T E S I   D I   D O T T O R A T O

Dipartimento di Scienze Economiche Aziendali e Statistiche

P-spline Quantile Regression:
a new algorithm for smoothing parameter selection

FEDERICO TORRETTA

Tutor: **Vito Muggeo**

Coordinatore Dottorato: **Marcello Chiodi**

Dottorato di Ricerca in "Statistica, Statistica Applicata e Finanza Quantitativa", XXVI Ciclo
Settore Scientifico Disciplinare: SECS/S01 - Statistica
Anno conseguimento titolo: 2016

Università degli Studi di Palermo

DSEAS

Dipartimento di Scienze Economiche, Aziendali e Statistiche

To my family

Acknowledgements

Contents

Introduction

This thesis is concerned with quantile regression (QR). This technique is becoming popular mainly because, compared to mean regression, it provides a complete description of the relationship between a response variable and a set of covariates. It is also attractive because it does not require any distributional assumption. With a non parametric approach it is possible to achieve an estimate of the so-called quantile function without specifying the form of the predictor. In this work, the attention is particularly focused on P-spline QR and the choice of the smoothing parameter. The main contribution of the thesis consists of the implementation of an iterative algorithm which allows to get an optimal smoothing parameter. Simulations are reported to see how the method performs compared with some alternatives; furthermore an application in agronomy is presented to show how it works in practice.

In this chapter, some motivations to move from mean regression to quantile regression technique and from parametric to non parametric approach are discussed. Since the thesis is going to deal with QR, some examples in which QR represents a better choice than mean regression are presented.

From mean regression to quantile regression

In several fields of applied sciences, researchers need to model data to achieve a description of the relationship among a response variable and a set of explanatory variables or covariates. The idea is that, using a set of statistical techniques, it is possible to validate (or not) some theoretical properties of the object of the study. This allows to get answers to some research questions aimed to scientist's decisions. The most common technique to study those relationships is mean regression analysis. Mean regression is considered a simple but powerful method because, relying on a small set of assumptions, it is computationally easy to estimate; furthermore, the interpretation of the parameters ruling the generator data process is generally quite simple. In the last two centuries mean regression was applied in almost every field of scientific knowledge, from Economics to Biology, from Psychology to Engineering. For many years mean regression was presented as the only tool to study dependence of variables but *sometimes it does not help to describe the phenomenon which the researchers are interested in*.

An example of possible shortcomings using mean regression is when one has to estimate growth curves. There are many ways to estimate a growth model using mean regression. Since growth curves are usually supposed to be non linear, Logistic, Gompertz, Richards or Weibull models can be used (see Zimmerman *et al.* (2001) for further details). According to these models, the growth rate changes only for the mean; however, it could be useful to know if the growth rate changes constantly varying quantile or not. For instance, in the work of de Onis (2006) one of the major purposes is to analyse some anthropometric variables (BMI, height, weight) for children in the early ages. Mean regression can be unreliable when it does not take into account that there are children who are born heavy (or tall) and

will naturally belong to the upper tail of the distribution and children who are born light (or short) and will naturally belong to the lower tail of the distribution. The use of the mean in this context implies the assumption that the growth rate is the same both for heavy and light children. So it is important to provide more specific information to get a real idea of how those variables grow along with time.

This problem is conceptually easy to solve with a quantile regression (QR) (Koenker, 2005; Davino *et al.*, 2013) approach. Quantile regression represents a more complete alternative to mean regression because it provides a description of the whole conditional distribution of the response given a set of covariates. It is easy to see how, looking at the plot in Figure 1.

Quantile regression analysis can help to avoid a too complex model specification and a set of heavy distributional assumptions which do not hold in many real situations. Moreover it is easy to incorporate in other frameworks: Engle and Manganelli (2004) compute a quantile autoregressive model for Var estimation, giving birth to CAViaR models. Muggeo *et al.* (2013) estimate a growth model for height and weight of *Posidonia oceanica* via non parametric quantile regression. Eilers and De Menezes (2005), in order to find changes in copy numbers along chromosomes, propose to shift the problem from L_2-norm (based on least squares) to L_1-norm (based on least absolute values); in other words, they move from mean regression to median regression.

The estimation procedure is done via linear programming. There are some interesting properties which make QR approach quite attractive. One of them is *robustness*. It is known that the mean is very sensible to outliers; in other words, large (or small) values in the sample affect the estimation of the mean. That does not happen for quantiles. Anyway, robustness of QR
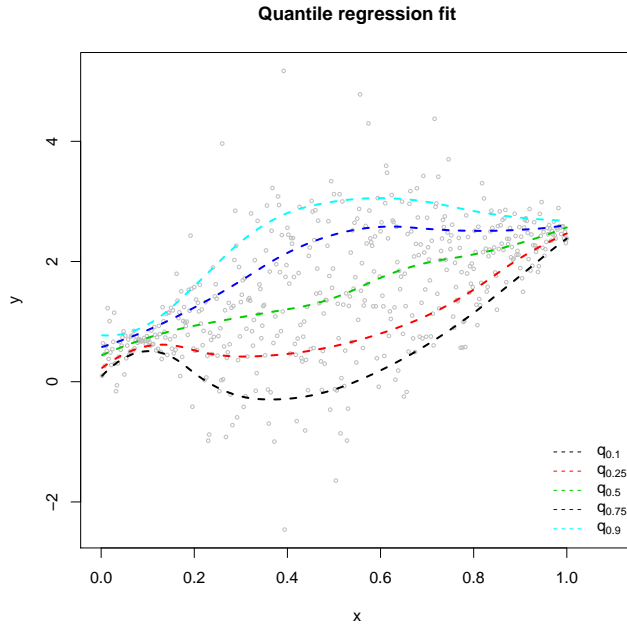
Figure 1: Quantile regression fit. It is possible to include a curve for each quantile function; in practise, it is possible to depict the whole conditional distribution instead of evaluating only the mean.

approach does not concern only outliers: in fact, it allows the researcher to get reliable estimates even when some distributional assumptions do not hold; QR is potentially distribution free. Another important feature of quantile regression is *equivariance* (see Koenker and Bassett (1978)). Equivariance permits to monotonically change the scale of the response or the parametrization of the model without affecting the results of the estimates. So there is no need to change the modelling approach like in case of Generalized Linear Model framework: one can just transform the data without any problem.

It is worthwhile to state that mean regression can be also used to get an estimate of the whole conditional distribution and hence the so-called distributional regression. In fact it is possible to achieve a quantile estimation from any mean regression model (Kneib (2013) for further information). Efron (1991) describes the procedure to assign regression percentiles from a regression model: given a linear regression model, $y_i = \mu(x_i) + \epsilon_i$ with i.i.d. errors, $\epsilon_i \sim N(0, \sigma^2)$, then

$$q_{100\alpha} = \hat{\mu}(x) + \hat{\sigma} z_\tau,$$

where z_τ is the 100τ-th quantile given a probability level τ of the normal distribution. It is easy to note that the fit will be represented by parallel lines, usually based on the same functional form, displaced by $\hat{\sigma} z_\alpha$. That can be a problem especially when data are characterized by heteroskedasticity; the simulated example plotted in Figure 2 shows how heteroskedasticity affects the reliability of the fitted curves. It turns out that a more flexible tool is needed to handle this kind of data and QR represents a valid alternative.

From parametric to non parametric approach

Many real phenomena result very difficult to model because it is not easy to find a prior functional form.

An example of importance of good model specification is when logistic regression is used instead of the linear model for binomial response. The relationship between response and covariates follows a logistic curve avoiding problems on the scale of the response, on coherence of results and so on. Unfortunately, it is not always obvious to decide a non linear functional form to model relationships. A common way to solve this problem is to
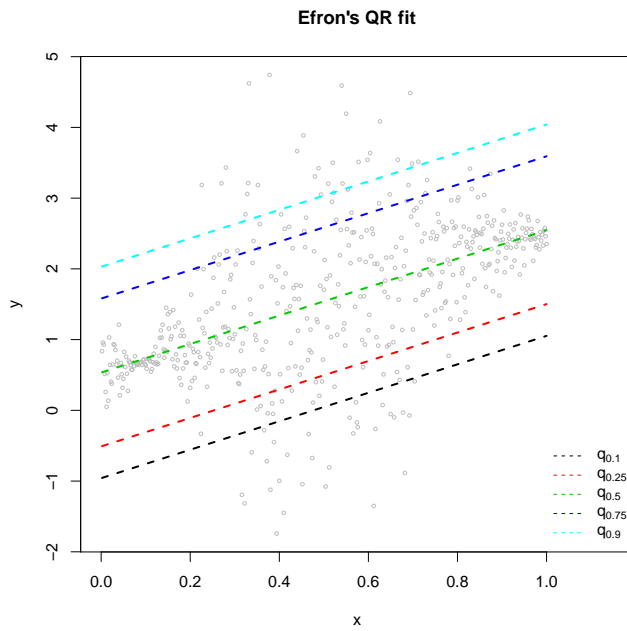
Figure 2: Quantile regression fit from a linear model. The approach proposed by Efron (1991) with heteroskedastic data does not work properly. Although the relationship is linear in the mean, the estimated quantile curves are characterized by a clear lack of fit.

use a *non parametric approach*. Instead of assuming a specific function, this flexible approach makes data speak for themselves using automatic procedures with the advantage of reducing the problem of model predictor specification to the choice of just one or few parameters. Since those parameters rule the regularity or the smoothness of a curve they are called *smoothing* parameters. On the other hand, the non parametric approach has the disadvantage of losing interpretation of parameters. Many interesting methodologies of non parametric Statistics are included in the books of Ruppert *et al.* (2003) and Wood (2006). In this work, P-spline (Eilers and Marx (1996)) will be the main smoother used to get a non parametric fit of QR. P-splines are a very flexible tool and very easy to incorporate in QR framework. P-spline smoother minimizes a "fidelity plus penalty" function. The penalty is weighted by the smoothing parameter.

The choice of smoothing parameter is crucial: according to this quantity, one can get a very regular (or smooth) fit or a very rough one. There are many ways to choice this parameter: first of all, according to the problem to deal with, the choice can be done just by visualizing data or using more technical procedures. In the second case, there are many alternatives such as information criteria (AIC or SIC) or cross validation and several extensions. Most of these methods are based on the specification of a grid of values. As we will see in the next chapters, these methods could make troubles especially when many covariates are involved in the analysis. The main problem is the computational burden derived from those approaches. In fact, to apply one of the aforementioned criteria, one has to estimate the model for any value of the grid and select the optimal model according to the best value of the criterion. An idea to fix this problem is to use an iterative algorithm which does not work on grid of values. This is the seminal

idea behind our research.

Contribution of the thesis

The main purpose of this thesis is to present an iterative algorithm aimed to smoothing parameter estimation for P-spline QR. The algorithm is an extension to QR of the one described by Schall (1991) which relies on the ratio of variances in the mixed model framework to estimate smoothing parameter in mean regression. This approach is advantageous respect to common grid search methods because it avoids the use of grids which can either lead to computational problems (especially with a large number of covariates) or to misleading results in case they are not properly set up.

To extend the algorithm, several issues has to be considered: first of all, until now Schall algorithm was used for L_2-norm problems but it is known that the estimation of QR comes from an L_1 problem. Hence, instead of considering the ratio of variances we use the standard deviation ratio.

Moreover, since the concept of standard deviation is based on the mean, usually it is not simple to establish a measure of a quantile-based variance. We propose several solutions mainly based on the use of the Asymmetric Laplace distribution, following the approach used by Geraci and Bottai (2007).

Finally we propose a new computation of the ED based on an approximated hat matrix explained in the work of Muggeo *et al.* (2012). Currently, the unique method to evaluate the residual ED in QR is based on the number of non-zero residuals in the model.

The outline of the work is as follows: in the next chapter an overview on parametric and non parametric quantile regression is presented: due to the

main purposes of the thesis, the non parametric approach will be dealt with reporting also some basics about P-splines. The second chapter concerns smoothing parameter selection: after explaining some grid search methods, Schall algorithm and our proposal will be discussed. The third chapter will be entirely focused on results derived from simulation studies with the aim to analyse the performance of the iterative algorithm as compared to grid-search criteria.

In the fourth chapter an application of the methodology to real data will be illustrated. The application regards the comparison between a traditional and a compost based fertilizer in terms of root length of Sorghum bicolor Moench x S. sudanense (Piper).

Finally the thesis will end with some conclusions and possible future works.

Chapter 1

Quantile regression model

This chapter is devoted to the introduction to Quantile Regression. It is divided into two main parts: the first part consists of a presentation of the parametric approach. In the second part, non parametric approach to QR will be discussed. To know properly how non parametric QR works, some considerations on P-splines and smoothing splines will be addressed in the chapter. Then, shape constraints to get reliable estimated curves will be described. Finally, some alternatives to QR, namely expectiles modelling (Waltrup *et al.*, 2014) and LMS method (Cole and Green, 1992) will be briefly discussed.

1.1 Quantile regression framework

The idea of using quantiles to model relationships was introduced by Koenker and Bassett (1978). In their seminal work, with the aim of fixing the problem of robustness of an estimator, the authors propose a minimization problem to produce the ordinary sample quantiles. In the same work, the authors

prove the asymptotic properties of the quantile regression estimator, the superiority of median regression respect to mean regression over a class of non-Gaussian distributions and several equivariance properties.

The common definition of quantile is, given a random variable Y with distribution function $F(y) = P(Y \leq y)$ and a probability level τ,

$$q_\tau = F^{-1}(\tau) = \inf\{y : F(y) \geq \tau)\}, \qquad 0 \leq \tau \leq 1.$$

Starting from an exercise of decision theory in the book of Ferguson (1967), Koenker and Bassett (1978) define sample quantile as minimizer of the function

$$\tau \sum_{e_i \geq 0} |e_i| + (1 - \tau) \sum_{e_i \leq 0} |e_i| = \sum_i^n e_i \left(\tau - I(e_i \leq 0)\right) = \sum_i \rho_\tau(e_i) \qquad (1.1)$$

where $e_i = y_i - q_\tau$ and $\rho_\tau(\cdot)$ is called *check function*. The definition of the quantile by minimization is attractive because it allows to incorporate a regression model for q_τ. For the sake of simplicity just one covariate x will be considered. The *linear quantile regression model* is the combination of the covariate x and the parameter β_τ, namely

$$q_\tau(Y|x_i) = x_i^T \beta_\tau, \qquad (1.2)$$

It is important to note that, only the assumption of τ-th quantile equal to zero is required.

If we have a model of the form

$$y_i = x_i^T \beta_\tau + \epsilon_{i\tau},$$

and considering the assumption mentioned above

$$F_{\epsilon_i}(0) = \tau \qquad \Rightarrow \qquad q_\tau(Y|x_i) = x_i^T \beta,$$

because

$$\tau = F_{\epsilon_{i\tau}}(0) = \mathbb{P}(\epsilon_{i\tau} \le 0) = \mathbb{P}(x_i^T \beta_\tau + \epsilon_i \le x_i^T \beta_\tau) = \mathbb{P}(y_i \le x_i^T \beta_\tau) = F_{Y_i}(x_i^T \beta_\tau).$$

The objective function (1.1) is based on least absolute deviation or L_1-norm and it is piecewise linear as shown in Figure 1.1. The special case of the median, i.e. when $\tau = 0.5$, weights positive and negative residuals in the same way; for a given $\tau \ne 0.5$ the weight is asymmetrical and changes along with the probability level: low probability levels correspond to low weights for positive residuals and high weights for negative ones. Looking at the picture, there is a sharp point in zero: it means that the right and left derivatives are different and this makes the $\rho_\tau(\cdot)$ function not differentiable for residuals equal to zero. Since the contribution to the objective function is given by a sum of many check functions, the objective is not differentiable as well. Therefore, standard methods usually based on least squares algorithms do not work. That is why, to minimize such kind of objective function, algorithm based on linear programming are required.

Afterwards, many works to improve the algorithm estimation initially based on simplex methods (Koenker and d'Orey, 1987) have been published. One of the most important, by Portnoy *et al.* (1997), describe a modification of
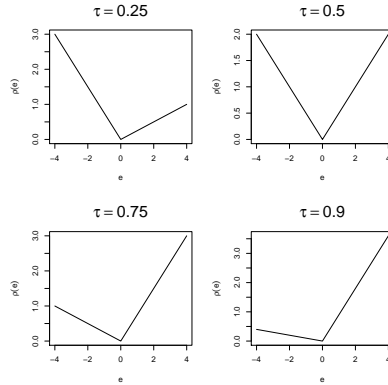
Figure 1.1: Plot of the check function $\rho_\tau(\cdot)$ for $\tau = \{0.25, 0.5, 0.75, 0.9\}$. As probability levels increase, weights to positive residuals increase. For the median, the weight for positive and negative residuals is the same therefore the function is symmetrical.

the Frisch-Newton algorithm to estimate QR which results advantageous for large problems (i.e many observations). All these algorithms are included in the R package *quantreg* (Koenker, 2013).

1.2   Non parametric quantile regression

As already sketched in the Introduction, it is possible to estimate the whole distribution via QR. Just like in case of mean regression a linear specification of the (1.2) could be not enough to describe properly the relationship. Several non-parametric approaches were developed to fix this issue, for instance Chaudhuri (1991) finds a method to estimate QR using local polynomial regression approach. However, kernel approaches are known being computationally heavy and having boundary problems. Koenker *et al.* (1994) use quantile smoothing splines with a total variation

penalty term to get a flexible estimation of the model and imposing some inequality constraints (Koenker, 2005) to guarantee particular features of the estimated curves such as monotonicity or convexity. The total variation penalty depends on the difference of the absolute values of the first derivative of adjacent coefficients of the splines. A similar penalization method is provided by Ng and Maechler (2007) to implement the so-called constrained B-splines (COBS). Yet another approach, the one used in this thesis, consists of the implementation of QR model via P-spline (Bollaerts *et al.*, 2006). A big advantage of non parametric techniques is that they can be easily included in other frameworks: for instance Andriyana *et al.* (2014) implement a varying coefficient QR model via P-spline.

With a non parametric approach, the expression of the linear term to model quantile function in (1.2) is replaced with

$$q_\tau(Y|x_i) = s(x_i), \tag{1.3}$$

where $s(\cdot)$ is an unknown and possibly non linear function. For simplicity, only one non linear covariate is considered in the model but it can be possible to include a semiparametric specification. Most of the differences among the aforementioned approaches concern just the methods to obtain a smooth estimation of the function $s(\cdot)$.

1.2.1 Quantile smoothing splines

One of the most known non parametric approach for QR is given by Koenker *et al.* (1994). In their work, they describe a method to include smoothing spline in a QR framework. A spline is defined as follows (Green and Silverman, 1993): given n points $x_1 < x_2 < \cdots < x_n$ in an inter-

val (a, b), a q-degree spline is a function s characterized by a q-degree function for each interval $(a, x_1), (x_1, x_2), \ldots, (x_n, b)$ such that s, s' and s'' are continuous. Because of optimality conditions natural cubic splines are often used. The main characteristic of natural cubic splines is that $s''(a) = s''(b) = s'''(a) = s'''(b) = 0$. It is always possible to interpolate a given set of values by natural cubic splines. This property helps to extend the approach in mean regression framework: to limit the fact that the estimated curves perfectly interpolate the points giving too rough estimates, a penalized least squares approach can be used. The penalty term is given by the integral of the squared second derivative of the fitted function. The idea is that a very smooth curve has a very low $\int [s''(x)]^2 \, dx$. The term is quadratic to consider in the penalty also the part of the curve where $s''(x) < 0$.

In QR framework, a smoothing spline model aims to minimize the following objective function

$$S = \sum_{i=1}^{n} \rho_\tau \Big(y_i - s(x_i)\Big) + \lambda \int [s''(x)]^2 \, dx. \qquad (1.4)$$

The amount of penalty is ruled by the *smoothing parameter* λ, with $\lambda \in [0, +\infty)$. The most important features of this quantity will be described in the next section of the chapter. An alternative way to express the penalty is based on the use of the so-called *total variation penalty*. The total variation function V of a generic function s is defined as

$$V(s) = \sup \sum_{i=1}^{n} |s(x_{i+1}) - s(x_i)|.$$

When s is absolutely continuous then

$$V(s) = \int |s'(x)|dx.$$

Applying this function to the first derivative of the $s(\cdot)$,

$$V(s') = \sup \sum_{i=1}^{n} |s'(x_{i+1}) - s'(x_i)|,$$

and for absolutely continuous s' it turns out

$$V(s') = \int |s''(x)|. \tag{1.5}$$

Therefore the objective function is

$$S = \sum_{i=1}^{n} \rho_\tau \Big(y_i - s(x_i) \Big) + \lambda V(s'(x_i)). \tag{1.6}$$

The function $s(\cdot)$ which minimizes (1.6) has a piecewise linear form with knots corresponding to the observations $(x_1, x_2, \ldots, x_n)$. This means that $\hat{s}(x) = \alpha_i + \beta_i(x - x_i)$ for $x \in [x_i, x_{i+1})$ and $i = 0, \ldots, n$. This implies

$$\begin{aligned}
\beta_i &= \frac{s(x_{i+1}) - s(x_i)}{x_{i+1} - x_i} \\
&= \frac{\alpha_{i+1} - \alpha_i}{h_i},
\end{aligned}$$

where $h_i = x_{i+1} - x_i$. So the penalty can be expressed as

$$V(s') = \sum_{i}^{n-1} |\beta_{i+1} - \beta_i|$$

$$= \sum_{i}^{n-1} |(\alpha_{i+2} - \alpha_{i+1})/h_{i+1} - (\alpha_{i+1} - \alpha_i)/h_i|$$

and the objective to minimise becomes

$$S = \sum_{i=1}^{n} \rho_\tau \left(y_i - \alpha_i \right) + \lambda \sum_{i}^{n-1} |d_j^T \alpha|, \tag{1.7}$$

where $d_j^T = (0, \ldots, 0, h_j^{-1}, -(h_{j+1}^{-1} - h_j^{-1}), h_{j+1}^{-1}, 0 \ldots, 0)$ and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$. Thus, according to (1.7) the model matrix of QR smoothing spline problem can be expressed as a set of dummy variables.

1.2.2   P-spline quantile regression

Smoothing splines are a good tool to estimate but they are quite inefficient for at least two reasons: first of all, with smoothing splines there is one knot for each observed x_i; in other words, the model matrix has a column for each unit in the sample. It means that the number of evaluated splines increases linearly with n. Nowadays most statistical application are based on big data so the computational burden of this tool can be excessive. Secondly, quantile smoothing splines results in a piecewise linear function. Unless there is a good theoretical reason to consider this pattern of data (for instance, see Eilers and De Menezes (2005)), this approach is not generally advisable.

An alternative idea is to consider a B-spline approach. Basically, the method consists of changing the basis of the space spanned by the model matrix

with q-degree overlapping spline basis. This basis results in a low rank matrix so that the number of columns will be smaller than the number of observations.

For instance, let consider a mean regression model of the form

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i$$

Thus $X = [1 \ x \ x^2 \ x^3]$ is a canonical basis for the vector space for all 3-rd degree polynomials in x. A linear combination of the basis with the parameters provides the fit. Instead of considering the X matrix, one can include a spline basis matrix B for $m+1$ knots. The plot in Figure 1.2 shows the basis for the linear model and some B-spline bases.

The number of B-splines used in the model is usually $J = m + q$ where m is the number of intervals of the domain split by knots and q is the degree of the polynomials. It is clear that the degree q of the spline does not affect too much the number of columns of B: a common optimality criterion is to choose 3-rd degree B-splines. A more important problem is to determine number and position of knots: as the B-spline is more flexible (i.e. many knots) to fit the data, the bias generally decreases, while the variance increases. So when the flexibility level is too high there is the risk to have meaningless and too wiggly estimated curves, a phenomenon called *undersmoothing*. There is a large amount of publications to find a procedure aimed to get an optimal number and position of knots, see for instance the work of Friedman and Silverman (1989). However it looks simpler to choose a high number of equally spaced knots (a common choice is $\min(40, n/4)$) and again penalize the integral of the second derivative of the fitted curve to avoid undersmoothing. Eilers and Marx (1996) replaced the penalty term with the sum of squares of the d-order differences among the
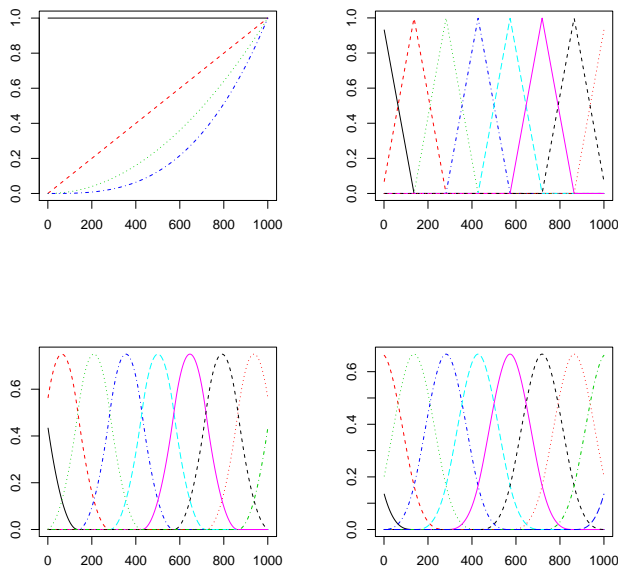
Figure 1.2: Plot of the basis for 4 given models. In the top left, basis for a cubic parametric model. Elsewhere, spline basis with degree $q = \{1(topright), 2(bottomleft), 3(bottomright)\}$. The number of columns for spline regression is $J = m + q$, where m is the number of intervals and q the degree of the spline basis. In this example, $m = 7$.

adjacent coefficients of the B-splines. This quantity is expressed via proper difference matrix D^d.

Then, in mean regression framework the objective function is L_2 norm and the penalty is usually L_2 norm.

$$\mu_i = B_1(x_i)a_1 + B_2(x_i)a_2 + \cdots + B_J(x_i)a_J = B(x_i)a \qquad (1.8)$$

where $\mu_i = \mathbb{E}[y|x_i]$, $B_j(x_i)$ is the $j-th$ spline basis and a_j is the $j-th$ coefficient.

The objective function is

$$S_2 = \sum_{i=1}^{n} (y_i - \mu_i(a))^2 + \lambda \sum_{j=1}^{J-d} (D^d a)_j^2, \qquad 0 \le \lambda \le +\infty \qquad (1.9)$$

The estimation of the vector of coefficients $\hat{a}$ can be computed via standard least squares method. It is important to underline that the estimated coefficients have no statistical meaning: they just scale the basis.

In QR context, according to the model in equation (1.3) and using a P-spline approach, we have

$$q_\tau(Y|x_i) = s(x_i) = B_1(x_i)a_1 + B_2(x_i)a_2 + \cdots + B_J(x_i)a_J = Ba. \qquad (1.10)$$

Of course the objective function is different respect to mean regression framework because is based on the L_1-norm for the fidelity term. The objective function could be

$$S_1 = \sum_{i=1}^{n} \rho_\tau \left(y_i - s(x_i) \right) + \lambda \sum_{j=1}^{J-d} (D^d a)_j^2. \qquad (1.11)$$

This formulation of the objective function has a L_1 norm fidelity term and a L_2 norm penalty term. Many authors tried to solve the quadratic programming problem in (1.11); for instance Pratesi *et al.* (2009), use a non parametric M-quantile approach and Nychka *et al.* (1995) provide an application in quality air analysis: these works are based on penalised iterative reweighted least squares (PIRLS) technique.

However, Koenker *et al.* (1994) stated that the resulting quadratic program poses serious computational obstacles. This situation leads to consider a quite natural solution: use a L_1 norm penalty which is based on the absolute differences between the adjacent coefficients, therefore

$$S_1 = \sum_{i=1}^{n} \rho_\tau \left(y_i - s(x_i) \right) + \lambda \sum_{j=1}^{J-d} |D^d a|_j. \qquad (1.12)$$

The order of the differences d does not affect the estimated curve for L_2 norm problems as illustrated in Figure 1.3.

On the other hand, the L_1-norm estimations are piecewise $d - 1$ order polynomials, so the order of the differences is not a negligible quantity. The plot in Figure 1.4 describes this situation: it is clear that $d \geq 3$ provides a rather smooth fit; for $d = 1$ a step function and for $d = 2$ a piecewise linear function are achieved. A plausible idea is to choose $d \geq 3$ unless particular reasons to use piecewise linear or step function (see Eilers and De Menezes (2005) for an example of piecewise constant QR).

The smoothing parameter λ is fundamental to determine the amount of smoothing provided in the estimation procedure. When λ is large, even
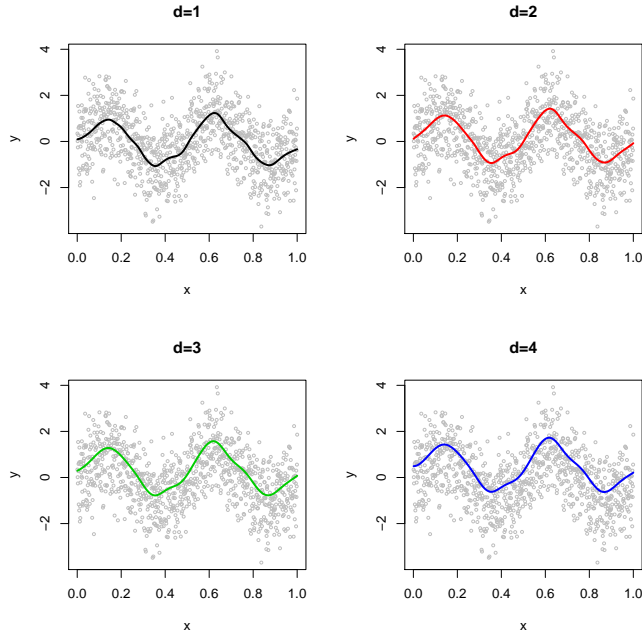
Figure 1.3: P-spline mean regression varying degree of differences of the penalty term. The pattern of estimated curves seems to be not sensible to changes in the order of differences, d. For graphical reasons, the fitted curves are shifted each other by 0.15.

small changes in adjacent coefficients are amplified; this fact determines a strong penalty on the second derivative of the estimated function which constrained the curve to be smooth.

As $\lambda \to \infty$ the fitted curve tends to a $d - 1$ polynomial. On the other hand, when λ is small the differences among the coefficients are considered negligible; in P-spline QR this means to have a piecewise $d - 1$ order polynomials as fitted curve. In the limit case of $\lambda = 0$ the fit will correspond to a B-spline QR. Figure 1.5 illustrates what happens varying the smoothing
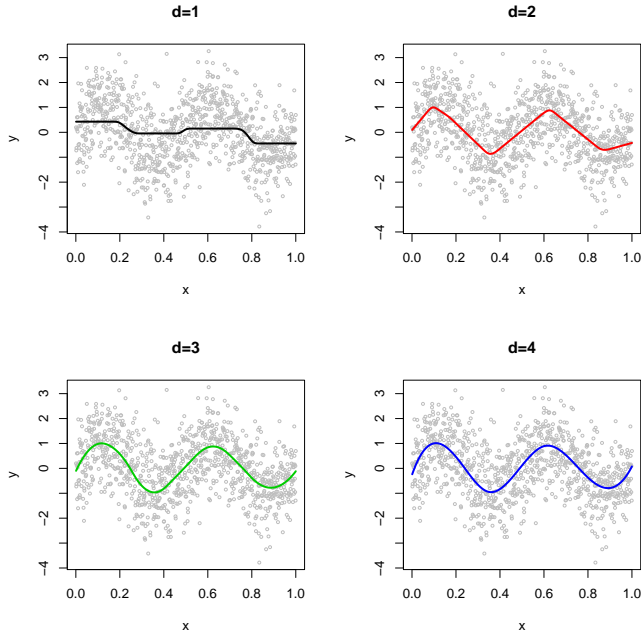
Figure 1.4: P-spline median regression varying degree of differences of the penalty term. Because of the L_1-norm fidelity plus L_1-norm penalty objective function, the fit is always piecewise and the polynomials change according to d.

parameter. In the simulated example, data present a clear heteroskedastic pattern despite of a constant relationship between the response and the co-variate. The order of the differences is 3: the B-spline fit (i.e with $\lambda = 0$) shows a very wiggly pattern but as λ increases, the estimated curves tend gradually to be quadratic polynomials.

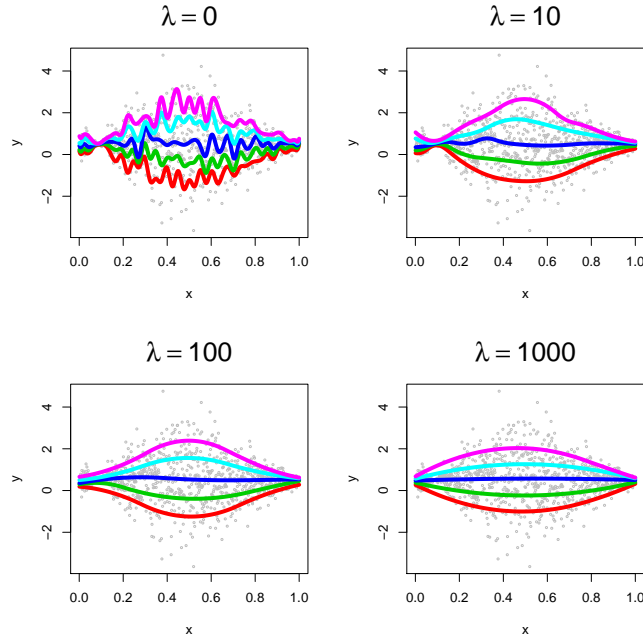Figure 1.5: P-spline quantile regression varying the smoothing parameter λ. The data were generated using $y_i = 0.5 + g(x)\epsilon_i$, $g(x) = x + \sin(1.5\pi x - .5)$. The asymptotic property $\lambda \to \infty \Rightarrow \hat{q}_\tau \to P_{d-1}(x)$ is shown.

1.3 Some alternatives to non-parametric QR

Expectiles

Expectile smoothing is one alternative to quantile regression. The advantage of using expectiles relies on the estimation method which allows to avoid linear programming. The first work on this topic dates back to Newey and Powell (1987) and deals with the asymmetric least squares estimation problem; other more recent papers on the topic are provided by Waltrup *et al.* (2014) and Sobotka *et al.* (2011). As already seen, quantiles are de-

fined by minimization of the (1.1) which can be expressed as

$$\sum_{i}^{n} w_i(\tau)|y_i - q_{\tau,i}|,$$

where w_i is τ for positive residuals and $1 - \tau$ for negative ones. It is easy to note that quantile regression solves a weighted least absolute problem. Instead of minimizing absolute values expectiles solve a weighted least squares problem. The definition of expectile relies on least asymmetrically weighted squares, namely

$$\arg\min_{\mu_{\tau,i}} \sum_{i}^{n} w_i(\tau)(y_i - \mu_{\tau,i})^2 \tag{1.13}$$

It is worthwhile to underline that expectiles are averages; the probability levels τ is called *asymmetry*. Since the estimation of the expectiles is based on the L_2-norm, $\hat{\mu}_{\tau,i}$ is more efficient than $\hat{q}_{\tau,i}$. Moreover, Jones (1994) shows that expectiles and M-quantiles are quantiles for a quite particular transformation of data. On the other hand, expectiles are not easy to interpret: in regression analysis, it can be seen that the expectile $\mu_{\tau,i}$ determines at $X = x$ the point such that $100\tau\%$ of the mean distance between it and Y comes from the mass below it. This interpretation is due to Yee (2004) and does not have the same intuitive meaning of quantiles. Furthermore, expectiles result too sensible to outliers because they are averages; hence, they do not work as quantiles but they can be considered an efficient tool to approximate them.

If one aims to estimate an expectile smoothing model via P-spline, namely

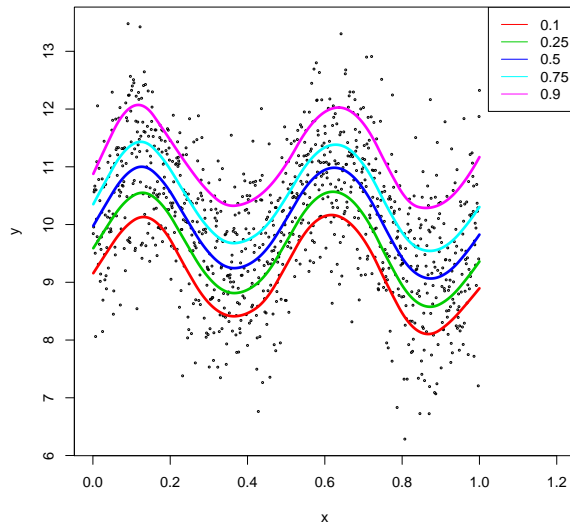$$\mu_\tau = s_\tau(x) = Ba, \tag{1.14}$$

Figure 1.6: Expectile smoothing with asymmetry τ = $\{0.1, 0.25, 0.5, 0.75, 0.9\}$. Implementation via R package *expectreg* (Sobotka *et al.*, 2011)

the estimation of coefficients is of the form

$$\hat{a} = \left(B^T W B + \lambda P\right)^{-1} B^T W y,$$

where W is a diagonal matrix with asymmetrical weights w_i. The estimation procedure consists of iterating two steps: in the first one, estimation of coefficients given an initial matrix $\tilde{W}$ is computed; after that, one can calculate $\tilde{W}$ according to the signs of the residuals. The plot in Figure 1.6 shows an expectile smoothing via P-splines for a sinusoidal signal plus a gaussian error. See Sobotka and Kneib (2012) for an application of the method and further details.

LMS

Another way to get quantile estimates is given in Cole and Green (1992). The technique consists of estimating quantiles using the Box-Cox λ-Power transformation. Hence, assuming that response variable y has mean μ, and that y^λ is normally distributed, Box-Cox Power transformation is

$$
z^* = \begin{cases} \frac{(y/\mu)^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log_e(y/\mu) & \lambda = 0 \end{cases} \tag{1.15}
$$

With this transformation $z^* \sim N(0, \sigma)$. Dividing z^* by its standard deviation σ, $z = z^*/\sigma$, on can get $z \sim N(0, 1)$:

$$
z = \begin{cases} \frac{(y/\mu)^\lambda - 1}{\lambda\sigma} & \lambda \neq 0 \\ \frac{\log_e(y/\mu)}{\sigma} & \lambda = 0. \end{cases} \tag{1.16}
$$

The three parameters in (1.16), μ, λ and σ are supposed to vary smoothly according to a covariate x and so

$$
\mu = M(x), \qquad \lambda = L(x), \qquad \sigma = S(x).
$$

Replacing the smooth functions in (1.16) and writing y as function of the covariate one can get the estimation of the quantile $q_\tau(Y|x)$ as

$$
q_\tau(x) = \begin{cases} M(x)(1 + L(x)S(x)z_\tau)^{1/L(x)} & L(t) \neq 0 \\ M(x)\exp[S(x)z_\tau] & L(t) = 0 \end{cases} \tag{1.17}
$$

The estimation of each smooth functions is achieved via penalized maximum likelihood, one can use either smoothing or P-spline (see plot in Figure 1.7 with P-spline as smoother). For this plot we used $L(t) = S(t) = k$
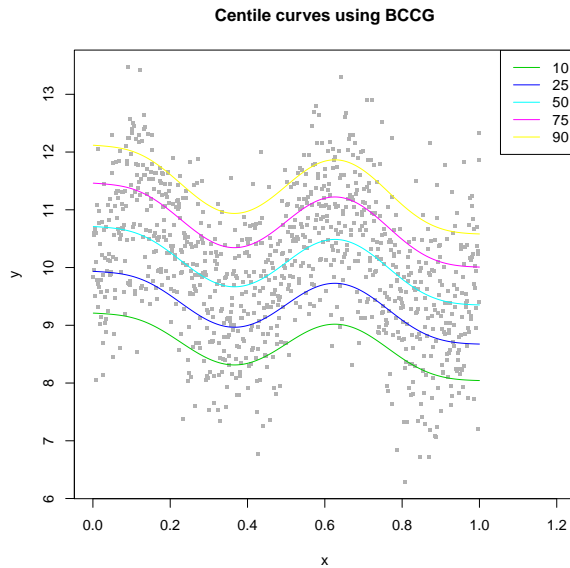
Figure 1.7: LMS fitting via GAMLSS function, for probability levels $\tau = \{0.1, 0.25, 0.5, 0.75, 0.9\}$

because a vector of i.i.d. gaussian errors was generated.

An implementation of the LMS method is provided within R package GAMLSS (Stasinopoulos and Rigby, 2007). GAMLSS allow to assume several distributions for the data in quantile modelling. For instance, in addition to Box-Cox Normal distribution, Box-Cox t_ν distribution or power exponential distribution can be used. A good advantage of using LMS methods is that the estimated quantile curves cannot cross each other. Furthermore, the results are easier to interpret than expectiles.

On the other hand, LMS models are based on the assumption of normality of (transformed) data which has to be checked while for QR it is not required.

Chapter 2

Smoothing parameter selection

This chapter includes the main contribution of the thesis. Smoothing parameter selection is a crucial point in any non parametric technique and setting up a "good" criterion is important for this purpose. A good method for smoothing parameter selection allows for a reliable fit of data and saves computational time. In non parametric QR framework, many approaches can be followed: Koenker *et al.* (1994) use a slightly modified version of the Schwarz Information Criterion (SIC); Oh *et al.* (2004) propose the so-called Robust Cross Validation (RCV) and Nychka *et al.* (1995) provide an approximation of this version (ACV) to reduce the computational burden. Then a generalized approach to ACV is given by Yuan (2006). Andriyana *et al.* (2014) select the smoothing parameter via L-curve (Frasso and Eilers, 2015), a function representing the trade-off between fidelity and penalty. Yet another interesting approach, given by Reiss and Huang (2012), provides a likelihood-based selection according to the link between penalized

and mixed models. However, all these criteria are based on grid search and so one has to compute the model for each value of the grid and then choosing the smoothing parameter, requiring an high number of computations to carry out. Further computational problems may arise when many covariates are involved in the analysis: the higher number of variables involved, the higher number of dimensions of the grid. At this aim, an iterative algorithm based on a mixed model approach due to Schall (1991) is presented. The algorithm is modified to allow for the estimation of smoothing parameter in a P-spline QR model. Schall algorithm is used to estimate mixed models (i.e in the context of L_2-norm problems) but we extend it to P-spline QR framework. The extension to the QR has never been discussed and guarantees a good fit of data without spending too much computational time.

This chapter is divided into two main parts: the first one concerns the description of the aforementioned grid search methods. Then, Schall algorithm for L_2-norm problems and its extension to P-spline QR model are introduced; in this section, the issues concerning the implementation of the algorithm and some theoretical aspects are also described. Furthermore, a method to determine the effective dimension of the model is provided.

2.1 Methods for smoothing parameter selection

2.1.1 The Schwarz information criterion

In the approach used by Koenker *et al.* (1994), Schwarz information criterion is proposed to find the optimal amount of penalty for a quantile smoothing spline model.

$$SIC_\lambda = \log\left[n^{-1}\sum_i^n \rho_\tau\{y_i - \hat{s}(x_i)\} + 0.5n^{-1}ED_\lambda \log(n)\right] \quad (2.1)$$

The optimal model will be the one with the lowest SIC. The formula of the criterion is slightly different from models based on L_2 norm. Since QR approach is distribution free, the check function replaces the common least squares objective of the model. Another relevant quantity is the measure of the effective dimension of the model ED_λ. The authors stated that in a quantile smoothing spline with objective function (1.6), the ED_λ can be considered as the number of points interpolated by the fitted curve: in other words, ED_λ is the number of zero residuals in the model.

2.1.2   Cross validation-based methods

Cross validation (CV) methods are based on the estimation of a model using a subset of the observations. A very popular technique for L_2-norm smoothing is the leave-one-out CV, namely $CV_\lambda = n^{-1}\sum_i^n [y_i - \hat{s}_{-i}(x_i, \lambda)]^2$, where $\hat{s}_{-i}(x_i, \lambda)$ is the estimation of the smooth function omitting the i-th point. Given a grid of λs, the optimal smoothing parameter is the one which minimize CV. CV is mainly used for L_2-norm problems; for L_1-norm smoothing, Oh *et al.* (2004) propose to use a robust version based on the check function.

$$RCV_\lambda = n^{-1}\sum_i^n \rho_\tau\left(y_i - \hat{s}_{-i}(x_i, \lambda)\right). \quad (2.2)$$

It is well-known that the computational load of the leave-one-out CV is very high. Basically, given a fixed smoothing parameter, one should estimate n different versions of $\hat{s}_{-i}(x_i, \lambda)$. The number of computation increases

dramatically along with the number of values included in the grid. This shortcoming can be reduced using the so-called Approximate CV (ACV) proposed by Nychka *et al.* (1995):

$$ACV_\lambda = n^{-1} \sum_i^n \rho_\tau \left(\frac{y_i - \hat{s}(x_i, \lambda)}{1 - h_{ii}} \right), \tag{2.3}$$

where $h_{ii} = \delta \hat{s}(x_i, \lambda)/\delta y_i$.

ACV allows to avoid the estimation of the $\hat{s}_{-i}(x_i, \lambda)$, reducing computational time: one has to compute one model (instead of n models) for each value of the grid. However, Yuan (2006) shows in his simulation that ACV does not perform well; so he suggested to compute the Generalized Approximate CV, GACV. Cross-validation was thought as the minimizer of the mean square error MSE_λ,

$$MSE_\lambda = n^{-1} \sum_i^n \left(s(x_i, \lambda) - \hat{s}_{-i}(x_i, \lambda) \right)^2.$$

However, in QR framework it is possible to use a different function to minimize, the Generalized Comparable Kullback-Leibler distance which depends on the check function,

$$GCKL_\lambda = n^{-1} \sum_i^n E_Z \left[\rho_\tau(y_i - \hat{s}(x_i, \lambda)) \right],$$

where $z_1, z_2, \ldots, z_n$ is a sample such that the conditional distribution of $z|X = x_i$ is the same of $y|X = x_i$. According to the author ACV is not a reliable estimate of GCKL. As an alternative he proposes

$$GACV_\lambda = n^{-1} \sum_i^n \rho_\tau \left(\frac{y_i - \hat{s}(x_i, \lambda)}{1 - tr(H)} \right) \tag{2.4}$$

where $tr(H)$ is the trace of the hat matrix, $tr(H) = \sum_i^n h_{ii}$.

2.1.3   L-curve

Another approach used by Andriyana *et al.* (2014) in P-spline with varying coefficient QR framework is based on the L-curve (Hansen, 1992). The smoothing parameter selection via L-curve for mean regression is computed as follows: given a grid of λs and the objective function in (1.9), the L-curve can be obtained plotting the fidelity term, $Fid_\lambda = \sum_{i=1}^n (y_i - \hat{\mu}_i)^2$ against the penalty term $Pen_\lambda = \sum_{j=1}^{J-d} (D^d a)_j^2$. As pictured in Figure 2.1 the optimal λ corresponds to the corner of the L-curve. This point has the minimum distance from the origin. Therefore one can search for the optimal λ using this approach.

The L-curve is a worthwhile method often used in Econometrics because it is robust in case of autocorrelation of the errors (Frasso and Eilers, 2015).

In P-spline QR framework, the computation of fidelity and penalty term is achieved via L_1-norm, according to the equation (1.12). Instead of computing directly the L-curve, the authors re-scale the fidelity and the penalty term. Then they select the optimal point using the minimum euclidean distance from the origin ; therefore the best λ will minimize

$$d(\lambda) = \sqrt{z_{Fid}^2(\lambda) + z_{Pen}^2(\lambda)}$$

where the scale fidelity and penalty are

$$z_{Fid}(\lambda) = \frac{Fid(\lambda) - \min Fid(\lambda)}{\max Fid(\lambda) - \min Fid(\lambda)}, \qquad z_{Pen}(\lambda) = \frac{Pen(\lambda) - \min Pen(\lambda)}{\max Pen(\lambda) - \min Pen(\lambda)}$$
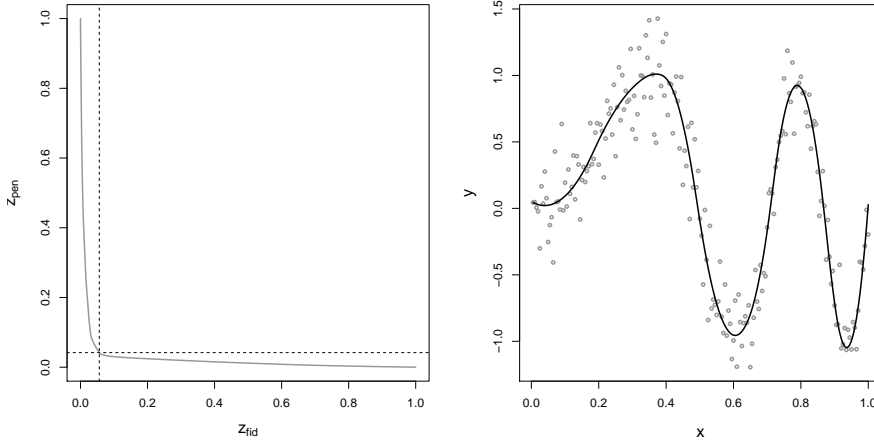
Figure 2.1: Plot of L-curve for simulated data. An optimal λ corresponds to the corner of the L-curve (left). The right panel shows a smooth fit.

2.1.4   Maximum likelihood from a mixed model

This method suggested by Reiss and Huang (2012), relies on the link between penalized and mixed model for P-spline QR with L_2-norm penalty (see Lee *et al.* (2006)). In other words, it is possible to re-parametrize a mixed model as function of λ. According to this link and referring to equation (1.10),

$$q_\tau(Y) = Ba = B\begin{pmatrix} Q_1 & Q_2 \end{pmatrix}\begin{pmatrix} Q_1^T & Q_2^T \end{pmatrix}^T a = X\beta + Zu \qquad (2.5)$$

where X is a matrix of covariates with fixed effects β, Z is a matrix of covariates with random effects $u \sim N(0, \psi)$; Q_1 and Q_2 are two matrices $K \times d$ and $K \times K - d$ (i.e. the number of penalized coefficients) such that $X = BQ_1, \beta = Q_1^T a, Z = BQ_2$ and $u = Q_2^T a$. Using this re-parametrization,

the L_2 penalty can be written as

$$a^T Pa = u^T Q_2^T P Q_2 u$$

The authors provide a mixed model representation of the (1.11) assuming a Laplace distribution for the error term and a normal distribution for the random effect term, namely $y_i|u \sim AL(x_i^T\beta + z_i^T u, \sigma, \tau)$ for $i = 1, \ldots, n$ where n is the number of subjects and $u \sim N(0, (\sigma/2\lambda)Q_2^T P^+ Q_2)$, with P^+ an generalized inverse of P. Standard theory about mixed models considers the joint distribution as $f(y_i, u) = f(y_i|u)f(u)$ and the maximization of the profile likelihood to get parameter estimates. The joint likelihood is

$$L(\beta, \sigma, \lambda) = \left[\frac{\tau(1-\tau)}{\sigma}\right]^n \frac{|(2\lambda/\sigma)Q_2^T P Q_2|^{1/2}}{2\pi^{(K-d)/2}} \times$$
$$\exp\left[-\frac{1}{\sigma}\left\{\sum_i^n \rho_\tau(y_i - b_i^T a) + \lambda a^T Pa\right\}\right]$$

The profile likelihood is not easy to compute because there is no closed form solution to get $\hat{\beta} = \arg\max L(\beta, \sigma, \lambda)$. So the authors use the parametric quantile regression estimate, $\tilde{\beta} = \arg\min \sum_i^n \rho_\tau(y_i - x_i^T\beta)$. The approximated profile likelihood is then $\hat{L}_p(\sigma, \lambda) = L(\tilde{\beta}, \sigma, \lambda)$. Also $\hat{L}_p(\sigma, \lambda)$ is not easy to compute because the integral to solve is intractable. To fix this issue, a simulated maximum likelihood is computed via Monte Carlo approximation. It is possible to estimate the nuisance parameter $\hat{\sigma}$ via numerical optimization. Finally, the λ selection can be done choosing the value of the grid which maximize the likelihood calculated in the points $\tilde{\beta}_\tau$ and $\hat{\sigma}$.

2.2 Our proposal: Schall algorithm

All the aforementioned methods are far from being efficient because they rely on criteria working on a pre-specified grid of smoothing parameters. In practise, one has to estimate the model for any value of the grid and select the final model according to the best value of the criterion. This means that there is the risk to choose a bad optimal value when the grid is not appropriate. For instance, if the grid is too sparse then there could be values out of the grid with a better performance than the values inside the grid. Furthermore, the computational burden becomes particularly expensive when the regression equation involves multiple additive components leading to a multidimensional grid of smoothing parameters. To overcome this shortcoming, the key idea is to use an iterative algorithm based Schall (1991). The estimation of variance components via Schall algorithm was used for smoothing parameter selection by Schnabel and Eilers (2009) within expectile smoothing framework and by Rodríguez-Álvarez *et al.* (2014) to get a smoothing parameter estimation in case of anisotropic penalty. It was never applied in QR framework.

In the next subsections, the application of the Schall algorithm for a L_2-norm smoothing problem is described; that is useful to show the differences respect to our algorithm for a P-spline QR model (based on $L_1 1$ norm) which will be discussed afterwards.

Schall algorithm in L_2 norm smoothing

Consider a random effect model of the form

$$y = X\beta + Zu + \epsilon \tag{2.6}$$

where $\epsilon \sim N(0, \Sigma)$ and $u \sim N(0, \Omega)$, X and Z are $m \times p$ and $m \times q$ for fixed and random part of the model, m is the total number of observations and n is the number of subjects.

- $cov(\epsilon) = \Sigma = \sigma_\epsilon^2 I_n$

- $cov(u) = \Omega = \psi^2 I_q$

- $cov(y) = \Sigma + Z\Omega Z^T$.

Schall algorithm in mixed models consists of iteratively computing the error variance σ_ϵ^2, the random effect variance ψ^2. Taking advantage of the link between penalized and mixed models, we can get $Ba = X\beta + Zu$ with Z such that $u = Da$ (D is the difference matrix). The smoothing parameter can be expressed as the ratio of the estimated variances

$$\hat{\lambda} = \frac{\hat{\sigma}_\epsilon^2}{\hat{\psi}^2} = \frac{\frac{\|y - Ba\|^2}{m - ED}}{\frac{\|Da\|^2}{ED}}$$

so one can iterate the computation of variances and λ until convergence.

2.2.1 Schall algorithm in P-spline QR

The procedure to select the smoothing parameter can be extended to QR framework. The algorithm is set up as follows:

1. Fix a (small) starting value for the smoothing parameter $\lambda^{(0)}$;

2. Fit the model minimising the objective

$$\sum_{i=1}^{n} \rho_\tau \left(y_i - s(x_i) \right) + \lambda \sum_{j=1}^{J-d} |D^d a|_j;$$

3. Compute the variances:

   $\hat{\sigma}_\epsilon^2$, and $\hat{\sigma}_b^2$

4. $\hat{\lambda} = \hat{\sigma}_\epsilon / \hat{\sigma}_b$;

5. Set $\hat{\lambda} \to \lambda^{(0)}$ and repeat steps 2. to 4. till convergence.

The algorithm presents some differences respect to the one used in L_2 norm smoothing. There are at least two issues to discuss: first, the estimation of variances has to be changed. Variance is a concept commonly based on mean and it is not simple to establish a measure of a quantile-based variance, especially when τ is close to 0 or 1. Second: unlike mixed models wherein an error belonging to the exponential class and Gaussian random effects is assumed, in QR framework the approach is distribution-free. This implies the need to choose a reliable distribution either for error and random effect to estimate the variance components. Furthermore, while the link between penalized and mixed models was proved for mean regression, there is no formal proof in QR framework. Since the approach is mainly empirical, this topic is not discussed in this thesis and represents a future challenging work.

A reasonable approach to fix the first issue is to include in the algorithm the ratio of standard deviations instead of variances which seems to be more sounded for L_1 framework. Moreover, the effective dimension of the model ED is computed via the trace of the hat matrix derived from a parametric smooth approximation of the objective (1.12). This procedure allows to partition the total model ED among the specified multiple smooth terms and then to evaluate the complexity of *each* variable in the model. Thus the term-specific variances and smoothing parameters can be obtained according to the ratios of step 3 and step 4 respectively. This makes the algorithm

very attractive in the multidimensional case where the multidimensional grid search gets substantially unfeasible.

The issue concerning the choice of the distribution for the error and the random effects and the estimation of the corresponding standard deviations is discussed in the next section.

2.2.2   Estimation of variance components

Error variance

Several methods can be followed to estimate standard deviations. From an empirical point of view, it is possible to estimate the error variance according to the asymmetrically weighted squared residuals (Schnabel and Eilers, 2009),

$$\hat{\sigma}^2_\epsilon = \frac{(y - \hat{q}_\tau)^T W (y - \hat{q}_\tau)}{m - ED}.$$

A second alternative to estimate σ_ϵ relies on the use of the check function, namely

$$\hat{\sigma}_\epsilon = \frac{\sum_i^n \rho_\tau (y_i - \hat{q}_{\tau,i})}{m - ED} \tag{2.7}$$

However, other approaches can be adopted. One of them relies on the Asymmetric Laplace (ALD) which is a known distribution to model quantiles, see Geraci and Bottai (2007) and Geraci and Bottai (2014) for further details. ALD can be described as a three-parameter (μ, σ, τ) distribution (Yu and Zhang, 2005) of the form

$$f(y|\mu, \sigma, \tau) = \frac{\tau(1 - \tau)}{\sigma} exp\left[-\rho_\tau\left(\frac{y - \mu}{\sigma}\right)\right] \tag{2.8}$$

where $0 < \tau < 1$, $\sigma > 0$ is a scale parameter and $-\infty < \mu < +\infty$ is a location parameter. For $\tau = 0.5$ one can get the symmetric Laplace or double exponential distribution. It is worthwhile to underline that many authors (Koltz *et al.*, 2001; Inusah and Kozubowski, 2006) present the symmetric Laplace distribution as

$$f(x|\mu, \phi) = \frac{1}{2\phi} exp\left[-\frac{|x - \mu|}{\phi}\right].$$

(2.9)

In our context the symmetric Laplace has a different parametrization of the scale parameter, namely $\phi = 2\sigma$.
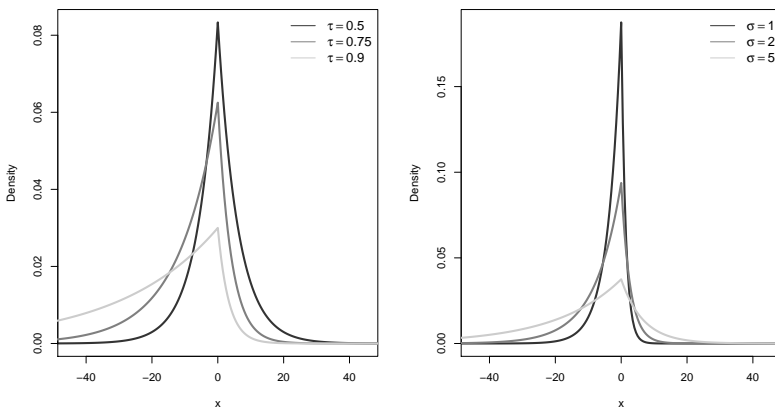

Figure 2.2: Density of Asymmetric Laplace distribution. it is possible to see how the density varies according to different τs (left panel) and σs (right panel, $\tau = 0.75$).

Looking at the plot in figure (2.2), it is easy to note that the distribution is not differentiable in the expected value (in this example $\mu = 0$). The scale parameter affects the variability of the variable: an increase of σ leads to an increase of $\mathbb{V}[y]$.

In QR context, it is possible to consider τ as the probability level and μ as the quantile of the conditional distribution and the expression in (2.10) as estimation of the error variance. According to (2.8), it is possible to see that the variance of the response is

$$\mathbb{V}[y] = \frac{\hat{\sigma}^2(1 - 2\tau + 2\tau^2)}{(1 - \tau)^2\tau^2}, \tag{2.10}$$

and that a ML estimation of the scale parameter is given by

$$\hat{\sigma} = n^{-1} \sum_i^n \rho_\tau(y_i - \mu_{\tau,i}).$$

Random effect variance: Laplace distribution

The argument of the exponential in (2.8) remind us the normal distribution but with an L_1 norm objective: it seems that the Laplace distribution plays for L_1 norm problems a similar role of the normal distribution for L_2 norm problems. This fact probably leads Geraci and Bottai (2007) to propose an estimation of the random effects variance using the symmetric Laplace distribution, $u_i \sim ALD(\mu_i = 0, \psi, \tau = 0.5)$. The join density of (y_i, u_i) is given by the product of the density for the i-th subject conditional on the random intercept u_i and the density of the random effects, namely

$$f(y_i, u_i) = f(u_i|\psi) \prod_j^{q_i} f(y_{ij}|u_i, \sigma_\epsilon)$$

$$= \left\{\frac{\tau(1 - \tau)}{\sigma_\epsilon}\right\}^{m_i} \frac{1}{4\psi} \exp\left[-\sum_{j=1}^{m_i}\left\{\rho_\tau\left(\frac{y_{ij} - q_{\tau,ij}}{\sigma_\epsilon}\right)\right\} - \frac{1}{2\psi}|u_i|\right] \tag{2.11}$$

Setting $\tau = 0.5$ and $\lambda = \frac{\sigma_\epsilon}{\psi}$, the above expression becomes

$$f(y_i, u_i) = \frac{1}{(4\psi)^{m_i+1} \lambda^{m_i}} \exp\left[-\frac{1}{2\sigma_\epsilon}\left\{\sum_j^{m_i}(|y_{ij} - \mu_{ij}|) + \lambda|u_i|\right\}\right].$$

The argument of the exponential has a similar form as the penalized quantile regression used by Koenker (2004). Therefore the joint distribution seems to be quite related to a penalized model.

The log-likelihood for the i-th subject related to (2.11) can be expressed as

$$l_i(\sigma_\epsilon; \mu_{ij}, \psi, \tau) \propto -m_i \log \sigma_\epsilon - \log \psi - \sum_{j=1}^{m_i}\left\{\rho_\tau\left(\frac{y_{ij} - q_{\tau,ij}}{\sigma_\epsilon}\right)\right\} - \frac{1}{2\psi}|u_i|$$

and the partial derivative respect to σ_ϵ is

$$\frac{\partial l_i(\sigma_\epsilon; \mu_{ij}, \tau)}{\partial \sigma_\epsilon} = -\frac{m_i}{\sigma_\epsilon} + \frac{1}{\sigma_\epsilon^2}\sum_{j=1}^{m_i}\left\{\rho_\tau\left(y_{ij} - q_{ij}\right)\right\}$$

then

$$\hat{\sigma}_\epsilon = \frac{\sum_i^n \sum_j^{m_i} \rho_\tau\left(y_{ij} - \hat{q}_{ij}\right)}{\sum_i^n m_i}. \tag{2.12}$$

For the random effect variance ψ,

$$\frac{\partial l_i(\psi; u_i)}{\partial \psi} = -\frac{1}{2\psi} - \frac{|u_i|}{2\psi}$$

then

$$\hat{\psi} = \frac{\sum_i^n |\hat{u}_i|}{n} \tag{2.13}$$

and

$$\hat{\lambda} = \frac{\hat{\sigma}_\epsilon}{\hat{\psi}} = \frac{\frac{\sum_i^n \sum_j^{m_i} \rho_\tau(y_{ij} - \hat{q}_{ij})}{\sum_i^n m_i}}{\frac{\sum_i^n |\hat{u}_i|}{n}}. \qquad (2.14)$$

Therefore, assuming that the distribution of the error is Asymmetric Laplace and the random effects distribution is symmetric Laplace, the smoothing parameter computed in the algorithm is approximately a Maximum Likelihood Estimator.

In ML framework, the dimension of a model is roughly determined by the number of the observations; hence the denominator of the error variance will be the total number of observations $m = \sum_i^n m_i$ and the denominator of the random effects will be the number of subjects n. It is known that ML variance estimations are biased because they ignore the complexity of a model. An intuitive way to fix this problem is to compute the Restricted ML estimates (REML) depending on the ED of the model. This means to replace the denominators of (2.12) and (2.13) with respectively, $n - ED$ for the error variance and ED for the random effect variance.

Random effect variance: normal distribution

Geraci and Bottai (2007) have also introduced a linear quantile mixed model with ALD error and normal random effects.

Assuming that the random effects are independent and identically normally distributed $u_i \sim N(0, \psi^2)$, the joint density for (y_i, u_i) for the i-th subject is

$$f(y_i, u_i) = f(u_i|\psi^2) \prod_{j}^{q_i} f(y_{ij}|u_i, \sigma_\epsilon)$$

$$= \left\{\frac{\tau(1-\tau)}{\sigma_\epsilon}\right\}^{m_i} \frac{1}{\sqrt{2\pi\psi^2}} \exp\left[-\sum_{j=1}^{m_i}\left\{\rho_\tau\left(\frac{y_{ij}-q_{\tau,ij}}{\sigma_\epsilon}\right)\right\} - \frac{1}{2\psi^2}u_i^2\right]$$

$$(2.15)$$

Now, setting $\tau = 0.5$ and $\lambda = \frac{\sigma_\epsilon}{\psi^2}$, the density becomes

$$\left(\frac{\tau(1-\tau)}{\lambda\psi^2}\right)\frac{1}{\sqrt{2\pi\psi^2}} \exp\left[-\frac{1}{\sigma_\epsilon}\left\{\sum_{j}^{m_i}|y_{ij}-\mu_{ij}| - \frac{\lambda}{2}u_i^2\right\}\right]$$

The argument of the exponential has a form similar to the objective function of a QR model with a L_2-norm penalty, which is out of our framework. It is also worthwhile to note that in this situation λ does not represent a pure number because the numerator and the denominator belong to different scales (numerator in linear scale, denominator in quadratic scale). The individual contributes to the part of log-likelihood related to (2.15) which depends on σ_ϵ and ψ^2 are

$$l_i(\sigma_\epsilon; \mu_{ij}, \tau) \propto -m_i \log \sigma_\epsilon - \frac{1}{2}\log\psi^2 - \sum_{j=1}^{m_i}\left\{\rho_\tau\left(\frac{y_{ij}-q_{\tau,ij}}{\sigma_\epsilon}\right)\right\} - \frac{1}{2\psi^2}u_i^2$$

and the correspondent partial derivatives are

$$\frac{\partial l_i(\sigma_\epsilon; \mu_{ij}, \tau)}{\partial \sigma_\epsilon} = -\frac{m_i}{\sigma_\epsilon} + \frac{1}{\sigma_\epsilon^2}\sum_{j=1}^{m_i}\left\{\rho_\tau\left(y_{ij}-q_{ij}\right)\right\}$$

then

$$\hat{\sigma}_\epsilon = \frac{\sum_i^n \sum_j^{m_i} \rho_\tau \left(y_{ij} - \hat{q}_{ij}\right)}{\sum_i^n m_i}.$$ (2.16)

The partial derivative for ψ^2 is

$$\frac{\partial l_i(\psi^2; u_i)}{\partial(\psi^2)} = -\frac{1}{2\psi^2} + \frac{u_i^2}{2(\psi^2)^2}$$

then

$$\hat{\psi}^2 = \frac{\sum_i^n \hat{u}_i^2}{n}$$ (2.17)

Therefore, the estimation of the smoothing parameter is of the form

$$\hat{\lambda} = \frac{\hat{\sigma}_\epsilon}{\hat{\psi}^2} = \frac{\frac{\sum_i^n \sum_j^{m_i} \rho_\tau(y_{ij} - \hat{q}_{ij})}{\sum_i^n m_i}}{\frac{\sum_i^n u_i^2}{n}}.$$ (2.18)

Approximation of the ED

The evaluation of ED is commonly based on the number of zero residuals. Here a new alternative based on an approximation of the hat matrix for a quantile regression model is presented. To achieve an estimate of the ED, the computation of the hat matrix is fundamental: a shortcoming is that for L_1 norm problems it is not possible to get it. A nice approach to fix this issue consists of estimating a quantile regression model using iterative weighted least squares (IWLS) which is based on the L_2 norm so that it is quite easy to obtain the hat matrix.

The approximation is described in the work of Muggeo *et al.* (2012) and it

is valid either for parametric and non-parametric quantile regression models. The idea is to smooth the objective function (1.1) using a parametric approximation.

Instead of use the objective in (1.1) one can use the following function:

$$
\begin{cases}
(\tau - 1)e_i & e_i \leq -c\tau \\
\frac{(1-\tau)e_i^2}{2c\tau} + \frac{c\tau(1-\tau)}{2} & -c\tau < e_i \leq 0 \\
\frac{\tau e_i^2}{2c(1-\tau)} + \frac{c\tau(1-\tau)}{2} & 0 < e_i < c(1-\tau) \\
\tau e_i & e_i \geq c(1-\tau)
\end{cases}
\tag{2.19}
$$

where c is a parameter which regulates the smoothness of the approximation. It is worthwhile to note that the approximation does not make the function change for values far from zero: the objective is not approximated for $e_i \leq -c\tau$ and $e_i \geq -c(1 - \tau)$. On the other hand, the smooth part of the approximation is close to zero. This is quite logic because the check function has a kink in correspondence of zero which is the unique not differentiable point.

The use of this new objective function allows for iterative weighted least squares estimation via path-following algorithm described by the authors. The iterative step for a P-spline QR model estimation is of the form

$$
\hat{\beta} = (B^T W(c)B + P)^{-1} B^T z(c)
$$

where B is a $(n \times J)$ B-spline matrix, $z(c)$ and $W(c)$ are, respectively, a working response and a diagonal matrix of weights (see Muggeo *et al.* (2012) for further details). P is the penalty matrix which depends the matrix of d-order differences D_d and on the smoothing parameter λ. Hence the hat matrix has the form

$$H = X(X^T W(c)X + P)^{-1} X^T W(c), \tag{2.20}$$

The smoothing parameter c determines the range of approximation: when c is small (big), the approximation will concern a small (big) portion of the domain of e_i. However when c is very small the estimation is more difficult because the gradient tends to be a step function more than for high values of c. This value can be chosen adaptively by taking, at each step of the estimation algorithm, 'the absolute value of the quantile corresponding to the percentile of the positive (negative) current residuals when $\tau \geq 0.5$ ($\tau \leq 0.5$) '.

The IWLS has a computational advantage respect to linear programming estimates especially when the number of covariates is quite large. On the other hand, the authors state that IWLS algorithm yields less efficient and more biased quantile estimates than the ones achieved from a linear program.

Our proposal is to use the approximation (2.19) of the objective and calculate the hat matrix and its trace without computing directly IWLS estimates. It is possible to do that by fitting the standard QR model, fixing the smoothing parameter c, computing the asymmetrical residuals $\tau \hat{e}_i$ for $\hat{e}_i > 0$ and $(\tau - 1)\hat{e}_i$ for $\hat{e}_i < 0$ and use them to calculate the diagonal matrix of weights $W(c)$. In this way one can get the hat matrix H according to (2.20) and its trace.

Since the trace is invariant under cyclic permutations,

$$tr(H) = tr\{(X^T W(c)X + P)^{-1} X^T W(c)X\}. \tag{2.21}$$

This form is more convenient because the dimension of H is no longer $(n \times n)$ but it becomes $(J \times J)$. This allows for an easy estimation of ED for single smooth term. It will suffice to sum over the elements of the diagonal of H which correspond to the number of columns used to create the spline basis for the single smooth term.

Chapter 3

Simulations

In this chapter results derived from the simulation study to compare the performances of different methods for smoothing parameter selection are reported. The results are compared in terms of Root Mean Integrated Square Error (RMISE), Root Mean Integrated Absolute Error (RMIAE) and Effective Degrees of Freedom (EDF) of the model. That is useful to analyse the performances of the methods looking either the at fidelity of the estimates (via RMISE and RMIAE) and at the complexity of the fitted model (via EDF). The best method has the highest fidelity and the lowest complexity.

3.1   Comparison among methods

We compared several smoothing parameter estimation methods. All versions of the iterative algorithm are summarized in Table 3.1.

1. *rqss*: it is the function used in the R package *quantreg* (Koenker, 2013). It computes smoothing spline quantile regression using a total variation penalty. The residual EDF are computed as the number of

non-zero residuals in the model. Smoothing parameter selection is done via SIC.

2. *laplace*: this method (and also the next ones) is referred to *gcrq* function in the R package *QuantregGrowth* (Muggeo, 2014). The computation of the smoothing parameter is performed via Schall-like algorithm using ALD error variance and normal random effects variance. The computation of the degrees of freedom is achieved via trace of the approximated hat matrix in (2.21).

3. *scale*: computation of the smoothing parameter as ratio of scale parameters of ALD for the error component and symmetric Laplace distribution for the random effect component. The computation of the degrees of freedom is achieved via trace of the approximated hat matrix (2.21).

4. *varscale*: computation of smoothing parameter via Schall-like algorithm using ALD error variance as numerator and the scale parameter of the symmetric Laplace distribution as denominator.

5. *variances*: computation of smoothing parameter via Schall-like algorithm using the variance estimation of the asymmetric Laplace distribution for the error variance and the symmetric Laplace variance for the random effect variance.

Method	Numerator	Denominator
laplace	$\sqrt{\dfrac{\hat{\sigma}^2(1-2\tau+2\tau^2)}{(1-\tau)^2\tau^2}}$	$\sqrt{\dfrac{\|Da\|^2}{ED}}$
scale	$\dfrac{\sum_i^n \sum_j^{m_i} \rho_\tau(y_{ij}-\hat{q}_{ij})}{\sum_i^n m_i - ED}$	$\dfrac{\sum_i^n |D\hat{a}|_i}{ED}$
varscale	$\sqrt{\dfrac{\hat{\sigma}^2(1-2\tau+2\tau^2)}{(1-\tau)^2\tau^2}}$	$\dfrac{\sum_i^n |D\hat{a}|_i}{ED}$
variances	$\sqrt{\dfrac{\hat{\sigma}^2(1-2\tau+2\tau^2)}{(1-\tau)^2\tau^2}}$	$\sqrt{8\hat{\psi}^2}$

Table 3.1: Set of possible computations of the iterative algorithm for λ selection

3.2 Simulation Plan

We run simulations of the scenarios derived from the following plan:

1. $B = 100$ number of replicates for each scenario;

2. sample size: $n \in \{100, 400\}$;

3. model: $y = f(x) + sig(x)\epsilon$;

4. signal: $f(x) \in \{0.2+0.4x, \log(x), \sin(2\pi x), g_0(x) = \sqrt{x(1-x)}\sin\left(\frac{2\pi(1+2^{-7/5})}{x+2^{-7/5}}\right)\}$

5. distribution: $\epsilon \in \{N(0,1), t_1, t_3, \chi_3^2 - 3\}$

6. scale: $sig(x) = \{0.2, 0.2(1+x)\}$

7. percentile: $\tau = \{0.5, 0.75, 0.9\}$

8. Difference order of the penalty matrix: $d \in \{2, 3\}$
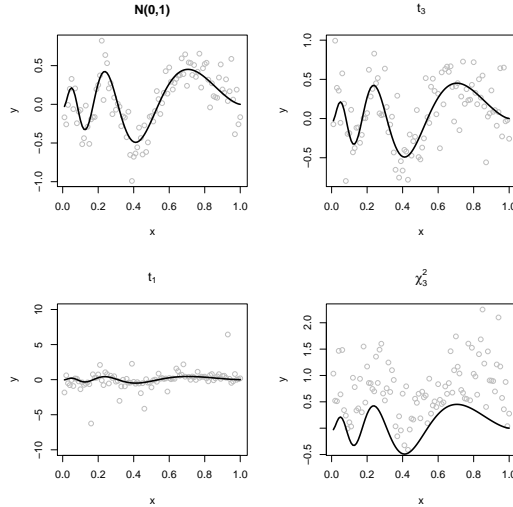
Figure 3.1: Data generated from 4 different distributions given the signal $g_0(x)$.

For each scenario we run simulations and estimate quantile regression for the given percentiles. To show the results we use boxplots showing RMISE, RMIAE, namely

$$RMISE = \sqrt{n^{-1} \sum_i^n \left(\hat{f}(x_i) - f(x_i) \right)^2},$$

$$RMIAE = \sqrt{n^{-1} \sum_i^n \left| \hat{f}(x_i) - f(x_i) \right|},$$

and effective dimension of the model (or Equivalent Degrees of Freedom, EDF) for the different estimation methods. Figure 3.1 describes how the choice of the distribution for the generator data process affects data in $g_0(x)$.

Looking at the y-scale of each panel, one can notice that Gaussian and t_3 errors have lower variability than t_3; furthermore, since $\chi_3^2 - 3$ distribution is positive skewed, all observation lie above the signal.

We include many $f(x)$s to see the behaviour of the fitted curves for different levels of linearity of the signal. The choice of 2 levels of heteroskedasticity is included to know whether the method is sensible to the homoskedasticity assumption and we aim to analyse all the scenarios for three quantiles because results could be worst when an extreme quantile is considered. Since the smoothing spline quantile regression results in a piecewise linear fit, the comparison when $d = 2$ is done only for the linear signal.

3.3 Results

Results concerning the whole simulation study are shown in the Appendix of the thesis. For the sake of simplicity we report the results relative to RMIAE and consider here *rqss* and *laplace* methods. The latter and *varscale* are considered the most performing versions of the iterative algorithm; since the performances of *varscale* and *laplace* are quite similar, we include in this part just the method relying on the gaussian random effect distribution.

3.3.1 Simulation with $d = 2$

Results are shown in Figures 3.2 and 3.3. For all the analysed scenarios, differences in terms of goodness of fit (i.e. RMIAE) are rather negligible. The boxplots referred to the degrees of freedom show that *laplace* method provides less complex models than *rqss*, which is a quite desirable characteristic occurring regardless of the scenario. Due to the different esti-

mation methods (smoothing spline has a major computational burden than P-spline) and to the iterative nature of the Schall-like algorithm, *laplace* can be considered preferable respect to *rqss*.

3.3.2   Simulation with $d = 3$

Results concerning RMIAE for $d = 3$ are reported in Figures 3.4, 3.5, 3.6 and 3.7.

There are not big differences between the methods considering a linear relationship. When the signal is $g_0(x)$, *laplace* works better than *rqss* for gaussian and t_3-distributed data; these differences decrease for extreme percentiles. Given a sinusoidal and logarithmic signal the differences between the methods are almost negligible except for gaussian and t_3-distributed data where *laplace* works slightly better than *rqss* especially in the IID case.

Results concerning the effective dimension of the model are provided in Figures 3.8, 3.9, 3.10 and 3.11. In general, models derived from *rqss* function are more complex than models achieved via *laplace*. Furthermore, the variability of the EDF for the latter models is very low. That is quite easy to see in almost all scenarios of the simulation (as counterpart see some scenarios of Figure 3.10 which is referred to a logarithmic signal).
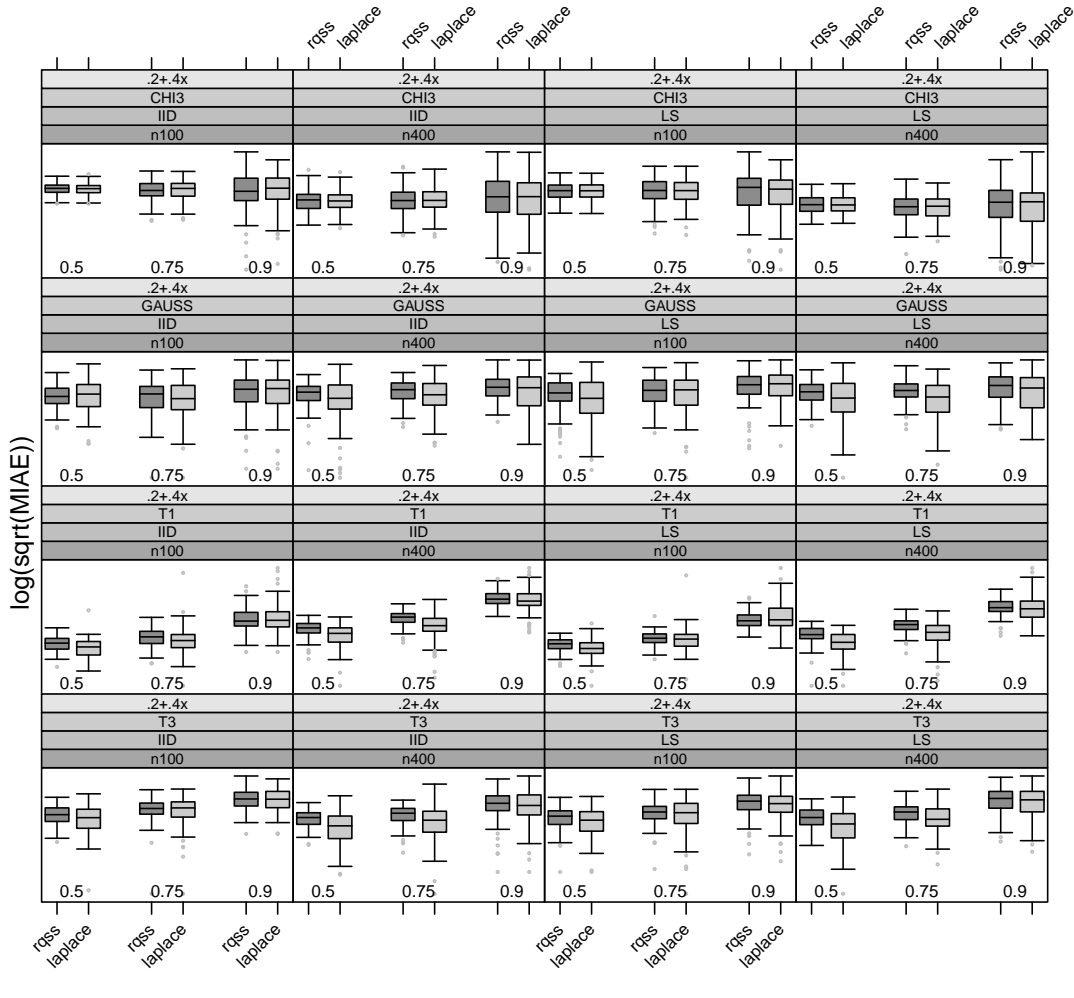
Figure 3.2: Root Mean Integrated Absolute Error to compare *rqss* versus *laplace* methods with $d = 2$
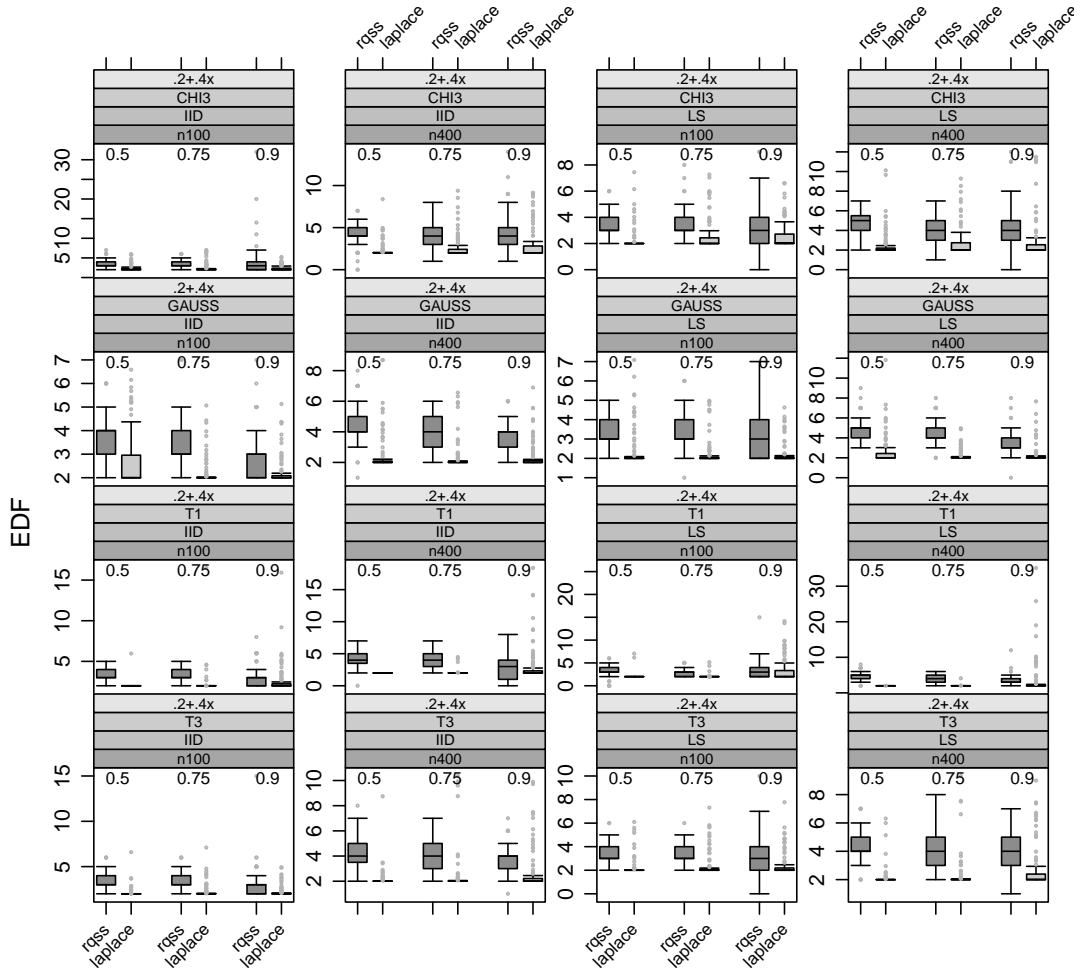
Figure 3.3: Effective Degrees of Freedom to compare *rqss* versus *laplace* methods with $d = 2$
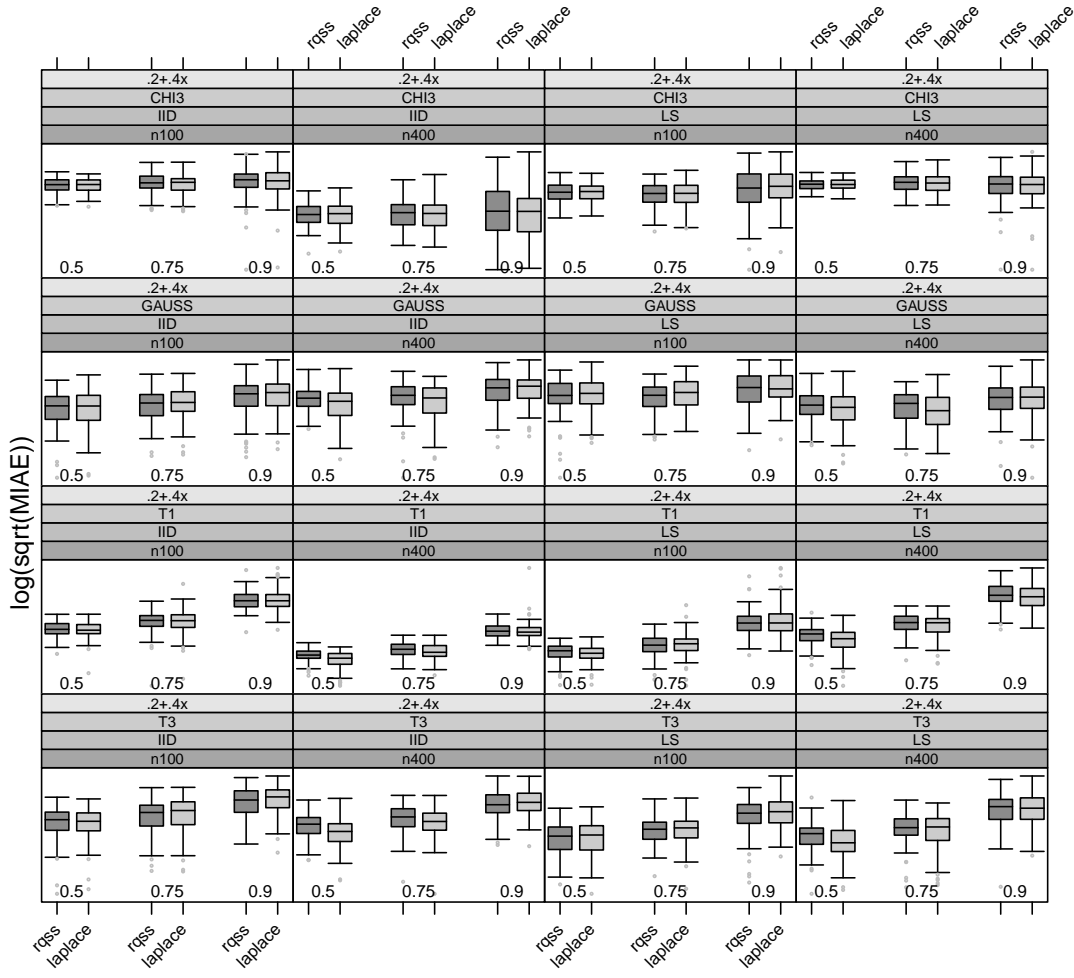
Figure 3.4: Root Mean Integrated Absolute Error to compare *rqss* versus *laplace* methods (part1) for $d = 3$. Linear signal
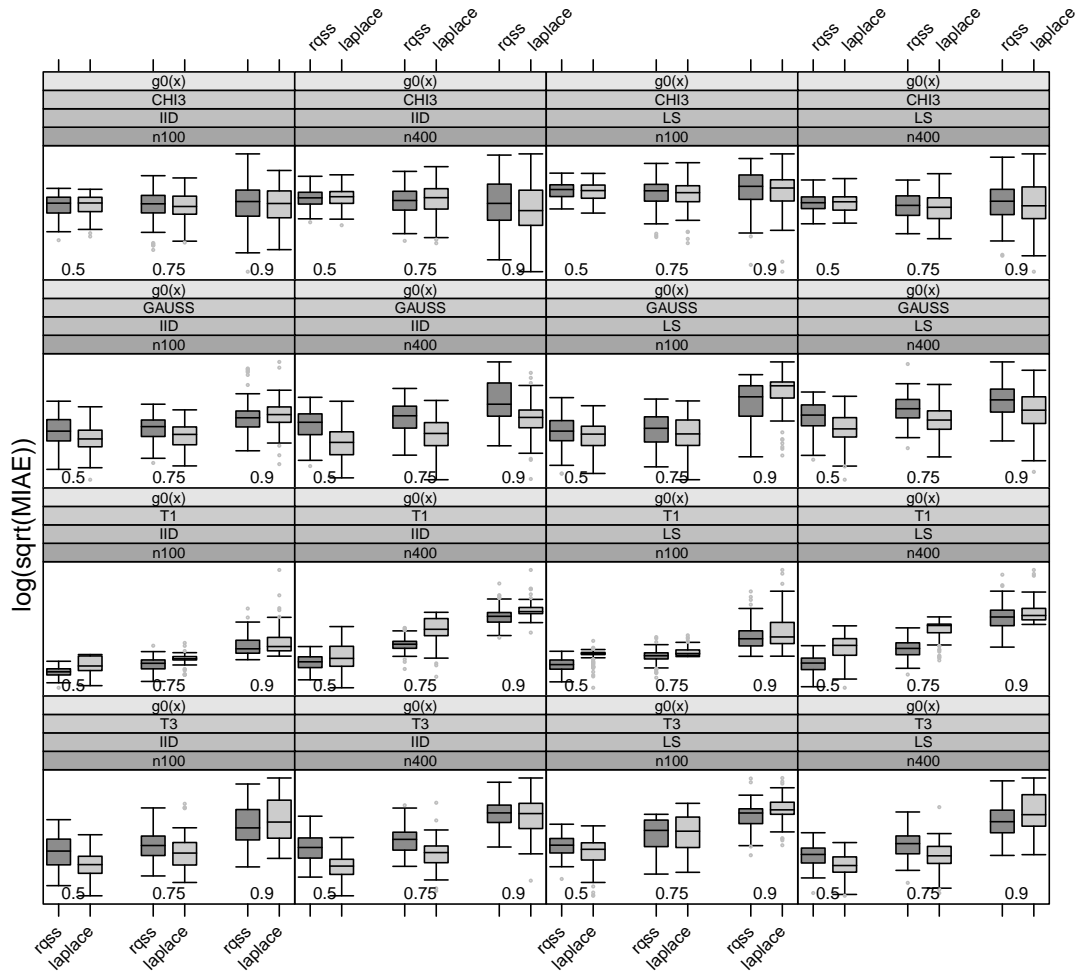
Figure 3.5: Root Mean Integrated Absolute Error to compare *rqss* versus *laplace* methods (part2) for *d* = 3. $g_0(x)$ signal
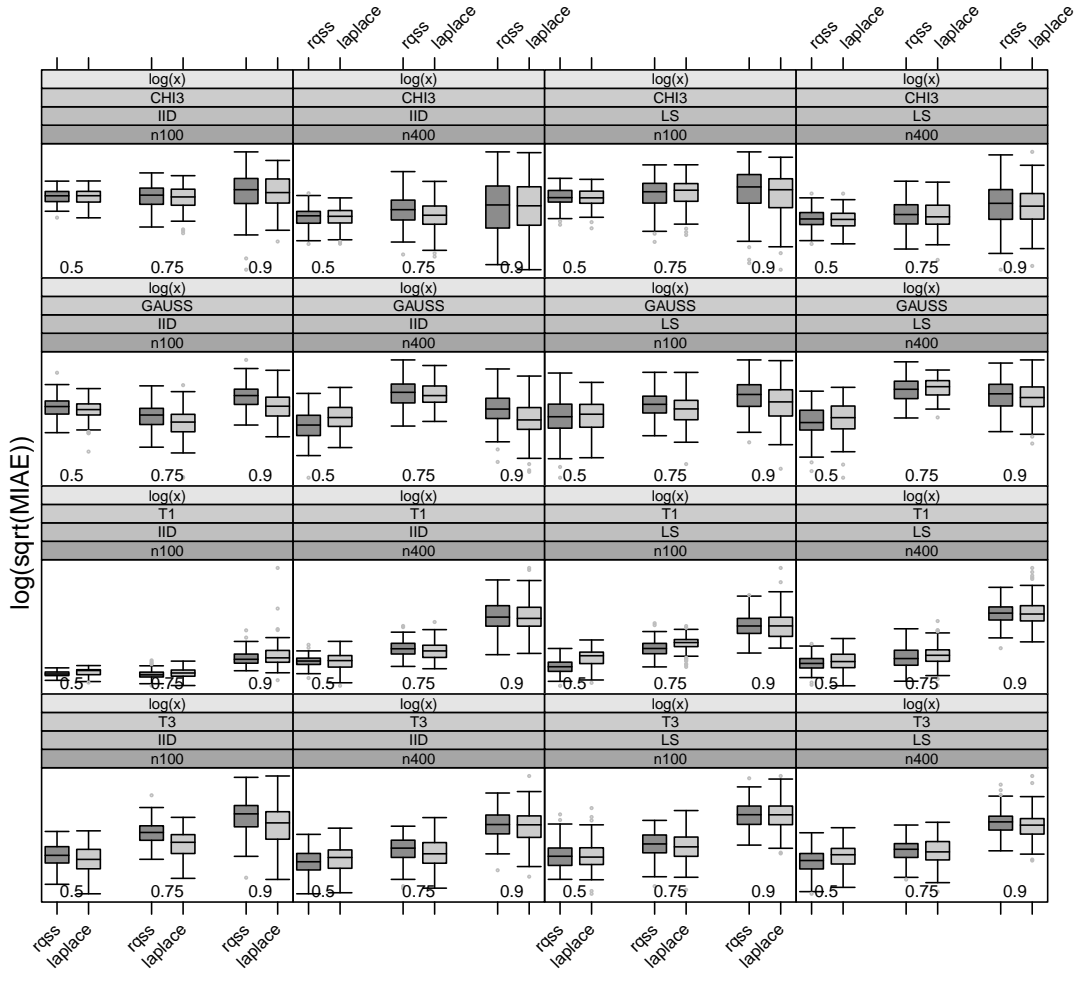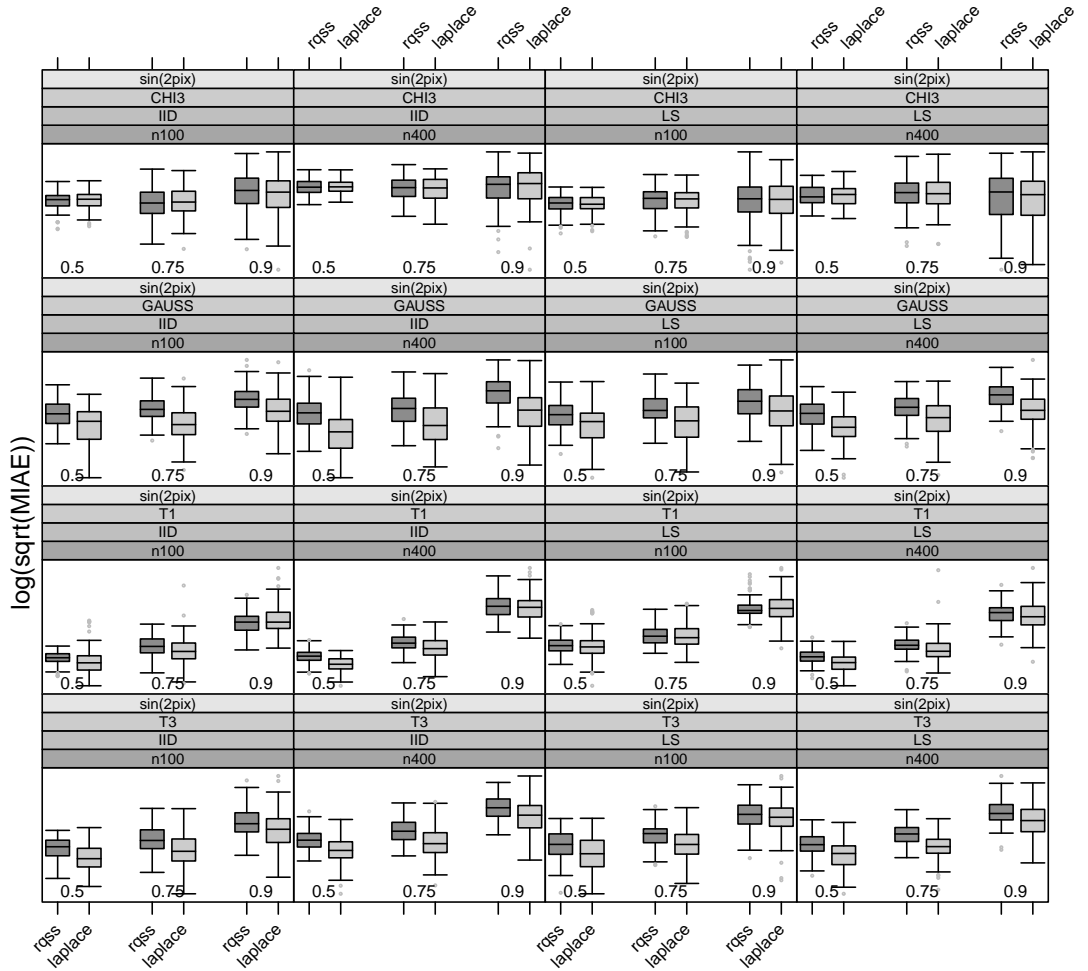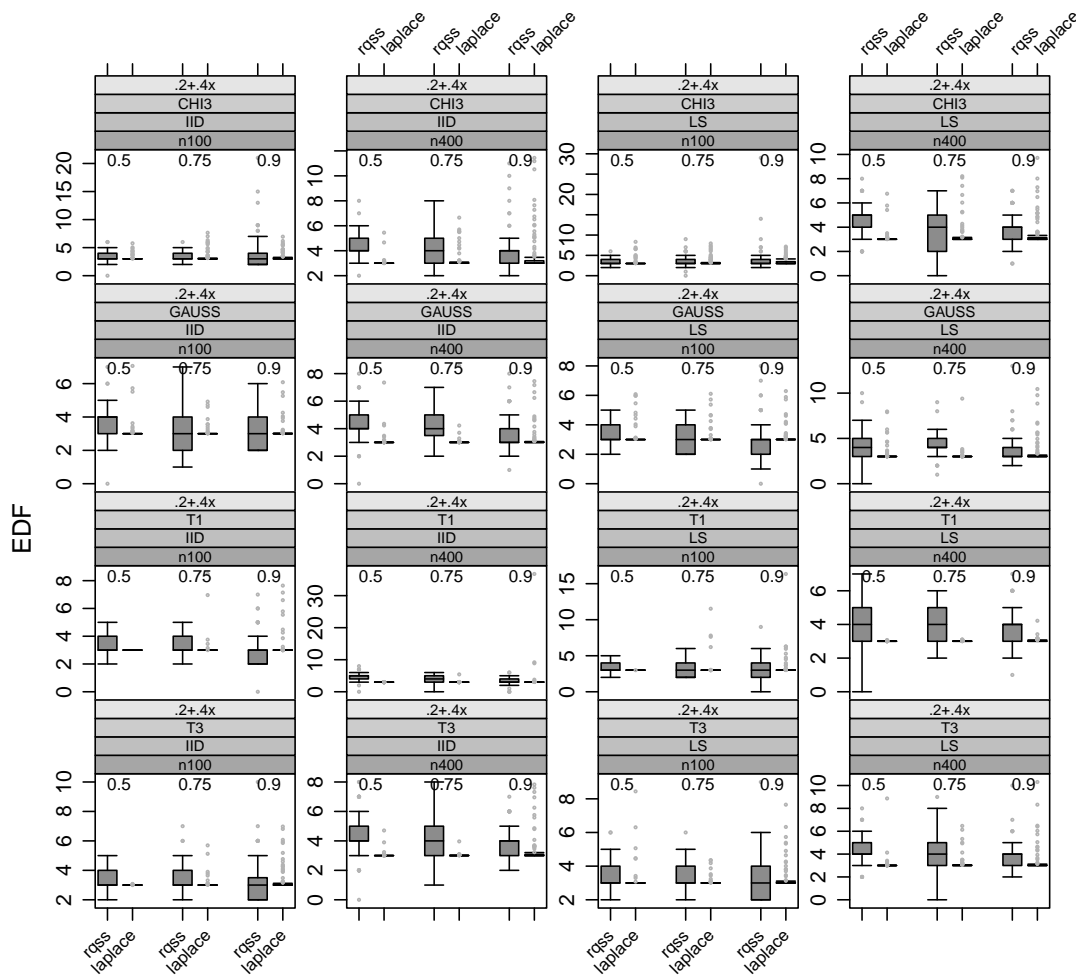
Figure 3.6: Root Mean Integrated Absolute Error to compare *rqss* versus *laplace* methods (part3) for *d* = 3. log(*x*) signal

Figure 3.7: Root Mean Integrated Absolute Error to compare *rqss* versus *laplace* methods (part4) for *d* = 3. sin(2π*x*) signal
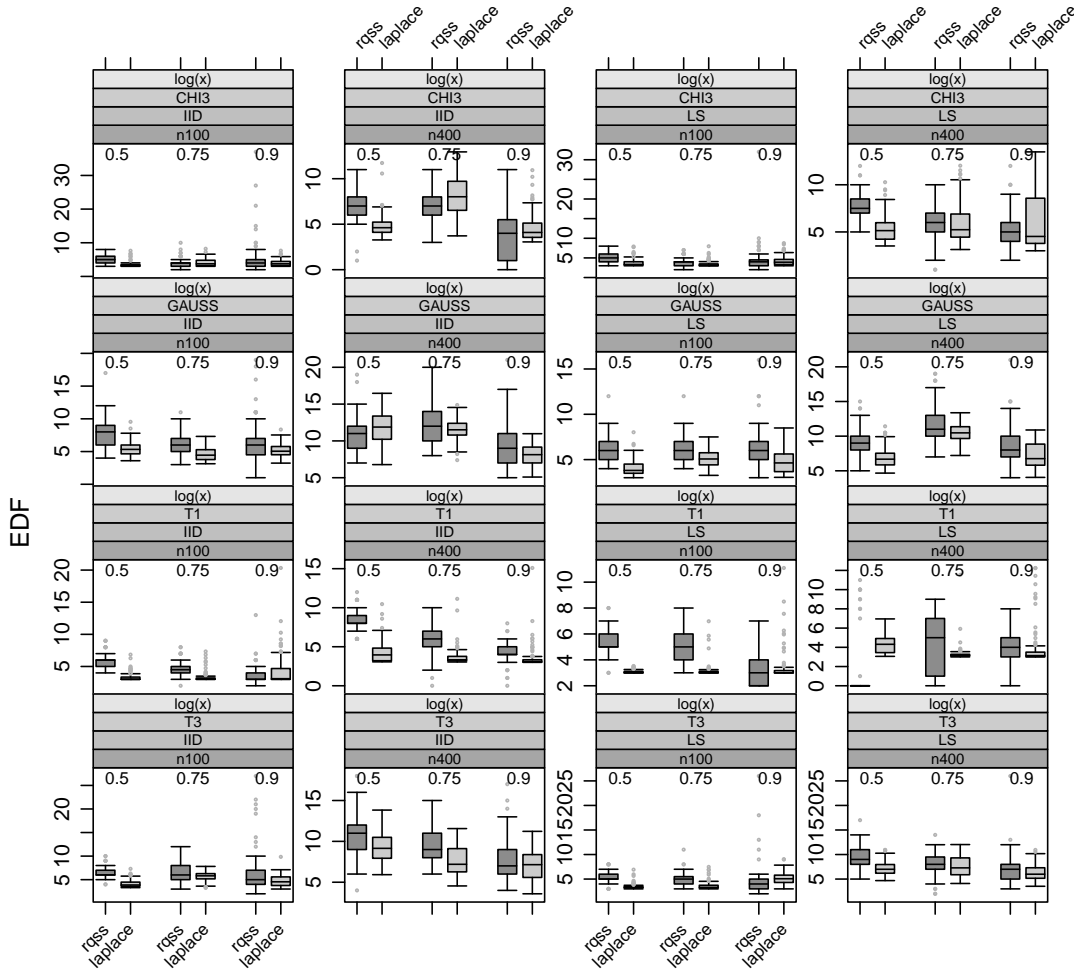
Figure 3.8: Effective Degrees of Freedom to compare *rqss* versus *laplace* methods (part1) for *d* = 3. Linear signal
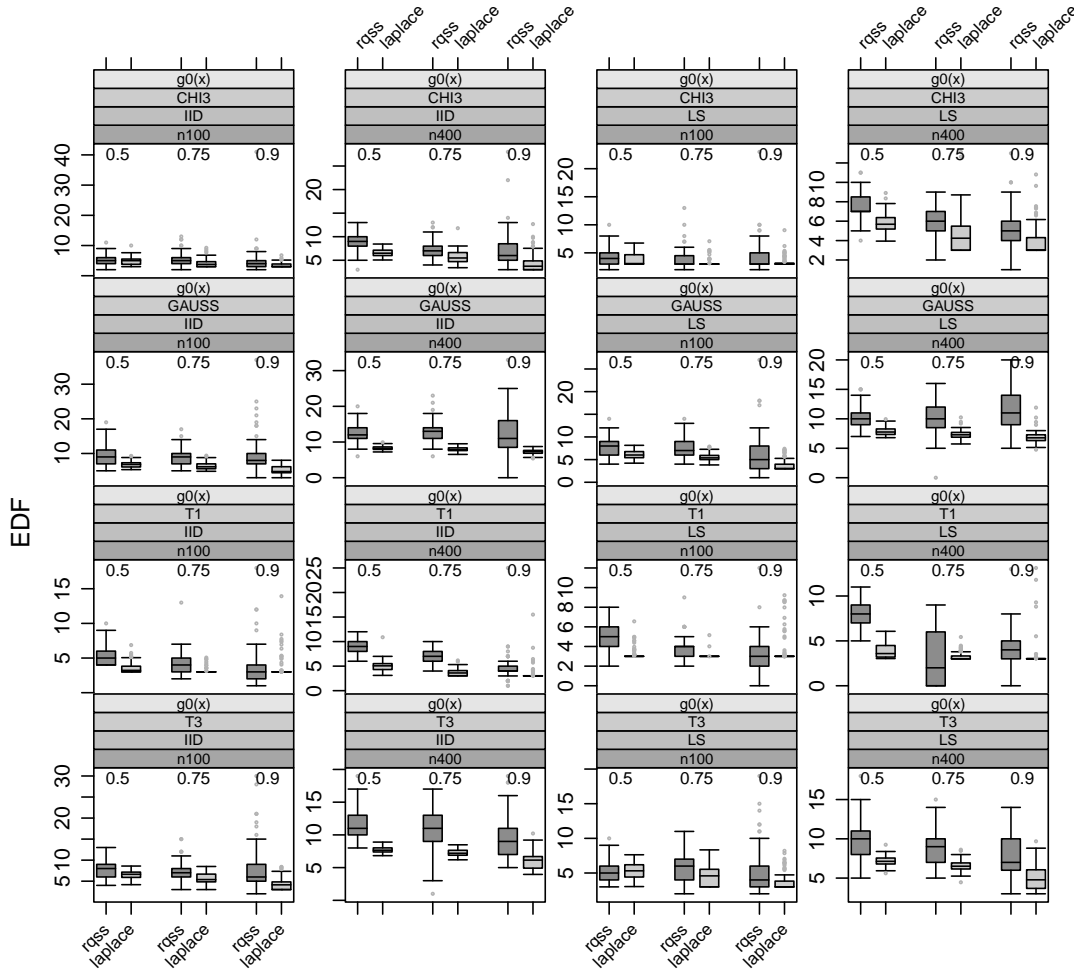
Figure 3.9: Effective Degrees of Freedom to compare *rqss* versus *laplace* methods (part2) for *d* = 3. $g_0(x)$ signal

Figure 3.10: Effective Degrees of Freedom to compare *rqss* versus *laplace* methods (part3) for $d = 3$. Signal log(x)

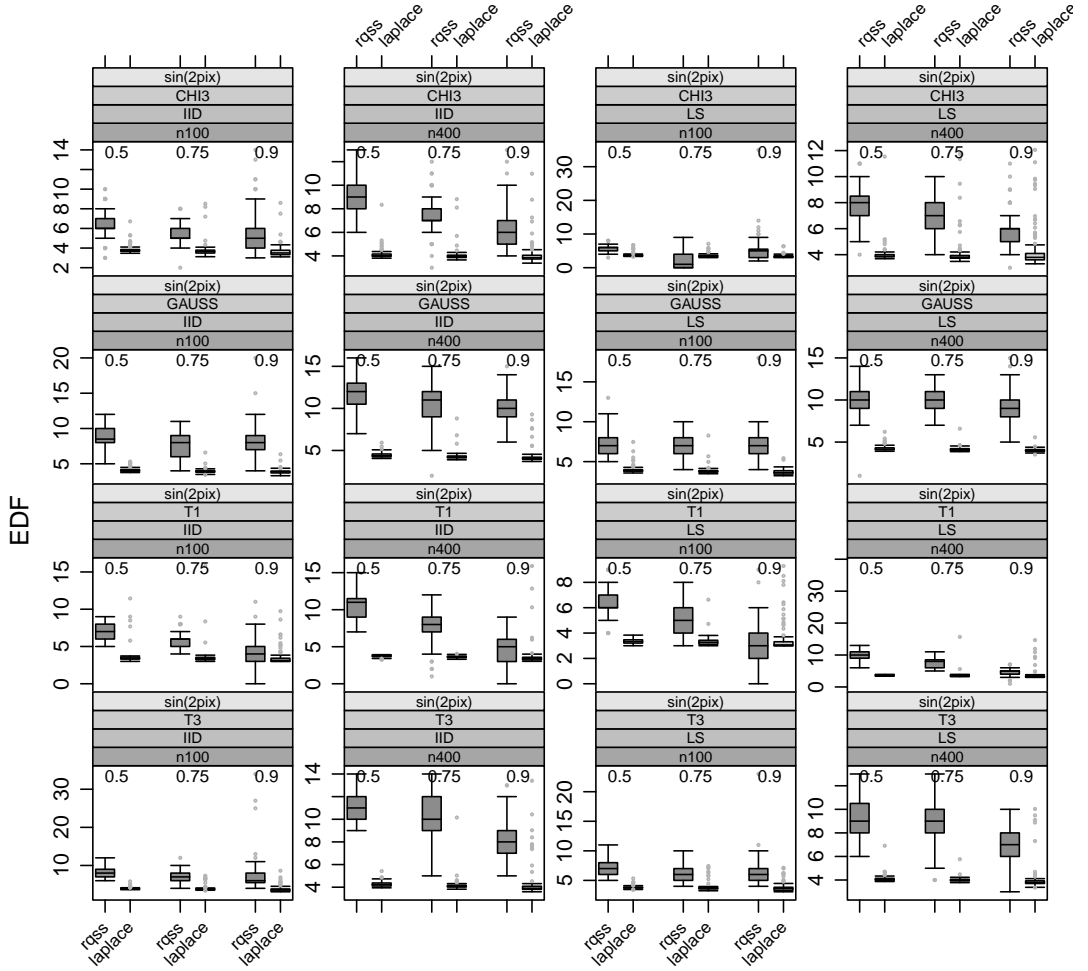Figure 3.11: Effective Degrees of Freedom to compare *rqss* versus *laplace* methods (part4) for $d = 3$. Signal $\sin(2\pi x)$

3.4   Summary

In this chapter a simulation study to compare performances of our algo-
rithm respect to SIC method used in Koenker's function "rqss" have been
provided.

Simulations show that in most of selected scenarios Schall-based algo-
rithms performs equal or sometimes better than "rqss" in terms of RMIAE.
This result is usually achieved regardless of sample size or percentile or
kind of error (IID/LS). The only case of better performance of "rqss" is
with $g_0(x)$ as signal and t_1-distributed heteroskedastic data (see the plot in
Figure 3.5).

"laplace" method seems to be the most reliable and robust in terms of re-
sults. At the moment it is considered the best method even when the per-
formance results equal to "rqss". In fact smoothing spline QR has clearly
an higher computational burden: firstly because smoothing spline involves
all points in the model matrix and secondly because "rqss" smoothing pa-
rameter selection relies on SIC which is a grid-search method.

Simulations computed with $d = 2$ show that even when the signal is linear,
RMIAE derived from the two methods are similar. That is not an obvious
result because SIC method used in "rqss" always results in a piecewise
linear fit and so it could seem more reliable for this situation.

The complexity of the model in "rqss" results higher than in "laplace" for
almost all of the scenarios regardless also of order of the differences d.
The approximation of the hat matrix described in the previous chapter pro-
vides lower degrees of freedom respect to the standard approach relying
on the number of non-zero residuals in the model. To conclude, simula-
tion show that Schall-type algorithms provide less complex models with
the same level of performance and then it can be considered an important

alternative to computationally demanding grid-search methods.

Chapter 4

Application

In this chapter an application for real data is provided. The main objective is to compare two different fertilizers in terms of root length of Sorghum (Bochicchio, 2013). Here we propose a quantile regression approach based on *P*-splines to assess, quantify and compare the root growth patterns in two treatment groups respectively undergoing compost and traditional fertilization.

4.1   QR for plant roots growth

4.1.1   Motivation

Plant roots are a major pool of total carbon in the planet, and their dynamics are directly relevant to greenhouse gas balance. Composted wastes are increasingly used in agriculture for environmental and economic reasons, but their role as a substitute for traditional fertilizers needs to be evaluated and tested on all plant components. At this aim a three-year experiment (2007-2009) was carried out by Dipartimento di Scienze dei Sistemi Colturali,

Forestali e dell'Ambiente, Università degli Studi della Basilicata, Potenza, Italy. Compost application was compared to traditional fertilization with regard to growth of roots of Sorghum bicolor Moench x S. sudanense (Piper) Stapf. in Battipaglia (Sa), Italy. After sowing and treatment of compost or traditional fertilization, plant roots were monitored through sequential images taken with a digital microscope from 4 transparent acrylic access the soil depth of 60 cm from the surface (see Figure 4.1).
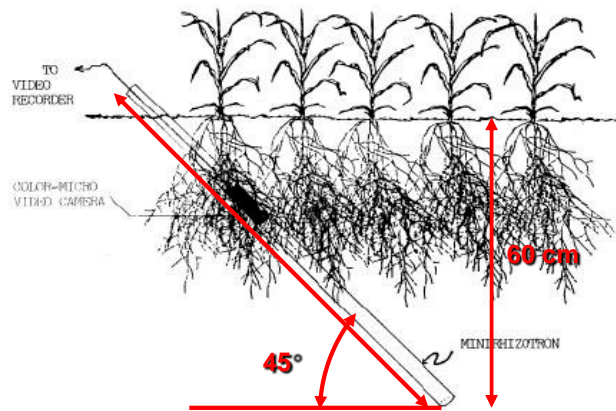

Figure 4.1: Figure showing the experimental procedure to get the images of the roots

A total of 18 images representing depths from 0 to 60 cm were analyzed from all 8 tubes at each date for a total of 3024 images (3 years x 7 dates x 18 depths x 8 tubes). Each image represents an investigated area of 207 mm^2. Image analysis was carried out through a dedicated software three root growth measurements were obtained for each image and four tracked root types: *total length*, *total surface area* and *average diameter* for *total*, *alive*, *white* and *dark* roots. The aim of the experiment was to assess root growth across *days after sowing*, by emphasizing differences due to two

treatment 'arms': compost vs. traditional fertilization.

The usual modeling framework for growth curves is via mean regression, namely by means of specification of a regression equation for the expected value of the response conditional distributions (Pollice *et al.*, 2013). However there are at least two issues that should be emphasized when modelling the data via mean regression. Firstly, the non-negligible portion of zeroes cannot be ignored and it needs to be modelled properly, for instance via mixture models; secondly, and more importantly, mean regression does not provide a complete picture of data when interest lies in studying growth patters, particularly with strongly heteroscedastic data. In order to analyze root growth we propose an approach based on quantile regression (QR); more specifically we aim at modelling the growth patterns, i.e. the growth curves for different quantiles, with respect to days after sowing by emphasizing possible differences due to the two aforementioned treatment groups. There are several additional advantages in using QR, including robustness to outliers and no need to specify the response distribution.

4.1.2 Methods

Let Y be the growth variable, here the total length of roots, $q_{\tau_k}(Y|t, x_i)$ the τ_kth quantile of Y conditional to covariates x_i and time t . We consider the following quantile regression model

$$q_{\tau_k}(Y|t, x_i) = x_i^T \beta_{\tau_k} + s_{\tau_k}(t) \tag{4.1}$$

where β_{τ_k} quantifies the linear effect of p covariates and $s_{\tau_k}(z_i)$ accounts for the growth pattern with respect to days after sowing. Since growth patters are typically nonlinear, $s_{\tau_k}(\cdot)$ is a smooth but unspecified function, and we

use B-splines at this goal, namely $s_{\tau_k}(\cdot) = \sum_j^J b_{jk} B_j(\cdot)$.

By setting $\theta_k = (\beta_k^T, b_k^T)^T$ and $w_i = (x_i^T, B_i^T)^T$, the objective function to be minimized can be written as

$$\sum_i \rho_k(y_i - w_i^T \theta_k) + \lambda \sum_{j=1}^{J-d} |\Delta^d b_k|_j, \qquad (4.2)$$

where $\rho_k(u) = u(\tau_k - I(u < 0))$ is the so-called check function and the penalty term $\lambda \sum_{j=1}^{J-d} |\Delta^d b_k|_j$ controls the wigglyness of fitted curve. Δ^d is the order d difference operator whereby d affects the curve behaviour. Here $d = 3$ then the fit will be piecewise quadratic and as $\lambda \to \infty$ the fitted tends to a quadratic polynomial. Finally objective (4.2) is extended to allow multiple estimation of several quantile curves with noncrossing constraints using appropriate augmented matrices.

4.1.3   Results

We propose the analysis of the growth of dark roots by a refined modelling of the distribution of their total length, rather summarized by six quantiles than by its expectation, using QR. The distributions of dark roots total lengths for DAS classes and treatment groups in Figure 4.2 show strong asymmetric behavior and marked zero inflation soon after sowing.

Notice that the zeroes excess in the continuous response variable corresponds to images containing no roots, and can be understood as roots with no growth. When modelling the expectation of zero-inflated responses common alternatives include mixture modelling (Zuur *et al.*, 2012) and the use of Tweedie distribution models (within the exponential dispersion family framework, see Pollice *et al.* (2013) and references therein). QR is robust to the presence of zeroes excess as in this framework we do not need
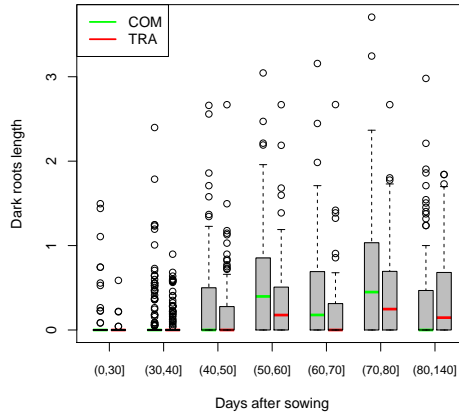
Figure 4.2: Boxplots of dark roots total lengths for DAS classes and treatment groups.

to specify a probability distribution for the response, but we only need to constrain the fitted curve to have only nonnegative values. At this aim we model the log values and then come back to the original growth scale by exponentiating the fitted values; this is legitimate as quantiles are invariant to monotone transformations.

Figure (4.3) displays the fitted quantiles at probability levels $(0.10, 0.25, 0.50, 0.75, 0.90, 0.95)$. The quantile curves at low probability levels are indistinguishable due to the presence of zero values in both treatment groups; however at higher probability levels the two treatments lead to quite different profiles: in the COM group quantile curves are higher and steeper suggesting better performance, particularly within $100 - 120$ days from sowing.

The plot in Figure 4.4 shows how the use of Schall algorithm provides a
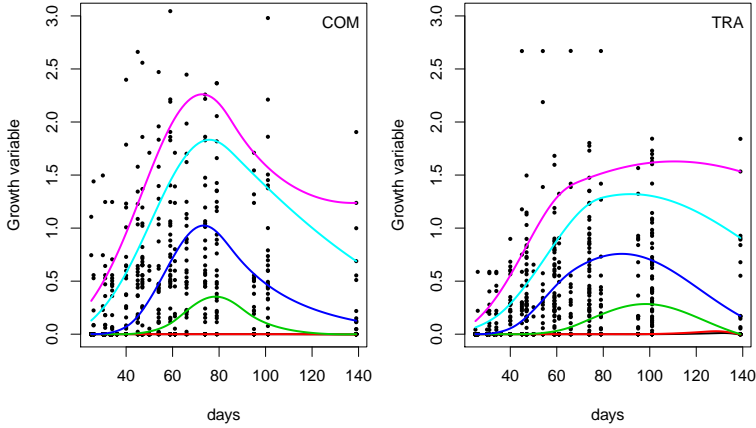
Figure 4.3:   Fitted regression quantiles at probability level
$(0.10, 0.25, 0.50, 0.75, 0.90, 0.95)$ for dark roots total length in the
two treatment groups. Due to the presence of zeroes in both treatments,
quantile curves at low probability level are indistinguishable.

smoother fit than B-spline QR model and smoothing spline QR model. In
particular, the fit provided by the smoothing spline model (green) results
strongly wiggly. To graphically compare the methods, the order of differ-
ences for the B-spline and P-spline models is $d = 2$.

In order to quantify the treatment effect on root growth we consider the
difference of estimated quantile curves at each probability level τ_k

$$\hat{s}_k^{COM}(t) - \hat{s}_k^{TRA}(t) = \sum_j (\hat{b}_{jk}^{COM} - \hat{b}_{jk}^{TRA})B_j(t). \tag{4.3}$$

The rationale is plain: if the two treatments do not make any difference the
difference profile should settle around zero. Asymptotic theory for penal-
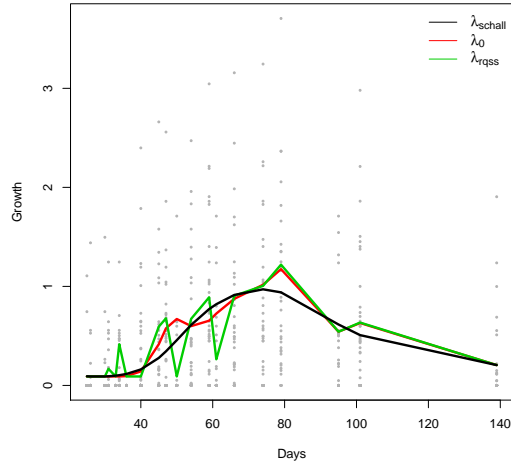ized quantile regression is far from being well established and it is instead a

Figure 4.4: Fitted regression quantiles ($\tau = 0.75$) using P-spline QR via Schall-like algorithm (black), B-spline QR (so with $\lambda = 0$, red line) and smoothing spline QR with selection of λ via SIC (green) using $rqss$ function (Koenker, 2013).

hot and challenging topic (Koenker, 2005); thus, in order to obtain a sample distribution for difference quantiles, we rely on bootstrap according to the following steps:

1. Resample data independently from the two treatment groups;

2. Fit two noncrossing quantile regressions with *P*-splines using the same basis;

3. Compute the difference quantiles (4.3) for each probability level τ_k.

By repeating these steps a large number of times we obtain a bootstrap distribution of the difference quantiles; Figure 4.5 reports the results for each

of six quantiles, showing remarkable differences between the two treatments. The differential evolution of dark root lengths along time for the two treatments highlights agronomic instances that are worth considering in deeper detail.
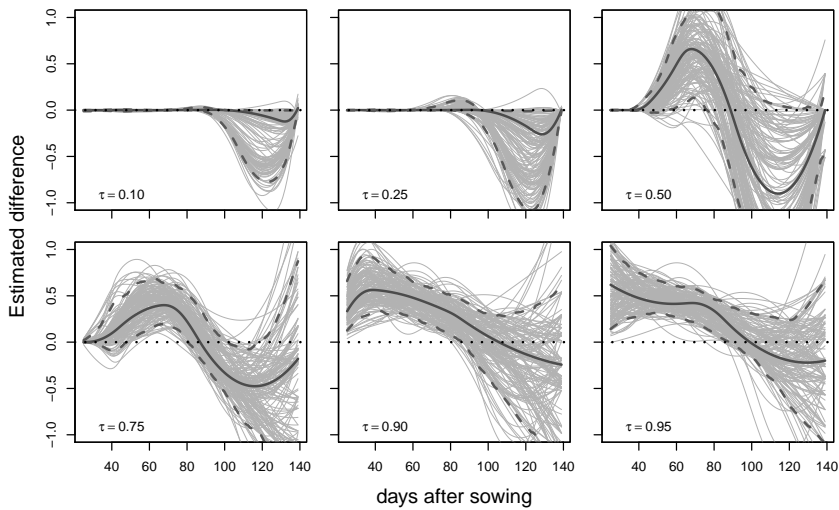

Figure 4.5: Bootstrap distributions of the difference quantile curves for each probability level. Within each panel, one grey line corresponds to the quantile difference (4.3) for one bootstrap replicate. Bold grey lines correspond to point estimates and bootstrap 95% point-wise bands.

Chapter 5

Conclusions

5.1 Summary

In this thesis we have been discussed non parametric Quantile Regression with emphasis on smoothing parameter selection. After explaining some motivations to the use of this technique either via parametric and non parametric approach in the Introduction, smoothing spline and P-spline QR framework have been presented in Chapter 1. In Chapter 2, several versions of smoothing parameter selection via Schall algorithm have been described. In Chapter 3 a simulation study to compare the aforementioned methods with the smoothing spline QR function "rqss" has been reported. Results suggest to conclude that the proposed algorithm is a very valid alternative to grid search methods. Finally, in Chapter 4 an application concerning the performance of two fertilizers in terms root length of sorghum has been reported.

In the implementation of the Schall-type algorithm several specifications of the error and random effect variance have been evaluated. Alternative

methods to grid-search have been provided. The advantage in terms of computational load of those methods respect to smoothing spline QR is quite clear. Simulations have shown that 2 of these methods provide less complex models than "rqss" with the same level of performances in terms of RMISE and RMIAE (or sometimes better).

The application of P-spline QR to compare the performance of the fertilizers allows to know the difference in terms of growth of the plants in the whole distribution.

However, many topics included in this thesis deserve to be studied further: for instance, a proof of the link between penalized and mixed quantile regression models would justify the use of the Schall-like algorithms also from a theoretical point of view. It also would be worthwhile to analyse the performance of the iterative algorithm in the context of longitudinal data: in the linear mixed model framework one could choose the amount of penalty on the random intercepts using the so-called PQL approach. Another relevant topic is the extension of the iterative algorithm to bivariate smoothing wherein there could be the need to consider anisotropic penalties. Moreover, the algorithm is still referred to just one quantile: it would be useful to extend the method in the simultaneous estimation of several quantile curves. That could represent a starting point for the implementation of an R package.

Appendix A

Appendix

In this appendix the results of the whole simulation study described in Chapter 3 are reported.

A.1 Simulation study with order of differences $d = 2$

RMISE

Results are shown in Figure A.1. It does not seem to be too big differences in terms of RMISE for different sample sizes and kind of error (*IID* or *LS*). However, it is possible to see that when $n = 400$ differences amongs methods are a little bit higher than with $n = 100$. When distribution is $\chi_3^2 - 3$ no difference among methods is detected. The "variances" and "scale" methods have the highest RMISE for most of distributions. "varscale" and "laplace" result preferable respect to other methods even when the performance is at the same level of "rqss". In fact the latter has an higher computational load than the formers.
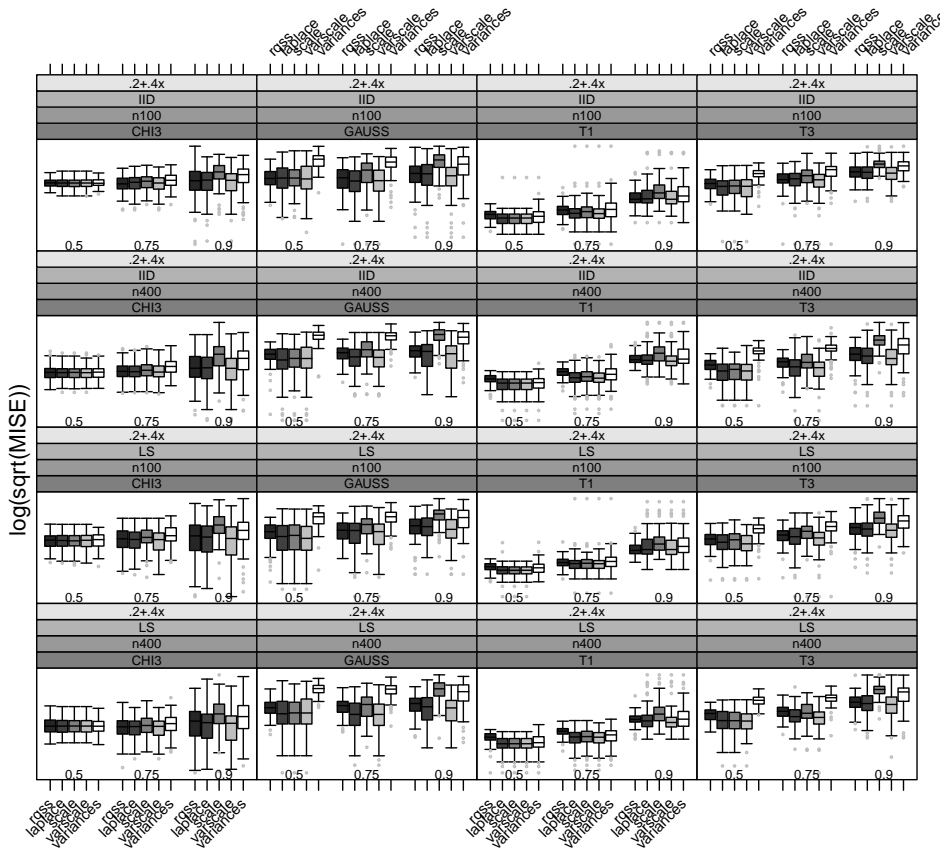
Figure A.1: Root Mean Integrated Squared Error to compare *rqss* versus *laplace* methods with $d = 2$

RMIAE

Results are shown in Figure A.2. There are no differences between methods in terms of RMIAE. The exception is given by "variances" and "scale" which are again the worst methods especially for extreme quantiles. "rqss"
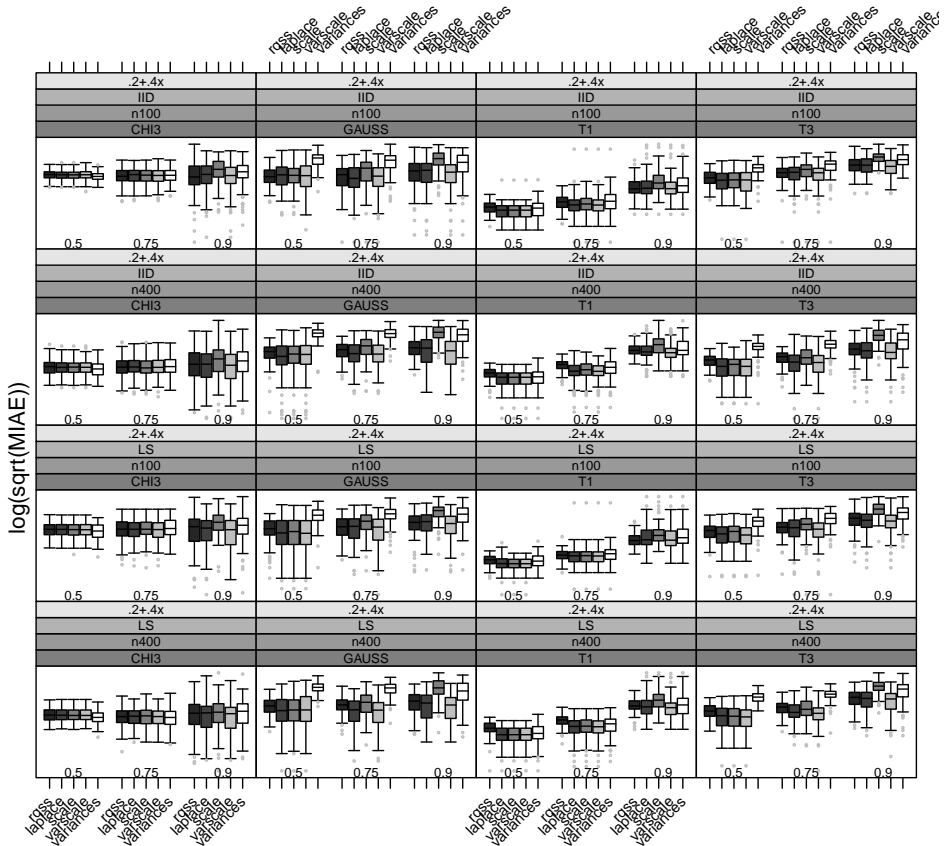
Figure A.2: Root Mean Integrated Absolute Error to compare *rqss* versus *laplace* methods with $d = 2$

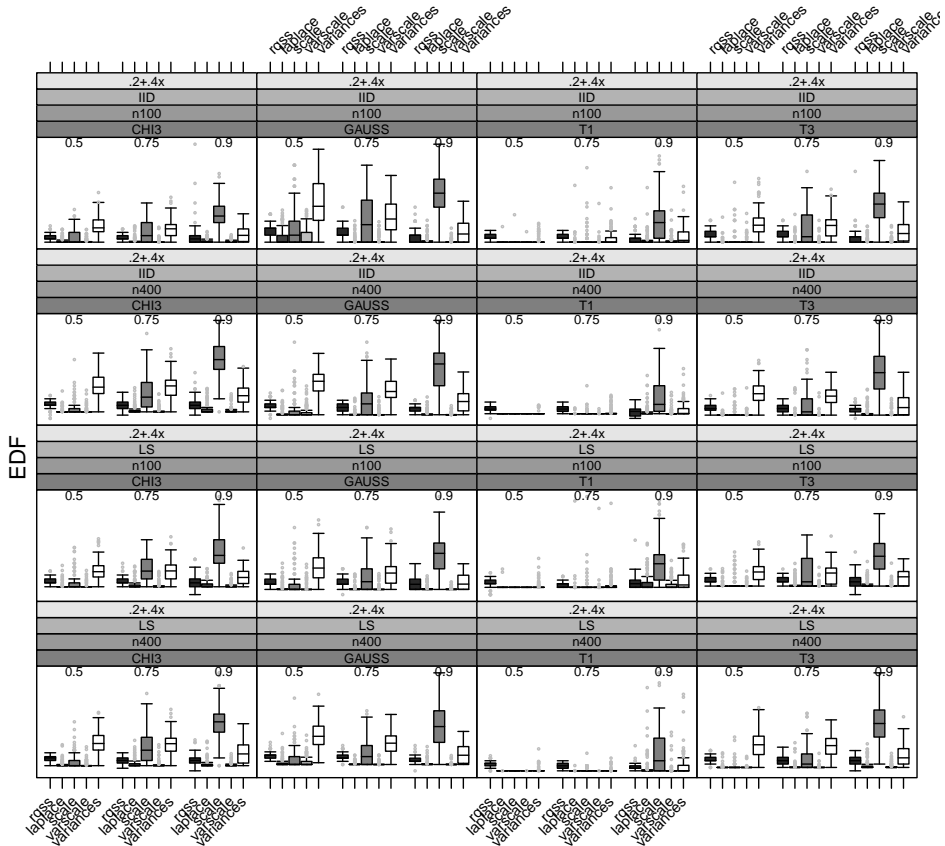performance is the same (or worst for t_1 distribution) respect to "laplace" and "varscale".

Figure A.3: Effective Degrees of Freedom to compare *rqss* versus *laplace* methods with $d = 2$

Effective dimension of the model

Results are shown in Figure A.3. It is easy to note that "laplace" and "varscale" provide models with lower effective dimension than the other methods. It means that model based on these methods are less complex

than the others. Furthermore, there is very low variability of results which validate the robustness of the computation of the ED via smooth approximation. It should be emphasized that "variances" and "scale" yield the most complex models.

A.2 Simulation study with order of differences $d = 3$

RMISE

The methods have the same performance varying kind of error and sample size. The plots in Figures 3.4, 3.5, 3.6, 3.7 suggest that no differences between methods are detected when distribution is $\chi_3^2 - 3$. The exception is for "scale" and "variances" methods which have an higher RMISE for $\tau = 0.9$.

For Gaussian data "rqss" often performs worst than the other methods while for linear signal "scale" and "variances" have the highest RMISE.

When distribution is t_1, "variances" seems to have the lowest RMISE, especially for $\log(x)$ and $g_0(x)$ and $\tau = 0.5, 0.75$. The other methods based on the iterative algorithm seems to perform slightly worst than "rqss" especially for signals $\log(x)$ and $g_0(x)$.

For data generated from a t_3 distribution "rqss" is often the worst method. However as the quantile becomes more extreme the differences are less evident. At the same time for "scale" and "variances" RMISE increases when τ increases.

RMIAE

The result shown in Figures 3.4, 3.5,3.6 and 3.7,are very similar to the RMISE but the differences seems to be less evident. In practise "rqss"
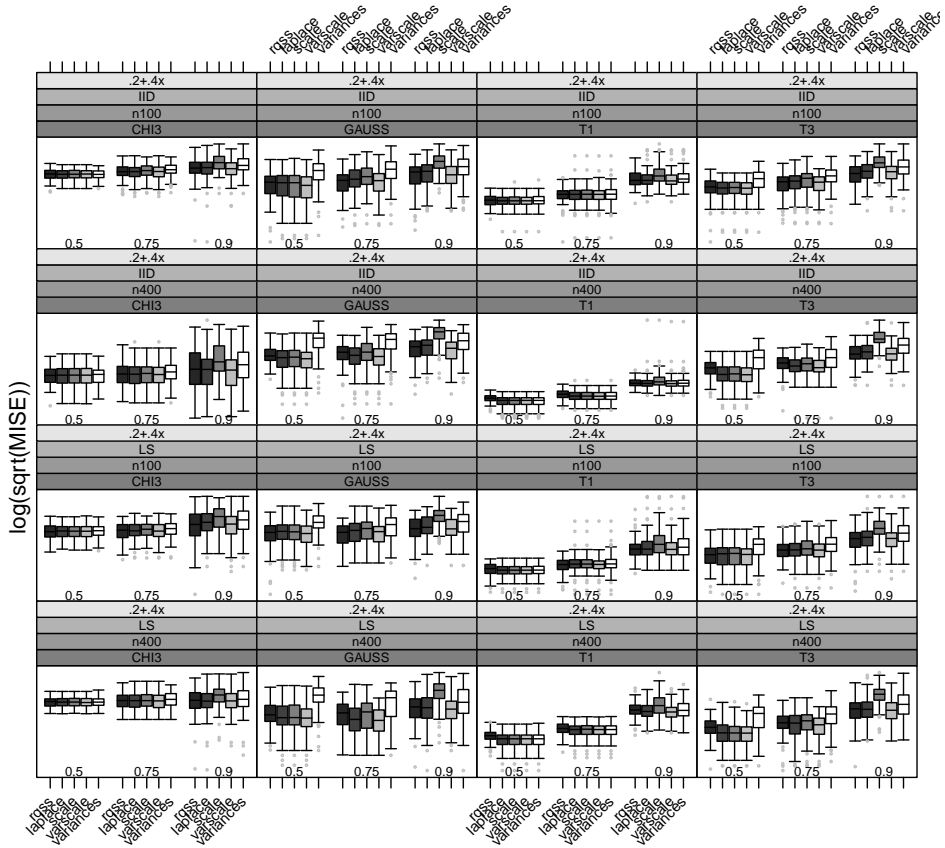
Figure A.4: Root Mean Integrated Squared Error to compare *rqss* versus *laplace* methods (part 1) for *d* = 3. Linear signal
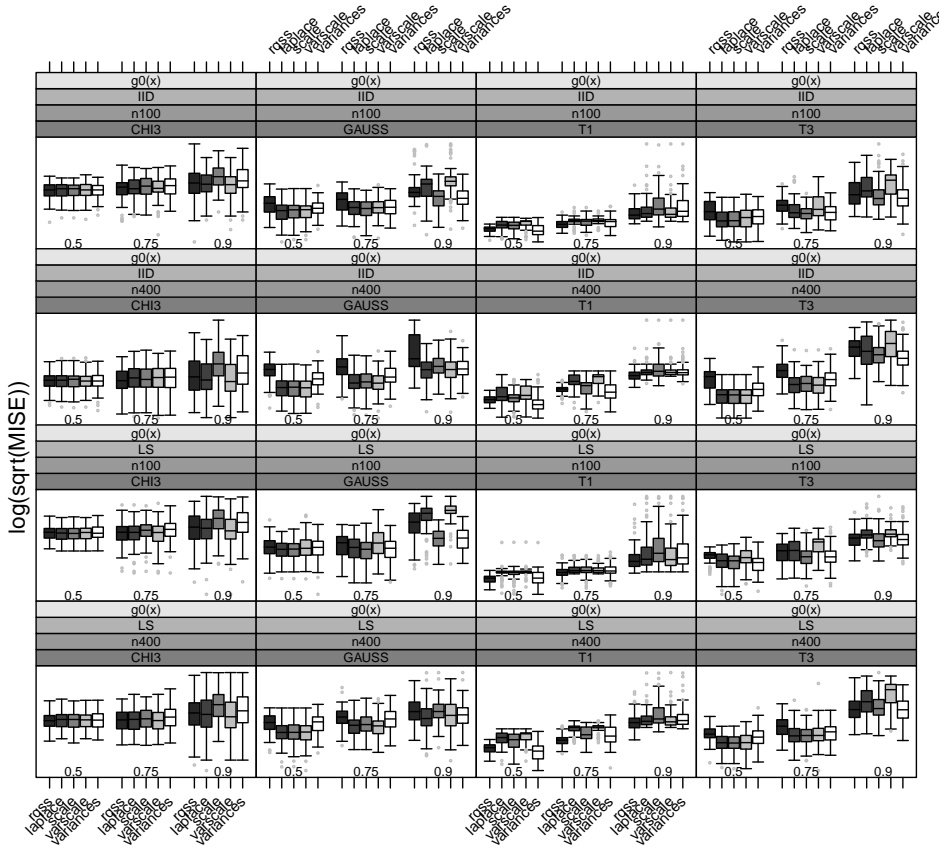
Figure A.5: Root Mean Integrated Squared Error to compare *rqss* versus *laplace* methods (part 2) for $d = 3$. $g_0(x)$ signal
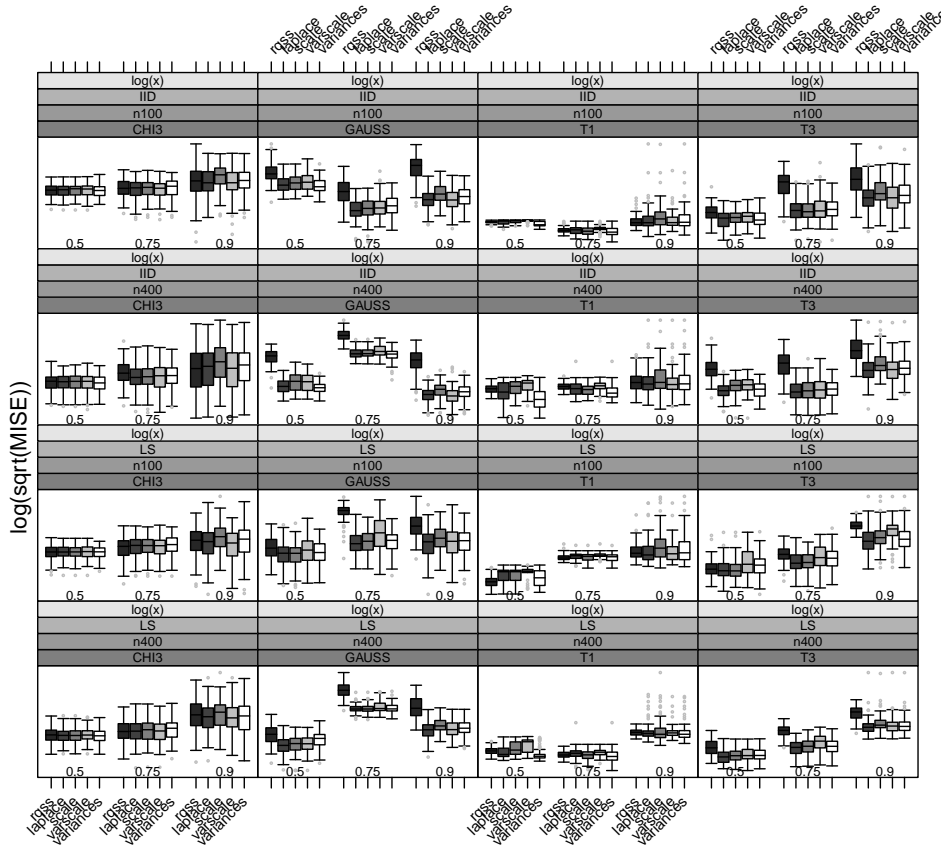
Figure A.6: Root Mean Integrated Squared Error to compare *rqss* versus *laplace* methods (part3) for $d = 3$. $\log(x)$ signal
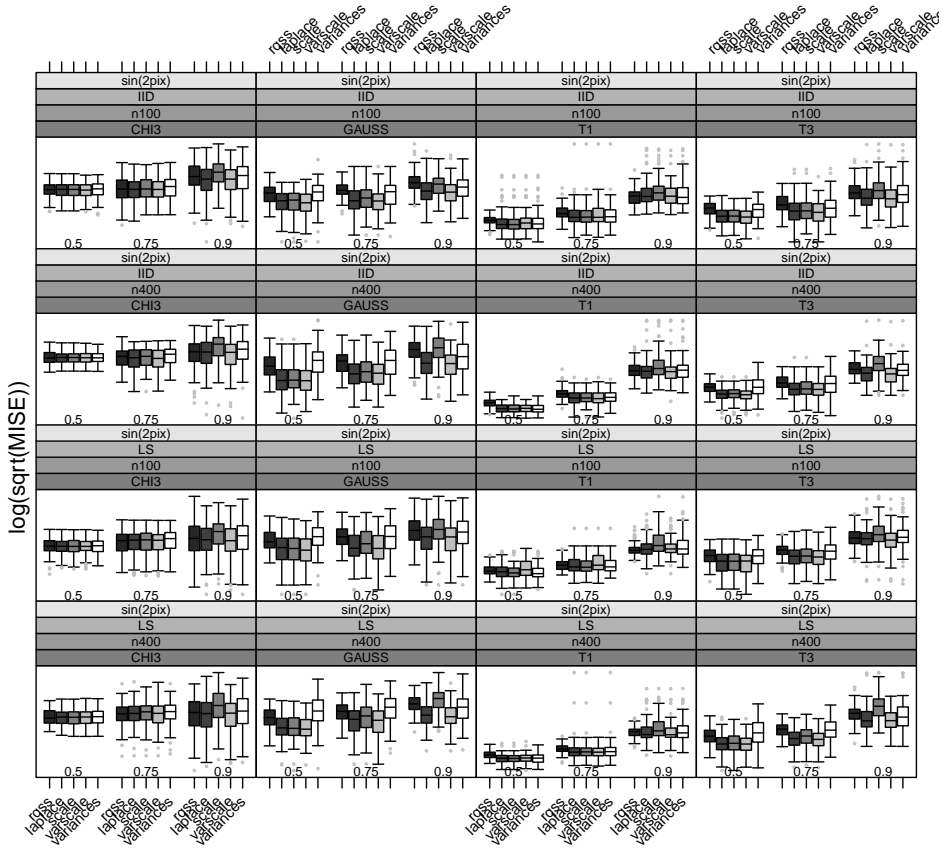
Figure A.7: Root Mean Integrated Squared Error to compare *rqss* versus *laplace* methods (part4) for $d = 3$. $\sin(2\pi x)$ signal

often works bad for Gaussian and t_3-distributed data when the signal is non linear and slightly better than the other methods (except "variances") when the error is distributed as t_1 and the signal is $g_0(x)$. For $\chi_3^2 - 3$ distributed data, differences among methods are negligible.
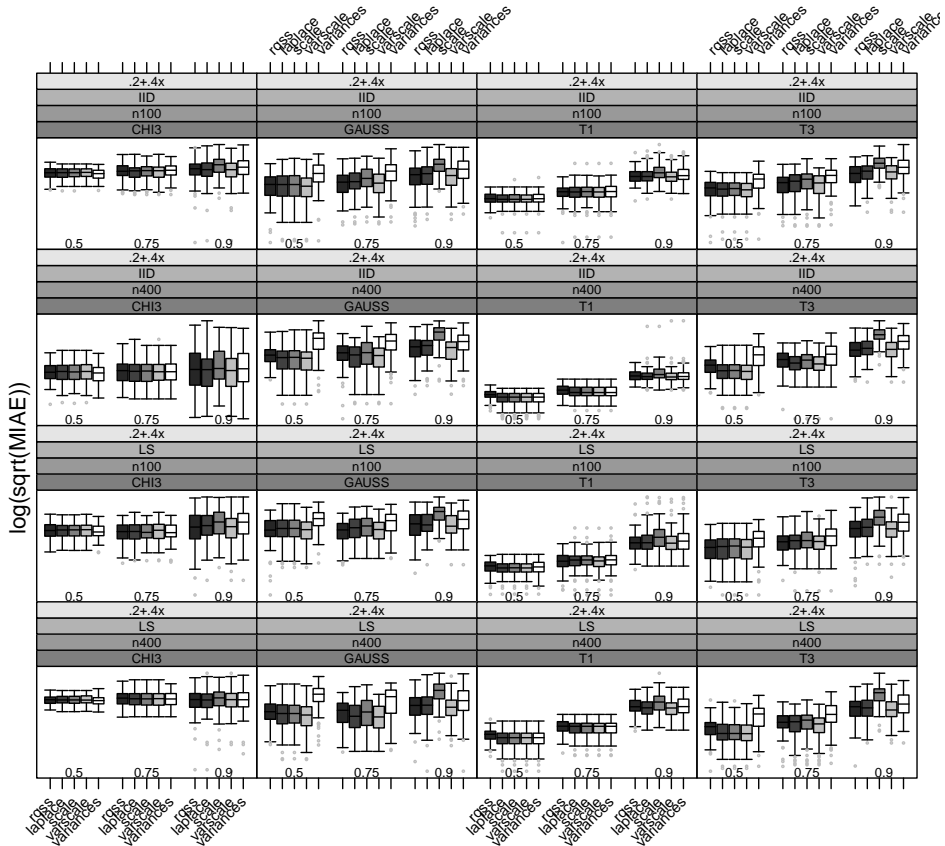

Figure A.8: Root Mean Integrated Absolute Error to compare *rqss* versus *laplace* methods (part1) for $d = 3$. Linear signal
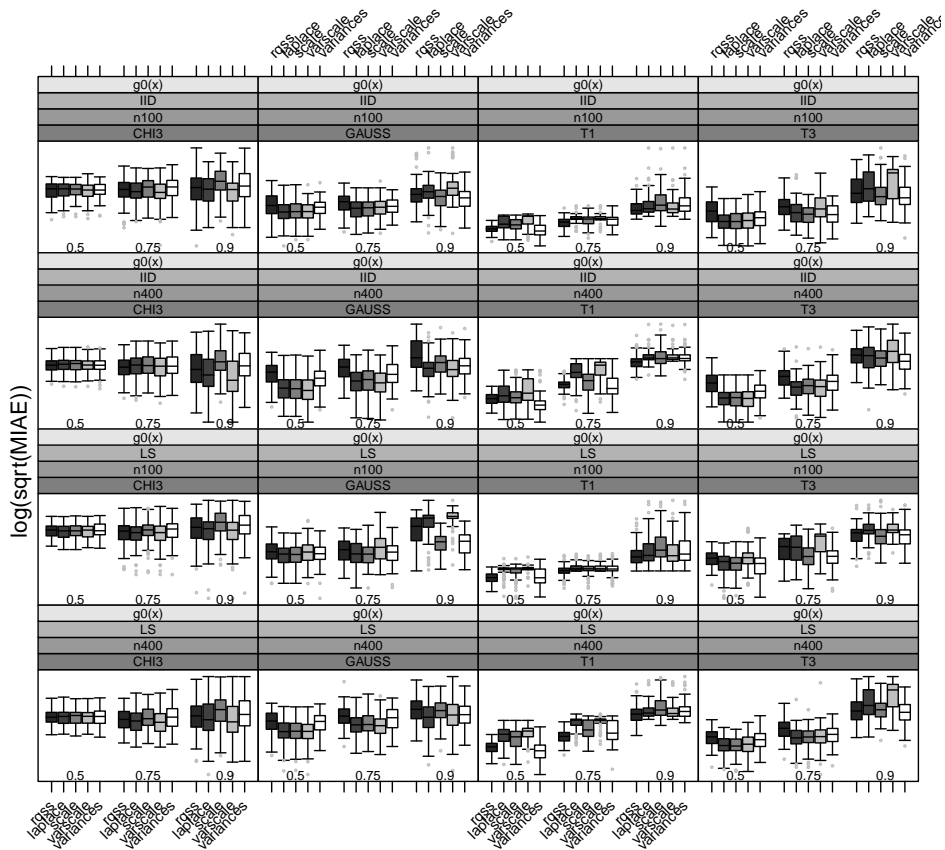
Figure A.9: Root Mean Integrated Absolute Error to compare *rqss* versus *laplace* methods (part2) for $d = 3$. $g_0(x)$ signal
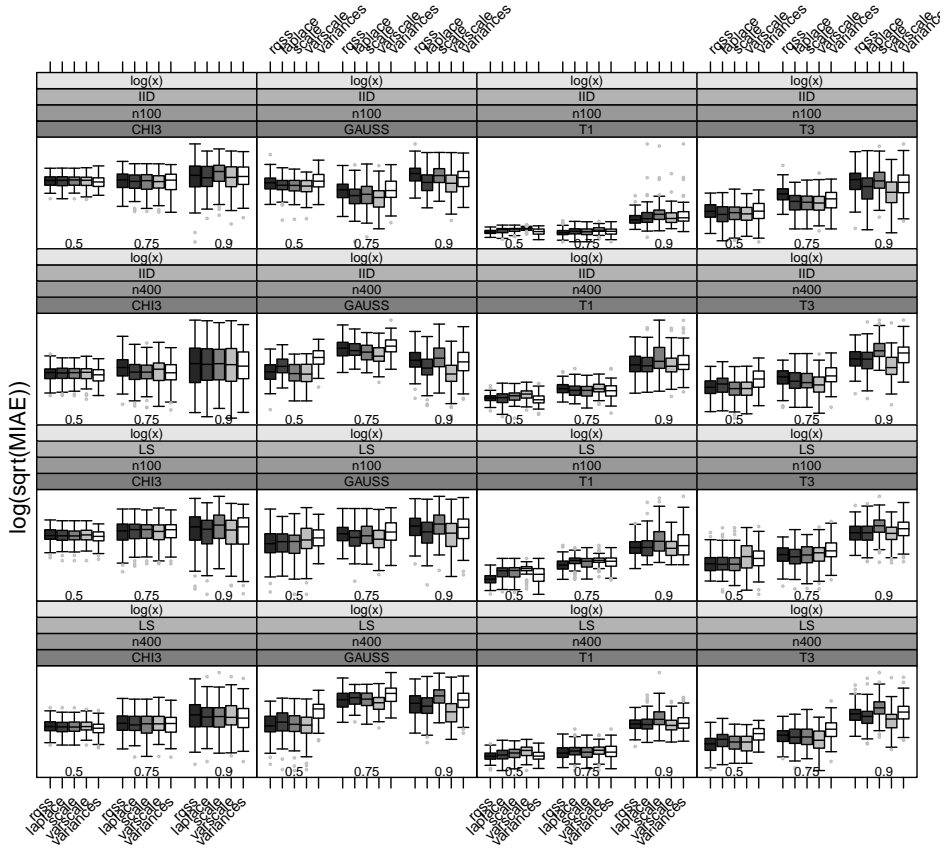
Figure A.10: Root Mean Integrated Absolute Error to compare *rqss* versus *laplace* methods (part3) for *d* = 3. log(*x*) signal
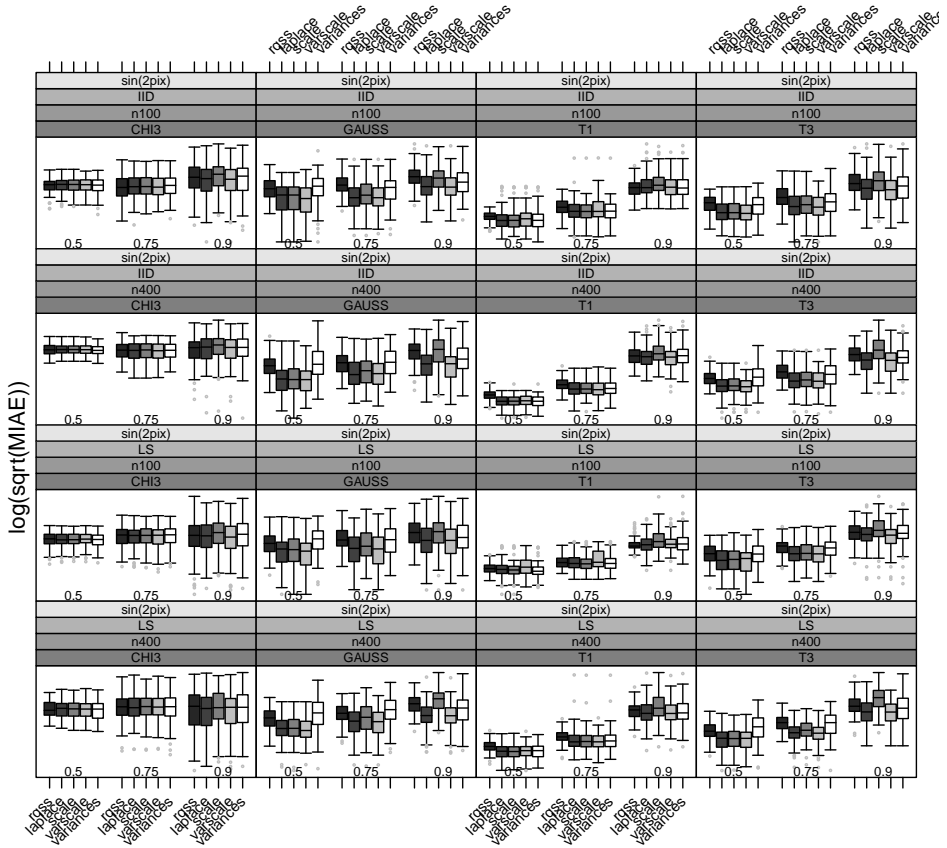
Figure A.11: Root Mean Integrated Absolute Error to compare *rqss* versus *laplace* methods (part4) for $d = 3$. $\sin(2\pi x)$ signal

EDF

Results are shown in Figures 3.8, 3.9, 3.10 and 3.11. With $\chi_3^2 - 3$ distributed data "rqss","scale" and "varscale" provide the most complex models. The effective dimension achieved with the other two methods seems to be more reliable. One can check on the very low variability of the boxplots referred to these methods to understand it.

A similar situation occurs for the other errors nad signals. "rqss" give less complex model than "laplace" only with Gaussian error and $g_0(x)$ signal.

Figure A.12: Effective Degrees of Freedom to compare *rqss* versus *laplace* methods (part1) for $d = 3$. Linear signal
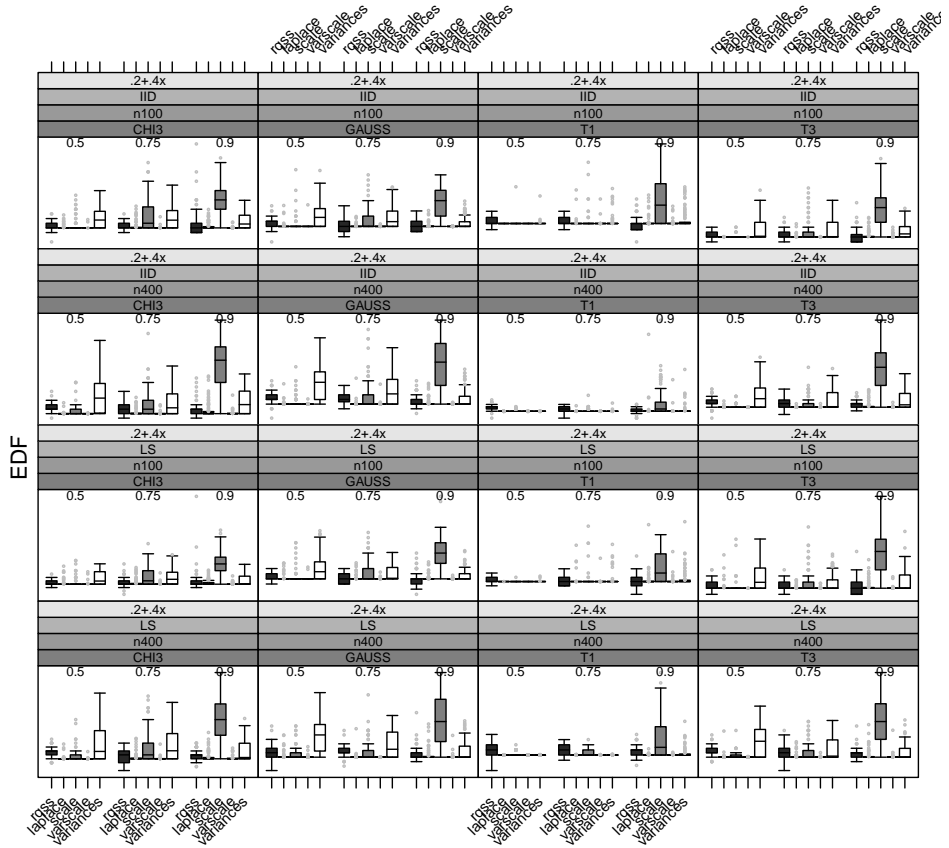
Figure A.13: Effective Degrees of Freedom to compare *rqss* versus *laplace* methods (part2) for $d = 3$. $g_0(x)$ signal

Figure A.14: Effective Degrees of Freedom to compare *rqss* versus *laplace* methods (part3) for $d = 3$. Signal $\log(x)$

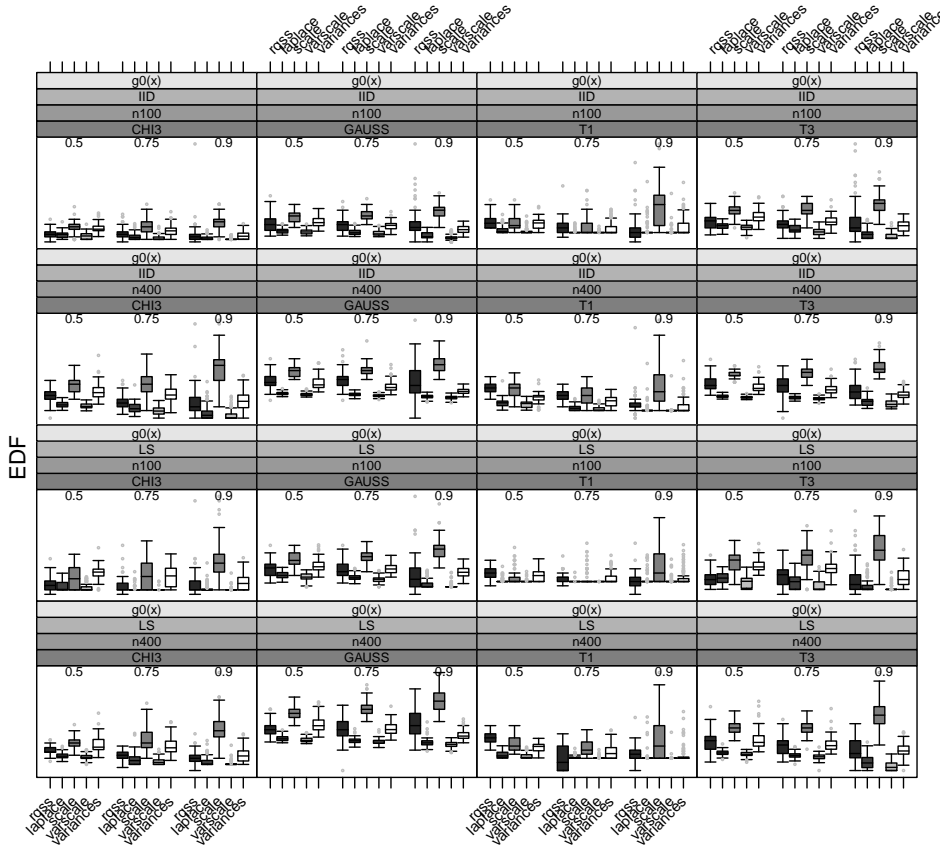Figure A.15: Effective Degrees of Freedom to compare *rqss* versus *laplace* methods (part4) for *d* = 3. Signal sin(2π*x*)

Bibliography

Andriyana, Y., Gijbels, I. and Verhasselt, A. (2014). P-splines quantile regression estimation in varying coefficients models. *Test*, **23**, 153–194.

Bochicchio, R. (2013). *The contribution of roots to the sustainable management of cerals in a changing climate: experimental evidence from Sorghum (Sorghum bicolor (L.) Moench x S. su- danense (Piper) Stapf) under biosolid amendment, and Wheat (Triticum aestivum L.) under simulated climate change*. Ph.D. thesis, Universitá degli Studi della Basilicata.

Bollaerts, K., Eilers, P. and Aerts, M. (2006). Quantile regression with monotonicity restrictions using p-splines and the l1-norm. *Statistical Modelling*, **6**(3), 189–207.

Chaudhuri, P. (1991). Global nonparametric estimation of conditional quantile functions and their derivatives. *Journal of Multivariate Analysis*, **39**(2), 246–269.

Cole, T. J. and Green, P. J. (1992). Smoothing reference centile curves: the lms method and penalized likelihood. *Statistics in medicine*, **11**(10), 1305–1319.

Davino, C., Furno, M. and Vistocco, D. (2013). *Quantile regression: theory and applications*. John Wiley & Sons.

de Onis, M. (2006). *WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age*. WHO.

Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, **1**(1), 93–125.

Eilers, P. and Marx, B. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, **11**(2), 89–121.

Eilers, P. H. and De Menezes, R. X. (2005). Quantile smoothing of array cgh data. *Bioinformatics*, **21**(7), 1146–1153.

Engle, R. F. and Manganelli, S. (2004). Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, **22**(4), 367–381.

Ferguson, T. S. (1967). Mathematical statistics.

Frasso, G. and Eilers, P. H. (2015). L-and v-curves for optimal smoothing. *Statistical Modelling*, **15**(1), 91–111.

Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics*, **31**(1), 3–21.

Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, **8**, 140–154.

Geraci, M. and Bottai, M. (2014). Linear quantile mixed models. *Statistics and computing*, **24**, 461–479.

Green, P. J. and Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press.

Hansen, P. C. (1992). Analysis of discrete ill-posed problems by means of the l-curve. *SIAM review*, **34**(4), 561–580.

Inusah, S. and Kozubowski, T. J. (2006). A discrete analogue of the laplace distribution. *Journal of statistical planning and inference*, **136**(3), 1090–1102.

Jones, M. (1994). Expectiles and m-quantiles are quantiles. *Statistics & Probability Letters*, **20**, 149–153.

Kneib, T. (2013). Beyond mean regression. *Statistical Modelling*, **13**(4), 275–303.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, **91**(1), 74–89.

Koenker, R. (2005). *Quantile regression*. Cambridge University Press.

Koenker, R. (2013). Quantreg: quantile regression. *R package version*, **5**.

Koenker, R. and Bassett, W. (1978). Regression quantiles. *Econometrica*, **46**, 33–50.

Koenker, R., Ng, P. and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, **81**(4), 673–680.

Koenker, R. W. and d'Orey, V. (1987). Algorithm as 229: Computing regression quantiles. *Applied Statistics*, pages 383–393.

Koltz, S., Kozubowski, T. and Podgorski, K. (2001). The laplace distribution and generalizations: A revisit with applications to communications, economics, engineering, and finance.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006). *Generalized linear models with random effects: unified analysis via H-likelihood*. CRC Press.

Muggeo, V. M. (2014). Quantreggrowth. *R package version*.

Muggeo, V. M., Sciandra, M. and Augugliaro, L. (2012). Quantile regression via iterative least squares computations. *Journal of Statistical Computation and Simulation*, **82**.

Muggeo, V. M., Sciandra, M., Tomasello, A. and Calvo, S. (2013). Estimating growth charts via nonparametric quantile regression: a practical framework with application in ecology. *Environmental and ecological statistics*, **20**(4), 519–531.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, pages 819–847.

Ng, P. and Maechler, M. (2007). A fast and efficient implementation of qualitatively constrained quantile smoothing splines. *Statistical Modelling*, **7**(4), 315–328.

Nychka, D., Gray, G., Haaland, P., Martin, D. and O'connell, M. (1995). A nonparametric regression approach to syringe grading for quality improvement. *Journal of the American Statistical Association*, **90**(432), 1171–1178.

Oh, H.-S., Nychka, D., Brown, T. and Charbonneau, P. (2004). Period analysis of variable stars by robust smoothing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **53**(1), 15–30.

Pollice, A., Bochicchio, R. and Amato, M. (2013). Nonlinear growth curves of sorghum roots under compost/traditional fertilization. In *TIES 2013 the 23th Annual Conference of The International Environmetrics Society*.

Portnoy, S., Koenker, R. *et al.* (1997). The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, **12**(4), 279–300.

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

Reiss, P. T. and Huang, L. (2012). Smoothness selection for penalized quantile regression splines. *The international journal of biostatistics*, **8**(1).

Rodríguez-Álvarez, M. X., Lee, D.-J., Kneib, T., Durbán, M. and Eilers, P. (2014). Fast smoothing parameter separation in multidimensional generalized p-splines: the sap algorithm. *Statistics and Computing*, pages 1–17.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge university press.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**(4), 719–727.

Schnabel, S. and Eilers, P. (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis*, **53**, 4168–4177.

Sobotka, F. and Kneib, T. (2012). Geoadditive expectile regression. *Computational Statistics & Data Analysis*, **56**(4), 755–767.

Sobotka, F., Schnabel, S., Schulze Waltrup, L., Eilers, P., Kneib, T. and Kauermann, G. (2011). expectreg: Expectile and quantile regression. *R package version 0.25*.

Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, **23**(7), 1–46.

Waltrup, L. S., Sobotka, F., Kneib, T. and Kauermann, G. (2014). Expectile and quantile regression—david and goliath? *Statistical Modelling*, page 1471082X14561155.

Wood, S. (2006). *Generalized Additive Models- An introduction with R.* Chapman & Hall.

Yee, T. W. (2004). Vgam family functions for quantile regression. *University of Auckland*.

Yu, K. and Zhang, J. (2005). A three-parameter asymmetric laplace distribution and its extension. *Communications in Statistics—Theory and Methods*, **34**(9-10), 1867–1879.

Yuan, M. (2006). Gacv for quantile smoothing splines. *Computational statistics & data analysis*, **50**(3), 813–829.

Zimmerman, D. L., Núñez-Antón, V., Gregoire, T. G., Schabenberger, O., Hart, J. D., Kenward, M. G., Molenberghs, G., Verbeke, G., Pourahmadi, M., Vieu, P. *et al.* (2001). Parametric modelling of growth curve data: an overview. *Test*, **10**(1), 1–73.

Zuur, A. F., Saveliev, A. A. and Ieno, E. N. (2012). *Zero inflated models and generalized linear mixed models with R*. Highland Statistics Limited Newburgh.