EX		0.3173	1.4302	0.1909	0.9024	0.1834	0.8213
AR	10	0.3145	1.3921	0.1890	0.8874	0.1774	0.7827
TO		0.3153	1.4352	0.1923	0.9044	0.1813	0.8117
BS		0.2730	1.1983	0.1903	0.8910	0.1805	0.7805
WI		0.4306	1.9953	0.2268	0.9978	0.1858	0.8891
EX		0.4146	1.9234	0.2323	1.0414	0.1857	0.8707
AR	5	0.4219	1.9758	0.2317	1.0377	0.1859	0.8647
TO		0.4250	1.9645	0.2360	1.0470	0.1854	0.8701
BS		0.3570	1.4985	0.2210	0.9626	0.1869	0.8874

Table 3.2: Mean squared error of the proposed estimator at three quantiles for different working correlation structures, cluster size T and true correlation structures. WI=independence, EX=exchangeable, AR=autoregressive, TO=Toeplitz, BS=cluster bootstrap.

		$\tau = 0.10$		$\tau = 0.25$		$\tau = 0.50$	
	T	Avg s.e.	Cover	Avg s.e.	Cover	Avg s.e.	Cover
True structure: Toeplitz							
WI		0.2036	0.796	0.1260	0.868	0.1217	0.864
EX	10	0.0965	0.856	0.0090	0.946	0.0067	0.922
TO		0.1039	0.848	0.0216	0.940	0.0209	0.936
WI		0.2932	0.750	0.1479	0.880	0.1399	0.890
EX	5	0.1667	0.786	0.0161	0.926	0.0035	0.940
TO		0.1694	0.792	0.0229	0.918	0.0124	0.938
True structure: Independence							
WI		0.0418	0.860	-0.0197	0.954	-0.0161	0.956
EX	10	0.0447	0.860	-0.0171	0.952	-0.0136	0.950
TO		0.0447	0.862	-0.0171	0.954	-0.0133	0.950
WI		0.1038	0.854	-0.0184	0.924	-0.0256	0.966
EX	5	0.1076	0.848	-0.0147	0.922	-0.0230	0.964
TO		0.1078	0.844	-0.0141	0.924	-0.0228	0.960
True structure: Exchangeable with $\rho = 30$							
WI		0.2887	0.776	0.2298	0.830	0.2274	0.800
EX	10	0.1173	0.844	0.0554	0.942	0.0570	0.936
TO		0.1184	0.846	0.0552	0.944	0.0571	0.938
WI		0.2715	0.774	0.1478	0.890	0.1473	0.866
EX	5	0.1541	0.804	0.0100	0.934	0.0079	0.946
TO		0.1564	0.804	0.0118	0.938	0.0078	0.942
True structure: Exchangeable with $\rho = 60$							
WI		0.5468	0.658	0.4402	0.714	0.4234	0.680
EX	10	0.2913	0.814	0.2358	0.856	0.2388	0.828
TO		0.2918	0.810	0.2373	0.848	0.2416	0.836
WI		0.5100	0.628	0.3237	0.786	0.2870	0.810
EX	5	0.2908	0.752	0.1190	0.904	0.0998	0.912
TO		0.2928	0.740	0.1216	0.904	0.1024	0.908

Table 3.3: Standard error difference between the robust and naive estimators (Avg s.e.) and empirical coverage of the naive estimators (Cover) for the parameter β_1 at three quantiles for different working correlation structures, cluster size T and true correlation structures. WI=independence, EX=exchangeable, TO=Toeplitz.

		$\tau = 0.10$				$\tau = 0.25$				$\tau = 0.50$			
True model	T	WI	EX	TO	AR	WI	EX	TO	AR	WI	EX	TO	AR
Independent	10	354	86	60	0	394	84	22	0	413	87	0	0
	5	369	87	44	0	393	84	23	0	415	76	9	0
Exchangeable 30	10	12	407	81	0	0	442	58	0	0	498	2	0
	5	30	407	63	0	3	452	45	0	0	494	6	0
Exchangeable 60	10	0	366	134	0	0	405	95	0	0	474	26	0
	5	0	403	97	0	0	445	55	0	0	459	41	0
Toeplitz	10	29	147	324	0	1	39	460	0	0	0	500	0
	5	24	278	198	0	1	142	357	0	0	17	483	0

Table 3.4: Frequency of correct selections of the true model in different scenarios and quantiles. T indicates the cluster size. WI=independence, EX=exchangeable, TO=Toeplitz, AR=autoregressive.

3.4 Application to internet-based cognitive therapy data set

The dataset consisted of ninety-five ($n = 95$) families with a child aged 8-12 years with a principal diagnosis of generalised anxiety, panic disorder, separation anxiety, social phobia and specific phobia Vigerland et al. (2016). Participants were randomised to ten weeks of internet-based cognitive theraphy (ICBT) with therapist support (70 families), or to a waitlist control condition (25 families). At weekly intervals, the amount of difficulties in emotion regulation scale (DERS) was measured on a 0-100 scale, where 0 indicated no difficulties and 100 extreme difficulties. The maximum number of repeated measurements is fourteen ($T = 14$). Similarly to previous studies, we assumed that data were missing completely at random. The main interest of the study was to assess the relationship between the treatment and DERS.

Figure 3.1 shows the residual errors of a classical mean regression and their kernel density. It was markedly skewed and there were several outliers. The mean was not an appropriate summary of it, and we considered quantiles. There seemed to be a time effect on the response variable, both in treated and untreated groups (Figure 3.2). The intraclass correlation was estimated to be 70%.

We considered three different quantiles, $\tau = (0.25, 0.50, 0.75)$, and estimated the following model:

$$Q(Ders_{ij}|\,Treat,\, Time) = \beta_0 + \beta_1\, Treat_i + \beta_2\, Time_{ij} + \beta_3\, Treat_i \times Time_{ij}. \quad (3.4)$$

We considered an exchangeable, Toeplitz, independence and treatment-dependent exchangeable association structures, along with a cluster bootstrap. Results are shown in Table 3.5.

The treatment coefficient, β_1, represents the difference in the considered quantile between treatment groups at the start of the study. It was statistically significant at the 50-th and 25-th percentiles in the exchangeable and treatment-dependent exchangeable structures. Considering the randomization, the observed difference at baseline can only be explained by chance. The time coefficient, β_2, indicates the change in the considered quantile per unit change in time. It was significant at the median with all correlation structures except the Toeplitz. At the 75-th percentile it was significant with all correlation structures. The interaction coefficient, β_3, represents the difference in the slope over time between the treatment groups. We found no significant difference in slope. The treatment-dependent exchangeable structure had the smallest AIC. Measurements from subjects that were treated were less strongly associated than

those from subjects that were not treated (Table 3.6). Standard errors obtained with cluster bootstrap were up to 50 percent greater than those obtained with the proposed method.

Figure 3.3 shows the relationship between DERS and Time, stratified by treatment. As the estimated quantile increases, the correlation between DERS and time decreases. The differences between treated and untreated are stronger at the median than at other percentiles.

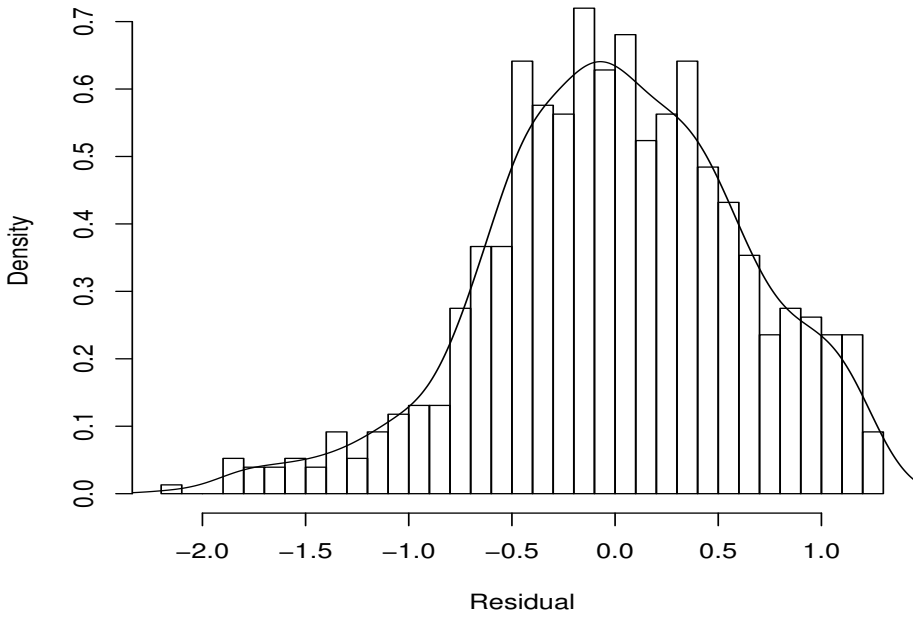

Figure 3.1: Mean regression residuals obtained from model (3.4) and their estimated kernel density.

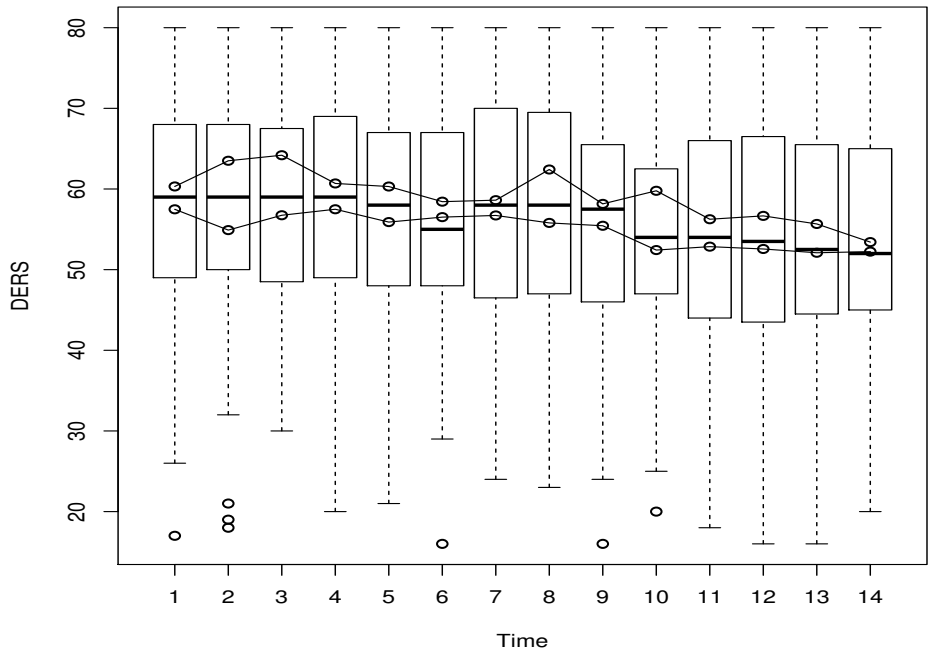

Figure 3.2: Boxplot of difficulties in emotion regulation scale (DERS) at different times. The two lines represents the mean of the response for that specific time in the group of treated and the untreated group.

	Coefficient	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
	Intercept	53.208 (0.908)	65.273 (3.243)	72.915 (2.739)
	Treatment	-5.988 (2.589)	-8.320 (3.908)	-6.406 (3.790)
EX	Time	-0.630 (0.648)	-0.608 (0.270)	-0.446 (0.218)
	TrxTime	0.202 (0.694)	0.081 (0.307)	0.265 (0.301)
	AIC	3165	4043	3045
	Intercept	53.152 (0.936)	65.282 (3.250)	72.933 (2.745)
	Treatment	-5.766 (2.576)	-8.221 (3.894)	-6.336 (3.728)
EX-Trt	Time	-0.649 (0.684)	-0.607 (0.270)	-0.445 (0.219)
	TrxTime	0.227 (0.728)	0.081 (0.307)	0.265 (0.300)
	AIC	3092	4031	3022
	Intercept	53.436 (0.611)	63.528 (6.809)	72.068 (2.467)
	Treatment	-6.964 (1.771)	-7.140 (6.950)	-6.307 (3.454)
TO	Time	-0.810 (0.658)	-0.472 (0.445)	-0.434 (0.198)
	Tr×Time	0.579 (0.689)	0.004 (0.478)	0.290 (0.300)
	AIC	3179	4037	3062
	Intercept	53.993 (3.392)	64.972 (3.560)	72.539 (2.639)
	Treatment	-6.079 (3.937)	-6.954 (4.097)	-5.240 (3.489)
WI	Time	-0.542 (0.426)	-0.679 (0.288)	-0.534 (0.214)
	TrxTime	0.120 (0.486)	0.211 (0.348)	0.319 (0.313)
	AIC	4207	5177	4207
	Intercept	54.430 (3.161)	65.059 (4.104)	72.229 (2.940)
	Treatment	-5.180 (3.854)	-6.171 (4.383)	-4.629 (3.914)
BT	Time	-0.503 (0.322)	-0.728 (0.366)	-0.627 (0.254)
	TrxTime	-0.080 (0.406)	0.187 (0.417)	0.391 (0.345)

Table 3.5: Estimated quantile regression coefficients, standard errors (in parentheses) and logistic models' AIC for three quantiles using exchangeable (EX), exchangeable varying with treatment (EX-Trt), Toeplitz (TO) and independence (WI) working correlation structures, along with a cluster bootstrap (BT).

Correlation Parameter	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
$\phi_{TRT=0}$	0.06	0.11	0.08
$\phi_{TRT=1}$	0.12	0.14	0.12
ϕ_{Exch}	0.10	0.13	0.11

Table 3.6: Estimated working correlation parameters for the exchangeable (ϕ_{Exch}) and exchangeable varying with treatment ($\phi_{TRT=1}$ and $\phi_{TRT=0}$) structures in the cognitive therapy application.

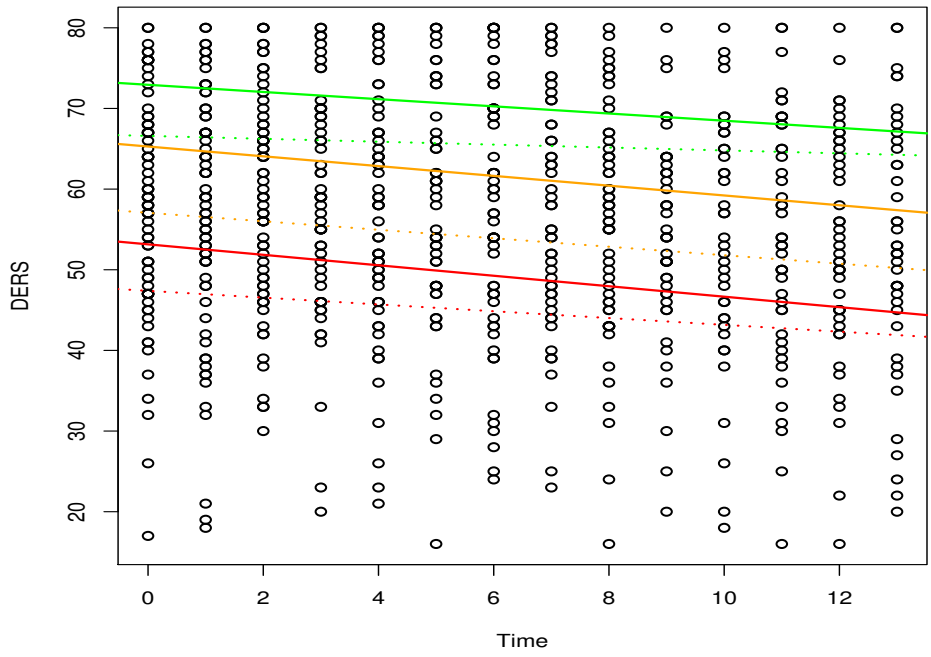

Figure 3.3: Estimated relationship between difficulties in emotion regulation scale (DERS) and time by treatment at three different quantiles $\tau = (0.25, 0.50, 0.75)$ with an exchangeable correlation structure varying with treatment. The dashed lines represent the treated group, while the solid lines represent the untreated group.

3.5 A summary

The present Chapter 3 chapter describes an alternative method to estimate the dependence within clusters for inference on population marginal quantile models. The association matrix of the residuals' sign is estimated by means of appropriate logistic regression models. The estimation is flexible and computationally fast. In our study the proposed method was approximately as fast as GEE applied to the mean. The best-fitting working structure can be selected by likelihood comparisons or, for non-nested models, by the Akaike's information criterion. Our simulations showed that this criterion could select the true dependence structure with high probability Possible extensions of this work include models for dependence with non-parametric quantile regression, M-estimators, LMS (Cole (1990)) and other classifiers.

Chapter 4

Penalized marginal quantile models

4.1 Introduction

This chapter presents two penalized versions of the marginal quantile regression estimator of Chapter 3, regularized smoothly clipped absolute deviation (SCAD) models for feature selection and reduced-rank penalized splines smoothing. Previous works in this field generally lack the possibility to either estimate quantiles (Wang et al. (2012), Chen et al. (2013)) or to account for the dependence structure of the data (Sherwood et al. (2016), Pratesi et al. (2009)).

Our proposed methods consist of solving estimating equations composed of two components. The first measures the fit of the model and it is the same as equation 3.2. The second is a penalization component that controls the model complexity through a tuning parameter. In regularized models, the model complexity is related to the coefficients' magnitude. Among the several penalty functions proposed in the literature, we chose the smoothly clipped absolute deviations (SCAD) penalty. Fan and Li (2001) showed that it has three desiderable properties of model selection: unbiasedness, sparsity and continuity. Other common penalty functions may lack any of these properties. For instance, the lasso penalty does not satisfies the unbiasedness property, the ridge penalty does not statisfied the sparsity property and elastic nets do not satisfy the continuity property.

In reduced-rank penalized spline smoothing, the model complexity is related to the smoothness of the estimated function. There are several choices for the spline basis. We focused on the truncated polynomial spline, because it provides analytical advantages.

In any given setting, other types of penalty functions or splines basis could perform better. The rest of the chapter is organized as follows. Section 4.2 presents the penalized quantile regression method and two special cases. Simulation studies are described in Section 4.3. We analyze data from a randomized trial on cognitive in behavior therapy for treatment of obsessive compulsive disorder in Section 4.4. The chapter is concluded with a discussion in Section 4.5.

4.2 Methods

4.2.1 Penalized quantile estimating equations

Let $\{y_{ij}, x_{ij}\}$, $i = 1, \ldots, n$, $j = 1, \ldots, T_i$ be longitudinal data, where $y_{ij} \in \mathbb{R}$ is the response variable and $x_{ij} \in \mathbb{R}^P$ is the covariate vector.
For brevity, we assume that the number of observations in a cluster T_i is constant across clusters, $T_i = T$. The general case when $T_i \neq T$ is a straightforward extension. As was shown in Chapter 3, marginal quantiles can be obtained by solving a smooth estimating equation,

$$\tilde{U}_Q(\beta) = \sum_{i=1}^{n} x_i^T \Gamma_i W_i^{-1}(\eta) \tilde{\psi}_\tau(\epsilon_i) = 0$$

The penalized quantile estimating equations are defined as

$$U(\beta, \lambda, \Omega) = \tilde{U}_Q(\beta, \Omega) - U_P(\beta, \lambda) = 0 \tag{4.1}$$

Note that if we take $\tilde{U}_Q(\beta, X) = \sum_{i=1}^{n} X_i^T W_i^{-1}(\eta)(Y_i - X_i^T \beta)$, the classical GEE for the mean, we obtain penalized marginal mean regression.

The tuning parameter λ controls the entire penalization component $U_P(\beta, \lambda)$. When it is zero, the penalization component disappears and the estimating equation reduces to equation 3.2. When it is infinite, the penalization is large. In penalized splines, this forces the estimated function to become linear. In regularized regression, it shrunk all the estimated parameters towards zero.

The complexity of the model might be controlled by a set of tuning parameters. In penalized splines, this allows the estimated function to be more flexible at knots where the function is strongly nonlinear, and less flexible at knots where the function is more linear. In regularized models, this allows to shrink coefficients with a different magnitude.

In this work, we focus on a single tuning parameter λ, selected through either cross-validation techniques or by minimizing a measure that balances goodness of fit and model complexity such as the Bayesian Information Criterion (BIC).

4.2.2 Regularized models for feature selection

There are several penalty functions available to perform feature selection, for instance lasso, elastic nets and smoothly clipped absolute deviation (SCAD). A good penalty function should result in an estimator with three properties: unbiasedness, sparsity (the estimator automatically sets small estimated coefficients to zero to reduce model complexity) and continuity (a relevant property to avoid instability in model prediction). Among the several penalty functions available, SCAD is one of the few that simultaneously achieves all these properties.

This penalty is defined as

$$U_P(\beta,\lambda) = q(\lambda,|\beta|)sgn(\beta), \tag{4.2}$$

$$q(\lambda,\beta) = n\lambda\left\{I(\beta\leq\lambda)+\frac{(a\lambda-\beta)_+}{\lambda(a-1)}I(\beta>\lambda)\right\}, \tag{4.3}$$

where a can be taken as 3.7 as recommended by Fan and Li (2001).

The non-differentiability of the SCAD penalty might cause computational problems. Hunter and Li (2005) proposed to solve this equations through the minorization-maximization (MM) algorithm. For a small $\epsilon < 0$, the MM algorithm suggests that the penalization term can be approximated by

$$\tilde{U}_P(\beta,\lambda) = q(\lambda,|\beta|)\ sign(\beta)\frac{|\beta|}{\epsilon+|\beta|} \tag{4.4}$$

In the numerical analyses, we take ϵ to be the fixed small number 10^{-6}. The estimation of $\tilde{\beta}$ and its covariance matrix is obtained through an algorithm, summarized by the following steps:

1. Initialization: set $\hat{\beta}^0 = (0,\ldots,0)_p$ and $\Omega^0 = n^{-1}I_p$ and $K = 0$.

2. Compute the covariance matrix $W_i(\eta)$ using logistic regression on the residuals sign of the current estimate $\hat{\beta}^K$.

3. Update $\hat{\beta}^{K+1}$ and $\hat{\Omega}^{K+1}$ by:

   $$\hat{\beta}^{K+1} = \hat{\beta}^K + \{H(\hat{\beta}^K,\lambda,\hat{\Omega}^K))\}^{-1}U(W_i^K,\hat{\beta}^K,\lambda,\hat{\Omega}^K)$$
   $$\hat{\Omega}^{K+1} = [H(\hat{\beta}^{K+1},\lambda,\hat{\Omega}^K)]^{-1}Cov\{U_Q(W_i^K,\hat{\beta}^{K+1},\hat{\Omega}^K)\}\{[H(\hat{\beta}^{K+1},\lambda,\hat{\Omega}^K)]^{-1}\}^T$$

   where
   $H(\beta,\lambda,\Omega) = \sum_{i=1}^n X_i^T\Gamma_i W_i^{-1}(\eta)\tilde{\Lambda}_i X_i - diag\left\{\frac{q(\lambda,|\beta|)}{\epsilon+|\beta_1|},\ldots,\frac{q(\lambda,|\beta|)}{\epsilon+|\beta_p|}\right\}$
   $Cov\{U_Q(W_i^K,\hat{\beta}^{K+1},\hat{\Omega}^K)\} = [U_Q(W_i^K,\hat{\beta}^{K+1},\hat{\Omega}^K)]^T U_Q(W_i^K,\hat{\beta}^{K+1},\hat{\Omega}^K)$

4. Repeat steps (2) and (3) until convergence.

The final values of $\hat{\beta}$ and $\hat{\Omega}$ are the estimator of β and its covariance matrix. In our simulations, we continued the iterative procedure until $||\beta^{K+1} -$

$\beta^K|| < 10^{-8}$. Our numerical experiment showed that the stopping criterion was achieved on average in 30 iterations.

4.2.3 Reduced-rank penalized splines smoothing

Let $B(x)$ denote an l-dimensional vector of spline basis functions such as B-splines or truncated polynomials. For the pth order truncated polynomial with K knots, $B(x) = [1, x, \ldots, x^p, (x-\tau_1)_+^p, \ldots, (x-\tau_K)_+^p]$, where $\tau_1, \ldots, \tau_k$ is a sequence of knots. Let $B_i = [B(X_{i1}), \ldots, B(X_{in_i})]$ denote the $n_i \times l$ matrix of basis functions.

Chen et al. (2013) showed the penalized mean regression splines are obtained by solving

$$U_M(\theta, \lambda, \Omega, B) = \sum_{i=1}^{n} B_i^T V_i^{-1}(\eta)(Y_i - B_i^T\theta) - \lambda D_q \theta = 0, \tag{4.5}$$

where V_i is the working covariance matrix of the regression residuals, θ is a vector of basis coefficients and D_q is an appropriate difference-based penalty matrix that depends on the chosen basis.

Similarly, the penalized quantile regression splines are obtained by substituting the unpenalized component of equation 4.5 with the unpenalized component of marginal quantile regression:

$$U(\theta, \lambda, \Omega, B) = \sum_{i=1}^{n} B_i^T \Gamma_i W_i^{-1}(\eta)\tilde{\psi}_\tau(Y_i - B_i^T\theta) - \lambda D_q \theta = 0,$$

where Γ_i and W_i are the same as equation 3.2. For the qth order truncated polynomial basis, we have $D_q = (0_{q+1}, 1_K)$.

The derivative of the penalized regression spline estimating equations are

$$H(\theta, \lambda, \Omega) = H_Q(\theta, \Omega) - H_P(\theta, \lambda), \tag{4.6}$$

where $H_Q(\theta, \Omega) = \sum_{i=1}^{n} B_i^T \Gamma_i W_i^{-1}(\eta)\tilde{\Lambda}_i B_i$ and $H_P(\theta, \lambda) = \lambda D_q$ are the derivatives of the unpenalized and penalized components of the penalized splines estimating equation.

The vector of basis coefficients θ and its covariance matrix Ω can be estimated with the following algorithm.

1. Initialization: Set $\hat{\theta}^0 = (0, \ldots, 0)_p$ and $\Omega^0 = n^{-1} I_q$ and $K = 0$.

2. Compute the covariance matrix $W_i(\eta)$ using logistic regression on the residuals sign of the current estimate $\hat{\theta}^K$.

3. Update $\hat{\theta}^{K+1}$ and $\hat{\Omega}^{K+1}$ by:

$$\begin{aligned}\hat{\theta}^{K+1} &= \hat{\theta}^{K} + \{H(\hat{\theta}^{K}, \lambda, \hat{\Omega}^{K}))\}^{-1} U(W_i^K, \hat{\theta}^{K}, \lambda, \hat{\Omega}^{K}) \\ \hat{\Omega}^{K+1} &= [H(\hat{\theta}^{K+1}, \lambda, \hat{\Omega}^{K})]^{-1} Cov\{U_Q(W_i^K, \hat{\theta}^{K+1}, \hat{\Omega}^{K})\}\{[H(\hat{\theta}^{K+1}, \lambda, \hat{\Omega}^{K})]^{-1}\}^T\end{aligned}$$

where

$$Cov\{U_Q(W_i^K, \hat{\theta}^{K+1}, \hat{\Omega}^{K})\} = [U_Q(W_i^K, \hat{\theta}^{K+1}, \hat{\Omega}^{K})]^T U_Q(W_i^K, \hat{\theta}^{K+1}, \hat{\Omega}^{K})$$

4. Repeat steps (2) and (3) until convergence.

The final values of $\hat{\theta}$ and $\hat{\Omega}$ are the estimator of θ and its covariance matrix. In our simulations, we continued the iterative procedure until $||\theta^{K+1} - \theta^{K}|| < 10^{-4}$. The stopping criterion was achieved on average in 20 iterations. The covergence rate of the algorithm was generally higher when we used a working independence working association structure W_i.

The estimated function $\hat{f}(x)$ and its sandwich variance are, respectively, $\hat{f}(x) = B^T(x)\hat{\theta}$ and $var(\hat{f}(x)) = B^T(x) cov(\hat{\theta}) B(x)$.

4.2.4 Selection of the tuning parameter

The selection criterion of the tuning parameter λ is different for penalized spline and regularized SCAD models. In the former, we selected it through cross-validation. The data was randomly splitted into several non-overlapping subsets of approximately equal size. We removed one subset and fit the model to the remaining data, and estimated the prediction error from the removed observations. This was repeated for each subset, the estimated prediction errors were aggregated, and the best tuning parameter was selected by minimizing the aggregated estimated prediction error over a fine grid. We used the negative loglikelihood of an asymmetric Laplace distribution as loss function. Alternatively, we could have used the quasi-likelihood under the independence model criterion (QIC), Pan (2001)):

$$QIC(R) = -2Q(\hat{\beta}(R); I) + 2trace(\hat{\Omega}_I \hat{V}_R), \tag{4.7}$$

where I is the working independence structure, R is a working covariance structure, V_R is the robust estimator of the covariance matrix of the working structure R, $\Omega_I = \frac{\partial}{\partial \beta} \tilde{U}_Q(\beta, X, I)$ and $Q(\hat{\beta}(R); I)$ is the quasi-likelihood constructed using a working independence structure evaluated at the estimate $\hat{\beta}$ obtained using the working covariance matrix R. In particular,

$$Q(\hat{\beta}(R); I) = \sum_{i,j} (y_{ij} - x_{1ij}^T \hat{\beta}(R))\ \psi(y_{ij} - x_{1ij}^T \hat{\beta}(R)),$$

where $\psi(y_{ij} - x_{1ij}^T\hat{\beta}(R)) = \tau - I(y_{ij} - x_{1ij}^T\hat{\beta}(R) < 0)$. The main advantage of this measure is that it takes into account the dependence structure of the data. However, it is computationally slow, since it requires the estimation of two models, one for the working structure R and another for the working independence structure. Because the negative loglikelihood of an asymmetric Laplace distribution is simple to implement and computationally fast, we selected it as loss function.

In regularized SCAD models, we selected the tuning parameter λ by optimizing some data-driven criteria which balances goodness of fit and model complexity. We used the BIC for high-dimensional data proposed by Ma et al. (2013):

$$BIC_\lambda = \log\left(\frac{\sum_i e_i^T R_i^{-1} e_i}{nt}\right) + \frac{\log(nt)}{nt} df,$$

where $e_i = y_i - x_i^T\beta$ for given λ, R_i is the working correlation matrix and df is the number of non-zero coefficients of the model (at the end of the estimation procedure, any coefficient that is below 10^{-4} was considered as zero). Alternatively, we could have used the high-dimensional BIC for linear quantile regression when p is much larger than n (Lee et al. (2014)),

$$QBIC(\lambda) = \log\sum_{i=1}^{n}\sum_{j=1}^{t}\psi(y_{ij} - x_{ij}^T\hat{\beta}_\lambda) + \nu_\lambda\frac{\log(p)\log(nt)}{2nt},$$

where ν_λ is the degrees of freedom of the fitted model, which is the number of interpolated fits for quantile regression. The main problem of this method is that the number of interpolated fits of a smoothed quantile regression is very different from that of classical quantile regression. Moreover, QBIC does not account for the dependence structure of the data.

4.3 Simulation studies

4.3.1 Regularized models

We conducted simulations to assess the performance of the regularized SCAD quantile model. The response variable was generated from the following model:

$$y_{ij} = \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \epsilon_{ij} \quad i = 1, \ldots, 200, \quad j = 1, \ldots, 9$$

The covariates x_{1ij}, x_{2ij}, x_{3ij} were sampled from a uniform distribution $U(0,1)$ and the regression coefficients were set to $\beta = (1, 2, -0.5)$. We added 200 independent predictors sampled from a uniform distribution $U(0,1)$ and standardized.

We considered the following dependence structures for $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})$:

- exchangeable, with correlation coefficient 0.5;
- independence;
- Toeplitz, with bands correlation coefficient: $(0.8, 0.6, 0.4, 0.2, 0.2, 0.1, 0.1, 0.1, 0)$.

The distribution of the error term was multivariate Gaussian with unit variance diagonal.

We conducted a simulation study with 500 independent realizations and estimated two different quantiles, $\tau = (0.50, 0.75)$. All computations were performed using R version 3.23. The multivariate Gaussian random variable were generated using the "mvtnorm" library.

Coverage probabilities obtained through the sandwich formula for $\beta_1, \beta_2, \beta_3$ were close to their nominal value in all settings (Table 4.1).

Table 4.2 shows the proportion of times the penalized SCAD estimator (P), the unpenalized estimator (UP) and the oracle estimator (OR) under-selected (U), over-selected (O) and exactly selected (EXACT) the nonzero coefficients. The regularized SCAD quantile regression models selected the relevant predictors with high probability. The penalized and oracle estimator correctly selected the true model 100 percent of the time. The unpenalized estimator never selected correctly the true model and always overfitted.

Table 4.3 report the mean squared error (MSE) of the function $f(x) = x_{ij}^T \beta$, i.e. $(1800)^{-1} \sum_{i=1}^{200} \sum_{j=1}^{9} (x_{ij}^T \hat{\beta} - x_{1ij} - 2x_{2ij} + 0.5x_{3ij})^2$ for the oracle estimator (O), penalized estimator (P) and unpenalized estimator (UP). When the true correlation structure was used, the mean squared error of the prediced function decreased. The mean squared error of the correctly specified working structure decreased up to 50 percent compared with that of the working independence estimator. The mean squared error of the penalized and oracle estimators was around 70 times smaller than that of the unpenalized estimator.

Model selection performed through Akaike's information criterion selected the true dependence structure with high probability (Table 4.4). The smallest percentage of correct selection was 85.6 percent, observed for the unpenalized estimator when the true dependence structure was independent and $\tau = 0.75$. The highest was 100 percent, obtained for all estimators and quantiles when the true dependence structure was either Toeplitz or exchangeable. In general, the percentage of correct selection was higher for the median than for the 75-th percentile.

We do not provide a comparison to penalized SCAD marginal mean regression. Sherwood et al. (2016) showed that penalized marginal quantiles out-

performed penalized marginal means in presence of heavy-tailed distribution or heteroscedastic errors. They also showed that the SCAD penalty tends to pick a smaller and more accurate model compared to lasso and other penalty functions.

		True Structure								
		Independent			Exchangeable			Toeplitz		
	Estimator	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
	EX_P	0.946	0.940	0.926	0.934	0.932	0.926	0.946	0.932	0.946
	TO_P	0.944	0.940	0.928	0.936	0.924	0.922	0.922	0.918	0.934
	WI_P	0.944	0.940	0.928	0.956	0.970	0.926	0.950	0.930	0.956
	EX_O	0.946	0.940	0.926	0.934	0.932	0.922	0.946	0.932	0.946
$\tau = 0.50$	TO_O	0.944	0.940	0.928	0.936	0.924	0.920	0.922	0.918	0.934
	WI_O	0.944	0.940	0.928	0.966	0.970	0.926	0.950	0.930	0.956
	EX_{UP}	0.930	0.934	0.916	0.930	0.916	0.922	0.940	0.934	0.918
	TO_{UP}	0.934	0.942	0.922	0.940	0.912	0.920	0.944	0.920	0.906
	WI_{UP}	0.932	0.936	0.916	0.946	0.928	0.930	0.934	0.924	0.940
	EX_P	0.970	0.968	0.958	0.938	0.932	0.930	0.948	0.934	0.952
	TO_P	0.966	0.970	0.958	0.942	0.934	0.928	0.928	0.922	0.940
	WI_P	0.970	0.968	0.958	0.974	0.982	0.966	0.968	0.964	0.974
	EX_O	0.970	0.968	0.958	0.938	0.932	0.928	0.948	0.934	0.952
$\tau = 0.75$	TO_O	0.966	0.970	0.958	0.942	0.934	0.926	0.928	0.924	0.940
	WI_O	0.970	0.968	0.958	0.974	0.982	0.966	0.968	0.964	0.974
	EX_{UP}	0.974	0.970	0.958	0.940	0.932	0.928	0.946	0.936	0.930
	TO_{UP}	0.978	0.972	0.956	0.944	0.920	0.926	0.958	0.942	0.938
	WI_{UP}	0.974	0.968	0.956	0.970	0.962	0.968	0.974	0.962	0.960

Table 4.1: Empirical coverage of the non-zero coefficients of the model obtained through the penalized SCAD estimator (P), the unpenalized estimator (UP) and the oracle estimator (OR) for different quantiles, working correlation structures and true correlation structures. WI=independence, EX=exchangeable, TO=Toeplitz.

		True Structure								
		Independent			Exchangeable			Toeplitz		
	Estimator	U	O	EXACT	U	O	EXACT	U	O	EXACT
	EX_P	0	0	1	0	0	1	0	0	1
	TO_P	0	0	1	0	0	1	0	0	1
	WI_P	0	0	1	0	0	1	0	0	1
	EX_O	0	0	1	0	0	1	0	0	1
$\tau = 0.50$	TO_O	0	0	1	0	0	1	0	0	1
	WI_O	0	0	1	0	0	1	0	0	1
	EX_{UP}	0	1	0	0	1	0	0	1	0
	TO_{UP}	0	1	0	0	1	0	0	1	0
	WI_{UP}	0	1	0	0	1	0	0	1	0
	EX_P	0	0	1	0	0	1	0	0	1
	TO_P	0	0	1	0	0	1	0	0	1
	WI_P	0	0	1	0	0	1	0	0	1
	EX_O	0	0	1	0	0	1	0	0	1
$\tau = 0.75$	TO_O	0	0	1	0	0	1	0	0	1
	WI_O	0	0	1	0	0	1	0	0	1
	EX_{UP}	0	1	0	0	1	0	0	1	0
	TO_{UP}	0	1	0	0	1	0	0	1	0
	WI_{UP}	0	1	0	0	1	0	0	1	0

Table 4.2: Proportion of times that the penalized SCAD estimator (P), the unpenalized estimator (UP) and the oracle estimator (OR) under-selected (U), over-selected (O) and exactly selected (EXACT) the covariates with nonzero coefficients for different quantiles, true correlation structures and working structure.

	$\tau = 0.50$			$\tau = 0.75$		
	True Structure			**True Structure**		
Estimator	Indep	Exch	Toeplitz	Indep	Exch	Toeplitz
EX_P	0.0252	0.0167	0.0190	0.0251	0.0175	0.0190
TO_P	0.0252	0.0175	0.0127	0.0250	0.0176	0.0130
WI_P	0.0252	0.0237	0.0241	0.0251	0.0236	0.0240
EX_O	0.0253	0.0176	0.0190	0.0251	0.0176	0.0190
TO_O	0.0252	0.0177	0.0127	0.0250	0.0177	0.0130
WI_O	0.0252	0.0237	0.0241	0.0251	0.0236	0.0240
EX_{UP}	1.3862	0.9812	1.0500	1.3496	0.9803	1.0477
TO_{UP}	1.3959	0.9886	0.7051	1.3556	0.9865	0.7523
WI_{UP}	1.3852	1.3818	1.3759	1.3497	1.3463	1.3413

Table 4.3: Mean squared error$\times$10 of $\hat{f}(x)$ for the penalized estimator (P), unpenalized estimator (UP) and oracle estimator (O) in different scenarios and quantiles.

	$\tau = 0.50$			$\tau = 0.75$		
	True Structure			**True Structure**		
Estimator	Indep	Exch	Toeplitz	Indep	Exch	Toeplitz
EX_P	72	500	0	75	500	0
TO_P	0	0	500	0	0	500
WI_P	428	0	0	425	0	0
EX_O	72	500	0	73	500	0
TO_O	0	0	500	0	0	500
WI_O	428	0	0	427	0	0
EX_{UP}	73	500	0	77	500	0
TO_{UP}	0	0	500	0	0	500
WI_{UP}	427	0	0	423	0	0

Table 4.4: Frequency of correct selections of the true dependence structure for the penalized estimator (P), unpenalized estimator (UP) and oracle estimator (O) in different scenarios and quantiles.

4.3.2 Reduced-rank penalized splines smoothing

We conducted simulations to assess the performance of the reduced-rank penalized spline method. The response variable was generated from the following model:

$$y_{ij} = f(x_{1ij}) + \epsilon_{ij} \quad i = 1, \ldots, 200, \;\; j = 1, \ldots, 9,$$

where $f(x_{1ij})$ was either $\log(x_{1ij})$, $\sin(2\pi x_{1ij})$ or $2\exp(x_{1ij})$. The covariate x_{1ij} was sampled from a uniform distribution $U(0,1)$ and standardized. We considered the following dependence structures for $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})$:

- exchangeable, with correlation coefficient 0.5;
- independence;
- Toeplitz, with bands correlation coefficient: $(0.8, 0.6, 0.4, 0.2, 0.2, 0.1, 0.1, 0.1, 0)$.

The distribution of the error term was multivariate Gaussian with unit variance diagonal. For penalized spline estimators, we used a truncated quadratic polynomial base with 20 knots.

We conducted a simulation study with 100 independent realizations and estimated two different quantiles, $\tau = (0.50, 0.75)$. All computations were performed using R version 3.23. The multivariate Gaussian random variable were generated using the "mvtnorm" library. The tuning parameter λ was chosen through cross-validation.

We compare the proposed P-spline approach with a regression spline approach (R-spline). In the penalized approach, the spline complexity was automatically selected by a tuning parameter chosen through two-fold cross-validation. Table 4.5 summarizes the mean of average MSE (MMSE), that is, $(1800)^{-1}\sum_{i=1}^{200}\sum_{j=1}^{9}(\hat{f}(x_{ij}) - f(x_{ij}))^2$ for the penalized (P) and unpenalized (UP) estimators. In all scenarios, the P-spline approach had a smaller MMSE than the R-spline approach. When the true correlation structure was used, we observed a decrease in MMSE of the estimated function $\hat{f}(x)$. The penalized estimator MMSE was up to three times smaller than its unpenalized counterpart.

Model selection performed through Akaike's information criterion selected the true dependence structure with high probability (Table 4.6). The smallest percentage of correct selection was 79 percent, obtained by the penalized estimator when the true dependence structure was independent and $\tau = 0.75$. The highest was 100 percent, obtained when the true dependence structure was Toeplitz or exchangeable and for any quantile and estimator type. In general,

the percentage of correct selection was higher for the median than for the 75-th percentile.

	$\tau = 0.50$			$\tau = 0.75$		
	True Structure			**True Structure**		
Estimator	Indep	Exch	Toeplitz	Indep	Exch	Toeplitz
			$\hat{f}(x) = \exp$			
EX_P	0.0133	0.0225	0.0227	0.0119	0.0241	0.0263
TO_P	0.0134	0.0226	0.0199	0.0121	0.0243	0.0231
WI_P	0.0131	0.0227	0.0220	0.0119	0.0251	0.0270
EX_{UP}	0.0448	0.0472	0.0492	0.0507	0.0568	0.0574
TO_{UP}	0.0448	0.0471	0.0477	0.0509	0.0570	0.0548
WI_{UP}	0.0447	0.0480	0.0493	0.0504	0.0571	0.0585
			$\hat{f}(x) = \log$			
EX_P	0.0316	0.0373	0.0376	0.0319	0.0404	0.0448
TO_P	0.0316	0.0371	0.0358	0.0320	0.0405	0.0405
WI_P	0.0314	0.0378	0.0380	0.0319	0.0414	0.0459
EX_{UP}	0.0476	0.0501	0.0515	0.0535	0.0588	0.0614
TO_{UP}	0.0477	0.0499	0.0501	0.0537	0.0591	0.0577
WI_{UP}	0.0474	0.0510	0.0514	0.0532	0.0594	0.0618
			$\hat{f}(x) = \sin$			
EX_P	0.0230	0.0296	0.0315	0.0238	0.0308	0.0377
TO_P	0.0230	0.0295	0.0287	0.0238	0.0308	0.0337
WI_P	0.0229	0.0315	0.0330	0.0237	0.0339	0.0382
EX_{UP}	0.0447	0.0472	0.0492	0.0507	0.0568	0.0574
TO_{UP}	0.0448	0.0471	0.0477	0.0509	0.0569	0.0547
WI_{UP}	0.0446	0.0480	0.0493	0.0504	0.0570	0.0585

Table 4.5: Mean squared error×10 of $\hat{f}(x)$ for the penalized estimator (P) and unpenalized estimator (UP) in different scenarios and quantiles.

	$\tau = 0.50$			$\tau = 0.75$		
	True Structure			**True Structure**		
Estimator	Indep	Exch	Toeplitz	Indep	Exch	Toeplitz
			$\hat{f}(x) = \exp$			
EX_P	14	100	0	21	100	0
TO_P	0	0	100	0	0	100
WI_P	86	0	0	79	0	0
EX_{UP}	15	100	0	15	100	0
TO_{UP}	0	0	100	0	0	100
WI_{UP}	85	0	0	85	0	0
			$\hat{f}(x) = \log$			
EX_P	17	100	0	18	100	0
TO_P	0	0	100	0	0	100
WI_P	83	0	0	82	0	0
EX_{UP}	17	100	0	17	100	0
TO_{UP}	0	0	100	0	0	100
WI_{UP}	83	0	0	83	0	0
			$\hat{f}(x) = \sin$			
EX_P	16	100	0	20	100	0
TO_P	0	0	100	0	0	100
WI_P	84	0	0	80	0	0
EX_{UP}	16	100	0	15	100	0
TO_{UP}	0	0	100	0	0	100
WI_{UP}	84	0	0	85	0	0

Table 4.6: Frequency of correct selections of the true depedence structure for the penalized estimator (P) and unpenalized estimator (UP) in different scenarios and quantiles. WI=independence, EX=exchangeable, TO=Toeplitz.

4.4 Application

We applied the proposed penalized methods to the cognitive behavior therapy data of Chapter 3. Recall that the dataset had ninety-five clusters and fourteen repeated measurements. To show the full potential of the penalized SCAD method, we added $q = 20$ uncorrelated standardized predictors sampled from a uniform distribution $U(0, 1)$. We also added a nonparametric component defined as $x_{NPij} = (log(Ders_{ij} - \epsilon_{ij}))$ that we modelled through reduced-rank penalized spline smoothing. The distribution of the error term $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})$ was taken as multivariate normal with a treatment-dependent exchangeable correlation matrix with parameters 0.6 (treated group) and 0.2 (untreated group). The relationship between the response variable and x_{NPij} can be seen in Figure 4.1.

We considered three different quantiles, $\tau = (0.25, 0.50, 0.75)$ and we estimated the following model:

$$Q(Ders_{ij}) = \beta_0 + \beta_1 Treat_i + \beta_2 Time_{ij} + \beta_3 Treat_{ij} \times Time_i + \sum_{z=5}^{q+5} \beta_z x_{zij}, \quad (4.8)$$

We estimated the nonparametric association $Ders_{ij} = f(x_{NPij})$ using reduced-rank penalized splines smoothing with a truncated cubic polynomial with 20 knots.

For both methods, we considered an exchangeable, independence and treatment-dependent exchangeable association structures.

Similarly to Chapter 3, the treatment-dependent exchangeable association structure has the lowest AIC (Table 4.7 and Table 4.8).

Table 4.7 shows also the results of the parametric component of the SCAD model (Equation 4.8). Among the zero coefficients that were added to the model, none were selected by our method. In the 50-th and 75-th quantiles, only the intercept and treatment coefficients were statistically significant. This result is in line to what we saw in Chapter 3, where these coefficients were far away from zero. The magnitude of these coefficients is comparable to that of the unpenalized method. In the 25-th percentile the treatment coefficient is no longer significant. This might be caused by some computational problems that arose when we estimated models with a relative small tuning parameter λ.

Table 4.8 shows the selection of the working covariance structure for the reduced-rank penalized spline smoothing approach. The treatment-dependent structure had the lowest AIC, which suggested that it was the best dependence structure. Figure 4.2 and Figure 4.3 show the estimated function for the non parametric association at different quantiles obtained through, respectively, the unpenalized marginal quantile regression splines and the reduced-rank spline approach. We show the results for the treatment-dependent exchangeable as-

sociation structures only. The nonlinear association was well-captured by the model. The unpenalized functions were wiggly and clearly overfitted the data. The penalized functions were smoothed and seemed to fit the data better.

	Coefficient	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
	Intercept	46.208 (1.521)	62.473 (2.946)	70.472 (2.056)
	Treatment	0	-8.331 (3.664)	-4.759 (3.146)
EX	Time	0	0	0
	TrxTime	0	0	0
	$x_5, \ldots, x_{25}$	0	0	0
	AIC	2815	3962	3166
	Intercept	46.480 (1.512)	62.501 (2.957)	70.494 (2.060)
	Treatment	0	-8.260 (3.654)	-4.727 (3.087)
EX-Trt	Time	0	0	0
	TrxTime	0	0	0
	$x_5, \ldots, x_{25}$	0	0	0
	AIC	2713	3948	3141
	Intercept	45.854 (1.509)	60.792 (3.226)	69.750 (1.810)
	Treatment	0	-7.762 (4.019)	-5.130 (2.955)
TO	Time	0	0	0
	Tr×Time	0	0	0
	$x_5, \ldots, x_{25}$	0	0	0
	AIC	2820	3950	3181
	Intercept	46.874 (1.462)	60.994 (3.077)	69.573 (2.144)
	Treatment	0	-5.551 (3.778)	-3.432 (3.053)
WI	Time	0	0	0
	TrxTime	0	0	0
	$x_5, \ldots, x_{25}$	0	0	0
	AIC	3757	5164	4313

Table 4.7: Estimated quantile regression SCAD coefficients, standard errors (in parentheses, provided for non-zero coefficients only) and logistic models' AIC for three quantiles and different working covariance structures.

Working Structure	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
EX	3885	4478	3482
EX-Trt	3819	4357	3432
TO	3920	4515	3516
WI	4288	5189	4107

Table 4.8: AIC of the logistic regression models in reduced-rank penalized spline smoothing for different working covariance structures and quantiles. EX = exchangeable, TO= Toeplitz, WI = working independence, Ex-Trt = treatment-dependent exchangeable.

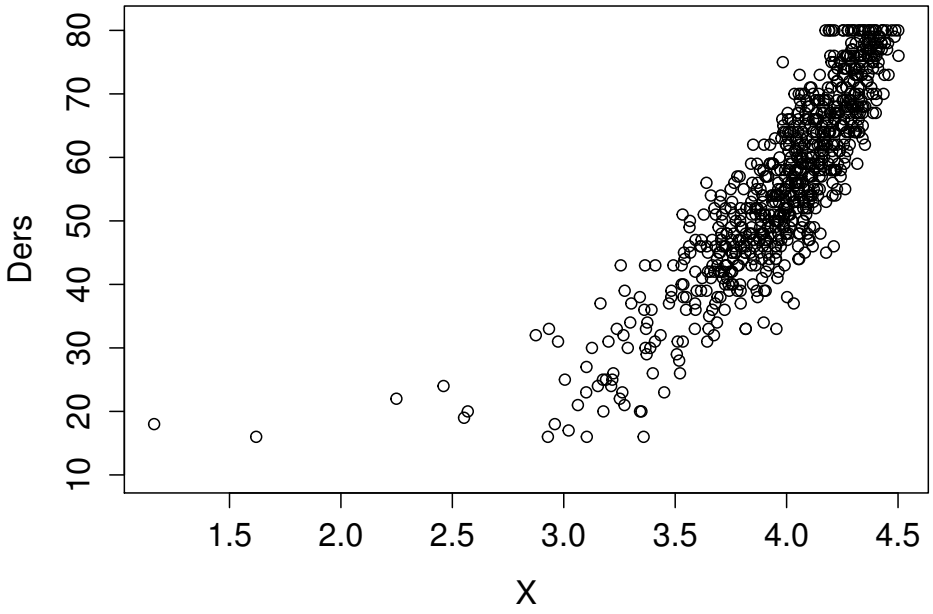

Figure 4.1: Relationship between the response variable Ders and the predictor x_{NP}.

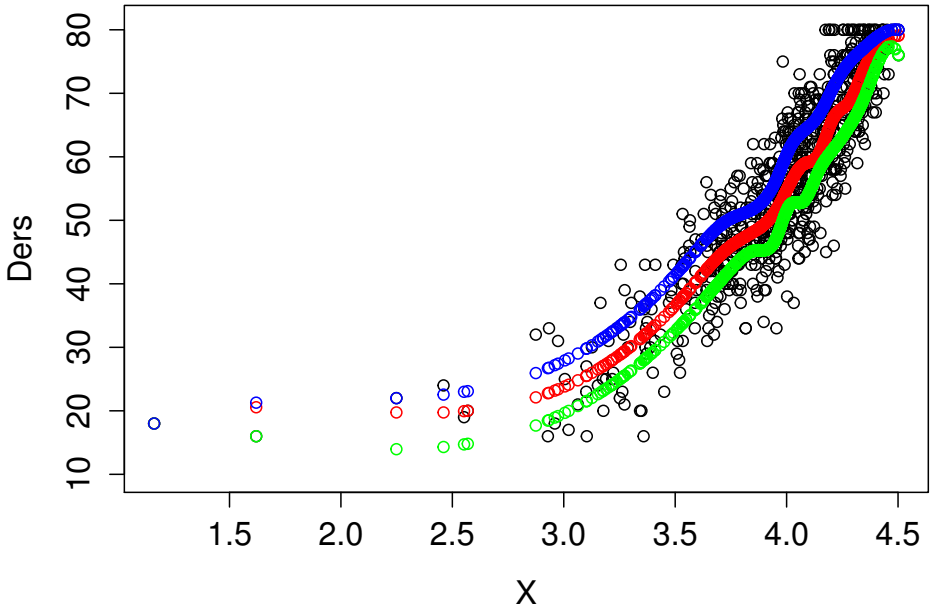

Figure 4.2: Predicted function of the nonlinear association between Ders and x_{NP} obtained through the unpenalized estimator at different quantiles.

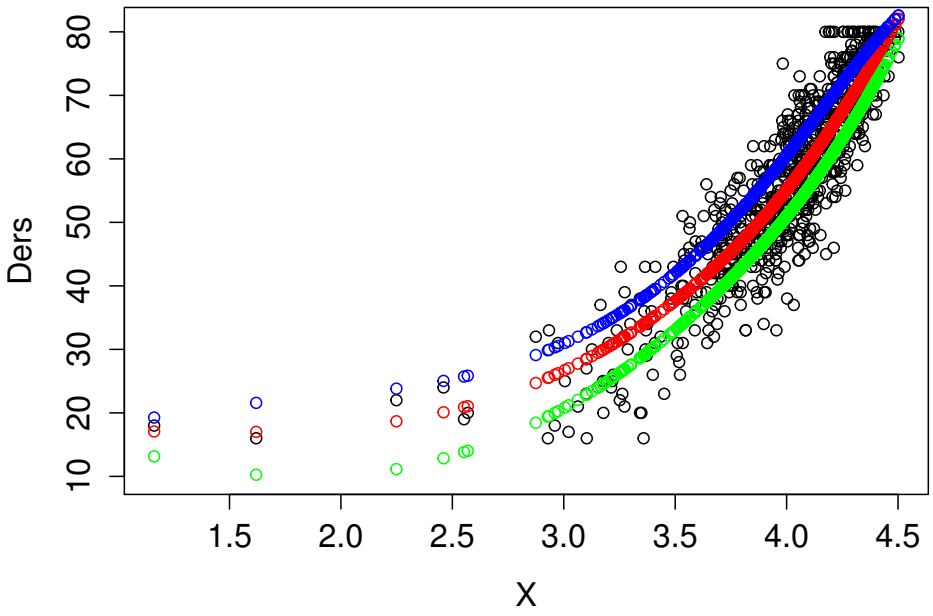

Figure 4.3: Predicted function of the nonlinear association between Ders and x_{NP} obtained through the reduced-rank penalized spline smoothing estimator at different quantiles.

4.5 A summary

This chapter describes a penalized estimating equation approach that accounts for the dependence induced by the clustering. The association matrix of the residuals' sign is estimated by means of appropriate logistic regression models. The estimation is flexible but more computationally intensive than the unpenalized method. The model flexibility allows the within-cluster association to depend on covariates and be multilevel. The best-fitting working structure can be selected by likelihood comparisons or the Akaike's information criterion. Our simulations showed that this criterion could select the true dependence structure with high probability. When the true correlation matrix was specified, the mean squared error of the estimated functions decreased considerably.

Penalized methods performed much better than their unpenalized counterparts. The MM algorithm for regularized SCAD marginal quantile regression was computationally fast and usually converged in less than 30 iterations. When we applied this algorithm to the cognitive-therapy data, we faced some computational issues. The algorithm did not converge for low quantiles ($\tau = 0.25$) when the values of the tuning parameter λ was small, thus we had to choose a tuning parameter that was larger than the ideal one. Therefore, the method shrunk some of the non-zero coefficients to zero. At the moment, the cause of this behavior is not clear to us. The reduced-rank penalized splines smoothing algorithm had a low convergence rate. When the algorithm did not converge, the quantity $||\theta^{K+1} - \theta^{K}||$ was not decreasing at each iteration. This suggests that there could be identifiability problems, perhaps caused by the high correlation among the spline basis of the truncated polynomial.

In general, other loss functions or data-driven measures could perform better when choosing the tuning parameter λ through cross-validation or by minimization. For instance, Chen et al. (2013) proposed to select the tuning parameter in penalized spline regression by minimizing an estimate of the asymptotic mean squared error as the sum of the squared shrinkage bias and the asymptotic variance.

We did not discuss how to obtain a range of values of the tuning parameter λ. Extreme values of λ may cause convergence problems, so it is important to find a range of values where the model can be estimated. The range of values depend on the penalty function used. For certain penalty function such as the lasso, the range of values can be obtained using data-driven methods. To our knowledge, this is not possible for the SCAD penalty. In our simulation study and application, we selected the range using interval methods.

Chapter 5

Conclusions

In this thesis we developed a novel marginal quantile regression method using a working odds ratio matrix. We proposed a penalized version of this method, with applications to reduced-rank penalized splines smoothing and regularized SCAD models. These methods extend the current literature on penalized quantile regression methods. In contrast to previous methods, the proposed quantile estimating equations accounts for the dependence structure of the data. In our simulations, we showed that correctly modeling the dependence induced by the clustering may increase the efficiency of the estimators. Furthermore, penalized methods might outperform unpenalized methods when the model complexity is high. The algorithm proposed in Chapter 3 was computationally fast and had good convergence rate. The MM algorithm of regularized SCAD model was computationally efficient and usually converged in less than 30 iterations. The quasi Newton-Raphson algorithm of reduced-rank penalized splines smoothing encountered several identifiability issues caused by the high correlation among the spline basis functions.

In regularized SCAD models, we did not discuss the more general setting of ultra-high dimensional data. The estimation procedure becomes more tedious because the design matrix is nonsingular and the computational burden may increase drammatically. The Sure Independence Screening method (Fan and Lv (2008)) reduces the dimensionality of the problem to $p < n$ and ensures that all the most relevant predictors are included in the new subset of predictors.

We did not present any theorem regarding the properties of the estimators. For the estimator of Chapter 3 this is not a concern, because we can apply the same theorem used by Fu and Wang (2012). As regards the penalized estimators, we could extend the proofs given by Fu and Wang (2012), Sherwood et al. (2016) and Chen et al. (2013).

Several extensions of the proposed methods might be explored in the future.

For instance, the simultaneous estimation of a set of quantiles. This approach overcomes the problem of quantile-crossing, which is frequent when estimating extreme quantiles. The penalized extension of simultaneous marginal quantile regression could be featured by an uniform tuning parameter for a set of quantile levels, that would avoid some of the potential problems with model selection at individual quantile levels. For independent data, this was proposed by Zheng et al. (2015).

Longitudinal data are often characterized by missing data. Our proposed methods require the data missingness mechanism to be missing completely at random (MCAR), i.e. the events that lead to any particular data being missing are independent both of observable variables and of unobservable parameters of interest, and occur entirely at random (Little and Rubin (2014)). This assumption is rather strong and can be relaxed to data coming from a missing at random (MAR) process. This process occurs when the missingness is not random, but where it can be fully accounted for by variables with complete information. For unpenalized marginal quantile regression, Yi and He (2009) proposed a weighted estimating equation approach that can handle missing at random data. This method can be easily extended to our approaches, even the penalized versions.

Another extension that could be explored are related to censored dependent data. Wang et al. (2013) recently proposed a method to perform variable selection in censored quantile regression that could be extended by modeling the dependence induced by the clustering.

Further extensions of this method are related to the small sample correction of the covariance matrix of the estimator, along with sample size and power calculations. The first occurs when the number of clusters is small and the covariance matrix of the estimator is underestimated. Wang et al. (2015) published a review of methods to correct the covariance matrix of the estimator in marginal mean regression. The sample size and power calculations are useful at the design stage of a study. In general, large samples have a larger chance to discover effects than small ones. However, gathering large samples might be expensive and sometimes impossible. Therefore, it is important to balance sample size and power. There are currently no published works on sample size and power calculations for quantile regression. Teerenstra et al. (2010) proposed a method to compute sample size for three-levels cluster randomized trials where the interest was on the marginal mean.

Bibliography

Bellavia, A., Åkerstedt, T., Bottai, M., Wolk, A., and Orsini, N. (2013). Sleep duration and survival percentiles across categories of physical activity. *American journal of epidemiology*, page kwt280.

Bellavia, A., Discacciati, A., Bottai, M., Wolk, A., and Orsini, N. (2015). Using laplace regression to model and predict percentiles of age at death when age is the primary time scale. *American journal of epidemiology*, page kwv033.

Bottai, M., Cai, B., and McKeown, R. E. (2010). Logistic quantile regression for bounded outcomes. *Statistics in medicine*, 29(2):309–317.

Brown, B. and Wang, Y.-G. (2005). Standard errors and covariance matrices for smoothed rank estimators. *Biometrika*, 92(1):149–158.

Buchinsky, M. (1994). Changes in the us wage structure 1963-1987: Application of quantile regression. *Econometrica: Journal of the Econometric Society*, pages 405–458.

Carey, V., Zeger, S. L., and Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, 80(3):517–526.

Centeno, M. and Novo, Á. A. (2006). The impact of unemployment insurance on the job match quality: a quantile regression approach. *Empirical Economics*, 31(4):905–919.

Chen, H., Wang, Y., Paik, M. C., and Choi, H. A. (2013). A marginal approach to reduced-rank penalized spline smoothing with application to multilevel functional data. *Journal of the American Statistical Association*, 108(504):1216–1229.

Cole, T. J. (1990). The lms method for constructing normalized growth standards. *European journal of clinical nutrition*, 44(1):45–60.

de Luca, F. and Boccuzzo, G. (2014). What do healthcare workers know about sudden infant death syndrome?: the results of the italian campaign genitoripiù. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 177(1):63–82.

Diggle, P. J., Heagerty, P. J., Liang, K.-Y., and Zeger, S. L. (2002). Analysis of longitudinal data second edition. *Oxford statistical science series*, 1(25):ALL–ALL.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008). *Longitudinal data analysis*. CRC Press.

Fu, L. and Wang, Y.-G. (2012). Quantile regression for longitudinal data with a working correlation model. *Computational Statistics & Data Analysis*, 56(8):2526–2538.

Fu, L., Wang, Y.-G., and Zhu, M. (2015). A gaussian pseudolikelihood approach for quantile regression with repeated measurements. *Computational Statistics & Data Analysis*, 84:41–53.

Hall, P. and Sheather, S. J. (1988). On the distribution of a studentized quantile. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 381–391.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2011). *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Hunter, D. R. and Li, R. (2005). Variable selection using mm algorithms. *Annals of statistics*, 33(4):1617.

Jung, S. (1996). Quasi-likelihood for median regression models. *Journal of the American Statistical Association*, 433:251–257.

Koenker, R. (2005). *Quantile regression*. Cambridge university press.

Koenker, R. and Bilias, Y. (2002). Quantile regression for duration data: a reappraisal of the pennsylvania reemployment bonus experiments. In *Economic Applications of Quantile Regression*, pages 199–220. Springer.

Lee, E. R., Noh, H., and Park, B. U. (2014). Model selection via bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, 109(505):216–229.

Leng, C. and Zhang, W. (2014). Smoothing combined estimating equations in quantile regression for longitudinal data. *Statistics and Computing*, 24(1):123–136.

Lipsitz, S. R., Laird, N. M., and Harrington, D. P. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, 78(1):153–160.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Ma, S., Song, Q., Wang, L., et al. (2013). Simultaneous variable selection and estimation in semiparametric modeling of longitudinal/clustered data. *Bernoulli*, 19(1):252–274.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57(1):120–125.

Pratesi, M., Ranalli, M. G., and Salvati, N. (2009). Nonparametric m-quantile regression using penalised splines. *Journal of Nonparametric Statistics*, 21(3):287–304.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, pages 1033–1048.

Seber, G. and Wild, C. (2003). *Nonlinear regression*. Wiley, New York.

Sherwood, B., Wang, L., et al. (2016). Partially linear additive quantile regression in ultra-high dimension. *The Annals of Statistics*, 44(1):288–317.

Teerenstra, S., Lu, B., Preisser, J. S., Van Achterberg, T., and Borm, G. F. (2010). Sample size considerations for gee analyses of three-level cluster randomized trials. *Biometrics*, 66(4):1230–1237.

Touloumis, A., Agresti, A., and Kateri, M. (2013). Gee for multinomial responses using a local odds ratios parameterization. *Biometrics*, 69(3):633–640.

Vigerland, S., Ljótsson, B., Thulin, U., Öst, L.-G., Andersson, G., and Serlachius, E. (2016). Internet-delivered cognitive behavioural therapy for children with anxiety disorders: A randomised controlled trial. *Behaviour research and therapy*, 76:47–56.

Wang, H. J., Zhou, J., and Li, Y. (2013). Variable selection for censored quantile regresion. *Statistica Sinica*, 23(1):145.

Wang, K.-M. (2011). Health care expenditure and economic growth: Quantile panel-type analysis. *Economic Modelling*, 28(4):1536–1549.

Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 68(2):353–360.

Wang, M., Kong, L., Li, Z., and Zhang, L. (2015). Covariance estimators for generalized estimating equations (gee) in longitudinal analysis with small samples. *Statistics in Medicine*.

Yi, G. Y. and He, W. (2009). Median regression models for longitudinal data with dropouts. *Biometrics*, 65(2):618–625.

Zheng, Q., Peng, L., and He, X. (2015). Globally adaptive quantile regression with ultra-high dimensional data. *Annals of statistics*, 43(5):2225.

Davide Bossoli

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 049 827 4174
e-mail: bossoli@stat.unipd.it

Current Position

Since January 2014; (expected completion: March 2017)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Extensions of quantile regression for dependent data
Supervisor: Prof. Matteo Bottai

Research interests

- Dependent data
- Quantile regression
- Generalized estimating equations
- Penalized methods

Education

October 2011 – July 2013
Master degree (*laurea specialistica/magistrale***) in Statistics** .
University of Bologna, Faculty of Statistics
Title of dissertation: "Logistic quantile mixed models - an application to health condition data "
Supervisor: Prof. Rossella Miglio
Cosupervisor: Prof. Matteo Bottai
Final mark: 110/110 cum laude

September 2008 – July 2011
Bachelor degree (*laurea triennale***) in Statistics**.
University of Bologna, Faculty of Statistics
Title of dissertation: "Statistical models for evaluating competences "
Supervisor: Prof. Stefania Mignani
Final mark: 110/110 cum laude.

Visiting periods

March 2015 – November 2016
Karolinska Institutet,
Stockholm, Sweden.

Supervisor: Prof. Matteo Bottai

April 2013 – December 2013
Karolinska Institutet,
Stockholm, Sweden.
Supervisor: Prof. Matteo Bottai

Further education

October 2015
Categorical data analysis
University of Padua
Instructor: Alan Agresti (University of Florida)

Work experience

September 2011 – March 2012
Municipality of Bologna.
Internship.

Awards and Scholarship

September 2013
University of Bologna's graduate scholarship.

April 2013
University of Bologna's graduate scholarship.

Computer skills

- R, Latex (advanced)
- Sas, Stata, Microsoft Office (intermediate)

Language skills

Italian: native; English: fluent.

Conference presentations

Bossoli, D., Bottai, M. (2016). Marginal quantile regression with a working association matrix. (contributed) *International society for clinical biostatistics conference*, Birmingham, United Kingdom, 22-26 August.

Bossoli, D., Bottai, M. (2016) Quantile regression for dependent data using a working odds ratios matrix. (contributed) *International conference on robust statistics*, Geneve, Switzerland, 4-8 July.

Bossoli, D., Bottai, M. (2016) Effect of internet-based cognitive theraphy on children anxiety disorders: results from a marginal logistic quantile regression. (poster) *Meeting of the statistical italian*

society, Salerno, Italy, 8-10 June.

Teaching experience

April 2016
Biostatistics I
Epidemiology
Lab sessions, 12
Karolinska Institutet
Instructor: Prof. Matteo Bottai

February 2016
Biostatistics II
Medicine
Lab sessions, 12
Karolinska Institutet
Instructor: Prof. Matteo Bottai

References

Prof. Matteo Bottai
Karolinska Institutet
Nobels vag 13, Stockholm, Sweden
Phone: +46 (0) 8-524 870 24
e-mail: matteo.bottai@ki.se