Modeling of Uncertainty in Wind Energy Forecast

Jan Kloppenborg Møller

Kongens Lyngby 2006
IMM-2006-10

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

Summary

The present work give a presentation of the theory of quantile regression and splines. Quantile regression and splines are combined to model the prediction error from Tunø Knob wind power plant. This data set is used as the basis for a discussion of performance parameters for quantiles.

An adaptive method for quantile regression is developed, this proves to give convincing results compared to the static model. The implementation of this also proves to be fast.

A method for restricted quantile regression for non crossings is implemented and analyzed. Further different approaches for solving the non crossing constraint problem is discussed.

Resumé

Fraktil regression er en metode til at modellere fraktiler i den betingede fordeling direkte. Ved linear fraktilregression forståes en linear model med en absolut og asymmetrisk tabs funktion.

Fraktil regression beskives og der gives specielt en simplex formulering af det tilknyttede linæere programmerings problem. Fraktilregression benyttes sammen med splines til at modellere fraktiler i forudsigelsesfejlene fra Tuno Knob vindmøllepark. I forbindelse med denne analyse diskuteres metoder til at beskrive kvaliteten af fraktilforudsigelser.

På basis af simplex formuleringen af fraktilregressions problemet, udvikles en adaptiv metode til fraktil regression, som er effektiv og relativt hurtig. Rapporten slutter af med at diskutere forskellige former for restriktioner på fraktilregression.

Preface

This Master thesis was prepared at Informatics Mathematical Modelling, the Technical University of Denmark during the period form February 1th 2005 to 1th February 2006.

The subject of the thesis is quantile regression and splines in the context of wind power modeling.

Lyngby, February 2006

Jan Kloppenborg Møller

Acknowledgements

I thank my supervisors Henrik Madsen and Henrik Aalborg Nielsen for guidance throughout this project.

Notation

General Notation

$\mathbb{R}$: The set of real numbers.

$\mathbb{Z}$: The set of all integers $\{..., -2, -1, 0, 1, 2, ...\}$.

$\mathbb{R}_+$: The set of positive real numbers.

$\mathbb{R}_0$: The set of non-negative real numbers.

$\mathbb{R}^n$: The n-dimensional real vectorspace.

$\mathbb{R}^{n\times m}$: The set of $m \times n$ real matrices.

$\mathbb{P}_n$: The space of polynumials of degree n, ie. $p_n \in \mathbb{P}_n \Rightarrow p_n(x) = \sum_{j=0}^{n} a_j x^j \quad a_j \in \mathbb{R}$.

CDF : Continuous Distribution Function

p.d.f : Probability Density Funftion

idd : Independent Identically Distributed

IQR : Inter Quartile Range, i.e. the difference of the 75% quantile and the 25% quantile.

Vectors and Matrices

$\mathbf{I}$: The identity matrix. The dimension will usually be clear form the context, otherwise it will be denoted $\mathbf{I}_k$, k being the dimension.

$\mathbf{e}$,$\mathbf{e}_k$: A vector consisting of ones, if the length is not cleare from the context then k will refer to this.

e_k : A vector consisting of zeros except for the k'th element which is one.

$\mathbf{0}$: A vector or matrix consisting of zeros. If the dimension is not clear it will be denoted $\mathbf{0}_k$ ($\mathbf{0}_{k\times l}$).

$\mathbf{x} = [x_i]$: Bold face lower case letter is used for vectors. A vector in this presentation is always a colum vector. $\mathbf{x}$, will be used for independent or explanatory variable. $\mathbf{y}$ will be used for dependent or respons variable. Constant vectors will be denoted $\mathbf{a}, \mathbf{b}, \mathbf{c}$ etc.

$\mathbf{A}, \mathbf{X}$: Bold face upper case letters is used for matrices. $\mathbf{X}$ will be used for the design matrix.

$\mathbf{A}_{:,s}$: The s'th column of the matrix, s can be a set of indexes $\mathbf{A}$.

$\mathbf{A}(h)$: The h'th row or rows of the matrix, h can be a set of indexes. $\mathbf{A}$.

$\mathbf{x} \odot \mathbf{y}$: $[x_i y_i]$ Elementwise multiplication.

$\text{diag}(\mathbf{A})$: A vector with the diagonal elements of the matrix $\mathbf{A}$.

$\text{diag}(\mathbf{a})$: A matrix with the diagonal elements equal to $\mathbf{a}$ and all other elements equal to zero.

$\mathbf{y}(h), \mathbf{y}_{\mathcal{B}}$: Vector containing the elements of index set h or $\mathcal{B}$.

Functions and Operators

A variable x in a function $f(x)$ such that $f : \mathbb{R} \to \mathbb{R}$ that is replaced with a vector $\mathbf{x}$ will be equivalent to

$$f(\mathbf{x}) = [f(x_i)]$$

Such that $f(\mathbf{x})$ is a vector of the function vaules of each element in $\mathbf{x}$.

$\mathbf{x} \leq \mathbf{y}$: $x_i \leq y_i \forall i$

$I(exp)$: Indicator function ie. it is one if the logical expression exp is true and zero otherwise.

$f_+(x)$: $f_+(x) = I(x \geq 0) f(x)$.

$f_-(x)$: $f_-(x) = I(x < 0) f(x)$.

$x^+ = x_+$: $x^+ = I(x_i \geq 0)x$.

$x^- = x_-$: $x^- = I(x_i < 0)x$.

$\delta(k)$: Kronekers delta-sequence $\delta(0) = 1$ and $\delta(k) = 0$ for $k \neq 0$ and $k \in \mathbb{Z}$.

$\text{sign}(x)$: $\text{sign}(x) = -I(x < 0) + I(x > 0)$, i.e. $\text{sign}(0) = 0$.

$\lceil x \rceil$: largest integer s.t. $\lceil x \rceil \leq x$ (also called ceil in e.g. matlab and R).

$\lfloor x \rfloor$: smallest integer s.t. $\lfloor x \rfloor \geq x$ (also called floor in e.g. matlab and R)

$x_{(i)}$: The order statistics of $\mathbf{x}$, i.e. $x_{(1)} \leq x_{(2)} \leq .. \leq x_{(N)}$.

$\wedge$: logic "and" s.t. $exp_1 \wedge exp_2 = I(exp_1) I(exp_2)$

$\overline{h}$: If h is a set then $\overline{h}$ is the complement of h.

$\overline{\mathbf{x}}$: If $\mathbf{x}$ is a vector then $\overline{\mathbf{x}}$ is the average of $\mathbf{x}$.

Reserved Characters

i : The counter i is reserved for counting observations.

N : N is reserved for sample size.

$\mathbf{r}$: The residuals of a model.

h : The index set which charactirize the solution of to the quantile regression problem, see Theorem 2.1.

$\rho(r)$: The loss functiuon of r, see Section 2.1

The Environments in the Text

Through the text there will be different environments. The purpose is to make the text more readable, and to highlight important passages of the text.

The environments are

Definition 0.1 *The Definition environment is used for basic definition of the theory.*

Theorem 0.1 *The theorems provide the theoretical foundation of the practical use of the theory. There have not been made any attempt to distinguish between, lemmas, theorems, proportions, etc. These all have the common label Theorem.*

Practical Summary 0.1 *Practical summary's are used for important computational recipes. These are not very deep in a theoretical sense.*

Example 0.1 The example environment provide simple example of use of the theory, the purpose of the example is to illustrate the theory, rather than to use the theory in the same way as it is done later on in the text.

PROOF. The proof environment is use for proofs of theorems, when the proof is given in connection with the theorem. □

Contents

Chapter 1

Introduction

As the share of the total energy production produced by wind power increase, so will the need for precise forecasts of the power production. This presentation discuss one approach to get better forecasts.

A good forecast will often be thought of as the mean value and efforts of forecasting will be concentrated on improving mean velum forecast by minimizing the variance of the prediction errors. If this is the right strategy really depend on the penalty for making a wrong forecast. If this penalty is not symmetric then the mean value might not be very important.

Energy is traded on a market called NordPool (see www.nordpool.com), the power suppliers put up energy for sales once a day and then decide how to act on this market. I.e. develop the optimal strategy for buying or selling energy. In such a market the penalty for making a wrong bet is not symmetric.

Therefore the mean value is not to develop an optimal market strategy enough, and a good or precise forecast will mean more information than we can get from the mean value.

In a more general setting if we know that the penalty p for choosing the strategy s in situation y, i.e. $p(s, y)$ is a known function of strategy and outcome, but y

is a random variable, then we would like to solve something like

$$\min_s \int p(s,y)dF(y) \tag{1.1}$$

where F is the distribution function for y. To solve something like (1.1) we need the distribution F, if we want to know the distribution F of y then it is not enough to have estimates of the the mean value and e.g possible the variance.

This presentation give a treatment of a possibly way to find this distribution function F. In the case of wind power production such a distribution will depend on the weather conditions or the meteorological forecasts. Thus what we need is an estimate of the distribution of power production given some meteorological forecasts. This is the conditional distribution of power production given a meteorological forecast.

Chapter 2 deal with this problem for the case, where we believe that the power production is a linear function of the weather forecast, this will be done in a general setting. Chapter 2 will enable us to model levels of the distribution of power production as a linear function of different meteorological forecasts $x_1 + ...x_p$. I.e. we can model quantiles of the power production y at level τ given our meteorological data by

$$\hat{F}^{-1}(\tau; x_1, ..., x_p) = \hat{Q}(x_1, ...x_p|\tau) = \alpha + \sum_j x_j\beta_j \tag{1.2}$$

Power production from a wind power plant is not a linear function of, e.g. the wind speed, and not surprisingly this is not the case for quantile levels in the distributions either. Therefore a combination of quantile regression and splines is used to model the the conditional distribution.

The spline function is the subject of Chapter 3, these will be treated in their own right, to give an over view of the properties of splines.

With the tools in place the presentation combine the two methods and uses quantile regression with splines to develop models for quantiles of prediction errors in wind energy production. The problem of evaluating the quality of a quantile forecast is not easy and many measures have been proposed the most common of these are discussed on the basis of a data set from a wind power plant located at Tunø Knob.

Adaptive models have proven to be effective for mean value prediction of wind power production. It is therefore obvious to attempt to do this for the quantile regression presented here as well. Such a procedure is set up in the last part of the presentation, and it have proven to be both fast and lead large to improvement of performance all the evaluated performance parameters.

Quantile estimation can give very unphysical estimates, especially a the setting where rare events need to be modeled, this is what we do here. Both high power production and high wind speeds are rare events. The estimates of the 75% quantile can e.g. in some situations be less than the 25% quantile, behavior like this is discussed in the last part of the presentation and some suggestions for solutions are proposed and discussed.

The report is divided in two parts, the first part cover the theory of quantile regression and splines, this is presented in its own right, with no special reference to wind power production. The second part of the report analyze a data set from Tunø Knob wind power plant. This analysis us used to motivate the development of the adaptive procedure and non crossing constraint. This dataset is further used for a discussion of performance parameters and their properties in general.

Part I

Theory: Quantile Regression and Splines

Chapter 2

Quantile Regression

2.1 Introduction

The most common statistic of datasets is the mean. The mean is found by minimizing a quadratic loss function. What is usually meant by linear regression is a linear model, with a quadratic loss function. Here linear regression will refer to a linear model with some loss function $\rho(r)$ from $\mathbb{R} \rightarrow \mathbb{R}_0$ with the following properties

$$\begin{aligned} \rho(0) &= 0 \qquad (2.1) \\ \rho(r_1) &< \rho(r_2) \quad \text{for} \quad |r_1| < |r_2|, \quad r_1 \cdot r_2 \geq 0 \qquad (2.2) \end{aligned}$$

The second condition ensure that the inequality only have to apply if r_1 and r_2 lie on the same half axis ($(-\infty, 0]$ or $[0, \infty)$). The aim in a linear regression is now to minimize the sum of the loss function, under the linear model.

The assumption in a linear model is that future observations of a response variable y_t can be written as a linear combination of observed (or forecasted) explanatory variables $\mathbf{x}_t$, where $\mathbf{x}_t \in \mathbb{R}^K$ is known, plus an error r_t. The model is

$$y_t = \mathbf{x}_t \hat{\beta} + r_t = \hat{y}_t + r_t \qquad (2.3)$$

given past observations of $\mathbf{y} = [y_i]$ and $\mathbf{X} = [\mathbf{x}_i^T]$ with $i = 1, ...N$, we can set up the observation equations

$$\mathbf{y} = \mathbf{X}\hat{\beta} + \mathbf{r} = \hat{\mathbf{y}} + \mathbf{r} \tag{2.4}$$

The matrix $\mathbf{X}$ is called the design matrix, the aim is now to estimate β s.t. the sum of loss functions $\rho(r_i)$ is minimized. The best estimate of β, with respect to this loss function is

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} \rho(r_i) = \arg\min_{\beta} S(\mathbf{r}) \tag{2.5}$$

If we use a quadratic loss function, then we have

$$S(\mathbf{r}) = \sum_{i=1}^{N} r_i^2 = \mathbf{r}^T\mathbf{r} \tag{2.6}$$

this leads to the conditional mean.

With the quadratic loss function we can write the best estimate of β as

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.7}$$

this is a nice and closed form for the estimates. The mean is often not the only statistic we would like to have on a set of random variables, since this does not really tell anything about randomness. We could now continue by estimating higher order moment of the distribution. If all moments are known, then the distribution is completely characterized. If we assume that data is normal distributed, then the mean and variance characterizes the distribution. If a distribution is completely characterized then we can calculate all quantiles.

Another approach is to find quantiles directly. If we know all the quantiles then the distribution is also completely characterized. This chapter deals with the idea of quantile regression. The main article of the subject seems to be Koenker and Basset's 1978 paper [2]. The idea is to replace the quadratic loss function with a piecewise linear, and asymmetric loss function, depending on the quantile we wish to estimate.

The idea of an using absolute loss function to find the sample median have been known before, but in this article the idea is generalized to other quantiles than the median.

For the mean estimator we could write down the parameters in a closed form, this is not the case when we go to the piecewise linear loss function. For this linear programming techniques are needed to find the best estimator.

This chapter give some of the fundamental definitions of quantile regression, and an introduction to linear programming in the quantile regression context. The linear programming formulation is useful from a implementation point of view, but it also gives the proof of some of the fundamental theorems of quantile regression.

2.2 Basic Definitions

The idea of quantile regression is to model the quantiles of a distribution directly, i.e. a regression of known variables. This offers an alternative to estimating conditional expectations and higher order moments. The focus is linear regression, which is the general linear model with an absolute loss function, but techniques for non-linear quantile regression have been develop (see e.g. [18]).

A regression quantile as presented in [2], is a linear regression with K explanatory variables

$$\begin{aligned}\hat{Q}(\tau;\mathbf{x}_t) &= \hat{\beta}_1(\tau)x_{1,t} + ... + \hat{\beta}_K(\tau)x_{K,t} \\ &= \mathbf{x}_t^T\hat{\beta}(\tau) \qquad (2.8)\end{aligned}$$

where $\hat{Q}(\tau)$ is the τ-quantile, $x_{1,t}$ would normally be constant equal to one s.t β_1 is an intercept. By introducing the loss

$$\rho_\tau(r) = \begin{cases} \tau r & , \quad r \geq 0 \\ (\tau - 1)r & , \quad r < 0 \end{cases} \tag{2.9}$$

where r is the residual i.e. $r_i = y_i - \hat{Q}(\tau;\mathbf{x}_i)$. The estimation of β is done by minimizing $\sum_i \rho(r_i)$ w.r.t. β, hereby we get the estimates

$$\hat{\beta}(\tau) = \arg\min_\beta \sum_{i=1}^N \rho_\tau(r_i) = \arg\min_\beta S(\beta;\tau,\mathbf{r}) \tag{2.10}$$

with the loss function $S(\beta)$

$$S(\beta) = \tau \sum_{r_i \geq 0} r_i + (\tau - 1)\sum_{r_i < 0} r_i = S_1(\beta) + S_2(\beta) \tag{2.11}$$

This is a linear optimization problem (LO), and gives the conditional τ-quantile. The proof of this fact is given through the linear programming formulations of the problem given in later sections.

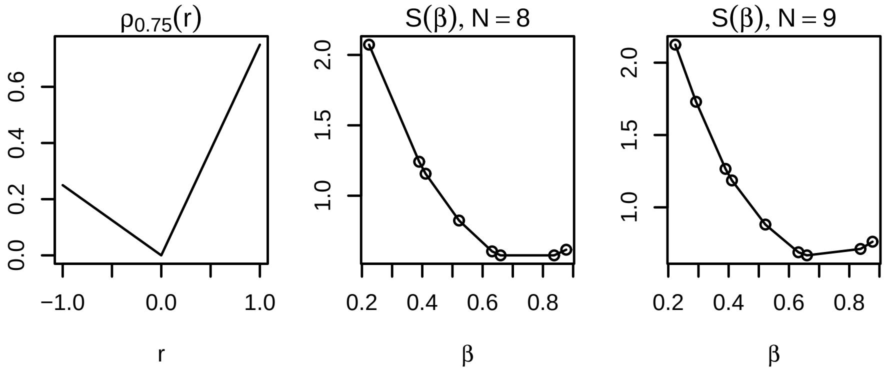

Figure 2.1: The figure show the asymmetric loss, when $\tau = 0.75$ and the loss function for 8 and 9 uniformly distributed numbers. Note that the optimum is not unique for $N = 8$, while we have a unique optimum for for $N = 9$, see also Example 2.2 for a treatment of the of the uniqueness property.

The definitions given so far are enough to show that (2.10) procedure a τ-quantile of a random sample, this is the subject of the next example.

Example 2.1 In (2.8) set $\mathbf{X} = \mathbf{e}$ this is the unconditional sample quantile, the parameter estimate is denoted $\hat{\beta} = \hat{\beta}_0$. We know that a sample quantile (it does not have to be unique) is $\hat{\beta}_0 = y_{(\lceil \tau N \rceil)}$, with $y_{(i)}$ being the order statistics, the loss function of (2.10) with $\hat{\beta}_0$ is

$$\begin{aligned}
S(\hat{\beta}_0;\tau) &= \tau \sum_{y_i \geq \hat{\beta}_0} r_i + (\tau - 1) \sum_{y_i < \hat{\beta}_0} r_i \\
&= \tau \sum_{i=\lceil \tau N \rceil}^{N} (y_{(i)} - y_{(\lceil \tau N \rceil)}) + (\tau - 1) \sum_{i=1}^{\lceil \tau N \rceil - 1} (y_{(i)} - y_{(\lceil \tau N \rceil)}) \\
&= \tau \sum_{i=1}^{N} (y_{(i)} - y_{(\lceil \tau N \rceil)}) - \sum_{i=1}^{\lceil \tau N \rceil - 1} (y_{(i)} - y_{(\lceil \tau N \rceil)}) \\
&= \tau \sum_{i=1}^{N} y_{(i)} - \sum_{i=1}^{\lceil \tau N \rceil - 1} y_{(i)} + (\lceil \tau N \rceil - 1 - \tau N) y_{(\lceil \tau N \rceil)}
\end{aligned}$$

$S(\hat{\beta};\tau)$ is a convex function, see also Figure 2.1 and 2.2, therefore in order to show that $\hat{\beta}_0 = y_{(\lceil \tau N \rceil)}$ is the optimal solution, it is enough to show that there exist two points $\beta_1 < \hat{\beta}_0$ and $\beta_2 > \hat{\beta}_0$ with $S(\hat{\beta}_1;\tau) \geq S(\hat{\beta}_0;\tau) \leq S(\hat{\beta}_2;\tau)$. To show this choose $\hat{\beta}_1 = y_{(\lceil \tau N \rceil - 1)}$ and $\hat{\beta}_1 = y_{(\lceil \tau N \rceil + 1)}$. In the same way as above we get

$$\begin{aligned}
S(\hat{\beta}_1;\tau) &= \tau\sum_{i=1}^{N} y_{(i)} - \sum_{i=1}^{\lceil\tau N\rceil-2} y_{(i)} + (\lceil\tau N\rceil - 2 - \tau N)y_{(\lceil\tau N\rceil)} \\
S(\hat{\beta}_2;\tau) &= \tau\sum_{i=1}^{N} y_{(i)} - \sum_{i=1}^{\lceil\tau N\rceil} y_{(i)} + (\lceil\tau N\rceil - \tau N)y_{(\lceil\tau N\rceil)}
\end{aligned}$$

Now look at the differences

$$\begin{aligned}
S(\hat{\beta}_1;\tau) - S(\hat{\beta}_0;\tau) &= y_{(\lceil\tau N\rceil-1)} + (\lceil\tau N\rceil - \tau N - 2)y_{(\lceil\tau N\rceil-1)} \\
&\quad -(\lceil\tau N\rceil - \tau N - 1)y_{(\lceil\tau N\rceil)} \\
&= (\lceil\tau N\rceil - \tau N - 1)(y_{(\lceil\tau N\rceil-1)} - y_{(\lceil\tau N\rceil)}) \\
&\geq 0
\end{aligned}$$

and

$$\begin{aligned}
S(\hat{\beta}_2;\tau) - S(\hat{\beta}_0;\tau) &= y_{(\lceil\tau N\rceil)} + (\lceil\tau N\rceil - \tau N)y_{(\lceil\tau N\rceil+1)} \\
&\quad -(\lceil\tau N\rceil - \tau N - 1)y_{(\lceil\tau N\rceil)} \\
&= (\lceil\tau N\rceil - \tau N)(y_{(\lceil\tau N\rceil+1)} - y_{(\lceil\tau N\rceil)}) \\
&\geq 0
\end{aligned}$$

This shows that the quantile regression produce the sample quantiles. □

The example shows that the quantile regression formulation produce the sample quantile. This is of course a minimal requirement for the quantile regression.

In the more general setting, let $\mathcal{J} = \{1, 2, ...N\}$ and let $\mathcal{H}$ denote the K-element subsets of $\mathcal{J}$, where K is the number of columns in X, let further $B^*(\tau)$ denote the set of solutions to the problem (2.10), finally set $H = \{h \in \mathcal{H}|\text{rank}X(h) = K\}$. Then the following theorem is due to [2]

Theorem 2.1 *If X has rank K then the set of regression quantiles, $B^*(\tau)$, has at least one element of the form,*

$$\beta^*(\tau) = \mathbf{X}(h)^{-1}\mathbf{y}(h) \tag{2.12}$$

for some $h \in H$. Moreover, $B^*(\tau)$ is the convex hull of all solutions having this form. If further the distribution function F of Y is continuous, then with probability one β^* is a unique solution if and only if

$$(\tau - 1)\mathbf{e}_K^T < \sum_{t\in\bar{h}}(\frac{1}{2}(1 - sign(y_t - x_t\beta^*)) - \tau)x_t\mathbf{X}(h)^{-1} < \tau\mathbf{e}_K^T \tag{2.13}$$

where $\mathbf{e}_K$ is a K-vector of ones and $\bar{h} = \mathcal{J} \setminus h$.

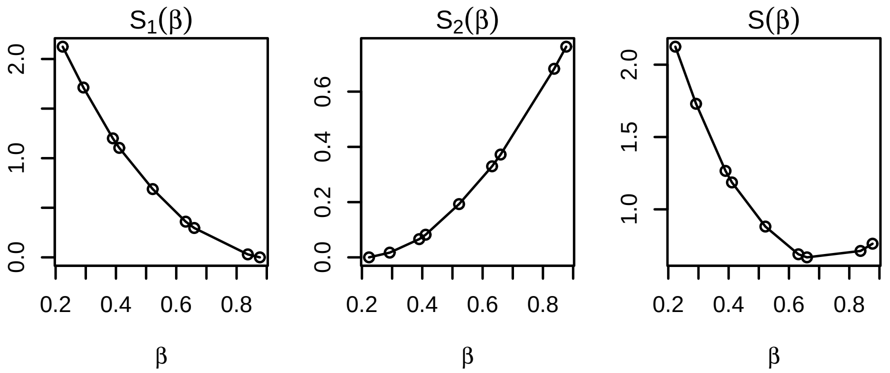

Figure 2.2: The figure show the loss function S and the two components S_1 and S_2, it is seen that they are all convex functions, the numbers used is the same as in Figure 2.1, with $N = 9$ (drawn from a uniform distribution).

The proof of Theorem 2.1 relay on the linear programming formulation of problem (2.10), which will be described in the next section and thereby provide a partial proof of the theorem. It is seen that the quantile regression interpolate K points of the pair $(\mathbf{X}, \mathbf{y})$, this was also shown in the sample quantile case in Example 2.1. A small example can again show the implication of the second part of the Theorem in the sample quantile case.

Example 2.2 As in example 2.1 look at the a random sample, we want to test if a solution to the quantile regression problem is unique. Example 2.1 shows that $\beta = y_{(\lceil \tau N \rceil)}$ is a optimum to the problem. Set $f(\beta) = \sum_{t \in \bar{h}} (\frac{1}{2}(1 - \text{sign}(y_t - x_t\beta)) - \tau) x_t \mathbf{X}(h)^{-1}$ then we get

$$\begin{aligned}
f(y_{(\lceil \tau N \rceil)}) &= \sum_{y_t < y_{(\lceil \tau N \rceil)}} (1-\tau) + \sum_{y_t > y_{(\lceil \tau N \rceil)}} (-\tau) \\
&= \sum_{t=1}^{\lceil \tau N \rceil - 1} (1-\tau) + \sum_{t=\lceil \tau N \rceil + 1}^{N} (-\tau) \\
&= (1-\tau)(\lceil \tau N \rceil - 1) - \tau(N - \lceil \tau N \rceil) \\
&= \lceil \tau N \rceil - \tau N - (1-\tau)
\end{aligned}$$

now $0 \leq \lceil \tau N \rceil - \tau N < 1$ so the theorem tells us that if $\lceil \tau N \rceil \neq \tau N$ then the solution to the problem is unique, if $\lceil \tau N \rceil = \tau N$ then we get $f(y_{(\lceil \tau N \rceil)}) = \tau - 1$ and the solution is not unique. This is also illustrated in Figure 2.1, for the sample quantile case. □

Now it has been establish that the quantile regression produce the sample quantile of a dataset, in this sense it seems promising, that it is somehow a generalizations of the sample quantile, i.e. that it is the conditional quantile.

2.3 Quantile Regression and Linear Programming

This section will describe the general setting of linear programming and write down the formulation for the specific case of quantile regression. As general references should be mentioned [14] and Chapter 6 of [3].

A linear programming problem consist of an objective function ($\mathbf{c}^T\mathbf{x}$) end a set of linear constraints ($\mathbf{A}\mathbf{x} = \mathbf{b}$ and $\mathbf{x} \geq \mathbf{0}$). In the so called standard form this is

$$(P_s) \quad \min\{\mathbf{c}^T\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \tag{2.14}$$

With $\mathbf{A} \in \mathbb{R}^{n\times m}$, a point $\mathbf{x}$ is called feasible if the constraints in (P_s) are met, the collection of all feasible points is called the feasible region and will be denoted $\mathcal{P}$. By reformulating (2.10) it is possible to bring the quantile regression to the form (2.14). We can write (2.10) as

$$\min\{\tau\mathbf{e}^T\mathbf{r}^+ + (1-\tau)\mathbf{e}^T\mathbf{r}^- : \mathbf{X}\beta + \mathbf{r}^+ - \mathbf{r}^- = \mathbf{y}, (\mathbf{r}^+, \mathbf{r}^-) \in \mathbb{R}_0^{2N}, \beta \in \mathbb{R}^K\}$$

with $r_i^+ = I(r_i \geq 0)r_i$ and $r_i^- = -I(r_i \leq 0)r_i$. In a more compact notation this is

$$\min\{\mathbf{c}^T\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{y}, (\mathbf{r}^+, \mathbf{r}^-) \in \mathbb{R}_0^{2N}, \beta \in \mathbb{R}^K\} \tag{2.15}$$

with

$$\mathbf{c} = \begin{bmatrix} \mathbf{0}_K \\ \tau\mathbf{e} \\ (1-\tau)\mathbf{e} \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} \beta \\ \mathbf{r}^+ \\ \mathbf{r}^- \end{bmatrix} \quad \mathbf{A} = [\mathbf{X}, \mathbf{I}, -\mathbf{I}] \tag{2.16}$$

A solution to problem (2.14) must lie in a vertex, i.e. at a point where the number of active constraints are equal to the dimension of $\mathbf{x}$. Or put another way $\mathbf{x}$ must be on the boundary of the feasible region.

$\mathbf{A}\mathbf{x} = \mathbf{y}$ give us N constraints in $2N + K$ unknown, so at least $N + K$ elements in the vector $[\mathbf{r}^+, \mathbf{r}^-]^T$ must be equal to zero (β can not be on the boundary). If $r_i^+ > 0$ then $r_i^- = 0$, since otherwise we can move an amount from one to the other without affecting the constraints, and at the same time improving the objective function. The vector $\mathbf{r} = \mathbf{r}^+ - \mathbf{r}^-$ will contain at least K zeros, let h be the index set s.t. $\mathbf{r}(h) = \mathbf{0}$ we can write

$$\mathbf{X}(h)\beta = \mathbf{y}(h) \tag{2.17}$$

If rank$\mathbf{X} = K$ then there exist an index set h with rank$\mathbf{X}(h) = K$, rank$\mathbf{X}$ should be K since otherwise one of the explanatory variables can be written as a linear combination of the other and the problem will be singular. With this first part of Theorem 2.1 is proved.

It is of course impossible to go through all these vertex points, since there will an extremely large number of these. In the context we use these later on, the dimension of $\mathbf{X}$ will be up to 10000×39 and the number of vertices is then $\binom{10000}{39} \sim 4 \cdot 10^{113}$. There are different ways to get faster to the optimal solution, in this presentation we will discuss the simplex method where the solution is iterated from vertex to vertex in the direction of a better objective function.

The simplex algorithm works because the objective function and the feasible region are both convex, this insures that a local optimum is also a global optimum. It should be mentioned that for large problems, one would normally use an interior point methods, where a penalty function is used to iterate through the interior of the feasible region to the optimal solution. This is also what is used by the statistical software "R".

Here the focus is on the simplex method because, this probably offers a more intuitively understanding of the problem, and more important the simplex method is considered superior if we have what is called a "*warm start*" in [14], i.e. we have a good guess on the solution. This will be used to develop an adaptive procedure for the quantile regression.

In [13] Koenker presents an algorithm for computing all quantiles of a distribution. This is done by first calculating the $\frac{1}{N}$ quantile (N being the sample size) and then step by step calculating all others quantiles, each of these requiring one simplex pivot. For large large sample sizes the number of different quantiles are very large a "typical" number being mentioned in the help function of the "R"-command "rq" is the order $N \log(N)$. Even with this in mind this should inspire to do something similar in an adaptive procedure for quantile regression.

Before going to the simplex algorithm we need some more background. In LP problems the so called dual problem plays a very important role. In the next section the dual problem is therefore explained and formulated for our quantile regression model. The dual problem is only used for analysis of the problem here.

2.3.1 The Dual Problem

Every LP problem have an associated dual problem, for a problem in the standard form (2.14) the dual problem is

$$(D) \quad \max\{\mathbf{b}^T\mathbf{z} : \mathbf{A}^T\mathbf{z} \leq \mathbf{c}\} \tag{2.18}$$

To see how we get this, look at the relaxed problem corresponding to (P_s)

$$(R) \quad \min\{\mathbf{c}^T\mathbf{x} + \mathbf{z}^T(\mathbf{b} - \mathbf{A}\mathbf{x}) : \mathbf{x} \geq \mathbf{0}\} \tag{2.19}$$

Here $\mathbf{z}$ is an arbitrary vector in $\mathbb{R}^m$, an optimal solution to (R) will always be less than or equal to the solution to the optimal solution to the primal problem, since otherwise we could just choose the optimal $\mathbf{x}^*$ as the solution and then get the same value of the objective function.

The relaxed problem therefore gives us a lower bound for the primal problem, what is then interesting is of course the maximal lower bound, so we want to solve

$$\max_{\mathbf{z}}\{\min_{\mathbf{x}}\{\mathbf{c}^T\mathbf{x} + \mathbf{z}^T(\mathbf{b} - \mathbf{A}\mathbf{x}) : \mathbf{x} \geq \mathbf{0}\}\} \tag{2.20}$$

Now rewrite the objective function of this problem as

$$\mathbf{c}^T\mathbf{x} + \mathbf{z}^T(\mathbf{b} - \mathbf{A}\mathbf{x}) = \mathbf{x}^T(\mathbf{c} - \mathbf{A}^T\mathbf{z}) + \mathbf{b}^T\mathbf{z} \tag{2.21}$$

If there exist any i s.t. $(\mathbf{c} - \mathbf{A}^T\mathbf{z})_i < 0$ then the minimization is easy, because then just let $(\mathbf{x})_i \to \infty$ and the objective function will be $-\infty$. To get a lower bound that is useful we therefore demand that $(\mathbf{c} - \mathbf{A}^T\mathbf{z})_i \geq 0$, but then the optimal value is $\mathbf{x} = \mathbf{0}$ and we get the dual problem (D).

Now we can write down the dual problem for the quantile regression problem. To make the steps clear we write the primal problem explicitly in the standard form, i.e. reformulate the variables and parameters in (2.15) as

$$\mathbf{c} = \begin{bmatrix} \mathbf{0}_{2K} \\ \tau\mathbf{e} \\ (1-\tau)\mathbf{e} \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} \beta^+ \\ \beta^- \\ \mathbf{r}^+ \\ \mathbf{r}^- \end{bmatrix} \quad \mathbf{A} = [\mathbf{X}, -\mathbf{X}, \mathbf{I}, -\mathbf{I}] \tag{2.22}$$

Then we have the standard form

$$(P_s) \quad \min\{\mathbf{c}^T\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{y}, \mathbf{x} \geq \mathbf{0}\} \tag{2.23}$$

This give the dual formulation

$$(D) \quad \max\{\mathbf{y}^T\mathbf{z} : \mathbf{A}^T\mathbf{z} \leq \mathbf{c}\} \tag{2.24}$$

but since

$$\mathbf{A}^T = \begin{bmatrix} & \mathbf{X}^T \\ - & \mathbf{X}^T \\ & \mathbf{I} \\ - & \mathbf{I} \end{bmatrix} \tag{2.25}$$

This can immediately be rewritten as

$$(D) \quad \max\{\mathbf{y}^T\mathbf{z} : \begin{bmatrix} & \mathbf{X}^T \\ - & \mathbf{X}^T \end{bmatrix} \mathbf{z} \leq \mathbf{0}, \mathbf{z} \in [\tau - 1, \tau]^N\} \tag{2.26}$$

which is the same as

$$(D) \quad \max\{\mathbf{y}^T\mathbf{z} : \mathbf{X}^T\mathbf{z} = \mathbf{0}, \mathbf{z} \in [\tau - 1, \tau]^N\}$$

The following theorem, which tells us about the existence of a solution, is due to [14]

Theorem 2.2 *For a given pair (P) and (D) there are three alternatives*

1. *Both (P) and (D) are feasible and bounded and there exist a strictly complementary optimal pair $(\tilde{\mathbf{x}} \in \mathcal{P}^*, \tilde{\mathbf{z}} \in \mathcal{D}^*)$ with $\mathbf{c}^T\tilde{\mathbf{x}} = \mathbf{b}^T\tilde{\mathbf{z}}$*
2. *Either (P) or (D) is unbounded and the other is infeasible.*
3. *Both (P) and (D) are infeasible.*

The proof of this will not be given for the general case here for this the reader is refereed to [14]. The proof of the quantile regression case will be given later on when the notation of the simplex method is establish.

The dual formulation (2.26) above is clearly bounded and feasible, to see that it is feasible just set $\mathbf{z} = \mathbf{0}$. So we are in situation one, and the optimal solution to the dual problem have the same solution as the optimal solution to the primal problem. So to prove Theorem 2.2 we just have to find the complementary pair.

To explain the complementary property we need to reformulate the problem (2.18) as

$$(D) \quad \max\{\mathbf{y}^T\mathbf{z} : \mathbf{A}^T\mathbf{z} + \mathbf{s} = \mathbf{c}, \mathbf{s} \geq \mathbf{0}\} \tag{2.27}$$

the vector $\mathbf{s}$ is called the *surplus vector* and the solution is said to be strictly complementary if it satisfy

$$\mathbf{x}^* \odot \mathbf{s}^* = \mathbf{0} \quad \text{and} \quad \mathbf{x}^* + \mathbf{s}^* > \mathbf{0} \tag{2.28}$$

This gives a partition $(\mathcal{B}, \mathcal{C})$ of the index set $\{1, ..., K + 2N\} = \Omega$, with $(\mathcal{B}, \mathcal{C})$ defined by

$$\mathcal{B} = \{j | x_j^* > 0\} \quad \mathcal{C} = \{j | s_j^* > 0\} \tag{2.29}$$

The splitting $(\mathcal{B}, \mathcal{C})$ can be found by using the simplex algorithm, this is the subject of the next section. If we have $\mathcal{B}$ then we can get the index set h and thereby $\hat{\beta}$ directly.

Note that h refer to rows of $\mathbf{X}$ while $\mathcal{B}$ and $\mathcal{C}$ refer to columns of $\mathbf{A}$, the connection is that if $i \wedge i + N \in \mathcal{C}$ then $i - K \in h$.

2.3.2 The Simplex Method

The idea of the simplex algorithm is to move through the vertices in an intelligent way. I.e. always move in the direction of a vertex with a better objective function. As stated this works because the objective function is convex and the feasible region is also a convex set.

The simplex algorithm assumes that we are at a vertex, so we have to get some method for getting to a vertex before starting the simplex algorithm, here we assume that we have such a solution. It is not important for the proofs how we get this and when we use the simplex method to develop an adaptive quantile regression method this solution is found with an interior point method algorithm in “R”.

Following [14] we define $\mathbf{B} = \mathbf{A}_{:\mathcal{B}}$ and $\mathbf{C} = \mathbf{A}_{:,\mathcal{C}}$ to easy notation. With such a splitting we can write down the constraints as

$$\mathbf{A}\mathbf{x} = \mathbf{B}\mathbf{x}_{\mathcal{B}} + \mathbf{C}\mathbf{x}_{\mathcal{C}} = \mathbf{y} \tag{2.30}$$

It have already been shown that, at a vertex $\mathbf{x}^{(k)}$ there exist a splitting $(\mathcal{B}, \mathcal{C})$ of the index set s.t.

$$\text{rank}(\mathbf{A}_{\mathcal{B}}) = m \tag{2.31}$$

$$\mathbf{x}_{\mathcal{C}}^{(k)} = \mathbf{0} \tag{2.32}$$

$$\mathbf{x}_{\mathcal{B}}^{(k)} = \mathbf{A}_{\mathcal{B}}^{-1}\mathbf{y} \geq \mathbf{0} \tag{2.33}$$

In the following it is assumed that $\mathcal{B}$ and $\mathcal{C}$ is ordered s.t. $\mathcal{B}(1) < \mathcal{B}(2) < ... < \mathcal{B}(N)$ and $\mathcal{C}(1) < \mathcal{C}(2) < ... < \mathcal{C}(N+K)$.

The objective function can now be written as

$$\begin{aligned}
\mathbf{c}^T\mathbf{x}^{(k)} &= \mathbf{c}_{\mathcal{B}}^T\mathbf{x}_{\mathcal{B}}^{(k)} + \mathbf{c}_{\mathcal{C}}^T\mathbf{x}_{\mathcal{C}}^{(k)} && (2.34)\\
&= \mathbf{c}_{\mathcal{B}}^T(\mathbf{x}_{\mathcal{B}}^{(k)} - \mathbf{B}^{-1}\mathbf{C}\mathbf{x}_{\mathcal{C}}^{(k)}) + \mathbf{c}_{\mathcal{C}}^T\mathbf{x}_{\mathcal{C}}^{(k)} && (2.35)\\
&= \mathbf{c}_{\mathcal{B}}^T\mathbf{x}_{\mathcal{B}}^{(k)} + (\mathbf{c}_{\mathcal{C}}^T - \mathbf{c}_{\mathcal{B}}^T\mathbf{B}^{-1}\mathbf{C})\mathbf{x}_{\mathcal{C}}^{(k)} && (2.36)\\
&= \mathbf{c}_{\mathcal{B}}^T\mathbf{x}_{\mathcal{B}}^{(k)} + (\mathbf{c}_{\mathcal{C}}^T - (\mathbf{C}^T(\mathbf{B}^{-T}\mathbf{c}_{\mathcal{B}})^T)^T)\mathbf{x}_{\mathcal{C}}^{(k)} && (2.37)\\
&= \mathbf{c}_{\mathcal{B}}^T\mathbf{x}^{(k)} + \mathbf{d}^T\mathbf{x}_{\mathcal{C}}^{(k)} && (2.38)
\end{aligned}$$

Since $\mathbf{x}_{\mathcal{C}} \geq \mathbf{0}$ we can not decrease the objective function if $\mathbf{d} \geq \mathbf{0}$ and $\mathbf{x}^{(k)}$ is therefore the optimal solution. From (2.37) we see that $\mathbf{d}$ is given by

$$\mathbf{d} = \mathbf{c}_{\mathcal{C}} - \mathbf{C}^T\mathbf{g} \quad ; \quad \mathbf{g} = \mathbf{B}^{-T}c_{\mathcal{B}} \tag{2.39}$$

If $\mathbf{x}^{(k)}$ is not optimal then choose a negative element d_s in $\mathbf{d}$, and change one element $(\mathbf{x}_{\mathcal{C}})_s$, while keeping the other elements of $\mathbf{x}_{\mathcal{C}}$ at zero. The basic variables are changed in direction $\mathbf{h} = \mathbf{B}^{-1}\mathbf{C}_{:,s}$, $\mathbf{x}_{\mathcal{B}}$ is changed in this direction until we meet a new vertex. This amount is given by $\alpha = \min\{\sigma_1, ..., \sigma_m\}$ with

$$\sigma_j = \begin{cases} (\mathbf{x}_{\mathcal{B}}^{(k)})_j/h_j & \text{if} \quad h_j > 0 \\ \infty & \text{if} \quad h_j \leq 0 \end{cases} \tag{2.40}$$

if $\alpha = \infty$ then the problem is unbounded, this can as stated in the Theorem 2.2 and the discussion thereafter not happen in our case. α can be zero and then the objective function is not improved, the step should be taken anyway since we could move to a position with a decent direction and an $\alpha > 0$. $\mathbf{x}_{\mathcal{B}}$ is now changed in two steps

$$\begin{aligned}
\mathbf{x}_{\mathcal{B}}^{(k+1)} &= \mathbf{x}_{\mathcal{B}}^{(k)} - \alpha\mathbf{h} && (2.41)\\
(\mathbf{x}_{\mathcal{B}}^{(k+1)})_q &= \alpha && (2.42)
\end{aligned}$$

where $q = \arg\min_j \sigma_j$. Further the q'th element of $\mathcal{B}$ is swapped with the sth of $\mathcal{C}$ and the algorithm starts over again. If α is zero, then it can happen that we move back and forth between two vertices with equal objective functions, so we should keep track where we have been and then be sure not to go back.

The expensive part of the simplex algorithm is to calculate the inverse of $\mathbf{B}$ (this should be done by a solve algorithm). However in the special case of quantile regression it is possible to write $\mathbf{B}^{-1}$ as products of known matrices and the inverse of $\mathbf{X}(h)$. To see this write down $\mathbf{B}$ as

$$\mathbf{B} = \begin{bmatrix} \mathbf{X}(h) & \mathbf{0} \\ \mathbf{X}(\bar{h}) & \mathbf{P} \end{bmatrix} \tag{2.43}$$

This is possible because we know that $\mathbf{X}(h)\hat{\beta} = \mathbf{y}(h)$ when we are at a vertex and therefore $\mathbf{r}(h) = \mathbf{0}$, $r_i^+ > 0 \Rightarrow r_i^- = 0$; $r_i^- > 0 \Rightarrow r_i^+ = 0$. Further we have $\mathbf{B}\mathbf{x}_{\mathcal{B}} = \mathbf{X}\hat{\beta} + \text{sign}(\mathbf{r}(\bar{h})) \odot |\mathbf{r}|$ so in summary we have

$$\mathbf{X}(h)\hat{\beta} = \mathbf{y}(h); \quad \mathbf{X}(\bar{h})\hat{\beta} + \text{sign}(\mathbf{r}(\bar{h})) \odot |\mathbf{r}(\bar{h})| = \mathbf{y}(\bar{h}) \tag{2.44}$$

We can therefore get (2.43) by interchanging rows in $\mathbf{B}$. $\mathbf{P}$ is a diagonal matrix with the diagonal elements $\text{sign}(r_i) + I(r_i = 0)$, $i \in \bar{h}$ implying $\mathbf{P} = \mathbf{P}^{-1}$. Now write $\mathbf{B}^{-1}$ as

$$\mathbf{B}^{-1} = \begin{bmatrix} \mathbf{B}_{11}^{-1} & \mathbf{B}_{12}^{-1} \\ \mathbf{B}_{21}^{-1} & \mathbf{B}_{22}^{-1} \end{bmatrix} \tag{2.45}$$

with $\mathbf{B}_{11}^{-1} \in \mathbb{R}^{K \times K}$, $\mathbf{B}_{21}^{-1} \in \mathbb{R}^{(N-K) \times K}$, $\mathbf{B}_{12}^{-1} \in \mathbb{R}^{K \times (N-K)}$ and $\mathbf{B}_{11}^{-1} \in \mathbb{R}^{(N-K) \times (N-K)}$. To find $\mathbf{B}^{-1}$, we have to solve the equations

$$\begin{aligned} \mathbf{X}(h)\mathbf{B}_{11}^{-1} + \mathbf{0}\mathbf{B}_{21}^{-1} &= \mathbf{I} & (2.46) \\ \mathbf{X}(h)\mathbf{B}_{12}^{-1} + \mathbf{0}\mathbf{B}_{22}^{-1} &= \mathbf{0} & (2.47) \\ \mathbf{X}(\bar{h})\mathbf{B}_{11}^{-1} + \mathbf{P}\mathbf{B}_{21}^{-1} &= \mathbf{0} & (2.48) \\ \mathbf{X}(\bar{h})\mathbf{B}_{12}^{-1} + \mathbf{P}\mathbf{B}_{22}^{-1} &= \mathbf{I} & (2.49) \end{aligned}$$

Which immediately give

$$\begin{aligned} \mathbf{B}_{11}^{-1} &= \mathbf{X}(h)^{-1} & (2.50) \\ \mathbf{B}_{12}^{-1} &= \mathbf{0} & (2.51) \\ \mathbf{B}_{21}^{-1} &= -\mathbf{P}\mathbf{X}(\bar{h})\mathbf{X}(h)^{-1} & (2.52) \\ \mathbf{B}_{22}^{-1} &= \mathbf{P} & (2.53) \end{aligned}$$

This is is very important, because in this context the matrix $\mathbf{B}$ is quite large, with several thousand elements in each direction. The formulation above make it possible to calculate the inverse of $\mathbf{B}$ by calculating the inverse of $\mathbf{X}(h)$ and then using matrix multiplication and element wise multiplication with the vector $\mathbf{p} = \text{diag}(\mathbf{P})$. In this presentation the number of elements in h will always be less than 40, so even if (and we would have to do so) we use the fact that $\mathbf{B}$ is sparse there are very large improvements w.r.t. calculation time and numerical stability compared with a standard sparse function. Further the implementation does not require sparse functions and does therefore not require sparse functions in the software.

The algorithm described in Practical Summary 2.1 is the same as the one described in [14], but the specialties w.r.t. quantile regression is used. The algorithm assumes that we are at a vertex i.e. we have $\mathbf{x}_{\mathcal{B}} = \mathbf{B}^{-1}\mathbf{y}$ and $\mathbf{x}_{\mathcal{C}} = \mathbf{0}$. An overview of the algorithm is given first and then we go through each step below.

Practical Summary 2.1 *The simplex algorithm*

1. *Compute* $\mathbf{d}$ *if* $\mathbf{d} \geq \mathbf{0}$ *stop* $\mathbf{x}$ *is optimal.*
   Otherwise choose s *s.t.* $d_s < 0$

2. *Compute* $\mathbf{h} = \mathbf{B}^{-1}\mathbf{A}_{:,\mathcal{C}(s)}$ *if* $\mathbf{h} \leq \mathbf{0}$*, stop the problem is unbounded.*

3. *Compute* α *and choose* q *such that* $\sigma_q = \alpha$

4. *Swap* $\mathcal{B}(q)$ *and* $\mathcal{C}(s)$ *and set* $\mathbf{x}_{\mathcal{B}} := \mathbf{x}_{\mathcal{B}} - \alpha\mathbf{h}$*;* $(\mathbf{x}_{\mathcal{B}})_q := \alpha$

Step one

The number of directions to calculate is equal to the number of elements in $\mathcal{C}$, this number is $N + K$, but the only directions which can be decent directions in the case of quantile regression is elements corresponding to h. It is therefore sufficient to examine $2K$ directions, corresponding to moving elements of $\mathbf{r}(h)$ in a positive or negative direction.

$\mathbf{d}$ is given by $\mathbf{d} = \mathbf{c}_{\mathcal{C}} - \mathbf{C}^T\mathbf{g}$ with $\mathbf{g} = \mathbf{B}^{-T}\mathbf{c}_{\mathcal{B}}$. If we have ordered $\mathbf{C}$ in the same way as $\mathbf{B}$ then the structure will be

$$\mathbf{C} = \begin{bmatrix} \mathbf{I}_K & -\mathbf{I}_K & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{P} \end{bmatrix} = \begin{bmatrix} \mathbf{P}_C & \mathbf{0} \\ \mathbf{0} & -\mathbf{P} \end{bmatrix} \tag{2.54}$$

Therefore we only need the first K elements of $\mathbf{g}$, these are

$$\begin{aligned} \mathbf{g}(h) &= \mathbf{B}_{1:K}^{-T}\mathbf{c}_{\mathcal{C}} && (2.55)\\ &= [\mathbf{X}(h)^{-T} \quad - (\mathbf{P}\mathbf{X}(\bar{h})\mathbf{X}(h)^{-1})^T]\mathbf{c}_{\mathcal{B}} && (2.56)\\ &= -\mathbf{X}(h)^{-T}\mathbf{X}(\bar{h})^T\mathbf{P}\rho_\tau(\mathrm{sign}(\mathbf{r}(\bar{h}))) && (2.57) \end{aligned}$$

this gives

$$\mathbf{d} = \begin{bmatrix} \tau\mathbf{e}_K - \mathbf{g}(h) \\ (1-\tau)\mathbf{e}_K + \mathbf{g}(h) \end{bmatrix} \tag{2.58}$$

This gives us at most K decent directions, since $\mathbf{d}_{K+1:2K} = \mathbf{e}_K - \mathbf{d}_{1:K}$ so if $d_1 < 0$ then $d_{K+1} > 1$, and in the optimal solution we have $0 \leq \mathbf{d} \leq 1$. This is actually the proof of the second part of Theorem 2.1, to see this rewrite $\mathbf{g}$

$$\begin{aligned}
\mathbf{g}^T(h) &= -\mathbf{P}\rho_\tau(\mathbf{r}(\bar{h}))^T\mathbf{X}(\bar{h})\mathbf{X}(h)^{-1} \\
&= -\mathbf{p}\odot\rho_\tau(\mathbf{r}(\bar{h}))^T\mathbf{X}(\bar{h})\mathbf{X}(h)^{-1} \\
&= \text{sign}\big(\mathbf{r}(\bar{h})\big)\odot\left(\frac{1}{2}\Big(\mathbf{e}-\text{sign}\big(\mathbf{r}(\bar{h})\big)\Big)+\tau\text{sign}\big(\mathbf{r}(\bar{h})\big)\right)^T\mathbf{X}(\bar{h})\mathbf{X}(h)^{-1} \\
&= \frac{1}{2}\Big(-\text{sign}\big(\mathbf{r}(\bar{h})+\mathbf{e}\big)-\tau\mathbf{e}\Big)^T\mathbf{X}(\bar{h})\mathbf{X}(h)^{-1} \\
&= \sum_{t\in\bar{h}}\left(\frac{1}{2}\Big(1-\text{sign}(\mathbf{r}_t)\Big)-\tau\right)\mathbf{x}_t\mathbf{X}(h)^{-1}
\end{aligned}$$

The demand $\mathbf{0}\leq\mathbf{d}\leq\mathbf{e}$ give the result from Theorem 2.1.

There are different ways to choose s one approach is to choose s as the steepest direction. Since $\mathbf{h}$ and therefore α depends on the direction we go in, this does not guarantee that we get the greatest improvement in the objective function.

Another approach is to compute $\mathbf{h}$ for all the decent directions, this is possible here because we can only have a very limited number of decent directions. This is the approach chosen here, we compute the improvements $(\alpha_s d_s)$ in all the decent direction and then choose the best direction.

Step one is completed by choosing $\tilde{s}=\{s|d_s<0\}$ and $\mathbf{P}_s=\mathbf{P}_{C,:,\tilde{s}}$.

Step two

Compute $\mathbf{h}$ according to the strategy in step one, i.e. $\mathbf{h}$ will be a matrix with each column being equal to one of the first K columns in $\mathbf{B}^{-1}$ times the sign of the residual produced by going in this direction. Set

$$\tilde{s}_i=\begin{cases}\tilde{s}_i & \text{if}\quad \tilde{s}_i\leq K\\ \tilde{s}_i-K & \text{if}\quad \tilde{s}_i>K\end{cases} \tag{2.59}$$

and $\mathbf{h}=\mathbf{B}^{-1}_{:,\tilde{s}}\mathbf{P}_s$.

Step three

The choice of α is done to prevent any of the variables to pass to the infeasible region, in our case this means that a residual can not change sign without passing through the index set h.

It should however be noted that an algorithm allowing residuals to change signs without passing through h could be implemented. We should just keep track of which of the residuals change sign, and the stop criterium should then be choose α s.t. the improvements in the objective function is maximized.

The improvements can in this case not be calculated directly from αd_s any more, since we have to take into account that the vector $\mathbf{c}_{\mathcal{B}}$ will now change every time we let a residual pass through zero.

To see how this would work define $\tilde{\mathbf{c}}_{\mathcal{B}} = \rho(-\mathbf{r}(\bar{h}))$, so $\tilde{\mathbf{c}}_{\mathcal{B}}$ is the loss when residuals is moved through zero. The decent direction was $\mathbf{h}$, the gain in the loss function is equal to $-\alpha \mathbf{h}\mathbf{c}_{\mathcal{B}} + \mathbf{c}_{\mathcal{C}(s)} = \alpha d_s$ when $\alpha = \sigma_{(1)}$. Let $\mathbf{q}$ be a index set s.t. $\sigma_{\mathbf{q}}$ is the order statistics of σ. Setting α equal to $\sigma_{(2)}$ correspond to moving one residual through zero, or equivalent to change $\mathbf{c}_{\mathcal{B}(\mathbf{q}_1)}$ to $-\tilde{\mathbf{c}}_{\mathcal{B}(\mathbf{q}_1)}$ (the minus is due to the fact that the direction of the gain in the loss function is reversed here), this change is equivalent to subtracting one from $\mathbf{c}_{\mathcal{B}(\mathbf{q}_1)}$. If we denote the gain in the loss function by moving in the direction $\mathbf{h}$ an amount σ_j by Lg_j, then we get the following recursive formula for the gain

$$\begin{aligned}
Lg_1 &= \sigma_{(1)}(\mathbf{h}\mathbf{c}_{\mathcal{B}} + \mathbf{c}_{\mathcal{C}(s)}) && (2.60)\\
&= \sigma_{(1)} d_s && (2.61)\\
Lg_2 &= Lg_1 + (\sigma_{(2)} - \sigma_{(1)})(\mathbf{h}\mathbf{c}_{\mathcal{B}} + \mathbf{h}(\mathbf{q}_1) + \mathbf{c}_{\mathcal{C}(s)}) && (2.62)\\
&= Lg_1 + (\sigma_{(2)} - \sigma_{(1)})(d_s + \mathbf{h}(\mathbf{q}_1)) && (2.63)\\
Lg_3 &= Lg_2 + (\sigma_{(3)} - \sigma_{(2)})(\mathbf{h}\mathbf{c}_{\mathcal{B}} + \mathbf{h}(\mathbf{q}_1) + \mathbf{h}(\mathbf{q}_2) + \mathbf{c}_{\mathcal{C}(s)}) && (2.64)\\
&= Lg_3 + (\sigma_{(2)} - \sigma_{(1)})(d_s + \mathbf{h}(\mathbf{q}_1) + \mathbf{h}(\mathbf{q}_2)) && (2.65)\\
&\vdots\\
Lg_j &= Lg_{j-1} + (\sigma_{(j)} - \sigma_{(j-1)})(d_s + \sum_{l=1}^{j-1} \mathbf{h}(\mathbf{q}_l)); \quad j > 1 && (2.66)
\end{aligned}$$

so the gain in the loss function will be better and better as long as $-d_s > \sum_{l=1}^{j-1} \mathbf{h}(\mathbf{q}_l)$. In this way we can skip some simplex steps and this can be done just by summing up a vector, which is of course much cheaper than inverting the matrices. As stated above this is not implemented, therefore how much this would save in calculation time is not examined. ■

This conclude the simplex steps and we are ready to go back to step one. The next example illustrate the technique of the simplex method applied to quantile regression, again in the sample quantile case.

Example 2.3 Assume that we want to find the sample median of $\mathbf{y} = [1, 2, 3, 4, 5, 6]^T$, we have the design matrix $\mathbf{X} = \mathbf{e}_6$ and $c_{\mathcal{B}} = \frac{1}{2}[0, \mathbf{e}_5^T]^T$, $c_{\mathcal{B}}$ is constant in this case since the loss function for the median regression is symmetric around zero. If we set out with $h = 1$ then $\bar{h} = \{1, ..., 6\} \setminus 1 = \{2, 3, 4, 5, 6\}$ we have

$$\begin{array}{rclcrcl} \mathbf{X}(h) & = & 1 & ; & \mathbf{X}(\bar{h}) & = & \mathbf{e}_5 \\ \mathbf{y}(h) & = & 1 & ; & \mathbf{y}(\bar{h}) & = & [2, 3, 4, 5, 6]^T \end{array} \quad (2.67)$$

This gives $\hat{\beta} = \mathbf{X}(h)^{-1}\mathbf{y}(h) = 1^{-1}1 = 1$, this gives

$$\mathbf{r}(\bar{h}) = \mathbf{y}(\bar{h}) - \mathbf{e}\beta = [2, 3, 4, 5, 6]^T - \mathbf{e}_5 = [1, 2, 3, 4, 5]^T \quad (2.68)$$

and so $\mathbf{p} = \text{diag}(\mathbf{P}) = \text{sign}(\mathbf{r}(\bar{h})) = \mathbf{e}_5^T$, $\mathbf{x}_{\mathcal{B}} = [1, 1, 2, 3, 4, 5]^T$, the objective function is $\frac{1}{2}\sum_{i=1}^{5} i = 7.5$ and

$$\mathbf{B}^{-1} = \left[\begin{array}{cc} \mathbf{X}(h)^{-1} & \mathbf{0} \\ -\mathbf{P}\mathbf{X}(\bar{h})\mathbf{X}(h)^{-1} & \mathbf{P} \end{array} \right] = \left[\begin{array}{cc} 1 & \mathbf{0} \\ -\mathbf{e}_5 & \mathbf{I} \end{array} \right] \quad (2.69)$$

Now we can find $\mathbf{g} = -\mathbf{X}(h)^{-T}\mathbf{X}(\bar{h})^T\mathbf{P}\mathbf{c}_{\mathcal{B}}(\bar{h}) = -1 \cdot \mathbf{e}_5^T \frac{1}{2}\mathbf{e}_5 = -\frac{1}{2} \cdot 5$ and

$$\mathbf{d} = \left[\begin{array}{c} \tau - \mathbf{g} \\ 1 - \tau + \mathbf{g} \end{array} \right] = \left[\begin{array}{c} 0.5 + 2.5 \\ 0.5 - 2.5 \end{array} \right] = \left[\begin{array}{c} 3 \\ -2 \end{array} \right] \quad (2.70)$$

This conclude step one.

The next step is now to find $\mathbf{h} = \mathbf{B}^{-1}\mathbf{C}_{:,2} = \mathbf{B}^{-1}[-1, \mathbf{0}_5^T]^T = -\mathbf{B}_{:,1}^{-1} = [-1, \mathbf{e}_5^T]^T$.

With this we get $\sigma = [\infty, 1, 2, 3, 4, 5]$ and $\alpha = 1$

Therefore we have $q = 2$ and we set $h = 2$, $\mathbf{x}_{\mathcal{B}} = [2, 1, 1, 2, 3, 4]^T$ the final step is to change $\mathbf{p}_1$ from 1 to -1 (because the decent direction was d_2). We are now ready to begin the next iteration, this is completely similar and will not be done here. The objective function was decreased to 5.5, in the first step.

If we use equation (2.66) then we get

$$\begin{array}{rcl} Lg_1 & = & \sigma_{(1)}d_s \\ & = & -2 \\ Lg_2 & = & Lg_1 + (\sigma_{(2)} - \sigma_{(1)})(d_s + \mathbf{h}(\mathbf{q}_1)) \\ & = & -2 + (2 - 1)(-2 + 1) \\ & = & -3 \\ Lg_3 & = & Lg_2 + (\sigma_{(3)} - \sigma_{(2)})(d_s + \mathbf{h}(\mathbf{q}_1) + \mathbf{h}(\mathbf{q}_2)) \\ & = & -3 + (3 - 2)(-2 + 1 + 1) \\ & = & -3 \\ Lg_4 & = & Lg_4 + (\sigma_{(3)} - \sigma_{(2)})(d_s + \mathbf{h}(\mathbf{q}_1) + \mathbf{h}(\mathbf{q}_2) + \mathbf{h}(\mathbf{q}_3)) \\ & = & -3 + (3 - 2)(-2 + 1 + 1 + 1) \\ & = & -2 \end{array}$$

which tells us to choose $\alpha = \sigma_{(3)}$ or $\alpha = \sigma_{(4)}$ corresponding to $h = 3$ or 4. So in this case the procedure takes us directly to the sample median in one step, with the loss function for the optimal solution being $7.5 - 3 = 4.5$. □

With the constructional description of the the quantile regression in place we go and study properties of the quantile regression.

2.4 Properties of Quantile Regression

The previous sections showed how to get the regression quantiles. It have been shown that in the case of the sample quantiles, the regression method introduced produce the sample quantile. We have not seen that this technique actually produce a quantile in the sense that the proportion of observations below the hyper plane produced by the regression is close to τN. That it does so should of course be the case, fortunately this is also the case. The next theorem state this and the condition needed to ensure this, the theorem is due to [3]

Theorem 2.3 *Let P_q, N_q and Z_q denote the proportion of positive, negative and zero elements of the residual vector $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta(\tau)$. If $\mathbf{X}$ contains an intercept, that is, if there exist $\alpha \in \mathbb{R}^K$ s.t. $\mathbf{X}\alpha = \mathbf{e}_N$, then for any $\hat{\beta}(\tau)$, solving (2.10) we have*

$$N_q \leq N\tau \leq N_q + Z_q \quad \textit{and} \quad P_q \leq N(1-\tau) \leq P_q + Z_q \tag{2.71}$$

The loss function defined in by (2.10) is not differentiable, at the points where one or more of the residuals r_i is zero. Therefore the directional derivative in direction w, have to be defined before we can go on to the the proof of the theorem. The directional derivatives is defined by

$$\begin{aligned}
\nabla S(\beta, \mathbf{w}) &= \frac{d}{dt} S(\beta + t\mathbf{w})\big|_{t=0} \\
&= \frac{d}{dt} \sum_{i=1}^{N} (y_i - \mathbf{x}_i^T\beta - \mathbf{x}_i^T t\mathbf{w})[\tau - I(y_i - \mathbf{x}_i^T\beta - \mathbf{x}_i^T t\mathbf{w})]\big|_{t=0} \\
&= -\sum_{i=1}^{N} \psi_\tau^*(y_i - \mathbf{x}_i^T\beta, -\mathbf{x}_i^T\mathbf{w})\mathbf{x}_i^T\mathbf{w} \qquad (2.72)
\end{aligned}$$

with

$$\psi_\tau^*(u, v) = \begin{cases} \tau - I(u < 0) & if \quad u \neq 0 \\ \tau - I(v < 0) & if \quad u = 0 \end{cases} \tag{2.73}$$

With this we are ready for the proof of Theorem 2.3

Proof. The condition for optimality is that $\nabla S(\beta, \mathbf{w}) \geq 0$ for all $\mathbf{w} \in \mathbb{R}^K$, therefore we can also choose $\mathbf{w} = \alpha$ and get

$$\begin{aligned} \nabla S(\beta, \alpha) &= -\sum_{i=1}^{N} \psi_\tau^*(y_i - \mathbf{x}_i^T \beta, -1) && (2.74) \\ &= -\tau P_q + (1-\tau) N_q + (1-\tau) Z_q \geq 0 && (2.75) \end{aligned}$$

and with $\mathbf{w} = -\alpha$ we get

$$\begin{aligned} \nabla S(\beta, -\alpha) &= \sum_{i=1}^{N} \psi_\tau^*(y_i - \mathbf{x}_i^T \beta, 1) && (2.76) \\ &= \tau P_q - (1-\tau) N_q + \tau Z_q \geq 0 && (2.77) \end{aligned}$$

with $P_q = N - N_q - Z_q$ and $N_q = N - P_q - Z_q$ these inequalities give the two conditions in the theorem and the proof is completed. □

A remark here is that by a re-parameterization of $\mathbf{w}$ s.t. $\mathbf{w} = \mathbf{X}(h)^{-1}\mathbf{v}$ and checking the directions $\pm e_j$ $j = 1, ...K$, where e_j is a vector of zeros except the k'th which is one, the demand on the directional derivatives will give the demands on the simplex vector $\mathbf{d}$ (see section 2.3.2) and thereby a proof of Theorem 2.1.

2.4.1 Quantile Crossings

We have now seen how to construct the regression quantiles and that these divide the data space in a manner that should be required by a quantile. The regression quantiles have however a quite serious problem, this is the problem of quantile crossings. These can occur because we model the quantiles individually. Koenker note in [3] that if our explanatory variables are assumed to take values in $\mathbb{R}$ then the only way to avoid these is to let all regression lines be parallel and only move the intercept, but this is quite restrictive and not really what we would expect here.

The hope is now that quantile crossings will only appear at remote areas of the data space, in [3] Koenker note that a significant number of crossings can be taken as evidence of misspecification of the covariate effects. So if we have many crossings we should probably examen our model. In [3] Koenker show that the sample path $\hat{Q}_Y(\tau|\overline{\mathbf{x}})$, where $\overline{\mathbf{x}}$ is the mean of x, is a nondecreasing function of τ. This says that for the "*typical*" value of $\mathbf{x}$ we do not have crossings, but we can not be sure not to have crossings at other values of $\mathbf{x}$ and it does not guarantee give us any area where there is no crossings.

In the context where we use quantile regression the variables can not take values in $\mathbb{R}$ but on in a subset of $\mathbb{R}$, this is probably also the case in most applications, when this is the case crossings can be avoided. In Section 6 some solutions to the crossing problem will be discussed.

2.4.2 Asymptotic Properties of Quantile Regression

Here we will not go deep into the asymptotic theory of quantile regression, but just give a few fundamental properties. A reference on the subject is Chapter 4 of [3]. A minimal asymptotic requirement on $\hat{\beta}$ is that it is consistent, the necessary and sufficient conditions for consistency is stated for $\tau \in (0,1)$ in [3] and are proved in [4] for the case $\tau = \frac{1}{2}$. These conditions are on the distribution of the errors and on the design matrix, the requirement on the distribution of the errors is that the distribution functions are strictly increasing in some neighborhood of the τ, in [3] Koenker note that the estimator can not be consistent if this is not the case since any estimator in this neighborhood would then be an estimator for the τ-quantile.

The conditions on the design matrix ensure that the explanatory variables is not concentrated in a subspace of $\mathbb{R}^K$ and that the rate of growth is not too large, this requirement will e.g. be satisfied if under the condition that $N^{-1}\sum \mathbf{X}^T\mathbf{X}$ converge to a positive definite matrix, but the conditions for convergence is weaker than this.

The conditions mentioned above only state that the quantile estimate will converge, not the rate of convergence or the distribution of the estimate. To say something about this the conditions must be strengthen, these conditions are also given in [3], these are mentioned here since they should be considered when we design the model, that is used later on. Further these give some indications of test procedures, which is the subject of the next chapter.

The condition on the conditional distribution of $Y|x$ will be denoted F, i.e. we have

$$Q_{Y_i}(\tau|x_i) = F_{Y_i}^{-1}(\tau|x_i) = \xi_i(\tau) \tag{2.78}$$

The distribution functions $F_{Y_i}(y|x_i)$ will be denoted F_i and the corresponding density function will be denoted f_i, note here that there is no assumption on identical distributions.

Conditions on F: The distribution functions $\{F_i\}$ are absolutely continuous, with continuous densities $f_i(\xi)$ uniformly bounded away from 0 and ∞ at the point $\xi_i(\tau)$, $i = 1, 2, ...$

As long as we consider continuous random variables this is not a strong requirement on the distribution functions. Even if it can be assumed that these conditions hold, then we will have difficulties if we want to use the estimates of $\hat{Q}$, since these does not necessarily fulfill these conditions, we can have crossings in which case $\hat{Q}$ is not on-to-one, and if the response variable is restricted to some interval then $\hat{Q}$ can also make forecast outside this interval, but here we know that f is zero. These complications make hypothesis testing very complicated (see e.g. [11]).

Conditions on X: There exist positive definite matrices $\mathbf{D}_0$ and $\mathbf{D}_1(\tau)$ s.t.

$$\lim_{n\to\infty} \quad N^{-1}\sum_i \mathbf{x}_i\mathbf{x}_i^T = \mathbf{D}_0 \tag{2.79}$$

$$\lim_{n\to\infty} \quad N^{-1}\sum_i f_i(\xi_i(\tau))\mathbf{x}_i\mathbf{x}_i^T = \mathbf{D}_1(\tau) \tag{2.80}$$

$$\max_i \quad N^{-\frac{1}{2}}||\mathbf{x}_i|| \to 0 \tag{2.81}$$

These conditions seems quite weak as well. The problem is more the complexity of these conditions, to estimate $\mathbf{D}_1$ is not easy under general f_i. The next theorem state that under the above conditions the quantile regression is consistent and efficient.

Theorem 2.4 *Under the above conditions*

$$\sqrt{N}(\hat{\beta}(\tau) - \beta(\tau)) \rightsquigarrow \mathcal{N}(0, \tau(1-\tau)D_1^{-1}D_0D_1^{-1}) \tag{2.82}$$

In the iid error model this reduces to

$$\sqrt{N}(\hat{\beta}(\tau) - \beta(\tau)) \rightsquigarrow \mathcal{N}(0, \omega^2 D_0^{-1}) \tag{2.83}$$

with $\omega^2 = \tau(1-\tau)/f_i^2(\xi_i(\tau))$.

The conditions does not seem very strong and the conditions on the design matrix is somewhat similar to the conditions in the general linear model.

As a small example we can look at the sample quantile case. If we have a realization of a stochastic variable from a distribution F then the observation $x_{(k)}$ with $k = \lceil \tau n \rceil$ will be asymptotically normally distributed with mean $F^{-1}(\tau)$

and variance

$$\omega^2(\tau) = \frac{\tau(1-\tau)}{Nf^2(F^{-1}(\tau))} \tag{2.84}$$

see [10]. We see that this is what we get from using 2.84 on the sample quantile case.

The results here give hope for some kind of hypothesis testing procedure, and these have also been developed under quite general conditions, the main problem is as we will see in the next section that these become very complicated if there is not very strong assumptions on the residuals. Even with complicated tests there are strong requirements on the residuals.

2.5 Hypothesis Tests

When having described the model above one of course want to perform some kind of test to be able to compare models, to do this we make partition of the model

$$Q_{y_i}(\tau;\mathbf{x}) = \mathbf{x}_i'\beta(\tau) = \mathbf{x}_{1i}^T\beta_1(\tau) + \mathbf{x}_{2i}^T\beta_2(\tau) \tag{2.85}$$

where $\beta_1 \in \mathbb{R}^K$ and $\beta_2 \in \mathbb{R}^p$ in this we want to test $H_0 : \beta_2 = 0$ against the alternative. Now set $\hat{S} = \min_\beta S$ and $\tilde{S} = \min_{\beta_1} S$

The most simple case is if r_i is iid from the asymmetric Laplacean density, this is described by

$$f(u) = \tau(1-\tau)e^{-\rho_\tau(u)} \tag{2.86}$$

The maximum likelihood estimate under the assumption that r_i come from this distribution yields the estimates described in (2.10). If r_i follow this distribution we can make the log likelihood ratio as $L_n = 2(\tilde{S}(\tau) - \hat{S}(\tau))$, this will be χ^2 distributed under the assumption that r_i follow the asymmetric Laplacean described above. Now it seems to be quite implausible that the residuals should follow this distribution so to make a test useful it should be more general.

If r_i is iid but follow some other distribution (with $f(F^{-1}(\tau)) > 0$) then the likelihood ratio is

$$L_N(\tau) = \frac{2(\tilde{S}(\tau) - \hat{S}(\tau))}{\tau(1-\tau)s(\tau)} \tag{2.87}$$

with $s(\tau) = 1/f(F^{-1}(\tau)$. L_N will converge weakly $\chi^2_{\eta(\tau)}(q)$, where the non centrality parameter $\eta(\tau)$ depend on the inverse of $\mathbf{D}_0$, the estimates and $\omega^2(\tau)$, further under the null hypothesis $\sup_{\tau\in\mathrm{T}} L_n(\tau)$ converges to the central χ^2_q distribution, where $\mathrm{T} = [\epsilon, 1-\epsilon]$ for some $\epsilon \in (0, \frac{1}{2})$. Already here we see the difficulties in the test procedure, since the test statistic involve the p.d.f. of the residuals, so to calculate this we have to estimate $f(F^{-1}(\tau))$ and we have to take the supremum over an interval of τ's.

[11] give tests in more general situations, the most general setting is that

$$y_i = \mathbf{x}_i^T\beta + \sigma_i u_i \tag{2.88}$$

with $\sigma_i = \mathbf{x}_i^T\gamma$ and u_i iid from a distribution F. Even though this might seem a general condition we still have to assume (or prove) independence of some sort.

The test statistics also becomes very complicated and besides the p.d.f of r_i we have to estimate γ or at least matrices depending on γ. The estimates of these parameters are themselves quite complicated. So we see that as soon as we leave the assumption that r_i is iid from the asymmetric Laplacean, then this gets quite complicated. [11] go through these kind of tests.

As we will discuss in Chapter 4 we can not assume that our residuals are independent. Even though the structure in equation (2.88) is quite general it still require independence of the u_i's and therefore that $\sigma_i u_i$ is uncorrelated. In the context of wind power forecast we will a priori assume correlation between the errors, further it is not clear how to determine if we can assume (2.88).

Hypothesis will be considered briefly in Chapter 4, for the data set used in the presentation. This chapter will also discuss other ways to measure performance of quantile regression models.

Chapter 3

About splines

The purpose of this chapter is to describe B-splines basis functions, which will be used in later chapters. First a formal definition of splines and motivations for using the special class of splines called B-splines is given. Then some fundamental properties of B-splines basis functions are given.

Section 3.4 concentrate on cubic B-splines basis functions and how to impose special boundary conditions for the cubic B-splines basis functions. This section and the Practical Summary's therein give a constructive guide to spline with special boundary conditions.

At the end of the chapter the hat-matrix, which takes the space of observations to the space of predictions, and thereby give a basis for comparison between different smoothers or kernels, is considered.

3.1 Introduction

The following definition of splines is taken from [8]. It is given to motivate the construction of the B-splines basis functions and the discussion of splines in more general.

Definition 3.1 *A spline function s of degree m ($s(x) \in \mathcal{S}_m(t_1, ..., t_n)$) is a polynomial of degree at most m on each of the intervals defined by the knot sequence $\{t_j\}_{i=1}^n$ and the intervals $(-\infty, t_1)$, (t_n, ∞), with the first $m-1$ derivatives varying continuously over the knots.*

Before going on with the construction of B-splines, we take a brief discussion of splines in general and what is so appealing about B-splines. In this context we are going to use the basis functions of $\mathcal{S}_m(t_1, ..., t_n))$, to estimate or approximate unknown functions f, so if $\{s_j\}_{j=1}^K$ is a basis for $\mathcal{S}_m(t_1, ..., t_n))$, then we estimate f by

$$\hat{f}(x) = \sum_{j=1}^{K} \hat{\alpha}_j s_j(x) \tag{3.1}$$

w.r.t. some loss function, e.g. the loss function as discussed in Chapter 2. Hence we have $\hat{f} \in \mathcal{S}_m(t_1, ..., t_n))$, so we can put some prior assumption on how many times f is differentiable into $\hat{f}$, by choosing the right m. By controlling the location of the knot sequence we can to some extend control how much local variation $\hat{f}$ can handle.

These arguments may not be very convincing especial not since we would properly normally assume that $f \in C^\infty$, and we can of course not choose $m = \infty$ and we can not search for functions in C^∞.

Traditionally $m = 3$ is chosen, one explanation for this is probably, as noted in [9] p. 22, that we are not able to see (from a graph), whether a function is C^2 or C^∞. So by choosing $m = 3$, the spline appears to be a C^∞ function.

The number of basis functions is $n + m + 1$, to see this simply count degrees of freedom, this is done explicitly below. This shows that there is a trade off between the number of basis functions and differentiability. Further when choosing B-splines the number of intervals where the basis functions have support is proportional to m. Other properties of the B-splines will be stated in the next sections of this chapter.

The arguments above is somewhat esthetic, a more formal argument for choosing $m = 3$ is to look at the minimization problem (3.2) below (see [9] p. 27). The use of N instead of n is to emphasize that we now look at every observation, and n is used for a knot sequence which does not necessarily have anything to do with the location of observations. It should also be emphasized that the minimization is to be done w.r.t. functions.

$$\arg\min_f \left(\sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int f''(t)^2 dt \right) \tag{3.2}$$

where λ is a fixed constant. The solution to this problem is a natural cubic spline (natural meaning $f''(x_1) = f''(x_N) = 0$), with the knots sequence $\{x_{(i)}\}_{i=1}^N$. Even though this is not what we do here it gives some support to the choice of $m = 3$. A remark here is that even though N is large the efficient degrees of freedom is far less than N. It is clear that there are many ways to make spline basis functions, one quit intuitively way is to use the truncated power series

$$\begin{aligned} s(x) &= \theta_{-m} + \theta_{-m+1}x + \cdots + \theta_0 x^m + \sum_{j=1}^{n} \theta_j (x - t_j)_+^m, \quad x \in \mathbb{R} \\ &= \sum_{j=-m}^{n} \theta_j s_j(x) \end{aligned} \tag{3.3}$$

this is clearly a spline of degree m. If there is no knots this is just a polynomial of degree m. It is also seen that there are $n+m+1$ degrees of freedom. In Definition 3.1 there are $n+1$ intervals and the polynomials on each interval have $m+1$ parameters, and there are m restrictions per knot. With $(n+1)(m+1) - mn = n + m + 1$ and since the s_j's are linearly independent, we actually have a basis for $\mathcal{S}_m(t_1, ..., t_n)$.

Each of the basis functions in (3.3) have support on a infinite interval and are strictly increasing as a function of the distance from t_j, so they grow fast. Therefore each of the basis functions can become very large, and as a consequence this can course numerical problems.

These problems are solved with the B-spline basis functions, which only have support in the interval $[t_j, t_{j+m+1}]$, so we don't have the problem of the basis functions going to ∞. Further the B-spline basis functions does not have values outside the interval $[0, 1]$.

The rest of this chapter will concentrate on the properties of B-spline basis functions. To be able to derive the basis functions and their properties in a quite simple notational setting, it is convenient to introduce *divided differences.* This presentation of divided differences follows [6], but only the results needed for the description of B-splines basis functions are stated, and some of the results are stated and proved in a less general setting than in [6].

3.2 Divided Differences and B-splines

When we are not interested in the actual construction of the B-spline basis functions, but merely in the properties of these, divided differences as defined below give a convenient way of introducing B-spline basis functions.

This sections starts by the definition of divided differences and a little on how to calculate these in order to get a feeling of what they are and how to work with them.

Definition 3.2 *The k-th divided difference of a function g at the points $t_j, ..., t_{j+k}$ is the leading coefficient (i.e. the coefficient of x^k) of the polynomial of degree k which agrees with g at $t_j, ..., t_{j+k}$. It is denoted*

$$[t_j, ..., t_{j+k}]g \tag{3.4}$$

It is maybe not very clear from the definition what this object is. By establishing the existence and uniqueness of this, we can get some feeling of what it is, especially the existence give a direct interpretation of this. The uniqueness and existence of this is of course important in its own right, because it ensures the existence and uniqueness of divided differences.

In order to do this define $\mathbb{P}_n$ as the space of all polynomials of degree n, i.e. $p_n(x) \in \mathbb{P}_n \Rightarrow p_n(x) = \sum_{j=0}^{n} a_j x^j$. For a given sequence of distinct points $\mathbf{t} = \{t\}_{j=1}^{n+1}$ and a function g, there is exactly one polynomial $p_n(x) \in \mathbb{P}_n$ for which $p_n(t_j) = g(t_j)$ for $j \in 1, 2..., n+1$. In order to see the uniqueness consider another polynomial $q_n(x) \in \mathbb{P}_n$, with $q_n(t_j) = g(t_j)$ $j = 1, ..., n+1$, then $p_n(x) - q_n(x)$ is a polynomial of degree n with $n+1$ roots, which must be the zero function.

To realize the existence of such a polynomial, define the set of integers between 1 and $n+1$ both included and take away the number j from this set, or more formal $\mathcal{J}_j = \{1, ..., j-1, j+1, .., n+1\}$ and look at the polynomial, which is called the Lagrange form

$$p_n(x) = \sum_{j=1}^{n+1} g(t_j) \prod_{l \in \mathcal{J}_j} \frac{(x - t_l)}{t_j - t_l} \tag{3.5}$$

This polynomial is clearly of degree n and $p_n(t_j) = g(t_j)$ for $j = 1, 2..., n+1$.

Now that we have established the existence and uniqueness of divided differences, we can actually write down the divide difference of a function g as

$$[t_1, ..., t_{n+1}]g = [t_1, ..., t_{n+1}]p_n = \sum_{j=1}^{n+1} \frac{g(t_j)}{\prod_{l \in \mathcal{J}_j} (t_j - t_l)} \tag{3.6}$$

So from the Lagrange form we can get the divided difference directly without calculating the entire polynomial.

In Definition 3.2 it is not required that the points $t_j, ..., t_{j+k}$ are distinct. If there are multiple points, then the term *agrees* means: if there is m points that coincide at t then p and g agree m-fold at t i.e.

$$p^{(j-1)}(t) = g^{(j-1)}(t) \quad \text{for } j = 1, ..., m$$

This property is used when working with the boundary conditions later on.

The next example gives complete calculation of the divided differences for a specific function and a knot sequence, and also offers a more complete derivation of (3.6).

Example 3.1 (Divided difference of $\sin(x)$) Look at the function $g(x) = \sin(x)$ and the knot sequence $\mathbf{t} = \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, 1\}$, this give $\{g(t_i)\}_{i=1}^{5} = \{0, \frac{\sqrt{2}}{2}, 1, \frac{\sqrt{2}}{2}, 0\}$. From the Lagrange form (3.5) we immediately get that the interpolating polynomial of degree 4 which agrees or interpolate $g(x)$ at the points in $\mathbf{t}$, can be written as ($\mathcal{O}(x^3)$ means terms of degree 3 or less)

$$\begin{aligned}
p_5(x) &= \sum_{j=1}^{5} g(t_j) \prod_{l \in \mathcal{J}_j} \frac{(x - t_l)}{t_j - t_l} \\
&= \sum_{j=1}^{5} \frac{g(t_j) x^4}{\prod_{l \in \mathcal{J}_j} (t_j - t_l)} + \mathcal{O}(x^3) && (3.7) \\
&= \frac{x^4}{\frac{\pi}{4}(-\frac{\pi}{4})(-\frac{\pi}{2})(-\frac{3\pi}{4})} \frac{\sqrt{2}}{2} + \frac{x^4}{\frac{\pi}{2}\frac{\pi}{4}(-\frac{\pi}{2})(-\frac{3\pi}{4})} \\
&\quad + \frac{x^4}{\frac{3\pi}{4}\frac{\pi}{2}\frac{\pi}{4}(-\frac{\pi}{4})} \frac{\sqrt{2}}{2} + \mathcal{O}(x^3) \\
&= \left(-\frac{\sqrt{2}}{3} + 1 - \frac{\sqrt{2}}{3} \right) \frac{4^3 x^4}{\pi^4} + \mathcal{O}(x^3) \\
&= \frac{4^3(3 - 2\sqrt{2})}{3\pi^4} x^4 + \mathcal{O}(x^3) && (3.8)
\end{aligned}$$

the divided difference is now given directly from (3.8) as $[0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, 1] \sin(x) = \frac{4^3(3-2\sqrt{2})}{3\pi^4}$. This example illustrates how to get (3.6). □

The definition of the B-spline is given in terms of divided differences, the definition can be given in other ways. Theorem 3.1 below gives an alternative but equivalent definition.

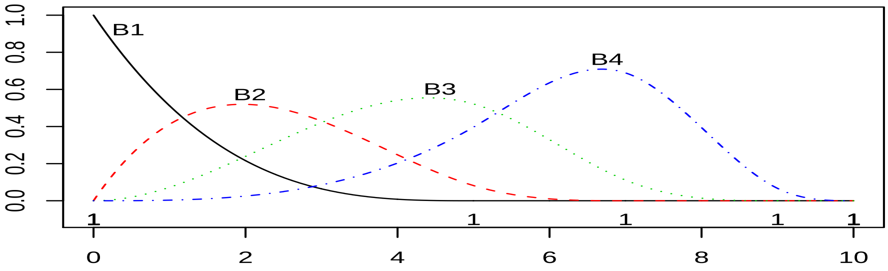

Figure 3.1: The figure illustrates how the B-spline basis functions behave when influenced by knots of different multiplicity. The knot sequence is $\mathbf{t} = \{0, 0, 0, 0, 5, 7, 9, 10\}$. So B_1 see a knot with multiplicity 4, B_2 see a knot with multiplicity 3, B_3 see a knot with multiplicity 2 and B_4 see only isolated knots. So this should illustrate the influence of knots with multiplicity greater than one. It should be noted that this is only the first 4 basis functions for the interval [0,10]. In order to span the space of all spline functions in this interval we would need 3 more basis functions, but only the 4 first have been plotted to make the properties of the single basis functions more clear.

Definition 3.3 *The j-th normalized B-spline of degree k+1 for a nondecreasing knot sequence* $\mathbf{t}$ *is defined as*

$$B_{j,k,\mathbf{t}}(x) = (t_{j+k} - t_j)[t_j, ..., t_{j+k}](\cdot - x)_+^{k-1} \forall x \in \mathbb{R} \tag{3.9}$$

The "$\cdot$" means that the divided difference is to be taken w.r.t. the "$\cdot$", which is a place holder for the considered function , while x is considered as a constant. The function $(x)_+^{k-1}$ should be read $((x)_+)^{k-1}$ is defined by

$$(x)_+ = \begin{cases} x & \text{for} \quad x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

Normalized here refers to the fact that $\sum_i B_{i,k,\mathbf{t}} = 1$ for $x \in [t_{j_1+k}, t_{j_n-k}]$. This fact will not be proved here, but it is quit straight forward by use of the same technique as used to calculate the value of the integral over one spline (see Section 3.3), or look at p. 110 in [6].

Since it will normally be clear from the context what k and $\mathbf{t}$ are, the B-splines basis functions will often just be denoted B_j. The normalized B-splines basis functions are often, see e.g. [8], denoted N_j.

Figure 3.1 shows the cubic ($k = 4$) B-spline basis functions for a knot sequence as indicated on the x-axis. Note that these do not span the space of all spline functions in the interval, since we would need 3 more basis functions to span this space or equivalent 3 knots greater than or equal to 10.

The knot at $x = 0$ have multiplicity 4, so the figure should illustrate how multiple knots influence the B-spline basis functions. It is seen that e.g. B_1 is not even continuous at $x = 0$.

Figure 3.2 goes into some more details about this point. It is seen that the basis functions are all less than one. They are zero outside the interval $[0, 10]$, and we therefore do not have the problems that was outlined for the truncated power series.

To prove that the B-spline is actually a spline, some properties of divided differences is needed. Some of these properties will also be used when deriving the derivative and integral of the B-spline. These properties can also be used to make a more constructive definition of B-splines, and hence this is done in the next theorem.

Consider $p_j \in \mathbb{P}_j$ (for the definition of $\mathbb{P}_j$ see the discussion right after Definition 3.2) the interpolating polynomial for the pair $(\mathbf{t}_j, g)$, $\mathbf{t}_j = \{t_l\}_{l=1}^{j}$. One representation of an interpolating polynomial of g on n point is

$$\begin{aligned} p_n(x) &= p_0(x) + (p_1(x) - p_0(x)) + \cdots + (p_n(x) - p_{n-1}(x)) \\ &= [t_1]g + (x - t_1)[t_1, t_2]g + \cdots \\ &\quad + (x - t_1)\cdots(x - t_{n-1})[t_1, ..., t_n]g \end{aligned} \tag{3.11}$$

where the second equality follows directly from the definition of divided differences, or it can be found by writing down (3.5) and (3.6) explicitly, as an illustration this is done for $p_1(x) - p_0(x)$ we get

$$\begin{aligned} p_1(x) - p_0(x) &= g(t_1)\frac{x - t_2}{t_1 - t_2} + g(t_2)\frac{x - t_1}{t_2 - t_1} - g(t_1) \\ &= x\left(\frac{g(t_1)}{t_1 - t_2} + \frac{g(t_2)}{t_2 - t_1}\right) - t_1\left(\frac{g(t_1)}{t_1 - t_2} + \frac{g(t_2)}{t_2 - t_1}\right) \\ &= (x - t_1)[t_1, t_2]g \end{aligned}$$

(3.11) is also called the Newton form. This representation will be independent of the order in which we take the points t_j, so p_n could also be written as

$$\begin{aligned} p_n(x) &= [t_n]g + (x - t_n)[t_n, t_1]g + \cdots \\ &\quad + (x - t_n)(x - t_1)\cdots(x - t_{n-2})[t_1, ..., t_n]g \end{aligned} \tag{3.12}$$

by equating coefficients of x^{n-2} in (3.11) and (3.12) we get

$$\begin{aligned}[t_1,...,t_{n-1}]g - (t_{n-2}+t_{n-1})[t_1,...,t_n]g &= [t_n,t_1,...,t_{n-2}]g \\ &\quad -(t_n+t_{n-2})[t_1,...,t_n]g\end{aligned}$$

Since n was arbitrary, this gives the formula for distinct points

$$[t_1,...,t_n]g = \frac{[t_1,...,t_{j-1},t_{j+1},...,t_n]g - [t_1,...,t_{l-1},t_{l+1},...,t_n]g}{t_l - t_j} \tag{3.13}$$

if $t_j = t_{j+1} = \cdots = t_{j+r}$ then the corresponding formula is $[t_j,...t_{j+1}]g = g^{(r)}(t_j)$. This follows from taking limits in (3.13) and by using the fact that $[t_j]g = g(t_j)$. For g differentiable at t_1, this give

$$\lim_{t_1\to t_2}[t_1,t_2]g = \lim_{t_1\to t_2}\frac{g(t_2)-g(t_1)}{t_2-t_1} = g'(t_1)$$

From (3.13) it follows immediately that B-splines can be written as

$$B_{j,k,\mathbf{t}}(x) = [t_{j+1},...t_{j+k}](\cdot - x)_+^{k-1} - [t_j,...t_{j+k-1}](\cdot - x)_+^{k-1} \tag{3.14}$$

Now by using (3.13), it is clear that the divided differences can be written as

$$[t_1,...,t_n]g = \sum_{j=1}^{n} c_j g(t_j) \tag{3.15}$$

where c_j depends on $\{t\}_{j=1}^n$ but not on g. This is useful to show properties of the B-spline basis functions and its derivatives without having to calculate the actual values of divided differences, which can be quite involved.

The presentation given so far, is rather abstract and does not really offer an intuitive feeling of the construction of B-spline basis functions. It is however convenient if we want to prove properties of the B-spline basis functions. Theorem 3.1 below gives, by the use of formula (3.6) and (3.14) a direct recursive formula for calculations of B-spline basis functions. As a remark this shows the correspondence between the introduction of the B-splines in [6] and [8].

Theorem 3.1 is not really used in this presentation, since the focus here is theoretical properties of B-splines and construction of splines with special boundary conditions from B-spline basis functions and, not the direct construction of B-splines. However the theorem perhaps offers a more direct and intuitive definition of B-spline.

Theorem 3.1 *Define* $M_{j,k} = \frac{B_{j,k}}{t_{j+k}-t_j}$, *then the following recursive formula* *holds*

$$\begin{aligned} M_{j,1}(x) &= \frac{I_{[t_j \le x < t_{j+1}]}(x)}{t_j - t_{j+1}} \\ M_{j,k}(x) &= \frac{(t_{j+k} - x)M_{j+1,k-1}(x) - (t_j - x)M_{j,k-1}(x)}{t_{j+k} - t_j} \qquad k > 1 \end{aligned}$$

and $B_{j,k} = (t_{i+k} - t_i)M_{i,k}$.

The proof of this is rather long, and is therefore placed in appendix A.1, but the idea is to use (3.6), (3.14) and the definition of divided differences.

With the tools developed above, we can prove the following theorem, which shows that B-spline basis functions (under some constraint) are actually splines. The theorem will be used to prove that B-spline basis functions are actually basis functions.

Theorem 3.2 *If* $\mathbf{t}$ *is a (strictly) increasing sequence of knots, then* $B_{j,k,\mathbf{t}}$ *is a spline of degree* $k-1$.

Proof. Using the properties of divided differences derived above it is easy to write down the derivatives of B_j

$$\begin{aligned} B_j(x) &= \sum_{l=1}^{k} c_{j+l}(t_{j+l} - x)_+^{k-1} \\ B_j'(x) &= -(k-1)\sum_{l=1}^{k} c_{j+l}(t_{j+1} - x)_+^{k-2} \\ &\vdots \\ B_j^{(k-1)}(x) &= (-1)^{k-1}(k-1)!\sum_{l=1}^{k} c_{j+l} I_{[x \le t_{j+l}]}(x) \end{aligned}$$

Since $(t_l - t_j^+)_+^h = (t_l - t_j^-)_+^h$ for $h > 0$, it follow directly that the first $k-2$ derivatives are continuous, and that B_j are polynomials of degree at most $k-1$ on each interval defined by the knot sequence and the intervals $(-\infty, t_1]$ and $[t_n, \infty)$. □

Theorem 3.2 tells us that the B-spline basis functions are really splines, under the condition that the knot sequence is strictly increasing. It does not, however prove that the B-spline basis functions form a basis of the space of all spline

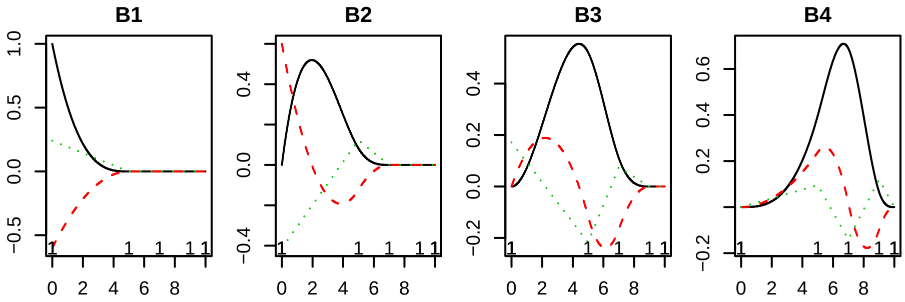

Figure 3.2: $B_{j,3,\mathbf{t}}$ for $\mathbf{t} = \{0, 0, 0, 0, 5, 7, 9, 10\}$. Each plot contains one of the basis functions from Figure 3.1 and its 2 first derivative, and illustrate in greater detail than figure 3.1 how the knot with multiplicity 4 at $x = 0$ effects the smoothness of the splines.

functions. This is the subject of the next theorem, but first a brief discussion of some of the consequences of the properties developed above is given.

Actually the B-spline basis functions does not span the space of all spline functions when just calculated from the knot sequence defined in Theorem 3.2. This can be seen from the fact that all the B-spline basis functions are zero outside the interval $[t_1, t_n]$. The trick is now to add some extra knots outside the interval (t_1, t_n), which will be referred to as outer knots. Exactly how this should be done will be discussed later, but an important point is that we allow knots of multiplicity greater than one for these outer knots.

The consequence of knots with multiplicity higher than one, is that the B-spline basis function is actually not a spline over this knot. This have already been seen in Figure 3.1. This is also the reason why we do not, calculate the B-splines outside the interval $[t_1, t_n]$ (and e.g "R" do not support values here). Figure 3.2, where the B-spline basis functions from Figure 3.1 and their first two derivative is plotted, goes into more details on this point.

To fully appreciate Figure 3.2 it should be noted that B_j and all its derivative is zero outside the interval $[t_j, t_{j+k}]$. This follows directly from the definition of divided differences and the fact that $(t_j - x)_+^{k-1}$ is a polynomial of degree $k - 1$ in the interval $(-\infty, t_j]$ and zero in the interval $[t_{j+k}, \infty)$.

With this in mind the figure shows how many of the derivatives are continuous for each of the basis functions. It is seen that B_1 is not continuous, B_2' is not continuous and B_3'' is not continuous, while $B_4^{(p)}$ is continuous for $p = 0, 1, 2$.

So B_1, B_2, B_3 are not spline functions over the knot at $x = 0$.

Hence the figure illustrate the effect of multiple knots, and that we can not allow multiple knots for a spline basis function over $\mathbb{R}$. Therefore we only define the B-spline basis function in the interval $[t_1, t_n]$ and then add outer knots, the B-spline basis functions will span the space of all spline functions in the interval $[t_1, t_n]$.

This is stated more precisely in the following theorem, which is a modification of Theorem IX.1 in [6].

Theorem 3.3 *If* $\mathbf{t}_1 = \{t_j\}_{j=1}^{n}$ *is a strictly increasing sequence and* $\mathbf{t}_- = \{t_j\}_{j=1-k+1}^{0}$ *and* $\mathbf{t}_+ = \{t_j\}_{j=n+1}^{n-1+k}$ *are two nondecreasing sequences with* $t_0 \leq t_1$ *and* $t_{n+1} \geq t_n$*, then the sequence* $\{B_{j,k,\mathbf{t}}\}_{j=2-k}^{n-1}$*, with* $\mathbf{t} = \{\mathbf{t}_-, \mathbf{t}_1, \mathbf{t}_+\}$*, of B-splines span the restriction of* $\hat{\mathcal{S}}_{k-1}(\mathbf{t})$ *to* $x \in [t_1, t_n]$

The restriction here means that the conditions in Definition 3.1 hold for $x \in (t_1, t_n)$. In the following $\mathbf{t}_1$ will be referred to as the defining knot sequence, $\{\mathbf{t}_+, \mathbf{t}_-\}$ will be referred to as outer knot sequences, $\{t_j\}_{j=2}^{n-1}$ the interior knots, and $\{t_1, t_n\}$ the boundary knots. Now the proof of Theorem 3.3 is given

Proof. That $\{B_{j,k,\mathbf{t}}\}_{j=2-k}^{n-1}$ belongs to the restriction of $\hat{\mathcal{S}}_{k-1}(\mathbf{t})$ to $x \in [x_1, x_n]$ follows from Theorem 3.2. The spline functions in each of the intervals defined by $\mathbf{t}_1$ can be written as $\sum_{l=1}^{k} a_l([t_j, t_{j+1}])x^{l-1}$ (i.e. $a_l([t_j, t_{j+1}])$, $j = 1, ...n-1$, is to be seen as a function of the interval), this gives $k(n-1)$, $(n-1 =$ the number of intervals) parameters. There are $k-1$ restrictions per interior knot and $n-2$ interior knot, this gives $k(n-1)-(k-1)(n-2) = n+k-2$ degrees of freedom, and the number of B-splines is $n-1-(2-k)+1 = n+k-2$. So if the B-splines are linearly independent, then the theorem is proved. To see this, take B_j and look at x in the interval $[t_j, t_{j+1}]$, $j > 0$, the only other nonzero B-splines in that interval is $B_{j-k+1}, ..., B_{j-1}$, i.e. $k-1$ of them.

Now it follows from (3.15) that B_j can be written as $B_j(x) = \sum_{l=1}^{k} b_l x^{l-1}$ for some b_l, this gives k b_l's. So we are short of one free parameter. If we should be able to write B_j as a linear combination of $B_{j-k+1}, ..., B_{j-1}$ then we would have to impose some restriction on each of the b_l's, but since these depend on different parts of the knot sequence, this would be the same as putting restrictions on the knot sequence, the conditions have to be true for all knot sequences. The theorem is hereby proved. □

Theorem 3.3 shows that the B-spline basis functions span the space $\hat{\mathcal{S}}_k$ when restricted to the interval defined by $\mathbf{t}_1$. Theorem 3.3 therefore provides the justification for using B-splines instead of e.g. the truncated power series. This justify the use of the term *basis functions*.

The advantage of the B-spline basis functions over the truncated power series have been implied through out the text. These are taken here again as a conclusion of why the B-spline are superior, at least from a numerical point of view. B-spline basis function have small support (over k+1 knots) while the basis function from the truncated power series have support over all knots to the right of the defining knot. $|B_j(x)| \leq 1$ for all x and j, while the basis function for the truncated power series can be very large (e.g. one of the basis functions is x^{k-1}).

The next part of this section looks at differentiations and integration of B-spline basis functions. It will be assumed that the knot sequence $\mathbf{t}$ is as in the Theorem 3.3, and that $x \in [t_1, t_n]$.

3.3 Differentiation and Integration of B-Splines

The next theorem provides rules for integration and differentiation of B-spline basis functions

Theorem 3.4 *The differential of $B_{j,k}$ is given by*

$$B'_{j,k}(x) = (k-1)\left(\frac{B_{j,k-1}(x)}{t_{j+k-1}-t_j} - \frac{B_{j+1,k-1}(x)}{t_{j+k}-t_{j+1}}\right) \tag{3.16}$$

The integral of B_j over the interval $[a,b] \subseteq [t_j, t_{j+k}]$ is given by

$$\int_a^b B_{j,k}(x)dx = \frac{t_{j+k}-t_j}{k}\sum_{l=0}^{k+1}\left(B_{j+l,k+1}(b) - B_{j+l,k+1}(a)\right)$$

Giving the special case

$$\int_{-\infty}^{\infty} B_{j,k}(x)dx = \int_{t_j}^{t_{j+k}} B_{j,k}(x)dx = \frac{t_{j+k}-t_i}{k} \tag{3.17}$$

for $j = -k+2, -k+3, ..., n-1$

The theorem state that we can write the differential and the integral of B-spline basis functions as a linear combination of lower and higher order B-spline basis functions, respectively. Especially (3.17) will be used later on, to restrict the B-spline basis functions. The proof again illustrates some techniques for manipulations with divided differences and hence B-spline basis functions.

Proof. To calculate the differential of $B_j(x)$ use the same fact as used in the proof of Theorem 3.2 and then use (3.13), to get

$$\begin{aligned} B'_{j,k}(x) &= -(k-1)\sum_{l=j}^{j+k} c_j(t_l - x)_+^{k-2} \\ &= -(k-1)(t_{j+k}-t_j)[t_j, ...t_{j+k}](\cdot - x)_+^{k-2} \\ &= -(k-1)(t_{j+k}-t_j)\frac{[t_{j+1}, ...t_{j+k}](\cdot - x)_+^{k-2} - [t_j, ...t_{j+k-1}](\cdot - x)_+^{k-2}}{t_{j+k}-t_j} \\ &= (k-1)\left(\frac{B_{j,k-1}(x)}{t_{j+k-1}-t_j} - \frac{B_{j+1,k-1}(x)}{t_{j+k}-t_{j+1}}\right) \end{aligned}$$

This can of course also be used for calculating the integral of a B-spline. For this we represent $B_{j,k}$ by $B_{j,k+1}$ for some j using 3.16 recursively and keeping in mind that $B_{j,k}(x) = 0$ for $x \notin [t_j, t_{j+k}]$, this gives

$$\begin{aligned} B_{j,k}(x) &= \frac{t_{j+k}-t_j}{k} D_x B_{j,k+1} + \frac{t_{j+k}-t_j}{t_{j+k+1}-t_{j+1}} B_{j+1,k} \\ &= \frac{t_{j+k}-t_j}{k} D_x B_{j,k+1} + \\ &\quad \frac{t_{j+k}-t_i}{t_{j+k+1}-t_{j+1}}\left(\frac{t_{j+k+1}-t_{j+1}}{k} D_x B_{j+1,k+1} + \frac{t_{j+k+1}-t_{j+1}}{t_{j+k+2}-t_{j+2}} B_{j+2,k}\right) \\ &\vdots \\ &= \frac{t_{j+k}-t_j}{k}\left(\sum_{l=0}^{k+1} D_x B_{j+l,k+1} + \frac{1}{t_{j+2k+2}-t_{j+k+2}} B_{j+k+1,k}\right) \\ &= \frac{t_{j+k}-t_j}{k}\sum_{l=0}^{k+1} D_x B_{j+l,k+1}(x) \quad x \in [t_j, t_{j+k}] \end{aligned}$$

With this it is clear that we can calculate the integral of a spline basis as

$$\int_a^b B_{j,k}(x)dx = \frac{t_{j+k}-t_j}{k}\sum_{l=0}^{k+1}(B_{j+l,k+1}(b) - B_{j+l,k+1}(a))$$

for $t_j \leq a \leq b \leq t_{j+k}$. Since $\int_{-\infty}^{\infty} B_{j,k}(x)dx = \int_{t_j}^{t_{j+k}} B_{j,k}(x)dx$, we start by calculating the following sums

$$\begin{aligned}
\sum_{l=0}^{k+1} B_{j+l,k+1}(t_j) &= B_{j,k+1}(t_j) \\
&= [t_{j+1}, ..., t_{j+k+1}](\cdot - t_j)_+^k - [t_j, ..., t_{j+k}](\cdot - t_j)_+^k \\
&= [t_{j+1}, ..., t_{j+k+1}](\cdot - t_j)^k - [t_j, ..., t_{j+k}](\cdot - t_j)^k \\
&= 1 - 1 = 0
\end{aligned}$$

and

$$\begin{aligned}
\sum_{l=0}^{k+1} B_{j+l,k+1}(t_{j+k}) &= \sum_{l=0}^{k+1} [t_{j+l+1}, ..., t_{j+l+k+1}](\cdot - t_{j+k})_+^k \\
&\quad - \sum_{l=0}^{k+1} [t_{j+l}, ..., t_{j+l+k}](\cdot - t_{j+k})_+^k \\
&= [t_{j+k+2}, ..., t_{j+2k+2}](\cdot - t_{j+k})_+^k - [t_j, ..., t_{j+k}](\cdot - t_{j+k})_+^k \\
&= [t_{j+k+2}, ..., t_{j+2k+2}](\cdot - t_{j+k})^k - [t_j, ..., t_{j+k}]0 \\
&= 1
\end{aligned}$$

This immediately gives

$$\int_{-\infty}^{\infty} B_{j,k}(x)dx = \int_{t_j}^{t_{j+k}} B_{j,k}(x)dx = \frac{t_{j+k} - t_j}{k}$$

for $j = -k+2, -k+3, ..., n-1$. □ Now the fundamental properties of B-spline

basis functions are proved, most important that they are really a basis for $\hat{\mathcal{S}}_k$, and we known how to differentiate and integrate these basis functions. The next section will show how to construct splines with different boundary conditions.

3.4 Construction of Special Splines

This section will treat two types of boundary conditions, namely natural boundary conditions and periodic boundary conditions, which are called *Natural splines* and *Periodic splines*. Further more the attention will be restricted to cubic splines, i.e. $k = 4$. These are constructed by controlling the outer knot sequence. The section will also treat the question of how to fix the level of a spline function.

The construction of these assumes that we have access to B-splines, either from some software or from, e.g. an implementation of Theorem 3.1.

3.4.1 Knots and Boundary Conditions

From an interpolating point of view one normally have some values of a function at points t_j, $j = 1, ..., n$, i.e. $f(t_j)$. Now choosing the knot sequence as in Theorem 3.3, with $\mathbf{t}_1 = \{t_j\}_{j=1}^n$ and $\mathbf{t}_-$ and $\mathbf{t}_+$ arbitrary, the number of B-spline basis functions are $n + k - 2$, so for $k > 2$ ($k = 2$ correspond to piecewise linear functions) there is fewer conditions than basis function. Therefore some extra conditions are needed if we want to interpolate the functions $f(t_j)$ with a unique spline.

Now the spline basis functions in this presentation is used for approximation not interpolating, but from some assumptions on the functions it can of course be reasonable to give some boundary conditions, or indeed unreasonable not to do so.

Boundary conditions can be controlled by the two outer knot sequences $\mathbf{t}_-$ and $\mathbf{t}_+$. The choice of knots in $\mathbf{t}_1$ and the boundary conditions will then determine these sequences, or at least put some restrictions on these. As it will become clear later there can be some degree of choice involved in making these sequences.

3.4.2 Natural Splines

The following definition of natural splines is taken from [8]

Definition 3.4 *The function $s(x) \in \mathcal{S}_m(t_1, ..., t_n)$ belongs to the set of natural splines of degree m, $\mathcal{S}_m^{\mathcal{N}}(t_1, ..., t_n)$, over the knots $t_1, ..., t_n$, if $m = 2j - 1$, $j \in \mathbb{N}$ and s is a polynomial of degree at most $j - 1$ for $x \notin [t_1, t_n]$.*

This is the same as demanding that $s^{(p)}(t_1) = s^{(p)}(t_n) = 0$ for $p \geq j, j + 1, ..., 2j - 2$. With $\mathbf{t}$ as in Theorem 3.3 and $k = 4$ ($m = 3$) this is the same as $B''_{j,4}(t_1) = B''_{j,4}(t_n) = 0$ for $j = -3, ..., n - 1$. Even though these are now label led $B_{j,4}$ these will technically not be B-spline basis functions. This is of course also the case for the periodic spline basis functions.

The term *natural* comes from the fact that if one takes a flexible rod and fix it along a number of point (the knots) then the resulting shape is described by a

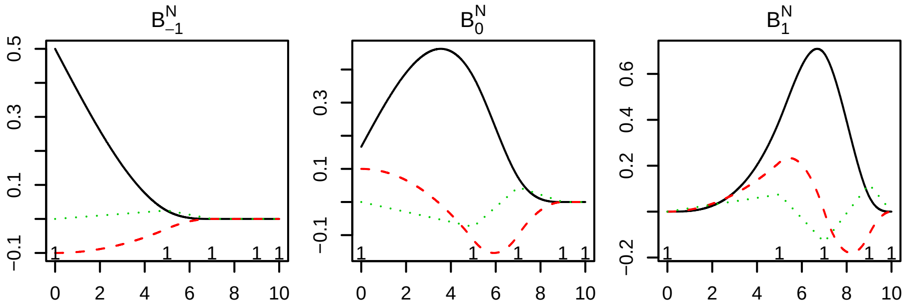

Figure 3.3: Natural B-spline basis functions and their 2 first derivative on the same interval as the B-spline of figure 3.1 and 3.2. The knot sequence is $\mathbf{t} = \{-5, -3, -1, 0, 5, 7, 9, 10\}$ and $B^N_{-1} = (2B_{-2} + B_{-1})/3$, $B^N_0 = (B_0 + 2B_{-1})/3$ and $B^N_1 = B_1$, by this a natural condition have been imposed at $x = 0$.

natural cubic spline. It was noted in the introduction that the solution to the minimization problem

$$\arg\min_f \left(\sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int f^{(q)}(t)^2 dt \right) \tag{3.18}$$

with $q = 2$, is a natural cubic spline. In general for $q \in \mathbb{N}$ the solution to (3.18) will be a natural spline of degree $m = 2q - 1$, see [6] p. 235.

Figure 3.3 shows natural splines and their first and second derivatives constructed from the B-spline basis function shown in Figure 3.1 and 3.2.

It is seen that there is only 3 basis functions. This is due to the fact that the first two natural splines is a linear combination of the first three B-splines basis functions. This is also quit natural since there have been made an extra constraint on the three first B-spline basis functions, by requiring that their second derivative is zero. How these are constructed is shown in details below.

When calculating natural spline basis functions, there is as stated above some degree of choice involved, implying that the way this is done below is mainly to be seen as an example of how this could be done. It should illustrate a technique of how boundary conditions can be imposed, and the technique for imposing other boundary condition.

As can be seen from (3.13) and the proof of Theorem 3.2 in the previous section, we have

$$\begin{aligned} B''_{j,4}(x) &= 6(t_{j+4}-t_j)[t_j,...,t_{j+4}](\cdot - x)_+ \\ &= 6([t_{j+1},...,t_{j+4}](\cdot - x)_+ - [t_j,...,t_{j+3}](\cdot - x)_+) \end{aligned}$$

at t_1 the only B-spline basis functions with nonzero second derivative is $B_{-2,4}, B_{-1,4}$ and $B_{0,4}$, and these will be

$$\begin{aligned} B''_{-2,4}(t_1) &= 6([t_{-1},...,t_2](\cdot - t_1)_+ - [t_{-2},...,t_1](\cdot - t_1)_+) \\ &= 6[t_{-1},...,t_2](\cdot - t_1)_+ \\ B''_{-1,4}(t_1) &= 6([t_0,...,t_3](\cdot - t_1)_+ - [t_{-1},...,t_2](\cdot - t_1)_+) \\ B''_{0,4}(t_1) &= 6([t_1,...,t_4](\cdot - t_1)_+ - [t_0,...,t_3](\cdot - t_1)_+) \\ &= -6[t_0,...,t_3](\cdot - t_1)_+ \end{aligned}$$

in generally $[t_j,...,t_{j+3}](\cdot - t_l)_+$ can be written as

$$[t_j,...,t_{j+3}](\cdot - t_l)_+ = \frac{1}{t_{j+3}-t_j}\left(\frac{\delta(l-j+2)}{t_{j+3}-t_{j+1}} - \frac{\delta(l-j+1)}{t_{j+2}-t_j}\right) \tag{3.19}$$

the proof of this is found in Appendix A.2, by the above we have

$$\begin{aligned} B''_{-2,4}(t_1) &= \frac{6}{(t_2-t_{-1})(t_2-t_0)} \\ B''_{-1,4}(x) &= \frac{-6}{(t_2-t_0)}\left(\frac{1}{t_3-t_0}+\frac{1}{t_2-t_{-1}}\right) \\ B''_{0,4}(x) &= \frac{6}{(t_3-t_0)(t_2-t_0)} \end{aligned}$$

Now forming two new functions $B^N_{-1,4}, B^N_{0,4}$ as linear combinations of $B_{-2,4}$, $B_{-1,4}$ and $B_{0,4}$ it is possible to create a basis for natural splines. This gives two equations in 6 unknown namely

$$\begin{aligned} B^N_{-1,4} &= a_1B_{-2,4}+b_1B_{-1,4}+c_1B_{0,4} \\ B^N_{0,4} &= a_2B_{-2,4}+b_2B_{-1,4}+c_2B_{0,4} \end{aligned}$$

since there are more free parameters than equations we have to make some choices here, these could e.g. be $c_1 = a_2 = 0$, $a_1 + b_1 = 1$, and $b_2 + c_2 = 1$, solving these gives

Practical Summary 3.1 (Natural B-splines) *From a sequence of B-spline basis functions $\{B_{j,4}\}_{j=-2}^{n-1}\}$, a set of spline basis functions $\{B^N_{-1,4}, B^N_{0,4}, \{B_{j,4}\}_{j=1}^{n-1}$ which are natural at t_1 can be constructed as*

$$\begin{aligned} B^N_{-1,4} &= \frac{t_2-t_{-1}+t_3-t_0}{t_2-t_{-1}+2(t_3-t_0)}B_{-2,4}+\frac{t_3-t_0}{t_2-t_{-1}+2(t_3-t_0)}B_{-1,4} \\ B^N_{0,4} &= \frac{t_2-t_{-1}}{t_3-t_0+2(t_2-t_{-1})}B_{-1,4}+\frac{t_3-t_0+t_2-t_{-1}}{t_3-t_0+2(t_2-t_{-1})}B_{0,4} \end{aligned}$$

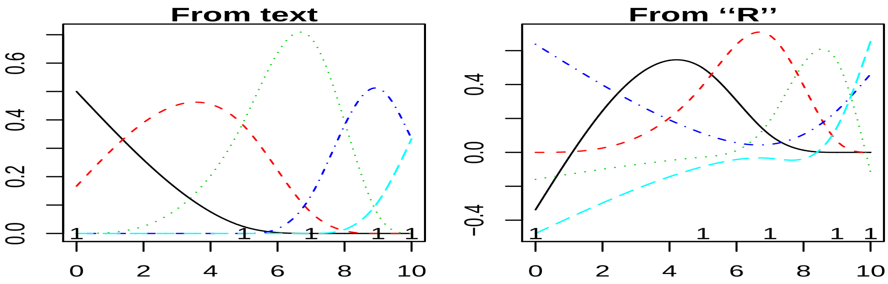

Figure 3.4: The left panel shows natural splines on the interval from $[0, 10]$ constructed from the recipe in the text, i.e. $\mathbf{t} = \{-5, -3, -1, 0, 5, 7, 9, 10, 11, 13, 15\}$, the right panel show natural splines as constructed from the build in function in "R". This illustrate that these natural splines are not unique.

something completely similar can be done at t_n.

For the special choice $t_3 - t_0 = t_2 - t_{-1}$, this becomes $a_1 = c_2 = \frac{2}{3}$ and $b_1 = b_2 = \frac{1}{3}$. This is what is shown in Figure 3.3.

This illustrates how natural splines basis functions can be constructed from B-spline basis functions and it is also clear that some choices have to be made as these are constructed.

Figure 3.4 illustrates this by showing a natural spline basis constructed from Practical Summary 3.1 and the natural spline basis functions constructed directly by "R". An advantaged of the one used in this presentation is that the basis functions in the opposite end of where the natural condition is imposed is not influenced, so in a similar way some other boundary condition could be impose at the opposite end.

3.4.3 Periodic Splines

In [8] it is stated that if the outer knots are chosen s.t.

$$\begin{array}{rclrcl} t_1 - t_0 &=& t_n - t_{n-1}, & t_0 - t_{-1} &=& t_{n-1} - t_{n-2} \\ t_2 - t_1 &=& t_{n+1} - t_n, & t_3 - t_2 &=& t_{n+2} - t_{n+1} \end{array} \tag{3.20}$$

Then the B-spline basis functions will be periodic. To see this it is actually enough to realize that these conditions imply that $t_{1+i} - t_{1+j} = t_{n+i} - t_{n+j}$ for $(i, j) \in \{-2, ..., 2\} \times \{-2, ..., 2\}$, and then use (3.13) to see that $[t_j, ..., t_{j+3}](\cdot -$

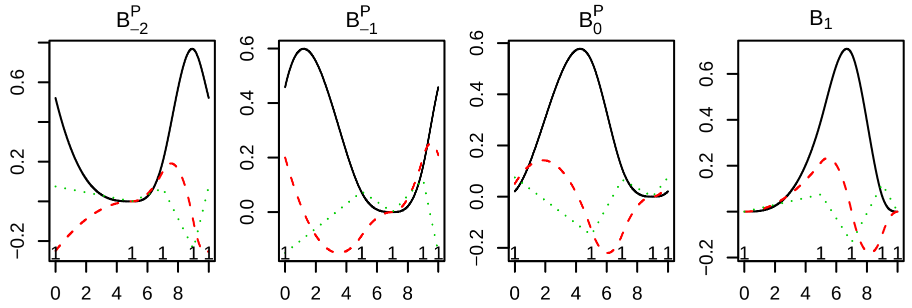

Figure 3.5: A periodic spline basis in the interval $[0, 10]$, constructed from the same inner knots as used in Figure 3.1 and 3.2, the knots sequence for the B-splines used to construct this is $\mathbf{t} = \{-5, -3, -1, 0, 5, 7, 9, 10, 15, 17, 19\}$. The two first derivative is also plotted in order to show that this is really a periodic spline basis.

$t_1)^k_+ = [t_{n-1+j}, ..., t_{n-1+j+3}](\cdot - t_n)^k_+$, $j = -1, 0$, $k \in \mathbb{Z}$. This leading to the conclusion $B^{(p)}_{1-l,4}(t_1) = B^{(p)}_{n-l,4}(t_n)$, $l = 1, 2, 3$ and $p = 0, 1, 2$.

Practical Summary 3.2 (Periodic splines) *Given a sequence of knots* $\mathbf{t}_1$ *and choosing the outer knots sequences* $\mathbf{t}_-$ *and* $\mathbf{t}_+$ *s.t. (3.20) is fulfilled. Then a set of periodic spline basis functions* $\{\{B^P_{1-l,4}\}^3_{l=1}, \{B_{j,4}\}^{n-4}_{j=1}\}$ *can be constructed as*

$$B^P_{1-l}(x) = B_{1-l,4}(x) + B_{n-l,4}(x), \quad l = 1, 2, 3 \tag{3.21}$$

this will give a periodic spline basis for the interval $[t_1, t_n]$.

Figure 3.5 shows a periodic spline basis constructed in this way, and for the same inner knots as in Figure 3.1 and 3.2 , the actual knot sequence is given in the figure text. The figure also shows the 2 first derivative of the basis basis functions.

As have been mentioned, the first and the last knot do not have any influence on the values inside the interval. This is true for both the natural and the periodic spline. These are necessary in the definition of the B-splines basis functions.

In conclusion we can control the B-splines by controlling the outer knots and then making a new basis as a linear combination of the resulting B-spline basis functions.

3.4.4 Fixing the Level

The purpose of the splines in this presentation is, as stated previously approximation, which is done through regression. The model for the quantile regression presented in Chapter 2, was

$$\hat{Q}(\tau; \mathbf{x}_t) = \mathbf{x}_t \hat{\beta}(\tau) \tag{3.22}$$

Now this should, as mentioned in the beginning of this chapter, correspond to an additive model of the type

$$\hat{Q}(\tau; \mathbf{x}_t) = \sum_{j=1}^{p} \hat{f}_j(x_{j,t}) \tag{3.23}$$

where $x_{j,t}$ is the explanatory variable j at time t. The functions $\hat{f}_j(x_j)$ have been approximated by splines as developed in the previous sections. The sum of functions in (3.23) is not unique unless some restrictions are put on $\hat{f}_j(x)$. One demand could be to force every function to go through zero or rather split the function $\hat{f}_j(x)$ into the two components $\hat{\alpha}_j$ and $\hat{g}_j(x)$, where the function $\hat{g}_j(x)$ is now fixed to some level. Now write down the sum of the $\hat{f}_j$

$$\sum_{j=1}^{p} \hat{f}_j(x_{j,t}) = \sum_{j=1}^{p} \hat{\alpha}_j + \sum_{j=1}^{p} \hat{g}_j(x_{j,t}) = \hat{\alpha} + \sum_{j=1}^{p} \hat{g}_j(x_{j,t}) \tag{3.24}$$

The last term will be unique.

The spline functions developed so far span the space of all spline functions of that type (natural, periodic or all spline functions), and for each function f_i we write

$$f_j(x_{j,t}) = \sum_{l=1}^{K} b_{j,l}(x_{j,t}) \tag{3.25}$$

where K is the number of degrees of freedom. As was seen in the case of the truncated power series this is of the same type as the functions in (3.24). So we have the same uniqueness problem.

To get around this, we can force the linear combination of spline functions to go through some specified value. If we choose this value to be zero, the demand becomes quit simple.

The following practical summary give a recipe for doing this in the case of natural spline basis functions, in this case it is very simple.

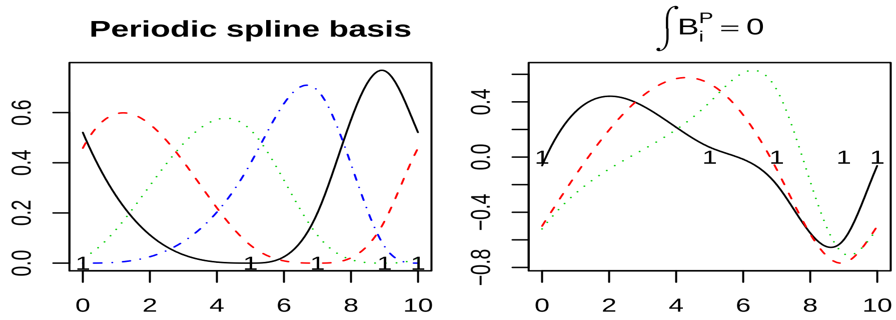

Figure 3.6: A periodic spline basis, and a periodic spline basis with integral zero in the interval $[0, 10]$, constructed from the same inner knots as used in figure 3.1 and 3.2, the knots sequence for the B-splines used to construct these is $\mathbf{t} = \{-5, -3, -1, 0, 5, 7, 9, 10, 15, 17, 19\}$. The spline basis with integral zero is constructed by subtracting $c_i B_0^P$ from each of the other periodic spline functions.

Practical Summary 3.3 (Fixed Natural Spline:) *By setting all the knots in the knot sequence* $\mathbf{t}_-$ *equal to* t_1 *and constructing a sequence of natural spline basis functions as described in Practical Summary 3.1, the sequence* $\{B_{0,4}^N, \{B_{j,4}\}_{j=1}^{n-1}\}$ *will be a basis for all spline* $s(x)$ *functions with* $s(t_1) = 0$ *and a natural condition at* t_1.

This is however not that simple for the periodic splines. Hence we use another approach, namely to demand that the integral over the period is zero, which is of course achieved with $\int B_i^P = 0$, $\forall i$. Since the spline basis functions are continuous, this also means that any linear combination of them is zero for some x_0. This x_0 will depend on the coefficients of the splines. This approach is also taken in [1].

By setting each of the basis functions equal to a linear combination of the original basis function and one of the other original basis function, this can be achieved. Again referring to the previous section, this is here done by choosing a new basis $B_j^* = B_j + c_k B_k$, $j \neq k$, by (3.17) the integral of B_j^* is

$$\int B_j^* = \frac{t_{j+4} - t_j + c_j(t_{k+4} - t_k)}{4} \tag{3.26}$$

Now this we can give the recipe for constructing a periodic spline basis with integral zero.

Practical Summary 3.4 (Periodic splines with integral zero) *From a sequence of periodic spline basis functions* $\{B_j^P\}_{j=1}^K$, *a sequence of periodic spline* *basis functions with integral zero* $\{B_j^*\}_{j\neq k}$ *can be constructed as*

$$B_j^* = B_j - \frac{t_{j+4} - t_j}{t_{k+4} - t_k} B_k \tag{3.27}$$

where k *is a fixed number.*

The proof follows directly from the previous discussion.

Figure 3.6 shows a periodic spline basis and the corresponding periodic spline basis with integral zero.

3.5 Smoothing

A smoother is a function which takes the pair of output or dependent variables and explanatory variables to the space of predictions. Such a function should be less variable than the output itself. For linear regression and least squares the smoother matrix is called the hat matrix.

The best LS estimate for the parameters in the general linear model

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} \tag{3.28}$$

is

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{3.29}$$

where $\mathbf{X}^T\mathbf{X}$ must have full rank for the expression to make sense, assuming this is the case we can write

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{S}\mathbf{y} \tag{3.30}$$

and in this case $\mathbf{S}$ is called the hat matrix, see [7].

In more general $\mathbf{S}$ is called the smoother matrix, when $\mathbf{S}$ does not depend on $\mathbf{y}$ it is called a linear smoother, this is the case for the hat matrix presented above.

The i'th row in $\mathbf{S}$ can be viewed as weights assigned to each of the observations when estimating $\hat{y}_i$. If two estimators have the same smoother matrix they are called equivalent kernels, see [9].

The matrix $\mathbf{S}$ is a function of the design matrix, and hence in the spline case a function of the spline basis function and the optimization criteria. For least

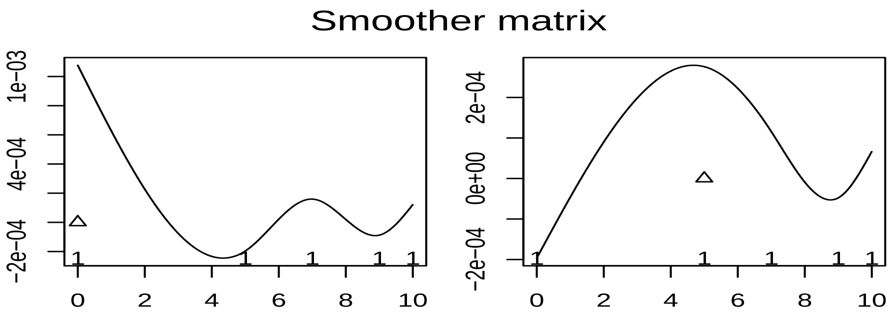

Figure 3.7: The figure shows two rows in the smoother matrix for least square estimation and the example used in the previous sections with natural boundary conditions. The triangle indicate the value of x in the row corresponding to the pair (x_i, y_i). The figure illustrates the quite slow decay in the values of the smoother matrix.

squares (LS) and linear regression, $\mathbf{S}$ is a linear smoother. This is however not the case for the quantile regression models presented in Chapter 2.

This section presents some properties of the LS estimator of the spline model, but not in details since this is not what is used in subsequent chapters. The LS smoother matrix is merely presented for comparing with the quantile regression model.

In this context the columns in the design matrix $\mathbf{X}$ will consist of the B-spline basis functions. The first column will be a vector of ones, due to the reasons discussed in Section 3.4.4.

As was pointed out in Section 3.2 the support of the cubic B-splines is small in the sense that it is zero outside an interval of 5 knots. So there will be many elements in the design matrix which are zero. This does not mean that the weights in the smoother matrix are zero. They will however be small far away from what could be the central point x_c, i.e. for $\hat{y}_i$ the central point is $x_c = x_i$. It makes good sense that the weight of y_i when constructing $\hat{y}_i$ should be the largest.

The analysis will be done with only one explanatory variable $x_{j,i} = x_i$, since otherwise we are not able to visualize $\mathbf{S}$.

First we take a short look at the design matrix in some more details, denote

$b_{j,i} = B_j(x_i)$. The the design matrix will be

$$\mathbf{X} = \begin{bmatrix} 1 & b_{1,1} & \dots & b_{k,1} \\ \vdots & \vdots & & \vdots \\ 1 & b_{1,N} & \dots & b_{k,N} \end{bmatrix} \tag{3.31}$$

So we can write down $\mathbf{X}^T\mathbf{X}$ directly as

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} N & \sum_{i=1}^N b_{1,i} & \sum_{i=1}^N b_{2,i} & \dots & \sum_{i=1}^N b_{k,i} \\ \sum_{i=1}^N b_{1,i} & \sum_{i=1}^N b_{1,i}^2 & \sum_{i=1}^N b_{1,i}b_{2,i} & \dots & \sum_{i=1}^N b_{1,i}b_{k,i} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^N b_{k,i} & \sum_{i=1}^N b_{1,i}b_{k,i} & \sum_{i=1}^N b_{2,i}b_{k,i} & \dots & \sum_{i=1}^N b_{k,i}^2 \end{bmatrix}$$

This implies that the first row, the first column and all diagonal elements of $\mathbf{X}^T\mathbf{X}$ will be different from zero. Therefore all elements in the inverse of this will (in general) be different from zero, this again leading to all elements in $\mathbf{S}$ being different from zero. However elements far away from x_C will be small, this can not be seen from $\mathbf{X}\mathbf{X}^T$, but can be seen if one calculate the rows in $\mathbf{S}$.

Figure 3.7 shows two different rows of the smoother matrix, the triangle is the x_i used to calculate the i'th row (i.e. $x_i = x_c$), in the sense that the i'th row of $\mathbf{S}$ is calculated from

$$\mathbf{S}_{i,\cdot} = [1, b_{1,i}, \dots, b_{k,i}](\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}^T \tag{3.32}$$

The smoothers in Figure 3.7 are made from the build in “R” natural spline function. These and the ones developed in previous sections of this chapter are very close, the absolute point wise difference being less than 3×10^{-17}, this difference can be explained by numerics so they are equivalent kernels as they should be.

The figure also illustrates the point that $\mathbf{S}$ is not zero far away from x_c. To illustrate this it would may be have been appropriate to give a plot with more knots. This is done in Figure 3.8 and 3.10 where an example from a real dataset is used to illustrate the implications of the quantile estimator producing a non-linear smoother. Even though the data have not been explained yet it should be clear what the effect of more knots is.

As was stated in Chapter 2 we can write the estimates of the parameters in the quantile regression model as

$$\hat{\beta} = \mathbf{X}^{-1}(h)\mathbf{y}(h) \tag{3.33}$$

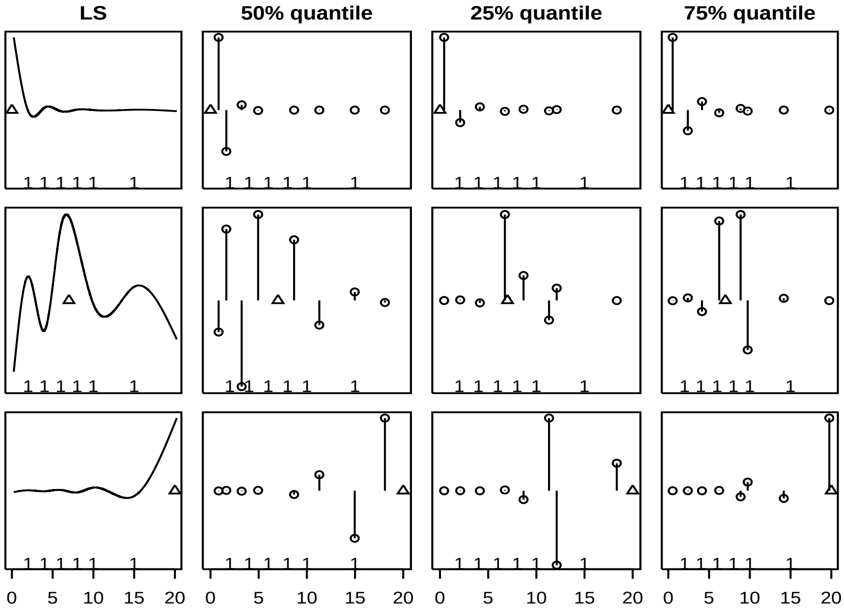

Figure 3.8: The Least square and quantile smoother matrix for the wind data, the least square smoother is plotted for reference. In most of the plots we see, that the LS smoother behave qualitatively like the LS. At least in the sense that weights close to x_c is large compared to weights far away. But for the 25% quantile and $x_c = 20$, there is a quite different picture, with weight far away being much larger than those close to.

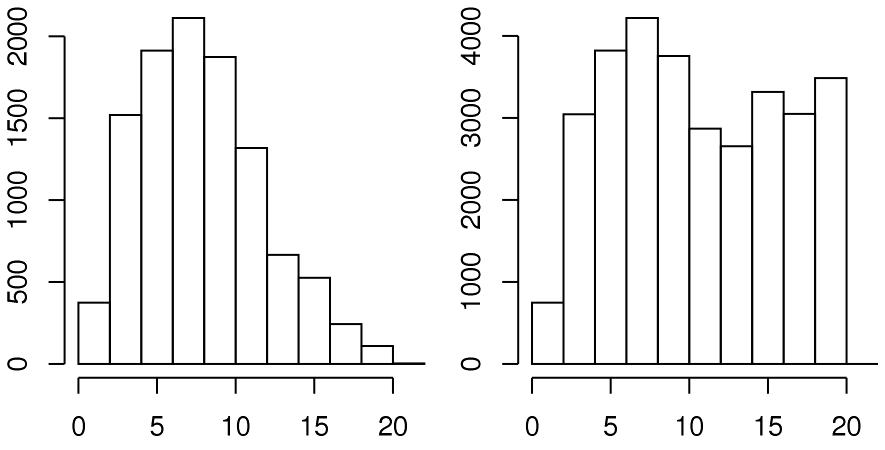

Figure 3.9: The distribution of wind speeds before and after the simulation of extra data points.

where h is some index set of length $K = no.\ elements\ in\ \hat{\beta}$. This means that we can create a smoother matrix, somewhat like the one above, by filling in vectors of zeros. In matrix form we can write this as

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{X}^{-1}(h)\mathbf{y}(h) = \mathbf{X}\mathbf{X}^{-1}(h)\mathbf{H}^T\mathbf{y} \tag{3.34}$$

where the elements in $\mathbf{H}$ is given by $H_{i,j} = \delta(i - h_j)$, $i = 1, 2, ..., N$ and $j = 1, 2, ..., K$. I.e. we can write down the smoother matrix as

$$\mathbf{S}_q = \mathbf{X}\mathbf{X}^{-1}(h)\mathbf{H}^T \tag{3.35}$$

$\mathbf{S}_q$ is now a function of y, since otherwise the parameter estimates are not a function of all of the y values and this is of course not the case. A simple implication of this formulation is that $S_{h_j,h_j} = 1$ and $S_{i,l} = 0$ for $j \notin h$.

This clearly illustrates that, there are some differences between the two smoother matrices. The hope could now be, that even though there are few weights compared to the number of observations, they would behave somehow as in the LS case in, e.g. that the largest weight were the one closest to x_c. This is however not the case in general.

Figure 3.8 shows the $\mathbf{S}$ matrix for the least squares fit and three different quantiles. Most of the quantile plots behave qualitatively similar to the LS case, but in the bottom panel for the 25% quantile we see a quit different behavior, where the largest weights are quite far away from x_c.

Now it is not very clear how to study the behavior of a nonlinear smoother. This is mainly because we are not able to control the index set h. If we increase the number of parameters to look at some kind of limit behavior, we will from time to time be very close to or indeed hit the points which produce the estimate and therefore generate one weight with value one and all other weights zero. To see this look at $\mathbf{S}_q(h)$, i.e. the rows of $\mathbf{S}_q$ defined by the index set h, this is

$$\mathbf{S}_q(h) = \mathbf{X}(h)\mathbf{X}^{-1}(h)\mathbf{H}^T = \mathbf{I}\mathbf{H}^T = \mathbf{H}^T \tag{3.36}$$

This shows that if $x_c = x_{h_j}$ for some $j \in \{1, 2, ...K\}$ then the weights in that case will consist of a vector with one element being one and all other being zero. This will of course happen if we increase the number of parameters far enough.

Therefore what is done here for studying $\mathbf{S}_q$ should be interpreted with some care and conclusions are not really made.

The data used for Figure 3.8, will be explained in greater details in subsequent chapters. For now it is enough to know that the input is wind speed and the output is an estimation error from a model to estimate wind power production. The number of observations is about 10000. In the data there are very few observation for large wind speeds, this could be one reason that this 25% quantile with $x_c = 20m/s$ looks so strange.

To study the effect of putting in more points at the height end of the wind scale new points are simulated according to the following scheme

$$x_{n+1} = \frac{1}{2}(x_{(i)} + x_{(i+1)}) + \epsilon$$
$$y_{n+1} = \frac{1}{2}(y_{(i)} + y_{(i+1)}) + \xi$$

$x_{(i)}$ refer to the order statistic and ϵ and ξ are normal random variables with small variances and mean zero. There are no physical meaning behind this simulation, but it permits us to put extra points in the position we want. In this way data points are filled in at the high end of the wind scale. Figure 3.9 shows histograms of the distribution of wind speeds before and after the simulation. So what has been done is to fill in data on the high end of the wind speed values.

What we would like is of course for $\mathbf{S}_q$ to behave somehow like the least squares smoother matrix. One approach is to let the number of observations go to infinity. Another approach is to impose a very large number of knots and hope that this will then look somehow the same as for the least square case.

Figure 3.10 shows rows in $\mathbf{S}$ and $\mathbf{S}_q$ for the simulated data, and different knots sequences. The upper left panel show the same knot sequence as Figure 3.8. It

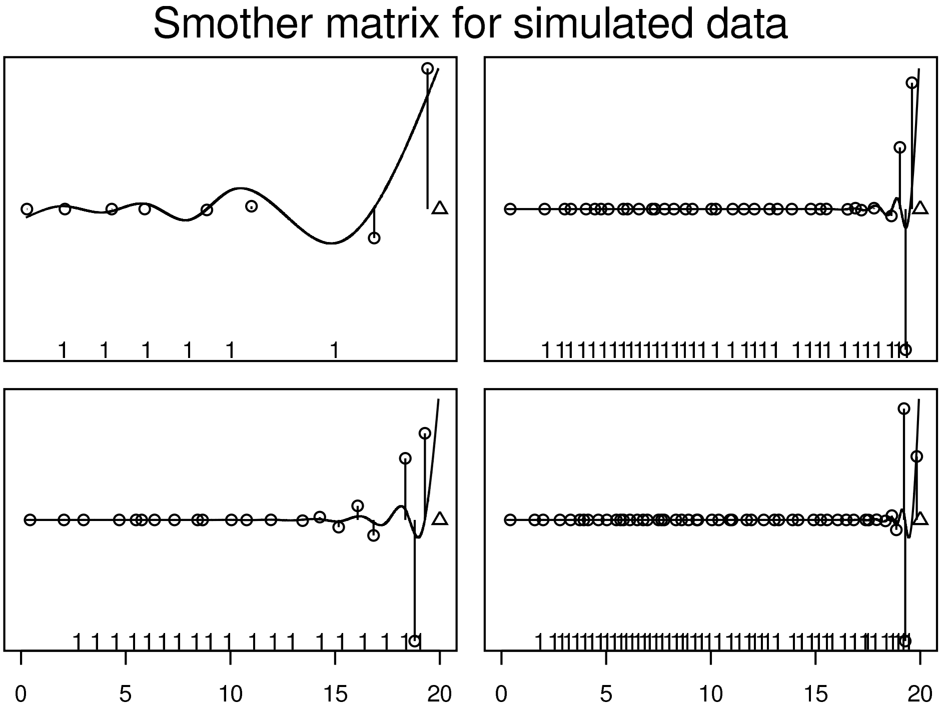

Figure 3.10: The figure show the smoother matrix for different knot placements for the simulated dataset, it is seen that the qualitative behavior is somehow like the one for the LS (the line), but the weight closest to x_c is not necessary the biggest.

is seen that we get quite good assembles with the LS. The 3 other panels show the same but for longer knot sequences. It is seen that we the location property is preserved, i.e. none of the smoothers "see" through a large number of knots. On the other hand we can not conclude that the largest weight are the one close to x_c.

A final remark on Figure 3.10 is that different simulations with the same scheme gave quite different results as to where the weight was located and how big they were.

Part II

Quantiles of Prediction Errors from Tunø Knob Wind Power Plant

Chapter 4

Quantile Models

4.1 Introduction

This chapter will use the theory of Chapter 2 and 3 to set up models for prediction of quantiles of the errors in predictions of wind power production from a wind power plant placed at Tunø Knob. The wind power is predicted with a program called WPPT (Wind Power Prediction Tool), see [19] for a description of WPPT. Figure 4.1 show the location of the Tunø knob wind power plant, along with the grid points where meteorological forecasts are available.

The aim is to get good information of the short term predictions. The most interesting horizons are those of 12-36 hours, since these are the horizons needed on the Nordic energy market called Nordpool. To obtain an optimal market strategy, it is not enough to know the mean and variance of prediction. Quantile regression is an approach to obtain additional information on the predictions.

We will use meteorological forecasts in the area around Tunø Knob to model these quantiles. It is not reasonable to assume that the conditional quantiles are linear functions of the meteorological forecast, hence to use the setting from Chapter 2, the quantile regression is done with respect to spline basis functions of the meteorological forecasts.

The first part of this chapter will define the general setting of the quantile regression with splines. Then some performance parameter for quantile regression are presented. In the last part of the chapter performance parameters are used to analyze some quantile regression models for the Tonø Knob data set. The aim of this analysis is twofold; on one hand to develop a good model for prediction of the quantiles, on the other hand to discuss the performance parameters in their own right.

4.2 The General Setting

The set up here follows the set up used in [1], where the same data set was analyzed. The most general quantile model $Q : \mathbb{R}^p \rightarrow \mathbb{R}$ we can think of is

$$Q(\mathbf{x};\tau) = g(\mathbf{x};\tau) \tag{4.1}$$

where g is any function of the vector $\mathbf{x} \in \mathbb{R}^p$. If we do not have a very clear idea of the behavior of the function g, then the space of all these functions is too large to search in. Therefore, some restrictions or approximations have to be made. The first approximation is to restrict the search to the space of additive models, i.e. the function g is replaced with a sum of functions $f_j : \mathbb{R} \rightarrow \mathbb{R}, \quad j = 1, ..., p$. With this we get the model

$$Q(\mathbf{x};\tau) = \alpha(\tau) + \sum_{j=1}^{p} f_j(x_j;\tau) \tag{4.2}$$

Models like this are described in [9], where a very thorough treatment of additive models is presented. Additive models can be any partition of the directions in $\mathbb{R}^p$, therefore strictly speaking the set up used here is only a subset of additive models.

The constant α, is a common intercept of the function. This is necessary because (as was discussed in Section 3.4.4), the functions f_j have to be fixed somehow to ensure uniqueness of the model. In [9] Hastie propose to force the average over observations to be zero. In the present presentation we approximate the functions f_j with natural and periodic spline functions and fix the level as described in Section 3.4.4.

The price to pay for these approximations is that mixed effects are ignored. So e.g. the effect of the wind speed is independent of the wind direction. This might or might not be reasonable, but some simplifications have to be done. The hope is just that the mixed effects are small compared to the isolated effect of the variables.

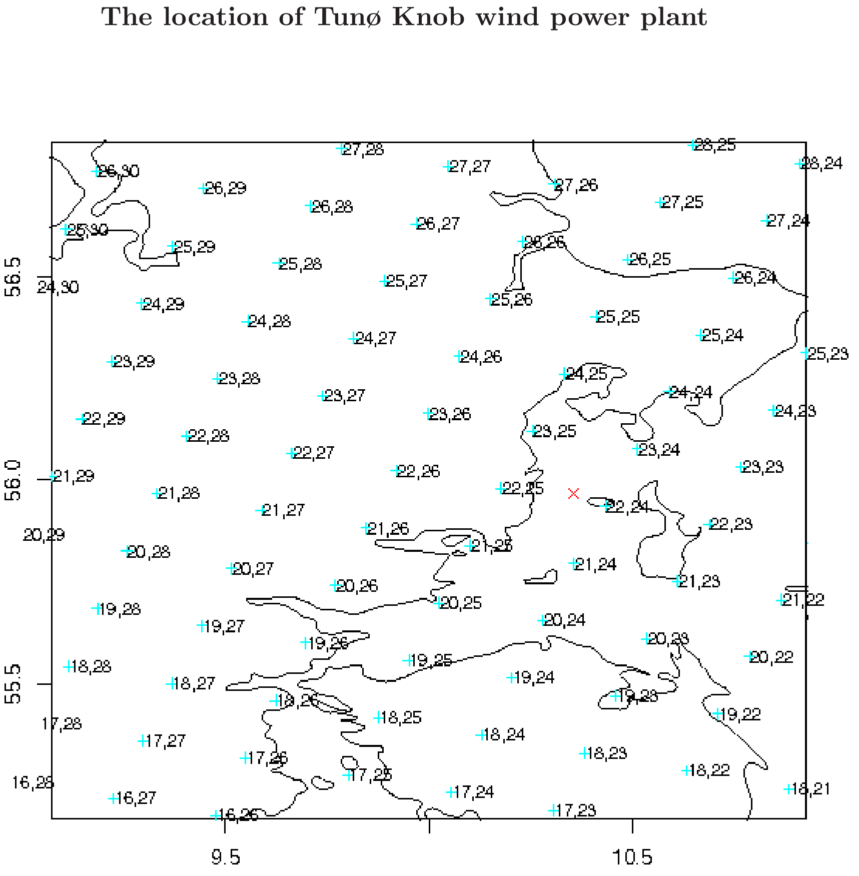

Figure 4.1: The grid points where meteorological forecasts are available. The Tunø Knob wind power plant is marked with the red x.

The aim here is to get a linear regression model as discussed in Chapter 2. The final step is to approximate the functions f_j with the spline basis functions described in Chapter 3. Hence we write

$$f_j(x_j) = \sum_{k=1}^{n_k} b_{jk}(x_j)\beta_{jk} \tag{4.3}$$

where b_{jk} is the k'th basis function of variable x_j and $\beta_{j,k}$ is the parameter to estimate ($\alpha = b_0\beta_0$ will denote the intercept and b_0 is therefore constant equal to one). The matrix formulation of the problem is now

$$\hat{Q}(\mathbf{x};\tau) = [\mathbf{b}(\mathbf{x}_i)^T] = \mathbf{X}\hat{\beta}(\tau) \tag{4.4}$$

where $\mathbf{b}(\mathbf{x}_i)$ is a vector and consists of known spline basis functions of the meteorological forecast. With this we are in the setting of linear quantile regression. In the rest of this presentation $\hat{Q}(\mathbf{x},\tau)$ will refer to this model.

To make sure that the solutions are unique the constraint $f_j(\min(x_j)) = 0$ is imposed for non-periodic functions, and for periodic functions the constraint is $\int_P b_{jk}(x_j) = 0$, where P is the period. The first column of the matrix $\mathbf{X}$ will as mentioned above, be a vector of ones. β will be given by $\beta = [\beta_0, \beta_{1,1}, ..., \beta_{1,n_k}, \beta_{2,1}, ..., \beta_{p,n_p}]$.

When visualizing the quantiles it is important to note that, one look at cuts in p-dimensional hyper planes in a $p+1$ dimensional space and crossings of the estimated quantile hyper planes can (and will) occur. So if we want to visualize the quantiles we have to make a choice of where to look. In this presentation the choice is “typical” values. Thus to see the effect of x_j choose $x_k = x_k^0$, and then plot $Q(x_j, x_1^0, ..., x_{j-1}^0, x_{j+1}^0, ..., x_k^0; \tau)$. Here the choice for non periodic variables is $x_k^0 = E(x_k)$ for $k \neq j$

Crossings should only occur in remote areas of the data set, and therefore these should not occur with the above choice of x_k^0. We saw in Chapter 2 that there will be no crossings at the mean value of the explanatory variables. This result will strictly speaking not apply here since the result apply to the average over the spline functions not the spline of the average.

When looking at plots of the quantile curves we should be careful when drawing conclusions. The focus should mainly be on the variation of the curves - not the level. The variation of the curve gives an indication of the explanatory power of that variable, while the level can be due to other parameters.

When quantiles have been modeled, some methods to measure performance should be available. The subject of such performance parameters is studied in the next section.

4.3 The Performance of Quantiles

When we model the quantiles it is of course important to have some performance parameters as a basis for model selection. This section will discuss some performance parameters from the literature on quantile regression. The main part of the literature used for this section comes from quantile regression in the context of wind power prediction.

The performance parameters presented here assumes that the data set has been divided into a test set and a training set, the training set is used for estimating the model parameters and the test is used to give some values of performance.

The loss function on the training set is also an option. This is also what is used in e.g. [11] to develop statistical tests for quantile regression. If performance parameters are calculated on the training, then we have to take the number of explanatory variables into account, because most performance parameter will get better on the training set, as we add more explanatory variables.

Here we will focus on the performance on the test set, the loss function is of course also a parameter we could look at. Since we know that the true quantile minimizes the loss function, a better loss function should tell us that we are in some sense closer to the true solution.

The meaning of the loss function is however not that clear and other performance parameters are therefore often used. This presentation will discuss the commonly used performance parameters for quantile regression and in the case of reliability a local measure is also suggested. A general problem for the measures is that we do not have a clear answer to the question of when is a performance parameter good enough or when two performance parameters are equal in some statistical sense.

4.3.1 Reliability

We know that minimizing the loss function described in Chapter 2, give the required quantile on the training set. Therefore it is of course more interesting what the performance is on a test set. We want the model to be able to predict the true quantile on the test set, a natural performance parameter is therefore the quantile that the model actually produce on the test set. Therefore a performance parameter is the fraction of responses below the model. This will be refereed to as reliability or overall reliability of the model and this should of course be close to the required quantile (τ). At times the difference or absolute difference to the required quantile will also be used.

Even if we see a good reliability for a model, this does not imply that this feature also holds locally. Therefore we should have some kind of local measure of reliability. The problem here is that with many explanatory variables the notion of local is problematic. If we want some local measure in all directions then there will be very few observations in each cell, thus a local measure in all directions is not really meaningful unless the number of observations is extremely large. To solve this we look at the reliability in one direction at a time, i.e. the residuals are grouped by some variable and the reliability is calculated as a function of this variable. A more formal definition is given in Definition 4.1 below. This definition also gives the precise recipe for calculating the local reliability.

Definition 4.1 *For a set of variables* $\mathbf{z}$ *corresponding to the observations in* $\mathbf{y}$ *let* $\mathcal{A}$ *be a perturbation of the index set* $1, 2, ...N$*, s.t.* $z_{\mathcal{A}_1} \leq z_{\mathcal{A}_2} \leq ... \leq z_{\mathcal{A}_N}$ *i.e.* $\mathbf{z}_{\mathcal{A}}$ *is the order statistics of* $\mathbf{z}$*. For a given value of* z *let*

$$p(z) = \begin{cases} 1 & \textit{if} \quad z_{\mathcal{A}_{(1)}} > z \\ \textit{arg}\max_i\{z_{\mathcal{A}_{(i)}} \leq z\} & \textit{otherwise} \end{cases} \tag{4.5}$$

and set $N_1 = \max\{1, p(z) - \lceil wN \rceil\}$*,* $N_2 = \min\{N, p(z) + \lceil wN \rceil\}$*, where* $0 \leq w \leq 1$*. The local reliability* $q(z)$ *is then defined as*

$$q(z; w, \tau) = \frac{1}{N_2 - N_1 + 1} \sum_{i=N_1}^{N_2} I(y_{\mathcal{A}_i} \leq \hat{Q}(\tau|\mathbf{x}_{\mathcal{A}_i}) \tag{4.6}$$

The parameter w is something like a bandwidth for the local reliability. If $w = 0$ then $q(z; w, \tau)$ is the indicator function $I(r_{\mathcal{A}_i} \leq 0)$ and with $w = 1$ it is constant equal to the overall reliability. So w indicates how local the local reliability curve is, if w is chosen too small then we can't see any trends because of variation in q and if w is to large these trends disappear. Definition 4.1 ensure, that we know how many elements q is based on. The price we pay for this is that the interval length in the direction z that q is based on will be a function of the z. In the following $w = 0.1$ is used.

From Definition 4.1 the local reliability is defined for all $z \in \mathbb{R}$, but it is constant outside the interval $[z_{(1)}, z_{(N)}]$ and it is constant between observations. The performance parameter $q(z)$ can now be plotted as a function of the explanatory variable z. This can as we will see later on reveal quite serious problems of the quantile estimator even though the overall reliability is good. Since the local reliability is a set of numbers it can be quite hard to compare local reliability for different quantile curves, therefore the squared distance over observations between $q(\mathbf{z}; w, \tau)$ and $\tau(\mathbf{z}) = \tau$ will also be used as a performance parameter, this is denoted $d(q(z;\tau))^2$ and calculated by

$$d(q(z;\tau))^2 = \frac{1}{N}\sum_i (q(z_i; w, \tau) - \tau)^2 \tag{4.7}$$

This will be refereed to as "*reliability distance*". This measure punish both variability and bias. Therefore it might be a little misleading, since the bias will punish in every direction. An alternative would be to look at the variance and the overall bias. It is however not obvious how the variance in each direction should be weighted, if we want to sum the variation for more directions. Even if it is not clear exactly how to interpret these distances, it is clear that if a model has better reliability distances in all directions then we should prefer this model.

To see the bias variance issue for the local reliability distance look at

$$\begin{aligned} d(q(z;\tau))^2 &= \frac{1}{N}\sum_i (q(z_i; w, \tau) - \tau)^2 \\ &= \frac{1}{N}\sum_i (q_i - E(q) + E(q) - \tau)^2 \\ &= \frac{1}{N}\sum_i ((q_i - E(q))^2 + (E(q) - \tau)^2 + 2(q_i - E(q))(E(q) - \tau)) \\ &= V(q) + (E(q) - \tau)^2 + 2(E(q) - \tau)\frac{1}{N}\sum_i (q_i - E(q)) \\ &= V(q) + (E(q) - \tau)^2. \end{aligned} \tag{4.8}$$

With this we see that $d(q(z,\tau))^2$ consists of the variance of q plus the squared distance from $E(q)$ to τ. This is what was referred to as bias. One argument for choosing $V(q)$ and (overall reliability$-\tau$)^2 could be that these variances and the squared bias could be added, and then give *one* reliability number. This would however mean that a model with high bias and low variance of the local reliability would perform better and better as we a add more and more directions.

Definition 4.1 have the effect that we know how many points each local reliability is based on, but we do not know how wide the interval is when measured as a function of the variable z_j. Therefore it can happen that local properties is not detected this procedure.

Before going on with defining other performance measures it should be noted that these measures may not be meaning full unless we have a good reliability. So in this sense reliability is the most important measure of performance.

4.3.2 Sharpness

Sharpness is a measure for how far symmetric (around 0.5) quantiles are separated. Given a good reliability we want these intervals to be as narrow as possible. Here the Inter Quartile Range (IQR), i.e. the difference between the 75% and 25% quantiles is used. This is the same as is used by Nielsen in [1] and [15]. This quantity should be minimized, since we would like to have small predicted intervals. Two different measures can be used here, namely the mean of IQR and the median of IQR. They do however seem to be quite close so it probably doesn't matter so much which one is used.

4.3.3 Resolution

Resolution is a measure of how well a model distinguish between different situations. The measure used for this is either the standard deviation of IQR or the MAD (Mean Absolute Deviation). In [15] both of these are used, they don't seem to differ very much, therefore we only consider the standard deviation here. In [1] the 5% and 95% quantiles of IQR is used as such a measure as well, i.e. we should look at the difference between these.

To maximize a standard deviation is problematic since random variation will be awarded with such a measure. I.e. if we have the true quantile and then add white noise then the model will perform better without affecting reliability. This stress the point that the measures should not be considered alone.

4.3.4 Spread / Skill Relationship

Spread / skill relationship refers to the relationship between some point value given by the forecasting system and some observed values. In [15] the absolute error from WPPT (Wind Power Prediction Tool, see [19]) is plotted as a function of the forecasted IQR. The mean or quantiles of the absolute error should now be an increasing function of IQR. This approach only say that this statistics (quantiles or mean) should be an increasing function of IQR. There is nothing about the optimal relationship in this procedure.

Here a different method is used. The residuals is grouped by the forecasted IQR, and the 50% sample quantile of observed errors in this group is plotted as a function of the 50% quantile of IQR in this group. With this definition of the spread / skill relationship the curve should be close to a straight line with a slope of 45°.

4.3.5 Skill Score

A skill score is one numerical value that gives a summary of the performance of the forecast. Such a value should collect the analysis from the above in one number, and award the best forecast or the forecast closest to the true model.

In [17] Gneiting defines the expected score under the probability P and the forecasted quantiles $\hat{q}_1, ..., \hat{q}_k$ as

$$S(\hat{q}_1, ..., \hat{q}_k; P) = \int S(\hat{q}_1, ..., \hat{q}_k; y) dP(y) \tag{4.9}$$

i.e. we use q_i as the forecast of P. A scoring rule S is (again following [17]) said to be *proper* if

$$S(q_1, ..., q_k; P) \geq \int S(\hat{q}_1, ..., \hat{q}_k; y) dP(y) \tag{4.10}$$

where $q_1, ..., q_k$ denote the true quantiles and $\hat{q}_1, ..., \hat{q}_k$ are any real numbers. Gneiting show that any rule of the form

$$S(\hat{q}_1, ..., \hat{q}_k; y) = \sum_{i=1}^{k} (\tau_i s_i(\hat{q}_i) + (s_i(y) - s_i(\hat{q}_i)) I(y \leq \hat{q}_i)) + h(y) \tag{4.11}$$

where τ_i is the prediction levels ($\tau_i \in (0, 1)$), s_i are non decreasing functions and h is any function, is proper for the quantiles τ_i. As is noted in [16] this scoring rule is a generalization of the loss function for the quantile regression. To see this set $s(x) = x$ and $h(x) = -\tau x$, then the scoring rule becomes

$$\begin{aligned} S(\hat{q}, y) &= \tau\hat{q} + (y - \hat{q}) I(y \leq \hat{q}) - \tau y && (4.12)\\ &= \tau \mathbf{x}^T \hat{\beta} + (y - \mathbf{x}^T \hat{\beta}) I(y \leq \mathbf{x}^T \hat{\beta}) - \tau y && (4.13)\\ &= -\tau r + r I(r \leq 0) && (4.14)\\ &= r(I(r \leq 0) - \tau) = -\rho_\tau(r) && (4.15) \end{aligned}$$

where ρ_τ is the loss function introduced in the beginning of Chapter 2. By (4.15) it we see that maximizing the skill score defined here correspond to minimizing the loss function defined in Chapter 2. Therefore the loss function is used as a skill score throughout this presentation. We should just keep in mind that we want to minimize this not maximize it.

In [17] Gneiting uses (4.11) to derive an interval score for a central $(1-\alpha) \times 100\%$ prediction interval. This is defined by calculating the score for the forecasts of

$\hat{q}_1$ and $\hat{q}_2$, with $\tau_1 = \frac{\alpha}{2}$, $\tau_2 = 1 - \frac{\alpha}{2}$, $s_1(x) = s_2(x) = 4x$ and $h(x) = -2x$. This gives the interval score

$$S_\alpha(\hat{q}_1, \hat{q}_2, y) = \begin{cases} -2\alpha(\hat{q}_2 - \hat{q}_1) - 4(\hat{q}_1 - y) & \text{if} \quad y \leq \hat{q}_1 \\ -2\alpha(\hat{q}_2 - \hat{q}_1) & \text{if} \quad \hat{q}_1 \leq y \leq \hat{q}_2 \\ -2\alpha(\hat{q}_2 - \hat{q}_1) - 4(y - \hat{q}_2) & \text{if} \quad y \geq \hat{q}_2 \end{cases} \tag{4.16}$$

This is an intuitively appealing form, but to write it like this ignores the fact that there can be crossings, since if $\hat{q}_1 > \hat{q}_2$ (4.16) is not unique. This is however not something that stems from the scoring rule, but only from writing it in the form (4.16). Thus a note for (4.16) is that the assumption $\hat{q}_1 \leq \hat{q}_2$ should be added. If $\hat{q}_1 > \hat{q}_2$ then $\hat{q}_1$ and $\hat{q}_2$ swap places in (4.16). Of course there is nothing intuitively appealing any more, but quantile crossings are by nature not intuitively appealing anyway. The final remark on the interval score is that by rewriting it in a similar way as in equation (4.15) we get

$$S_\alpha(\hat{q}_1, \hat{q}_2, y) = -4(\rho_{\tau_1}(y - \hat{q}_1) + \rho_{\tau_2}(y - \hat{q}_2)) \tag{4.17}$$

again this shows that we can use the sum of the loss function as well as the scoring rule, and again note that we minimize this.

As we saw in Chapter 2 crossings can (and will) occur, so this will also be considered a performance parameter and something we aim to avoid.

Now that we have some ways of evaluating the performance of quantile models and the tools to understand the quantile regression with splines, we go on by using a data set from Tunø Knob to analyze some quantile regression models.

As should be clear by now it is not simple to quantify the performance of a quantile or a set of quantiles. There are many different parameters to look at. We will discuss these performance parameters on the basis of the Tunø data set and the models developed in the rest of this chapter and the next chapter. Before doing so the data is presented.

4.4 The Tunø Data Set

The data set consists of prediction error (**pow.pe**) from WPPT (Wind Power Prediction Tool), and meteorological data from DMI's (Danish Meteorological Institute) meteorological forecasting system DMI-HIRLAM. The data set is the same as is used in [1] but with some extra meteorological data. These extra data are only avaiable from archive, this essentially means that they are not given as often as the rest.

The installed power at Tunø Knob is 5000kW and the prediction from WPPT is in the range from 0 to 5000kW so this is also the range of (absolute) prediction errors.

The aim is to model conditional quantiles of **pow.pe** conditioning on the explanatory variables, which are

pow.fc: Forecasted power from WPPT (kW).

horizon: The prediction horizon (hours).

ad: Forecasted air density from DMI-HIRLAM (g/m^3).

fv: Forecasted friction velocity (m/s).

wd10m (wd30m): Forecasted wind direction 10 (30) meter above ground level from DMI-HIRLAM in degrees.

ws10m (ws30m): Forecasted wind speed at 10 (30) meter above ground level (m/s).

wdL··: Forecasted wind direction in model level ··, the levels in the data set are 31, 38, 39 and 40 (degrees), data from archive.

wsL··: Forecasted wind speed in model level ··, the levels in the data set are 31, 38, 39 and 40, data from archive.

tkeL··: Forecasted turbulent kinetic energy in model level··, the levels in the data set are 38, 39 and 40, data from archive.

r··: Meteorological risk index of data .., these are given for the explanatory variables **ad**, **wd10m**, **wd30m** **ws10m**, **wd30m** and **fv**.

The model levels are different levels of the atmosphere, the higher the level number the closer to the ground level. The risk index is the same as was used in [1] which is based on the definition in [12].

The forecasted power is given every 15th minute, the meteorological data which have a risk index is given every hour. The rest of the meteorological data is only accessible every 3rd hour (the data from archive). The meteorological data from DMI is given in a number of grid points throughout Denmark. The grid points around Tunø is seen in Figure 4.1. To get the forecast at the location of Tunø Knob, a bilinear interpolation between the four points around the location is performed. To get the forecast every 15th minute a linear interpolation between time points is performed.

The forecasted power is based on meteorological forecast at the time point 06, and meteorological forecasts made at this point. The meteorological data from archive is only given at the times 12 or 00. In the present context the ones from 00 is used, which imply that horizon only apply for forecasted power and some of the meteorological data. The meteorological risk index is presented in [12]. This requires at least two forecasted values and since all meteorological values are given with 48 hours horizon this is only given for the meteorological forecast which are given at 06.

The data is divided into a train and a test set. The train set is the period from January 1th to June 1th 2003, the test set is the period from June 1th to October 31th 2003. The training set consists of 10658 data points and the test set consists of 11095 data points. At September 2rd DMI introduced a model change in DMI-HIRLAM (the forecast system), which was expected to have large influence on **wd10m**. The test set is therefore further divided into two parts, before and after September 2. These sets have 6861 and 4234 data points respectively.

Figure 4.2 shows the pairwise scatter plot of some of the explanatory variables. The plot shows that forecasted power and wind speed is very correlated, and that in this sense friction velocity can be thought of as a wind speed. Thus we have to choose either the forecasted power or a wind speed as explanatory variable. The plot does not show the scatter plot between meteorological data from different levels. These are however as could be imagined also very correlated.

The aim is to predict quantiles of a horizons between 12 and 36 hours, with the combination of splines and quantile regression described previously. However the available data from WPPT has only horizons between 18 and 36 hours, these are therefore the horizons studied in this presentation.

Figure 4.3 shows histograms of some of the explanatory variables. The rest of the variables is showed in Appendix B. It is seen that there are areas in all the data with very few observations. Especially the turbulent kinetic energy have very few observations with high values. To deal with this Turbulent kinetic energy is log transformed, the histogram for this is also shown in Figure 4.3.

4.4.1 The Three Periods

As was mentioned earlier, DMI introduced a change in the prediction model on September 2nd 2003. This change was expected to have a large impact on **ws10m**. The data is therefore divided into the two test periods and a training period. Figure 4.4 shows histogram plots of wind speed, predicted and observed

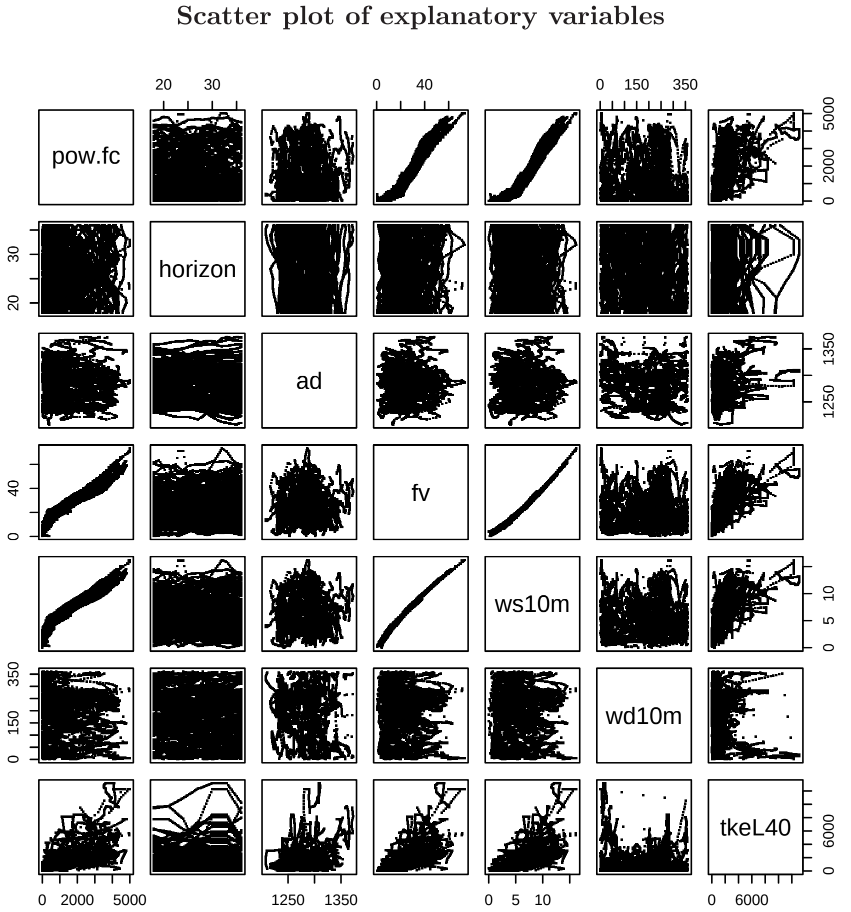

Figure 4.2: Pairwise scatter plot of meteorological data and the forecasted power curve from WPPT for the training data.

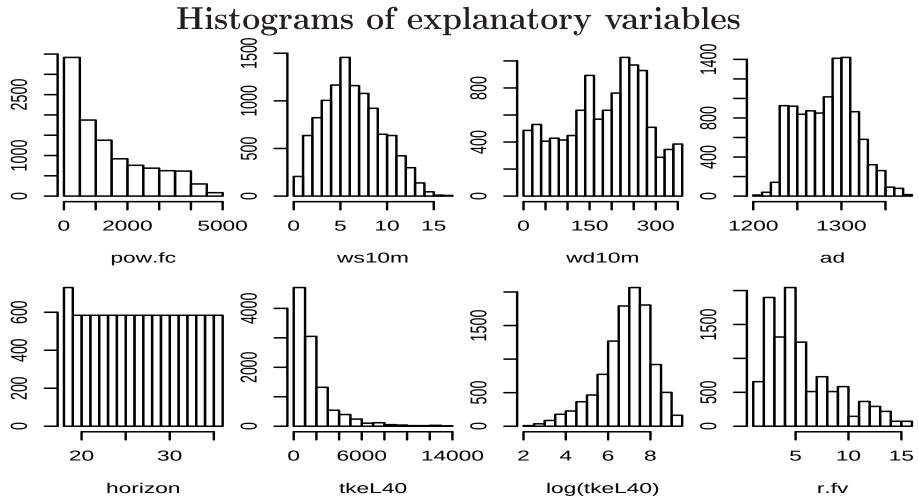

Figure 4.3: Histogram of some of the explanatory variables. In appendix B histograms of all explanatory variables are displayed.

Key numbers for the three periods

Data	E(**ws10m**)	E(**pow.fc**)	E(**pow.obs**)
Train	6.37	1408.54	1519.88
Test	6.37	1290.32	1153.37
Test 1	6.00	1146.24	1000.23
Test 2	6.98	1530.44	1408.57

Table 4.1: Mean of wind speed and forecasted and observed power for the three periods in the dataset.

power, and the prediction error from WPPT for the three periods.

The figure shows some differences especially as was expected in predicted wind speed. Table 4.1 gives some key numbers for the data presented in the figure. The table essentially tells us the same as the figure, namely that there are differences in the wind speed and predicted and observed power. The figure and the table give some support to the point of analyzing each of the test period separately. As will be seen there are differences in the performance of the two periods when we use **pow.fc** as input. It is however not clear that the variation could not be explained as annual variation in the data.

With the data presented we will go on and present the model.

Histogram plots for the three different periods in the model

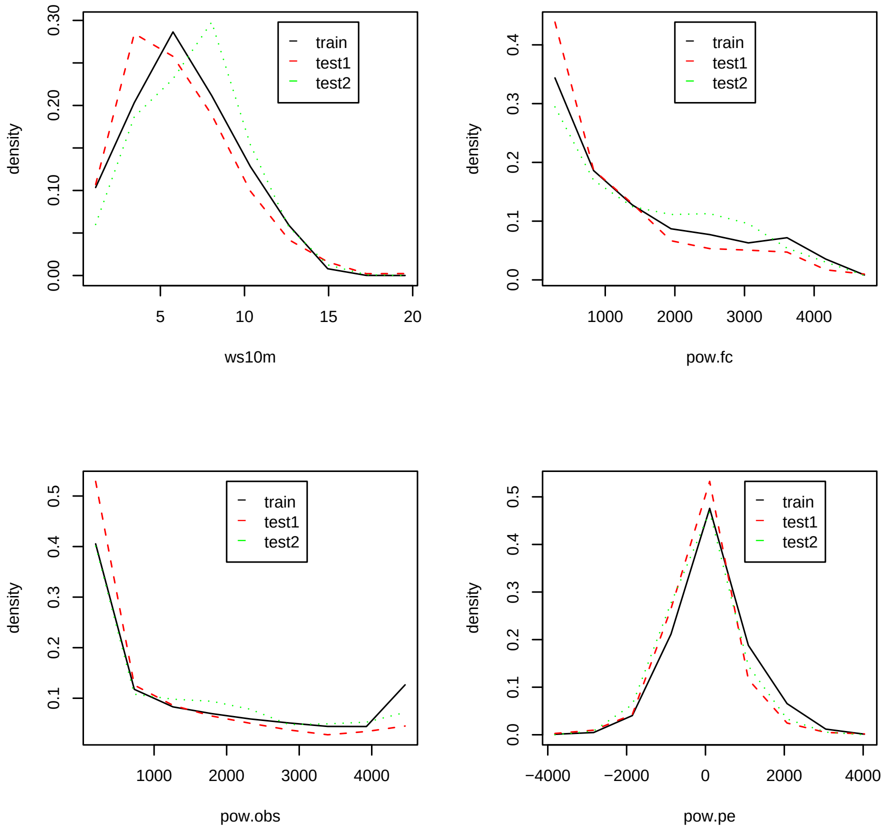

Figure 4.4: Histogram plot for the 3 different periods in the data.

4.5 Quantile Regression Models for the Tunø Knob Data Set

The combined quantile regression and spline models are now used on the Tunø data set. In this chapter we only consider the 25% and 75% quantiles. These are the same as considered in [1] and the model in Figure 4.5 is also considered in this article. The visualizations in [1] is a little different than the one chosen here.

In [1] the components of the additive model is plotted one by one. This is the same as choosing cuts (see Section 4.2) as points where each of the other functions in the additive model is zero. In the natural spline case this is at the leftmost knot and for the periodic splines this depend on the parameters.

The method used here is to choose "*typical*" values of the other parameters, these typical values are the mean of each of the explanatory variables except for the wind direction, where the mean does not really make sense. Here the choice is 250 degrees, which is taken on basis of the histograms in Figure 4.3.

An overview of the construction of the models shown in Figure 4.5-4.8 is given below.

Model 1: The knots are placed as 10% quantiles of the explanatory variables in the training set, except in the wind direction where 8 knots have been placed with equal distances. This model is the only model with air density as input. The number of parameters to estimate is 39.

Model 2: Air density has been taken out of Model 1 and fewer knots have been placed. These knots have been placed manually to capture the overall variation in Model 1. This model leave out some of the very local variation in Model 1. The number of parameters to estimate is 16.

Model 3: The forecasted power has been replaced with the wind speed in level 40 as input. Wind direction is also from this level. In addition turbulent kinetic energy and the risk index of friction velocity have been taken into account. The model only has 14 parameters and is in this respect the most simple model considered here. The knots have been placed manually, by trial and error.

Model 4: As Model 3 but with some more knots. The number of parameters to estimate is 24.

4.5.1 Remarks on Figure 4.5-4.8

Figure 4.5-4.8 show the quartile curves for Model 1-4, along with the local reliability in the direction of each of the explanatory variables for the training and the test period. In the case of Model 1 also for the periods Test 1 and Test 2. Since these seem to be alike the local reliability for Test 1 and 2 has been left out in the three other plots.

The IQR as a function of forecasted power would be expected to be large for moderate wind speeds and small at the endpoints of the interval. The reason for this is that the power curve, i.e. the forecasted power as a function of wind speed, is steep for moderate wind speeds and becomes more constant close to the endpoints. Thus in the middle of the forecast interval small changes in the wind speed will course large changes in the wind power production.

This is also the qualitative behavior of Model 1 and 2, which uses the wind power as an explanatory variable. There is however not a clear explanation for the shoulder at around $700kW$ for Model 1. This shoulder is also seen in the local reliability on the test period, so it seems that the model might be over fitted here. This behavior can be controlled by the knots, and it seems that we have to many knots for low forecasted power in Model 1. Some of these knots have been taken out in Model 2 and it is seen from the two figures that Model 2 perform a little better on the test set.

It is seen that there is some variation of the local reliability. For Model 2 this is even seen for the training period, because the model is not able to describe the local behavior in the training set any more. In Model 2 the shoulder have moved from the 75% quantile curve to the local reliability for the training period. The general picture of the reliability on the training period is that there is some variation. It is of course much smaller that for test period. We saw earlier that quantile regression split response variable in two sets according to the required probability. This is seen not to be a strict local property.

The horizon does not seem to explain much variation of the curves in any of the models.

The wind direction seems to explain much more variation in Model 1 and 2 than in Model 3 and 4. The two different set up also uses different wind directions, which could be the explanation for this. The wind direction have a greater explanatory power for the 75% quantile than for the 25% quantile, but there is a large variation in the local reliability.

It seems that there is not really a good physical explanation for the estimated

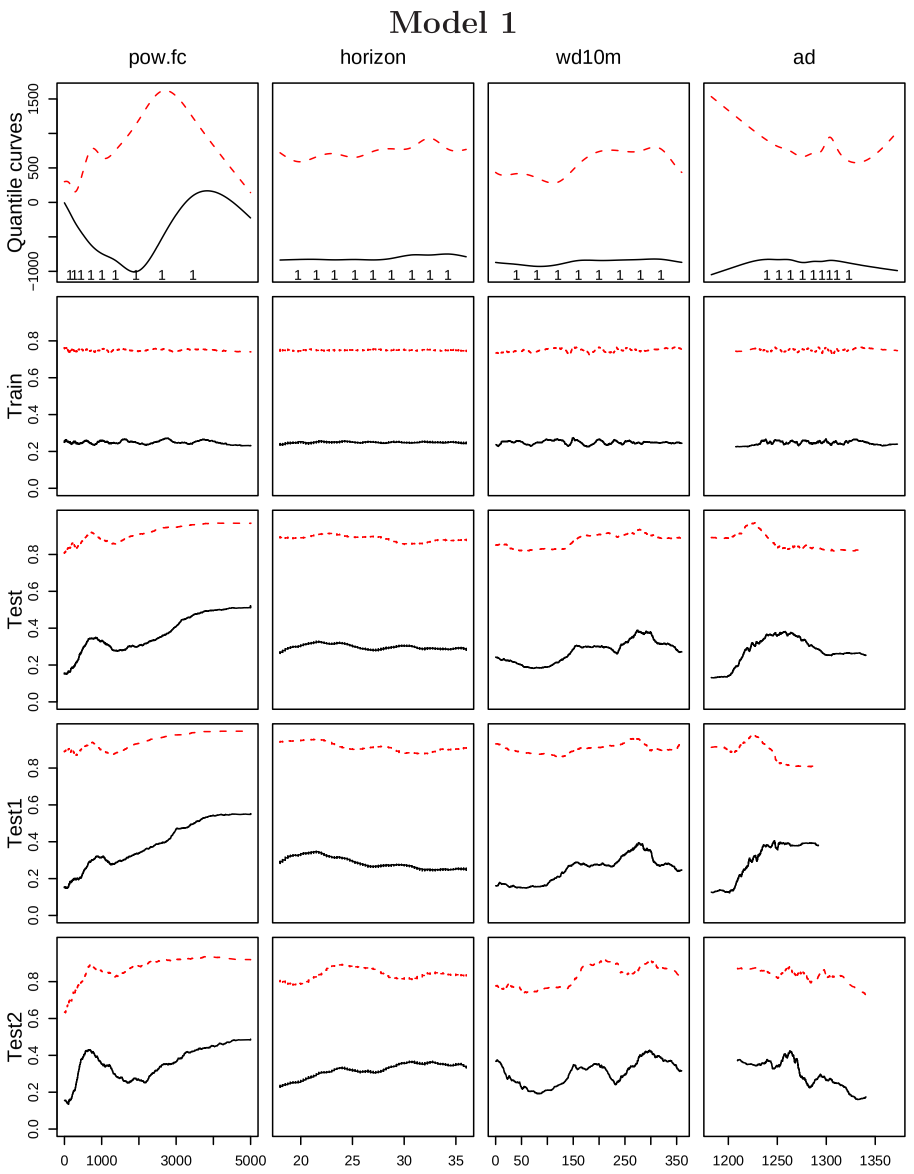

Figure 4.5: Plots of a quantile model for the Tunø data set, the first row shows the dependencies of the explanatory variables. Second row shows local reliability for the training data. Third, fourth and fifth row show the local reliability for the whole test period, test period one and two respectively. The rugs at the first axis of row one indicate the placement of the knots.

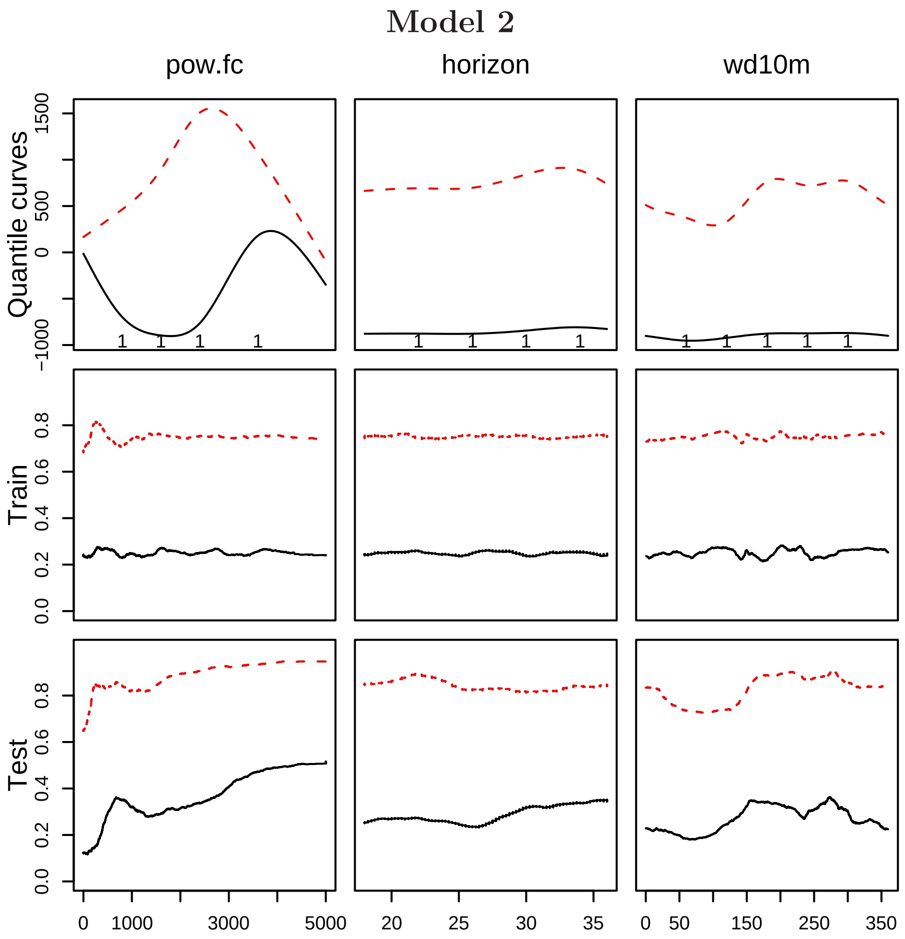

Figure 4.6: Plots of a quantile model for the Tunø data set. The first row show the dependencies of the explanatory variables. Second row shows local reliability for the training data. Third row shows the local reliability for the test period. The rugs at the first axis of row one indicate the placement of the knots.

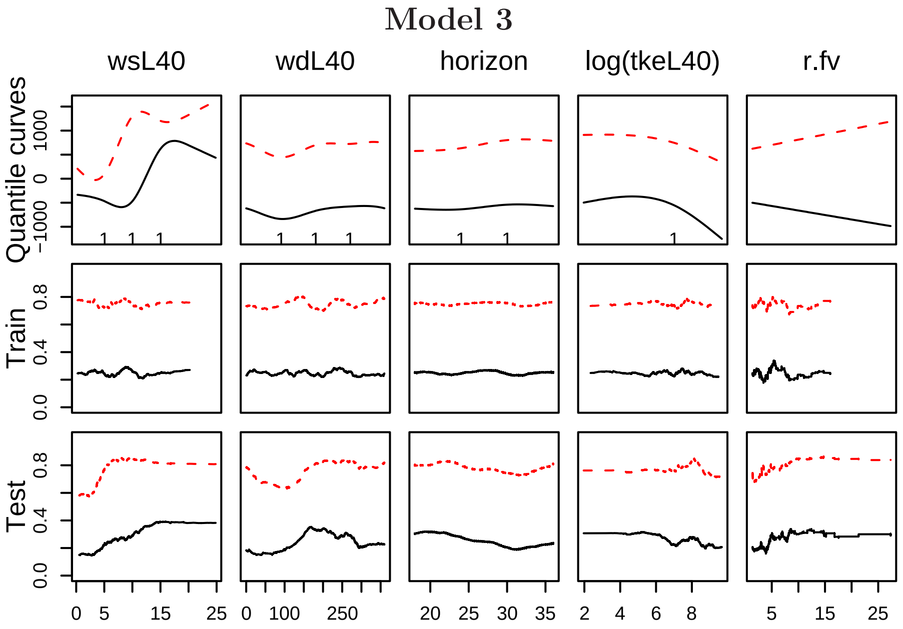

Figure 4.7: Plots of a quantile model for the Tunø data set. The first row show the dependencies of the explanatory variables. Second row shows local reliability for the training data. Third row shows the local reliability for the test period. The rugs at the first axis of row one indicate the placement of the knots.

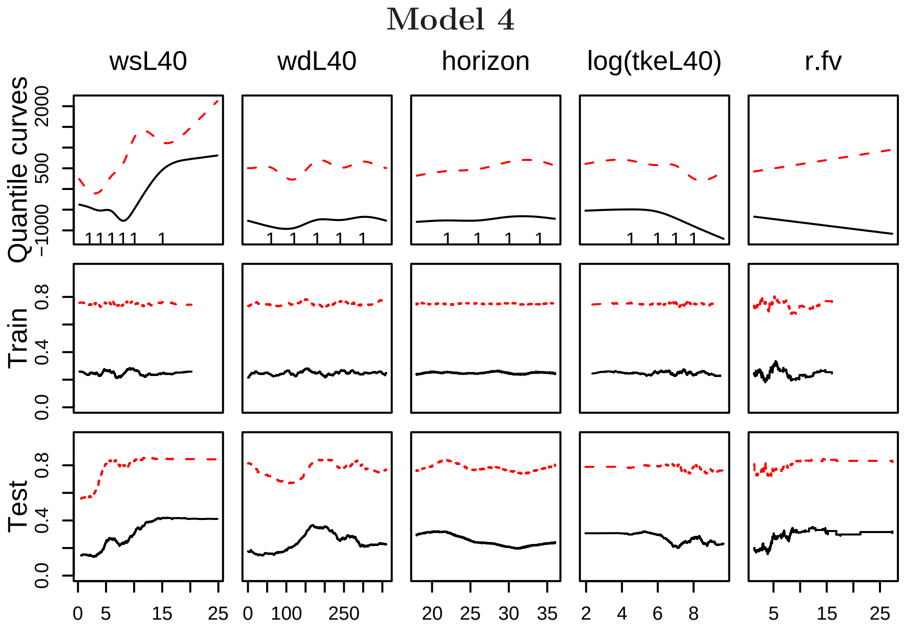

Figure 4.8: Plots of a quantile model for the Tunø data set. The first row show the dependencies of the explanatory variables. Second row shows local reliability for the training data. Third row shows the local reliability for the test period. The rugs at the first axis of row one indicate the placement of the knots.

quantile curve for the air density. Another point which is illustrated by Figure 4.5 (see the reliability for the different periods) is that there is a large annual variation in the air density. This means that the data in the training set does not span the range of data in the test set. When something like this happens linear regression models will have difficulties, and the spline models probably make things worse.

Model 3 and 4 use wind speed instead of forecasted power. The local reliability of this is also quite bad on the test set. The surprising behavior at high wind speeds, can be explained by the fact that there are no observations in this area for in the training set.

The curve for the turbulent kinetic energy varies a lot and at the same time the local reliability on the test set is fairly constant. This is actually a property that we are looking for.

The risk index of the meteorological data is not taken into account in Model 1 and 2. In [1] the dependence of the risk index for Model 1 was found to be small. This has also been checked here for these models and the same result was found.

Models with different knot placement have been tried out, but it does not affect the models greatly. For Model 3 and 4 the risk index shows a behavior which we would expect, but this also seems to have quite large annual variation, and the local reliability plot shows large variations.

Appendix B go through the model with different explanatory variables in a more systematic way. The models in Appendix B is with wind speed instead of forecasted power.

4.5.2 Hypothesis Test

The question that have to be answered if we want to make hypothesis tests is: can we assume that the requirements for using the hypothesis test presented in Section 2.5 is fulfilled? The answer to this question is no. To argue for this conclusion look at Figure 4.9, where a one lag correlation plot of the residuals from Model 4 is presented. The conclusion from this is clearly that we can not assume that these are iid. The assumption should then be that the residuals can be modeled by

$$r_i = \mathbf{x}_i^T \gamma e_i \tag{4.18}$$

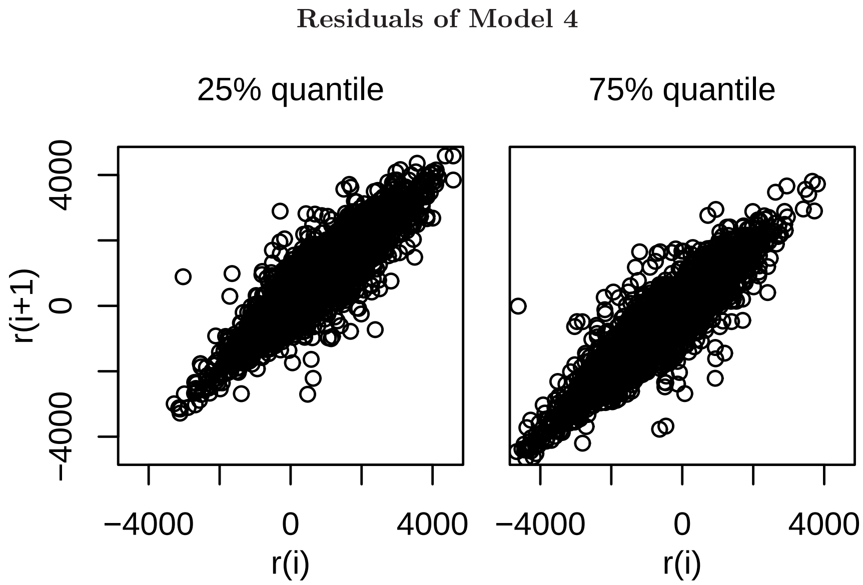

Figure 4.9: Correlation plot for Model 4.

With e_i iid, we do not really have any reason to assume this. In fact the prior assumption would be that the e_i's are correlated. The reason for this prior assumption is that, given the meteorological forecast is wrong at time i then it would also be expected to be wrong (in the same way) at time $i + 1$. Therefore we would expect that the e_i's are positively correlated. Therefore to do any hypothesis testing we should prove that the assumption in 4.18 is true or likely. Since it is not clear how to prove something like this, hypothesis testing will not be considered further in this presentation.

Plots like Figure 4.9 have been considered for the other models and the result is approximately the same.

4.5.3 Reliability

Table 4.2 gives the overall reliability for Model 1-4. From the perspective of Table 4.2 the obvious choice would be Model 3 or 4. Model 3 is very close to the required reliability for the 25% quantile, but we already saw that the local reliability as a function of the explanatory variables is not very convincing. Figure 4.10 shows the local reliability for the four models plotted together as functions

Reliability for the quantile models in Figure 4.5-4.8

Period	Model	1	2	3	4
Test	Below **75%**	88.6%	84.2%	**78.0%**	78.3%
Test	Below **25%**	29.7%	28.7%	**25.3%**	**25.3%**
Test 1	Below **75%**	91.3%	83.7%	**76.5%**	77.4%
Test 1	Below **25%**	27.7%	27.9%	**25.4%**	25.6%
Test 2	Below **75%**	84.2%	85.0%	80.5%	**79.6%**
Test 2	Below **25%**	33.0%	30.0%	**25.1%**	24.8%

Table 4.2: The table contains the overall reliability for the models shown in Figure 4.5- 4.8 The best parameter in each row is marked with bold face letters.

of forecasted power, horizon and time. The plot also shows the reliability of the center 50% interval. We see that the very nice results in the overall reliability is not a local property, and from these plots it is not obvious that, e.g. Model 3 is better than Model 2. The two models are simply wrong in different ways.

Table 4.3 gives the reliability distance in the direction of pow.fc, horizon and time. The table gives the reliability distance for the first and third quartile and for the distance of the two quartiles, i.e. how close to 50% of the observations fall between the two quartiles. This number does not punish translations of the interval, i.e. a true interval of 0 to 50% quantile would give the same number as the true interval of the 25% to 75% interval. From the perspective of this table we should choose Model 4. The row with total reliability is $d^2_{total}(x) = (d(q(x,0.25))^2 + d(q(x,0.75))^2 + d_{be}(x)^2)/3$ with $d_{be}(x)^2$ being the reliability distance for the IQR.

As can be seen from Figure 4.10 none of the models have a very convincing performance of local reliability. Here we will try to solve this with an adaptive model. Before we go on with this, a discussion of the other performance parameters will be given, even though we should have in mind that the reliability (at least locally) is quite bad.

4.5.4 Skill Score and Crossings

Table 4.4 gives the average loss function and the number of crossings for the four models. From the skill score point of view the conclusion is clearly that we should choose Model 2 for the 25% quantile and Model 4 for the 75% quantile. The interval score (the sum of the loss functions) tells us to choose Model 2. This is a different conclusion than what we got from the reliability discussion, where the conclusion was to use Model 3 or 4.

Numbers related to the reliability for the models shown in Figure 4.5-4.7

Model	1	2	3	4
$d(q(\mathbf{pow.fc}, 0.25))$	**0.102**	0.110	0.142	0.145
$d(q(\mathbf{pow.fc}, 0.75)$	0.144	0.114	0.119	**0.102**
$d(\mathbf{q}, \mathbf{pow.fc}, 0.5))$	0.107	0.090	**0.081**	0.085
$dq_{total}(pow.fc)$	0.119	**0.105**	0.117	0.113
$d(q(\mathbf{hor}, 0.25))$	0.050	0.052	0.043	**0.041**
$d(q(\mathbf{hor}, 0.75))$	0.137	0.095	0.042	**0.040**
$d(q(\mathbf{hor}, 0.50))$	0.089	0.073	**0.034**	0.039
$dq_{total}(\mathbf{hor})$	0.098	0.076	**0.040**	**0.040**
$d(q(\mathbf{time}, 0.25))$	0.080	0.055	0.047	**0.046**
$d(q(\mathbf{time}, 0.75))$	0.143	0.096	0.054	**0.050**
$d(q(\mathbf{time}, 0.5))$	0.127	0.068	0.046	**0.040**
$dq_{total}(\mathbf{time})$	0.120	0.075	0.049	**0.046**

Table 4.3: Reliability distance in the direction of some explanatory variables. These all refer to the test set. The best parameter in each row is marked with bold face letters.

Skill score and crossings for the quantile models in Figure 4.5-4.7

Period	Model	1	2	3	4
Test	$\overline{\rho}_{0.75}(\mathbf{r})$	297.3	270.7	254.4	**253.2**
Test	$\overline{\rho}_{0.25}(\mathbf{r})$	219.7	**217.6**	241.0	243.5
Test	$\overline{\rho}_{0.75}(\mathbf{r}) + \overline{\rho}_{0.25}(\mathbf{r})$	517.0	**488.3**	495.4	496.7
Test 1	$\overline{\rho}_{0.75}(\mathbf{r})$	301.2	262.8	245.5	**243.9**
Test 1	$\overline{\rho}_{0.25}(\mathbf{r})$	206.3	**204.8**	241.0	236.8
Test 1	$\overline{\rho}_{0.75}(\mathbf{r}) + \overline{\rho}_{0.25}(\mathbf{r})$	507.5	**467.6**	486.5	480.4
Test 2	$\overline{\rho}_{0.75}(\mathbf{r})$	291.1	284.0	269.3	**268.7**
Test 2	$\overline{\rho}_{0.25}(\mathbf{r})$	241.1	**238.9**	254.9	254.7
Test 2	$\overline{\rho}_{0.75}(\mathbf{r}) + \overline{\rho}_{0.25}(\mathbf{r})$	532.2	**522.9**	524.2	523.4
Train	**Crossings**	119	138	**0**	**0**
Test	**Crossings**	76	188	**0**	**0**
Test 1	**Crossings**	13	128	**0**	**0**
Test 2	**Crossings**	63	60	**0**	**0**

Table 4.4: The table contains some key numbers for the models shown in Figure 4.5-4.7. The best number in each row is printed in bold face.

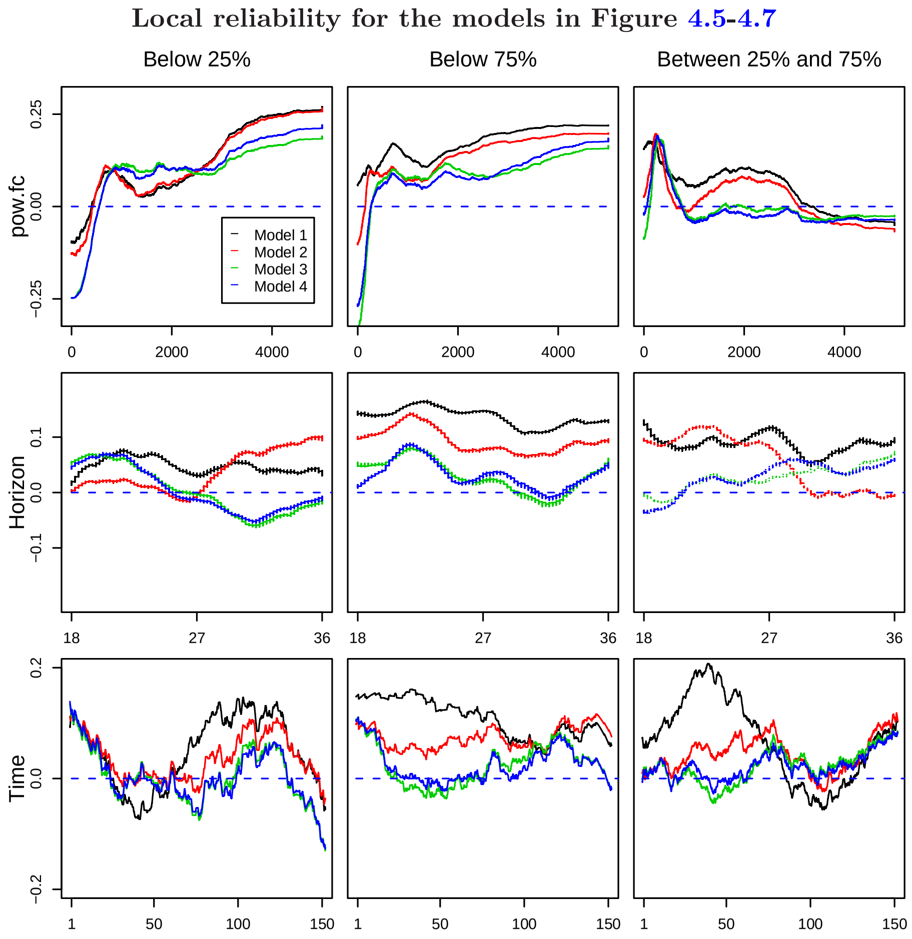

Figure 4.10: Local reliability for the four models in the direction of forecasted power, horizon and time from the beginning of the test period. All the plots are for the test period and give a picture of how well the four models perform in the reliability sense. The dotted line is the perfect line.

Numbers related to the Inter Quartile Range for the quantile models in Figure 4.5-4.7

Model	**1**	**2**	**3**	**4**
$E(\mathbf{IQR})$	1274.7	1086.0	1069.3	**1041.9**
$sd(\mathbf{IQR})$	**668.6**	655.5	592.7	526.2
$Q(\mathbf{IQR}); 0.5)$	1305.9	**1026.7**	1068.4	1051.1
$Q(\mathbf{IQR}); 0.05)$	247.3	143.8	226.7	262.0
$Q(\mathbf{IQR}); 0.95)$	2313.5	2222.4	1961.6	1848.6

Table 4.5: Numbers related to IQR for the models in figure 4.5-4.7, the numbers is for the test period. Best performer in each row is marked with bold face letter.

A very nice property of Model 3 and 4 is that they does not have any crossings, so this again points in the direction of Model 4 rather than Model 2.

4.5.5 Sharpness and Resolution

Table 4.5 and Figure 4.11 deals with sharpness and resolution of the four models described above. The table gives the overall numbers while the figure gives sharpness and resolution as a function of horizon and forecasted power. From the point of the reliability distance and loss function it was not clear if we should choose Model 2 or 4. From the table we see that Model 4 have the best sharpness if the mean of IQR is used as measure and Model 2 has the best sharpness if the median of IQR is used as measure, while Model 1 has the best resolution and Model 2 has a better resolution than Model 4.

We disregard Model 1 since it gives bad results in the analysis made so far. When grouped by horizon the sharpness of the three models remaining perform quite similar, while the resolution is much better for Model 2 than for Model 3 and 4.

When mean and standard deviation of IQR are grouped by forecasted power, we see a quite large difference in sharpness and the resolution has switched place, so Model 3 and 4 perform better than Model 2. This is because Model 2 use the forecasted power as input and the standard deviation therefore becomes small as a function of forecasted power, so the local performance actually gets worse when grouping by the explanatory variable. So we should be careful with these kind of plots, since they can be quite misleading. Actually we could look at the sharpness plot as a form of resolutions, since tells something about the model ability to distinguish between different situations, i.e. different forecasted power in this case.

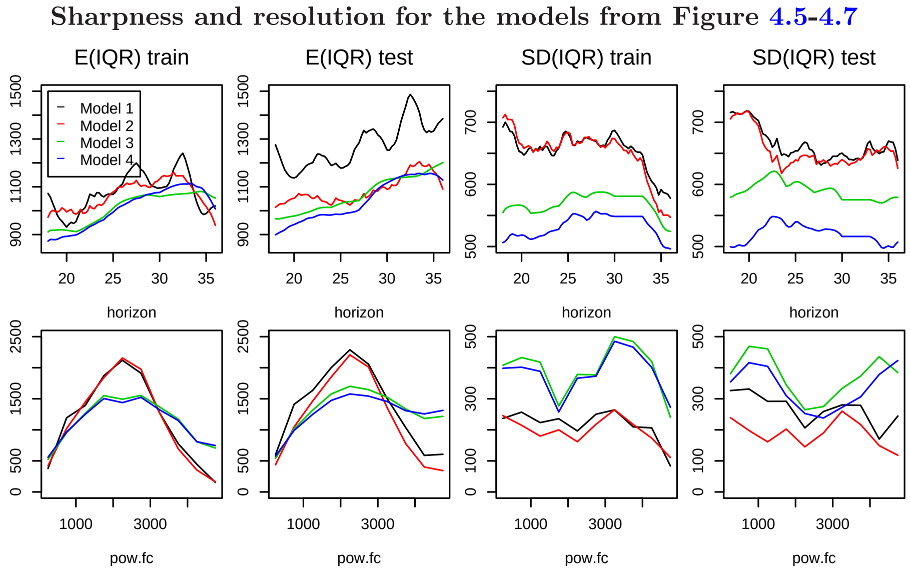

Figure 4.11: The mean and standard deviation of IQR for the test and training periods as functions of horizon and forecasted power. In the horizon case variables is grouped by the same horizon. In the forecasted power direction IQR is grouped by 10 intervals in forecasted power, with each interval having the length of 500kW.

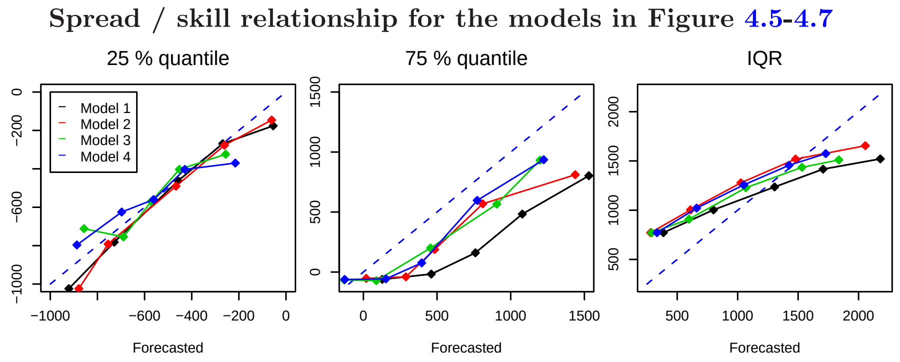

Figure 4.12: Prediction error is grouped by forecasted quantiles and in each group the realized quantile is plotted against the median of the forecasted quantile. All plots are for the test set. The groups are 20% quantile of forecasted quantiles.

4.5.6 Spread / Skill Relationship

Figure 4.12 shows observed quantiles and IQR as a function of predicted quantiles and IQR. The plot shows that all models seem to overestimate the 75% quantile, while the 25% quantile fits quite well. IQR is underestimated for low estimations and overestimated for high estimations of IQR. So this plot suggest that we should concentrate improvement on the 75% quantile.

4.6 Discussion/ Conclusion

This chapter have given an analysis of four different quantile regression models. None of these gave satisfactory results. The analysis does however also give a discussion of how we can use the performance parameter presented in the first part of the chapter. From this discussion it is clear that reliability and possible skill score should be considered the most important performance parameters, which should be good before we go on and consider other performance parameter.

An important conclusion w.r.t reliability is that a good overall reliability does not imply good local reliability. In fact we saw that Model 3, had an overall reliability of 25.3% on the test set, when the 25% quantile was required. This must be considered very good, however the local reliability given a forecasted power of $0kW$ was 0% (see Figure 4.10), which must be considered very bad.

As stated above the models does not really show satisfactory results, this might be due to the fact that the data set is too small. We actually require a model trained in a period of winter and spring to perform well in summer and autumn. It can very well be that the models would improve if the available data set covered e.g. two years.

If we follow the assumption, that there are big annual variation in the data, then the natural conclusion would be to follow an adaptive approach. This is the subject for the next chapter.

Chapter 5

Adaptive Quantile Regression

5.1 Introduction

This chapter will describe an adaptive procedure for quantile regression. This will be used to analyze data from the Tunø Knob wind power plant. Adaptive versions of the models in Chapter 4 will be analyzed as well as simpler models.

In adaptive least square estimation, the estimates at time $t + 1$ can be written as a function of the estimate at time t and the residual at time $t + 1$. In this case the only knowledge of the past we need is what parameter estimates in the model was at time t. This is unfortunately not the case for quantile regression.

We would however expect that the solution to the quantile regression problem at time $t + 1$ is close to the solution to the quantile regression problem at time t. The idea is to use the simplex algorithm described in Section 2.3.2 to get the solution at time $t + 1$, given that we know the solution at time t.

What is meant by close is that the differences of the estimates are small or possibly that the difference over the quantile hyper planes are small in the two solutions. To use the simplex algorithm we need solutions to be close, in the sense that only few simplex steps are needed to get from the solution at time t to the solution at time $t+1$. The hope is now that this first sense of close imply

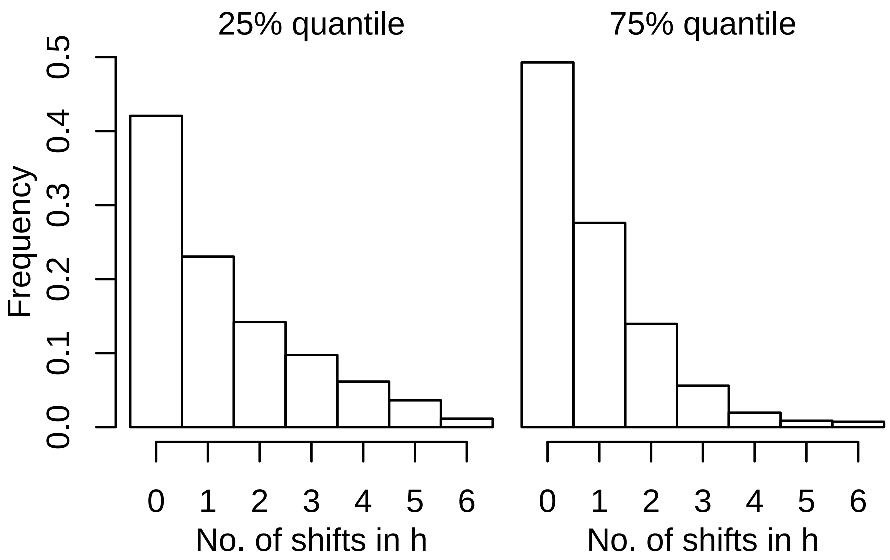

Figure 5.1: The figure show the number of elements shifted in each iteration for a simple model with only 6 parameters and for the 1th and 3th quartile.

the other.

In Chapter 2 we saw that the solution to the quantile regression problem could be written as $\hat{\beta} = \mathbf{X}(h)^{-1}\mathbf{y}(h)$, we now introduce time and write

$$\hat{\beta}_t = \mathbf{X}_t(h_t)^{-1}\mathbf{y}_t(h_t) \tag{5.1}$$

The procedures that will be examined here, is that $[\mathbf{X}_t \ \mathbf{y}_t]$ and $[\mathbf{X}_{t+1} \ \mathbf{y}_{t+1}]$ are the same except for one row. This is a sort of a gliding window where past observations either have the weight one or zero.

To answer the question of how close these solutions are in the simplex sense, a very simple model with only forecasted power as input and 5 knots is constructed. The number of elements in h which was replaced in each iteration is then counted. If the number of elements replaced in h is small it is taken as evidence of the solutions being close in the simplex sense.

This is not done with an adaptive procedure. The model is simply re-estimated at time $t + 1$ without using the knowledge of the solution at time t. In this analysis a simple gliding window was used, i.e. the oldest element in $\mathbf{X}_t$ and $\mathbf{y}_t$ was simply replaced with the observation at time $t + 1$.

This analysis is carried out for the 25% of 75% quantiles. The number of elements in each of the data sets was 10000. The result of this analysis is shown in Figure 5.1. In 40-50% of the steps, there are no replacements in h. In these cases the number of simplex steps will be zero. The mean number of replacements were 1.30 and 0.88 for the two quartiles respectively. So it seems that there should be some possibilities in an adaptive procedure.

Figure 5.1 was constructed by comparing the location of residuals of size 0. This means that the numbers can be a little off, but it is still clear that the solutions are close and that information of the model at time t can be used to get the solution at time $t+1$.

The next section presents the algorithm for the adaptive quantile regression method. The notation is as in Section 2.3.2.

5.2 The Algorithm

As has been seen the number of elements in h to be replaced when we take one step forward is small. It does therefore seem reasonable to use the information about the solution to the quantile regression problem at time t as a guess of the solution of the quantile problem at time $t+1$. The updating procedure in the preliminary analysis in the previous section used the simple procedure to take out the oldest observation each time a new observation become available. The problem of this is the structure of the data set from Tunø, namely that there are very few observations in some areas of the sample space. E.g. there are very few observations of high wind speed or high forecasted power. Therefore as we move our window, the number of observations in these areas of the sample space will at times be very small. This may lead to bad estimations and in extreme cases that the problem become singular i.e. rank$\mathbf{X} < K$.

The algorithm can therefore be divided into two parts, namely the updating procedure and the simplex steps from the updated solution. The second part have already been described in Section 2.3.2, therefore this will not be described here.

As we saw in Chapter 2 the solution to the quantile regression problem is characterized by an index set h and a vector $\mathbf{p}$ of diagonal elements of $\mathbf{P}$. Or rather these are what we need to perform the next simplex steps. In the following we write the design matrix $\mathbf{X}$ as

$$\mathbf{X} = \begin{bmatrix} \mathbf{b}(\mathbf{x_1})^T \\ \vdots \\ \mathbf{b}(\mathbf{x_N})^T \end{bmatrix} \tag{5.2}$$

where $\mathbf{b}(\mathbf{x})$ is the spline basis functions of $\mathbf{x}$, so the set up is like in Section 4.2. Note that $\mathbf{b}(\mathbf{x}) \neq \mathbf{b}$. $\mathbf{b}$ is the restrictions in the LO problem presented in Section 2.3.2.

Practical summary 5.1 gives an overview of the adaptive procedure. The individual steps is then described below

Practical Summary 5.1 *The adaptive procedure for quantile regression consist of the following steps:*

1. *Decide which $\mathbf{b}(\mathbf{x}_l)$ that have to leave the design matrix.*
2. *If $l \in h_t$ take one simplex step s.t. $l \notin h_t$.*
3. *Update $\mathbf{X}$, $\mathbf{P}$, h and $\mathbf{c}_{\mathcal{B}}$*
4. *Perform the simplex steps needed to get to the optimal solution characterized by h_{t+1}.*

The steps of the algorithm is described in details below.

Step 1

In this presentation two different updating procedures are considered, the gliding window and an approach where the data set is divided into a number of bins in the direction of one of the explanatory variables. The first approach correspond to one bin.

Check what bin the new observation belongs to, and then let the oldest observation in that bin leave the design matrix. In this presentation these bins will be in the direction of wind speed or forecasted power.

The bins will be denoted I_j, and should be a partition of the sample space of this direction. Assume that the vector $\mathbf{x}$ is ordered s.t. x_1 is the direction of this partition and $\mathbf{x} \in \mathbb{R}^p$ then $\mathbf{x}_{t+1} \in I_k \times \mathbb{R}^{p-1} = \mathcal{I}_k$. For each $\mathcal{I}_j$ a maximal number of elements n_j is decided before hand. If the number of elements in $\mathcal{I}_k$ is less than n_k, then set $l = \emptyset$ and go to step three, otherwise choose $l = \min(i|\mathbf{x}_i \in \mathcal{I}_k)$ and proceed to step two.

Step 2

If the leaving variable is in h, then we can not remove it, since the solution depend on the inverse of $\mathbf{X}(h)$. Therefore if $l \in h$ we perform one simplex step with a new objective function, where the loss on r_l is set equal to zero. In the terminology of Chapter 2 this corresponds to setting $\mathbf{c}_{\{l,l+N\}} = \mathbf{0}$. This result is a change of two elements in the simplex vector $\mathbf{d}$ (see section 2.3.2). The elements we have to change is the elements corresponding to l. Denote these $\mathbf{d}^l$, we then have

$$\mathbf{d}^l = \begin{bmatrix} -\mathbf{g}(l) \\ \mathbf{g}(l) \end{bmatrix} \tag{5.3}$$

Therefore we only have to calculate the two elements of $\mathbf{d}^l$ corresponding to l. This determines the decent direction and gives us s (see section 2.3.2). If $\mathbf{g}(l) = 0$ then this will not be a decent direction. We can however take one simplex step anyway, the direction is then not important. With this we can find $\mathbf{h}$ and thereby σ and q, which was the variable that had to enter h. Having swapped l and q we are ready to update the design matrix and the other matrices needed for the simplex algorithm.

Step 3

If we let $\Omega = \{1, 2, ..., N\}$ and $r_{t+1} = y_{t+1} - \mathbf{b}(\mathbf{x}_{t+1})^T \hat{\beta}_t$ then the updated versions of the simplex variables will be

$$\begin{aligned} \mathbf{X}^{(t+1)} &= \begin{bmatrix} X^{(t)}_{\Omega\backslash l} \\ \mathbf{b}(\mathbf{x})_{t+1} \end{bmatrix} && (5.4) \\ \mathbf{P}^{(t+1)} &= \begin{bmatrix} \mathbf{P}^{(t)}_{\Omega\backslash l} \\ \mathrm{sign}(r_{t+1}) \end{bmatrix} && (5.5) \\ \mathbf{c}^{(t+1)}_{\mathcal{B}} &= \begin{bmatrix} \mathbf{c}^{(t)}_{\mathcal{B};\Omega\backslash l} \\ \rho_\tau(\mathrm{sign}(r_{t+1})) \end{bmatrix} && (5.6) \end{aligned}$$

For $j = 1...K$, we update the index set h as

$$h_j^{t+1} = \begin{cases} h_j^t & \text{if} \quad h_j < l \\ h_j^t - 1 & \text{if} \quad h_j > l \end{cases} \tag{5.7}$$

With the elements needed for the simplex algorithm in place, we go on to Step 4.

Step 4

Two changes or modifications of the simplex algorithm described in Section 2.3.2 are implemented. The first change is that a maximum of simplex steps is set. This is set to 24. This number is quite arbitrary, but for the models that are stable there are very few iteration where the number of simplex steps is equal to 24.

The second change is that in each step the condition number of the matrix $\mathbf{X}(h^{(t+1)})$ is calculated and the simplex step will only be taken if this number is less than a specified value. It is not obvious what such a value should be, but as long as we just have some value, the algorithm will not terminate. It will however not update the solution in this step either. Here the maximum condition number is set to 10^6.

The simplex algorithm will sum up small numerical errors in each step. It is therefore necessary to fix this once in a while. [14] give an algorithm for doing this. In the quantile regression setting we just recalculate $\hat{\beta}$ from $\mathbf{X}(h)$ after each time step and find $\mathbf{r}$, $\mathbf{P}$ etc. from this solution.

5.2.1 Remarks on the Algorithm

For a real on-line implementation of this algorithm, the point of why $\mathbf{X}(h)$ becomes singular should probably be investigated further. The point is here that based on the theory on the simplex method it should not be possible to go to such a solution. One practical problem is to determine when a number is zero and when it is a small number different from zero. This problem is clear in Step 3 of the simplex algorithm, where we divide with elements of $\mathbf{h}$, and some of the elements in $\mathbf{h}$ will be zero. In this step a decision of when h_i is zero should be made, i.e. a tolerance have to be set. What the size of this tolerance should be is not clear. If it is set too large then the algorithm can terminate early (because we know that the problem is bounded the algorithm terminates if $\alpha = \infty$), and if it is too small we can get very large gain in the objective function in directions where it should have been zero.

5.3 The Performance of Adaptive Models

The performance parameters for the adaptive models will be the performance parameters described and discussed in Chapter 4. But in addition the maximum and mean of crossings will also be considered. For an adaptive procedure, the timing is also something that should be considered and this will also be discussed here. The performance parameter are considered on test sets as defined below.

5.3.1 The Test Set for Adaptive Models

The adaptive models are updated each time a new observation becomes available. This means that we would have a new model every quarter of an hour if the model was running online. The data set available only has 18-36 hour forecast based on the meteorological data from time 06 and the corresponding observations.

The performance data is based on the adaptive model at a forecast horizon of 24 hours. I.e. when the observations of the 24 hour forecast becomes available, the model is used to forecast the next 18-36 hour ahead and the residuals from this is used to calculate the performance parameters.

The test set is still the last part of the data set. The first 10000 points are the "training" set and the rest of the data set is the test set. So performance parameters can be compared with numbers from Chapter 4 even though the test sets are not identical.

5.4 Four Simple Models

To study the effect of the adaptive procedure and the effect of different updating strategies a simple model, which only uses forecasted power as input and with knots placed at 20% quantiles of forecasted power, is constructed. This model is studied with different updating strategies.

The models in this basic analysis will be referred to as Basic 1-4. The updating procedure in the models are described below

Reference: A static model based on the first 10000 data points, and used to forecast on the rest of the data points.

Basic 1: An adaptive model with a gliding window with 10000 data points.

Basic 2: An adaptive model with a gliding window with 5000 data points.

Basic 3: An adaptive model with bins placed at the knots, i.e. the sequence of borders is $\{-\infty, 308, 789, 1465.5, 2701, \infty\}$. The number of elements allowed in each bin is 1200.

Basic 4: An adaptive model with the borders of the bins placed at $\{-\infty, 800, 1600, 2400, 3600, \infty\}$. The number of elements allowed in each bin is 1000.

These models are now examined for the 25% and 75% quantile.

Figure 5.2 shows time plots of the two quartiles of Basic 4. The plot shows the quartile curves at each time step. It illustrate how the quartile curves varies with time. The top row of the figure shows the 75% quantile. In the left end of the plot we see that the quartile curve have a shoulder like the one we saw in Model 1 of Chapter 4. We see that this slowly disappears and after about 40 days it is gone. In the bottom row we see the 25% quantile. At around 100 days the forecasted quartile at high values of **pow.fc** drops very rapidly down to about $-1500kW$. This behavior can probably be explained by cut off effects, i.e. the wind power plant shots down to avoid damage on the plant due to very high wind speeds. If this happens there is a great possibility of forecasting $5000kW$, when actual production becomes $0kW$. It could be argued that, since we use an updating strategy with several bins a few cut offs should not affect the quantile curves so dramatically. The problem is however that the forecasted power within the bins is not equally distributed. So it can very well be that at this point there were no forecasted power close to $5000kW$.

5.4.1 Reliability

Figure 5.3 shows local reliability as a function of forecasted power, horizon, and time. The plots clearly shows that we get a very large improvement for this simple model when we make it adaptive. The adaptive models are clearly better in all plots, except for the horizon for the 25% quantile, where all models seems to perform equal. Table 5.1 gives the overall and local reliability for the same variables as used in Figure 5.3.

In the reliability sense these simple adaptive models perform better than the more advanced, but static, model analyzed in Chapter 4. The reliability distance of the model Basic 4 in the direction of forecasted power is less than 1/3 of the best reliability distances in the static models.

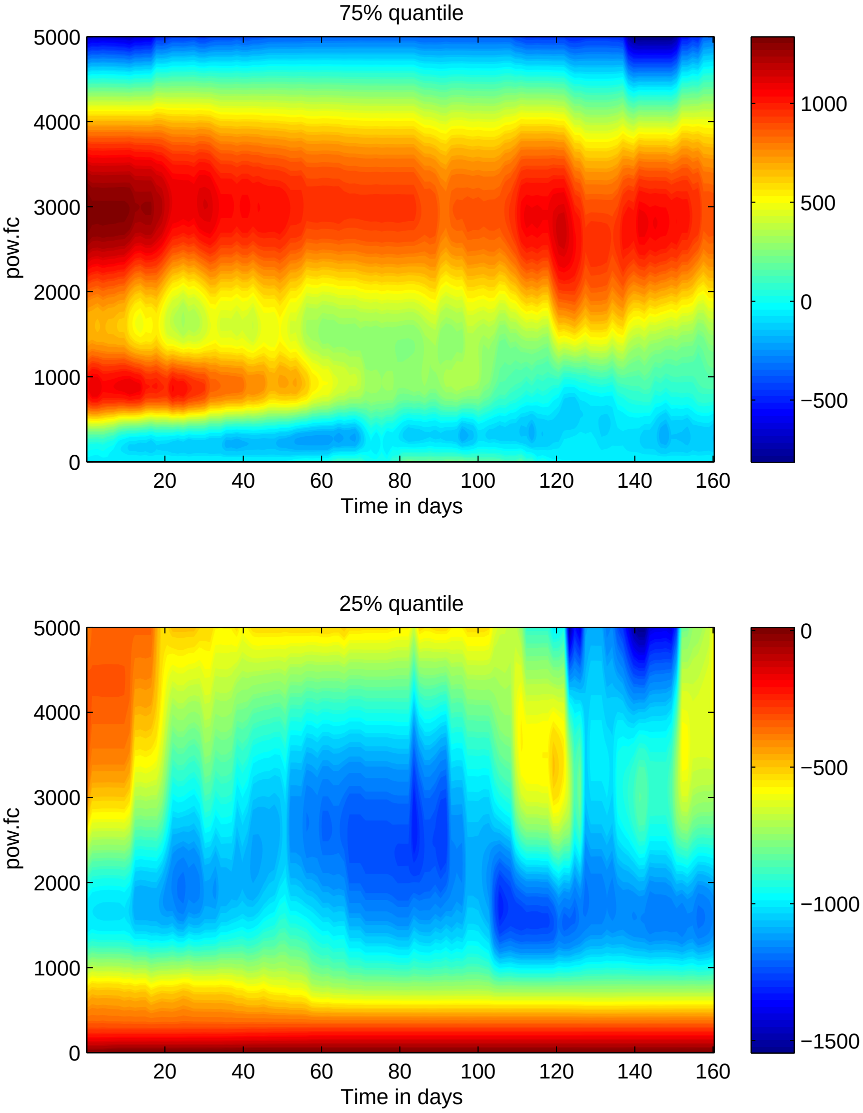

Figure 5.2: Quartile curves for model Basic 4, as a function of time, for the test set. The figure illustrate how the quantile curve changes at each time step. The top panel is the 75% quantile and bottom panel is the 25% quantile.

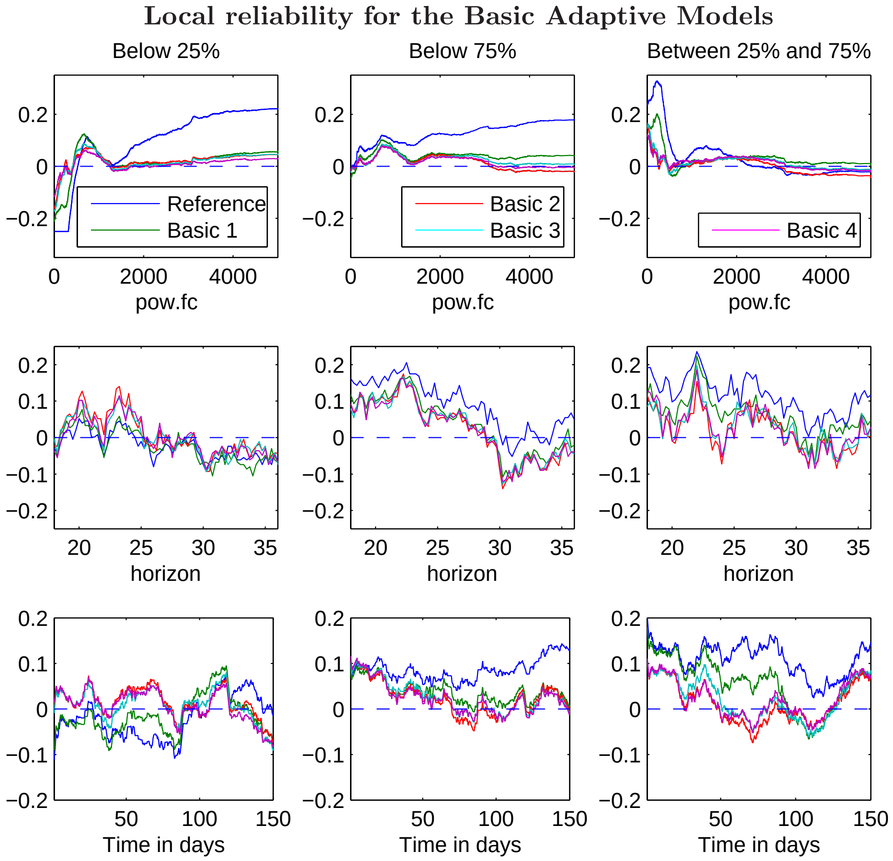

Figure 5.3: Local reliability for the four Basic adaptive models in the direction of pow.fc, horizon and time.

Local reliability measure

Model	**Reference**	**Basic 1**	**Basic 2**	**Basic 3**	**Basic4**
Below **75% (test)**	83.9%	78.6%	**77.2%**	77.6%	**77.2%**
Below **25% (test)**	22.6%	22.9%	25.7%	25.1%	**25.0%**
$d(q(\mathbf{pow.fc}, 0.25))$	0.165	0.095	0.046	0.053	**0.033**
$d(q(\mathbf{pow.fc}, 0.5))$	0.172	0.094	0.040	0.047	**0.037**
$d(q(\mathbf{pow.fc}, 0.75))$	0.100	0.048	0.034	0.036	**0.033**
$dq_{tatal}(\mathbf{pow.fc})$	0.149	0.082	0.040	0.046	**0.034**
$d(q(\mathbf{hor}, 0.25))$	**0.043**	0.049	0.056	0.050	0.050
$d(q(\mathbf{hor}, 0.5))$	0.125	0.080	**0.056**	0.061	0.062
$d(q(\mathbf{hor}, 0.75))$	0.112	0.086	0.083	**0.081**	**0.081**
$dq_{total}(\mathbf{hor})$	0.100	0.073	0.066	**0.065**	0.066
$d(q(\mathbf{time}, 0.25))$	0.048	0.045	**0.036**	0.039	0.037
$d(q(\mathbf{time}, 0.5))$	0.118	0.079	**0.048**	0.054	0.049
$d(q(\mathbf{time}, 0.75))$	0.093	0.044	0.041	0.042	**0.040**
$dq_{total}(\mathbf{time})$	0.091	0.058	0.042	0.045	**0.042**

Table 5.1: Reliability distance in the direction of pow.fc, horizon and time for the four basic adaptive models.

The static models perform better than the adaptive model in direction of horizon, and for the adaptive model we also see a systematic deviation from the required reliability in the direction of horizon. Especially for the 75% quantile. This could indicate that we need to have horizon in the models. In the direction of horizon we see that the reference model actually performs better than the adaptive models.

In the direction of time the adaptive models perform better than the static model, but there is actually not much difference between them.

Looking at the reliability performance of the adaptive model the oblivious choice would be the model Basic 4. This is based on 5000 points while Basic 3 is based on 6000 points. The question of how many points we should base our model on is addressed in Section 5.7.

5.4.2 Skill Score and Crossings

Table 5.2 gives the loss functions for the Basic models and the Reference model, from the skill score point of view we should choose Basic 1 for the 75% quantile and Basic 4 for the 25% quantile. We see a large improvement in the skill score if we compare with the reference model, and all adaptive models perform

Skill Score and Crossings for Basic 1-4

Model	Reference	Basic 1	Basic 2	Basic 3	Basic 4
$\overline{\rho}_{0.75}(\mathbf{r})$	260.3	**251.3**	251.8	251.8	251.4
$\overline{\rho}_{0.25}(\mathbf{r})$	209.6	201.3	199.5	199.6	**198.0**
$\overline{\rho}_{0.75}(\mathbf{r}) + \overline{\rho}_{0.25}(\mathbf{r})$	469.9	452.6	451.3	451.4	**449.4**
Crossings	113	**84**	147	123	180
min(**IQR**)	-346.8	-415.4	-454.9	-454.0	**-253.2**
$E(\mathbf{IQR} < 0)$	-176.6	-254.5	-160.1	-191.0	**-73.0**

Table 5.2: Numbers related to skill score and crossing for the Basic models and for the test period

better than the static model in this sense and are quite close. Actually even the Reference model perform better than the static models from Chapter 4 in the sense of Skill score for the 25% quantile.

With respect to crossings we see that Basic 1 have the fewest number of crossings, but on the other hand the maximum size of the crossings is much larger than from Basic 4. The mean size of crossings are also smaller for Basic 4 than for the other model.

The top row of Figure 5.5 shows realized IQR as a function of forecasted power. From this plot we see that all large crossings are realized at forecasted power close to $5000kW$. The bottom row shows a picture of possible IQR. These plots are constructed by calculating the possible outcomes at each time point, and then taking quantiles of these values, something like a projection of Figure 5.2. From this plot it is seen that Basic 4 actually at some points have been able to produce very large crossings. These are just not realized as is seen from table 5.2 and the top row in the same figure.

5.4.3 Sharpness and Resolution

Figure 5.5 shows realized and possible IQR for the four basic models. We see both realized and possible IQR is quite different from the reference model.

Figure 5.5 and Table 5.3 deals with sharpness and resolution of the Basic adaptive models. The table indicate that we should choose Basic 2, and that all models have improved in this aspect when we go to an adaptive approach. It also shows large difference between the 50% quantile and the mean, and that using the two different measures would lead to different conclusions.

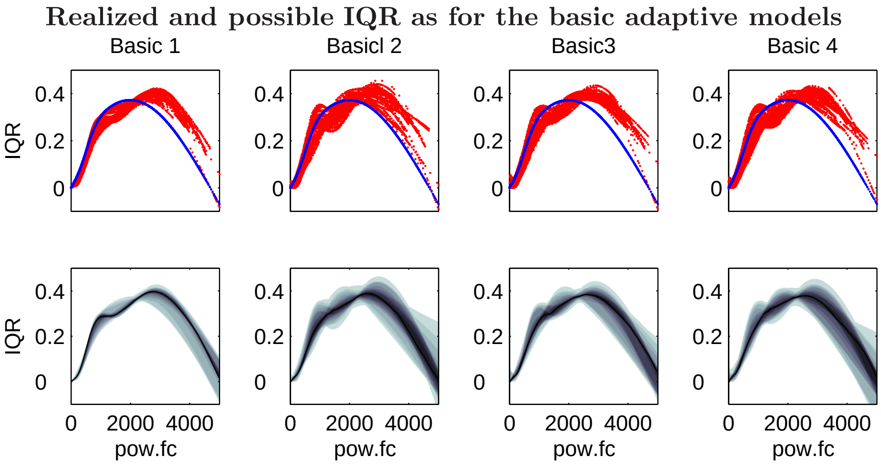

Figure 5.4: IQR plots. The first row shows realized IQR for the four models. The blue line is the reference model. Second row shows quantile curves, i.e. for each time step. The quantile curve is calculated and then 0 to 100% (in steps of 5%) quantiles is calculated for each value of forecasted power. This gives an idea of possible outcomes

The local sharpness and resolution is plotted in figure 5.5. The main conclusion is that there is a large difference between the static and adaptive models, but that the adaptive models behave alike.

5.4.4 Spread / Skill Relationship

Figure 5.6 shows observed quantiles as a function of predicted quantiles. The plot is constructed in the same way as Figure 4.12. For the 75% quantile we see that the adaptive models perform better than the Reference model, but for the 25% quantile and IQR we see that the curves are very close. In the IQR case the Reference model perform better than the adaptive models. This probably have to do with the way the data is grouped.

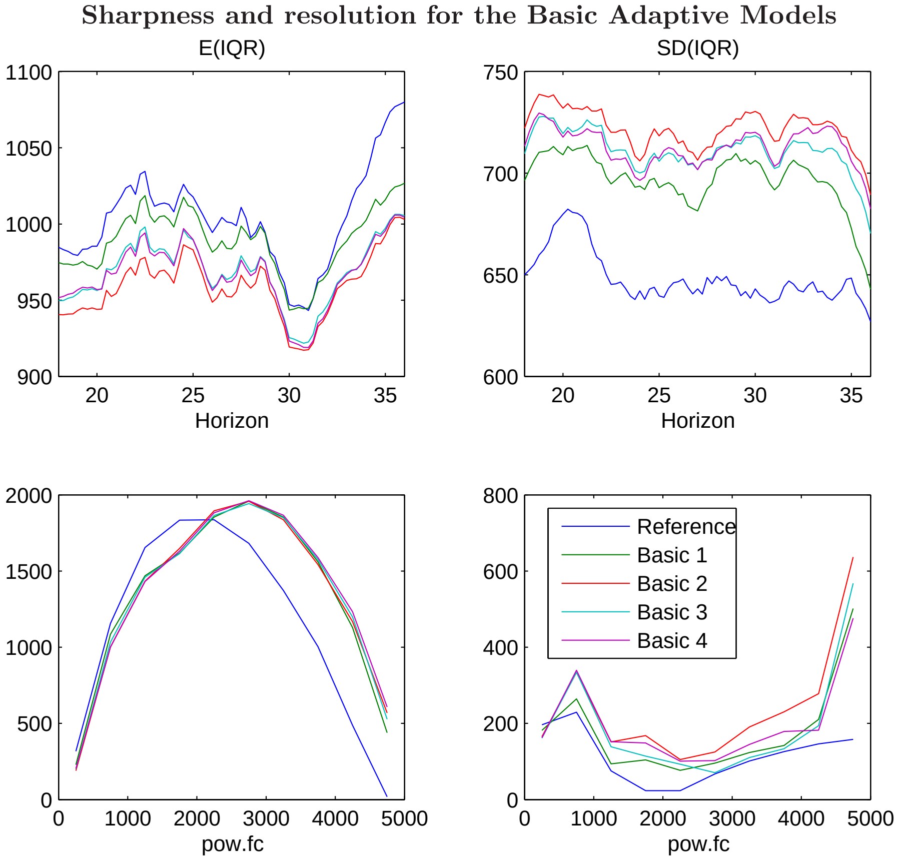

Figure 5.5: The mean value and standard deviation of IQR for the four basic adaptive models and the reference model as a function of horizon and forecasted power.

Sharpness and Resolution for the four Basic models

Model	**Reference**	**Basic 1**	**Basic 2**	**Basic 3**	**Basic 4**
$E(\mathbf{IQR})$	1015.4	998.7	**968.6**	977.9	975.7
$sd(\mathbf{IQR})$	648.6	693.2	**716.2**	706.3	706.9
$Q(\mathbf{IQR}; 0.5)$	**1090.4**	1177.0	1099.9	1120.9	1100.7
$Q(\mathbf{IQR}; 0.05)$	85.1	39.7	19.1	25.9	25.9
$Q(\mathbf{IQR}; 0.95)$	1850.1	1951.7	1981.0	1939.4	1979.8

Table 5.3: Numbers related to IQR and the over all performance for the four basic model used to illustrate the adaptive approach

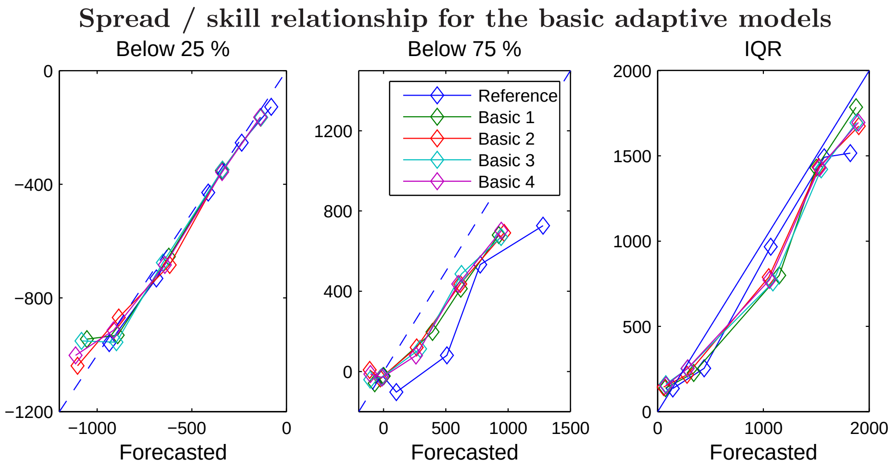

Figure 5.6: Observed quantiles and IQR as a function of predicted quantiles and IQR.

5.5 Performance of the Adaptive Versions of Model 1-4

Adaptive versions of the models in Chapter 4 are examined in this section. These will be referred to as Model A1-A4. The updating procedure is in all cases that the knots in the direction of wind speed or **pow.fc** defines the bins. For Model A1 1100 points in each bin are allowed, so the estimates will be based on 11000 points for this model. It is necessary to have this number of points in each bin for the adaptive procedure to be stable. This is probably because of the large annual variation in the air density as discussed in Chapter 4 and we will see that the model performance is still very poor. The other models allow between 800 and 1300 points in each bin. This means that the estimates are based in approximately 5000 point for all other models.

5.5.1 Reliability

Table 5.4 and Figure 5.7 deal with reliability Model A1-A4. Both the plot and the table clearly show very large improvements in reliability. The strength of the adaptive procedure is again stressed, but the reliability of these models is actually not better than for the Basic models analyzed in the previous section. So the reliability alone can not lead to choosing one of the more advanced

Local reliability measure

Model	A1	A2	A3	4
Below **75% (test)**	77.8%	77.7%	76.5%	**75.7%**
Below **25% (test)**	27.7%	27.5	**25.5%**	26.2%
$d(q(\mathbf{pow.fc}, 0.25))$	0.045	**0.032**	0.092	0.101
$d(q(\mathbf{pow.fc}, 0.5))$	**0.026**	0.033	0.064	0.081
$d(q(\mathbf{pow.fc}, 0.75))$	0.045	0.042	0.050	**0.040**
$dq_{total}(\mathbf{pow.fc})$	0.040	**0.036**	0.071	0.078
$d(q(\mathbf{hor}, 0.25))$	0.036	0.072	**0.027**	0.036
$d(q(\mathbf{hor}, 0.5))$	0.033	0.068	**0.032**	0.036
$d(q(\mathbf{hor}, 0.75))$	0.039	0.042	0.035	**0.032**
$dq_{total}(\mathbf{hor})$	0.036	0.062	**0.032**	0.035
$d(q(\mathbf{time}, 0.25))$	0.057	0.053	**0.032**	0.039
$d(q(\mathbf{time}, 0.5))$ 50%	**0.027**	0.028	0.042	0.039
$d(q(\mathbf{time}, 0.75))$ 75%	0.063	0.056	0.049	**0.042**
$d_{total}q(\mathbf{time})$ total	0.051	0.045	0.042	**0.040**

Table 5.4: Overall reliability and reliability distance in the direction of pow.fc, horizon and time for the adaptive models.

models. Table 5.4 suggest that we should choose Model A3 or A4 if we should choose one of these adaptive models.

5.5.2 Skill Score and Crossings

Table 5.5 shows the skill score and the number of crossings. From a skill score perspective we should choose Model A3 for the 75% quantile and Model A2 for the 25% quantile. If the interval score is considered we should choose Model A2. Model A3 and A4 have a better skill score for the 75% quantile than the Basic models from the previous section. This supports the conclusion that the Basic models seemed too simple to model the 75% quantile.

Table 5.5 shows that Model A1 have 378 crossings. That is far more than the other adaptive models. A more serious problem is that it produces crossings of the magnitude of 21000kW. This should be compared with the fact that absolute value of the maximum error from WPPT is 5000kW. The problem is probably, as was also discussed in Section 4.5, the very large annual variation in air density. Which results in few observations to support estimates in some areas of the data.

A point in connections with this is that this was not really punished in the

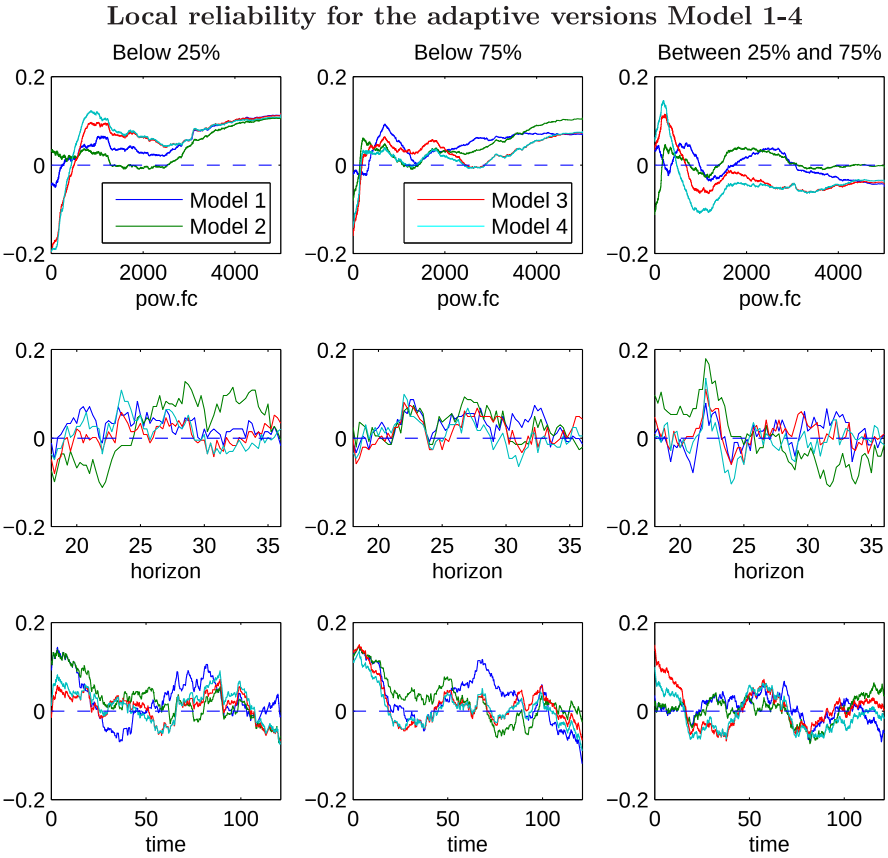

Figure 5.7: Reliability as a function of pow.fc, horizon and time for adaptive versions of the models presented in Section 4.5.

Skill Score and Crossing of Model A1-4

Model	A1	A2	A3	A4
$\overline{\rho}_{0.75}(\mathbf{r})$	295.9	254.2	**243.1**	244.8
$\overline{\rho}_{0.25}(\mathbf{r})$	226.7	**216.9**	231.8	234.1
$\overline{\rho}_{0.75}(\mathbf{r}) + \overline{\rho}_{0.25}(\mathbf{r})$	522.3	**471.1**	474.9	478.9
Crossings (test)	378	114	**7**	39
$min(\mathbf{IQR})$	-21252.0	-307.5	**-24.0**	-112.6
$E(\mathbf{IQR} < 0)$	-1139.2	-67.3	**-15.1**	-53.1

Table 5.5: Numbers related to IQR for Model A1-A4

Sharpness and resolution for Model A1-A4

Model	A1	A2	A3	A4
$E(\mathbf{IQR})$	1013.3	966.0	967.9	**926.5**
$sd(\mathbf{IQR})$	**1100.2**	**609.4**	548.8	552.1
$Q(\mathbf{IQR}, 0.5)$	1044.8	896.9	947.4	**851.9**
$Q(\mathbf{IQR}, 0.05)$	63.7	147.2	206.8	188.0
$Q(\mathbf{IQR}, 0.95)$	2091.4	1993.3	1849.3	1840.0

Table 5.6: Numbers related to IQR for Model A1-A4

reliability measures. This mean that we can not let reliability stand alone as a measure. This behavior is however punished in the loss function.

From a crossing perspective we should prefer Model A3.

5.5.3 Sharpness and Resolution

Table 5.3 gives numbers related to resolution and sharpness. Note that the resolution Model A1 is very good compared to the rest of the model. This is however due to the extreme crossings so in this case the measure award a very undesirable behavior of the model. We also see that the extreme crossings of Model A1 is not punished much by sharpness.

Figure 5.8 shows sharpness and resolution for Model A2-A4. the behavior of sharpness as a function of horizon is surprising since we would expect IQR to be an increasing function of prediction horizon. We see that it drops down at the largest prediction horizon. A possible explanation for this is that what we see is actually daily variation. We have however no way of checking this with the available data.

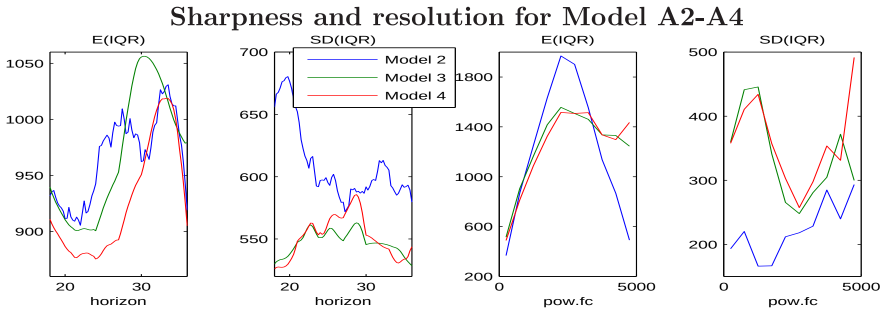

Figure 5.8: Observed quantiles and IQR as functions of predicted quantiles and IQR, for Model A2-A4. Model A1 is left out since it prevents us from seing the variation of the other models.

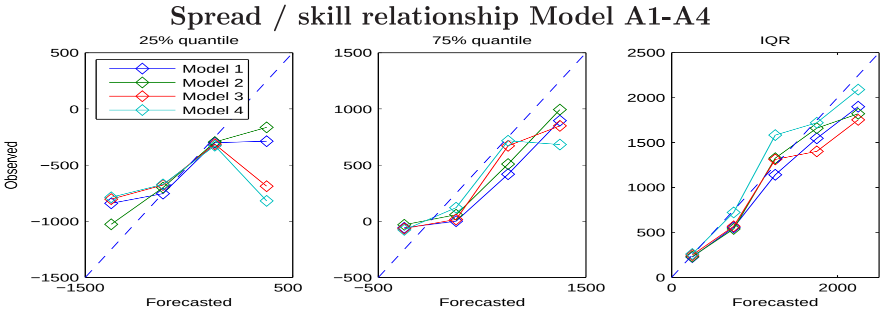

Figure 5.9: Observed quantiles and IQR as functions of predicted quantiles and IQR for Model A1-A4.

5.5.4 Spread / Skill Relationship

Figure 5.9 shows observed IQR as a function of forecasted IQR. This plot is constructed in a different way than the plot for the Basic models. In these plots we group in bins of forecasted IQR with a constant length. It is seen that all the models follow the perfect line quite well, except for extreme values. This can be explained with the fact that there are few observations here and that the medians therefore not really are well defined.

The plots are also difficult to interpret because there will not be equally many observations in each bin for the different models.

Mean time used per iteration and mean number of simplex steps for the adaptive model

Model	B1	B2	B3	B4	A1	A2	A3	A4
E(**Time**) 25%	0.15	0.09	0.06	0.07	3.06	0.23	0.16	0.48
E(**Time**) 75%	0.08	0.06	0.06	0.05	2.79	0.24	0.17	0.46
E(**n**) 25%	3.15	2.93	1.44	1.78	11.77	4.51	3.87	6.45
E(**n**) 75%	1.36	1.75	1.24	1.17	10.41	4.63	3.95	6.86

Table 5.7: Mean time used per iteration and mean number of simplex steps (n) for the adaptive model, B1-B4 refer to the Basic adaptive models, while A1-A4 refer to the adaptive versions of Model 1-4

5.6 Time Consumption the Adaptive Models

If an adaptive procedure should produce online estimation, then the time it uses would also be a performance parameter. The algorithms presented have been implemented in Matlab and the focus in the implementation has been stability rather than optimizing with respect to time. Therefore it will probably be possible to reduce the timings presented here.

The timing only measures the simplex steps, so there will be some additional effort to calculate the splines basis functions from the forecasted meteorological data as they become available.

As a reference the model Basic 1 was also calculated with the "rq" method in the statistical software"R", i.e. this was not really an adaptive model. The whole model was just re-estimated on new training set. The adaptive approach should be faster than this, at least when they the adaptive procedure is optimized. The run i "R" used in average 0.83 and 0.41 seconds per iteration for the 25% and 75% quantiles respectively.

Table 5.7 gives the average time consumption per iteration and the mean number of simplex steps used in each time step for the models considered so far. It is seen that all models except for model A1 uses less than the timing in "R". As was stated above A1 is also badly conditioned, and inverting $\mathbf{X}(h)$ is a time consuming task.

5.7 How Adaptive?

The previous sections treated adaptive models. This section will address the question of how adaptive a models should be. To study this three different model are examined. These are characterized with the same knot placement as the Basic models in Section 5.4, but with three different updating procedures. These models are now examined with different sizes of the bins. The models are referred to as

Bin1: One gliding window

Bin2: Bins defined by the knots.

Bin3: Bins defined by the knots and in the midpoint between each knot.

These models are now studied with respect to the performance parameter that we can visualize. These are reliability distance, skill score, crossings, sharpness, and resolution. The performance parameters can however only by visualized as the overall measure. Such measures are visualized as a function of the number of elements in the design matrix at the end of the test period. The number of elements in the design matrix will not be constant over the test period if many observations are allowed in the bins. This is due to the size of the available data set.

Figure 5.10 shows local reliability distance for the three updating procedures. There is not a very big difference between these. This figure indicate that we should have quite few element in the design matrix, about 2000. This correspond to about one month in the gliding window case. In the other cases it the will be different from each of the bins.

The reason why the updating procedures looks so similar in the reliability distance, is that the different updating strategies take care of rare events and rare events will not affect reliability distance. Model A1 was an extreme example of this.

The effect of choosing different updating strategies is illustrated in Figure 5.11, where the number of crossings, the size of extreme crossings and the mean size of the crossings are plotted. This figure clearly shows unacceptable behavior when the design matrix is small. With few elements in the design matrix the absolute size the crossings are of the same size as the possible interval of forecast. It is seen that the mean size of crossings and the size of extreme crossings stabilizes earlier for Bin 2 and 3, so if we want few elements in the design matrix then an

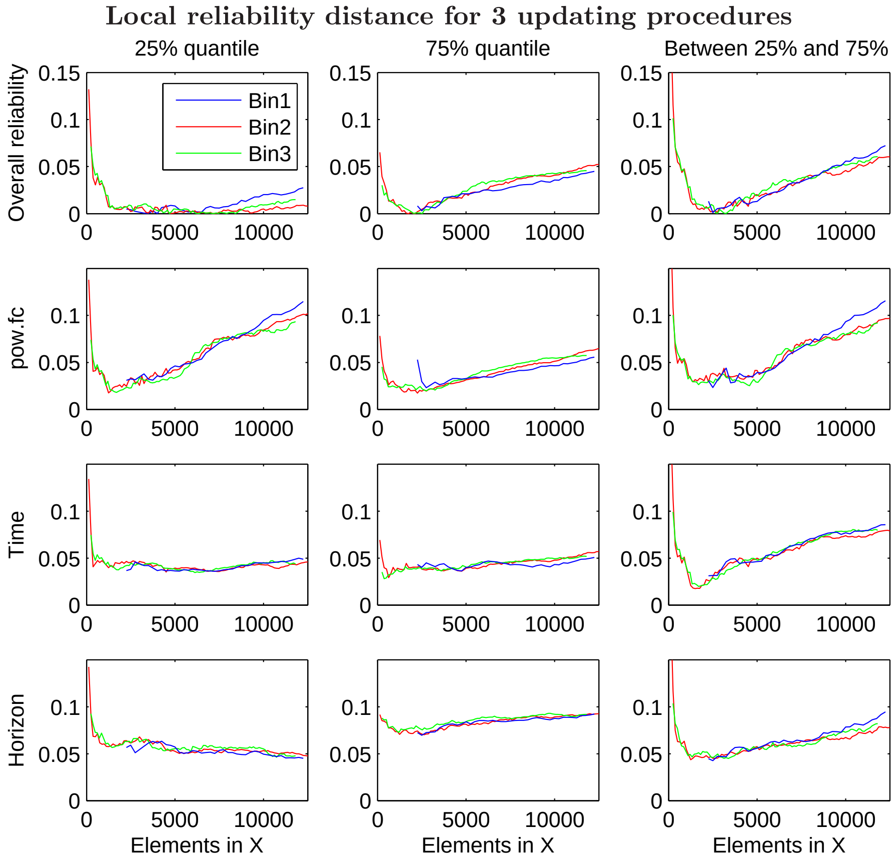

Figure 5.10: Local reliability distance as a function of the number of elements in the design matrix $\mathbf{X}$, for the 3 different updating procedures

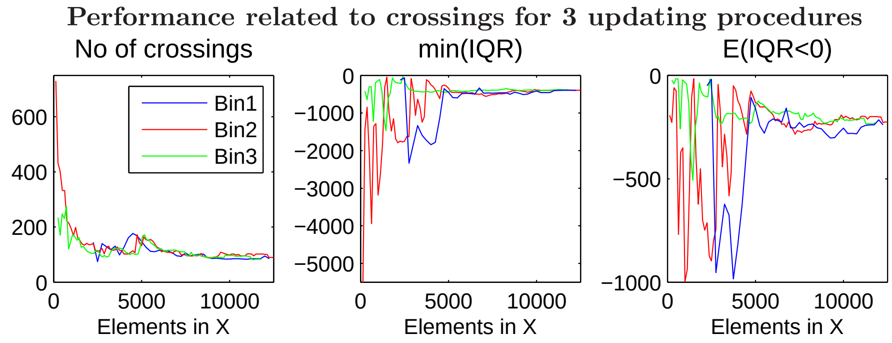

Figure 5.11: Number of crossings, the most extreme crossing in the test set and the average size of the observed crossing as a function of the number of rows in the design matrix.

updating procedure with bins should be chosen. It is also noted that both the mean size of crossings and the size of extreme crossing display random behavior for small design matrices. This means that we could be mislead by a good performance in this sense simply by chance.

Figure 5.11 suggests that we should choose the number of elements in the bins such that the number of elements in the design matrix is about 5000. For Bin 3 this could be chosen a somewhat smaller maybe around 2500-3000. So comparing with the reliability plot we should choose Bin 3 and with the number of elements in each bin such that the size of the design matrix become about 3000.

Figure 5.12 shows the skill score for each of the quantiles. The interval score is not shown here, but this is just the sum of the quantile scores. The figure suggest to choose Bin 2 with about 4000 elements in the design matrix. This is also where the crossings begin to stabilize for this model, but the conclusion is quite different from the reliability plot in Figure 5.10. The skill score is higher for Bin 3 which has better performance with respect to crossings and reliability combined. I.e. we can choose a smaller design matrix and still avoid extreme crossings.

Figure 5.13 shows sharpness and resolution for the three updating procedures. For small design matrices we can not relay on the sharpness measure since the large number, and extreme size, of the crossings give an unrealistic picture of the size of sharpness. Therefore we should disregard sharpness for design matrices with few elements. The only information we really get from sharpness is that we should choose Bin 2 or 3 and with as few elements as the crossing analysis

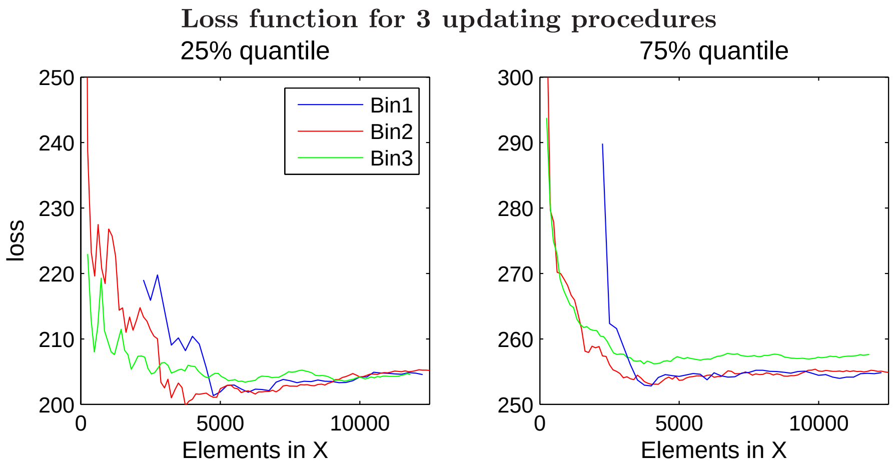

Figure 5.12: Average loss on the test set for the 3 different updating procedures as a function of the number of rows in the design matrix.

allow. Resolution also reward extreme behavior of the crossings, but we see an extrema of this at around 4000 elements for Bin 2 and 3. So this conclusion is similar to the conclusion from the skill score, and the behavior of the crossing is not extreme any more.

Figure 5.14 shows the average cpu time per time step and the mean number of simplex steps per time step average is also taken over the two quartiles. The timing is quite close to a linear function and it is fast enough for a real time implementation, for all sizes of the design matrix. For both timing and number of simplex steps we see that Bin 2 and 3 perform better than Bin 1. Figure 5.15 shows standard deviation for the number of simplex steps and the cpu timing. The standard deviation of time displays the same behavior as the mean time, while the standard deviations of the number of simplex steps becomes smaller as we get more elements in the design matrix **X**.

5.8 The Prediction Interval for Tunø

In the analysis of the models we have presented so far, the number of quantile crossings has been used as a performance parameter. We want to avoid these because we can not give any physical interpretation of phenomena like this.

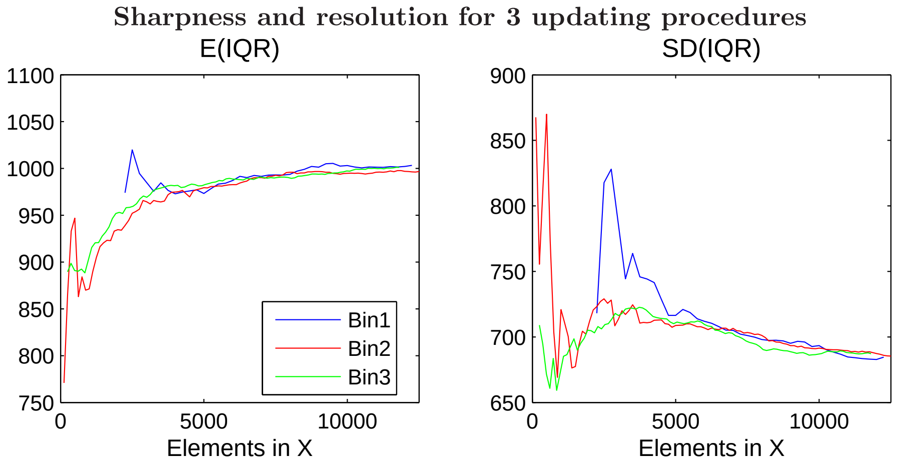

Figure 5.13: The mean and standard deviation of IQR as a function of the number of row in the design matrix **X**.

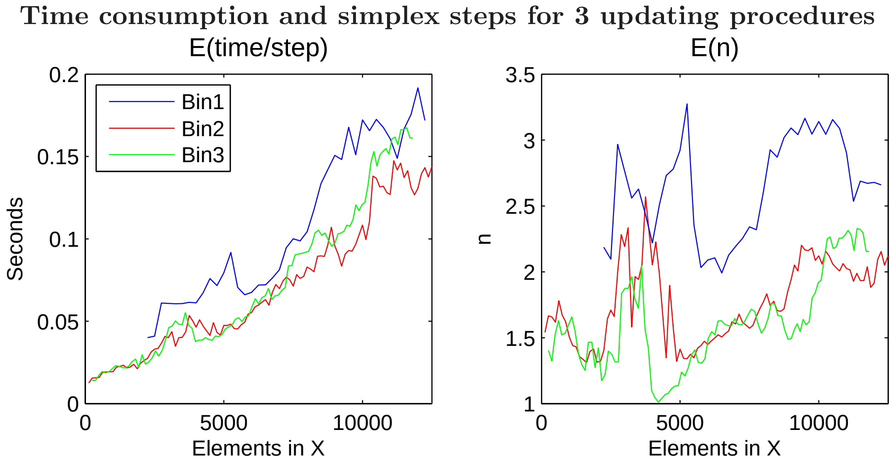

Figure 5.14: The mean time per time step and the mean number of simplex iterations per time step as a function of the number of rows in the design matrix **X**

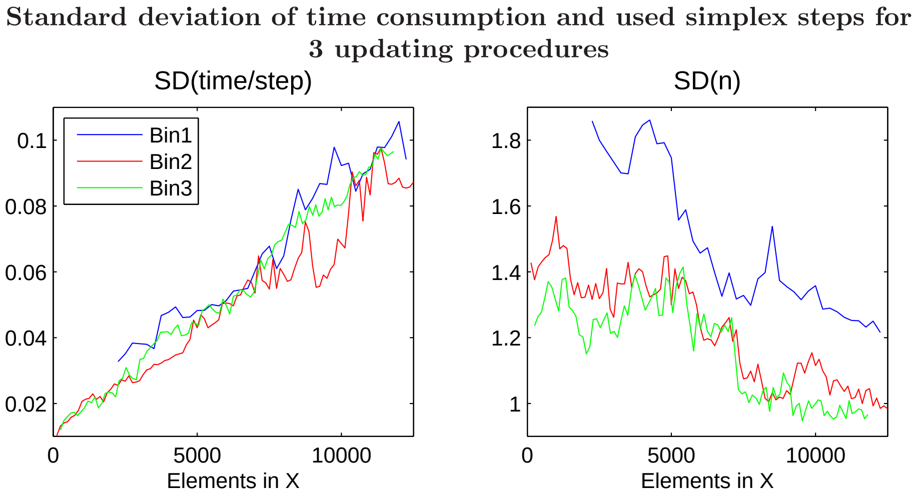

Figure 5.15: The standard deviation of the cpu time used per time step and the standard deviation of the number of simplex iterations per time step as functions of the number of elements in the design matrix **X**.

A similar problem is that the models can produce forecast outside the interval of prediction. The installed power at the Tunø Knob wind power plant is $5000kW$, so predictions of the quantiles of prediction error plus the predicted power above this value is also undesirable. This sum is also bounded below, but not by zero as one should think. The observed power can actually be negative.

The smallest observed power in the data set is $-72kW$ and the largest is $4726kW$. To analyze the performance of the models in this aspect we count the number of observation outside the interval $[-100; 5000]$. The choice of the lower boundary is quite arbitrary, but since we do not have any observations outside this interval, neither the 25% or the 75% quantile should produce predictions outside this interval.

None of the Basic models produces such forecasts. Table 5.8 gives the number of forecasts outside this interval for Model A1-A4. As can be seen from this table Model A3 and A4 have bad performance in this aspect, so the property that these had few crossings should be weighted against the point that they produce more forecasts outside the interval of definition.

As mentioned above the basic model does not produce forecasts outside the interval of definition. This is because these do not add up effects of different explanatory variables. The problem is in the additive model, where effects of

Number of forecast outside the interval $[-100; 5000]$

Model	**A1**	**A2**	**A3**	**A4**
$Q(\mathbf{x}; 0.25)$+**pow.fc**< -100	978	1281	3189	2974
$Q(\mathbf{x}; 0.25)$+**pow.fc**> 5000	13	0	87	48
$Q(\mathbf{x}; 0.75)$+**pow.fc**< -100	371	191	1023	128
$Q(\mathbf{x}; 0.75)$+**pow.fc**> 5000	87	23	0	243

Table 5.8: The number of elements outside the interval $[-100; 5000]$. The probability of getting observations outside this interval is zero.

different components are added and mixed effects are ignored.

A possible solution is to model a transformation of the power. This transformation should be a monotone transformation of the interval of definition into $\mathbb{R}$ and then model the transformed power. With such a set up we can not predict anything outside the interval of definition.

Another possibility is to set up the restrictions in the linear programming problem. How this could be done will be shown in the next chapter where a similar solution to the problem of quantile crossing are also suggested.

5.9 Discussion and Conclusion

The main conclusion of this chapter is that the adaptive procedure introduced in the beginning of the chapter works, and that it performs very convincing compared to equivalent static models. In general all performance parameters were improved, with this procedure. Table 5.9 gives a summary of the models considered so far. For every performance parameter the models are ordered and the 3 best performing models are given their number in the table. This table is primarily given to argue for the adaptive approach. We see from this table that the adaptive models perform better than the static model in nearly all parameters.

The simple adaptive models with only forecasted power as input perform very well compared to the more complicated model. A point here is that models of different complexity should be checked. In this presentation the models that have been analyzed are either very simple or very complex.

For the simple models the conclusion is that the estimates should be based on about 4000 data points. This is equivalent to a little less than two month.

Unfortunately the horizons we can model with the available data is only 18-36 hours, while the required horizons are 12-36 hours. So this should of course also be checked.

The timing for the adaptive procedure is good enough for an online implementation. A possible problem for this procedure is that it require past observations to be stored, since these are needed for the simplex algorithm. A Basic model as mentioned above require access to a matrix with $4000 \cdot 6$ elements. The bright side is that we only need one matrix to model all the quantiles we are interested in. When we have the matrix we only need the index set h to characterize each quantile model. This index set is a small vector, and with the Basic model this have the length 6.

In this chapter we saw that the performance parameters should be combined. It is not enough to look at one of these and then conclude on the basis of that one. The extreme behavior of Model A1 was only detected when we really looked for large crossings or at the Skill score. The extreme crossings were not really punished in sharpness, and it was awarded in resolution. A possibility could be to use the absolute value of IQR to construct sharpness. Then negative value in IQR would be punished by sharpness. By using absolute of negative IQR we would not award resolution either as is the case without taking absolute value first.

Summary of performance parameters

Model	**1**	**2**	**3**	**4**	**B1**	**B2**	**B3**	**B4**	**A1**	**A2**	**A3**	**A4**
Below **25%**			3	3			2	1				
Below **75%**						3		3			2	1
$d(q(\mathbf{pow.fc}, 0.25))$								2	3	1		
$d(q(\mathbf{pow.fc}, 0.5))$								3	1	2		
$d(q(\mathbf{pow.fc}, 0.75))$						2	3	1				
$dq_{total}(\mathbf{pow.fc})$						3		1	3	2		
$d(q(\mathbf{hor}, 0.25))$				3					2		1	2
$d(q(\mathbf{hor}, 0.5))$									2		1	3
$d(q(\mathbf{hor}, 0.75))$			2								3	1
$dq_{total}(\mathbf{hor})$									3		1	2
$d(q(\mathbf{time}, 0.25))$						2		3			1	
$d(q(\mathbf{time}, 0.5))$									1	2		3
$d(q(\mathbf{time}, 0.75))$						2	3	1			3	3
$dq_{total}(\mathbf{time})$						2		2			2	1
$\overline{\rho}_{0.25}(\mathbf{r})$						2	3	1				
$\overline{\rho}_{0.75}(\mathbf{r})$					1						1	2
$\overline{\rho}_{0.75}(\mathbf{r}) + \overline{\rho}_{0.75}(\mathbf{r})$						2	3	1				
$E(\mathbf{IQR})$										2	3	1
$Q(\mathbf{IQR}; 50)$										2	3	1
$sd(\mathbf{IQR})$						2		3	1			
Crossings			1	1							2	3
$\min(\mathbf{IQR})$			1	1							2	3
$E(\mathbf{IQR} < 0)$			1	1							2	3

Table 5.9: The table gives a summary of the performance parameters discussed in this chapter. The table marks the three best performing models in each parameter and these are given their number in the table (1 is the best performer).

Chapter 6

Solutions to the Crossing Problem

6.1 Introduction

As we saw in Chapter 4 and 5 crossings even between quantiles far away from each other such as, the 25% and 75% quantile can occur. Such crossings are of course not desirable either from a theoretical point of view and from a forecaster point of view. In [2] and [3] Koenker notes that in linear quantile regression crossings will always occur in some areas of the sample space. The underlying assumption in such a statement is that the independent variables can take values in all $\mathbb{R}$. If this is the case then all quantile hyper planes have to be parallel if there should be no quantile crossings. In the setting of the data we have analyzed, such a statement is clearly not reasonable.

In this chapter we will propose and analyze some solutions to this problem.

6.2 A Simple Approach

We use quantile regression on cubic splines and therefore our regression is not on variables with a sample space equal to $\mathbb{R}^K$, but a subset $\mathbb{R}^K$. We can therefore in principle demand that there is no crossings in the samples pace.

If we estimate the quantiles at two levels τ_1 and τ_2 with $\tau_1 \neq \tau_2$, then the demand of no crossing is that for every $\mathbf{x} \in \mathcal{P}$, where $\mathcal{P} \subset \mathbb{R}^p$ with p being the number of explanatory variables and $\mathcal{P}$ the sample space of explanatory variables, then following have to hold true

$$\mathbf{b}(\mathbf{x})^T \text{sign}(\tau_1 - \tau_2)(\hat{\beta}(\tau_1) - \hat{\beta}(\tau_2)) \geq 0 \quad \forall \mathbf{x} \in \mathcal{P} \tag{6.1}$$

as soon as we choose a point $\mathbf{x}$ this is a linear constraint and we can incorporate this in the simplex algorithm presented in Chapter 2.

As described in Section 4.2, the basic assumption in our model is that we can write the quantile model as

$$Q(\mathbf{x}; \tau) = \alpha(\tau) + \sum_{j=1}^{p} f_j(x_j; \tau) \tag{6.2}$$

with the assumption that the variables x_j take values independently of each other the non crossing demand becomes (with $\tau_2 > \tau_1$)

$$\begin{aligned}
\min_{\mathbf{x}}(Q(\mathbf{x};\tau_2) - Q(\mathbf{x};\tau_1)) &= \min_{\mathbf{x}} \left(\alpha(\tau_2) - \alpha(\tau_1) + \sum_{j=1}^{p} \Big(f_j(x_j;\tau_2) - f_j(x_j;\tau_1) \Big) \right) \\
&= \alpha(\tau_2) - \alpha(\tau_1) + \sum_{j=1}^{p} \min_{x_j} \left(\Big(f_j(x_j;\tau_2) - f_j(x_j;\tau_1) \Big) \right) \\
&= \alpha(\tau_2) - \alpha(\tau_1) + \sum_{j=1}^{p} \min_{x_j} \left(\Big(f_j(x_j;\tau_2) - f_j(x_j;\tau_1) \Big) \right) \\
&\geq 0
\end{aligned} \tag{6.3}$$

Since the functions f_j was approximated with splines, the minimums described above are 3th degrees polynomials of x_j between knots with coefficients that are linear combinations of elements from β. Therefore solutions to then minimum could be found as functions of β, these would however be nonlinear in β and it is therefore not possible to introduce in the LO problem.

The procedure examined here is to choose a number of points from $\mathcal{P}$ and then use the demand (6.1) in these points, the hope is now that these demands result in non crossing quantiles on the test set. In [21] Takeuchi propose to avoid crossings in the training set, and then hope that this will imply no crossing in the test set. We will use a different approach, namely to avoid crossings in a discrete subset (independent of the training set) of the sample space. This results in fewer constraints, than no crossings in the observations would, at least when we have only one explanatory variable.

Assuming we have chosen a number of points from $\mathcal{P}$, these points are then collected in a matrix $\mathbf{X}_{nc}$ the rows of $\mathbf{X}_{nc}$ are the spline basis functions of the points chosen from $\mathcal{P}$. With this the constraints are

$$\mathbf{X}_{nc}\text{sign}(\tau_2 - \tau_1)(\beta(\tau_2) - \beta(\tau_1)) \geq \mathbf{0} \tag{6.4}$$

this can be put directly into our LO problem, as a first approach we will use the estimate of one quantile to calculate the next. I.e. we keep $\beta(\tau_1)$ constant and calculate $\beta(\tau_2)$ s.t. the will be no crossings between the two quantile curves. With the notation of Chapter 2 we get

$$\mathbf{A} = \begin{bmatrix} \mathbf{X} & \mathbf{I} & -\mathbf{I} & \mathbf{0} \\ \mathbf{X}_{nc} & \mathbf{0} & \mathbf{0} & \text{sign}(\tau_2 - \tau_1)\mathbf{I} \end{bmatrix} \tag{6.5}$$

$\mathbf{y}$ is expanded with $\mathbf{y}_{nc} = \mathbf{X}_{nc}\beta(\tau_1)$, $\mathbf{x}$ is expanded with N_{nc} extra rows, which dependt on the start guess and $\mathbf{c}$ is expanded with $\mathbf{0}_{N_{nc}}$. $h(\tau_1)$ is used as a start guess for $h(\tau_2)$ and the objective function in $\mathbf{c}$ is simply changed according to τ_2.

The principles in solving this problem is exactly the same as what was presented in Chapter 2. The only difference is that now we have to deal with infeasible points, an infeasible point is a point where $x_i < 0$, until now this have just been the same $r^+ < 0$ or $r^- < 0$. This can be fixed by multiplying one entry of $\mathbf{P}$ by -1 and changing one element in $\mathbf{c}_{\mathcal{B}}$.

In the setting with $\mathbf{X}_{nc}$ we have to deal with infeasible points in some other way, infeasible points will occur as we iterate through the solutions, this is due to small errors which are summed up in each iteration. Such problems can be solved in different ways, the one used here is essentially the simplex algorithm, but with a objective function that punish infeasible points, this is described in [14] p. 75-76. This procedure does not seem to be well suited if we get many crossings and we get many crossings in the first simplex steps in this set up.

A possible improvement for this is to use a dual simplex algorithm to solve for infeasible points, Appendix C discuss the set up for dual simplex in the context of non crossing quantile regression.

The first approach we use is to find the optimal solution at level τ_1 as a start guess for the optimal solution of τ_2, the solution at τ_2 is now used to iterate to the next solution at level τ_3 etc.. The point is simply that the solution τ_{i-1} is used as starting point for the algorithm with a new objective function corresponding to τ_i, with the non crossing constraint fulfilled at all times during the iterations.

If we only have one explanatory variable, i.e. $\mathcal{P} \subset \mathbb{R}$, then it is not a problem to choose $\mathbf{X}_{nc}$, since in this case we can just choose a series of numbers in $\mathcal{P}$. E.g. if we use only **pow.fc** as input then we can let $\mathbf{x}_{nc} = \{0, 10, ..., 5000\}$ and $\mathbf{X}_{nc} = [\mathbf{b}(\mathbf{x}_{nc})]$. If we do not have crossings at $\mathbf{x}_{nc}$ then the chance of having crossings between these points will also be small. If we use more explanatory variables it becomes difficult to choose the non crossing constraints, since the number of constraints is the product of constraints in each direction.

We use the simple model set up used earlier, with **pow.fc** as explanatory variable, 5 knots at 20% sample quantiles of **pow.fc**, 10^4 data points in the training set and $\mathbf{x}_{nc} = \{0, 10, ..., 5000\}$

Figure 6.1 show conditional CDF as a function of all values of forecasted power, calculated from $\tau_0 = 0.5$ in each direction (of τ) demanding no crossings in each step. Even though it is not easy to see if there is crossings in a plot like this, things seems to be in the right order. This is confirmed by Table 6.1 where statistics related to IQR is listed for different choices of $\Delta\tau$, Reference refer to $\mathbf{x}_{nc} = \emptyset$. What is seen here is that we still have 52 crossings in the test set for $\Delta\tau = 0.05$ and $\Delta\tau = 0.01$ but that the size of these crossings have been reduced to the order of 10^{-12}. Which in this context must be considered to be equal zero.

The table also gives the timing of the models. We see that it is a very time consuming task to calculate many non crossing quantiles in this way. We should however consider that this can not be compared to the adaptive model or something like that. We can not expect to be close to the solution, with the start guess we use here. If we made an adaptive implementation for this procedure and used the solutions as a start guess then we could expect to see much better performance with respect to timing.

That the model does not change much is also seen in Figure 6.2, where the loss function is plotted as a function of τ for the Reference model (left column). The relative difference to the non crossing models is plotted in the right column. It is seen that these does not differ much, the non crossing quantiles have larger loss functions on the training set (as they must have), but this is not necessarily the case on the test period. The loss functions are close in all cases.

The shape of these loss functions might be a little surprising, it is seen that the loss functions goes to zero and are not symmetric around $\tau = 0.5$. To

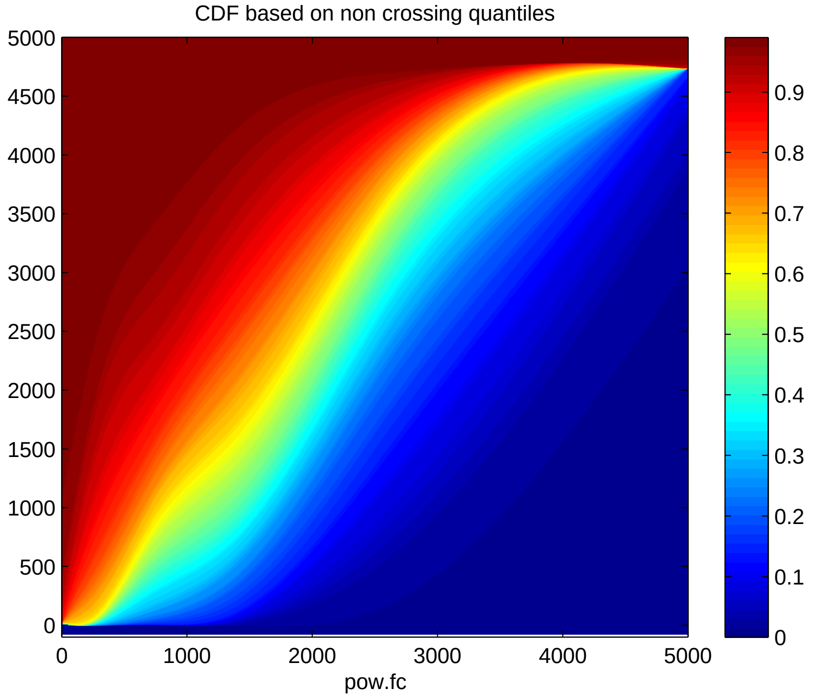

Figure 6.1: Conditional CDF as a function of pow.fc, for the basic model of the foregoing section and based on the non crossing algorithm. The calculations is done in steps of 0.01 in τ.

Key numbers for the models

Model	**Reference**	$\boldsymbol{\Delta\tau = 0.05}$	$\boldsymbol{\Delta\tau = 0.01}$	$\boldsymbol{\Delta\tau = 0.005}$	$\boldsymbol{\Delta\tau = 0.001}$
$E(\mathbf{IQR})$	1020.3	1021.4	1021.3	1020.0	1020.1
$sd(\mathbf{IQR})$	648.8	641.2	641.2	639.4	639.3
$Q(\mathbf{IQR}; 0.5)$	1100.5	1085.9	1084.7	1085.9	1086.0
$Q(\mathbf{IQR}; 0.05)$	86.3	100.5	100.6	100.5	100.7
$Q(\mathbf{IQR}; 0.95)$	1850.7	1872.7	1873.9	1867.2	1867.1
Crossings	113	52	52	0	0
$\min(\mathbf{IQR})$	-346.8	$-0.9 \cdot 10^{-12}$	$-3 \cdot 10^{-12}$	$2.3 \cdot 10^{-3}$	$17.6 \cdot 10^{-3}$
$E(\mathbf{IQR} < 0)$	-176.6	$-0.9 \cdot 10^{-12}$	$-3 \cdot 10^{-12}$	-	-
Time (minutes)		14.95	34.46	51.53	159.49

Table 6.1: Numbers related to IQR for the reference model and the non crossing model, the timing is the accumulated time to calculated the whole distribution.

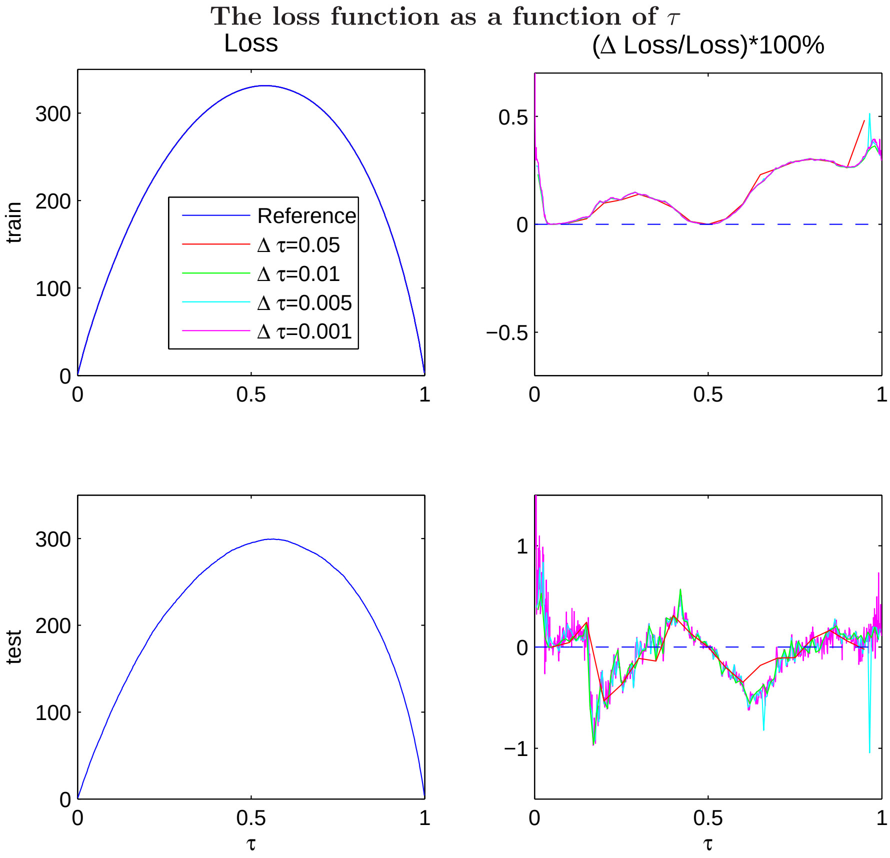

Figure 6.2: The first column show the Loss function based on the reference model, i.e. there are no constraints and the quantiles can cross. The second column show the difference from the reference model to the models calculated with the non-crossing algorithm.

understand this we analyze the sample quantile case and look at the expected loss as a function of τ given that we know the true quantile, i.e. $\hat{Q}(\tau) = F^{-1}(\tau)$ and y have the p.d.f. f. Let (a, b) be the support of f (it is not important if this interval is closed or open and a and b can be equal to $-\infty$ or ∞ respectively), with this we have

$$\begin{aligned}
E(\rho_\tau(y)|F(y)) &= \tau \int_{F^{-1}(\tau)}^{b} (y - F^{-1}) f(y)dy + (1-\tau) \int_{a}^{F^{-1}(\tau)} (F^{-1} - y) f(y)dy \\
&= \tau \left(\int_a^b y f(y)dy - F^{-1}(\tau) \int_a^b f(y)dy \right) && (6.6) \\
&\quad + F^{-1}(\tau) \int_a^{F^{-1}(\tau)} f(y)dy - \int_a^{F^{-1}(\tau)} y f(y)dy && (6.7) \\
&= \tau(E(y) - F^{-1}(\tau)) + F^{-1}(\tau)\tau - \int_a^{F^{-1}(\tau)} y f(y)dy && (6.8) \\
&= \tau E(y) - \int_a^{F^{-1}(\tau)} y f(y)dy && (6.9)
\end{aligned}$$

since $F^{-1}(\tau \to 0) = a$ and $F^{-1}(\tau \to 1) = b$, we can write down the value of $E(\rho_\tau|F(y))$ as τ approaches the endpoints

$$\begin{aligned}
\lim_{\tau \to 1} E(Loss|\tau) &= E(y) - E(y) = 0 && (6.10) \\
\lim_{\tau \to 0} E(Loss|\tau) &= 0E(y) - \int_a^a y f(y)dy = 0 && (6.11)
\end{aligned}$$

technically the arguments above require that the distribution f to have a expectation, and quantile regression does not require that. This is however a very special case so it is fair to conclude that the loss function have to go to zero at $\tau = 0$ and $\tau = 1$. This could also be realized by looking at the loss function of τ.

We can also write down the conditions for symmetry of the loss function. To demand symmetry is the same as demanding that $\Delta E(\rho_{0.5+\epsilon}(y) - \rho_{0.5-\epsilon}(y)|F(y)) = E(\rho_{0.5+\epsilon}(y)|F(y)) - E(\rho_{0.5-\epsilon}(y)|0.5 - \epsilon) = 0$ for all $\epsilon \in [0, 0.5)$, using (6.9) we get

$$\Delta E(\rho_{0.5+\epsilon}(y) - \rho_{0.5-\epsilon}(y)|\epsilon) = 2\epsilon E(y) - \int_{F^{-1}(0.5-\epsilon)}^{F^{-1}(0.5+\epsilon)} y f(y)dy \quad (6.12)$$

so if the loss function should be symmetric around $\tau = 0.5$ then

$$2\epsilon E(y) = \int_{F^{-1}(0.5-\epsilon)}^{F^{-1}(0.5+\epsilon)} y f(y)dy \quad (6.13)$$

we see that if we have symmetry then the integral on the right hand side should be a linear function of the expectation of y, so the differential of this integral should be constant equal to $2E(y)$. To differentiate such an integral we need Leibniz integration rule, this is

$$\frac{\partial}{\partial z}\int_{a(z)}^{b(z)} f(x,z)dx = \int_{a(z)}^{b(z)} \frac{\partial f}{\partial z}dx + f(b(z),z)\frac{\partial b}{\partial z} - f(a(z),z)\frac{\partial a}{\partial z} \tag{6.14}$$

with $f(x,z) = f(x)$ this reduce to

$$\frac{\partial}{\partial z}\int_{a(z)}^{b(z)} f(x)dx = f(b(z))\frac{\partial b}{\partial z} - f(a(z))\frac{\partial a}{\partial z} \tag{6.15}$$

we can therefore write

$$\begin{aligned} 2E(x) &= \frac{\partial}{\partial \epsilon}\int_{F^{-1}(0.5-\epsilon)}^{F^{-1}(0.5+\epsilon)} yf(y)dy \\ &= F^{-1}(0.5+\epsilon)f(F^{-1}(0.5+\epsilon)\frac{dF^{-1}(0.5+\epsilon)}{d\epsilon} \\ &\quad -F^{-1}(0.5-\epsilon)f(F^{-1}(0.5-\epsilon)\frac{dF^{-1}(0.5-\epsilon)}{d\epsilon} \\ &= F^{-1}(0.5+\epsilon)\frac{dF(F^{-1}(0.5+\epsilon))}{d\epsilon} - F^{-1}(0.5-\epsilon)\frac{dF(F^{-1}(0.5-\epsilon))}{d\epsilon} \\ &= F^{-1}(0.5+\epsilon)\frac{d(0.5+\epsilon)}{d\epsilon} - F^{-1}(0.5-\epsilon)\frac{d(0.5-\epsilon)}{d\epsilon} \\ &= F^{-1}(0.5+\epsilon) + F^{-1}(0.5-\epsilon) \end{aligned} \tag{6.16}$$

by setting $\epsilon = 0$ we get $F^{-1}(0.5) = E(Y)$ so a first requirement a symmetric loss function is that the mean the expectation is equal to the median. With this we get the requirement that

$$F^{-1}(0.5) = \frac{1}{2}\left(F^{-1}(0.5+\epsilon) + F^{-1}(0.5-\epsilon)\right) \quad \forall \epsilon \in [0, 0.5) \tag{6.17}$$

This is the same as complete symmetry aruond the median which was equal to the expectation. These arguments show that we can only expect the loss as a function of τ to be symmetric around $\tau = 0.5$ in very special situations. Therefore the picture in Figure 6.2 should be expected.

With the construction of non crossing quantiles as above, the assumption in the Theorems of Chapter 2 does not necessarily hold true any more. E.g. Theorem 2.3 tells us that the quantile curve split the data set in two parts and that the number of elements in these part are close to τN and $(1-\tau)N$. The central argument in the proof of Theorem 2.3 was that the directional derivative should

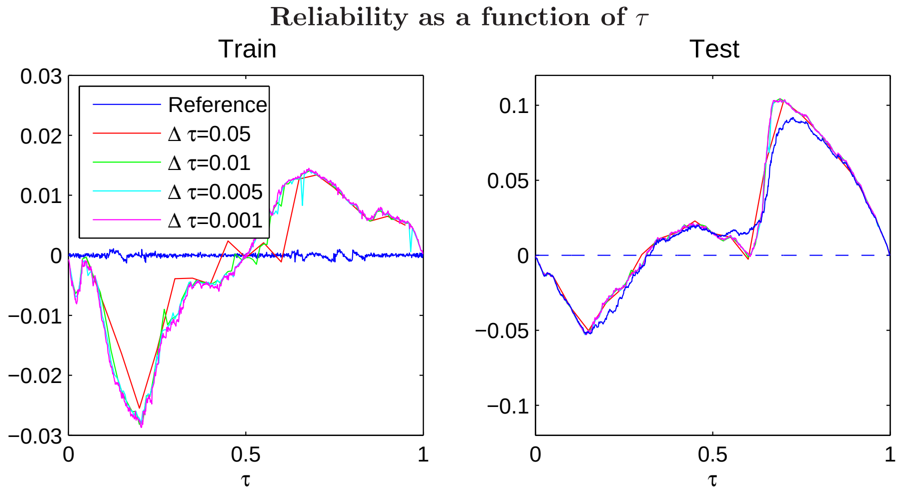

Figure 6.3: The overall reliability for the reference model and the different non crossing models, first column is the training set and the second column is the test set.

be zero in all directions, this can not be assumed any more, since there can now be decent direction out of the feasible region.

Figure 6.3 show the reliability as a function of τ. We see that the Reference model split the data space as it should on the training set, while it seems like the other quantiles are pushed out by the non-crossing constraints.

From a theoretical point of view something like Figure 6.3 is problematic, if this also holds asymptotically then the non-crossing quantile estimator is not consistent, which would normally be a minimum requirement for an estimator. From a practically point of view Figure 6.3 is not so problematic, since the performance in the reliability sense is very close for the reference model and the other models on the test set. We can of course not draw asymptotic conclusions from Figure 6.3. We can however say that Theorem 2.3 does not hold under these conditions.

In [20] Zhao analyze 2 different restricted regression quantiles (RRQ), in this paper a RRQ is a modified version of the quantile regression, where restrictions that guarantees no crossings in some points or areas of the sample space are imposed. The restrictions are imposed in two different ways. The first model is a linear model where parallel quantile plans are obtained, and thereby guarantees no crossing in the sample space. The second model is a linear hesteroscedastic model

$$y = \mathbf{x}^T\beta + (\mathbf{x}^T\gamma)\epsilon \tag{6.18}$$

with a three step procedure that guarantee no crossings in at the training set.

In this paper Zhao show that that both the models are consistent and with parameter estimates asymptoticly normal (with a very complicated variance structure). The restrictions in the hesteroscedatic model is build into the estimation procedure, where every quantile is estimated separately but with restrictions that make ensure they do not cross at the training points. In both cases Zhao assumes iid errors and this seems to be important for the consistency results. As was discussed in Section 4.5.2, we can not assume something like that in our data.

Even though we can not use these results directly, they give indications that the behavior of Figure 6.3 is somewhat strange and that the result probably does not apply asymptoticly.

6.3 Simultaneous Estimation of Several Quantiles

In the set up for non crossing quantiles in the previous section, the silent assumption was that the 50% quantile estimate is correct or at least more correct than other quantile estimates. This assumption is used when we choose the 50% quantile as the restriction curve for the other estimates.

There are good reasons to have better confidence in central quantiles, one thing is that the estimates of the central quantiles does not depend on tails in the conditional distributions, events. While the quantiles near zero or one depend on the tails in the conditional distributions.

This can also be realized from the asymptotic theory of quantile regression. We saw in Chapter 2 that the asymptotic distribution of the estimates was normal with mean zero and in the idd case a variance that depended on $\tau(1-\tau)/f^2(F^{-1})$. In the general setting the variance was much more complicated. To understand how this variance function depend on τ and the distribution we study the iid case in Example 6.1.

Example 6.1 We look at the sample quantile case from an iid sample. The variance of $\sqrt{N}x_{(k)}$ with $k = \lceil \tau N \rceil$ as an estimation of $\sqrt{N}F^{-1}(\tau)$ (see Section

2.4.2) is

$$\omega^2(\tau) = \frac{\tau(1-\tau)}{f^2(F^{-1}(\tau))} \tag{6.19}$$

if x_i come from a standard normal distribution then this variance is

$$\omega^2(\tau) = 2\pi\tau(1-\tau)e^{(\Phi^{-1}(\tau))^2} \tag{6.20}$$

this variance is shown in Figure 6.4, the figure also shows the variance for a non central student t distribution and an Exponential distribution, these are shown to study the dependence between the variance and the distribution.

Figure 6.4 show theoretical asymptotic variance of estimates of the sample quantiles taken from these distributions, what is seen is that these variances depend on the sample distribution. We see that heavy tails give large variance, and that asymmetries in the distribution also give asymmetries in the variance. For the exponential distribution we see that the variance go to zero as τ goes to zero.

Having noted this it is of course difficult to take this into account when we choose the first quantile to estimate in the non crossing algorithm. □

If we had knowledge like what is needed to construct plots like Figure 6.4, we would not really have to estimate quantiles. Therefore the choice of starting with $\tau = 0.5$ is simple and for distributions with some symmetry we should not be that far from the best choice. Another approach could be to estimate the unconditional distribution of the respons variable and then use that to make a variance plot like in Figure 6.4 and from this choose the first quantile to estimate.

A better approach might be to estimate several quantiles simultaneous under the non crossing constraints. Here we will set up the problem and do some analysis of the structure that we might be able to exploit in a practical set up.

Consider $\mathbf{t} = [\tau_1, \tau_2, ..., \tau_l]$ with $\mathbf{t}$ being the quantiles we want to estimate, assume that $\mathbf{t}$ is ordered s.t. $0 < \tau_1 < \tau_2 < ... < \tau_l < 1$. The model is now

$$\hat{\mathbf{Q}}(\mathbf{t}, \mathbf{x}) = \mathbf{x}^T \hat{\Theta}(\mathbf{t}) \tag{6.21}$$

where Θ is a matrix given by

$$\Theta = [\beta(\tau_1) \quad \beta(\tau_2) \quad ... \quad \beta(\tau_l)] \tag{6.22}$$

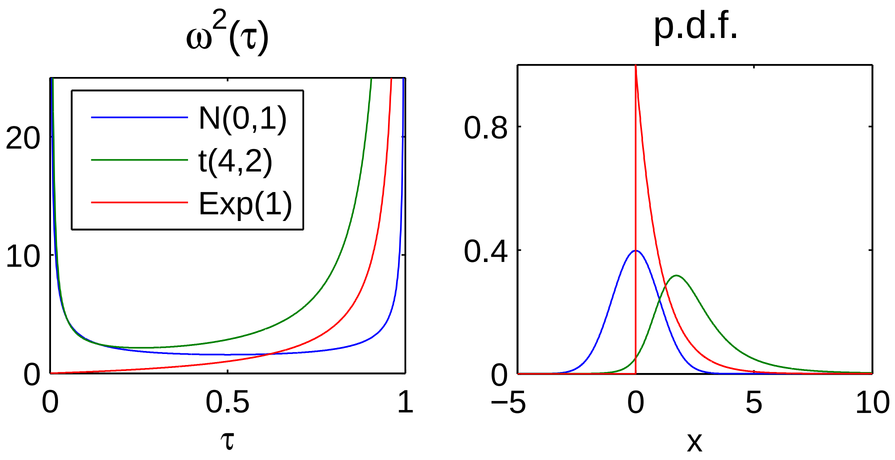

Figure 6.4: Theoretical asymptotic variance for quantile estimators of 3 different p.d.f's as a function of the quantile τ to estimate

so with the point $\mathbf{x}$ we get a vector of quantiles estimates. The estimate of $\hat{\Theta}$ is the solution to the problem

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{j=1}^{l} \sum_{i=1}^{N} \rho_{\mathbf{t}_j}(y_i - \mathbf{x}_i^T \Theta_{:,j}) \tag{6.23}$$

$$= \arg\min_{\Theta} \sum_{j=1}^{l} \sum_{i=1}^{N} \rho_{\tau_j}(y_i - \mathbf{x}_i^T \beta_j) \tag{6.24}$$

$$= \arg\min_{\Theta} \sum_{j=1}^{l} S(\beta_j; \tau_j, \mathbf{r}_j) \tag{6.25}$$

subject to some constraints, in the case of no crossings these are

$$\mathbf{X}_{nc}(\beta_{j+1} - \beta_j) \geq \mathbf{0}; \quad j = 1, 2..., l \tag{6.26}$$

compared to the previous section we now let both β_j and β_{j+1} vary to get a better solution.

6.3.1 A Weight Function

If the variance of each $\hat{\beta}_j$ was constant we would like the expected loss to be constant as a function of τ as well. Now we know that neither the expected loss function or the variance of $\hat{\beta}$ are constant as functions of τ.

To argue for a weight function for ρ_τ we look at the least square estimation in the linear model

$$y_i = \mathbf{x}_i^T \beta + r_i \tag{6.27}$$

if r_i is assumed to be iid then the loss function is

$$S_{ls} = \sum_{i=1}^{N} r_i^2 = \sum_{i=1}^{N} \rho_{ls}(r_i) \tag{6.28}$$

if the r_i's are independently distributed with $V(r_i) = \sigma_i^2$ (σ not constant) then we should use the weighted least square, now the loss function becomes

$$S_{wls} = \sum_{i=1}^{N} \frac{r_i^2}{\sigma_i^2} = \sum_{i=1}^{N} \rho_{wls}(r_i) \tag{6.29}$$

Now look at the loss function for the median, this is

$$S(\beta; 0.5, \mathbf{r}) = 0.5 \sum_{i=1}^{N} |r_i| = 0.5 \sum_{i=1}^{N} \sqrt{\rho_{ls}(r_i)} \tag{6.30}$$

if we replace ρ_{ls} with ρ_{wls} we get

$$S_w(\beta; 0.5, \mathbf{r}) = 0.5 \sum_{i=1}^{N} \frac{|r_i|}{\sigma_i} \tag{6.31}$$

this can, with the above argumentation, be consider as a weighted median loss function. We now propose to minimize the following sum of loss functions

$$\sum_{j=1}^{l} w(\tau_j) S(\beta_j; \tau_j, \mathbf{r}_j) \tag{6.32}$$

with

$$w(\tau_i) = \frac{1}{\sigma_\beta(\tau_i) E(\rho(\tau_i(y)|F)} \tag{6.33}$$

with this the quantiles are weighted acording to the variance, since both $\sigma_\beta(\tau_i)$ and $E(\rho(\tau_i(y)|F)$ depend on the distribution function, we can not really find the weights unless we make some assumptions on the distribution F. To do something simple we calculate $w(\tau_i)$ under the assumption that we have drawn

y_i from a standard normal distribution and wish to estimate sample quantiles, under this assumption and using (6.9)we get

$$\begin{aligned}
E(\rho_\tau(y)|F) &= \tau E(y) - \int_{-\infty}^{F^{-1}(\tau)} y f(y) dy && (6.34)\\
&= -\int_{\infty}^{\frac{1}{2}F^{-1}(\tau)^2} f(y) d\frac{y^2}{2} && (6.35)\\
&= -\frac{1}{\sqrt{2\pi}} \int_{\infty}^{\frac{1}{2}F^{-1}(\tau)^2} e^{-t} dt && (6.36)\\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}F^{-1}(\tau)^2} && (6.37)\\
& && (6.38)
\end{aligned}$$

and

$$\begin{aligned}
\sigma_{\beta(\tau)} &= \frac{\sqrt{\tau(1-\tau)}}{f(F^{-1}(\tau))} && (6.39)\\
&= \frac{\sqrt{\tau(1-\tau)}}{\sqrt{2\pi}} e^{\frac{1}{2}(F^{-1}(\tau))^2} && (6.40)
\end{aligned}$$

with the assumptions above we get

$$w(\tau_i) = \frac{1}{\sqrt{2\pi\tau_i(1-\tau_i)}} \qquad (6.41)$$

The proposed weight is therefore

$$\tilde{\rho_i}(r) = \frac{\rho_i(r)}{2\pi\sqrt{\tau_i(1-\tau_i)}} \qquad (6.42)$$

This loss function should now be used in the estimation for $\hat{\mathbf{\Theta}}$.

6.3.2 The Simplex Formulation

As long as we do not add any constraints the solution to this problem is exactly the same as if that quantiles would be estimated one by one. Following the formulation in Section 2.3.2 we write down the LO problem in the standard form

$$\begin{aligned}
(P) \quad \min\{\mathbf{c}^T\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}, (\mathbf{r}_1^+, \mathbf{r}_1^-, ...\mathbf{r}_l^+, \mathbf{r}_l^-) \in \mathbb{R}_0^{2lN},\\
(\beta(\tau_1), ..., \beta(\tau_l)) \in \mathbb{R}^{lK}\} \qquad (6.43)
\end{aligned}$$

with

$$\mathbf{c} = \begin{bmatrix} \mathbf{0}_{lK} \\ \tilde{\rho_1}(1)\mathbf{e} \\ \tilde{\rho_1}(-1)\mathbf{e} \\ \tilde{\rho_2}(1)\mathbf{e} \\ \tilde{\rho_2}(-1)\mathbf{e} \\ \vdots \\ \tilde{\rho_l}(1)\mathbf{e} \\ \tilde{\rho_l}(-1)\mathbf{e} \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_l \\ \mathbf{r}_1^+ \\ \mathbf{r}_1^- \\ \vdots \\ \mathbf{r}_l^+ \\ \mathbf{r}_l^- \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} \mathbf{y} \\ \vdots \\ \mathbf{y} \end{bmatrix} \tag{6.44}$$

and

$$\mathbf{A} = \begin{bmatrix} \mathbf{X} & \mathbf{0} & \cdots & \mathbf{0} & [\mathbf{I}, -\mathbf{I}] & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X} & \mathbf{0} & \vdots & \mathbf{0} & [\mathbf{I}, -\mathbf{I}] & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{X} & \mathbf{0} & \mathbf{0} & \mathbf{0} & [\mathbf{I}, -\mathbf{I}] \end{bmatrix} = [\mathbf{X}_l, \mathbf{L}] \tag{6.45}$$

with the same notational setting and argumentation as in, Section 2.3.2, we can write

$$\mathbf{B} = \begin{bmatrix} \mathbf{X}_l(h) & \mathbf{0} \\ \mathbf{X}_l(\bar{h}) & \mathbf{P} \end{bmatrix} \tag{6.46}$$

in Section 2.3.2 we saw that the expensive part of the simplex steps was to calculate $\mathbf{X}_l(h)^{-1}$, this is still true. $\mathbf{X}_l(h)$ will be a $(lK) \times (lK)$ matrix, the only elements that will be different from zero are l, $(K \times K)$ blocks along the diagonal. These blocks are as in characterized by index sets h_j, $j = 1...l$, there will be no coupling terms, so $\mathbf{X}_l(h)^{-1}$ will also be a matrix with only l $(K \times K)$ blocks along the diagonal being different from zero and they are equal to $\mathbf{X}(h_j)^{-1}$, with the index sets h_j referring to $\mathbf{X}$ in the same way as in Section 2.3.2.

The case discussed above could the called an uncoupled situation, in this case the matrices $\mathbf{X}_l(h)$, $\mathbf{X}_l(h)^{-1}$ and $\mathbf{X}_l(\bar{h})$ are all sparse if l is large. The relative number of non zero elements in both $\mathbf{X}_l(h)$ and $\mathbf{X}_l(h)^{-1}$ is less than or equal to

$$\frac{K^2 l}{(lK)^2} = \frac{1}{l} \tag{6.47}$$

and the inverse of $\mathbf{B}$ can be calculated as we did in Section 2.3.2, but this relay on the fact that we have not imposed the coupling terms, this will be done below.

With this setting in place we can impose the non crossing constraints, at the points $\mathbf{X}_{nc}$, we require

$$\mathbf{X}_{nc}(\beta_{j+1} - \beta_j) \geq \mathbf{0}; \quad j = 1, ..., l-1 \tag{6.48}$$

We saw in Section 5.8 that the quantile estimator could make forecasts outside the interval of definition, we can avoid this in the same way. Let $[y_{max}, y_{min}]$ be the interval of definition, then we use the constraints

$$\begin{aligned} \mathbf{X}_{nc}\beta_1 &\geq y_{min}\mathbf{e} && (6.49) \\ -\mathbf{X}_{nc}\beta_l &\geq -y_{max}\mathbf{e} && (6.50) \end{aligned}$$

Now we can set up the restricted quantile problem by adding rows and columns to $\mathbf{A}$ and rows to $\mathbf{c}$, $\mathbf{x}$ and $\mathbf{b}$, the LO formulation of the problem is

$$(P) \quad \min\{\mathbf{c}_{nc}^T\mathbf{x}_{nc} : \mathbf{A}_{nc}\mathbf{x}_{nc} = \mathbf{b}_{nc}\} \tag{6.51}$$

subject to

$$\begin{bmatrix} \mathbf{r}_1^+ \\ \mathbf{r}_1^- \\ \vdots \\ \mathbf{r}_l^+ \\ \mathbf{r}_l^- \\ \mathbf{s} \end{bmatrix} \in \mathbb{R}_0^{2lN+(l+1)N_{nc}}; \quad \begin{bmatrix} \beta(\tau_1) \\ \vdots \\ \beta(\tau_l)) \end{bmatrix} \in \mathbb{R}^{lK} \tag{6.52}$$

with

$$\mathbf{c}_{nc} = \begin{bmatrix} \mathbf{c} \\ \mathbf{0} \end{bmatrix} \quad \mathbf{x}_{nc} = \begin{bmatrix} \mathbf{x} \\ \mathbf{s} \end{bmatrix} \quad \mathbf{b}_{nc} = \begin{bmatrix} \mathbf{b} \\ y_{min}\mathbf{e}_{nc} \\ \mathbf{0}_{nc} \\ \vdots \\ \mathbf{0}_{nc} \\ -y_{max}\mathbf{e}_{nc} \end{bmatrix} \quad \mathbf{A}_{nc} = \begin{bmatrix} \mathbf{X}_l & \mathbf{L} & \mathbf{0} \\ \mathbf{X}_{NC} & \mathbf{I}_{l\cdot nc} & \mathbf{0} \end{bmatrix}$$

and

$$\mathbf{X}_{NC} = \begin{bmatrix} \mathbf{X}_{nc} & \mathbf{0} & & & \\ -\mathbf{X}_{nc} & \mathbf{X}_{nc} & \mathbf{0} & & \\ \mathbf{0} & -\mathbf{X}_{nc} & \mathbf{X}_{nc} & \ddots & \\ & \ddots & \ddots & \ddots & \mathbf{0} \\ & & \mathbf{0} & -\mathbf{X}_{nc} & \mathbf{X}_{nc} \\ & & & \mathbf{0} & -\mathbf{X}_{nc} \end{bmatrix} \tag{6.53}$$

In principle we have the matrix formulation of the basic solution as was constructed above, but the matrix $\mathbf{X}_{NC}$ give rise to off diagonal elements that course all elements (worse case) of the inverse of $\mathbf{X}(h)$ to be different from zero. With

$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_l \\ \mathbf{X}_{NC} \end{bmatrix} \tag{6.54}$$

we can write $\mathbf{B}$ as

$$\mathbf{B} = \begin{bmatrix} \tilde{\mathbf{X}}(h) & \mathbf{0} \\ \tilde{\mathbf{X}}(\bar{h}) & \mathbf{P} \end{bmatrix} \tag{6.55}$$

The dimension of $\tilde{\mathbf{X}}$ is $l(N+N_{nc}) \times (l \cdot K)$. In the sitting of Section 6.2 with non crossing quantiles calculated from the median, we had $N = 10^4$, $N_{nc} = 500$ and $K = 6$. If we want to estimate the quantiles in steps of 5% then the dimension of $\tilde{\mathbf{X}}$ in this setting is $19 \cdot 10500 \times 19 \cdot 5$ or about $2 \cdot 10^5 \times 95$, the problem is not this matrix since this is sparse and the non zero blocks are $\mathbf{X}$ or $\mathbf{X}_{nc}$ so we just have to keep track of the placement of these. The problem is that $\tilde{\mathbf{X}}(h)^{-1}$ is not sparse and that $\tilde{\mathbf{X}}(\bar{h})\tilde{\mathbf{X}}(h)^{-1}$ is therefore not sparse either, and this have the same dimension as $\tilde{\mathbf{X}}$. There is really not any chance of working with such a matrix, even with relatively few quantiles to estimate.

We could hope that it is possible to exploit the structure of $\tilde{\mathbf{X}}(h)$ to make some recursive formula for elements in $\tilde{\mathbf{X}}(\bar{h})\tilde{\mathbf{X}}(h)^{-1}$. The point is that we know the structure of $\tilde{\mathbf{X}}(h)$, and we do not necessarily need the full matrix $\tilde{\mathbf{X}}(\bar{h})\tilde{\mathbf{X}}(h)^{-1}$, but only the vector $\mathbf{d}$ and $\mathbf{h}$. Unfortunately there is not time enough to study this further, so this will stand as a suggestion for future work in this field.

6.4 Discussion and Suggestions

This Chapter have discussed an implementation of a non crosssing procedure, the analysis shows that this works in the sense that it produces curves that do not cross. The discussion and analysis raises the question if this is actually a quantile, since it does not split the training set in a proper way. The curves does however perform similar to a Reference quantile on the test set, i.e. a set of quantile curves estimated without the non crossing constraints. The performance analysis had a very narrow focus on Skill Score and overall reliability. Focus have been on the understanding of the results rather than comparing models.

The last part of the Chapter set up the demands and formulate the LO problem for estimation of several quantile simultaneously, even though we are able to give the formulation an implementations for many quantile seems unrealistic, since the set up requires us to work with very large matrices.

In the discussion on the result of the reliability it was mentioned that Zhao in [20] have shown that parallel quantiles are actually consistent. In such a regression the slope is constant over all quantiles and the only difference of these quantiles are their intercept.

In this presentation we have used spline basis functions, and we would not expect that moving only the intercept could lead to anything usefull. The plots we have seen throughout the presentation also support this. We can not bring the 75% quantile curve of prediction error to the 25% quantile curve of prediction error, just by moving the intercept.

The natural spline basis functions that we use throughout this presentations is the ones given by "R"'s spline function, these can take both positive and negative values. Had we used the definition of natural splines given in Section 3.4 the the natural splines would have been functions from $\mathbb{R}$ into a subset of the interval $[0, 1]$, with linear regression on such a set of functions then the requirement $\beta_j(\tau_1) < \beta_j(\tau_2) < ... < \beta_j(\tau_l), \quad j = 1, 2, ..., K$, would lead to global noncrossing estimate. Of course we should then show that such restrictions still give us the flexibility we get from the spline functions. Such a set would solve the problem of choosing the non crossing restrictions, since there would be only K of these, i.e. equal to the number of basis functions. A set up like that does however not solve the problem with off diagonal elements in $\mathbf{X}(h)$, and this would still not be sparse. It might however be simpler to analyze since the off diagonal rows from the $K \times K$ identity matrix.

Chapter 7

Conclusion

This presentation have treated quantile regression with splines. The set up for an implementation of the simplex algorithm in the quantile regression case, was developed in Chapter 2. In the case of quantile regression the simplex set up becomes very simple, because of the structure of the linear constraints. This formulation does not relay on the spline set up.

Quantile regression and splines have been used to model the prediction error from WPPT at the Tunø Knob wind power plant. The data set seems too small to model the phenomena we are interested in, so static models performed very poorly on the test set. The performance parameters of quantiles have been discussed through out the presentation. From the discussion of these it is clear that none of the performance parameters discussed are able to give a clear picture of what model to choose and most of these were not able to punish all undecidable behavior. To illustrate this point, we saw that an adaptive model with very large quantile crossings performing very well in most of the other performance parameters. Reliability is often thought of as the key performance parameter, it is of course important for a model to have good reliability, but it does not punish e.g. crossings.

This point makes it difficult to distinguish between different models, simply because they will be good and bad in different ways. The skill score seems to be able to punish extreme behavior such as very large crossings at least to some extend. A good example of this is the analysis of three different adaptive model in Section 5.7 reliability suggested models with very large crossing, while the skill score punished these models. The skill score did however not react to increase in reliability distance. The combined discussions of this kind makes model selections very hard, what might be clear is that looking at sharpness and resolution before other parameters are considered will lead to wrong conclusions. Selecting a quantile model require us to look at numbers and curves, this is in itself a problem since it make it hard to compare models, these will simply be wrong in different ways.

With this in mind, the analysis of the adaptive models in Chapter 5 showed clear and superior performance compared to the static models. Further it is by far fast enough for an online implementation, with a time use of less than 0.5 second per time step for all models, with the exception of Model A1, which broke down due to the structure of the data.

The data analyzed here cover about one 10 month of time in 2003, this seems to be insufficient for the static model. Even though we performance for the static model an analysis of larger dataset would be a good idea. Especially since the different updating procedure are not studied to a full extend, more data is simply needed.

Non crossing constraints for quantile regression were analyzed in the Chapter 6, in this analysis an implementation for the non crossing constraints was analyzed. This implementation was slow, a central point in this connection is however that we would not expect solution to quantiles at different levels to be close in a simplex sense or any other sense for that matter. With the implementation we would however expect to be close to the set of non crossing quantile at the next time step, therefore an adaptive version of these would be expected to have far better timing.

The implementation of the non crossing quantile, uses the median to calculate the rest of the quantiles with respect to the non crossing constraint. The set up for simultaneous estimation of several quantiles was analyzed, unfortunately this does not seems to be possible even for a moderate number of quantiles since the matrix structure of the problem makes the problem extremely computational expensive.

Appendix A

Proofs

A.1 Proof of Theorem 3.1

The inspiration for the substitutions in the following can be found in theorem 2.1 of [8]. First we define the basis functions $M_{j,k} = \frac{B_{j,k}}{t_{j+k}-t_j}$, and then calculate $(t_{j+k} - x)M_{j+1,k-1}$, with $\mathcal{J}_{l,j} = j, j+1, ..., l-1, l+1, ..., k-1$ and using (3.6) we get

$$\begin{aligned}
(t_{j+k} - x)M_{j+1,k-1}(x) &= (t_{j+k} - x)[t_{j+1}, ..., t_{j+k}](\cdot - x)_+^{k-2} \\
&= (t_{j+k} - x) \sum_{l=j+1}^{j+k} \frac{(t_l - x)_+^{k-2}}{\prod_{m\in\mathcal{J}_{l,j+1}}(t_l - t_m)} \\
&= \sum_{l=j+1}^{j+k} \frac{(t_l - x)_+^{k-2}(t_{j+k} + t_l - t_l - x)}{\prod_{m\in\mathcal{J}_{l,j+1}}(t_l - t_m)} \\
&= \sum_{l=j+1}^{j+k} \frac{(t_l - x)_+^{k-1} + (t_l - x)_+^{k-2}(t_{j+k} - t_l)}{\prod_{m\in\mathcal{J}_{l,j+1}}(t_l - t_m)} \\
&= [t_{j+1}, ..., t_{j+k}](\cdot - x)_+^{k-1} \\
&\quad + \sum_{l=j+1}^{j+k} \frac{(t_l - x)_+^{k-2}(t_{j+k} - t_l)}{\prod_{m\in\mathcal{J}_{l,j+1}}(t_l - t_m)} \qquad \text{(A.1)}
\end{aligned}$$

defining $\mathcal{J}^*_{l,j} = j+1, ..., l-1, l+1, ..., j+k-1$, the second term can be rewritten as

$$\begin{aligned}
\sum_{l=j+1}^{j+k} \frac{(t_l - x)_+^{k-2}(t_{j+k} - t_l)}{\prod_{m \in \mathcal{J}_{l,j+1}}(t_l - t_m)} &= \sum_{l=j+1}^{j+k-1} \frac{(t_l - x)_+^{k-2}(t_{j+k} - t_l)}{\prod_{m \in \mathcal{J}_{l,j+1}}(t_l - t_m)} \\
&= \sum_{l=j+1}^{j+k} \frac{(t_l - x)_+^{k-2}(t_{j+k} - t_l)}{(t_l - t_{j+1}) \cdots (t_l - t_{l-1})(t_l - t_{l+1}) \cdots (t_l - t_{j+k})} \\
&= -\sum_{l=i+1}^{j+k} \frac{(t_l - x)_+^{k-2}}{(t_l - t_{j+1}) \cdots (t_l - t_{l-1})(t_l - t_{l+1}) \cdots (t_l - t_{j+k-1})} \\
&= -\sum_{l=j+1}^{j+k-1} \frac{(t_l - x)_+^{k-2}}{\prod_{m \in \mathcal{J^*}_{l,j}}(t_l - t_m)}
\end{aligned}$$

in the exact same way we find

$$\begin{aligned}
(t_j - x)M_{j,k-1}(x) &= (t_j - x)[t_j, ..., t_{j+k-1}](\cdot - x)_+^{k-2} \\
&= (t_j - x) \sum_{l=j}^{j+k-1} \frac{(t_l - x)_+^{k-2}}{\prod_{m \in \mathcal{J}_{l,j}}(t_l - t_m)} \\
&= \sum_{l=j}^{j+k-1} \frac{(t_l - x)_+^{k-2}(t_j + t_l - t_l - x)}{\prod_{m \in \mathcal{J}_{l,j}}(t_l - t_m)} \\
&= \sum_{l=j}^{j+k-1} \frac{(t_l - x)_+^{k-1} + (t_l - x)_+^{k-2}(t_i - t_l)}{\prod_{m \in \mathcal{J}_{l,j}}(t_l - t_m)} \\
&= [t_j, ..., t_{j+k-1}](\cdot - x)_+^{k-1} \\
&\quad + \sum_{l=j}^{j+k-1} \frac{(t_l - x)_+^{k-2}(t_j - t_l)}{\prod_{m \in \mathcal{J}_{l,j}}(t_l - t_m)} \\
&= [t_{j+1}, ..., t_{j+k}](\cdot - x)_+^{k-1} - \\
&\quad \sum_{l=j+1}^{j+k-1} \frac{(t_l - x)_+^{k-2}}{\prod_{m \in \mathcal{J^*}_{l,j}}(t_l - t_m)} \qquad \text{(A.2)}
\end{aligned}$$

now combining (A.1) and (A.2) and using (3.15) we can write

$$\begin{aligned}
M_{j,k} &= \frac{B_{j,k}}{t_{j+k} - t_j} \\
&= [t_{j+1}, ..., t_{i+k}](\cdot - x)_+^{k-1} - [t_j, ..., t_{j+k-1}](\cdot - x)_+^{k-1} \\
&= \frac{(t_{j+k} - x)M_{j+1,k-1} - (t_j - x)M_{j,k-1}}{t_{j+k} - t_j} \qquad \text{(A.3)}
\end{aligned}$$

now we just need $M_{j,1}$, but this is very easy to calculate as

$$\begin{aligned} M_{j,1}(x) &= \frac{B_{j,1}}{t_{j+1}-t_j} = \frac{[t_{j+1}](\cdot - x)_+^0 - [t_j](\cdot - x)_+^0}{t_{j+1}-t_j} \\ &= \frac{I_{[t_j \leq x < t_{j+1}]}(x)}{t_{j+1}-t_j} \end{aligned} \tag{A.4}$$

(A.3) and (A.4) together give the recursive formula in Theorem 3.1 and the theorem is thereby proved.

A.2 Proof of equation (3.19)

$$\begin{aligned} [t_j, ..., t_{j+3}](\cdot - t_l)_+ &= \frac{1}{t_{j+3}-t_l}\left([t_{j+1}, t_{j+2}, t_{j+3}](\cdot - t_l)_+ - [t_j, t_{j+1}, t_{j+2}](\cdot - t_l)_+\right) \\ &= \frac{1}{t_{j+3}-t_j}\left[\frac{[t_{j+2}, t_{j+3}](\cdot - t_l)_+ - [t_{j+1}, t_{j+2}](\cdot - t_l)_+}{t_{j+3}-t_{j+1}}\right. \\ &\quad \left. - \frac{[t_{j+2}, t_{j+1}](\cdot - t_l)_+ - [t_j, t_{j+1}](\cdot - t_l)_+}{t_{l+2}-t_j}\right] \\ &= \frac{1}{t_{j+3}-t_j}\left[\frac{1}{t_{j+3}-t_{j+1}}\left(\frac{(t_{j+3}-t_l)_+ - (t_{j+2}-t_l)_+}{t_{j+3}-t_{j+2}}\right.\right. \\ &\quad \left. - \frac{(t_{j+2}-t_l)_+ - (t_{j+1}-t_l)_+}{t_{j+2}-t_{j+1}}\right) \\ &\quad - \frac{1}{t_{j+2}-t_j}\left(\frac{(t_{j+2}-t_l)_+ - (t_{j+1}-t_l)_+}{t_{j+2}-t_{j+1}}\right. \\ &\quad \left. - \frac{(t_{j+1}-t_l)_+ - (t_j-t_l)_+}{t_{j+1}-t_j}\right) \\ &= \frac{1}{t_{j+3}-t_j}\left(\frac{I_{[x \leq j+2]}(l) - I_{[x \leq j+1]}(l)}{t_{j+3}-t_{j+1}}\right. \\ &\quad \left. - \frac{I_{[x \leq j+1]}(l) - I_{[x \leq j]}(l)}{t_{j+2}-t_j}\right) \\ &= \frac{1}{t_{j+3}-t_j}\left(\frac{\delta(l-j-2)}{t_{j+3}-t_{j+1}} - \frac{\delta(l-j-1)}{t_{j+2}-t_j}\right) \end{aligned}$$

Appendix B

Analasis of Data From Tunø

B.1 Histograms of DMI-HIRLAM Data

Figure B.1-B.4 give histogram plots of available explanatory variables. It is seen that wind speed and forecasted power even though very correlated, diplay a very different density function.

The plots of winddirection display very similar behaivor. For turbolunt kinetic energi only the log transformed is displayed, and we see similar behavior in all levels. The risk indices show quite different behavior, and all of them have areas with very few obeservations.

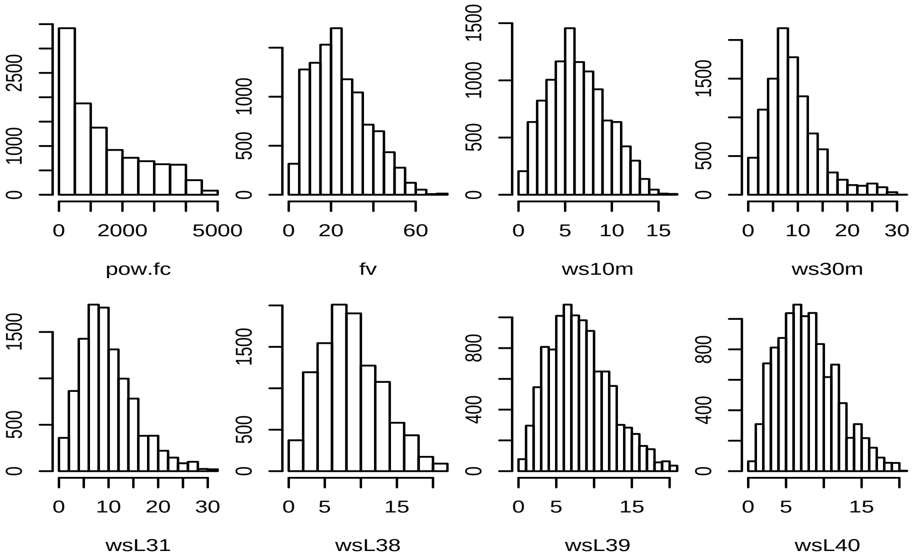

Figure B.1: Histograms of windsspeeds predicted from DMI Hirlam and the power curve predicted by WPPT.

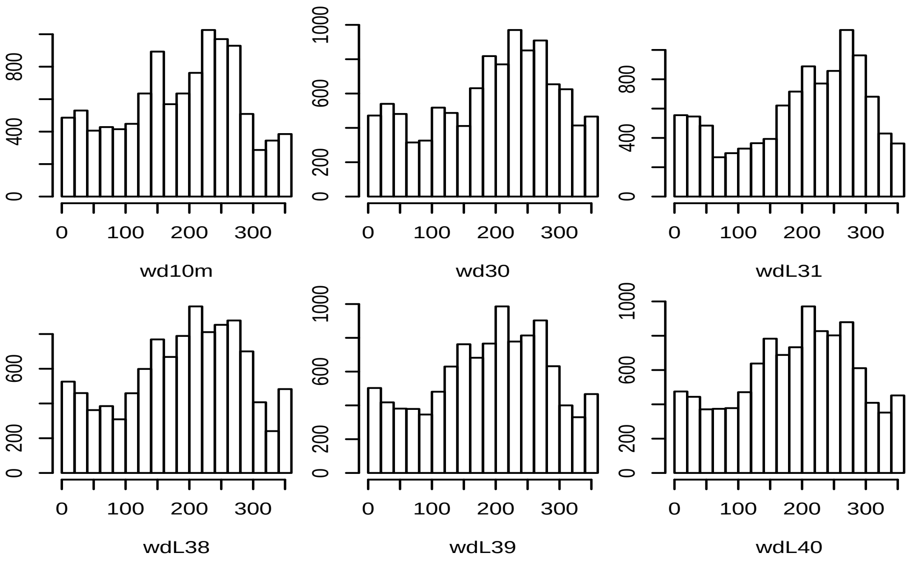

Figure B.2: Histograms of winddirections predicted from DMI Hirlam.

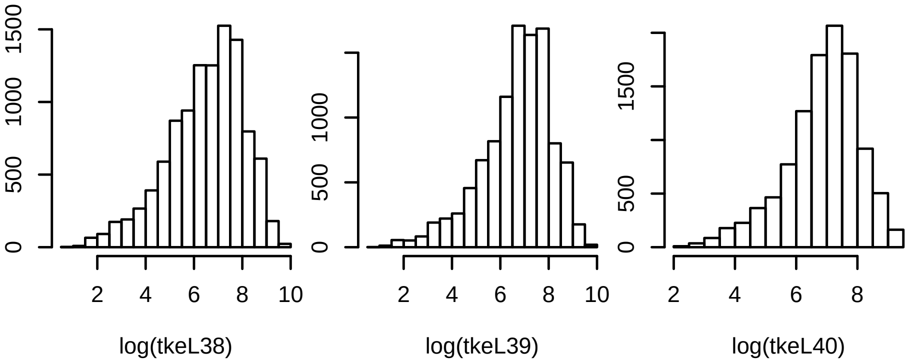

Figure B.3: Histograms of turbulent kinetic energy predicted from DMI Hirlam.

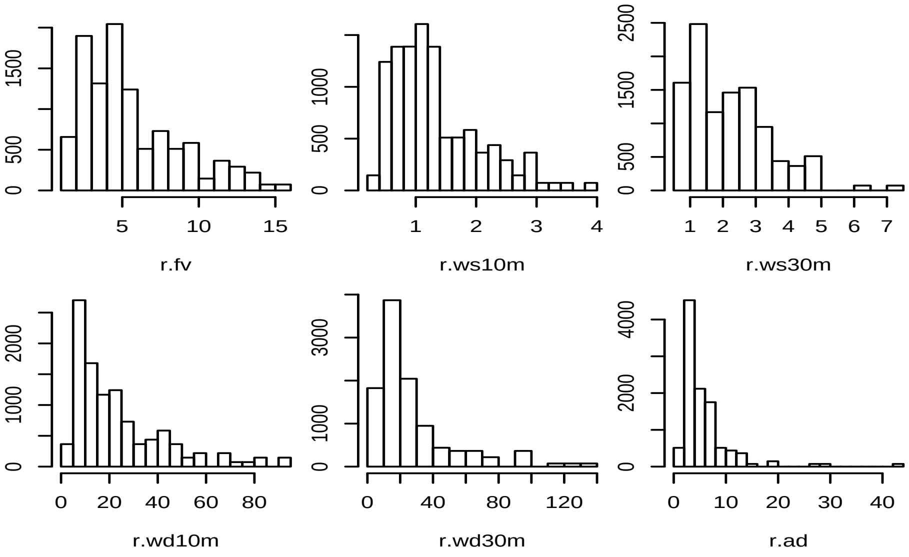

Figure B.4: Histograms of risk indicies calculated from the predictions from DMI Hirlam.

Key numbers for the models of the 25% and 75% quantile

Model	pow.fc	Ws10m	Ws30	Fv	fv.avg	WsL31	WsL38	WsL39	WsL40
Df	6	6	6	6	6	6	6	6	6
below 25%	23.9%	29.9%	33.2%	28.7%	31.1%	32.5%	29.5%	29.9%	29.8%
below 75%	83.6%	83.0%	82.7%	83.7%	84.9%	81.8%	81.7%	82.1%	82.5%
Loss25	215.8	232.3	241.1	234.4	251.0	250.1	244.0	244.0	244,3
Loss75	266.4	269.1	269.4	270.5	275.7	263.0	254.3	252.5	253.0
cross	95	40	0	46	41	1	12	0	0

Table B.1: The table contains some key numbers for models only contaning wind speed of some sort, the knots are in all cases placed as 20% sample quantiles of the traning data.

Key numbers for the models of the 25% and 75% quantiles

Model	Wd10m	Wd30	WdL31	WdL38	WdL39	WdL40
DF	11	11	11	11	11	11
below 25%	33.3%	33.0%	31.8%	33.2%	33.0%	33.0%
below 75%	83.1%	83.8%	83,3%	82.2%	82.0%	81.8%
Loss25	248.5	247.3	246.7	249.8	249.4	249.3
Loss75	261.5	262.3	259.5	258.9	258.9	258.9
cross	22	0	0	0	0	0

Table B.2: The table contains some key numbers for models contaning WSL40 and different wind direction, the knots are in all cases placed as 20% sample quantiles of the traning data.

B.2 Model Construction - An Example

Table B.1 to B.4 give an example of an model building, where one new variable is added at the time. In this analysis overall reliability, the loss function and crossings are considered the key performance parameters. Even with this few parameters it is not clear which model to choose in each step.

It is seen from the figure tables performance does not really improve dramaticly as we add more variables, and at times it even gets worse. The example indicate that the availiable data may be insufficient for the models. I.e. we would need more data to give a conviencing model.

It should be noted that procedures like this can very well lead to wrong conclussions.

Key numbers for the models of the 25% and 75% quantiles

Model	log(tkeL40)	log(tkeL39)	log(tkeL38)
DF	16	16	16
Below 25%	26.9%	28.4%	29.3%
Below 75%	78.6%	79.6%	81.5%
Loss25	249.8	249.2	249.7
Loss75	256.1	257.9	258.6
cross	0	9	8

Table B.3: The table contains some key numbers for models contaning WSL40, wdL40 and turbulent kinetic energy in different lags, the knots are in all cases placed as 20% sample quantiles of the traning data.

Key numbers for the models of the 25% and 75% quantiles

Model	r.ad	r.fv	r.wd10m	r.wd30	r.ws10m	r.ws30
DF	17	17	17	17	17	17
%below	26.8%	26.0%	27.6%	27.0%	26.2%	26.8%
%below	78.2%	78.4%	79.2%	78.6%	78.5%	78.1%
Loss25	249.9	245.6	249.6	249.8	246.6	247.9
Loss75	256.7	257.0	254.3	255.1	256.6	255.8
cross	0	0	0	0	0	0

Table B.4: The table contains some key numbers for models contaning WSL40, wdL40, log(tkeL40) and the risk indices added one by one the knots are in all cases placed as 20% sample quantiles of the traning data

Appendix C

Dual Simplex For Non Crossing Constraints

We already saw in Chapter 2 that the strictly complementary property means that we have a splitting of the index set s.t.

$$\mathcal{B} = \{j|x_j^* > 0\} \quad \mathcal{C} = \{j|s_j^* > 0\} \tag{C.1}$$

and $s_j \odot x_j = 0$, with this we can write down the solution to the dual problem as $\mathbf{z}^* = \mathbf{B}^{-T}\mathbf{c}_{\mathcal{B}}$, with this we can write down the solution to the dual problem as

$$\begin{aligned} \mathbf{z}(\bar{h}) &= \mathbf{P}\mathbf{c}_{\mathcal{B}}(\bar{h}) && \text{(C.2)} \\ \mathbf{z}(h) &= -\mathbf{P}\mathbf{X}(h)^{-T}\mathbf{X}^T(\bar{h})\mathbf{c}_{\mathcal{B}}(\bar{h}) && \text{(C.3)} \end{aligned}$$

so from the solution to the original problem we can write down the solution to the dual problem directly. When we add the non crossing constraint, there are two possibilities, either the problem is still feasible, and then we are done. The other possibility is that the primal problem becomes infeasible, but then the dual problem is feasible and we just have to find the solution to this to find the solution to the original problem.

The non crossing constraints can be introduced directly into the $\mathbf{B}$ matrix, with the solution to the original problem for τ_2 just expand $\mathbf{B}$ and $\mathbf{y}$ in the following way

$$\tilde{\mathbf{B}} = \begin{bmatrix} \mathbf{X}(h) & \mathbf{0} & \mathbf{0} \\ \mathbf{X}(\bar{h}) & \mathbf{P} & \mathbf{0} \\ \mathbf{X}_{nc} & \mathbf{0} & sign(\tau_2 - \tau_1)\mathbf{I} \end{bmatrix} \quad \tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y}(h) \\ \mathbf{y}(\bar{h}) \\ \mathbf{X}_{nc}\beta(\tau_1) \end{bmatrix} \tag{C.4}$$

now our $\mathbf{x}_{\mathcal{B}}$ is expanded to

$$\tilde{\mathbf{x}} = \begin{bmatrix} \beta(\tau_2) \\ |\mathbf{r}(\bar{h})| \\ \mathbf{X}_{nc}(\beta(\tau_1) - \beta(\tau_2) \end{bmatrix} \tag{C.5}$$

if $\tilde{\mathbf{x}} \geq \mathbf{0}$ then the solution is optimal and we are done, if there exist an element in $\tilde{\mathbf{x}}$ that is less than zero, then the point we are in is infeasible, but this corresponding point in the dual problem is feasible. So we just have to set up the dual problem and find the optimal solution for this then we have the optimal solution for the primal problem.

The cost vector is expanded with a vector of zeros with the same number of elements as $\mathbf{X}_{nc}$ and we therefore have

$$\tilde{\mathbf{B}}^T\tilde{\mathbf{z}} = \begin{bmatrix} \mathbf{X}(h)^T & \mathbf{X}(\bar{h})^T & \mathbf{X}_{nc}^T \\ \mathbf{0} & \mathbf{P} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & sign(\tau_2 - \tau_1)\mathbf{I} \end{bmatrix} \tilde{\mathbf{z}} = \begin{bmatrix} \mathbf{c}_{\mathcal{B}} \\ \mathbf{0} \end{bmatrix} \tag{C.6}$$

from this we can immediately get a starting point for $\tilde{\mathbf{z}}$ as $\tilde{\mathbf{z}} = [\mathbf{z} \quad \mathbf{0}]^T$. Now we have a starting point for the dual simplex algorithm, which will be expained now.

If we do not have the optimal solution then there is at least one element in $\tilde{\mathbf{x}}_{\mathcal{B}}$ which is less than zero, choose such an element $(\tilde{\mathbf{x}}_{\mathcal{B}})_q$ and change the dual solution in this direction, i.e. change $\tilde{\mathbf{z}}^{k+1} = \tilde{\mathbf{z}}^k - \nu\tilde{\mathbf{B}}^{-T}\mathbf{e}_q$, if $\nu > 0$ this will increase the dual objective. The amount we can change the objective is determined by the surplus vector, since we must have $\mathbf{s} \geq \mathbf{0}$. Since we have $\tilde{\mathbf{s}}_{\mathcal{B}} = \tilde{\mathbf{c}}_{\mathcal{B}} - \tilde{\mathbf{B}}^T\tilde{\mathbf{z}}$ and $\tilde{\mathbf{s}}_{\mathcal{C}} = \tilde{\mathbf{c}}_{\mathcal{C}} - \tilde{\mathbf{C}}^T\tilde{\mathbf{z}}$ then the amount we can change $\mathbf{s}$ is determined by $\alpha = \min_j\{\sigma_j\}$ with

$$\sigma_j = \begin{cases} -(\mathbf{s}_{\mathcal{C}}^{(k)})/h_j & \text{if} \quad h_j < 0 \\ +\infty & \text{otherwise} \end{cases} \tag{C.7}$$

with $\mathbf{h} = \tilde{\mathbf{C}}^T\mathbf{B}^{-T}\mathbf{e}_q$, the updating of $\mathbf{B}$ and $\mathbf{C}$ is now as described in section 2.3.2. When we have the optimal solution to this problem then we have the optimal solution to the primal problem.

Bibliography

[1] Henrik Aalborg Nielsen, Henrik Madsen, Torben Skov Nielsen *Using Quantile Regression to an Existing Wind Power Forecasting System with Probabilistic Forecats*.

[2] Roger Koenker, Gilbert Bassett Jr *Regression Quantile* Econometrica, Vol. 46, No 1, Jan 1978 (33-50).

[3] Roger Koenker*Quantile Regression* Cambridge University Press 2005.

[4] Faouzi El Bantli and Marc Hallin *L_1-Estimation in Linear Models with Heterogeneous White Noise* Statistics and Probanility Letter Vol 45, 1999 (305-315).

[5] Roger Koenker, Pin NG *Inequality Constrained Quantile Regression*

[6] Carl de Boor *A Practical Guide to Splines* Springer-Verlag 1978

[7] Bent Jørgensen *The Theory of Linear Models* Chapman and hall 1993

[8] Hans Bruun Nielsen *Cubic Splines* IMM Department of Mathematical Modeling, DTU

[9] T. Hastie and R. Tibshirani *Genralized Additive Models* Chapman and Hall, 1990.

[10] Bernard W. Lindgren *Statistical Theory 4th Edition* Chapman and Hall

[11] Roger Koenker, Jose A. F. Machado *Goodness of Fit and Related Inference Processes for Quantile Regression* Journal of the Statistical Association, Vol 94, No 448 (dec. 1999), p. 1296-1310

[12] P. Pinson, G. Karinnotakis *On-line Assessment of Prediction Risk for Wind Power Production Forecast*

[13] Roger Koenker, Vasco d'Orey *Algorithm AS 229: Computing Regression Quantiles* Applied Statistics, Vol. 36, No 3 (1987), 383-393.

[14] Hans Bruun Nielsen *Algorithms for Linear Optimization, an Introduction* (1999) Course note for the DTU course *Optimimization and Data fitting 2*

[15] Henrik Aalborg Nielsen, Devon Yates, Henrik Madsen, Torben Skov Nielsen, Jake Badger, Gregor Giebel, Lars Landberg, Kai Satler, Henrik Feddersen *Ensemble-forecast for Wind Power, Analysis of the Results of an On-line Wind Power Quantile Forecasting System* October 2005

[16] P. Pinson,G. Kariniotakis, H. Aa. Nielsen, T.S. Nielsen, H. Madsen *Properties of Quantiles and Their Forecasts of Wind Generation and their Evaluation* 2005

[17] Tilmann Gneiting and Adrian E. Raftery *Strictly Proper Scoring Rules, Prediction, and Estimation* Technical Reprot no. 463, Department of Statistics, University of Washington, September 2005

[18] Roger Koenker, Breum J. Parks *An Interor Point Algorithm for Nonlinear Regression* Journal of Econometrics 71 (1996)

[19] Henrik Madsen, Henrik Aalborg Nielsen, Torben Skov Nielsen *A Toll for Predicting the Wind Power Production of Off-Shore Wind Plants*

[20] Quanshui Zhao *Restricted Regression Quantiles* Journal of Multivariate Analysis 72 (1998)

[21] Ichiro Takeuchi, Takeshi Furuhashi *Non-crossing Quantile Regression by SVM* (2004)